Confidence Interval and Hypothesis Testing

Confidence Interval estimate the Population Mean μ (with σ unknown)

- the population standard deviation is unknown
- the underlying population is approximately normal,
- we use the *t* distribution.

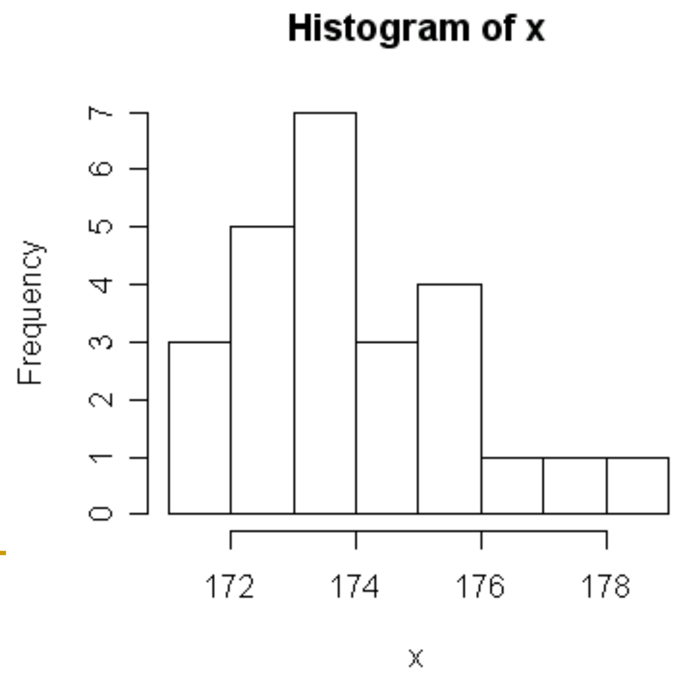
$$\overline{X} \pm t_{df,\alpha/2} \frac{s}{\sqrt{n}}$$

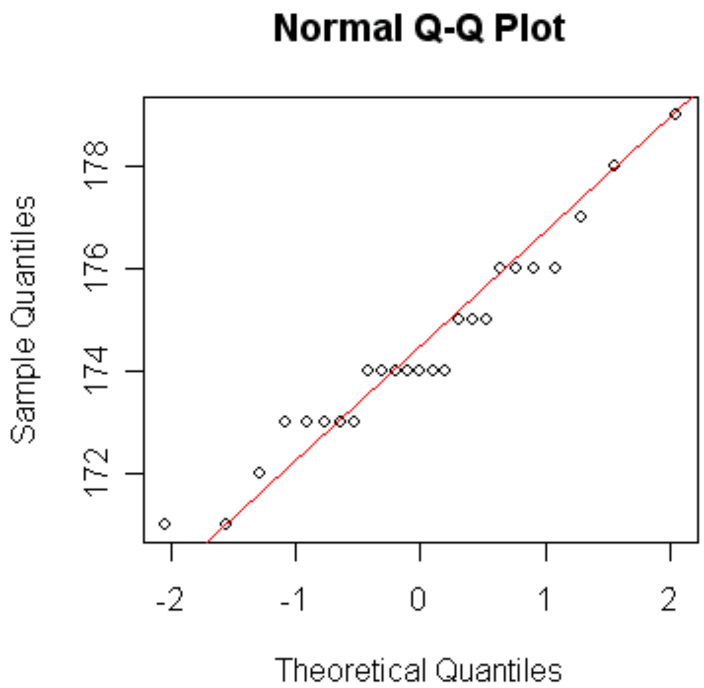
The value of *t* for a given confidence level depends upon its degrees of freedom.

Example: Suppose a person weighs himself on a regular basis and finds his weight to be
174 173 174 174 173 174 176 174 177 171 174 173 173 175 175 172 178 173 174 176 176 171 176 175 179

Construct a 95% confidence interval estimate for weighs. (Use R)

Solution: When use t distribution, the data should be normal or approximately normal.





```
x<-c(174, 173, 174, 174, 173, 174, 176, 174, 177, 171, 174, 173, 173, 175, 175,
172, 178, 173, 174, 176, 176, 171, 176, 175, 179)
n<-length(x)
xbar<-mean(x)
se<-sd(x)/sqrt(n)
alpha<-1-0.95
zvalue<-qt(1-alpha/2,n-1)
ci<-xbar+c(-zvalue*se,zvalue*se)
```

Alternative:

```
t.test(x)
```

```
 One Sample t-test

data:  x
t = 445.3774, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 173.5918 175.2082
sample estimates:
mean of x
    174.4
```

Confidence Interval for a Population Proportion

If the sample size n is large*, then a **(1–α)100% confidence interval for p** is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ $=$ **sample proportion** $= \dfrac{X}{n}$

$= \dfrac{\textbf{number of items having the characteristic}}{\textbf{sample size}}$

p $=$ **population proportion**

Z $=$ **critical value from the standardized normal distribution**

n $=$ **sample size**

assuming both X and (1-p) $\times\, n$ **are greater than 5**

Example: A sample of 500 executives who own their own home revealed 175 planned to sell their homes and retire to Arizona. Develop a 98% confidence interval for the proportion of executives that plan to sell and move to Arizona

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.35 \pm 2.33\sqrt{\frac{(.35)(.65)}{500}} = .35 \pm .0497$$

Use R?

Alternative: (try it!)

Hypothesis Testing: One-Sample Tests

Five Step Model for Hypothesis Tests

Step 1: State null and alternate hypotheses

Step 2: Select a level of significance

Step 3: Identify the test statistic

Step 4: Formulate a decision rule

Step 5: Take a sample, arrive at a decision

Do not reject null

Reject null and accept alternate

Example:

Suppose a car manufacturer claims a model gets 25 mpg. A consumer group asks 10 owners of this model to calculate their mpg and the mean value was 22 with a standard deviation of 1.5. Is the manufacturer's claim supported?

```
## Compute the t statistic. Note we assume mu=25 under H_0
> xbar=22;s=1.5;n=10
> t = (xbar-25)/(s/sqrt(n))
> t
[1] -6.324555
## use pt to get the distribution function of t
> 2*pt(t,df=n-1)
[1] 0.0001369366
```

This is a small p-value (9). The manufacturer's claim is suspicious.

t.test() In R

There is an example concerning daily energy intake in KJ for 11 women. The value are placed in a data vector

```
daily.intake<-c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)
```

Investigate whether the women's energy intake deviates systematically From a recommended value of 7725KJ.
Assuming that data comes from A normal distribution, the object is to test whether this distribution might have mean 7725. This is done with **t.test**, as follow

```
t.test(daily.intake,mu=7725)
```

Example:

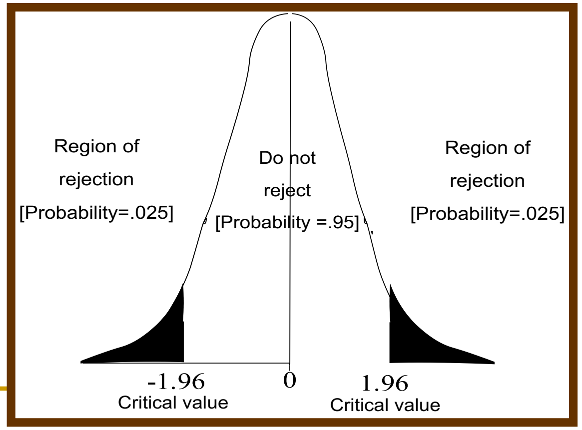
There is an example concerning daily energy intake in KJ for 11 women. The value are placed in a data vector

daily.intake<-c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)

Investigate whether the women's energy intake deviates systematically from a recommended value of 7725KJ.

Solution:

$H_0 = 7725 \qquad H_1 \neq 7725$



t.test() In R

> t.test(daily.intake,mu=7725)

```
        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
```

Three arguments in function t.test() are relevant in one-sample problems

t.test(x, mu=, alternative=“”,conf.level=)

x: a numeric vector of data values

mu: a number indicating the vaule of the mean under null hypothesis

alternative: a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less".

Conf.level: confidence level of the interval

Exercise:

In recent years the interest rate on home mortgages has declined to less than 6.0 percent. However, according to a study by Federal Reserve Board the rate charge on credit card debit is more than 14 percent. Listed below is the interest rate charged on a sample of 10 credit cards.

14.6 16.7 17.4 17.0 17.8 15.4 13.1 15.8
14.3 14.5

Is it reasonable to conclude the mean rate charged is greater than 14 percent? Use the 0.01 significance level.

One-Sample Proportion Test

The fraction or percentage that indicates the part of the population or sample having a particular trait of interest.

$$\hat{p} = \frac{\text{Number of successes in the sample}}{\text{Number sampled}}$$

The sample proportion $\hat{p}$ is and p is the population proportion.

Test Statistic for Testing a Single Population Proportion

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Example: NSC

For a Christmas and New Year′s week, the National Safety Council estimated that 500 people would be killed and 25,000 injured on the nation′s roads. The NSC claimed that 50% of the accidents would be caused by drunk driving.

A sample of 120 accidents showed that 67 were caused by drunk driving. Use these data to test the NSC′s claim with $\alpha = 0.05$.

Example: NSC

- Hypothesis H_0: $p = .5$ H_1: $p \neq 0.5$
- Test Statistic

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{.5(1-.5)}{120}} = .045644$$

$$z = \frac{\bar{p}-p_0}{\sigma_{\bar{p}}} = \frac{(67/120)-.5}{.045644} = 1.278$$

- Rejection Rule

  Reject H_0 if $z < -1.96$ or $z > 1.96$

- Conclusion

  Do not reject H_0. For $z = 1.278$, the p-value is .201.

How to get the p-value using R?

prop.test() in R

```
> prop.test(67,120,p=.5,correct =F)
```

```
  1-sample proportions test without continuity correction

data:  67 out of 120, null probability 0.5
X-squared = 1.6333, df = 1, p-value = 0.2012
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4690452 0.6440025
sample estimates:
        p
0.5583333
```

X-squared = 1.6333= $(1.278)^2 = z^2$

prop.test(x, p=, alternative="",conf.level=)

Hypothesis Testing: Two-Sample Tests

Two-Sample Mean Tests

Two Population Means, Independent Samples

Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
i.e.,	i.e.,	i.e.,
$H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$	$H_0: \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 - \mu_2 > 0$	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$

Under two respective normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.

If **σ_1, σ_2 are unknown**, a *t*-test is suggested.

Comparing Means of Two **Independent** Populations

$\sigma_1 = \sigma_2$ are unknown and assume equal

- In most cases the variances are unknown. The test statistic for $\mu 1 - \mu 2$ is:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where $\quad s_p^2 = \text{pooled variance} = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$

$\overline{X}_1 = \text{mean of the sample taken from population 1}$

$s_1^2 = \text{variance of the sample taken from population 1}$

$n_1 = \text{size of the sample taken from population 1}$

$\overline{X}_2 = \text{mean of the sample taken from population 2}$

$s_2^2 = \text{variance of the sample taken from population 2}$

$n_2 = \text{size of the sample taken from population 2}$

Comparing Means of Two **Independent** Populations

σ1 and σ2 unknown, not assumed equal. The test statistic for μ1 – μ2 is:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

This t- test statistics approximately follows a t distribution with the degrees of freedom is the integer portion of:

$$\nu = \frac{\left(\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}\right)^2}{\frac{\left(\frac{S_1^{\,2}}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^{\,2}}{n_2}\right)^2}{n_2 - 1}}$$

Example: Recovery time for new drug

Suppose the recovery time for patients taking a new drug is measured (in days). A placebo group is also used to avoid the placebo effect. The data are as follows

```
with drug:  15 10 13  7  9  8 21  9 14  8
placebo:    15 14 12  8 14  7 16 10 15 12
```

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 < \mu_2$$

```
> x = c(15, 10, 13, 7, 9, 8, 21, 9, 14, 8)
> y = c(15, 14, 12, 8, 14, 7, 16, 10, 15, 12)
> t.test(x,y,alt="less")
```

Welch Two Sample t-test
data: x and y
t = -0.5331, df = 16.245, p-value = 0.3006
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf 2.044664
sample estimates:
mean of x mean of y
11.4 12.3

We accept the null hypothesis because p-value is much larger than 10%

```
> x = c(15, 10, 13, 7, 9, 8, 21, 9, 14, 8)
> y = c(15, 14, 12, 8, 14, 7, 16, 10, 15, 12)
> t.test(x,y,alt="less",var.equal=T)
```

```
Two Sample t-test

data:  x and y
t = -0.5331, df = 18, p-value = 0.3002
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf 2.027436
sample estimates:
mean of x mean of y
    11.4      12.3
```

Example

Refer to the coutry.txt data, Conduct a *t* test of hypothesis to determine if there is a difference in the mean birthrate of developed and developing country. **>country<-read.table("a:/country.txt",header=T)** **>attach(country)**

Code:
```
>birthrate.developed<-birthrate[develop==0]
>birthrate.developing<-birthrate[develop==1]
>t.test(birthrate.developed, birthrate.developing)
```

Or
```
>t.test(birthrate~develop,data=country)
```

```
Welch Two Sample t-test
data:  birthrate.developed and birthrate.developing
t = -11.4663, df = 60.953, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -24.34388 -17.11388
sample estimates:
mean of x mean of y
 15.35714  36.08602
```

Reject H_0

Comparing Means of Two **Dependent** Populations

- **Independent samples are samples that are not related in any way.**
- **Dependent samples are samples that are paired or related in some fashion.**

Two Dependent Populations: Examples

- 1. In order to measure the effectiveness of a new diet, we would weigh the dieters at the start and at the finish of the program.
- 2. Nike wants to see if there is a difference in durability of 2 sole materials. One type is placed on one shoe, the other type on the other shoe of the same pair.

Comparing Means of Two **Dependent** Populations

Two types of dependent samples

The same subjects measured at two different points in time.

Matched or paired observations

Comparing Means of Two **Dependent** Populations

Use the following test when the samples are **dependent:**

$$\boxed{t = \frac{\bar{d}}{s_d / \sqrt{n}}}$$

where $\bar{d}$ is the mean of the differences

s_d is the standard deviation of the differences

n is the number of pairs (differences)

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}}$$

Example:

An independent testing agency is comparing the daily rental cost for renting a compact car from Hertz and Avis. A random sample of eight cities revealed the following information. At the .05 significance level can the testing agency conclude that there is a difference in the rental charged?

City	Hertz	Avis	*d*	d^2
Atlanta	42	40	2	4
Chicago	56	52	4	16
Cleveland	45	43	2	4
Denver	48	48	0	0
Honolulu	37	32	5	25
Kansas City	45	48	-3	9
Miami	41	39	2	4
Seattle	46	50	-4	16

Step 1 H_o: $m_d = 0$ H_1: $m_d \neq 0$

Step 2 The stated significance level is .05.

Step 3 The appropriate test statistic is the paired t-test.

Step 4 $H0$ is rejected if $t < -2.365$ or $t > 2.365$; or if p-value < .05.

We use the t distribution with 7 degrees of freedom.

Step 5 Perform the calculations and make a decision.

$$\bar{d} = \frac{\Sigma d}{n} = \frac{8.0}{8} = 1.00 \qquad s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{78 - \frac{8^2}{8}}{8-1}} = 3.1623$$

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{1.00}{3.1623 / \sqrt{8}} = 0.894$$

P(|t|>.894) = .40 for two one-tailed t-test at 7 degrees of freedom.

Because 0.894 is less than the critical value, the p-value of .40 > .05, do not reject the null hypothesis. There is no difference in the mean amount charged by Hertz and Avis.

t.test(…,paired=T) In R

```
> x<-c(42,56,45,48,37,45,41,46)
> y<-c(40,52,43,48,32,48,39,50)
> t.test(x,y,paired=T)
```

```
 Paired t-test

data:  x and y
t = 0.8944, df = 7, p-value = 0.4008
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.643730  3.643730
sample estimates:
mean of the differences
                      1
```

Comparing Two Population Proportions

Hypothesis for Two Population Proportions:

Lower-tail test:	Upper-tail test:	Two-tail test:
H_0: $p_1 \geq p_2$ H_a: $p_1 < p_2$ i.e., H_0: $p_1 - p_2 \geq 0$ H_a: $p_1 - p_2 < 0$	H_0: $p_1 \leq p_2$ H_a: $p_1 > p_2$ i.e., H_0: $p_1 - p_2 \leq 0$ H_a: $p_1 - p_2 > 0$	H_0: $p_1 = p_2$ H_a: $p_1 \neq p_2$ i.e., H_0: $p_1 - p_2 = 0$ H_a: $p_1 - p_2 \neq 0$

Testing the Difference of Two Population Proportions

The test statistic for $p_1 - p_2$ is a Z statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sigma_{\hat{p}_1 - \hat{p}_2}}$$

- Estimate $\sigma_{\hat{p}_1 - \hat{p}_2}$ by $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$$\hat{p} = \frac{\text{the total number of units in the two samples that fall into the category of interest}}{\text{the total number of units in the two samples}}$$

Example

- In a random sample, 36 of 72 men and 31 of 50 women indicated they would vote Yes
- Is there a significant difference between the proportion of men and the proportion of women who will vote Yes on Proposition A?
- Test at the .05 level of significance

- The hypothesis test is:

H_0: $p_1 - p_2 = 0$ (the two proportions are equal)
H_a: $p_1 - p_2 \neq 0$ (there is a significant difference between proportions)

Two population Proportions (continued)

- The sample proportions are:
  - Men: $\hat{p}_1$ = 36/72 = .50
  - Women: $\hat{p}_2$ = 31/50 = .62

- The pooled estimate for the overall proportion is:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{36+31}{72+50} = \frac{67}{122} = .549$$

The test statistic for $p_1 - p_2$ is:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{(.50 - .62) - (0)}{\sqrt{.549(1-.549)\left(\frac{1}{72} + \frac{1}{50}\right)}} = -1.31$$

Decision: Do not reject H_0

Conclusion: There is not significant evidence of a difference in proportions who will vote yes between men and women.

The p-value for this test is

$2 \times P(z<-1.31) = 0.1902$

prop.test(c(36,31),c(72,50), correct=F)

```
        2-sample test for equality of proportions without continuity correction

data:  c(36, 31) out of c(72, 50)
X-squared = 1.7163, df = 1, p-value = 0.1902
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.29731146  0.05731146
sample estimates:
prop 1 prop 2
  0.50   0.62
```

prop.test(x, n, p=, alternative="",conf.level=)

x, vector of counts of successes
n, a vector of counts of trials;

Summary

Setting	test on mean	test on proportion
One sample	One sample *t*-test t.test(x, mu, alt=…)	One sample proportion test prop.test (x,n,p,alt=…)
Two independent sample	Two sample *t*-test t.test(x,y,alt…)	Two sample proportion test prop.test (x,n,alt=…)
Matched pairs	Two sample paired t-test t.test(x,y,paired=T,alt=)	