Hypothesis Testing

- What is hypothesis testing
- Standard procedures
- Examples

What the data refer to?

- If we flip a coin 100 times, and 45 come up heads

  this could easily occur by chance. There is not sufficient evidence to suggest that the coin is unfair.

- If we flip a coin 100 times, and 25 come up heads

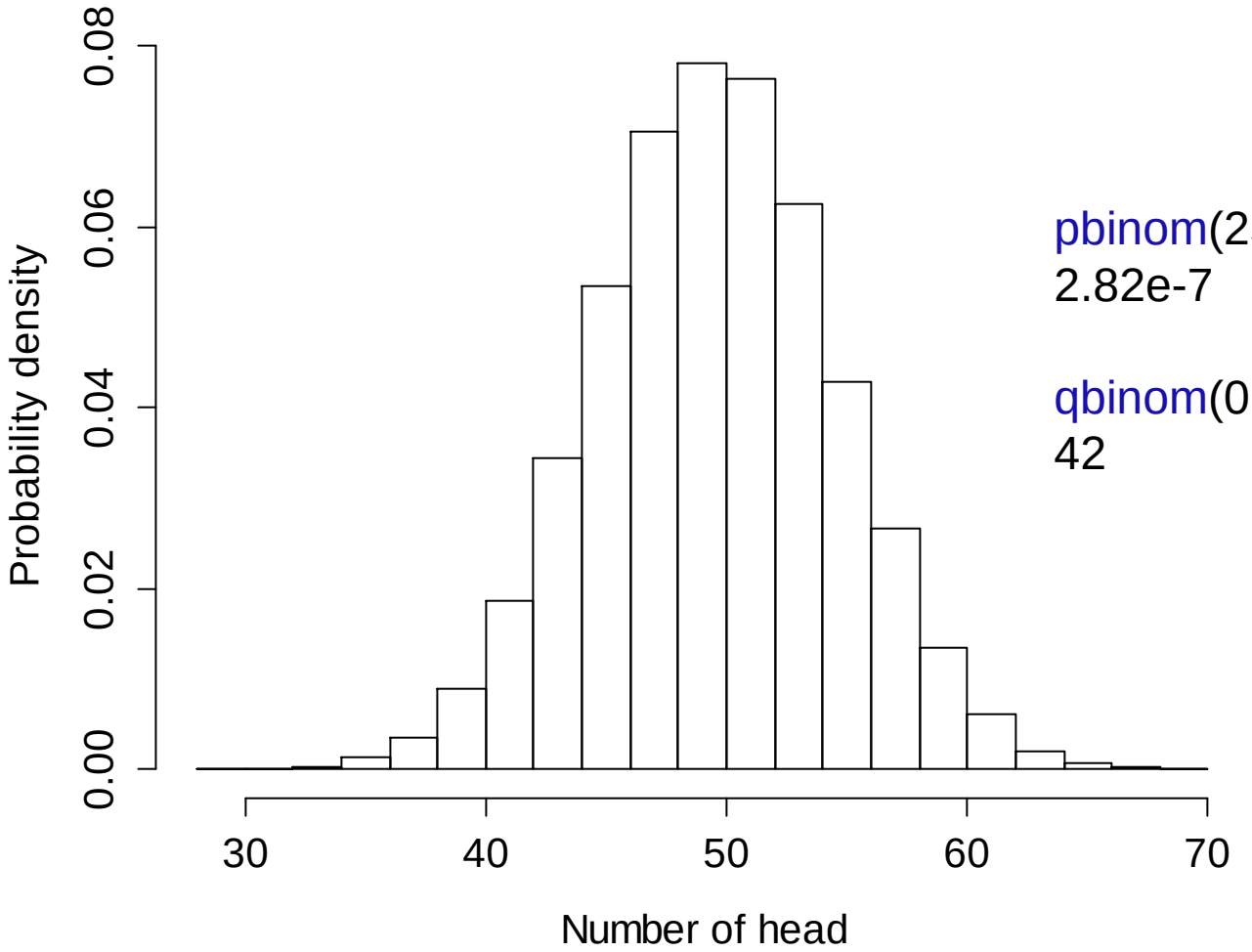
  this would be an rare event if the coin was fair. The low probability is evidence that the coin may not be fair.

  The coin is definitely unfair?

  Although rare, 25 heads is still possible by chance from a fair coin.

Use probability density

```
hist(rbinom(100000, 100, .5), freq = F, main = “”,
     xlab = 'Number of head', ylab = 'Probability density')
```




```
pbinom(25, 100, .5)
2.82e-7

qbinom(0.05, 100, .5)
42
```

A story: Lady tasting tea

- The lady tasting tea is a famous randomized experiment devised by Ronald A. Fisher and reported in his book The Design of Experiments (1935).
- The lady (Muriel Bristol) be able to tell whether the tea or the milk was added first to a cup.
- The experiment provided the Lady with 8 randomly ordered cups of tea – 4 prepared by first adding milk, 4 prepared by first adding the tea. She was to select the 4 cups prepared by one method.
- The null hypothesis was that the Lady had no such ability.

A story: Lady tasting tea

Tea-Tasting Distribution

Success count	Permutations of selection	Number of permutations
0	oooo	1
1	ooox, ooxo, oxoo, xooo	16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	36
3	oxxx, xoxx, xxox, xxxo	16
4	xxxx	1
Total		70

If and only if the Lady properly categorized all 8 cups was Fisher willing to reject the null hypothesis – effectively acknowledging the Lady's ability at a 1.4% significance level.

Hypothesis testing – confirmatory data analysis

- A hypothesis is a claim or statement about a property of a population (e.g. the mean or a proportion of the population)

- A hypothesis test (or test of significance) is a standard procedure for testing a claim or statement about a property of a population.

  It is extremely important to realize that we are not making definitive conclusions. We are giving probabilistic conclusions. We are either concluding that the results we get are likely due to chance, or unlikely.

(Zar 1999)

Origins

- Hypothesis testing is the product of Ronald Fisher, Jerzy Neyman, Karl Pearson and Egon Pearson.
- Fisher emphasized rigorous experimental design and methods to extract a result from few samples assuming Gaussian distributions.
- Neyman and E. Pearson emphasized mathematical rigor and methods to obtain more results from many samples and a wider range of distributions (Neyman and Pearson 1933).
- Modern hypothesis testing is a hybrid of the Fisher vs. Neyman/Pearson formulation, methods and terminology developed in the early 20th century.

Neyman, Jerzy; Pearson, Egon S. (1933). "On the problem of the most efficient tests of statistical hypotheses". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 231: 289–337.

Other options rather than frequentist hypothesis testing

- Confidence interval (CI) is a particular kind of interval estimate of a population parameter and is used to indicate the reliability of an estimate

- The Bayesian approach to hypothesis testing is to base rejection of the hypothesis on the posterior probability

- Other approaches to reaching a decision based on data are available via decision theory and optimal decisions

t-test using R

```
# t-test daily energy intake in kJ for 11 women (Altman, 1991, p. 183)
daily.intake = c(5260,5470,5640,6180,6390,6515,
                 6805,7515,7515,8230,8770)
mean(daily.intake)
sd(daily.intake)
quantile(daily.intake)

t.test(daily.intake, mu = 7725)
# Nonparametric
wilcox.test(daily.intake, mu = 7725)
# Two samlpes
x1 = rnorm(300, 0, 1)
x2 = sample(0:100, 300, rep = T)
t.test(x1, x2)
# Check normality
plot(x1); hist(x1); qqnorm(x1)
shapiro.test(x1)
```

One Sample t-test

data: daily.intake
t = -2.8208, df = 10, **p-value = 0.01814**
alternative hypothesis:
true mean is not equal to 7725

95 percent confidence interval:
5986.348 7520.925

sample estimates:
mean of x
6753.636

Wilcoxon signed rank test

data: daily.intake
V = 8, **p-value = 0.0293**
alternative hypothesis: true location is not equal to 7725

D.G. Altman (1991), *Practical Statistics for Medical Research*, Chapman & Hall.

Example

sample mean v.s. population mean

Fish in polluted water are larger?

http://texascoastgeology.com/passes/san_bernard_10_29_10%20%20004a2.jpg

Research proposition

- Water pollution usually cause eutrophication, resulting in sizes of fish being larger than normal.
- We draw a random sample of 30 fish individuals and calculate their mean length (5.3 cm)
- Based on surveys at upstream, we know the mean length of the fish is 5.1 cm
- By comparing the means, we are asking whether it is reasonable to consider the sample of the fish that is larger than the normal size.

Fish	Length
1	5.76
2	5.21
3	5.44
4	5.46
5	5.45
6	5.07
7	5.14
8	5.64
9	5.10
10	4.62
11	5.08
12	5.45
13	5.21
14	5.03
15	5.40
16	5.52
17	5.81
18	5.54
19	5.35
20	5.36
21	5.69
22	4.57
23	5.83
24	5.18
25	5.01
26	5.45
27	5.31
28	5.34
29	5.13
30	5.34

Sample and population

Downstream	Upstream
$\bar{x}$ = 5.3	μ = 5.1
n = 30	σ = 0.3

The sample mean is 0.2 cm larger than the population mean

What are the possibilities?

- The average fish lengths at downstream area are about the same as the normal size, and this sample happens to show a really large mean.
  difference = sampling variability

- The average fish lengths at downstream area are indeed larger than the normal size.
  difference = real

How do we decide which explanation makes more sense?

- The traditional way: hypothesis tests
  - Five steps

One sample hypothesis test

- **A random sample:** 30 fish with an average of length of 5.3 cm.
- **Population:** The average fish length at upstream is 5.1 cm with a standard deviation of 0.3 cm.

Do fish in polluted water were significantly larger than the population?

Step one: checking assumptions

Hypothesis testing involves several assumptions that must be met for the results of the test to be valid

For the one sample hypothesis test, we assume:

- random sampling
- the level of measurement is interval-ratio
- the sampling distribution is normal

Step two: stating hypotheses

- Null hypothesis (H_0): a statement of 'no difference'
- Alternative hypothesis (H_a): a statement that reflects the research question
- Both are expressed in terms of population parameters

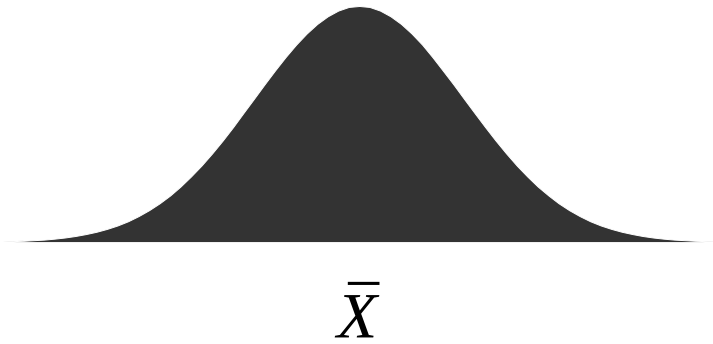
Step two: stating hypotheses

$$H_0: \mu_{polluted} = 5.1$$

$$H_a: \mu_{polluted} > 5.1$$

Step three: select the sampling distribution and establish the critical region (1/3)

- Select sampling distribution:

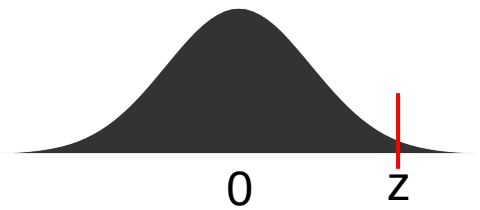


$\mu = 5.1$
$\sigma = 0.3$

Step three: selecting sampling distribution and establish the critical region (2/3)

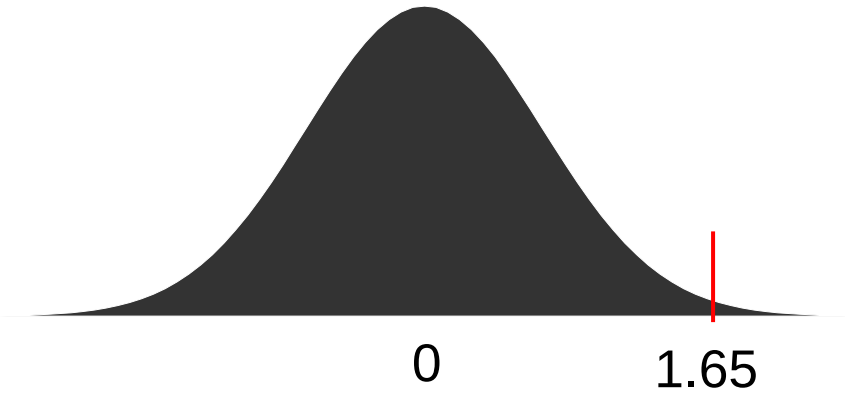
- Sampling distribution $\bar{X}$ -> statistic distribution Z. using Z statistic and normal distribution when the sample is large
- $Z_{critical}$: the score associated with a particular α level and marking the beginning of the critical region
- Critical region: area under the sampling distribution that includes all unlikely sample results
- P-value: the “chance” of getting the observed sample mean **NOT** further away from the hypothesized population mean?

  P-value small: reject $\mathbf{H_o}$
  P-value large: fail to reject $\mathbf{H_o}$



Step three: selecting sampling distribution and establish the critical region (3/3)

- Confidence level: $1-\alpha = 0.95$
- $Z_{critical} = 1.65$



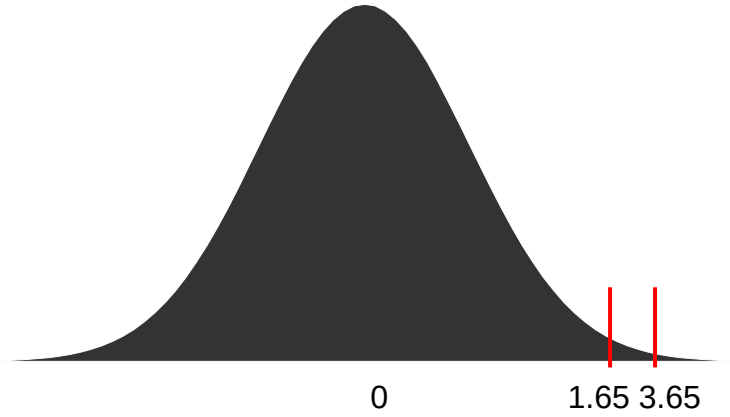
Step four: compute the test statistic

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

$$Z = \frac{5.3 - 5.1}{0.3/\sqrt{30}} = 3.65$$

Step five: make a decision

- Plot the test statistic on the sampling distribution
- If the test statistic is in the critical region, our decision is: reject the null (“statistically significant”)
- If the test statistic is *not* in the critical region, our decision is: fail to reject the null



Step five: make a decision

- We reject the null hypothesis.
- We are 95% confident that fish in polluted water have significantly larger sizes than normal fish.

Formal hypothesis testing

1. Making/checking assumptions
2. Convert your claim into a symbolic null and alternative hypothesis
3. Select the sampling distribution and establish the critical region
4. Calculate a test statistic
5. Compare the test statistic to critical values OR a probability, write a conclusion

For the one sample hypothesis test, we assume

- Random sampling
- The level of measurement is interval-ratio
- The sampling distribution is normal

When can you assume the shape of the sampling distribution is normal?

- When the population distribution is normal
- When you check sample mean (Based on the Central Limit Theorem)
  - shapiro.test(x) p>0.05
  - qqnorm(x) straight line

Test of normality shapiro.test

- Shapiro-Wilk test of normality

  published in 1965 by Samuel Sanford Shapiro and Martin Wilk

- The test may be statistically significant from a normal distribution in any large samples. Thus a Q–Q plot is required for verification in addition to the test.

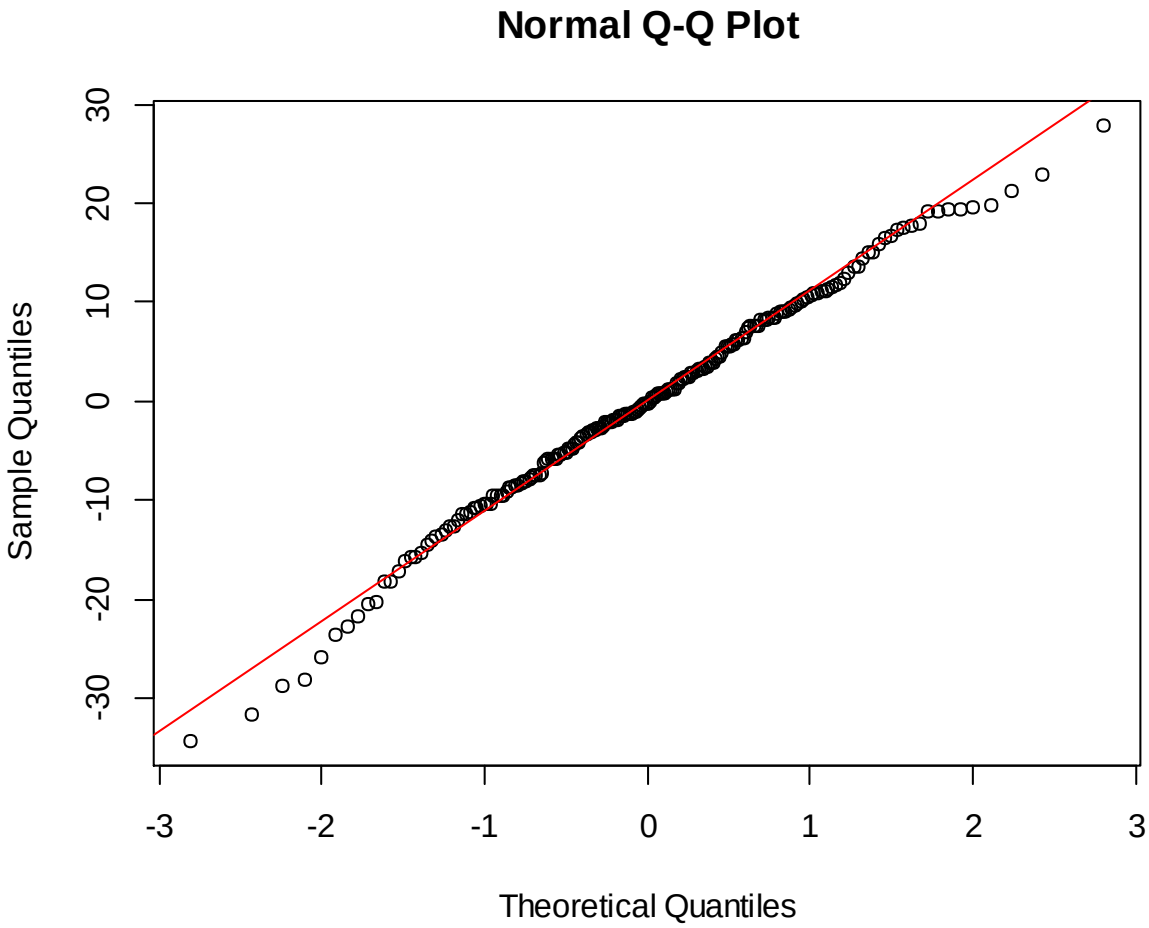
```
shapiro.test(rnorm(5000, mean = 5, sd = 100))
shapiro.test(runif(30, min = 2, max = 4))
```

```
Shapiro-Wilk normality test
data: rnorm(5000, mean = 5, sd = 100)
W = 0.99934, p-value = 0.06477
```

Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3–4): 591–611.

Quantile-Quantile Plots

```
y <- rnorm(200, sd = 10)
qqnorm(y)
qqline(y, col = 2)
```



Central limit theorem

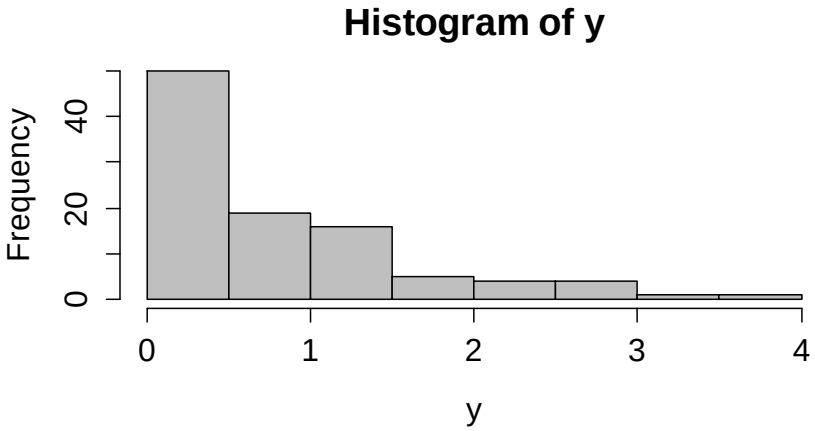
- Sampling distribution of means becomes normal as N increases, regardless of shape of original distribution.

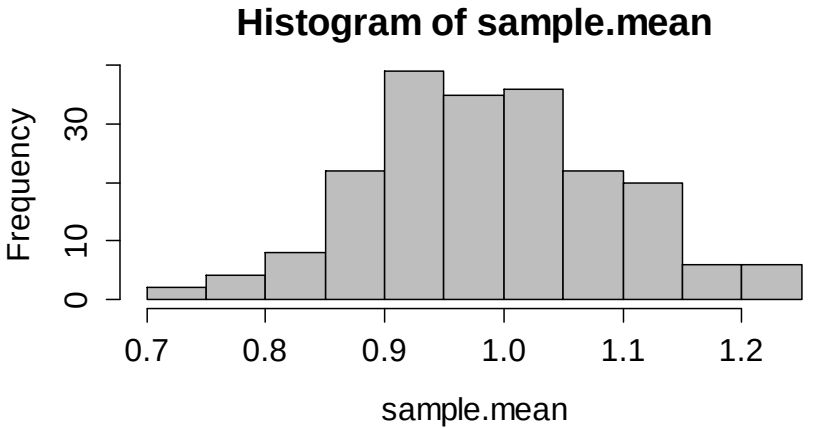
```
# The central limit theorem
# Exponential distribution is a skewed distribution
y <- rexp(100); hist(y, col = 'grey')

# Define a vector with length 200
sample.mean <- numeric(200)

# The mean of 200 samples
for(i in 1:200) sample.mean[i] <- mean(rexp(100))

# A normal distribution appears
hist(sample.mean , col = 'grey')
shapiro.test(sample.mean)
```





Null hypothesis (denoted H_0)

- A statement that the value of a population parameter (such as proportion or mean) is equal to some claimed value, or has no change.

- The original claim includes equality

  (<=, =, or >=)

Null hypotheses

- Some null hypotheses may be:
  - there is no difference between the height of the male and female students
  - there is no difference in the location (distance to downtown) of superstores and small grocers shops

Alternative hypothesis (denoted H_1 or H_a)

- A statement that the value of a population parameter differs from the null hypothesis.
- The symbolic form must be a >, < or ≠ statement.

Understanding hypothesis

- Null hypothesis
  - Results are due to “chance” (H_0)
- Alternative hypothesis
  - Results are due to a true “effect” (H_1)
- Assess
  - Assuming H_0 is true, what is the probability or “chance” of obtaining the data we did?
- Decide
  - If the chance is small enough, reject H_0 and infer the “effect” is real.

Test statistics

A value computed from the sample data, used in making the decision whether or not to reject the null hypothesis.

Test statistics

Z value for proportion

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Z value for mean (sigma known)

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

T value for mean (sigma unknown)

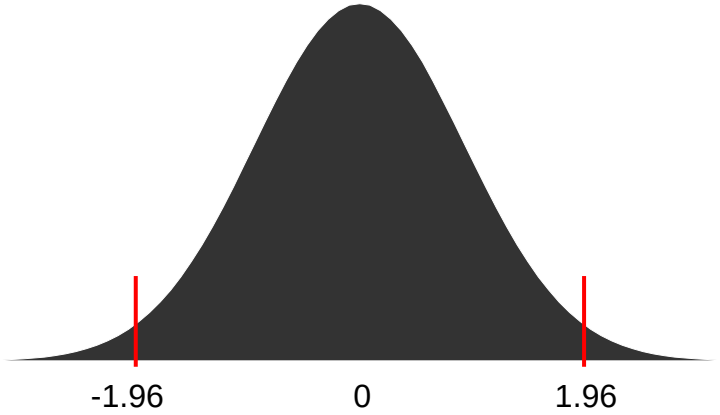
$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

The test statistic indicates how far our sample deviates from the assumed population parameter.

Z test statistic

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

- Test continuous outcome
- Known variance
- Under H_0 $\quad \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$
- Therefore,



$$\text{Reject } H_0 \text{ if } \left|\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}\right| > 1.96 \text{ (gives a 2-sided } \alpha = 0.05 \text{ test)}$$

$$\text{Reject } H_0 \text{ if } \bar{X} > \mu_0 + 1.96\frac{\sigma}{\sqrt{n}} \text{ or } \bar{X} < \mu_0 - 1.96\frac{\sigma}{\sqrt{n}}$$

pnorm(z)
t.test(x, mu=0, var.equal = T)

T test statistic

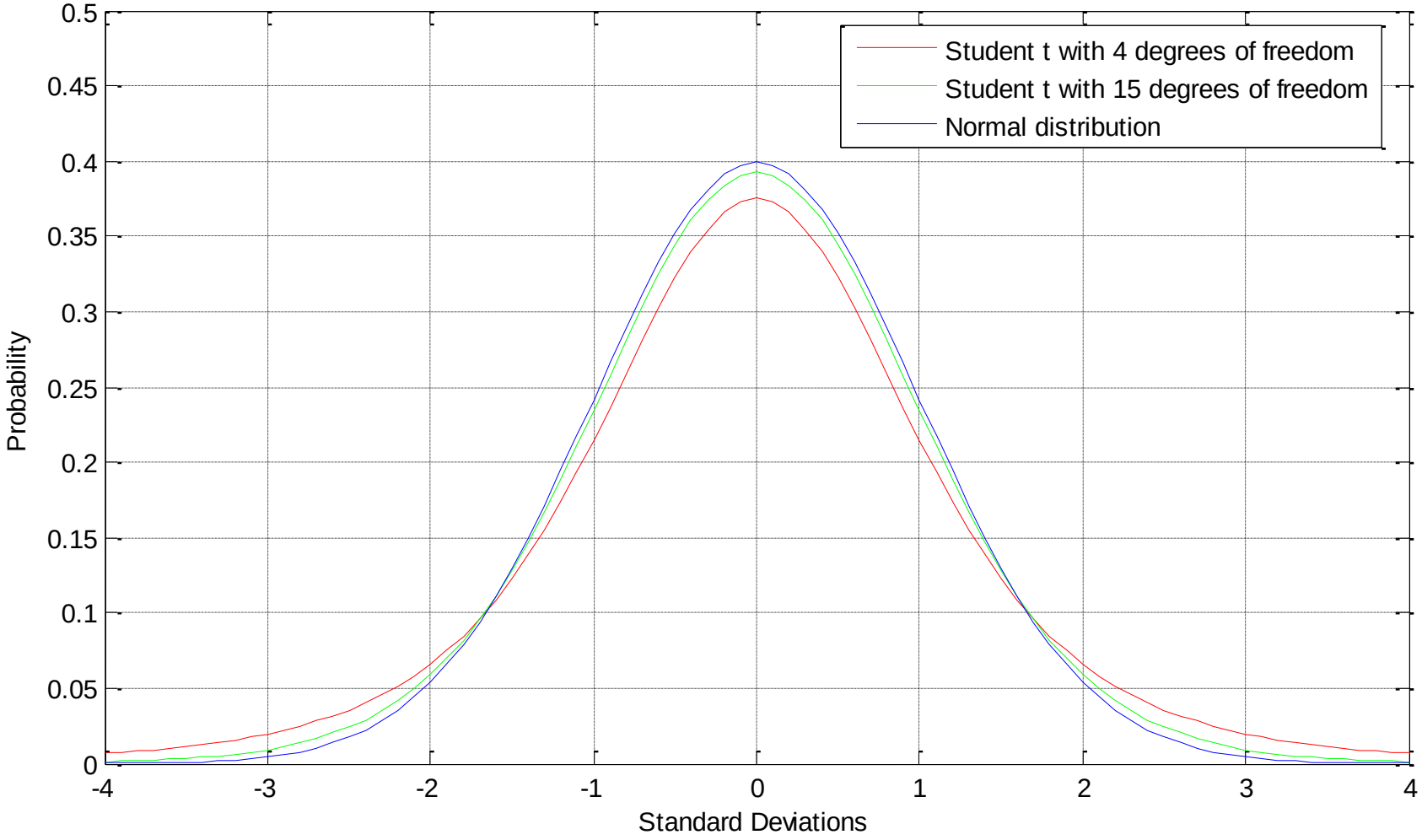
- Test continuous outcome
- Unknown variance
- Under H_0

$$\frac{\bar{X} - \mu_0}{s / \sqrt{n}} \sim t_{(n-1)}$$

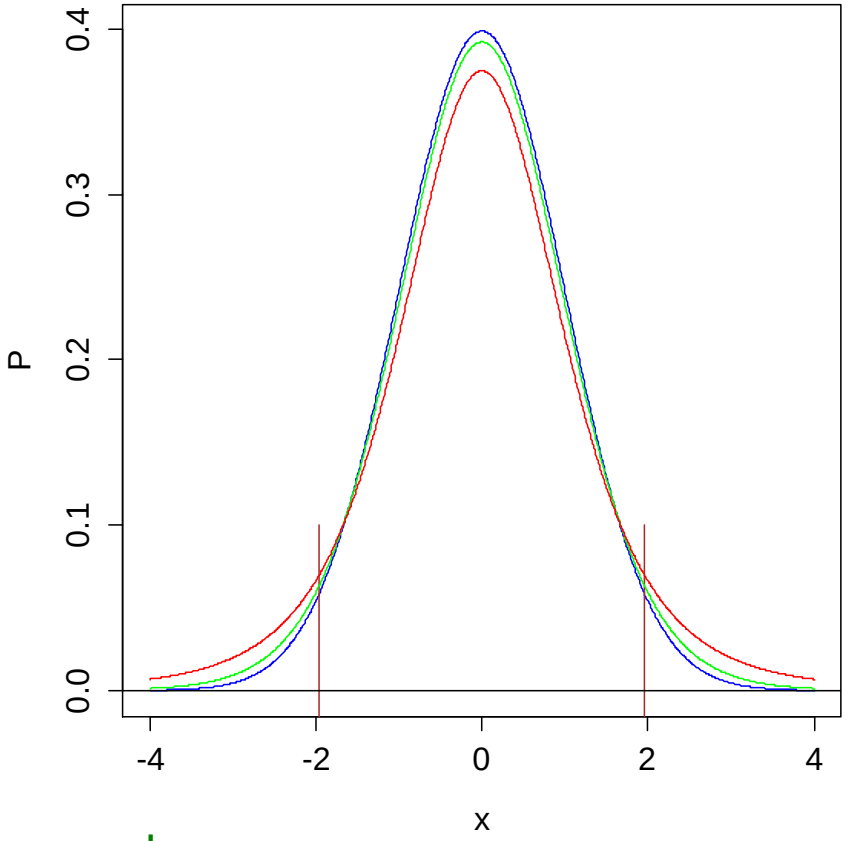
- Critical values: depend on degree of freedom, from computer or tables in statistics books
- t-distribution approximately normal for degrees of freedom (df) >30

t.test(x, mu=0, var.equal = F)

Comparing the student-t distribution to the normal distribution



Plot




```
# Plot a normal distribution and t distributions
x <- seq(-4, 4, len = 1000)
z <- dnorm(x, 0, 1)
plot(x, z, type = 'l', col = 'blue') # the normal curve
segments( 1.96, -0.05,  1.96, 0.1, col = 'brown')
segments(-1.96, -0.05,-1.96, 0.1,  col = 'brown')
abline(0,0)
lines(x, dt(x, 15), col = 'green')
lines(x, dt(x, 4),   col = 'red')
```

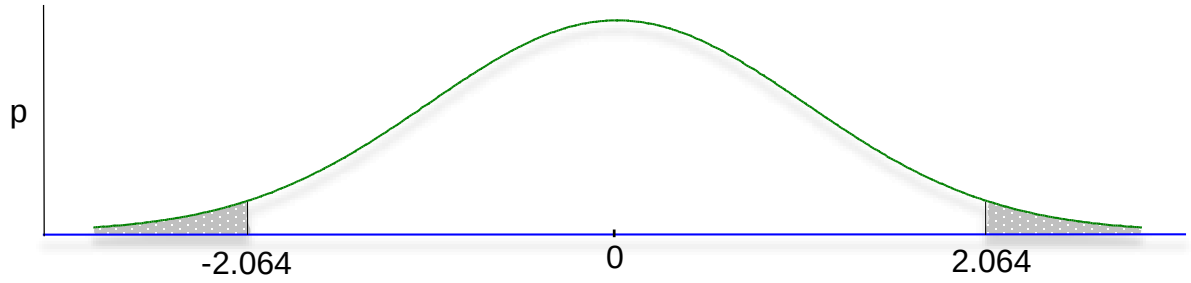
Critical values (quartiles) of the *t* distribution

The *t* table

DF		*P*									
	One tail	0.25	0.20	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
v	Two tails	0.50	0.40	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1		1.000	1.376	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2		0.816	1.061	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3		0.765	0.978	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4		0.741	0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5		0.727	0.920	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6		0.718	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7		0.711	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8		0.706	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9		0.703	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10		0.700	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
21		0.686	0.859	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22		0.686	0.858	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23		0.685	0.858	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24		0.685	0.857	1.318	1.711	**2.064**	2.492	2.797	3.091	3.467	3.745
25		0.684	0.856	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725

$t_{0.05,24} = 2.064$

$P(|t| \geq 2.064) = 0.05$



Two-tailed, right-tailed, left-tailed tests

The tails in a distribution are the extreme regions bounded by critical values.

Two-tailed Test

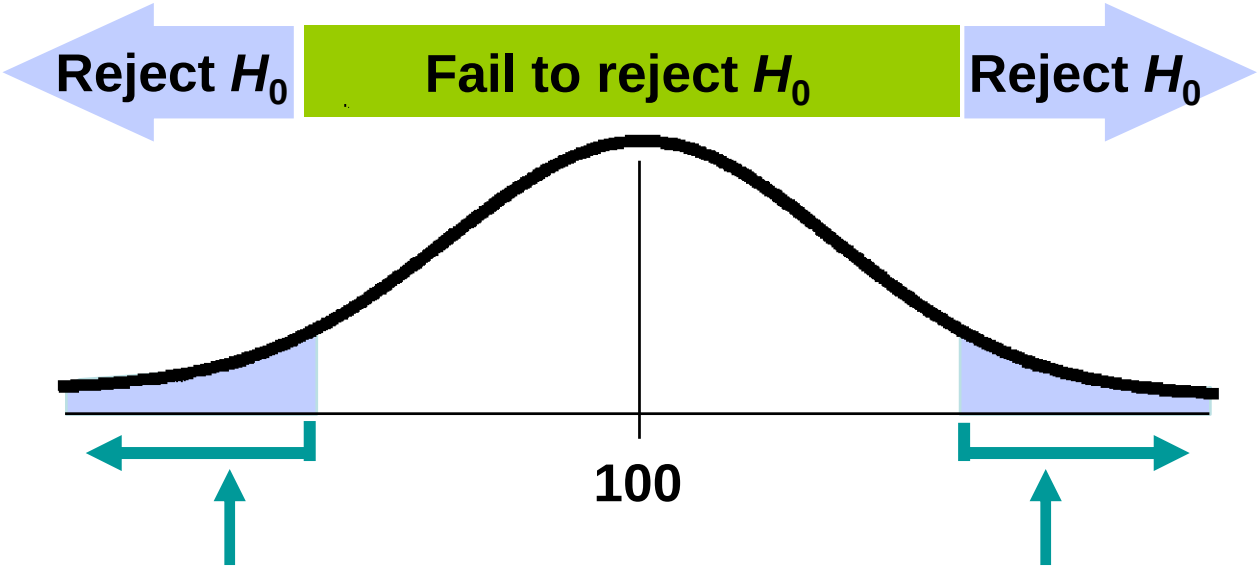
```
t.test(x, mu = 100, alt = "two.sided")
```

H_0: μ = 100

H_1: μ ≠ 100

α is divided equally between the two tails of the critical region

UNEQUAL means less than or greater than

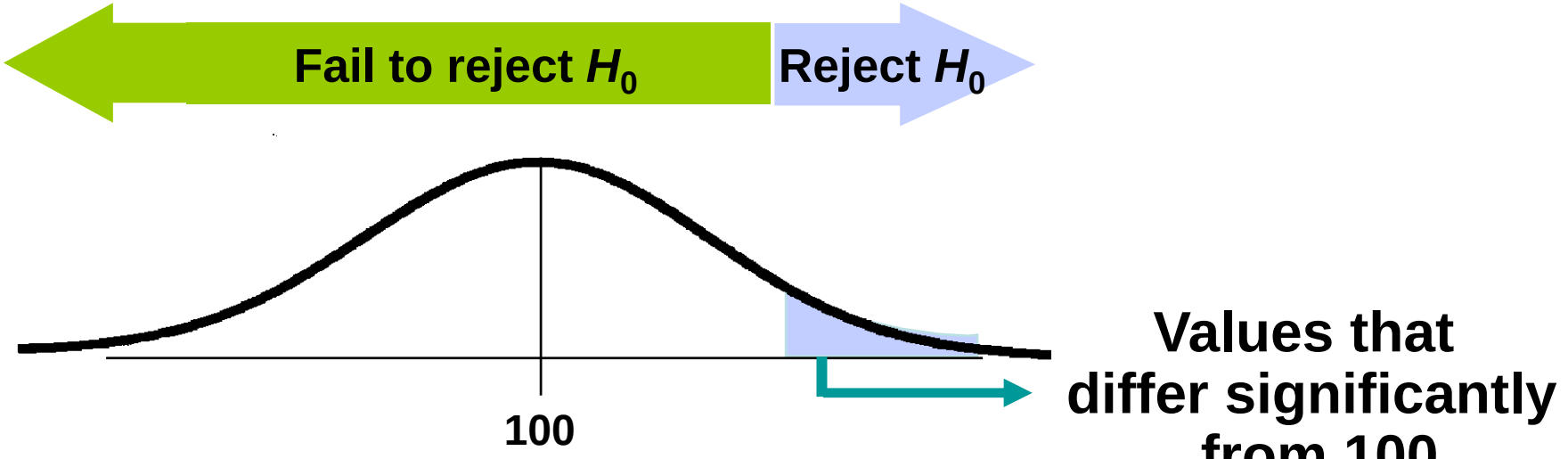


Right-tailed Test

t.test(x, mu=100, alt = “greater")

H_0: μ ≤ 100

H_1: μ > 100



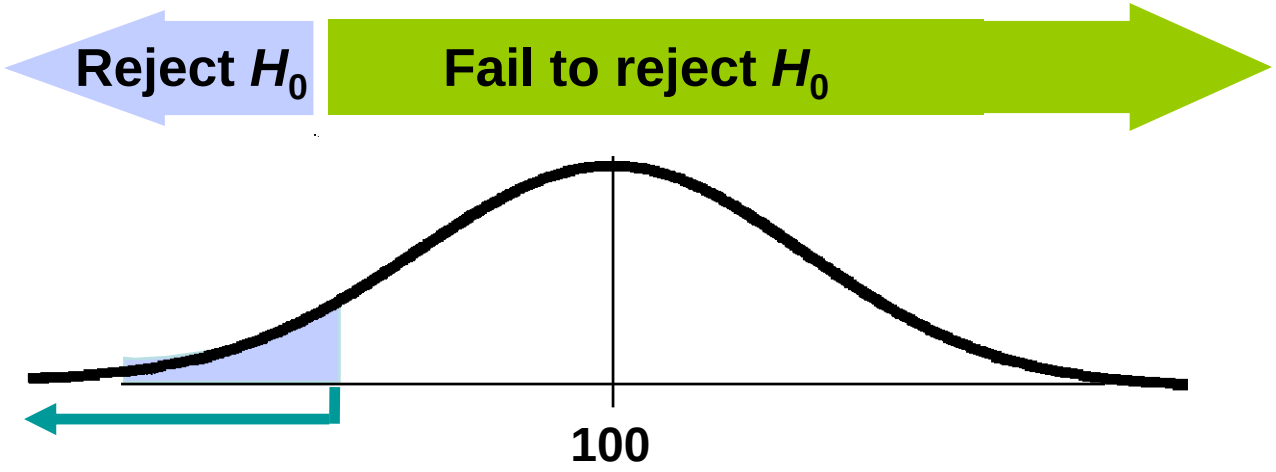
Values that differ significantly from 100

Left-tailed Test

t.test(x, mu=0, alt = “less")

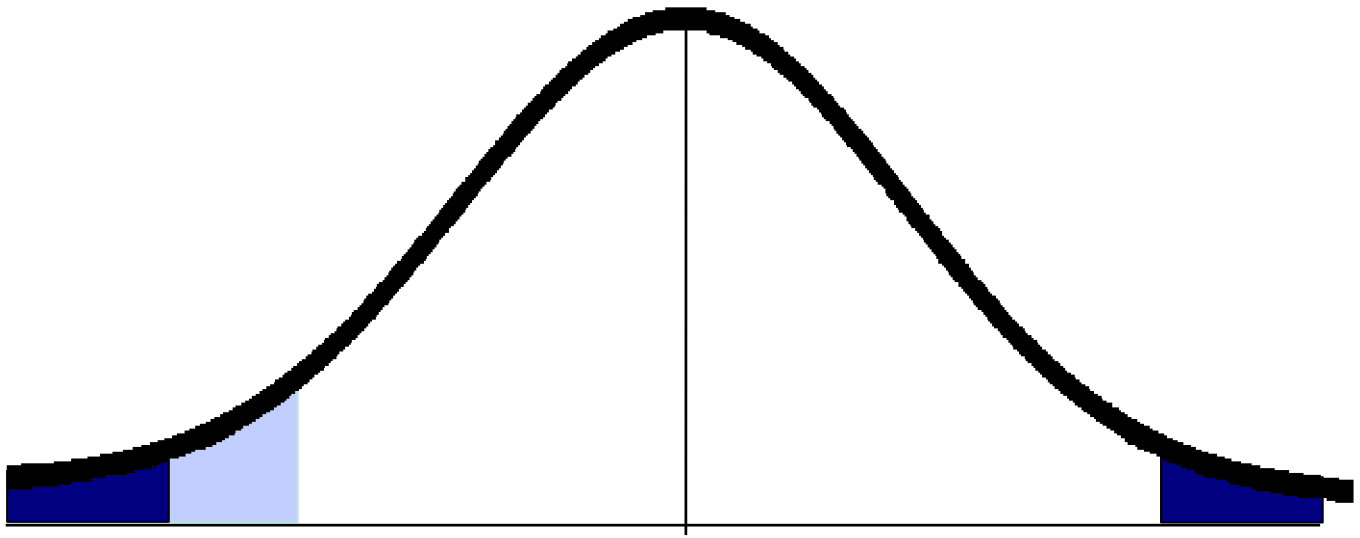
H_0: μ ≥ 100

H_1: μ < 100



Values that differ significantly from 100

Advantage of one-tailed test



P value

- The probability of getting a value of the test statistic that is at least as extreme as the one obtained for the sample data.
- If the P value is very small (such as less than 0.05), we will reject the null hypothesis.
- Measure of the strength of evidence in the data that the null is not true
- A variable whose value lies between 0 and 1

Significance level

- Denoted by α
- The probability that the test statistic will fall in the critical region when the null hypothesis is actually true.
- Common choices are 0.05, 0.01, and 0.10

How to state

If the original claim contains equality (became H_0)

- Reject H_0: “There is sufficient evidence to reject the claim that…”
- Fail to Reject H_0: “There is not sufficient evidence to reject the claim that…”

If the original claim does not contain equality (was H_1)

- Reject H_0: “The sample data support the claim that…”
- Fail to Reject H_0: “There is not sufficient sample evidence to support the claim that…”

Don’t use “Accept”

- Some texts use “accept the null hypothesis”, which is misleading, implying incorrectly that the null has been proven
- We are not proving the null hypothesis (can’t **PROVE** equality)
- The phrase ‘fail to reject’ represents the result more correctly.

If the sample evidence is not strong enough to warrant rejection, then the null hypothesis may or may not be true (just as a defendant found NOT GUILTY may or may not be innocent)

Decisions and conclusions

P-value method

Reject H_0 if P-value ≤ α

Fail to reject H_0 if P-value > α

Other methods

Give P-value, and leave conclusion to the reader

Look at whether population parameter falls in confidence interval estimate

One sample hypothesis test

- Faculty salary equity is a hot political issue: some argue that **earnings of unionized faculty** are considerably **higher** than the **earnings of faculty nationwide**.
- Assume that we know the national mean salary of a full-time faculty member is $45,000 and that salaries are normally distributed.
- We take a random sample of 23 unionized faculty members and find that they make on average $47,000 with a standard deviation of $13,200.
- Do unionized faculty members make significantly more a year than all faculty members in the population? Set α = .01

Parameters

$$\mu = \$45{,}000$$

$$\bar{x} = \$47{,}000$$

$$s = \$13{,}200$$

$$N = 23$$

Step one

We assume

- random sampling,
- level of measurement is interval-ratio,
- the sampling distribution is normal

Step two

$$H_0: \mu_{unionized} = \$45{,}000$$

$$H_1: \mu_{unionized} > \$45{,}000$$

Step three

- Sampling distribution = t distribution
- α = .01
- df = N – 1 = 22
- $t_{critical}$ = 2.508

```
qt(0.01, 22) # -2.508
```

Step four

$$t = \frac{\bar{x} - \mu}{s / \sqrt{N}}$$

$$t = \frac{47{,}000 - 45{,}000}{13{,}200 / \sqrt{23}}$$

$$t = .73$$

Step five

- We fail to reject the null hypothesis.
- We are 99% confident that there is no difference between the income of unionized faculty members and all faculty members in the population.

One sample hypothesis tests: proportions

- We know that 80% of the population has at least a high school diploma (U.S. Census Bureau 1990).
- Previous studies have shown that the South is economically disadvantaged and that Southerners have less overall education than the rest of the population.
- We randomly sampled 300 Southerners and found that 77% had a high school diploma at least .
- Are Southerners less educated than the population at large? Set $\alpha = .01$

Parameters

$$P_u = .80 \qquad P_s = .77$$

$$N = 300$$

Step one

We assume

- random sampling
- the proportion is interval-ratio
- the sampling distribution is normal

Step two

$$H_0 : P_s = .80$$

$$H_1 : P_s < .80$$

Step three

- Sampling distribution = Z distribution
- α = .01
- $Z_{critical}$= -2.33

qnorm(0.01)

Step four

$$Z = \frac{P_s - P_u}{\sqrt{P_u(1-P_u)/N}}$$

$$Z = \frac{.77 - .80}{\sqrt{.80(1-.80)/300}}$$

$$Z = -1.30$$

R script - one sample hypothesis tests: proportions

```
prop.test(x, n, p=null,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

```
prop.test(0.77*300, 300, p=0.8)
```

```
 1-sample proportions test with continuity correction

data:  0.77 * 300 out of 300, null probability 0.8
X-squared = 1.5052, df = 1, p-value = 0.2199
alternative hypothesis: true p is not equal to 0.8
95 percent confidence interval:
 0.7173811 0.8155549
sample estimates:
   p
0.77
```

Step five

- We fail to reject the null hypothesis.
- We are 99% confident that there is no difference between Southerners and the rest of the U.S. population with respect to the proportion owning a high school diploma.

Assignment

General objectives: **learn** about hypothesis testing, t test, and R.

You have a number of **specific** objectives in this assignment. I will provide you with the programs that you will use to test hypotheses and data that **you** develop.

For the report, you will provide a **brief introduction** to the data set, formally state the hypotheses that you are going to test (Ho's and Ha's), and provide a print out's of the data set, programs and their output. Indicate in your **results and discussion** section what you found, i.e. did you reject your null, and the conclusions that you have drawn from the analysis.

Tasks

Develop a t test experimental design. Generate your own data and FORMALIZE your hypotheses.

Make sure that you go through all of the steps; esp. satisfy assumptions:

- random sampling
- the level of measurement is interval-ratio
- the sampling distribution is normal

R script

```
# input data
lines <-
"ID  x  y
1  208  197
2  202  150
3  203  255
4  200  134
5  205  266
6  206  200
7  207  189
8  208  186
9  203  215
10  210  199"
weight <- read.table(con <- textConnection(lines), header=TRUE)
close(con)
# check data
weight
head(weight)
```

```
# t test
t.test(weight$x, weight$y, var.equal = F) # two samples
t.test(weight$x, mu = 200) # one sample
shapiro.test(weight$x) # check normal distribution
shapiro.test(weight$y)
hist(weight$y)
```