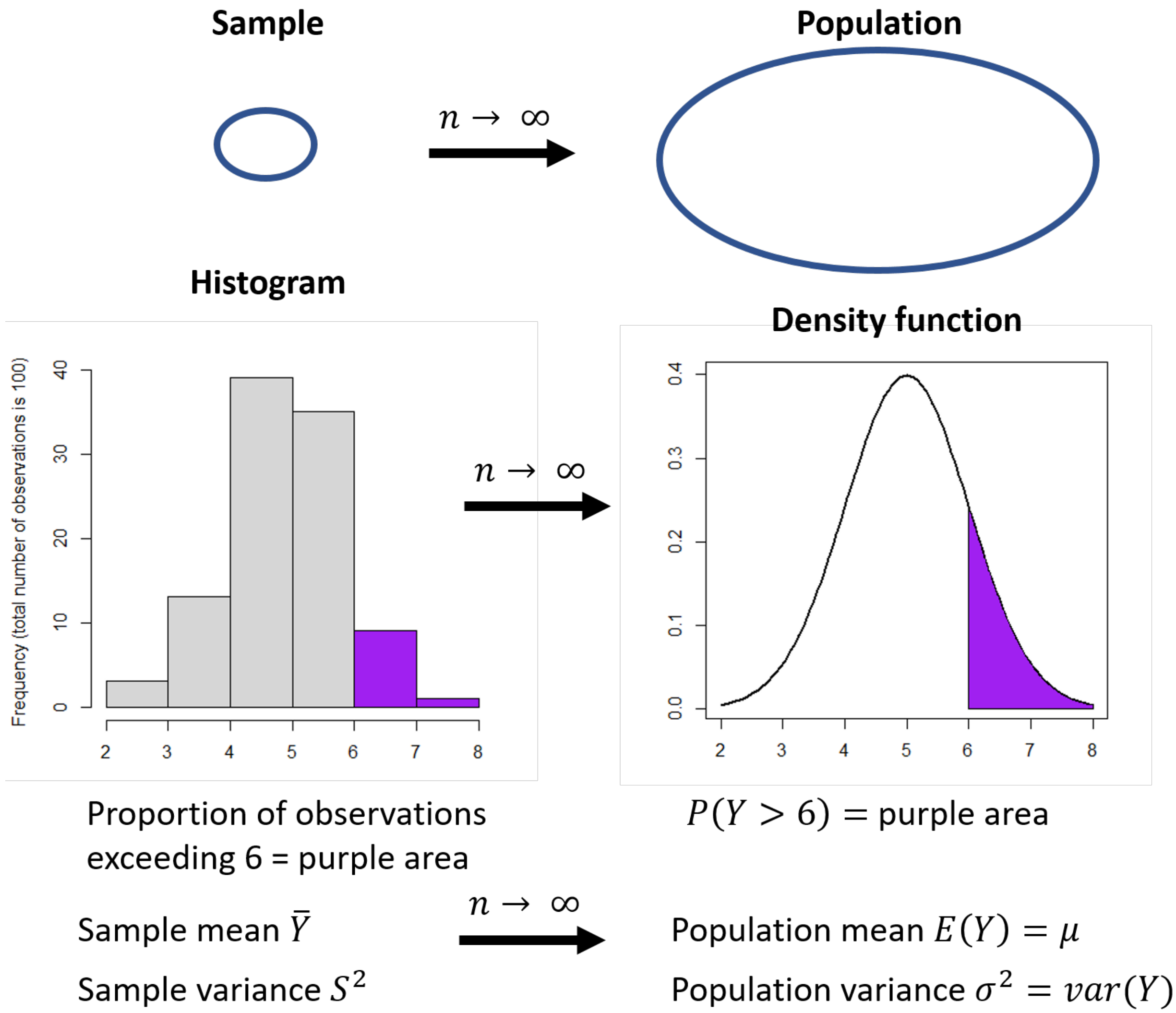
Chapter 3: Important Distributions

Contents

1 Sample versus Population



A **population** consists of the totality of observations with which we are concerned. It can be finite (number of bottles produced by a company on a daily base) or infinite (CO concentration measured at a daily base).

Usually we cannot observe the complete population but have to make inferences about the distribution of the population by making a random sample. A **sample** is a subset of observations selected from the population. The sample must be representative for the population. Therefore, the sample must be random and not biased.

2 Discrete and Continuous random variable.

A **random variable** (denoted by Y) is a variable whose values depend on outcomes of a random phenomenon.

2.1 Discrete variable:

Let Y be a random variable. Y is said to be a ***discrete random variable*** *if Y can take at most a countable number of values.*

- **Example** ***Decathlon***:
  Y = Competition → Only two possible values: Olympic games or Decastar competition.

- **Example *Temperature***:
  Y = area of the city → Only four possible values: North, South East or West.

2.2 Continuous variable:

*Let Y be a random variable. Y is said to be a **continuous random variable** if Y can take any possible real value over some interval.*

- **Example *Decathlon***:
  Y = time on 100 meters.

- **Example *Temperature***:
  Y = October temperature of a city.

Remark:
We use *uppercase letters* (X, Y, ...) to denote random variables and *lowercase letters* (x, y, ...) to denote specific values for these variables.

3 Discrete distributions

3.1 Introduction.

Y is said to be a **discrete random variable** if Y can take at most a countable number of values.

Every discrete random variable Y has a **probability function**

$$f_Y(y) = p(y) = P(Y = y).$$

Two properties of $p(y)$ are:

- $0 \leq P(Y = k) \leq 1$
- $\sum_k P(Y = k) = 1$

3.2 Binomial distribution

We often must model the random behavior of data that we can classify as either a success or a failure. We repeat the event a number of times and count the number of successes.

- **Example *Albino***:
  If two carriers of the gen for albinism get children, then each of the children has probability of 0.25 of being albino.
  A binomial random variable Y can be found in this example:
  Y = number of Albino children when the couple has 3 children.

In general for a binomial random variable:

1. Each trial results in one of two outcomes: one outcome that is considered a **success** and the other is considered as a **failure**.
2. The **probability of a success on a single trial is** p and remains the same from trial to trial. The probability of a failure is denoted by $q = (1 - p)$.
3. The experiment consists of n ***identical attempts or "trials"***.
4. The trials are *independent*.

If these four conditions hold and we are **interested in Y, the total number of successes among the n trials**, then Y is called a **binomial random variable**.

A binomial random variable Y follows a binomial distribution. This is notated as:

$$Y \sim B(n, p)$$

The probability function is

$$f_Y(y) = \binom{n}{y} p^y q^{n-y} = \frac{n!}{y!(n-y)!} p^y q^{n-y} \text{ for } y = 0, 1, ..., n.$$

If Y follows a binomial distribution ($Y \sim B(n,p)$), we can show that:

- $\mu = E(Y) = np$
- $\sigma^2 = npq = np(1-p)$
- $\sigma = \sqrt{npq}$

3.2.1 Examples

Example *Albino*:
If two carriers of the gen for albinism get children, then each of the children has probability of 0.25 of being albino.
Y = number of albino children when the couple has 3 children.
$Y \sim B(n = 3, p = 0.25)$.

We want to complete the next table:

y = observed number of albinos	$y = 0$	$y = 1$	$y = 2$	$y = 3$
$P[Y = y]$				
$P[Y \leq y]$				

1. What is the probability that one child out of 3 children is albino?
Solving $P[Y = 1]$ corresponds to solving the question: What is the probability that one child out of 3 children is albino?
$P[Y = 1] = \frac{3!}{1! \cdot 2!}(0.25^1 \cdot 0.75^2) = 3 \cdot 0.1406 = 0.42$
Analogue calculations can be used to complete the first row of the table.

In R we can use the function `dbinom` for completing first row of table:

```
k <- c(0:3)
densbin <- dbinom(k, 3, 0.25)
names(densbin) <- k
densbin
```

```
##        0        1        2        3
## 0.421875 0.421875 0.140625 0.015625
```

Filling in these numbers in the table:

y = observed number of albinos	$y = 0$	$y = 1$	$y = 2$	$y = 3$
$P[Y = y]$	0.42	0.42	0.14	0.015
$P[Y \leq y]$				

2. What is the probability that at most 1 child will be an albino?
Solving $P[Y \leq 1]$ corresponds to solving the question: What is the probability that at most 1 child will be an albino?
$P[Y \leq 1] = 0.42 + 0.42 = 0.84$

In R, we can use the function `pbinom`:

```
# Compute cumulative density
cumdens <- pbinom(k, 3, 0.25)
```

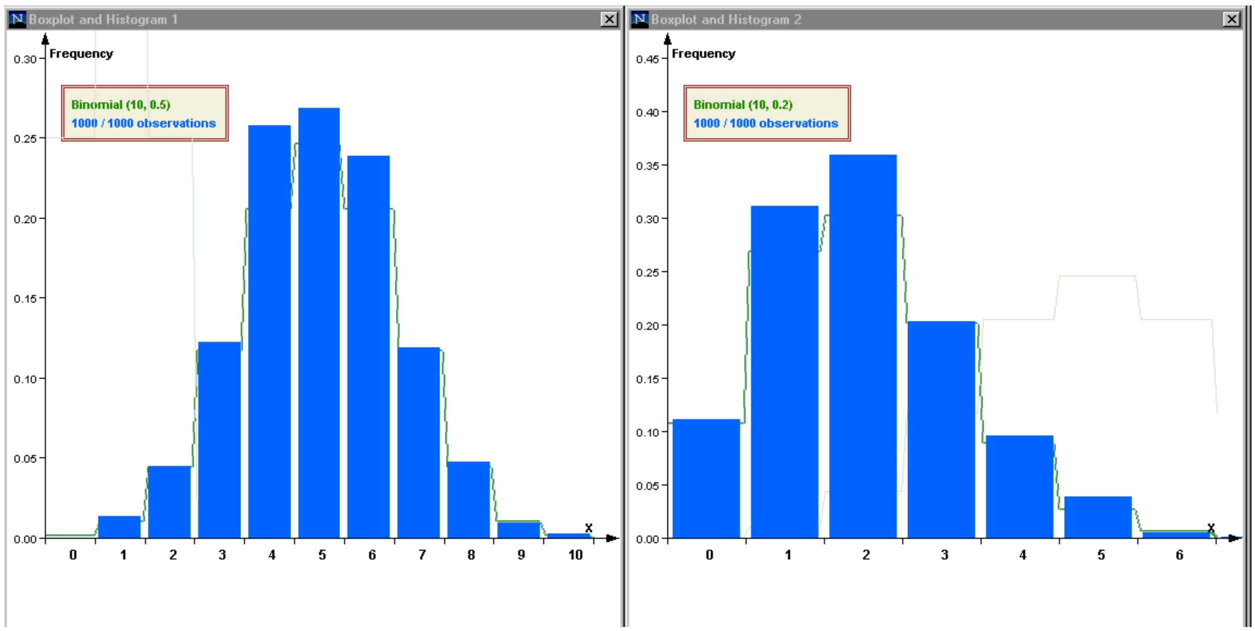
```r
names(cumdens) <- k
cumdens
```

```
##        0        1        2        3
## 0.421875 0.843750 0.984375 1.000000
```

With this, the table can be completed:

y = observed number of albinos	$y = 0$	$y = 1$	$y = 2$	$y = 3$
$P[Y = y]$	0.42	0.42	0.14	0.015
$P[Y \leq y]$	0.42	0.84	0.98	1

3.2.2 Visualization of some binomial distributions

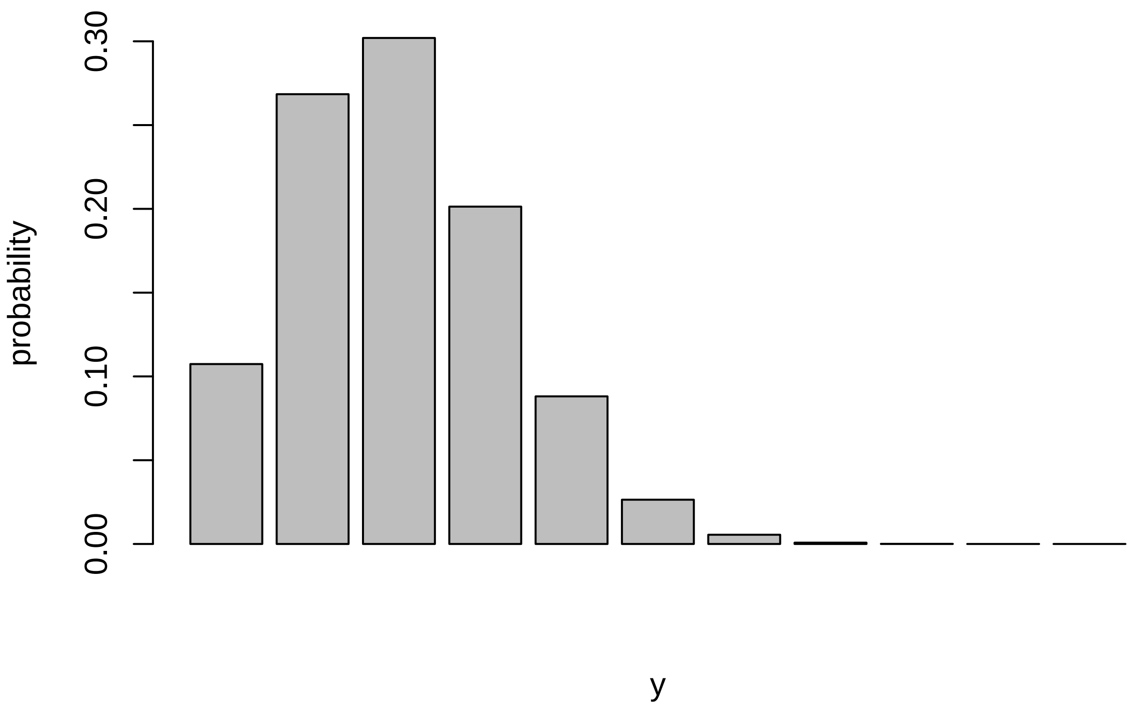


3.2.3 Use of the Binomial distribution in R

Function in R	Symbolic notation	Description
`dbinom(y, size = , p = )`	$P(Y = y)$	Binomial probability distribution
`pbinom(y, size = , p = )`	$P(Y \leq y)$	Cumulative binomial probability distribution

Plot the binomial distribution for $n = 10$ and $p = 0.20$ as follows:

```r
y <- c(0:10)
dens <- dbinom(y, size = 10, prob = 0.2)
barplot(dens, xlab = "y", ylab = "probability")
```



3.3 The Poisson distribution

Example *cars per minute*
There are 12 cars crossing a bridge per minute on average.

1. Find the probability of having 17 cars crossing the bridge in a particular minute.
2. Find the probability of having 17 cars or more crossing the bridge in a particular minute.

We often model **counts per unit or counts per interval by a Poisson distribution**. Let Y be the random variable associated with such a count, and let λ be the appropriate expected rate of occurrences.

$$Y \sim Poisson(\lambda)$$

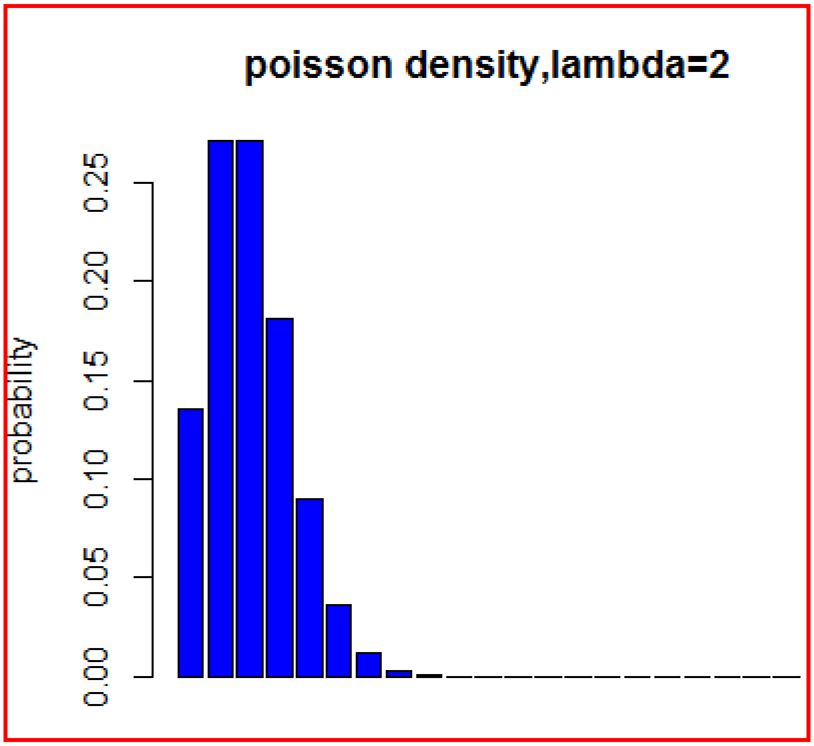
The probability function of Y is

$$f_Y(y) = \frac{\lambda^y}{y!} \exp(-\lambda) \text{ for } y = 0, 1, 2, ... \text{ and } \lambda > 0.$$

If Y follows a Poisson distribution ($Y \sim Poisson(\lambda)$), we can show that:

- $\mu = E(Y) = \lambda$
- $\sigma^2(Y) = \lambda$
- $\sigma(Y) = \sqrt{\lambda}$

3.3.1 Visualization of a Poisson distribution



3.3.2 Example

Example *cars per minute*
There are 12 cars crossing a bridge per minute on average.
Y = number of cars crossing the bridge per minute.
$\lambda = 12$; the average number of cars crossing that bridge per minute.
$Y \sim Poisson(12)$

1. Find the probability of having 17 cars crossing the bridge in a particular minute.

```
dpois(17, lambda = 12)
```

```
## [1] 0.03832471
```

Thus, $P[Y = 17] = 0.038$.

2. Find the probability of having 17 cars or more crossing the bridge in a particular minute.

$P[Y \geq 17] = 1 - P[Y \leq 16]$

```
cumprob <- ppois(16, lambda = 12)
1-cumprob
```

```
## [1] 0.101291
```

Thus, $P[Y \geq 17] = 1 - P[Y \leq 16] = 1 - 0.898 = 0.101$

If there are 12 cars crossing a bridge per minute on average, the probability of having 17 or more cars crossing the bridge in a particular minute is 10.1%.

3.3.3 Use of the Poisson distribution in R

Function in R	Symbolic notation	Description
`dpois(y, lambda = )`	$P(Y = y)$	Poisson probability distribution function
`ppois(y, lambda = )`	$P(Y \leq y)$	Cumulative Poisson distribution function

4 Continuous distributions

Let Y be a random variable. Y is said to be a **continuous random variable** if Y can take any possible real value over some interval.

For any continuous density function, the following property holds:
$\int_{-\infty}^{+\infty} f_Y(x)dx = 1$

For continuous variable, we usually work with **cumulative distribution functions**:

$$F_Y(y_0) = P(Y \leq y_0) = \int_{-\infty}^{y_0} f(x)dx$$

4.1 The normal distribution

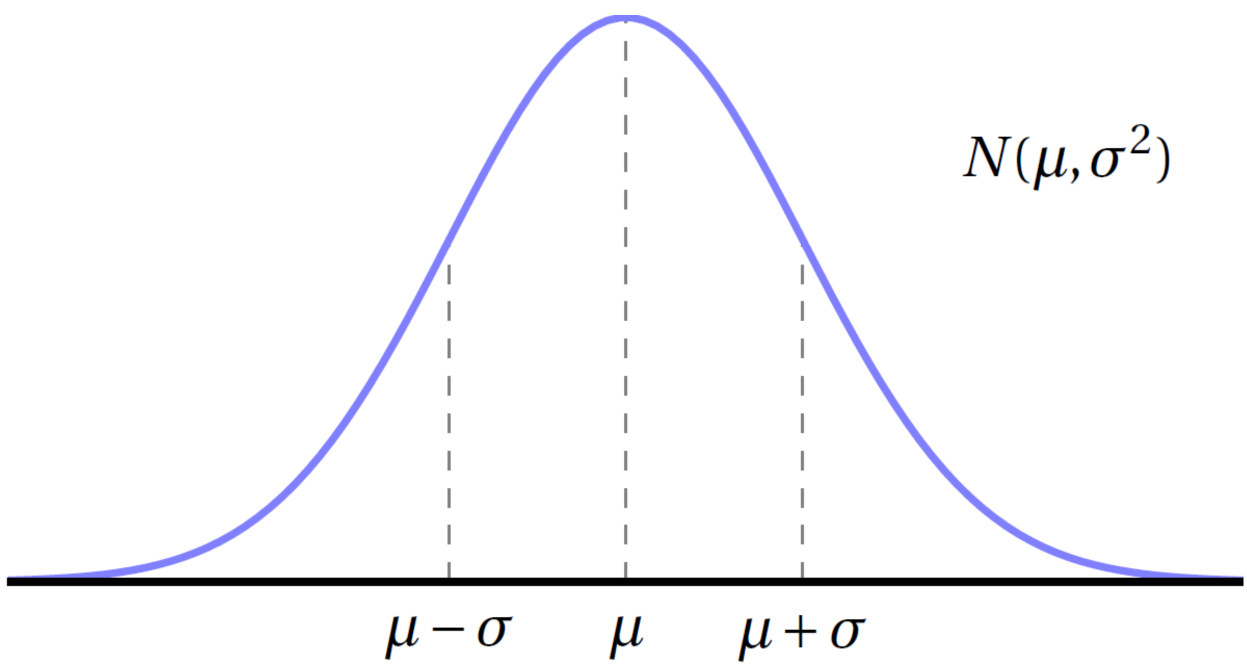
The **normal distribution** is the most frequently used distribution. It is characterized by a bell shaped curve which is symmetric around the mean. The mean, median and mode are equal.

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

The probability density function for the normal distribution is given by

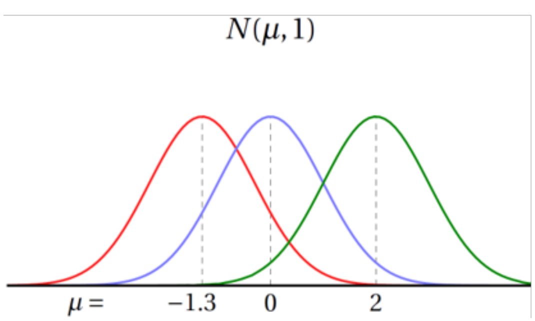
$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{y-\mu}{\sigma})^2] \text{ with } -\infty < y < \infty$$

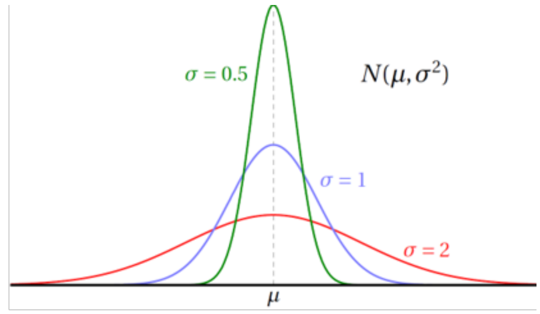
For a normal distribution, $\mu \pm \sigma$ represents the points of inflection for the probability density function.



4.1.1 Interpretation of the parameters

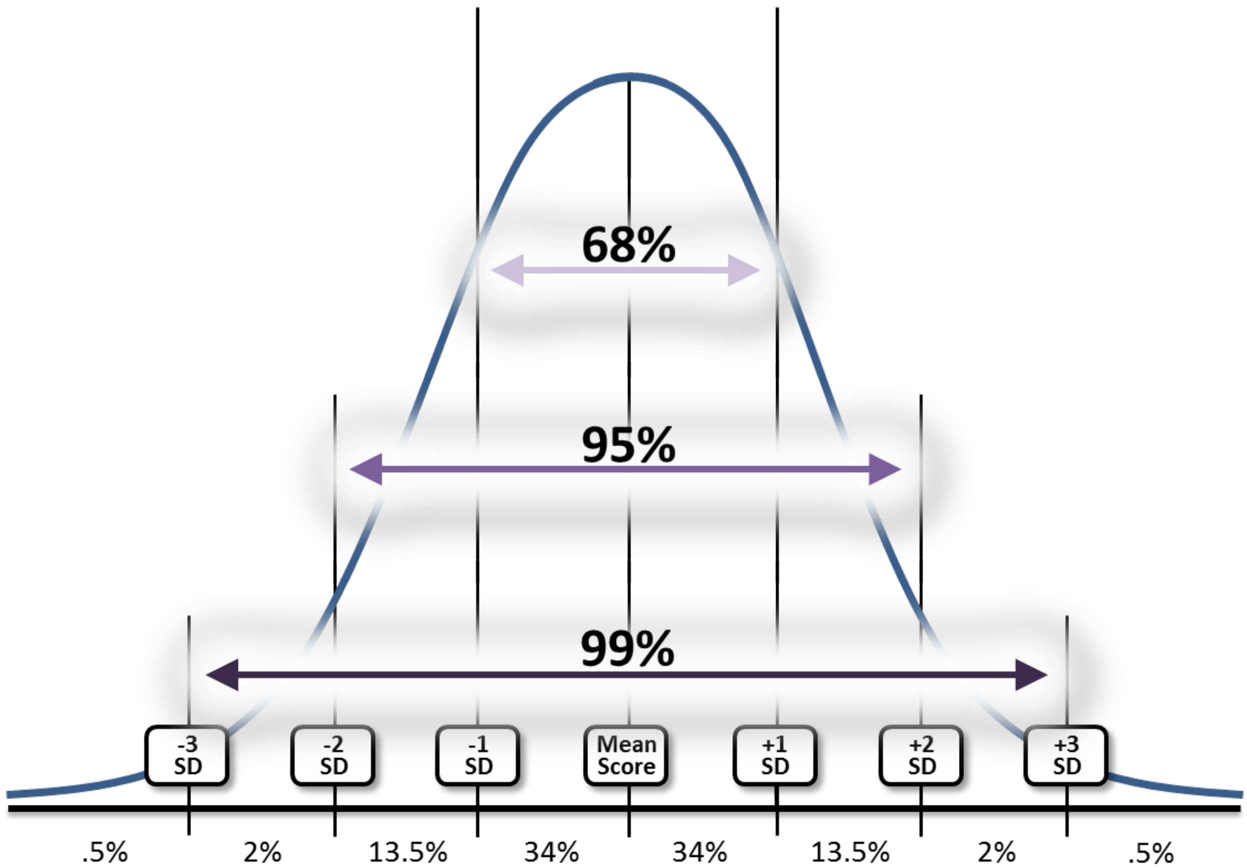
The **mean** μ is the point of symmetry. The **standard deviation** σ controls the spread of the curve.





4.1.2 "68-95-99.7" rule

The "68-95-99.7" rule describes the percentage of values that lie within a band around the mean in a normal distribution. The "68-95-99.7" rule can be broken down into three parts:
$\mu \pm 1\sigma$ includes about 68.3% of the data
$\mu \pm 2\sigma$ includes about 95.4% of the data
$\mu \pm 3\sigma$ includes about 99.7 % of the data



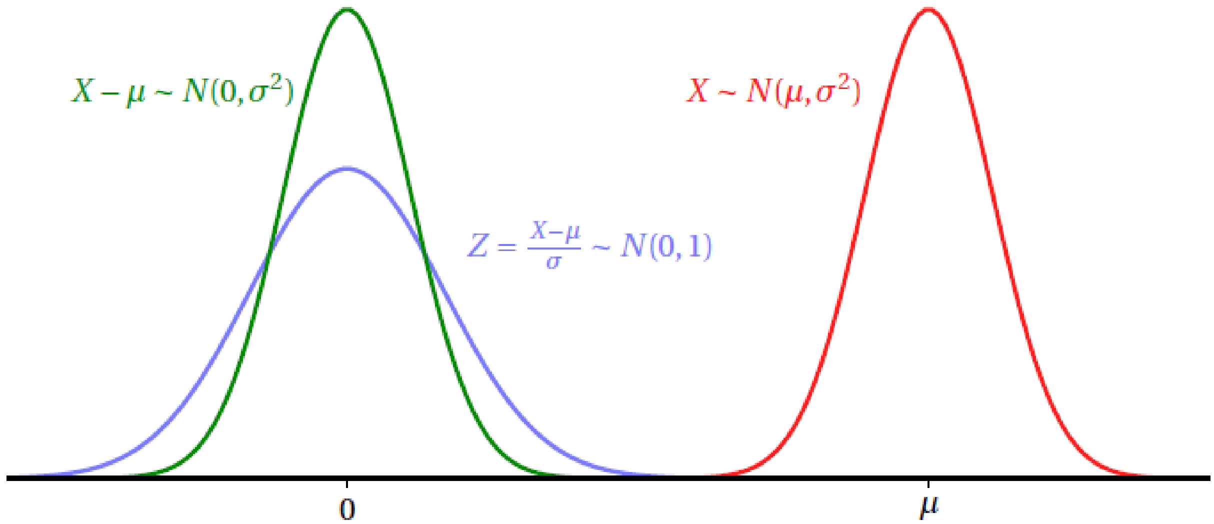
4.1.3 Standard normal distribution

The **standard normal distribution** is a normal distribution with mean of 0 and standard deviation of 1, i.e. $N(0,1)$.

The transformation from $N(\mu,\sigma^2)$ into the standard normal form $N(0,1)$ is obtained as follows:

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

A key property of a normal distribution is that any normal random variable can be standardized.



4.1.4 Use of normal distribution in R

The family name is `norm`

Function in R	Symbolic notation	Description
`dnorm(y, mean = , sd = )`	$P(Y = y)$	Normal probability function
`pnorm(y, mean = , sd = )`	$P(Y \leq y)$	Normal cumulative distribution function

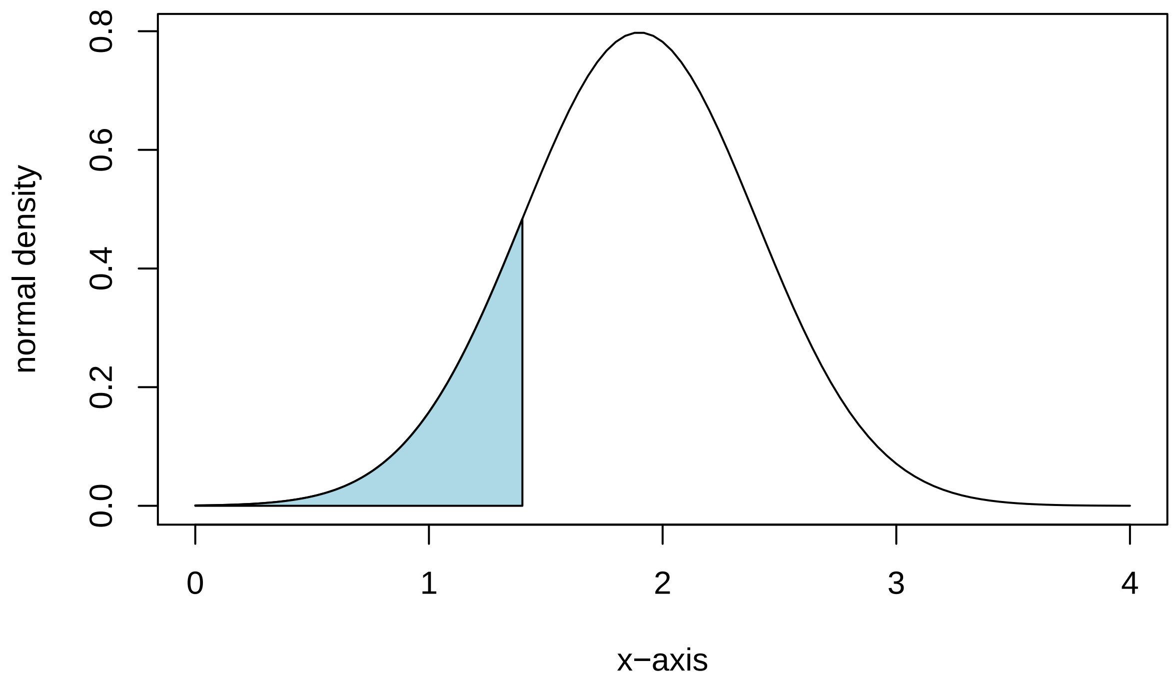
4.1.5 Examples

4.1.5.1 Example *gene expression* Suppose that the expression values of gene CCND3 Cyclin D3 can be represented by $Y \sim N(\mu, \sigma^2)$ with $\mu = 1.9$ and $\sigma = 0.5$.

1. What is the probability that the expression values are less than or equal to 1.4?
   $P[Y \leq 1.4] = 0.1586$

```
pnorm(1.4, 1.9, 0.5)
```

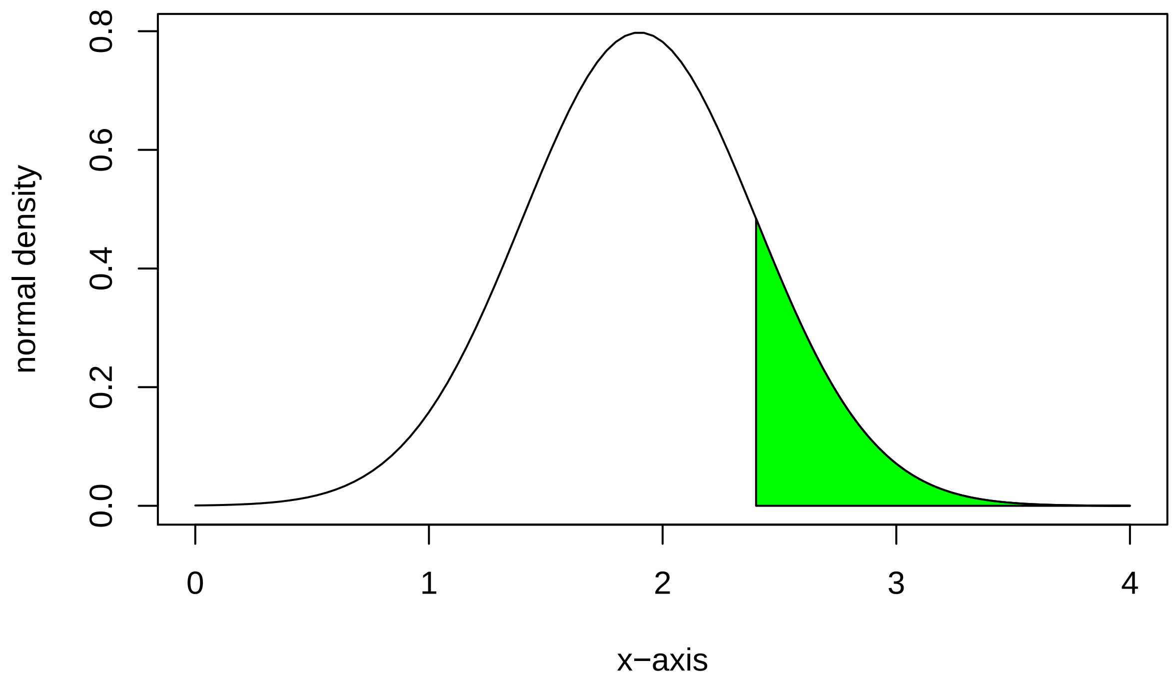
```
## [1] 0.1586553
```



2. What is the probability that the expression values are larger than 2.4?
   $P(Y > 2.4) = 0.1586$

```
1-pnorm(2.4, 1.9, 0.5)
```

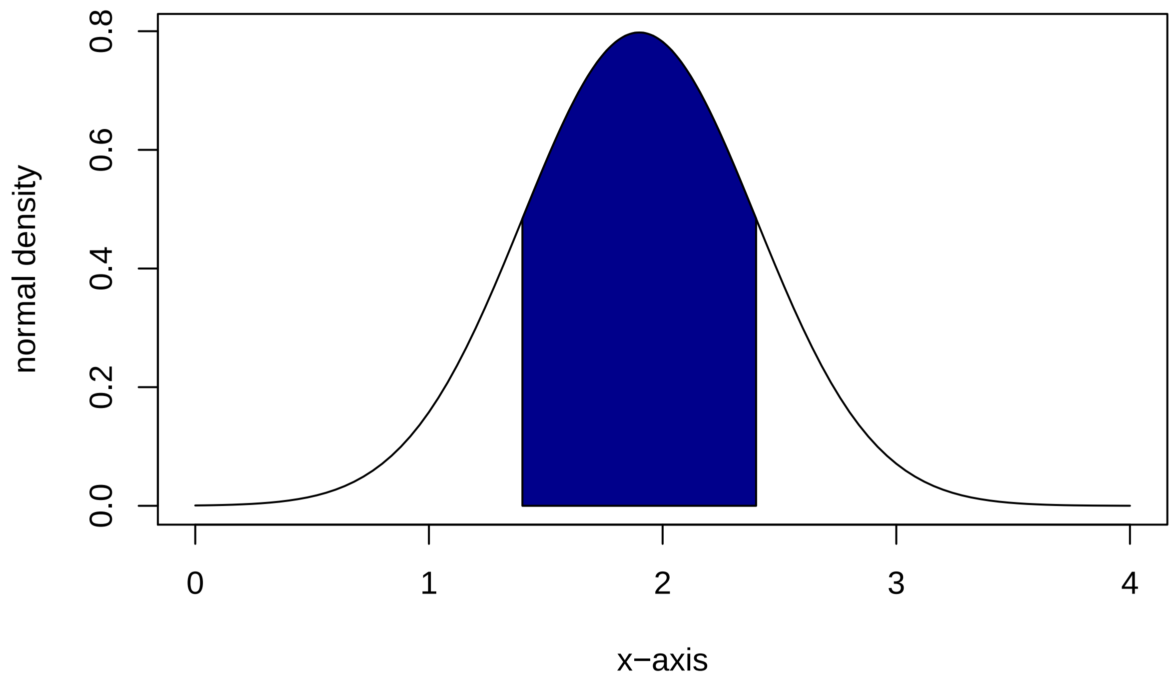
```
## [1] 0.1586553
```



3. What is the probability that the expression values are between 1.4 and 2.4?
   $P(1.4 < Y \leq 2.4) = 0.68$

```
pnorm(2.4, 1.9, 0.5) - pnorm(1.4, 1.9, 0.5)
```

```
## [1] 0.6826895
```



4.1.5.2 Example standard normal distribution $N(0, 1)$

1. How much area is no more than one standard deviation from the mean?
   $P(-1 < Y \le 1)$
   $= P(Y \le 1) - P(Y \le -1)$
   $=$ `pnorm(1, mean = 0, sd = 1) - pnorm(-1, mean = 0, sd = 1)` $= 0.68$
2. How much area is no more than two times the standard deviation from the mean?
   $P(-2 < Y \le 2)$
   $= P(Y \le 2) - P(Y \le -2)$
   $=$ `pnorm(2, mean = 0, sd = 1) - pnorm(-2, mean = 0, sd = 1)` $= 0.95$
3. How much area is no more than three times the standard deviation from the mean?
   $P(-3 < Y \le 3)$
   $= P(Y \le 3) - P(Y \le -3)$
   $=$ `pnorm(3, mean = 0, sd = 1) - pnorm(-3, mean = 0, sd = 1)` $= 0.997$

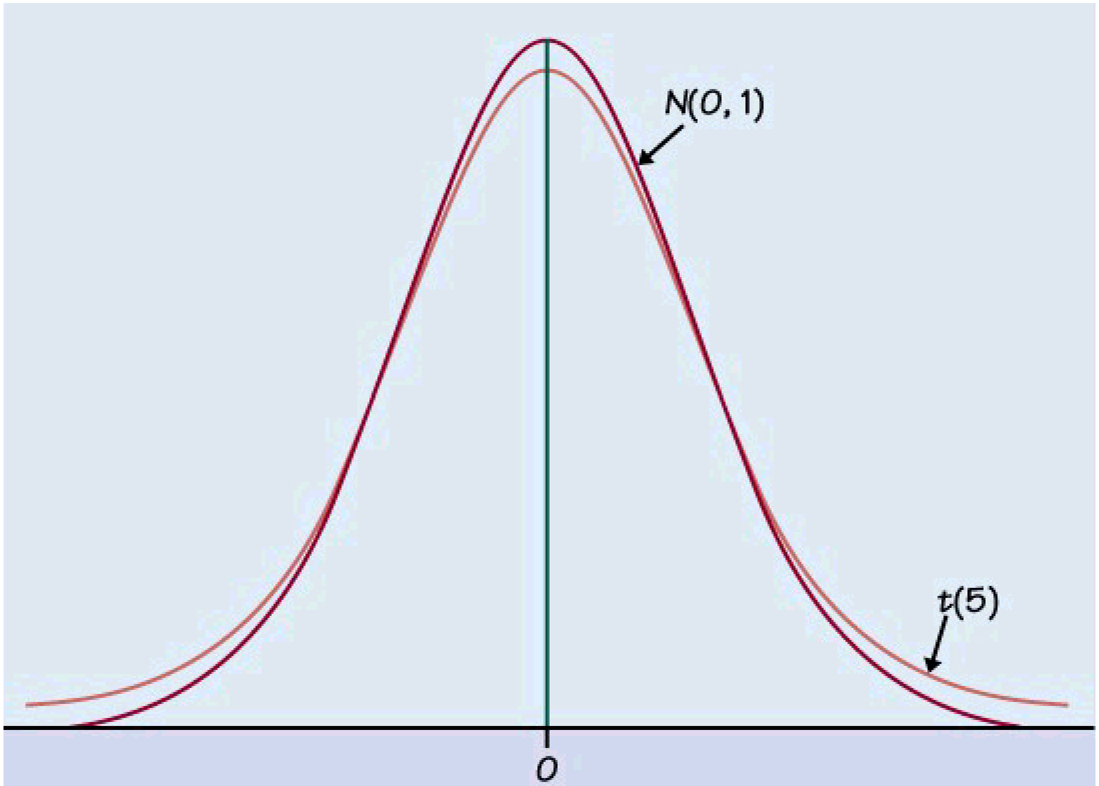
```r
diff(pnorm(c(-3, 3), mean = 0, sd = 1))
```

```
## [1] 0.9973002
```

4.2 The T distribution

The **T distribution** has many useful applications for testing hypotheses about means, in particular when the sample size is small.

It will be discussed later than when the data is normally distributed, then $\sqrt{n}\frac{\overline{X}-\mu}{S} \sim t_{n-1}$



The shape of the t-distribution is similar to the standard normal curve. As the degrees of freedom, k, increase, the $t(k)$ density curve approaches the $N(0, 1)$ curve even more closely, since S approaches σ as n increases.

The density curve of the t-distribution has the following characteristics:

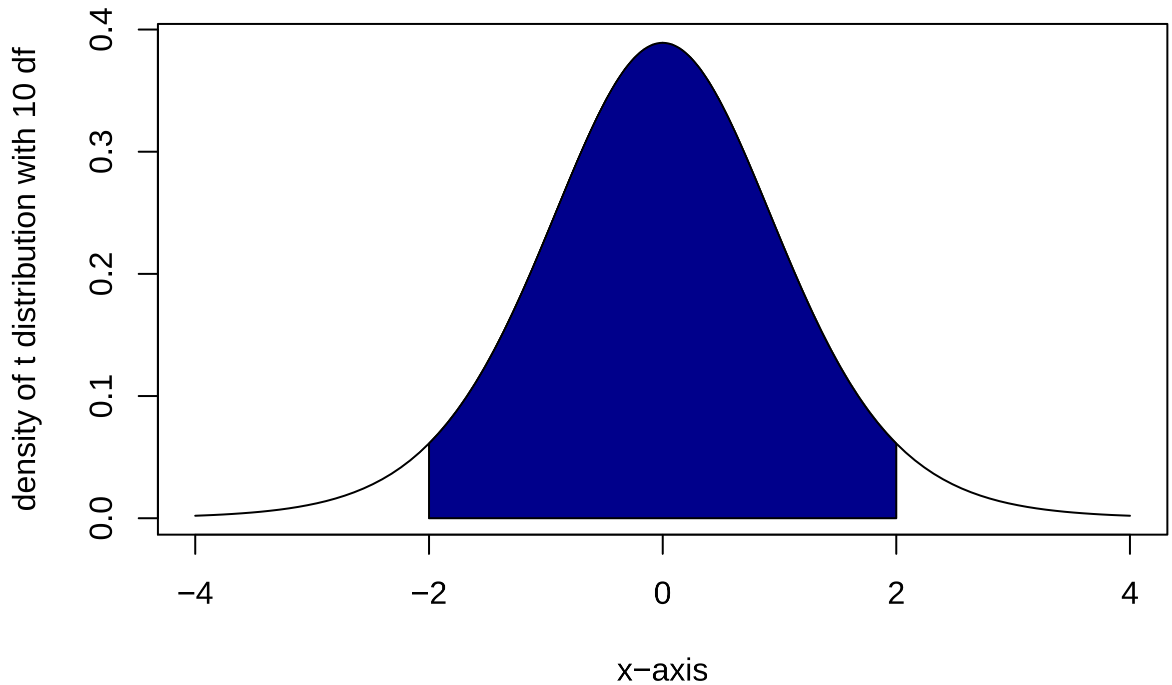
- it is bell shaped
- it is symmetric around 0
- the spread is larger than for the standard normal curve due to extra variability caused by substituting the random variable S for the fixed parameter σ
- It has heavier tails than the density curve of the standard normal distribution.

Example
What is the probability that a random variable $Y \sim t_{10}$ takes values between -2 and $+2$?

$Y \sim t_{10}$ $P(-2 < Y \leq 2) = 0.92$

```r
pt(2, 10) - pt(-2, 10)
```

```
## [1] 0.926612
```



Remark:
Note that this area is smaller than 0.95!

4.3 The F distribution

The **F distribution** is important for testing the equality of two variances. It can be shown that the ratio of variances from two independent sets of normally distributed random variables follows an F distribution.

It will be discussed later that if two population variances are equal ($\sigma_1^2 = \sigma_2^2$) then $\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}$.

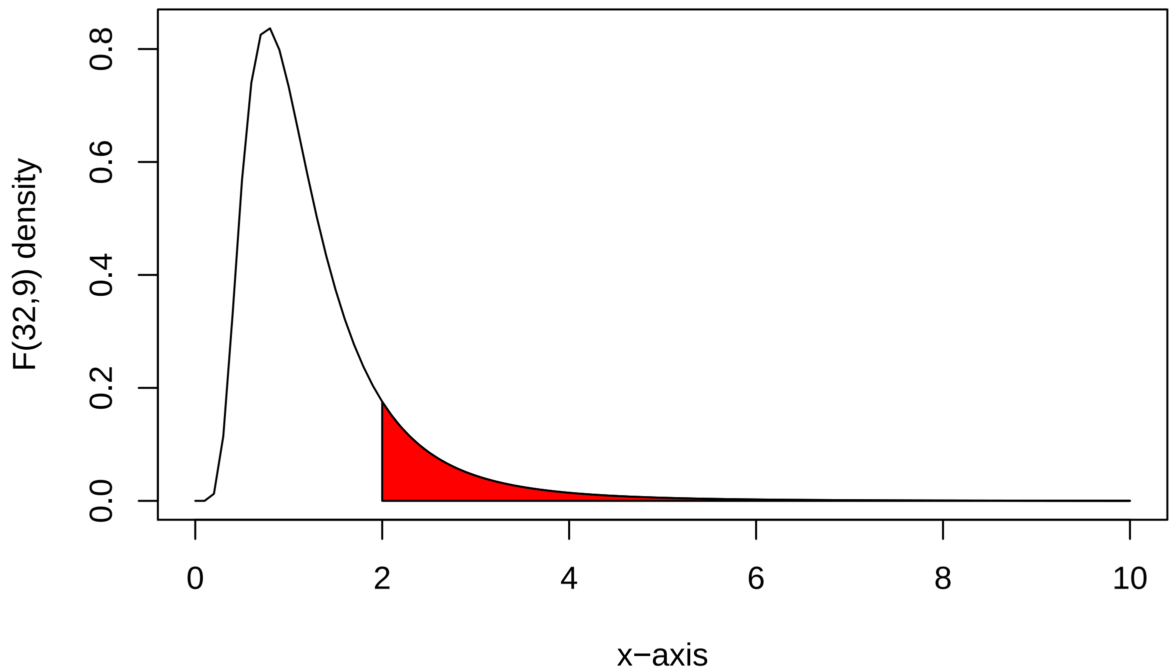
S_1^2 and S_2^2 are the sample variances of the first and second set (with corresponding sample size n_1 and n_2).

Properties of the F distribution
The F-distribution is not symmetric, but is right-skewed.
Because sample variances cannot be negative, the F-statistic takes only positive values.

Example
Let's consider the F distribution with 32 and 9 degrees of freedom. Compute the probability of values larger than 2.
$P(F_{32,9} > 2) = 0.137$

```
1-pf(2, 32, 9)
```

```
## [1] 0.1366314
```



4.4 The chi-square distribution

The chi-square distribution plays an important role in testing hypotheses about frequencies.
Like the t-distribution, the χ^2 distributions form a family described by a single parameter, the number of degrees of freedom (df).

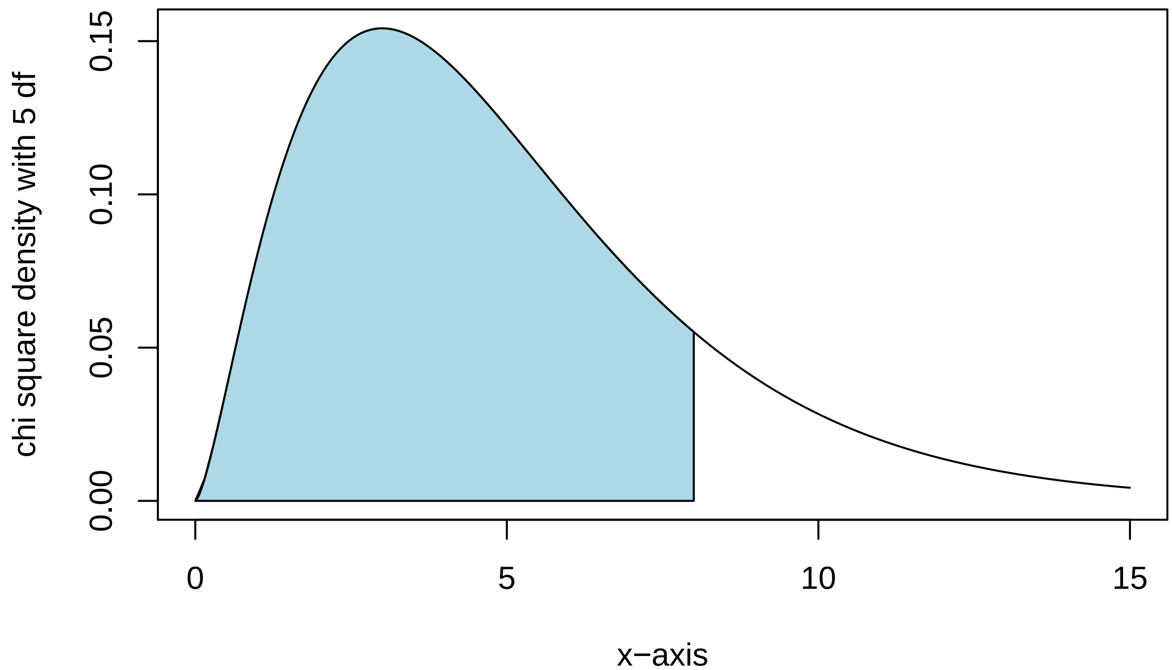
We use χ^2_{df} to indicate a particular member of this family.
The χ^2 distributions take only positive values and are skewed to the right.

Example
Let's consider a random variable Y with a chi-square distribution with 5 degrees of freedom: $Y \sim \chi^2_5$
Compute the probability of values smaller than or equal to 8.

$P(Y \leq 8) = 0.84$

```
pchisq(8,5)
```

```
## [1] 0.8437644
```



4.5 Remarks

With all the above distributions:

- The associated random variables are continuous random variables
- The density curves of the distributions are smooth curves
- Areas under the density curve represent probabilities
- Given the value of the statistic, use R to calculate the respective probability

5 The Central Limit Theorem

5.1 Introduction

Example
Suppose X= expenditure of a customer on Saturday morning in a specific grocery in Phnom Penh in 2013. Assume that $X \sim N(\mu = \$50, \sigma^2)$ (this is an assumption about the population distribution).

Each Saturday morning, we go to this grocery and take a sample of 30 customers at random ($n = 30$). We compute for each sample the average expenditures of these 30 customers (expressed in \$).

Week of 2013	$\overline{X}$
1	$\overline{x_1} = 48.2$
2	$\overline{x_1} = 51.8$
...	...
52	$\overline{x_{52}} = ...$

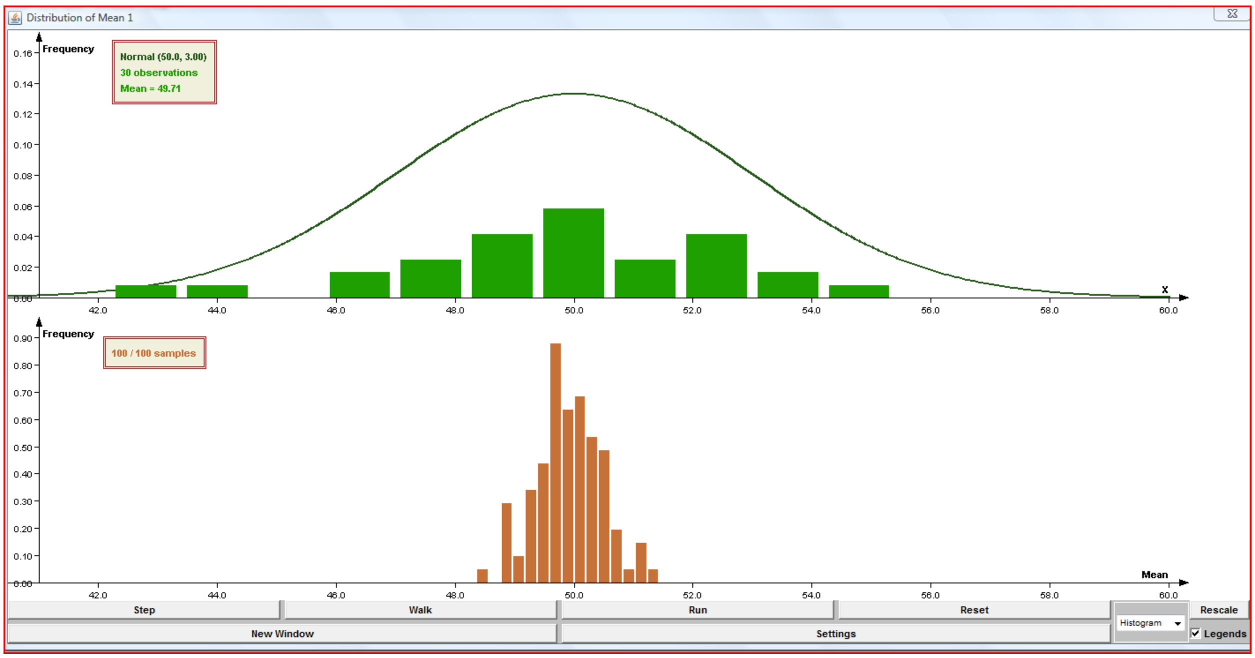
Questions of interest:

1. If we average these 52 average expenditures. What do you expect? They will be centered around ...

2. What do you expect about the variability of these average expenditures? The standard deviation of these averages will be
   a. the same as in the population ($= \sigma$).
   b. smaller than in the population ($< \sigma$).
   c. larger than in the population ($> \sigma$).
3. What is the distribution of $\overline{X}$?

Take a look at http://lstat.kuleuven.be/java/index.htm. Here you can find some JAVA applets for the visualization of statistical concepts.

By selecting *BASICS* → *Distribution of Mean (Continuous Distributions)*, and running the applet, you can find the answers on these questions.



Hence:

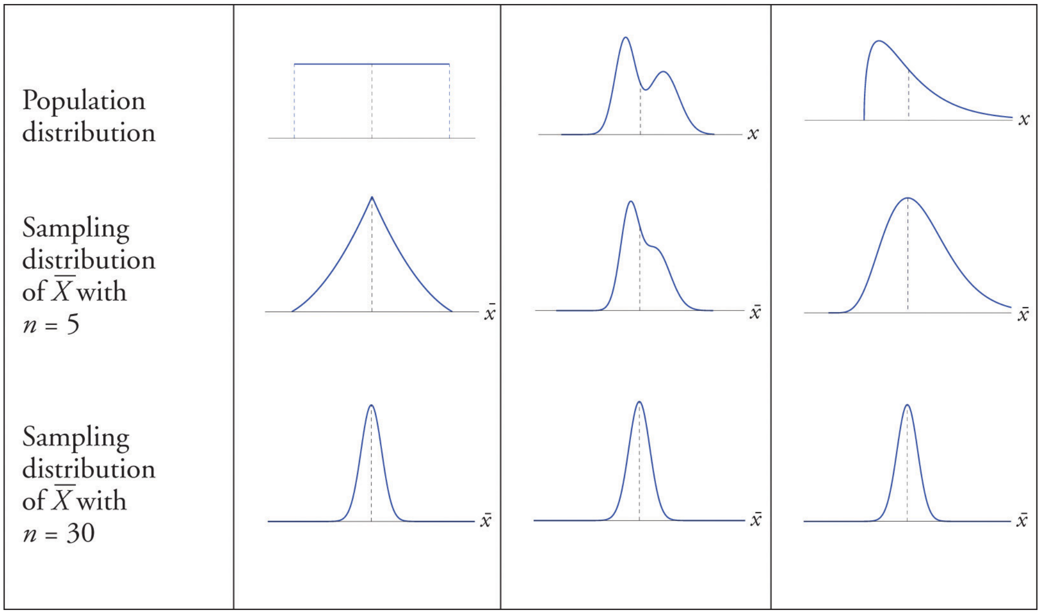
> If $X_1, X_2, ..., X_n$ are independent and identically distributed ($X_i \sim N(\mu, \sigma^2)$ for $i =, ..., n$), then $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

Question of interest:
What if the condition $X_1, X_2, ..., X_n$ *are independent and identically distributed (i.i.d.)* $N(\mu, \sigma^2)$ does not hold?

5.2 The central limit theorem (CLT)

Consider a series of random samples all containing n observations, each from an underlying population with mean μ and variance σ^2. If n is sufficiently large, then the sampling distribution of $\overline{X}$ is approximately normal with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$

> **Whatever the population from which we sample looks like, the distribution of the sample mean $\overline{X}$ approaches, as n tends to infinity, a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$**



This figure shows the sampling distribution of $\overline{X}$ contrasted with the parent population distribution. The last row show that even though the population is not normal distributed, the sampling distribution of $\overline{X}$ still becomes approximately normal.

Some consequences:

- The normal distribution is very important
- $\overline{X}$ can be used to estimate μ. The larger n, the better the estimate
- Replication increases precision (precision is inversely related to the width of the distribution of the estimator). The standard deviation of the estimator is called *standard error*.

  If n is sufficiently large, then $\frac{\overline{X}-\mu}{\sigma_{\overline{X}}} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ follows a standard normal distribution $N(0, 1)$.

5.3 What is sufficiently large?

- If the underlying population follows a normal distribution, then $n = 1$ is sufficiently large.
- If the underlying population is symmetric and single-peaked and the tails die out rapidly, then n around 5 is sufficiently large.
- Most textbooks suggest that whenever $n \geq 25$ we can use the CLT.

Example

The time it takes to check out at a grocery store varies widely. A certain checker has a historic average of one minute service time per customer, with a one minute standard deviation. If she sees 40 customers, what is the probability that her average check out time is 0.9 minutes or less.

Solution:
We assume that the each service time has an unspecified parent population with $\mu = 1$ and $\sigma = 1$ and that the sequence of service times is i.i.d. As well, we assume that n is large enough so that the distribution of $\overline{X}$ is approximately $N(1, \frac{1}{40})$.
Then $P(\overline{X} \leq 0.9) = 0.26$ is given by `pnorm(0.9, mean = 1, sd = 1/sqrt(40))`

6 Testing for normality

The idea of hypothesis testing will be dealt with in the next chapter.

6.1 Kolmogorov-Smirnov test

One possibility to check for normality is using the **Kolmogorov-Smirnov test** (which will not be used in this course). The Kolmogorov-Smirnov test uses the statistic

$D_n = \max_x |\hat{F}_n(x) - F_0(x)|$

$F_0(x)$ is the assumed population distribution (here the normal cumulative distribution function).
$\hat{F}_n(x)$ is the empirical distribution function and provides an approximation to the theoretical distribution function of the population.
$\hat{F}_n(x) = \frac{1}{n}$ (the number of data not larger than x)
$(\hat{F}_n(x_{(i)}) = \frac{i}{n})$

6.2 Shapiro-Wilk test

The **Shapiro-Wilk test** is based on the correlation coefficient computed from the normal quantile plot. We will use the Shapiro-Wilk test in the chapter on hypothesis testing.

Set r_i be the normal quantiles, then the correlation coefficient is calculated by

$r_Q = \frac{\sum(r_i - \overline{r})(x_{(i)} - \overline{x})}{\sqrt{\sum(x_{(i)} - \overline{x})^2 \sum(r_i - \overline{r})^2}}$

How to perform a test for normality in R and how to interpret the result will be seen in the chapter on hypothesis testing.