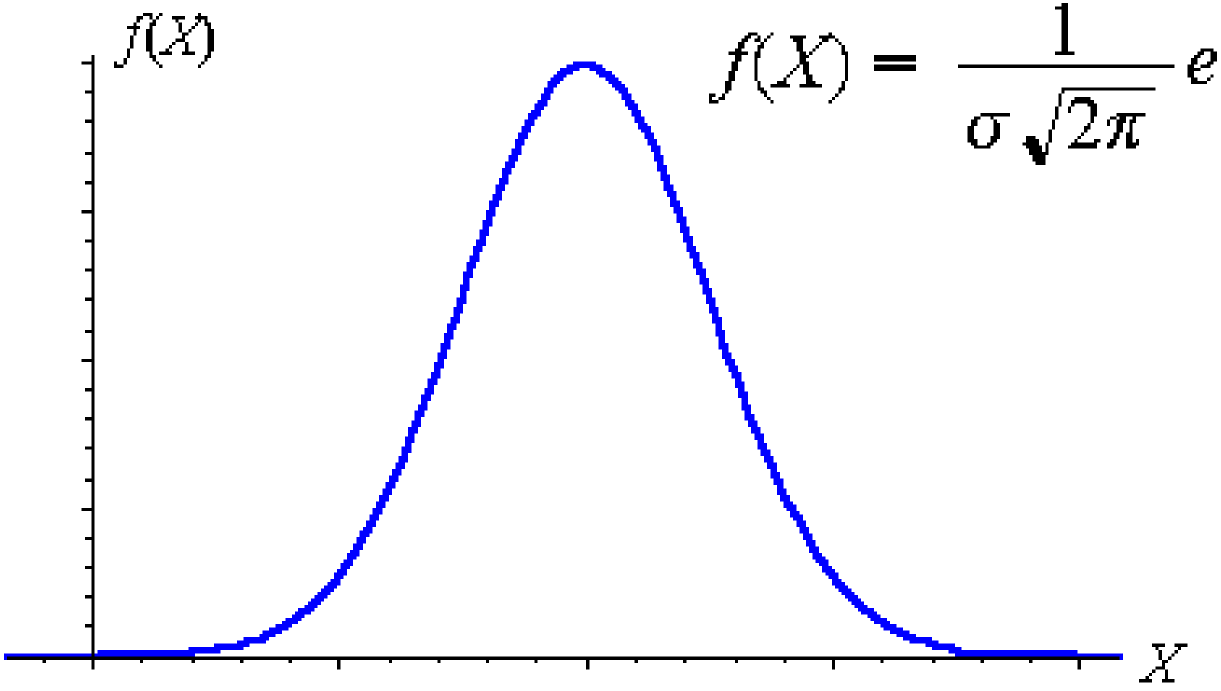
Probability, Distributions, and Hypothesis Test



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

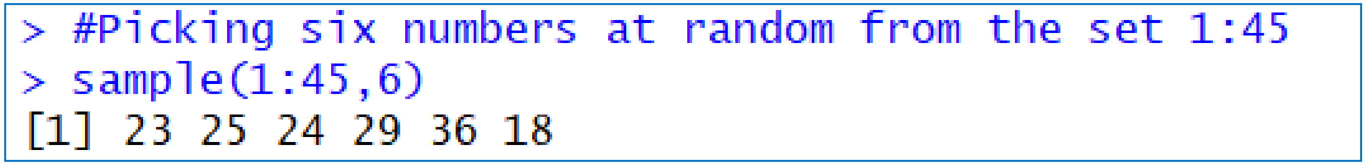
Normal Curve

1. Basic Rules of Probability
2. Probability Distributions
3. The Normal Curve
4. Standard Scores and the Normal Curve
5. Percentile Ranks from the Normal Curve
6. Confidence Intervals
7. Hypothesis Testing

1. Random Sampling

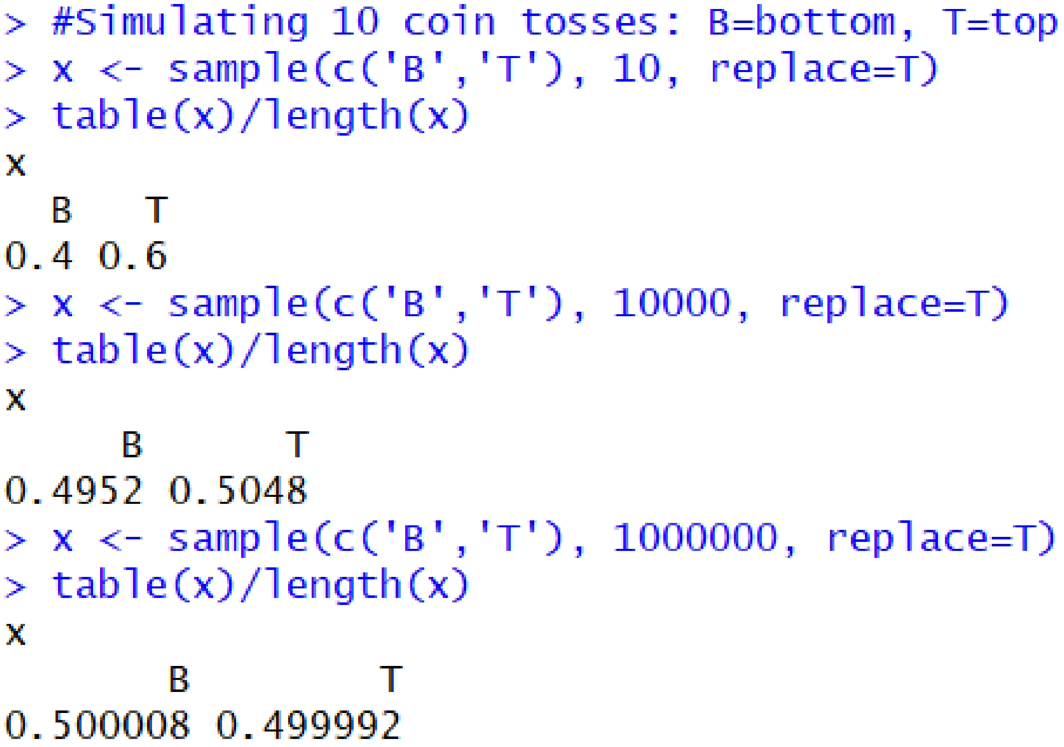
In R, we can simulate dealing from a well-shuffled pack of cards or picking numbered balls from a well-stirred urn.

sample(x, size, replace=FALSE, prob=NULL)
It takes a sample of the specified size from the elements of x using either with or without replacement.



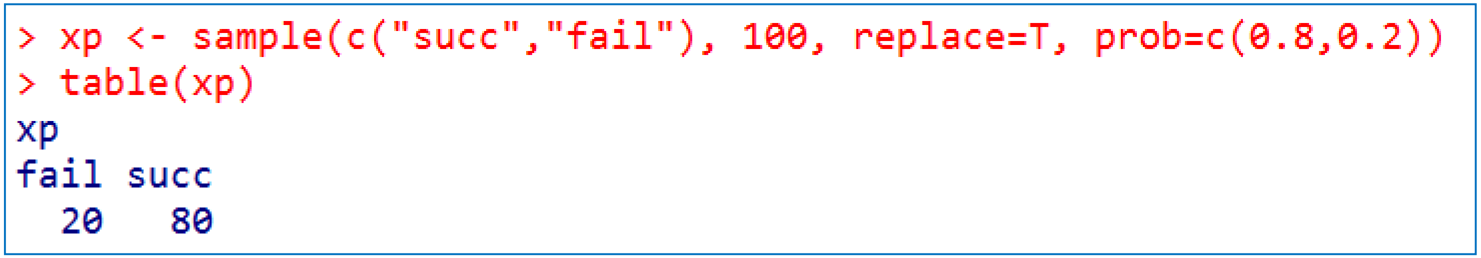
```
> #Picking six numbers at random from the set 1:45
> sample(1:45,6)
[1] 23 25 24 29 36 18
```

Sampling with replacement is suitable for modelling coin tosses.



```
> #Simulating 10 coin tosses: B=bottom, T=top
> x <- sample(c('B','T'), 10, replace=T)
> table(x)/length(x)
x
  B   T 
0.4 0.6 
> x <- sample(c('B','T'), 10000, replace=T)
> table(x)/length(x)
x
     B      T 
0.4952 0.5048 
> x <- sample(c('B','T'), 1000000, replace=T)
> table(x)/length(x)
x
       B        T 
0.500008 0.499992 
```

You can simulate data with nonequal probabilities for the outcomes (say, a 80% chance of success) by using the prob argument to sample, as in



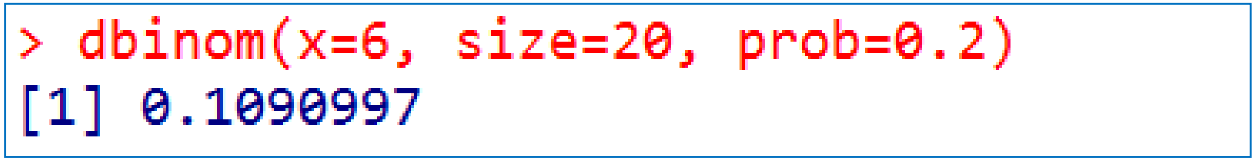
```
> xp <- sample(c("succ","fail"), 100, replace=T, prob=c(0.8,0.2))
> table(xp)
xp
fail succ
  20   80
```

dbinom(x, size, prob, ...)

Density for the binomial distribution with parameters size and prob.

This is conventionally interpreted as the number of 'successes' in size trials.

Suppose there are 20 multiple choice questions in an science class quiz. Each question has five possible answers, and only one of them is correct. The probability of having exactly six correct answers if a student attempts to answer every question at random as follows.



```
> dbinom(x=6, size=20, prob=0.2)
[1] 0.1090997
```

rbinom(n,size,prob) # *n* random generation for the binomial distribution with parameters *size* and *prob*.

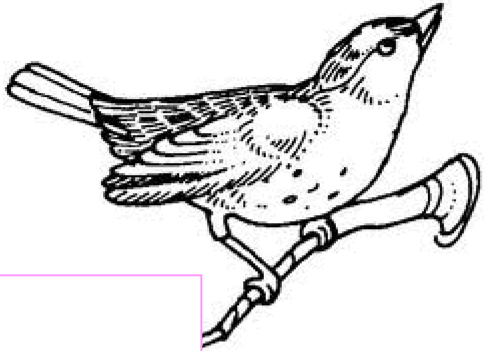
2. Normal Distribution

► Probability density function of the normal distribution:

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

where

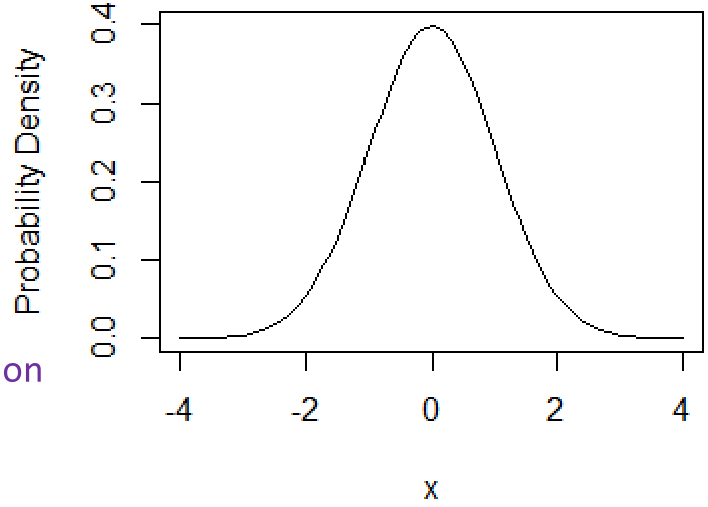
- $\sigma > 0$ is the standard deviation
- μ is the exepected value



dnorm(x, …)

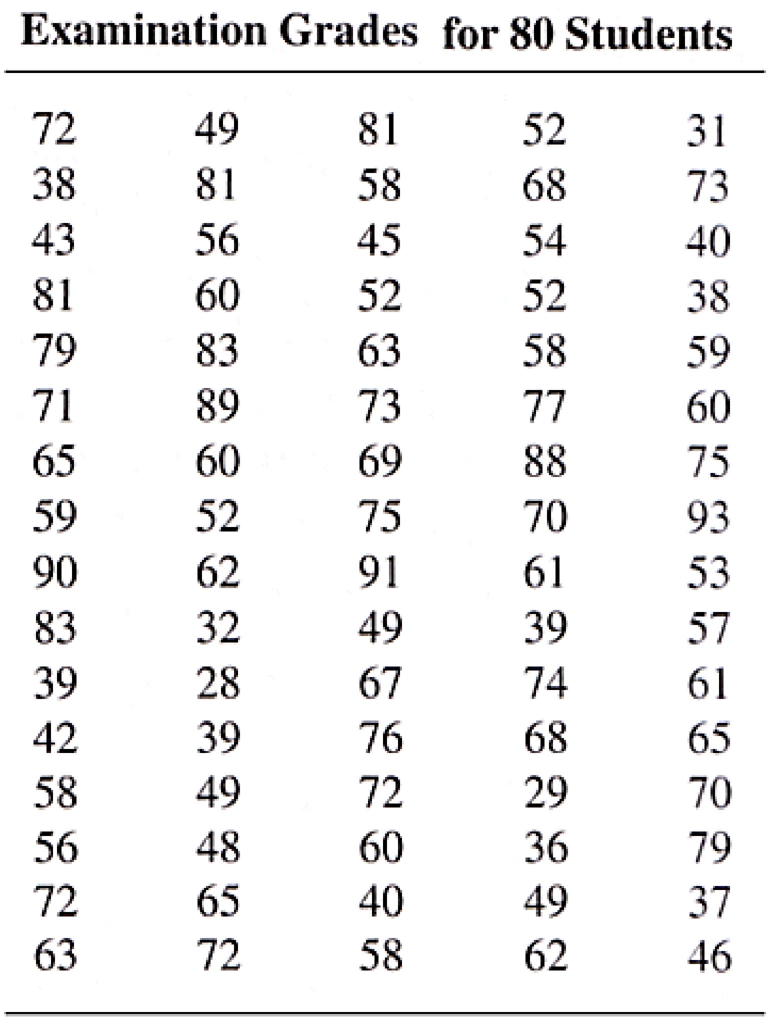
Density distribution function for the normal distribution.

```
# A simple normal curve #
x1=-4; x2=4
x <- seq(x1, x2, 0.1); y <- dnorm(x)
plot(x, y, type='l', xlim=c(x1,x2), ylim=c(0,0.4),
 xlab="x", ylab="Probability Density")
```



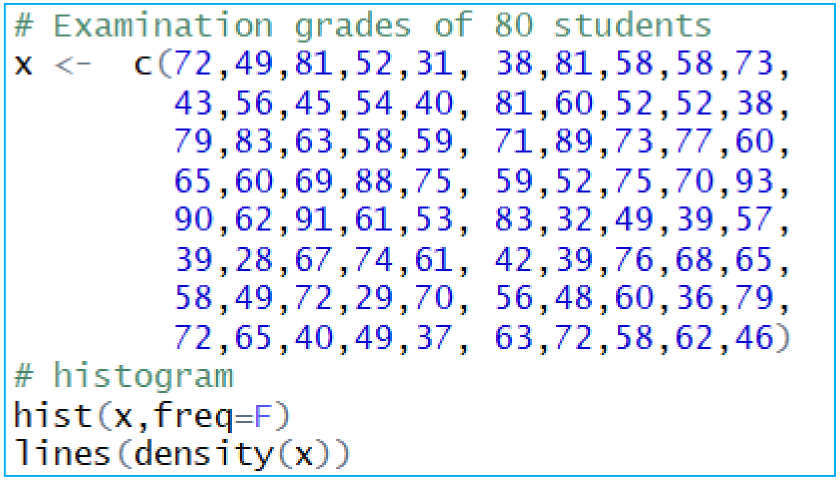
#**dnorm** gives the density and **pnorm** gives the distribution

Normality test is used to determine if a data set is well-modeled by a normal distribution. See example below on examination grades of 80 students and determine whether the test scores follow a normal distribution.

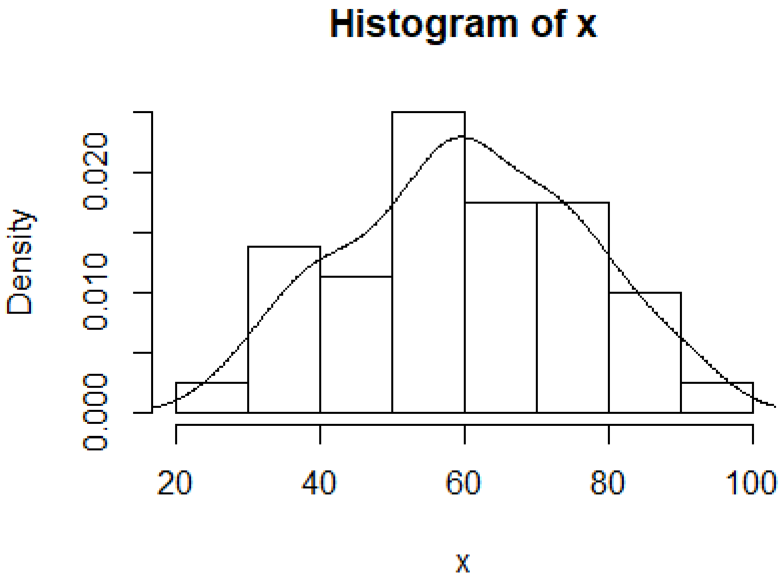


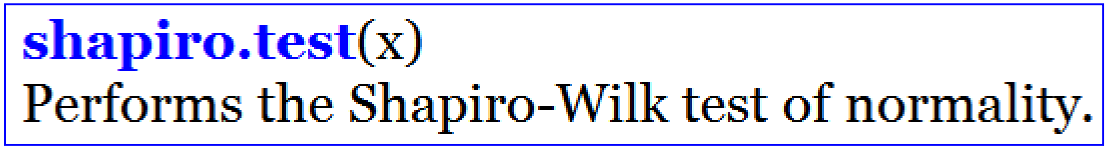
Examination Grades for 80 Students

72	49	81	52	31
38	81	58	68	73
43	56	45	54	40
81	60	52	52	38
79	83	63	58	59
71	89	73	77	60
65	60	69	88	75
59	52	75	70	93
90	62	91	61	53
83	32	49	39	57
39	28	67	74	61
42	39	76	68	65
58	49	72	29	70
56	48	60	36	79
72	65	40	49	37
63	72	58	62	46

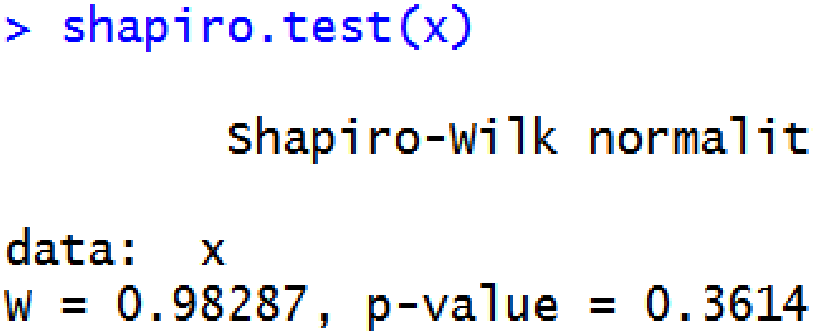


```
# Examination grades of 80 students
x <-  c(72,49,81,52,31, 38,81,58,58,73,
        43,56,45,54,40, 81,60,52,52,38,
        79,83,63,58,59, 71,89,73,77,60,
        65,60,69,88,75, 59,52,75,70,93,
        90,62,91,61,53, 83,32,49,39,57,
        39,28,67,74,61, 42,39,76,68,65,
        58,49,72,29,70, 56,48,60,36,79,
        72,65,40,49,37, 63,72,58,62,46)
# histogram
hist(x,freq=F)
lines(density(x))
```





shapiro.test(x)
Performs the Shapiro-Wilk test of normality.

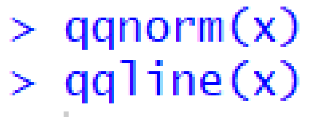


```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.98287, p-value = 0.3614
```

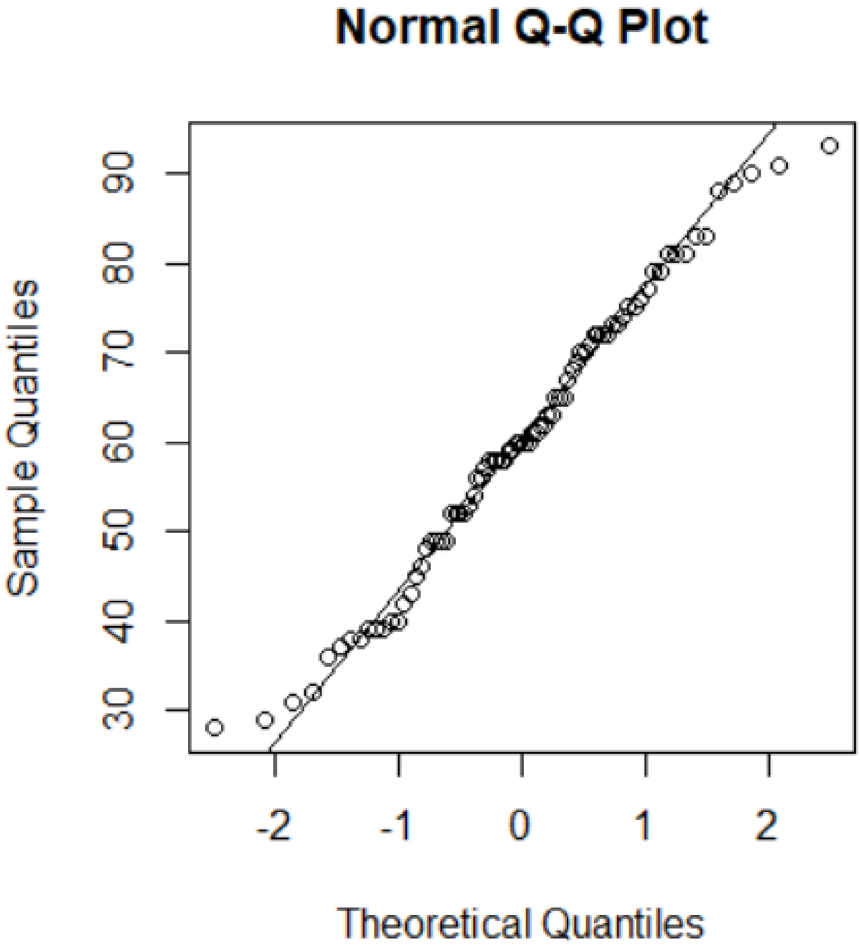
The null-hypothesis (Ho) of this test is that the population is normally distributed.

The p-value is greater than **0.05** implying that we **fail to reject** the null hypothesis that the sample comes from a population which has a normal distribution.

P-value = 0.3614 indicates that we have a 36.14% probability of seeing the same kind of data from this process.



```
> qqnorm(x)
> qqline(x)
```

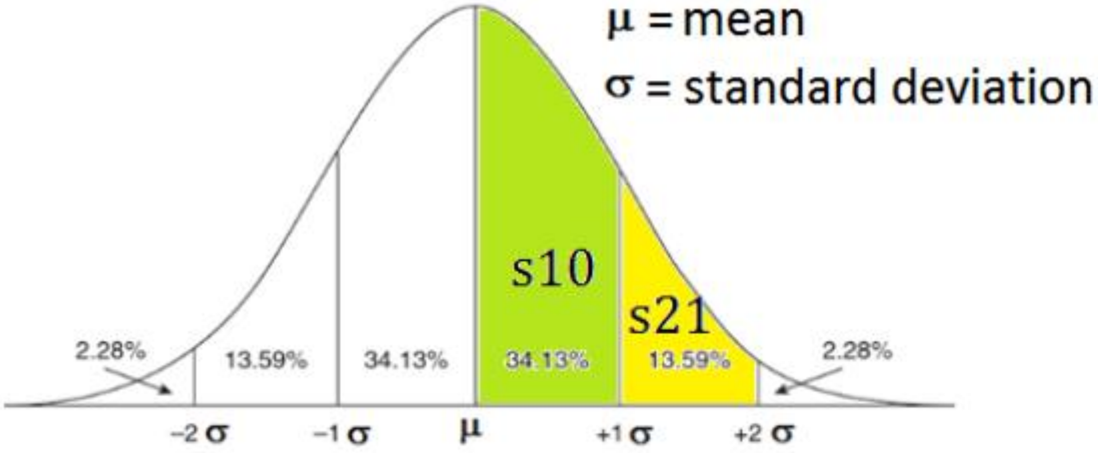


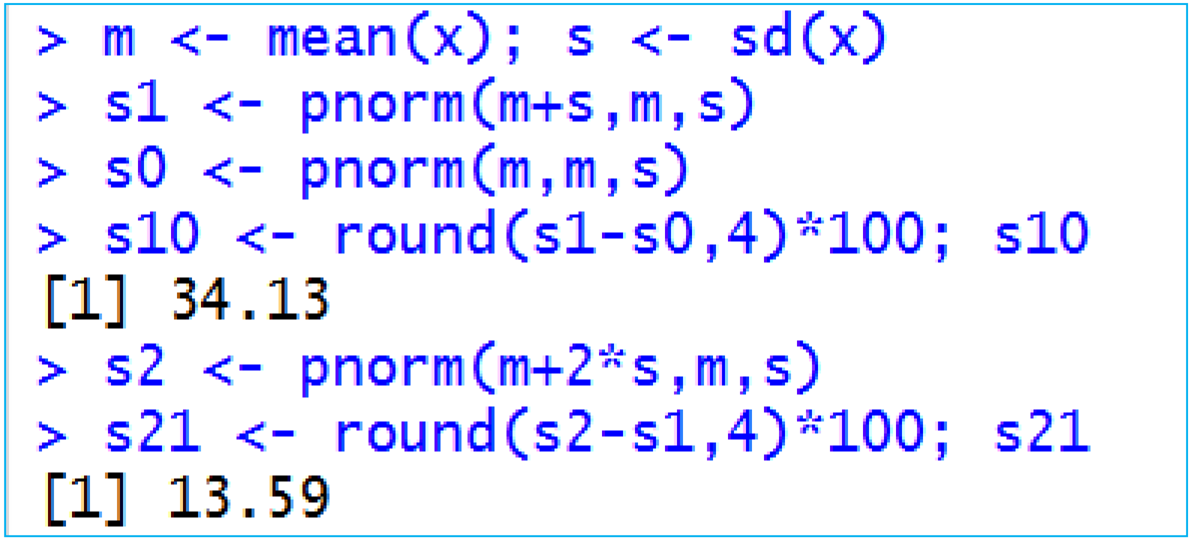
qqnorm compares a sample with a theoretical sample that comes from a certain distribution – in this case, the normal distribution.

qqline adds a line for visual inspection. The closer all points lie to the line, the closer the distribution of your sample comes to the normal distribution.

▪ Area Under the Normal Curve

pnorm(q, …)
Area (probability) under the normal curve to the left of q





```
> m <- mean(x); s <- sd(x)
> s1 <- pnorm(m+s,m,s)
> s0 <- pnorm(m,m,s)
> s10 <- round(s1-s0,4)*100; s10
[1] 34.13
> s2 <- pnorm(m+2*s,m,s)
> s21 <- round(s2-s1,4)*100; s21
[1] 13.59
```

```
#dnorm gives the density and pnorm gives the distribution
```

3. Sampling Distribution of Means

The social researcher’s methods of sampling:

Sampling members are **representative enough of the entire population** to permit making accurate generalizations about that population.

The symbols for the descriptive terms:

	Mean	Standard Deviation	Variance
Population	μ	σ	σ^2
Sample	$\bar{x}$	s	s^2

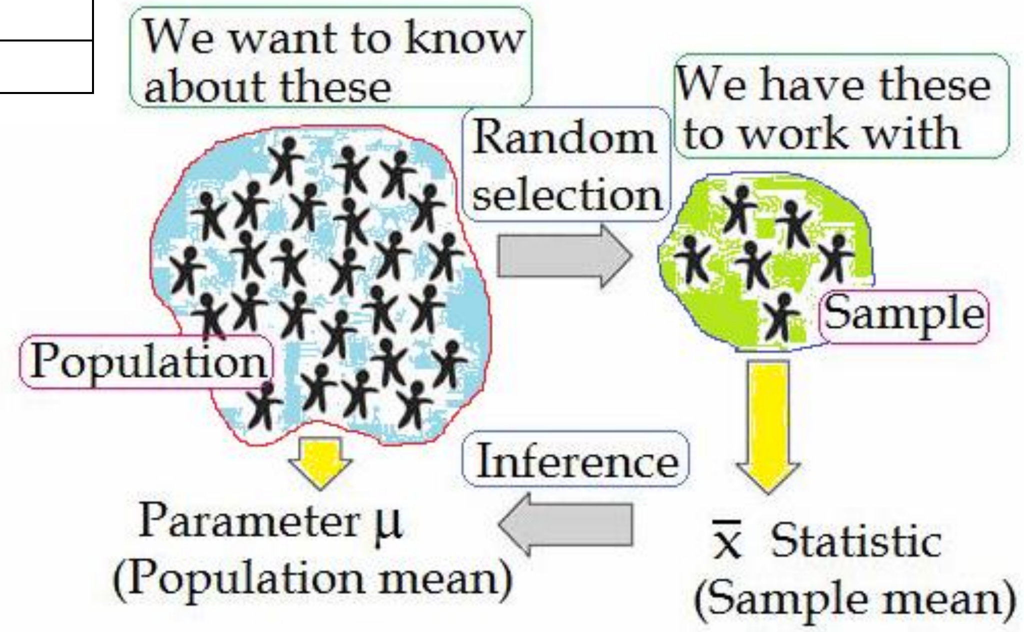


Figure: Illustration of the relationship between samples and populations.

Generalization from a Sample to a Population

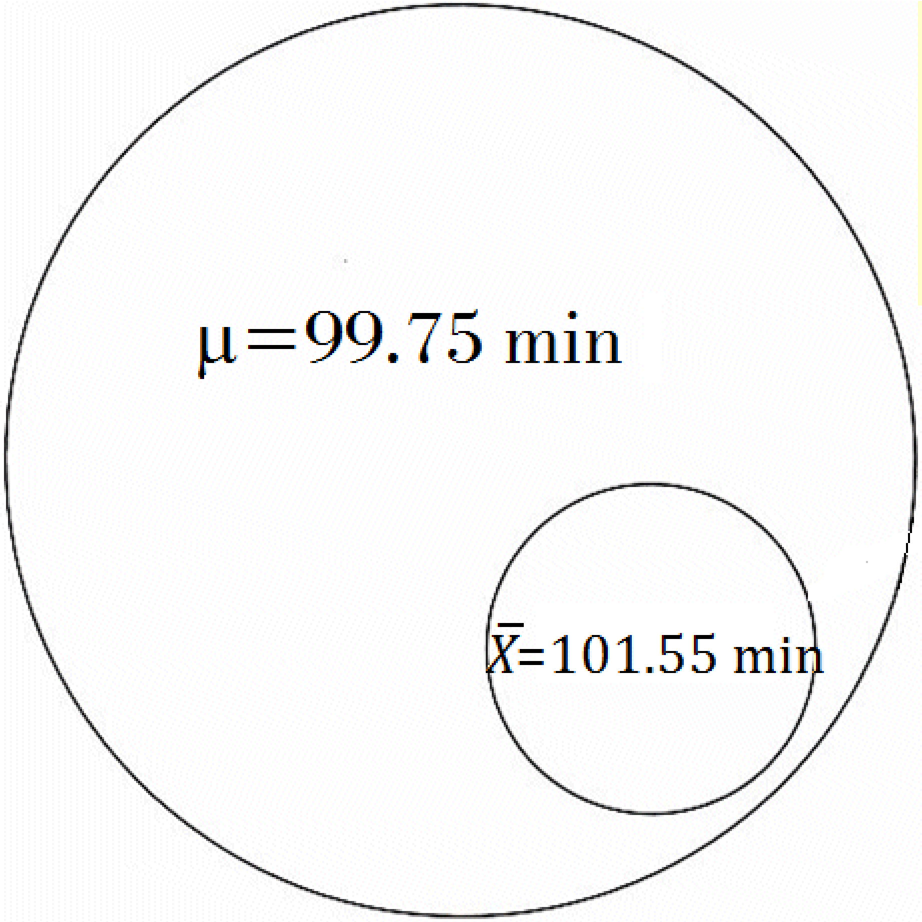


FIGURE: Mean Long-Distance Phone Time for a Random Sample Taken from the Hypothetical Population L&F p186

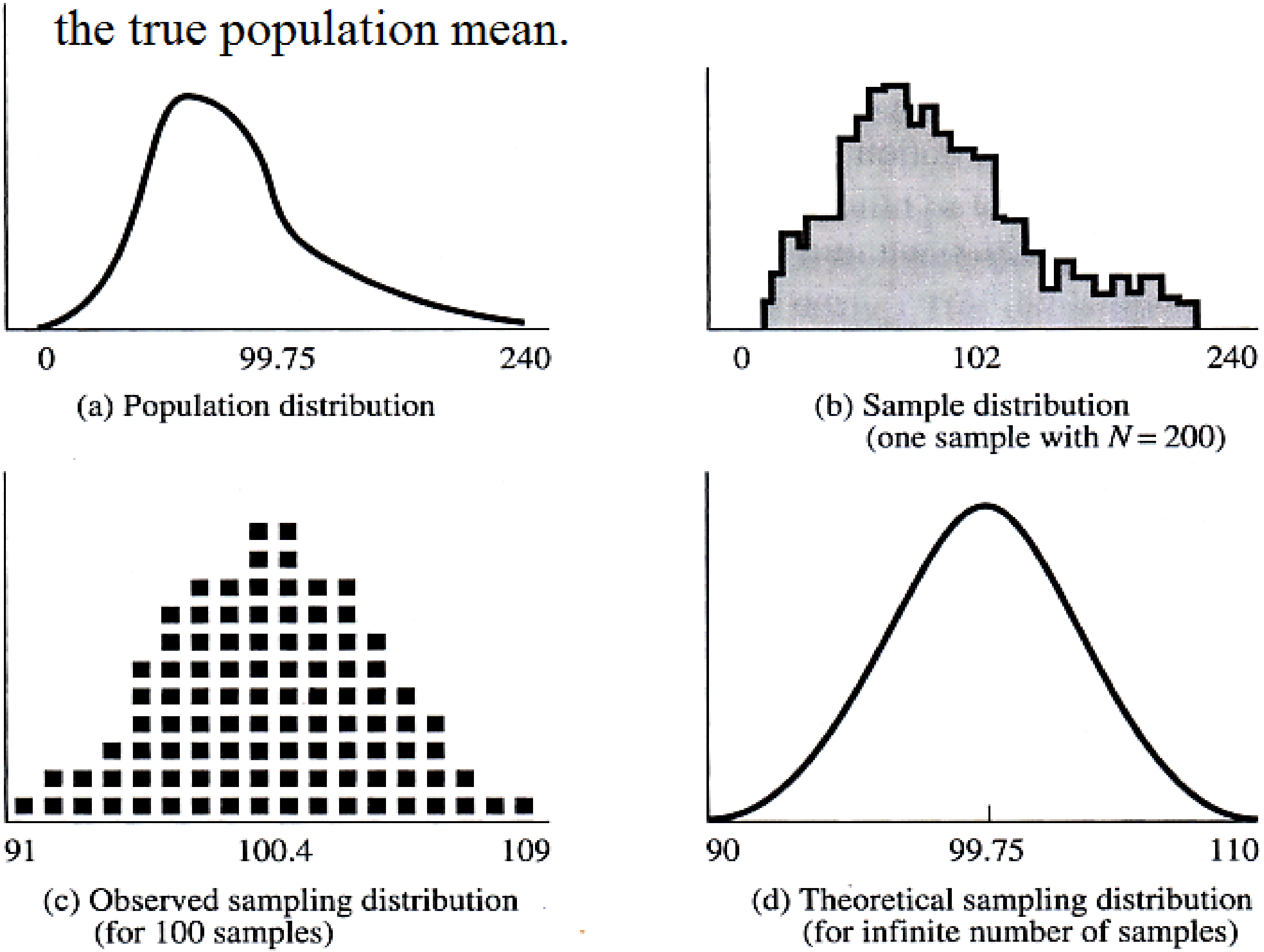
- Assume that a hypothetical sociologist monitors the long-distance calls of a sample of 200 households taken at random from the entire population.

- Sample range: 0 ~ 240 min.
- Mean: 101.55 min.

Characteristics of a Sampling Distribution of Means

(1) The sampling distribution of means will be approximately normally distributed.

(2) The mean of a sampling distribution of means is equal to the true population mean.



(a) Population distribution

(b) Sample distribution
(one sample with $N = 200$)

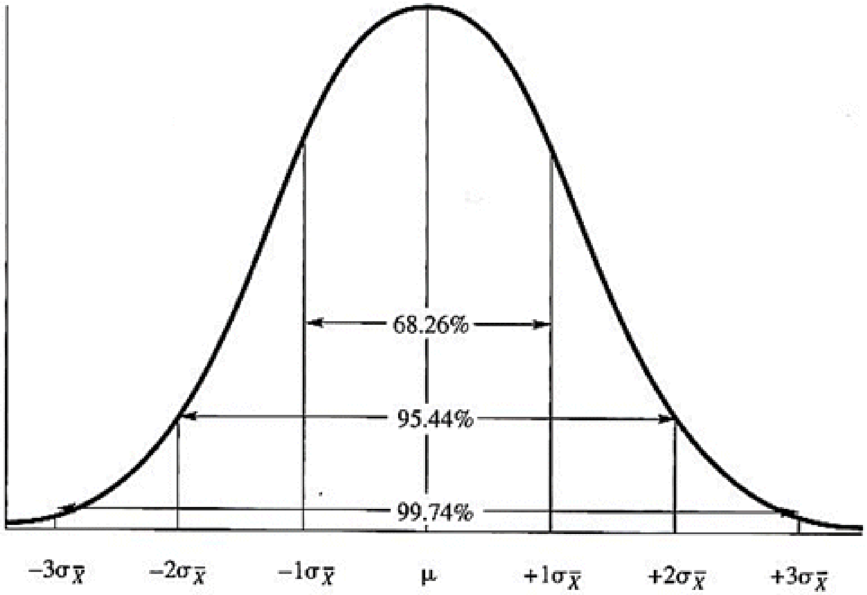
(c) Observed sampling distribution
(for 100 samples)

(d) Theoretical sampling distribution
(for infinite number of samples)

Population, Sample, and Sampling Distributions L&F p189

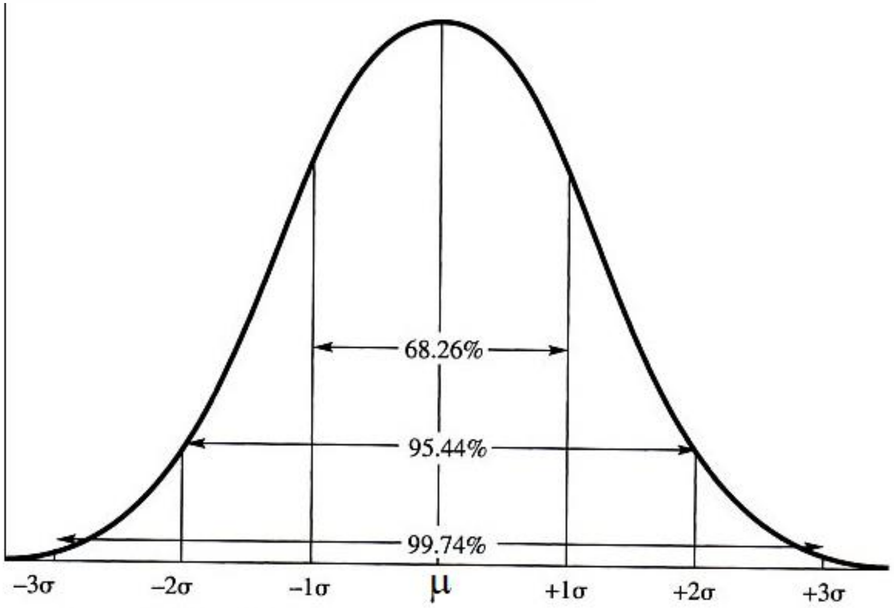
Sampling Distribution of Means as a Normal Curve

- Using the μ and σ from a normal curve, we can then assess the probability of finding specific scores along this distribution



Sampling Distribution of Means as a Probability Distribution

L&F p191

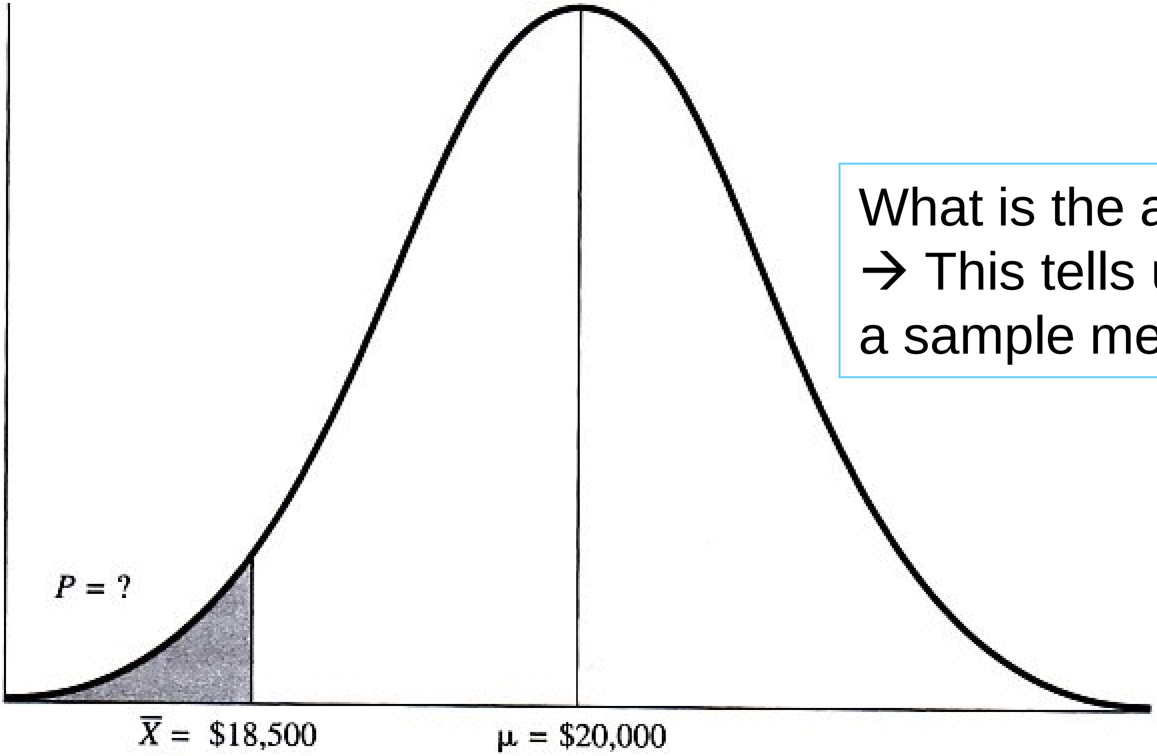


Percent of Total Area under the Normal Curve between -1σ and $+1\sigma$, -2σ and $+2\sigma$, and -3σ and $+3\sigma$

- The same applies to the distribution of sampling means, since theory tells us that this will be normally distributed

▶ Example

Imagine that University-X claims its graduates earn an average (m) annual income of $20,000. We decide to test this claim by sampling 100 alumni and measuring their incomes. In the process we get a sample mean of $18,500. Has the UNC told the truth?



What is the area of the shaded region?
→ This tells us the probability of obtaining a sample mean of $18,500 or less.

Probability Associated with Obtaining a Sample Mean of $18,500 or Less If the True Population Mean Is $20,000 and the Standard Deviation Is $700

L&F p192

Sampling means example cont.

(1) Obtain the z score for this value, using

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

- $\bar{x}$ is the sample mean (\$18,500)
- μ is the mean of means (the university's claim of \$ 20,000)
- $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of means

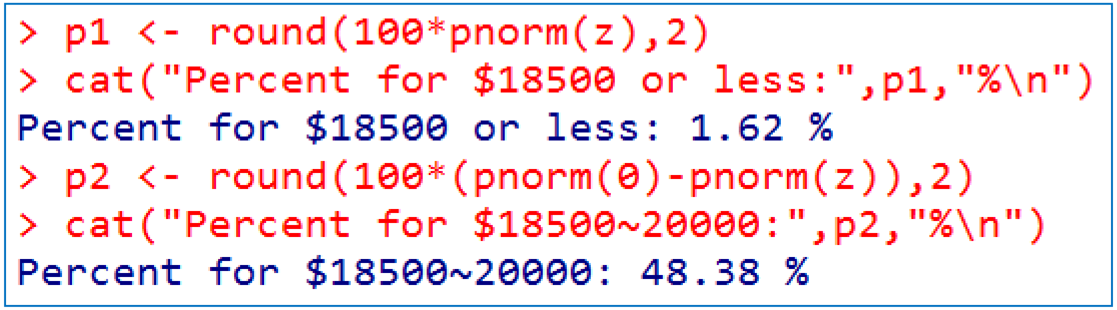
(2) Suppose we know that the standard deviation of the sampling procedure is $\sigma_{\bar{x}} = \$700$. Then we translate to a z score as:

$$z = \frac{18{,}500 - 20{,}000}{700} = -2.14$$

```
> X=18500; m=20000; sx=700
> z <- round((X-m)/sx,2)
> z
[1] -2.14
```

Sampling means example cont.

(3) Then we can consider the probability up to this value:



```
> p1 <- round(100*pnorm(z),2)
> cat("Percent for $18500 or less:",p1,"%\n")
Percent for $18500 or less: 1.62 %
> p2 <- round(100*(pnorm(0)-pnorm(z)),2)
> cat("Percent for $18500~20000:",p2,"%\n")
Percent for $18500~20000: 48.38 %
```

(4) What do we conclude then about the original reported value of $20,000 from the university claim?

Answer:

The percent of the distribution that represents sample means of $18,500 or less is 1.62%.
With such a small probability, we reject the university's claim.

Confidence Interval

- Using the standard error of the mean, we can determine a range of values within which our population mean is most likely to fall. This is the concept of a confidence interval.

- Confidence intervals are usually calculated so that the confidence interval percentage is 95%, but can be others (e.g. 99%).

$$\boxed{95\%\text{ confidence interval} = \overline{X} \pm 1.96\sigma_{\overline{X}}}$$

where $\overline{X}$ = sample mean
$\sigma_{\overline{X}}$ = standard error of the sample mean

- Confidence intervals are more informative than the simple results of hypothesis tests (where we decide "reject the null" or "don't reject the null"), since they provide a range of plausible values for the unknown parameter.

▪ Confidence Interval

```
> z=1.0   # 68% CI
> pnorm(z)-pnorm(-z)
[1] 0.6826895
>
> z=1.96  # 95% CI
> pnorm(z)-pnorm(-z)
[1] 0.9500042
>
> z=2.58  # 99% CI
> pnorm(z)-pnorm(-z)
[1] 0.99012
```

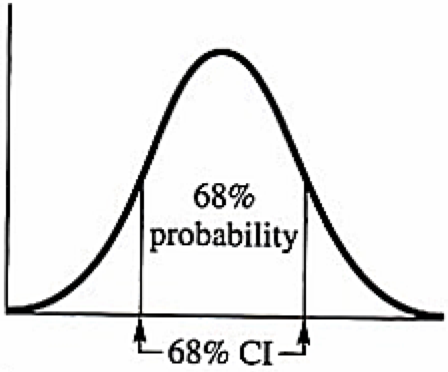
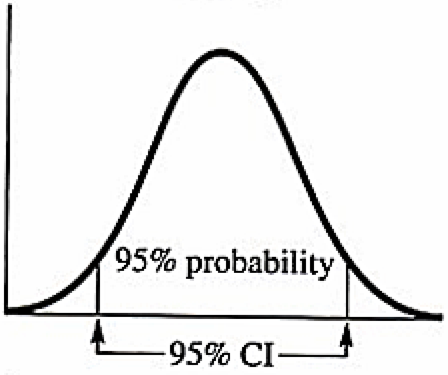
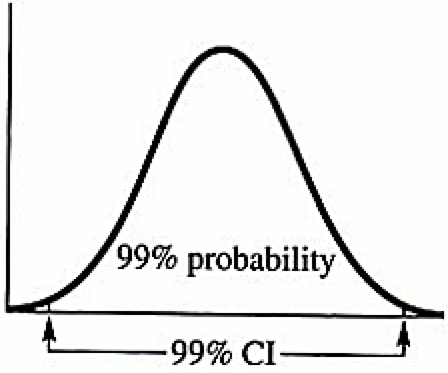
Level of confidence	z value	Confidence interval (CI)
68%	$z = \pm 1.00$	
95%	$z = \pm 1.96$	
99%	$z = \pm 2.58$	

FIGURE: Levels of Confidence L&F p198

4. t-Distribution

Student's $t-$ Distribution

- The sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a *z*-score, and use the normal distribution to evaluate probabilities with the sample mean.

- But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the ***t*** score:

Confidence intervals for a sample mean $(\overline{x})$

$$\overline{x} \pm t_{1-a/2,n-1}\frac{s}{\sqrt{n}}$$

s = standard deviation of the sample

n = number of observations

$t_{1-a/2,n-1}$ = critical t-value

When the sample size is small and population std is unknown.

- t-distribution probability density function (PDF):

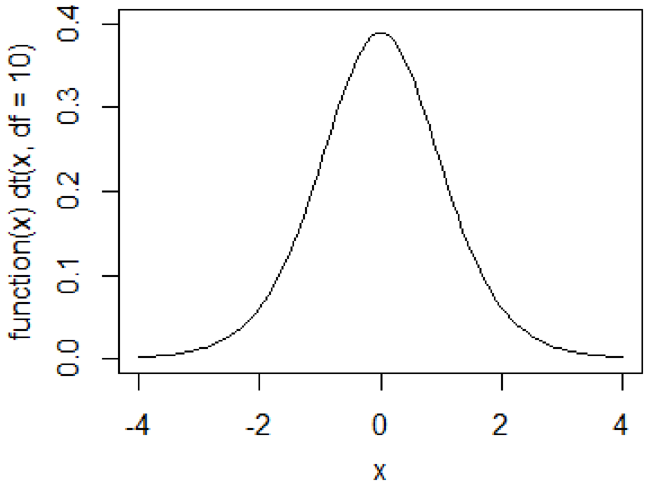
$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

k : the number of degree of freedom $\Gamma(\frac{k}{2})$: gamma function

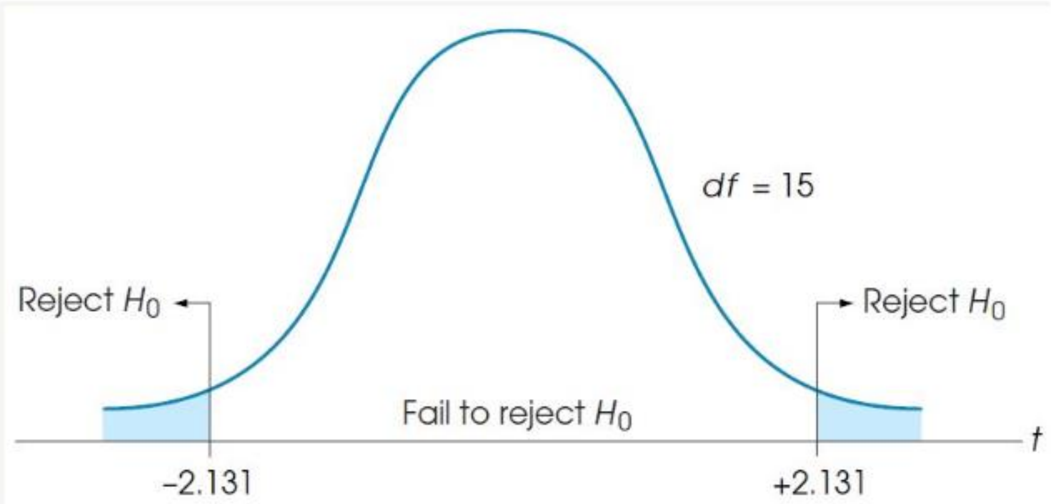
dt(x, df, ncp, ...) {stats}

Density function for the t distribution with df degrees of freedom (and optional non-centrality parameter ncp).

```
# t-distribution density curve for df=10 #
plot(function(x) dt(x,df=10),-4,4,ylim=c(0,0.4))
```



The critical region in the *t* distribution for alpha= .05 and *df*=15.

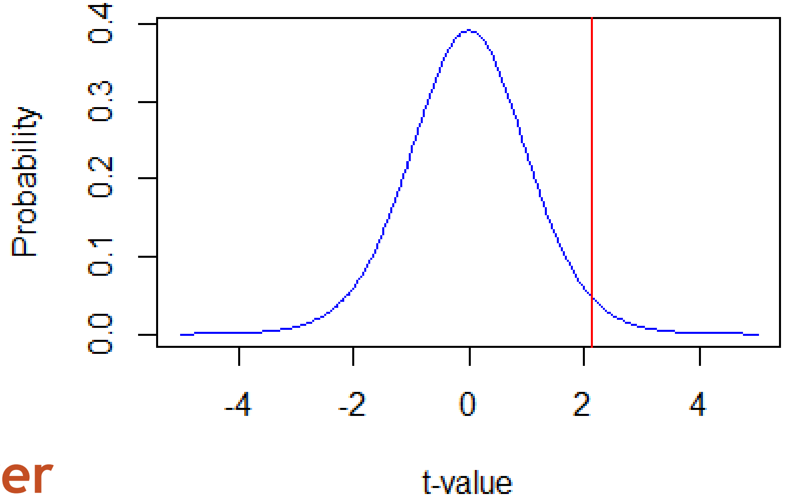


Problem

Find the 97.5th percentile of the t-distribution with 15 degrees of freedom.

Solution

```
> #critical t with a=0.05
> a=0.05; (tc=qt(1-a/2,df=15))
[1] 2.13145
> #t distribution
> curve(dt(x,df=15),-5,5,200,
+       col=4,ylab="Probability",
+       xlab="t-value")
> abline(v=tc,col=2)
```

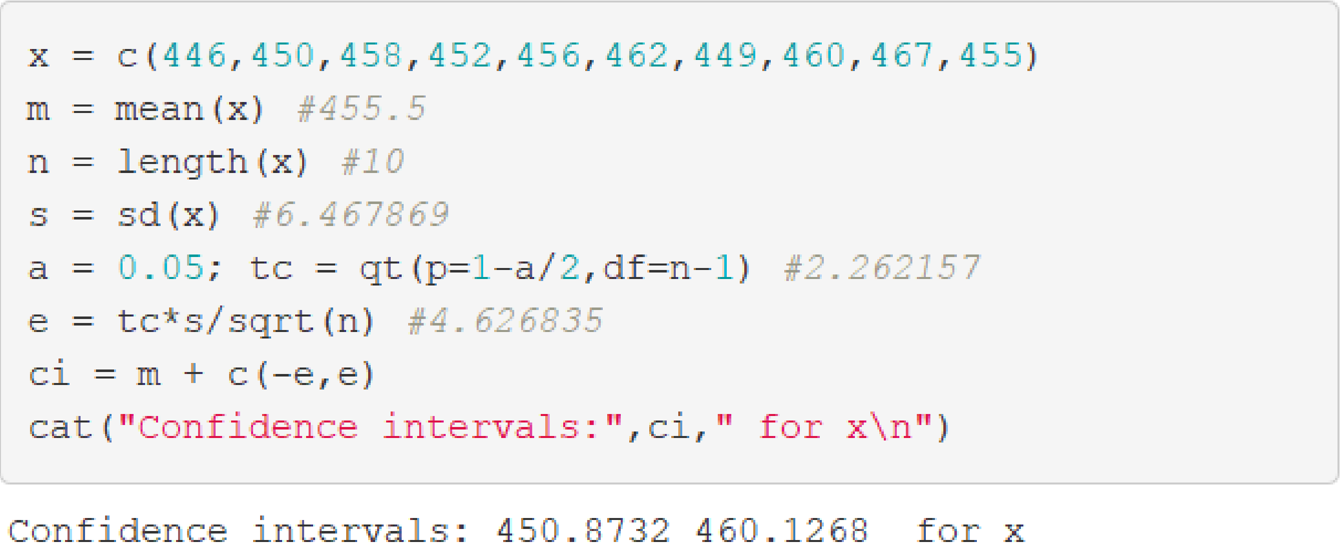


Answer

The 97.5th percentile of the t- distribution with 15 degrees of freedom is 2.131.

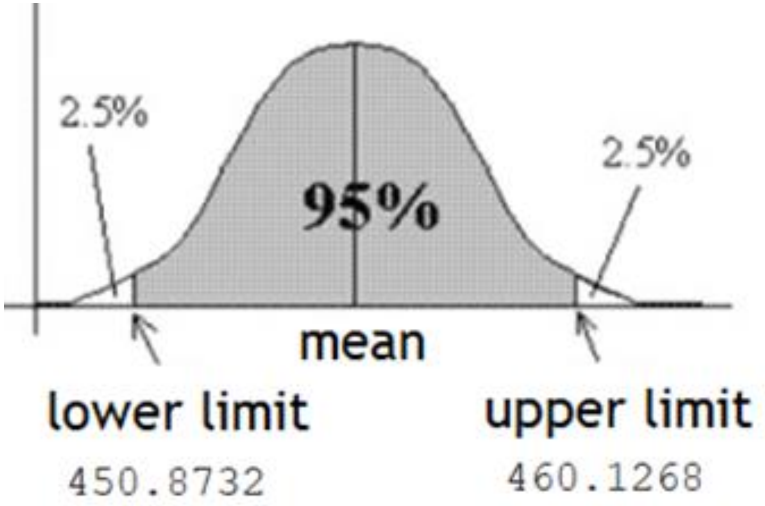
A confidence interval is the probability that a value will fall between an upper and lower bound of a probability distribution.

[Example]



```
x = c(446,450,458,452,456,462,449,460,467,455)
m = mean(x) #455.5
n = length(x) #10
s = sd(x) #6.467869
a = 0.05; tc = qt(p=1-a/2,df=n-1) #2.262157
e = tc*s/sqrt(n) #4.626835
ci = m + c(-e,e)
cat("Confidence intervals:",ci," for x\n")
```

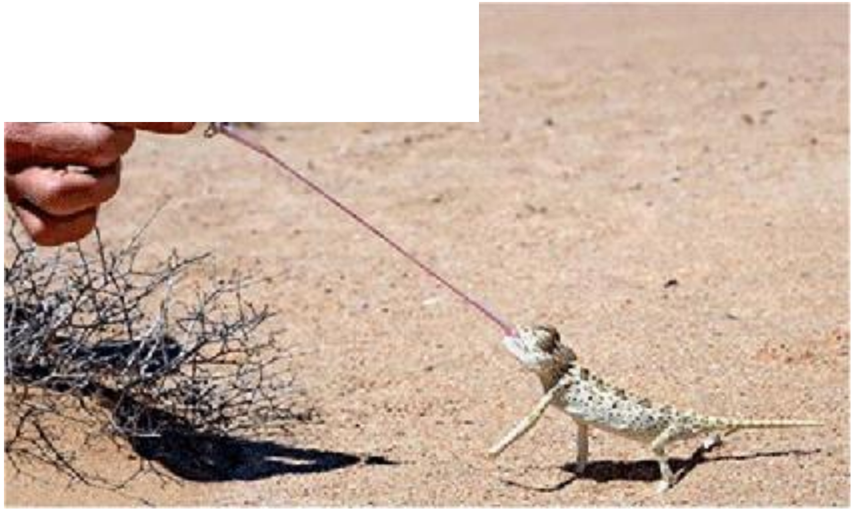
```
Confidence intervals: 450.8732 460.1268  for x
```



▪ t.test

```
> # Simply we can evaluate CI(95%) by using t.test
> t.test(X)
```

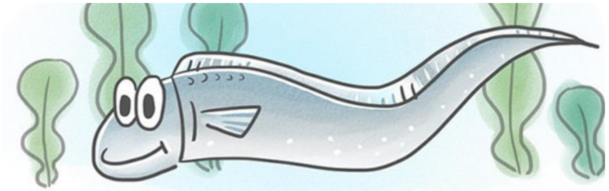
```
> cat("Confidence intervals:",ci," for x\n")
Confidence intervals: 450.8732 460.1268  for x
> t.test(x)

        One Sample t-test

data:  x
t = 222.7, df = 9, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 450.8732 460.1268
sample estimates:
mean of x
    455.5
```



5. Hypothesis Testing

- **Hypothesis: a proposition that is consistent with known data, but has been neither verified nor shown to be false**

- **Hypothesis Testing: the use of statistics to determine the probability that a given hypothesis is true**



● **Null Hypothesis (귀무가설)** H_o

The null hypothesis is a statistical hypothesis that is tested for possible rejection under the assumption that is true.
Usually, it is a statement of 'no effect' or 'no difference'.

● **Alternative Hypothesis (대립가설)** H_{a}

Contrary to the null hypothesis, the alternative hypothesis shows that observations are the result of a real effect.

In Statistics, **significant** means **probably true.**

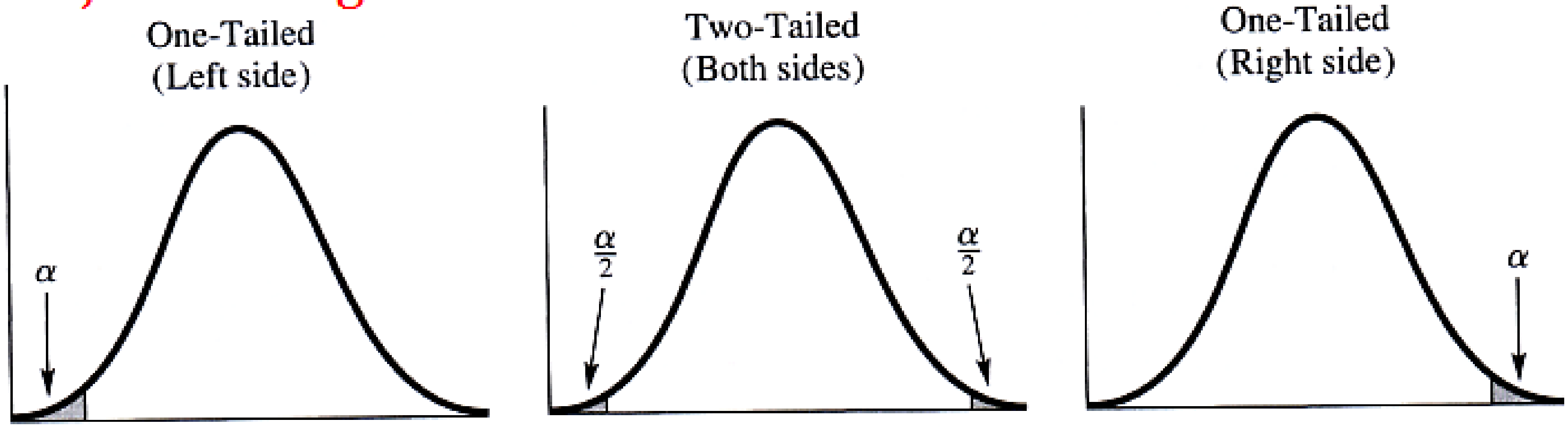
When statisticians say a result is **highly significant**, they mean it is very probably true.

- α : level of significance

  The level of probability at which H_o can be rejected with confidence and H_a accepted with confidence

- There are several levels of significance that researchers can choose: α=0.01 or 0.05

Rejection Regions:



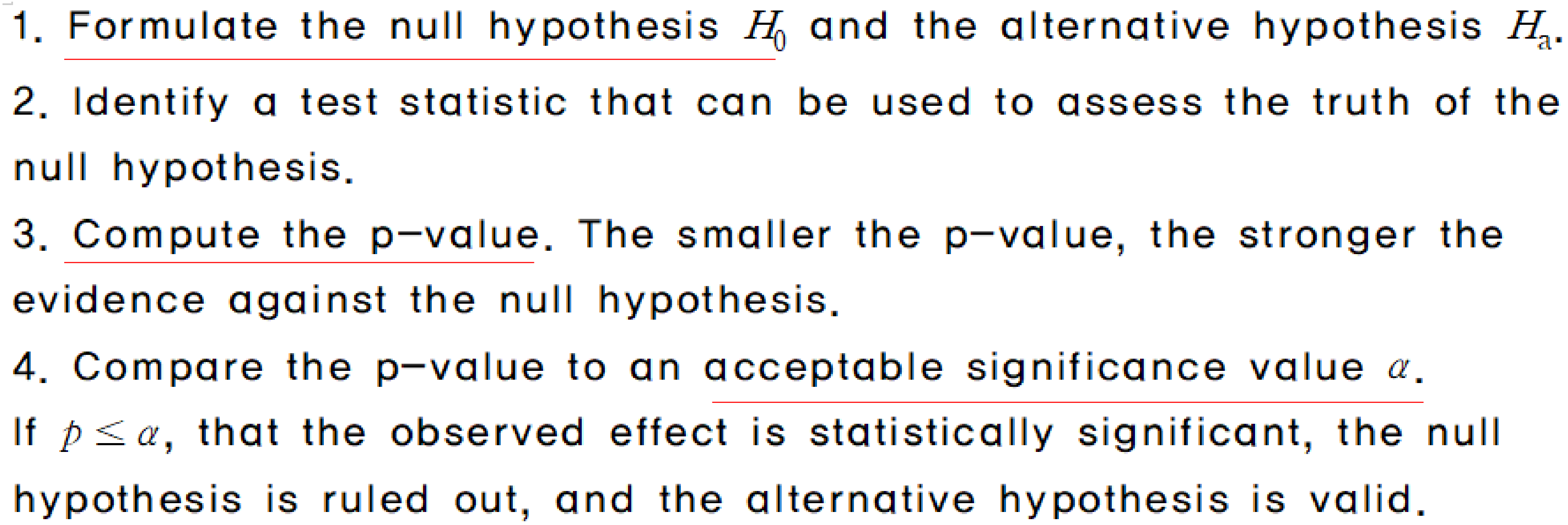
One-Tailed (Left side)

Two-Tailed (Both sides)

One-Tailed (Right side)

The usual process of hypothesis testing:

[Ref] http://mathworld.wolfram.com/HypothesisTesting.html



1. Formulate the null hypothesis H_0 and the alternative hypothesis H_a.

2. Identify a test statistic that can be used to assess the truth of the null hypothesis.

3. Compute the p-value. The smaller the p-value, the stronger the evidence against the null hypothesis.

4. Compare the p-value to an acceptable significance value α.

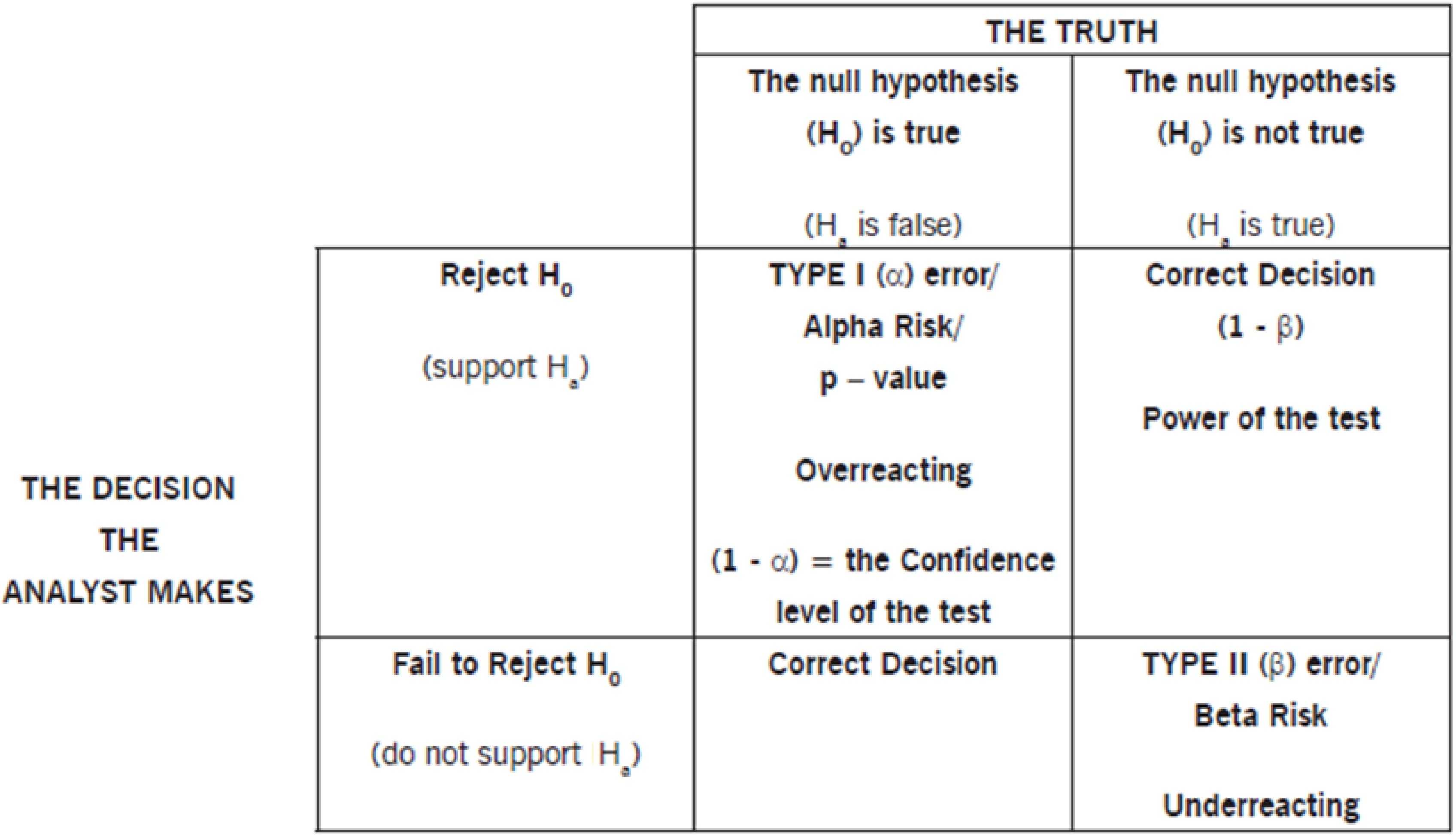
If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

Type I and type II errors

https://tinyurl.com/y58vu3co

Type I error : Failing to retain the null hypothesis
(Error of rejecting the null hypothesis)

Type II error : Failing to reject a false null hypothesis
(Error of retaining the null hypothesis)



		THE TRUTH	
		The null hypothesis (H_0) is true (H_a is false)	The null hypothesis (H_0) is not true (H_a is true)
THE DECISION THE ANALYST MAKES	Reject H_0 (support H_a)	TYPE I (α) error/ Alpha Risk/ p – value Overreacting ($1 - \alpha$) = the Confidence level of the test	Correct Decision ($1 - \beta$) Power of the test
	Fail to Reject H_0 (do not support H_a)	Correct Decision	TYPE II (β) error/ Beta Risk Underreacting

Type I and type II errors

https://tinyurl.com/y58vu3co

Type I error : Failing to retain the null hypothesis
(Error of rejecting the null hypothesis)

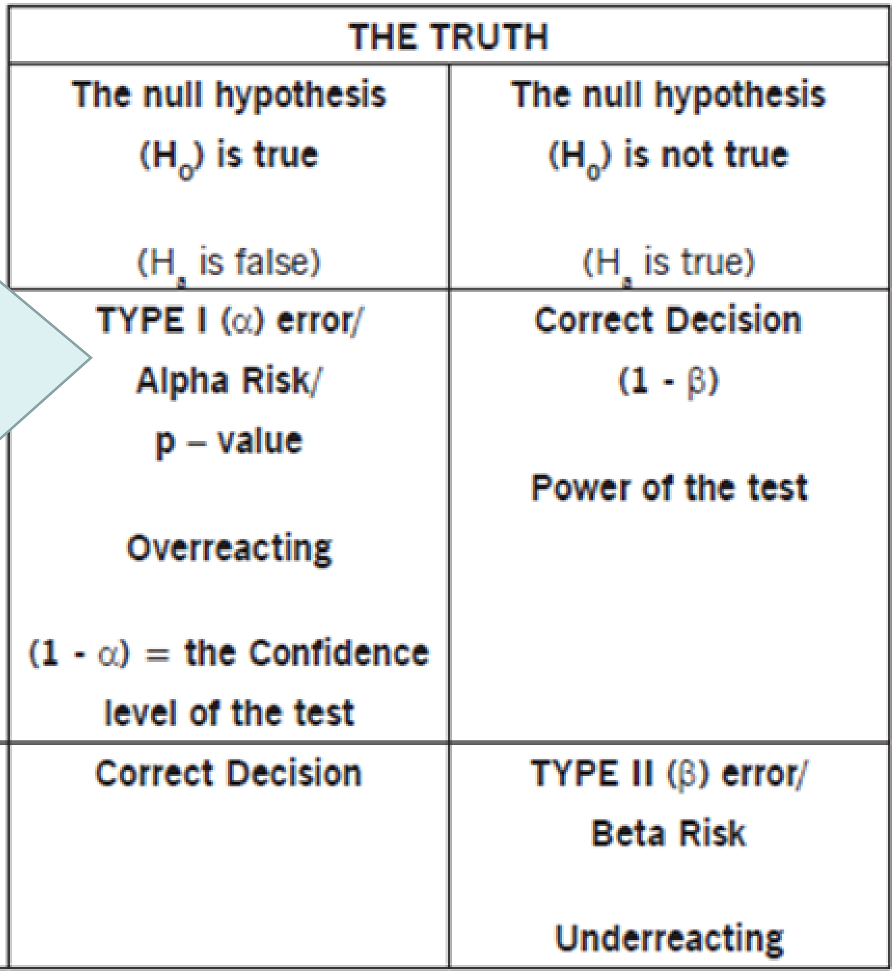
Type II error : Failing to reject a false null hypothesis
(Error of retaining the null hypothesis)

```
#false means incorrect
#positive means rejecting Ho
```

Type I error (False Positive)

Same as **false alarm**. When the null hypothesis is actually true, but was rejected as false at testing.

There is a 5% chance that the sample results are due to chance alone, so there is a 5% chance that rejecting the null hypothesis (and supporting the alternative hypothesis) will be an incorrect decision.



THE TRUTH	
The null hypothesis (H_0) is true (H_a is false)	The null hypothesis (H_0) is not true (H_a is true)
TYPE I (α) error/ Alpha Risk/ p – value Overreacting (1 - α) = the Confidence level of the test	Correct Decision (1 - β) Power of the test
Correct Decision	TYPE II (β) error/ Beta Risk Underreacting

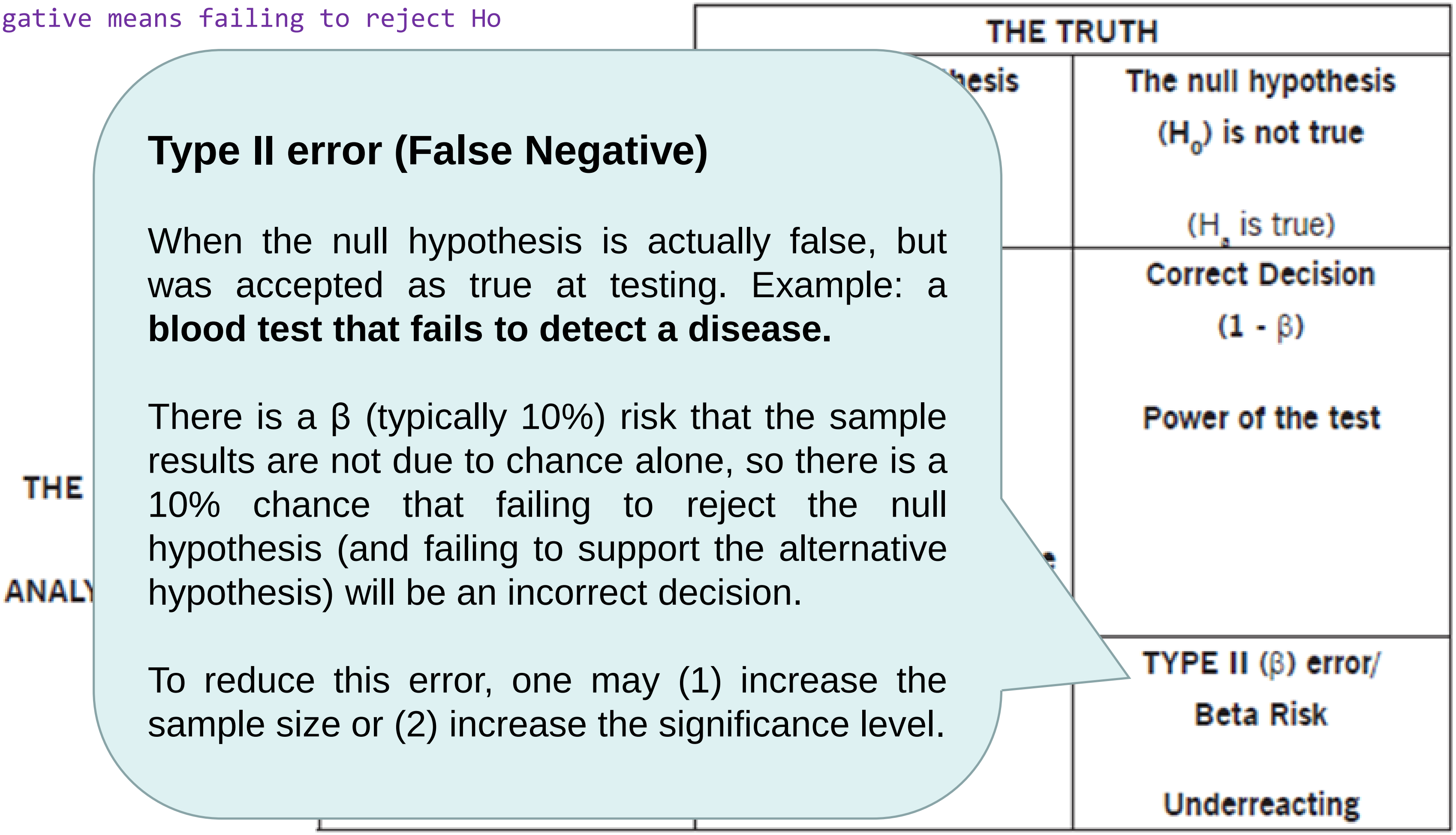
Type I and type II errors

https://tinyurl.com/y58vu3co

Type I error : Failing to retain the null hypothesis
(Error of rejecting the null hypothesis)

Type II error : Failing to reject a false null hypothesis
(Error of retaining the null hypothesis)

```
#false means incorrect
#negative means failing to reject Ho
```



THE TRUTH

hesis

The null hypothesis (H_0) is not true

(H_a is true)

Correct Decision

(1 - β)

Power of the test

TYPE II (β) error/

Beta Risk

Underreacting

THE

ANALY

e

Type II error (False Negative)

When the null hypothesis is actually false, but was accepted as true at testing. Example: a **blood test that fails to detect a disease.**

There is a β (typically 10%) risk that the sample results are not due to chance alone, so there is a 10% chance that failing to reject the null hypothesis (and failing to support the alternative hypothesis) will be an incorrect decision.

To reduce this error, one may (1) increase the sample size or (2) increase the significance level.

Type I and type II errors

https://tinyurl.com/y58vu3co

Ha: The evidence produced before the court proves that this man is guilty.
Null Hypothesis (Ho): This man is innocent.

		THE TRUTH	
		The null hypothesis (H_0) is true (H_a is false)	**The null hypothesis (H_0) is not true** (H_a is true)
THE DECISION THE ANALYST MAKES	**Reject H_0** (support H_a)	**Type I error** Convicting an innocent person	Convicting a guilty person
	Fail to Reject H_0 (do not support H_a)	Let go an innocent person	**Type II error** Let go a guilty person

● Example 1: Independent two-sample t-test

Two-Tailed Test (Both sides)

Null hypothesis $H_0 : \bar{x}_1 = \bar{x}_2$

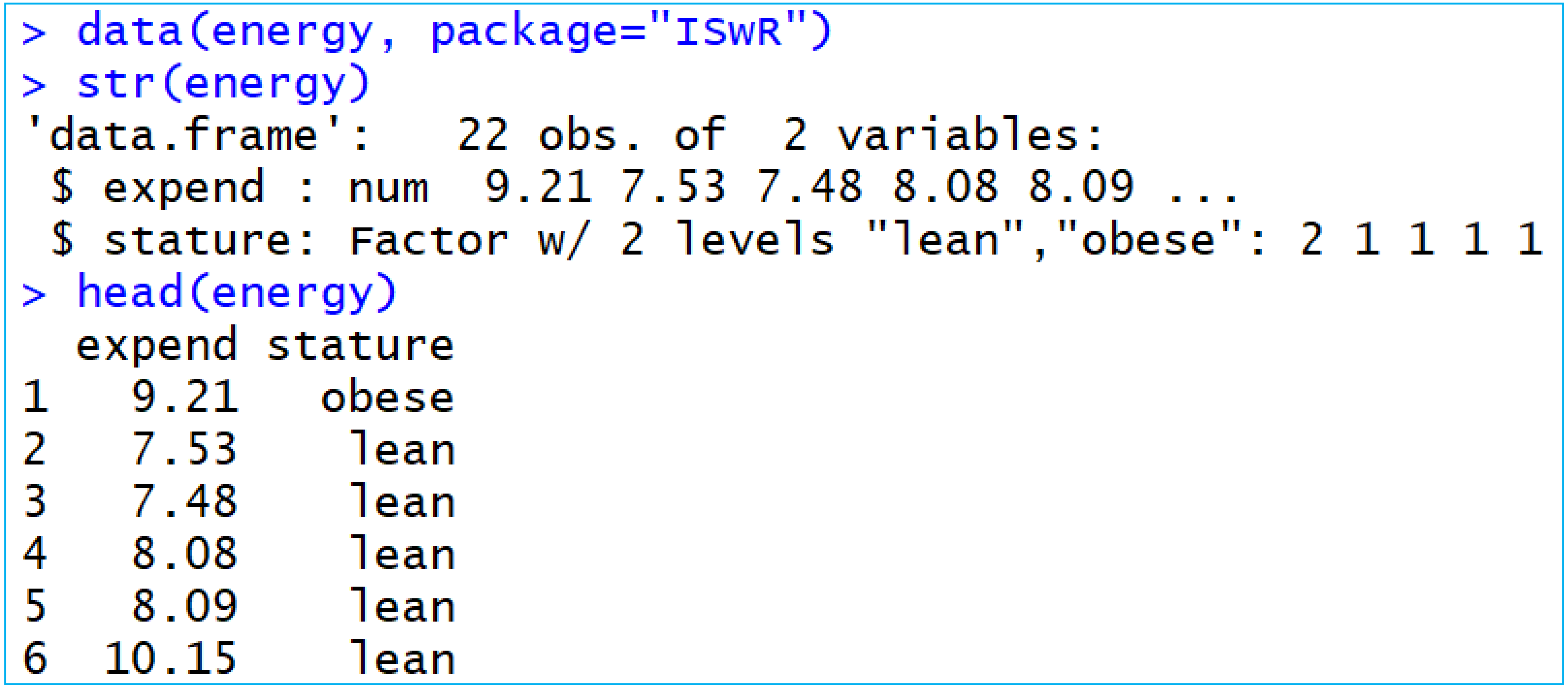
Alternative hypothesis $H_a : \bar{x}_1 \neq \bar{x}_2$

Test Statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + \sqrt{s_2^2/n_2}}}$$

where n_1 and n_2 are the sample sizes, $\bar{x}_1$ and $\bar{x}_2$ are the sample means, and s_1^2 and s_2^2 are the sample variances.

[Sample Dataset] Energy expenditure



```
> data(energy, package="ISwR")
> str(energy)
'data.frame':   22 obs. of  2 variables:
 $ expend : num  9.21 7.53 7.48 8.08 8.09 ...
 $ stature: Factor w/ 2 levels "lean","obese": 2 1 1 1 1
> head(energy)
  expend stature
1   9.21   obese
2   7.53    lean
3   7.48    lean
4   8.08    lean
5   8.09    lean
6  10.15    lean
```

Hypotheses:

Null hypothesis $H_0 : \bar{x}_1 = \bar{x}_2$

Alternative hypothesis $H_a : \bar{x}_1 \neq \bar{x}_2$ (1: lean, 2: obese)

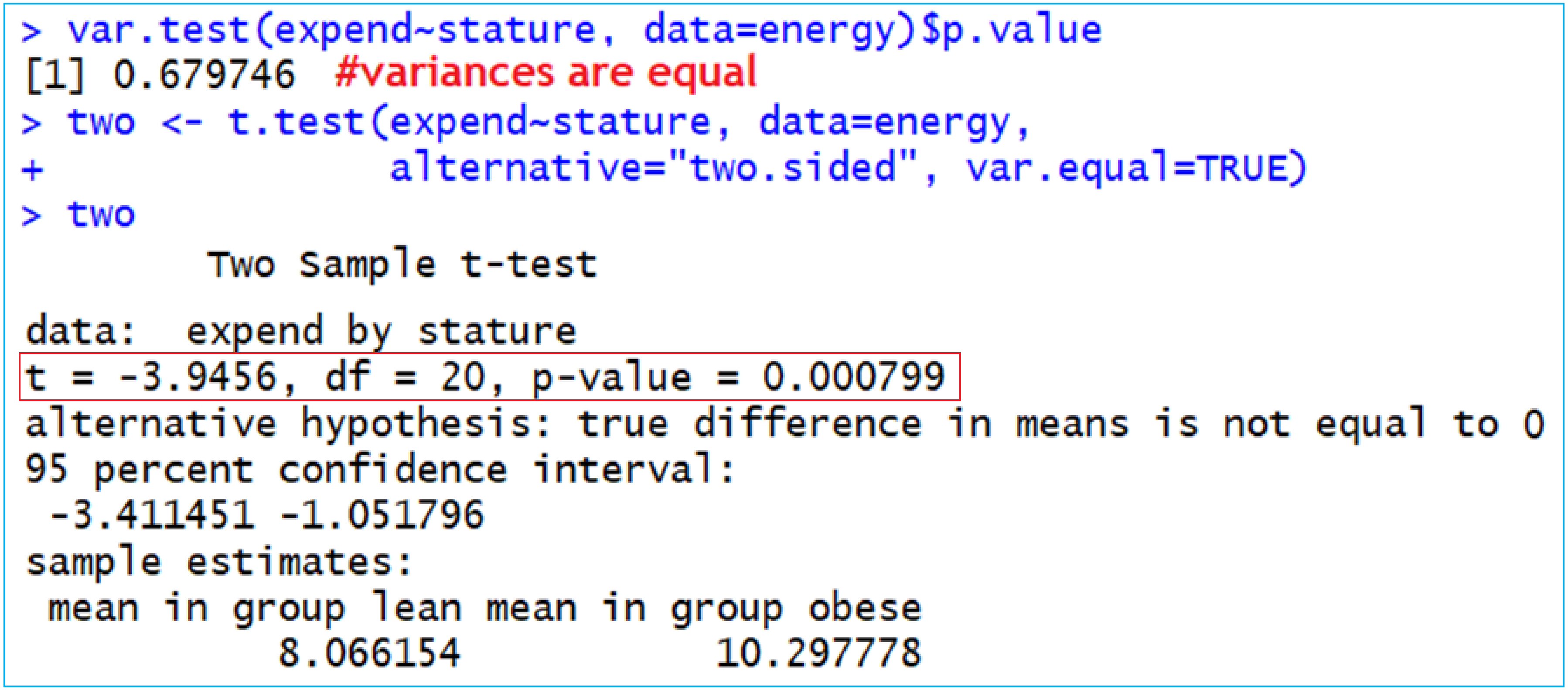
Example 1: Independent two-sample t-test

The t-test is used to determine whether the means of two groups are equal to each other.

t.test(formula, data, alternative, var.equal, …)

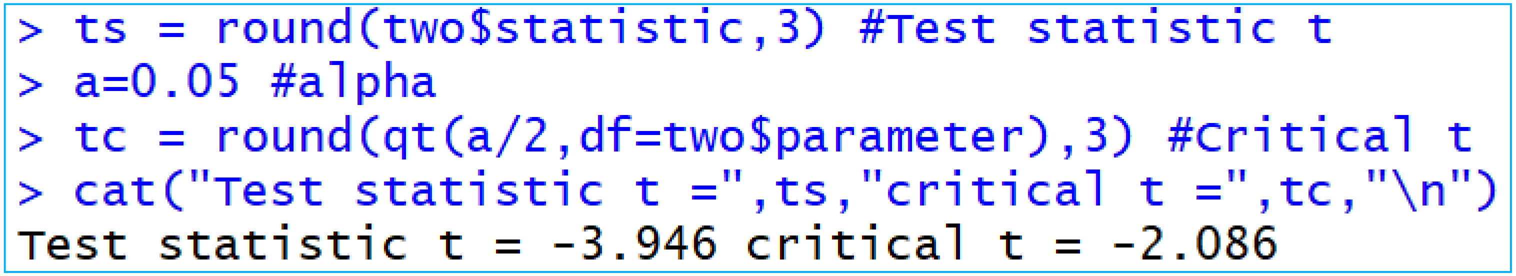
Variances of the two samples are equal: var.equal=TRUE

Variances of the two samples are not equal: var.equal=FALSE (Welch's test)

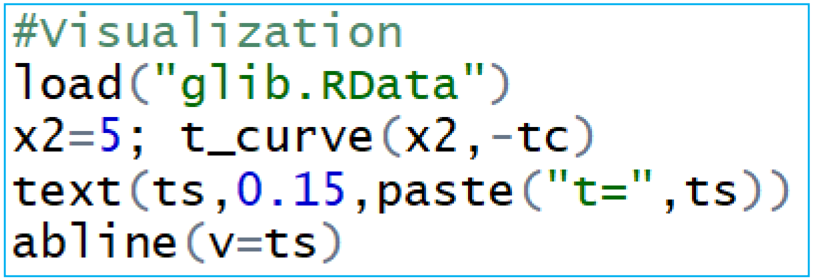


```
> var.test(expend~stature, data=energy)$p.value
[1] 0.679746  #variances are equal
> two <- t.test(expend~stature, data=energy,
+               alternative="two.sided", var.equal=TRUE)
> two

        Two Sample t-test

data:  expend by stature
t = -3.9456, df = 20, p-value = 0.000799
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.411451 -1.051796
sample estimates:
 mean in group lean mean in group obese
           8.066154           10.297778
```

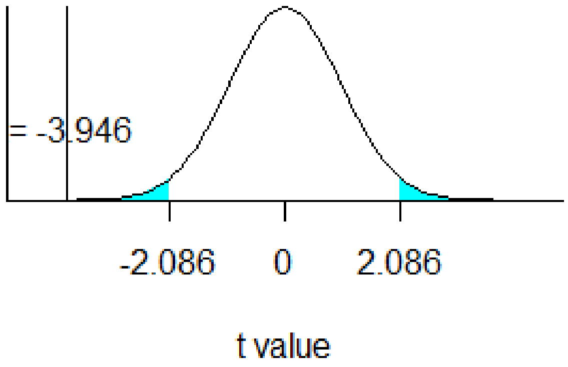
Test statistic t and critical t:



```
> ts = round(two$statistic,3) #Test statistic t
> a=0.05 #alpha
> tc = round(qt(a/2,df=two$parameter),3) #Critical t
> cat("Test statistic t =",ts,"critical t =",tc,"\n")
Test statistic t = -3.946 critical t = -2.086
```



```
#Visualization
load("glib.RData")
x2=5; t_curve(x2,-tc)
text(ts,0.15,paste("t=",ts))
abline(v=ts)
```

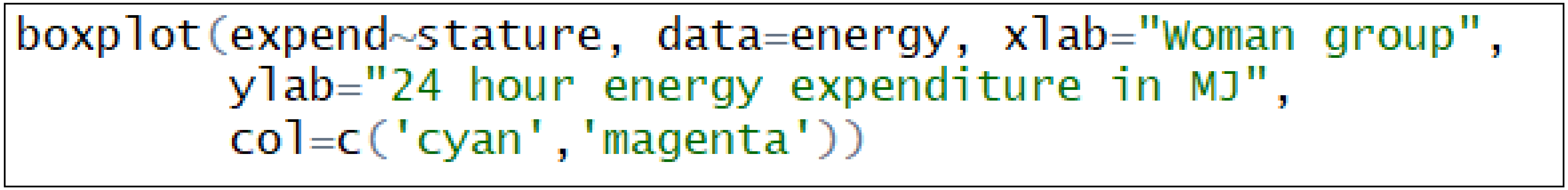


Interpretation:

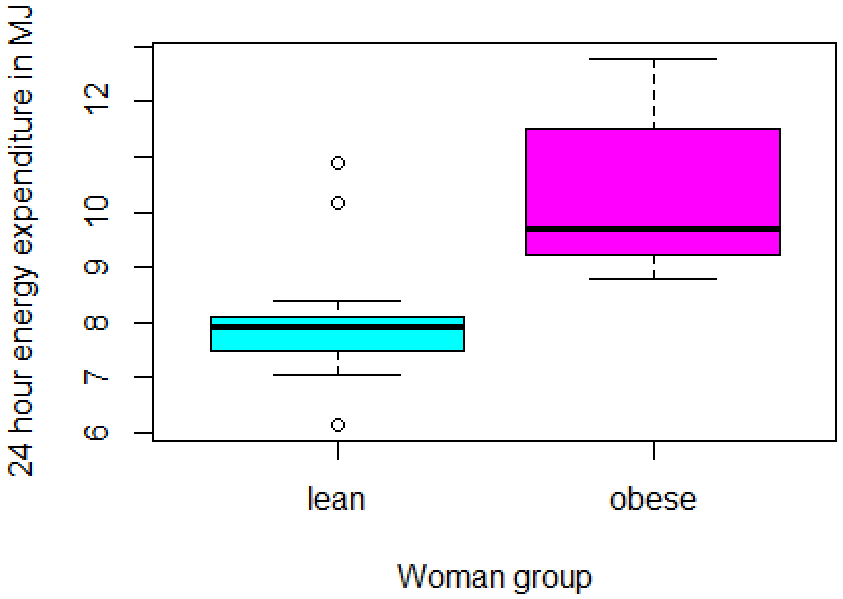
Test statistic t=-3.945 is less than critical t=-2.086.
P-value=0.001 is less than α=0.05.

We reject the null hypothesis. There is a significant difference between lean and obese women in energy expenditure.

Example 1: Independent two-sample t-test



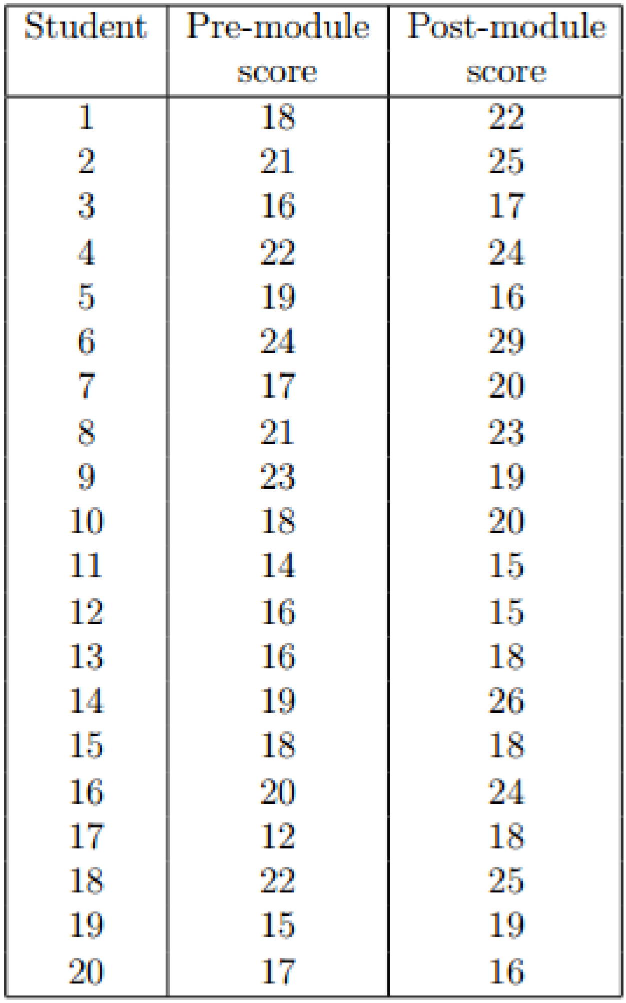
```
boxplot(expend~stature, data=energy, xlab="Woman group",
        ylab="24 hour energy expenditure in MJ",
        col=c('cyan','magenta'))
```



This box plot shows that the energy expenditures of the obese women are greater than the lean women.

Example 2: Dependent t-test for paired samples

Suppose a sample of 20 students were given a diagnostic test before studying a particular module and then again after completing the module. We want to find our teaching leads to improvements in students' knowledge/skills.



Student	Pre-module score	Post-module score
1	18	22
2	21	25
3	16	17
4	22	24
5	19	16
6	24	29
7	17	20
8	21	23
9	23	19
10	18	20
11	14	15
12	16	15
13	16	18
14	19	26
15	18	18
16	20	24
17	12	18
18	22	25
19	15	19
20	17	16

[Sample Dataset] PrePost.csv

```
> dt <- read.csv("PrePost.csv", header=T)
> head(dt)
  x1 x2
1 18 22
2 21 25
3 16 17
4 22 24
5 19 16
6 24 29
```

● Example 2: Dependent t-test for paired samples

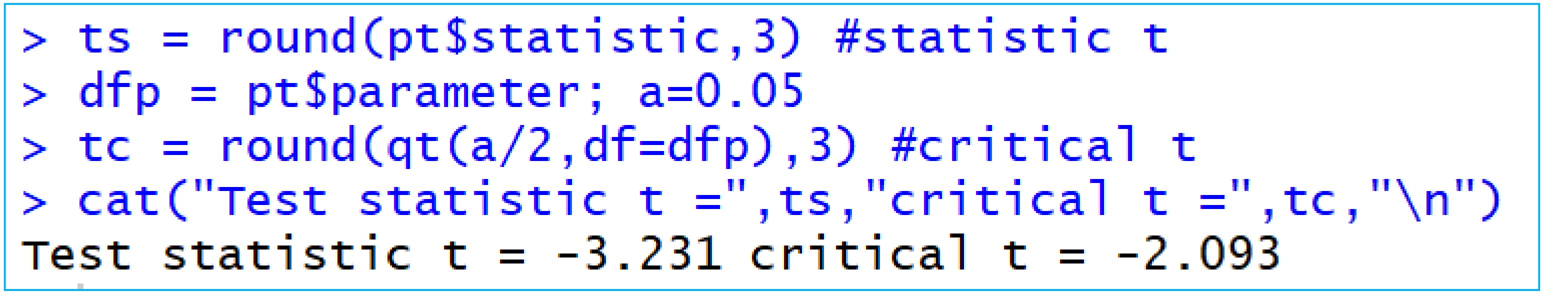
Hypotheses:

Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypothesis $H_a : \mu_1 \neq \mu_2$ (1: Pre, 2: Post)

```
> #paired t-test
> pt <- t.test(dt$x1,dt$x2,paired=TRUE)
> pt

        Paired t-test

data:  dt$x1 and dt$x2
t = -3.2313, df = 19, p-value = 0.004395
95 percent confidence interval:
 -3.3778749 -0.7221251
sample estimates:
mean of the differences
                  -2.05
```

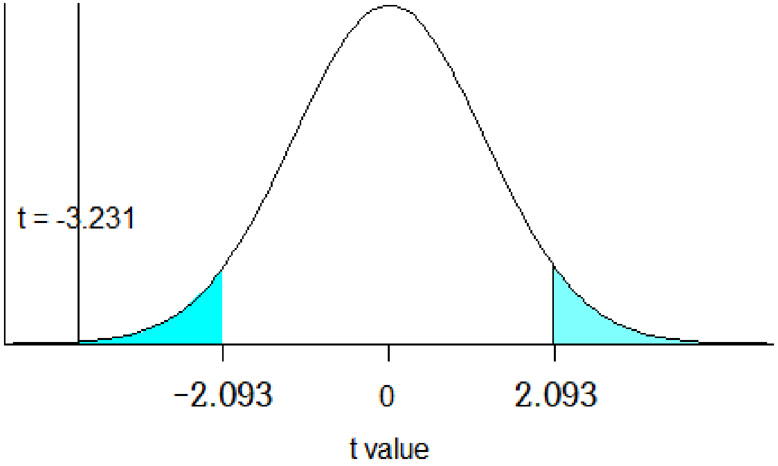
Example 2: Dependent t-test for paired samples



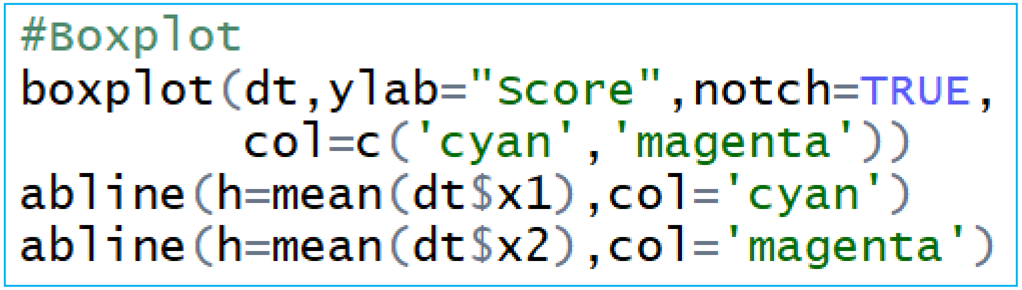
```
> ts = round(pt$statistic,3) #statistic t
> dfp = pt$parameter; a=0.05
> tc = round(qt(a/2,df=dfp),3) #critical t
> cat("Test statistic t =",ts,"critical t =",tc,"\n")
Test statistic t = -3.231 critical t = -2.093
```

Interpretation:

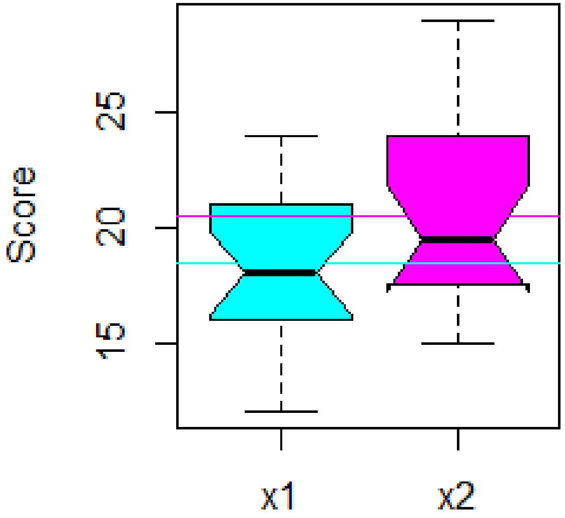
Test statistic t=-3.231 is less than critical t=-2.093. P-value=0.004 < 0.05.



We reject the null hypothesis. There is a significant difference between pre- and post-module scores. The teaching module is considered effective.



```
#Boxplot
boxplot(dt,ylab="Score",notch=TRUE,
        col=c('cyan','magenta'))
abline(h=mean(dt$x1),col='cyan')
abline(h=mean(dt$x2),col='magenta')
```


Score
x1
x2