STAT 139

Unit 8: Analysis of Variance (ANOVA)

Chapter 5, Sec. 13.1-13.2 in the Text

Unit 8 Outline

- Analysis of Variance (ANOVA)
  - General format and ANOVA's *F*-test
  - Assumptions for ANOVA *F*-test
  - Contrast testing
  - Other post-hoc tests
  - Two-way ANOVA (and Multi-way ANOVA)
- Kruskal-Wallis Test

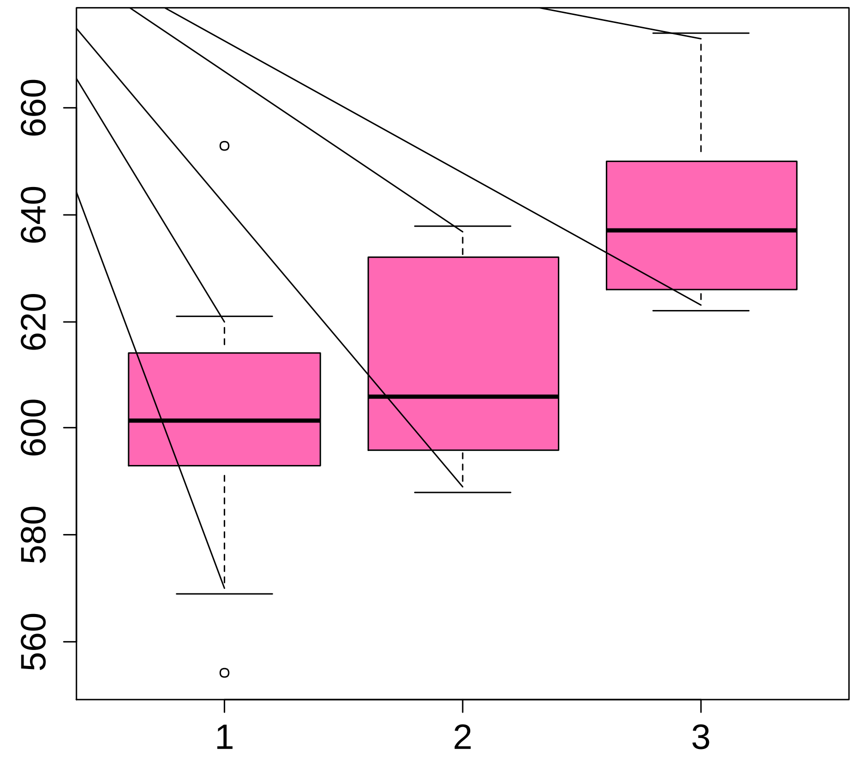
2

Example: Inference for 3+ Means – Bone Density

- Studies suggest a link between exercise and healthy bones
- A study of 30 rats examined the effect of *jumping* on the bone density of growing rats
- **Three** treatment groups
  - No jumping (10 rats - group 1)
  - 30 cm jump (10 rats - group 2)
  - 60 cm jump (10 rats - group 3)
- 10 jumps per day, 5 days per week for 8 weeks
- Bone density measured after 8 weeks
- Test to see if the jumping treatments affect bone density (measured in mg/cm^3)

Visualizing the *Jumping Rats* Example

- As always, first visualize the data:



Groups
1 – No jumping
2 – 30 cm jump
3 – 60 cm jump

Means & SD's

1	601	27.4
2	613	19.3
3	639	16.6

- We'd like to do a *t*-test, but there's no formula for 3 groups ☹
- Solution: Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA)

- ANOVA extends hypotheses that we have seen already
- With one population:

$$H_0\colon \mu = \mu_0$$

- With two populations:

$$H_0\colon \mu_1 = \mu_2$$

- Now consider inferences for I = 3 or more populations (technically, $I \geq 2$):

$$H_0\colon \mu_1 = \mu_2 = \ \ldots\ = \mu_I$$

- Analysis technique for this situation is ANOVA

Main concept of ANOVA

- With *I* populations (groups), there are two types of variability in the data:

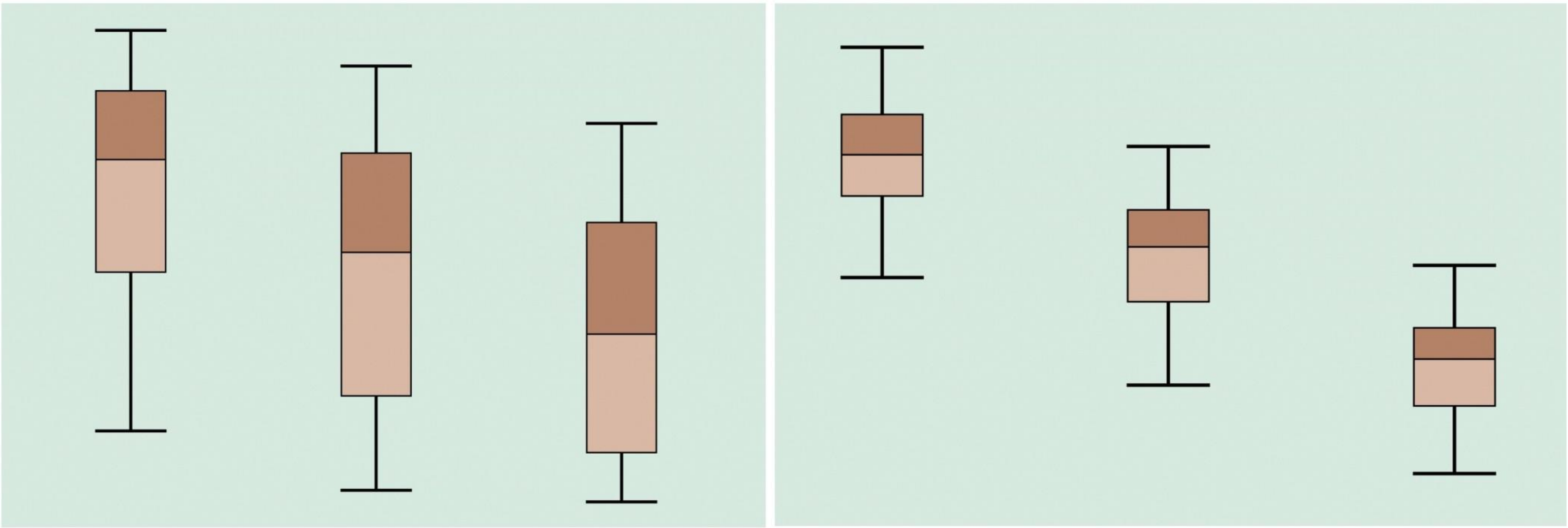
  (A) Variation of individual values around their group means
  (variability within groups)

  (B) Variation of group means around the overall mean
  (variability between groups)

- Main concept: If (A) is small relative to (B), this implies the group means are different.
- ANOVA determines whether variability in data is mainly from variation within groups or variation between groups. That is, does labelling the data into these groups lead to '*more than chance'* differences in the group means
- Think Boxplots split by groups…

6

Within group variability relatively large, adds noise, obscures group differences



Within group variability relatively small, group differences not obscured by noise

One-way ANOVA – the model

- The one-way ANOVA is used when independently for *I* groups:

$$Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2), \ldots, Y_{Ij} \sim N(\mu_I, \sigma^2)$$

- This can be written as:

  $Y_{ij} = \mu_i + \varepsilon_{ij}$

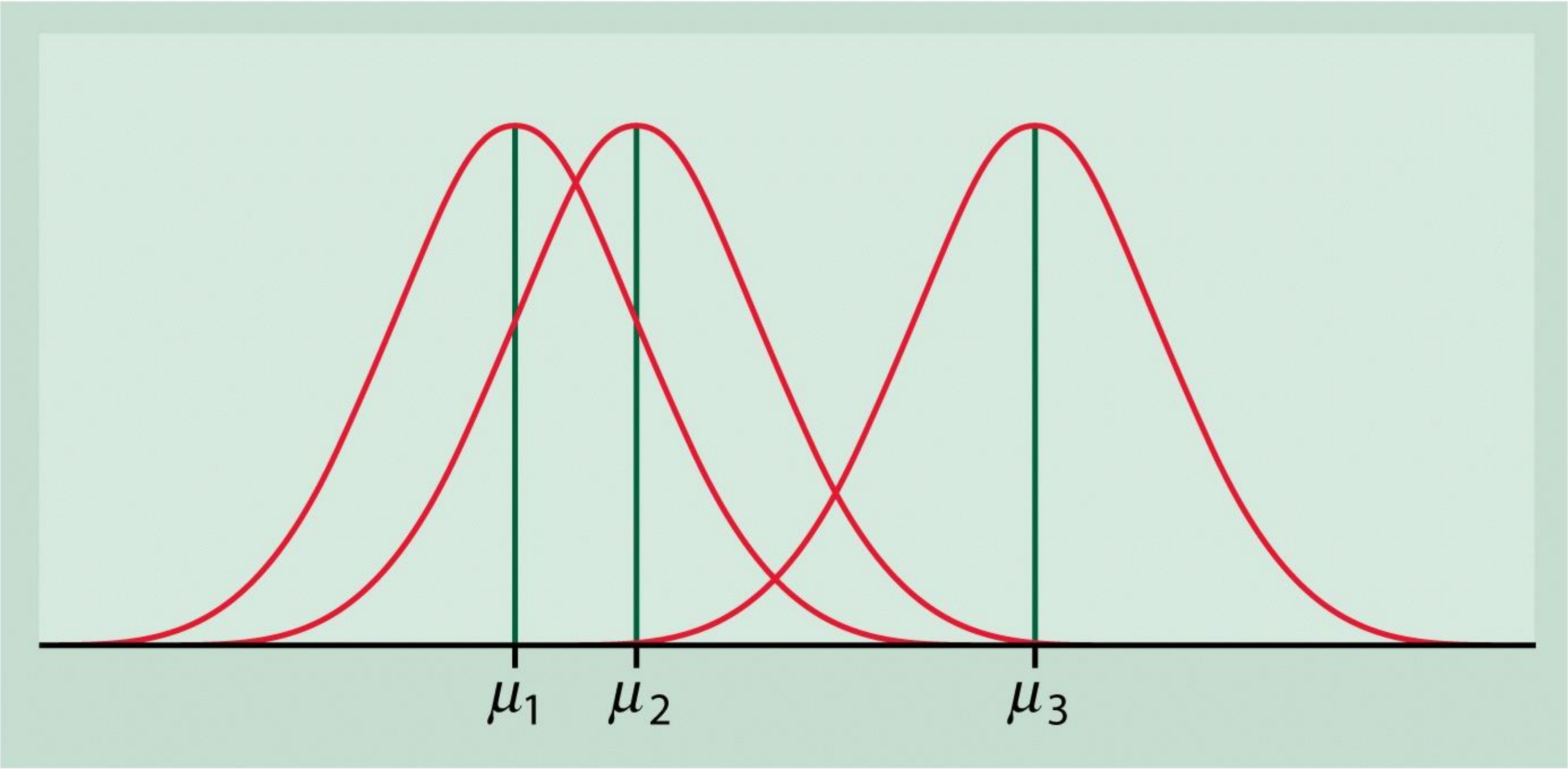
  for $i = 1, 2, \ldots, I$ groups (subpopulations) for a specific *factor*

  and $j = 1, 2, \ldots, n_i$ individuals sampled in each group

- The ε_{ij} are assumed to be $\sim N(0, \sigma^2)$
- Model parameters are $\mu_1, \mu_2, \ldots, \mu_I$ and σ^2 (one common variance!)
- Preview of regression models: can be written as (just like regression!)

  DATA = MODEL + RESIDUAL/ERROR

Model for one-way ANOVA with $I = 3$ groups



Errors/Residuals ε_{ij} are assumed to be ~ $N(0, \sigma^2)$

So only difference among the groups are differing means

9

One-way ANOVA – the data

- SRS's from each of the I populations (groups):

Sample size n_1 n_2 … n_I

Observations Y_{1j} Y_{2j} … Y_{Ij}

for $j = 1, 2, \ldots, n_i$

Sample means $\bar{Y}_1$ $\bar{Y}_2$ … $\bar{Y}_I$

Sample Variances S^2_1 S^2_2 … S^2_I

- Practical rule for examining Variances in ANOVA (checking assumption of a common σ^2). If largest S^2 is less than twice the smallest S^2, the ANOVA results will be approximately correct. Check to see if the ratio holds: $\boxed{\dfrac{S^2_{\text{largest}}}{S^2_{\text{smallest}}} \le 2}$

Variance within groups

- Since we assumed equal σ^2's for all I groups, all sample variances should be estimating the common σ^2.
- Thus we combine them in a pooled estimate (same idea as for the 2-sample pooled t-test).
- Pooled estimate of σ^2 (the variance within groups) is

$$S_p^2 = S_W^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \ldots + (n_I-1)S_I^2}{n-I} = \frac{\sum_{i=1}^{I}(n_i-1)S_i^2}{n-I}$$

$$= \frac{SSError}{dfError} = MSE$$

- The subscript W refers to this within groups estimate
- Also known as the "mean square within groups (MSW)" and "mean square error (MSE)"

Variance between groups

- If the null hypothesis, H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$, is true then examining individual sample means is as if we are sampling I times from the same population, with mean μ and variance σ^2.
- Recall from the sampling distribution of sample means: $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- So under H_0, σ^2 can be estimated by

$$S_B^2 = \frac{n_1(\bar{Y}_1 - \bar{Y})^2 + n_2(\bar{Y}_2 - \bar{Y})^2 + \ldots + n_I(\bar{Y}_I - \bar{Y})}{I-1} = \frac{\sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2}{I-1}$$

$$= \frac{SSGroups}{dfGroups} = MSG = MSB$$

- The subscript B (on S_B) refers to this between groups estimate
- Also known as the “mean square between groups (MSB or MSG)” or “mean square of the model” (MSM)
- Only a valid estimate of σ^2 if H_0: true, otherwise inflated

12

Concept behind the test

- If H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$ is true then S^2_W and S^2_B both estimate σ^2 and should be of similar magnitude
- If H_0 is not true, the between groups estimate of σ^2, S^2_B, will in general, be larger than the within groups estimate of σ^2, S^2_W
- Therefore a test of H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$ can be based on a comparison (ratio) of the between groups and within groups estimates of σ^2.
- Examine this ratio of variance estimates as an F test

$$F = \frac{S_B^2}{S_W^2} = \frac{MS\ between\ groups}{MS\ within\ groups}$$

$$= \frac{MSB}{MSW} = \frac{SSB / df_B}{SSW / df_W}$$

13

One-way ANOVA: F-test

- To test H_0: $\mu_1 = \mu_2 = \ldots = \mu_{\mathrm{I}}$
  H_{A}: not all μ_i are equal,
  use the F statistic

$$F = \frac{MSB}{MSW} = \frac{SSB / df_B}{SSW / df_W} = \frac{SSB/(I-1)}{SSW/(N-I)}$$

- This test statistic has an F distribution with I – 1 and N – I degrees of freedom when H_0 is true

- Reject H_0: for large values of F; if $F > F^*_{\alpha,\, I-1,\, N-I}$ or the corresponding p-value is less than α.

14

Analysis of variance (ANOVA)

- As we'll see for linear regression, for one-way ANOVA the total sum of squares is decomposed by:

$$\text{SST} = \text{SSB} + \text{SSW}$$

- The degrees of freedom are now partitioned

$$\begin{array}{ccccc} \text{DFT} & = & \text{DFB} & + & \text{DFW} \\ (n - 1) & = & (I - 1) & + & (n - I) \end{array}$$

- The mean squares (MS) are formed the same way in every partition:

$$MS = \frac{SS}{df} = \frac{\text{Sum of Squares}}{\text{degrees of freedom}}$$

One-way ANOVA table

Source	SS	DF	MS	F
Groups (Between)	SSB	$I - 1$	S_B^2 = SSB/DFB	F = MSB/MSW
Error (Within)	SSW	$n - I$	S_W^2 = SSW/DFW	
Total	SST	$n - 1$	SST/DFT	

- The F statistic tests if there is a difference among any of the I population means

- MSW is still our estimate of σ^2: the variance of the residuals or the average variance of the observations from their group means (also called the *pooled* variance estimate).

Solution:

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

H_A: at least one μ_i is different

```
> summarystats
           mean  n       sd
control   601.1 10 27.36360
lowjump   612.5 10 19.32902
highjump  638.7 10 16.59351
```

$$SSB = \sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2$$

$$= \left(10 \times (601.1 - 617.43)^2\right) + \left(10 \times (612.5 - 617.43)^2\right) + \left(10 \times (638.7 - 617.43)^2\right)$$

$$= 7433.867 \rightarrow MSB = SSB / df_B = 7433.867/(3-1) = 3716.9$$

$$SSW = \sum_{i=1}^{I} (n_i - 1)S_i^2 = (9 \times 27.364^2) + (9 \times 19.329^2) + (9 \times 16.594^2)$$

$$= 12579.84 \rightarrow MSW = SSW / df_W = 12579.84/(10+10+10-3) = 465.9$$

F = 3716.9 / 465.9 = **7.978**

F* = 3.35. Since our $F > F^*$, we reject H_0. The rat's have differing mean bone densities in the different treatment groups.

```
> qf(0.95,2,27)
[1] 3.354131
```

ANOVA Results from R

In R:

```
> model1 = aov(data$bone.density~data$group)
> anova(model1)
Analysis of Variance Table

Response: data$bone.density
           Df  Sum Sq Mean Sq F value   Pr(>F)
data$group  2  7433.9  3716.9  7.9778 0.001895 **
Residuals  27 12579.5   465.9
```

Unit 8 Outline

- Analysis of Variance (ANOVA)
  - General format and ANOVA's *F*-test
  - Assumptions for ANOVA
  - Contrast testing
  - Other post-hoc tests
  - Two-way ANOVA (and Multi-way ANOVA)
- Kruskal-Wallis Test

19

F-test assumptions

- Let's assume the data follow the assumptions. That is:

  1a) Independence between groups

  1b) Independence of observations within groups.

  2) Equal variances within group

  3) Each observation is Normally dist. around that group's mean:

  $$Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2), \ldots, Y_{Ij} \sim N(\mu_I, \sigma^2)$$

- Under H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$, so they are all equal to some common mean μ. And thus can be thought of as ALL individuals coming from just one Normal distribution.
- Then what are the distributions of the two sums of squares terms (SSB and SSW)?

$$SSB = \sum_{i=1}^{I} n_i (\bar{Y}_i - \bar{Y})^2 \sim \sigma^2 \chi^2_{I-1}$$

$$SSW = \sum_{i=1}^{I} (n_i - 1) S_i^2 \sim \sigma^2 \chi^2_{n_1 - 1} + \ldots + \sigma^2 \chi^2_{n_I - 1} = \sigma^2 \chi^2_{N-I}$$

F-test assumptions

- Also, these two r.v.s, SSB and SSW, are independent from one another.
- Recall from Unit 4, that we define an F r.v. as:

$$F = \frac{X / df_1}{Y / df_2} \sim F_{df_1, df_2}$$

- If $X \sim \chi^2_{df_1}$ and $Y \sim \chi^2_{df_2}$, independent of each other.
- Thus, if all of our assumptions are true, and the null hypothesis is also true, then the sampling distribution of the F-test statistic will be exactly F-distributed!
- What happens to this test statistic if our assumptions fail?

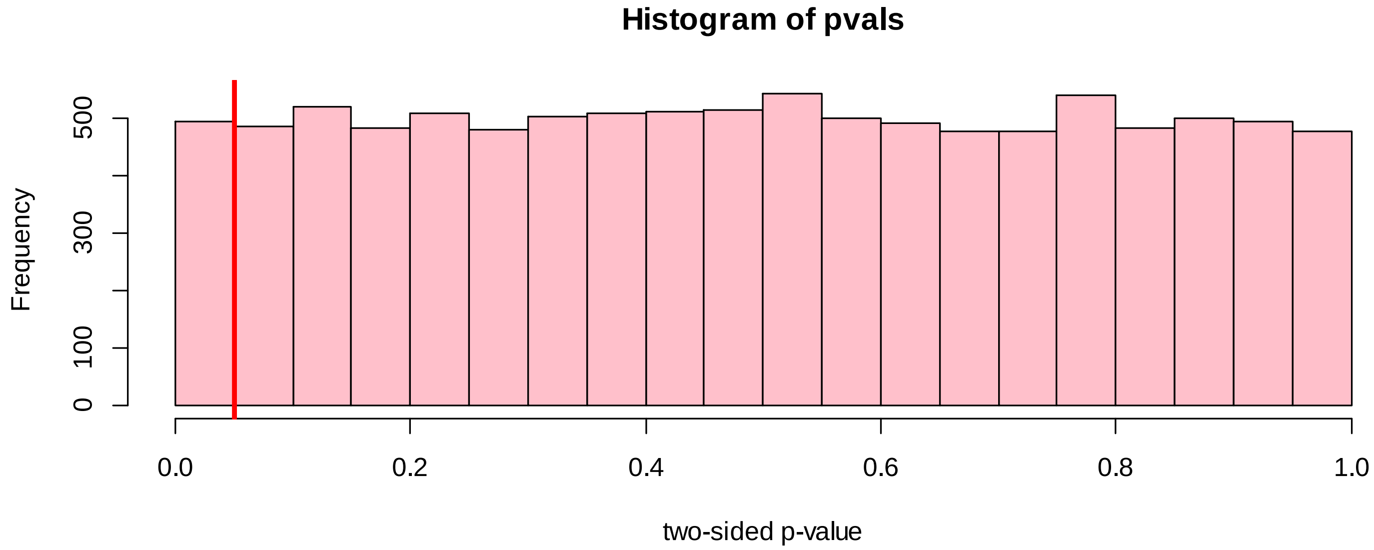
Assumptions all met

$Y_1 \sim N(0, 1)$
$n_{Y_1} = 10$

$Y_2 \sim N(0, 1)$
$n_{Y_2} = 10$

$Y_3 \sim N(0, 1)$
$n_{Y_3} = 10$

$$F = \frac{\left(\sum_{i=1}^{3} n_i (\bar{Y}_i - \bar{Y})^2\right) \Big/ (3-1)}{\left(\sum_{i=1}^{3} (n_i - 1) S_i^2\right) \Big/ (n_1 + n_2 + n_3 - 3)} \sim F_{3-1, n_1+n_2+n_3-3}$$



Independence Violations

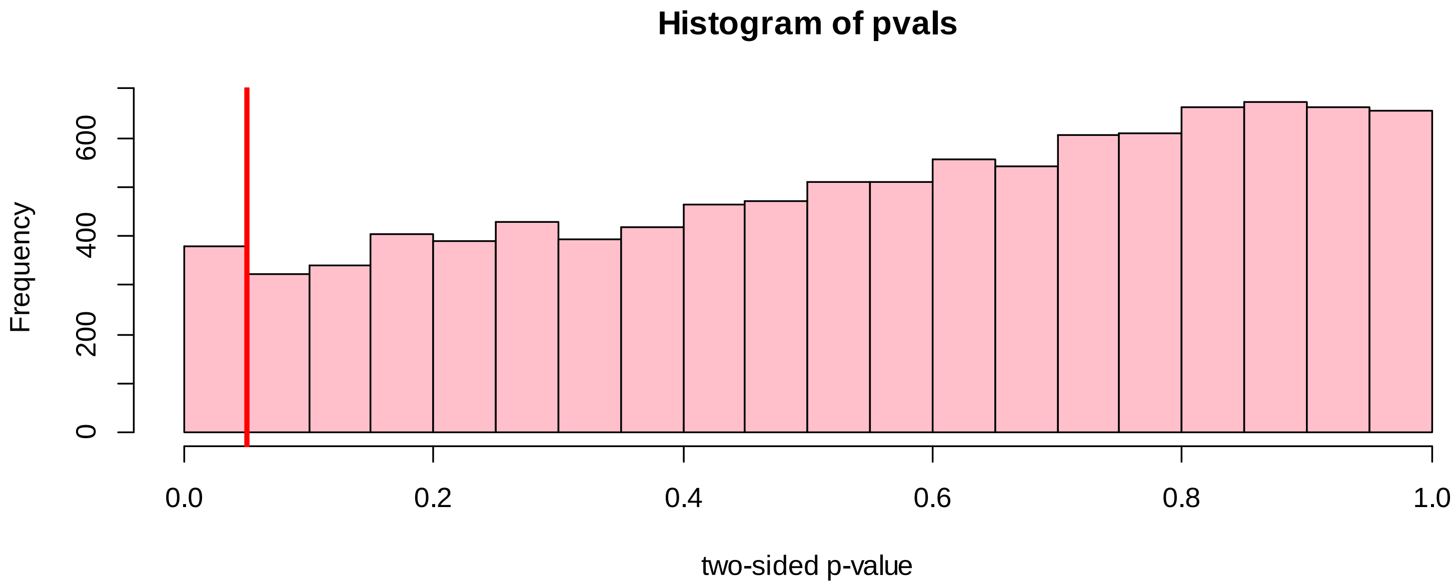
$Y_1 \sim N(0, 1)$
$n_{Y_1} = 10$

$Y_2 \sim N(0, 1)$
$n_{Y_2} = 10$

$Y_3 \sim N(0, 1)$
$n_{Y_3} = 10$

$$F = \frac{\left(\sum_{i=1}^{3} n_i(\bar{Y}_i - \bar{Y})^2\right) \Big/ (3-1)}{\left(\sum_{i=1}^{3} (n_i - 1)S_i^2\right) \Big/ (n_1 + n_2 + n_3 - 3)} \sim F_{3-1, n_1+n_2+n_3-3}$$

$\sigma_{Y_2,Y_3} = 0.577$



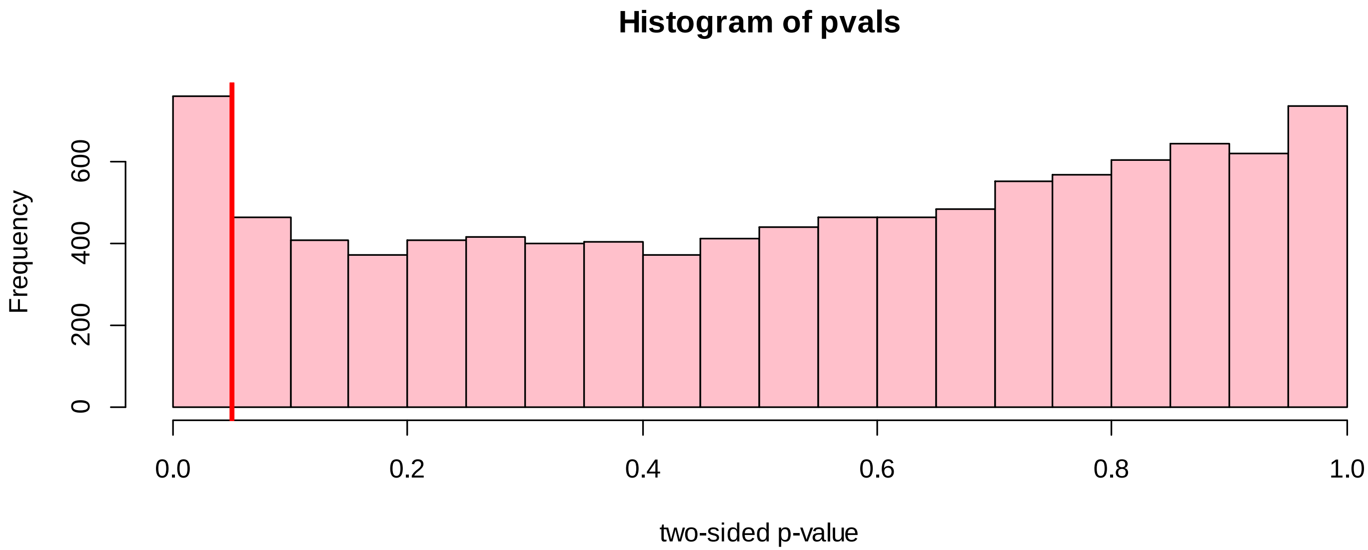
Constant Variance Violation #1

$Y_1 \sim N(0, 1)$
$n_{Y_1} = 10$

$Y_2 \sim N(0, 1)$
$n_{Y_2} = 10$

$Y_3 \sim N(0, 9)$
$n_{Y_3} = 10$

$$F = \frac{\left(\sum_{i=1}^{3} n_i(\bar{Y}_i - \bar{Y})^2\right) \Big/ (3-1)}{\left(\sum_{i=1}^{3} (n_i - 1)S_i^2\right) \Big/ (n_1 + n_2 + n_3 - 3)} \sim F_{3-1, n_1+n_2+n_3-3}$$



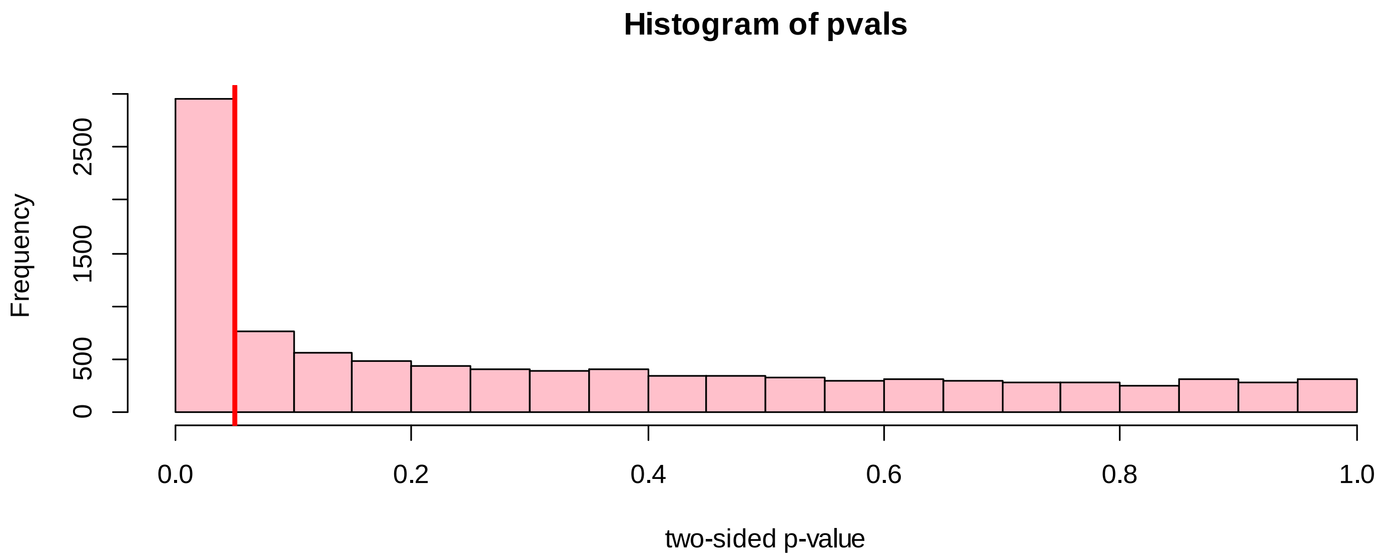
Constant Variance Violation #2

$Y_1 \sim N(0, 1)$
$n_{Y_1} = 50$

$Y_2 \sim N(0, 1)$
$n_{Y_2} = 50$

$Y_3 \sim N(0, 9)$
$n_{Y_3} = 10$

$$F = \frac{\left(\sum_{i=1}^{3} n_i(\bar{Y}_i - \bar{Y})^2\right) \Big/ (3-1)}{\left(\sum_{i=1}^{3} (n_i - 1)S_i^2\right) \Big/ (n_1 + n_2 + n_3 - 3)} \sim F_{3-1, n_1+n_2+n_3-3}$$



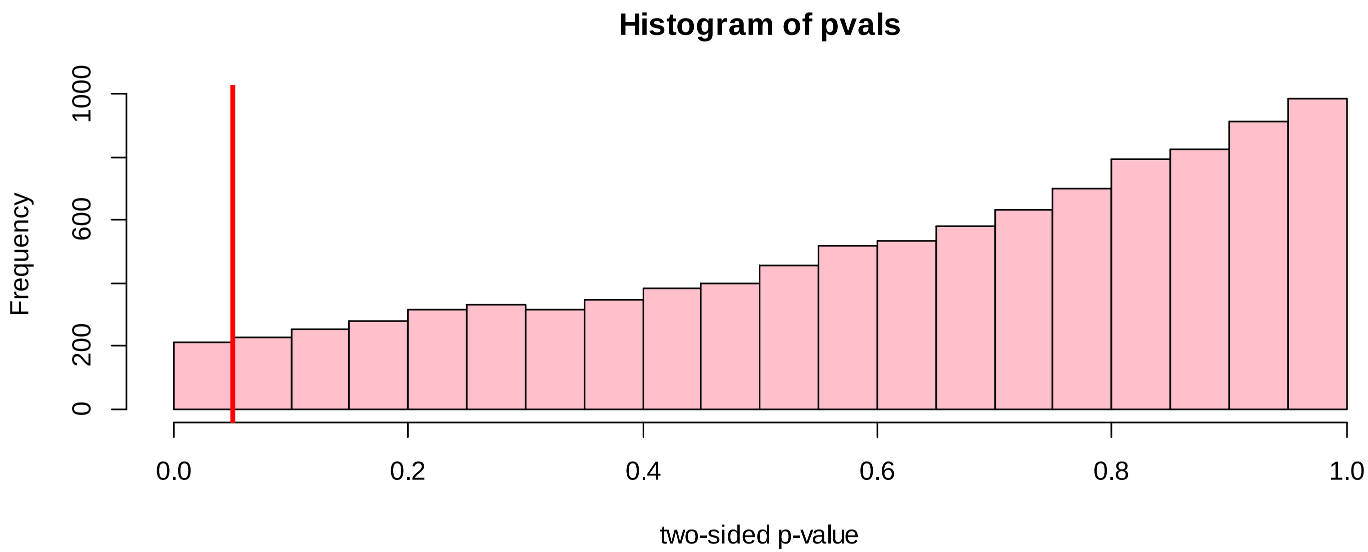
Constant Variance Violation #3

$Y_1 \sim N(0, 1)$
$n_{Y_1} = 50$

$Y_2 \sim N(0, 1)$
$n_{Y_2} = 50$

$Y_3 \sim N(0, 1/9)$
$n_{Y_3} = 10$

$$F = \frac{\left(\sum_{i=1}^{3} n_i(\bar{Y}_i - \bar{Y})^2\right) \Big/ (3-1)}{\left(\sum_{i=1}^{3} (n_i - 1)S_i^2\right) \Big/ (n_1 + n_2 + n_3 - 3)} \sim F_{3-1, n_1+n_2+n_3-3}$$



Normality Violation: All Skewed

$Y_1 \sim \text{Expo}(1) - 1$

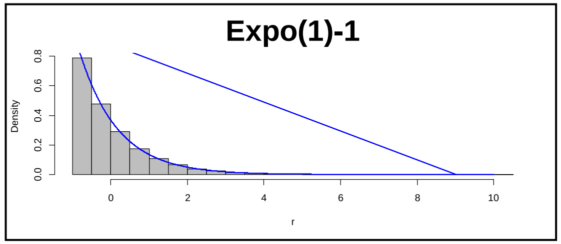
$n_{Y_1} = 10$

$Y_2 \sim \text{Expo}(1) - 1$

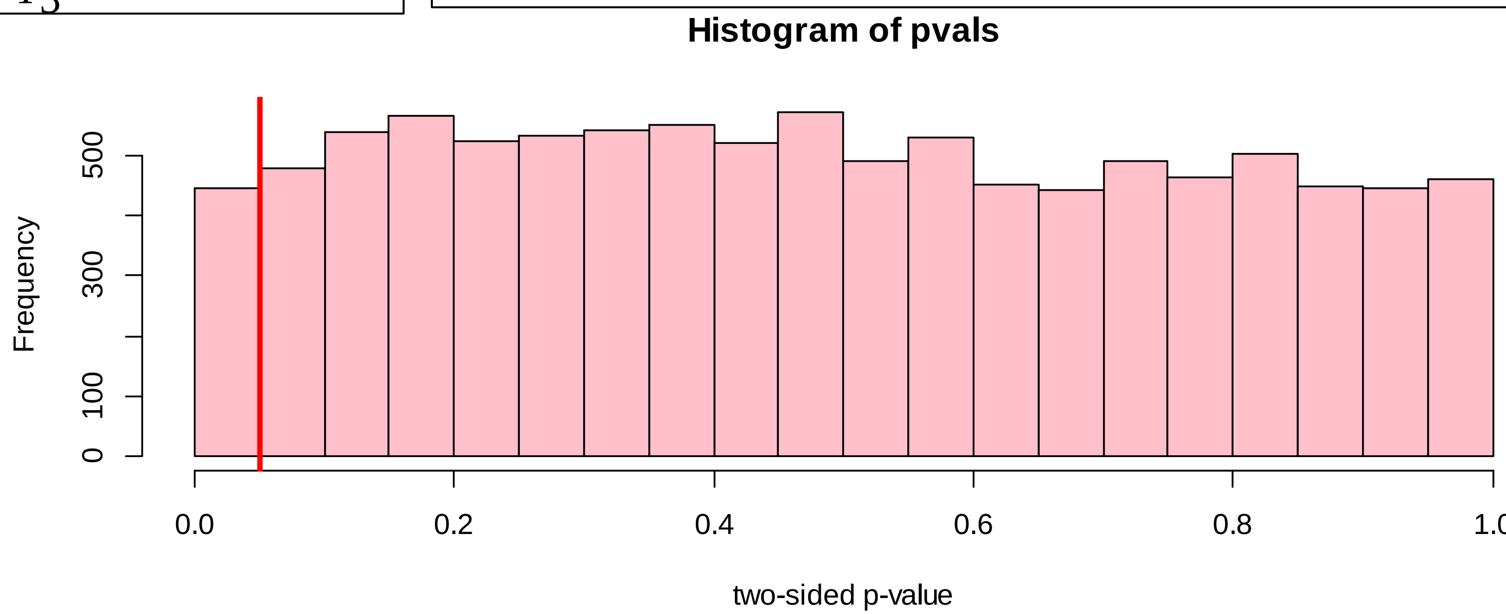
$n_{Y_2} = 10$

$Y_3 \sim \text{Expo}(1) - 1$

$n_{Y_3} = 10$



$$F = \frac{\left(\sum_{i=1}^{3} n_i (\bar{Y}_i - \bar{Y})^2\right) \Big/ (3-1)}{\left(\sum_{i=1}^{3} (n_i - 1) S_i^2\right) \Big/ (n_1 + n_2 + n_3 - 3)} \sim F_{3-1, n_1+n_2+n_3-3}$$



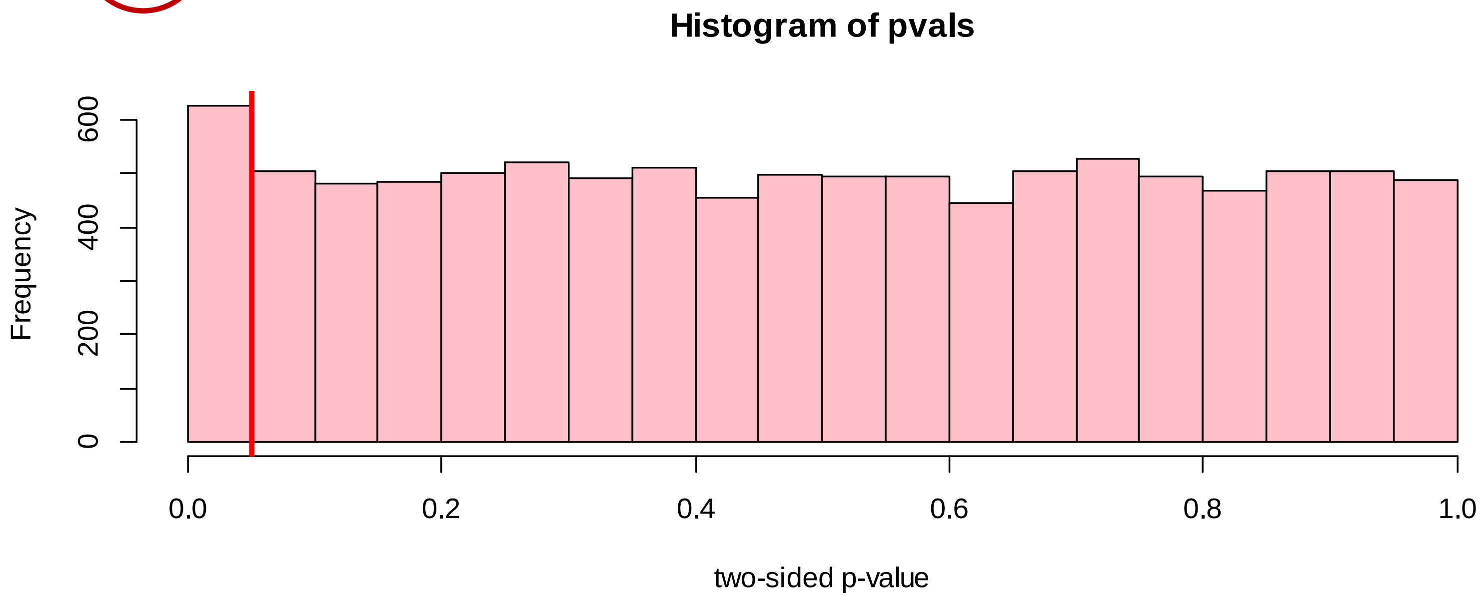
Normality Violation: Some Skewed

$Y_1 \sim N(0, 1)$
$n_{Y_1} = 10$

$Y_2 \sim N(0, 1)$
$n_{Y_2} = 10$

$Y_3 \sim \text{Expo}(1)-1$
$n_{Y_3} = 50$

$$F = \frac{\left(\sum_{i=1}^{3} n_i(\bar{Y}_i - \bar{Y})^2\right) \Big/ (3-1)}{\left(\sum_{i=1}^{3} (n_i - 1)S_i^2\right) \Big/ (n_1 + n_2 + n_3 - 3)} \sim F_{3-1, n_1+n_2+n_3-3}$$



Assumption Violation Summary

- The *F*-test in ANOVA has the following assumptions:
  1a) Independence between groups
  1b) Independence of observations within groups.
  2) Equal variances within group
  3) Each observation is Normally dist. around that group's mean:
- The *F*-test is robust (conservative at least) to positive correlation between groups (though, there may be a better approach) but not to positive correlation within groups (we did not show).
- The *F*-test is sensitive (not robust) to violations of constant variance, especially if sample sizes are different in the groups
- The *F*-test is somewhat sensitive to skewedness, especially if the groups have different skewedness and/or different sample sizes.
- What should we do if they all have the same skew?

Unit 8 Outline

- Analysis of Variance (ANOVA)
  - General format and ANOVA's *F*-test
  - Assumptions for ANOVA *F*-test
  - Contrast testing
  - Other post-hoc tests
  - Two-way ANOVA (and Multi-way ANOVA)
- Kruskal-Wallis Test

30

Steps in a Complete ANOVA Procedure

1. Examine the data, checking assumptions
2. If possible, formulate in advance some working (alternative) hypotheses about how the population group means might differ.
3. Check the evidence against the global null hypothesis of no differences among the groups by calculating the F-test
4. If the F-test is significant (that is, if it leads to a rejection of the null hypothesis of equal population group means):
   - Test the individual hypotheses specified at the second step
   - If there was not enough information to formulate working hypotheses, test all pairwise comparisons of means, adjusting for multiple comparisons (example also coming)

31

Contrasts

- After the omnibus F test has shown overall significance investigate other comparisons of groups using *contrasts* can be performed
- A contrast is a linear combination of μ_i's

$$\psi = \sum a_i(\mu_i) \quad \text{where} \quad \sum a_i = 0 \quad (\psi \text{ is the Greek "psi"}).$$

The corresponding contrast of sample means is $c = \sum a_i(\bar{Y}_i)$

- In this rats example, what might be an interesting comparison of the 3 groups involved (no jump, low jump, high jump)?
- For example, consider the bone density study.
- To compare control (group 1) versus treatment (groups 2 & 3) use:

$$\psi = \mu_1 - (\mu_2 + \mu_3)/2 = (1)\mu_1 + (-{}^1/_2)\mu_2 + (-{}^1/_2)\mu_3$$

- To compare levels of jumping (group 2 versus group 3) use

$$\psi = \mu_2 - \mu_3 = (0)\mu_1 + (1)\mu_2 + (-1)\mu_3$$

32

Contrasts – testing hypotheses

- To test the hypotheses:

  $H_0 : \psi = 0$ versus $H_A : \psi \neq 0$

  we use a *t*-test much like the pooled *t*-test:

$$T = \frac{\sum a_i \bar{Y}_i}{S_p \sqrt{\sum \frac{a_i^2}{n_i}}}$$

  This *t*-test has a *t* distribution with degrees of freedom associated with s_p under H_0 (test can be 1-sided or 2-sided)

Example – bone density

- Analysis of variance for bone density

S_p^2

```
Analysis of Variance Table
Response: data$bone.density
             Df  Sum Sq Mean Sq F value    Pr(>F)
data$group    2  7433.9  3716.9  7.9778 0.001895 **
Residuals    27 12579.5   465.9
```

- Reject H_0: $\mu_1 = \mu_2 = \mu_3$ if our calculated $F > F^*_{0.05,\mathrm{df}=2,27} = 3.35$
- Omnibus test: $p = 0.002$, reject H_0 and conclude there is a difference in bone density among the three groups.
- Where does that difference lie? Let's look at the comparisons (via contrasts):
  - (Group 1) versus (Groups 2 and 3)
  - (Group 2) versus (Group 3)

34

Example – bone density

- Does jumping (any level) affect bone density?
- Set-up hypotheses:

$$H_0 : \mu_{Control} - 0.5(\mu_{lowjump} + \mu_{highjump}) = 0$$

$$H_A : \mu_{Control} - 0.5(\mu_{lowjump} + \mu_{highjump}) \neq 0$$

- Calculate *t*-test

$$T = \frac{\sum a_i \bar{Y}_i}{S_p \sqrt{\sum \frac{a_i^2}{n_i}}} = \frac{(1)601.1 + (-0.5)612.5 + (-0.5)638.7}{\sqrt{465.9}\sqrt{\frac{1^2}{10} + \frac{(-0.5)^2}{10} + \frac{(-0.5)^2}{10}}}$$

$$= \frac{-24.5}{8.36} = -2.93$$

- Calculate the *p*-value: $df = N - I = 30 - 3 = 27$.
  p-value = $2*P(t<-2.93) < 2(0.005) = 0.01$.

- So reject H_0 and conclude any kind of jumping improves bone density over no jumping at all.

35

Example – bone density

Does the level of jumping affect bone density?

Look at group 2 versus group 3 using the contrast

$$H_0 : (0)\mu_{Control} + (1)\mu_{lowjump} + (-1)\mu_{highjump} = 0$$

$$H_A : (0)\mu_{Control} + (1)\mu_{lowjump} + (-1)\mu_{highjump} \neq 0$$

$$T = \frac{\sum a_i \overline{Y}_i}{S_p \sqrt{\sum \frac{a_i^2}{n_i}}} = \frac{(0)601.1 + (1)612.5 + (-1)638.7}{\sqrt{465.9}\sqrt{\frac{0^2}{10} + \frac{(1)^2}{10} + \frac{(-1)^2}{10}}}$$

$$= \frac{-26.2}{9.653} = -2.714$$

So reject H_0 (since our p-value $< 0.02 < \alpha$) and conclude higher jumping affects bone density differently than moderate jumping (in fact, it makes bones denser).

36

Unit 8 Outline

- Analysis of Variance (ANOVA)
  - General format and ANOVA's *F*-test
  - Assumptions for ANOVA *F*-test
  - Contrast testing
  - Other post-hoc tests
  - Two-way ANOVA (and Multi-way ANOVA)
- Kruskal-Wallis Test

37

Recall: Multiple Comparisons

- If we perform multiple contrast tests or pairwise *t*-tests after getting a significant F-test, then Type I can be inflated due to ***multiple comparisons***: we have looked at multiple tests at once (each with $\alpha = 0.05$), and thus it will likely lead to significant results that are not truly there (simply by chance alone).
- What conservative correction approach have we seen already?
- The Bonferroni correction says to protect the overall level of α we must perform each individual test or confidence interval at level:

  $\alpha^* = \alpha$/(# tests being performed).
- Why this choice? Why is it conservative?

Compounded Uncertainty

- For N simultaneous tests that are independent, the overall type-I error (i.e., the probability of *at least one* type-I error among N tests) is $\boxed{1-(1-\alpha)^N}$.

- For N simultaneous tests that are perfectly positively dependent, the overall type-I error is just $\boxed{\alpha}$.

- For N simultaneous tests that have unknown dependence, the maximum overall type-I error is $\boxed{\min(N\alpha, 1)}$.

$$\alpha \leq 1-(1-\alpha)^N \leq N\alpha$$

Simultaneous Inferences for Confidence Intervals

- Individual confidence level is a probability that a *single* confidence interval covers the true value.

- Overall (Familywise) confidence level is a probability that *all* confidence intervals cover the corresponding true values.

- Analogously, if the success rate of a (1-α)100% confidence interval is (1-α), the simultaneous success rate of several (1-α)100% confidence intervals is *less than* (1-α).

Methods for Multiple Comparison

- Differ by types of a multipliers for CIs or modifications to reference distribution.

$$Estimate \pm \boxed{(Multiplier)}SE$$

Interval half-width;
aka, *margin of error*

- Bonferroni;
- Tukey's HSD (or, in general, Tukey-Kramer procedure);
- Fisher's (protected) least significant difference (LSD);
- Scheffe's procedure.

Bonferroni

- Most popular and the simplest (but, also, most conservative).

$$\alpha \le 1-(1-\alpha)^N \le N\alpha$$

- Main Idea: Use the upper bound for the probability that at least one test falsely rejects under the null.

- Set individual significance levels for each of N tests at α/N. Then, the overall level will be α.

- For pairwise mean comparisons:

Margin of Error for $(1-\alpha)100\%$ CI: $t_{n-I,\,1-\alpha/(2N)} \cdot SE$

Why conservative?

Tukey HSD (Honest Significant Difference)

- Main Idea: Consider the largest difference between any two sample means for I groups.
- Appropriate when interested in differences between all pairs of group means.
- If we assume normality, equal variances, equal sample sizes ($\bar{n}$), under H_0 (equal means):

$$Q = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{S_p / \sqrt{\bar{n}}} \sim \boxed{q(1-\alpha, I, n-I)}, \text{ where } n = \bar{n} I$$

called *Studentized Range Distribution*
Use `qtukey()` in R to obtain quantiles.

- Margin of error for CI:

$$\left(\frac{q(1-\alpha, I, n-I)}{\sqrt{2}} \right) SE \text{, where } SE = S_p \sqrt{\frac{1}{\bar{n}} + \frac{1}{\bar{n}}}$$

Tukey-Kramer Procedure

- Extension to unequal sample sizes.
- Margin of error for CI:

$$\left(\frac{q(1-\alpha, I, n-I)}{\sqrt{2}}\right) SE \text{ , where } SE = S_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

and $n = n_1 + n_2 + \ldots + n_I$.

- Best illustrated with an example in R:

```
model <- aov(…)
TukeyHSD(model)
> qtukey(0.95,2,1000)/sqrt(2)
[1] 1.959964
```

Unit 8 Outline

- Analysis of Variance (ANOVA)
  - General format and ANOVA's *F*-test
  - Assumptions for ANOVA *F*-test
  - Contrast testing
  - Other post-hoc tests
  - Two-way ANOVA (and Multi-way ANOVA)
- Kruskal-Wallis Test

45

Extension to Two-way ANOVA (and beyond!)

- We can expand ANOVA to include an additional factor (a second *grouping* variable) – call it two-way ANOVA
  - For example, we could predict text messaging based on class year (with I = 4 groups) and sex (with I = 2 groups), and even a third variable like house (I = 12?) or a fourth like concentration (with I = ???).
- Algebraically, things get a little more complicated, but we won't focus on that in this course
- Having multiple grouping variables (just like predictor variables) also allows us to include interaction terms as well
- Easy to perform in R!!! Math is not the easiest to go through: decompositions of the Total Sums of Squares
- They are connected to multiple regression, which we explore further.

Two-way ANOVA in R

- The example we will use is trying to predict the *logtexts* messages by Harvard College students based on their *class year* and *sex*.
- In R:

Oneway ANOVA of *logtexts* by *female*

```
model1=aov(logtexts~female)
```

Oneway ANOVA of *logtexts* by *classyear*

```
model2=aov(logtexts~classyear)
```

Twoway ANOVA of *logtexts* by *classyear* and by *female*

```
model3=aov(logtexts~classyear+female)
```

Twoway ANOVA with ***interaction*** of *logtexts* by *classyear* and by *female*

```
model4=aov(logtexts~classyear*female)
```

- What in the world is an ***interactive effect***?

Separate Oneway ANOVAs in R

- Here are the results in R for this sample of n = 169 students for the two separate oneway ANOVAs.
- What should be the 2 different sets of degrees of freedoms here?

```
> summary(model1)
              Df Sum Sq Mean Sq F value Pr(>F)
female         1   0.01  0.0137   0.012  0.913
Residuals    167 190.84  1.1427

> summary(model2)
              Df Sum Sq Mean Sq F value  Pr(>F)
classyear      3  15.42   5.139   4.833 0.00299 **
Residuals    165 175.43   1.063
```

- What do you notice?

48

Two-way ANOVA in R
(without interaction)

- Here are the results in R for this sample of n = 169 students for the twoway ANOVA model without interaction.
- What should be the 2 different sets of degrees of freedoms here?

```
> summary(model3)
              Df Sum Sq Mean Sq F value  Pr(>F)
classyear      3  15.42   5.139   4.822 0.00304 **
female         1   0.66   0.661   0.620 0.43212
Residuals    164 174.77   1.066
```

- What do you notice? How does this compare to the separate model?
- Why are the sums of squares for the *female* variable different here than in the separate oneway ANOVAs?

49

Two-way ANOVA in R
(with interaction)

- We can also consider the interaction between the two grouping variables here (sex and class year)
- What would the interaction term represent?
  - The effect of class year on # text messages sent may be different for men than women
  - Equivalent to saying that the effect of women on # text messages sent may be different in the 4 class years
    - Maybe men's average decreases over the 4 years, but women's average stays the same/increases.
- What should be the degrees of freedom for the interaction between these two variables?

$(I_1\text{-}1)*(I_2\text{-}1) = (4\text{-}1)*(2\text{-}1) = 3$

50

Two-way ANOVA in R
(with interaction)

- Here are the results in R for this sample of n = 169 students for the twoway ANOVA including interaction.
- What should be the 3 different sets of degrees of freedoms here?

Note: *df* may not be "full" if there is a missing group:

```
> summary(model4)
                   Df Sum Sq Mean Sq F value  Pr(>F)
classyear           3  15.42   5.139   4.806 0.00311 **
female              1   0.66   0.661   0.618 0.43288
classyear:female    3   2.65   0.882   0.825 0.48163
Residuals         161 172.13   1.069
```

- What do you notice? Is the interaction variable important?
- This is called a *full factorial* model since all interactions are included.
- This is just a preview: we will get into this more deeply once we get to regression modeling.

ANOVA: Main Points

- Simple idea behind ANOVA: significant difference among 2 or more groups if variability between groups (differences among the group means) is significantly larger than variability within groups (differences from mean within each group) [*F*-test].

- If there is evidence of difference among groups, then an *a priori* hypothesis can be tested via a contrast *t*-test, or some care must be taken in searching for the group pairs that are significantly different (using Bonferroni or other correction for multiple testing, like the Tukey adjustment).

- ANOVA can be expanded to include multiple grouping variables (Multi-way ANOVA). Algebra is hard, but R does the work for us ☺. Interactions can then be considered.

52

Unit 8 Outline

- Analysis of Variance (ANOVA)
  - General format and ANOVA's *F*-test
  - Assumptions for ANOVA *F*-test
  - Contrast testing
  - Other post-hoc tests
  - Two-way ANOVA (and Multi-way ANOVA)
- Kruskal-Wallis Test

53

Kruskal-Wallis Test

- If either of the Normality or constant variance assumptions fail, then another test to consider is the ***Kruskal-Wallis Test.***
- The KW test is just an extension of the Wilcoxon Rank Sum test to 3 or more groups.
- Remember the procedure to calculate the test statistic?

  1) Rank all the combined data ignoring groups from 1 to N. (treating them like one sample). For any ties, average those ranks.

  2) Then calculate an [F-like] χ^2 test statistic:

$$K = (N-1)\frac{\sum_{i=1}^{I} n_i(\bar{R}_i - \bar{R})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2} \overset{approx}{\sim} \chi^2_{I-1}$$

- Hypotheses? Just like for the Rank Sum test.

Kruskal-Wallis Test

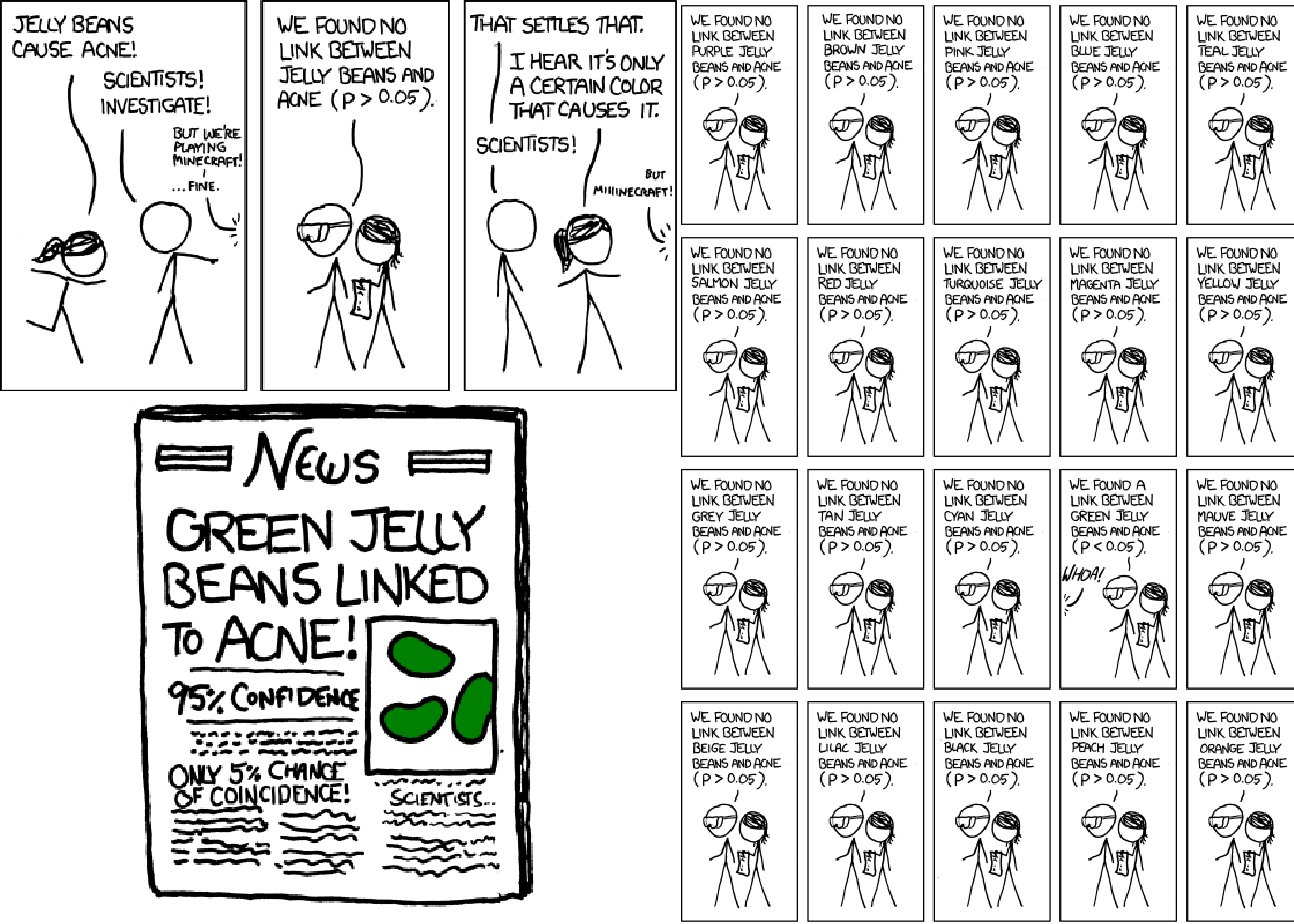
Here's some R code to do the calculations:

```
ranks=rank(logtexts)
ni=as.vector(table(classyear))
ribar=as.vector(by(ranks,classyear,mean))
rbar=mean(ranks)
N=length(ranks)
I=length(ni)
K=(N-1)*sum(ni*(ribar-rbar)^2)/sum((ranks-rbar)^2)
K
1-pchisq(K,df=I-1)
```

Or use:

```
kruskal.test()
```


55



http://xkcd.com/882/

Data snooping.