MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, we finished Module 4 (hypothesis testing)

- Today we will cover Chapter 11 Two Population Means Comparison.

Chapter11 Comparison of Two Means

(A) Paired Samples

- Definition: An experiment is paired if each observation in the first sample is associated with or similar to a single observation in the second sample.

"similar" means, in the absence of systematic difference between the two populations being paired, the values of two measurements within a pair are expected to be closer on average than two measurements on different pairs.

- Examples: Two animals from the same litter; Two persons matched to have the same gender, age, approximate IQ, etc.; Same person on two occasions; et al.

(A) Paired Samples

Example. How do learning disabled children (with otherwise high IQ) differ from normal children in short term memorization ability? Each learning disabled child was matched with a normal child of the same gender, age and approximate IQ. Then they took a test on short-term memorization ability (higher score indicate better memory).

Pair	Learning disabled	Normal	
1	42	65	
2	59	72	
3	64	62	
4	55	80	
5	31	49	
6	40	57	
7	51	71	
	Y_i	X_i	

(A) Paired Samples Inference

How to make the inference here?

- Let μ_X= mean memorization score of normal children,

  μ_Y= mean memorization score of L.D. children

- $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$ then using the property of linear combination of normal RVs: $D_i = X_i - Y_i \sim N(\mu_D, \sigma_D^2)$, where $\mu_D = \mu_X - \mu_Y$ and σ_D^2 =? (=$\sigma_X^2+\sigma_Y^2$ if X independent of Y)

  In fact, we do not know a formula of σ_D^2 except that it is a constant. We do not need the formula of σ_D^2 for inference.

- Hence we can use the **univariate inference** on $D_1,\ldots,D_n$ for the inference on paired comparisons of X and Y.

(A) Paired Samples Inference

Memorization Example.

Pair	Learning disabled	Normal	Difference
1	42	65	23
2	59	72	13
3	64	62	-2
4	55	80	25
5	31	49	18
6	40	57	17
7	51	71	20
	Y_i	X_i	D_i

- (A.1) Point estimator for μ_X - μ_Y.
  Use $\hat{\mu}_D = \overline{D} = \sum_{i=1}^{n}(X_i - Y_i) = \overline{X} - \overline{Y}$
- For this example, $\hat{\mu}_D = \overline{D}$=16.3

(A) Paired Samples Inference

- (A.2) Confidence interval for μ_X - μ_Y.

  Use the one-sample t-interval for μ_D: $\overline{D} \pm t_{n-1,\, \alpha/2} \frac{s_D}{\sqrt{n}}$

  where $s_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \overline{D})^2$

- For this Memorization Example, the 95% CI for μ_D is

$$\overline{D} \pm t_{6,\, 0.025} \frac{s_D}{\sqrt{7}} = 16.3 \pm 2.447 \frac{8.98}{\sqrt{7}} = (8.0, 24.6)$$

- We are 95% confident that μ_X - μ_Y lies within (8.0, 24.6).

(A) Paired Samples Inference

- (A.3) Test $\mathbf{H_0}: \mu_X = \mu_Y$ versus $\mathbf{H_A}: \mu_X \neq \mu_Y$.
  $\Leftrightarrow \mathbf{H_0}: \mu_X - \mu_Y = 0$ versus $\mathbf{H_A}: \mu_X - \mu_Y \neq 0$.
- Use the one-sample t-test for μ_D:

  Reject at α level if $T_{obs} = \left|\frac{\overline{D}}{s_D/\sqrt{n}}\right| > t_{n-1,\alpha/2}$
- For this Memorization Example,

  $T_{obs} = \left|\frac{16.3}{8.98/\sqrt{7}}\right| = 4.8$, $t_{6,\,0.025} = 2.447 < T_{obs}$

  Reject $\mathbf{H_0}$ at $\alpha = 0.05$ level.
- Conclusion: L.D. children do differ from normal children in short-term memorization.

(B) Independent Samples

Example. Two diets are tested on mice to determine if they lead to different weight gains.
Data: weight gains in grams.

#1 cereal diet	#2 pork diet
98	94
74	79
56	96
111	98
95	
$X_{1,i}$	$X_{2,i}$

Can not be paired.

New inference methods are needed.

(B) Independent Samples Inference

Mathematical setup

		Group 1	Group 2
	Observations	$X_{1,1},\ldots,X_{1,n1}$	$X_{2,1},\ldots,X_{2,n2}$
Population parameters	mean	μ_1	μ_2
	variance	σ_1^2	σ_2^2
Sample statistics	mean	$\bar{X}_1$	$\bar{X}_2$
	variance	s_1^2	s_2^2
	Sample size	n_1	n_2

Inference question: $\mu_1 = \mu_2$?

(B) Independent Samples Inference

Inference question: $\mu_1 = \mu_2$?

- (B.1) What is the point estimator for $\mu_1 - \mu_2$?
- (B.2) What is the confidence interval for $\mu_1 - \mu_2$?
- (B.3) Test $\mathbf{H_0}: \mu_1 = \mu_2$ versus $\mathbf{H_A}: \mu_1 \neq \mu_2$.

  $\Leftrightarrow \mathbf{H_0}: \mu_1 - \mu_2 = 0$ versus $\mathbf{H_A}: \mu_1 - \mu_2 \neq 0$.

We focus on presenting formulas for (B.3) hypothesis test. The other formulas follows due to the equivalence introduced before.

(B) Independent Samples Inference

- If n_1=n_2, we can just pair the data randomly and used paired inference.
- If $n_1 \neq n_2$, we do not want to pair (that will lose some data). How to do inference then?

What is the statistic for μ_1-μ_2?

Obviously, we should use $\bar{X}_1$-$\bar{X}_2$.

If we have theoretical distribution of $\bar{X}_1$-$\bar{X}_2$, we can get inferences like before.

(B) Independent Samples Inference

- What is the distribution of $\bar{X}_1$-$\bar{X}_2$?
- From the property of linear combination of normal RVs: since $X_{1,i} \sim N(\mu_1, \sigma_1^2)$ is <u>independent</u> of $X_{2,j} \sim N(\mu_2, \sigma_2^2)$, then

$$\bar{X}_1\text{-}\bar{X}_2 \sim N(\mu_1\text{-}\mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

Notice that $\text{Var}(\bar{X}_1\text{-}\bar{X}_2) = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

where the sign is + not -.

(B) Independent Samples Inference

(B.3) Test $\mathbf{H_0}: \mu_1 = \mu_2$ versus $\mathbf{H_A}: \mu_1 \neq \mu_2$.

$\Leftrightarrow \mathbf{H_0}: \mu_1 - \mu_2 = 0$ versus $\mathbf{H_A}: \mu_1 - \mu_2 \neq 0$.

- (B.3a) Know σ_1^2 and σ_2^2, then use z-test.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

Reject $\mathbf{H_0}$ when $Z_{obs} = \left| \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right| > z_{\alpha/2}$

(B) Independent Samples Inference

- (B.3b) Unknown σ_1^2 and σ_2^2 but assume $\sigma_1=\sigma_2=\sigma$

Then we estimate σ^2 by pooled sample variance

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \sim \frac{\sigma^2\chi^2_{n_1-1}+\sigma^2\chi^2_{n_2-1}}{n_1+n_2-2} \sim \frac{\sigma^2\chi^2_{n_1+n_2-1}}{n_1+n_2-2}$$

So $$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$$

Reject $\mathbf{H_0}$ when $T_{obs} = \left|\frac{\bar{X}_1-\bar{X}_2}{s_p\sqrt{1/n_1+1/n_2}}\right| > t_{n_1+n_2-2,\alpha/2}$

(B) Independent Samples Inference

Mice Diet Example.

$\bar{X}_1$=86.8, $\bar{X}_2$=91.75,

s_1^2=472.7, s_2^2=74.91,

n_1=5, n_2=4,

$$s_p^2 = \frac{4(472.7) + 3(74.91)}{7}$$
$$=320.22$$

$s_p=\sqrt{320.22}=17.38$

#1 cereal diet	#2 pork diet
98	94
74	79
56	96
111	98
95	
$X_{1,i}$	$X_{2,i}$

(B) Independent Samples Inference

- (B.3b) Unknown σ_1^2 and σ_2^2 but assume $\sigma_1=\sigma_2=\sigma$

Mice example:

$\bar{X}_1=86.8,\ \ \bar{X}_2=91.75,\ s_1^2=472.7,\ s_2^2=74.91, n_1=5,\ \ n_2=4,\ s_p=17.38$

For $\alpha=0.1$, $t_{7,0.05}$ =1.415

$$T_{obs} = \left|\frac{\bar{X}_1-\bar{X}_2}{s_p\sqrt{1/n_1+1/n_2}}\right| = \left|\frac{86.8-91.75}{17.38\sqrt{1/5+1/4}}\right| = 0.4245$$

Hence we do NOT reject **H_0** ($T_{obs} < t_{n_1+n_2-2,\alpha/2}$).

Conclusion: There is not sufficient evidence in these data to conclude that weight gains differ for diet 1 and diet 2.

Question: Do we reject at $\alpha=0.05$?

(B) Independent Samples Inference

- (B.3c) Unknown σ_1^2 and σ_2^2 generally use test statistic $\dfrac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$ which is *approximately* distributed as t_{df} with degree of freedom

$$\text{df}=\frac{(s_1^2/n_1)^2+(s_2^2/n_2)^2}{\sqrt{(s_1^2/n_1)^2/(n_1-1)+(s_2^2/n_2)^2/(n_2-1)}} \le n_1+n_2-2$$

Notice that $(s_1^2/n_1)^2+(s_2^2/n_2)^2 \sim \frac{\sigma_1^2}{n_1(n_1-1)}\chi^2_{n_1-1}+\frac{\sigma_2^2}{n_2(n_2-1)}\chi^2_{n_2-1}$.

We cannot get exact chi-square by adjusting weights since is σ_1^2/σ_2^2 unknown.

(B) Independent Samples Inference

- (B.3c) Unknown σ_1^2 and σ_2^2 and **NO** assumption $\sigma_1=\sigma_2$

Mice example: df= $\dfrac{(s_1^2/n_1)^2+(s_2^2/n_2)^2}{\sqrt{(s_1^2/n_1)^2/(n_1-1)+(s_2^2/n_2)^2/(n_2-1)}}$ = 5.456.

Use the next lower value df=5. For α=0.1, $t_{5,0.05}$ =1.476

T_{obs} =$\left|\dfrac{86.8-91.75}{\sqrt{472.5/5+74.91/4}}\right|$ =$\left|\dfrac{-4.95}{10.64}\right|$ =0.465

Hence we do NOT reject $\mathbf{H_0}$ at α=0.1 level.

90% CI for μ_1-μ_2 is

(-4.95 - 1.476*10.64, -4.95 + 1.476*10.64) = (-20.7, 10.8)

R commands for Two Samples Comparison

- **Using the response times from the mini-project in Lab1.**

I collected response times in Lab 1 with “small” and “xlarge” boxes for 15 times and put the measured times in a text file ResponseTime2.txt

Small	Xlarge
531	532
843	515
578	547
563	469
625	625
704	578
657	781
875	594
672	656
672	766
812	703
813	547
954	515
657	500
609	500

R commands for Two Samples Comparison

(a) Paired t-test.

```
> # Import data set. This is formatted as two columns/variables
> MyTime <- read.table(file="ResponseTime2.txt", header=TRUE)
> # For paired t-test, simply create the new variable of difference,
> #  then do one sample t-test as before
> diff.time<- MyTime$Small-MyTime$Xlarge
> t.test(diff.time)  #At default alpha=0.05 level
    One Sample t-test
data:  diff.time
t = 2.847, df = 14, p-value = 0.01293
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  28.56146 203.03854
sample estimates:
mean of x
    115.8
```

- According to the paired t-test, the null hypothesis will be rejected at 0.05 level but not 0.01 level.

R commands for Two Samples Comparison

(a) Paired t-test.

We can also do the paired t-test directly in R without the new variable.
The outputs are almost exactly the same.

```
> t.test(MyTime$Small, MyTime$Xlarge, paired=T)
    Paired t-test

data:  MyTime$Small and MyTime$Xlarge
t = 2.847, df = 14, p-value = 0.01293
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  28.56146 203.03854
sample estimates:
mean of the differences
                 115.8
```

- According to the paired t-test, the null hypothesis will be rejected at 0.05 level but not 0.01 level.

R commands for Two Samples Comparison

(b) The independent two samples t-test.

We now illustrate how to do the two sample t-tests in R, pretending the data is not paired. We can simply put these two variables in t.test() without the 'paired' option.

```
> # Unpaired t-test
> t.test(MyTime$Small, MyTime$Xlarge)
    Welch Two Sample t-test
data:  MyTime$Small and MyTime$Xlarge
t = 2.8081, df = 26.433, p-value = 0.009247
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  31.10277 200.49723
sample estimates:
mean of x mean of y
 704.3333  588.5333
```

- The p-value of the unpaired t-test here is a bit smaller than that of the paired t-test. (This actually is unusual. The pairing here does not affect the timing much.)

R commands for Two Samples Comparison

(b) The independent two samples t-test.

While the above commands are simple, the two groups often have different sample sizes, data is often presented in a different format:

```
        time    group
1       531     Small
2       843     Small
3       578     Small
4       563     Small
...
13      954     Small
14      657     Small
15      609     Small
16      532     Xlarge
17      515     Xlarge
18      547     Xlarge
...
28      515     Xlarge
29      500     Xlarge
30      500     Xlarge
```

R commands for Converting the Format

We first illustrate how to convert the data set into this new format using R commands.

```
> ## Rearrange the data
> small.time<-data.frame(time=MyTime$Small, group='Small') #two columns: time and group
> xlarge.time<-data.frame(time=MyTime$Xlarge, group='Xlarge')
> new.data<- rbind(small.time, xlarge.time) #merge two sets above
```

```
Small  Xlarge
531      532
843      515
578      547
563      469
625      625
704      578
657      781
875      594
672      656
672      766
812      703
813      547
954      515
657      500
609      500
```

```
          time       group
1          531       Small
2          843       Small
3          578       Small
4          563       Small
...
13         954       Small
14         657       Small
15         609       Small
16         532      Xlarge
17         515      Xlarge
18         547      Xlarge
...
28         515      Xlarge
29         500      Xlarge
30         500      Xlarge
```

R commands for Two Samples Comparison

(b) The independent two samples t-test.

In this new format, we can do the independent two sample t-test as

```
> # Unpaired t-test
> t.test(time~group, data=new.data)
        Welch Two Sample t-test

data:  time by group
t = 2.8081, df = 26.433, p-value = 0.009247
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  31.10277 200.49723
sample estimates:
 mean in group Small mean in group Xlarge
            704.3333             588.5333
```

The output is almost exactly same as before.

R commands for Two Samples Comparison

- (b.1) The independent two samples t-test assuming $\sigma_1=\sigma_2=\sigma$

In R, this is simple, just set the option of "var.equal"

```
> # Unpaired t-test
> t.test(time~group, data=new.data, var.equal=T)
        Two Sample t-test
data:  time by group
t = 2.8081, df = 28, p-value = 0.008978
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  31.32887 200.27113
sample estimates:
 mean in group Small mean in group Xlarge
            704.3333             588.5333
```

In this case, the p-value is very close to the p-value of t-test of unequal variance.

R commands for Two Samples Comparison

- (b.1) The independent two samples t-test assuming $\sigma_1=\sigma_2=\sigma$

We check if $\sigma_1=\sigma_2=\sigma$ is reasonable by F-test.

```
> # Testing equal variance
> var.test(time~group, data=new.data)
        F test to compare two variances
data:  time by group
F = 1.6435, num df = 14, denom df = 14, p-value = 0.3636
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5517881 4.8954551
sample estimates:
ratio of variances
          1.643549
```

p-value is big here, so we accept the null hypothesis of equal variance. In this data set, the two group variances are not very different, so the two versions of t-test (equal/unequal variance) give very similar results.

R commands for Two Samples Comparison

- Above we gave the R commands for doing the independent two samples t-test both assuming $\sigma_1=\sigma_2=\sigma$ and without assuming it. In this data set, the two versions of t-test (equal/unequal variance) give very similar results.

But generally in practice, which version should we use?

Summary

Module 5 done. You should know:

- When to use paired test and when to use unpaired test.
- R commands for doing the tests.
- For unpaired tests: assuming equal variance gives exact theoretical distribution. But in practice, we should use the unequal variance t-test.
- Homework 4 is due in one week. Project proposal is due in two weeks.
- Next lecture we start the next module Analysis of Variance (chapter 12).