# Designing Experiments, ANOVA and Correlation Coefficients

## Biology 683

## Lecture 6

Heath Blackmon

# Last week

1. How do you choose the number of permutations to run?

2. Which t-test doesn’t assume the variables being studied have a normal distribution in the population?

# Today

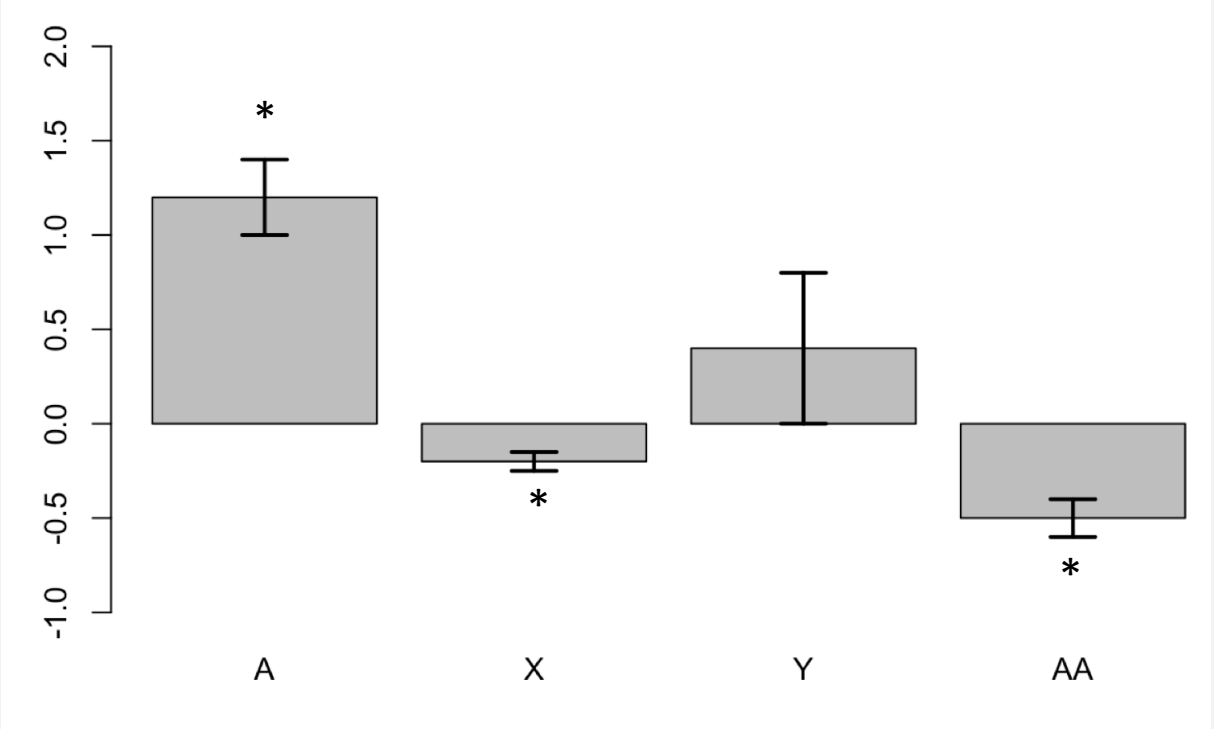
1) Experimental design
2) ANOVA
3) Correlation and Covariance

# Statistical Significance versus Biological Importance

A *p*-value is not a sufficient description of the effect of a treatment or variable

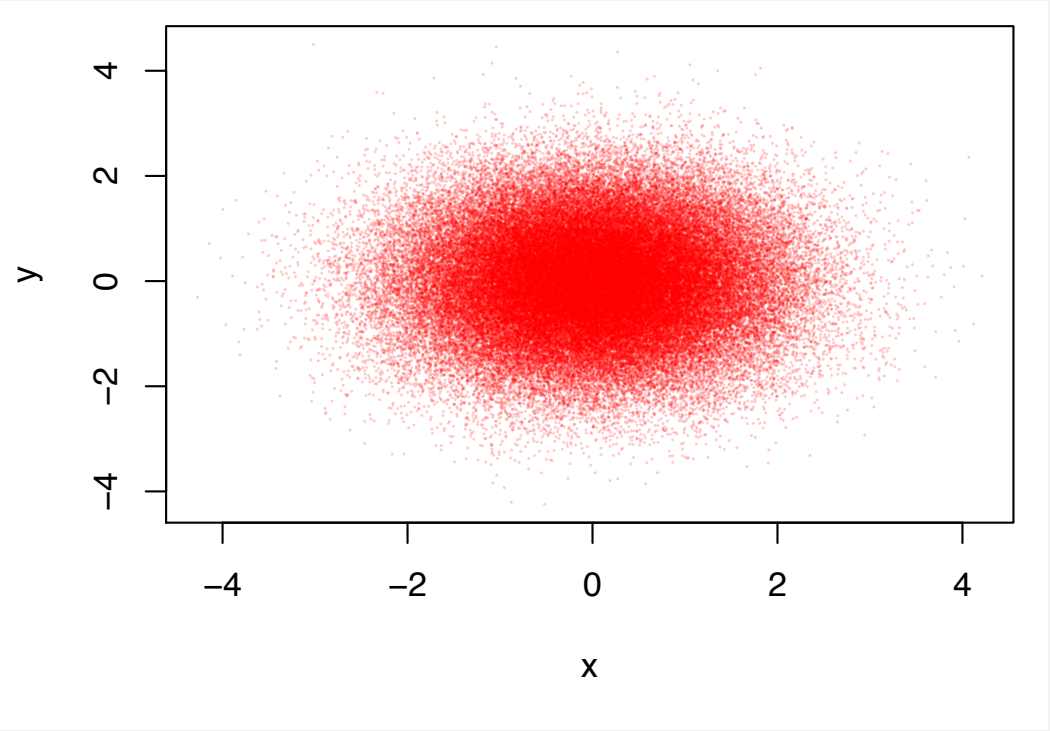
Even very small effects can be shown to be highly statistically significant with large enough sample sizes

This issue is especially problematic in the new era of big data (GWAS, Expression studies)



# Statistical Significance versus Biological Importance

```
data:  x and y
t = 2.154, df = 99998, p-value = 0.03124
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.00061360 0.01300895
sample estimates:
        cor
0.006811538
```

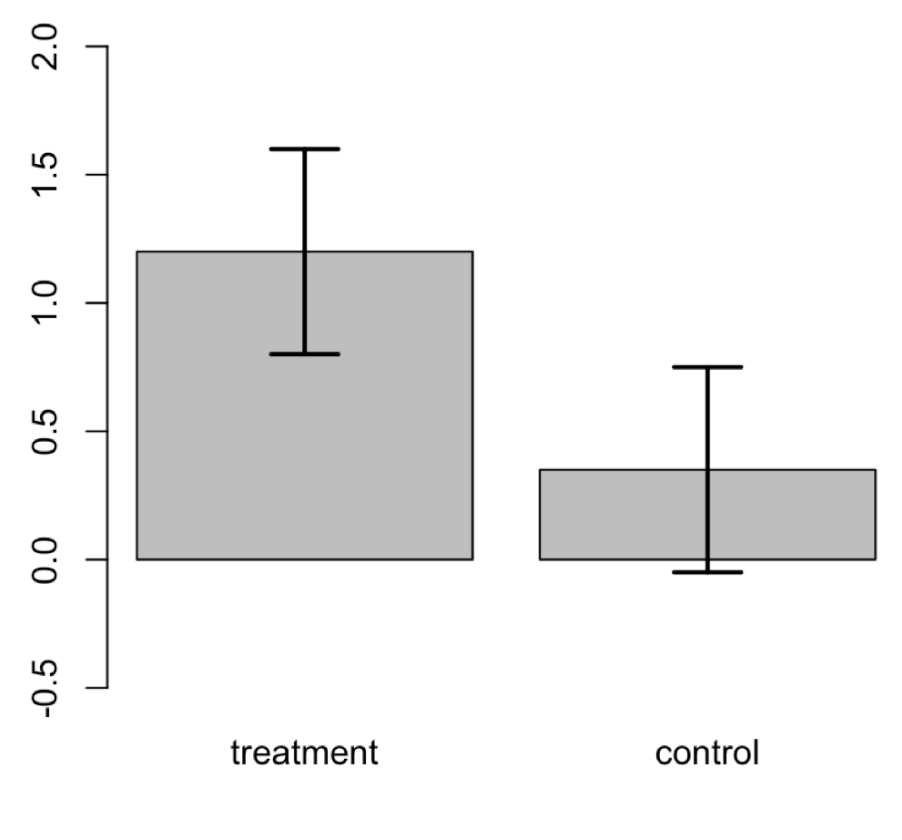


# Effect size!

Statistical significance is not enough – you also need to calculate and report the magnitude of the effects. If the effect is exceedingly small, then it might not be interesting. Biological significance can only be determined in light of the system under study.

Why is this important in basic research / applied research?

# Common mistake




- The treatment is significantly different from zero
- The control is not significantly different from zero
- So… the treatment must be different from the control, right?

- **No. You must directly compare the treatment to the control. Lots of possible reasons for control not to be zero!**

# Designing Experiments

Some considerations we've already discussed:

1. Confounding variables
2. Experiment bias
3. Multiple control groups
4. Randomization
5. Single-blind and double-blind experiments

# Some New Concepts

**Balance**
A balanced design has an equal sample size in each treatment and control group – balanced designs have the most power.

**Blocking**
Similar to a paired, but possibly with multiple treatments and controls. Example on next slide.

**Extreme Treatments**
By using extremely high values of the treatment, we might be able to show an effect. We would have to consider whether these extreme values have biological relevance

**Factors**
A factor is a single treatment variable whose effects are of interest. An experiment can have more than one factor (light and humidity, for instance, would be two different factors). We sometimes wish to test multiple factors simultaneously, because they might interact

# Block what you can, randomize what you cannot.

Shelf 1

| | |
|---|---|
| Treatment A | Treatment B |
| Treatment C | Control |

Shelf 2

| | |
|---|---|
| Treatment C | Treatment A |
| Control | Treatment B |

Shelf 3

| | |
|---|---|
| Control | Treatment C |
| Treatment B | Treatment A |

Shelf 4

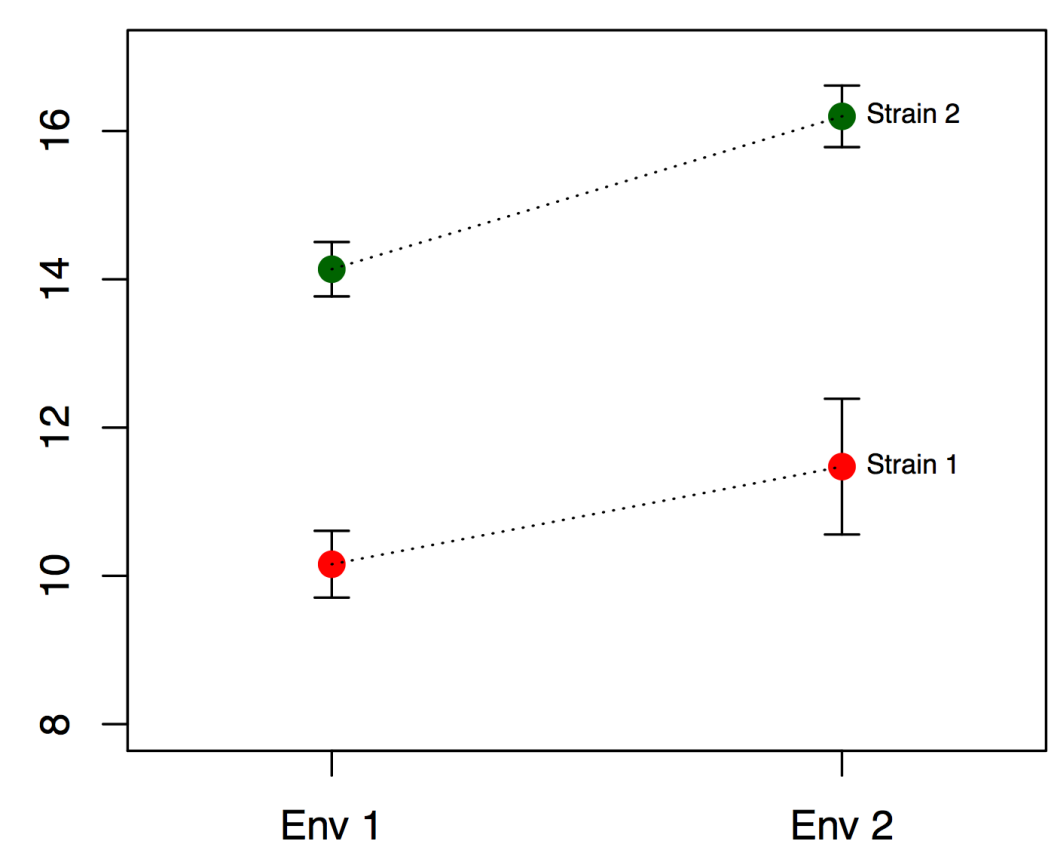
| | |
|---|---|
| Treatment B | Control |
| Treatment A | Treatment C |

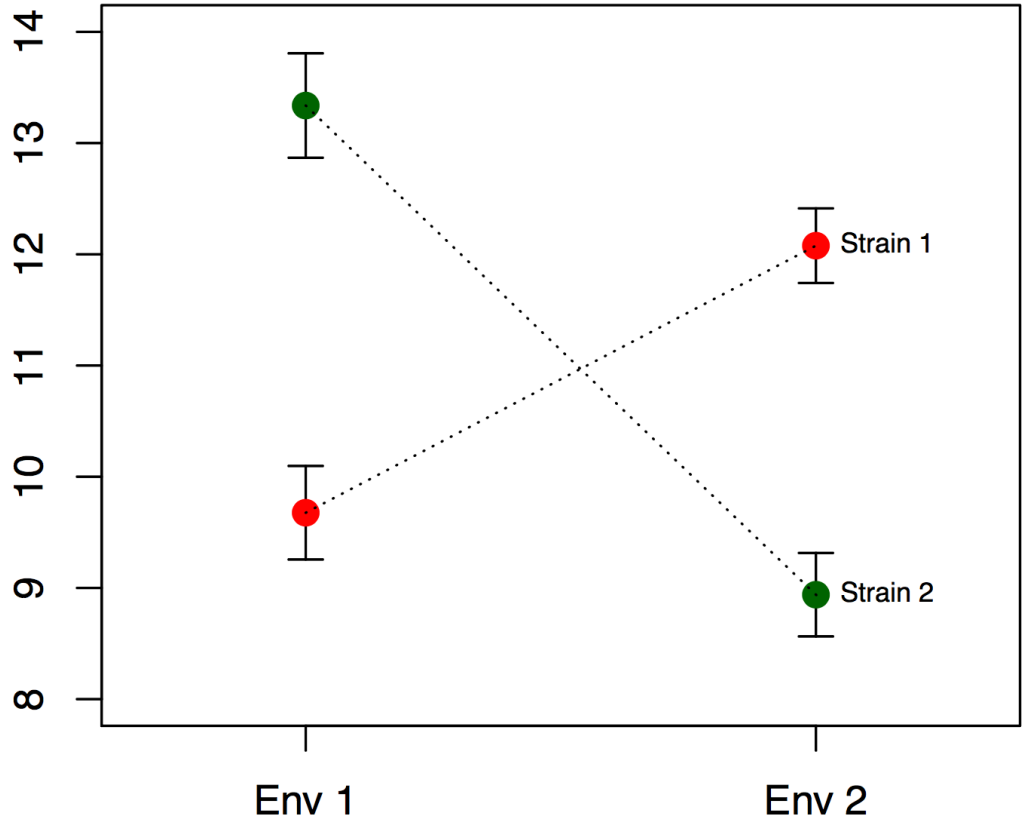
What is being blocked?
What is being randomized?
What else might we randomize?

# Example of Multiple Factors

- An experiment has multiple factors when you are interested in the effects of more than one variable.
- An example might be the role of genotype and environment on growth rate.
- Imagine there are two strains and you want to test their growth in two environments.
- A factorial design would require testing each strain in each environment for a total of four treatments
- This experiment would be called a **full factorial design**

# Multiple factors and interactions



# Precision

- It is possible to plan your experiment to achieve a desired level of precision
- The catch is that you have to know something about the expected standard deviation of your response variable
- The other catch is that you will often find the sample sizes needed to be shockingly high

# Precision

- Formula for necessary sample size is test specific see page WS 447-449
- For comparing two means

$$n \approx 8\left(\frac{\sigma}{margin\ of\ error}\right)^2$$

Where n is the sample size required for each group and margin of error is the desired half width

- A smaller margin of error requires a larger sample size
- A larger standard deviation requires a larger sample size

# Power

- A more useful and common way to estimate the necessary sample size is to focus on power

- **Power**: the probability of rejecting the null hypothesis for a hypothesized difference in means

- **Power calculations are often necessary**

  - When you fail to reject the null, was it because you had insufficient power to detect a reasonable departure?

  - When you apply for a grant or institutional approval, the agency often wants to know that your experimental design will have the power to detect a difference if one exists.

# Calculating Power

- Imagine an experiment testing the difference between two means in an unpaired design
- The appropriate test is a two-sample *t*-test

- Example:
- A 6 week old chick weighs around 250 grams, s.d. around 50
- Feeding them a supplement results in about a 10-20 percent increase in weight at 6 weeks.
- What sample size would I need to detect an effect of a supplement (i.e., reject the null hypothesis that it has no effect)

# Power for a Two-sample t-test

For a power of 0.8 (i.e., an 80 percent chance of rejecting the null hypothesis when it's not true), I need a sample size of:

$$n = 16\left(\frac{\sigma}{D}\right)^2$$

*D* is the expected difference in means.

$$n = 16\left(\frac{50}{275 - 250}\right)^2$$

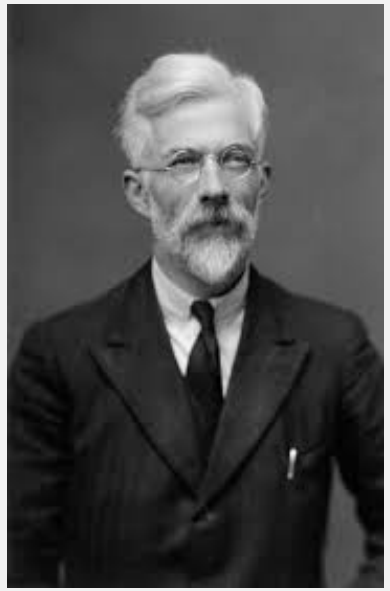
$$n = 64$$

*For each treatment*

# Power for a Two-sample t-test

```
> power.t.test(delta=25,sd=50, power=.80)

     Two-sample t test power calculation

              n = 63.76576
          delta = 25
             sd = 50
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

# Fisher



**R. A. Fisher**

Already heard about “Fisher’s Exact Test” Among many other contributions, Fisher also invented analysis of variance and the F-distribution

He also was one of the figures in the Modern Synthesis – reconciling the facts of genetics with the idea of natural selection.

Popularized the p <= 0.05 as general rule of thumb (not a hard fast rule as applied today)

FYI he was a firm believer in eugenics, racist, and spoke against inference that smoking caused cancer.

**The Correlation between relatives on the supposition of Mendelian Inheritance**

By R. A. FISHER, B.A.

*Communicated by* Professor J. ARTHUR THOMSON

With Four Figures in Text

(*MS. received* 15 *June* 1918. *Read* 8 *July* 1918. *Issued separately* 1 *October* 1918)

ON THE "PROBABLE ERROR" OF A COEFFICIENT OF CORRELATION DEDUCED FROM A SMALL SAMPLE

Author's Note (CMS 1. 2a)

This is the second of three papers dealing with the sampling errors of correlation coefficients covering the cases (*i*) "The frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, Vol. 10, pp. 507–521, 1915.

# Analysis of Variance

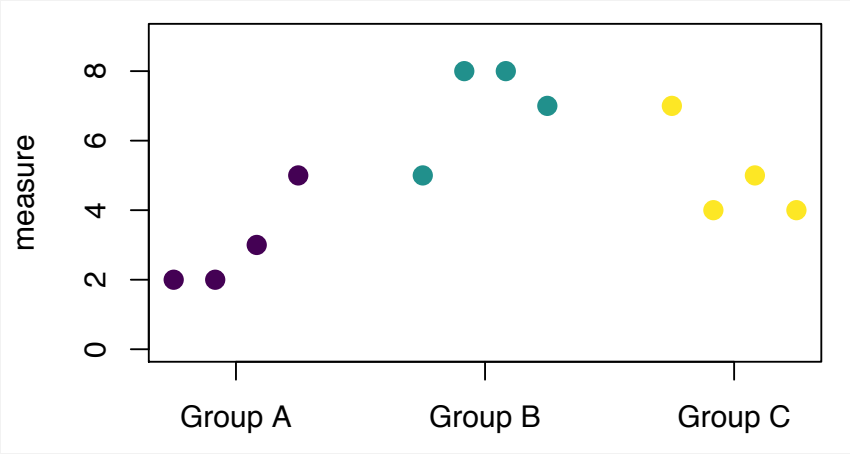
- Used to compare the means among more than two groups

- If you are comparing three groups, for instance, you cannot just do three pair-wise *t*-tests – this approach would cause too many false positives

- ANOVA takes into account the fact that you are comparing multiple groups and controls the false positive rate.

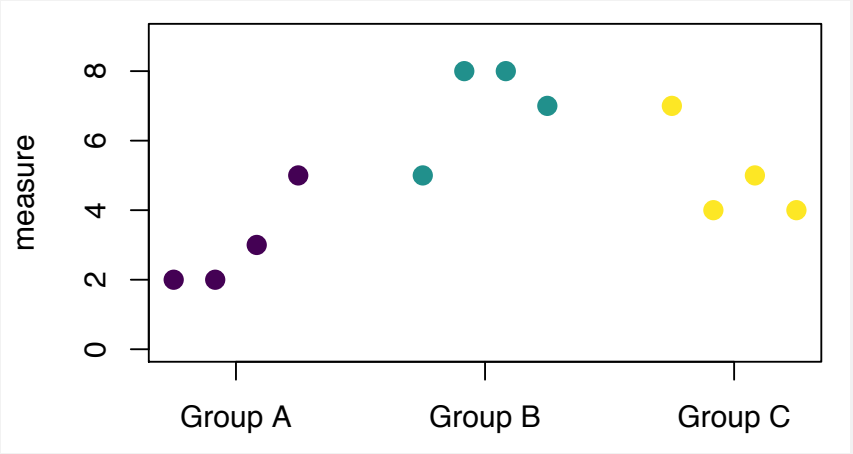
# Analysis of Variance

| A | B | C |
|---|---|---|
| 2 | 5 | 7 |
| 2 | 8 | 4 |
| 3 | 8 | 5 |
| 5 | 7 | 4 |



# Analysis of Variance

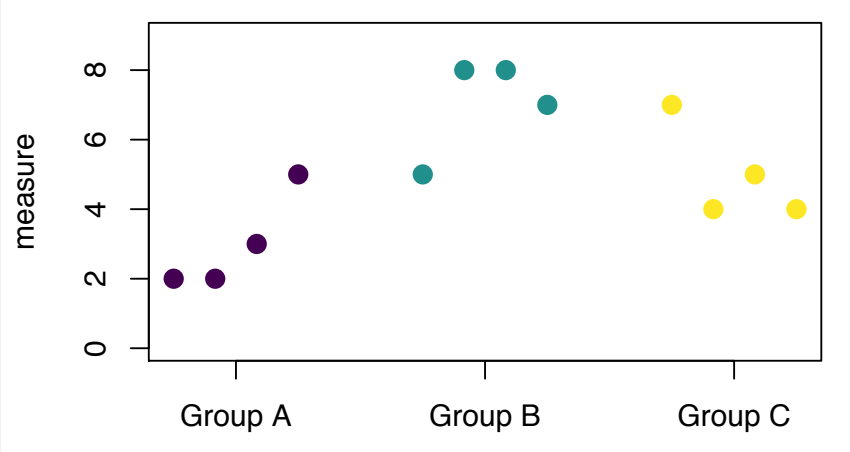
| A | B | C |
|---|---|---|
| 2 | 5 | 7 |
| 2 | 8 | 4 |
| 3 | 8 | 5 |
| 5 | 7 | 4 |




$$f\ statistic = \frac{\frac{\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

# Analysis of Variance

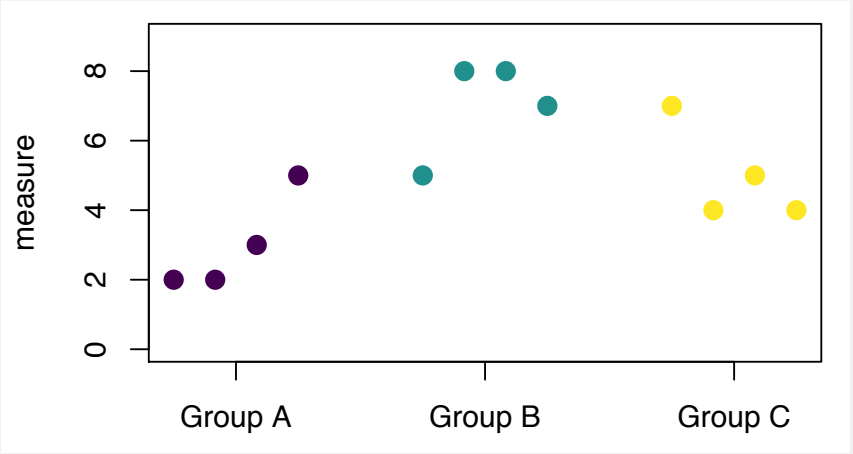
| A | B | C |
|---|---|---|
| 2 | 5 | 7 |
| 2 | 8 | 4 |
| 3 | 8 | 5 |
| 5 | 7 | 4 |

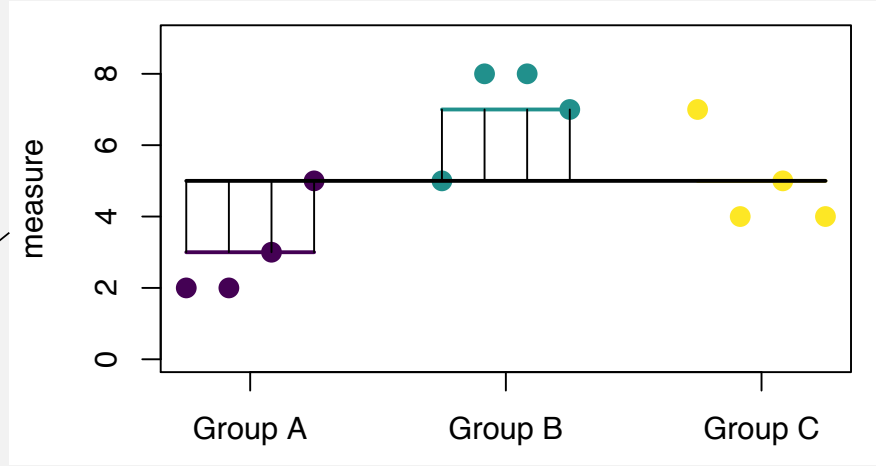



$$f\ statistic = \frac{\frac{\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

# Analysis of Variance

| A | B | C |
|---|---|---|
| 2 | 5 | 7 |
| 2 | 8 | 4 |
| 3 | 8 | 5 |
| 5 | 7 | 4 |

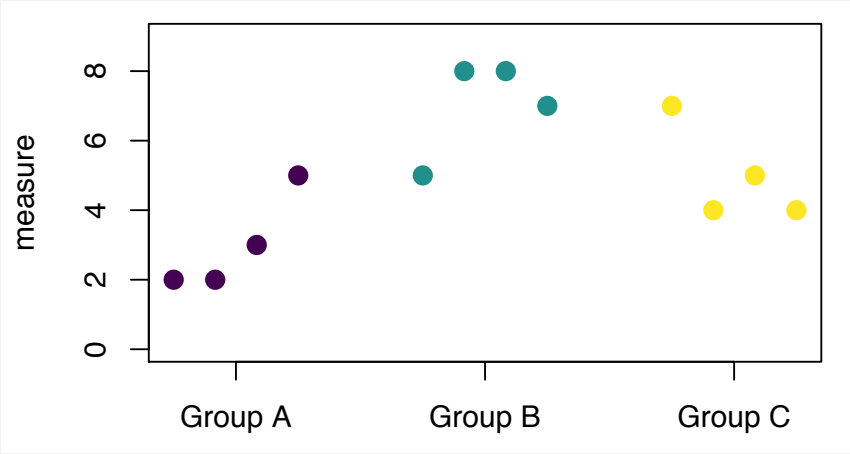


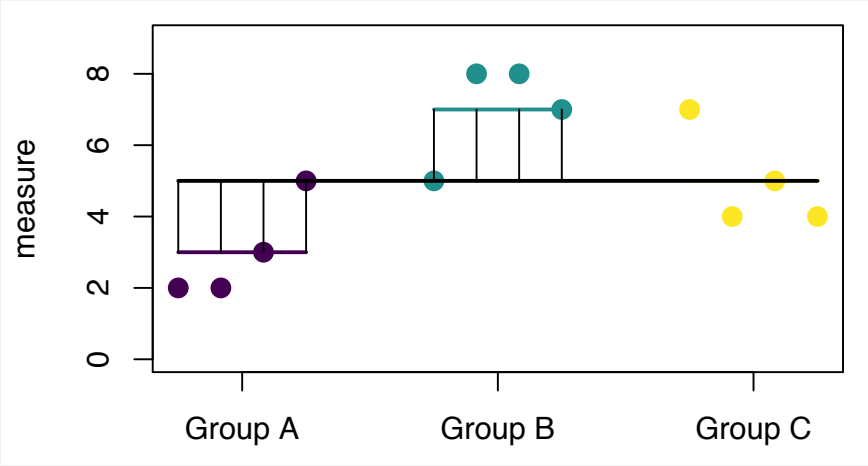




$$f\ statistic = \frac{\frac{\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

# Analysis of Variance

| A | B | C |
|---|---|---|
| 2 | 5 | 7 |
| 2 | 8 | 4 |
| 3 | 8 | 5 |
| 5 | 7 | 4 |

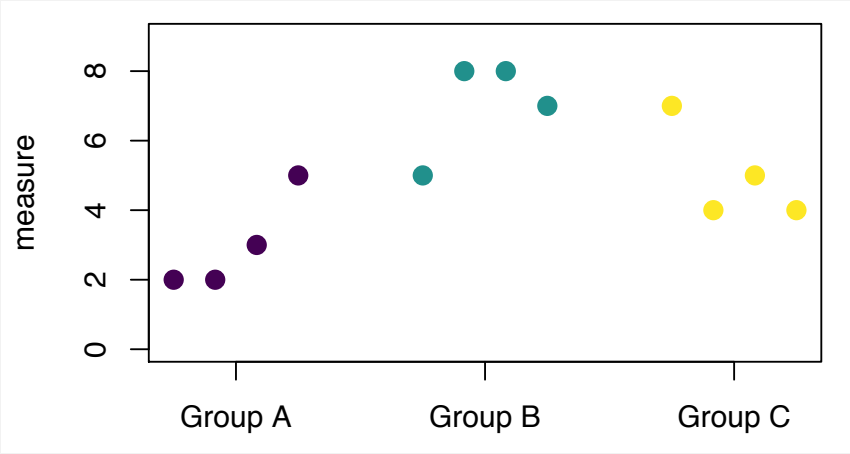


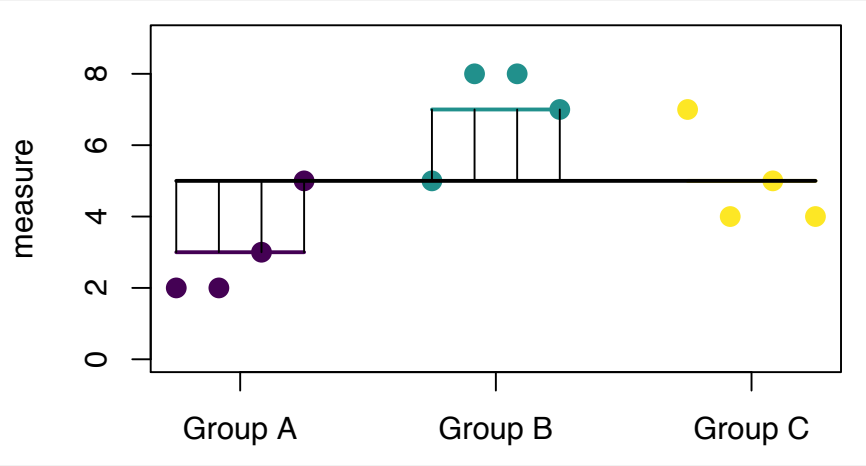




$$f\ statistic = \frac{\frac{\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

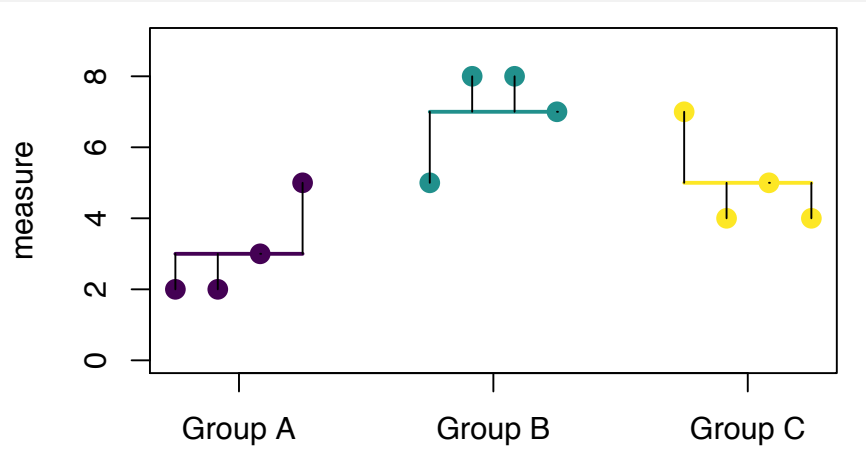
# Analysis of Variance

| A | B | C |
|---|---|---|
| 2 | 5 | 7 |
| 2 | 8 | 4 |
| 3 | 8 | 5 |
| 5 | 7 | 4 |



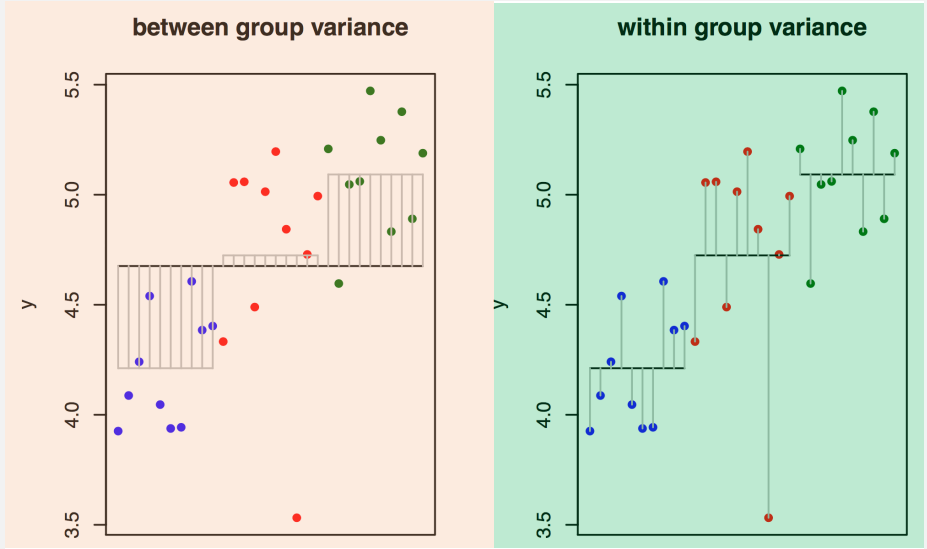




$$f\ statistic = \frac{\frac{\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

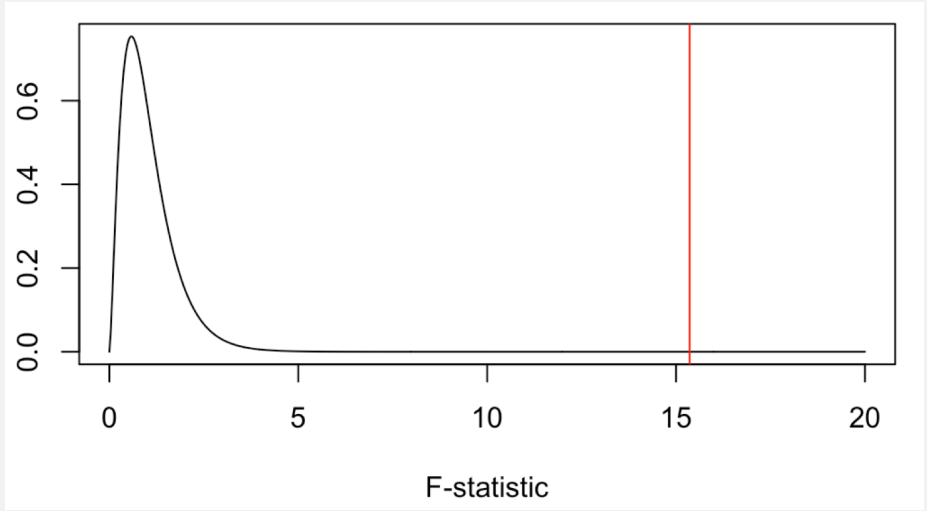


# Running ANOVA in R

```
> data("chickwts")
> feed.types <- lm(weight~feed, data=chickwts)
> anova(feed.types)
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
feed       5 231129   46226  15.365 5.936e-10 ***
Residuals 65 195556    3009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



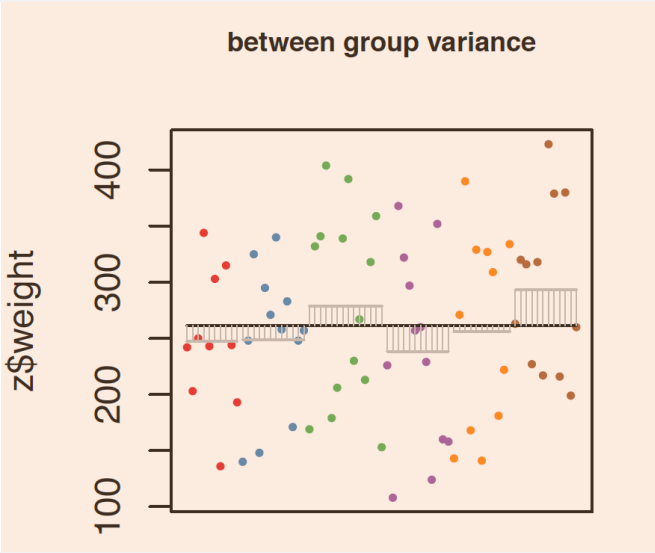

This significant result tells us that at least one of the groups of chickens have significantly different mean weights than other groups. (significant ANOVA result allows us to reject the null that they are all the same)

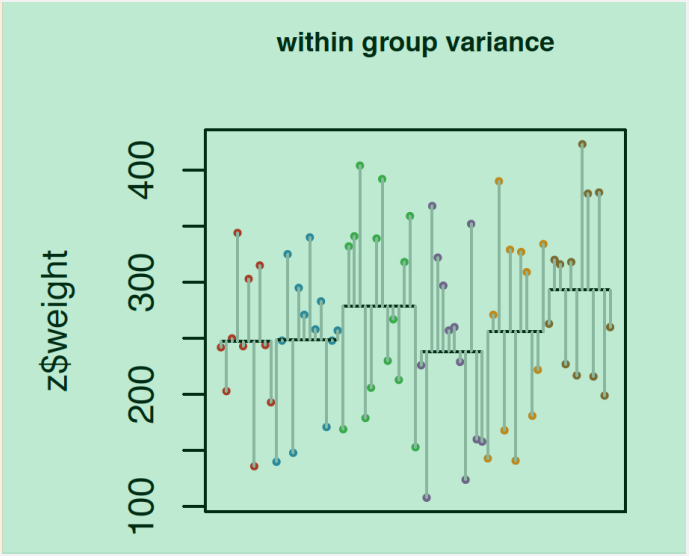


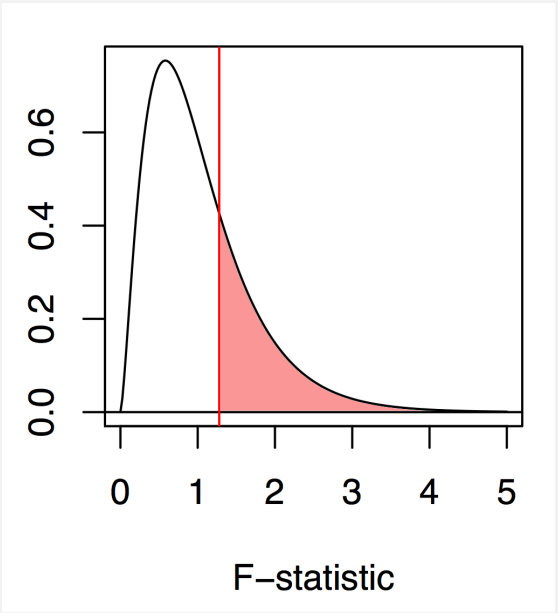
# Running ANOVA in R

```
z <- chickwts
z$weight <- sample(z$weight)
feed.types <- lm(weight~feed, data=z)
anova(feed.types)
```

```
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value Pr(>F)
feed       5  38225  7644.9  1.2792 0.2837
Residuals 65 388461  5976.3
```



# Post-hoc tests

If your ANOVA is significant, you may be interested in discovering which groups are different from one another

A variety of post-hoc comparisons of the means can be used

**Fisher's LSD**

- Least conservative test, basically uses *t*-tests to compare the means

**Scheffe's method**

- Performs all comparisons simultaneously, but has relatively low power

**Tukey-Kramer method**

- A pair-wise method, like a *t*-test, but corrected for multiple comparisons

# Post-hoc tests

```
> feed.anova <- anova(feed.types)
> feed.aov <- aov(feed.types)
> TukeyHSD(feed.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = feed.types)

$feed
                          diff        lwr       upr     p adj
horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

anova and aov functions will both perform an ANOVA but the results are stored slightly differently. For this posthoc test we want the aov format

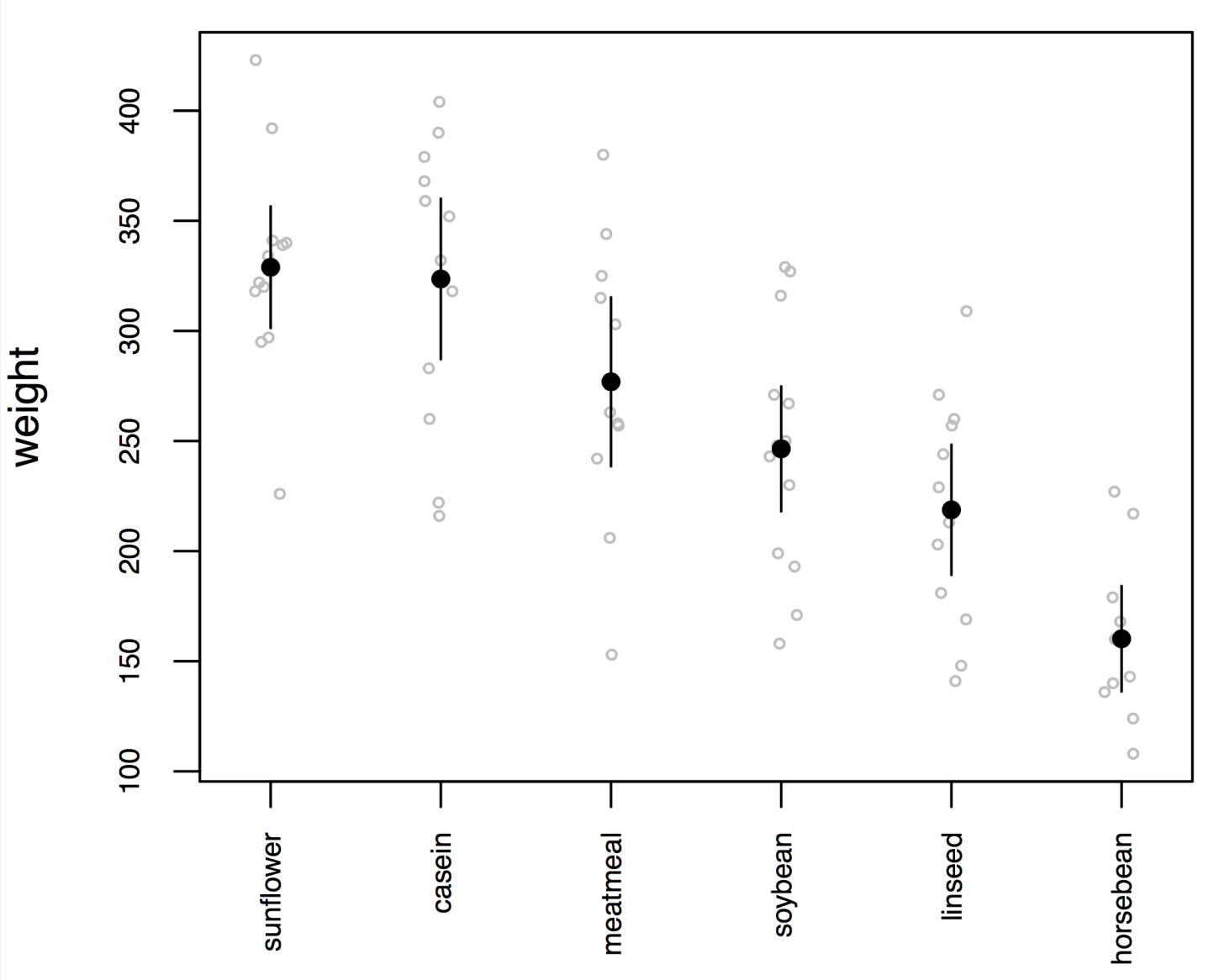
# Interpreting post-hoc tests

```
> feed.anova <- anova(feed.types)
> feed.aov <- aov(feed.types)
> TukeyHSD(feed.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = feed.types)

$feed
                          diff        lwr       upr     p adj
horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

When we examine all the significantly different ones we can draw several conclusions:

1) Chicks fed casein are significantly heavier than those fed horsebean, linseed, and soybean.
2) Chicks fed horsebean are significantly lighter than those fed meatmeal, soybean, and sunflower.
3) Chicks fed sunflower are significantly heavier than those fed linseed or soybean.

# Plotting this kind of data




Our results from the ANOVA and Tukey match up pretty well with our rules of thumb about 95% CI overlaps

# Example of code

```
#sets the order of the treatments
chickwts$feed <- factor(chickwts$feed,
                        levels=c("sunflower", "casein",
                                 "meatmeal","soybean",
                                 "linseed", "horsebean"))
stripchart(weight ~ feed, data=chickwts,
           method = "jitter", vertical = TRUE, cex.axis =.7,
           col = "gray", pch = 1, cex = .5, las = 3)
#Add error bars:
#First calculate means and SDs
meanShift <- tapply(chickwts$weight, chickwts$feed, mean)
sdevShift <- tapply(chickwts$weight, chickwts$feed, sd)
n <- tapply(chickwts$weight, chickwts$feed, length)
feed_table <- data.frame(mean = meanShift,
                         std.dev = sdevShift, n = n)
#Now add the SEM for each group:
seShift <- 1.96 * sdevShift / sqrt(n)
segments(1:6, meanShift - seShift,
         1:6, meanShift + seShift)
points(meanShift ~ c(1:6), pch = 16)
```

# Assumptions of the ANOVA

- The variable is normally distributed within each group
- The variance is the same in the different groups
- The design is balanced – you have the same sample size for each group
- But… ANOVA is fairly robust to violations of these assumptions

# A Non-Parametric Alternative

- Kruskal-Wallis Test
- Based on ranks
- The multiple-group version of the Mann-Whitney U-test

R-implementation:

```
> kruskal.test(weight ~ feed, data = chickwts)

        Kruskal-Wallis rank sum test

data:  weight by feed
Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

p-value suggests this test has lower power than ANOVA

# A Non-Parametric post-hoc

- Dunn's test – is the non-parametric equivalent of the Tukey

Not in base R need to install:
install.packages("dunn.test", dependencies=TRUE)

```
> dunn.test(chickwts$weight, g=chickwts$feed,
+           altp = T, method = "bonferroni")
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 37.3427, df = 5, p-value = 0
                           Comparison of x by group
                                 (Bonferroni)
Col Mean-|
Row Mean |     casein   horsebea    linseed   meatmeal    soybean
---------+-------------------------------------------------------
horsebea |   4.813069
         |    0.0000*
         |
 linseed |   3.308292  -1.658736
         |    0.0141*     1.0000
         |
meatmeal |   1.415755  -3.364059  -1.819817
         |     1.0000    0.0115*     1.0000
         |
 soybean |   2.499922  -2.602093  -0.933255   0.974144
         |     0.1863    0.1390     1.0000     1.0000
         |
sunflowe |  -0.182969  -4.987524  -3.491262  -1.594703  -2.689798
         |     1.0000    0.0000*    0.0072*     1.0000     0.1072

alpha = 0.05
Reject Ho if p <= alpha
```

# ANOVA Summary

- ANOVA is the foundation of essentially all tests comparing multiple means
- Don’t make it too complicated – the null hypothesis is simple: they are all the same.
- Post-hoc tests are important for determining which means are the source of a significant ANOVA.
- You can only justify a post-hoc test if the ANOVA is significant in the first place.
- Before applying ANOVA, check that your data fit the assumptions (consider transforming the data lots of times this will be based on your biological knowledge because you will have insufficient data to say much about the observed distribution)

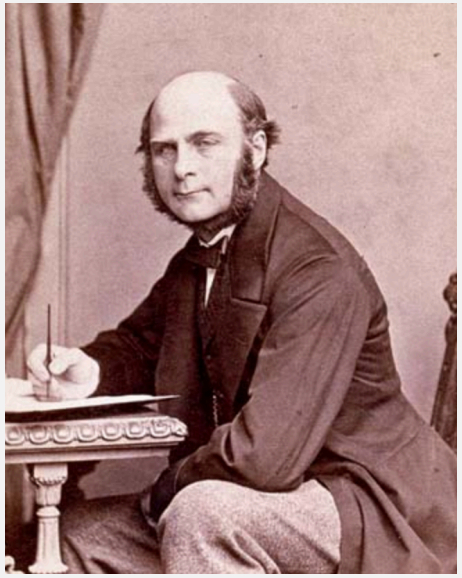
# Correlation and Causation

The correlation is a foundational concept in the biological sciences

Pearson's correlation coefficient:

$$r(X,Y) = \frac{cov(X,Y)}{s_x s_y}$$

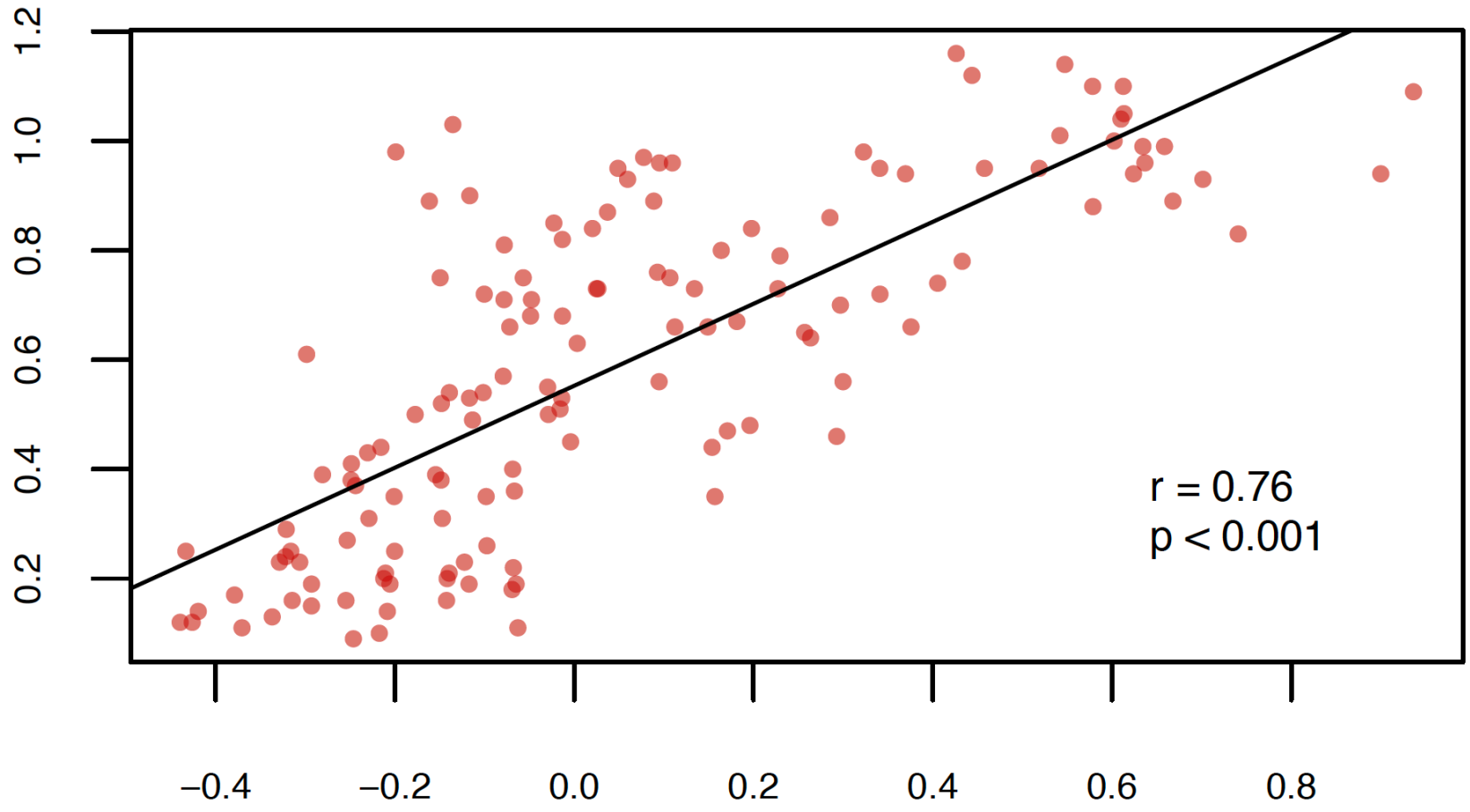
$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$



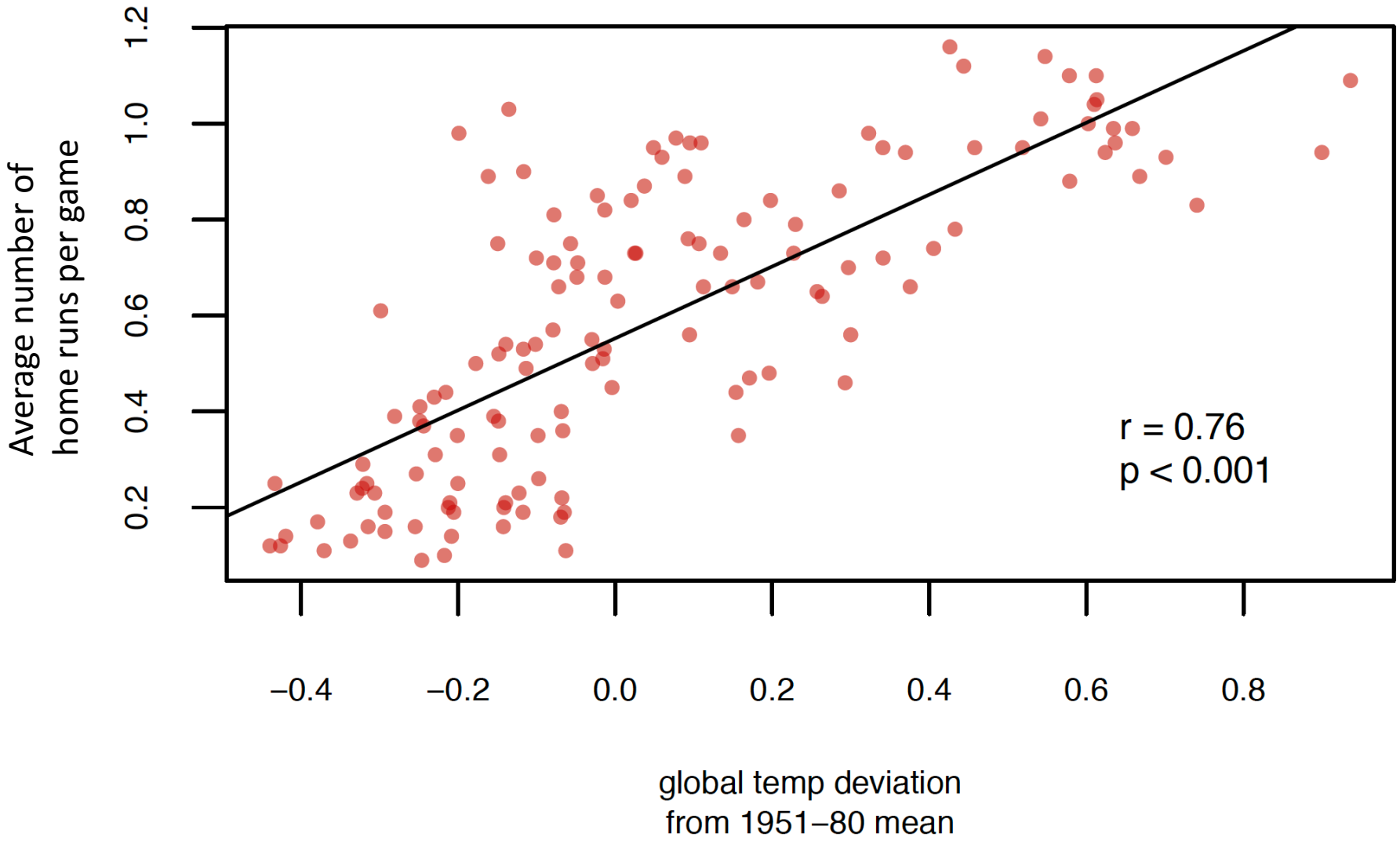
Francis Galton
1822 - 1911

Karl Pearson
1857 - 1936

# Covariance and Correlation



# Covariance and Correlation



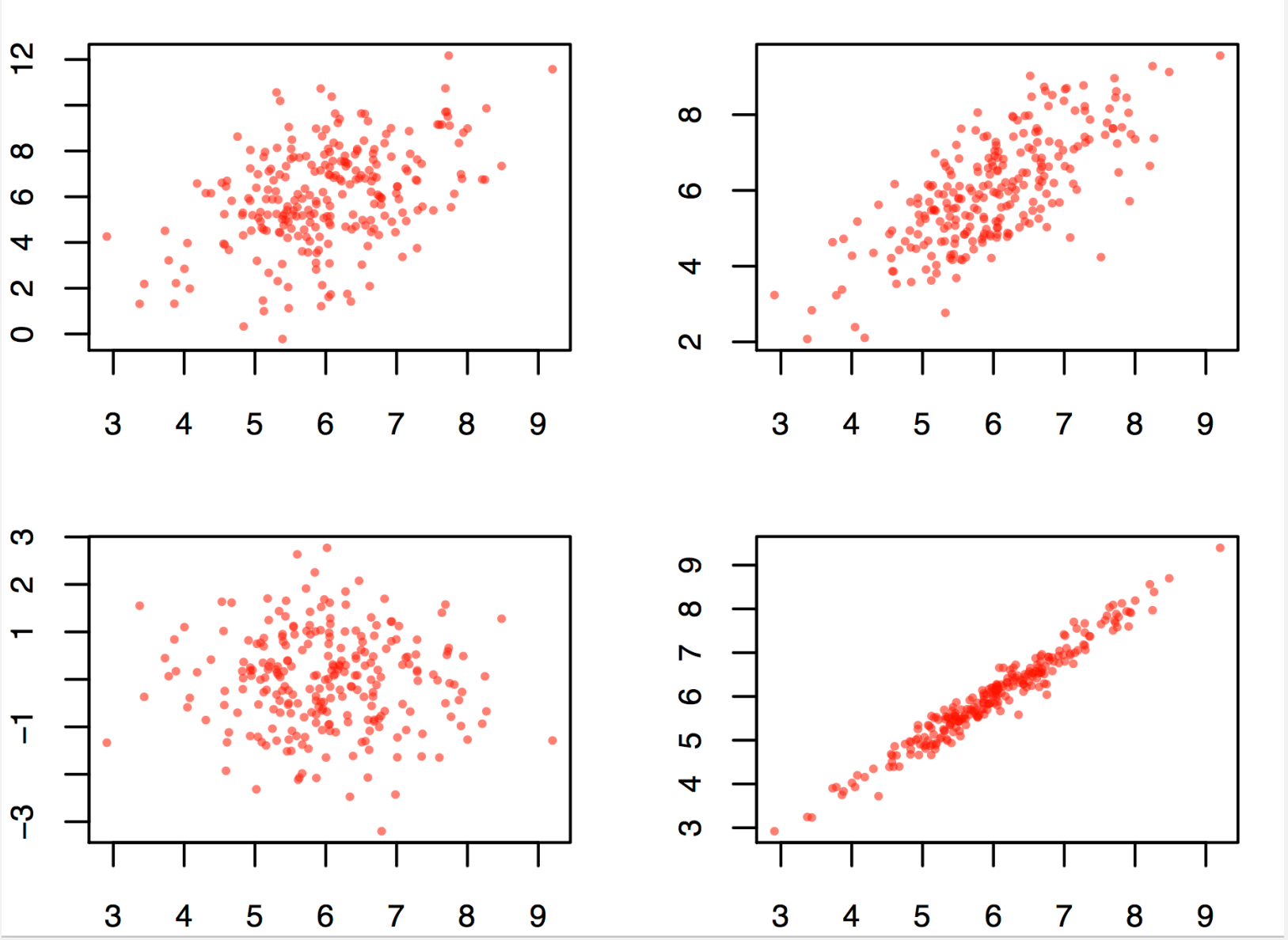
# Pearson's Correlation Coefficient

Assumes that the data are roughly from a bivariate normal distribution

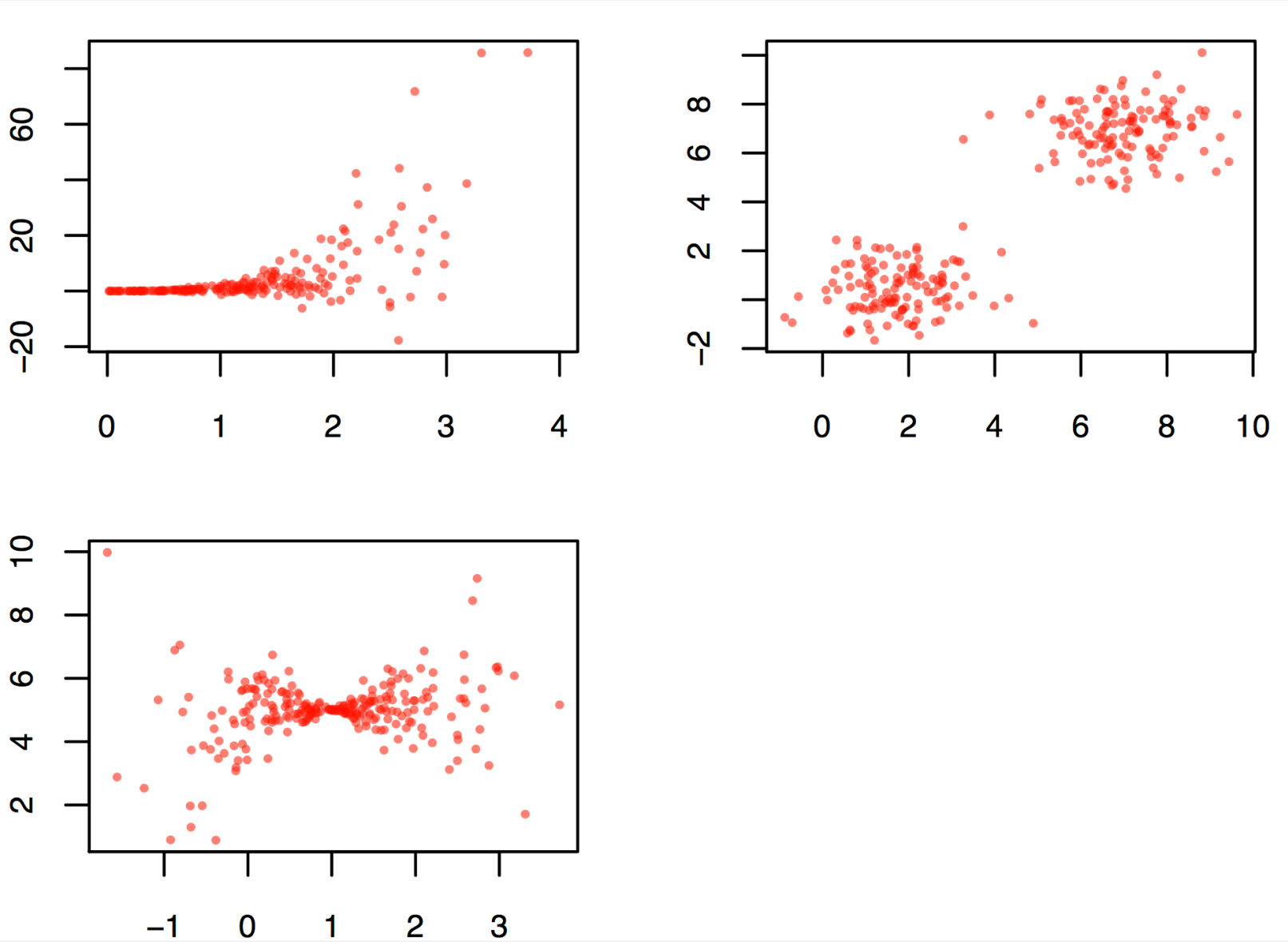
Would **not** be appropriate when:

- Relationship is non-linear (possible solutions?)
- Dataset contains extreme outliers (evaluate leverage)
- Data are in multiple distinct clouds of points (what could cause this?)

# Good data


0 2 4 6 8 12
3 4 5 6 7 8 9
2 4 6 8
3 4 5 6 7 8 9
−3 −1 1 2 3
3 4 5 6 7 8 9
3 4 5 6 7 8 9
3 4 5 6 7 8 9

# Bad data



# Spearman's Rank Correlation

An alternative to Pearson's correlation coefficient for data that depart substantially from bivariate normality

Based on ranks

```
> cor.test(x, y, method = "pearson")

        Pearson's product-moment correlation

data:  x and y
t = 3.8345, df = 18, p-value = 0.001214
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3241974 0.8583527
sample estimates:
      cor
0.6705223

> cor.test(x, y, method = "spearman")

        Spearman's rank correlation rho

data:  x and y
S = 684, p-value = 0.03148
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4857143
```

# Correlation Summary

- Correlations are very important
- We are interested in correlations between values of treatments and response variables
- We are interested in correlations between various factors that could affect the response variable
- We need to think about correlations between explanatory variables and potentially confounding variables
- **Many types of experiments call for the calculation and consideration of correlation coefficients**

# For Thursday

Read chapter WS 14-16

Bring laptop to class!

Heath Blackmon
BSBW 309
coleoguy@gmail.com
@coleoguy