Expectations, moments and transformations

It is often convenient to consider a set of simple numbers that describe some dominant features of the random variable.

One such number is the **expected value or mean**, a measure of **location**. It gives the center of mass of a probability distribution.

- For **discrete** variables, the expected value is defined as

$$E(X) = \sum_{x \in \mathbb{X}} x p(x)$$

- For **continuous** variables, it is

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

3.2 Existence of expected values for discrete variables

The expectation of a random variable X is not always defined. For a discrete random variable X taking on values in the set

$$\{x_1, x_2, ...\}$$

to have a defined expected value,

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i)$$

we must check the condition that the series is **absolutely convergent**, that is

$$\sum_{i=1}^{\infty} |x_i| P(X = x_i) < \infty$$

Example. Infinite expected value.

Let X be the random variable taking values $2, 2^2, 2^3, \ldots$ with probabilities

$$P(X = 2^i) = \frac{1}{2^i} \quad for \quad i = 1, 2, \ldots$$

Then p_X is a valid probability function, because

$$p_X(x) \geq 0 \quad \forall x \quad and \quad \sum_x p_X(x) = 1$$

On the other hand

$$E(X) = \sum_{i=1}^{\infty} 2^i p(X = x_i) = \sum_{i=1}^{\infty} 2^i \frac{1}{2^i} = \sum_{i=1}^{\infty} 1 = \infty$$

so the expected value of X is infinite.

Let Y be a discrete random variable with probability function

$$p_Y(y) = \begin{cases} \frac{1}{2y} & \text{for} \quad y = 2, 4, 8, 16, \dots \\ \frac{1}{2|y|} & \text{for} \quad y = -2, -4, -8, -16, \dots \\ 0 & otherwise \end{cases}$$

So

$$p_Y(2) = p_Y(-2) = 1/4$$
$$p_Y(4) = p_Y(-4) = 1/8$$
$$p_Y(8) = p_Y(-8) = 1/16$$

and so on.

p_Y is indeed a valid probability function. But the expected value of Y does not exist

$$E(Y) = \sum_y y p_Y(y) = \sum_{k=1}^{\infty} (2^k) \frac{1}{2 \times 2^k} + \sum_{k=1}^{\infty} (-2^k) \frac{1}{2 \times 2^k} =$$

$$= \sum_{k=1}^{\infty} (1/2) - \sum_{k=1}^{\infty} (1/2) = \infty - \infty$$

3.5 Existence of expected values for continuous variables

Example. Expected value is infinite. Suppose we are given a continuous random variable X with density

$$f_X(x)) = \begin{cases} 0 & \text{for} \quad x < 1 \\ \frac{1}{x^2} & \text{for} \quad x \geq 1 \end{cases}$$

First let us check whether it is a proper density. This is a well-defined density because $f_X(x) \geq 0$ is piecewise continuous and

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_1^{\infty} \frac{1}{x^2} dx = \left[-\frac{1}{x} \right]_{x=1}^{\infty} = 1$$

However, the expectation of X is infinite:

$$\int_{-\infty}^{\infty} x f_X(x) dx = \int_1^{\infty} \frac{1}{x} dx = \infty$$

3.6 Moments and central moments

The moments of a distribution form an important class of expectations.

- The n-th moment of X is defined as

$$m_n = E(X^n) = \sum_{x \in \mathbb{X}} x^n p_X(x)$$

  for **discrete** variable, and

- $$m_n = E(X^n) = \int_{-\infty}^{\infty} x^n f_X(x) dx$$

  for **continuous** variables, where $p_X(x)$ is the probability mass function and $f_X(x)$ is the probability density function.

The **expected value** is therefore the first moment of a distribution

Central moments are defined in a very similar way after substracting the first moment.

Aside from the expected value (first moment) $E(X)$, a very important feature of a probability distribution is captured by the second central moment or **variance**

$$\sigma^2 = Var(X) = E[(X - E(X))^2]$$

For **discrete** random variables,

$$Var(X) = \sum_{x \in \mathbb{X}} (x - \mu)^2 p_X(x)$$

and for **continuous** random variables

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

where $\mu = E(X)$

Variance measures the **degree of spread** of a distribution, i.e., how widely all values are distributed around the mean

Two density functions may have the same mean but different variances.

A working expression for the variance is given by:

$$\sigma^2 = E[(X - E(X))^2] = E(X^2) - E(X)^2 \geq 0$$

The variance cannot be negative!

Example. A continuous random variable X has probability density

$$f_X(x)) = \begin{cases} 2x & \text{for} \quad 0 < x \leq 1 \\ 0 & else \end{cases}$$

Then

$$E(X) = \int_0^1 x(2x)dx = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

$$E(X^2) = \int_0^1 x^2(2x)dx = 2\left[\frac{x^4}{4}\right]_0^1 = \frac{1}{2}$$

$$Var(X) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{8}$$

3.9 The variance may be undefined

Let X be a random variable with probability mass function

$$P\left(X = \frac{3^k}{2^k}\right) = \frac{1}{2^k}, \quad k = 1, 2, 3...$$

The distribution is well defined since

$$\sum_{k=1}^{\infty} \frac{1}{2^k} = 1$$

The expectation converges:

$$E(X) = \sum_{k=1}^{\infty} \frac{3^k}{2^k} \frac{1}{2^k} = \sum_{k=1}^{\infty} \left(\frac{3}{4}\right)^k = 3$$

but

$$E(X^2) = \sum_{k=1}^{\infty} \left(\frac{3^k}{2^k}\right)^2 \frac{1}{2^k} \rightarrow \textit{diverges!}$$

so the variance is not defined.

3.10 SD and variation

The **standard deviation** σ is the positive square root of the variance, that is

$$\sigma = \sqrt{Var(X)}$$

Note that

The measurement unit forthe standard deviation is the same as that for the original variable X

The measurement unit on the variance is the *square* of the original unit.

The ratio σ/μ is called **coefficient of variation** and is the measure of dispersion of a probability distribution

- It is useful when comparing the degree of variation from one data set to another even when the means of the distributions are different

- It is often reported as a percentage

We may be interested to obtain an **upper bound** for the probability

$$P(X \geq x)$$

by using some well-known inequalities. Let X be a non-negative continuous random variable with finite expectation $E(X)$.

The **Markov's inequality** states that

$$P(X \geq a) \leq \frac{E(X)}{a} \quad \forall a > 0$$

Proof. Consider a p.d.f. with $f_X(x) = 0$ for $x < 0$. Remember that

$$P(X \geq a) = \int_a^\infty f_X(x)dx$$

then

$$E(X) = \int_0^\infty x f_X(x)dx \geq \int_a^\infty x f_X(x)dx \geq a \int_a^\infty f_X(x)dx$$

Hence the inequality

$$\int_a^\infty f_X(x)dx = P(X \geq a) \leq \frac{E(X)}{a}$$

Assuming that X has a finite expectation $\mu = E(X)$, **Chebyshev's inequality** states that

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}$$

- Thsi inequality allows us to estimate the likelihood of a deviation of a random variable X from its mean value even if little information is available about the distribution of X.
- It is a simple application of Markov's inequality. Recall that

$$P(X \geq a) \leq \frac{E(X)}{a} \quad \forall a > 0$$

Substract the mean $\mu = E(X)$ and take the absolute value, then

$$P(|X - E(X)| \geq a) = P(|X - E(X)|^2 \geq a^2) \leq \frac{E([X - \mu]^2)}{a^2} = \frac{Var(X)}{a^2}$$

Let X be a random variable with mean 15 and variance 9.

Compute:

(a) the **maximum** value for $P(|X - 15| \geq 10)$

(b) the **maximum** value for $P(X \leq 10 \cup X \geq 20)$

(c) the **minimum** value for $P(|X - 15| \leq 10)$

These cases can be solved by applying Chebyshev's inequality:

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}$$

Some univariate discrete and continuous distributions

Permutations

Suppose we have a set S with n elements and we want to count the number of permutations of length $k \leq n$ obtained from S, that is the number of elements of the set

$$\{(s_1, \ldots, s_k) : s_i \in S, s_i \neq s_j \text{ when } i \neq j\}$$

- We have n choices for the first elements s_1
- $n-1$ choices for the second element s_2
- ... and finally $n-(k-1) = n-k+1$ choices for the last element s_k

Therefore there are

$$n(n-1)\cdots(n-k+1)$$

permutations of length k from a set of n elements, which can be written as

$$\frac{n!}{(n-k)!}$$

Notice that, when $k = n$ there are

$$n! = n(n-1)\cdots 2 \cdot 1$$

permutations of length n

Three genes have been previosuly classified as beloging to three different biological pathways. Suppose that the correct classification scheme is lost and a random classification is attempted. What is the probability that each of the three genes is assigned to the correct class?

Here we have a set $S = \{1, 2, 3\}$, and there are

$$3! = 3 \times 2 \times 1 = 6$$

permutations of its elements, that is

$$\{1,2,3\}, \{1,3,2\}, \{2,1,3\}$$

$$\{2,3,1\}, \{3,1,2\}, \{3,2,1\}$$

Only one of these assignments is correct, so the probability is $1/6$

3.17 Summary of counting rules

Summary of counting rules

- Very often we will need to count the number of ways of arranging r elements from a larger set of n elements.
- This count depends on whether we need to keep the elements in **order**, and whether we sample with or without **replacement**.

Here is a summary:

	With Replacement	Without Replacement
Ordered	n^r	$\frac{n!}{(n-r)!}$
Unordered	$\binom{n+r-1}{r}$	$\binom{n}{r}$

3.18 Bernoulli distribution

A **Bernoulli experiment or trial** is an experiment having only *two* possible outcomes, e.g. $0/1$ or A/B. Usually the outcomes represent success or failure of an event.

Consider the random variable X defined as

$$X = \{\ \textbf{outcome is a success}\ \}$$

A random variable X is called **Bernoulli** if the probability mass function is given by

$$p_X(k) = \mathsf{P}(X = k|p) = p^k(1-p)^{1-k} \quad k = 0, 1$$

where $0 \leq p \leq 1$

Or more simply, when $\mathsf{P}(X = 1) = p$ and $\mathsf{P}(X = 0) = 1 - p$

- The **cumulative distribution function** is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

- **Expected value and variance** are given by

$$\mathsf{E}(X) = \sum_{x=0}^{1} x\, p_X(x) = (0)(1-p) + (1)p = p$$

and

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 = p - p^2 = p(1-p)$$

3.19 Binomial distribution

- Consider a sequence of n **independent Bernoulli trials** in each of which there are only two possible results:

$$\mathrm{P}(\text{success}) = p \qquad \mathrm{P}(\text{failure}) = 1 - p$$

- We are interested in computing the probability that we observe r successes in n trials
- For a given n, let X be the random variable defined as

$$X = \{ \textbf{ number of successes in } n \textbf{ trials } \}$$

- The probability of observing **one particular sequence** with r successes and $n - r$ failures is given by

$$p^r (1-p)^{n-r}$$

- The event $\{X = r\}$ contains

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

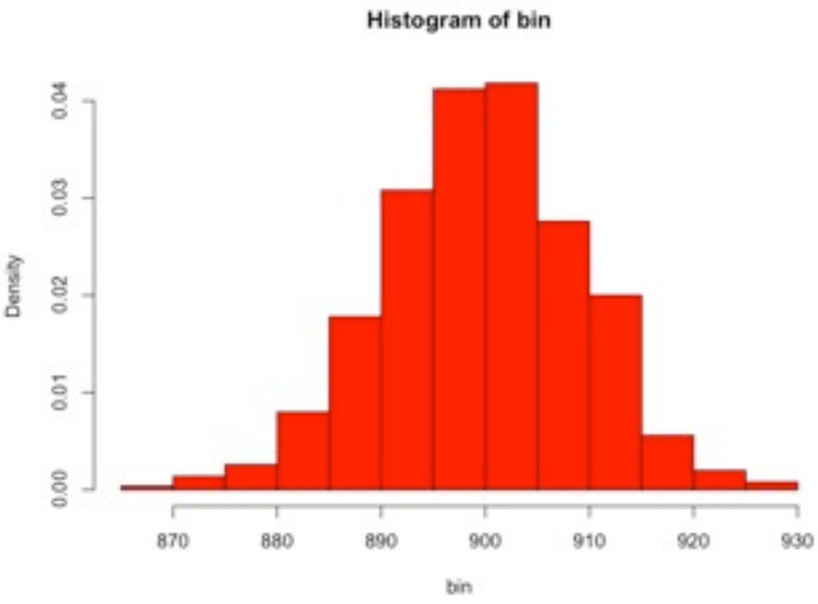
such sequences. Therefore

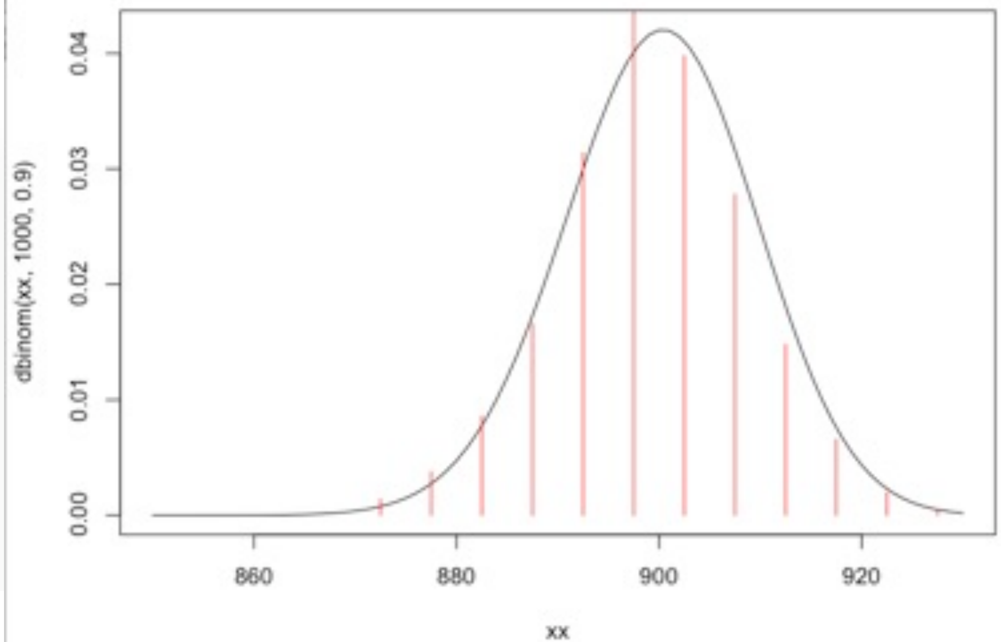
A random variable with probability mass function

$$p_X(r) = \mathrm{P}(X = r | n, p) = \binom{n}{r} p^r (1-p)^{n-r} \quad r = 0, 1, ..., n$$

is called **Binomial**. It provides the probability of observing r successes in n (independent) trials. It has two parameters and we refer to it as $\mathrm{Bin}(n, p)$.

```
bin=rbinom(2000,1000,0.9)
h=hist(bin,freq=F,col="red")
xx=seq(850,930,by=1)
plot(xx,dbinom(xx,1000,0.9),type="l")
points(h$mids,h$intensities,type="h",col="red")
```





It can be easily checked that the probability mass function given before indeed defines a proper probability distribution.

- First, recall the **binomial theorem**: for any real numbers x and y and any integer $n \geq 0$,

$$\sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r} = (x+y)^n$$

- We can check that the Binomial p.m.f. $p_X(x)$ is indeed a genuine probability function, that is

$$\sum_{r=0}^{n} p_X(x) = \sum_{r=0}^{n} \mathrm{P}(X=r) = 1$$

by taking $x = p$ and $y = 1-p$. Then apply the binomial theorem

$$\sum_{r=0}^{n} p_X(x) = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r}$$
$$= \{p + (1-p)\}^n = 1$$

- Consider a collection X_i, $i = 1, \ldots, n$, of independent **Bernoulli** random variables all having equal success probability p. That is, they are **independent and identically distributed**.
- Recall that mean and variance of each X_i are given by

$$\mathsf{E}(X_i) = p$$

and

$$\mathsf{Var}(X_i) = p(1-p)$$

- Let us now define a random variable Y as

$$Y = \sum_{i=1}^{n} X_i$$

then Y is distributed as a Binomial random variable with parameters n and p,

$$Y \sim \mathsf{Bin}(n, p)$$

- Using the linearity property of the expectation operator, mean and variance of Y are then easily derived as

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = nE(X_i) = np$$

and

$$Var(Y) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = n\,Var(X_i) = np(1-p)$$

3.23 Poisson distribution

The number of random occurrences that may occur in a given unit of reference (such as time or space) can be modeled by the **Poisson distribution**.

We assume that events occur in a given time period with an **constant rate** $\lambda > 0$. We define a random variable X as

$$X = \{\textbf{number of occurrences in a given interval}\}$$

The probability mass function of a **Poisson** random variables is given by

$$p_X(x) = \mathsf{P}(X = k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!} \quad k = 0, 1, 2, \ldots$$

where $\lambda > 0$ is the **intensity parameter.** In short, Pois(λ)

```
plot(0:10,dpois(0:10,0.7),type="l")
```

The mechanism that gives rise to this density involves the following assumptions:

- Events are rare (the probability of an event in a unit of reference is small)
- Events are independent
- Events are equally likely to occur at any interval of the reference unit
- The probability that events happen simultaneously is negligible (for all practical purposes it is zero).

When the reference is time, this is considered another important **waiting time distribution**. In genetics it arises in many situations,

- Model the distribution of mutations
- Model the distribution of recombination rates
- ...

Moments of the Poisson distribution

- As usual, we may want to check that

$$\sum_{x=0}^{\infty} \mathsf{P}(X = x) = 1$$

- Recalling the Taylor series expansion

$$e^y = \sum_{i=0}^{\infty} \frac{y^i}{i!}$$

it follows that

$$\begin{aligned}\sum_{x=0}^{\infty} \mathsf{P}(X = x) &= e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right) \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} e^{\lambda} = 1\end{aligned}$$

- Mean and variance of a Poisson random variable are

$$\mathsf{E}(X) = \mathsf{Var}(X) = \lambda$$

so this may not be a suitable model when we expect sample mean and variance to be *different*.

The Poisson distribution can be used as a good approximation to the binomial distribution, i.e.

$$e^{-\lambda}\frac{\lambda^k}{k!} \approx \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \ldots, n$$

when

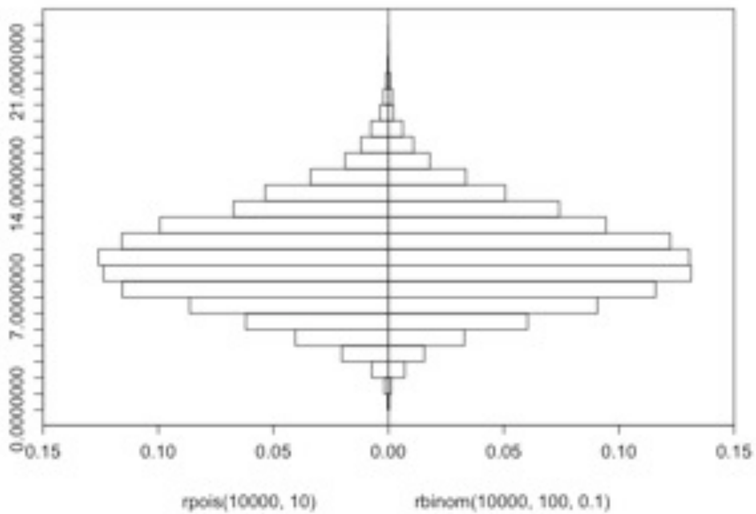
- the number of trials n is very large, so $\binom{n}{k}$ is **hard to compute**
- the probability of success p is very small, so that $np = \lambda$ is fixed

Therefore the Poisson approximation is valid exactly when it is most useful, freeing us from calculation of binomial coefficients and powers for large n.

Example 5.8. Poisson vs binomial

Probability mass functions of a Binomial$(100, 1/10)$ (triangles) and a Poisson(10) (circles) evaluated at points $0, 1, \ldots, 20$

```
> histbackback(rpois(10000,10),rbinom(10000,100,0.1),probability=T,brks=seq(0,24,by=1))
```



3.26 Geometric distribution

So far we have assumed a **fixed** number of trials n. Now suppose n is not known a priori: we continue to run the experiment until the **first success** is observed. Again we assume that at each trial a success occurs independently with fixed probability p and a failure with probability $1-p$.

Therefore now we are interested in studying a random variable X defined as

$$X = \{\textbf{number of trials untill the first success}\}$$

The probability mass function of X is given by

$$p_X(x) = \mathsf{P}(X = x|p) = (1-p)^{x-1}p \qquad x = 1, 2, \ldots$$

which defines the **geometric distribution** or Geom(p).

- This is the simplest of many **waiting time distributions** for discrete variables.
- In order to prove that $p_X(x)$ is a proper p.m.f., recall that for any number r with $|r| < 1$ we have

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{(1-r)}$$

Then it is easily checked that

$$\sum_{x=1}^{\infty} p_X(x) = \sum_{x=1}^{\infty}(1-p)^{x-1}p = p\sum_{x=0}^{\infty}(1-p)^x = p\frac{1}{1-(1-p)} = 1$$

We review the most common and most important continuous univariate random variables. They will be encountered numerous times later and form the basis of a large part of **statistical inference**

When encountering a new distribution it is valuable to know particular characteristics of importance such as

- **Its theoretical importance.** For instance, the normal distribution arises in a plethora of applications because of its association with the so-called central limit theorem, discussed later
- **Its use in applications** Many distributions are often associated with a specific application, but are in fact used in many other contexts as well.
- **How its functional form came about.** For instance it could be
  - be a base distribution arising from mathematical simplicity (e.g. uniform and exponential)
  - be strongly associated with a particular application (e.g. F distribution in the analysis of variance, discussed later)
  - be a generalization of a simpler distribution (e.g. the Gamma generalizes the the Exponential)

Uniform distribution

The uniform is arguably the simplest continuous distribution and is used for modeling situations in which events of equal length in (a, b) are equally likely to occur.

The **uniform distribution** assigns equal probabilities to all outcomes in the interval $[a, b] \subset \Re$. It has probability density function

$$f_X(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{elsewhere} \end{cases}$$

We write that X is Uniform(a, b).

- Alternatively, using the **indicator function**, we could write

$$f_X(x) = \frac{1}{b-a}\mathbf{I}_{[a,b]}(x)$$

where

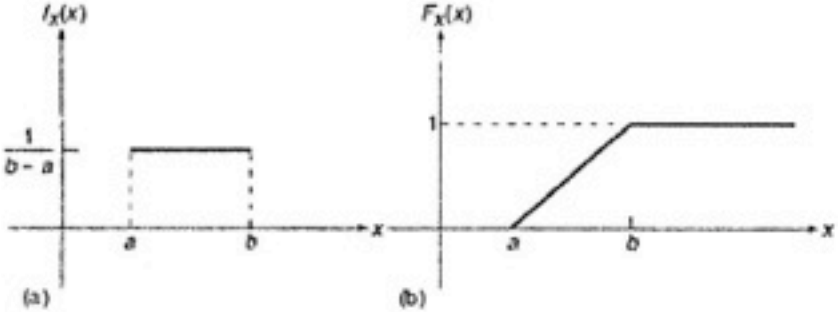
$$\mathbf{I}_{[a,b]}(x) = \begin{cases} 1 & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

- The cumulative distribution function is

$$F_X(x) = \mathrm{P}(X \leq x|a,b) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

- The uniform with $a = 0$ and $b = 1$ is called the **standard uniform distribution**

3.30 Moments of uniform distribution



- The expected value is

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) = \frac{1}{b-a}\int_a^b x dx$$
$$= \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{1}{2}(b+a)$$

- Having obtained

$$E(X^2) = \frac{1}{b-a}\frac{b^3-a^3}{3}$$

the variance is given by

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{1}{12}(b-a)^2$$

- An important special case is the standard uniform. If $X \sim \mathrm{Uniform}(0,1)$, then

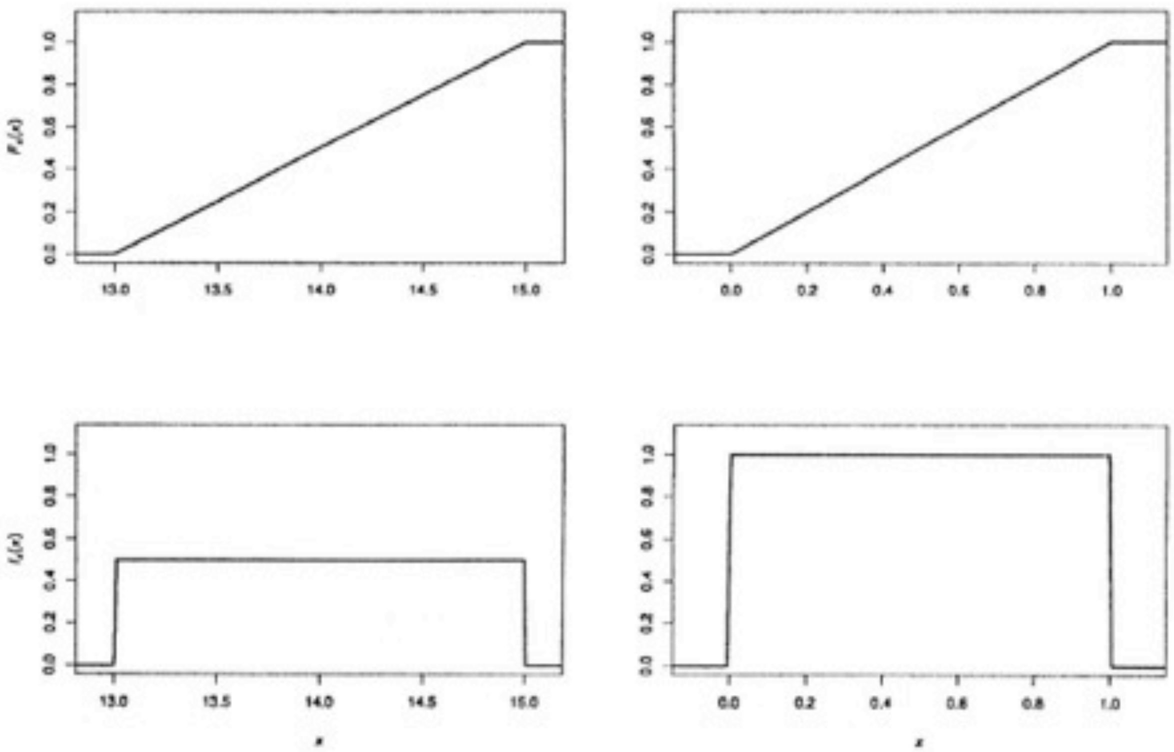
$$f_X(x) = \mathbf{I}_{[0,1]}(x)$$

and

$$F_X(x) = P(X \le x) = x\,\mathbf{I}_{[0,1]}(x)$$

On the left, $X \sim \text{Uniform}(13, 15)$. On the right, $Y \sim \text{Uniform}(0, 1)$.

Shown are the cumulative distribution function (top) and density function (bottom).



For instance,

$$E(X) = \frac{1}{2}(a+b) = \frac{1}{2}(13+15) = 14$$

and

$$\text{Var}(X) = \frac{1}{12}(b-a)^2 = \frac{4}{12} = \frac{1}{3}$$

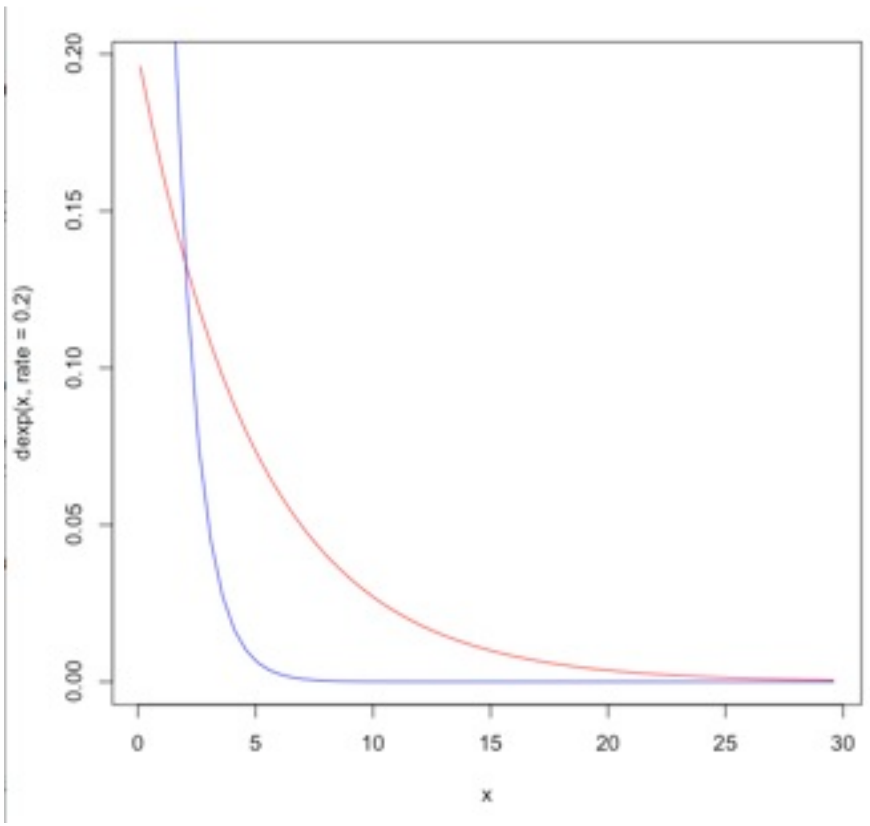
If X is an exponential random variable, then its density function is given by

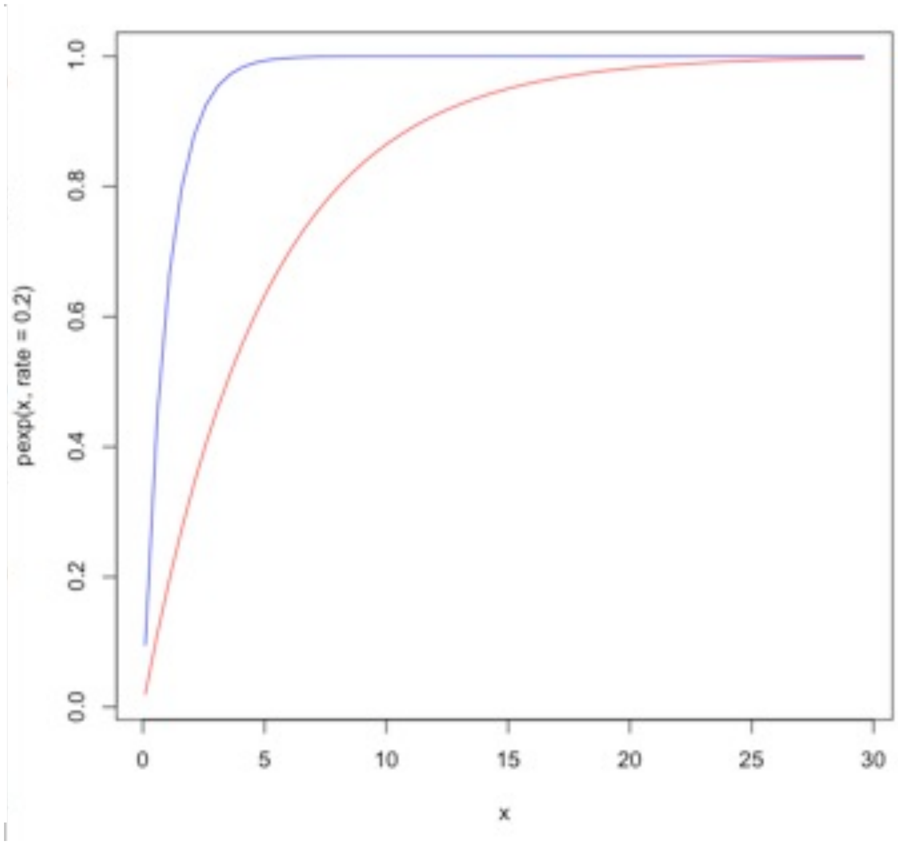
$$f_X(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

for $\lambda > 0$. Write $X \sim \text{Exp}(\lambda)$.

- The **exponential distribution** may be seen as the continuous counterpart of the geometric distribution: it is used to model the *time between independent events* happening at a constant *average rate*
- There are numerous applications of this distribution. Under the assumption of constant rate λ, it can be used to model
  - the time until a radioactive particle decays
  - the time until the next customer enters a queue
  - the time until the next recombination event
  - ...

```
> x=seq(0.1,30,by=0.5)
> plot(x,dexp(x,rate=0.2),type="l",col="red")
> points(x,dexp(x,rate=1),type="l",col="blue")
> plot(x,pexp(x,rate=0.2),type="l",col="red")
> points(x,pexp(x,rate=1),type="l",col="blue")
```





3.34 Link between Poisson and exponential distribution

There is a very close tie between the Poisson distribution and the exponential distribution

- Let X be a Poisson random variable counting the number of occurrences in the interval $[0, t]$
- Our interest is in the **time between two arrivals**, which obviously is also a random variable
- Let the arrival time be denoted by T. Its probability distribution function by definition is

$$F_T(t) = \begin{cases} P(T \leq t) = 1 - P(T > t) & t \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

- In terms of X, the event $T > t$ is equivalent to the event that there are no arrivals during the time interval $[0, t)$, which means $X = 0$.
- Hence, since

$$P(X = 0) = e^{-\lambda t}$$

we have

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

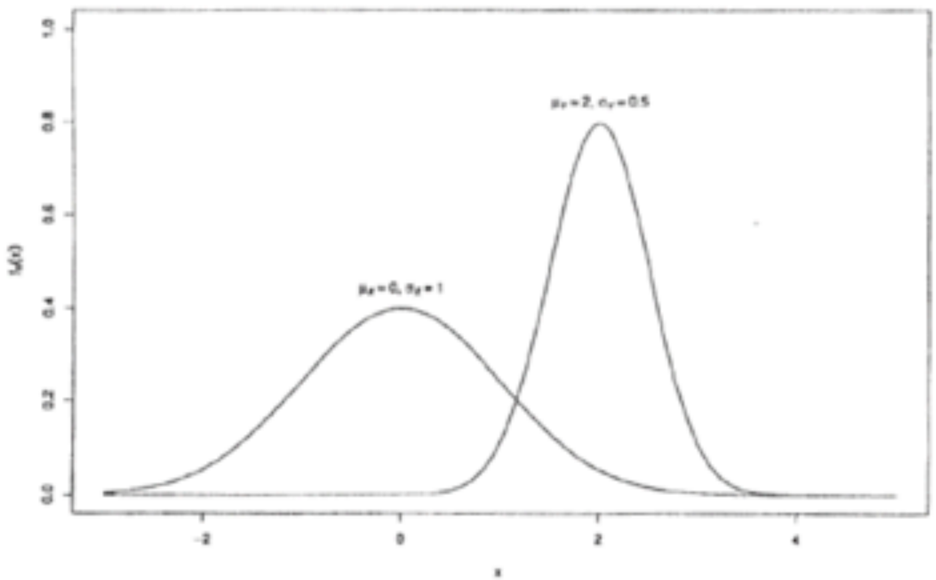
- Hence the interarrival times between Poisson events has an exponential distribution.
- The parameter λ in the distbution of T is the mean arrival rate associated with Poisson arrivals

3.35 Normal distribution

The probability density function of the **normal distribution**

$$f_X(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{on} \quad (-\infty, \infty)$$

which depends on a location parameter (mean) μ and scale parameter (standard deviation) σ. The distribution will be denoted by $N(\mu, \sigma^2)$



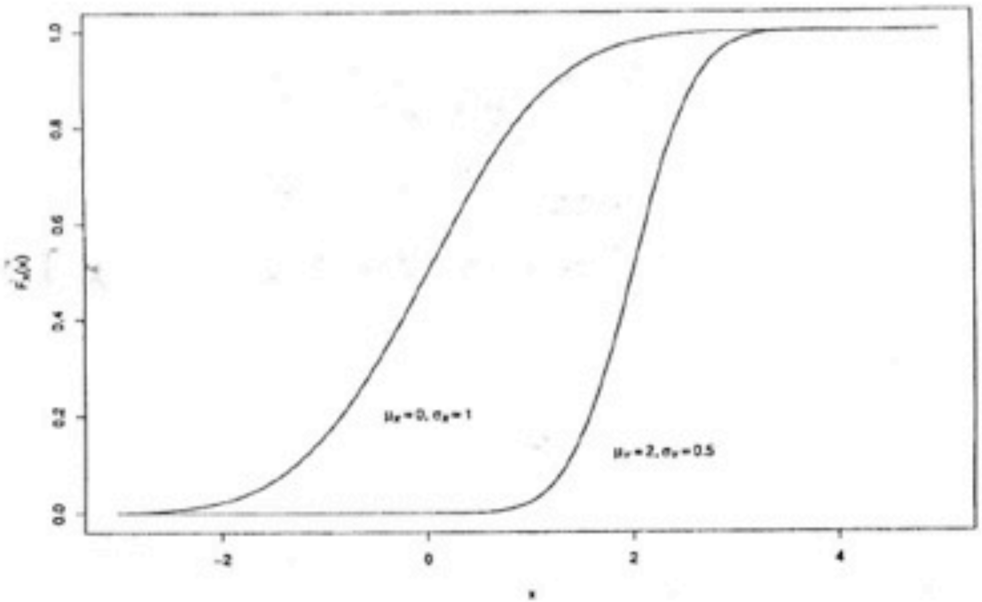
- Some basic facts:
  - The density is a symmetric, bell-shaped curve with points of inflection at $\mu - \sigma, \mu + \sigma$
  - The factor $1/\sqrt{2\pi}\sigma$ is the normalizing constant
  - The mean μ corresponds to the maximum value, $(1/\sqrt{2\pi}\sigma)$
  - The mean is also the *median* (i.e. the point x such that $F_X(x) = 1/2$ is μ) and the *mode* (the peak of the density)

3.36 Cumulutative function

Cumulative distribution function

The cumulative distribution function of X is

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(v-\mu)^2}{2\sigma^2}} dv$$



Problem: the integral above cannot be evaluated in a closed form and must be computed numerically. These calculations have been done and collected in tables for a *standard* normal distributiuon

The standard normal distribution

If $X \sim N(\mu, \sigma^2)$ and we take the transformation

$$Z = \frac{X - \mu}{\sigma}$$

then $Z \sim N(0,1)$, a normal distribution with $\mu = 0$ and $\sigma^2 = 1$. The distribution function of the standadized random variable Z is

$$F_Z(z) = \mathrm{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{v^2}{2}}\, dv = \Phi(z)$$

Note that $\Phi(z)$ denotes the distribution function of a standard normal distribution.

How do we compute probabilities for any normal distribution?

Suppose we have a normally distributed random variable $X \sim \mathrm{N}(\mu, \sigma^2)$. How do we solve probabilities queries about X?

(a) Re-write the problem in terms of $Z = \frac{X-\mu}{\sigma}$ so that $Z \sim \mathrm{N}(0,1)$

(b) Use the tabulated values of $\Phi(z) = \mathrm{P}(Z \leq z)$

Example 6.4. Mass spectra

- The peak intensity of a mass sprectrum is modeled as a normally distributed random variable X, say N$(2, 9)$ (in cm.).
- What is the probability of observing a peak between 1 and 4 cm.?
- First, let us rewrite the general problem in terms of Z, the standard normal

$$\mathrm{P}(a \leq X \leq b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right)$$
$$= \mathrm{P}\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

where $Z \sim$ N(0,1), for which we know $\Phi(z)$ from the tables.

- In our example,

$$P(1 \leq X \leq 4) = P\left(\frac{1-2}{3} \leq Z \leq \frac{4-2}{3}\right)$$
$$= P\left(Z \leq \frac{2}{3}\right) - P\left(Z \leq -\frac{1}{3}\right)$$
$$= \phi\left(\frac{2}{3}\right) - \left[1 - \phi\left(\frac{1}{3}\right)\right]$$
$$\approx 0.7486 - (1 - 0.6253) = 0.3775$$

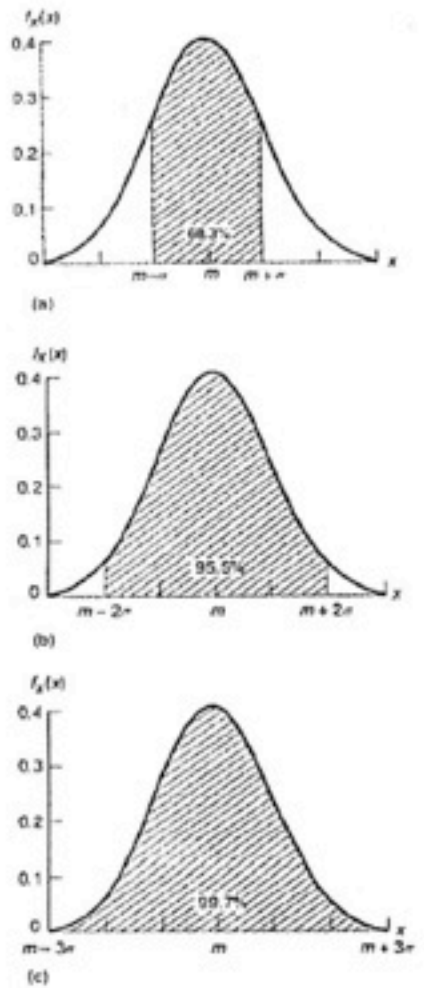
3.39 Computing probabilities under the normal curve

Note that

$$P(\mu - z\sigma < X \leq \mu + z\sigma) = P(-z < Z \leq z) = \Phi(z) - \Phi(-z)$$

is independent on μ and σ and it is only a functic z. The probability that X takes values within z standard deviations about its expected value depends only on z.

For instance, illustrated here are the probabilities that X is within $\mu \pm \sigma$ (a), $\mu \pm 2\sigma$ (b) and $\mu \pm 3\sigma$ (c)



Sampling distributions

and

main large scale sample theorems

In many statistical applications we will be confronted with the following:

- Suppose that $X_1, X_2, \ldots, X_n$ is a sequence of n **inpedendent and identically distributed** random variables. That is, $X_1, X_2, \ldots, X_n$ is a sample from some distribution, e.g.$X_i \sim f_X(x)$ where $f_X(x)$ can be either a mass function or a density function

- We want to find the distribution of a new random variable

$$Y = h(X_1, X_2, \ldots, X_n)$$

  for some function h, and for some n. For instance, we may need the distribution function of Y or just some moments of this distribution, such as mean and variance.

- The distribution of Y is called the **sampling distribution** because Y is derived as a transformation of a sample coming from some underlying distribution

Some of the methods we have studied so far could be used for solving this problem in situations when it is possible to compute an **exact solution**. In this chapter we also consider some known results when $f_X(x)$ is the density of a normally distributed random variable, and for various functions h.

When an exact solution is not available we have two options:

(a) Since Y is defined for each sample size n, we can consider a sequence of random variables $Y_1, Y_2, \ldots,$ and therefore study the **limiting distribution** of such a sequence so that, when n is large, we can approximate the distribution of Y_n by the limit

(b) When n is small or the problem is particularly difficult, we can set up Monte Carlo simulations

- In statistical applications, we typically do not know much about the *underlying distribution* of the X_i from which we are sampling – these are observed data
- We then collect a sample $x_1, \ldots, x_n$, from which we calculate a value of $y = h(x_1, \ldots, x_n)$, and use y to estimate a characteristic of the (generally unknown) underlying distribution of X (often we *assume* that the underlying distribution of the data is normal)
- For instance, we may collect a sample to estimate the mean or variance of the underlying distribution – these are functions of the observed random samples.
- Then we want to know what happens to these estimates when n (the sample size) grows.
- If we have chosen our **estimator** well, then the estimates will *converge* to the quantities we are estimating as n increases – *how do we define convergence?*
- When this happens, the estimate is called **consistent**

Suppose we obtain a sample X_1, X_2 of sample size $n = 2$ from the discrete distribution with probability mass function given by

$$p_X(x) = \begin{cases} 1/2 & x = 1 \\ 1/4 & x = 2 \\ 1/4 & x = 3 \\ \text{otherwise} & \end{cases}$$

What is the distribution of $Y_2 = (X_1 X_2)^{\frac{1}{2}}$ (the geometric mean)? Since $n = 2$, the sampling distribution of Y_2 is easily found by inspecting the following table.

Y	sample	$p_{Y_2}(y)$
1	$(1,1)$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
$\sqrt{2}$	$\{(1,2),(2,1)\}$	$\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{4}$
$\sqrt{3}$	$\{(1,3),(3,1)\}$	$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{4}$
2	$\{2,2\}$	$\frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$
$\sqrt{6}$	$\{2,3\}\{3,2\}$	$\frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{8}$
3	$\{3,3\}$	$1/4 \cdot 1/4 = 1/16$

If $n = 20$, what is the distribution of $Y_{20} = (X_1 X_2 \cdots X_{20})^{\frac{1}{20}}$?

There are now $3^{20} = 3,486,784,401$ possible samples.

Computing $p_{Y_{20}}$ directly would be hard, even with a computer.

Example 8.2. Geometric mean example continued

Two options are:

(a) Look at the distribution of $Y_n = (X_1 X_2 \cdots X_n)^{\frac{1}{n}}$ when n is large.

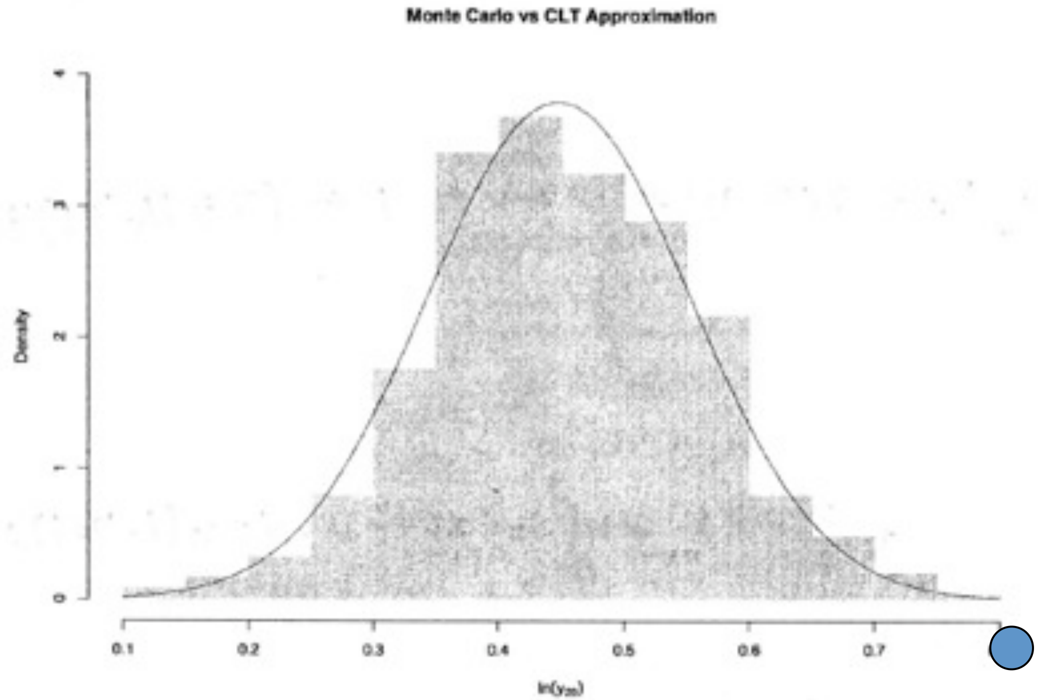
For instance, the Central Limit Theorem (CLT) will show that

$$\ln Y_n = \frac{1}{n} \sum_{i=1}^{n} \ln X_i$$

is well approximated by a normal distribution when n is large

(the $\ln$ transform is 1-1 and is used to avoid potentially large values here)

(b) Set up a Monte Carlo simulation



4.7 Convergence in probability

Notions of convergence are fundamentals in mathematics. However, if the values taken by a variable are random, then how can they *converge* to anything?

- We will consider the *probabilities* associated to the random variables, and check whether they converge, in some sense.
- Let $X_1, X_2, \ldots$ be an infinite sequence of random variables and let Y be another random variable

A sequence $\{X_n\}$ is said to **converge in probability** to Y if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} \mathsf{P}(| X_n - Y | \geq \epsilon) = 0$$

or, equivalently,

$$\lim_{n \to \infty} \mathsf{P}(| X_n - Y | < \epsilon) = 1$$

Convergence in probability is often denoted by adding the letter P over an arrow indicating convergence, i.e. $X_n \xrightarrow{P} Y$

4.8 Convergence in probability

Example 8.3. (Degenerate distributions) Suppose that

$$\mathrm{P}\left(X_n = 1 - \frac{1}{n}\right) = 1 \qquad \text{and} \qquad \mathrm{P}(Y = 1) = 1$$

then

$$\mathrm{P}(|X_n - Y| \leq \epsilon) = 0 \quad \text{whenever } n > \frac{1}{\epsilon}$$

Hence

$$\mathrm{P}(|X_n - Y| \leq \epsilon) \to 0 \text{ as } n \to \infty \text{ for all } \epsilon > 0$$

One of the most important applications of the convergence in probability is the weak law of large numbers.

- We are given a sequence $X_1, X_2, \ldots$ of independent random variables each one having the **same mean**, $E(X_i) = \mu$, and **finite variance** less than or equal v
- For large n, define the **sample average**

$$M_n = \frac{1}{n}(X_1 + ... + X_n)$$

- The weak low of large numbers provides a precise sense in which sample average values of M_n get closer to μ for large n

Suppose we have the sequence $X_1, X_2, \ldots$ of independent variables with $\mathrm{E}(X_i) = \mu$ and variance $v < \infty$. For any $\epsilon > 0$

$$\lim_{n \to \infty} \mathrm{P}(| M_n - \mu | \geq \epsilon) = 0$$

In other words, for ϵ an arbitrary positive quantity, the **weak law of large numbers** states that the probability that

$$|M_n - \mu| < 0$$

approaches 1 for a sufficiently large n, i.e. $M_n \xrightarrow{P} \mu$

The law of large numbers is a fundamental notion that describes how likely the average of a *randomly selected sample* from a population is to be close to the average of the whole population

4.9 Weak law of large numbers

4.10 Proof

We are assuming that $X_1, X_2, \ldots$ are independent, each one having same mean μ, and each one having variance less than or equal $v < \infty$. Remember that the sample average is,

$$M_n = \frac{1}{n}(X_1 + \ldots + X_n)$$

Using the linearity of the expected value, we see that

$$\mathsf{E}(M_n) = \frac{1}{n}\mathsf{E}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}(n\mu) = \mu$$

Using independence, we have

$$\begin{aligned} \mathsf{Var}(M_n) &= \frac{1}{n^2}\left(\mathsf{Var}(X_1) + \ldots + \mathsf{Var}(X_n)\right) \\ &\leq \frac{1}{n^2}(v + v + \ldots + v) \\ &= \frac{1}{n^2}(nv) = \frac{v}{n} \end{aligned}$$

Applying Chebychev's inequality

$$\mathsf{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\mathsf{Var}(M_n)}{\epsilon^2} \leq \frac{v}{\epsilon^2 n}$$

and note that this converges to 0 as $n \to \infty$.

Let $X_1, X_2, \ldots$ be i.i.d. with distribution $\mathsf{N}(3,5)$. Then

$$\mathsf{E}(M_n) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(X_i) = 3$$

and by the weak law of the large numbers,

$$\mathsf{P}(|M_n - 3| < \epsilon) \to 1$$

as $n \longrightarrow \infty$, or alternatively

$$\mathsf{P}(3 - \epsilon < M_n < 3 + \epsilon) \to 1$$

Therefore, for large n, the average value M_n will be very close to the known mean 3 with certainty.

Consider flipping a sequence of identical fair coins, where $X_i = 0$ (tail) and $X_i = 1$ (head)

4.12 Fair coin

Let M_n be the fraction of the first n coins that are heads.

$$M_n = (X_1 + X_2 + \cdots + X_n)/n$$

Hence by the weak law of lage numbers we have

$$\begin{aligned}
&\lim_{n\to\infty} \mathsf{P}(M_n < 0.49) \\
&\quad = \lim_{n\to\infty} \mathsf{P}(M_n - 0.5 < -0.01) \\
&\quad \leq \lim_{n\to\infty} \mathsf{P}(M_n - 0.5 < -0.01 \text{ or } M_n - 0.5 > -0.01) \\
&\quad = \lim_{n\to\infty} \mathsf{P}(|M_n - 0.5| > 0.01) = 0
\end{aligned}$$

And similarly

$$\lim_{n\to\infty} \mathsf{P}(M_n > 0.51) = 0$$

Then, for large n, it is very luckely that M_n is very close to 0.5

4.13 Almost sure convergence

A notion of convergence of random variables that is very similar to the notion of convergence of a sequence of real numbers if provided by the concept of **almost sure convergence** or convergence with probability 1

Let $X_1, X_2, \ldots,$ be an infinite sequence of random variables. The sequence $\{X_i\}$ is said to **converge almost surely** (or with probability 1) to a random variable Y if,

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = Y\right) = 1$$

and we write $X_n \xrightarrow{a.s.} Y$

- It is similar to *point-wise convergence* of a sequence of functions, but convergence need not occur on a set of events with probability 0, hence its name *almost sure*
- Alternatively, we say that X_n converges to Y **almost everywhere** or **strongly** towards Y - these are all identical definitions.
- It is a stronger type convergence than the convergence in probability: if $\{X_i\} \to Y$ almost surely, then $\{X_i\} \to Y$ in probability. The converse is not true.

4.14 Example

Let $U \sim \text{Uniform}[0,1]$. As in the previous example, define X_n by

$$X_n = \begin{cases} 3 & U \leq \frac{2}{3} - \frac{1}{n} \\ 8 & \text{otherwise} \end{cases}$$

and define Y by

$$Y = \begin{cases} 3 & U \leq \frac{2}{3} \\ 8 & \text{otherwise} \end{cases}$$

Note that

- If $U > 2/3$ then $Y = 8$ and also $X_n = 8$ for all n, so clearly $X_n \to Y$
- If $U < 2/3$ then $Y = 3$ and for large enough n we will also have

$$U \leq \frac{2}{3} - \frac{1}{n}$$

  so again $X_n \to Y$.
- If $U = 2/3$ then we will always have $X_n = 8$, even though $Y = 3$

Hence, $X_n \to Y$ exept when $U = 2/3$. Because

$$\mathsf{P}(U = 2/3) = 0$$

we have that $X_n \to Y$ with probability 1.

4.15 Strong law of large numbers

We can now state a stronger result than the weak law of large numbers. It is stronger because it concludes almost surely convergence instead of just convergence in probability.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each one having finite mean μ. Then

$$\mathsf{P}\left(\lim_{n\to\infty} M_n = \mu\right) = 1$$

That is, the average converges with probability 1 to the common mean μ, or $M_n \overset{a.s.}{\to} \mu$

- This large sample result says that sample averages converge with probability 1 to the common mean μ
- Analogously to the previous result, it says that for large n the averages M_n are arbitrarily close to $\mu = \mathsf{E}(X_i)$
- In addition, it says that if n is large enough, then the averages will **all** be close to μ, for all sufficienylu large n
- In a statistical sense, the *sample* mean is consistent for the mean μ (more on this later).

4.16 Convergence in distribution

There is another notion of convergence of a sequence of random variables that is important in applications of probability and statistics.

A sequence of random variables $X_1, X_2, \ldots, X_n$ converges in distribution to a random variable X if, for all $x \in \Re$ such that

$$\mathrm{P}(X = x) = 0$$

we have

$$\lim_{n\to\infty} \mathrm{P}(X_n \leq x) = \mathrm{P}(X \leq x) = F_X(x)$$

and we write $X_n \xrightarrow{D} X$

- Intuitively, $\{X_n\}$ converges in distribution to X if, for large n, the distribution of X_n is close to that of X
- The importance of this result is that it allows us to work with the distribution of X, which may be much easier to work with than the distribution of X_n. We can approximate the distribution of X_n by that of X
- On the other hand, the fact that X_n converges in distribution to X says nothing about the underlying relationship between X_n and X, it only says something about their distribution.
- Note that convergence in probability implies convergence in distribution, but the converse is not always true

Central Limit Theorem

The Central Limit Theorem (CLT) is one of the most important results in probability theory. Intuitively, it says that a large sum of i.i.d. random variables, properly normalized, will always have approximately a *normal* distribution.

Suppose we are given a collection $X_1, X_2, \ldots$ of *i.i.d.* random variables each having *finite* mean μ and variance σ^2. Let

$$S_n = X_1 + \ldots + X_n$$

be the sample sum, and

$$M_n = S_n/n$$

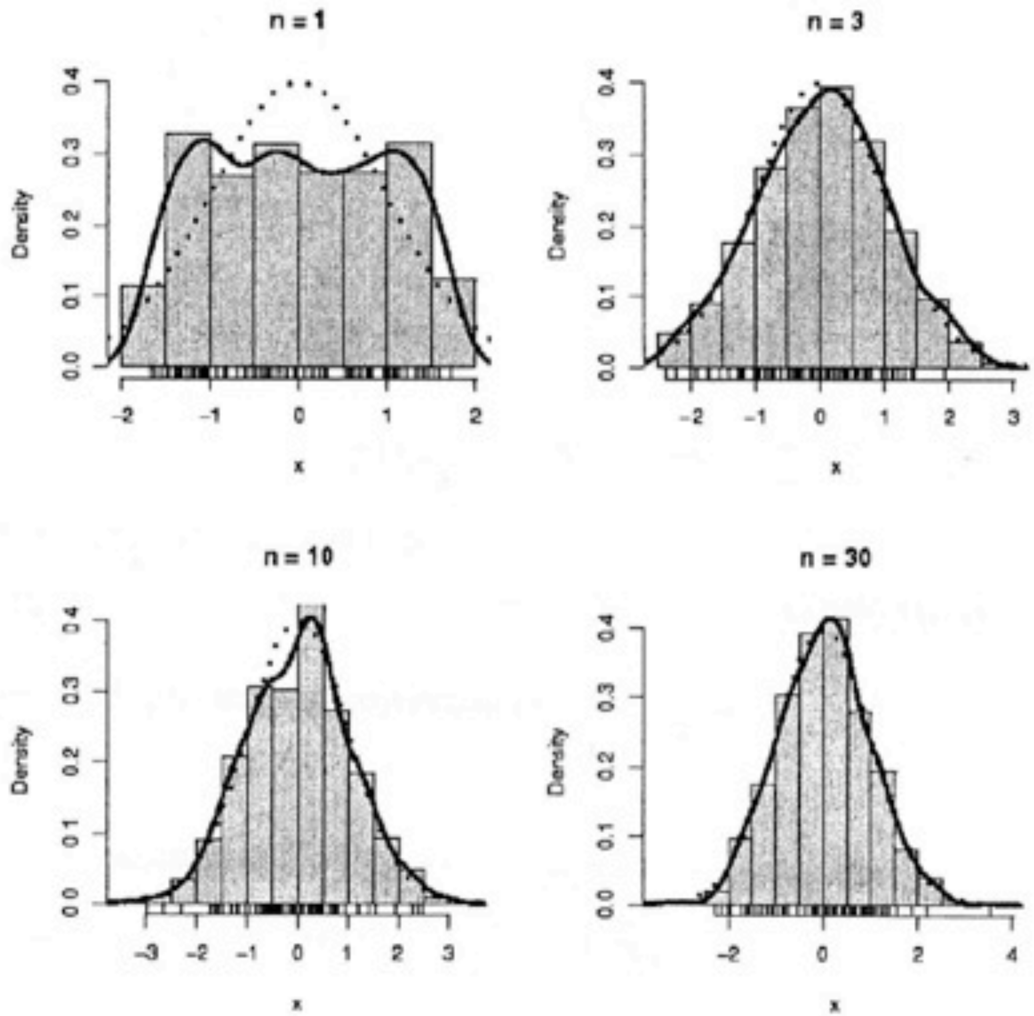
the sample mean. Let the standardized sample mean be

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{M_n - \mu}{\sigma/\sqrt{n}}$$

The CLT states that, as $n \to \infty$, the sequence $\{Z_n\}$ converges in distribution to Z, i.e. $Z_n \xrightarrow{D} Z$ N(0,1)

- Note that the result holds regardless of the form of the distribution of the individuals X_i–the only requirement is that each X_i has finite mean and variance.

In this example, $X_1, X_2, \ldots, X_n$ are i.i.d. from a uniform distribution. We look at the distribution of Z_n as n increases.



```
runif2x<-function(n, min1, max1, min2, max2) replicate(n,
ifelse(runif(1)>0.5, runif(1,min1,max1),
runif(1,min2,max2)))

x1 <- runif2x(1000,0,1,2,3)for (i in 1:100) {x1 <- x1 +
runif2x(1000,0,1,2,3)}
hist(x1)
```

OR

```
x1=c(runif(1000,0,1),runif(1000,2,3))
hist(x1,breaks=100,col='red')
x2=0
for(i in 1:1000){x2[i]=mean(sample(x1, 1000, replace = TRUE, prob = NULL))}
hist(x2,breaks=100,col='red')
```

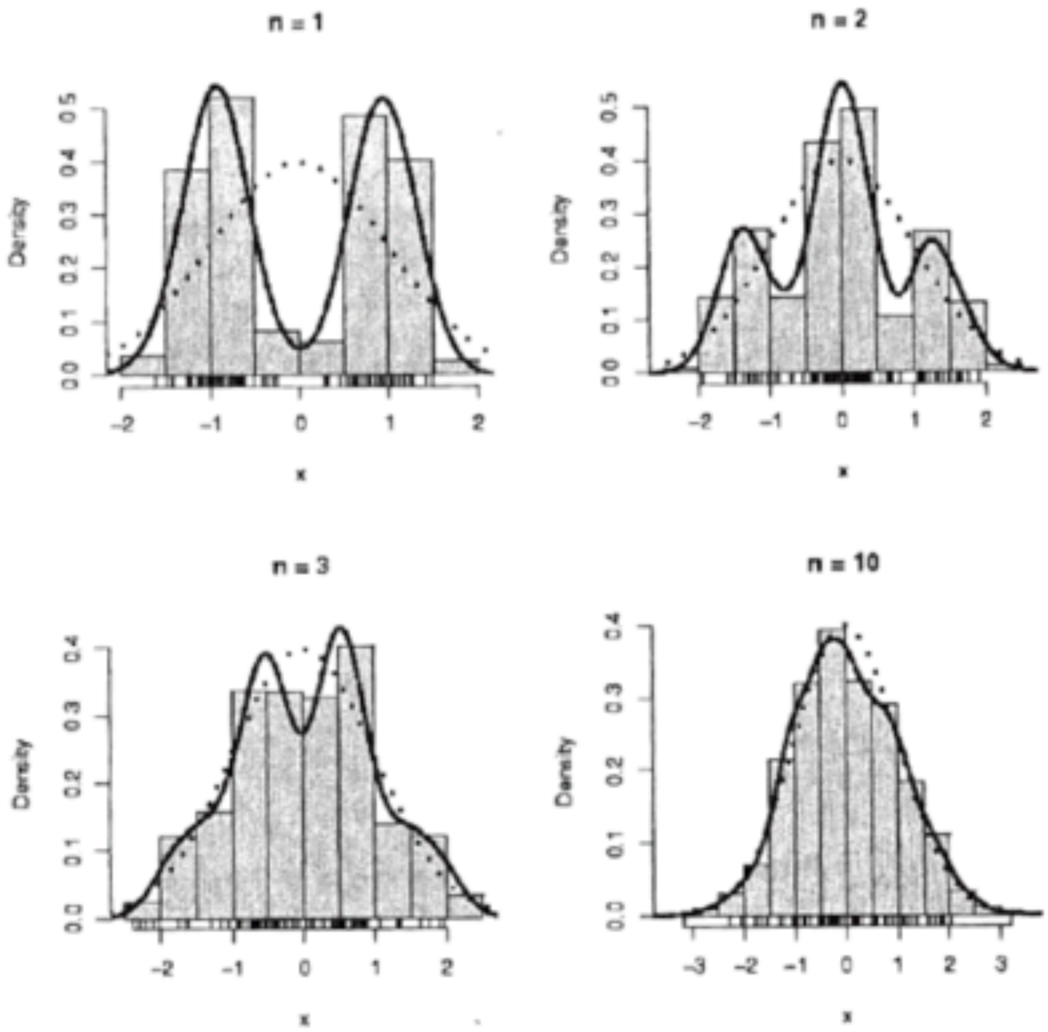
Not completely correct. Why?

OR

```
x1 <- sample(c(runif(1000,0,1),runif(1000,2,3)))
for (i in 1:10) {x1 <- x1 + sample(c(runif(1000,0,1),runif(1000,2,3)))}
hist(x1,col="red",breaks=20)
```

4.19 Another illustration

In this example, $X_1, X_2, \ldots, X_n$ are i.i.d. from a bimodal distribution.



Monday, 18 November 13

Alternative ways to express the CLT

The CLT ican be restated in alternative ways. For each fixed $x \in \Re$, then

(a) $\lim_{n\to\infty} \mathsf{P}(Z_n \le x) = \Phi(x)$

(b) $\lim_{n\to\infty} \mathsf{P}(S_n \le n\mu + x\sqrt{n}\sigma) = \Phi(x)$

(c) $\lim_{n\to\infty} \mathsf{P}(M_n \le \mu + x\sigma\sqrt{n}) = \Phi(x)$

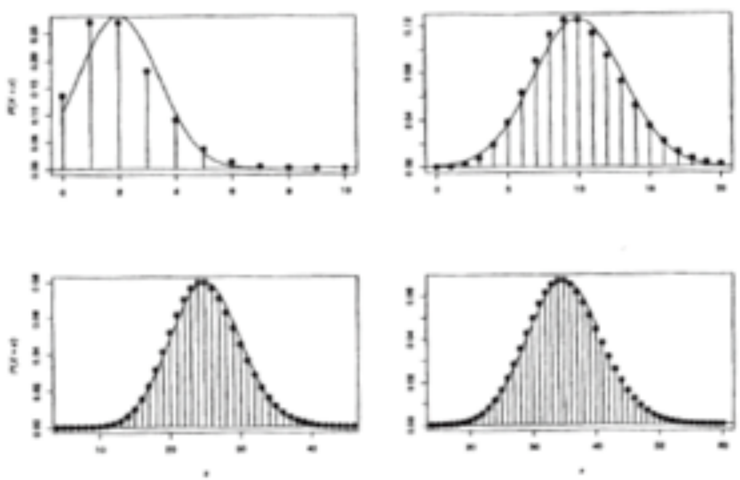
where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution.

Suppose $X_1, X_2, \ldots$ are i.i.d. random variables each with the Poisson(5) distribution. Recall that this implies

$$\mu = \mathsf{E}(X_i) = 5 \qquad \sigma^2 = \mathsf{Var}(X_i) = 5$$

Hence, for a fixed x, we have

$$\mathsf{P}(S_n \le 5n + x\sqrt{5n}) \to \Phi(x) \qquad \text{as } n \to \infty$$



4.21 Normal approximation to the Binomial

Suppose $X_1, X_2, \ldots$ are i.i,d. random variables each with a Bernoulli(θ) distribution

Recall that this implies E$(X_i) = \theta$ and $v =$ Var$(X_i) = \theta(1-\theta)$

Hence, for each fixed x,

$$\mathrm{P}(S_n \le n\theta + x\sqrt{n\theta(1-\theta)}) \to \Phi(x) \qquad \text{as } n \to \infty$$

We also know that

$$Y_n = S_n = X_1 + X_2 + \ldots + X_n \sim \text{Binomial}(n, \theta)$$

So the previous statement implies that

$$\mathrm{P}(Y_n \le y) = \mathrm{P}\left(\frac{Y_n - n\theta}{\sqrt{\theta(1-\theta)}} \le \frac{y - n\theta}{\sqrt{\theta(1-\theta)}}\right) \approx \Phi\left(\frac{y - n\theta}{\sqrt{\theta(1-\theta)}}\right)$$

for large n.

Note how, again, we are approximating a discrete distribution by a continuous distribution. A small improvement is often made to the previous approximation when y is a non-negative integer. We use

$$\mathrm{P}(Y_n \le y) \approx \Phi\left(\frac{y + 0.5 - n\theta}{\sqrt{\theta(1-\theta)}}\right)$$

Adding 0.5 to y is called the **correction for continuity**. In effect, this allocates all the relevant normal probability in the interval

$$(y - 0.5, y + 0.5)$$

to the non-negative integer y, which generally improves the approximation.

Normal distribution theory

Suppose that $X_1 \sim \mathsf{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$, and are independent.

Two important poperties of the normal distribution are

$$cX_1 + d \sim \mathsf{N}(c\mu_1 + d, c^2\sigma^2)$$

$$X_1 + X_2 \sim \mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Combining these results, we can

see that linear combinations of independent normal random variables are also normal

Suppose $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ and that they are independent random variables. Let

$$Y = (\sum_i a_i X_i) + b$$

for some constants $\{a_i\}$ and b. Then

$$Y \sim \mathsf{N}\left((\sum_i a_i \mu_i) + b, \sum_i a_i^2 \sigma_i^2 \right)$$

Sampling distribution of the sample mean

The result above immediately implies that

Suppose $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ and that they are independent random variables. If

$$\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$$

then

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Another property of linear combination of normal random variables

Two linear combinations of the same collection of independent normal random variables are independent if and only if they covariance equals 0

Suppose $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ and that they are independent random variables. Let

$$U = \sum_i a_i X_i$$

and

$$V = \sum_i b_i X_i$$

for some constants $\{a_i\}$ and $\{b_i\}$. Then

$$\mathrm{Cov}(U, V) = \sum_i a_i b_i \sigma_i^2$$

Furthermore, $\mathrm{Cov}(U, V) = 0$ if and only if U and V are independent.

Note how, for normal distributions, if $\mathrm{Cov}(U, V) = 0$ then U and V are independent. It is important to remember that that this is not generally true for all random variables.

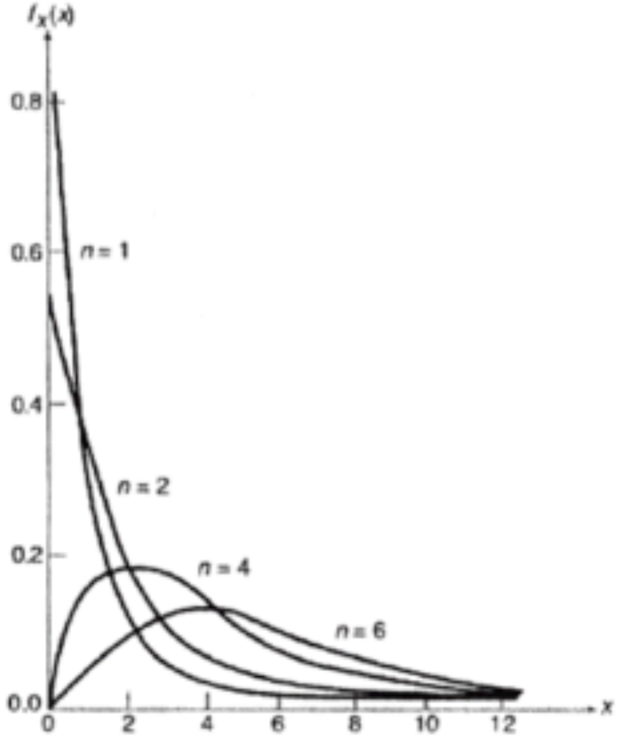
$$\mathrm{Cov}(X, Y) = \mathrm{E}\big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\big],$$

4.25 Chi-squared distribution revisited

The χ^2 distribution with n **degrees of freedom** or χ^2_n is the distribution of the random variable

$$Z = X_1^2 + X_2^2 + \cdots + X_n^2$$

where $X_1, \ldots, X_n$ are i.i.d., each with standard normal distribution $N(0, 1)$



Observe that the χ^2 distributions are **asymmetric** and **skewed** to the right. As the degrees of freedom increase, the central mass of probability moves to the right.

As another special case, note that $n = 2$ given an Exp(2) distribution

3.40 Chi-squared distribution

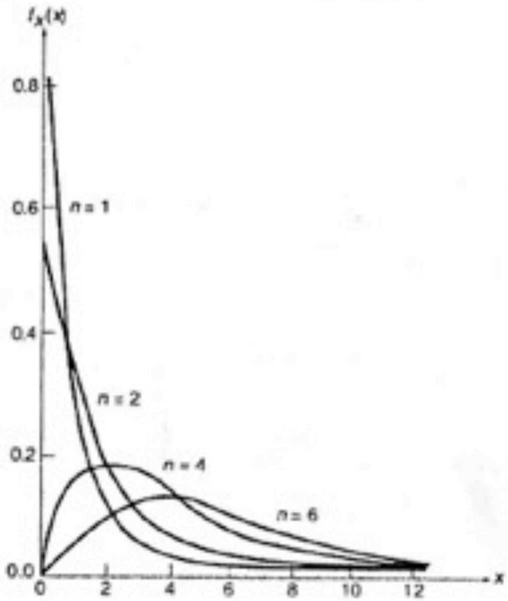
Another special case of the gamma distribution is given by the χ^2 distribution, which is very important in statistical inference and hypothesis testing.

The density function of a random variable X with chi-squared distribution is

$$f_X(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} & x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

which is obtained as special case of the gamma with $\lambda = 1/2$ and $\eta = n/2$

- The n parameter is usually referred to as the **degrees of freedom**
- The utility of this distribution arises from the fact that a sum of the squares of n independent standardized normal random variables has a χ^2 distribution with n degrees of freedom.
- The density of a χ^2 distribution for $n = 1, 2, 4, 6$



4.26 Sampling distribution of the sampling variance

The expected value of this distribution can be easily computed

If $Z \sim \chi^2_n$ then $\mathsf{E}(Z) = n$

To see this result

- Write
$$Z = X_1^2 + \cdots + X_n^2$$
where $\{X_i\}$ are i.i.d. $\mathsf{N}(0,1)$.
- Then $\mathsf{E}((X_i)^2) = 1$.
- The result follows by linearity.

The following result is important in statistical inference

Let $X_1, \ldots, X_n$ be i.i.d. $\mathsf{N}(\mu, \sigma^2)$, and put

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

and furthermore S^2 and $\bar{X}$ are independent.

It follows that

$$\mathsf{E}\left(\frac{(n-1)S^2}{\sigma^2}\right) = n - 1$$

and therefore

$$\mathsf{E}(S^2) = \sigma^2$$

The t distribution with n degrees of freedom or t_n arises as the distribution of a standard normal divided by the square-root of $1/n$ times an independent chi-squared random variable with n degrees of freedom

The t distribution with n degrees of freedom or t_n is the distribution of the random variable

$$Z = \frac{X}{\sqrt{(X_1^2 + X_2^2 + \ldots + X_n^2)/n}}$$

where $X, X_1, \ldots, X_n$ are i.i.d. with standard normal distribution. Equivalently,

$$Z = X/\sqrt{Y/n}$$

where $Y \sim \chi_n^2$

The F distribution with m and n degrees of freedom arises as the distribution of m/n times a chi-squared distribution with m degrees of freedom divided by an independent chi-squared distribution with n degrees of freedom.

The F distribution with m and n degrees of freedom or $F_{m,n}$ is the distibution of the random variable

$$Z = \frac{(X_1^2 + X_2^2 + \ldots + X_m^2)/m}{(Y_1^2 + Y_2^2 + \ldots + Y_n^2)/n}$$

where $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are i.i.d. each with the standard normal distribution. Equivalently,

$$Z = \frac{X/m}{Y/n}$$

where $X \sim \chi_m^2$ and $Y \sim \chi_n^2$