Probability models are mathematical representations used to study and understand uncertain phenomena.

Random Experiment

A **random experiment** is a process that leads to at least two possible outcomes with uncertainty as to which will occur.

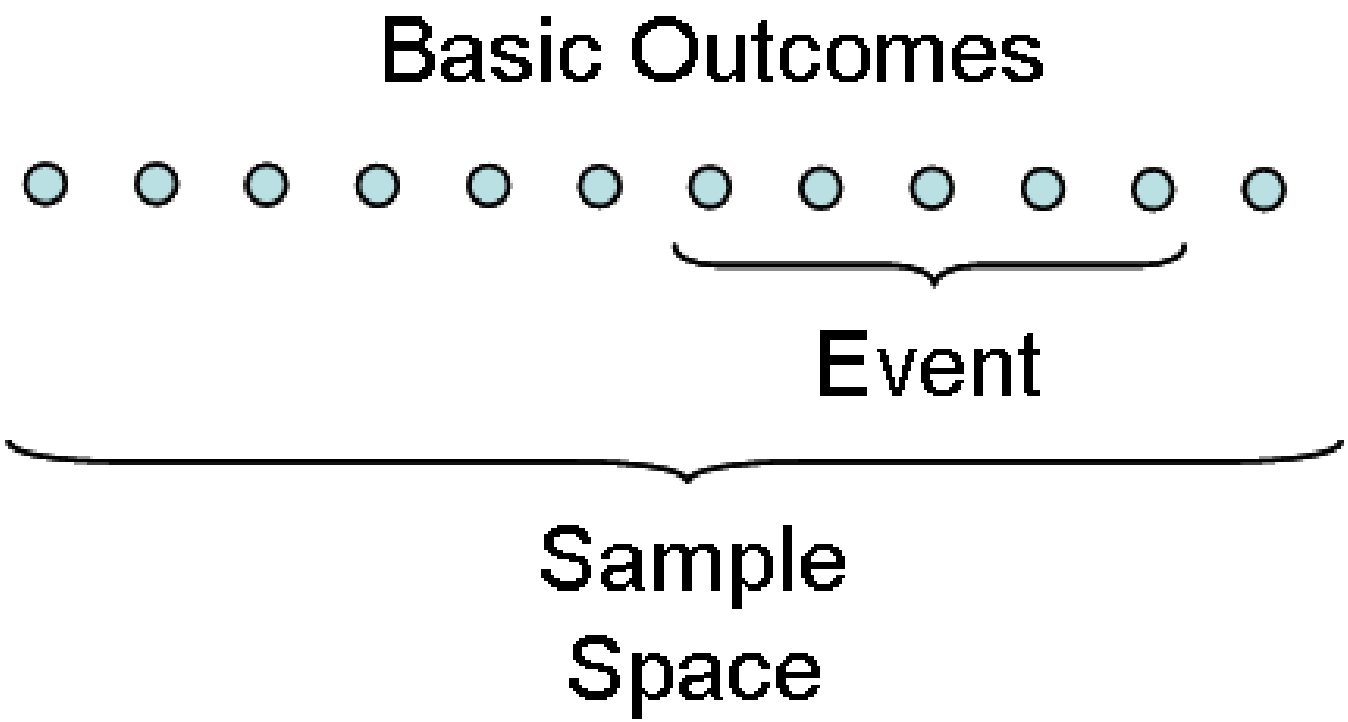
The possible outcomes are the **basic outcomes** of the experiment.

The **sample space** is the set of all basic outcomes

Examples:
Toss a coin twice. Sample space: {HH, HT, TH, TT}.
Roll a dice. Sample space: {1,2,3,4,5,6}

An **event** is a subset of the sample space. It is a list of basic outcomes.

Examples:
"At least one head": {HH, HT, TH}
"No more than one head": {TT, HT, TH}
When rolling dice, get an even number: {2,4,6}


Basic Outcomes
Event
Sample
Space

Let A and B be events in the sample space S.

- $A \cap B$ means that both A and B occur
- $A \cup B$ means that either A or B occurs

Example: Rolling a die.

- Event A: outcome is even ($\{2,4,6\}$).
- Event B: outcome is greater than 3 ($\{4,5,6\}$).

$A \cap B = \{4,6\}$

$A \cup B = \{2,4,5,6\}$

If A and B are two events in the sample space such that $A \cap B$ is the empty set (i.e. they share no common outcome), then they are said to be **mutually exclusive.**

Given k events $E_1, E_2, ...E_k$ in the sample space S, if $E_1 \cup E_2 \cup ...E_k = S$ then these k events are **collectively exhaustive**.

The set of elements belonging to S but not to A is the **complement** of A. It is denoted by $\overline{A}$.

The **probability** of an event is the relative frequency of that outcome as the experiment is repeated an arbitrarily large number of times

$$P(A) = n_A / n$$

This interpretation relies on the possibility of repeating an experiment a large number of times.

In the case of a nonrepeatable event, probability is a subjective assessment of the likelihood of that event occurring

Axioms of probability ("probability postulates")
Suppose that
A is an event
O_i is a basic outcome in the random experiment

Then

- $0 \leq P(A) \leq 1$
- $P(A) = \Sigma_A P(O_i)$
- $P(S) = 1$

These axioms deliver the following consequences:

I. If there are n equally likely outcomes in the sample space, then each of them has probability 1/n.
(Q: Why? A: Because otherwise they would not sum up to 1).

II If there are n equally likely basic outcomes in the sample space and event A consists of n_A of these outcomes, then $P(A) = n_A/n$
(Q. Why? A. Previous consequence and postulate two.)

Coin toss example. Assume that a fair coin is tossed twice. Here is the probability of each event

Event A	P(A)
{HH}	¼
{HT}	¼
{TH}	¼
{TT}	¼
S	1
{HH,HT,TH}	¾
{HH,HT}	½

What are the probabilities of the following two events?

- To get at least one six when tossing a dice six times
- To get at least two sixes when tossing a dice 12 times

Sample space for tossing a dice six times

{1,1,1,1,1,1}

{1,1,1,1,1,2}

and so on

Contains 6^6 elements

Of these 5^6 have no six

All are equally likely

So the probability of no six is 0.335

The probability of at least one six is 0.665

Sample space for tossing a dice 12 times

Contains 6^{12} elements

Of these 5^{12} have no six

$12 * 5^{11}$ have one six

All are equally likely

So the probability of 0 or 1 six is $(5^{12} + 12 * 5^{11}) / 6^{12} = 0.381$

The probability of at least 2 sixes is 0.619

Rules of Probability

1. Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$\Rightarrow$Bonferroni Bound: $P(A \cap B) \geq P(A) + P(B) - 1$

2. Complement Rule $P(A) + P(\bar{A}) = 1$

3. If $A \subseteq B$ then $P(A) \leq P(B)$

   Equivalently, if $A \Rightarrow B$ then $P(A) \leq P(B)$

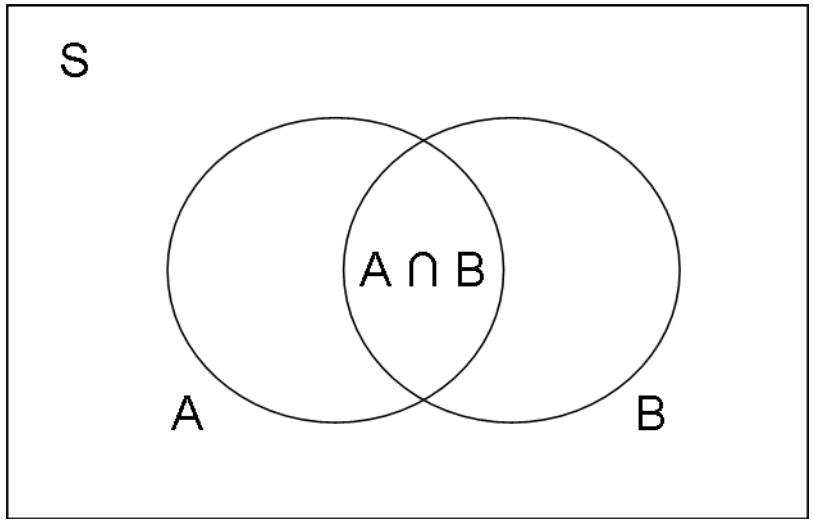
4. If $\{B_1, B_2, ..\}$ is set of mutually exclusive and collectively exhaustive events then $P(A) = \Sigma_i P(A \cap B_i)$

   $\Rightarrow P(A) = P(A \cap C) + P(A \cap \bar{C})$

5. If $\{A_1, A_2, ..\}$ is a set of events, then $P(\bigcup_i A_i) \leq \Sigma_i P(A_i)$

Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Intuition



Addition Rule Example. A restaurant found that 75% of its customers use mustard, 80% use ketchup and 65% use both.

What is the probability that a customer will use at least one of these?

Let A be the event "Customers use mustard"
Let B be the event "Customers use ketchup".
Then
P(A) = 0.75, P(B) = 0.80 and $P(A \cap B) = 0.65$
$\therefore P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.75 + 0.8 - 0.65 = 0.9$

Complement Rule: $P(A) + P(\overline{A}) = 1$

Proof: A and $\overline{A}$ are mutually exclusive and collectively exhaustive

$P(A \cup \overline{A}) = P(A) + P(\overline{A}) - P(A \cap \overline{A})$

$\therefore P(A \cup \overline{A}) = P(A) + P(\overline{A})$

$P(A \cup \overline{A}) = 1$

So $P(A) + P(\overline{A}) = 1$

Terminology

Probability of two (or more) events occurring is a **joint** probability
$P(A \cap B)$ is a joint probability

Probability of one event occurring is a **marginal** probability
$P(A)$ is a marginal probability.

Let $A_1,...A_h$ be a set of events that are mutually exclusive and collectively exhaustive.

Let $B_1,...B_k$ be another such set of events

A way to represent the relation between these events:

	B_1	B_2	..	B_k
A_1	$A_1 \cap B_1$	$A_1 \cap B_2$		$A_1 \cap B_k$
A_2	$A_2 \cap B_1$	$A_2 \cap B_2$		$A_2 \cap B_k$
A_h	$A_h \cap B_1$	$A_h \cap B_2$		$A_h \cap B_k$

We can assign probabilities to these cells.

Example: A study with 150 heart attack patients was performed on four alternative treatments: A, B, C and D.

	Treatment Group				
	A	B	C	D	Total
Died	18	13	22	24	77
Survived	22	25	16	10	73
Total	40	38	38	34	150

Probability of randomly chosen patient being in each cell

	Treatment Group				
	A	B	C	D	Total
Died	0.120	0.087	0.147	0.160	0.513
Survived	0.147	0.167	0.107	0.067	0.487
Total	0.267	0.253	0.253	0.227	1.000

Elements in red are joint probabilities

Elements in blue are marginal probabilities

Q. How do we go from joint probabilities to marginals?

A. Add them up.

$$P(A_i) = \Sigma_j P(A_i \cap B_j)$$

Permutations

Five letters A, B, C, D, E

Pick 2 different elements and arrange in order

AB	BA	CA	DA	EA
AC	BC	CB	DB	EB
AD	BD	CD	DC	EC
AE	BE	CE	DE	ED

Twenty ways of doing this

In general, if I have n objects and I want to pick k out of them and arrange them in order, then the number of ways of doing so is

$$P_k^n = \frac{n!}{(n-k)!} = n(n-1)...(n-k+1)$$

$$P_2^5 = 5*4 = 20$$

If I don't care about the order, the number of ways is

$$C_k^n = \frac{n!}{(n-k)!k!}$$

This allows us to do probability calculations

Example: What is the probability the no two individuals in a group of k people have birthdays on the same date?

Assume births are equally distributed throughout the year
Q. What is sample space?
A. Since there are 365 possibilities of birthday for each of the k people there are 365^k possibilities. Each is equally likely.

Q. In how many of these do no birthdays coincide?
A. P_k^{365}

Therefore the desired probability is $P_k^{365} / 365^k$

Here are the numbers

k	P
5	0.973
10	0.883
15	0.747
20	0.580
22	0.524
23	0.493
25	0.431
30	0.294
40	0.109
50	0.030
60	0.006

Conditional Probability

Suppose A and B are two events. Typically we will be interested in the probability of event A once we know that B has occurred. For this we define conditional probabilities.

Conditional Probability

Let A and B be two events. Then the **conditional probability** of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example. The probability that it rains and I have an umbrella is 0.4. The probability that it rains is 0.8. Therefore

$$P(Umbrella \mid Rains) = \frac{P(Rains \cap Umbrella)}{P(Rains)} = \frac{0.4}{0.8} = 0.5$$

Example: A study with 150 heart attack patients was performed on four alternative treatments: A, B, C and D.

	Treatment Group				
	A	B	C	D	Total
Died	18	13	22	24	77
Survived	22	25	16	10	73
Total	40	38	38	34	150

If we were to randomly sample patients in this study,
What is the probability that a patient dies when
he or she receives treatment D?

P(“D”) = 34/150

P(“Dies and D”) = 24/150

P(“Dies”|“D”) = [24/150]/[34/150] = 0.706

If A and B are mutually exclusive then $P(A \mid B) = 0$

If $A \subseteq B$ then $P(A \mid B) \leq 1$

If $B \subseteq A$ then $P(A \mid B) = 1$

When you condition on B, it is as though B becomes the new sample space.

Two events A and B are statistically independent if
$P(A \cap B) = P(A)P(B)$

If A and B are independent then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Similarly $P(B \mid A) = P(B)$

Intuition: the occurrence of B carries no information about A and vice-versa.

Example: Toss a fair coin twice. Are the events
A: “get a head on the first toss” and
B: “get a head on the second toss”
statistically independent?

$P(A) = 0.5$
$P(B) = 0.5$
$P(A \cap B) = 0.25$

$P(A \cap B) = P(A) * P(B)$ so A and B are indeed statistically independent.

Bayes rule

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

$$\therefore P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$\therefore P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$ Bayes Rule

Often in applying Bayes rule, we aren't given $P(B)$. But

$$P(B) = P(B \cap A) + P(B \cap \overline{A}) = P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})$$

So we can compute the denominator in Bayes rule

Or if $A_1, A_2, \ldots A_n$ is a mutually exclusive and collectively exhaustive set of events

$$P(B) = P(B \cap A_1) + P(B \cap A_2) \ldots + P(B \cap A_n)$$

$$\therefore P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) \ldots + P(B \mid A_n)P(A_n)$$

Bayes Rule: Example 1

A car dealership knows that 10% of people who come into the showroom purchase a car. The dealership offers a promotion where free lunch is offered to anyone who listens to a sales pitch. 40% of people who purchased cars had a free lunch. 10% of the people who did not purchase cars had a free lunch.

How likely are people who accept the lunch to buy a car?

$$P(Buy \mid Lunch) = \frac{P(Lunch \mid Buy)P(Buy)}{P(Lunch \mid Buy)P(Buy) + P(Lunch \mid Not)P(Not)}$$

$$= \frac{0.40 * 0.10}{(0.40 * 0.10) + (0.10 * 0.90)} = 0.308$$

Bayes Rule: Example 2

Suppose one has a test for a disease

P(Disease)=0.005 --- prevalence of the disease in the population

P(Negative|Disease)=0.01 (prob. of false negative)

P(Positive|No Disease)=0.05 (prob. of false positive)

Questions.

1. Is this a good test?
2. If you test positive, how worried should you be?

We are interested in

$$P(\text{Disease}|\text{Positive}) = \frac{P(\text{Positive}|\text{Disease})P(\text{Disease})}{P(\text{Positive})}$$

$$P(\text{Disease}|\text{Positive}) = \frac{0.99*0.005}{P(\text{Positive})}$$

$$P(\text{Positive}) = P(\text{Positive}|\text{Disease})P(\text{Disease}) + P(\text{Positive}|\text{No Disease})P(\text{No Disease})$$

$$\therefore P(\text{Positive}) = 0.05*0.995 + 0.99*0.005 = 0.0547$$

$$\therefore P(\text{Disease}|\text{Positive}) = \frac{0.99*0.005}{0.0547} = 0.0905$$

Bayes Rule: Example 3

Suppose you are in a game show. *Y*ou are given the choice of three doors – one of which conceals a valuable prize and the others conceal a goat.

After you make a choice, the host opens one of the **other** doors (–one **without** a prize).

He then gives you the option of staying with the initial choice of door or switching to the other door.

The door finally chosen is then opened.

Should you switch, not switch, or does it make no difference what the contestant does?

Call:
A: The door you initially choose;
B: The door the host, opens;
C: The remaining door.
We want to compare:

P(“Prize is behind A”|“Host opens B”)
and
P(“Prize is behind C”|“Host opens B”).

Bayes' Rule: P(Prize is behind A|Host opens B) is

$$\frac{P(\text{Host opens B}|\text{Prize behind A}) \times P(\text{Prize behind A})}{P(\text{Host opens B})}$$

Now, notice that:

P($ is behind A) = P($ is behind B) = P($ is behind C) =1/3

P(Opens B|Prize behind A) = ½

P(Opens B|Prize behind is B) = 0

P(Opens B|Prize behind is C) = 1

P(Prize behind A|Host opens B) is $\frac{(1/2)*(1/3)}{P(\text{Host Opens B})}$

P(Prize behind C|Host opens B") is $\frac{1*(1/3)}{P(\text{Host Opens B})}$

P(Host Opens B)=
P(Host Opens B|Prize Behind A)*P(Prize Behind A)
P(Host Opens B|Prize Behind B)*P(Prize Behind B)
P(Host Opens B|Prize Behind C)*P(Prize Behind C)

=(1/2)*(1/3) + 0*(1/3) + 1*(1/3)=1/2

P("Prize behind A"|"Host opens B") is $\frac{(1/2)*(1/3)}{1/2} = \frac{1}{3}$

P("Prize behind C"|"Host opens B") is $\frac{1*(1/3)}{1/2} = \frac{2}{3}$

The contestant should switch!

A **random variable** is a variable that takes on one of a number of different numerical values with uncertainty as to which one occurs

More formally, a random variable is a function that maps the sample space to R

e.g. Number that I get when I roll a die is 1, 2, 3, 4, 5 or 6. This is a random variable.

Notation: X is the random variable and x denotes the values that it can take on

A random variable is **continuous** if it can take on any value in an interval

Otherwise it is **discrete**

Examples of continuous random variables
The maximum temperature on any day
The time between the arrival of two buses

Examples of discrete random variables
The number of heads when I toss a coin 10 times
The number of claims on an insurance policy in a given year.

Probability Distribution (Mass) Function for a discrete random variable gives the probability that the random variable takes on each possible value, P(X=x) for each x

We write it as P(x)

For example, if a die is thrown, the probability distribution function is

P(1)=1/6
P(2)=1/6
P(3)=1/6
P(4)=1/6
P(5)=1/6
P(6)=1/6

Properties of a probability mass function

1. $P(x) \geq 0$ for all x

2. $\Sigma_x P(x) = 1$

The cumulative probability function $F(x_0)$ gives the probability that a discrete random variable X does not exceed x_0

$$F(x_0) = P(X \leq x_0) = \Sigma_{x \leq x_0} P(x)$$

Also

$$P(a \leq X \leq b) = \Sigma_{a \leq x \leq b} P(x)$$

Properties of cumulative probability function

1. $0 \leq F(x) \leq 1$ for every x

2. If $x_0 < x_1$ then $F(x_0) \leq F(x_1)$

Example. The number of computers sold per day is a random variable with the following probability distribution function

x	P(x)
0	0.05
1	0.1
2	0.2
3	0.2
4	0.2
5	0.15
6	0.1

$$F(4) = P(X \le 4) = 0.05 + 0.1 + 0.2 + 0.2 + 0.2 = 0.75$$

$$F(6) = P(X \le 6) = 0.05 + 0.1 + 0.2 + 0.2 + 0.2 + 0.15 + 0.1 = 1$$

$$P(4 \le X \le 6) = \Sigma_{4 \le x \le 6} P(x) = 0.2 + 0.15 + 0.1 = 0.45$$

Example of a discrete distribution: the binomial distribution

Suppose that an experiment is repeated n times and on each time, the probability of success is p. Let *X* be the **number** of successes.

X is said to have a binomial distribution with parameters n and p

$$P(X = x) = C_x^n p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Note: A Binomial with parameters 1 and p is called a *Bernoulli* random variable. It is 1 w.p. p and 0 otherwise.

Example: You toss a fair coin 10 times. What is the probability of exactly 3 heads? What is the probability of at least 3 heads?

Let X be the number of heads. On each experiment, there is a head with probability 0.5 and a tail with probability 0.5

$$P(3) = \frac{10!}{3!7!}(0.5)^3(1-0.5)^7 = 0.1172$$

Probability of at least 3 heads is 1-P(0)-P(1)-P(2)=0.945

Another example of a discrete distribution: Poisson

Suppose that we are interested in the number of successes that happen in an interval of time

Examples.
1. # of customers to arrive at a checkout aisle in an hour.
2. # of failures on a computer network during a day.
3. # of jumps (large price movements) in the price of a stock during a year.

Suppose that

1. The probability of a success at any one instant is constant.

2. The probability of two successes happening at exactly the same time is small.

3. The timing of successes are independent.

Then the number of successes in **any** interval of time follows a **Poisson** distribution.

The Poisson Distribution takes on the values 0,1,2,.....

The probability distribution function for the Poisson is

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where λ is a parameter that determines how frequently events occur and $e \simeq 2.71828$

Example: The number of failures on a computer network per day is Poisson with parameter $\lambda = 1$.

What is the probability that more than 2 computers fail on a given day?

$$P(X = 0) = \frac{e^{-1}1^0}{0!} = 0.368$$

$$P(X = 1) = \frac{e^{-1}1^1}{1!} = 0.368$$

$$P(X = 2) = \frac{e^{-1}1^2}{2!} = 0.184$$

Probability more than 2 is 1-0.368-0.368-0.184=0.08

A continuous random variable can take on any value in an interval

A discrete random variable cannot

For a continuous random variable, the probability that it takes on any specific value is zero.

We have to use integral calculus to work with continuous random variables.

Let X be a continuous random variable, and let x be any number in the range of possible values for this random variable. The **cumulative distribution function** of this random variable is defined as

$$F(x_0) = P(X \leq x_0)$$

Note that since $P(X = x_0) = 0$, it does not matter whether the definition states $P(X \leq x_0)$ or $P(X < x_0)$

Given the cumulative distribution function $F(x)$, the **probability density function** is

$$f(x) = \frac{dF(x)}{dx}$$

Properties of a continuous probability density function $f(x)$:

1. $f(x) \geq 0$ for all x
2. The area under the probability density function $f(x)$ over all values of the random variable X equals 1: $\int_{-\infty}^{\infty} f(x) = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$

Properties of a continuous cumulative dist. function $F(x)$

1. $P(X \leq x_0) = F(x_0) = \int_{-\infty}^{x_0} f(x)dx$
2. $0 \leq F(x) \leq 1$ for every x
3. If $x_0 < x_1$ then $F(x_0) \leq F(x_1)$

We can construct the following parallels

Discrete	Continuous
$P(X=x)$	$f(x)$
$\Sigma_x P(X=x)=1$	$\int_{-\infty}^{\infty} f(x)=1$
$P(X \le x_0)=\Sigma_{x \le x_0} P(x)$	$F(x_0)=\int_{-\infty}^{x_0} f(x)dx$
$P(a \le X \le b)=\Sigma_{a \le x \le b} P(x)$	$P(a \le X \le b)=F(b)-F(a)$

Example of a continuous distribution: the uniform distribution

The uniform distribution can take on any value between a and b. Each value is equally likely.

$$F(x) = \frac{x - a}{b - a}$$

$$f(x) = \frac{1}{b - a}$$

X is uniform on the unit interval (i.e. between 0 and 1).

What is the probability that *X* is between ½ and ¾?

F(½)=(½-0)/(1-0)= ½

F(¾)=(¾-0)/(1-0)= ¾

The probability that *X* is between ½ and ¾ is F(¾)-F(½)=¼

Example of a continuous distribution: the logistic distribution

cdf $F(x) = \dfrac{1}{1+e^{-x}}$

pdf $f(x) = \dfrac{e^{-x}}{(1+e^{-x})^2}$

Expectations: For a discrete random variable

The expected value for a random variable X, E(X), is

$$E(X) = \mu_x = \Sigma_x P(x)x$$

The expected value is also known as the mean

Example: The probability distribution function for the number of errors on each page of a book is

P(0)=0.81

P(1)=0.17

P(2)=0.02

The expected number of errors is

(0.81*0)+(0.17*1)+(0.02*2)=0.21

Saint Petersburg Paradox (Bernoulli in the 18th century)

Suppose that there is a lottery with possible payoffs:

Payoff	Probability
1	½
2	¼
4	1/8
8	1/16
16	1/32
etc....	etc...

What is the expected payoff from this lottery?

$$E(X) = \Sigma P(x)x = \frac{1}{2} + \frac{2}{4} + \frac{4}{8} + \frac{8}{16} .... = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} ... = \infty$$

How much would you pay for one of these lottery tickets?

They have an infinite expected value

Expectations also exist for functions of a random variable

The expected value for g(X), E(g(X)), is given by

$$E(g(X)) = \mu_{g(x)} = \Sigma_x P(x) g(x)$$

Example. There is a project which will take a number of days to complete. Here is the probability distribution of the number of days

P(1)=0.5

P(2)=0.25

P(3)=0.15

P(4)=0.1

The cost to the contractor is the number of days squared (costs go up because the 3rd and 4th days are weekends). What is the expected cost to the contractor?

(0.5*1)+(0.25*4)+(0.15*9)+(0.1*16)=4.45

Example: binomial distribution

An experiment is repeated n times and the probability of success each time is p. X is the number of successes

$$P(X = x) = C_x^n p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$E(X) = \Sigma_{x=0}^{n} P(X = x)x = \Sigma_{x=0}^{n} \frac{n!x}{x!(n-x)!} p^x (1-p)^{n-x} = np$$

(after a few lines of algebra)

Example: Poisson distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$E(X) = \Sigma_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!} x$$

$$\therefore E(X) = \Sigma_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!} x = \Sigma_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{(x-1)!} = \lambda e^{-\lambda} \Sigma_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda e^{-\lambda} \Sigma_{x=0}^{\infty} \frac{\lambda^{x}}{x!}$$

$$e^{z} = 1 + z + \frac{z^{2}}{2!} + \frac{z^{3}}{3!} \ldots = \Sigma_{i=0}^{\infty} \frac{z^{i}}{i!} \text{ (Defn. of exponential)}$$

$$\therefore E(X) = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

Expectations: For a continuous random variable

$$E(X) = \int xf(x)dx$$

where the integral is taken over the "support" of the random variable....i.e. the set of possible outcomes.

Example. X is uniform between a and b.

$$E(X)=\int_a^b xf(x)dx=\int_a^b \frac{x}{b-a}dx=\frac{1}{b-a}\int_a^b xdx=\frac{1}{b-a}[\frac{x^2}{2}]_a^b$$

$$\therefore E(X)=\frac{1}{b-a}(\frac{b^2}{2}-\frac{a^2}{2})=\frac{b^2-a^2}{2(b-a)}=\frac{(b-a)(b+a)}{2(b-a)}=\frac{b+a}{2}$$

Example. $f(x)=\frac{1}{\sqrt{2}\sigma}\exp(-\frac{|x-\mu|}{\sigma/\sqrt{2}})$. Find $E(X)$

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx=\int_{-\infty}^{\infty}\frac{x}{\sqrt{2}\sigma}\exp(-\frac{|x-\mu|}{\sigma/\sqrt{2}})dx$$

$$=\int_{-\infty}^{\mu}\frac{x}{\sqrt{2}\sigma}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})dx+\int_{\mu}^{\infty}\frac{x}{\sqrt{2}\sigma}\exp(\frac{\mu-x}{\sigma/\sqrt{2}})dx$$

Use integration by parts

$$\int_a^b f(x)g'(x)dx=[f(x)g(x)]_a^b-\int_a^b f'(x)g(x)dx$$

$$\therefore\int_{-\infty}^{\mu}\frac{x}{\sqrt{2}\sigma}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})dx=[\frac{x}{\sqrt{2}\sigma}\frac{\sigma}{\sqrt{2}}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})]_{-\infty}^{\mu}$$

$$-\int_{-\infty}^{\mu}\frac{1}{\sqrt{2}\sigma}\frac{\sigma}{\sqrt{2}}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})dx$$

$$= [\frac{x}{2}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})]_{-\infty}^{\mu} - \int_{-\infty}^{\mu} \frac{1}{2}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})dx$$

$$= [\frac{x}{2}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})]_{-\infty}^{\mu} - [\frac{1}{2}\frac{\sigma}{\sqrt{2}}\exp(\frac{x-\mu}{\sigma/\sqrt{2}})]_{-\infty}^{\mu}$$

$$= \frac{\mu}{2} - [\frac{1}{2}\frac{\sigma}{\sqrt{2}}] = \frac{\mu}{2} - \frac{\sigma}{2\sqrt{2}}$$

Similarly, $\int_{\mu}^{\infty} \frac{x}{\sqrt{2}\sigma}\exp(\frac{\mu-x}{\sigma/\sqrt{2}})dx = \frac{\mu}{2} + \frac{\sigma}{2\sqrt{2}}$

$$\therefore E(X) = \frac{\mu}{2} - \frac{\sigma}{2\sqrt{2}} + \frac{\mu}{2} + \frac{\sigma}{2\sqrt{2}} = \mu$$

For a continuous random variable

$$E(g(X)) = \int g(x) f(x) dx$$

For a discrete or continuous random variable, the **nth uncentered moment** is $\mu_n' = E(X^n)$

The **nth centered moment** is $\mu_n = E[(X - \mathrm{E}(X))^n]$
(It is centered around the mean, $E(X)$)

For $n = 2: \mu_2 = E[(X - E(X))^2]$ is the variance of X, $\mathrm{var}(X)$

$\sqrt{\mu_2}$ is the standard deviation

Properties of expectation and variance

$$E(a+bX)=a+bE(X)$$

$$E(a+\Sigma b_i X_i)=a+\Sigma b_i E(X_i)$$

$$\mathrm{Var}(a+bX)=b^2\mathrm{Var}(X)$$

$$\mathrm{Var}(X)=E(X^2)-(E(X))^2$$

Variance of a uniform random variable

Let X be uniform from a to b.

$$f(x) = \frac{1}{b-a} \text{ and } E(X) = \frac{a+b}{2}$$

$$E(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^3-a^3}{3(b-a)}$$

$$Var(X) = E(X^2) - E(X)^2 = \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4}$$

$$\therefore Var(X) = \frac{4(b^3-a^3)}{12(b-a)} - \frac{3(b-a)(a+b)^2}{12(b-a)}$$

$$\therefore Var(X) = \frac{b^3-a^3-3ab^2+3a^2b}{12(b-a)} = \frac{(b-a)^3}{12(b-a)} = \frac{(b-a)^2}{12}$$

Expectations and variances of the rvs we've seen

	Expectation	Variance
Binomial	np	$np(1-p)$
Poisson	λ	λ
Uniform a to b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Jensen's Inequality

For a nonlinear function g(.)

$E(g(X)) \neq g(E(X))$

- If g is concave then $E(g(X)) \leq g(E(X))$
- If g is convex then $E(g(X)) \geq g(E(X))$

Jensen's Inequality can be used to prove an inequality between arithmetic and geometric means

If $\bar{X} = \frac{1}{n}(X_1 + X_2 ... + X_n)$ is the arithmetic mean of $X_1,...X_n$
and $\bar{X}_G = [X_1 X_2 ... X_n]^{1/n}$ is their geometric mean
then $\bar{X}_G \leq \bar{X}$.

Proof: Define X as a discrete random variable with support $X_1,...X_n$ such that $P(X = X_i) = 1/n$.

$$\log(\bar{X}_G) = \frac{1}{n}\Sigma_{i=1}^{n} \log(X_i) = E(\log(X))$$

$$E(\log(X)) \leq \log(E(X)) \qquad \text{(Jensen's Inequality)}$$

$$\log(\bar{X}) = \log(\frac{1}{n}\Sigma_{i=1}^{n} X_i) = \log(E(X))$$

$$\therefore \log(\bar{X}_G) \leq \log(\bar{X})$$

$$\therefore \bar{X}_G \leq \bar{X}$$

An individual maximizes the expectation of the utility function $u(c)$ where the function is concave and c is consumption. Which will this individual prefer, $c = 1$ or c is a random variable that is uniform from 0 to 2?

By Jensen's inequality, $E(u(E(c))) = u(E(c)) \geq E(u(c))$ and so the individual will always prefer $c = 1$.

Moment generating function

The **moment generating function** of a random variable X is $M(t) = E(e^{tX})$. So

$M(t) = \Sigma e^{tx} P(X = x)$ (for discrete X)

$M(t) = \int e^{tx} f(x) dx$ (for continuous X)

The uncentered moments of X are generated from the moment generating function by

$$E(X^n) = \frac{d^n}{dt^n} M(t)\Big|_{t=0}$$

Example: Poisson random variable $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$M(t) = \Sigma e^{tx} P(X = x)$

$\therefore M(t) = \Sigma_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\Sigma_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}$

$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} \ldots = \Sigma_{i=0}^{\infty} \frac{z^i}{i!}$ (Defn. of exponential)

$\therefore M(t) = e^{-\lambda} e^{\lambda e^t} = \exp(\lambda(e^t - 1))$

Now $M'(t) = \lambda e^t \exp(\lambda(e^t - 1)) \Rightarrow M'(0) = \lambda \Rightarrow E(X) = \lambda$

$M''(t) = (\lambda^2 e^{2t} + \lambda e^t)\exp(\lambda(e^t - 1))$

$\Rightarrow M'(0) = \lambda^2 + \lambda \Rightarrow E(X^2) = \lambda^2 + \lambda$

$\mathrm{Var}(X) = E(X^2) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

Moment generating functions are useful because they can characterize moments, and especially, because in many cases, they characterize a distribution.

Theorem. If $M_X(t)$ and $M_X(t)$ are the mgfs of X and Y respectively, then $M_X(t) = M_Y(t)$ for all t in some neighborhood of zero, then $F_X(u) = F_Y(u)$ for all u.

Result: If $Y = aX + b$ and $m(t)$ is the mgf of X, then $e^{bt}m(at)$ is the mgf of Y.

Characteristic function

The **characteristic function** of a random variable X is

$$\phi_X(t) = E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f_X(x)dx$$

The moment generating function uniquely determines a distribution, if the mgf exists.

But a characteristic function always exists and uniquely characterizes a distribution (every cdf has a unique characteristic function).

Inversion formula.

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)dt$$

Transformations of random variables

Suppose that X is a continuous random variable with cdf F_X, pdf f_X and $Y=g(X)$. Y is also a random variable. What is its cdf and pdf?

Suppose g is monotone

1. If g is monotone increasing $F_Y(y) = F_X(g^{-1}(y))$
2. If g is monotone decreasing $F_Y(y) = 1 - F_X(g^{-1}(y))$
3. $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$
4. If the support of X is from $X_{\min}$ to $X_{\max}$, the support of Y is from $g(X_{\min})$ to $g(X_{\max})$

Example 1: X is uniform between -1 and +1 and $Y = \exp(X)$

$F_X(x) = (x+1)/2$ and $f_X(x) = 1/2$

g(.) is a monotone increasing transformation

$g(.) = \exp(.) \Rightarrow g(.)^{-1} = \ln(.)$

$\therefore F_Y(y) = (\ln(y)+1)/2$

The support of Y is $[1/e, e]$.

The pdf is $f_Y(y) = F_Y'(y) = \dfrac{1}{2y}$

Or $f_Y(y) = f_X(g(y)^{-1}) \left| \dfrac{dg^{-1}(y)}{dy} \right| = \dfrac{1}{2}\dfrac{1}{y} = \dfrac{1}{2y}$

Example 2: X is uniform between 0 and 1 and $Y = -\lambda \log(X)$

$F_X(x) = x$ and $f_X(x) = 1$

g(.) is a monotone decreasing transformation

$g(.) = -\lambda \log(.) \Rightarrow g(.)^{-1} = \exp(-y/\lambda)$

$\therefore F_Y(y) = 1 - \exp(-y/\lambda)$

The support of Y is from $-\infty$ to 0.

The pdf is $f_Y(y) = F_Y'(y) = \frac{1}{\lambda} e^{-y/\lambda}$

Or

$$f_Y(y) = f_X(g(y)^{-1}) \left| \frac{dg^{-1}(y)}{dy} \right| = 1 \left| -\frac{1}{\lambda} \exp(-y/\lambda) \right| = \frac{1}{\lambda} e^{-y/\lambda}$$

Example 3: X is uniform between -1 and +1 and $Y = X^2$

$F_X(x) = (x+1)/2$

g(.) is NOT monotone transformation..formula doesn't work

$$F_Y(y) = P(X^2 \le y)$$

$$= P(-\sqrt{y} \le X \le \sqrt{y})$$

$$= F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

$$= \frac{\sqrt{y}+1}{2} - \frac{-\sqrt{y}+1}{2}$$

$$= \sqrt{y}$$

The support of Y is from 0 to 1.The pdf is $f_Y(y) = \frac{1}{2\sqrt{y}}$

Example 4: X has the pdf $f(x) = e^{-x}$ for $x > 0$. Find the cdf and pdf of $Y = 1/X$.

$F(x) = 1 - e^{-x}$

$g(X)$ is a monotone decreasing transformation

$g^{-1}(Y) = 1/Y$

$\therefore F_Y(y) = 1 - \{1 - e^{-1/y}\} = e^{-1/y}$

The support of Y is 0 to ∞.

The pdf is $f_Y(y) = F_Y'(y) = \frac{1}{y^2} e^{-1/y}$

Or $f_Y(y) = f_X(g(y)^{-1}) \left| \frac{dg^{-1}(y)}{dy} \right| = e^{-1/y} \left| -\frac{1}{y^2} \right| = \frac{1}{y^2} e^{-1/y}$

Example 5: X has the pdf $f(x) = e^{-x}$ for $x > 0$. Find the cdf and pdf of $Y = \log(X)$.

$F(x) = 1 - e^{-x}$

$g(X)$ is a monotone increasing transformation

$X = g^{-1}(Y) = \exp(Y)$

$\therefore F_Y(y) = 1 - e^{-\exp(y)}$

The support of Y is $-\infty$ to ∞.

The pdf is $f_Y(y) = F_Y'(y) = e^{-\exp(y)}e^y = \exp(y - e^y)$

Or $f_Y(y) = f_X(g(y)^{-1}) \left| \frac{dg^{-1}(y)}{dy} \right| = e^{-e^y} \mid e^y \mid = \exp(y - e^y)$

Example 6: X has the pdf $f(x) = e^{-x}$ for $x > 0$. Find the cdf and pdf of $Y = 1 - e^{-X}$.

$F(x) = 1 - e^{-x}$

$g(X)$ is a monotone increasing transformation

$X = g^{-1}(Y) = -\log(1 - Y)$

$\therefore F_Y(y) = 1 - e^{\log(1-y)}$

The support of Y is from 0 to 1.

The pdf is $f_Y(y) = F_Y'(y) = \dfrac{e^{\log(1-y)}}{1-y} = 1$

Or $f_Y(y) = f_X(g(y)^{-1}) \left| \dfrac{dg^{-1}(y)}{dy} \right| = e^{\log(1-y)} \left| \dfrac{1}{1-y} \right| = \dfrac{e^{\log(1-y)}}{1-y} = 1$

Theorem. If X has a continuous cdf $F_X(x)$ and $Y = F_X(X)$ then Y is uniform on the unit interval, i.e. $F_Y(y) = y$.

Proof. g(.) is monotone

$F_Y(y) = F_X(g^{-1}(y))$

$g = F_X \Rightarrow g^{-1} = F_X^{-1}$

$\therefore F_Y(y) = F_X(F_X^{-1}(y)) = y$

Intuition of why this is useful. Suppose I think I know the cdf of X, but I want to check.

Define $Y = F_X(X)$ and look to see if Y seems uniform.

Specific Discrete Distributions

- Binomial
- Poisson
- Negative Binomial
- Geometric

The Poisson Distribution

We've already seen the Poisson Distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Useful facts about Poisson distribution

- The sum of K Poisson Random Variables with parameter λ is a Poisson Random Variable with parameter $K\lambda$.
- If the number of events occurring in a period of time is Poisson with parameter λ then the number of events in t periods is Poisson with parameter λt

Example. A computer manager reports that the expected number of component failures on a system is 3 every 100 days. The number of component failures is Poisson.

(a) What is the probability of no failures in a given day?
(b) What is the probability of **one or more** component failures on a given day?
(c) What is the probability of **at least two** failures in a three day period?

Let X be the number of computer failures per day.

The number of computer failures in 100 days is Poisson with expectation 3.

Therefore the number of computer failures in 100 days is Poisson with parameter 3.

Therefore the number of computer failures in 1 day is Poisson with parameter 0.03.

(a) The probability of no failure on one day is

$$P(0) = \frac{e^{-0.03} 0.03^0}{0!} = 0.97$$

(b) The probability of one or more failures is

$$1 - P(0) = 1 - 0.97 = 0.03$$

Therefore the number of computer failures in 100 days is Poisson with parameter 3.

Therefore the number of computer failures in 3 days is Poisson with parameter 0.09.

The probability of no failure is $P(0) = \dfrac{e^{-0.09}0.09^0}{0!} = 0.914$

The probability of 1 failure is $P(1) = \dfrac{e^{-0.09}0.09^1}{1!} = 0.082$

The probability of 2 or more failures is
$1 - P(0) - P(1) = 1 - 0.914 - 0.082 = 0.004$

The Poisson can be used to approximate the binomial

Let X be binomial with parameters n and p.

If n is large, X also has approximately a Poisson distribution with parameter $\lambda = np$

Sketch of Proof:

mgf of a binomial is $M_X(t) = [pe^t + (1-p)]^n$

mgf of a Poisson is $M_Y(t) = \exp(\lambda(e^t - 1))$

Let $p = \lambda / n$

$$\lim_{n\to\infty} M_X(t) = \lim_{n\to\infty}[\frac{\lambda}{n}e^t + 1 - \frac{\lambda}{n}]^n$$

$$= \lim_{n\to\infty}[1 + \frac{\lambda e^t - 1}{n}]^n = \exp(\lambda(e^t - 1))$$

(using $e^x = \lim_{n\to\infty}(1 + \frac{x}{n})^n$)

Negative binomial distribution

The binomial distribution gives the number of successes when a trial is repeated n times (p: prob. of success each time)

The *negative* binomial distribution turns the question around and gives the number of trials needed for r successes to occur.

$$P(X = x) = C_{r-1}^{x-1} p^r (1-p)^{x-r}, \; x = r, r+1, r+2, \ldots$$

Example: What is the probability that I need to toss a fair coin exactly 5 times to get 2 heads?

Answer: x=5, r=2

$$P(X = x) = C_{r-1}^{x-1} p^r (1-p)^{x-r}$$

$$P(X = 5) = C_1^4 p^1 (1-p)^4 = \frac{4!}{1!3!} * 0.5 * 0.5^4 = 4 * 0.5^5 = \frac{1}{8}$$

Geometric distribution

The geometric distribution is the number of trials needed for the first success.

It is a special case of the negative binomial distribution with r=1

$$P(X = x) = p(1-p)^{x-1},\ x = 1,2,3,...$$

Specific Continuous Distributions

- Uniform (already)
- Normal (now)
- Log-Normal
- Beta
- Truncated Normal (later in the class)
- Exponential
- Gamma
- Chi-Squared
- T-distribution
- Cauchy
- F-distribution

The normal distribution

The most important distribution is the normal.

It was discovered by Gauss and is also called the Gaussian distribution.

The probability density function for a normal random variable is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \text{ for } -\infty < x < \infty$$

If X has a normal distribution then

$E(X) = \mu$

and

$Var(X) = \sigma^2$

The parameters of the normal distribution are μ (the "mean") and σ^2 (the "variance")

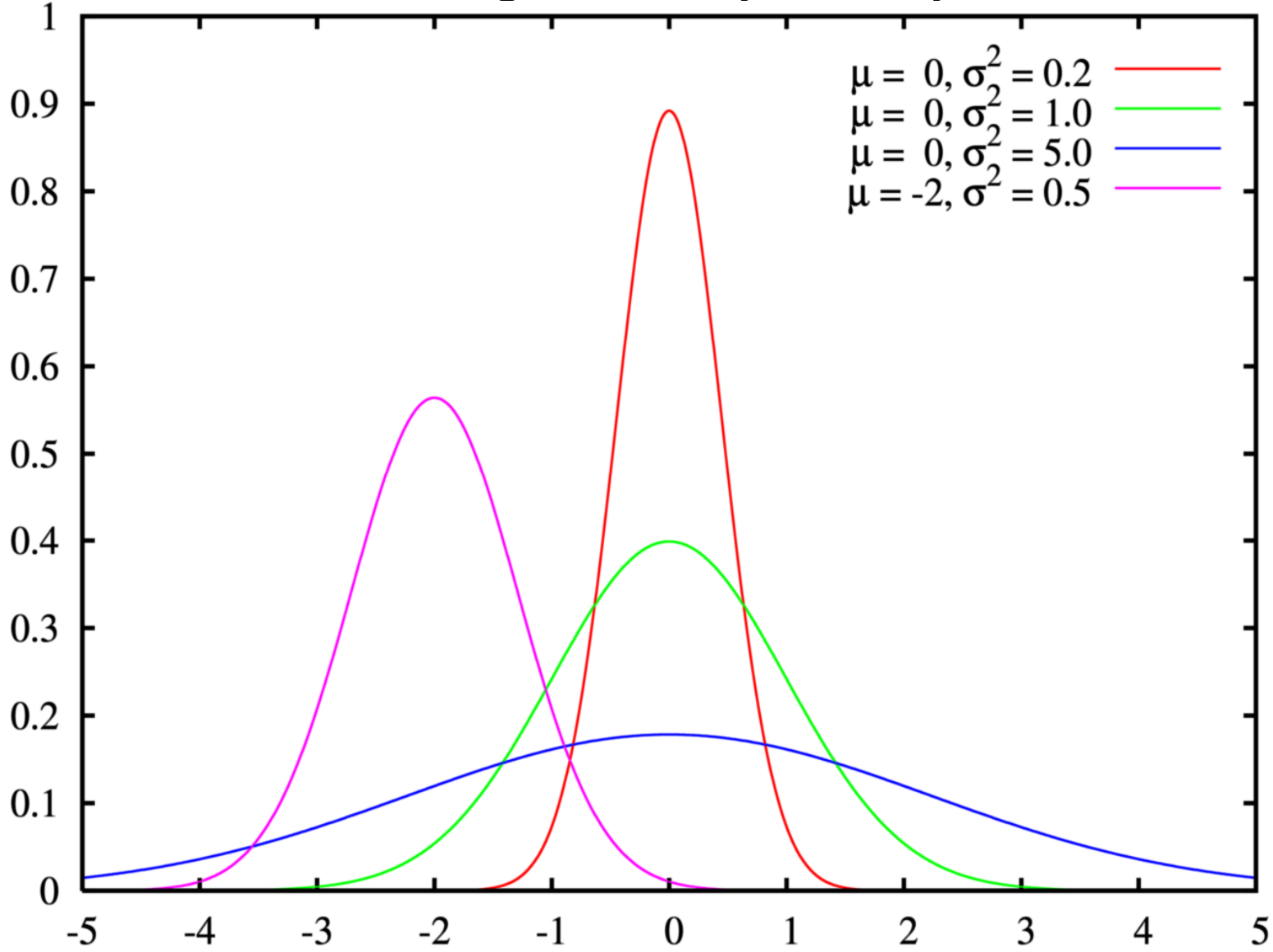
σ^2 must be positive, which is why we write it as sigma-squared

We write a normal random variable with mean μ and variance σ^2 as an $N(\mu,\sigma^2)$ random variable

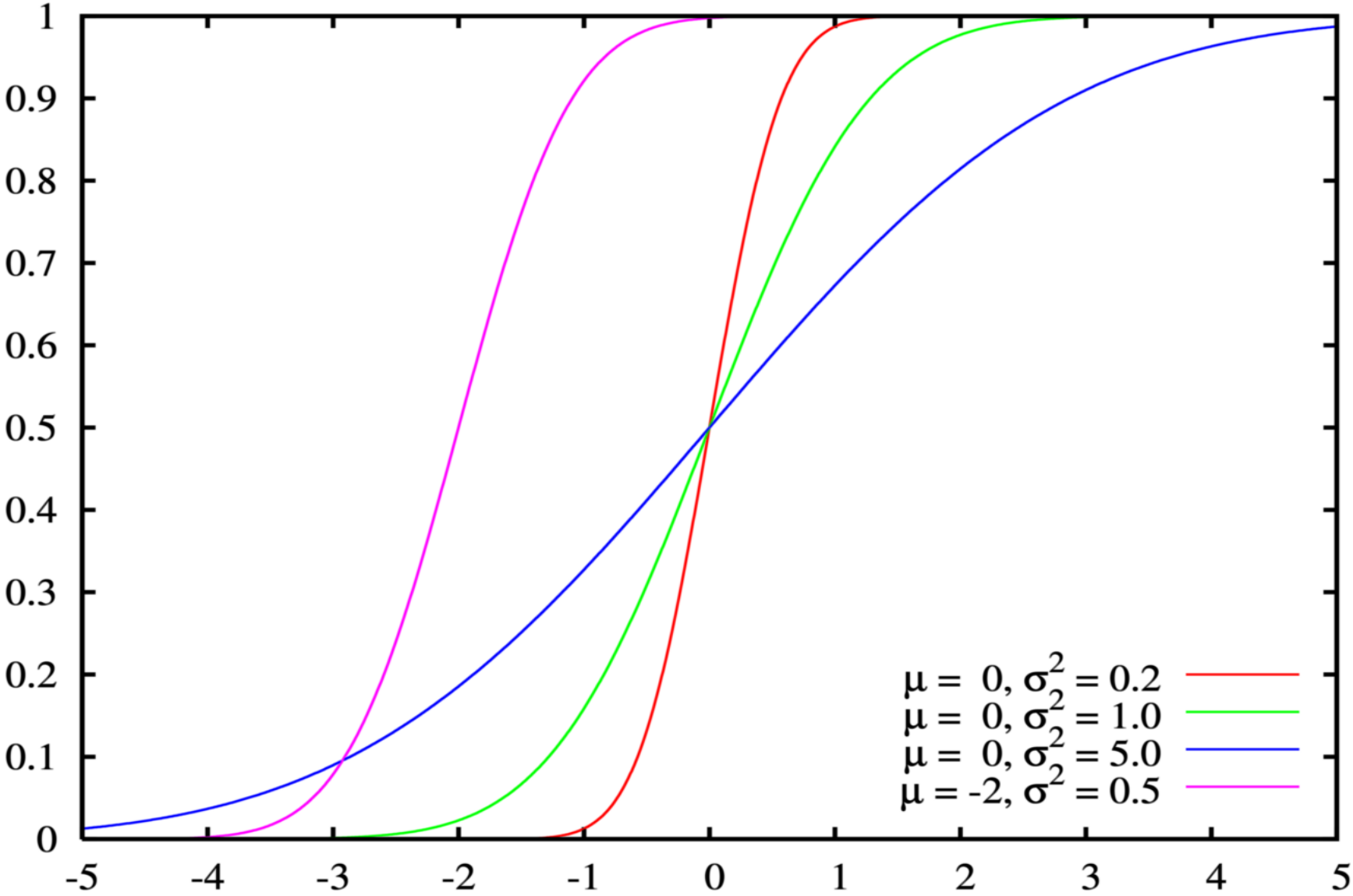
The special case in which $\mu = 0$ and $\sigma^2 = 1$ is a **standard** normal random variable.

We write its pdf as $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and its cdf as $\Phi(x)$

Normal probability density functions



Normal cumulative distribution functions



Properties of the normal distribution

- The normal distribution is symmetric about zero
  $P(X \le x) = P(X \ge -x) = 1 - P(X \le -x)$

- There is no closed form expression for the cdf of a normal random variable.
- But there are tables for looking up the cdf of a standard normal random variable.
- These give $P(X \le x)$ where X is a standard normal random variable for $x \ge 0$

- If $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$

Tables of the Normal Distribution

Probability Content from $-\infty$ to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

Gives $P(X \leq x)$ for $x \geq 0$

But $P(X \leq x) = 1 - P(X \leq -x)$

So $P(X \leq -1) = 1 - P(X \leq 1) = 1 - 0.8413 = 0.1587$

Classic cases

$P(X \geq 1.64) = P(X \leq -1.64) = 0.05$

$P(X \geq 1.28) = P(X \leq -1.28) = 0.1$

$P(X \geq 1.96) = P(X \leq -1.96) = 0.025$

If X is $N(\mu,\sigma^2)$ then $Z=\dfrac{X-\mu}{\sigma}$ is N(0,1)

P($X\le$k)=$P(X-\mu\le k-\mu)$=P($\dfrac{X-\mu}{\sigma}\le\dfrac{k-\mu}{\sigma}$)=P(Z$\le\dfrac{k-\mu}{\sigma}$)

If Φ denotes the standard normal cdf and $X\sim N(\mu,\sigma^2)$ then the cumulative distribution function of X is $F(x)=\Phi(\dfrac{x-\mu}{\sigma})$

Example. Suppose that X is N(4,5). Find $P(X\le -1)$

$P(X\le -1)$== $\Phi(-2.236)$=0.013

If $\phi(.)$ denotes the standard normal probability density and $X \sim N(\mu, \sigma^2)$ then the probability density function of X is

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$

Follows from $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ by differentiating both sides

The normal moment generating function is

$$M(t) = E(e^{tX}) = \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{tx} dx = \exp(\mu t + \frac{\sigma^2 t^2}{2})$$

The standard normal mgf is $\exp(t^2 / 2)$

Higher moments of a normal distribution

If $X \sim N(\mu, \sigma^2)$, then

- Skewness: $E((X-\mu)^3) = 0$
- Kurtosis: $E((X-\mu)^4) = 3\sigma^4 \Rightarrow E(\frac{(X-\mu)^4}{\sigma^4}) = 3$

The log-normal distribution

Another important distribution is the log-normal.

If $X \sim N(\mu, \sigma^2)$, then $\exp(X)$ is log normal.

The support is from zero to infinity and it is skewed to the right.

It is useful for modeling variables such as individual incomes or interest rates.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$ is the pdf of X

$$Y = \exp(X) \Rightarrow g^{-1}(.) = \log(.)$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\log(y)-\mu)^2/2\sigma^2} \frac{1}{y}$$

The normal mgf was

$$E(e^{tX}) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$$

So $E(Y) = \exp(\mu + \frac{\sigma^2}{2})$ is the mean of a lognormal

$E(Y) > \exp(E(X))$ (consistent with Jensen's inequality)

The beta distribution

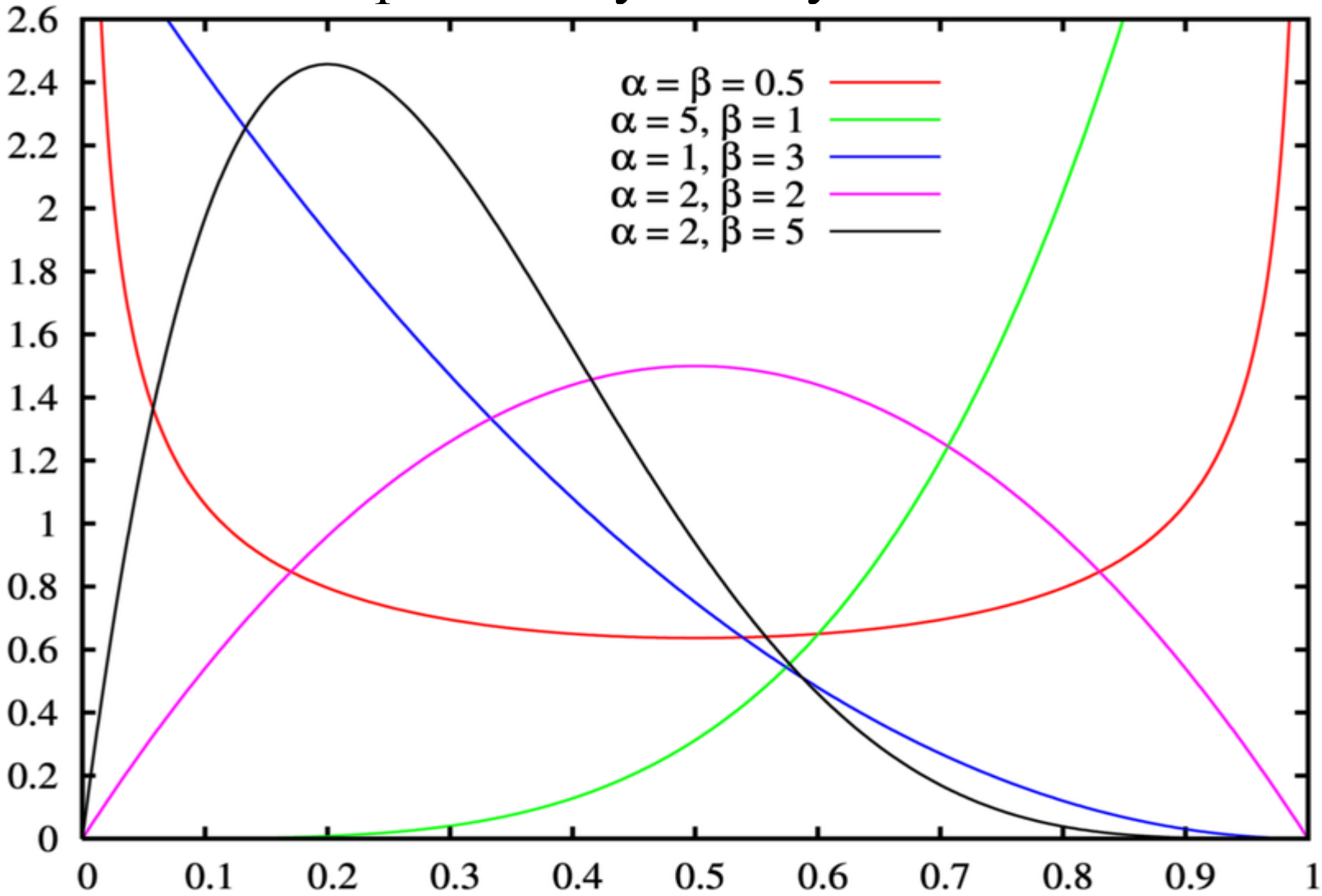
The beta distribution has support on the unit interval.

It has two parameters, α and β that determine the shape of the density.

$E(X) = \alpha / (\alpha + \beta)$

$\alpha = \beta = 1$ is the uniform, but the beta distribution can be much more flexible and is useful for modeling random variables must lie in given intervals.

Beta probability density functions



The beta pdf is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \alpha > 0, \beta > 0$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.

Note: $\Gamma(\alpha) = (\alpha-1)!$ for positive integer α

Chebychev's Inequality.

For any nonnegative random variable X and constant c>0

$$P(X \geq c) \leq \frac{E(X)}{c}$$

Proof: $E(X) = \int_0^\infty xf(x)dx = \int_0^c xf(x)dx + \int_c^\infty xf(x)dx$

$$\therefore E(X) \geq \int_c^\infty xf(x)dx$$

$$\geq \int_c^\infty cf(x)dx = c\int_c^\infty f(x)dx = cP(X \geq c)$$

$$\therefore E(X) \geq cP(X \geq c) \Rightarrow P(X \geq c) \leq E(X)/c$$

Simpler form of Chebychev (often quoted). For any random variable Y with mean μ and variance σ^2 and constant c>0

$$P((Y-\mu)^2 \geq c\sigma^2) \leq \frac{1}{c}$$

Proof: Let $X = (\frac{Y-\mu}{\sigma})^2$

$$P(X \geq c) \leq \frac{E(X)}{c}$$

$$\therefore P((\frac{Y-\mu}{\sigma})^2 \geq c) \leq \frac{E((\frac{Y-\mu}{\sigma})^2)}{c} = \frac{\frac{1}{\sigma^2}E((Y-\mu)^2)}{c} = \frac{1}{c}$$

$$\therefore P((Y-\mu)^2 \geq c\sigma^2) \leq \frac{1}{c}$$

Example of Chebychev

Let Y be a N(0,1) random variable.

$$P(Y^2 \geq c) \leq \frac{1}{c}$$

$$\therefore P(Y \geq \sqrt{c}) + P(Y \leq -\sqrt{c}) \leq \frac{1}{c}$$

$$\therefore 2P(Y \geq \sqrt{c}) \leq \frac{1}{c}$$

$$\therefore P(Y \geq \sqrt{c}) \leq \frac{1}{2c}$$

e.g. $P(Y \geq 1.96) \leq \frac{1}{2 * 1.96^2} = 0.13$

True, but not a “sharp” bound.

Multiple Random Variables: Discrete

An **n-dimensional random variable** is a function from the sample space to R^n.

The **joint cumulative distribution function** for an n-dimensional random variable is
$P(X_1 \leq x_1, X_2 \leq x_2, ... X_n \leq x_n)$

The **joint probability mass function** for an n-dimensional random variable is
$P(X_1 = x_1, X_2 = x_2, ... X_n = x_n)$

Let X and Y denote two discrete random variables.

Here is a way to represent them:

	y_1	y_2	..	y_k
x_1	$x_1 \cap y_1$	$x_1 \cap y_2$		$x_1 \cap y_k$
x_2	$x_2 \cap y_1$	$x_2 \cap y_2$		$x_2 \cap y_k$
x_h	$x_h \cap y_1$	$x_h \cap y_2$		$x_h \cap y_k$

$P(X = x_i, Y = y_j)$ is a joint probability

$P(X = x_i) = \Sigma_j P(X = x_i, Y = y_j)$ is a marginal probability

We can define conditional probabilities for pairs of random variables

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

Example. Toss a fair coin three times. The sample space is

HHH,HHT,HTH,HTT,THH,THT,TTH,TTT

each of which is equally likely.

X: Number of heads in the first two tosses

Y: Number of heads in the second two tosses

	Y=0	Y=1	Y=2
X=0	TTT	TTH	
X=1	HTT	HTH,THT	THH
X=2		HHT	HHH

Joint Probabilities

	Y=0	Y=1	Y=2	
X=0	1/8	1/8	0	1/4
X=1	1/8	1/4	1/8	1/2
X=2	0	1/8	1/8	1/4
	1/4	1/2	1/4	

- What is $P(Y \leq 1, X \leq 1)$?

  5/8

- What is $P(Y = 2 \mid X = 1)$?

$$P(Y = 2 \mid X = 1) = \frac{P(Y = 2, X = 1)}{P(X = 1)} = \frac{1/8}{1/2} = \frac{1}{4}$$

Example. Here is the joint distribution of two random variables. What is P(Y=0|X=0)?

	Y=0	Y=1
X=0	0.3	0.5
X=1	0.1	0.1

$$P(Y=0 \mid X=0) = \frac{P(Y=0, X=0)}{P(X=0)} = \frac{0.3}{0.8} = 0.375$$

Multiple Random Variables: Continuous

Let X and Y denote two continuous random variables.

The joint cumulative distribution function of X and Y is defined as

$$F(x,y) = P(X \le x, Y \le y)$$

The joint probability density function is

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$$

So $F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t)dtds$

$$P((X,Y) \in A) = \iint_{(s,t)\in A} f(s,t)dtds$$

Marginal densities from integrating rather than summing

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

Example: $f(x,y) = 6xy^2, \quad 0 < x < 1$ and $0 < y < 1$
and 0 otherwise. What is the marginal density of X?

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 f(x,y)dy = \int_0^1 6xy^2 dy$$

$$\therefore f(x) = 6x[\frac{y^3}{3}]_0^1 = \frac{6x}{3} = 2x$$

Example. Suppose that the joint pdf of X and Y is given by

$ce^{-x}e^{-2y}$ if $0 < x < \infty$ and $0 < y < \infty$

0 otherwise

1. What is c?

$$\int_0^\infty \int_0^\infty ce^{-x}e^{-2y}dxdy = \int_0^\infty [-ce^{-x}]_0^\infty e^{-2y}dy = \int_0^\infty ce^{-2y}dy = -\frac{c}{2}[e^{-2y}]_0^\infty = \frac{c}{2}$$

So $c = 2$

2. What is the marginal density of X?

$$f_X(x) = \int_0^\infty 2e^{-x}e^{-2y}dy = 2e^{-x}[-\frac{e^{-2y}}{2}]_0^\infty = e^{-x}$$

3. What is $F(x,y)$?

$$\int_0^x \int_0^y 2e^{-s}e^{-2t}dtds = \int_0^x 2e^{-s}[-\frac{e^{-2t}}{2}]_0^y ds = \int_0^x 2e^{-s}[\frac{1-e^{-2y}}{2}]ds$$

$$= (\frac{1-e^{-2y}}{2})[-2e^{-s}]_0^x = 2(\frac{1-e^{-2y}}{2})(1-e^{-x}) = (1-e^{-2y})(1-e^{-x})$$

4. What is P(X>1,Y<1)?

$$\int_0^1 \int_1^\infty 2e^{-x}e^{-2y}dxdy = \int_0^1 2e^{-2y}[-e^{-x}]_1^\infty dy = \int_0^1 2e^{-2y}e^{-1}dy = e^{-1}(1-e^{-2})$$

5. What is P(X<Y)?

$$\int_0^\infty \int_0^y 2e^{-x}e^{-2y}dxdy = \int_0^\infty 2e^{-2y}[-e^{-x}]_0^y dy = \int_0^1 2e^{-2y}(1-e^{-y})dy$$

$$= \int_0^\infty 2e^{-2y}dy - \int_0^\infty 2e^{-3y}dy = [-e^{-2y}]_0^\infty - [-\frac{2}{3}e^{-3y}]_0^\infty = 1-\frac{2}{3} = \frac{1}{3}$$

The conditional probability density is defined as

$$f(y \mid x) = \frac{f(x,y)}{f(x)}$$

Example. Suppose that $f(x,y) = e^{-y}, \quad 0 < x < y < \infty$ and 0 otherwise.

What is $f(y \mid x)$?

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{x}^{\infty} f(x,y)dy = \int_{x}^{\infty} e^{-y}dy$$

$$\therefore f(x) = -[e^{-y}]_{x}^{\infty} = -(0 - e^{-x}) = e^{-x}$$

$$\therefore f(y \mid x) = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)} \text{ if } y > x \text{ and 0 otherwise}$$

$$f(y \mid X \varepsilon A) = \frac{\int_{x \varepsilon A} f(x, y) dx}{P(X \varepsilon A)}$$

$$f(x \mid X \varepsilon A) = \frac{f(x) 1(x \varepsilon A)}{P(X \varepsilon A)}$$

[1(.) is the indicator function, 1 iff the argument is true]

Example 1. X is uniform in the unit interval. What's the density of X conditional on $X>1/2$?

$$f(x)=1 \Rightarrow f(x \mid X>1/2)=\begin{cases} 2 & \text{if } X>1/2 \\ 0 & \text{otherwise} \end{cases}$$

Example 2. X is binomial with parameters 5 and p. What's the density of X conditional on X>3?

$$P(X=4)=\frac{5!}{1!4!}0.5^4(1-0.5)^1=0.156$$

$$P(X=5)=\frac{5!}{0!5!}0.5^4(1-0.5)^1=0.031$$

$$\therefore P(X>3)=0.187$$

$$P(X=4\mid X>3)=\frac{0.156}{0.187}=\frac{5}{6}$$

$$P(X=5\mid X>3)=\frac{1}{6}$$

Bayes Rule applies to random variables (of course)

$$f(x \mid y) = \frac{f(y \mid x) f(x)}{f(y)}$$ (and similarly for discrete random variables).

Example. Suppose that $f(x, y) = e^{-y}, \quad 0 < x < y < \infty$ and 0 otherwise.

What is $f(x \mid y)$?

We saw earlier that

$$f(x) = e^{-x}$$

$$f(y \mid x) = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)} \text{ if } y > x \text{ and 0 otherwise}$$

$$f(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_{0}^{y} f(x,y)dx = \int_{0}^{y} e^{-y}dx = ye^{-y}$$

$$\therefore f(x \mid y) = \frac{e^{-x}e^{-(y-x)}1(y > x)}{ye^{-y}} = \frac{e^{-y}1(y > x)}{ye^{-y}}$$

So $f(x \mid y) = 1/y$ if $x > y$ and 0 otherwise.

The pattern…..

For events, discrete random variables or continuous random variables

- Conditional = Joint/Marginal

- To get the marginal density you add up (or integrate) the joint density

- Bayes Rule

Conditional expectation

Discrete case: $E(g(Y) \mid x) = \Sigma g(y) f(y \mid x)$

Continuous case: $E(g(Y) \mid x) = \int_{-\infty}^{\infty} g(y) f(y \mid x) dy$

Example. Suppose that $f(x,y) = e^{-y}, \quad 0 < x < y < \infty$ and 0 otherwise. What is $E(Y \mid x)$?

$$E(Y \mid x) = \int_{-\infty}^{\infty} y f(y \mid x) dy = \int_{x}^{\infty} y e^{-(y-x)} dy$$

Use integration by parts

$$\int_a^b f(y) g'(y) dy = [f(y) g(y)]_a^b - \int_a^b f'(y) g(y) dy$$

$$\int_x^{\infty} y e^{-(y-x)} dy = [-y e^{-(y-x)}]_x^{\infty} - \int_x^{\infty} -e^{-(y-x)} dy$$

$$= x + \int_x^{\infty} e^{-(y-x)} dy$$

$$= x + [-e^{-(y-x)}]_x^{\infty}$$

$$= x + 1$$

$$\therefore E(Y \mid x) = 1 + x$$

Two discrete random variables X and Y are said to be **independent** if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Two continuous random variables X and Y are said to be **independent** if

$$f(x, y) = f(x) f(y)$$

If X and Y are independent then $f(y \mid x) = f(y)$ and $f(x \mid y) = f(x)$

Example. Consider the discrete bivariate random vector with joint probability mass function

	Y=1	Y=2	Y=3	
X=10	0.1	0.2	0.3	0.5
X=20	0.1	0.1	0.2	0.5
	0.2	0.3	0.5	

Are X and Y independent?

$P(X=10,Y=3)=0.3\neq 0.5*0.5$

No, they are not independent.

Example: Consider the continuous random variables X and Y with joint pdf

$f(x,y) = 1$ for $0 \le x \le 1$ and $0 \le y \le 1$

Are X and Y independent?

$$f(x) = \int_0^1 f(x,y)dy = \int_0^1 1dy = [y]_0^1 = 1$$

$$f(y) = \int_0^1 f(x,y)dx = \int_0^1 1dx = [x]_0^1 = 1$$

$$\therefore f(x,y) = f(x)f(y)$$

*Y*es, they are independent

If X and Y are independent, then $E(XY) = E(X)E(Y)$

Proof: (continuous random variables)

$$E(XY) = \int\int xyf(x,y)dxdy = \int\int xyf(x)f(y)dxdy$$
$$= \int xf(x)\int yf(y)dydx$$
$$= \int xf(x)dx\int yf(y)dy$$
$$= E(X)E(Y)$$

If $X_1, X_2, ...X_n$ are independent random variables with pdfs $f_1, f_2, ... f_n$, then the joint pdf of $X_1, X_2, ...X_n$ is

$$f(x) = \Pi_{i=1}^{n} f_i(x)$$

If $X_1, X_2, ...X_n$ are independent random variables with the **same** pdf $\overline{f}$, then the joint pdf of $X_1, X_2, ...X_n$ is

$$f(x) = \overline{f}(x)^n$$

These are said to be independently and identically distributed (iid).

Covariance

The covariance between two random variables X and Y is

$$Cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

Note: $Cov(X,X) = E((X - E(X))^2) = Var(X)$

Useful rules on variances and covariances

- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- $Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y)$
- $Cov(X + Y, Z) = Cov(X,Z) + Cov(Y,Z)$

Correlation

The correlation between two random variables X and Y is

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

which is between -1 and +1 (we'll prove this later..)

Covariance and independence

If X and Y are independent, then their covariance (and correlation) is zero. They are "uncorrelated".

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

But the converse is not true. Counterexample.
X=-1, 0, 1: each with probability 1/3
$Y = X^2$

The joint density of X and Y is

	X=-1	X=0	X=1	
Y=0	0	1/3	0	1/3
Y=1	1/3	0	1/3	2/3
	1/3	1/3	1/3	

Clearly X and Y are NOT independent. But

$$E(XY) = \frac{1}{3} * (-1 * 1) + \frac{1}{3} * (0 * 0) + \frac{1}{3} * (1 * 1) = 0$$

$$E(X) = \frac{1}{3} * (-1) + \frac{1}{3} * 0 + \frac{1}{3} * 1 = 0. \quad E(Y) = \frac{1}{3} * 0 + \frac{2}{3} * 1 = \frac{2}{3}$$

$$\therefore Cov(X,Y) = E(XY) - E(X)E(Y) = 0$$

X and Y are uncorrelated

More results on independent random variables

1. If X and Y are independent, g is a function only of X and h is a function only of Y then g(X) and g(Y) are independent

2. If X and Y have a joint probability density $f(x,y)$ then they are independent if and only if there exist functions $g(x)$ and $h(y)$ such that $f(x,y) = g(x)h(y)$. If these satisfy the conditions for pdfs, then they are the marginal pdfs of X and Y.

3. If $X_1, X_2, ...X_n$ are mutually independent, then
$E(g(X_1)g(X_2)...g(X_n)) = E(g(X_1))E(g(X_2))...E(g(X_n))$

4. If $X_1, X_2, ...X_n$ are mutually independent random variables with moment generating functions $M_{X_1}(t), M_{X_2}(t), ..M_{X_n}(t)$, respectively and $Z = \Sigma_{i=1}^{n} X_i$, the moment generating function of Z is

$$M_Z(t) = M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t)$$

Proof of result 3:

$$E(g(X_1)g(X_2)...g(X_n))$$

$$= \int\int..\int g(x_1)g(x_2)...g(x_n)\,f(x_1,x_2,....x_n)dx_1dx_2..dx_n$$

$$= \int\int..\int g(x_1)g(x_2)...g(x_n)\,f(x_1)\,f(x_2)...\,f(x_n)dx_1dx_2..dx_n$$

$$= \int g(x_1)\,f(x_1)dx_1 \int g(x_2)\,f(x_2)dx_2...\int g(x_n)\,f(x_n)dx_n$$

$$= E(g(X_1))E(g(X_2))...E(g(X_n))$$

Proof of result 4:

$$M_Z(t) = E(e^{Zt}) = E(e^{(X_1+X_2..+X_n)t}) = E(e^{X_1t}e^{X_2t}...e^{X_nt})$$

$$\therefore M_Z(t) = E(e^{X_1t})E(e^{X_2t})...E(e^{X_nt})$$

$$\therefore M_Z(t) = M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t)$$

Moment generating function of independent random vars

Result 4: If $X_1,....X_n$ are independent rvs with mgf $m_1(t),....m_n(t)$, then the mgf of $\Sigma_{i=1}^{n} X_i$ is $\Pi_{i=1}^{n} m_i(t)$

If $X_1,....X_n$ are independent rvs all with the same mgf $m(t)$, then the mgf of $\Sigma_{i=1}^{n} X_i$ is $m(t)^n$

If $X_1, X_2, ...X_n$ are independent Bernoulli random variables then $\Sigma_{i=1}^{n} X_i$ is Binomial with parameters n and p.

The mgf of a Bernoulli is $E(e^{tX}) = pe^t + (1-p)$

The mgf of $\Sigma_{i=1}^{n} X_i$ is $[pe^t + (1-p)]^n$ and this is the Binomial mgf.

More practice with multiple random variables

Example 1. (X,Y) are uniformly distributed on the unit interval such that

$f(x,y)=1,\ 0\le x\le 1$ and $0\le y\le 1$

Q. What is $f(x)$?

$f(x)=\int_0^1 f(x,y)dy=\int_0^1 1dy=[y]_0^1=1$

Q. What is $f(x,y\,|\,x\ge 1/2)$?

$$f(x,y\,|\,x\ge 1/2)=\frac{f(x,y)}{P(x\ge 1/2)} \text{ for } X\ge 1/2$$

$$0 \qquad \text{otherwise}$$

$\therefore\ f(x, y \mid x \geq 1/2) = 2$ for $X \geq 1/2$

Q. What is $f(y \mid x \geq 1/2)$?

$$f(y \mid x \geq 1/2) = \int_{1/2}^{1} f(x, y \mid x \geq 1/2)dx = \int_{1/2}^{1} 2dx = [2x]_{1/2}^{1} = 1$$

Q. What is $P(X \geq Y)$?

$$P(X \geq Y) = \int_0^1 \int_y^1 f(x, y)dxdy = \int_0^1 \int_y^1 1dxdy$$

$$\therefore P(X \geq Y) = \int_0^1 [x]_y^1 dy = \int_0^1 (1 - y)dy = [y - \frac{y^2}{2}]_0^1$$

$$\therefore P(X \geq Y) = 1 - \frac{1}{2} = \frac{1}{2}$$

Q. What is $f(x, y \mid x \geq y)$?

$$f(x, y \mid x \geq y) = \frac{f(x, y)}{P(X \geq Y)} \quad \text{for } X \geq Y$$

$$0 \qquad \text{otherwise}$$

$$\therefore f(x, y \mid x \geq y) = 2 \quad \text{for } X \geq Y$$

Q. What is $f(y \mid x \geq y)$?

$$f(y \mid x \geq y) = \int_y^1 f(x, y \mid x \geq y) dx = \int_y^1 2 dx = [2y]_y^1 = 2(1 - y)$$

Example 2: (X,Y) are uniformly distributed with support
$\{(x,y): 0 < x < 2\theta, 0 < y \le 1 - x/2\theta\}$
$f(x,y) = c$

Q. What is c?

$$\int_0^{2\theta}\int_0^{1-x/2\theta} c\,dy\,dx = \int_0^{2\theta}[cy]_0^{1-x/2\theta}\,dx = \int_0^{2\theta} c(1-\frac{x}{2\theta})dx$$

$$= [cx]_0^{2\theta} - \frac{c}{2\theta}[\frac{x^2}{2}]_0^{2\theta} = 2\theta c - \frac{c}{2\theta}\frac{4\theta^2}{2} = 2\theta c - \theta c = \theta c$$

Since this must integrate to 1, $\theta c = 1 \Rightarrow c = 1/\theta$

Q. What is f(*X*)?

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{0}^{1-\frac{x}{2\theta}} \frac{1}{\theta} dy = \frac{1}{\theta}[y]_{0}^{1-\frac{x}{2\theta}} = \frac{1}{\theta}(1-\frac{x}{2\theta})$$

Q. What is E(*X*)?

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{2\theta} x\frac{1}{\theta}(1-\frac{x}{2\theta})dx = \frac{1}{\theta}\int_{0}^{2\theta} xdx - \frac{1}{2\theta^2}\int_{0}^{2\theta} x^2 dx$$

$$= \frac{1}{\theta}[\frac{x^2}{2}]_{0}^{2\theta} - \frac{1}{2\theta^2}[\frac{x^3}{3}]_{0}^{2\theta} = \frac{1}{\theta}\frac{4\theta^2}{2} - \frac{1}{2\theta^2}\frac{8\theta^3}{3} = 2\theta - \frac{8}{6}\theta = \frac{2}{3}\theta$$

Example 3: Two bidders participate in an auction for a white elephant. Each bidder has the same underlying valuation for the elephant, given by the *same* random variable $V \sim U[0,1]$.

Neither bidder knows V. Each bidder receives an independent signal about V: $X_i \mid V \sim U[0,V]$

Assume each bidder submits a bid equal to her conditional expectation: for bidder 1, this is $E(V \mid X_1)$. How much does she bid?

Bayes Rule: $f(v \mid x_1) = \dfrac{f(x_1 \mid v) f(v)}{f(x_1)}$

$f(x_1 \mid v) = 1/v \qquad \text{for } 0 < x_1 < v$

$f(v) = 1 \qquad \text{for } 0 < v < 1$

$f(x_1, v) = \dfrac{1}{v} \qquad \text{for } x_1 < v < 1$

$$\therefore f(x_1) = \int_{x_1}^{1} f(x_1, v)\,dv = \int_{x_1}^{1} v^{-1}\,dv = \ln(1) - \ln(x_1) = -\ln(x_1)$$

$$\therefore f(v \mid x_1) = -\frac{v^{-1}}{\ln(x_1)} = -\frac{1}{v \ln(x_1)}$$

$$\therefore E(V \mid x_1) = \int_{x_1}^{1} -\frac{1}{v \ln(x_1)} v\,dv = -\frac{1}{\ln(x_1)} \int_{x_1}^{1} dv = -\frac{1 - x_1}{\ln(x_1)} = \frac{x_1 - 1}{\ln(x_1)}$$

Specific Continuous Distributions

- Uniform (already)
- Normal
- Log-Normal
- Beta
- Truncated Normal (now)
- Exponential
- Gamma
- Chi-Squared (later in the class)
- T-distribution
- Cauchy
- F-distribution

The Truncated Normal Distribution

This can be useful in empirical micro. Suppose that $X \sim N(\mu, \sigma^2)$ but we condition on $a < X < b$. Then the conditional density of X is truncated normal with the density

$$f(x) = \begin{cases} \dfrac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} & \text{if } a < X < b \\ 0 & \text{otherwise} \end{cases}$$

which follows from the definition of conditional probability.

There are closed from expressions for the mean and variance of a truncated normal. For example,

$$E(X \mid a < X < b) = \mu + \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\sigma$$

$$E(X \mid X > a) = \mu + \frac{\phi(\frac{a-\mu}{\sigma})}{1 - \Phi(\frac{a-\mu}{\sigma})}\sigma$$

(special case of the first with $b = \infty$)

Example. An individuals desired consumption of tobacco is $X = N(0,1)$. But no individual can consume a negative quantity of tobacco. So the observed tobacco consumption is

$Y = X$ if $X > 0$

0 otherwise

What is $E(Y)$?

$$E(Y) = E(Y \mid X < 0)P(X < 0) + E(Y \mid X > 0) * P(X > 0)$$

$$\therefore E(Y) = E(X \mid X > 0) * 1/2$$

$$E(X \mid X > 0) = \frac{\phi(0)}{1 - \Phi(0)} = \frac{\phi(0)}{1/2} = 2\phi(0)$$

$$\therefore \boxed{E(Y) = \phi(0) = \frac{1}{\sqrt{2\pi}}} \quad \text{(remember: } \phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})\text{)}$$

The Exponential Distribution

Useful for modeling waiting times (times between events)

cdf: $F(x) = 1 - e^{-x/\lambda} \quad 0 < x < \infty$

pdf: $f(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 < x < \infty$

$E(X) = \lambda$

$Var(X) = \lambda^2$

Example: The time between incoming phone calls in a call center is exponential with parameter 3 (in minutes).

A call comes in at 14:00. What is the probability that the next call will not come until AFTER 14:02?

The probability that it comes in LESS than 2 minutes is

$$F(2) = 1 - e^{-2/3} = 0.487$$

The answer is 0.513

If X has an exponential distribution then if $s > t$

$$P(X > s \mid X > t) = P(X > s - t)$$

This property is called "memoryless"

Proof: $P(X > s \mid X > t) = \dfrac{P(X > s, X > t)}{P(X > t)}$

$$= \frac{P(X > s)}{P(X > t)} = \frac{\int_s^\infty \lambda^{-1} \exp(-x/\lambda) dx}{\int_t^\infty \lambda^{-1} \exp(-x/\lambda) dx} = \frac{[-e^{-x/\lambda}]_s^\infty}{[-e^{-x/\lambda}]_t^\infty} == \frac{e^{-s/\lambda}}{e^{-t/\lambda}} = e^{(t-s)/\lambda}$$

$$P(X > s - t) = \int_{s-t}^\infty \lambda^{-1} \exp(-x/\lambda) dx = [-e^{-x/\lambda}]_{s-t}^\infty = e^{-(s-t)/\lambda} = e^{(t-s)/\lambda}$$

$$\therefore P(X > s \mid X > t) = P(X > s - t)$$

The gamma distribution

Memorylessness is often unappealing. The gamma distribution also has support from 0 to ∞ but is more general, and not memoryless. It has two parameters: α and β.

The pdf is $f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$

If $\alpha = 1$, then this reduces to an exponential.

Distribution of the max and the min

Suppose that $X_1, X_2, ...X_n$ are iid random variables with a probability density $f(x)$ and cdf $F(x)$. What's the distribution of the max and the min of these random variables?

The cdf of the maximum is

$$F_{\max}(x) = P(\max(\{X_i\}) \le x) = \Pi_{i=1}^{n} P(X_i \le x) = F(x)^n$$

The corresponding pdf is

$$f_{\max}(x) = nF(x)^{n-1} f(x)$$

The cdf of the minimum is

$$F_{\min}(x) = P(\min(\{X_i\}) \le x) = 1 - \Pi_{i=1}^{n} P(X_i > x) = 1 - (1 - F(x))^n$$

The corresponding pdf is

$$f_{\min}(x) = n(1 - F(x))^{n-1} f(x)$$

Example. If I draw n uniform random numbers on the unit interval, how many do I have to draw to have a 99 percent chance that at least one is above 0.8.

$$F(x) = x$$
$$\therefore F_{\max}(x) = x^n \qquad \text{(cdf of the max of } n \text{ uniforms)}$$
$$\therefore P(\max\{X_i\}_{i=1}^{n} > x) = 1 - x^n$$
$$\therefore P(\max\{X_i\}_{i=1}^{n} > 0.8) = 1 - 0.8^n$$

I need to solve for the smallest n such that $1 - 0.8^n > 0.99$

The answer is 21

Distribution of order statistics.

Continue with this problem. Let the variables be ordered as $X_{(1)} < X_{(2)} ... < X_{(n)}$. What is the cdf and pdf of $X_{(n-1)}$?

$P(X_{(n-1)} \le x) = nF(x)^{n-1} - (n-1)F(x)^n$ (Convince yourself)

The pdf of $X_{(n-1)}$ is

$n(n-1)F(x)^{n-2} f(x) - (n-1)nF(x)^{n-1} f(x)$

$= n(n-1)F(x)^{n-2} f(x)(1-F(x))$

In general, the pdf for $X_{(j)}$ (the jth smallest statistic) is

$$\frac{n!}{(j-1)!(n-j)!} f(x)F(x)^{j-1}(1-F(x))^{n-j}$$

Example. If I draw n uniform random numbers on the unit interval, how many do I have to draw to have a 99 percent chance that at least two are above 0.8.

$$P(X_{(n-1)} \le x) = nF(x)^{n-1} - (n-1)F(x)^n$$
$$\therefore P(X_{(n-1)} \le x) = nx^{n-1} - (n-1)x^n$$
$$\therefore P(X_{(n-1)} > x) = 1 - \{nx^{n-1} - (n-1)x^n\}$$
$$\therefore P(X_{(n-1)} > 0.8) = 1 - \{n0.8^{n-1} - (n-1)0.8^n\}$$

The smallest n that satisfies this is 31.

Variance-Covariance Matrix

Suppose that we have *n* random variables: $X_1, X_2, ...X_n$

The variance-covariance matrix of these is as follows

$$\begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & & \\ \vdots & & \ddots & \vdots \\ Cov(X_n, X_1) & & \dots & Var(X_n) \end{pmatrix}$$

It is square and symmetric and positive definite.

Example: X and Y are both normal with mean zero and variances 1 and 4, respectively. Their correlation is 0.3. What is the variance-covariance matrix of these two random variables?

Answer: $\begin{pmatrix} 1 & 0.6 \\ 0.6 & 4 \end{pmatrix}$ (Cov(X,Y)=2*1*0.3=0.6)

The correlation matrix is

$$\begin{pmatrix} 1 & Corr(X_1, X_2) & \cdots & Corr(X_1, X_n) \\ Corr(X_2, X_1) & 1 & & \\ \vdots & & \ddots & \vdots \\ Corr(X_n, X_1) & & \cdots & 1 \end{pmatrix}$$

More properties of expectations and variances

Suppose that the nx1 vector $X = (X_1, X_2, ...X_n)'$ has mean μ and variance-covariance matrix Σ. Then

$$\boxed{E(a + b'X) = a + b'\mu}$$

$$\boxed{Var(a + b'X) = b'\Sigma b}$$

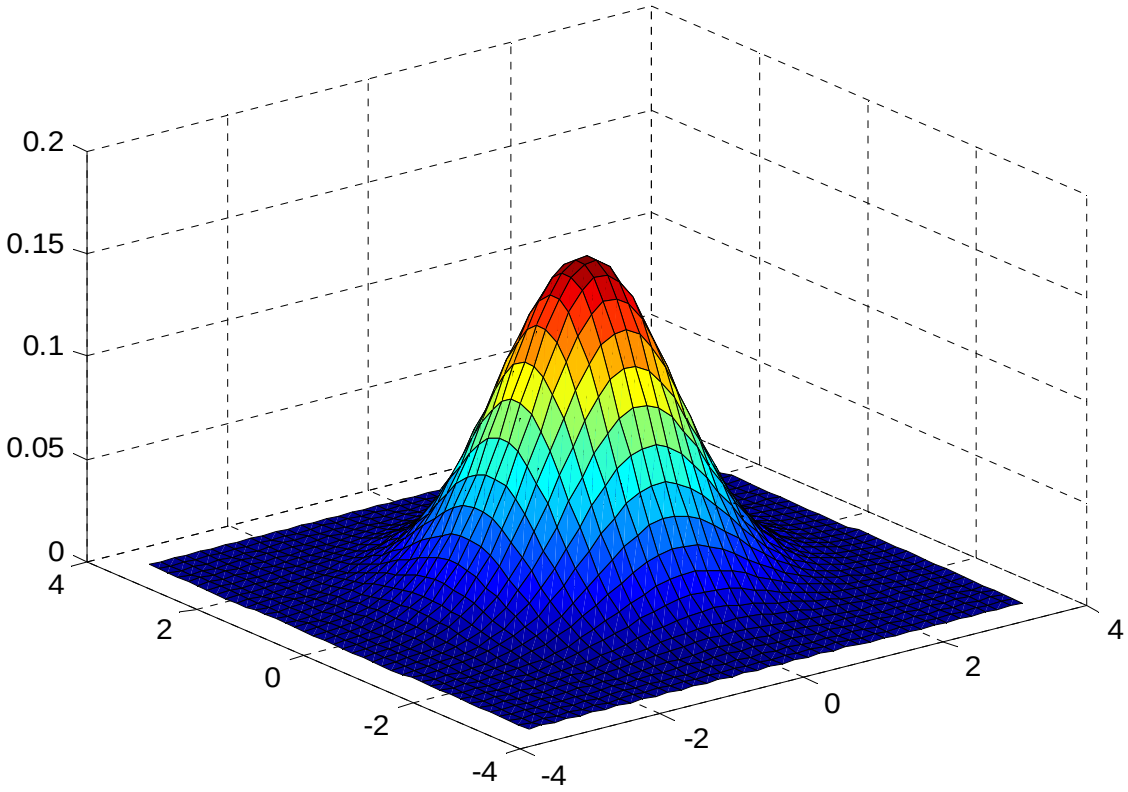
The bivariate normal density

Suppose that X_1 is $N(\mu_1, \sigma_1^2)$ and X_2 is $N(\mu_2, \sigma_2^2)$ and the two random variables have a correlation of ρ. Then the vector $X = (X_1, X_2)'$ has a bivariate normal distribution with pdf

$$f(x) = (2\pi)^{-1} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu))$$

where $\mu = (\mu_1, \mu_2)'$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ is the variance-covariance matrix of X_1 and X_2.

Graph of bivariate normal pdf



Properties of the bivariate normal density

- If X_1 and X_2 are uncorrelated then they are independent.

- $aX_1 + bX_2 + c \sim N(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2)$

The conditional normal density

The distribution of X_2 conditional on $X_1 = x_1$ is

$$N\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2)\right)$$

Proof: Let $Z = \frac{1}{\sigma_2}X_2 - \frac{\rho}{\sigma_1}X_1$

$$Cov(Z, X_1) = \frac{1}{\sigma_2}Cov(X_1, X_2) - \frac{\rho}{\sigma_1}Var(X_1)$$

$$= \frac{\rho\sigma_1\sigma_2}{\sigma_2} - \frac{\rho\sigma_1^2}{\sigma_1} = \rho\sigma_1 - \rho\sigma_1 = 0$$

Z and X_1 are uncorrelated and so independent (by property 1).

From property 2, Z is $N(\frac{\mu_2}{\sigma_2} - \frac{\rho\mu_1}{\sigma_1}, 1-\rho^2)$

As Z and X_1 are independent the distribution of Z conditional on $X_1 = x_1$ is the same as the unconditional distribution.

Rearranging the definition of Z, $X_2 = \frac{\rho\sigma_2}{\sigma_1} X_1 + \sigma_2 Z$

So the distribution of X_2 conditional on $X_1 = x_1$ is

$$N(\mu_2 + \frac{\rho\sigma_2 x_1}{\sigma_1} - \frac{\rho\mu_1\sigma_2}{\sigma_1}, \sigma_2^2(1-\rho^2))$$

$$= N(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2))$$

Example: Suppose that annual US and Canadian stock returns are both normally distributed with $\mu = 8$ and $\sigma = 16$. Suppose that the correlation between them is 0.7. If in a given year, U.S. stock returns are 10 percent, what is the probability that Canadian returns are at least 10 percent?

Let X_1 and X_2 be US and Canadian returns. The conditional distribution of X_2 given that $X_1 = 10$ is:

$$N(8 + \frac{0.7 * 16}{16}(10 - 8), 16^2(1 - 0.7^2)) = N(9.4, 130.56)$$

So the probability that Canadian returns are at least 10 percent is 0.479.

How to simulate correlated normal random variables

Suppose we want to generate standard normal random variables X and Y with correlation ρ (common situation).

- Let X be standard normal.
- Let U be standard normal (independent of X).
- Let $Y = \rho X + \sqrt{1-\rho^2}U$

$E(Y) = 0,\ Var(Y) = \rho^2 + 1 - \rho^2 = 1,$
$Cov(X,Y) = E(XY) = \rho Var(X) = \rho$

So X and Y are standard normal with correlation ρ

Matlab code

Generate random variables Y and X with correlation 0.9.

```
x=randn(100000,1);
a1=0.9; a2=sqrt(0.19);
y=(a1*x)+(a2*randn(100000,1));
mean(x.*y)
```

Gibbs sampling

General method for simulating from joint distn of 2 rvs

Suppose I can't draw directly from the distn of Y and X.
I can draw from the distribution of Y|X and X|Y

Algorithm. Pick an X.
1. Take a draw from Y given X.
2. Take a draw from X given Y.
3. Repeat 1 and 2 many times.

The draws from 1 and 2 after discarding an initial "burnin" are draws from the joint distribution of Y and X.

Gibbs sampling example.

Suppose that $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}]$ but we cannot draw from the joint distribution. We know that

$X \mid Y = y \sim N(\rho y, 1-\rho^2)$

$Y \mid X = x \sim N(\rho y, 1-\rho^2)$

Then we can use the Gibbs sampler.

The Gibbs sampler is widely used in Bayesian methods (more on this later).

The multivariate normal density

Suppose that X is an nx1 vector with pdf

$$f(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu))$$

where μ is an nx1 vector and Σ is an nxn variance-covariance matrix then $X \sim N(\mu, \Sigma)$

Properties of the multivariate normal density

- If two elements of X are uncorrelated, then they are independent
- $a + b'X \sim N(a + b'\mu, b'\Sigma b)$

The conditional normal density

Suppose that $X \sim N(\mu, \Sigma)$ where $X = (X_1', X_2')'$, $\mu = (\mu_1', \mu_2')'$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then the distribution of X_2 conditional on $X_1 = x$ is

$$N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

Multivariate Transformation Formula

Let X and Y have a joint pdf $f_{X,Y}(x,y)$ and U and V be random variables such that $U = g_1(X,Y)$ and $V = g_2(X,Y)$ is a one-to-one transformation.

Define inverse transforms $X = h_1(U,V)$ and $Y = h_2(U,V)$

Let J denote the determinant of the matrix

$$\begin{pmatrix} \dfrac{\partial h_1(u,v)}{\partial u} & \dfrac{\partial h_1(u,v)}{\partial v} \\ \dfrac{\partial h_2(u,v)}{\partial u} & \dfrac{\partial h_2(u,v)}{\partial v} \end{pmatrix}$$

Then the density of U and V is

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v),h_2(u,v))\,|J|$$

Example 1. $f(x,y)=2,\ \ 0<x<1, 0<y<1, x+y<1$ and zero otherwise. Let $U=X+Y$ and $V=X-Y$. What is the joint density of U and V?

$X=(U+V)/2$ and $Y=(U-V)/2$

$$\begin{pmatrix} \partial X/\partial U & \partial X/\partial V \\ \partial Y/\partial U & \partial Y/\partial V \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix} \Rightarrow |J| = 1/2$$

$$\therefore f_{U,V}(u,v) = f_{X,Y}\left(\frac{1}{2}(u+v), \frac{1}{2}(u-v)\right) * \frac{1}{2} = 2 * \frac{1}{2} = 1$$

Support: $U \in [0,1],\ V \in [-1,1],\ U > V$

Example 2. Suppose that X and Y are independent standard normal random variables, so that

$$f(x,y)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}=\frac{1}{2\pi}\exp(-\frac{x^2+y^2}{2})$$

Let $U=X+Y$ and $V=X-Y$. What is the joint density of U and V?

$$X=(U+V)/2 \text{ and } Y=(U-V)/2$$

$$\begin{pmatrix}\partial X/\partial U & \partial X/\partial V\\ \partial Y/\partial U & \partial Y/\partial V\end{pmatrix}=\begin{pmatrix}1/2 & 1/2\\ 1/2 & -1/2\end{pmatrix}\Rightarrow|J|=1/2$$

$$\therefore f_{U,V}(u,v)=\frac{1}{2\pi}\exp(-\frac{(u+v)^2/4+(u-v)^2/4}{2})\frac{1}{2}$$

$$\therefore\ f_{U,V}(u,v) = \frac{1}{4\pi}\exp\left(-\frac{(u+v)^2+(u-v)^2}{8}\right)$$

$$\therefore\ f_{U,V}(u,v) = \frac{1}{4\pi}\exp\left(-\frac{u^2+v^2+2uv+u^2+v^2-2uv}{8}\right)$$

$$= \frac{1}{4\pi}\exp\left(-\frac{u^2+v^2}{4}\right)$$

$$\therefore\ f_{U,V}(u,v) = \frac{1}{\sqrt{4\pi}}e^{-u^2/4}\,\frac{1}{\sqrt{4\pi}}e^{-v^2/4}$$

So U and V are independent with

$$f(u) = \frac{1}{\sqrt{4\pi}}e^{-u^2/4} \text{ and } \quad f(v) = \frac{1}{\sqrt{4\pi}}e^{-v^2/4}$$

The marginal densities of U and V are both $N(0,2)$.

In fact it can be shown that if X and Y are independent with a common distribution function F then U=X+Y and V=X-Y are independent **if and only if** F is normal.

Example 3. Suppose that X and Y are independent gamma random variables with params $(\alpha, 1/\lambda)$ and $(\beta, 1/\lambda)$ so that

$$f(x,y) = \frac{x^{\alpha-1}e^{-x\lambda}}{\Gamma(\alpha)(1/\lambda)^{\alpha}} \frac{x^{\beta-1}e^{-x\lambda}}{\Gamma(\beta)(1/\lambda)^{\beta}} = \frac{\lambda^{\alpha}x^{\alpha-1}e^{-x\lambda}}{\Gamma(\alpha)} \frac{\lambda^{\beta}y^{\beta-1}e^{-y\lambda}}{\Gamma(\beta)}$$

$$= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda(x+y)} x^{\alpha-1} y^{\beta-1}$$

Let $U = X + Y$ and $V = X/(X+Y)$. What is the joint density of U and V?

$X = UV$ and $Y = U(1-V)$

$$\begin{pmatrix} \partial X/\partial U & \partial X/\partial V \\ \partial Y/\partial U & \partial Y/\partial V \end{pmatrix} = \begin{pmatrix} V & U \\ 1-V & -U \end{pmatrix} \Rightarrow |J| = U$$

$$\therefore f(u,v) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda(uv+u(1-v))} (uv)^{\alpha-1} (u(1-v))^{\beta-1} u$$

$$= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda u} u^{\alpha+\beta-1} v^{\alpha-1} (1-v)^{\beta-1}$$

$$\therefore f(u,v) = \frac{\lambda e^{-\lambda u} (\lambda u)^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha-1} (1-v)^{\beta-1}$$

So U and V are independent where U has a gamma distribution with parameters $(\alpha+\beta, 1/\lambda)$ and V has a beta distribution with parameters (α, β).

Law of Iterated Expectations

Useful result

If X and Y are any two random variables then
$E(E(X \mid Y)) = E(X)$

Let's check this in one case: the bivariate normal

$$E(X_2 \mid X_1 = x_1) = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

$$E(X_2) = E(E(X_2 \mid X_1 = x_1)) = E(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1 - \mu_1)) = \mu_2$$

Works in this case……

More substantive example.

Suppose that X is binomial with parameters n and p and p is beta-distributed with parameters α and β ("hierarchical" model).

$$E(X \mid P) = np$$

$$E(X) = E(E(X \mid p)) = E(np) = nE(p) = n\frac{\alpha}{\alpha + \beta}$$

Conditional variance identity

If X and Y are any two random variables then

$$Var(X) = E(\text{var}(X \mid Y)) + Var(E(X \mid Y))$$

Stein's Lemma

If X is $N(\theta,\sigma^2)$ and g is any function that satisfies $E(g'(X))<\infty$ then $E((g(X)(X-\theta))=\sigma^2 E(g'(X))$

Proof:

$$E((g(X)(X-\theta))=\int_{-\infty}^{\infty} g(x)(x-\theta)\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\theta)^2}{2\sigma^2})dx$$

$$=[-g(x)\sigma^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\theta)^2}{2\sigma^2})]_{-\infty}^{\infty}$$

$$-\int_{-\infty}^{\infty}-\sigma^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\theta)^2}{2\sigma^2})g'(x)dx$$

$$= [-g(x)\sigma^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\theta)^2}{2\sigma^2})]_{-\infty}^{\infty}$$

$$+\sigma^2 \int_{-\infty}^{\infty} g'(x) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\theta)^2}{2\sigma^2}) dx$$

$$= \sigma^2 E(g'(X))$$

Useful inequalities

- Cauchy-Schwarz Inequality

Suppose X and Y are two random variables

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$$

Let $X = U - E(U)$ and $Y = V - E(V)$

$$|E((U-E(U))(V-E(V))| \leq \sqrt{E((U-E(U))^2)E((V-E(V))^2)}$$

$$\therefore |Cov(U,V)| \leq \sqrt{Var(U)Var(V)}$$

This implies that correlations must be between -1 and +1

- Hölder's Inequality

Suppose X and Y are two random variables

$$|E(XY)| \leq E(X^p)^{1/p} E(Y^q)^{1/q}$$

where $\frac{1}{p} + \frac{1}{q} = 1.$

Cauchy-Schwarz is a special case with p=q=2.

- Liapounov's Inequality

Suppose X is a random variable

$$\{E(|X|^r)\}^{1/r} \leq \{E(|X|^s)\}^{1/s}, \quad 1 < r < s$$

- Minkowski's Inequality

Suppose X and Y are two random variables

$$E(|X+Y|^p)^{1/p} \le E(|X|^p)^{1/p} + E(|Y|^p)^{1/p},\ p \ge 1$$

- Covariance Inequality

Let X be any random variable and g and h are functions such that E(g(X)), E(h(X)) and E(g(X)h(X)) exist. Then

- If g is nondecreasing and h is nonincreasing

$$E(g(X)h(X)) \le E(g(X))E(h(X))$$

- If g and h are both nondecreasing or both nonincreasing

$$E(g(X)h(X)) \ge E(g(X))E(h(X))$$

Stochastic Processes

A **stochastic process** $\{X_t\}_{t=1}^{T}$ is a collection of random variables, where the index t refers to time.

Example. Consider flipping a coin forever. Let $X_t = 1$ if the tth toss is heads and 2 otherwise. This defines a stochastic process where X_s is independent of X_r for $r \neq s$.

The index t can be continuous (continuous time) or discrete (discrete time).

Consider a stochastic process in discrete time that takes on a finite number of possible values (e.g. 1 and 2).

If

$$P(X_{t+1} = j \mid X_t = i, X_{t-1}, \ldots X_1) = P(X_{t+1} = j \mid X_t = i) \forall t = 1, 2, ..$$

then the stochastic process is said to be a **Markov Chain**.

"Transition probabilities": $p_{ij} = P(X_{t+1} = j \mid X_t = i)$

"Transition matrix": $P = \begin{pmatrix} p_{11} & .. & p_{k1} \\ : & & \\ p_{1k} & .. & p_{kk} \end{pmatrix}$

Transition probabilities have the property that $\Sigma_j p_{ij} = 1$

In other words, each column of P adds up to 1.

Recall: If $A_1, A_2, \ldots A_h$ is a mutually exclusive and collectively exhaustive set of events

$$P(B) = P(B \cap A_1) + P(B \cap A_2) \ldots + P(B \cap A_h)$$

$$\therefore P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) \ldots + P(B \mid A_h)P(A_h)$$

Hence

$$P(X_{t+1} = j) = \Sigma_{i=1}^{k} P(X_{t+1} = j \mid X_t = i) * P(X_t = i)$$

Hence

If $P_t^X = (P(X_t = 1), P(X_t = 2), \ldots P(X_t = k))'$ then

$$P_{t+1}^X = P * P_t^X$$

Let $P^{(n)}$ denote the n-step transition matrix, the ijth element of which is

$P(X_{t+n} = j \mid X_t = i)$

Then $P^{(n)} = P^n$

Under some conditions

$\lim_{n\to\infty} P(X_{t+n} = j \mid X_t = i) = \pi_j$

exists and is independent of i. If $\pi = (\pi_1, \ldots \pi_k)'$, then it solves the equation $(P - I)\pi = 0$

This means that π is the unit-length eigenvector of P corresponding to the unit eigenvalue.

π gives the "steady state" probabilities of the Markov-Chain. π_j is the "proportion of time" that the process is in state j.

Example: The coin-flip process is a Markov chain with

$$P = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

Example. Suppose that the economy can be in two states: expansion and recession with transition matrix

$$\begin{pmatrix} 0.9 & 0.4 \\ 0.1 & 0.6 \end{pmatrix}$$

Given that the economy is in expansion in quarter t, what is the probability that it will be in expansion in quarter $t+2$?

Given that the economy is in recession in quarter t, what is the probability that it will be in expansion in quarter $t+3$?

What is the steady state probability of expansion/recession?

Given that the economy is in expansion in quarter t, what is the probability that it will be in expansion in quarter $t+2$?

$$P = \begin{pmatrix} 0.9 & 0.4 \\ 0.1 & 0.6 \end{pmatrix} \Rightarrow P^2 = \begin{pmatrix} 0.85 & 0.6 \\ 0.15 & 0.4 \end{pmatrix}$$

Answer: 0.85

Given that the economy is in recession in quarter t, what is the probability that it will be in expansion in quarter $t+3$?

$$P = \begin{pmatrix} 0.9 & 0.4 \\ 0.1 & 0.6 \end{pmatrix} \Rightarrow P^3 = \begin{pmatrix} 0.825 & 0.7 \\ 0.175 & 0.3 \end{pmatrix}$$

Answer: 0.7.

What is the steady state probability of expansion/recession?

$$(P - I)\pi = 0$$

$$\begin{pmatrix} -0.1 & 0.4 \\ 0.1 & -0.4 \end{pmatrix}\begin{pmatrix} \pi_1 \\ 1-\pi_1 \end{pmatrix} = 0$$

$$-0.1\pi_1 + 0.4(1-\pi_1) = 0 \Rightarrow 0.5\pi_1 = 0.4 \Rightarrow \pi_1 = 0.8$$

80 percent of the time, the economy is in expansion
20 percent of the time, the economy is in recession

Unusual transition matrices for Markov Chains

Periodic: $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ (no “steady state”)

Absorbing: $\begin{pmatrix} 0.9 & 0 \\ 0.1 & 1 \end{pmatrix}$ (once in state 2, you never leave)

More Markov Chain Definitions.

A Markov chain is said to be **aperiodic** if $p_{ii} > 0 \ \forall i$.

A Markov chain is said to be **positive recurrent** if
$E(\inf\{n : X_{t+n} = i \mid X_t = i\}) < \infty \ \ \forall i$

A Markov chain is said to be **ergodic** if it is both aperiodic and positive recurrent.

A counting process is a stochastic process in continuous time.

Intuitively: How many events have happened by time t

Formally: A **counting process** is a stochastic process $\{N(t)\}$ where $t \in [0,\infty)$ such that

(i) $N(t) \geq 0$

(ii) $N(t)$ is integer-valued

(iii) $s < t \Rightarrow N(s) \leq N(t)$

(iv) If $s < t$ then $N(t) - N(s)$ is the number of events between s and t.

A counting process $N(t)$ is a **Poisson process** with rate $\lambda > 0$ if

(i) $N(0) = 0$

(ii) If $t_1 < t_2 \le t_3 < t_4$ then $N(t_2) - N(t_1)$ is independent of $N(t_4) - N(t_3)$

(iii) If $s < t$ then the number of events between s and t, $N(t) - N(s)$, is Poisson distributed with mean $\lambda(t-s)$. That is

$$P(N(t) - N(s) = n) = \frac{e^{-\lambda(t-s)}(\lambda(t-s))^n}{n!}$$

for $n = 0,1,2,...$

A discrete time stochastic process, X_t, is said to be **white noise** if

(i) $E(X_t) = 0 \ \forall t$

(ii) $Var(X_t) = \sigma^2 \ \forall t$

(iii) $Cov(X_s, X_t) = 0$ for $s \neq t$

A discrete time stochastic process, X_t, is said to be **Gaussian white noise** if $X_t \sim N(0, \sigma^2)$ and are independent of each other.

A discrete time stochastic process, X_t, is said to be a **random walk** if $X_t = X_{t-1} + \varepsilon_t$ where ε_t is white noise.

A discrete time stochastic process, X_t, is said to be a **martingale** if $E_{t-1}(X_t) = X_{t-1}$ where $E_{t-1}(.)$ denotes the expectation conditional on the information set at time $t-1$.

A discrete time stochastic process, X_t, is said to be a **martingale difference sequence** if $E_{t-1}(X_t) = 0$.

A random walk is a martingale, but the converse is not true.

$$X_t = X_{t-1} + \varepsilon_t \Rightarrow E(X_t \mid X_{t-1}) = X_{t-1}$$

Suppose that $X_t = X_{t-1} + \sigma_t \varepsilon_t$ where $\log(\sigma_t^2) = \log(\sigma_{t-1}^2) + u_t$ and u_t and ε_t are independent white noise processes.

X_t is not a random walk

$$E_{t-1}(X_t) = X_{t-1} + E_{t-1}(\sigma_t \varepsilon_t) = X_{t-1} + E_{t-1}(\sigma_t) E_{t-1}(\varepsilon_t) = X_{t-1}$$

So X_t is a martingale.

Financial asset prices are often thought to be martingales, but not random walks, because of clustering in volatility.

A white noise process is a martingale difference sequence, but the converse is not true.

By the law of iterated expectations, if X_t is a martingale then $E_t(X_{t+h}) = X_t$ for all h.

$$E_t(X_{t+2}) = E_t(E_{t+1}(X_{t+2})) = E_t(X_{t+1}) = X_t$$
$$E_t(X_{t+3}) = E_t(E_{t+1}(X_{t+3})) = E_t(X_{t+1}) = X_t$$

and so on…

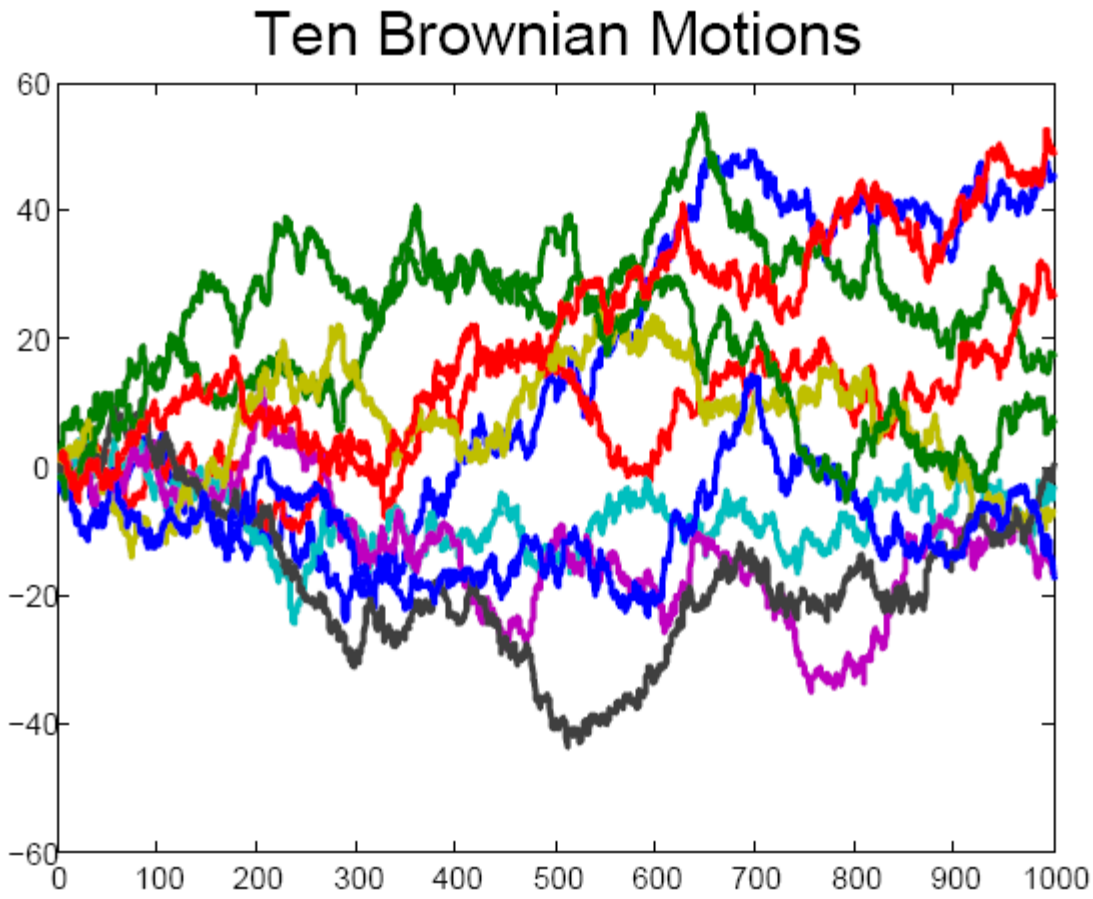
Brownian motion.

A Brownian motion is the most important continuous time stochastic process in macro and finance.

A Brownian motion is the continuous time analog of a Gaussian random walk.

The stochastic process $B(t)$ is a **Brownian motion** if

1. $B(0) = 0$
2. $B(t) - B(s) \sim N(0, \sigma^2(t-s))$ for any $t > s$
3. If $t_1 < t_2 \leq t_3 < t_4$ then
   $B(t_2) - B(t_1)$ is independent of $B(t_4) - B(t_3)$


Ten Brownian Motions
60
40
20
0
−20
−40
−60
0
100
200
300
400
500
600
700
800
900
1000

Some Brownian motion properties

- A Brownian motion is a martingale: $EB(t+\Delta t) \mid B(t) = B(t)$

- $Cov(B(s), B(t)) = \min(s,t)$

Proof: Suppose wlog that $s<t$

$$Cov(B(s), B(t)) = E(B(s)B(t)) = E((B(t) - B(s) + B(s))B(s))$$

$$\therefore Cov(B(s), B(t)) = E[(B(t) - B(s))(B(s) - B(0))] + E(B(s)^2)$$

$$\therefore Cov(B(s), B(t)) = s$$

- If T_a is the first time that $B(t)$ hits a ("first hitting time")

$$P(T_a \le t) = \sqrt{\frac{2}{\pi}} \int_{|a|/\sqrt{t}}^{\infty} e^{-y^2/2} dy$$

$$\therefore P(T_a < \infty) = \lim_{t\to\infty} P(T_a \le t) = \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-y^2/2} dy = 1$$

- $E(T_a) = \infty$

- $P(\max_{0\le s\le t} B(s) \le a) = 1 - \sqrt{\frac{2}{\pi}} \int_{a/\sqrt{t}}^\infty e^{-y^2/2} dy$ for $a > 0$

  Proof: $P(\max_{0\le s\le t} B(s) \le a) = 1 - P(T_a \le t)$

  $$= 1 - \sqrt{\frac{2}{\pi}} \int_{a/\sqrt{t}}^\infty e^{-y^2/2} dy$$

- P(Goes up A before going down B)$= \frac{B}{A+B}$

Generating a Brownian Motion on the computer

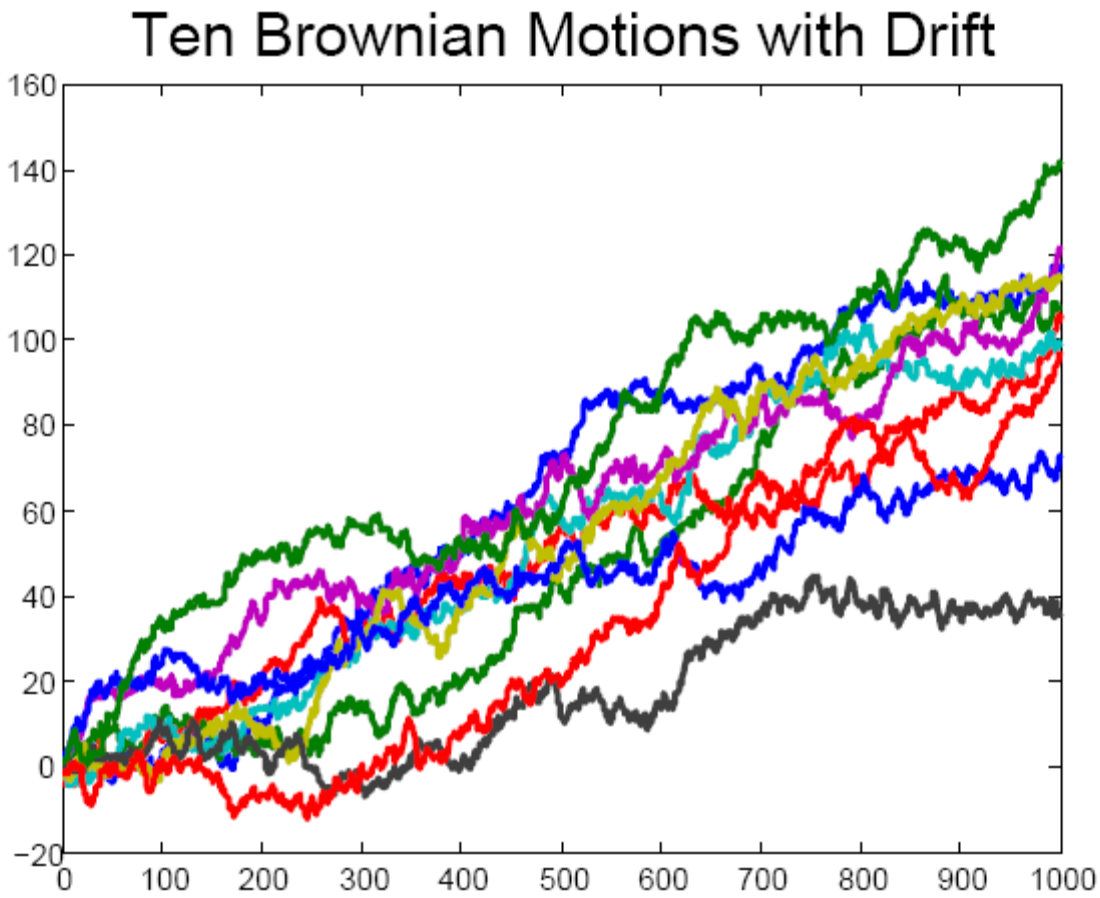
```
randn('seed',123);
p=1;
n=1000;              %Some large number
x=cumsum(randn(n*p,1)/sqrt(n));
```

x is now $B(1/n), B(2/n), ...B(p)$

Brownian Motion with drift

The stochastic process $B(t)$ is a **Brownian motion** if

1. $B(0) = 0$
2. $B(t) - B(s) \sim N(\mu, \sigma^2(t-s))$ for any $t > s$
3. If $t_1 < t_2 \le t_3 < t_4$ then
   $B(t_2) - B(t_1)$ is independent of $B(t_4) - B(t_3)$


Ten Brownian Motions with Drift
160
140
120
100
80
60
40
20
0
−20
0
100
200
300
400
500
600
700
800
900
1000

Some Brownian motion with drift properties

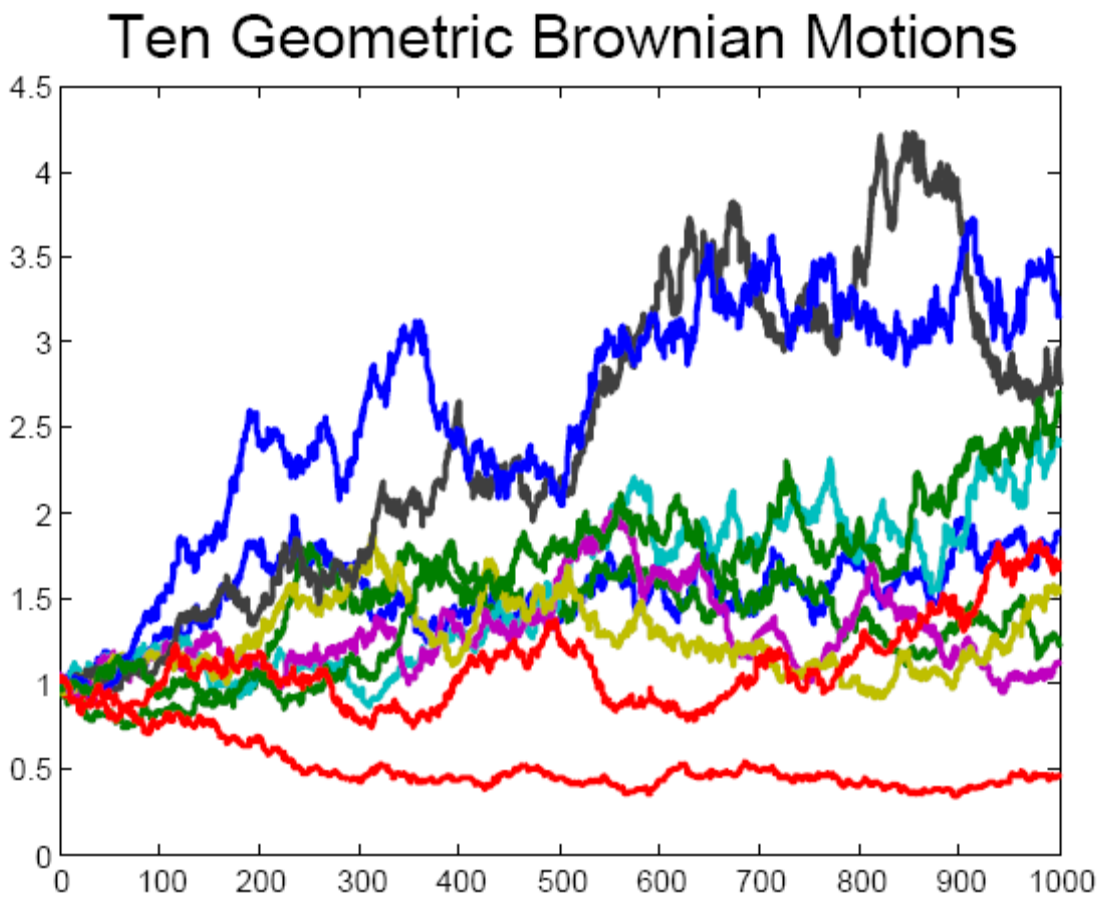
- If $\mu > 0$, $P(T_a < \infty) = 1 \quad \text{if } a > 0$
  $= e^{2\mu a} \quad \text{if } a < 0$

- If $\mu < 0$, $P(T_a < \infty) = 1 \quad \text{if } a < 0$
  $= e^{2\mu a} \quad \text{if } a > 0$

- P(Goes up A before going down B)$= \dfrac{e^{2\mu B} - 1}{e^{2\mu B} - e^{2\mu A}}$

Geometric Brownian Motion

If $B(t)$ is a Brownian motion, then $Y(t) = \exp(B(t))$ is a **geometric Brownian motion**. It is useful for modeling financial asset prices.


Ten Geometric Brownian Motions
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
100
200
300
400
500
600
700
800
900
1000

Example: Suppose that a stock price $Y(t)$ follows a geometric Brownian motion with $\sigma^2 = 1$. At time zero, the stock price is $Y(0) = 1$ and an investor has an option to buy the stock at time T at a price K. The investor will exercise the option if and only if $Y(T) > K$. What is the probability that the investor will exercise the option?

$$P(Y(T) > K) = P(\exp(B(T)) > K) = P(B(T) > \log(K))$$

But $B(T) \sim N(0,T)$ and so the probability of exercising is

$$P(N(0,T) > \log(K)) = P(N(0,1) > \frac{\log(K)}{\sqrt{T}}) = 1 - \Phi(\frac{\log(K)}{\sqrt{T}})$$

Behavior of the sample average and sample variance.

Suppose that $X_1, X_2, ...X_n$ are independent random variables all drawn from the same $N(\mu, \sigma^2)$ distribution (independently and identically distributed).

$\bar{X} = n^{-1}\Sigma_{i=1}^{n} X_i$ is a natural estimator of μ

$s^2 = (n-1)^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2$ is a natural estimator of σ^2

What are their sampling distributions?

Results:

- $E(\bar{X}) = \mu$
- $\bar{X} \sim N(\mu, \sigma^2 / n)$
- $E(s^2) = \sigma^2$
- $\bar{X}$ and s^2 are independent

Proof that $E(s^2) = \sigma^2$:

$$s^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2 = \frac{\Sigma_{i=1}^{n}X_i^2 - n\bar{X}^2}{n-1}$$

$$\therefore E(s^2) = \frac{nE(X_i^2) - nE(\bar{X}^2)}{n-1}$$

$E(X_i^2) = \mu^2 + \sigma^2$ and $E(X_i X_j) = \mu^2$ for $i \neq j$

$$E(\bar{X}^2) = \frac{1}{n^2}E((X_1 + X_2 ... + X_n)^2) = \frac{n(\mu^2 + \sigma^2) + n(n-1)\mu^2}{n^2}$$

$$\therefore E(\bar{X}^2) = \frac{\mu^2 + \sigma^2 + (n-1)\mu^2}{n} = \frac{\sigma^2}{n} + \mu^2$$

$$\therefore E(s^2) = \frac{n(\mu^2 + \sigma^2) - \sigma^2 - \mu^2 n}{n-1} = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

For more results, need to introduce the remaining continuous distributions

Chi-squared distribution

Suppose that $Z_1, Z_2 ... Z_p$ are independent $N(0,1)$ random variables.

Then $Z = \Sigma_{i=1}^{p} Z_i^2$ is χ^2 distributed on p degrees of freedom

$E(Z) = p$
$Var(Z) = 2p$
$0 \leq Z < \infty$

Two more results about the chi-squared distribution

- If X and Y are two independent χ^2 random variables on n_X and n_Y degrees of freedom, respectively, then $X + Y$ is χ^2 on $n_X + n_Y$ degrees of freedom.

- If $X \sim N(0, \Sigma)$ is kx1 with a multivariate normal distribution, then $X'\Sigma^{-1}X$ is $\chi^2(k)$ distributed.

The t distribution

If $Y \sim N(0,1)$ and $X \sim \chi^2(p)$ are independent then

$Z = Y / \sqrt{X / p}$ is t distributed on p degrees of freedom

$E(Z) = 0$ if $p > 1$ (and infinite if $p = 1$)

$Var(Z) = \dfrac{p}{p-2}$ if $p > 2$ (and infinite if $p \leq 2$)

A t distribution on 1 degree of freedom is also known as a *Cauchy* distribution.

Result: If $Y \sim N(0,1)$ and $X \sim N(0,1)$ are independent, then Y / X is Cauchy distributed.

Proof. Let $Z = X^2$

$\therefore Z \sim \chi^2(1)$

$\therefore Y / \sqrt{Z/1} \sim t(1)$

$\therefore Y / X \sim t(1)$ (a.k.a. Cauchy)

A Cauchy distribution has infinite mean and infinite variance

The F distribution

If V_1 and V_2 are independent $\chi^2(d_1)$ and $\chi^2(d_2)$ variables, respectively, then $\frac{V_1 / d_1}{V_2 / d_2}$ is $F-$ distributed on d_1 and d_2 degrees of freedom.

Now back to the problem….

Suppose that $X_1, X_2, ...X_n$ are independent random variables all drawn from the same $N(\mu, \sigma^2)$ distribution (independently and identically distributed).

We already had

- $E(\bar{X}) = \mu$
- $\bar{X} \sim N(\mu, \sigma^2 / n)$
- $E(s^2) = \sigma^2$
- $\bar{X}$ and s^2 are independent

Now we can add two new results:

- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$
- $\frac{(\bar{X}-\mu)}{s/\sqrt{n}} \sim t(n-1)$

Suppose that $X_1, X_2, ...X_n$ are independent random variables all drawn from the same distribution (independently and identically distributed) with mean μ and variance σ^2 but are **not necessarily normal**.

What can we say about the sampling distribution of $\bar{X}$?

As *n* gets big, $\bar{X}$ "converges" to μ.

What exactly does this mean?

Nonstochastic limits.

$\lim_{n\to\infty} X_n = a$ means
$\forall \varepsilon > 0, \exists n_0$ such that $n > n_0 \Rightarrow | X_n - a | < \varepsilon$

$\lim_{n\to\infty} X_n = a$ and $\lim_{n\to\infty} Y_n = b$ implies that

- $\lim_{n\to\infty} (X_n + Y_n) = a + b$
- $\lim_{n\to\infty} X_n Y_n = ab$
- $\lim_{n\to\infty} g(X_n) = g(a)$ for any cts fn $g(.)$

Nonstochastic orders of magnitude

We say that X_n is of **order of magnitude** f_n, $X_n = O(f_n)$ if $\lim_{n\to\infty} X_n / f_n = C, 0 < C < \infty.$

We say that X_n is of **smaller order of magnitude** than f_n, $X_n = o(f_n)$ if $\lim_{n\to\infty} X_n / f_n = 0$

Example: $\Sigma_{i=1}^{n} i = O(n^2)$

$$\Sigma_{i=1}^{n} i = \frac{n(n+1)}{2} \Rightarrow \frac{\Sigma_{i=1}^{n} i}{n^2} = \frac{n(n+1)}{2n^2} \to \frac{1}{2} \text{ as } n \to \infty.$$

$$X_n = O(f_n) \text{ and } Y_n = O(g_n) \Rightarrow X_n + Y_n = O(\max(f_n, g_n))$$

$$X_n = O(f_n) \text{ and } Y_n = O(g_n) \Rightarrow X_n Y_n = O(f_n g_n)$$

$$X_n = o(f_n) \text{ and } Y_n = o(g_n) \Rightarrow X_n + Y_n = o(\max(f_n, g_n))$$

$$X_n = o(f_n) \text{ and } Y_n = o(g_n) \Rightarrow X_n Y_n = o(f_n g_n)$$

Suppose that X_n is a random sequence.

X_n **converges in probability** to μ if for all $\varepsilon > 0$
$\lim_{n\to\infty} P(|X_n - \mu| > \varepsilon) = 0.$

- We write this as $X_n \to_p \mu$

X_n **converges almost surely** to μ if for all $\varepsilon > 0$
$P(\lim_{n\to\infty} X_n = \mu) = 1.$

- We write this as $X_n \to_{as} \mu$

X_n converges **in quadratic mean** to μ if
$\lim_{n\to\infty} E((X_n - \mu)^2) = 0$

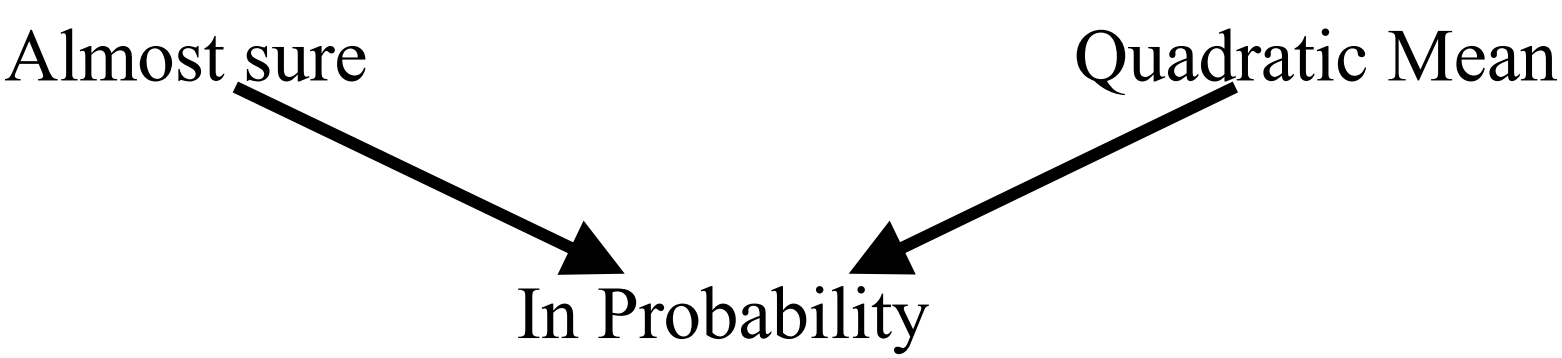
- We write this as $X_n \to_{qm} \mu$

X_n is **unbiased** for μ if $E(X_n) = \mu$

X_n is **asymptotically unbiased** for μ if $\lim_{n\to\infty} E(X_n) = \mu$

Theorems.

- If $X_n \to_{as} \mu$ then $X_n \to_p \mu$ but the converse is not true.
- If $X_n \to_{qm} \mu$ then $X_n \to_p \mu$ but the converse is not true.



Example that convergence in probability does not imply convergence almost surely.

Let S be uniform on the unit interval. Consider the random sequence

$X_1(s) = s + 1(0 \le s \le 1)$

$X_2(s) = s + 1(0 \le s \le 1/2)$ $X_3(s) = s + 1(1/2 < s \le 1)$

$X_4(s) = s + 1(0 \le s \le 1/3)$ $X_5(s) = s + 1(1/3 < s \le 2/3)$

$X_6(s) = s + 1(2/3 < s \le 1)$

etc. and let $X(s) = s$

$\lim_{n\to\infty} P(|X_n(s) - X(s)| > \varepsilon) = 0 \ \forall \varepsilon > 0 \Rightarrow X_n(s) \to_p X(s)$

But for every s, $X_n(s)$ alternates between s and $s+1$

So $X_n(s)$ does not converge almost surely to $X(s)$

Proof that convergence in q.m. implies convergence in probability.

By Chebychev, for any $\varepsilon > 0$

$$P[|X_n - \mu| \geq \varepsilon] = P[(X_n - \mu)^2 \geq \varepsilon^2] \leq \frac{E((X_n - \mu)^2)}{\varepsilon^2}$$

Suppose that $X_n \rightarrow_{qm} \mu$.

$\Rightarrow \lim_{n\to\infty} E((X_n - \mu)^2) = 0$

$\Rightarrow \lim_{n\to\infty} P[|X_n - \mu| \geq \varepsilon] = 0$

$\Rightarrow X_n \rightarrow_p \mu$

Counterexample that the converse is not true

Suppose that $X_n = \mu$ w.p $1 - 1/n$

$X_n = n$ w.p. $1/n$

$\lim_{n\to\infty} P(|X_n - \mu| > \varepsilon) = 0$

$\therefore X_n \to_p \mu$

$$E((X_n - \mu)^2) = 0 * (1 - \frac{1}{n}) + (n - \mu)^2 * (\frac{1}{n}) = \frac{(n-\mu)^2}{n}$$

$\therefore \lim_{n\to\infty} E((X_n - \mu)^2) \neq 0$ (actually it is ∞)

X_n does not converge in quadratic mean to μ

More results

- If $X_n \to_p \mu_x$ and $Y_n \to_p \mu_y$ then $X_n + Y_n \to_p \mu_x + \mu_y$

- If $X_n \to_p \mu_x$ and $Y_n \to_p \mu_y$ then $X_n Y_n \to_p \mu_x \mu_y$

- If $X_n \to_p \mu$ and $g(.)$ is a continuous function then $g(X_n) \to_p g(\mu)$

Three ways of saying the same thing:

- $X_n \rightarrow_p \mu$
- The probability limit (or "plim") of X_n is μ
- X_n is consistent for μ

Weak Law of Large Numbers

If $\{X_1, X_2, ...X_n\}$ is iid with mean μ and variance $\sigma^2 < \infty$ then $\bar{X} = n^{-1}\Sigma_{i=1}^{n} X_i \rightarrow_p \mu$

Proof:

$$P(|\bar{X}_n - \mu| \geq \varepsilon) = P((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \leq \frac{E((\bar{X}_n - \mu)^2)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

$$\therefore \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \varepsilon) = 0$$

$$\therefore \bar{X}_n \rightarrow_p \mu$$

"You can…never fortell what any one man will do, but you can say with precision what the average number will be up to. Individuals vary but percentages remain constant. So says the statistician."
Sherlock Holmes.

Implication of WLLN

If $\{X_1, X_2, ...X_n\}$ is iid with mean μ and variance $\sigma^2 < \infty$
then $s^2 = (n-1)^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2 \rightarrow_p \sigma^2$

Proof: $s^2 = (n-1)^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}(\frac{1}{n}\Sigma_{i=1}^{n}X_i^2 - n\bar{X}^2)$

$n^{-1}\Sigma_{i=1}^{n}X_i^2 \rightarrow_p E(X_i^2) = \mu^2 + \sigma^2$

$\bar{X} \rightarrow_p \mu \Rightarrow \bar{X}^2 \rightarrow_p \mu^2$

$\therefore \frac{1}{n}\Sigma_{i=1}^{n}X_i^2 - n\bar{X}^2 \rightarrow_p \mu^2 + \sigma^2 - \mu^2 = \sigma^2$

$\therefore s^2 \rightarrow_p \sigma^2$

Strong Law of Large Numbers

If $\{X_1, X_2, ...X_n\}$ is iid with mean μ and variance $\sigma^2 < \infty$
then $\bar{X} = n^{-1}\Sigma_{i=1}^{n} X_i \rightarrow_{as} \mu$

Convergence in distribution

If X_n is a sequence of random variables each with cdf F_n, then X_n **converges in distribution** to X, written $X_n \to_d X$ if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x where F is the cdf of X.

Results on convergence in distribution

- Slutsky's Theorem

If $Z_n \to_d Z$ and $Y_n \to_p \mu$ (a constant), then

$Z_n Y_n \to_d \mu Z$

$Z_n + Y_n \to_d Z + \mu$

- If $X_n \to_p X$ then $X_n \to_d X$

- If c is a constant then $X_n \to_p c$ if and only if $X_n \to_d c$

Central Limit Theorem

$\bar{X} \to_p \mu$…just tells us that the distribution of $\bar{X}$ is degenerate in a large enough sample.

The central limit theorem. If $X_1, X_2 ... X_n$ are independently and identically distributed with mean μ, variance σ^2 and $2+\delta$ finite moments for some $\delta > 0$ then:

$$\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$$

Amazing result….the $\{X_i\}$s are not normal, but average is.

The approximation to the sampling distribution of $\bar{X}$ obtains by "flipping over" the results of the CLT

$$\bar{X} \sim_{\text{approx}} N(\mu, \sigma^2 / n)$$

This is not an approximation if the $\{X_i\}$s are normal

Illustration of the central limit theorem

Let X_i be a Bernoulli random variable (1 with probability p and 0 otherwise).

$X = \Sigma_{i=1}^{n} X_i$ is binomial with parameters n and p

$$P(X = x) = C_x^n p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$\{X_i\}$ are independent, $E(X_i) = p$ and $Var(X_i) = p(1-p)$

$$\Rightarrow \sqrt{n}(\bar{X} - p) \rightarrow_d N(0, p(1-p)) \quad \text{(CLT)}$$

$$\Rightarrow \bar{X} \sim_{\text{approx}} N(p, \frac{p(1-p)}{n})$$

$$\Rightarrow \Sigma_{i=1}^{n} X_i \sim_{\text{approx}} N(np, np(1-p))$$

The central limit theorem implies the normal approximation to the binomial. A binomial random variable with parameters n and p is approximately $N(np, np(1-p))$.

Example. Suppose that 52 percent of voters in fact plan to vote for candidate *X*. A random sample of 1,000 voters is chosen. What is the probability that at least 500 people in the sample plan to vote for candidate *X*?

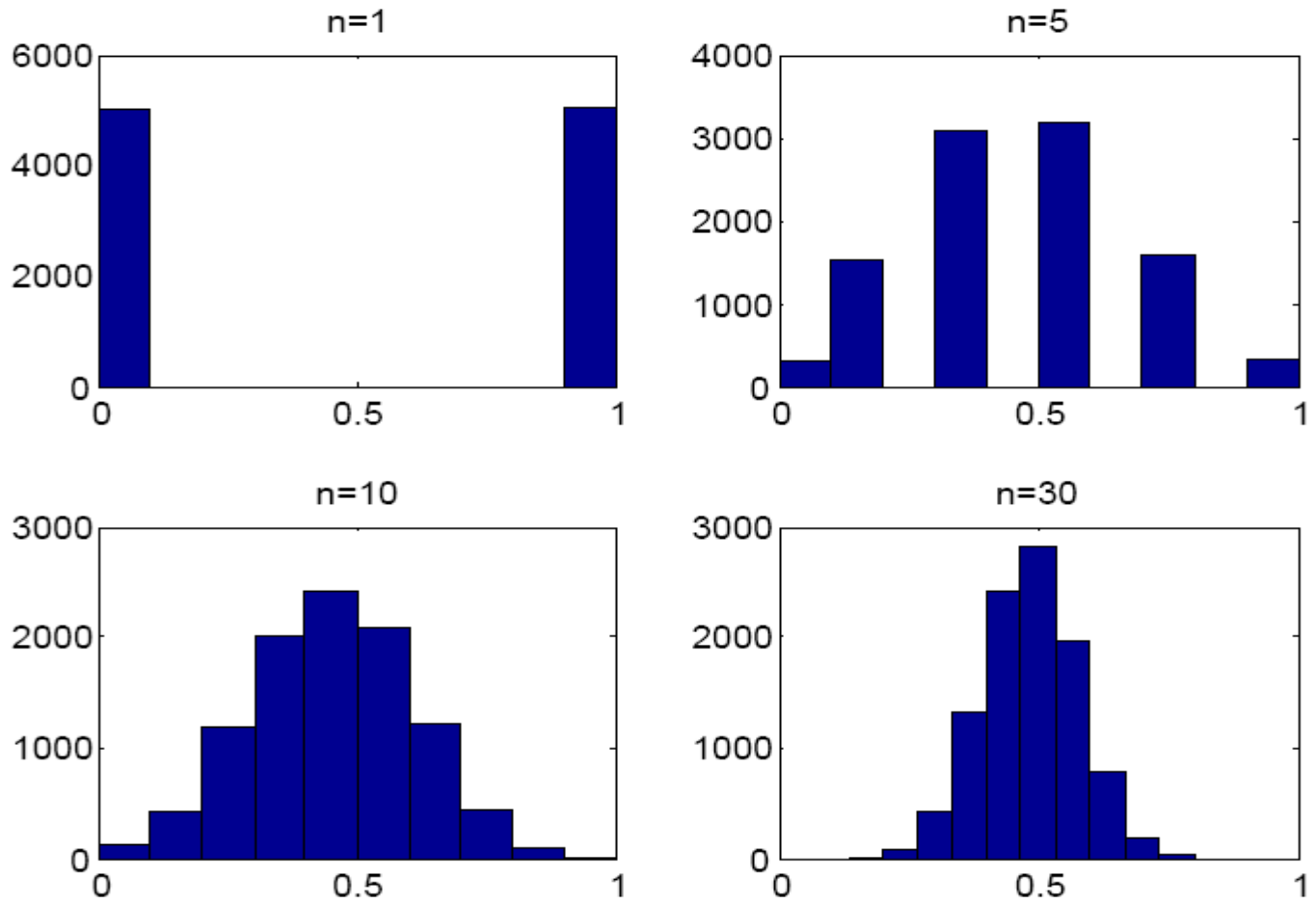
The exact answer is the probability that a Binomial with parameters 1,000 and 0.52 is greater than or equal to 500. Hard to calculate.

But the number in the sample who vote for *X*
$\approx N(0.52*1000, 0.52*0.48*1000) = N(520, 249.6)$

Probability that this is at least 500=0.897

$\{X_i\}$ is Bernoulli.

Monte-Carlo Simulated distribution of $\overline{X} = n^{-1}\Sigma_{i=1}^{n} X_i$



Sketch of proof of the Central Limit Theorem

Let $Y_i = \frac{X_i - \mu}{\sigma}$. Then $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \Sigma_{i=1}^{n} Y_i$.

Let $M_Y(t)$ denote the moment generating function of Y_i.

The mgf of $\frac{1}{\sqrt{n}} \Sigma_{i=1}^{n} Y_i$ is $M_Y(\frac{t}{\sqrt{n}})^n$

$$M_Y(\frac{t}{\sqrt{n}}) \approx M_Y(0) + \frac{t}{\sqrt{n}} M_Y'(0) + \frac{1}{2}\frac{t^2}{n} M_Y''(0)$$

$$M_Y(0) = E(\exp(0 * Y_i)) = 1$$

$$M_Y'(0) = E(Y_i) = 0$$

$$M_Y''(0) = E(Y_i^2) = 1$$

$$\therefore M_Y(\frac{t}{\sqrt{n}}) \approx 1 + \frac{1}{2}\frac{t^2}{n}$$

$$\therefore M_Y(\frac{t}{\sqrt{n}})^n \approx [1 + \frac{1}{2}\frac{t^2}{n}]^n$$

$$\therefore \lim_{n\to\infty} M_Y(\frac{t}{\sqrt{n}})^n \approx \exp(t^2/2) \quad (\text{using } e^x = \lim_{n\to\infty}(1+\frac{x}{n})^n)$$

The limiting mgf of $\frac{1}{\sqrt{n}}\Sigma_{i=1}^n Y_i$ is $\exp(t^2/2)$

$$\therefore \frac{1}{\sqrt{n}}\Sigma_{i=1}^n Y_i \to_d N(0,1) \Rightarrow \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \to_d N(0,1)$$

$$\therefore \sqrt{n}(\bar{X}-\mu) \to_d N(0,\sigma^2)$$

Application of Slutsky's Theorem

If $X_1, X_2 ... X_n$ are independently and identically distributed with mean μ, variance σ^2, what's the distribution of $\frac{\sqrt{n}(\bar{X}-\mu)}{s}$?

$$\frac{\sqrt{n}(\bar{X}-\mu)}{s} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\frac{\sigma}{s}$$

CLT: $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \rightarrow_d N(0,1)$

Showed Earlier: $s^2 \rightarrow_p \sigma^2 \Rightarrow s/\sigma \rightarrow_p 1$

Using these + Slutsky: $\frac{\sqrt{n}(\bar{X}-\mu)}{s} \rightarrow_d N(0,1)$

Cumulants

- Define the "cumulant generating function" as $\log(M(t))$

where $M(t)$ is the moment generating function.

- Define the "cumulants" of a random variable as

$$\kappa_j = \left. \frac{d^j \log(M(t))}{dt^j} \right|_{t=0}$$

- For a random variable X

$$\kappa_1 = E(X) = \mu$$

$$\kappa_2 = E((X-\mu)^2) = \sigma^2$$

$$\kappa_3 = E((X-\mu)^3)$$

$$\kappa_4 = E((X-\mu)^4) - 3\sigma^4$$

Note: If X is normal, then $\kappa_3 = \kappa_4 = 0$

Edgeworth Expansion: Refinement on the CLT

Suppose that $X_1, X_2, \ldots X_n$ are iid with a cdf F with mean μ and variance σ^2. Let be the ith cumulant of F.

- Define the "standardized cumulants": $\rho_i = \kappa_i / \sigma^i$
- Let $F_n(x)$ denote the cdf of $\dfrac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$

Edgeworth Expansion (Two Terms)

$$F_n(x) = \Phi(x) - \frac{\rho_3 \Phi^{(3)}(x)}{6\sqrt{n}} + O(\frac{1}{n})$$

where $\Phi(x)$ is the $N(0,1)$ cdf and $\Phi^{(j)}(x) = \dfrac{d^j \Phi(x)}{dx^j}$

Edgeworth Expansion (Three Terms)

$$F_n(x) = \Phi(x) - \frac{\rho_3 \Phi^{(3)}(x)}{6\sqrt{n}} + \frac{1}{n}\left[\frac{\rho_4 \Phi^{(4)}(x)}{24} + \frac{\rho_3^2 \Phi^{(6)}(x)}{72}\right] + O\left(\frac{1}{n^{3/2}}\right)$$

Note: If the "parent" distribution, F, is normal, $\rho_3 = \rho_4 = 0$ and $F_n(x) = \Phi(x)$.

Immediate implication of the Edgeworth expansion

If $X_1, X_2, ...X_n$ are iid with a cdf F with mean μ and variance σ^2 then

$$\boxed{F_n(x) = \Phi(x) + O(n^{-1/2})}$$

Tells us about speed of convergence.

Berry-Esseen Bound

$$| F_n(x) - \Phi(x) | \leq \frac{CE(| X_i |^3)}{\sigma^3} n^{-1/2}$$

Central limit theorems for non iid random variables

- Suppose that $X_1, X_2 ... X_n$ are independently distributed with different means $\mu_1, ... \mu_n$ and different variances $\sigma_1^2, ... \sigma_n^2$. Then (under regularity conditions)

$$\sqrt{n}(\bar{X} - \mu) \rightarrow_d N(0, \sigma^2)$$

where $\mu = \lim_{n\rightarrow\infty} n^{-1} \Sigma_{i=1}^{n} \mu_i$ and $\sigma^2 = \lim_{n\rightarrow\infty} n^{-1} \Sigma_{i=1}^{n} \sigma_i^2$.

- Suppose that $X_1, X_2 ... X_T$ is a martingale difference sequence and $E(X_t^2) = \sigma^2$. Then (under some conditions)

$$\sqrt{T}\,\bar{X} \rightarrow_d N(0, \sigma^2)$$

Stochastic Orders of Magnitude

Stochastic orders of magnitude can be useful in asymptotic statistical theory.

Suppose that X_n is a random sequence.

We say that X_n is of **order of magnitude** f_n, $X_n = O_p(f_n)$ if $\forall \varepsilon > 0, \exists C < \infty$ and n_0 such that $P(|X_n| / f_n > C) < \varepsilon$ for all $n > n_0$.

We say that X_n is of **smaller order of magnitude** than f_n, $X_n = o_p(f_n)$ if $X_n / f_n \to_p 0$

$$X_n = O_p(f_n) \text{ and } Y_n = O_p(g_n) \Rightarrow X_n + Y_n = O_p(\max(f_n, g_n))$$

$$X_n = O_p(f_n) \text{ and } Y_n = O_p(g_n) \Rightarrow X_n Y_n = O_p(f_n g_n)$$

$$X_n = o_p(f_n) \text{ and } Y_n = o_p(g_n) \Rightarrow X_n + Y_n = o_p(\max(f_n, g_n))$$

$$X_n = o_p(f_n) \text{ and } Y_n = o_p(g_n) \Rightarrow X_n Y_n = o_p(f_n g_n)$$

$$X_n = o_p(f_n) \Rightarrow X_n = O_p(f_n)$$

$$\text{If } f_n / g_n \to 0 \text{ then } X_n = O_p(f_n) \Rightarrow X_n = o_p(g_n)$$

$$X_n = O_p((E \mid X_n \mid^r)^{1/r}) \text{ for } r > 0 \text{ (from Chebychev Inequality)}$$

$$X_n = O_p(f_n) \Rightarrow X_n g_n = O_p(f_n g_n)$$

Suppose that $X_1, X_2, ...X_n$ are iid with mean μ and variance σ^2.

Example 1.

By the CLT, $\sqrt{n}(\bar{X} - \mu) \rightarrow_d N(\mu, \sigma^2)$

$\therefore \bar{X} - \mu = O_p(n^{-1/2})$

Example 2.

Let $\hat{\sigma}^2 = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \bar{X})^2$. What is the asymptotic Distribution of $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$?

$$\hat{\sigma}^2 = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2$$

$$= \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \mu)^2 + \frac{1}{n}\Sigma_{i=1}^{n}(\mu - \bar{X})^2 + \frac{2}{n}\Sigma_{i=1}^{n}(X_i - \mu)(\mu - \bar{X})$$

$$= \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \mu)^2 - (\mu - \bar{X})^2 = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \mu)^2 + O_p(n^{-1})$$

$$\therefore \sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{1}{\sqrt{n}}\{\Sigma_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\} + O_p(n^{-1/2})$$

$$\therefore \sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \frac{1}{\sqrt{n}}\{\Sigma_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\} + o_p(1)$$

By the CLT

$$\frac{1}{\sqrt{n}}\Sigma_{i=1}^{n}\{(X_i-\mu)^2-\sigma^2\} \rightarrow_d N(0, Var((X_i-\mu)^2))$$

So by Slutsky's Theorem

$$\sqrt{n}(\hat{\sigma}^2-\sigma^2) \rightarrow_d N(0, Var((X_i-\mu)^2))$$

If the data were normal, this would reduce to

$$\sqrt{n}(\hat{\sigma}^2-\sigma^2) \rightarrow_d N(0, 2\sigma^4)$$

"Stable" random variables

If $X_1, X_2, ...X_n$ are iid random variables with some distribution and $\Sigma_{i=1}^{n} X_i$ has the same distribution, then these random variables are **stable**.

- The normal distribution is stable

- By the CLT, most distributions are **not** stable

- A Brownian motion has normal increments. Would it make sense to change the increments to be (say) t(10)?

- But the Cauchy distribution is stable

To recap…we have a random sample $X_1, X_2 ... X_n$ that is drawn from a distribution with mean μ and variance σ^2

Parameter	μ	σ^2
Estimator	$\bar{X} = n^{-1}\Sigma_{i=1}^{n} X_i$	$s^2 = (n-1)^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2$
Expectation	$E(\bar{X}) = \mu$	$E(s^2) = \sigma^2$
Variance	$Var(\bar{X}) = \frac{\sigma^2}{n}$	$Var(s^2) = \frac{2\sigma^4}{n-1}$
If $X_1, X_2 ... X_n$ are normal	$\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$	$\frac{(n-1)s^2}{\sigma^2}$ is $\chi^2(n-1)$
In large samples	$\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$	$\sqrt{n}(s^2 - \sigma^2) \rightarrow_d N(0, \kappa^2)$ $\kappa^2 = Var((X_i - \mu)^2)$

Convergence results for nonlinear transformations

Continuous Mapping Theorem

If $X_n \to_d X$ and $g(.)$ is a continuous function then $g(X_n) \to_d g(X)$

Delta Method

If $\sqrt{n}(X_n - \mu) \to_d N(0, \sigma^2)$ and $g(.)$ is a continuous function s.t. $g'(\mu) \neq 0$ then $\sqrt{n}(g(X_n) - g(\mu)) \to_d N(0, g'(\mu)^2 \sigma^2)$

Proof: $g(X_n) \approx g(\mu) + (X_n - \mu) g'(\mu)$

$\therefore \sqrt{n}(g(X_n) - g(\mu)) \approx g'(\mu) \sqrt{n}(X_n - \mu)$

$\therefore \sqrt{n}(g(X_n) - g(\mu)) \to_d g'(\mu) N(0, \sigma^2) = N(0, g'(\mu)^2 \sigma^2)$

Example 1. If $\{X_1, X_2, ...X_n\}$ is iid with mean μ and variance σ^2, what is the distribution of $\frac{1}{\bar{X}}$?

$$\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$$
$$g(z) = 1/z \to g'(z) = -1/z^2$$
$$\sqrt{n}(\frac{1}{\bar{X}} - \frac{1}{\mu}) \to_d N(0, \{-\frac{1}{\mu^2}\}^2 \sigma^2) = N(0, \frac{\sigma^2}{\mu^4})$$

Example 2. If $\{X_1, X_2, ...X_n\}$ is iid with mean μ and variance σ^2, what is the distribution of $e^{\bar{X}}$?

$$\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$$
$$g(z) = e^z \to g'(z) = e^z$$
$$\sqrt{n}(e^{\bar{X}} - e^{\mu}) \to_d N(0, \{e^{\mu}\}^2 \sigma^2) = N(0, \sigma^2 e^{2\mu})$$

Example 3. If $\{X_1, X_2, ...X_n\}$ is iid with mean $\mu > 0$ and variance σ^2, what is the distribution of $\log(\bar{X})$?

$$\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$$
$$g(z) = \log(z) \to g'(z) = 1/z$$
$$\sqrt{n}(\log(\bar{X}) - \log(\mu)) \to_d N(0, \{\frac{1}{\mu}\}^2 \sigma^2) = N(0, \frac{\sigma^2}{\mu^2})$$

Multivariate central limit theorem

Suppose that $\{X_1, X_2, ...X_n\}$ are kx1 random vectors that are independently and identically distributed with mean μ and variance-covariance matrix Σ. Then

$$\sqrt{n}(\bar{X} - \mu) \rightarrow_d N(0, \Sigma)$$

Cramer-Wold device

If X_n and X are kx1 random vectors, then if $\lambda' X_n \rightarrow_d \lambda' X$ for all fixed kx1 vectors λ, then $X_n \rightarrow_d X$.

Immediately gives a proof of the multivariate CLT as a consequence of the univariate CLT.

Multivariate delta method

If $\sqrt{n}(X_n - \mu) \rightarrow_d N(0, \Sigma)$ and $g(.)$ is a continuous function, then

$$\sqrt{n}(g(X_n) - g(\mu)) \rightarrow_d N(0, \frac{\partial g}{\partial \mu'} \Sigma \frac{\partial g}{\partial \mu})$$

Example. Suppose that $U_1,...U_n$ and $V_1,...V_n$ are random variables that are iid with mean μ_U and μ_V, variance σ_U^2 and σ_V^2 and correlation ρ. What is the distribution of $\bar{U} / \bar{V}$?

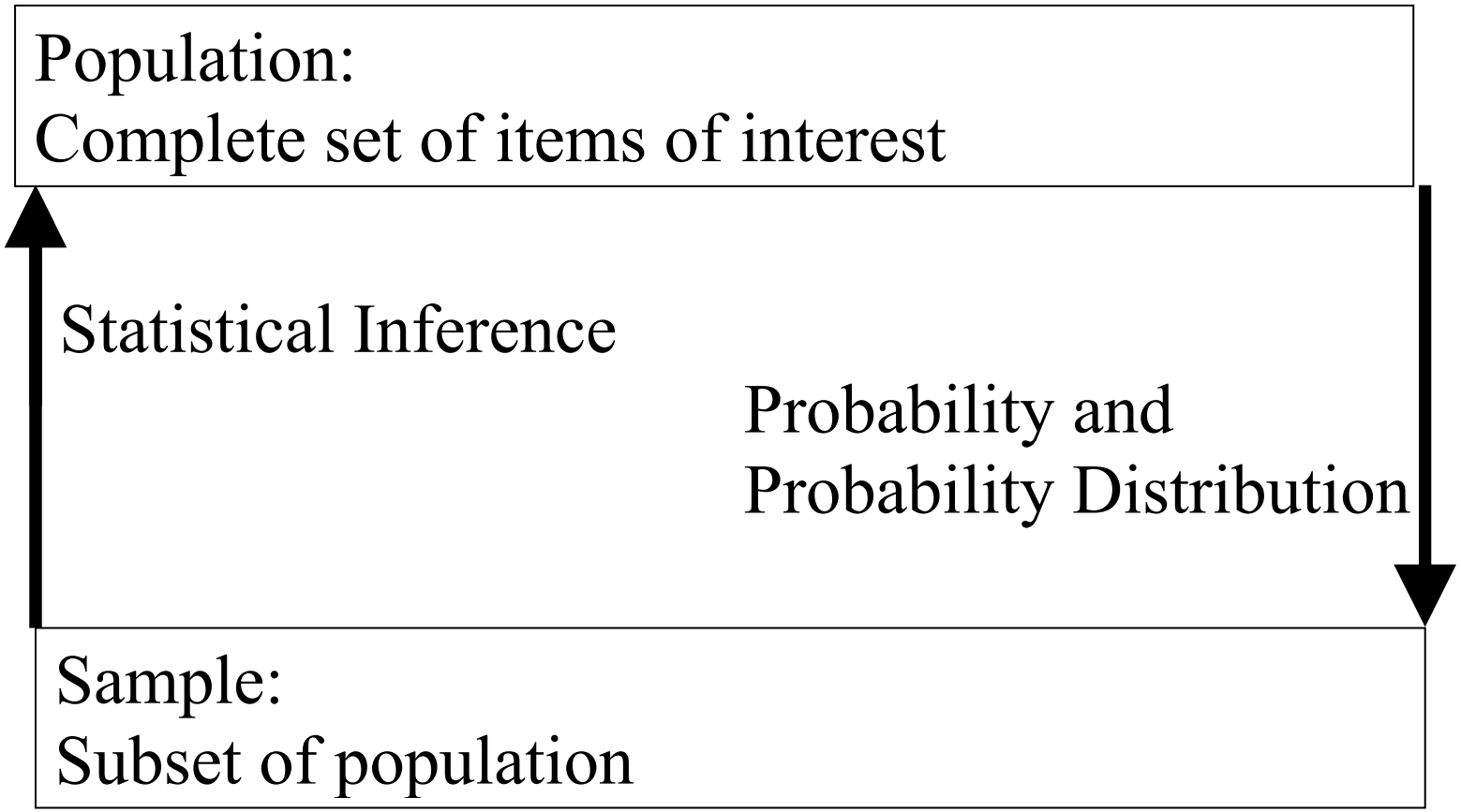
$$\sqrt{n}\begin{pmatrix} \bar{U} - \mu_U \\ \bar{V} - \mu_V \end{pmatrix} \rightarrow_d N(0, \begin{pmatrix} \sigma_U^2 & \rho\sigma_U\sigma_V \\ \rho\sigma_U\sigma_V & \sigma_V^2 \end{pmatrix})$$

$$g : R^2 \rightarrow R^1 : g(u,v) = u / v$$

$$\frac{\partial g}{\partial u} = \frac{1}{v} \text{ and } \frac{\partial g}{\partial v} = -\frac{u}{v^2}$$

$$\therefore \sqrt{n}\left(\frac{\bar{U}}{\bar{V}} - \frac{\mu_U}{\mu_V}\right) \to_d N\left(0, \begin{pmatrix} \frac{1}{\mu_V} & -\frac{\mu_U}{\mu_V^2} \end{pmatrix} \begin{pmatrix} \sigma_U^2 & \rho\sigma_U\sigma_V \\ \rho\sigma_U\sigma_V & \sigma_V^2 \end{pmatrix} \begin{pmatrix} \frac{1}{\mu_V} \\ -\frac{\mu_U}{\mu_V^2} \end{pmatrix}\right)$$

$$\therefore \sqrt{n}\left(\frac{\bar{U}}{\bar{V}} - \frac{\mu_U}{\mu_V}\right) \to_d N\left(0, \frac{\sigma_U^2}{\mu_V^2} + \frac{\sigma_V^2\mu_U^2}{\mu_V^4} - 2\frac{\rho\sigma_U\sigma_V\mu_U}{\mu_V^3}\right)$$


Population:
Complete set of items of interest
Statistical Inference
Probability and
Probability Distribution
Sample:
Subset of population

Parameter Estimation

Say I have an unknown parameter. For example $X_1, X_2, ...X_n$ are iid with some mean μ and variance σ^2. We want to estimate the parameters.

Three methods:

1. Method of moments.
2. Maximum Likelihood.
3. Bayesian estimation.

Method of moments

Say $X_1, X_2, ...X_n$ is iid from a density $f(x,\theta)$ where θ is a kx1 vector of parameters.

Rule. Set the first k sample uncentered moments to their population counterparts and solve the equations for $\hat{\theta}$

$$\frac{1}{n}\Sigma_{i=1}^{n} X_i = \int x f(x,\hat{\theta})dx \quad (\text{or } \Sigma x P(X_i = x))$$

$$\frac{1}{n}\Sigma_{i=1}^{n} X_i^2 = \int x^2 f(x,\hat{\theta})dx$$

...

$$\frac{1}{n}\Sigma_{i=1}^{n} X_i^k = \int x^k f(x,\hat{\theta})dx$$

Method of Moments Example 1:

Suppose that $X_1, X_2, \ldots X_n$ is iid $N(\mu, \sigma^2)$.

$$\bar{X} = \hat{\mu}$$

$$\frac{1}{n}\Sigma_{i=1}^{n} X_i^2 = \hat{\mu}^2 + \hat{\sigma}^2$$

$$\Rightarrow \frac{1}{n}\Sigma_{i=1}^{n} X_i^2 - \bar{X}^2 = \hat{\sigma}^2$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\Sigma_{i=1}^{n} (X_i - \bar{X})^2$$

Gives estimators $\hat{\mu}$ and $\hat{\sigma}^2$

…the sample mean and (almost) sample variance

Method of Moments Example 2:

Suppose that $X_1, X_2, \ldots X_n$ is iid Binomial with parameters k and p. We want to estimate k and p.

$$n^{-1}\Sigma_{i=1}^{n} X_i = \hat{k}\hat{p}$$
$$n^{-1}\Sigma_{i=1}^{n} X_i^2 = \hat{k}\hat{p}(1-\hat{p}) + \hat{k}^2\hat{p}^2$$

Solving these equations for $\hat{k}$ and $\hat{p}$ yields:

$$\hat{k} = \frac{\overline{X}^2}{\overline{X} - n^{-1}\Sigma_{i=1}^{n}(X_i - \overline{X})^2}$$
$$\hat{p} = \overline{X} / \hat{k}$$

Method of Moments Example 3:

Suppose that $X_1, X_2, ...X_n$ is iid uniform from 0 to θ.

$n^{-1}\Sigma_{i=1}^{n} X_i = \hat{\theta}/2$

$\therefore \hat{\theta} = 2\bar{X}$

Q. Do you think this is a sensible estimator of θ?

Maximum Likelihood estimation

Say $X_1, X_2, \ldots X_n$ is iid from a density $f(x,\theta)$ where θ is a kx1 vector of parameters.

The joint probability density of the data is $\Pi_{i=1}^{n} f(X_i,\theta)$

Idea of maximum likelihood estimation. View this as a function of θ called the likelihood function:

$L(\theta) = \Pi_{i=1}^{n} f(X_i,\theta)$

The MLE is the value of θ that maximizes the likelihood function:

$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta)$

Because it is easier to work with sums than products, we generally write the MLE as

$$\hat{\theta}_{MLE} = \arg\max_{\theta} l(\theta)$$

where

$$l(\theta) = \log L(\theta) = \Sigma_{i=1}^{n} \log f(X_i, \theta)$$

Maximum Likelihood Example 1:

Suppose that $X_1, X_2, \ldots X_n$ is iid $N(\mu, 1)$.

$$f(x,\mu) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

$$\therefore \log f(x,\mu) = -\frac{\log(2\pi)}{2} - \frac{(x-\mu)^2}{2}$$

$$\therefore l(\mu) = -\frac{n\log(2\pi)}{2} - \frac{1}{2}\Sigma_{i=1}^{n}(X_i - \mu)^2$$

FOC: $l'(\mu) = 0 \Rightarrow -\Sigma_{i=1}^{n}(X_i - \mu) * (-1) = 0$

$$\therefore \Sigma_{i=1}^{n}(X_i - \hat{\mu}_{MLE}) = 0$$

$$\therefore \hat{\mu}_{MLE} = \bar{X}$$

Maximum Likelihood Example 2:

Suppose that $X_1, X_2, \ldots X_n$ is iid $N(\mu, \sigma^2)$.

$$f(x, \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\therefore \log f(x, \mu, \sigma^2) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\therefore l(\mu, \sigma^2) = -\frac{n\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(X_i - \mu)^2$$

FOC: $l'(\mu) = 0 \Rightarrow -\frac{1}{\sigma^2}\Sigma_{i=1}^{n}(X_i - \mu) * (-1) = 0 \Rightarrow \hat{\mu}_{MLE} = \bar{X}$

$$l'(\sigma^2) = 0 \Rightarrow -\frac{n}{2}\frac{1}{2\pi\sigma^2}2\pi + \frac{1}{2\sigma^4}\Sigma_{i=1}^{n}(X_i - \mu)^2 = 0$$

$$\therefore -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma_{i=1}^{n}(X_i - \mu)^2 = 0$$

$$\therefore -n + \frac{1}{\sigma^2}\Sigma_{i=1}^{n}(X_i - \mu)^2 = 0$$

$$\therefore \frac{1}{\sigma^2}\Sigma_{i=1}^{n}(X_i - \mu)^2 = n$$

$$\therefore \hat{\sigma}^2_{MLE} = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \hat{\mu}_{MLE})^2$$

Maximum Likelihood Example 3:

Suppose that $X_1, X_2, \ldots X_n$ is iid uniform from 0 to θ.

$$f(x) = \frac{1}{\theta} 1(0 \le x \le \theta)$$

$$\therefore l(\theta) = \Sigma_{i=1}^{n} \log\left(\frac{1(0 \le X_i \le \theta)}{\theta}\right)$$

Suppose $\theta < \max_{i=1,\ldots n} X_i$.

Then $l(\theta) = -\infty$.

Suppose $\theta \ge \max_{i=1,\ldots n} X_i$.

Then $l(\theta) = \Sigma_{i=1}^{n} \log(1/\theta) = n \log(1/\theta)$.

This is monotonically decreasing in θ.

$$\therefore \hat{\theta}_{MLE} = \max_{i=1,\ldots n} X_i$$

Maximum Likelihood Example 4:

Suppose that $X_1, X_2, \ldots X_n$ is iid Bernoulli with parameter p

$$f(x) = p^x (1-p)^{1-x}$$

$$\therefore \log f(x) = x \log(p) + (1-x)\log(1-p)$$

$$\therefore l(p) = \Sigma_{i=1}^{n} \{X_i \log(p) + (1-X_i)\log(1-p)\}$$

$$\therefore l(p) = \log(p)\Sigma_{i=1}^{n} X_i + \log(1-p)(n - \Sigma_{i=1}^{n} X_i)$$

$$\therefore l'(p) = 0 \Rightarrow \frac{\Sigma_{i=1}^{n} X_i}{p} - \frac{n - \Sigma_{i=1}^{n} X_i}{1-p} = 0$$

$$\therefore (1-p)\Sigma_{i=1}^{n} X_i = p(n - \Sigma_{i=1}^{n} X_i)$$

$$\therefore \Sigma_{i=1}^{n} X_i = pn$$

$$\therefore \hat{p}_{MLE} = \Sigma_{i=1}^{n} X_i / n$$

Maximum Likelihood Example 5:

Suppose that $X_1, X_2, \ldots X_n$ is iid Poisson with parameter λ

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$\therefore \log f(x) = x\log(\lambda) - \lambda - \log(x!)$$

$$\therefore l(\lambda) = \Sigma_{i=1}^{n}(X_i \log(\lambda) - \lambda - \log(X_i!))$$

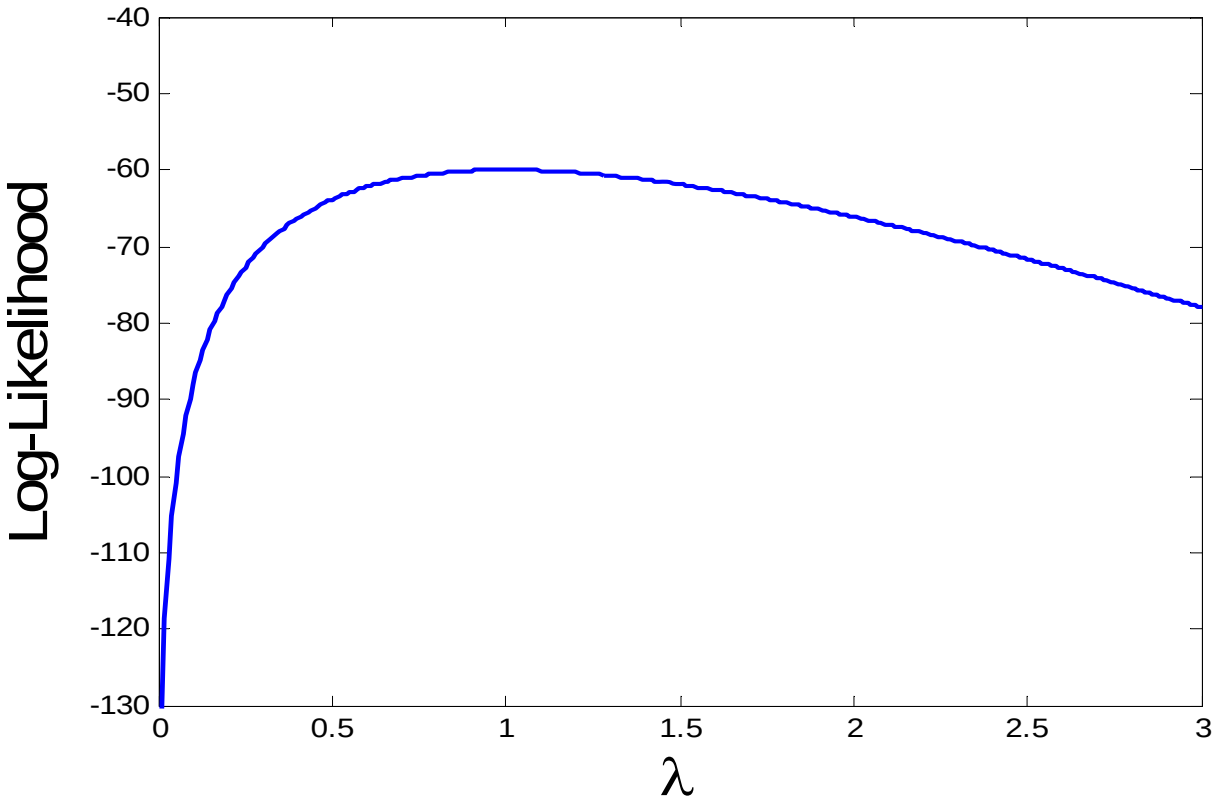
$$l'(\lambda) = \Sigma_{i=1}^{n}(\frac{X_i}{\lambda} - 1) = \frac{\Sigma_{i=1}^{n} X_i}{\lambda} - n = \frac{n\bar{X}}{\lambda} - n$$

$$l'(\lambda) = 0 \Rightarrow \frac{n\bar{X}}{\lambda} - n = 0 \Rightarrow \frac{n\bar{X}}{\lambda} = n \Rightarrow \frac{\bar{X}}{\lambda} = 1$$

$$\therefore \hat{\lambda}_{MLE} = \bar{X}$$

For the last example, suppose that $n = 20$, $\bar{X} = 1$ and $\Sigma_{i=1}^{n} \log(X_i!) = 40$.

Here's what the log-likelihood function looks like.



Often maximize log-likelihood numerically.

EM Algorithm

Method for numerically maximizing log-likelihood.

Sometimes we can find "missing data" such that
$f(y_i, \theta) = \int f(y_i, z_i, \theta) dz_i$ and $\Pi_{i=1}^{n} Ef(y_i, \theta \mid z_i)$ is easy to maximize.

EM Algorithm.
1. Take a draw of θ and work out the distribution of z.
2. Maximize $\Pi_{i=1}^{n} Ef(y_i, \theta \mid z_i)$ wrt θ.

On any iteration, the likelihood can only go up

Example. Suppose that $X_1, X_2, ...X_n$ is drawn from a mixture of normals, $N(\mu_0, 1)$ wp ½ and $N(\mu_1, 1)$ otherwise. There are two parameters: μ_0 and μ_1 and the log-likelihood is

$$\Sigma_{i=1}^{n} \log\{\frac{1}{2}[\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_i-\mu_0)^2}{2})]+\frac{1}{2}[\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_i-\mu_1)^2}{2})]\}$$

Let z_i be the indicator that x_i is $N(\mu_0, 1)$

$$f(x_i) = f(x_i, z_i = 1) + f(x_i, z_i = 0)$$

$$f(x_i, z_i) = \frac{1}{2}[\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_i-\mu_0)^2}{2})]^{z_i}[\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x_i-\mu_1)^2}{2})]^{1-z_i}$$

The "complete" log-likelihood is, apart from a constant:

$$\Sigma_{i=1}^{n} \log(f(x_i, z_i)) = -\frac{1}{2}\Sigma_{i=1}^{n} z_i (x_i-\mu_0)^2 - \frac{1}{2}\Sigma_{i=1}^{n}(1-z_i)(x_i-\mu_1)^2$$

EM Algorithm

1. For any μ_0 and μ_1

$$P(z_i = 1 \mid x_i, \mu_0, \mu_1) = \frac{f(x_i \mid z_i = 1, \mu_0, \mu_1)}{f(x_i \mid z_i = 1, \mu_0, \mu_1) + f(x_i \mid z_i = 0, \mu_0, \mu_1)}$$

$$= \frac{\phi(y_i, \mu_0, 1)}{\phi(y_i, \mu_0, 1) + \phi(y_i, \mu_1, 0)} = p_i$$

The expected log-likelihood is (apart from a constant)

$$-\frac{1}{2}\Sigma_{i=1}^{n} p_i (x_i - \mu_0)^2 - \frac{1}{2}\Sigma_{i=1}^{n} (1 - p_i)(x_i - \mu_1)^2$$

2. Maximizing this yields

$$\tilde{\mu}_0 = \frac{\Sigma_{i=1}^{n} p_i x_i}{\Sigma_{i=1}^{n} p_i}, \quad \tilde{\mu}_1 = \frac{\Sigma_{i=1}^{n} (1 - p_i) x_i}{\Sigma_{i=1}^{n} (1 - p_i)}$$

Bayesian Inference

Fundamentally different thought-experiment.
θ is not a fixed parameter; it is a random variable

Researchers beliefs about θ are summarized in a **prior** probability density function $f(\theta)$

By Bayes rule, $f(\theta \mid X_1,...X_n) = \dfrac{f(X_1,...X_n \mid \theta) f(\theta)}{f(X_1,...X_n)}$

$$\therefore f(\theta \mid X_1,...X_n) = \frac{f(X_1,...X_n \mid \theta) f(\theta)}{\int f(X_1,...X_n \mid \theta) f(\theta) d\theta}$$

$f(\theta \mid X_1,...X_n)$ is called the posterior density.

A Bayesian estimator of θ is just some moment of the posterior density.

Posterior mean: $\int \theta f(\theta \mid X_1,...X_n) d\theta$
Posterior mode: $\max_\theta f(\theta \mid X_1,...X_n)$

Bayesian researchers often report the whole posterior density, not just a moment of it

Bayesian Inference: Example 1.

$X_1, X_2, \ldots X_n$ are iid $N(\mu, \sigma^2)$ where σ^2 is known
Prior for μ is flat

Posterior distribution is:

$$\mu \mid X_1, \ldots X_n \sim N(\bar{X}, \frac{\sigma^2}{n})$$

Posterior mean (Bayes estimator of μ) is $\bar{X}$

Bayesian Inference: Example 2.

$X_1, X_2, \ldots X_n$ are iid $N(\mu, \sigma^2)$ where σ^2 is known

Prior: $\mu \sim N(\theta, \tau^2)$

Posterior distribution is:

$$\mu \mid X_1, \ldots X_n \sim N\left(\frac{n\tau^2}{n\tau^2+\sigma^2}\bar{X} + \frac{\sigma^2}{n\tau^2+\sigma^2}\theta, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right)$$

Posterior mean (Bayes estimator of μ) is

$$\boxed{\frac{n\tau^2}{n\tau^2+\sigma^2}\bar{X} + \frac{\sigma^2}{n\tau^2+\sigma^2}\theta}$$

Posterior mean (Bayes estimator of μ) is

$$\boxed{\frac{n\tau^2}{n\tau^2+\sigma^2}\bar{X}+\frac{\sigma^2}{n\tau^2+\sigma^2}\theta}$$

Intuition: Weighted average of $\bar{X}$ (the data) and θ (the prior belief).

- The larger is n, the more weight I put on the data
- The larger is τ^2, the more weight I put on the data
- The larger is σ^2, the more weight I put on the prior

In the limit as $n \to \infty$, the posterior mean of μ is $\bar{X}$ ….just the same thing as MLE.

Bayes estimator: Example 3.

X is a Bernoulli random variable with parameter p.
The prior for p is uniform on the unit interval.

$$f(p) = 1$$
$$f(x \mid p) = p^x (1-p)^{1-x}$$
$$P(X = 0) = P(X = 1) = 1/2$$

$$\therefore\ f(p \mid x) = \frac{p^x (1-p)^{1-x}}{1/2} = 2p^x (1-p)^{1-x}$$

If I observe X=1, then the posterior is $2p$.

The posterior mean is $\int_0^1 2p^2 dp = [\frac{2p^3}{3}]_0^1 = 2/3$

If I observe X=0, then the posterior is $2(1-p)$.

The posterior mean is

$$\int_0^1 2p(1-p)dp = \int_0^1 2p - 2p^2 dp = [p^2 - \frac{2p^3}{3}]_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

Bayes estimator: Example 4.

Again X is a Bernoulli random variable with parameter p.

But we want the prior for p to be more general. The prior for p is beta distributed with parameters α and β.

$E(p) = \dfrac{\alpha}{\alpha + \beta}$ is the prior mean

The uniform prior is a special case $\alpha = \beta = 1$

The posterior of p is beta with parameters $\alpha + X$ and $1 - X + \beta$.

So the posterior mean is $\dfrac{\alpha + X}{\alpha + X + 1 - X + \beta} = \dfrac{\alpha + X}{\alpha + 1 + \beta}$

If we observe $X = 1$, the posterior mean is above the prior mean.

If we observe $X = 0$, the posterior mean is below the prior mean.

Bayes estimator: Example 5.

X is a Binomial random variable with parameters n and p. We observe s successes.

n is known. The prior for p is beta distributed with parameters α and β.

The posterior of p is beta with params $\alpha + s$ and $n - s + \beta$.

So the posterior mean is $\dfrac{\alpha + s}{\alpha + n + \beta}$

As $n, s \to \infty$ this becomes s / n, the sample proportion of successes

"Conjugate" priors

Suppose that $X_1,...X_n$ are iid with a density $f(x,\theta)$. A prior for θ is said to be **conjugate** for f if the posterior for θ is the same type of density.

- For the normal density with known variance, the normal prior is conjugate

- For the binomial density, the beta prior is conjugate

Conjugate priors are often relatively easy to work with analytically.

Priors: Uninformative versus Informative

- Can be thought of as a “necessary evil” and trying to introduce as little information as possible

- Or can be informative, based on earlier studies

Priors: Proper and improper

A proper prior is a prior that is itself a well defined density (integrates to one).

But you can write down a prior that is not a well defined density (such as $p(\theta) = k$ where θ is unbounded which does not integrate to 1) and still apply Bayes rule and get a posterior density.

This could be useful as an uninformative prior.

It is called an improper prior

Gibbs sampling and computing the posterior

The cases listed above are ones in which the posterior is available in closed form.

Most of the time, that isn't the case: computational advances have made Bayesian approaches much more widely used.

Gibbs sampling (discussed earlier) is very useful in simulating posteriors. An example is in the HW.

Random Walk Metropolis Hastings Algorithm

Suppose I want to draw from a posterior $f(\theta) = f(\theta \mid Y)$ and don't have a way of breaking it down into conditionals.

1. Take some candidate θ_1.

2. Set $t = 2$ and define $\theta^* = \theta_{t-1} + \varepsilon_t$ where ε_t is an iid random disturbance (probably Gaussian).

3. Set $\theta_t = \theta^*$ w.p. $\alpha = \min(1, \frac{f(\theta^*)}{f(\theta_{t-1})})$ and $\theta_t = \theta_{t-1}$ otherwise.

4. Repeat (2) and (3) for $t = 3, 4, ...$

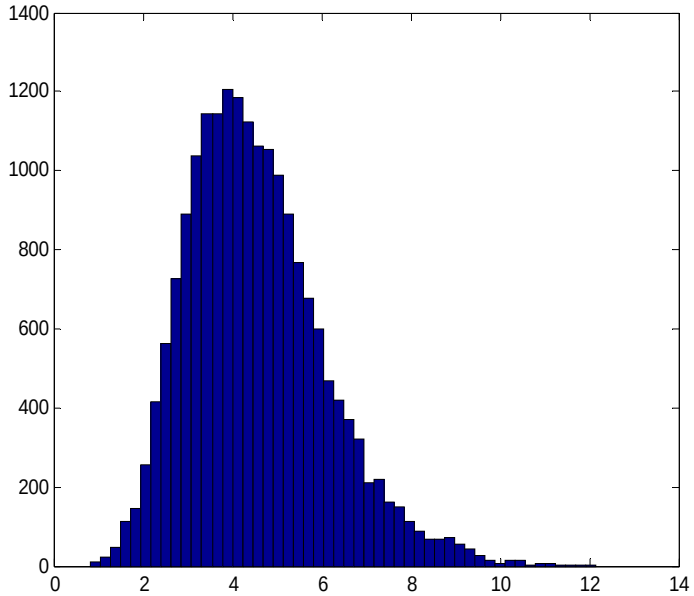
5. Discard the first (say) 1,000 draws. The remaining distribution is a random sample from the posterior.

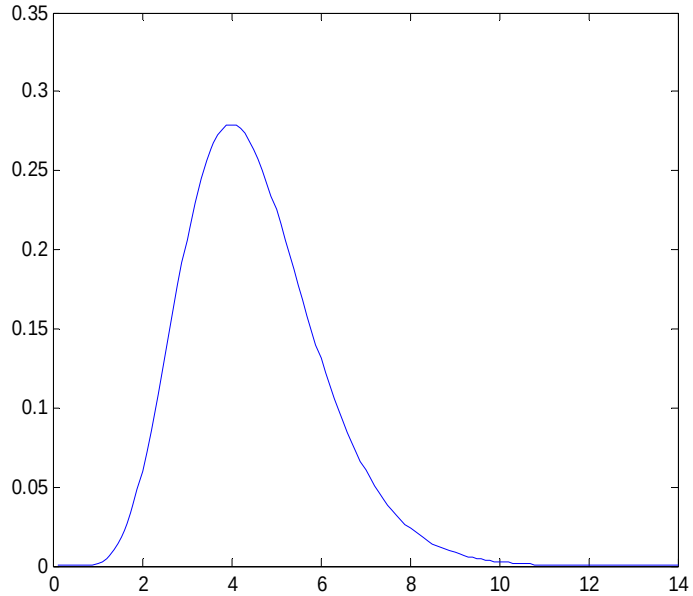
Simple Example of Metropolis-Hastings

Suppose I want to take draws from a posterior that is gamma (9,0.5). Here is code for doing it.

```
x(1)=1;
for t=2:20000;
    xstar=x(t-1)+randn(1,1);
    alpha=min(1,gampdf(xstar,9,0.5)/gampdf(x(t-1),9,0.5));
    if rand(1,1)<alpha; x(t)=xstar; else; x(t)=x(t-1); end;
end;
x=x(1001:end);
hist(x,50);
```

The resulting histogram and the actual gamma pdf:





Works well here and is a reliable general method for simulating posteriors.

Methods of evaluating estimators

Suppose that W is some estimator of a parameter θ.

- The **mean square error** of W is $E((W-\theta)^2)$
  - or $E(\|W-\theta\|^2)$ for a vector estimator

- The **bias** of W is $E(W-\theta)$

- If $E(W)=\theta$, then W is an **unbiased** estimator of θ.

The mean square error can be decomposed as

$$
\begin{aligned}
&E((W-\theta)^2) = E((W-E(W)+E(W)-\theta)^2) \\
&= E((W-E(W))^2 + (E(W)-\theta)^2 + 2(E(W)-\theta)E(W-E(W)) \\
&= E((W-E(W))^2 + (E(W)-\theta)^2 \\
&= Var(W) + (E(W)-\theta)^2
\end{aligned}
$$

MSE = Variance + Bias^2

Mean square error is the most common criterion for an estimator: want to select the W that mimimizes MSE.

But it isn't the only criterion.

Could pick an estimator to minimize any *loss* function

Mean absolute error loss: $E(|W-\theta|)$
"Utility based" loss function.

We say that W is the **minimum variance unbiased estimator (MVUE)** for a parameter θ if $E(W) = \theta$ and $Var(W) \leq Var(\tilde{W})$ where $\tilde{W}$ is any other unbiased estimator of θ.

Example: I have two estimators of a parameter θ.

$E(\hat{\theta}) = \theta - 0.1$ and $Var(\hat{\theta}) = 1$

$E(\tilde{\theta}) = \theta$ and $Var(\tilde{\theta}) = 2$

Q. Could $\hat{\theta}$ be the MVUE?

A. No. $\hat{\theta}$ is not even unbiased. But $\tilde{\theta}$ could be.

Example: Suppose that $X_1,...X_n$ are iid uniform from 0 to θ.
Consider the following estimators of θ: $\hat{\theta} = 2\bar{X}$

$$\tilde{\theta} = \max(X_1,...X_n)$$

What of these estimators could be the MVUE?

$$E(\hat{\theta}) = 2E(\bar{X}) = \frac{2}{n}E(\Sigma_{i=1}^{n}X_i) = \frac{2}{n}nE(X_i) = 2\frac{\theta}{2} = \theta.$$

So $\hat{\theta}$ is unbiased.

$$Var(\hat{\theta}) = 4Var(\bar{X}) = \frac{4}{n^2}Var(\Sigma_{i=1}^{n}X_i) = \frac{4}{n^2}nVar(X_i) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

$\tilde{\theta} = \max(X_1, \ldots X_n)$

The cdf of $\tilde{\theta}$ is $(x/\theta)^n$.

The pdf of $\tilde{\theta}$ is $n(\frac{x}{\theta})^{n-1}\frac{1}{\theta}$

$$\therefore E(\tilde{\theta}) = \int_0^\theta xn(\frac{x}{\theta})^{n-1}\frac{1}{\theta}dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{\theta^n}\frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta$$

So $\tilde{\theta}$ is <u>not</u> unbiased.

$$E(\tilde{\theta}^2) = \int_0^\theta x^2 n(\frac{x}{\theta})^{n-1}\frac{1}{\theta}dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1}dx = \frac{n}{\theta^n}\frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2$$

$$\therefore Var(\tilde{\theta}) = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \theta^2[\frac{n}{n+2} - \frac{n^2}{(n+1)^2}]$$

$$= \frac{\theta^2 n}{(n+1)^2(n+2)}$$

$Var(\hat{\theta}) = O(\frac{1}{n})$ and $Var(\tilde{\theta}) = O(\frac{1}{n^2})$

$\tilde{\theta}$ has a smaller variance than $\hat{\theta}$, but is biased.

$\tilde{\theta}$ cannot be the MVUE.

But what about a bias-adjusted counterpart? $\tilde{\theta}_{BA} = \tilde{\theta}\frac{(n+1)}{n}$

$E(\tilde{\theta}_{BA}) = \theta$ and

$$\therefore Var(\tilde{\theta}_{BA}) = \frac{(n+1)^2}{n^2}\frac{\theta^2 n}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$$

This *could* be the MVUE.

Optimal choice of estimators depends on the loss function.

Suppose that $\hat{\theta}$ is an estimator and that the loss function is $E((\hat{\theta}-\theta)^2)$.

Then clearly the estimator with minimum mean square error will be optimal.

But we could pick an estimator to minimize $E(|\hat{\theta}-\theta|)$ or an asymmetric loss function.

Admissible Estimators

An estimator $\hat{\theta}$ is said to be **admissible** if there does not exist any other estimator $\tilde{\theta}$ such that

$$L(\theta,\tilde{\theta}) < L(\theta,\hat{\theta}) \ \ \forall\, \theta$$

where $L(.,.)$ denotes the loss function that is being used. Most often, this is MSE loss.

Stein Estimator

Suppose that θ is an mx1 parameter vector, and $x_i \sim_{iid} N(\theta, \sigma^2 I)$ for i=1,2,…n.

Estimator of θ is $\hat{\theta} = \bar{x}$. This is the MVUE.

Now for any fixed μ consider a new estimator **for m>2**

$$\hat{\theta}_S = (1 - \frac{(m-2)\sigma^2}{n \| \bar{x} - \mu \|^2})\bar{x} + \frac{(m-2)\sigma^2}{n \| \bar{x} - \mu \|^2} \mu$$

It is biased, but amazing thing is: $MSE(\hat{\theta}_S) < MSE(\hat{\theta}) \ \forall \theta$
(improvement is biggest if μ is close to θ)

So MLE is not admissible for m>2
It turns out that MLE *is* admissible for m≤2

Admissibility of Bayes estimators

Suppose that we have a posterior density $f(\theta \mid X)$ and we select θ to solve

$$\hat{\theta}^* = \arg\min_{\hat{\theta}} \int L(\theta, \hat{\theta}) f(\theta \mid X) d\theta$$

For example

$$\hat{\theta}^* = \arg\min_{\hat{\theta}} \int (\theta - \hat{\theta})^2 f(\theta \mid X) d\theta \Rightarrow \hat{\theta}^* = \int \theta f(\theta \mid X) d\theta$$

This is Bayes estimator optimizing a particular loss function

Admissibility of Bayes estimators

Under some mild conditions we have:

Result 1: With a proper prior a Bayes estimator must be admissible

Result 2: Any admissible rule is Bayes for some prior distribution.

In contrast, MLE may not be admissible (see Stein estimator)

Sufficient statistics

A statistic $T(X)$ is a **sufficient statistic** for a parameter θ if the distribution of X given T does not depend on θ.

Intuition: A sufficient statistic contains all the information in the data for a parameter of interest.

Given the sufficient statistic, the data are irrelevant for telling us what the parameter is.

Example. If $X_1,...X_n$ are Bernoulli random variables, their joint pmf is

$$f(x_1, x_2,...x_n) = p^{\Sigma_{i=1}^n X_i}(1-p)^{n-\Sigma_{i=1}^n X_i}$$

Let $T = \Sigma_{i=1}^n X_i$

The conditional pmf of $X_1,...X_n$ given T is

$$f(x_1, x_2,...x_n \mid t) = 1/C_t^n \quad \text{if } t = \Sigma_{i=1}^n x_i$$

$$0 \quad \text{otherwise}$$

So the conditional distribution does not depend on p and so $\Sigma_{i=1}^n X_i$ is a sufficient statistic.

The factorization theorem.

If a joint pdf $f(x,\theta)$ can be factorized as

$$f(x,\theta) = g(t(x),\theta)h(x)$$

Then $T(X)$ is a sufficient statistic for θ.

Example. Suppose that $X_1, X_2, ...X_n$ are iid $N(\mu,\sigma^2)$ where σ^2 is known. The joint pdf is

$$\Pi_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$$

$$=(2\pi\sigma^2)^{-n/2}\exp(-\Sigma_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2})$$

$$= (2\pi\sigma^2)^{-n/2} \exp(-\Sigma_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2})$$

$$= (2\pi\sigma^2)^{-n/2} \exp(-\frac{\Sigma_{i=1}^{n} x_i^2 + n\mu^2 - 2\mu\Sigma_{i=1}^{n} x_i}{2\sigma^2})$$

$$= (2\pi\sigma^2)^{-n/2} \exp(\frac{-\Sigma_{i=1}^{n} x_i^2 - n\mu^2 + 2\mu n\bar{x}}{2\sigma^2})$$

$$= \underbrace{(2\pi\sigma^2)^{-n/2} \exp(\frac{2\mu n\bar{x} - n\mu^2}{2\sigma^2})}_{g(t(x),\theta)} \underbrace{\exp(-\frac{\Sigma_{i=1}^{n} x_i^2}{2\sigma^2})}_{h(x)}$$

By the factorization theorem, this means that $\bar{X}$ is a sufficient statistic for μ.

Techniques for finding the MVUE

Rao Blackwell Theorem

Let $W = W(X)$ be any unbiased estimator for θ and $T = T(X)$ be a sufficient statistic for θ. Then

$$\tilde{W} = E(W \mid T)$$

must have smaller variance than W. Under further conditions, it is the MVUE of θ.

So we just have to look for *any* unbiased estimator and take its conditional expectation to find the MVUE.

Example of Rao-Blackwell Theorem.

Suppose that $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$ where σ^2 is known.

$E(X_1) = \mu$ so X_1 is an unbiased, if ridiculous, estimator for μ.

$\bar{X}$ is a sufficient statistic for μ

$E(X_1 \mid \bar{X}) = \bar{X}$ (intuitive, but formal derivation next slide)

By Rao-Blackwell, this means $\bar{X}$ is the MVUE of μ

Formal proof that $E(X_1 \mid \bar{X}) = \bar{X}$:

$$\begin{pmatrix} \bar{X} \\ X_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2/n & \sigma^2/n \\ \sigma^2/n & \sigma^2 \end{pmatrix}\right)$$

$$Corr(X_1, \bar{X}) = 1/\sqrt{n}$$

$$\therefore X_1 \mid \bar{X} \sim N\left(\mu + \frac{(1/\sqrt{n})\sigma}{\sigma/\sqrt{n}}(\bar{X} - \mu), \sigma^2\left(1 - \frac{1}{n}\right)\right)$$

$$\therefore E(X_1 \mid \bar{X}) = \mu + \bar{X} - \mu = \bar{X}$$

$$\therefore E(X_1 \mid \bar{X}) = \bar{X}$$

Cramer-Rao Bound

Let X be one or more random variables with a joint pdf $f(x,\theta)$. Let W be *any* unbiased estimator of θ. Under suitable regularity conditions

$$Var(W) \geq \frac{1}{E\{[\frac{\partial \log(f(x,\theta_0))}{\partial \theta}]^2\}}$$

where θ_0 denotes the true parameter value.

Or, in the vector case

$$Var(W) \geq [E\{\frac{\partial \log(f(x,\theta_0))}{\partial \theta}\frac{\partial \log(f(x,\theta_0))}{\partial \theta'}\}]^{-1}$$

An estimator that reaches the Cramer-Rao bound is MVUE.

Example: Let X be a single exponential random variable with pdf $f(x,\lambda)=\frac{1}{\lambda}\exp(-\frac{x}{\lambda})$. What is the Cramer-Rao bound for an estimator of λ?

$$f(x,\lambda)=\frac{1}{\lambda}\exp(-\frac{x}{\lambda}) \Rightarrow \log f(x,\lambda)=-\log(\lambda)-\frac{x}{\lambda}$$

$$\frac{\partial \log f(x,\lambda)}{\partial \lambda}=-\frac{1}{\lambda}+\frac{x}{\lambda^2}=\frac{(x-\lambda)}{\lambda^2}$$

$$\therefore E\{[\frac{\partial \log f(x,\lambda)}{\partial \lambda}]^2\}=E\{\frac{(x-\lambda)^2}{\lambda^4}\}=\frac{1}{\lambda^4}E(x-\lambda)^2$$

$E(x)=\lambda$ and $Var(x)=\lambda^2$

$$\therefore E\{[\frac{\partial \log f(x,\lambda)}{\partial \lambda}]^2\} = \frac{1}{\lambda^4}\lambda^2 = \frac{1}{\lambda^2}$$

So the Cramer-Rao bound is $\frac{1}{1/\lambda^2} = \lambda^2$

Does the method of moments estimator reach the bound?

$X = \hat{\lambda}$

$Var(X) = \lambda^2$

This estimator <u>does</u> attain the bound.

Sketch of the derivation of the CR bound

Let $W(x)$ be any estimator of θ and $V = \frac{\partial \log(f(x,\theta))}{\partial\partial}$

$$E(VW) = \int W(x) \frac{\partial \log(f(x,\theta))}{\partial\partial} f(x,\theta) dx = \int W(x) \frac{\partial f(x,\theta)}{\partial\partial} dx = \frac{\partial}{\partial\partial} E(W(x))$$

Suppose that $W(x)$ is unbiased.

Then $E(W(x)) = \theta$ and $\frac{\partial}{\partial\partial} E(W(x)) = 1$

$$Cov(V,W) = E(VW) - E(V)E(W) = E(VW) = 1 \qquad \text{(E(V)=0)}$$

From the Cauchy-Schwarz Inequality

$$Var(W(x)) \geq \frac{Cov(V,W)}{Var(V)} = \frac{1}{Var(V)} = \frac{1}{E[(\frac{\partial \log(f(x,\theta))}{\partial\theta})^2]}$$

Cramer-Rao Bound for iid random variables

Suppose that $X_1, X_2, ...X_n$ are iid random variables each with the same pdf $f(x,\theta)$. Let W be *any* unbiased estimator of θ. Under suitable regularity conditions

$$Var(W) \geq \frac{1}{nE\{[\frac{\partial \log(f(x,\theta_0))}{\partial \theta}]^2\}}$$

where θ_0 denotes the true parameter value.

Or, in the vector case

$$Var(W) \geq \frac{1}{n}[E\{\frac{\partial \log(f(x,\theta_0))}{\partial \theta}\frac{\partial \log(f(x,\theta_0))}{\partial \theta'}\}]^{-1}$$

Is the bound for *n* iid variables consistent with the more general bound? *Y*es.

$$E\{[\frac{\partial \log(f(x_1,...x_n,\theta_0))}{\partial\theta}]^2\} = E\{[\Sigma_{i=1}^{n}\frac{\partial \log(f(x_i,\theta_0))}{\partial\theta}]^2\}$$

$$= \Sigma_{i=1}^{n} E\{\frac{\partial \log(f(x_i,\theta_0))^2}{\partial\theta}\}$$

$$+\Sigma_{i=1}^{n}\Sigma_{j=1,\,j\neq i}^{n} E\{\frac{\partial \log(f(x_i,\theta_0))}{\partial\theta}\frac{\partial \log(f(x_j,\theta_0))}{\partial\theta}\}$$

$$= nE\{\frac{\partial \log(f(x_i,\theta_0))^2}{\partial\theta}\}$$

Example 1: Suppose that $X_1, X_2, ...X_n$ are iid random variables each with a $N(\mu, \sigma^2)$ distribution where σ^2 is known. What is the Cramer-Rao bound for an estimator of μ?

$$\log(f(x,\mu)) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\therefore \frac{\partial \log(f(x,\mu))}{\partial \mu} = \frac{1}{\sigma^2}(x-\mu)$$

$$\therefore E\{[\frac{\partial \log(f(x,\mu))}{\partial \mu}]^2\} = \frac{1}{\sigma^4}E((x-\mu)^2) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

So the Cramer-Rao bound for an estimator of μ is

$$\frac{1}{n(1/\sigma^2)} = \frac{\sigma^2}{n}$$

In this case, the method of moments estimator and the MLE are both $\bar{X}$.

Both attain the Cramer-Rao bound because $Var(\bar{X}) = \sigma^2 / n$.

Example 2: Suppose that $X_1, X_2, ...X_n$ are iid random variables each with a $N(\mu, \sigma^2)$ distribution where μ and σ^2 are unknown. What is the Cramer-Rao bound for an estimator of $\theta = (\mu, \sigma^2)'$?

$$\log(f(x,\theta)) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\therefore \frac{\partial \log(f(x,\theta))}{\partial \mu} = \frac{1}{\sigma^2}(x-\mu)$$

$$\frac{\partial \log(f(x,\theta))}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2$$

$$\therefore E\{[\frac{\partial \log(f(x,\theta))}{\partial \mu}]^2\} = \frac{1}{\sigma^4}E((x-\mu)^2) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

$$\therefore E\{[\frac{\partial \log(f(x,\theta))}{\partial \sigma^2}]^2\} = \frac{1}{4\sigma^4} + \frac{E((x-\mu)^4)}{4\sigma^8} - \frac{2E((x-\mu)^2)}{4\sigma^6}$$

$$= \frac{1}{4\sigma^4} + \frac{3\sigma^4}{4\sigma^8} - \frac{2\sigma^2}{4\sigma^6} = \frac{1+3-2}{4\sigma^4} = \frac{1}{2\sigma^4}$$

$$\therefore E\{\frac{\partial \log(f(x,\theta))}{\partial \mu}\frac{\partial \log(f(x,\theta))}{\partial \sigma^2}\} = -\frac{E(x-\mu)}{2\sigma^4} + \frac{E((x-\mu)^3)}{2\sigma^6} = 0$$

So the Cramer-Rao bound for an estimator of θ is

$$\frac{1}{n}\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}$$

Example 3 Suppose that $X_1, X_2, ...X_n$ is iid Poisson with parameter λ. What is the Cramer-Rao bound for an estimator of λ?

$$f(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$\therefore \log f(x,\lambda) = x\log(\lambda) - \lambda - \log(x!)$$

$$\therefore \frac{\partial \log f(x,\lambda)}{\partial \lambda} = \frac{x}{\lambda} - 1$$

$$\therefore E\{[\frac{\partial \log f(x,\lambda)}{\partial \lambda}]^2\} = E\{[\frac{x}{\lambda} - 1]^2\} = \frac{E(X^2)}{\lambda^2} - 2\frac{E(X)}{\lambda} + 1$$

$$= \frac{Var(X) + E(X)^2}{\lambda^2} - 2\frac{E(X)}{\lambda} + 1 = \frac{\lambda + \lambda^2}{\lambda^2} - 2\frac{\lambda}{\lambda} + 1 = \frac{1}{\lambda}$$

So the Cramer-Rao bound for an estimator of λ is $\frac{\lambda}{n}$

Under regularity conditions, we have the "information equality"

$$\boxed{E\{[\frac{\partial \log(f(x,\theta))}{\partial \theta}]^2\} = -E\{\frac{\partial^2 \log(f(x,\theta))}{\partial \theta^2}\}}$$

This means that the Cramer-Rao bound can also be written as

$$Var(W) \geq -\frac{1}{nE\{\frac{\partial^2 \log(f(x,\theta))}{\partial \theta^2}\}}$$

Example 1: Again $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$ random variables where σ^2 is known. What is the Cramer-Rao bound for an estimator of μ using the information equality?

$$\log(f(x,\mu)) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\therefore \frac{\partial \log(f(x,\mu))}{\partial \mu} = \frac{1}{\sigma^2}(x-\mu)$$

$$\therefore \frac{\partial^2 \log(f(x,\mu))}{\partial \mu^2} = -\frac{1}{\sigma^2} \Rightarrow E\{\frac{\partial^2 \log(f(x,\mu))}{\partial \mu^2}\} = -\frac{1}{\sigma^2}$$

So the Cramer-Rao bound for an estimator of μ is (as before)

$$-\frac{1}{n*(-1/\sigma^2)} = \frac{\sigma^2}{n}$$

Example 2: Suppose that $X_1, X_2, \ldots X_n$ are iid random variables each with a $N(\mu, \sigma^2)$ distribution with known μ. What is the Cramer-Rao bound for an estimator of σ^2?

$$\log(f(x,\sigma^2)) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\therefore \frac{\partial \log(f(x,\sigma^2))}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2$$

$$\therefore \frac{\partial^2 \log(f(x,\sigma^2))}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x-\mu)^2$$

$$\therefore E\{\frac{\partial^2 \log(f(x,\sigma^2))}{\partial (\sigma^2)^2}\} = \frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} = -\frac{1}{2\sigma^4}$$

So the Cramer Rao bound is $-\frac{1}{n*(-1/2\sigma^4)} = \frac{2\sigma^4}{n}$

Sketch of proof of information equality

$$E\left(\frac{\partial \log(f(x,\theta))}{\partial \theta}\right) = \int \frac{\partial \log(f(x,\theta))}{\partial \theta} f(x,\theta)dx$$

$$= \int \frac{1}{f(x,\theta)} \frac{\partial f(x,\theta)}{\partial \theta} f(x,\theta)dx = \int \frac{\partial f(x,\theta)}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int f(x,\theta)dx$$

$$= \frac{\partial 1}{\partial \theta} = 0$$

Differentiating both sides

$$\frac{\partial}{\partial \theta} E\left(\frac{\partial \log(f(x,\theta))}{\partial \theta}\right) = 0$$

But

$$\frac{\partial}{\partial \theta} E\left(\frac{\partial \log(f(x,\theta))}{\partial \theta}\right) = \frac{\partial}{\partial \theta} \int \left[\frac{\partial \log(f(x,\theta))}{\partial \theta} f(x,\theta)\right] dx$$

$$= \int \frac{\partial}{\partial\theta}[\frac{\partial \log(f(x,\theta))}{\partial\theta} f(x,\theta)]dx$$

$$= \int \frac{\partial^2 \log(f(x,\theta))}{\partial\theta^2} f(x,\theta) + \frac{\partial \log(f(x,\theta))}{\partial\theta}\frac{\partial f(x,\theta)}{\partial\theta}dx$$

$$= \int \frac{\partial^2 \log(f(x,\theta))}{\partial\theta^2} f(x,\theta) + \frac{\partial \log(f(x,\theta))}{\partial\theta}\frac{\partial \log(f(x,\theta))}{\partial\theta} f(x,\theta)dx$$

$$= E(\frac{\partial^2 \log(f(x,\theta))}{\partial\theta^2}) + E\{[\frac{\partial \log(f(x,\theta))}{\partial\theta}]^2\}$$

So

$$E(\frac{\partial^2 \log(f(x,\theta))}{\partial\theta^2}) + E\{[\frac{\partial \log(f(x,\theta))}{\partial\theta}]^2\} = 0$$

$$\boxed{\Rightarrow E(\frac{\partial^2 \log(f(x,\theta))}{\partial\theta^2}) = -E\{[\frac{\partial \log(f(x,\theta))}{\partial\theta}]^2\}}$$

If we find that an estimator reaches the Cramer-Rao bound, we know that there is no other unbiased estimator with smaller variance.

But if it doesn't then there *could* be a better estimator, though there doesn't have to be.

But we have an important result. An unbiased estimator with variance equal to the Cramer-Rao bound exists if and only if

$$\frac{\partial \log(f(x,\theta))}{\partial \theta} = a(\theta)(W(x)-\theta)$$

for some functions $a(.)$ and $W(.)$. Moreover, $W(X)$ is the MVUE and the MLE.

Example: Suppose that $X_1, X_2, ...X_n$ are iid random variables each with a $N(\mu, \sigma^2)$ distribution where σ^2 is known.

$$\log(f(x,\mu)) = -\frac{1}{2}\log(2\pi\sigma^2) - \Sigma_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\therefore \frac{\partial \log(f(x,\mu))}{\partial \mu} = \frac{1}{\sigma^2}\Sigma_{i=1}^{n}(x_i-\mu) = \frac{n(\overline{x}-\mu)}{\sigma^2} = \frac{n}{\sigma^2}(\overline{x}-\mu)$$

So $\overline{X}$ reaches the Cramer-Rao bound.

It is the MVUE and the MLE.

(we already knew this, but this is another way of showing it)

Cramer-Rao intuition

$\log(f(x,\theta))$ is just the log-likelihood function

The steeper it is near θ_0, the more accurate an estimator should be.

A high value of $-\frac{\partial^2 \log(f(x,\theta_0))}{\partial \theta^2}$ means a steep log-likelihood and a small variance estimator

Large sample properties of estimators

Suppose that $\{W_n\}$ is a sequence of estimators that is a function of data $\{X_1, X_2, ...X_n\}$ that are iid with pdf $f(x,\theta)$.

$\{W_n\}$ is said to be **consistent** for θ if $W_n \rightarrow_p \theta$.

If $\sqrt{n}(W_n - \theta) \rightarrow_d N(0,V)$, then V is the **asymptotic variance** of $\{W_n\}$.

$\{W_n\}$ is said to be **asymptotically efficient** for θ if $\sqrt{n}(W_n - \theta) \rightarrow_d N(0,V)$ where

$$V = \frac{1}{E\{[\frac{\partial \log f(x,\theta)}{\partial \theta}]^2\}}$$

which is the Cramer-Rao bound

If $\{V_n\}$ and $\{W_n\}$ are two alternative estimators of θ such that $\sqrt{n}(V_n - \theta) \rightarrow_d N(0, \sigma_V^2)$ and $\sqrt{n}(W_n - \theta) \rightarrow_d N(0, \sigma_W^2)$, then the **asymptotic relative efficiency** of $\{V_n\}$ relative to $\{W_n\}$ is $\frac{\sigma_W^2}{\sigma_V^2}$.

Useful properties of maximum likelihood estimators

Suppose $\{X_1, X_2, ...X_n\}$ that are iid with pdf $f(x,\theta)$ and let

$$\tilde{\theta} = \arg\max_\theta \Sigma_{i=1}^n \log f(X_i,\theta)$$

denote the MLE. Then we have

- Consistency. $\tilde{\theta} \to_p \theta$
- Asymptotic distribution. $\sqrt{n}(\tilde{\theta} - \theta) \to_d N(0, 1/I)$ where

$$\boxed{I = E\{[\frac{\partial \log f(x,\theta)}{\partial \theta}]^2\} = -E\{\frac{\partial^2 \log f(x,\theta)}{\partial \theta^2}\}}.$$

- The MLE is asymptotically efficient.
- Invariance. $\tau(\tilde{\theta})$ is the MLE of $\tau(\theta)$ for any function $\tau(.)$

Asymptotic distribution of MLE $\sqrt{n}(\tilde{\theta}-\theta) \rightarrow_d N(0,1/I)$ where

$$\boxed{I = E\{[\frac{\partial \log f(x,\theta)}{\partial \theta}]^2\} = -E\{\frac{\partial^2 \log f(x,\theta)}{\partial \theta^2}\}}.$$

We "flip this around" to get the approximate variance of $\tilde{\theta}$.

$\tilde{\theta}$ is approximately $N(\theta, \frac{1}{n} I^{-1})$

But what we call the "asymptotic variance" of $\tilde{\theta}$ is I^{-1}.

Example 1: Suppose that $X_1, X_2, ...X_n$ is iid $N(\mu, \sigma^2)$ where σ^2 is known.

$$f(x,\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

We know that the MLE is $\bar{X}$

$$\log f(x,\mu) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\therefore \frac{\partial \log(f(x,\mu))}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

$$\therefore I = E\{[\frac{\partial \log(f(x,\mu))}{\partial \mu}]^2\} = \frac{1}{\sigma^4} E((x-\mu)^2) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

$$\therefore \sqrt{n}(\bar{X} - \mu) \rightarrow_d N(0, \sigma^2)$$

Example 2: Suppose that $X_1, X_2, \ldots X_n$ is iid Poisson.

$$f(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

We know that the MLE is $\bar{X}$

$$\log(f(x,\lambda)) = x\log(\lambda) - \lambda - \log(x!)$$

$$\therefore \frac{\partial \log(f(x,\lambda))}{\partial \lambda} = \frac{x}{\lambda} - 1$$

$$\therefore I = E\{[\frac{\partial \log(f(x,\lambda))}{\partial \mu}]^2\} = E((\frac{x}{\lambda} - 1)^2) = \frac{E(x^2)}{\lambda^2} - \frac{2}{\lambda}E(x) + 1$$

$$= \frac{Var(x) + E(x)^2}{\lambda^2} - \frac{2}{\lambda}E(x) + 1 = \frac{\lambda + \lambda^2}{\lambda^2} - \frac{2}{\lambda}\lambda + 1 = \frac{1}{\lambda}$$

$$\therefore \sqrt{n}(\bar{X} - \lambda) \to_d N(0,\lambda)$$

Example 3: Suppose that $\{X_1, X_2, ...X_n\}$ is iid Bernoulli with parameter p

$$f(x, p) = p^x (1-p)^{1-x}$$

We know that the MLE is $\bar{X}$

$$\log(f(x, p)) = x\log(p) + (1-x)\log(1-p)$$

$$\frac{\partial \log(f(x, p))}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x(1-p) - p(1-x)}{p(1-p)}$$

$$= \frac{x - xp - p + px}{p(1-p)} = \frac{x-p}{p(1-p)}$$

$$\therefore E\{[\frac{\partial \log(f(x, p))}{\partial p}]^2\} = \frac{E(x-p)^2}{p^2(1-p)^2} = \frac{E(x^2) - 2pE(x) + p^2}{p^2(1-p)^2}$$

$$= \frac{p - 2p^2 + p^2}{p^2(1-p)^2} = \frac{p-p^2}{p^2(1-p)^2} = \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)}$$

$$\therefore \sqrt{n}(\bar{X} - p) \rightarrow_d N(0, p(1-p))$$

We knew this before, but now can derive it from a general formula for the MLE distribution.

Sketch of the proof that $\sqrt{n}(\tilde{\theta}-\theta) \rightarrow_d N(0, I^{-1})$

$$0 = l'(\tilde{\theta}) \approx l'(\theta) + (\tilde{\theta}-\theta) l''(\theta) \Rightarrow \tilde{\theta}-\theta \approx -l''(\theta)^{-1} l'(\theta)$$

$$\therefore \sqrt{n}(\tilde{\theta}-\theta) \approx [-\frac{1}{n} l''(\theta)]^{-1} \frac{1}{\sqrt{n}} l'(\theta)$$

$$\frac{1}{n} l''(\theta) = \frac{1}{n} \Sigma_{i=1}^{n} \frac{\partial^2 \log f(x_i, \theta)}{\partial \theta^2} \rightarrow_p E\{\frac{\partial^2 \log f(x, \theta)}{\partial \theta^2}\}$$

$$-\frac{1}{n} l''(\theta) \rightarrow_p -E\{\frac{\partial^2 \log f(x, \theta)}{\partial \theta^2}\}$$

$$\frac{1}{\sqrt{n}} l'(\theta) = \frac{1}{\sqrt{n}} \Sigma_{i=1}^{n} \frac{\partial \log(f(x_i, \theta))}{\partial \theta} \rightarrow_d N(0, E\{[\frac{\partial \log f(x, \theta)}{\partial \theta}]^2\})$$

But $I = E\{[\frac{\partial \log f(x, \theta)}{\partial \theta}]^2\} = -E\{\frac{\partial^2 \log f(x, \theta)}{\partial \theta^2}\}$

$$\therefore \sqrt{n}(\tilde{\theta}-\theta) \rightarrow_d N(0, I^{-1} I I^{-1}) = N(0, I^{-1})$$

Multivariate case

If θ is a vector $\sqrt{n}(\tilde{\theta}-\theta) \rightarrow_d N(0, I^{-1})$ where

$$\boxed{I = E\{[\frac{\partial \log f(x,\theta)}{\partial \theta}][\frac{\partial \log f(x,\theta)}{\partial \theta}]'\} = -E\{\frac{\partial^2 \log f(x,\theta)}{\partial \theta \partial \theta'}\}}.$$

Example: Suppose that $X_1, X_2, ...X_n$ is iid $N(\mu, \sigma^2)$
The MLE of μ is $\bar{X}$ and the MLE of σ^2 is $n^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2$

$$I = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}$$

$$\therefore \sqrt{n}(\tilde{\theta}-\theta) \rightarrow_d N(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix})$$

Pseudo-maximum likelihood estimation (PMLE)

Maximum-likelihood estimation requires the probability density to be completely specified.

What if it is wrong?

For example, we assume that $\{X_1, X_2, ...X_n\}$ are iid $N(\mu, \sigma^2)$ when in fact they are just iid with mean μ and variance σ^2?

Maximizing the Gaussian log-likelihood is still an estimator, but it is the quasi- or pseudo- maximum likelihood estimator.

Theorem. If $\tilde{\theta}$ is the pseudo-maximum likelihood estimator of θ, then $\tilde{\theta} \rightarrow_p \theta$ (i.e. is consistent) and

$$\sqrt{n}(\tilde{\theta} - \theta) \rightarrow_d N(0, A^{-1}BA^{-1})$$

where $A = E\{\frac{\partial^2 \log f(x,\theta)}{\partial \theta^2}\}$ and $B = E\{[\frac{\partial \log f(x,\theta)}{\partial \theta}]^2\}$.

But when the density is misspecified, the information equality does <u>not</u> hold, so $A \neq -B$.

Estimating the mean of a population (again)

Suppose that $X_1, X_2, ...X_n$ is iid with pdf $f(x)$, mean μ and variance σ^2. We want to estimate μ.

A natural estimator is $\bar{X}$

$$\sqrt{n}(\bar{X} - \mu) \rightarrow_d N(0, \sigma^2)$$

An alternative estimator is X_m, the median of $X_1, X_2, ...X_n$

$$\sqrt{n}(X_m - \mu) \rightarrow_d N(0, \frac{1}{4 f(\mu)^2})$$

Asymptotic distribution depends on data distribution

Now suppose that the data are normal, which means that $\bar{X}$ is the MLE

$$f(x) = \frac{1}{(2\pi\sigma^2)^{-1/2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

$$\therefore f(u) = \frac{1}{(2\pi\sigma^2)^{-1/2}}$$

$$\sqrt{n}(\bar{X} - \mu) \rightarrow_d N(0, \sigma^2)$$

$$\sqrt{n}(X_m - \mu) \rightarrow_d N(0, \frac{2\pi\sigma^2}{4}) = N(0, \frac{\pi}{2}\sigma^2)$$

So the asymptotic efficiency of $\bar{X}$ relative to X_m is $\pi / 2 \approx 1.57$.

Q. So why would anyone use the sample median rather than the sample mean?

A. The sample median may be more efficient with certain heavy-tailed distributions.

For example, suppose that $f(x) = \frac{1}{\sqrt{2}\sigma}\exp(-\frac{|x-\mu|}{\sigma/\sqrt{2}})$

$E(X) = \mu$ and $Var(X) = \sigma^2$ (earlier in lectures and HW)

$f(\mu) = \frac{1}{\sqrt{2}\sigma}$

$\sqrt{n}(\bar{X} - \mu) \rightarrow_d N(0, \sigma^2)$

$\sqrt{n}(X_m - \mu) \rightarrow_d N(0, \frac{2\sigma^2}{4}) = N(0, \frac{\sigma^2}{2})$

So the asymptotic efficiency of $\bar{X}$ relative to X_m is 0.5.

Now the median is *more* efficient

Asymptotic distribution of extremum estimates

Suppose that $\hat{\theta} = \arg\min_{\theta} Q_n(\theta)$

Then we say that $\hat{\theta}$ is an extremum estimate (defined to maximize of minimize an objective function)

Most estimators are extremum estimates (e.g. MLE, PMLE)

If $n^{-1/2} \frac{\partial Q_n(\theta_0)}{\partial \theta} \rightarrow_d N(0, D)$ and $n^{-1} \frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'} \rightarrow_p E$ then

$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, E^{-1} D E^{-1})$

Sketch of the proof of the asymptotic distribution of extremum estimates

$$0 = Q_n'(\hat{\theta}) \approx Q_n'(\theta_0) + (\hat{\theta} - \theta) Q_n''(\theta_0)$$

$$\therefore Q_n'(\theta_0) + (\hat{\theta} - \theta) Q_n''(\theta_0) = 0$$

$$\therefore \sqrt{n}(\hat{\theta} - \theta) = -[\frac{1}{n} Q_n''(\theta_0)]^{-1} n^{-1/2} Q_n'(\theta_0)$$

$$\therefore \sqrt{n}(\hat{\theta} - \theta) = \rightarrow_d N(0, E^{-1} D E^{-1})$$

Hypothesis testing.

We have seen how we can use sample information to make "guesses" about population parameters through the use of appropriate estimators.

We will now investigate what to do when we want to test hypotheses (claims) about population parameters.

Setup: We have a claim about a **population** parameter that we want to assess using a randomly chosen **sample**.

Examples:

- A coin is fair (p=1/2)
- A treatment has no effect (mean effect of drug is zero, on average for the human population)
- Average returns on bonds and stocks are the same ($\mu_{STOCK} = \mu_{BOND}$, on average over all time periods)
- The height of the average human (in the world population) is 2 feet.

Claim: The average human is 2 feet tall.

Randomly chosen sample of 10 humans

6’2	6’0	5’11	5’10	5’8
5’7	5’7	5’5	5’3	5’1

Logically, is it possible that the average human is 2 feet tall (using this information alone)? Yes.

Do we want to believe the claim? Seems suspicious.

Making this more precise is what hypothesis testing is about.

Setup is that there is a parameter θ, a null hypothesis (usually $\theta = \theta_0$) and some alternative (usually $\theta \neq \theta_0$).

In assessing the claim, there are four possible outcomes

		Truth	
		True	False
Decision	Accept Claim	Correct Decision	Type 2 Error
	Reject Claim	Type 1 Error	Correct Decision

Statistics has a particular way of testing the claim.

1. Pick a test statistic (function of the data and parameter).
2. Tentatively suppose that the hypothesis is true.
3. Work out the sampling distribution of the test statistic if the hypothesis really is true.
4. Identify a rejection region C_X which is a set of values of the test statistic which has small probability of occurring if the hypothesis is true but large probability under the alternative.
5. Then we look at the test statistic from our sample, reject the hypothesis if it is C_X and accept otherwise.

The probability of Type 1 error is the probability of being in C_X if the null hypothesis is true.

This is the **size** or **significance level** of the test.

We set C_X so that the size is some pre-set level either in all sample sizes (exact test) or in the limit as the sample size goes to infinity (asymptotic test).

The **power** is the probability of rejecting a false hypothesis. It is 1 minus the probability of type 2 error.

Example 1. A college admits 25 percent of applicants. In a pool of 20 Canadian applicants, 3 were admitted. Test the hypothesis that the probability of acceptance for Canadians is 0.25. Let the alternative be that the probability is less than 0.25.

Suppose the hypothesis is true.
X, the number of applicants admitted is binomial with parameters 20 and 0.25.

$$P(X = 0) = C_0^{20} 0.25^0 0.75^{20} = 0.0032$$

$$P(X = 1) = C_1^{20} 0.25^1 0.75^{19} = 0.0211$$

$$P(X = 2) = C_2^{20} 0.25^1 0.75^{19} = 0.0669$$

etc.

A natural shape for the critical regions is $C_X=\{0,1,\ldots K\}$

If $C_X=\{0,1\}$ then the probability of being in C_X under the null is 0.0243.

So the test accepts.

We could define C_X as $\{0,1\}$ and including 2 with probability 0.38.

But again the test accepts.

The power of the test.

Stick with the rejection region $C_X = \{0,1\}$.

The power of the test is the probability of rejection if the true probability of a Canadian being admitted is p.

$$\text{Power}(p) = (1-p)^n + C_1^{20} p(1-p)^{n-1}$$

Here is the power for some values of p

p	Power(p)
0.25	0.0243
0.1	0.3917
0.05	0.7358

Example 2. We have a random sample $X_1,...X_n$ drawn from a population that is normal with unknown mean μ and known variance σ^2.

If $\mu = \mu_0$ then $\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$ is $N(0,1)$

Rejection rule: Reject if $\left| \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| > 1.96$

An auto manufacturer claims that gas mileage for a new model averages 31.5 mpg. The population is normally distributed with standard deviation 2.4 mpg. A random sample of 16 cars has been taken. The sample mean is $\bar{X} = 30.6$ mpg. We want to test the null hypothesis that the population mean is 31.5 mpg with a 5% significance level.

Suppose that the claim is true.

Rejection rule: Reject if $\left|\frac{\bar{X}-31.5}{2.4/\sqrt{16}}\right|>1.96$

Reject if $\bar{X}$<30.324 or>32.676

In our sample, $\bar{X}$ = 30.6 mpg.

Claim is accepted

Interpretation: the sample mean is 30.6 which is not the claim, 31.5. However, it is close enough to the claim that we can believe that the difference can be explained by chance alone.

The higher is the significance level

- The more likely we are to reject a hypothesis

- The more likely we are to make Type 1 errors (reject a true hypothesis)

- The less likely we are to make Type 2 errors (accept a false hypothesis)

- If $\bar{X} = \mu_0$, then there is no significance level at which we reject the hypothesis.

- If $\bar{X} \neq \mu_0$, there will be some cutoff significance level, $\bar{\alpha}$, such that the hypothesis will be **rejected** at all significance levels **higher** than $\bar{\alpha}$ and **accepted** at all levels **lower** than $\bar{\alpha}$.

- $\bar{\alpha}$ solves the equation

$$\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| = z_{\bar{\alpha}/2}$$

where $\bar{X}$ is the observed sample mean.

In our example, $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$=-1.5

P(N(0,1)>1.5)=0.067

So $\bar{\alpha}$=0.134.

The hypothesis will be rejected at all significance levels higher than 13.4 percent and accepted at all levels lower than 13.4 percent.

$\bar{\alpha}$ is called the **p-value** of the test.

Definition: The **p-value** is the significance level at which we are just on the point of accepting or rejecting the hypothesis.

A hypothesis test has a

- Size: Probability of falsely rejecting a true null
- Rejection region: values of the test statistic which reject
- Power: Probability of correctly rejecting a false null

Example: An auto manufacturer claims that gas mileage for a new model averages 31.5 mpg. The population is normally distributed with standard deviation 2.4 mpg. A random sample of 16 cars has been taken. If the true gas mileage is 30 mpg, what is the power of the test?

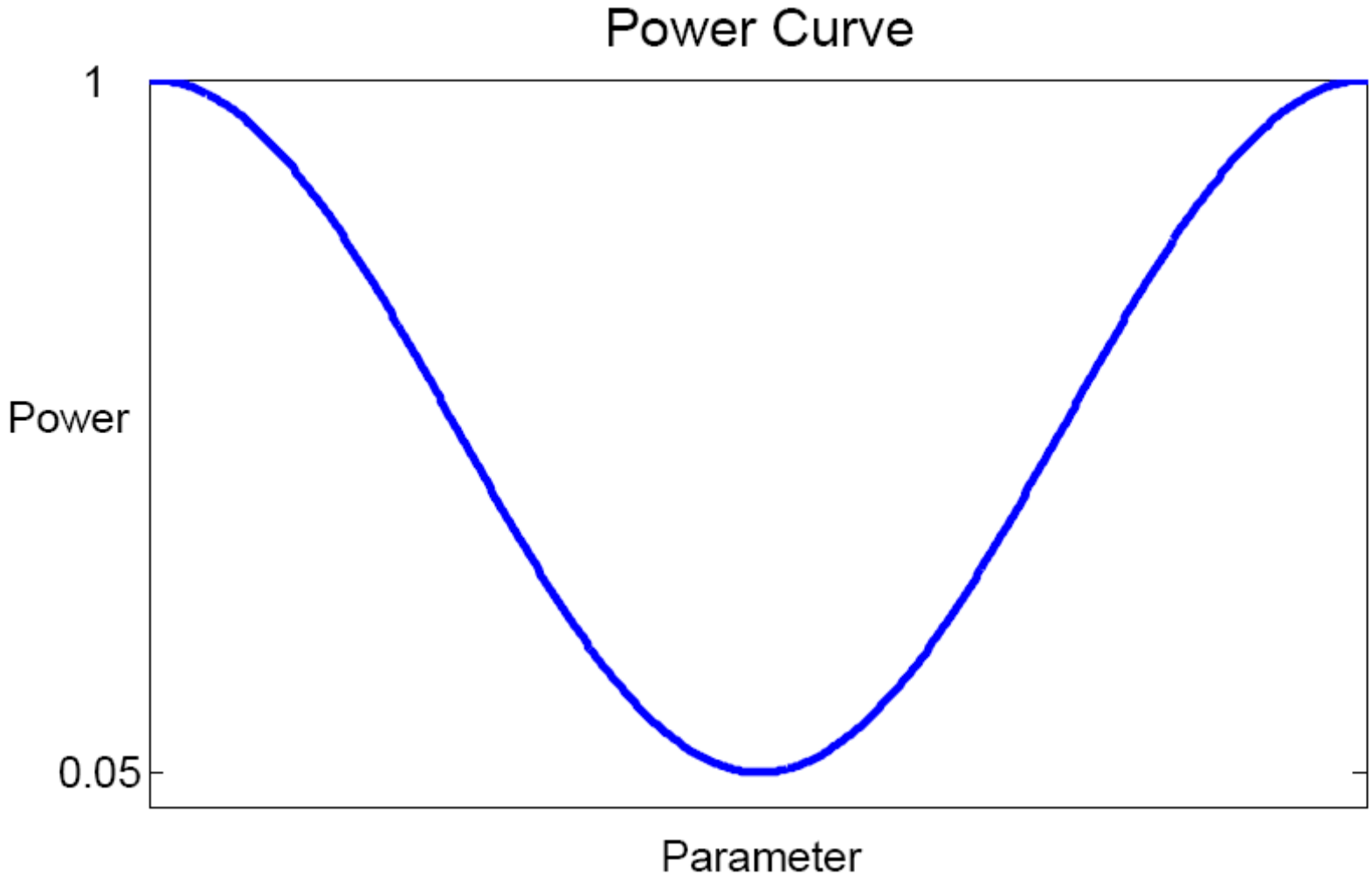
The test rejects if $\bar{X}$ is less than 30.324 or more than 32.676.

$\bar{X}$ is $N(30, \frac{2.4^2}{16})$ which is $N(30, 0.36)$

$P(\bar{X} < 30.324) = 0.705$ and $P(\bar{X} > 32.676) \approx 0$

The power of the test is 0.705.

A plot of the power of a test against the alternative is called a power curve.



Formula for the power curve

If $X_1,...X_n$ is drawn from a population that is normal with unknown mean μ and known variance σ^2, the rejection region for a 5 percent test of $\mu = \mu_0$ is $|\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}| > 1.96.$

What is the power function of the test?

$$\text{Power}(\mu) = 1 - P(\mu_0 - \frac{1.96\sigma}{\sqrt{n}} < \bar{X} < \mu_0 + \frac{1.96\sigma}{\sqrt{n}})$$

$$= 1 - P(\mu_0 - \frac{1.96\sigma}{\sqrt{n}} < N(\mu, \frac{\sigma^2}{n}) < \mu_0 + \frac{1.96\sigma}{\sqrt{n}})$$

$$= 1 - P(\mu_0 - \mu - \frac{1.96\sigma}{\sqrt{n}} < N(0, \frac{\sigma^2}{n}) < \mu_0 - \mu + \frac{1.96\sigma}{\sqrt{n}})$$

$$= 1 - P(\frac{\mu_0 - \mu}{\sigma / \sqrt{n}} - 1.96 < N(0,1) < \frac{\mu_0 - \mu}{\sigma / \sqrt{n}} + 1.96)$$

$$= 1 - \Phi(\frac{\mu_0 - \mu}{\sigma / \sqrt{n}} + 1.96) + \Phi(\frac{\mu_0 - \mu}{\sigma / \sqrt{n}} - 1.96)$$

If $\mu = \mu_0$, this is 1-0.975+0.025=0.05, which reduces to the size. For other values of μ, it is the power, and will be bigger.

Problem. If $\sigma = 1$ and $\mu_0 - \mu = 0.1$ (i.e. the difference between the hypothesized and true values is 1, how big a sample size do I need to get a power of at least 50 percent.

$$0.5 = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma / \sqrt{n}} + 1.96\right) + \Phi\left(\frac{\mu_0 - \mu}{\sigma / \sqrt{n}} - 1.96\right)$$

$$\therefore 0.5 = 1 - \Phi\left(\frac{0.1}{1 / \sqrt{n}} + 1.96\right) + \Phi\left(\frac{0.1}{1 / \sqrt{n}} - 1.96\right)$$

$$\therefore 0.5 = 1 - \Phi(0.1\sqrt{n} + 1.96) + \Phi(0.1\sqrt{n} - 1.96)$$

$$\therefore n = 385$$

A test is said to be **consistent** if the probability of rejecting any false hypothesis converges to 1 in the limit as the sample size goes to infinity.

Consistency is a minimal property of a test.

The power function we just had is

$$\text{Power}(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma / \sqrt{n}} + 1.96\right) + \Phi\left(\frac{\mu_0 - \mu}{\sigma / \sqrt{n}} - 1.96\right)$$

As $n \to \infty$, this converges to $1 - \Phi(\infty) + \Phi(\infty) = 1$

So the test is consistent.

Pitman Alternatives (Local Asymptotic Power)

Consistency means that we can't use asymptotic theory to say much about the relative power of different tests.

We can instead adopt the thought experiment that $\mu_n = \mu_0 + \frac{m}{\sqrt{n}}$ so that the hypothesis is "nearly" true

This is called a sequence of Pitman alternatives.

The limit as $n \to \infty$ of the probability of rejecting the null that $\mu = \mu_0$ when in fact $\mu_n = \mu_0 + \frac{m}{\sqrt{n}}$ is called the local asymptotic power.

For the power function we just had

$$\text{Power}(\mu_n) = 1 - \Phi\left(\frac{\mu_0 - \left(\mu_0 + \frac{m}{\sqrt{n}}\right)}{\sigma / \sqrt{n}} + 1.96\right) + \Phi\left(\frac{\mu_0 - \left(\mu_0 + \frac{m}{\sqrt{n}}\right)}{\sigma / \sqrt{n}} - 1.96\right)$$

$$= \boxed{1 - \Phi\left(-\frac{m}{\sigma} + 1.96\right) + \Phi\left(-\frac{m}{\sigma} - 1.96\right)}$$

This is the local asymptotic power (and does not equal 1).

More common testing situations

1. $X_1, X_2, ...X_n$ are iid with some distribution with mean μ and variance σ^2. σ^2 is known. Test null $\mu = \mu_0$.
2. $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$. σ^2 is not known. Test null $\mu = \mu_0$.
3. $X_1, X_2, ...X_n$ are iid with some distribution with mean μ and variance σ^2. σ^2 is not known. Test null $\mu = \mu_0$.
4. $X_1, X_2, ...X_n$ and $Y_1, Y_2, ...Y_n$ are iid with some distributions with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2. But the *X*s and *Y*s are independent. Test null $\mu_X = \mu_Y$
5. $\hat{p}$ is a sample proportion from a binomial. Test null $p = p_0$.
6. $\hat{\rho}$ is a sample correlation. Test null $\rho = 0$.

1. $X_1, X_2, ...X_n$ are iid with some distribution with mean μ and variance σ^2. σ^2 is known. Test null $\mu = \mu_0$.

Suppose the hypothesis is true

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \rightarrow_d N(0,1)$$

Test statistic: $\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$

Rejection Region: Reject if absolute value exceeds 1.96.
Size: Asymptotically 5 percent (by CLT).

2. $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$. σ^2 is not known. Test null $\mu = \mu_0$.

Suppose the hypothesis is true

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim t(n-1)$$

Test statistic: $\frac{\sqrt{n}(\bar{X} - \mu_0)}{s}$

Rejection Region: Reject if absolute value exceeds the upper 2-1/2 percentile of the t distribution on *n-1* degrees of freedom.

Size: Exact test. 5 percent.

3. $X_1, X_2, ...X_n$ are iid with some distribution with mean μ and variance σ^2. σ^2 is not known. Test null $\mu = \mu_0$.

Suppose the hypothesis is true

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \rightarrow_d N(0,1)$$

Test statistic: $\frac{\sqrt{n}(\bar{X} - \mu_0)}{s}$

Rejection Region: Reject if absolute value exceeds 1.96.

Size: Asymptotically 5 percent (by CLT).

4. $X_1, X_2, ...X_n$ and $Y_1, Y_2, ...Y_n$ are iid with some distributions with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2. The *X*s and *Y*s are independent. Test null $\mu_X = \mu_Y$

By the CLT,

$\sqrt{n}(\bar{X} - \mu_X) \rightarrow_d N(0, \sigma_X^2)$ and $\sqrt{n}(\bar{Y} - \mu_Y) \rightarrow_d N(0, \sigma_Y^2)$

$\therefore \bar{X} - \bar{Y}$ is approximately $N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n})$

Suppose the null hypothesis is true

$\bar{X} - \bar{Y}$ is approximately $N(0, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n})$

$$\therefore \left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{n}}}\right) \rightarrow_d N(0,1)$$

Test statistic: $\frac{\sqrt{n}(\bar{X}-\bar{Y})}{\sqrt{s_X^2+s_Y^2}}$

Rejection Region: Reject if absolute value exceeds 1.96.

Size: Asymptotically 5 percent.

Example: Shirley Brown, an agricultural economist wants to compare cow manure and turkey dung as fertilizers. She wants to test the hypothesis that they give equal crop yield.

She applies cow manure to 25 randomly chosen fields. The mean crop yield in this sample is 100 with a sample variance of 400. She applies turkey dung to another 25 randomly chosen fields. The mean crop yield in this sample is 115 with a sample variance of 225. Assuming that yields are normal, test the hypothesis of equal crop yield at the 5 percent level.

Test statistic:

$$\left|\frac{\sqrt{n}(\bar{X}-\bar{Y})}{\sqrt{s_X^2+s_Y^2}}\right| = \left|\frac{\sqrt{25}(\bar{X}-\bar{Y})}{\sqrt{400+225}}\right| = \left|\frac{\sqrt{25}(\bar{X}-\bar{Y})}{\sqrt{625}}\right|$$

$$= \left|\frac{5(\bar{X}-\bar{Y})}{25}\right| = \left|\frac{\bar{X}-\bar{Y}}{5}\right|$$

So the rejection region is $|\bar{X}-\bar{Y}| > 5*1.96 = 9.8$

In fact, $\bar{X}-\bar{Y} = -15$ and so the hypothesis is rejected.

5. $\hat{p}$ is a sample proportion from a binomial. Test null $p = p_0$.

Suppose the hypothesis is true $\hat{p} \rightarrow_d N(p_0, \frac{p_0(1-p_0)}{n})$

Test statistic: $\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Rejection Region: Reject if absolute value exceeds 1.96.
Size: Asymptotically 5 percent (by CLT).

Example: An opinion poll asked 1,000 randomly sampled voters whether they would vote for Obama. 48 percent replied yes. Test the hypothesis that 50 percent of all voters will vote for Obama.

Test statistic: $\frac{0.48 - 0.5}{\sqrt{\frac{0.5 * (1 - 0.5)}{1000}}} = -1.26$

The hypothesis is not rejected.

In this last example, what would the power of the test be against the alternative that 45 percent of all voters will vote for Obama.

The rejection region of the test is for p below $0.5 - 1.96\sqrt{0.5 * 0.5 / 1000} = 0.469$ and for p above 0.531.

If in fact $p=0.45$, the probability of rejecting is

$$P(N(0.45, \frac{0.45 * 0.55}{1000}) > 0.531) + P(N(0.45, \frac{0.45 * 0.55}{1000}) < 0.469)$$

=0+0.886=0.886.

6. $\hat{\rho}$ is a sample correlation. Test null $\rho = 0$.

Suppose that ρ denotes the correlation between *X* and *Y*.

Let $\hat{\rho} = \dfrac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{n}(X_i - \bar{X})^2 \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}}$

denote the sample correlation.

If $\rho = 0$ then $\hat{\rho} \rightarrow_d N(0, 1/n)$.

Useful and simple result.

Test statistic: $\hat{\rho} / \sqrt{1/n} = \hat{\rho}\sqrt{n}$

Rejection Region: Reject if absolute value exceeds 1.96.

Size: Asymptotically 5 percent.

Example. Over 60 days, the correlation between stock returns and oil returns is -0.2. Is this significantly different from zero?

Test statistic: $-0.2 / \sqrt{1/60} = -0.2 * \sqrt{60} = -1.549$

At the 5 percent level, the hypothesis is NOT rejected.

p-value: 0.121

Uniformly Most Powerful Tests

Consider testing the hypothesis $\theta \in \Theta_0$ against the alternative $\theta \in \Theta_1$. Let C_α be the set of rejection regions such that the probability of rejecting under the null is α.

A test is said to be **uniformly most powerful** if it is in C_α and has higher power than any other test in C_α, for all $\theta \in \Theta_1$.

Finding a uniformly most powerful test is a tall order.

But in one special case, we can…..

The Neyman-Pearson Lemma

Suppose that we have data X with joint pdf $f(x,\theta)$. We wish to test the hypothesis that $\theta = \theta_0$ against the alternative that $\theta = \theta_1$, then the test with the rejection region $\{x : \frac{f(x,\theta_1)}{f(x,\theta_0)} \geq k_\alpha\}$ where k_α is such that the probability of rejecting under the null is α is the uniformly most powerful test.

Neyman-Pearson Lemma Example 1:

Suppose that X has an exponential distribution with parameter λ. We wish to test the hypothesis that $\lambda = 1$ against the alternative $\lambda = 0.5$.
The Neyman-Pearson rejection region is

$$\{x : \frac{2e^{-2x}}{e^{-x}} \geq k_\alpha\} = \{x : 2e^{-2x} \geq k_\alpha e^{-x}\} = \{x : -2x \geq k'_\alpha - x\} = \{x : x \leq k''_\alpha\}$$

Now we choose k''_α so as to ensure that $P(x \leq k''_\alpha) = \alpha$ under the null (with $\alpha = 0.05$).

$$\int_0^{k''_\alpha} e^{-x} = 0.05 \Rightarrow [-e^{-x}]_0^{k''_\alpha} = 0.05 \Rightarrow 1 - e^{-k''_\alpha} = 0.05 \Rightarrow e^{-k''_\alpha} = 0.95$$

and so $k''_\alpha = -\log(0.95)$. The rejection region is *[0,-log(0.95)]*.

Neyman-Pearson Lemma Example 2:

Suppose that $X_1,...X_n$ are iid $N(\mu,1)$. We want to test the null that $\mu = \mu_0$ against the alternative $\mu = \mu_1$ for some $\mu_1 > \mu_0$. The Neyman-Pearson rejection region is

$$\{x : \frac{\Pi_{i=1}^{n}(2\pi)^{-1/2}\exp(-\frac{(X_i-\mu_1)^2}{2})}{\Pi_{i=1}^{n}(2\pi)^{-1/2}\exp(-\frac{(X_i-\mu_0)^2}{2})} \geq k_\alpha\} = \{x : \frac{\exp(-\Sigma_{i=1}^{n}\frac{(X_i-\mu_1)^2}{2})}{\exp(-\Sigma_{i=1}^{n}\frac{(X_i-\mu_0)^2}{2})} \geq k_\alpha\}$$

$$= \{x : \frac{\exp(-\Sigma_{i=1}^{n}\frac{X_i^2}{2} + \Sigma_{i=1}^{n}X_i\mu_1 - \frac{n\mu_1^2}{2})}{\exp(-\Sigma_{i=1}^{n}\frac{X_i^2}{2} + \Sigma_{i=1}^{n}X_i\mu_0 - \frac{n\mu_0^2}{2})} \geq k_\alpha\}$$

$$= \{x : \exp(\Sigma_{i=1}^{n} X_i(\mu_1 - \mu_0) + \frac{n\mu_0^2}{2} - \frac{n\mu_1^2}{2}) \geq k_\alpha\}$$

$$= \{x : (\mu_1 - \mu_0) n\bar{X} + \frac{n\mu_0^2}{2} - \frac{n\mu_1^2}{2}) \geq k'_\alpha\}$$

$$= \{x : (\mu_1 - \mu_0)\bar{X} + \frac{\mu_0^2 - \mu_1^2}{2} \geq k''_\alpha\} = \{x : \bar{X} \geq k'''_\alpha\}$$

Suppose we want to test at the 0.05 level. Then

$P(\bar{X} \geq k'''_\alpha) = 0.05$

$$\bar{X} \sim N(\mu_0, 1/n) \Rightarrow P(\bar{X} > \mu_0 + \frac{1.64}{\sqrt{n}}) = 0.05$$

So the rejection region is $\bar{X} \geq \mu_0 + \frac{1.64}{\sqrt{n}}$

In this case, the rejection region doesn't depend on μ_1

Likelihood Ratio Test

A very general method. Almost always applicable and sometimes optimal.

Let $L(\theta)$ denote the likelihood function. To test the hypothesis that $\theta \in \Theta_0$ against the alternative $\theta \notin \Theta_0$, the likelihood ratio test is

$$\lambda = \frac{\sup_{\Theta_0} L(\theta)}{\sup_{\Theta} L(\theta)}$$

The rejection region is of the form $\{x : \lambda \leq c\}$.
Note that by construction $0 < \lambda < 1$.

Example. Let $X_1, X_2, ...X_n$ be an iid random sample from a population with pdf

$$f(x) = e^{-(x-\theta)} \quad x \geq \theta$$
$$0 \qquad \text{otherwise}$$

What is a LR test of $\theta = \theta_0$ against the alternative $\theta \neq \theta_0$.

$\log(f(x)) = -(x-\theta) = \theta - x$ if $x \geq \theta$ and $-\infty$ otherwise.

$\therefore l(\theta) = n\theta - \Sigma_{i=1}^{n} X_i$ if $\min(X_1, ...X_n) \geq \theta$ and $-\infty$ otherwise.

The maximum likelihood estimator is $\tilde{\theta} = \min(X_1, ..X_n)$.

$$L(\theta_0) = \Pi_{i=1}^{n} e^{-(X_i - \theta_0)} = \exp(-\Sigma_{i=1}^{n}(X_i - \theta_0))$$
$$\therefore L(\theta_0) = \exp(n\theta_0 - \Sigma_{i=1}^{n} X_i)$$
if $\min(X_1,...X_n) \geq \theta_0$ and $-\infty$ otherwise.

Meanwhile,
$$L(\tilde{\theta}) = \Pi_{i=1}^{n} e^{-(X_i - \min(X_1,...X_n))} = \exp(-\Sigma_{i=1}^{n}(X_i - \min(X_1,...X_n)))$$
$$\therefore L(\tilde{\theta}) = \exp(n\min(X_1,...X_n) - \Sigma_{i=1}^{n} X_i)$$

Hence the LR statistic is
$$\lambda = \exp(n(\theta_0 - \min(X_1,...X_n))$$

The rejection region is
$$\{x : \lambda \leq c\} = \{x : \log(\lambda) \leq \log(c)\}$$
$$= \{x : \min(X_1,...X_n) \geq c'\} \bigcup \{x : \min(X_1,...X_n) < \theta_0\}$$

To work out the rejection region (c') need to know the cdf of $\min(X_1,...X_n)$

$$F(x) = 1 - e^{-(x-\theta)}$$

$$\therefore F_{\min}(x) = 1 - (1 - F(x))^n = 1 - e^{-n(x-\theta)}$$

$$\therefore P(\min(X_1,...X_n) \geq c') = 1 - [1 - e^{-n(c'-\theta)}] = e^{-n(c'-\theta)}$$

To set this to (say) 5 percent

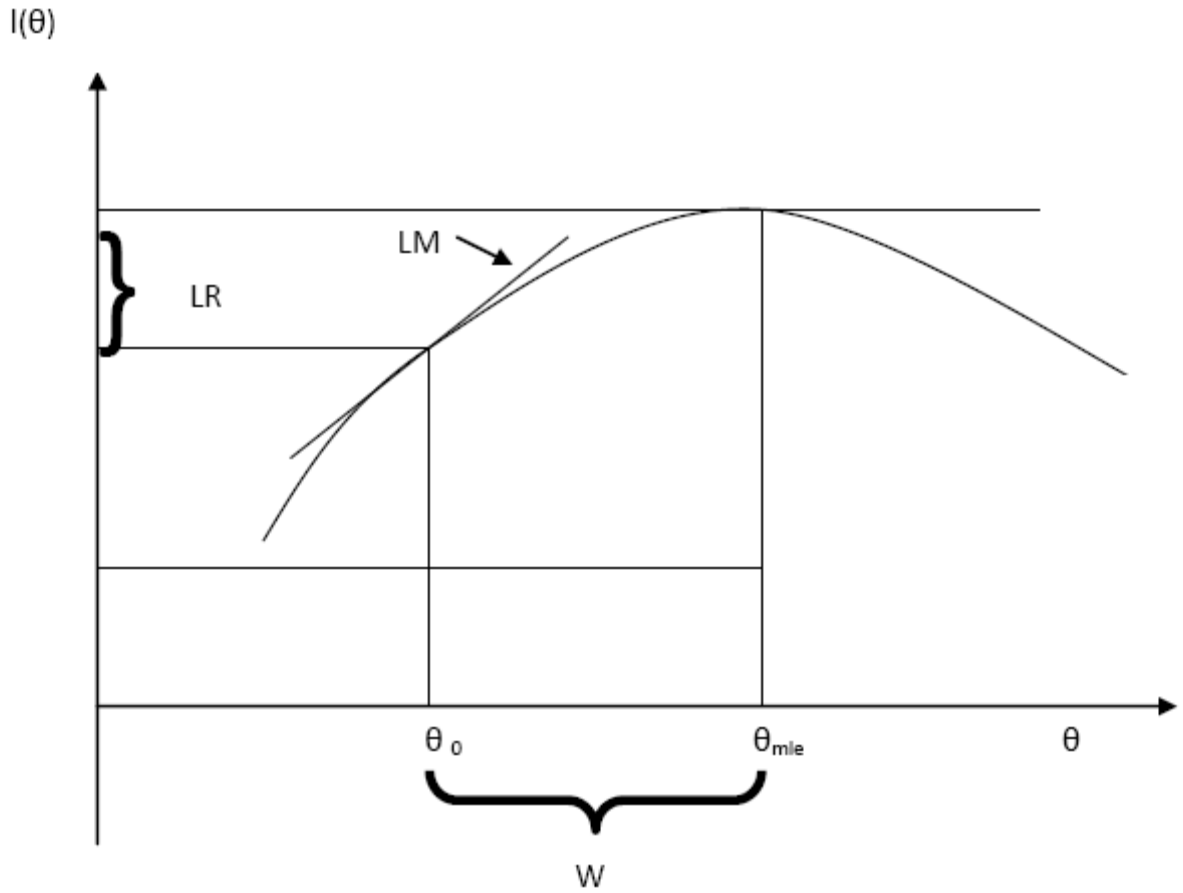
$$e^{-n(c'-\theta)} = 0.05$$

which can be solved for c'.

The holy trinity of tests: Wald, LR and LM.

Suppose that θ is a parameter, $l(\theta)$ is the log-likelihood function and $\tilde{\theta} = \arg\max_{\theta} l(\theta)$ is the MLE.

We want to test $\theta = \theta_0$ against $\theta \neq \theta_0$



Wald: Compare $\tilde{\theta}$ and θ_0

LR: Compare $l(\tilde{\theta})$ and $l(\theta_0)$

LM: See how close $l'(\theta_0)$ is to zero

Suppose that p is the dimension of θ. Under the null $\theta = \theta_0$,

- LR statistic: $2(l(\tilde{\theta}) - l(\theta_0)) \rightarrow_d \chi^2(p)$
- Wald statistic: $n(\tilde{\theta} - \theta_0)' I(\tilde{\theta})(\tilde{\theta} - \theta_0) \rightarrow_d \chi^2(p)$
- LM statistic: $\frac{1}{n} l'(\theta_0) I(\theta_0)^{-1} l'(\theta_0) \rightarrow_d \chi^2(p)$

Note: The LR test is minus twice the log of what was called the likelihood ratio earlier.

Example 1 of the LR, Wald and LM statistics.

Suppose that $X_1, X_2, ...X_n$ are iid Poisson with parameter λ.

$$P(X_i = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

We want to test the null $\lambda = 6$ against the alternative $\lambda \neq 6$
Suppose that $n = 100$ and $\Sigma_{i=1}^{n} X_i = 500$.

The log-likelihood is

$$l(\lambda) = \Sigma_{i=1}^{n} \{x_i \log(\lambda) - \lambda - \log(x_i !)\}$$

The MLE is $\tilde{\lambda} = \bar{X} = 5$ (showed this earlier)

The LR test is

$$LR = 2 * [\Sigma_{i=1}^{n} \{x_i \log(5) - 5 - \log(x_i !)\} - \{\Sigma_{i=1}^{n} x_i \log(6) - 6 - \log(x_i !)\}]$$
$$= 2 * \{\Sigma_{i=1}^{n} x_i \log(5) - 500 - \Sigma_{i=1}^{n} x_i \log(6) + 600\}$$
$$= 2 * \{\Sigma_{i=1}^{n} x_i (\log(5) - \log(6)) + 100\}$$
$$= 2 * \{500 * (\log(5) - \log(6)) + 100\} = 17.6$$

$I = 1/\lambda$ (derived earlier too) and our estimate of I is $1/5$

The Wald test is

$$W = 100 * (5-6) * I * (5-6) = 100 * (5-6) * \frac{1}{5} * (5-6) = 20$$

The LM test is derived as follows:

$$\frac{\partial l(\lambda)}{\partial \lambda} = \Sigma_{i=1}^{n}(\frac{x_i}{\lambda_0} - 1) = \frac{\Sigma_{i=1}^{n} x_i}{\lambda_0} - n = \frac{500}{6} - 100 = -16.7$$

$$\therefore LM = \frac{1}{100} * -16.7 * \frac{1}{1/6} * -16.7 = 16.73$$

All three statistics are different, but they all reject (critical value: 3.84 for a 5 percent test)

Example 2 of the LR, Wald and LM statistics.

Suppose that $X_1, X_2, ...X_n$ are iid $N(\mu, 1)$.

We want to test the null $\mu = 0$ against the alternative $\mu \neq 0$

The log-likelihood is

$$l(\mu) = -\frac{n}{2}\log(2\pi) - \Sigma_{i=1}^{n} \frac{(X_i - \mu)^2}{2}$$

and the MLE is $\tilde{\mu} = \bar{X}$.

The LR test is

$$LR = 2*(l(\bar{X}) - l(0))$$

$$= 2\{-\frac{n}{2}\log(2\pi) - \Sigma_{i=1}^{n}\frac{(X_i - \bar{X})^2}{2} + \frac{n}{2}\log(2\pi) + \Sigma_{i=1}^{n}\frac{X_i^2}{2}\}$$

$$= 2\{-\Sigma_{i=1}^{n}\frac{(X_i - \bar{X})^2}{2} + \Sigma_{i=1}^{n}\frac{X_i^2}{2}\}$$

$$= 2\{\frac{\Sigma_{i=1}^{n}X_i^2}{2} - \frac{\Sigma_{i=1}^{n}X_i^2 - n\bar{X}^2}{2}\}$$

$$= n\bar{X}^2$$

The Wald test is

$$W = n(\tilde{\mu} - 0)^2 I$$

$$\tilde{\mu} = \bar{X}$$

$$I = \frac{1}{\sigma^2} = 1$$

$$\therefore W = n\bar{X}^2$$

This is a squared t-test….the t-test would be

$$\sqrt{n}\left(\frac{\bar{X} - 0}{1}\right) = \sqrt{n}\,\bar{X}$$

As for the LM test…

$$l(\mu) = -\frac{n}{2}\log(2\pi) - \Sigma_{i=1}^{n}\frac{(X_i - \mu)^2}{2}$$

$$\therefore l'(\mu) = \Sigma_{i=1}^{n}(X_i - \mu) \Rightarrow l'(0) = \Sigma_{i=1}^{n}X_i = n\bar{X}$$

$$I = 1$$

$$\therefore LM = \frac{n^2\bar{X}^2}{n} = n\bar{X}^2$$

In this case it just so happens that all three tests are the same.

Example 3 of the LR, Wald and LM statistics

Suppose that $\{X_1, X_2, ...X_n\}$ is iid Bernoulli with parameter p

We want to test the null $p = 1/2$ against the alternative $p \neq 1/2$

The MLE of p is $\hat{p} = \bar{X}$.
Suppose this is 0.55 in a sample of 100 observations.

The log-likelihood is

$$l(p) = \Sigma_{i=1}^{n} X_i \log(p) + (1 - X_i)\log(1 - p)$$
$$= \log(p)\Sigma_{i=1}^{n} X_i + \log(1 - p)\Sigma_{i=1}^{n}(1 - X_i)$$

$$= \log(p) n\hat{p} + \log(1-p) n(1-\hat{p}) = n(\log(p)\,\hat{p} + \log(1-p)(1-\hat{p}))$$

The LR test is

$$2(l(\hat{p}) - l(p)) = 2n(\log(\hat{p})\,\hat{p} + \log(1-\hat{p})(1-\hat{p})$$

$$-\log(p_0)\,\hat{p} - \log(1-p_0)(1-\hat{p}))$$

$$= 2n(\hat{p}\log(\frac{\hat{p}}{p_0}) + (1-\hat{p})\log(\frac{1-\hat{p}}{1-p_0}))$$

In this case

$$LR = 2*100*(0.55\log(\frac{0.55}{0.5}) + 0.45\log(\frac{0.45}{0.5})) = 1.00$$

So the test accepts

$$I = \frac{1}{p(1-p)}$$ (showed this earlier)

The Wald statistic is

$$n\frac{(\hat{p}-p)^2}{\hat{p}(1-\hat{p})} = 100 * \frac{(0.55-0.5)^2}{0.55*0.45}$$

With our numbers, this is 1.01.
So the test accepts.

As for the LM test…

$$l'(p) = n(\frac{\hat{p}}{p} - \frac{1-\hat{p}}{1-p})$$

$$\therefore LM = \frac{1}{n} n^2 (\frac{\hat{p}}{p} - \frac{1-\hat{p}}{1-p})^2 \, p(1-p)$$

$$= n(\frac{\hat{p}}{p} - \frac{1-\hat{p}}{1-p})^2 \, p(1-p)$$

With our numbers, this is 1.00.
So the test accepts.

Example 4 of the LR, Wald and LM statistics.

Suppose that $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$.

We want to test the null $\theta = \theta_0$ against the alternative $\theta \neq \theta_0$ where $\theta = (\mu, \sigma^2)'$.

The log-likelihood is

$$l(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \Sigma_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}$$

and the MLE is $\tilde{\mu} = \bar{X}$ and $\tilde{\sigma}^2 = n^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})^2$.

The LR statistic is

$$2(l(\tilde{\theta}) - l(\theta_0)) = -n\log(2\pi\tilde{\sigma}^2) - \Sigma_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\tilde{\sigma}^2}$$

$$+n\log(2\pi\sigma_0^2) + \Sigma_{i=1}^{n} \frac{(X_i - \mu_0)^2}{\sigma_0^2}$$

The Wald statistic is

$$n(\tilde{\mu}-\mu_0 \;\; \tilde{\sigma}^2-\sigma_0^2)\begin{pmatrix} 1/\tilde{\sigma}^2 & 0 \\ 0 & 1/2\tilde{\sigma}^4 \end{pmatrix}\begin{pmatrix} \tilde{\mu}-\mu_0 \\ \tilde{\sigma}^2-\sigma_0^2 \end{pmatrix}$$

$$= n\frac{(\tilde{\mu}-\mu_0)^2}{\tilde{\sigma}^2} + n\frac{(\tilde{\sigma}^2-\sigma_0^2)^2}{2\tilde{\sigma}^4}$$

Sketch of the proof of the asymptotic distribution

$$LR = 2(l(\tilde{\theta}) - l(\theta_0))$$
$$W = n(\tilde{\theta} - \theta_0)' I(\tilde{\theta})(\tilde{\theta} - \theta_0)$$
$$LM = \frac{1}{n} \frac{dl(\theta_0)}{d\theta'} I(\theta_0)^{-1} \frac{dl(\theta_0)}{d\theta'}$$

Assume that $I(\theta_0) \approx I(\tilde{\theta}) \approx I$

Results from earlier

- $\frac{1}{n}\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'} \to_p E\{\frac{\partial^2 \log f(x,\theta)}{\partial\theta^2}\} = -I$
- $\frac{1}{\sqrt{n}}\frac{dl(\theta_0)}{d\theta} \to_d N(0,I)$
- $\sqrt{n}(\tilde{\theta}-\theta_0) \to_d N(0,I^{-1})$
- If $X \sim N(0,\Sigma)$ has a multivariate normal distribution, then $X'\Sigma^{-1}X$ is $\chi^2(k)$ distributed where k is the dimension of Σ

Hence,

$W = \sqrt{n}(\tilde{\theta}-\theta_0)'I\sqrt{n}(\tilde{\theta}-\theta_0) \Rightarrow W \to_d \chi^2(p)$

Similarly,

$$LM = \frac{1}{\sqrt{n}} \frac{dl(\theta_0)}{d\theta'} I^{-1} \frac{1}{\sqrt{n}} \frac{dl(\theta_0)}{d\theta} \Rightarrow LM \rightarrow_d \chi^2(p)$$

Finally,

$$l(\theta_0) \approx l(\tilde{\theta}) + (\theta_0 - \tilde{\theta})' \frac{\partial l(\tilde{\theta})}{\partial \theta} + \frac{1}{2}(\theta_0 - \tilde{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} (\theta_0 - \tilde{\theta})$$

$$\frac{\partial l(\tilde{\theta})}{\partial \theta} = 0 \Rightarrow l(\theta_0) \approx l(\tilde{\theta}) + \frac{1}{2}(\theta_0 - \tilde{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} (\theta_0 - \tilde{\theta})$$

$$\therefore 2(l(\tilde{\theta}) - l(\theta_0)) \approx -(\theta_0 - \tilde{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} (\theta_0 - \tilde{\theta})$$

$$\therefore LR \approx n(\theta_0 - \tilde{\theta})' I (\theta_0 - \tilde{\theta}) \rightarrow_d \chi^2(p)$$

Hypothesis tests and the delta method

If $\{X_1, X_2, ...X_n\}$ is iid with mean μ and variance σ^2, how can I test the hypothesis that $\frac{1}{\mu} = 2$?

We know from the delta method that

$$\sqrt{n}(\frac{1}{\overline{X}} - \frac{1}{\mu}) \rightarrow_d N(0, \frac{\sigma^2}{\mu^4})$$

$$\therefore \sqrt{n}(\frac{1}{\overline{X}} - 2) \rightarrow_d N(0, 16\sigma^2)$$

$$\therefore \sqrt{n}(\frac{\overline{X}^{-1} - 2}{4s}) \rightarrow_d N(0,1)$$

This allows us to do a test.

t-tests are not invariant to nonlinear reparameterization.

Example. $\bar{X} = 0.9$, $s = 1$ and $n=25$.
First want to test $\mu = 0.5$

$$\frac{\bar{X} - 0.5}{s/\sqrt{n}} = \frac{0.9 - 0.5}{1/\sqrt{25}} = 2 \ldots \text{reject}$$

Next test $\frac{1}{\mu} = 2 \ldots$ **same** hypothesis, parameterized differently

$$\frac{\bar{X}^{-1} - 2}{4s/\sqrt{n}} = \frac{0.9^{-1} - 2}{4/\sqrt{25}} = -1.11 \ldots \text{don't reject}$$

Hypothesis tests and the delta method

Suppose that $U_1,...U_n$ and $V_1,...V_n$ are random variables that are iid with mean μ_U and μ_V, variance σ_U^2 and σ_V^2 and correlation ρ. Let $\mu^* = \mu_U / \mu_V$. We want to test the hypothesis that $\mu^* = \mu_0^*$.

We know from the delta method that

$$\sqrt{n}(\frac{\bar{U}}{\bar{V}} - \frac{\mu_U}{\mu_V}) \to_d N(0, \lambda^2)$$

where $\lambda^2 = \frac{\sigma_U^2}{\mu_V^2} + \frac{\sigma_V^2 \mu_U^2}{\mu_V^4} - 2\frac{\rho \sigma_U \sigma_V \mu_U}{\mu_V^3}$. Hence

$$\sqrt{n}(\frac{\bar{U}/\bar{V} - \mu_0^*}{\hat{\lambda}}) \to_d N(0,1)$$

Fieller's Method

Check the size of tests about the ratio of two means

Simple Matlab code

```
n=50;
randn('seed',123);
for imc=1:10000;
      u=randn(n,1)+1; v=randn(n,1)+0.1;
      ubar=mean(u); vbar=mean(v);
      lambdasq=(1/(vbar^2))+((ubar^2)/(vbar^4));
      tstat(imc)=sqrt(n)*((ubar/vbar)-10)/sqrt(lambdasq);
end;
mean(abs(tstat)>1.96)
```

Has poor size

Fieller (1954) proposed a simple trick to do **exact** inference on a ratio of two means if the data are normal.

Suppose that $U_1,...U_n$ and $V_1,...V_n$ are random variables that are iid normal with mean μ_U and μ_V, known variances σ_U^2 and σ_V^2 and correlation 0.

Let $\mu^* = \mu_U / \mu_V$. We want to test the hypothesis that $\mu^* = \mu_0^*$.

$$\sqrt{n}(\bar{U} - \mu_U) \sim N(0, \sigma_U^2)$$
$$\sqrt{n}(\bar{V} - \mu_V) \sim N(0, \sigma_V^2)$$

Consider the variable $\mu_U - \mu_0^* \mu_V$.

$$\sqrt{n}(\bar{U} - \mu_0^* \bar{V} - (\mu_U - \mu_0^* \mu_V)) \sim N(0, \sigma_U^2 + \mu_0^{*2} \sigma_V^2)$$

Under the null hypothesis $\mu_U - \mu_0^* \mu_V = 0$. So under the null $\sqrt{n}(\bar{U} - \mu_0^* \bar{V}) \sim N(0, \sigma_U^2 + \mu_0^{*2} \sigma_V^2)$

This gives an exact test.

Test statistic is $\dfrac{\sqrt{n}(\bar{U} - \mu_0^* \bar{V})}{\sqrt{\sigma_U^2 + \mu_0^{*2} \sigma_V^2}}$

It works even where there is correlation between the variables, or when the variances are not known (here the distribution becomes t, not normal).

See page. 464 of Casella and Berger.

Bayesian approach to tests

Recall that the Bayesian paradigm treats the parameter θ as a random variable.

Suppose that $f(\theta)$ is the prior density and

$$f(\theta \mid X_1,\ldots X_n) = \frac{f(X_1,\ldots X_n \mid \theta) f(\theta)}{f(X_1,\ldots X_n)}$$

is the posterior density.

In assessing the hypothesis that $\theta \in \Theta_0$ and the alternative $\theta \notin \Theta_0$, the **posterior odds ratio** is

$$\frac{P(\theta \in \Theta_0 \mid X_1,...X_n)}{P(\theta \notin \Theta_0 \mid X_1,...X_n)} = \frac{\int_{\theta \in \Theta_0} f(\theta \mid X_1,...X_n)d\theta}{\int_{\theta \notin \Theta_0} f(\theta \mid X_1,...X_n)d\theta}$$

$$= \frac{\int_{\theta \in \Theta_0} f(X_1,...X_n \mid \theta) f(\theta)d\theta}{\int_{\theta \notin \Theta_0} f(X_1,...X_n \mid \theta) f(\theta)d\theta}$$

Special case. Comparing hypothesis that $\theta = \theta_0$ against the alternative that $\theta = \theta_1$, the posterior odds ratio is

$$\frac{f(X_1,...X_n \mid \theta_0) f(\theta_0)}{f(X_1,...X_n \mid \theta_1) f(\theta_1)} = \frac{L(\theta_0)}{L(\theta_1)} \frac{f(\theta_0)}{f(\theta_1)}$$

where $L(.)$ is the likelihood function.

If the prior is equal for the two points, $f(\theta_0) = f(\theta_1) = 1/2$ then the posterior odds is $\dfrac{L(\theta_0)}{L(\theta_1)}$

This is the likelihood ratio…but the interpretation is different.

Bayesians statisticians don't look at p-values.

For comparing the hypothesis against the alternative, Jeffreys proposed the following scale for odds ratios

Posterior Odds Ratio	Conclusion
<1	No evidence for hypothesis
1-3	Barely worth a mention
3-10	Substantial
10-30	Strong
30-100	Very Strong
>100	Decisive

Interval estimation.

In interval estimation, we give up on trying to get a single estimate, but rather instead look for range.

Suppose that $[L(X),U(X)]$ is a confidence interval for a parameter θ. Then we say that the **coverage** of this confidence interval is $P(L(X) \leq \theta \leq U(X))$.

Of course $[-\infty,\infty]$ is a confidence interval with coverage 1. But it isn't very useful.

The **width** of a confidence interval is $U(X) - L(X)$.

Example. $X_1, X_2, ...X_n$ are iid uniform between 0 and θ. $Y_n = \max(X_1, X_2, ...X_n)$. Consider two confidence intervals for θ.

$[aY_n, bY_n]$ where $1 \le a < b$

$[Y_n + c, Y_n + d]$ where $0 \le c < d$

What is the coverage of each confidence interval?

The coverage of the confidence interval $[aY_n, bY_n]$ is

$$P(aY_n \le \theta \le bY_n) = P(\frac{\theta}{b} \le Y_n \le \frac{\theta}{a})$$

The pdf of Y_n is $\frac{1}{\theta^n} n y^{n-1}$

$$\therefore P(\frac{\theta}{b} \le Y_n \le \frac{\theta}{a}) = \int_{\theta/b}^{\theta/a} \frac{1}{\theta^n} n y^{n-1} dy = \frac{1}{\theta^n} \int_{\theta/b}^{\theta/a} n y^{n-1} dy$$

$$\therefore P(\frac{\theta}{b} \le Y_n \le \frac{\theta}{a}) = \frac{1}{\theta^n} [y^n]_{\theta/b}^{\theta/a} = \frac{(\theta/a)^n - (\theta/b)^n}{\theta^n}$$

$$\therefore P(\frac{\theta}{b} \le Y_n \le \frac{\theta}{a}) = (1/a)^n - (1/b)^n$$

is the required coverage.

The coverage of the confidence interval $[Y_n + c, Y_n + d]$ is

$$P(Y_n + c \le \theta \le Y_n + d) = P(\theta - d \le Y_n \le \theta - c)$$

The pdf of Y_n is $\frac{1}{\theta^n} n y^{n-1}$

$$\therefore P(\theta - d \le Y_n \le \theta - c) = \int_{\theta-d}^{\theta-c} \frac{1}{\theta^n} n y^{n-1} dy = \frac{1}{\theta^n} \int_{\theta-d}^{\theta-c} n y^{n-1} dy$$

$$\therefore P(\theta - d \le Y_n \le \theta - c) = \frac{1}{\theta^n} [y^n]_{\theta-d}^{\theta-c} = \frac{(\theta - c)^n - (\theta - d)^n}{\theta^n}$$

is the required coverage.

General method for forming a confidence set

Consider a test of the hypothesis $\theta = \theta_0$. Let $C(X)$ denote the set of values of θ_0 for which this test does not reject.

Q. What is the probability that $C(X)$ includes the true value of θ, $\overline{\theta}$?

A. It is the probability that the test does not reject the hypothesis $\theta = \overline{\theta}$. This is 1 minus the probability that the test rejects the hypothesis $\theta = \overline{\theta}$.

- If the test has a size of exactly α, then $C(X)$ is a confidence set with coverage exactly 1-α.
- If the test has asymptotic size of α, then$C(X)$ is a confidence set with asymptotic coverage 1-α.

From the first principals construction of a confidence set, it could have a wierd shape (e.g. disjoint). Sometimes that's true. Usually not.

Most typically, let $\hat{\theta}$ be an estimator such that $\sqrt{n}(\frac{\hat{\theta}-\theta}{V}) \rightarrow_d N(0,1)$

The test of the hypothesis $\theta = \theta_0$ will accept (at the 5 percent level) if $|\sqrt{n}(\frac{\hat{\theta} - \theta_0}{V})| \leq 1.96$

So the set of θ_0 for which the test accepts is

$$\{\theta_0 : -1.96 \leq \sqrt{n}(\frac{\hat{\theta} - \theta_0}{V}) \leq 1.96\}$$

$$= \{\theta_0 : -\frac{1.96V}{\sqrt{n}} \leq \hat{\theta} - \theta_0 \leq \frac{1.96V}{\sqrt{n}}\}$$

$$= \{\theta_0 : -\frac{1.96V}{\sqrt{n}} \leq \theta_0 - \hat{\theta} \leq \frac{1.96V}{\sqrt{n}}\}$$

$$= \{\theta_0 : \hat{\theta} - \frac{1.96V}{\sqrt{n}} \leq \theta_0 \leq \hat{\theta} + \frac{1.96V}{\sqrt{n}}\}$$

which is $\hat{\theta} \pm \frac{1.96V}{\sqrt{n}}$. This confidence set has asymptotic coverage of 95 percent.

If $\hat{\theta}$ were an estimator such that $\sqrt{n}(\frac{\hat{\theta} - \theta}{V}) \sim N(0,1)$, then $\hat{\theta} \pm \frac{1.96V}{\sqrt{n}}$ would be a confidence set for θ with coverage of exactly 95 percent.

Example 1. $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$ where σ^2 is known. We want to form a 95 percent confidence interval for μ.

$$\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \sim N(0,1)$$

$$\therefore \bar{X} \pm \frac{1.96\sigma}{\sqrt{n}}$$ is a confidence interval for μ with coverage of exactly 95 percent.

Example 2. Same but the parent distribution is not necessarily normal.

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \rightarrow_d N(0,1)$$

$\therefore \bar{X} \pm \frac{1.96\sigma}{\sqrt{n}}$ is a confidence interval for μ with asymptotic coverage of 95 percent.

Example 3: $\hat{p}_n$ is a sample proportion when an experiment is repeated n times. We want to form a 95 percent confidence interval for p, the probability of success.

$$\frac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightarrow_d N(0,1)$$

$\therefore \hat{p}_n \pm 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$ is a confidence interval for p with asymptotic coverage of 95 percent.

Opinion polls often say that they have a "±3 percent margin of error"

This comes from the formula

$$\therefore \hat{p}_n \pm 1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

$\hat{p}_n \approx 0.5$ and $n = 1{,}000$ (typically) so

$$1.96\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} = 1.96\sqrt{\frac{0.5*0.5}{1000}} = 0.0310$$

If $\sqrt{n}(\frac{\hat{\theta}-\theta}{V})$ has a distribution that does not depend on the parameter θ, then it is said to be **pivotal**.

If $\sqrt{n}(\frac{\hat{\theta}-\theta}{V})$ has an asymptotic distribution that does not depend on the parameter θ, then it is said to be **asymptotically pivotal**.

Suppose that $\sqrt{n}(\frac{\hat{\theta}-\theta}{V})$ has a pivotal distribution with lower and upper α percentiles F_α and $F_{1-\alpha}$, respectively.

The test of the hypothesis $\theta = \theta_0$ will accept (at the 2α significance level) if $\{\theta_0 : F_\alpha \le \sqrt{n}(\frac{\hat{\theta}-\theta_0}{V}) \le F_{1-\alpha}\}$

The confidence set is $\{\theta_0 : F_\alpha \le \sqrt{n}(\frac{\hat{\theta}-\theta_0}{V}) \le F_{1-\alpha}\}$

$$= \{\theta_0 : \frac{VF_\alpha}{\sqrt{n}} \le \hat{\theta} - \theta_0 \le \frac{VF_{1-\alpha}}{\sqrt{n}}\} = \{\theta_0 : -\frac{VF_{1-\alpha}}{\sqrt{n}} \le \theta_0 - \hat{\theta} \le -\frac{VF_\alpha}{\sqrt{n}}\}$$

$$= \{\theta_0 : \hat{\theta} - \frac{VF_{1-\alpha}}{\sqrt{n}} \le \theta_0 \le \hat{\theta} - \frac{VF_\alpha}{\sqrt{n}}\}$$

Similarly, if $\sqrt{n}(\frac{\hat{\theta}-\theta}{V})$ has an asymptotically pivotal distribution with lower and upper α percentiles F_{α} and $F_{1-\alpha}$, respectively, the confidence set

$$\{\theta_0 : \hat{\theta} - \frac{VF_{1-\alpha}}{\sqrt{n}} \leq \theta_0 \leq \hat{\theta} - \frac{VF_{\alpha}}{\sqrt{n}}\}$$

has asymptotic coverage $1-2\alpha$.

Example. $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$ where σ^2 is unknown and $n = 10$. We want to form a 95 percent confidence interval for μ.

$\sqrt{n}(\frac{\hat{\mu} - \mu}{s})$ is t-distributed on 9 degrees of freedom.

It is pivotal.

The 95 percent confidence interval is

$$\{\mu : \hat{\mu} - \frac{2.26s}{\sqrt{n}} \leq \mu \leq \hat{\mu} + \frac{2.26s}{\sqrt{n}}\}$$

If $X_1, X_2, ...X_n$ are iid $N(\mu, \sigma^2)$ then $\frac{(n-1)s^2}{\sigma^2}$ is χ^2 distributed on 9 degrees of freedom. This is also pivotal and enables a confidence set to be constructed for σ^2.

Let the lower and upper α percentiles of the $\chi^2(n-1)$ distribution be F_α and $F_{1-\alpha}$, respectively.

The $1-2\alpha$ confidence interval is

$$\{\sigma^2 : F_\alpha \le \frac{(n-1)s^2}{\sigma^2} \le F_{1-\alpha}\}$$

$$= \{\sigma^2 : \frac{1}{F_{1-\alpha}} \le \frac{\sigma^2}{(n-1)s^2} \le \frac{1}{F_\alpha}\}$$

$$= \{\sigma^2 : \frac{(n-1)s^2}{F_{1-\alpha}} \leq \sigma^2 \leq \frac{(n-1)s^2}{F_{\alpha}}\}$$

"flipped around" again.

For example, if $n = 10$ and $s^2 = 1$, the 95 percent confidence interval for σ^2 is

$\{\sigma^2 : \frac{9}{19.02} \leq \sigma^2 \leq \frac{9}{2.70}\}$ which is from 0.47 to 3.33.

It's much easier to form confidence intervals with pivotal test statistics. So we do that when possible (occasionally it's not).

The important (asymptotically) pivotal distributions are

$\sqrt{n}(\frac{\bar{X}-\mu}{s})$ t/normal

$\frac{(n-1)s^2}{\sigma^2}$ χ^2

in the usual notation.

The regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} .... + \beta_k x_{ik} + u_i = \beta' x_i + u_i \ , \ \ i = 1,2,...n$$

Can be written as $Y = X\beta + u$ where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_{11} & x_{12} & .. & x_{1k} \\ 1 & x_{21} & x_{22} & .. & x_{2k} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \text{ and } u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Assumptions:

1. x_i is fixed (hence uncorrelated with the error term)
2. u_i is iid with mean zero and variance σ^2
3. x_is are not perfectly multicollinear

OLS is an intuitive estimator

$$\hat{\beta} = (\Sigma_{i=1}^{n} x_i x_i')^{-1} \Sigma_{i=1}^{n} x_i y_i = (X'X)^{-1} X'y$$

Assuming the errors are normal and σ^2 is known, the log-likelihood function is

$$l(\beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(y_i - \beta' x_i)^2$$

and the MLE is OLS

$$l'(\beta) = \frac{1}{\sigma^2}\Sigma_{i=1}^{n} u_i x_i$$

$$l''(\beta) = -\frac{1}{\sigma^2}\Sigma_{i=1}^{n} x_i x_i'$$

Basic statistical results applying to linear regression

1. $n^{-1/2}\Sigma_{i=1}^{n} u_i x_i \rightarrow_d N(0, \sigma^2 n^{-1}\Sigma_{i=1}^{n} x_i x_i')$
2. $n^{-1}\Sigma_{i=1}^{n} x_i x_i' \rightarrow M$

So $\hat{\beta} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'u$

$$n^{1/2}(\hat{\beta} - \beta) = [n^{-1}\Sigma_{i=1}^{n} x_i x_i']^{-1} n^{-1/2}\Sigma_{i=1}^{n} x_i u_i$$

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2 M^{-1})$$

So in different ways, we can see that

$$I = \frac{1}{\sigma^2} M$$

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2 M^{-1})$$

Wald, LR and LM tests can then be done. For example, the LM test of $\beta = \beta_0$ uses the statistic

$$\frac{1}{\sigma^2}\{\Sigma_{i=1}^{n}(y_i - \beta_0' x_i)^2 - \Sigma_{i=1}^{n}(y_i - \hat{\beta}' x_i)^2\}$$

and the Wald test statistic is

$$n(\hat{\beta} - \beta_0)' \sigma^{-2} M (\hat{\beta} - \beta_0)$$

MLE works as well if the error variance σ^2 is unknown.

The log-likelihood function is

$$l(\beta,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(y_i - \beta' x_i)^2$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2 = \frac{1}{n}\hat{u}'\hat{u}$$

where $\hat{u} = y - X\hat{\beta}$. This is not the unbiased estimator of σ^2, which is $s^2 = \frac{1}{n-k}\hat{u}'\hat{u}$. Also

$$I = \begin{pmatrix} (1/\sigma^2)M & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}$$

The Inverse-Gamma distribution

The inverse-gamma distribution arises in the context of Bayesian analysis of the regression model.

We say that if $\sigma^2 \sim IG(a,b)$ then $\sigma^2 = \dfrac{ab}{\chi^2(b)}$

(Sometimes parameterized differently)

Bayesian approach. We'll consider two priors, one non-informative and one informative.

Prior 1. Diffuse Prior

$p(\beta,\sigma^2) \propto 1/\sigma^2$ for $-\infty < \beta < \infty$ and $\sigma^2 > 0$

With this prior, the posterior is

$$p(\beta \mid \sigma^2, y, X) \sim N(\hat{\beta}, \sigma^2 (X'X)^{-1})$$

$$p(\sigma^2 \mid \beta, y, X) \sim IG(\frac{1}{n-k}(y - X\hat{\beta})'(y - X\hat{\beta}), n-k)$$

Recall that in the "location" model with the diffuse prior and known variance:

$$p(\mu,\sigma^2 \mid x) \sim N(\overline{x}, \sigma^2 / n)$$

Prior 2: "Informative" prior

$$p(\beta \mid \sigma^2) \sim N(\overline{\beta}, \sigma^2 A)$$
$$p(\sigma^2) \sim IG(s_0^2, v_0)$$

With this prior, the posterior is

$$p(\beta \mid \sigma^2, y, X) = N(\tilde{\beta}, \sigma^2 (X'X + A^{-1})^{-1})$$
$$\tilde{\beta} = (X'X + A^{-1})^{-1} A^{-1} \overline{\beta} + (X'X + A^{-1})^{-1} X'X \hat{\beta}$$
$$p(\sigma^2 \mid y, X) \sim IG(\frac{1}{v_0 + n}[v_0 s_0^2 + (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta} - \overline{\beta})'(A + (X'X)^{-1})^{-1}(\hat{\beta} - \overline{\beta})], v_0 + n)$$

Recall that in the "location" model with an informative $N(\theta, \tau^2)$ prior and known variance:

$$\mu \mid X \sim N\left(\frac{n\tau^2}{n\tau^2+\sigma^2}\bar{X} + \frac{\sigma^2}{n\tau^2+\sigma^2}\theta, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right)$$

This is the same as in the regression model with $\bar{\beta} = \theta$ and $A = \tau^2 / \sigma^2$

The bootstrap

The bootstrap is a simulation method for forming confidence intervals and obtaining standard errors using only information from the sample.

Like a Monte-Carlo simulation, but uses only the data.

Advantages

1. Applicable in a wide range of contexts
2. Easy.
3. The bootstrap in some cases may produce better approximations

Basic idea of bootstrap

Suppose that $X_1, X_2, ...X_n$ is iid with a parameter θ and an estimator $\hat{\theta}$.

We can resample from $X_1, X_2, ...X_n$ with replacement (otherwise just reordering) and get a new estimate $\hat{\theta}_i^{boot}$.

Can repeat n_B times

Estimate standard error of $\hat{\theta}$ as $\sqrt{\frac{1}{n_B - 1} \Sigma_{i=1}^{n_B} (\hat{\theta}_i^{boot} - \hat{\theta})^2}$

Bootstrapping the binomial mean: a simple illustration

Suppose that we add up two independent Bernoulli random variables, each of which is 1 with probability p and 0 otherwise.

Let $\hat{p}$ be the average of this sample of 2…an unbiased estimate of p.

Suppose that we observe 1 and 0

Then $\hat{p} = 1/2$ and the standard error is $\sqrt{\frac{\hat{p}(1-\hat{p})}{2}} = \frac{1}{\sqrt{8}}$.

Now let's work this out by the bootstrap.

We observe 1 and 0.

Resampling with replacement will give the four following outcomes each with equal probability

Draw 1	Draw 2	Estimate of p
0	0	0
0	1	0.5
1	0	0.5
1	1	1

Across the bootstrap samples, the variance of the estimates of p is $(0 + 0.5^2 + 0.5^2 + 0) / 4 = 1/8 \Rightarrow SE = 1/\sqrt{8}$.

Exactly right!

Say I want a 95 percent confidence interval for θ.

Sort the bootstrap estimates $\hat{\theta}_i^{boot}$.

Let F_α^{boot} denote the percentiles of the distribution of $\hat{\theta}_i^{boot}$.

A confidence interval for θ is $[F_{0.025}^{boot}, F_{0.975}^{boot}]$.

Hall justifies the bootstrap with a “Russian Dolls” analogy

- Doll zero: population that we do not get to see
- Doll one: sample we observe
- Doll two: bootstrap sample

- Key idea: Doll 0 is to doll 1 as doll 1 is to doll 2
- Population is to sample as sample is to bootstrap sample

Pretend that we know the distribution of $\hat{\theta} - \theta$ and assume that this is pivotal (does not depend on θ). Call it G.

If we knew G, a 1-2α confidence interval for θ would be

$$\{\theta : G_{\alpha} \le \hat{\theta} - \theta \le G_{1-\alpha}\}$$
$$= \{\theta : -G_{1-\alpha} \le \theta - \hat{\theta} \le -G_{\alpha}\}$$
$$= \{\theta : \hat{\theta} - G_{1-\alpha} \le \theta \le \hat{\theta} - G_{\alpha}\}$$

But we don't know G

We do however know the distribution of $\hat{\theta}_i^{boot}$ and can assume that the distribution of $\hat{\theta}_i^{boot} - \hat{\theta}$ is the same as the distribution of $\hat{\theta} - \theta$ (assuming pivotalness again).

$G_\alpha = F_\alpha^{boot} - \hat{\theta}$ and $G_{1-\alpha} = F_{1-\alpha}^{boot} - \hat{\theta}$

The confidence interval is then

$$\{\theta : \hat{\theta} - (F_{1-\alpha}^{boot} - \hat{\theta}) \le \theta \le \hat{\theta} - (F_\alpha^{boot} - \hat{\theta})\}$$
$$= \{\theta : 2\hat{\theta} - F_{1-\alpha}^{boot} \le \theta \le 2\hat{\theta} - F_\alpha^{boot}\}$$

But alas: in most cases $\hat{\theta} - \theta$ is not pivotal

Suppose that $\frac{\hat{\theta}-\theta}{s_{\hat{\theta}}}$ is pivotal (at least asymptotically).

Let $\tilde{F}$ be its distribution.

If $\tilde{F}$ is known, then

$\{\theta : \hat{\theta} - s_{\hat{\theta}}\tilde{F}_{1-\alpha} \le \theta \le \hat{\theta} - s_{\hat{\theta}}\tilde{F}_{\alpha}\}$

is a $1-2\alpha$ confidence interval for θ.

Now let $\tilde{F}^{boot}$ denote the bootstrap distribution of $\frac{\hat{\theta}_i^{boot}-\hat{\theta}}{s_{\hat{\theta}}}$.

This should be the same (asymptotically) as $\tilde{F}$

CI: $\{\theta : \hat{\theta} - s_{\hat{\theta}}\tilde{F}^{boot}_{1-\alpha} \le \theta \le \hat{\theta} - s_{\hat{\theta}}\tilde{F}^{boot}_{\alpha}\}$

Three ways of getting a bootstrap confidence interval in Hall's terminology:

1. Other percentile:
   $[F_{\alpha}^{boot}, F_{1-\alpha}^{boot}]$.
2. Percentile:
   $[2\hat{\theta} - F_{1-\alpha}^{boot}, 2\hat{\theta} - F_{\alpha}^{boot}]$
3. Percentile-t:
   $[\hat{\theta} - s_{\hat{\theta}}\tilde{F}_{1-\alpha}^{boot}, \hat{\theta} - s_{\hat{\theta}}\tilde{F}_{\alpha}^{boot}]$

From the Edgeworth expansion, we saw that the cdf of $\frac{\bar{X} - \mu}{\sigma}$ was $F_n(x) = \Phi(x) + O(n^{-1/2})$.

An asymptotic confidence interval should have coverage that is the nominal coverage plus $O(n^{-1/2})$.

True for the other percentile and percentile bootstraps too.

But the percentile-t bootstrap gives a "higher order" refinement: error in coverage probability is $O(n^{-1})$.

In practice though, the percentile and other percentile methods can work well in small samples.

Example of the bootstrap

Suppose that these are the speeds of a sample of seven cars on a stretch of highway:
79 73 68 77 86 71 69

Form a 95 percent confidence interval for the population mean (a) asymptotically, and (b) via the bootstrap.

$\bar{X} = 74.71,\ s = 6.40,\ n = 7$

The 95 percent asymptotic confidence interval is

$$74.71 \pm \frac{2.45 * 6.4}{\sqrt{7}}$$

which is from 68.78 to 80.64.

Bootstrap: Simple Matlab code

```
x=[79  73  68  77  86  71  69]';
xbar=mean(x); se=std(x)/sqrt(7);
nboot=1000;
for imc=1:nboot;
    xboot=x(ceil(7*rand(7,1)));
    xbarboot(imc)=mean(xboot);
    tstatboot(imc)=(mean(xboot)-xbar)/se;
end;
xbarboot=sort(xbarboot); tstatboot=sort(tstatboot);
```

[xbarboot(25) xbarboot(975)]

[xbar-(tstatboot(975)*se) xbar-(tstatboot(25)*se)]

The last two lines give the OP and Percentile t CIs for mean

Confidence Intervals

	Lower Bound	Upper Bound
Asymptotic	68.78	80.64
Other Percentile	70.57	79.71
Percentile-t	69.71	78.86

In this case, they are all fairly similar.

But the bootstrap ones needed no asymptotic theory

Bootstrap test

Suppose that we have a test statistic for a hypothesis $\theta = \theta_0$, based on iid data $X_1,...X_n$, such as $t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$

We don't know it's distribution.

We can approximate it's distribution by creating bootstrap samples of the same test, but of the hypothesis $\theta = \hat{\theta}$

$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{SE(\hat{\theta}^*)}$

We can use the bootstrap in linear regression. Model is $y_i = \beta x_i + \varepsilon_i$

A standard bootstrap would resample pairs (y_i, x_i)

Can then get a bootstrap standard error for the OLS estimate of β or a bootstrap test or confidence interval

Some alternative bootstraps

- Hold X fixed and resample the residuals.
- Hold X fixed and multiply the randomize the sign of the residuals (wild bootstrap).

Density estimation

Suppose that $X_1, X_2, ...X_n$ are iid with some density f. We want to estimate f.

One approach is to specify a model, like that the density is normal and estimate the parameters.

A nonparametric approach avoids writing down a particular model.

Goal is minimizing integrated risk: $\int E(\hat{f}(x) - f(x))^2 dx$

With a parametric model, risk is $O(\frac{1}{n})$.

Example: $f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$ and apply delta method to get

$\hat{f}(x) - f(x) = O_p(n^{-1/2})$

Hence $(\hat{f}(x) - f(x))^2 = O_p(n^{-1})$

A histogram is a simple non-parametric density estimator.

Of the n observations, let n_j be the number falling in bin j and let h be the width of each bin.

The estimate of the density is equal for all points in the bin and is $\frac{n_j}{nh}$.

Let $p_j = \int_{B_j} f(u)du$ be the probability of being in bin B_j which we normalize to go from $(j-1)h$ to jh.

If B_j is the bin containing a fixed x, then

$$E(\hat{f}(x)) = \frac{p_j}{h}$$

$$Var(\hat{f}(x)) = \frac{p_j(1-p_j)}{nh^2}$$

If the bin width is small, $\int_{B_j} f(u)du \simeq hf(x)$ and hence

$$E(\hat{f}(x)) \simeq \frac{hf(x)}{h} = f(x)$$

Suppose that for any u in B_j $\quad f(u) \simeq f(x) + (u-x) f'(x)$

Then

$$p_j = \int_{B_j} f(u)du = \int_{B_j} [f(x) + (u-x) f'(x)]du$$

$$= f(x)h + f'(x)[\frac{(u-x)^2}{2}]_{h(j-1)}^{hj}$$

$$= f(x)h + f'(x)[\frac{(hj-x)^2 - (h(j-1)-x)^2}{2}]$$

$$= f(x)h + f'(x)[\frac{h^2 j^2 + x^2 - 2xhj - h^2 j^2 - h^2 + 2jh^2 - x^2 + 2xhj - 2xh}{2}]$$

$$= f(x)h + f'(x)[\frac{-h^2 + 2jh^2 - 2xh}{2}]$$

$$= f(x)h + f'(x)h[h(j - \frac{1}{2}) - x]$$

The bias of the estimator in general is

$$\frac{f(x)h + hf'(x)(h(j-\frac{1}{2}) - x)}{h} - f(x) = f'(x)(h(j-\frac{1}{2}) - x)$$

MSE of the estimator is bias-squared plus variance.

Average bias squared is $\frac{h^2}{12}\int f'(x)^2\,dx$

Variance is $\frac{1}{nh}$

So risk is $\frac{h^2}{12}\int f'(x)^2\,dx + \frac{1}{nh}$

The choice of h that minimizes risk is $\frac{1}{n^{1/3}}\left(\frac{6}{\int f'(u)^2\,du}\right)^{1/3}$.

With this choice of binwidth, the risk is $\frac{C}{n^{2/3}}$.

A practical way to choose the binwidth is by cross-validation.

The loss function is

$$\int (\hat{f}(x) - f(x))^2 dx$$

$$= \int \hat{f}(x)^2 dx - 2\int \hat{f}(x) f(x) dx + \int f(x)^2 dx$$

The last term doesn't depend on the binwidth, so minimizing binwidth amounts to minimizing

$$L = \int \hat{f}(x)^2 dx - 2\int \hat{f}(x) f(x) dx = \int \hat{f}(x)^2 dx - 2E[\hat{f}(x)]$$

The cross-validation estimator of this is

$$\hat{L} = \int \hat{f}(x)^2 dx - \frac{2}{n}\Sigma_{i=1}^{n} \hat{f}_{(i)}(X_i)] = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\Sigma(\frac{n_j}{n})^2$$

The kernel density estimator is

$$\hat{f}(x) = \frac{1}{n}\Sigma_{i=1}^{n}\frac{1}{h}K(\frac{x - X_i}{h})$$

$K()$ is any nonnegative smooth function such that $\int K(x)dx = 1$ and $\int xK(x)dx = 0$.

A histogram is like a kernel estimator where $K(.)$ is the indicator that the argument is in a bin.

Two widely used kernels are

$$K(x) = \frac{3}{4}\frac{1 - 0.2x^2}{\sqrt{5}}1(x^2 < 5)$$ Epanechnikov

$$K(x) = \phi(x)$$ Gaussian

For general density f and kernel K

Risk is

$$\frac{1}{4}[\int x^2 K(x)dx]^2 h^4 \int f''(x)^2 dx + \frac{\int K^2(x)dx}{nh}$$

Minimizing this with respect to the bandwidth h gives $h^* = O(n^{-1/5})$ and a risk $\frac{C}{n^{4/5}}$.

Asymptotic distribution if $h = Cn^{-1/5}$

$$\sqrt{nh}(\hat{f}(x) - f(x)) \rightarrow_d N(B,V)$$

A sequence of functions $\phi_1, \phi_2, \ldots$ is orthonormal on [a,b] if $\int_a^b \phi_j^2(x)dx = 1$ and $\int_a^b \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$.

The cosine basis is orthonormal on [0,1]. They are $\phi_j(x) = \sqrt{2}\cos(\pi jx)$

Any density with support on [a,b] can be written as
$f(x) = \Sigma_{j=0}^{\infty} \beta_j \phi_j(x)$ where $\phi_0(x) = 1$

Define $\hat{\beta}_j = \frac{1}{n} \Sigma_{i=1}^{n} \phi_j(X_i)$

$E(\hat{\beta}_j) = \beta_j$

$Var(\hat{\beta}_j) = \int (\phi_j(x) - \beta_j)^2 \, f(x) dx = \sigma_j^2$

The orthonormal density estimator is

$\hat{f}(x) = 1 + \Sigma_{j=1}^{J} \hat{\beta}_j \phi_j(x)$

The number of terms J is like the bandwidth. Increasing it reduces bias but increase variance.

The risk of the estimator is

$$\Sigma_{j=1}^{J} \frac{\sigma_j^2}{n} + \Sigma_{j=J+1}^{\infty} \beta_j^2$$

An estimator of the risk is

$$\Sigma_{j=1}^{J} \frac{\hat{\sigma}_j^2}{n} + \Sigma_{j=J+1}^{\bar{J}} \max(\hat{\beta}_j^2 - \frac{\hat{\sigma}_j^2}{n}, 0)$$

$$\hat{\sigma}_j^2 = \frac{1}{n-1} \Sigma_{i=1}^{n} (\phi_j(X_i) - \hat{\beta}_j)^2$$

Maximum likelihood estimator of the density

Likelihood function: $L(\hat{f}) = \Pi_{i=1}^{n} \hat{f}(X_i)$

This is obviously maximized by having all the mass at the observed points, which is a silly density estimator.

A variant on this introduces a smoothness penalty.

$$\hat{f} = \arg\max_f \log L(f) - \frac{\lambda}{2} \int \frac{f'^2}{f}$$

Density estimation methods have regression analogs

$$y_i = f(x_i) + \varepsilon_i$$

$$\hat{f}(x) = \frac{\Sigma 1(|\, x - X_i \,| < h) Y_i}{\Sigma 1(|\, x - X_i \,| < h)}$$

Like a histogram, and called a “binning estimator”.

Can have

$$\hat{f}(x) = \frac{\Sigma_{i=1}^{n} K(\frac{x - X_i}{h}) Y_i}{\Sigma_{i=1}^{n} K(\frac{x - X_i}{h})}$$

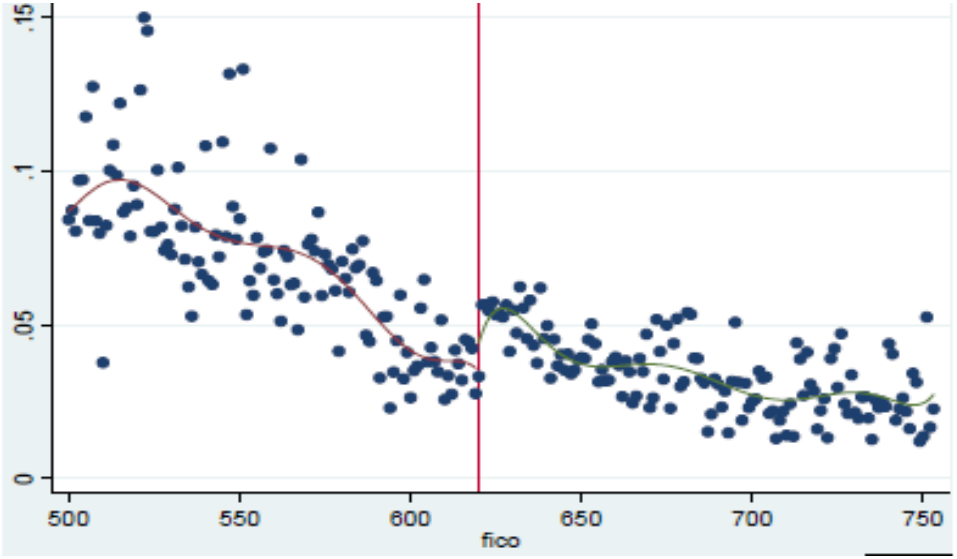
for a kernel function as before. This is called the Nadaraya-Watson estimator.

The risk of the Nadaraya-Watson estimator is $O_p(n^{-4/5})$.

We would typically pick h by a cross-validation criterion minimizing

$\Sigma_{i=1}^{n}(Y_i - \hat{f}_{(-i)}(X_i))^2$

Delinquency rates and FICO scores for loans made in 2002



We can alternatively write $f(x) = \Sigma \beta_j \phi_j(x)$ where $\phi_j(x)$ is an orthonormal series estimator. We then estimate $\hat{\beta}_j = \frac{1}{n} \Sigma_{i=1}^{n} Y_i \phi_j(X_i)$ and use a series estimator.

Principal components

Often in economic applications of statistics, there will be several series that have some common component and we would like to estimate that common component.

e.g. many indicators of the business cycle.

Let $X_1(t), X_2(t), ...X_n(t)$ denote a set of n time series variables with mean zero and covariance matrix Σ.

Task: Find the linear combination of these n series with the biggest variance.

$Var(\Sigma_{i=1}^{n}\lambda_i X_i(t)) = \lambda'\Sigma\lambda$ where $\lambda = (\lambda_1, \lambda_2, ...\lambda_n)'$.

So the problem is to find

$$\lambda^{(1)} = \arg\max_{\lambda:||\lambda||=1} \lambda'\Sigma\lambda$$

Solution: $\lambda^{(1)}$ is the eigenvector of Σ corresponding to the largest eigenvalue of Σ, normalized to have length 1.

$\Sigma_{i=1}^{n}\lambda_i^{(1)} X_t = X\lambda^{(1)}$ is then the "first" principal component.

Now let

$$\lambda^{(2)} = \arg\max_{\lambda:\|\lambda\|=1, Cov(X\lambda^{(1)}, X\lambda=0)} \lambda'\Sigma\lambda$$

Solution: $\lambda^{(2)}$ is the eigenvector of Σ corresponding to the second largest eigenvalue of Σ, normalized to have length 1.

$\Sigma_{i=1}^{n}\lambda_i^{(2)}X_t = X\lambda^{(2)}$ is then the "first" principal component.

If $\lambda^{(j)}$ is the eigenvector of Σ corresponding to the *j*th largest eigenvalue of Σ, normalized to have length 1, then

$$\Sigma_{i=1}^{n}\lambda_i^{(j)}X_t = X\lambda^{(2)}$$

is the *j*th principal component.

All principal components are uncorrelated with each other.

- For principal component analysis, the series must have mean zero. We subtract the mean off from each series first.

- As defined, principal component analysis will not be scale invariant (changing units changes principal components).

- If the variables are divided by their standard deviation first, then it will be scale invariant.

- This is principal component analysis based on the *correlation* matrix, not the *covariance* matrix.

If many series that seem to have much in common.

What's the first principal component?

Recipe.

1. Subtract the mean from each series.
2. Divide each by the standard deviation.
3. Compute the variance-covariance matrix.
4. Find the eigenvector corresponding to the largest eigenvalue.
5. Weight each series by the element of this eigenvector.

Factor analysis

A related (but distinct) method. Let $X_1(t), X_2(t), ...X_n(t)$ denote a set of n times series. Suppose that

$$X_i(t) = \beta_i f(t) + \varepsilon_i(t) \quad \text{for } i = 1, ...n$$

where $\varepsilon_1(t), \varepsilon_2(t), ...\varepsilon_n(t)$ are n white noise processes with variance σ_ε^2 that are all mutually independent and $f(t)$ is iid $N(0, \sigma_f^2)$ that is independent of all the εs.

Example: The Xs are returns on individual stocks and $f(t)$ is the market return.

Let $X_t = (X_1(t), X_2(t), \ldots X_n(t))'$ and $\beta = (\beta_1, \beta_2, \ldots \beta_n)'$

X_t is iid $N(0, \Sigma)$ where $\Sigma = \beta\beta'\sigma_f^2 + diag(\sigma_\varepsilon^2)$.

We could always double β and halve σ_f. So we adopt the normalization $\sigma_f^2 = 1$. This means that $\Sigma = \beta\beta' + diag(\sigma_\varepsilon^2)$ and we just have two parameters to estimate.

The pdf of X_t (assuming normality) is

$$(2\pi)^{-n/2} \, |\Sigma|^{-1/2} \, \exp(-\frac{1}{2} X_t' \Sigma^{-1} X_t)$$

The log pdf of X_t is

$$-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2} X_t' \Sigma^{-1} X_t$$

So the log-likelihood function is

$$-\frac{nT}{2}\log(2\pi) - \frac{T}{2}\log(|\Sigma|) - \frac{1}{2}\Sigma_{t=1}^{T} X_t' \Sigma^{-1} X_t$$

No closed form maximum in this case, but it can be maximized numerically wrt β and σ_ε^2.

We could have multiple factors.

Suppose that

$$X_i(t) = \beta_{i1} f_1(t) + \beta_{i2} f_2(t) \ldots + \beta_{ik} f_k(t) + \varepsilon_i(t) \quad \text{for } i = 1, \ldots . n$$

where the $f(t)$ is iid $N(0, \Sigma_f)$. Let β be the *n*x*k* matrix $[\beta_{ij}]$. Then X_t is iid $N(0, \Sigma)$ where $\Sigma = \beta \Sigma_f \beta' + diag(\sigma_\varepsilon^2)$

We need to adopt a normalization (e.g. diagonal elements of Σ_f are all 1) and can then again estimate by MLE.