MATH 3339
Statistics for the Sciences
Chapter 10: Inferences on Two Groups or Populations

Wendy Wang
wwang60@central.uh.edu

Lecture 16 - 3339

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Outline

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Matched Pairs

Matched Pairs T test

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Matched Pairs

In low-speed crash test of five BMW cars, the repair costs were computed for a factory-authorized repair center and an independent repair facility. The results are as follows.

Authorized repair center	\$797	\$571	\$904	\$1147	\$418
Independent repair center	\$523	\$488	\$875	\$911	\$297

We want to estimate the mean of the difference between the two repair centers.

Inference for Matched Pairs

- The previous question is a matched pair.
- We are looking at the same car. The subject units are exactly the same for both responses.
- We calculate the differences first and find the mean and standard deviation of the differences.
- Then this problem is the same as a one-sample confidence interval.
  - We first find the differences from each observation.
  - The point estimate is $\bar{x}_d$ = mean of the differences.
  - The standard deviation is s_d = the standard deviation of the differences.
  - Then the margin of error is $m = t^* \left(\frac{s_d}{\sqrt{n}} \right)$.
  - The confidence interval is $\bar{x}_d \pm t^* \left(\frac{s_d}{\sqrt{n}} \right)$.
- If we want a hypothesis test, the test statistic is: $t = \frac{\bar{x}_d - \mu_d}{s_d / sqrtn}$

Matched Pairs Assumptions

- Matched pairs is a special test when we are comparing corresponding values in data.
- This test is used only when our data samples are DEPENDENT upon one another (like before and after results).
- Matched pairs t – test assumptions:
  1. Each sample is an SRS of size n from the same population.
  2. The test is conducted on paired data (the samples are NOT independent).
  3. Unknown population standard deviation.
  4. Either a Normal population or large samples ($n \geq 30$).
- Hypotheses - $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$ or $\mu_d < 0$ or $\mu_d > 0$. Where μ_d is the mean of the differences.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Crash Test Repair Costs

We want to determine a 95% confidence interval for the difference in the repair cost of the authorized repair center and the independent repair center.

Authorized repair center	\$797	\$571	\$904	\$1147	\$418
Independent Repair center	\$523	\$488	\$875	\$911	\$297
Differences	\$274	\$83	\$29	\$236	\$121

$\$274 = \$797 - \$523$

⋮

$\$121 = \$418 - \$297$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

R code

sample1

sample2

matched pairs t test

```
> auth=c(797,571,904,1147,418)
> indep=c(523,488,875,911,297)
> t.test(auth,indep,conf.level = 0.95, paired = TRUE)

Paired t-test

data:  auth and indep
t = 3.2148, df = 4, p-value = 0.03244
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
20.26155 276.93845
sample estimates:
mean of the differences
148.6
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example 2

A new law has been passed in a city. For six neighborhoods, the numbers of reported crimes one year before and one year after the new law were given. Does this indicate that the number of reported crimes have dropped?

$\mu_d =$ after $-$ before

Neighborhood	1	2	3	4	5	6
Before	18	35	44	28	22	37
After	21	23	30	19	24	29

$H_0: \mu_d = 0$

$H_a: \mu_d < 0$

if $\mu_d =$ before $-$ after

$H_a: \mu_d > 0$

R code

"two tailed"
"greater": right-tailed
"less" : left-tailed

```
> t.test(before,after,alternative="greater",paired=TRUE)

Paired t-test

data:  before and after
t = 2.1624, df = 5, p-value = 0.04147
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.4316912       Inf
sample estimates:
mean of the differences
6.333333
```

t.test(after, before, alternative = "less", paired=T)

Two-population inference

- Is the mean miles per gallon of automobiles significantly different depending on the manufacturer of the automobile?
- Is the mean price of a business textbook significantly lower than the mean price of a general course textbook?
- What is the difference between the mean height of men and mean height of women?
- We want to estimate: $\mu_1 - \mu_2$

not μ_d

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Notations Used

- For the population

Population	variable	Mean	Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

- For the sample

Population	Sample Size	Sample Mean	Sample Standard deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Two Population problems

- The goal of inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.

Assumptions for Difference of Two Means

1. Both samples must be independent SRSs from the populations of interest.
2. Both sets of data must come form normally distributed populations.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Two-sample t

- If the population standard deviations σ_1 and σ_2 is unknown the sample standard deviations s_1 and s_2 is used.
- When we use the sample standard deviations we use the **two-sample** t **statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with k degrees of freedom approximated by software or the smaller value of $n_1 - 1$ or $n_2 - 1$.

$$df = \min(n_1 - 1, n_2 - 1)$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Approximate Degrees of Freedom

- The reality is that the previous model is not really Student's t, but only something close.
- So the calculators and other software such as R uses an approximate degrees of freedom called **Satterthwaite** degrees of freedom.
- Calculated

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

- This is only to show what degrees of freedom R and the calculators are using. If we do this by hand use the smaller of $n_1 - 1$ or $n_2 - 1$.

smaller of $n_1 - 1$ $n_2 - 1$

UNIVERSITY of HOUSTON DEPARTMENT OF MATHEMATICS

Interval Estimation of $\mu_1 - \mu_2$

$\bar{x} \pm t_{(\frac{\alpha}{2}, df)} \cdot \frac{s}{\sqrt{n}},$

1. **Point Estimate**: $\bar{x}_1 - \bar{x}_2$
2. **Confidence level**: $1 - \alpha = C$

   $qt(\frac{1+C}{2}, df)$

3. **Critical value**: t^* with degrees of freedom of $n_1 - 1$ or $n_2 - 1$ whichever is smaller. In R: t^* = **qt**~~($C + \alpha/2$,**df**~~**)**.
4. **Margin of Error**:

$$E = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

5. **Confidence Interval**: point estimate $\pm$ margin of error

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example: Check out

A well known grocery store chain performed a study to determine whether the average purchase through a self-checkout facility was less than the average purchase at the traditional checkout stand. To conduct the test, a random sample of 125 customer transactions at the self-checkout was obtained and a second random sample of 125 transactions from customers using traditional checkout process was obtained. The following statistics were computed from each sample

Self-Checkout	Traditional Checkout
$\bar{x}_1 = \$45.68$	$\bar{x}_2 = \$78.49$
$s_1 = \$58.20$	$s_2 = \$62.45$
$n_1 = 125$	$n_2 = 125$

$C = 0.90$

$df = 124$

Develop a 90% confidence interval of the difference between the different checkouts.

$$(\bar{x}_1 - \bar{x}_2) \pm qt\left(\frac{1+c}{2}, df\right) * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

R Code

In R:
$(\bar{x}_1 - \bar{x}_2)$-qt($(1 + C)/2$)*sqrt($s_1^2/n_1 + s_2^2/n_2$) for lower value of confidence interval,
$(\bar{x}_1 - \bar{x}_2)$+qt($(1 + C)/2$)*sqrt($s_1^2/n_1 + s_2^2/n_2$) for upper value of confidence interval.

```
> (45.68-78.49)-qt(1.9/2,124)*sqrt(58.2^2/125+62.45^2/125)
[1] -45.4635
> (45.68-78.49)+qt(1.9/2,124)*sqrt(58.2^2/125+62.45^2/125)
[1] -20.1565
```

90% C.I: (−45.46, −20.16)

if 0 is in between.

Two - Sample t-Test

- Compare the responses to two treatments or characteristics of two populations.
- These tests are different than the matched pairs t-test.
- Hypotheses
  - Null - $H_0 : \mu_1 = \mu_2$
  - Alternative - $H_a : \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

For matched pairs

$H_0: \mu_d = 0$

$H_a: \mu_d \neq 0$, $\mu_d < 0$ or $\mu_d > 0$

Assumptions for a Two-Sample t-Test

The goal of inference is to compare the responses in two groups.

1. Each group is considered to be a **simple random sample** from two **distinct** populations.
2. The responses in each group are **independent** of those in the other group.
3. The distribution of the variables are **Normal** or have a large sample $n_1 \geq 30$ and $n_2 \geq 30$.

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

With degrees of freedom equal to the smaller of $n_1 - 1$ or $n_2 - 1$.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Comparing mean MPG

- From a random sample of 45 Prius automobiles and 45 Civic automobiles we get the following statistics:

Automobile	n	Sample mean $\bar{x}$	Sample SD s
Prius	45	47.62	2.430
Civic	45	49.4	7.226

- Can we say from this information that the Civic has a different mean mpg than the Prius?

MPG hypothesis

Is the mean MPG for Prius automobiles different from mean MPG for Civic automobiles?

- Null hypothesis: $H_0 : \mu_{\text{Prius}} = \mu_{\text{Civic}}$
- Alternative hypothesis: $H_A : \mu_{\text{Prius}} \neq \mu_{\text{Civic}}$

Two-sample t test statistic

Formula:

$$t = \frac{\text{estimate} - \text{hypothesized mean of estimate}}{\text{SE of estimate}}$$
$$= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$= \frac{47.62 - 49.4}{\sqrt{\frac{2.430^2}{45} + \frac{7.226^2}{45}}} = -1.5662$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Two-sample t test statistic

Formula:

$$
\begin{aligned}
t &= \frac{\text{estimate} - \text{hypothesized mean of estimate}}{\text{SE of estimate}} \\
&= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{47.62 - 49.4}{\sqrt{\frac{2.430^2}{45} + \frac{7.226^2}{45}}} = -1.5662
\end{aligned}
$$

Formula:

$$
\begin{aligned}
t &= \frac{\text{estimate} - \text{hypothesized mean of estimate}}{\text{SE of estimate}} \\
&= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{47.62 - 49.4}{\sqrt{\frac{2.430^2}{45} + \frac{7.226^2}{45}}} = -1.5662
\end{aligned}
$$

$H_a: \mu_1 \neq \mu_2$

$n_1 = 45$

$n_2 = 45.$

P-value = 2*pt(-1.5662,44)

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

P-value and Conclusion

$$P - \text{value} = 2P(T < -1.5663)$$

In R:

```
2*pt(-1.5663,df=44)
[1] 0.1244429
```

P-value = 0.1244, which is greater than 0.1 (10%). Thus we fail to reject the null hypothesis. Thus we **cannot conclude** that the mean MPG of Honda Civic automobiles is significantly different than the mean MPG of a Toyota Prius automobile.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Fats

Solid fats are more likely to raise blood cholesterol levels than liquid fats. Suppose a nutritionist analyzed the percentage of saturated fat for a sample of 6 brands of stick margarine (solid fat) and for a sample of 6 brands of liquid margarine and obtained the following results:

Stick:[25.5,26.7,26.5,26.6,26.3,26.4]
Liquid:[16.5,17.1,17.5,17.3,17.2,16.7]

We want to determine if there a significant difference in the average amount of saturated fat in solid and liquid fats.

stick = c(, , , , ,)

liquid = c( )

t.test(stick, liquid, paired= F)

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Comparing Two Proportions

What is the difference between the proportion of m&ms that are blue in the plain m&ms compared to the peanut m&ms?

- From a random sample of plain m&ms and peanut m&ms we get the following results.

Candy type	n	Number of Blue	Sample proportion ($\hat{p}$)
plain	81	28	$\hat{p}_{\text{plain}} = \frac{28}{81} = 0.3458$
peanut	100	20	$\hat{p}_{\text{peanut}} = \frac{20}{100} = 0.2$

- We want to know what is the difference of the proportion of m&ms that are blue for all of plain and peanut m&ms. That is, estimate:

$$p_{\text{plain}} - p_{\text{peanut}}$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Two-sample problems assumptions

The goal of inference is to compare the responses in two groups.

1. Each group is considered to be a **simple random sample** from two **distinct** populations.
2. The population sizes are both at least ten times the sizes of the samples.
3. The number of successes and failures in **both** samples must all be ≥ 10.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Confidence intervals for comparing two proportions

Choose an SRS of n_1 from a large population having proportion p_1 of successes and and independent SRS of size n_2 from another population having proportion p_2 of successes.

1. Point estimate: $D = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$ $= \frac{28}{81} - \frac{20}{100}$
2. Confidence level: C a percent predetermined in the problem if not use 95%.
3. Critical value: z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*.
4. Confidence interval:

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

z^* = qnorm$\left(\frac{1+C}{2}\right)$

LL UL

5. Interpret

Determine a 95% confidence interval for the difference of the proportion of m&ms that are blue for all of plain and peanut m&ms.

From a random sample of plain m&ms and peanut m&ms we get the following results.

Candy type	n	Number of Blue	Sample proportion ($\hat{p}$)
plain	81	28	$\hat{p}_{\text{plain}} = \frac{28}{81} = 0.3458$
peanut	100	20	$\hat{p}_{\text{peanut}} = \frac{20}{100} = 0.2$

$$(0.5458 - 0.2) \pm \text{qnorm}\left(\frac{1.95}{2}\right) * \sqrt{\frac{0.3458(1-0.3458)}{81} + \frac{0.2(1-0.2)}{100}}$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

R code

of blue mm

prop.test(x=c(x_1,x_2),n=c(n_1,n_2),conf.level = C, correct = FALSE)

```
prop.test(x=c(28,20),n=c(81,100),conf.level = 0.95,correct=FALSE)

2-sample test for equality of proportions without continuity
correction

data:  c(28, 20) out of c(81, 100)
X-squared = 4.8738, df = 1, p-value = 0.02727
alternative hypothesis: two.sided
95 percent confidence interval:
0.01578192 0.27557610
sample estimates:
prop 1   prop 2
0.345679 0.200000
```

Assumptions for Two-Sample Proportion Test

1. Both samples must be independent SRSs from the populations of interest.
2. The population sizes are both at least ten times the sizes of the samples.
3. The number of successes and failures in both sample must all be at least 10.

$H_0: p_1 = p_2$ $\qquad H_a: p_1 \neq p_2$

$p_1 > p_2$

$p_1 < p_2$

Test statistic:

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

RR. $\qquad$ p-value,

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Left-handedness

Is the proportion of left-handed students higher in honors classes than in academic classes? Two hundred academic and one hundred honors students from grades 6 - 12 were selected throughout a school district and their left handedness was recorded. The sample information is:

	Honors	Academic
Sample Size	$n_1 =$ 100	$n_2 =$ 200
Number of left-handed students	$x_1 =$ 18	$x_2 =$ 32

Is there sufficient evidence at the 1% significance level to conclude that the proportion of left-handed students is greater in honor classes?

$\hat{p}_H = \frac{x_H}{n_H} = \frac{18}{100} = 0.18$ $\hat{p}_A = \frac{x_A}{n_A} = \frac{32}{200} = 0.16$

$H_0: p_H = p_A$ $H_a: p_H > p_A$

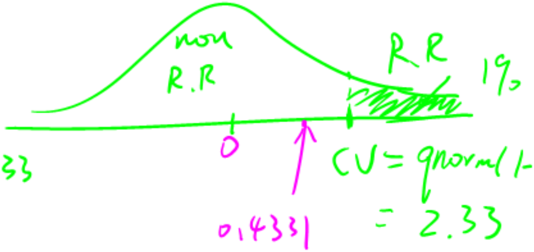
check # of successes & failures in each sample

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

$\alpha = 1\%$.

Rejection Region:

Reject H_0 if $z \geq 2.33$



$CV = \text{qnorm}(1-.01) = 2.33$

$$z = \frac{(\hat{p}_H - \hat{p}_A) - \cancel{(p_H - p_A)}}{\sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n_H} + \frac{\hat{p}_A(1-\hat{p}_A)}{n_A}}}$$

$$= \frac{(0.18-0.16) - 0}{\sqrt{\frac{0.18*(1-0.18)}{100} + \frac{(0.16)(1-0.16)}{200}}} = 0.4331$$

$z = 0.4331$ is in the Non-rejection region, so we fail to reject H_0.

Conclusion: The data does not provide sufficient evidence to conclude that

H_a _ _ _ _ _ _ _ _