Chapter 10: Comparing multiple independent populations

(Ott & Longnecker Sections: 14.2 and 14.5)

https://dzwang91.github.io/stat324/

Outline

What we have learnt...

- One sample tests: test population mean/median/proportion
- Two independent sample tests: compare two independent populations
- Two paired sample tests: compare two dependent populations

What we have learnt...

- One sample tests: test population mean/median/proportion
- Two independent sample tests: compare two independent populations
- Two paired sample tests: compare two dependent populations

A natural extension: How do we compare multiple independent populations?

Example

- Four new formulations of rat poison are being tested, call them 1, 2, 3 and 4. All of the poisons work by thinning the blood, so the response of interest is the time it takes for the blood to coagulate. A longer blood coagulation time indicates a more effective poison.
- 24 rats were randomly selected, and then randomized to the four poisons. They were fed the poison, and then after a specified length of time, their blood was drawn and the time to blood coagulation was measured. The data is below:

Treatment								Sample Mean
1	62	60	63	59				61
2	63	67	71	64	65	66		66
3	68	66	71	67	68	68		68
4	56	62	60	61	63	64	63	59	61

Example continued

- We'd like to know if any of these poisons results in a different coagulation time than any of the others.

Example continued

- We'd like to know if any of these poisons results in a different coagulation time than any of the others.
- Notations:
  - μ_1: population mean for poison 1
  - μ_2: population mean for poison 2
  - μ_3: population mean of poison 3
  - μ_4: population mean of poison 4

Example continued

- We'd like to know if any of these poisons results in a different coagulation time than any of the others.
- Notations:
  - μ_1: population mean for poison 1
  - μ_2: population mean for poison 2
  - μ_3: population mean of poison 3
  - μ_4: population mean of poison 4
- Hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_A: At least one mean differs from one other mean.

Example continued

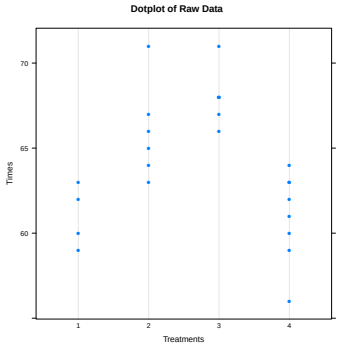
- We'd like to know if any of these poisons results in a different coagulation time than any of the others.
- Notations:
  - μ_1: population mean for poison 1
  - μ_2: population mean for poison 2
  - μ_3: population mean of poison 3
  - μ_4: population mean of poison 4
- Hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

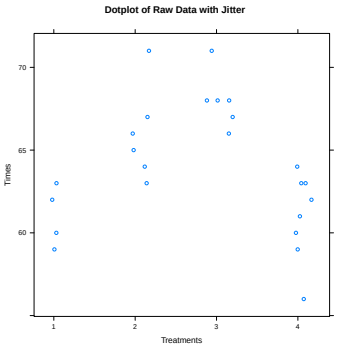
H_A: At least one mean differs from one other mean.

- Question: how can we do hypothesis testing in this setting? What is the test statistic?

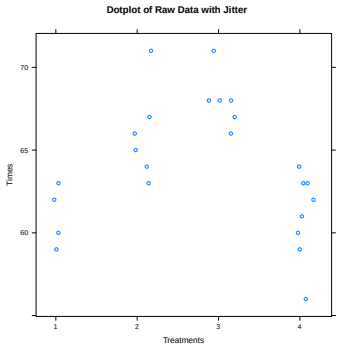
Example continued



Example continued

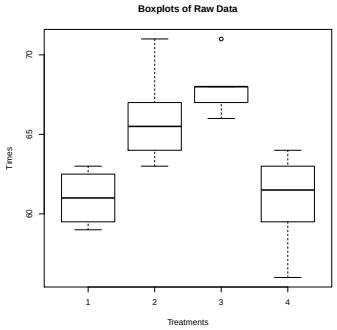


Example continued



The treatments seem to differ somewhat. Treatments 2 and 3 seem to have generally higher means than 1 and 4.

Example continued



It is clear that treatment 3 might be slightly higher than treatment 2.

Example continued

- Can we use the paired t test in this example? If we can, how do we implement the test?

Example continued

- Can we use the paired t test in this example? If we can, how do we implement the test?
  Yes we can, test each pair of means.

Example continued

- Can we use the paired t test in this example? If we can, how do we implement the test?
  Yes we can, test each pair of means.
- What's the limit of testing each pair of means?

Example continued

- Can we use the paired t test in this example? If we can, how do we implement the test?
  Yes we can, test each pair of means.
- What's the limit of testing each pair of means?
  In the worst case, need test $\binom{m}{2} = O(m^2)$ times for m treatments.

Example continued

- Can we use the paired t test in this example? If we can, how do we implement the test?
  Yes we can, test each pair of means.
- What's the limit of testing each pair of means?
  In the worst case, need test $\binom{m}{2} = O(m^2)$ times for m treatments.

A new approach: analysis of variance (ANOVA)

Example continued

- Can we use the paired t test in this example? If we can, how do we implement the test?
  Yes we can, test each pair of means.
- What's the limit of testing each pair of means?
  In the worst case, need test $\binom{m}{2} = O(m^2)$ times for m treatments.

A new approach: analysis of variance (ANOVA)

- By the end of this lecture, think about why we call it analysis of variance.

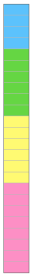
Outline

Notations

- Let t be the number of treatments.
- Let i index the treatments.
- Let n_i be the number of observations in treatment i.
- Let $N = \sum_{i=1}^{t} n_i$ be the total sample size.
- Let y_{ij} be observation j from treatment i.
- Let $\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ be the sample mean for treatment i.
- Let $\bar{y}_{\cdot\cdot} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}}{N}$ be the sample grand mean.

Sample mean for each treatment

Grand mean



$\bar{y}_{\cdot\cdot}$

Key idea of ANOVA

Where does the variability come from?

Key idea of ANOVA

Where does the variability come from?

- The key decomposition:

  Observation = Grand Mean + Deviation of Treatment Mean from Grand Mean + Deviation of Observation from Treatment Mean
- In notation,

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

Key idea of ANOVA

Where does the variability come from?

- The key decomposition:

  Observation = Grand Mean + Deviation of Treatment Mean from Grand Mean + Deviation of Observation from Treatment Mean
- In notation,

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

- Thus,

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

Ingredient 1: Sum of squares

- Variability decomposition:

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

Ingredient 1: Sum of squares

- Variability decomposition:

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

  - $\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$: sum of squares total
  - $\sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2$: sum of squares treatment (SS between)
  - $\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$: sum of squares error (SS within)
- In words,

  SSTot = SSTrt (SS between) + SSE (SS within)
- Proof in next chapter

Ingredient 2: Degrees of freedom

- For SSTot, degrees of freedom is $N - 1$
- For SSTrt, degrees of freedom is $t - 1$
- For SSE, degrees of freedom is $N - t$.

Ingredient 2: Degrees of freedom

- For SSTot, degrees of freedom is $N - 1$
- For SSTrt, degrees of freedom is $t - 1$
- For SSE, degrees of freedom is $N - t$.

What do you find?

Ingredient 2: Degrees of freedom

- For SSTot, degrees of freedom is $N - 1$
- For SSTrt, degrees of freedom is $t - 1$
- For SSE, degrees of freedom is $N - t$.

What do you find?

$$df_{Tot} = df_{Trt} + df_E$$

Ingredient 3: Mean squares and test staistic

- Mean squares:
  - $MSTrt = \frac{SSTrt}{df_{Trt}}$
  - $MSE = \frac{SSE}{df_E}$

Ingredient 3: Mean squares and test staistic

- Mean squares:
  - $MSTrt = \frac{SSTrt}{df_{Trt}}$
  - $MSE = \frac{SSE}{df_E}$
- Test statistic:
  - $F = \frac{MSTrt}{MSE}$
  - It is the ratio of the between variability to the within variability

Summarizing in a table

Source	SS	df	MS	F	p-value
Treat	SSTrt	$t-1$	MSTrt$=\frac{SSTrt}{df_{Trt}}$	F $=\frac{MSTrt}{MSE}$	?
Error	SSE	$N-t$	MSE $=\frac{SSE}{df_E}$		
Total	SSTot	$N-1$			

Summarizing in a table

Source	SS	df	MS	F	p-value
Treat	SSTrt	$t-1$	MSTrt$=\frac{SSTrt}{df_{Trt}}$	F $=\frac{MSTrt}{MSE}$	?
Error	SSE	$N-t$	MSE $=\frac{SSE}{df_E}$		
Total	SSTot	$N-1$			

- How do we calculate the p-value?
- What is the distribution of the test statistic given the null hypothesis is true?
- What are assumptions we need to assume?

F distribution

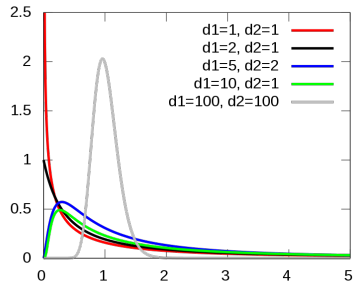
If we assume

- The data are independent within and between treatments
- The variances are the same for all treatments
- Each treatment has a normal distribution

then

- the distribution of the test statistic F is called an F distribution
- it has two parameters, called the numerator df and denominator df.
- The numerator df is df_{Trt}, and the denominator df is df_E.

Pdf of F distribution



Here $d1$ is the numerator df, $d2$ is the denominator df.

ANOVA table

If we are willing to assume:

- The data are independent within and between treatments
- The variances are the same for all treatments
- Each treatment has a normal distribution

then

Source	SS	df	MS	F	p-value
Treat	SSTrt	$t-1$	MSTrt$=\frac{SSTrt}{df_{Trt}}$	$F=\frac{MSTrt}{MSE}$	$P(F_{df_{Trt},df_E} > F)$
Error	SSE	$N-t$	MSE $=\frac{SSE}{df_E}$		
Total	SSTot	$N-1$			

- Use F table to calculate the p-value

Example

For our blood coagulation data, the ANOVA table is:

Source	SS	df	MS	F	p-value
Treat (between)	228	3	76	13.57	
Error (within)	112	20	5.6		
Total	340	23			

F table

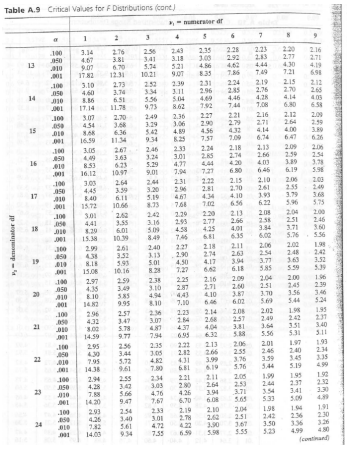


Table A.9 Critical Values for F Distributions (cont.)

Back to example

For our blood coagulation data, the ANOVA table is:

Source	SS	df	MS	F	p-value
Treat (between)	228	3	76	13.57	< 0.001
Error (within)	112	20	5.6		
Total	340	23			

Back to example

For our blood coagulation data, the ANOVA table is:

Source	SS	df	MS	F	p-value
Treat (between)	228	3	76	13.57	< 0.001
Error (within)	112	20	5.6		
Total	340	23			

Since the p-value is quite small, we would reject the null, and conclude that at least one poison has a different mean coagulation time than another.

Outline

Assumptions of ANOVA

- The data are independent within and between treatments. (check from the story.)

Assumptions of ANOVA

- The data are independent within and between treatments. (check from the story.)
- Each treatment has a normal distribution. (check using QQ plot.)

Assumptions of ANOVA

- The data are independent within and between treatments. (check from the story.)
- Each treatment has a normal distribution. (check using QQ plot.)
- The variances are the same for all treatments

Assumptions of ANOVA

- The data are independent within and between treatments. (check from the story.)
- Each treatment has a normal distribution. (check using QQ plot.)
- The variances are the same for all treatments

How do we check equal variance?

Assumptions of ANOVA

- The data are independent within and between treatments. (check from the story.)
- Each treatment has a normal distribution. (check using QQ plot.)
- The variances are the same for all treatments

How do we check equal variance?

Approach 1: Use residuals vs. fitted values plot

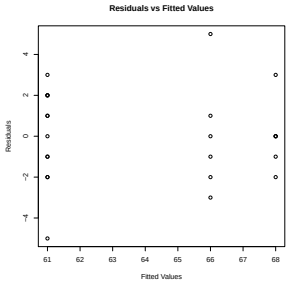
Approach 2: Ratio of SDs

Fitted values, residuals

- The fitted values are the treatment means $\bar{y}_{i\cdot}$.
- The residuals are the differences between the observed data(y_{ij}) and the treatment means $y_{ij} - \bar{y}_{i\cdot}$.
- In fact, the sum of squares error (SSE) is the sum of squares of the residuals.
- **Key idea:** If the spreads of residuals are about the same for each treatment, then we are safe to assume equal variance.

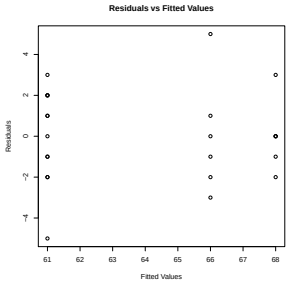
Residuals vs Fitted values plot

Assume there are 3 treatments. The y-axis is called 'Residuals' and the x-axis is called 'Fitted Values'.



Residuals vs Fitted values plot

Assume there are 3 treatments. The y-axis is called 'Residuals' and the x-axis is called 'Fitted Values'.



There might be slightly less spread in the last group, but they're close enough.

Equal variance does not always hold

Equal variance does not always hold

Equal variance does not always hold

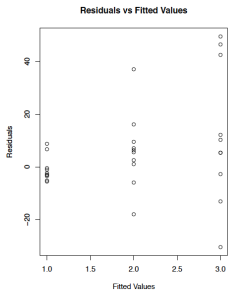
The SD of weights of 5 elephants will tend to be larger than the SD of weights of 5 hamsters.

Equal variance does not always hold

- It is often the case that variability will increase with increasing fitted values.

Equal variance does not always hold

- It is often the case that variability will increase with increasing fitted values.
- An example: we see the classic funnel pattern where the variability increases with the fitted value.



Ratio of SDs

- Another option for checking equal variance is to use the ratio of SDs guideline first mentioned when comparing two populations.
- Since there are now more than two groups, it is typical to take the ratio of the largest and smallest sample SDs - if this ratio passes the test, then every other pair will as well.
- For our example, the sample SDs for the four groups are 1.83, 2.83, 1.67, and 2.62. The ratio of the smallest to the largest is $1.67/2.83 = 0.59$, which falls between 0.5 and 2.0, so assuming the variances equal should be safe.

Outline

Multiple comparisons following significant ANOVA

- Consequence of ANOVA:
  - if we do not reject the null, we're done.
  - if we reject, we only know that at least one mean differs from at least one other mean, but not how many means differ, or which ones, or by how much.

Multiple comparisons following significant ANOVA

- Consequence of ANOVA:
  - if we do not reject the null, we're done.
  - if we reject, we only know that at least one mean differs from at least one other mean, but not how many means differ, or which ones, or by how much.
- Two sample t test for i-th treatment and j-th treatment:
  - $H_0 : \mu_i = \mu_j$ vs. $H_A : \mu_i \neq \mu_j$
  - use $t = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{S_p\sqrt{(1/n_i + 1/n_j)}}$ (t distribution with df=$n_i + n_j - 2$ given H_0 is true)
  - compute rejection region or p-value to make a conclusion

Multiple comparisons following significant ANOVA

- Consequence of ANOVA:
  - if we do not reject the null, we're done.
  - if we reject, we only know that at least one mean differs from at least one other mean, but not how many means differ, or which ones, or by how much.
- Two sample t test for i-th treatment and j-th treatment:
  - $H_0 : \mu_i = \mu_j$ vs. $H_A : \mu_i \neq \mu_j$
  - use $t = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{S_p\sqrt{(1/n_i + 1/n_j)}}$ (t distribution with df$=n_i + n_j - 2$ given H_0 is true)
  - compute rejection region or p-value to make a conclusion
  - $100(1-\alpha)\%$ CI on $\mu_i - \mu_j$:
    $$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) \pm t_{n_i+n_j-2,\alpha/2} S_p \sqrt{(1/n_i + 1/n_j)}$$
    If the CI contains 0, then we don't reject the null hypothesis.

Why does CI approach work for two-sided tests?

If 0 is in $100(1-\alpha)\%$ CI, then

$$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - t_{n_i+n_j-2,\alpha/2} S_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \leq 0 \leq (\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) + t_{n_i+n_j-2,\alpha/2} S_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$\downarrow$$

$$t_{obs} = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{S_p\sqrt{(1/n_i + 1/n_j)}} \leq t_{n_i+n_j-2,\alpha/2} \qquad t_{obs} = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{S_p\sqrt{(1/n_i + 1/n_j)}} \geq -t_{n_i+n_j-2,\alpha/2}$$

$$\downarrow$$

$$|t_{obs}| \leq t_{n_i+n_j-2,\alpha/2}$$

Therefore, we do not reject H_0 at significance level α.

Back to example

- Suppose we want 95% CIs, $t_{20,0.025} = 2.086$, and MSE $= 5.6$, so
  - Trt 1 vs Trt 2:
    $61 - 66 \pm 2.086\sqrt{5.6(1/4 + 1/6)} = -5 \pm 3.19 = (-8.19, -1.81)$
  - Trt 1 vs Trt 3:
    $61 - 68 \pm 2.086\sqrt{5.6(1/4 + 1/6)} = -7 \pm 3.19 = (-10.19, -3.81)$
  - Trt 1 vs Trt 4:
    $61 - 61 \pm 2.086\sqrt{5.6(1/4 + 1/8)} = 0 \pm 3.02 = (-3.02, 3.02)$
  - Trt 2 vs Trt 3:
    $66 - 68 \pm 2.086\sqrt{5.6(1/6 + 1/6)} = -2 \pm 2.85 = (-4.85, 0.85)$
  - Trt 2 vs Trt 4:
    $66 - 61 \pm 2.086\sqrt{5.6(1/6 + 1/8)} = 5 \pm 2.67 = (2.33, 7.67)$
  - Trt 3 vs Trt 4:
    $68 - 61 \pm 2.086\sqrt{5.6(1/6 + 1/8)} = 7 \pm 2.67 = (4.33, 9.67)$

Back to example

- Suppose we want 95% CIs, $t_{20,0.025} = 2.086$, and MSE $= 5.6$, so
  - Trt 1 vs Trt 2:
    $61 - 66 \pm 2.086\sqrt{5.6(1/4 + 1/6)} = -5 \pm 3.19 = (-8.19, -1.81)$
  - Trt 1 vs Trt 3:
    $61 - 68 \pm 2.086\sqrt{5.6(1/4 + 1/6)} = -7 \pm 3.19 = (-10.19, -3.81)$
  - Trt 1 vs Trt 4:
    $61 - 61 \pm 2.086\sqrt{5.6(1/4 + 1/8)} = 0 \pm 3.02 = (-3.02, 3.02)$
  - Trt 2 vs Trt 3:
    $66 - 68 \pm 2.086\sqrt{5.6(1/6 + 1/6)} = -2 \pm 2.85 = (-4.85, 0.85)$
  - Trt 2 vs Trt 4:
    $66 - 61 \pm 2.086\sqrt{5.6(1/6 + 1/8)} = 5 \pm 2.67 = (2.33, 7.67)$
  - Trt 3 vs Trt 4:
    $68 - 61 \pm 2.086\sqrt{5.6(1/6 + 1/8)} = 7 \pm 2.67 = (4.33, 9.67)$
- The conclusion is that treatments 2 and 3 are the same, and 1 and 4 are the same, but 2 and 3 differ from 1 and 4.

Letter code

- The information is summarized by sorting the treatment means from largest to smallest, and then adding letter codes. Two treatments share a letter if they do not differ significantly:

Treatment	Sample Mean	Letter Code
3	68	A
2	66	A
1	61	B
4	61	B

Summary

If we end up doing a bunch of pairwise tests, why do we use ANOVA?

Summary

If we end up doing a bunch of pairwise tests, why do we use ANOVA?

- Practically, if there are a lot of treatments, if there are really no differences between the treatments, doing one ANOVA could save time over doing many pairwise tests.
- Theoretically, the F-test is the most powerful test for the hypotheses we specified, provided all of our assumptions are met.

Outline

ANOVA in R

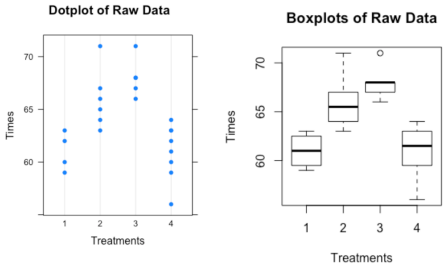
```
> #enter data
> times <- c(62, 60, 63, 59, 63, 67, 71, 64, 65, 66, 68,
+            66, 71, 67, 68, 68, 56, 62, 60, 61, 63, 64,
63, 59)
> diets <- c(rep(1, 4), rep(2, 6), rep(3, 6), rep(4, 8))
>
> #make diets an explicit factor
> dietsf <- factor(diets)
```


ANOVA in R

```
> #Run ANOVA: use aov
> mod <- aov(times ~ dietsf)
> summary(mod)
            Df Sum Sq Mean Sq F value   Pr(>F)
dietsf       3    228    76.0   13.57 4.66e-05 ***
Residuals   20    112     5.6
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

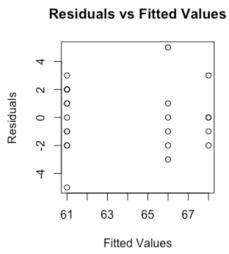

ANOVA in R

```
> #dotplot and boxplot
> library(lattice)
> dotplot(times ~ dietsf, ylab = "Times", xlab = "Treatme
nts", main = "Dotplot of Raw Data")
> boxplot(times ~ diets, ylab = "Times", xlab = "Treatmen
ts", main = "Boxplots of Raw Data")
```



ANOVA in R

```
> #residuals vs fitted plot
> plot(residuals(mod) ~ fitted(mod), ylab = "Residuals",
xlab = "Fitted Values", main = "Residuals vs Fitted Value
s")
```



What's the next?

We'll introduce linear regression in next lecture.