Stat 104: Quantitative Methods

Class 26: Comparing Two Groups

Comparing Two Groups

- ❑ Test the claim that the proportion of children who contract polio is less for children given the Salk vaccine than for children given a placebo
- ❑ Test the claim that subjects treated with Lipitor have a mean cholesterol level that is lower than the mean cholesterol level for subjects given a placebo.
- ❑ Test the claim that when college students are weighed at the beginning and end of their freshman year, the differences show a mean weight gain of 15 pounds (as in the "Freshman 15" belief).

Comparing Two Proportions

- We have two independent samples from populations of interest.
- Ask them the same question;
  - ❑ Did you floss this morning? (for example)

Notation for Two Proportions

For population 1, we let:

p_1 = population proportion

n_1 = size of the sample

x_1 = number of successes in the sample

$\hat{p}_1 = \dfrac{x_1}{n_1}$ (the sample proportion)

The corresponding notations apply to

$p_2, n_2, x_{2,}\ \hat{p}_2$, which come from population 2.

Requirements

1. We have proportions from two independent simple random samples.
2. For each of the two samples, the number of successes is at least 5 and the number of failures is at least 5.

The Confidence Interval

- From each sample we calculate the proportion of yes's:

$\hat{p}_1$

$\hat{p}_2$

Calculating the Confidence Interval

- Using the Central Limit Theorem, the 95% confidence interval for p_1-p_2 is given by

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- This assumes large sample sizes of course.

Example: Snap Fingers/Gender

- From a class survey we have data on male/female ability to snap fingers
- 40 out of 45 women can snap
- 119 out of 126 men can snap
- The R command is `prop.test`

R Output Using prop.test

```
> prop.test(c(40,119),c(45,126))

        2-sample test for equality of proportions with
continuity correction

data:  c(40, 119) out of c(45, 126)
X-squared = 0.83253, df = 1, p-value = 0.3615
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.17078906  0.05967795
```

Interpretation??

Example

- In 1982 and 1994, respondents in the General Social Survey were asked: “Do you agree or disagree with this statement? ‘Women should take care of running their homes and leave running the country up to men.’”

Year	Agree	Disagree	Total
1982	122	223	345
1994	268	1632	1900
Total	390	1855	2245

R output

```
> prop.test(c(122,268),c(345,1900))

2-sample test for equality of proportions with continuity
correction

data:  c(122, 268) out of c(345, 1900)
X-squared = 90.44, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1580372 0.2671039
```

Interpretation?

Example

- The Gallup organization surveyed 1100 adult Americans on May 6–9, 2002, and conducted an independent survey of 1100 adult Americans on May 3–6, 2010.
- In both surveys they asked the following: “Right now, do you think the state of moral values in the country as a whole is getting better or getting worse?” On May 3–6, 2010, 836 of the 1100 surveyed responded that the state of moral values is getting worse; on May 6–9, 2002, 737 of the 1100 surveyed responded that the state of moral values is getting worse.
- Construct and interpret a 95% confidence interval for the difference between the two population proportions.

R Output

```
> prop.test(c(836,737),c(1000,1000))

        2-sample test for equality of proportions with continuity correction

data:  c(836, 737) out of c(1000, 1000)
X-squared = 28.597, df = 1, p-value = 8.91e-08
alternative hypothesis: two.sided
95 percent confidence interval:
 0.06234508 0.13565492
```

- We are 95% confident that the percentage of adult Americans who believe that the state of moral values in the country as a whole was getting worse increased between 6% and 14% from 2002 to 2010.
- Because this interval does not contain 0, we might conclude that a higher proportion of the country believed that the state of moral values was getting worse in the United States in 2010 than in 2002.

Concussions in the NCAA

- Game exposures among college soccer players 1997-1999

Outcome	Concussion	No Concussion	Total
Gender			
Female	158	74924	75082
Male	101	75633	75734
Total	259	150557	150816

Interpret Output

- Conclusion?

```
> prop.test(c(158,101),c(75082,75734))

        2-sample test for equality of proportions with continuity correction

data:  c(158, 101) out of c(75082, 75734)
X-squared = 12.619, df = 1, p-value = 0.0003818
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0003391654 0.0012023363
```

Hypothesis Testing Two Proportions

We are often interested in comparing the proportions of two distinct groups relative to some attribute.

For example,

- ❑ We may be interested in comparing the proportion of defective units of a given product produced by two competing manufacturers.
- ❑ Or we may be interested in comparing the proportions of high school graduates from two different schools who attended college.

Decision Rules for Testing Two Proportions

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}, \; where \; \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$$

$H_0: p_1 = \mathrm{p}_2$ If $|T| > 1.96$ $reject\, H_o$

$H_a: p_1 \neq p_2$

$H_0: p_1 = p_2$ If $T < -1.64$ $reject\, H_o$

$H_a: p_1 < p_2$

$H_0: p_1 = p_2$ If $T > 1.64$ $reject\, H_o$

$H_a: p_1 > p_2$

Example: In July 1987 the Canadian parliament debated the reinstatement of the death penalty. One of the factors in this debate was the amount of public support for the death penalty. In 1982, a sample of 1500 Canadians reveled that 1015 favored the death penalty. In 1987, 915 in a sample of 1500 supported the death penalty. Do these data provide sufficient evidence at the 5% significance level to indicate that support has fallen between 1982 and 1987 ?

We want to test

$$H_o: p_{82} = p_{87} \quad vs. \quad H_o: p_{82} > p_{87}$$

The test statistic is

$$T = \frac{(.7 - .61)}{\sqrt{.655(1 - .655)(\frac{1}{1500} + \frac{1}{1500})}} = 5.18$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

We reject the null hypothesis if T>1.64 and 5.18 certainly is.

Hence we may conclude with 95% confidence that the level of support for the death penalty has declined from 1982 to 1987.

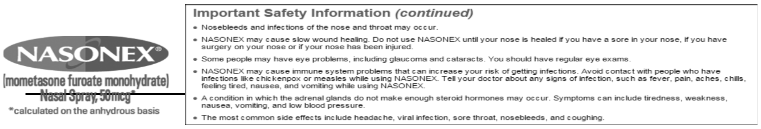
R Output

```
> prop.test(c(1015,915),c(1500,1500),alt="greater")

        2-sample test for equality of proportions with continuity
correction

data:  c(1015, 915) out of c(1500, 1500)
X-squared = 14.238, df = 1, p-value = 8.054e-05
alternative hypothesis: greater
95 percent confidence interval:
 0.03729934 1.00000000
sample estimates:
   prop 1    prop 2
0.6766667 0.6100000
```

Conclusion??

Example: Nasonex

- In clinical trials of Nasonex, 3774 adult and adolescent allergy patients (patients 12 years and older) were randomly divided into two groups.
- The patients in group 1 (experimental group) received 200 *μ*g of Nasonex, while the patients in group 2 (control group) received a placebo.
- Of the 2103 patients in the experimental group, 547 reported headaches as a side effect. Of the 1671 patients in the control group, 368 reported headaches as a side effect.


NASONEX®
(mometasone furoate monohydrate)
Nasal Spray, 50mcg*
*calculated on the anhydrous basis

Important Safety Information *(continued)*

- Nosebleeds and infections of the nose and throat may occur.
- NASONEX may cause slow wound healing. Do not use NASONEX until your nose is healed if you have a sore in your nose, if you have surgery on your nose or if your nose has been injured.
- Some people may have eye problems, including glaucoma and cataracts. You should have regular eye exams.
- NASONEX may cause immune system problems that can increase your risk of getting infections. Avoid contact with people who have infections like chickenpox or measles while using NASONEX. Tell your doctor about any signs of infection, such as fever, pain, aches, chills, feeling tired, nausea, and vomiting while using NASONEX.
- A condition in which the adrenal glands do not make enough steroid hormones may occur. Symptoms can include tiredness, weakness, nausea, vomiting, and low blood pressure.
- The most common side effects include headache, viral infection, sore throat, nosebleeds, and coughing.

Example: Nasonex

- Is there significant evidence to conclude that the proportion of Nasonex users that experienced headaches as a side effect is greater than the proportion in the control group?

```
> prop.test(c(547,368),c(2013,1671),alt="greater")

        2-sample test for equality of proportions with continuity correction

data:  c(547, 368) out of c(2013, 1671)
X-squared = 12.701, df = 1, p-value = 0.0001827
alternative hypothesis: greater
95 percent confidence interval:
 0.0276344 1.0000000
sample estimates:
   prop 1    prop 2
0.2717337 0.2202274
```

Practical vs Statistical Significance

- Looking back at the results, we notice that the proportion of individuals taking 200 mg of Nasonex who experience headaches is *statistically significantly* greater than the proportion of individuals 12 years and older taking a placebo who experience headaches.
- However, we need to ask ourselves a pressing question. Would you not take an allergy medication because 26% of patients experienced a headache taking the medication versus 22% who experienced a headache taking a placebo?
- Most people would be willing to accept the additional risk of a headache to relieve their allergy symptoms. While the difference of 5% is statistically significant, it does not have any *practical significance*.

Concussion Example Again

- Is there evidence that a higher percentage of female soccer players get concussions than male soccer players?

Look at the output-conclusion?

```
> prop.test(c(158,101),c(75082,75734),alt="greater")

        2-sample test for equality of proportions with continuity correction

data:  c(158, 101) out of c(75082, 75734)
X-squared = 12.619, df = 1, p-value = 0.0001909
alternative hypothesis: greater
95 percent confidence interval:
 0.0004064209 1.0000000000
sample estimates:
     prop 1      prop 2
0.002104366 0.001333615
```

Comparing Two Means (large sample)

- We want to compare the average age among two populations:

μ_1

μ_2

We will show how to construct a CI for $\mu_1 - \mu_2$

Notation

μ_1 = population mean

σ_1 = population standard deviation

n_1 = size of the first sample

$\overline{x}_1$ = sample mean

s_1 = sample standard deviation

Corresponding notations apply to population 2. (don’t need equal sample sizes)

Requirements

1. σ_1 and σ_2 are unknown and no assumption is made about the equality of σ_1 and σ_2.
2. The two samples are independent.
3. Both samples are simple random samples.
4. The two sample sizes are both large (over 30) and/or both samples come from populations having normal distributions.

Some technical issues

- A guess of μ_1-μ_2 is $\overline{\text{x}}_1$-$\overline{\text{x}}_2$
- Because we assume the two groups are independent,

$$Var(\overline{X}_1 - \overline{X}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

- We then turn to the Central Limit Theorem as before to form the confidence interval.

The Confidence Interval

A 95% confidence interval for $\mu_1 - \mu_2$ is given by

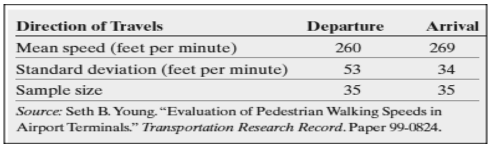
$$(\overline{x}_1 - \overline{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We assume that both sample sizes are bigger than 30.

(in general, the computer uses the t distribution, which is why this is often called a “t test”)

Example

- Do people walk faster in the airport when they are departing (getting on a plane) or when they are arriving (getting off a plane)?
- Researcher Seth B. Young measured the walking speed of travelers in San Francisco International Airport and Cleveland Hopkins International Airport. His findings are summarized in the table.



Direction of Travels	Departure	Arrival
Mean speed (feet per minute)	260	269
Standard deviation (feet per minute)	53	34
Sample size	35	35

Source: Seth B. Young. "Evaluation of Pedestrian Walking Speeds in Airport Terminals." *Transportation Research Record*. Paper 99-0824.

Confidence interval by hand

- The math isn't bad but a bit cumbersome

$$(\bar{x}_1-\bar{x}_2)\pm 1.96\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}=(260-269)\pm 1.96\sqrt{\frac{(53)^2}{35}+\frac{(34)^2}{35}}$$
$$=(-29.86,11.86)$$

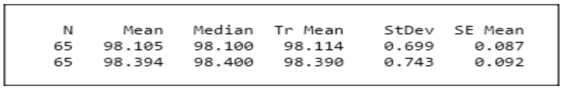
- Interpretation?

Using R

```
> tsum.test(n.x=35,mean.x=260,s.x=53,n.y=35,mean.y=269,s.y=34)

        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = -0.84558, df = 57.931, p-value = 0.4013
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -30.30597  12.30597
```

Example

- In a recent study, 65 men and 65 women were studied and their body temperature was recorded (everyone self reported being healthy).
- The data was as follows



	N	Mean	Median	Tr Mean	StDev	SE Mean
men	65	98.105	98.100	98.114	0.699	0.087
women	65	98.394	98.400	98.390	0.743	0.092

R

```
> tsum.test(n.x=65,mean.x=98.105,s.x=0.699,n.y=65,mean.y=98.394,s.y=0.734)

        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = -2.2988, df = 127.7, p-value = 0.02314
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53776395 -0.04023605
sample estimates:
mean of x mean of y
   98.105    98.394
```

Example: Cell Phone Impairment

- Article in *Psych. Science* describes experiment that randomly assigned 64 Univ. of Utah students to cell phone group or control group (32 each). Driving simulation machine flashed *red* or *green* at irregular periods. Instructions: Press brake pedal as soon as possible when detect red light.
- Cell phone group: Carried out conversation about a political issue with someone in separate room.
- Control group: Listened to radio broadcast

Purpose of Study

- Analyze whether (conceptual) population mean response time differs significantly for the two groups, and if so, by how much.

❑ Data on “Response times” has

32 using cell phone with sample mean 585.2, *s* = 89.6

32 in control group with sample mean 533.7, *s* = 65.3

Interpretation?

```
> tsum.test(n.x=32,mean.x=585.2,s.x=89.6,n.y=32,mean.y=533.7,s.y=65.3)

        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = 2.6276, df = 56.685, p-value = 0.01104
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 12.24834 90.75166
```

Two Samples with Small Sample Size

- What to do with samples with small n?
- As in the one sample case, we need to use the *t* distribution to adjust for greater uncertainty in estimating the population standard deviations.
- One way to get the degrees of freedom is to take df=n1+n2-2 (several methods)
- R does this automatically when you run the tsum.test command

Example

Do male and female college students differ with respect to their **fastest reported driving speed**?

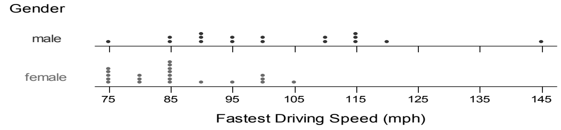
Population of all male college students

Population of all female college students

Sample of n_1 = 17 males report average of 102.1 mph

Sample of n_2 = 21 females report average of 85.7 mph

Graphical summary of sample data



Numerical summary

```
Gender    N     Mean   Median  TrMean   StDev
female   21    85.71    85.00   85.26    9.39
male     17   102.06   100.00  101.00   17.05

> tsum.test(n.x=21,mean.x=85.71,s.x=9.39,n.y=17,mean.y=102.06,s.y=17.05)

        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = -3.5427, df = 23.68, p-value = 0.001682
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.881849  -6.818151
```

The 2 Sample Hypotheses

- For the two-sample problem, the possible hypothesis are:
  - $H_o: \mu_1 = \mu_2 \qquad H_a: \mu_1 \neq \mu_2$
  - $H_o: \mu_1 = \mu_2 \qquad H_a: \mu_1 < \mu_2$
  - $H_o: \mu_1 = \mu_2 \qquad H_a: \mu_1 > \mu_2$

Decision Rules for Testing Two Samples

$$T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ the test statistic

$H_0: \mu_1 = \mu_2$ If $|T| > 1.96$ $reject\, H_o$
$H_a: \mu_1 \neq \mu_2$

$H_0: \mu_1 = \mu_2$ If $T < -1.64$ $reject\, H_o$
$H_a: \mu_1 < \mu_2$

$H_0: \mu_1 = \mu_2$ If $T > 1.64$ $reject\, H_o$
$H_a: \mu_1 > \mu_2$

Assuming both sample sizes > 30

Example

- Sony would like to test the hypothesis that the average age of a PlayStation user is different from the average age of an Xbox user.
- A random sample of 36 PlayStation users had an average age of 34.2 years while a random sample of 30 Xbox users had an average age of 32.7 years.
- Assume that the population standard deviation for the age of PlayStation and Xbox users is 3.9 and 4.0 years, respectively.

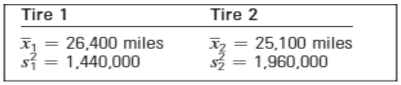
Output

$H_o: \mu_1 = \mu_2 \quad H_a: \mu_1 \neq \mu_2$

```
> tsum.test(n.x=36,mean.x=34.2,s.x=3.9,n.y=30,mean.y=32.7,s.y=4)

        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = 1.5343, df = 61.281, p-value = 0.1301
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.4547852  3.4547852
```

Example

- The wearing qualities of two types of automobile tires were compared by road-testing samples of 100 tires for each type and recording the number of miles until they wear out, defined as a specific amount of tire wear.
- Can we conclude that tire 1 lasts longer than tire 2?



Tire 1	Tire 2
$\bar{x}_1 = 26{,}400$ miles	$\bar{x}_2 = 25{,}100$ miles
$s_1^2 = 1{,}440{,}000$	$s_2^2 = 1{,}960{,}000$

R Output

$H_o: \mu_1 = \mu_2 \quad H_a: \mu_1 > \mu_2$

```
 > tsum.test(n.x=100,mean.x=26400,s.x=1200,n.y=100,mean.y=25100,s.y=1400,alt="greater")

       Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = 7.0502, df = 193.47, p-value = 1.538e-11
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 995.2447       NA
```

Conclusion??

Example: Calcium and Blood Pressure

Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure. The relationship was strongest for black men. Such observational studies do not establish causation. Researchers therefore designed a randomized experiment. The subjects were 21 healthy black men who volunteered to take part in the experiment. They were randomly assigned to two groups: 10 of the men received a calcium supplement for 12 weeks, while the control group of 11 men received a placebo pill that looked identical. The experiment was double-blind. The response variable is the decrease in systolic (top number) blood pressure

Group 1 (calcium):	7	−4	18	17	−3	−5	1	10	11	−2	
Group 2 (placebo):	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

We want to perform a test of

H_0: μ_1 - μ_2 = 0
H_a: μ_1 - μ_2 > 0

where μ_1 = the true mean decrease in systolic blood pressure for healthy black men like the ones in this study who take a calcium supplement, and μ_2 = the true mean decrease in systolic blood pressure for healthy black men like the ones in this study who take a placebo.

R Output

```
> group1=c(7,-4,18,17,-3,-5,1,10,11,-2)
> group2=c(-1,12,-1,-3,3,-5,5,2,-11,-1,-3)
> t.test(group1,group2,alt="greater")

        Welch Two Sample t-test

data:  group1 and group2
t = 1.6037, df = 15.591, p-value = 0.06442
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.476678       Inf
```

Matched Pairs

- Another two sample problem is **matched pairs.**
- One example is a group of people who decided to try Weight Watchers.
- You have their before and after weights, say, after two months of dieting.
- In this case we again have two different samples, but they are not independent, but rather matched.

Matched Pairs

- Sometimes the two samples that are being compared are matched pairs (not independent)
- Example: students matched on reading IQ, want to compare two different reading method on post exam

Pair	**New Method**	**Standard Method**
1	77	72
2	74	68
3	82	76
4	73	68
5	87	84
6	69	68
7	66	61
8	80	76

$\overline{x}_1 = 76 \quad \overline{x}_2 = 71.625$

$s_1^2 = 48 \quad s_2^2 = 49.1$

- We could test for the mean difference between X_1 = new method and X_2 = standard method
- However, we realize that these data are paired: each row of pairs are matched on a reading IQ
- Our *t*-test for two independent samples ignores this relationship

What if we ignore the matching?

- Does the data support the hypothesis that the new method is better than the old method?

$$H_0 : \mu_{new} = \mu_{old} \qquad H_a : \mu_{new} > \mu_{old}$$

```
> new=c(77,74,82,73,87,69,66,80)
> standard=c(72,68,76,68,84,68,61,76)
> t.test(new,standard,alt="greater")

        Welch Two Sample t-test

data:  new and standard
t = 1.2556, df = 13.998, p-value = 0.1149
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.762061       Inf
```

Interpret the R Output

- The p-value of 0.1149 says to fail to reject the null; there is no evidence the new method is better than the old method.
- Huh! But the new method is higher than the old method for every one of the eight pairs!
- There seems to be strong evidence the new method is better-what is happening?
- We ignored the matching which was a mistake.

Analyzing Paired Data

- The solution is simple; create a new column of the difference in values for each pair:

Pair	New Method	Standard Method	diff
1	77	72	5
2	74	68	6
3	82	76	6
4	73	68	5
5	87	84	3
6	69	68	1
7	66	61	5
8	80	76	4

$\bar{x}_D = 4.375$

$s_D = 1.69$

We then can do any hypothesis of interest on the difference; this is treated as a one sample t-test and doesn’t require any special handling.

Test if New Method > Old Method

- Define the difference

$$D = NewScore - OldScore$$

- We want to test

$$H_0 : \mu_D = 0 \qquad H_a : \mu_D > 0$$

```
> t.test(new,standard,alt="greater",paired=TRUE)

        Paired t-test

data:  new and standard
t = 7.3438, df = 7, p-value = 7.838e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.246316      Inf
sample estimates:
mean of the differences
                  4.375
```

Paired Difference Experiment

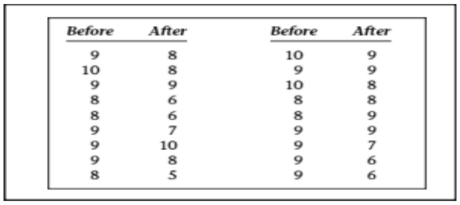
- Aha! Now we can conclude the new method is better.
- This is called a paired difference experiment.
- If your two samples are not independent, you need to adjust and not use the regular t-test.

Example

- A group of 18 concertgoers was selected at random. Before the concert they were given a hearing test, and then were given another one after the concert. (The volume varied during the test, and the person also had to state which ear the sound was in.)

Example

- Here are the number of correctly identified sounds out of 10, both before and after the concert.



Before	After	Before	After
9	8	10	9
10	8	9	9
9	9	10	8
8	6	8	8
8	6	8	9
9	7	9	9
9	10	9	7
9	8	9	6
8	5	9	6

Test if a person’s hearing is adversely effected by the concert noise.

R Output

```
 > before=c(9,10,9,8,8,9,9,9,8,10,9,10,8,8,9,9,9,9)
> after=c(8,8,9,6,6,7,10,8,5,9,9,8,8,9,9,7,6,6)
> t.test(before,after,paired=TRUE,alt="greater")

        Paired t-test

data:  before and after
t = 3.9626, df = 17, p-value = 0.0005027
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6856604       Inf
```

Conclusion??

Things you should know

❑Null and alternative hypothesis

❑Type I and II error

❑Decision rules for testing a mean, proportion, two means and two proportions.

❑Interpreting P-values