MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, we covered ANOVA
- Need to understand the ANOVA table, and how to get it from R.
- How to do contrasts after F-test rejects; When are multiple testing adjustments needed, and how to do it.
- ANOVA with blocking.

- Today we start with reviewing the handout of R commands for two-way ANOVA, then proceed to next topic in Module 8 Nonparametric methods.

ANOVA with blocking

- Compare the decomposition for sums of squares in one-way ANOVA without and with blocking:

  a treatment levels and b blocks

One-way (CR)	RB	DF	
Treatment	Treatment	a-1	
Error	Blocks	b-1	ab-a
	Error	(a-1) (b-1)	
Total	Total	ab-1	

ANOVA with blocking

- The ANOVA with blocking basically analyze the variance due to two factors: treatment and blocks, but then ignore blocking variable and focus on treatment variable only. If the blocking variable is of equal importance as the treatment variable, then we should do the full two-way ANOVA:

Treatment A: a levels

Treatment B: b levels

Each combination (ab in total) has n measurements.

Source		DF
Treatment	A (factor effect)	a-1
	B (factor effect)	b-1
	A x B (interaction)	(a-1) (b-1)
Error		(n-1)ab
Total		nab-1

- In the cheese example, n=1 and no degree of freedom left for MSE. Then use the AxB part of MS as the error MS.

R commands for ANOVA with blocking/two-way ANOVA

- See the handout (Two-ANOVA.pdf). The cheese data

```
yield temp lot
2.89 1 1
2.95 2 1
3.10 3 1
3.23 4 1
2.86 1 2
3.20 2 2
3.03 3 2
3.18 4 2
3.18 1 3
3.06 2 3
3.15 3 3
3.18 4 3
2.92 1 4
3.15 2 4
3.26 3 4
3.32 4 4
3.09 1 5
3.25 2 5
3.22 3 5
3.26 4 5
> # Import data set. This is formatted 3columns/variables
> Cheese.data <- read.table(file="CheeseYield.txt", header=TRUE)
```

R commands for ANOVA with blocking/two-way ANOVA

- Important: You need to inform R which are the categorical (factor) variables, otherwise you will not get correct ANOVA outputs.

```
> Cheese.data$temp<-as.factor(Cheese.data$temp)
> Cheese.data$lot<-as.factor(Cheese.data$lot)
> # One-way ANOVA of yield over temperature.
> summary(aov(yield~temp, data=Cheese.data))
            Df Sum Sq Mean Sq F value Pr(>F)
temp         3 0.1569 0.05231   4.589 0.0168 *
Residuals   16 0.1824 0.01140
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is the one-way ANOVA table in previous slides

```
> # Two-way ANOVA with temperature and lot (block)
> summary(aov(yield~temp+lot, data=Cheese.data))
            Df  Sum Sq Mean Sq F value Pr(>F)
temp         3 0.15692 0.05231   5.733 0.0114 *
lot          4 0.07288 0.01822   1.997 0.1591
Residuals   12 0.10948 0.00912
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is the table for ANOVA with blocking in previous slides.
Notice that the line of block (lot here) is simply ignored in the analysis.

R commands for ANOVA with blocking/two-way ANOVA

The R model getting the ANOVA with blocking is y~A+B. This fits the two -way ANOVA without interaction AxB. The full two-way ANOVA includes the AxB interaction term, in R, fit the model y~A*B.

```
> # Full two-way ANOVA with temperature and lot
> summary(aov(yield~temp*lot, data=Cheese.data))
             Df  Sum Sq Mean Sq
temp          3 0.15692 0.05231
lot           4 0.07288 0.01822
temp:lot     12 0.10948 0.00912
```

In this case, n=1 and no degree of freedom left for MSE. So in fact we use the previous ANOVA table where the interaction term used as error

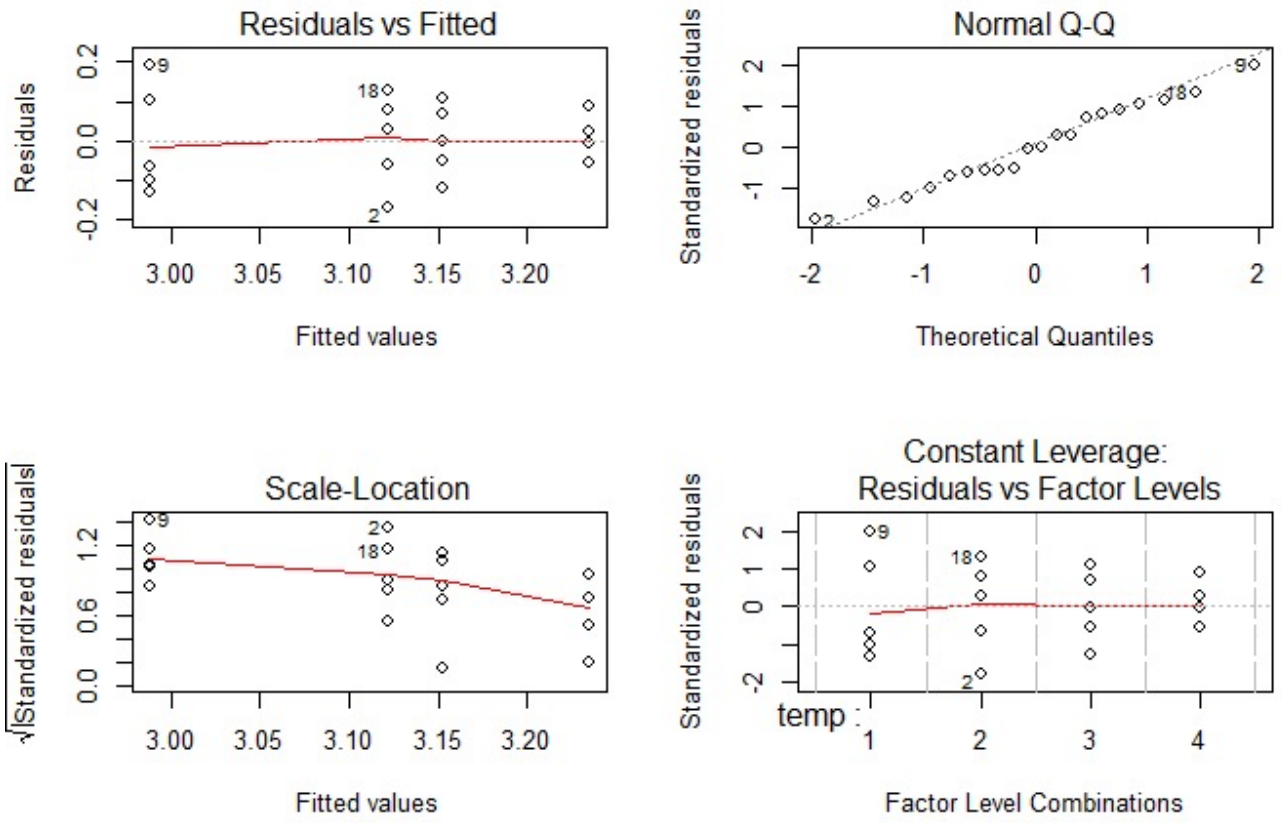
```
> summary(aov(yield~temp+lot, data=Cheese.data))
             Df  Sum Sq Mean Sq F value Pr(>F)
temp          3 0.15692 0.05231   5.733 0.0114 *
lot           4 0.07288 0.01822   1.997 0.1591
Residuals    12 0.10948 0.00912
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA with blocking

- Treatment A: a levels; Treatment B: b levels
  Each combination (ab in total) has n measurements.
  If n=1, use the interaction term AxB as the error. Then it is the same R command to get the two-way ANOVA and one-way ANOVA with blocking: both use y~A+B. The only difference is that you ignore the blocking (factor B) line when doing inference in one-way ANOVA with blocking.

Other R commands in the handout

```
> # Check one-way ANOVA fit
> ##### Diagnostic plots, put 4 in one page (2 rows by 2 columns).
> par(mfrow=c(2,2)) #set layout, 2 rows by 2 columns
> plot(aov(yield~temp, data=Cheese.data))
```



Other R commands in the handout

```
> # Mean yields for lots are different, consider blocking for lots.
> vapply(levels(Cheese.data$lot), FUN=function(x) mean(Cheese.data$yield[C
heese.data$lot==x]), FUN.VALUE = 0)
     1      2      3      4      5
3.0425 3.0675 3.1425 3.1625 3.2050
```

Here we defined a function

```
function(x) mean(Cheese.data$yield[Cheese.data$lot==x])
```

Then apply it to each level (discrete value) of variable lot

```
vapply(levels(Cheese.data$lot), FUN= Defined function)
```

Chapter 13 Nonparametric Methods

- Up to now, most of our statistical methods made the normality assumption:
- Single population: $X_1, \ldots, X_n$ i.i.d ~ N(μ,σ^2),
- Two populations: $X_1, \ldots, X_{n1}$ i.i.d ~ N(μ_1,σ_1^2),
  $Y_1, \ldots, Y_{n2}$ i.i.d ~ N(μ_2,σ_2^2),
- ANOVA: $X_{ij} = \mu_i + \epsilon_{ij}$ where ϵ_{ij} i.i.d ~ N(0,σ^2).
- If data are not normal, what to do?

  (a) Big sample size, then approximately normal $\overline{X}$ (CLT)

  (b) Use the nonparametric methods (today's topic)

Chapter 13 Nonparametric Methods

- We focus on deriving 2-sample test here.
- (1) **Paired** test
- Recall how did we get paired t-test.

$\mathbf{H_0}$: $X_i \sim N(\mu_1, \sigma_1^2)$, $Y_i \sim N(\mu_2, \sigma_2^2)$, $\mu_1 = \mu_2$

$\Rightarrow \mathbf{H_0}$: $X_i - Y_i \sim N(\mu_1 - \mu_2, \sigma^2)$, $\mu_1 - \mu_2 = 0$

Hence we used one-sample t-test on X_i-Y_i's instead.

- Similarly for nonparametric paired test, we can focus on X_i-Y_i's but remove the normality assumption.

Nonparametric Paired Test

- (1) **Paired** test

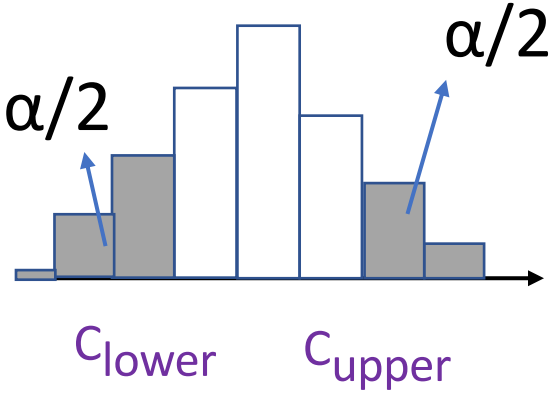
  $\mathbf{H_0}$: X_i-Y_i i.i.d. with median=0 $\Leftrightarrow$ $P(X_i - Y_i \geq 0) = P(X_i \geq Y_i) = 1/2$

versus $\mathbf{H_A}$: $P(X_i \geq Y_i) \neq 1/2$

Let $T = \#(X_i \geq Y_i)$ = # $P(X_i - Y_i \geq 0)$ out of the n pairs.

Under $\mathbf{H_0}$, $X_i \geq Y_i$ occurs with ½ probability. Thus

$T \sim Bin(n, p=½)$ under $\mathbf{H_0}$.



- A nonparametric paired test reject $\mathbf{H_0}$ when

  $\#(X_i \geq Y_i) \geq C_{upper}$ or $\#(X_i \geq Y_i) \leq C_{lower}$, where

$P(Bin(n, p=½) \geq C_{upper}) \leq \alpha/2$ and $P(Bin(n, p=½) \leq C_{lower}) \leq \alpha/2$

(1) Nonparametric Paired Test

Example. (p303 of textbook) We want to compare the resting energy expenditure (REE) for patients with cystic fibrosis (CF) and healthy matched on age, sex, height and weight.

	Pair	1	2	3	4	5	6	7	8	9	10	11	12	13
REE (kcal/day)	X(CF)	1153	1132	1165	1460	1634	1493	1358	1453	1185	1824	1793	1930	2075
	Y(healthy)	996	1080	1182	1452	1162	1619	1140	1123	1113	1463	1632	1614	1836

Since there is no guarantee that the data is normally distributed, and the sample size 13 is small for CLT to apply, we need to apply nonparametric methods.

(1) Nonparametric Paired Test

REE Example. We need to compute the difference D=X-Y.

	Pair	1	2	3	4	5	6	7	8	9	10	11	12	13
REE (kcal/day)	X(CF)	1153	1132	1165	1460	1634	1493	1358	1453	1185	1824	1793	1930	2075
	Y(healthy)	996	1080	1182	1452	1162	1619	1140	1123	1113	1463	1632	1614	1836
	X-Y	157	52	-17	8	472	-126	281	330	72	361	161	316	239
	Sign	+	+	-	+	+	-	+	+	+	+	+	+	+

$T = \#(X_i \geq Y_i) =$ 11 out 13.

One-sided p-value=$P(T \geq 11 \mid T \sim Bin(13, \frac{1}{2}))$

$$= \binom{13}{11}\left(\frac{1}{2}\right)^{11}\left(\frac{1}{2}\right)^{2} + \binom{13}{12}\left(\frac{1}{2}\right)^{12}\left(\frac{1}{2}\right)^{1} + \binom{13}{11}\left(\frac{1}{2}\right)^{13}\left(\frac{1}{2}\right)^{0}$$

=0.0112 (Or from R: 1-pbinom(10.5, size=13, prob=1/2))

Two-side p-value=2(0.0112) = 0.0224.

Hence we reject $\mathbf{H_0}$ at α=0.05 level.

(1.1) Nonparametric Paired Test -- Sign Test

- The above test is a special case of the Binomial test for univariate $\mathbf{H_0}$: median(D)=m_0. It use the fact that the test statistic T=#($D_i \geq m_0$)~Bin(n, p=½) under $\mathbf{H_0}$.
- We may apply this test to values of m_0 other than 0. For example, to test $\mathbf{H_0}$: median(X-Y)=130 in the REE example, T=#(X_i-$Y_i \geq 130$)=4 out of 13. Hence p-value= 2 P(T ≥ 4 | T ~Bin(13, ½)).

- When we applied this univariate test on D=X-Y to test $\mathbf{H_0}$: median(X-Y)=0, the test statistic only depends on the signs of the pair differences as show in the last row of the table. Hence this paired test is called the sign test in textbook.

(1.2) Wilcoxon Signed-rank Test (paired test)

- The sign test only uses the signs, and ignores the magnitudes, of the pair differences X_i-Y_i.

- The **Wilcoxon Signed-rank Test** would use the magnitude.

H_0: X_i-Y_i i.i.d. from a symmetric distribution with mean=median=0

H_A: X_i-Y_i i.i.d. from a symmetric distribution with mean=median≠0

Use test statistics $W=\sum_{i:X_i-Y_i\geq 0} rank(|X_i-Y_i|)$

We reject **H_0** if W is too big or too small.

(1.2) Wilcoxon Signed-rank Test

- The **Wilcoxon Signed-rank Test** would use the magnitude.

H_0: X_i-Y_i i.i.d. from a symmetric distribution with mean=median=0

H_A: X_i-Y_i i.i.d. from a symmetric distribution with mean=median≠0

Test statistics $W=\sum_{i:X_i-Y_i\geq 0} rank(|X_i-Y_i|)$

We reject **H_0** if W is too big or too small.

- The distribution of W is given in Table A.6 for n≤12.

When n>12, a normal approximation is used.

mean μ_W=n(n+1)/4 and variance σ_W^2= n(n+1)(2n+1)/24,

$$Z_W = \frac{W-\mu_W}{\sigma_W} \approx N(0,1)$$

(1.2) Wilcoxon Signed-rank Test

REE Example.

	Pair	1	2	3	4	5	6	7	8	9	10	11	12	13
REE (kcal/day)	X(CF)	1153	1132	1165	1460	1634	1493	1358	1453	1185	1824	1793	1930	2075
	Y(healthy)	996	1080	1182	1452	1162	1619	1140	1123	1113	1463	1632	1614	1836
	X-Y	157	52	-17	8	472	-126	281	330	72	361	161	316	239
	Sign	+	+	-	+	+	-	+	+	+	+	+	+	+
	Rank of $\lvert X_i - Y_i \rvert$	6	3	2	1	13	5	8	11	4	12	7	10	9

W= (6+3+1+13+8+11+4+12+7+10+9) =(1+2+…+13)-2-5 =91-7 =84

μ_W=13(14)/4=45.5, σ_W^2= 13(14)(27)/24=819/4=204.75

So $Z_W = \frac{84-45.5}{\sqrt{204.75}}$ =2.69. From Table A.3

Two-side p-value=2(0.004) = 0.008.

Hence we reject $\mathbf{H_0}$ at α=0.05 level.

(2) Wilcoxon Rank-sum Test (unpaired test)

- For samples from two independent populations.
- Assumptions:

  X_i i.i.d. from a distribution **F** are independent of

  Y_i+a i.i.d. from a distribution **F**, where a is a constant.
- $\mathbf{H_0}$: $a=0$ versus $\mathbf{H_A}$: $a\neq 0$
- Methods: rank X_i's and Y_i's together $\{X_1,\ldots,X_n,Y_1,\ldots,Y_m\}$.

Rank-sum test statistics $W=\begin{cases}\sum_{i=1}^{n} rank(X_i) & if\ n \le m \\ \sum_{j=1}^{m} rank(Y_j) & if\ n > m\end{cases}$

We reject $\mathbf{H_0}$ if W is too big or too small.

(2) Wilcoxon Rank-sum Test (unpaired test)

- For samples from two independent populations, we should use the **Wilcoxon Rank-sum** test statistics

$$W=\begin{cases}\sum_{i=1}^{n} rank(X_i) & if\ n \le m \\ \sum_{j=1}^{m} rank(Y_j) & if\ n > m\end{cases}$$

We reject $\mathbf{H_0}$ if W is too big or too small.

The distribution of W under $\mathbf{H_0}$ is in Table A.7(m ≤10, n≤10)

For other cases, we use the normal approximation where mean μ_W=(n+m+1)min(n,m)/2 and variance σ_W^2= nm(nm+1)/12,

$$Z_W = \frac{W-\mu_W}{\sigma_W} \approx N(0,1)$$

(2) Wilcoxon Rank-sum Test (unpaired test)

- Example Data: Two groups A: 11,32 and B: 28,52,42
- Together {11,28,32,42,52}

	A	B	A	B	B
Rank	1	2	3	4	5

Hence W = 1+3 = 4. Notice that μ_W=6(2)/2 =6.

One-sided p-value = P(W≤4), from Table A.7, since n=2, m=3 (larger one is n_2=3, and n_1=2) no normal approximation.

One-sided p-value = 0.2.

Two-sided p-value = P(W≤4)+P(W≥8) =2(0.2) =0.4.

We fail to reject $\mathbf{H_0}$ at α=0.05 level on this data set.

(2) Wilcoxon Rank-sum Test (unpaired test)

TABLE A.7

Distribution functions of W, the Wilcoxon rank sum test

	$n_2 = 3$		
W_0	$n_1 = 1$	2	3
1	0.25		
2	0.50		
3		0.10	
4		0.20	
5		0.40	
6		0.60	0.05
7			0.10
8			0.20
9			0.35
10			0.50

(2) Wilcoxon Rank-sum Test (unpaired test)

- The **Wilcoxon Rank-sum Test** do make assumptions:
  X_i i.i.d. from a distribution F are independent of
  Y_i+a i.i.d. from a distribution F, where a is a constant.
- $\mathbf{H_0}$: $a=0$ versus $\mathbf{H_A}$: $a\neq 0$
- It is assumed that samples from the two populations have the same shape, and are testing that they also have the same median. If the shapes are different, the size (Type I error rate) may be incorrect.
- **Nonparametric methods**: **No normality assumption**. But still make assumptions! (NOT no assumptions)

R commands for Nonparametric tests

All the nonparametric tests covered in this chapter are already programmed into the R base package. We can simply call their commands.

- Sign test:

```
diff<- x-y
binom.test( x=sum(diff>0), n=length(diff), p=0.5, alternative="two.sided")
```

- Wilcoxon signed-rank test:

```
wilcox.test (x, y, paired=TRUE, alternative="two.sided")
```

- Wilcoxon rank-sum test:

```
wilcox.test (x, y, paired=FALSE, alternative="two.sided")
```

R commands for Nonparametric tests

- **Using the response times from the mini-project in Lab1. This is the same data used in the two-sample t-tests handout, the data importation and processing are similar.**

```
>MyTime <- read.table(file="ResponseTime2.txt", header=TRUE)
>#  For paired data, create the new variable of difference,
>#   then do one sample test
>diff<- MyTime$Small-MyTime$Xlarge
>#  Sign-test: check the binomial distribution
>binom.test(x=sum(diff>0), n=length(diff), p=0.5, alternative="two.sided")

        Exact binomial test

data:  sum(diff > 0) and length(diff)
number of successes = 11, number of trials = 15, p-value = 0.1185
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4489968 0.9221285
sample estimates:
probability of success
             0.7333333
```

R commands for Nonparametric tests

```
> # Wilcoxon Signed-rank test (paired)
> wilcox.test (x= MyTime$Small, y= MyTime$Xlarge, paired=TRUE, alternative="two.sided")

        Wilcoxon signed rank test with continuity correction

data:  MyTime$Small and MyTime$Xlarge
V = 91.5, p-value = 0.0156
alternative hypothesis: true location shift is not equal to 0

Warning messages:
1: In wilcox.test.default(x = MyTime$Small, y = MyTime$Xlarge, paired = TRUE,  :
  cannot compute exact p-value with ties
2: In wilcox.test.default(x = MyTime$Small, y = MyTime$Xlarge, paired = TRUE,  :
  cannot compute exact p-value with zeroes
```

- At $\alpha=0.05$ level, the signed-rank test rejects the null hypothesis of two equal distributions. But the sign-test fail to reject at $\alpha= 0.05$ level.

R commands for Nonparametric tests

- Although the data is paired, we ignore the pairing to illustrate the syntax for unpaired test next.

```
> # Wilcoxon Rank-sum test (unpaired)
> wilcox.test (x= MyTime$Small, y= MyTime$Xlarge, paired=FALSE, alternative="two.sided")
        Wilcoxon rank sum test with continuity correction

data:  MyTime$Small and MyTime$Xlarge
W = 178, p-value = 0.006972
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = MyTime$Small, y = MyTime$Xlarge, paired = FALSE,  :
  cannot compute exact p-value with ties
```

- At $\alpha=0.05$ level, the rank-sum test rejects the null hypothesis of two equal distributions.

R commands for Nonparametric tests

- Notice the p-values differ here for different tests: Sign test 0.1185, Sign-rank test 0.0156, rank-sum test 0.006972.

- Generally, when the two sample means do differ (alternative hypothesis is true), often the tests with stronger assumptions have smaller p-values (since their power is higher). But when the pairings do contribute to reduce variation, the paired test usually has smaller p-value (not the case on this data) due to its higher power.

R commands for Nonparametric tests

The unpaired data often is put in a format with one column of group indicator and one column of the variable of interest. (See the two samples t-test notes). We show the syntax of Wilcoxon test for data in that format also.

```
> ## Rearrange the data
> small.time<-data.frame(time=MyTime$Small, group='Small') #two colum
ns: time and group
> xlarge.time<-data.frame(time=MyTime$Xlarge, group='Xlarge')
> new.data<- rbind(small.time, xlarge.time) #merge two sets above
>
> # Wilcoxon Rank-sum test (unpaired)
> wilcox.test(time~group, data=new.data, paired=FALSE, alternative="t
wo.sided")

        Wilcoxon rank sum test with continuity correction

data:  time by group
W = 178, p-value = 0.006972
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(531L, 843L, 578L, 563L, 625L, 704L,  :
  cannot compute exact p-value with ties
```

Summary

Today, we finished ANOVA R commands and
then went over the nonparametric tests in chapter 13

- Two-sample tests: Sign test, Wilcoxon signed-rank test and Wilcoxon rank-sum test.
- Know when to use which (paired versus independent two samples).
- Can use R to do them.

- Homework 5 is due in one week (which covers topics in Module 7 ANOVA)