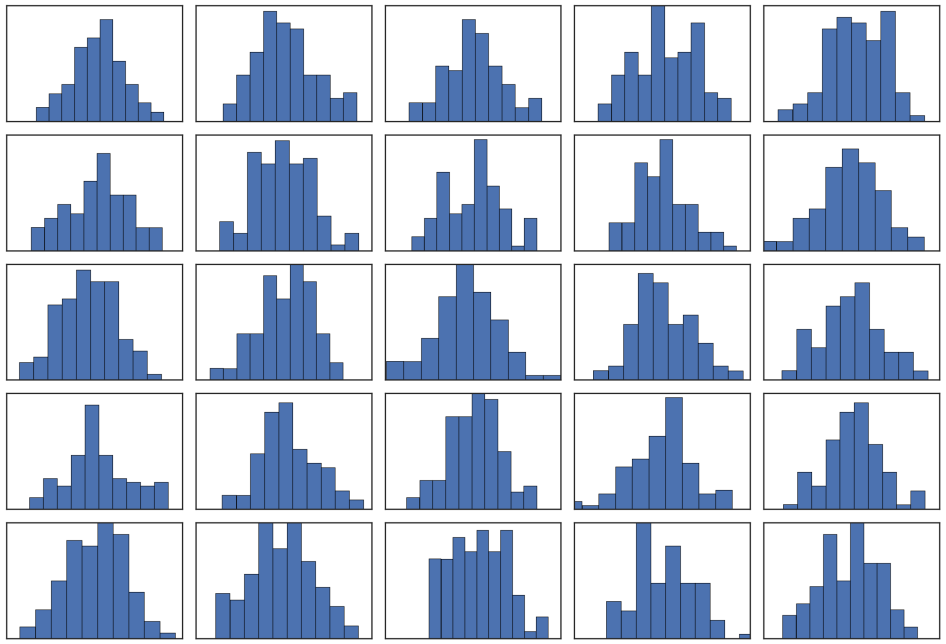


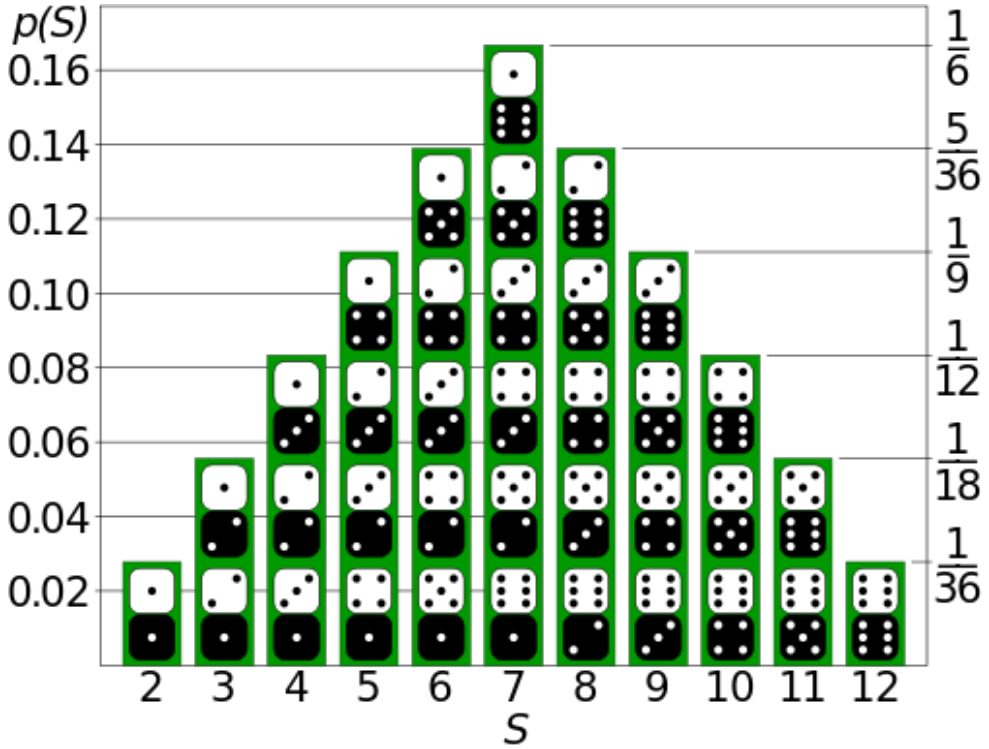
Lecture 2
Probability Distributions

Theophanis Tsandilas

What is a probability distribution?

Consider the population of all possible outcomes when throwing two dice.

How probable is each sum *S* of counts from the two dice?



The probability distribution provides the probability of occurrence of all possible outcomes in an experiment

Probability distribution of a population

It is generally **not known**. However, we may have sufficient information about this distribution to meet the goals of our research.

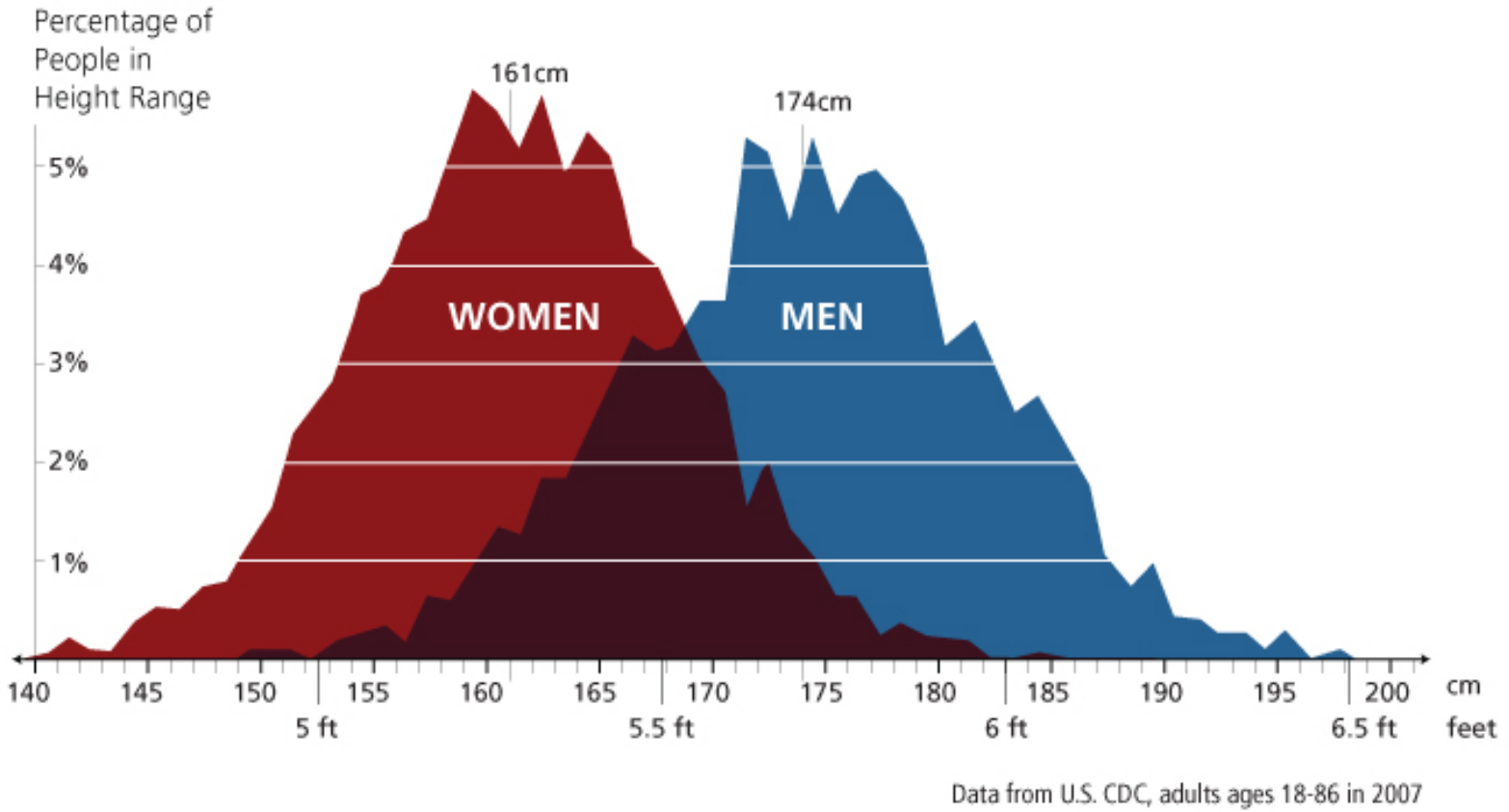
Statistical modeling: It is possible to rely on a small set of probability distributions that capture the key characteristics of a population.

We can then **infer** the parameters of these model probability distributions from our sample.

> Why? We expect that the sample contains some information about the population from which it was sampled.

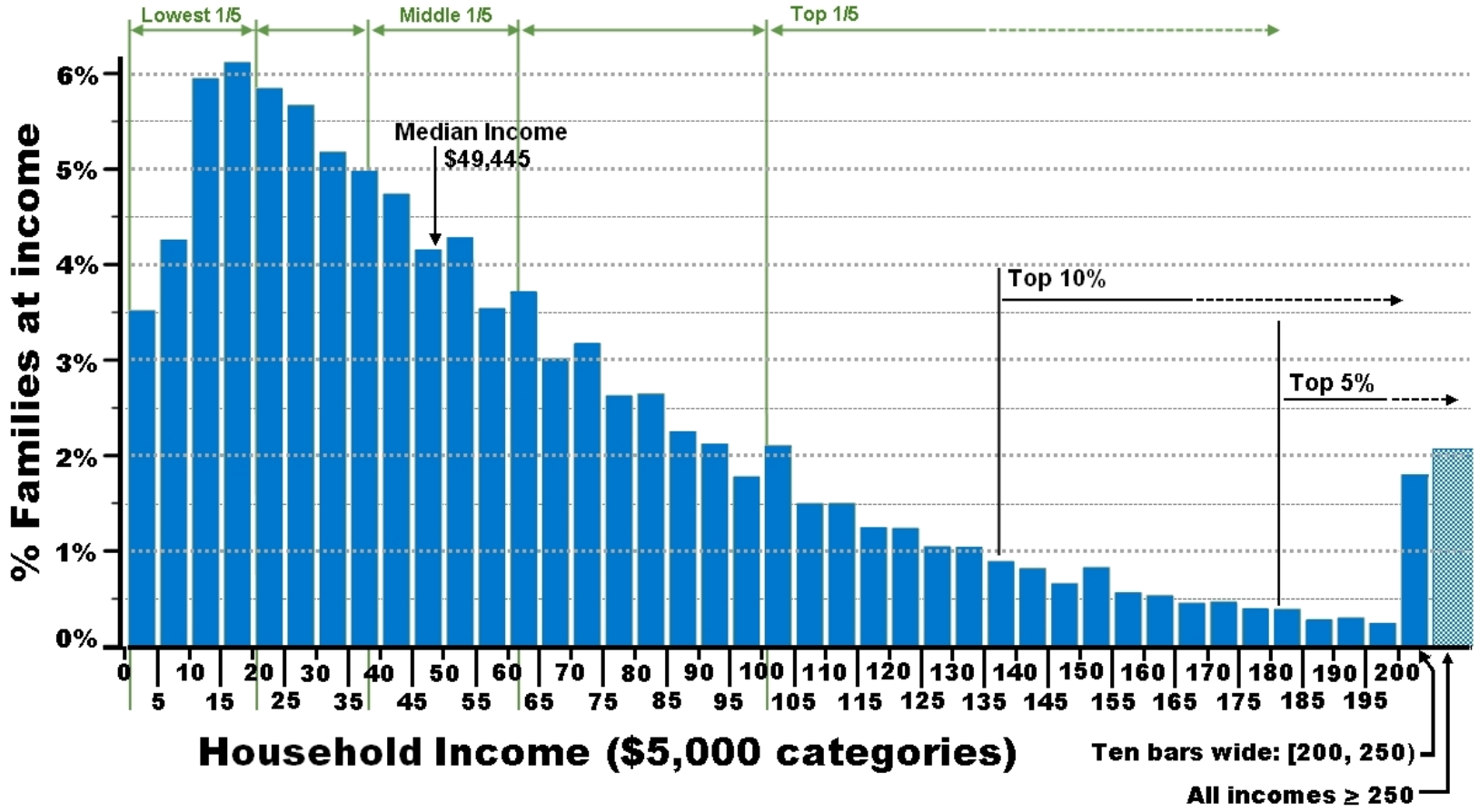
Example distributions

Height: Women vs. Men



Example distributions

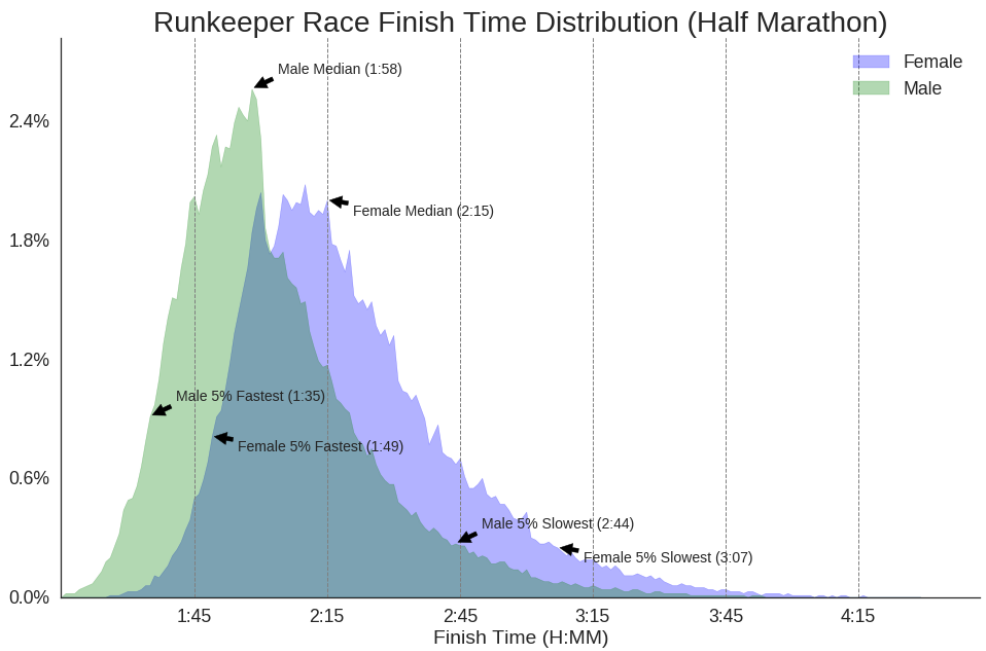
US Income distribution (older than 2013)

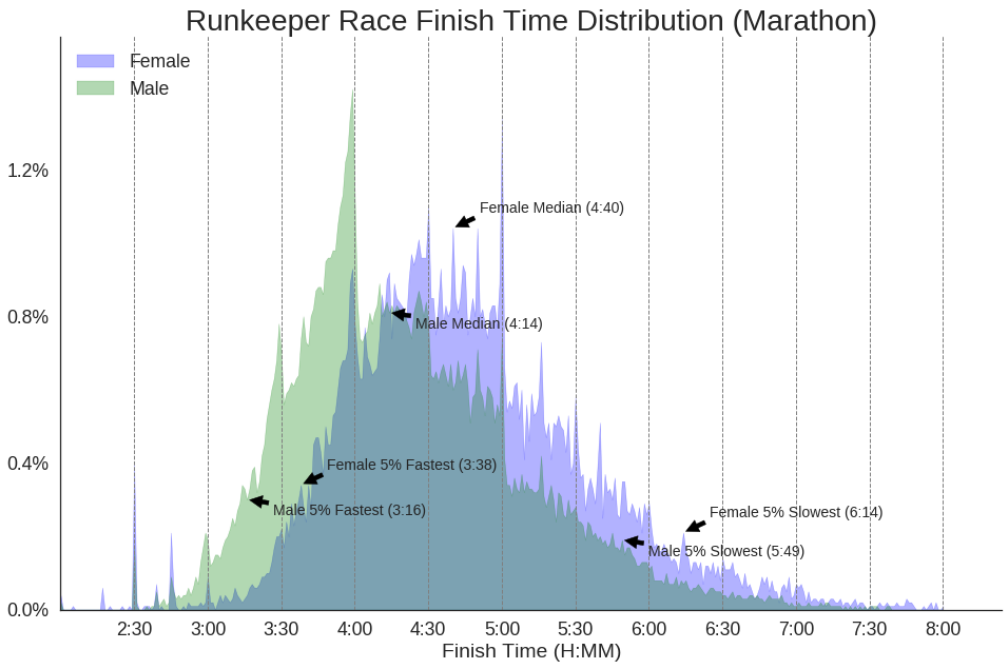


Data source: http://www.census.gov/hhes/www/cpstables/032011/hhinc/new06_000.htm

Example distributions

Half Marathon and Marathon race finish time

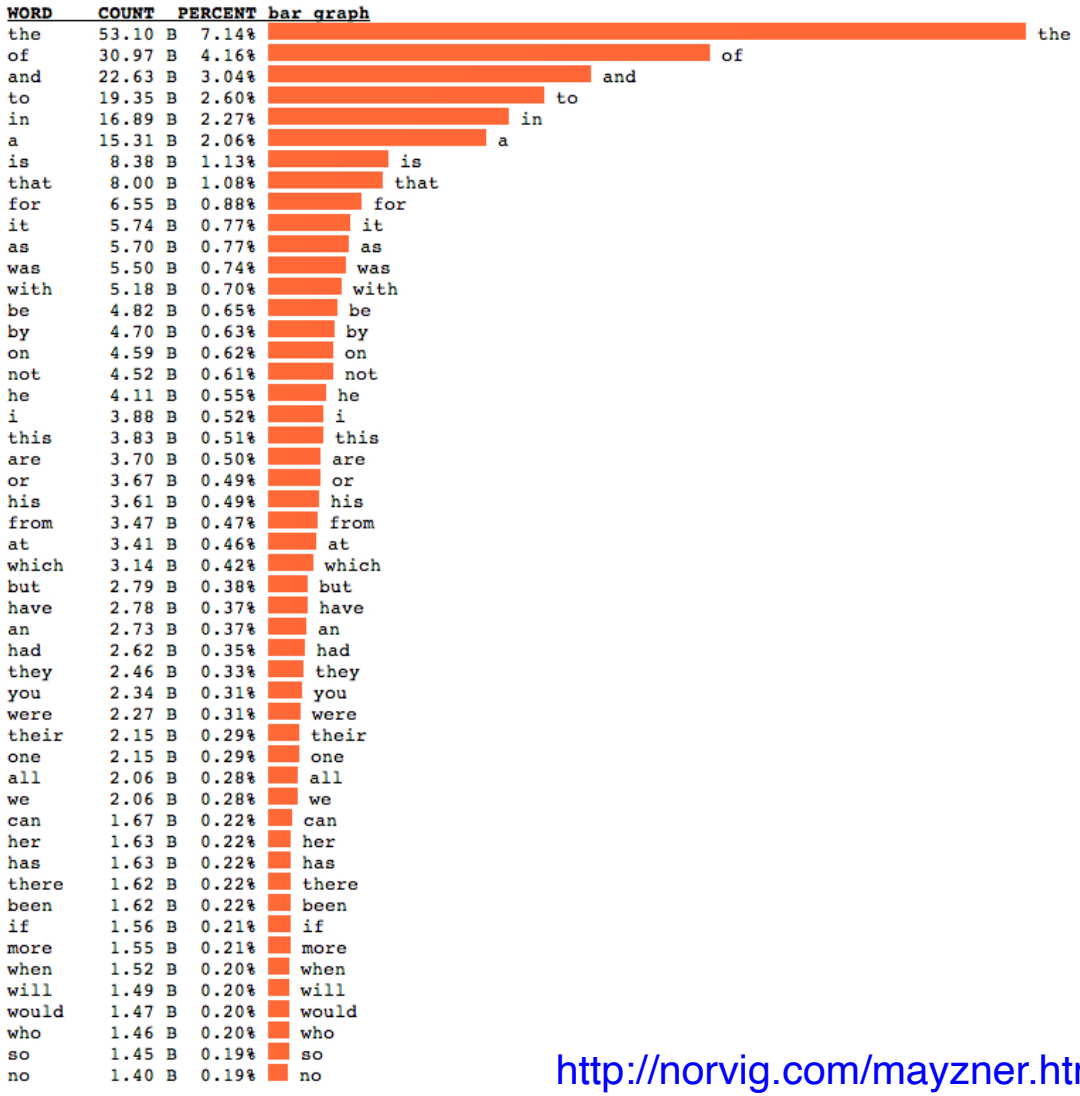




Source: https://medium.com/runkeeper-everyone-every-run/how-long-till-the-finish-line-494361cc901b

Example distributions

Distribution of most frequent English words



WORD	COUNT	PERCENT	bar graph
the	53.10 B	7.14%	the
of	30.97 B	4.16%	of
and	22.63 B	3.04%	and
to	19.35 B	2.60%	to
in	16.89 B	2.27%	in
a	15.31 B	2.06%	a
is	8.38 B	1.13%	is
that	8.00 B	1.08%	that
for	6.55 B	0.88%	for
it	5.74 B	0.77%	it
as	5.70 B	0.77%	as
was	5.50 B	0.74%	was
with	5.18 B	0.70%	with
be	4.82 B	0.65%	be
by	4.70 B	0.63%	by
on	4.59 B	0.62%	on
not	4.52 B	0.61%	not
he	4.11 B	0.55%	he
i	3.88 B	0.52%	i
this	3.83 B	0.51%	this
are	3.70 B	0.50%	are
or	3.67 B	0.49%	or
his	3.61 B	0.49%	his
from	3.47 B	0.47%	from
at	3.41 B	0.46%	at
which	3.14 B	0.42%	which
but	2.79 B	0.38%	but
have	2.78 B	0.37%	have
an	2.73 B	0.37%	an
had	2.62 B	0.35%	had
they	2.46 B	0.33%	they
you	2.34 B	0.31%	you
were	2.27 B	0.31%	were
their	2.15 B	0.29%	their
one	2.15 B	0.29%	one
all	2.06 B	0.28%	all
we	2.06 B	0.28%	we
can	1.67 B	0.22%	can
her	1.63 B	0.22%	her
has	1.63 B	0.22%	has
there	1.62 B	0.22%	there
been	1.62 B	0.22%	been
if	1.56 B	0.21%	if
more	1.55 B	0.21%	more
when	1.52 B	0.20%	when
will	1.49 B	0.20%	will
would	1.47 B	0.20%	would
who	1.46 B	0.20%	who
so	1.45 B	0.19%	so
no	1.40 B	0.19%	no

http://norvig.com/mayzner.html

Discrete probability distributions

The population (and hence its samples) contain discrete values, **either finite or infinite in number**

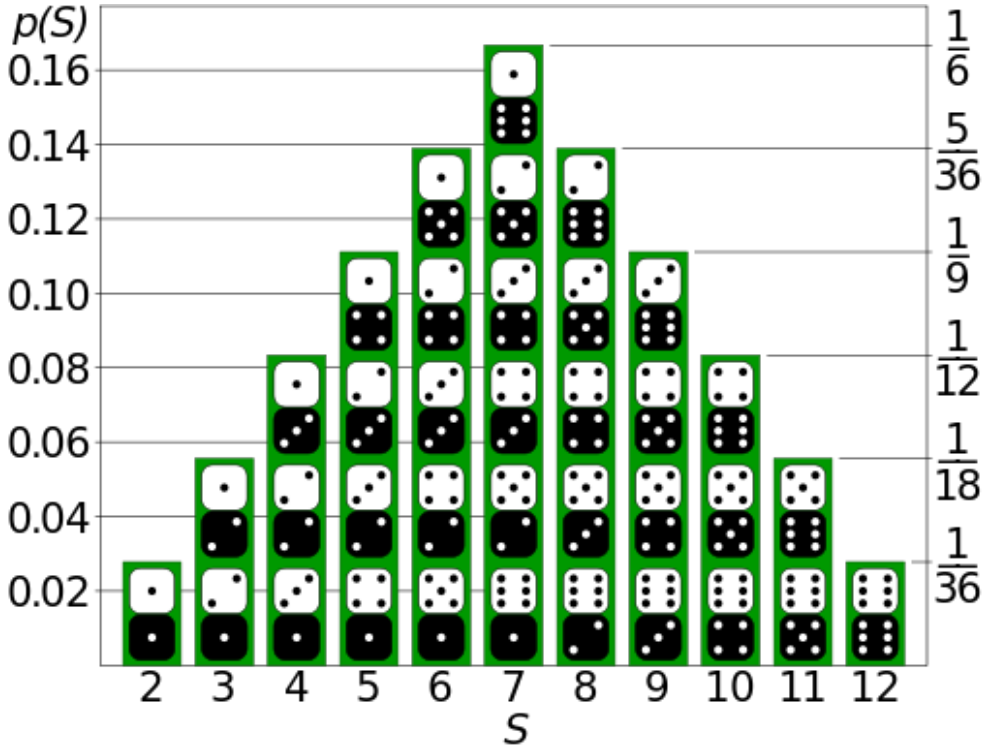
e.g., {-3, 1, 0, 1, 2}, {‘blue’, ‘brown’, ‘green’}, or {1, 2, 3, ...}

The probability of a value *x* in a population can be expressed as a function: $f(x) = p(X = x)$
(the probability of random variable *X* taking value *x*)

The function *f(x)* is called a **probability mass function (pmf)**

Discrete probability distributions

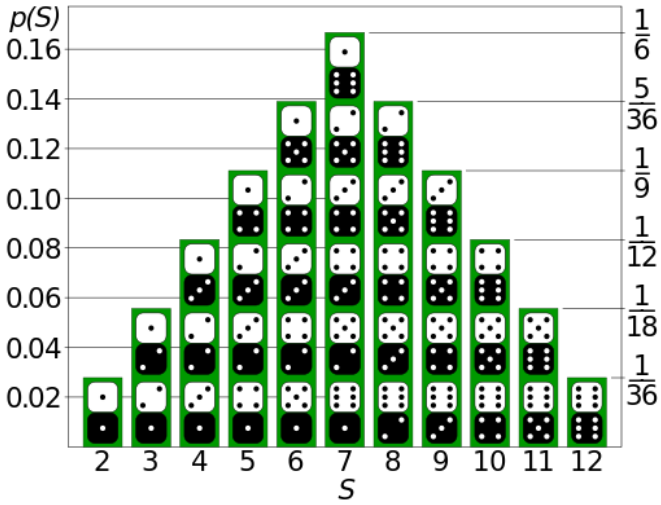
This is a discrete probability distribution



Discrete probability distributions

Probabilities should sum to 1:

$$\sum_{x \in S} f(x) = 1$$



The **expected value (or mean)**:

$$E(x) = \sum_{x \in S} f(x)x$$

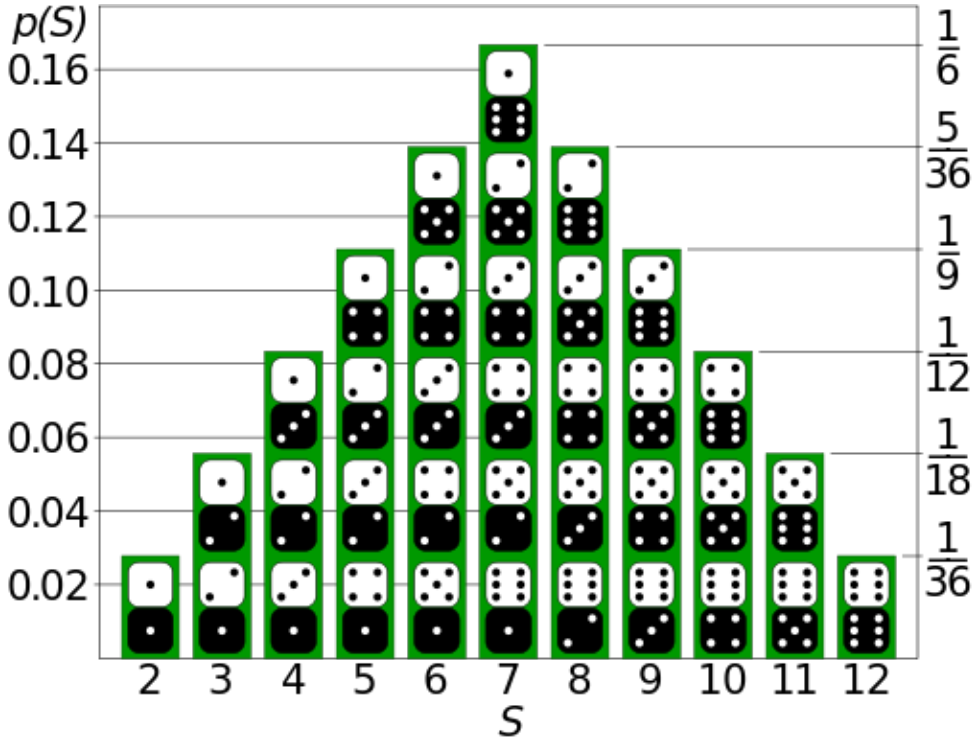
$$= 2\frac{1}{36} + 3\frac{1}{18} + 4\frac{1}{12} + 5\frac{1}{9} + 6\frac{5}{36} + 7\frac{1}{6} + 8\frac{5}{36} + 9 * \frac{1}{9} + 10\frac{1}{12} + 11\frac{1}{18} + 12\frac{1}{36} = 7$$

The **mode** is the most frequent value: which one?

The **median** is the middle value: which one?

Symmetrical probability distributions

When the mean coincides with the median



The above is a symmetrical, unimode distribution

The binomial distribution

Consider a population containing two values: *1* and *0*

A single sample ($n = 1$) from such a population is known as a **Bernoulli trial.**

> A coin flipping trial with a fair coin is a Bernoulli trial with $Pr(Head) = .5$ and $Pr(Tail) = .5$

If we perform n independent Bernoulli trials, the number of successful outcomes (or failures) will follow a **binomial distribution.**

The binomial distribution

If a random variable *X* follows a binomial distribution, we write: $X \sim B(n, P)$

where *X* is the number of successes given:

n is the number of Bernoulli trials

P is the chance (probability) of a success

If we know the parameters *n* and *P*, we can fully describe the distribution

The binomial distribution

Probability Mass Function (pmf)

$$f(x; n, P) = \binom{n}{x} P^x (1 - P)^{n-x}$$

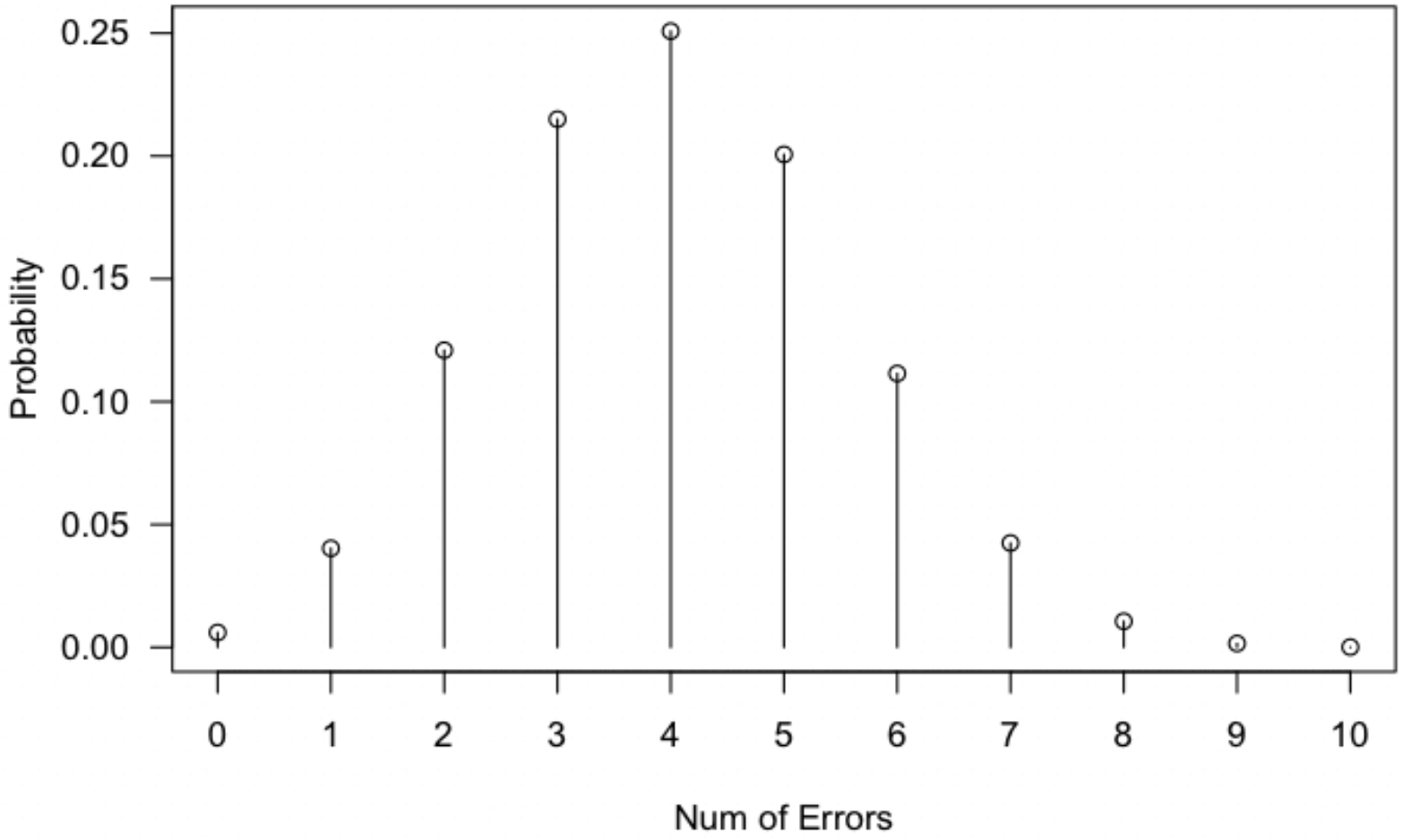
The probability of exactly n - *x* failures

The probability of exactly *x* successes

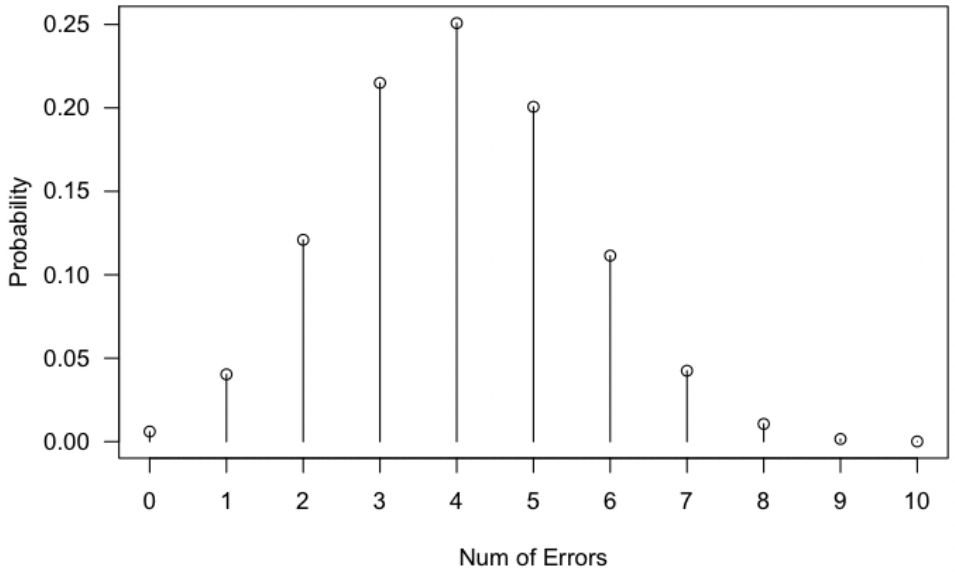
The number of possible ways
that *n* Bernoulli trials lead to *x* successes

The binomial distribution

The distribution of errors for 10 trials, when for each trial, errors occur with a probability of 40%



R code

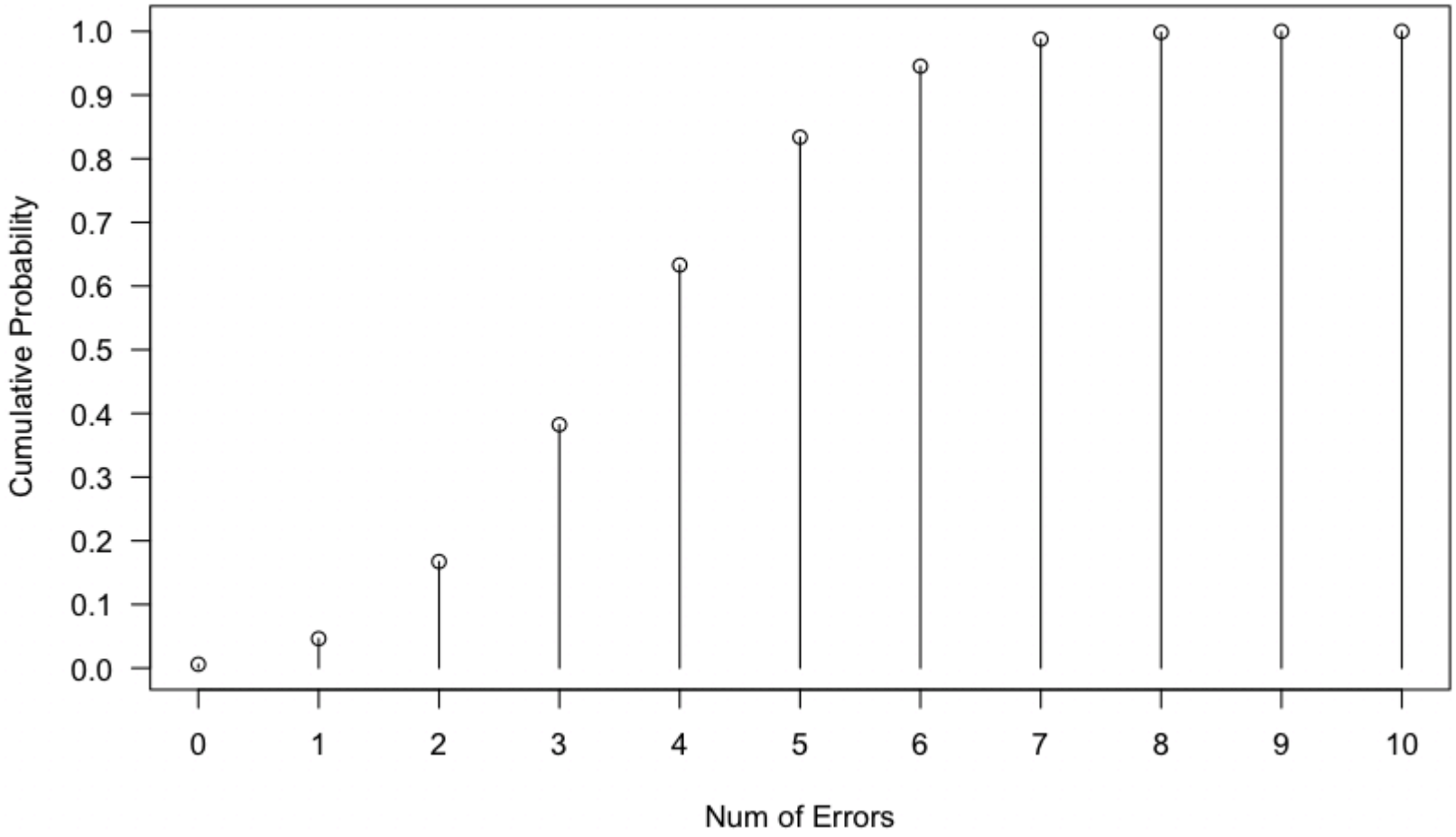



```
> errors <- 0:10
> prob.mass <- dbinom(errors, 10, .4)
> plot(errors, prob.mass, pch = 1, xlab = 'Num of Errors',
ylab = 'Probability', xaxt = "n", yaxt = "n")
> segments(x0 = errors, y0 = 0, x1 = errors, y1 = prob.mass)
> axis(1, at = seq(0, 10, by = 1), las=1)
> axis(2, at = seq(0, 1, by = .05), las=1)
```

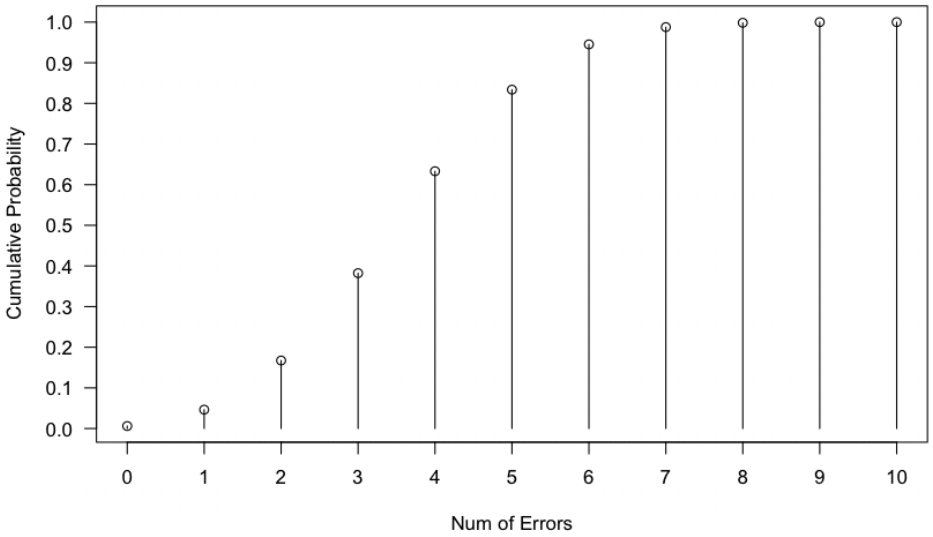
plot the values on the axes

The binomial distribution

Cumulative Distribution Function (cdf)



R code




```
> errors <- 0:10
> cprob.mass <- pbinom(errors, 10, .4)
> plot(errors, cprob.mass, pch = 1, xlab = 'Num of Errors',
ylab = 'Cumulative Probability', xaxt = "n", yaxt = "n")
> segments(x0 = errors, y0 = 0, x1 = errors, y1 = cprob.mass)
> axis(1, at = seq(0, 10, by = 1), las=1)
> axis(2, at = seq(0, 1, by = .1), las=1)
```

Continuous distributions

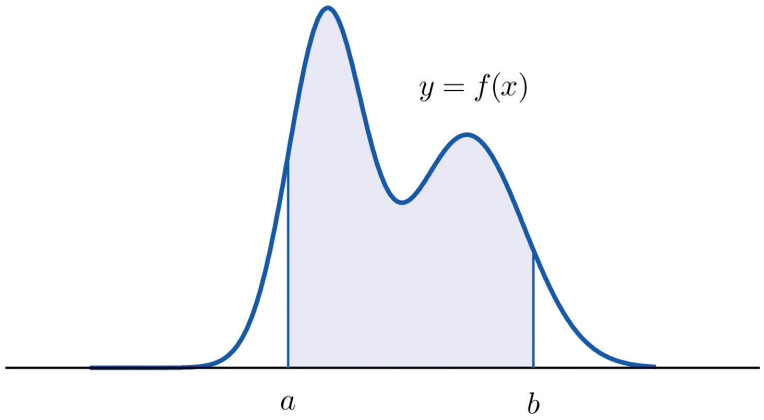
Not restricted to specific values. They can take any value between the lower and upper bound of a population

(of course, **populations can be unbounded**).

Continuous distributions

The probability of any particular value is zero. Probabilities can only be obtained for intervals (i.e., a range of values):

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

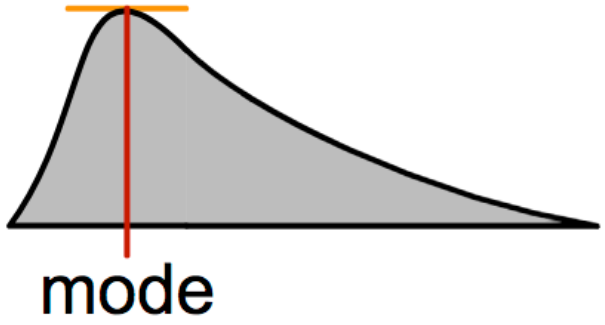


where *f(x)* is the **probability density function** (pdf). It provides the relative (rather than absolute) likelihood that a random variable *X* takes the value *x*

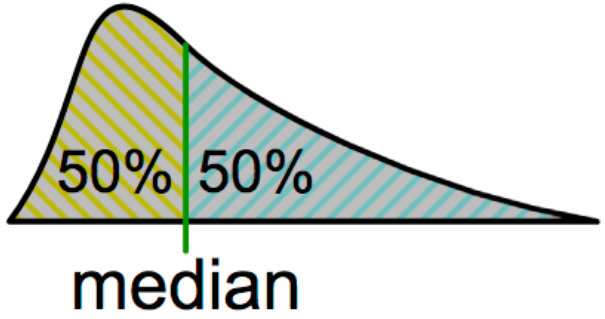
$$Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

Continuous distributions

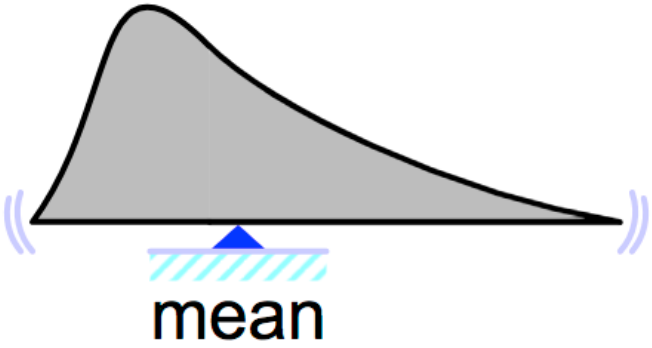
Mode: value of highest peak



Median: value that divides the area exactly in half

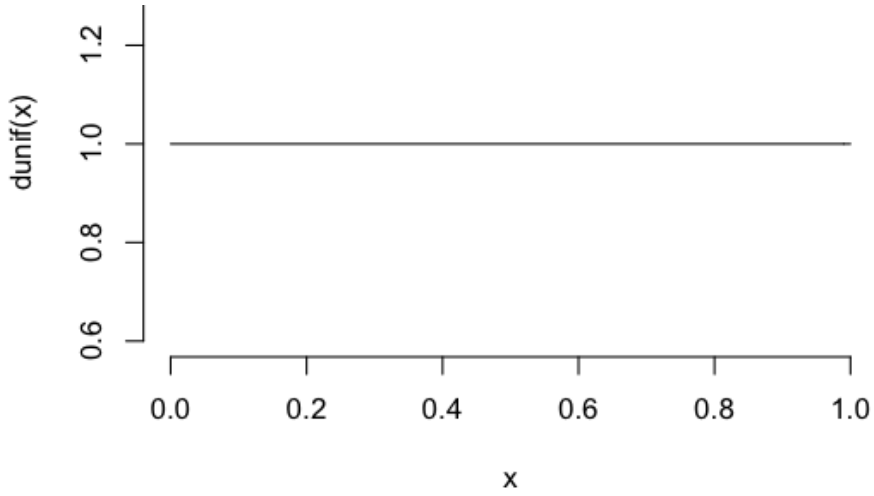


Mean: $\mu = \int_{-\infty}^{\infty} x f(x) dx$



The uniform distribution

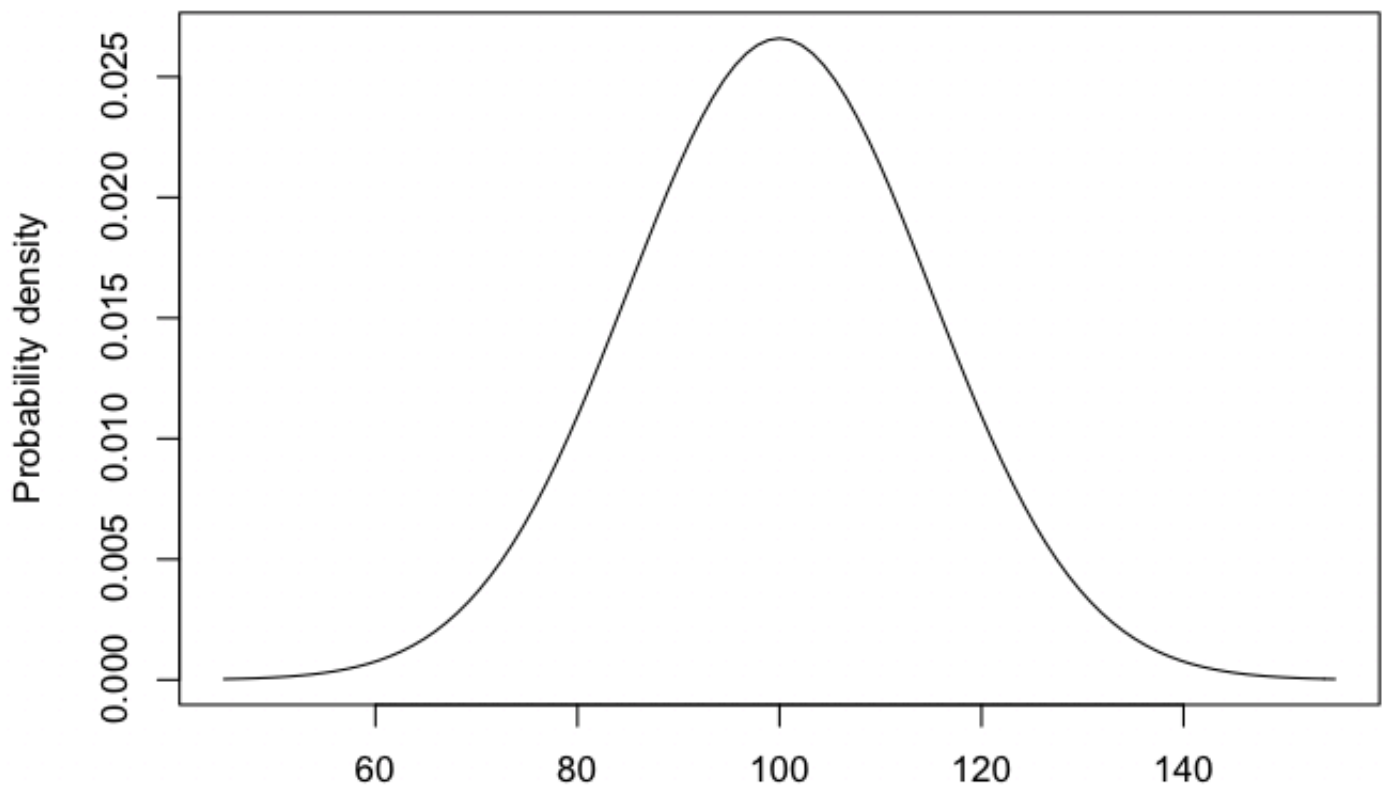
All values appear with the same probability



R code: `curve(dunif(x), xlim = c(0,1), bty="n")`

The normal distribution

Also known as the Gaussian distribution



The normal distribution

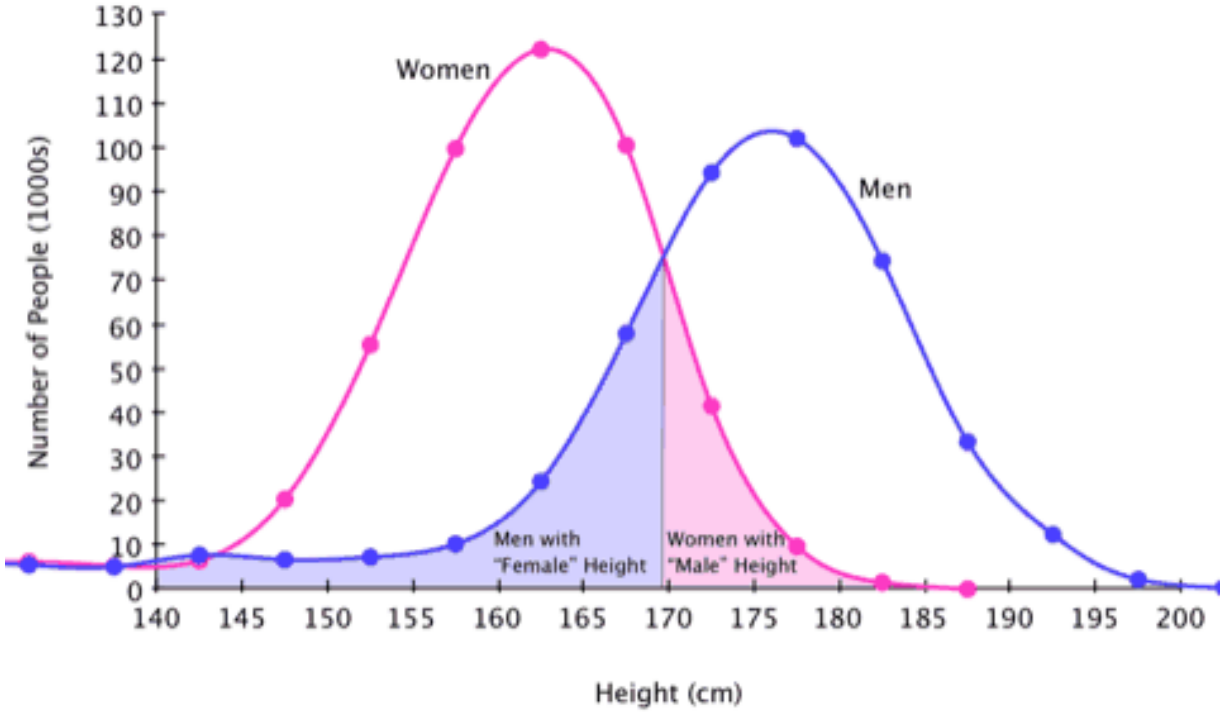
Symmetrical, unimodal and continuous

Can be derived as a sum of an infinite number of independent random variables

Thus, it is appropriate when data arise from a process that involves adding together the contributions from a large number of independent, random events

Example

The human height can be considered as the outcome of many independent random genetic and environmental influences



Normal distribution parameters

A normal distribution can be fully described by two only parameters: its mean μ and its variance σ^2

A normally distributed variable X can be written as

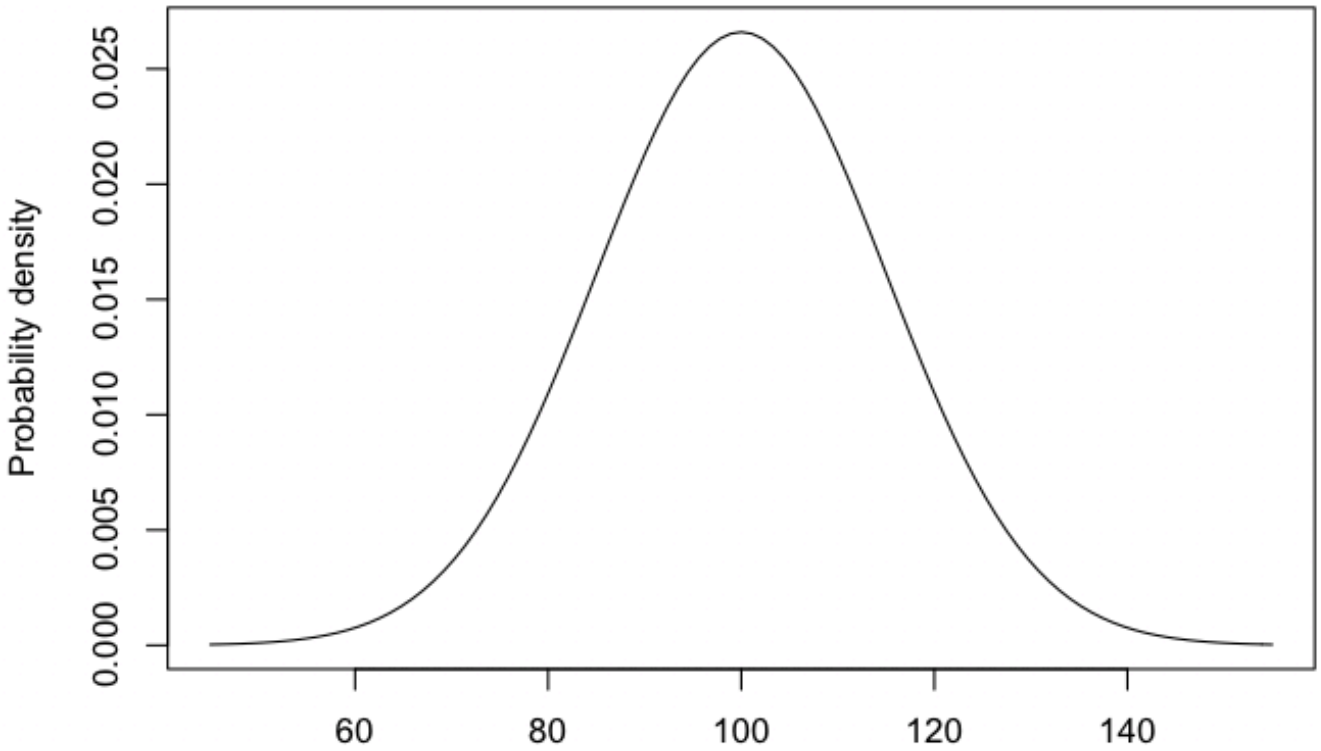
$$X \sim N(\mu, \sigma^2)$$

Its probability density function (pdf) is as follows:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu^2)}{2\sigma^2}}$$

Example

The following normal distribution has mean $\mu = 100$ and a standard deviation $\sigma = 15$




```
> curve(dnorm(x, 100, 15), xlim=c(45,155), ylab="Probability density")
```

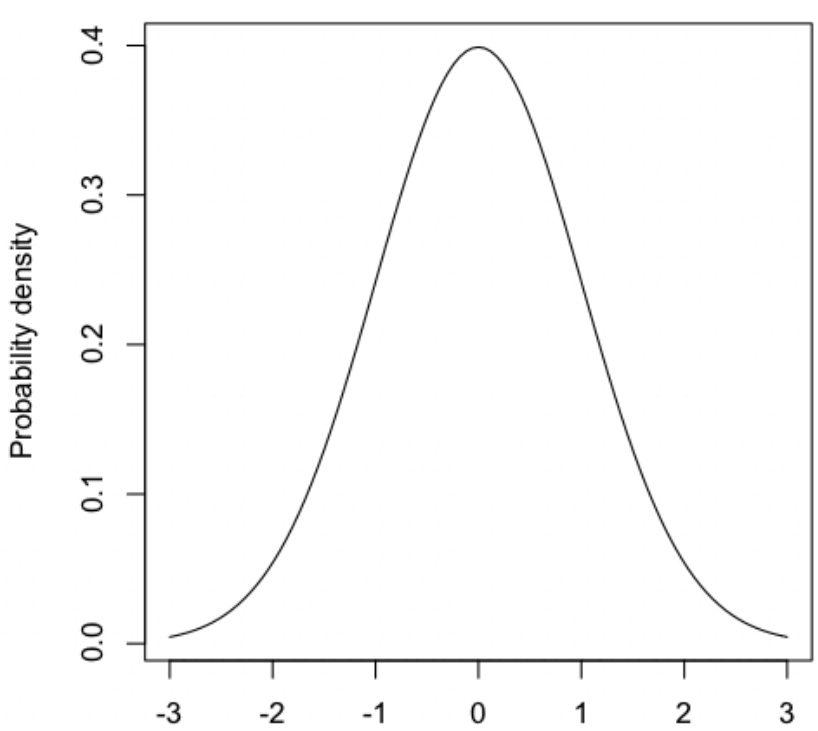
Standard normal distribution

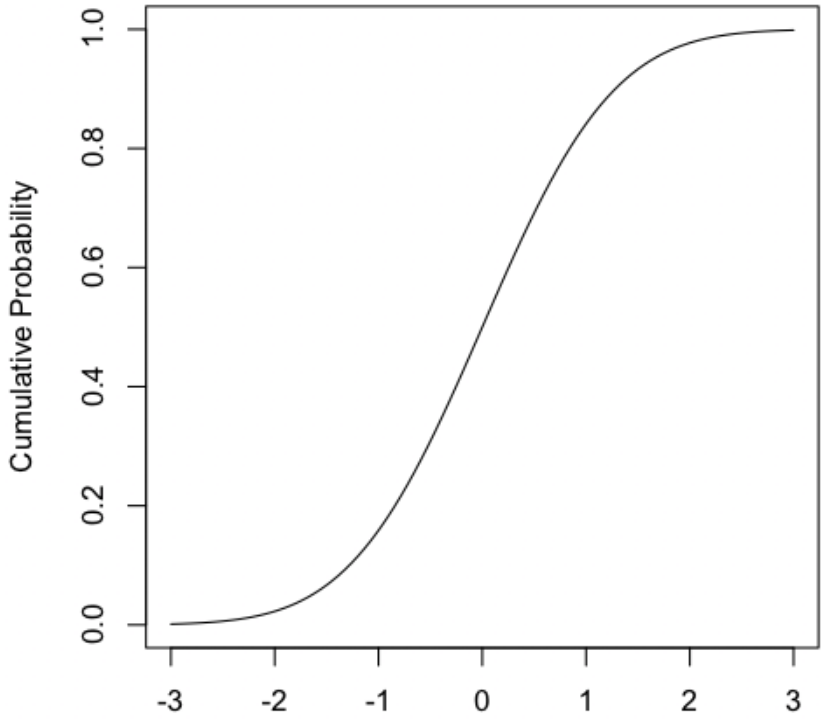
It is the normal distribution with a mean equal to 0 and a standard deviation (also variance) equal to 1:

$$z \sim N(0, 1)$$

The **standard normal distribution** is often abbreviated to *z*. It is frequently used to simplify working with normal distributions.

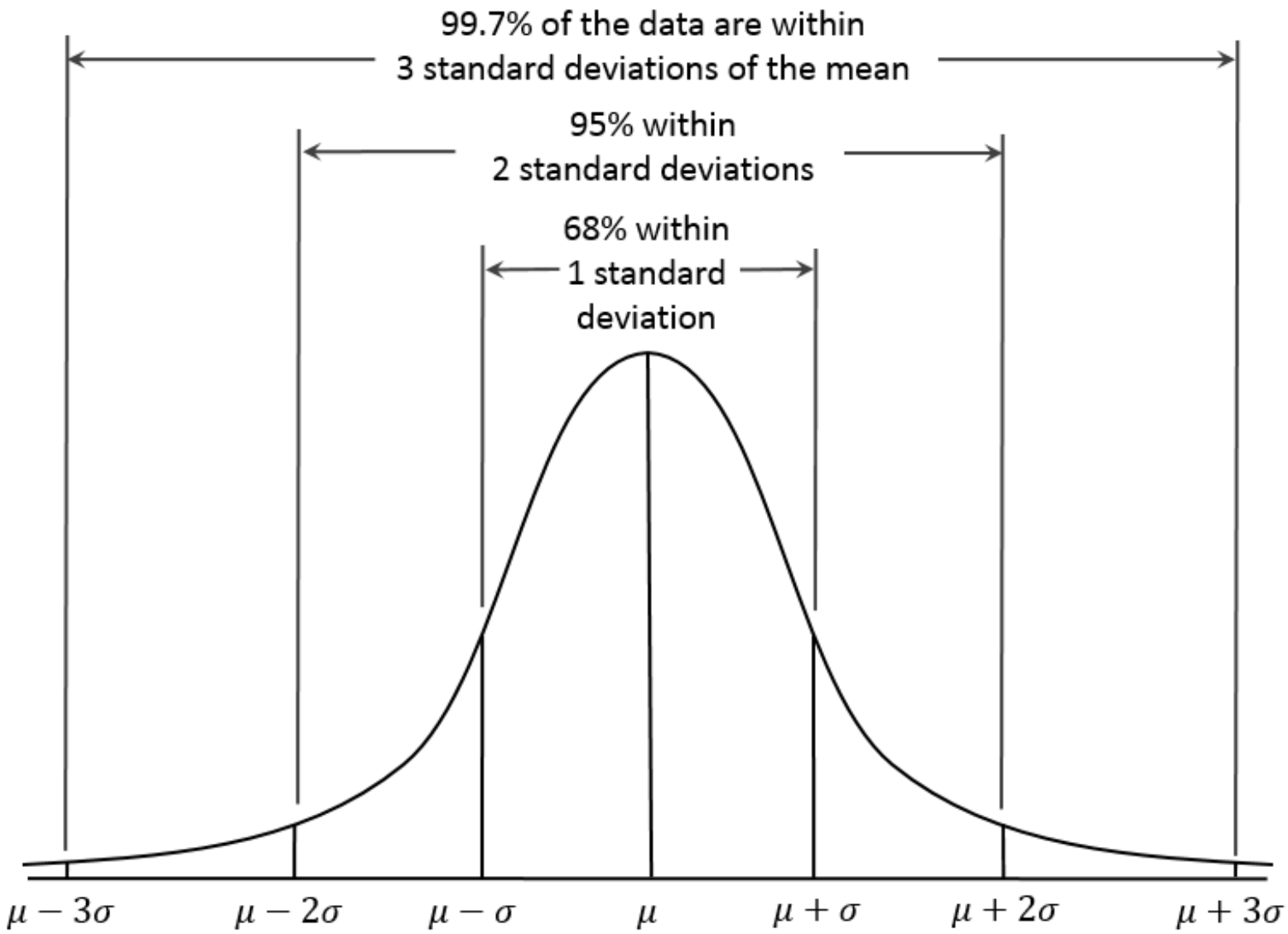
Standard normal distribution






```
> curve(dnorm(x, 0, 1), xlim=c(-3,3), ylab="Probability density")
> curve(pnorm(x, 0, 1), xlim=c(-3,3), ylab="Cumulative Probability")
```

Reading a normal distribution



Biased variance estimators (exercise)

1. Get a sample from a normal distribution with known variance and compute the variance of the sample (how?)

2. Repeat this process a large number of times and take the average to see how well you estimate approximates the true population variance

R Code

```r
biasedSD <- function(sample){
    M <- mean(sample)

    var <- sum((sample-M)^2)/length(sample)

    sqrt(var)
}

n <- 10 #Sample size
R = 100000 #Number of repetitions

biased <- replicate(R, biasedSD(rnorm(n, mean = 10, sd = 2)))
print(mean(biased))

unbiased <- replicate(R, sd(rnorm(n, mean = 10, sd = 2)))
print(mean(unbiased))
```

R Code

```r
biasedSD <- function(sample){
    M <- mean(sample)

    var <- sum((sample-M)^2)/length(sample)

    sqrt(var)
}

n <- 10 #Sample size
R = 100000 #Number of repetitions

biased <- replicate(R, biasedSD(rnorm(n, mean = 10, sd = 2)))
print(mean(biased))

unbiased <- replicate(R, sd(rnorm(n, mean = 10, sd = 2)))
print(mean(unbiased))
```

There is still bias, but it is lower…

Exercise

Suppose a random variable X is the average of 100 independent random variables $X_1 \ldots X_{100}$ that follow a uniform distribution.

Write an R program that displays the distribution of X

Solution

For each of the 100 random variables, take a random sample of 10000 values (from 0 to 1):

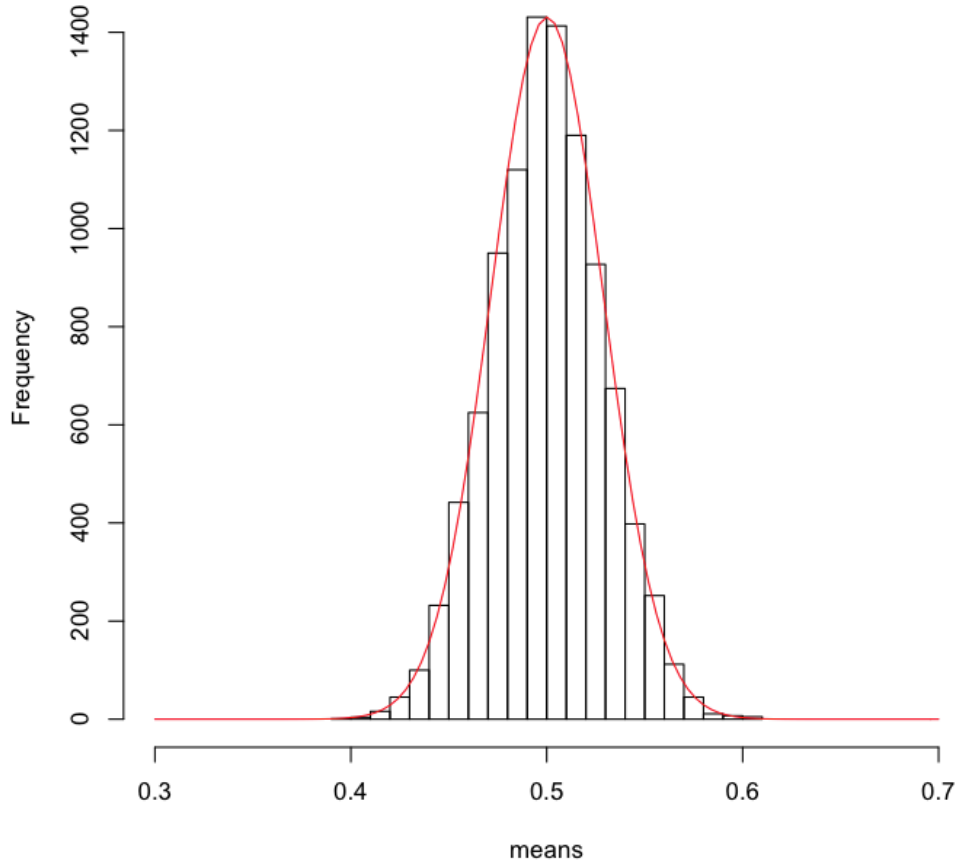
```
x1 <- runif(10000)
x2 <- runif(10000)
...
x100 <- runif(10000)
```

Then, take the average of the generated 100 vectors and plot the distribution of their values.

Solution

```
# We replicate the sampling process 100 times
matr <- replicate(100, runif(10000))

# We then take the average of the 100 columns of the created matrix
means <- rowMeans(matr)

# We plot the histogram of the means
hist(means, xlim = c(.3, .7), bty="n")

# Let's check if this is a normal distribution
par(new=TRUE)   # This is to say that we overdraw a new graph
curve(dnorm(x, mean = mean(means), sd = sd(means)), xlim = c(.3, .7),
      bty='n', col = 'red', axes = FALSE, ann = FALSE)
```

Solution



The distribution looks normal!

Sampling distribution of a statistic

It is the distribution obtained by calculating the statistic (e.g. the mean) from an infinite number of independent samples of size *n*

Example

An experiment measures the time it takes **n = 10** people to visually locate a target on a computer screen.

The same experiment is repeated a large (or infinite) number of times, where each time, we draw a new sample of size **n.**

For each experiment, we compute the **mean** time:

Experiment 1: M = 11.4 sec
Experiment 2: M = 12.6 sec
Experiment 3: M = 10.2 sec
...

What's the distribution of these mean values?

Sampling distribution of a statistic

Such distributions are interesting as they determine the probability of observing a particular value of the statistic, e.g., the mean.

It is often very different than the distribution of the data used to calculate the statistic.

distribution of the data

≠ sampling distribution of their means

Sampling distribution of the mean

Its mean value is also the mean (expected value) of the original population the samples were drawn from

Its standard deviation (SD) is known as the **standard error of the mean (SEM)**

The central limit theorem (CLT)

States that the **sampling distribution of a statistic** approaches the **normal distribution** as ***n* approaches infinity**

It applies to statistics computed by summing or averaging quantities (means, variances) but not to standard deviations (squared root of an average)

The central limit theorem (CLT)

States that the **sampling distribution of a statistic** approaches the **normal distribution** as ***n* approaches infinity**

It applies to statistics computed by summing or averaging quantities (means, variances) but not to standard deviations (squared root of an average)

central = fundamental to probabilities and statistics

limit = refers to a limit condition $n \to \infty$

The central limit theorem (CLT)

History

Back in 1733, **De Moivre** used the normal distribution to approximate the number of heads resulting from many tosses of a fair coin *(which as we said follows a binomial distribution)*.

Laplace gives a general proof of the theorem in 1809.

Practical importance of the CLT

If the size of the sample is **sufficiently large**, then the sampling distribution of the statistic will be **approximately normal**

(no matter what the distribution of the original population was)

But which sample size is **sufficiently large**?

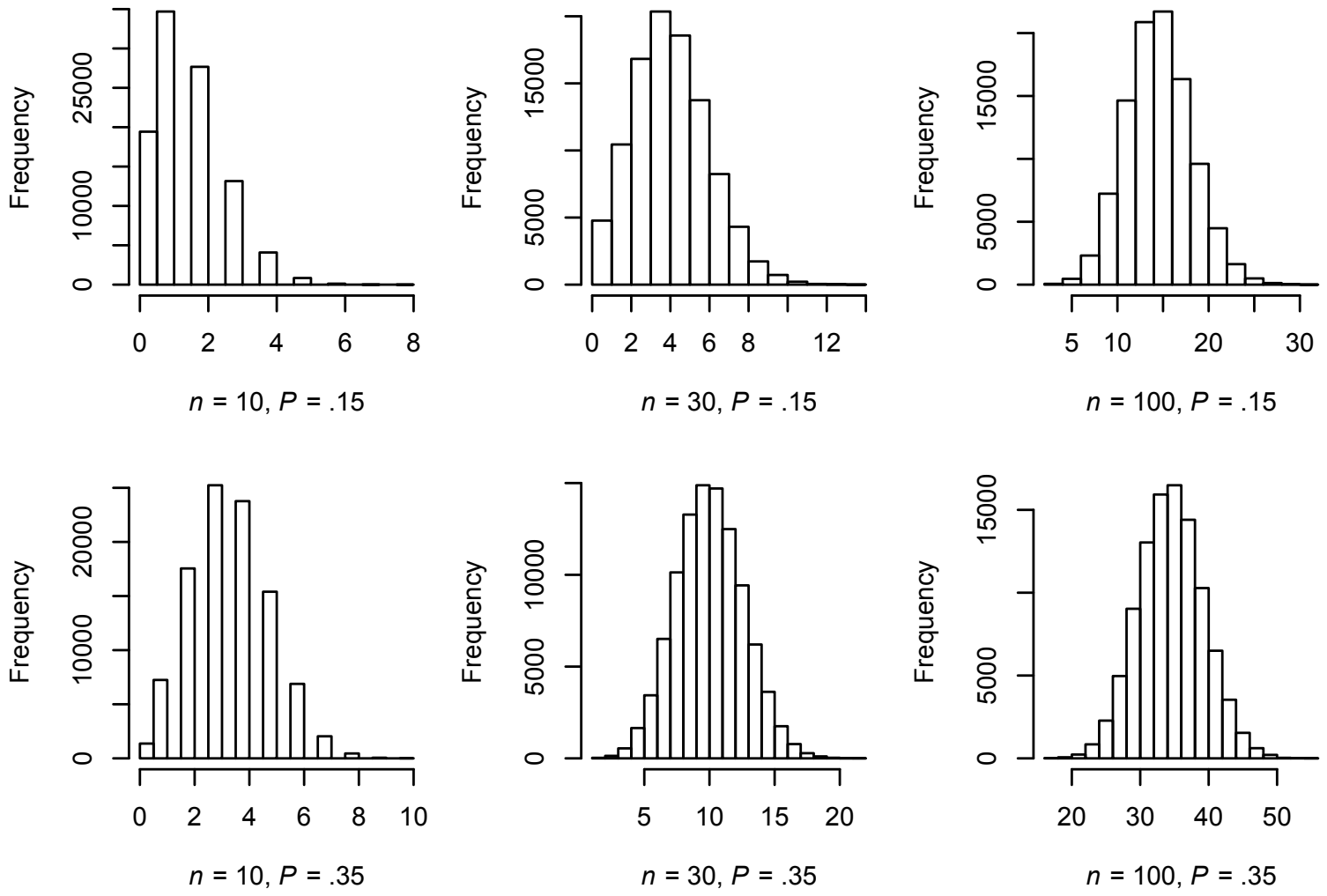
Sampling from normal distributions

If the original population is normal, then the CLT will always hold, even if the sample size is as low as $n = 1$

The further the original population moves away from a normal distribution, the larger the sample size n should be

Sampling from binomial distributions

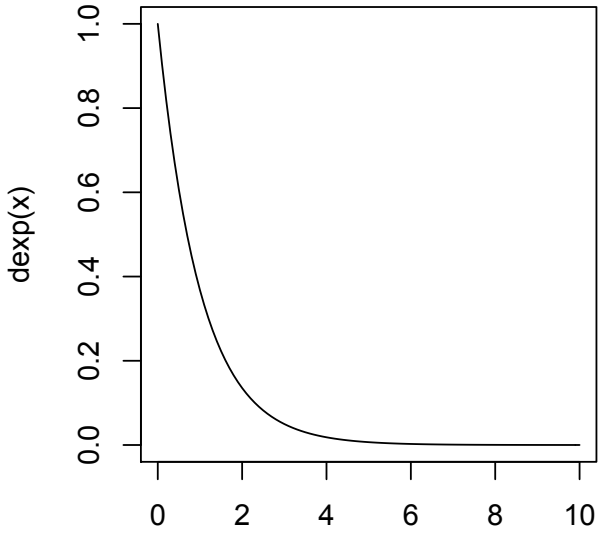
Statistic of interest: Count of successes from *n* Bernoulli trials



R code

```r
# P = .15
x1 <- rbinom(100000, 10, .15)
x2 <- rbinom(100000, 30, .15)
x3 <- rbinom(100000, 100, .15)

# P = .35
x4 <- rbinom(100000, 10, .35)
x5 <- rbinom(100000, 30, .35)
x6 <- rbinom(100000, 100, .35)

par(mfrow=c(2,3), mar = c(4,4,1,1), pty='s', cex.main = 1.1) # Show the histgrams in a 2x3 grid

hist(x1,xlab=expression(paste(italic(n), ' = 10, ', italic(P), ' = .15')), main = NULL)
hist(x2, xlab=expression(paste(italic(n), ' = 30, ', italic(P), ' = .15')), main = NULL)
hist(x3, xlab=expression(paste(italic(n), ' = 100, ', italic(P), ' = .15')), main = NULL)

hist(x4, xlab=expression(paste(italic(n), ' = 10, ', italic(P), ' = .35')), main = NULL)
hist(x5, xlab=expression(paste(italic(n), ' = 30, ', italic(P), ' = .35')), main = NULL)
hist(x6, xlab=expression(paste(italic(n), ' = 100, ', italic(P), ' = .35')), main = NULL)
```

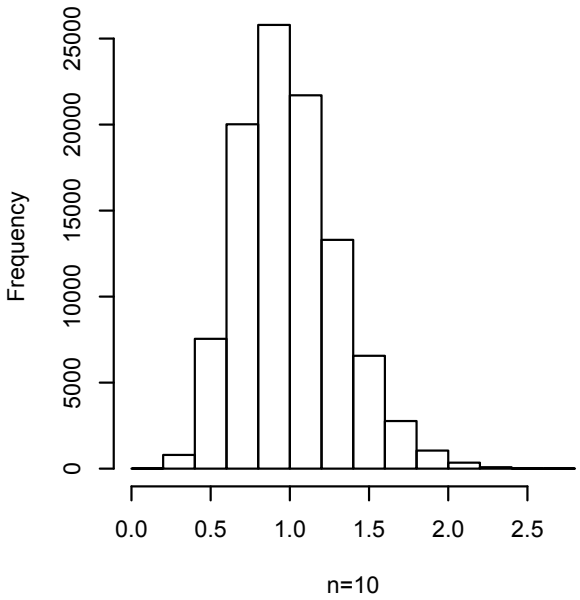
Sampling from exponential distributions

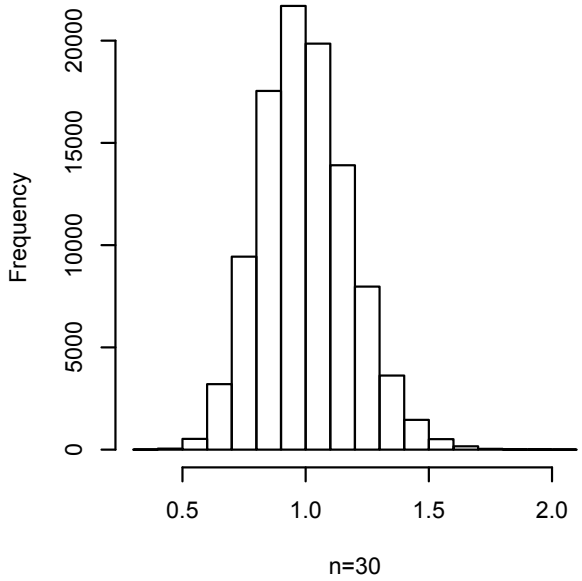


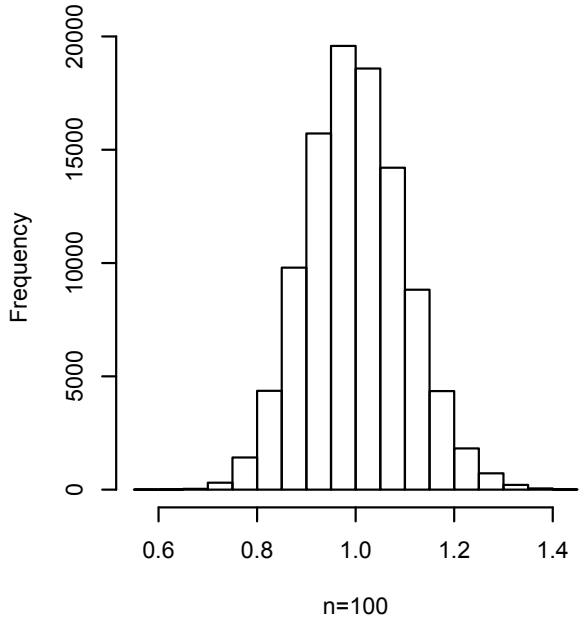
Distribution of source population

Statistic of interest: Mean of a sample of *n* drawn from an exponential distribution

Sampling distributions of the mean







R code

```r
# Show the histograms in a 1x3 grid
par(mfrow=c(1,3), mar = c(4,4,1,1), pty='s', cex.main = 1.1)

x <- replicate(100000, mean(rexp(10)))
hist(x,xlab='n=10', main = NULL)

x <- replicate(100000, mean(rexp(30)))
hist(x,xlab='n=30', main = NULL)

x <- replicate(100000, mean(rexp(100)))
hist(x,xlab='n=100', main = NULL)
```

What n is sufficiently large?

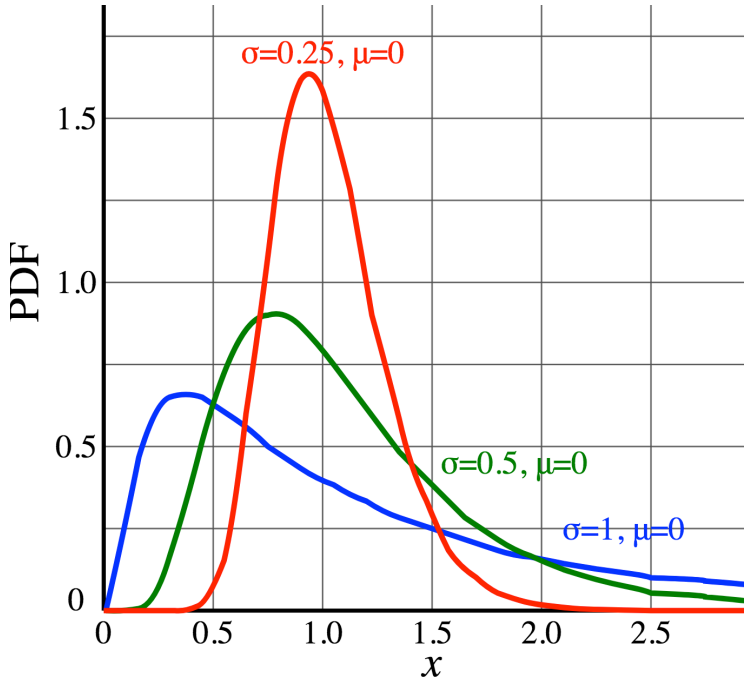
Several textbooks claim that n = 30 is enough to assume that a sampling distribution is normal, irrespective of the shape of the source distribution.

« **This is untrue** » [Baguley]
There is no magic number to guarantee that.

Log-normal distribution

A random variable *X* is log-normally distributed if the logarithm of *X* is normally distributed:

$$X \sim LogN(\mu, \sigma^2) \iff ln(X) \sim N(\mu, \sigma^2)$$



Simple math with logarithms

$$log_b(x) = a \iff b^a = x$$

$$log_b(1) = 0 \iff b^0 = 1$$

$$log_b(b) = 1 \iff b^1 = b$$

If the base of the logarithm is equal to the Euler number e = 2.7182, we write: $ln(x) = log_e(x)$

Which base to use is not important, but it is common to use *e* as a base.

Log-normal distribution

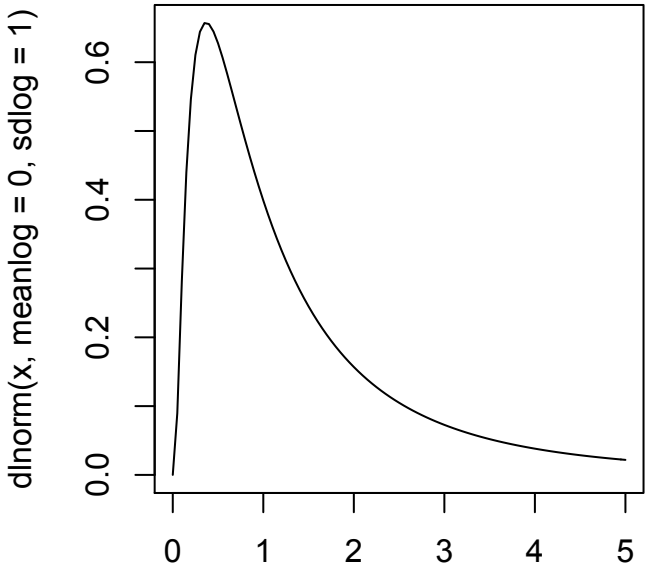
A common choice for real-world data bounded by zero, e.g., response time or task-completion time

> « *The reasons governing frequency distributions in nature usually favor the log-normal, whereas people are in favor of the normal* »

> « *For small coefficients of variation, normal and log-normal distribution both fit well.* »

[Limpert et al. 2001]
https://stat.ethz.ch/~stahel/lognormal/bioscience.pdf

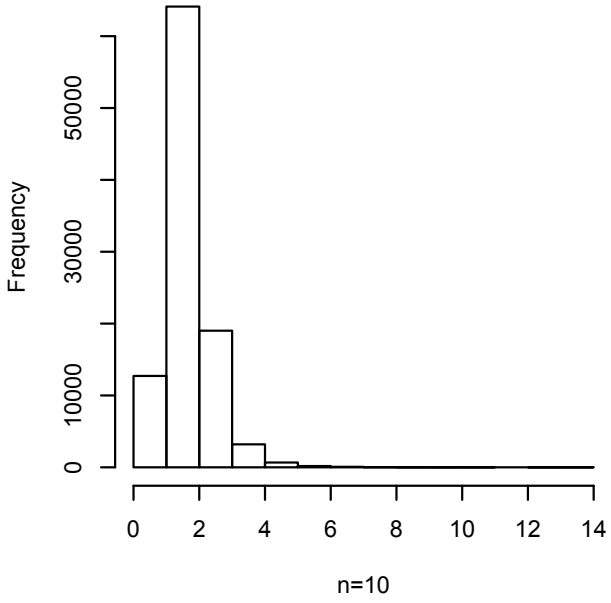
Sampling from lognormal distributions

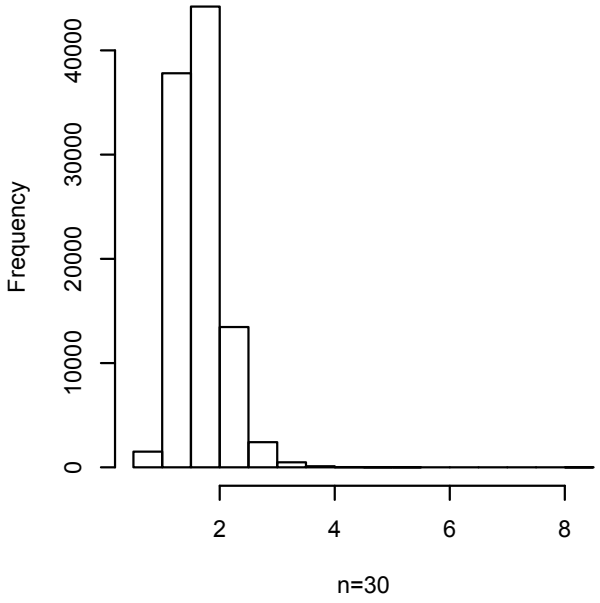


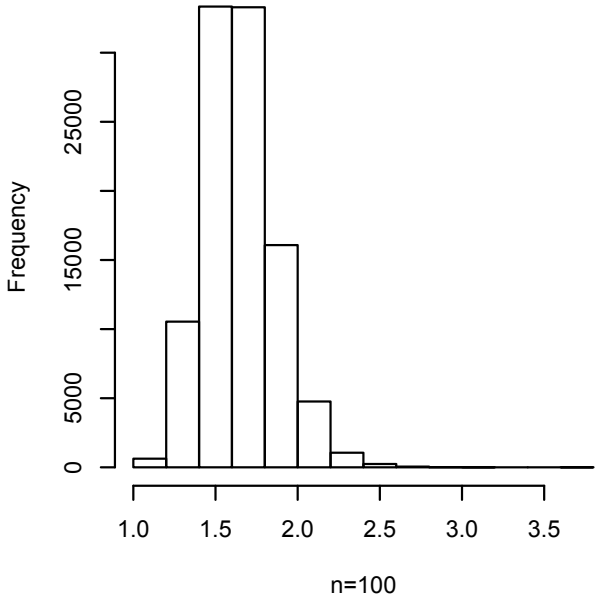
Distribution of source population

$\mu = 0, \sigma = 1$

Sampling distributions of the mean



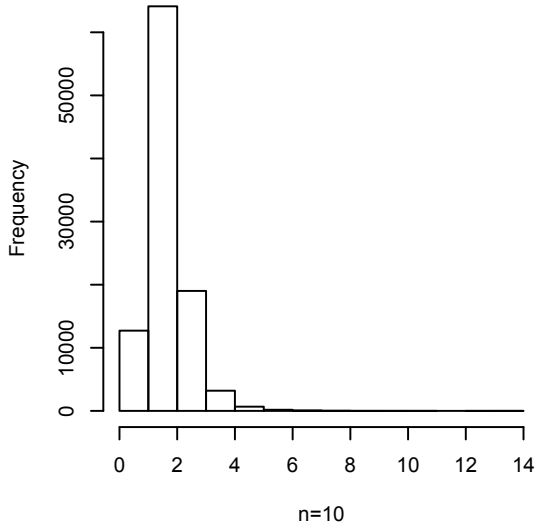


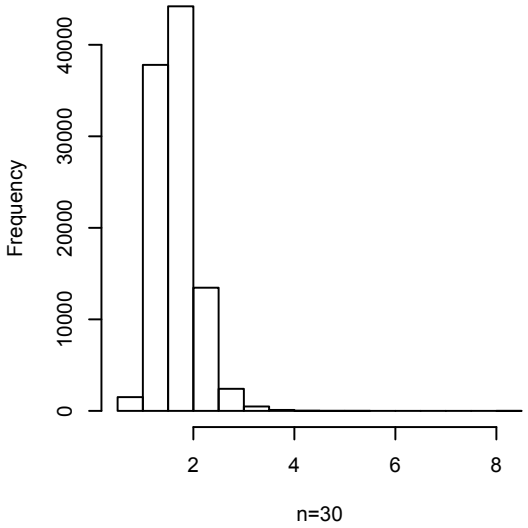


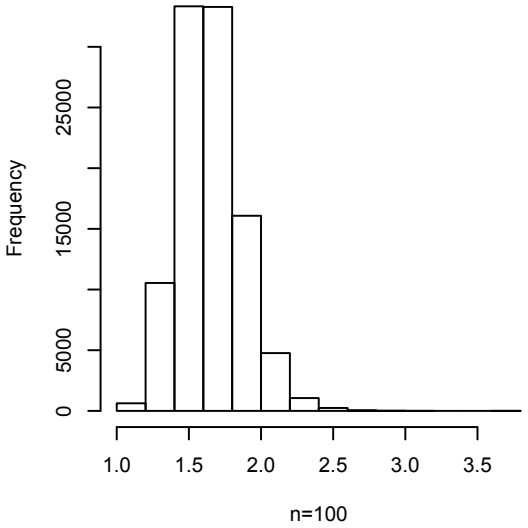
Sampling from lognormal distributions

Before...

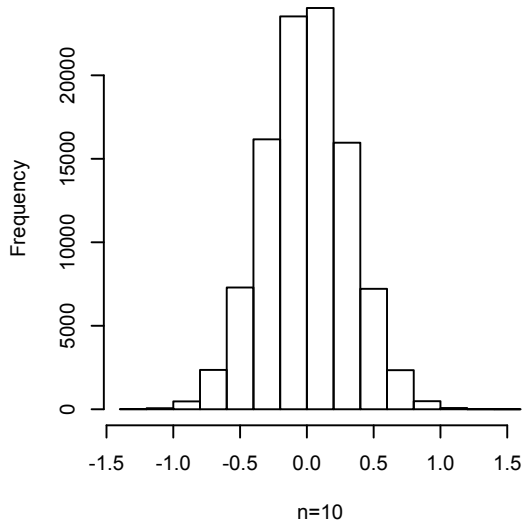
Sampling distributions of the mean

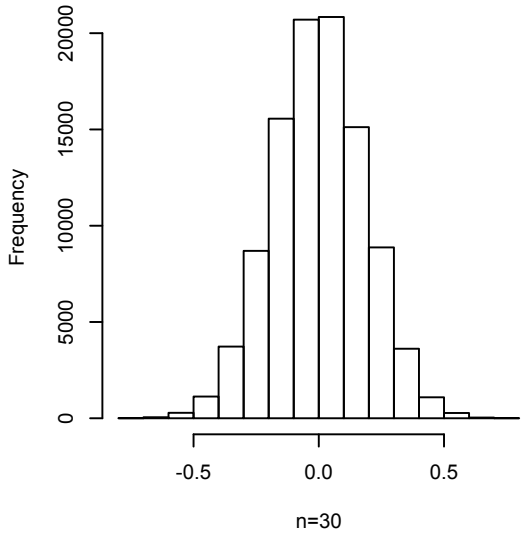


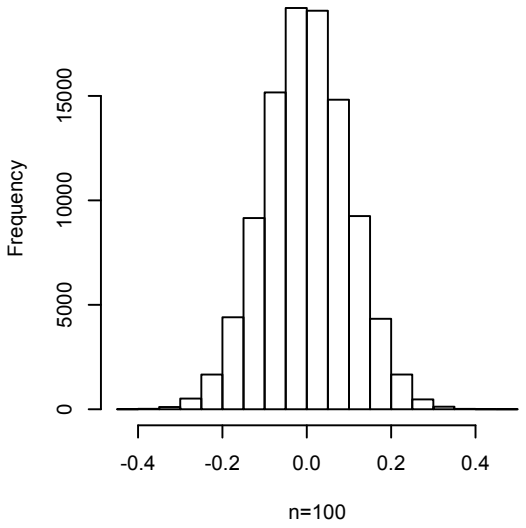




...and after applying a log transformation on the data





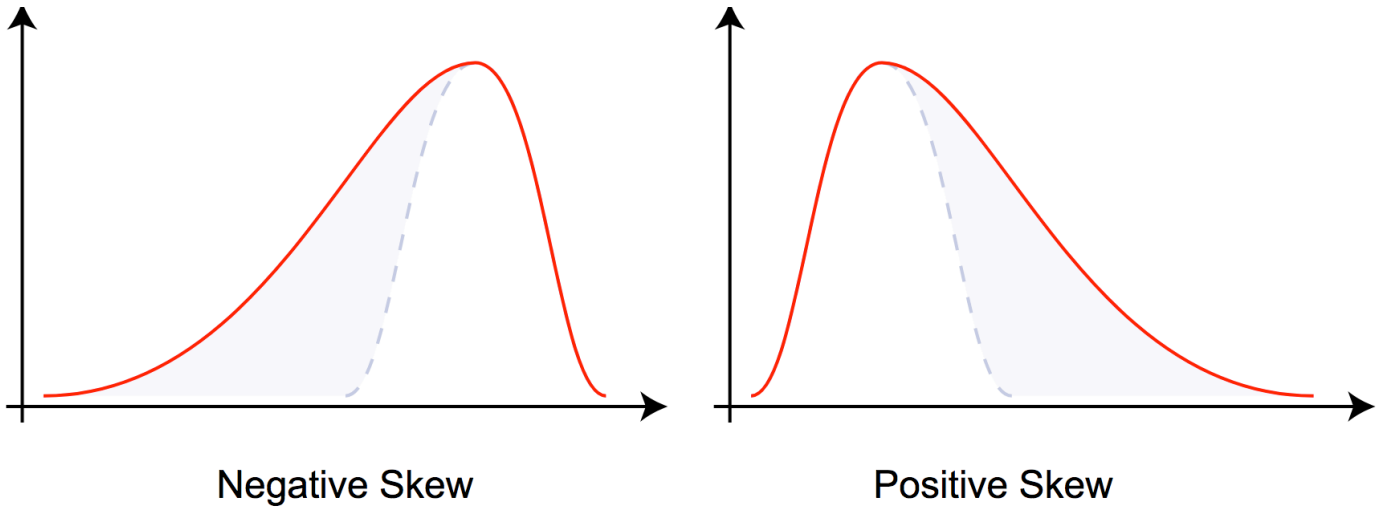


R code

```r
# Show the histograms in a 2x3 grid
par(mfrow=c(2,3), mar = c(4,4,1,1), pty='s', cex.main = 1.1)

x <- replicate(100000, mean(rlnorm(10, meanlog = 0, sdlog = 1)))
hist(x,xlab='n=10', main = NULL)

x <- replicate(100000, mean(rlnorm(30, meanlog = 0, sdlog = 1)))
hist(x,xlab='n=30', main = NULL)

x <- replicate(100000, mean(rlnorm(100, meanlog = 0, sdlog = 1)))
hist(x,xlab='n=100', main = NULL)


# Applying a log transformation to the values
x <- replicate(100000, mean(log(rlnorm(10, meanlog = 0, sdlog = 1))))
hist(x,xlab='n=10', main = NULL)

x <- replicate(100000, mean(log(rlnorm(30, meanlog = 0, sdlog = 1))))
hist(x,xlab='n=30', main = NULL)

x <- replicate(100000, mean(log(rlnorm(100, meanlog = 0, sdlog = 1))))
hist(x,xlab='n=100', main = NULL)
```

Skewed distributions

Asymmetrical distributions are said to be **skewed**



The chi-square (χ^2) distribution

Consider a squared observation z^2 drawn at random from the standard normal (z) distribution

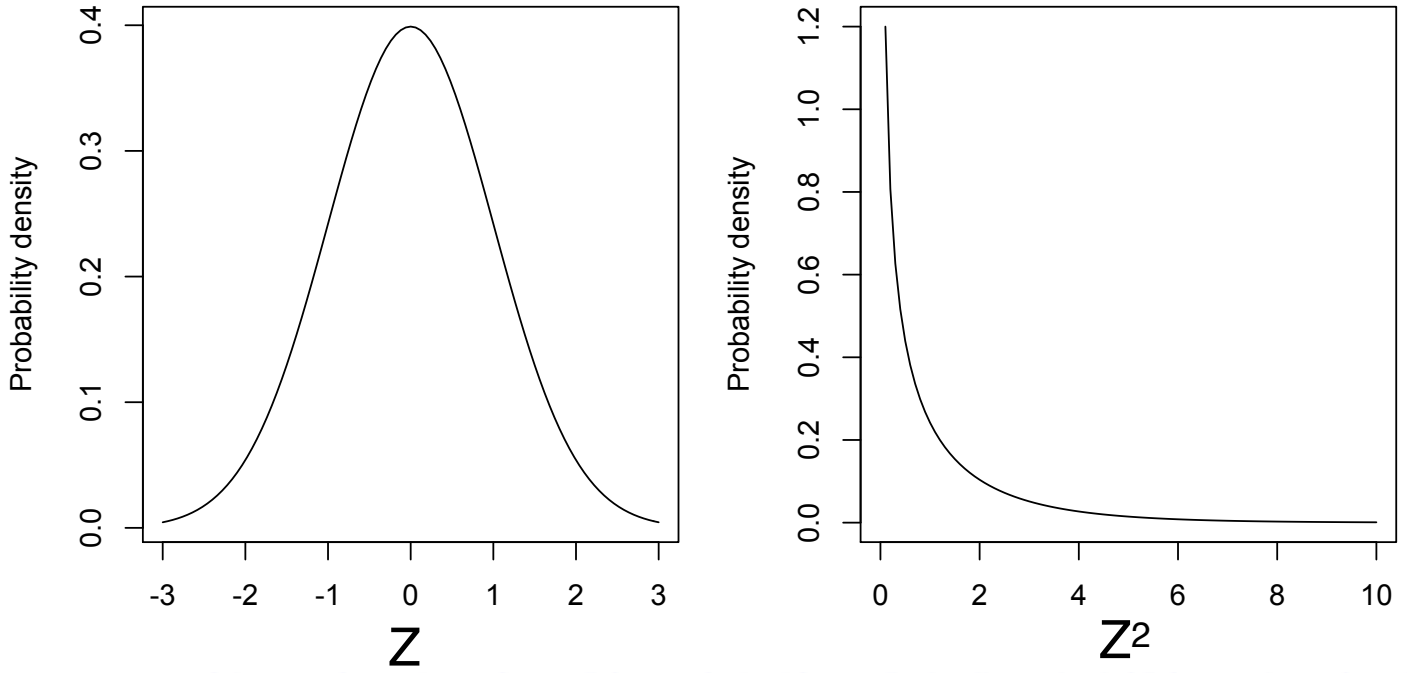
The distribution of z^2 will follow a χ^2 distribution with 1 degree of freedom (df)

To check how the distribution looks with R

```
> z<-rnorm(100000)
> hist(z^2)
```

The chi-square (χ^2) distribution

The distribution of z^2 will follow a χ^2 distribution with 1 degree of freedom (df)

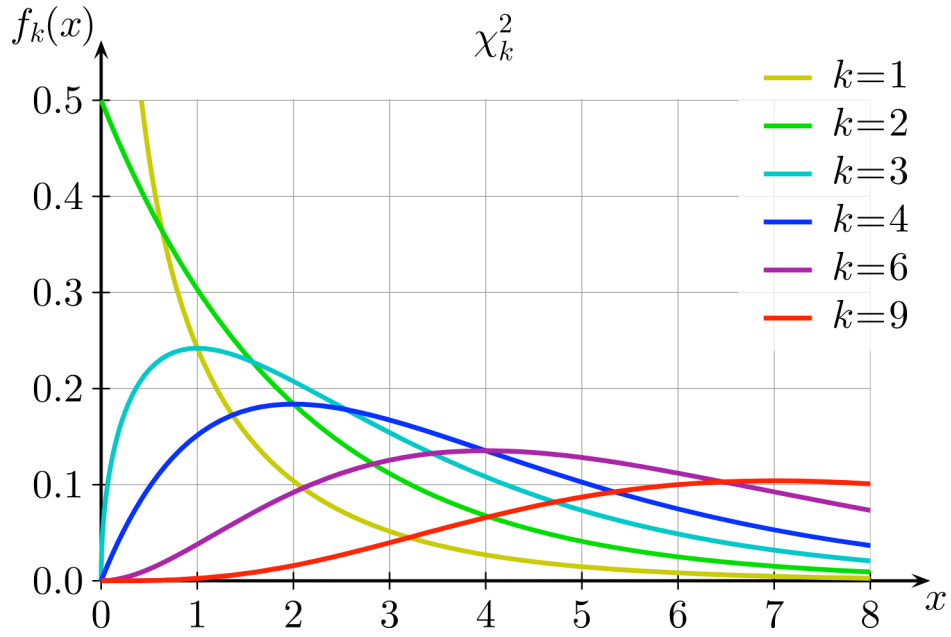



```
> curve(dnorm(x, 0, 1), xlim=c(-3,3), ylab="Probability density")
> curve(dchisq(x, 1), xlim=c(0,10), ylab="Probability density")
```

The chi-square (χ^2) distribution

A χ^2 distribution with **k degrees of freedom** is the distribution of a sum of squares of k independent variables that follow a standard normal distribution

$$Q = \sum_{i=1}^{k} Z_i^2 \Longrightarrow Q \sim \chi^2(k)$$



The chi-square (χ^2) distribution

Given the link between variances and sums of squares, **the chi-square distribution is useful for modeling variances of samples** from normal (or approximately normal) distributions.

Let's verify how the sampling distribution of the variance of n = 10 samples looks like

```
> samples <- replicate(10000, var(rnorm(10)))
> hist(samples)
```

R distribution functions

Binomial distribution

dbinom(x, n, P)
Provides the **probability mass function** for the binomial distribution *B(n,P)*

Examples:
dbinom(4, 20, .2): It will return the probability of *x = 4* successes for *n = 20* Bernoulli trials with a *P=.2* probability of success.

dbinom(c(1,2,3,4), 10, .2): It will return a vector with the probabilities of *x = {1, 2, 3, 4}* successes for *n =10* Bernoulli trials with a *P =.2* probability of success.

R distribution functions

Binomial distribution

pbinom(x, n, P)
Provides the **cumulative** probability mass function for the binomial distribution *B(n,P)*

Example:
pbinom(4, 20, .2): It will return the **cumulative** probability **up to** $x = 4$ successes for $n = 20$ Bernoulli trials with a $P = .2$ probability of success.

R distribution functions

Binomial distribution

rbinom(size, n, P)
It will generate a **random sample** of size *size* from the binomial distribution *B(n,P)*

Example:
rbinom(10, 20, .2): It will return a random sample of *size = 10* from the binomial distribution *B(n = 20, P = .2)*

```
> rbinom(10, 20, .2)
 [1] 3 6 1 5 3 4 5 2 4 3
```

R distribution functions

Normal distribution

dnorm(x, mean, sd)
Provides the probability density function for the normal distribution with a mean value equal to *mean* and a standard deviation equal to *sd.*

Examples:
dnorm(.2, 0, 1):
It will return the **relative** likelihood of the value *x = .2,* for the standard normal distribution.

curve(dnorm(x, 100, 15), xlim = c(60, 140)):
It will plot the probability density function from x = 60 to x = 140 for the normal distribution with *mean = 100* and *sd = 15*.

R distribution functions

Normal distribution

pnorm(x, mean, sd)
Provides the **cumulative** probability density function for the normal distribution with a mean value equal to *mean* and a standard deviation equal to *sd.*

Example:
pnorm(100, 100, 15):
It will return the **cumulative** probability **up to** *x = 100* for the normal distribution with *mean = 100* and *sd = 15*.
(What do you expect it to be?)

R distribution functions

Normal distribution

rnorm(size, mean, sd)
It will generate a random sample of size *size* from the normal distribution with a mean value equal to *mean* and a standard deviation equal to *sd.*

Example:
rnorm(10, 0, 1): It will return a random sample of *size = 10* from the standard normal distribution.

```
> rnorm(10, 0, 1)
 [1] -0.4517647  0.8649485 -0.6705683 -1.1822377 -0.4995629
 [6]  0.9161862  0.1604768 -0.7899337 -1.0221835 -0.4047310
```

R distribution functions

	Distribution function (pmf or cdf)	Cumulative distr. function	Random sampling
Binomial	dbinom(x, n, P)	pbinom(x, n, P)	rbinom(size, n, P)
Uniform	dunif(x)	punif(x)	runif(size)
Normal	dnorm(x, mean, sd)	pnorm(x, mean, sd)	rnorm(size, mean, sd)
Log-normal	dlnorm(x, meanlog, sdlog)	plnorm(x, meanlog, sdlog)	rlnorm(size, meanlog, sdlog)
chi-squared	dchisq(x, k)	pchisq(x, k)	rchisq(size, k)