HUDM4122
Probability and Statistical Inference

May 4, 2015

Today…

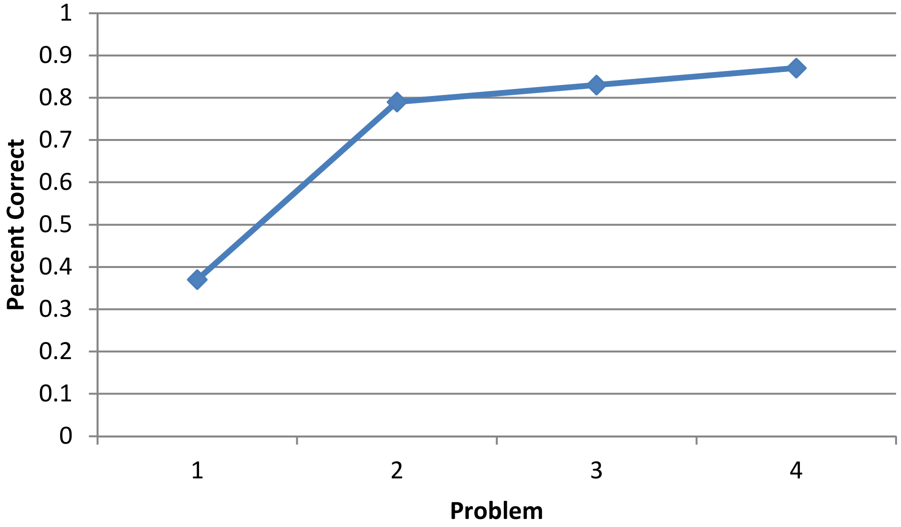
- We have a special guest

- Professor Young-Sun Lee

- If you enjoyed this class, you may enjoy taking
  - HUDM5122: Applied Regression Analysis
  - HUDM6051: Psychometric Theory I
  - HUDM6052: Psychometric Theory II
- With professor Lee

HW11: χ^2

- Very nice work!
- Folks struggled with the first problem, but brought it together in the later problems



Data as of 9:27 pm, 5/3/2015

HW11: χ^2

- Furthermore, most of the errors on problems 3-4 were minor mathematical errors rather than conceptual errors

So let’s continue with our discussion of ANOVA

Where we left off…

- … We were just getting rolling with discussing the mathematics behind Analysis of Variance (ANOVA)
- For the single-factor case
- Comparing a set of several means to each other

Review: Big Idea

- Compute whether there’s a difference between groups

- By comparing between-group variance to other variance

Review: Single-factor ANOVA

- H_0 : All groups have the same mean μ
- H_a : At least one group has a mean that is statistically significantly different than the other means

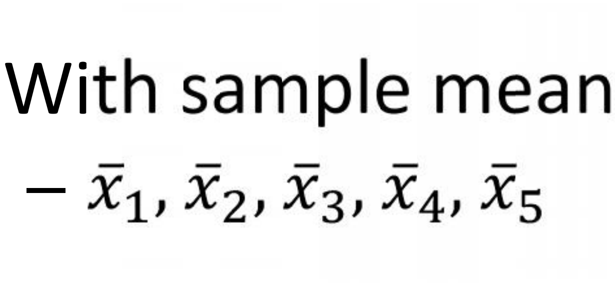
Example We Were Discussing

- The Mt. Vernon City School District is considering 5 science curricula
  - Interactive Science
  - Holt Science Spectrum
  - McDougal Littell Science
  - CK-12 Science
  - Bob’s Discount Science Curriculum

- They randomly divide students into five groups, and each classroom uses one curriculum

You have k samples from k populations

- In this case 5 samples from 5 populations

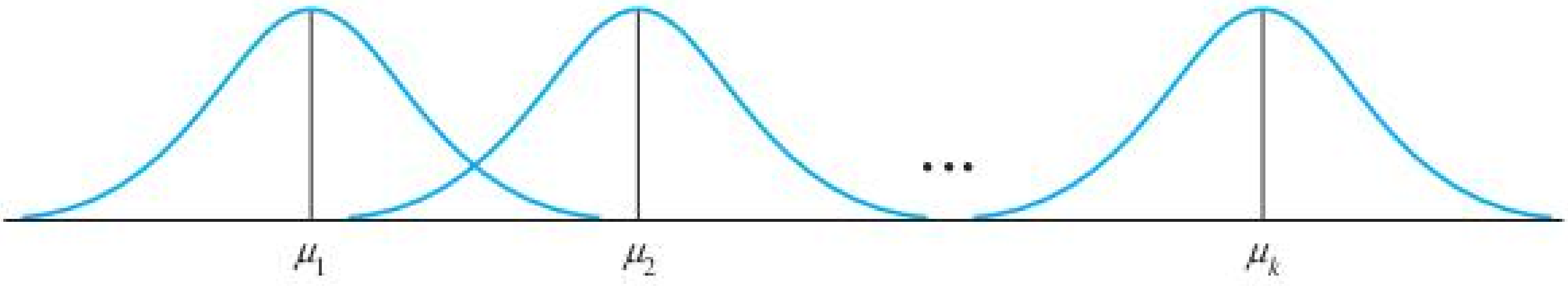


- With sample means
  - $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5$

- And sample standard deviations are close enough to hypothesize that there is a common σ^2

- Is at least one mean higher or lower than the rest?

Common variance but different means



Review: What we do

- Take x_{ij}, the j-th data point for the i-th sample
- And take the overall sample mean, $\bar{x}$

- In that case, we can assess the total variation in the experiment as the *total sum of squares*

Review: Total Sum of Squares

- Total SS = $\sum(x_{ij} - \bar{x})^2$

Review: Total Sum of Squares

- Is made up of two components
  - The sum of squares for treatments (SST)
  - The sum of squares for errors (SSE)

- Total SS = SST + SSE

Review:
Sum of Squares for Treatments (SST)

- The variance attributable to the difference between treatments

- SST = $\sum n_i(\bar{x}_i - \bar{x})^2$

Review:
Sum of Squares for Error (SSE)

- The pooled variation in the *k* samples

- SSE = $(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2$ + … + $(n_k - 1){s_k}^2$

Questions? Comments?

Let’s move forward

Once you have Total SS, SST, SSE

- You can find the degrees of freedom for each

- And then compute the mean squares

- Which are used to conduct an ANOVA

Degrees of freedom

- Degrees of freedom on total SS = (n-1)
  - Based on total data set size, ignoring groups
- Degrees of freedom on SST = (k-1)
  - Based on number of groups
  - Number of parameters we get from having that many groups
- Degrees of freedom on SSE = n-k
  - What's left over

Mean squares

- Degree of variance predicted per degree of freedom

- MSS = TSS/df(TSS)
- MST = SST/df(SST)
- MSE = SSE/df(SSE)

- MSE is a pooled estimate of σ^2
  - The estimated variance across the whole data set, regardless of whether or not H_0 is true

Now we can test our null hypothesis

- H_0 : All groups have the same mean μ

- H_a : At least one group has a mean that is statistically significantly different than the other means

How do we test it?

- Well, if H_0 is true

- Then MST = MSE

- Because the variation between groups will be the same as the variation within all groups together

How do we test it?

- But if H_0 is false

- Then MST > MSE

- Because the variation between groups will be bigger than the variation within all groups together

So we can compute

- F = $\frac{MST}{MSE}$

- Where F is a new distribution that we haven’t seen before

F Distribution

- Is the ratio of two χ^2 distributions
- F = $\chi_1{}^2$ (df$_1$)/ $\chi_2{}^2$ (df$_2$)

F Distribution

- And as such, it has two types of degrees of freedom

numerator df

denominator df

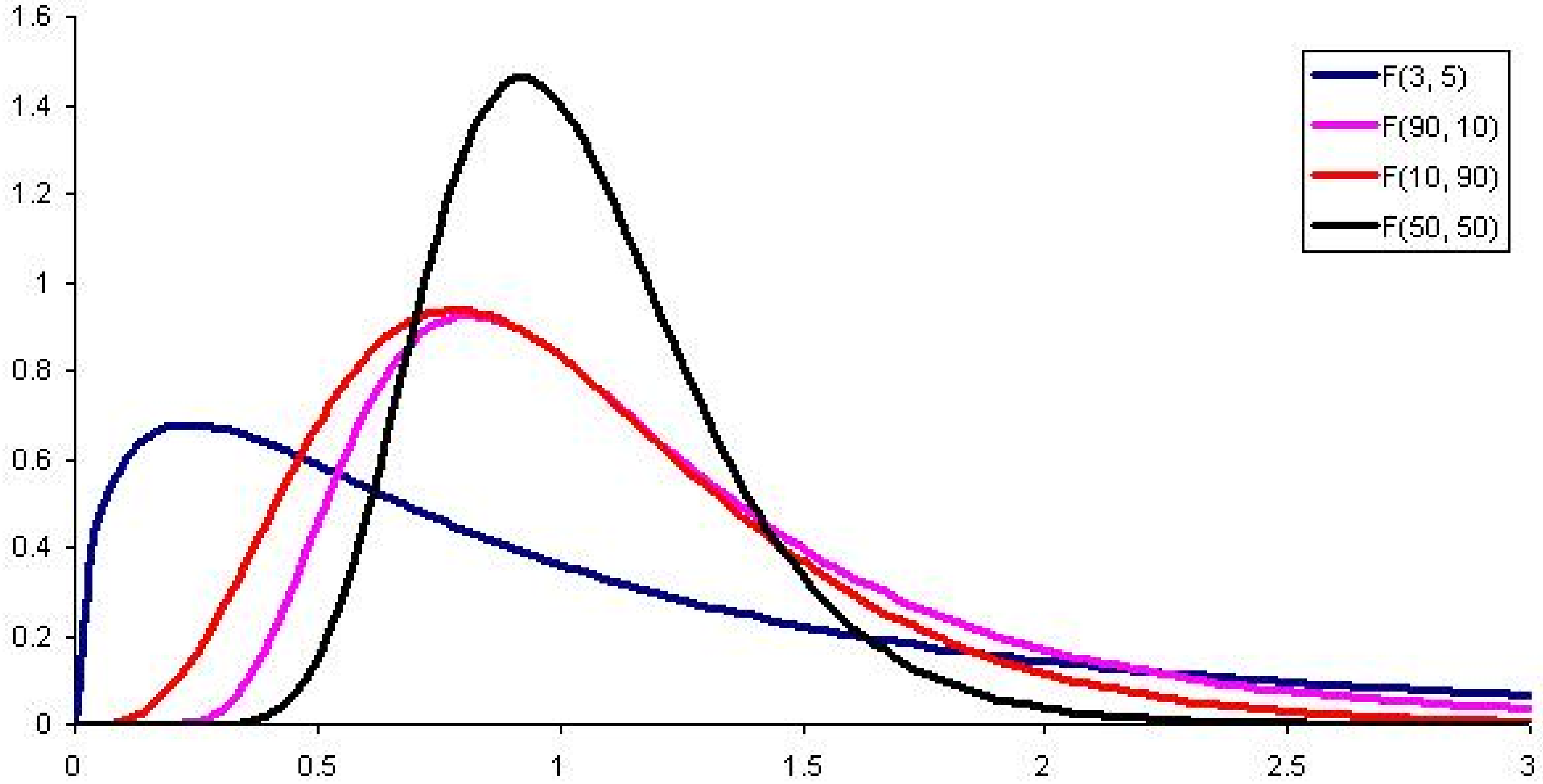
- $F = \chi_1^2\ (df_1)/\ \chi_2^2\ (df_2)$

F distribution

- Numerator degrees of freedom
  - MST degrees of freedom
  - k-1

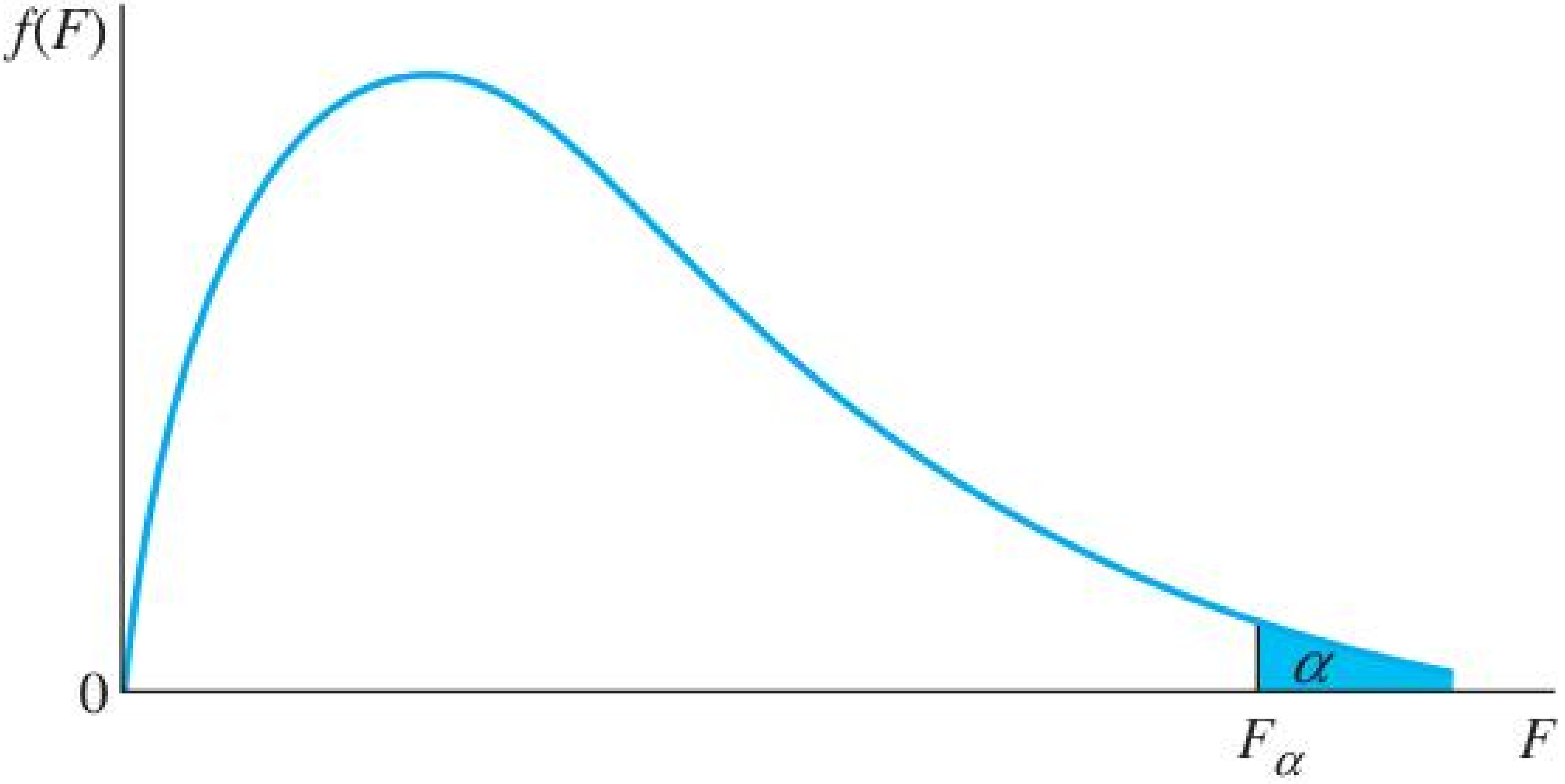
- Denominator degrees of freedom:
  - MSE degrees of freedom
  - n-k

F Distribution



From www.epixanalytics.com

Rejection Region



Finding the p value

- For a given F value, denominator degrees of freedom df (MST), and numerator degrees of freedom df (MSE)
- You can find the p value using a calculator
- For example, in Excel =FDIST(F,df(MST),df(MSE)
- Written F(df(MST),df(MST))=f, p =

Comments? Questions?

Example We Were Discussing

- The Mt. Vernon City School District is considering 5 science curricula
  - Interactive Science
  - Holt Science Spectrum
  - McDougal Littell Science
  - CK-12 Science
  - Bob’s Discount Science Curriculum

- They randomly divide students into five groups, and each classroom uses one curriculum

Let’s Say We Measure Student Attitudes

- Interactive Science: 3, 4, 5
- Holt Science Spectrum: 5, 5, 4
- McDougal Littell Science: 4, 4, 5
- CK-12 Science: 3, 3, 4
- Bob’s Discount Science Curriculum: 1, 1, 2

$$\bar{x} = 3.5333$$

- Interactive Science: 3, 4, 5
- Holt Science Spectrum: 5, 5, 4
- McDougal Littell Science: 4, 4, 5
- CK-12 Science: 3, 3, 4
- Bob’s Discount Science Curriculum: 1, 1, 2

$$\bar{x} = 3.5333$$

- Interactive Science: 3, 4, 5: $\bar{x}_1$
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$
- CK-12 Science: 3, 3, 4: $\bar{x}_4$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$

$$\bar{x} = 3.5333$$

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.33$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.33

Total SS = $\sum (x_{ij} - \bar{x})^2$

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.33$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.33

Total SS = $\sum (x_{ij} - \bar{x})^2$

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.33$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.33

- $(3-3.533)^2 + (4-3.533)^2 +(5-3.533)^2 +(5-3.533)^2+(5-3.533)^2+(4-3.533)^2 + ...$

Total SS = $\sum (x_{ij} - \bar{x})^2 = 25.7333$

- Interactive Science: 3, 4, 5: $\bar{x}_1 = 4$
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2 = 4.667$
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3 = 4.333$
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.33$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5 = 1.33$

- $(3-3.533)^2 + (4-3.533)^2 + (5-3.533)^2$
  $+ (5-3.533)^2 + (5-3.533)^2 + (4-3.533)^2 + ….$

SST = $\sum n_i(\bar{x}_i - \bar{x})^2$

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.333$
- Bob's Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.333

- $3(4-3.533)^2 + 3(4.667-3.533)^2 + 3(4.333-3.533)^2 + 3(3.333-3.533)^2 + 3(1.333-3.533)^2$

SST = $\sum n_i(\bar{x}_i - \bar{x})^2$ = 21.0667

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.333$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.333

- $3(4-3.533)^2 + 3(4.667-3.533)^2 + 3(4.333-3.533)^2 + 3(3.333-3.533)^2 + 3(1.333-3.533)^2$

Total SS = SST + SSE

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.333$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.333

Total SS = SST + SSE
25.7333 = 21.0667 + SSE

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.333$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.333

25.7333 = 21.0667 + SSE
SSE = 4.6667

- Interactive Science: 3, 4, 5: $\bar{x}_1$= 4
- Holt Science Spectrum: 5, 5, 4: $\bar{x}_2$ = 4.667
- McDougal Littell Science: 4, 4, 5: $\bar{x}_3$ = 4.333
- CK-12 Science: 3, 3, 4: $\bar{x}_4 = 3.333$
- Bob’s Discount Science Curriculum: 1, 1, 2: $\bar{x}_5$ = 1.333

DF

- Degrees of freedom on SST = (k-1)
  - DF(SST) = (5-1)=4
- Degrees of freedom on SSE = n-k
  - DF(SSE) = 15-4 = 11

Mean Squares

- MST = SST/df(SST) = 21.0667/4 = 5.2667
- MSE = SSE/df(SSE) = 4.6667/11 = 0.4242

F

- MST = SST/df(SST) = 21.0667/4 = 5.2667
- MSE = SSE/df(SSE) = 4.6667/11 = 0.4242

- F = $\frac{MST}{MSE}$ = 5.2667/0.4242 = 12.4143

- F(4,11) = 12.4143, $p<0.01$

- So there is an overall difference between groups

Questions? Comments?

Please Try This In Pairs

- Student Numbers of Complaints by Class

- Interactive Science: 2, 4, 4
- Holt Science Spectrum: 4, 4, 5
- McDougal Littell Science: 3, 2, 3
- CK-12 Science: 3, 4, 3
- Bob’s Discount Science Curriculum: 4, 3, 4

Questions? Comments?

A little more on violations of assumptions

Violations of assumptions

- We’ve talked about violations of assumptions throughout the semester

- Which ones are more serious
- Which ones are less serious

Sample Sizes

- You can’t use Z or χ^2 with insufficiently large sample sizes

- As we’ve discussed, t is an alternative to Z for small samples

Sample Sizes

- Fisher’s Exact Test is an alternative to χ^2 for small samples

- Either small total sample or badly unbalanced samples

Violations of distribution

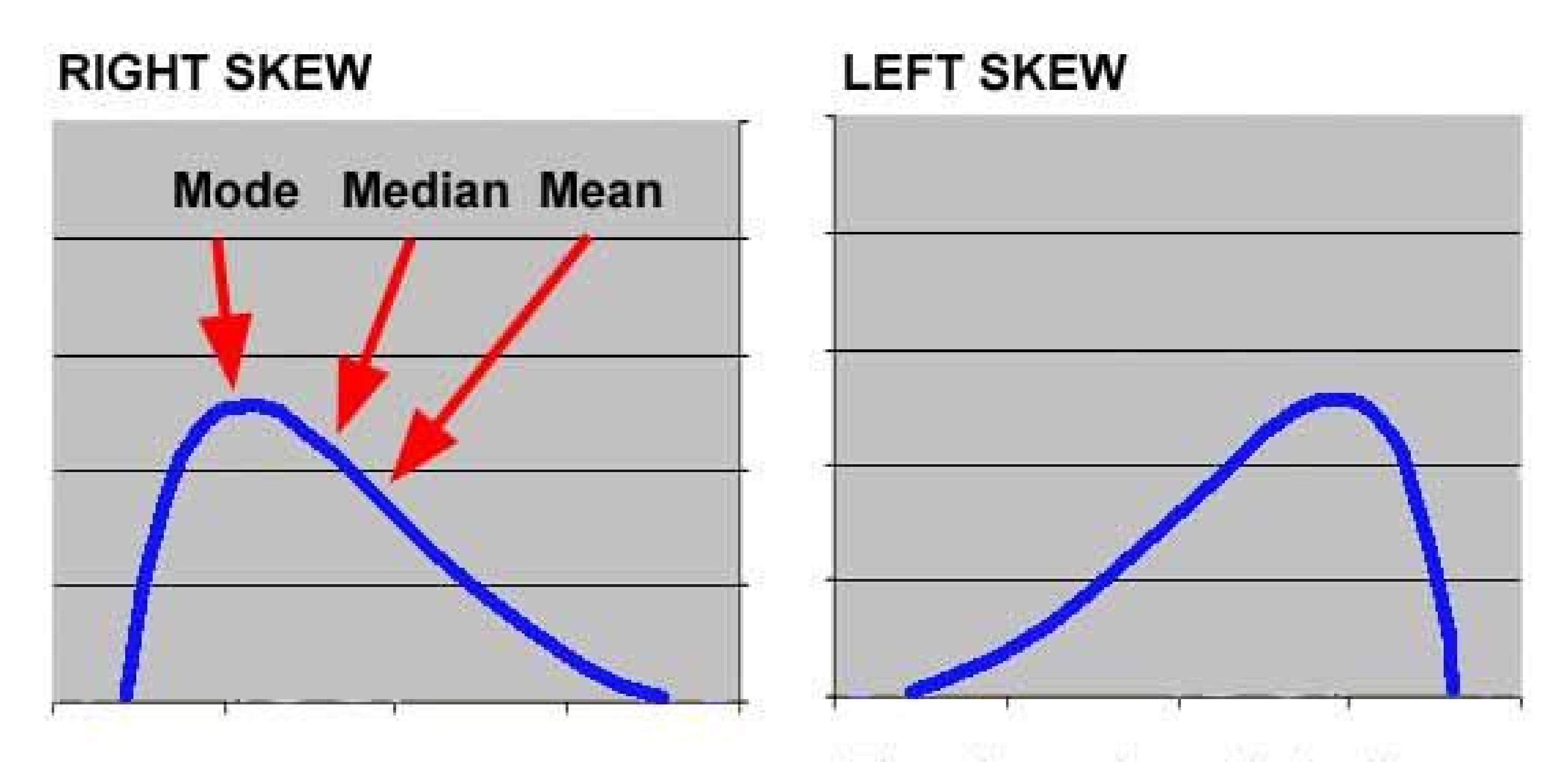
- All the tests we’ve discussed this semester assume that your data is approximately distributed according to the expected distribution

- Often called a “violation of normality” when the data is expected to be normal

Robustness

- ANOVA (and other tests we’ve discussed) are reasonable robust to violations of normality

Skew



Skew

- Not a huge problem
- You can usually transform the data by taking the logarithm or exponentiating, to cure this
- There are “tests of skewness” that can provide guidelines on whether you ought to be doing this

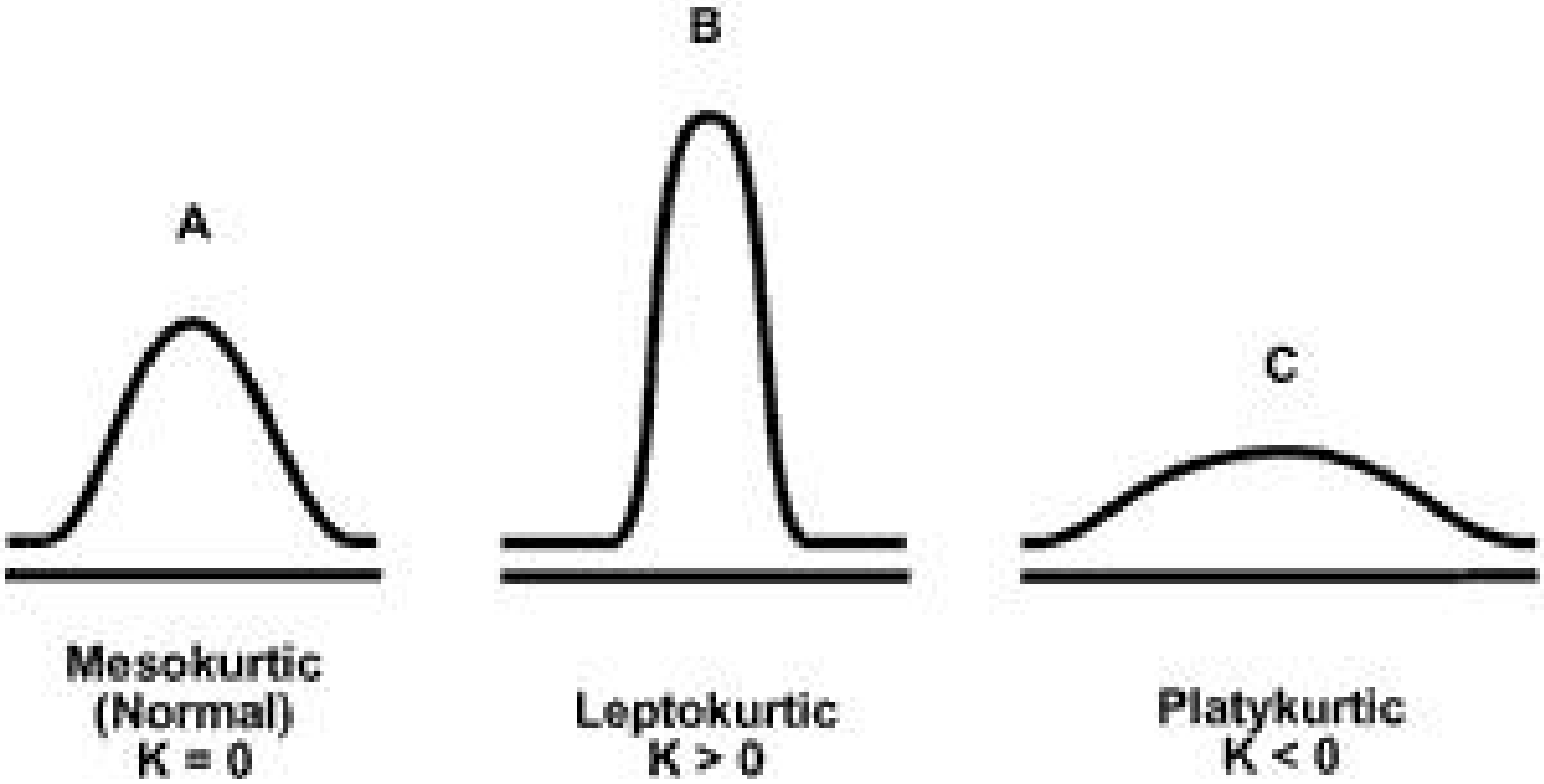
Tests of Skewness

- Best known is perhaps Pearson’s Moment Coefficient of Skewness

- Can be used to assess whether data is too skew to use tests we’ve discussed in this class (at least not without transformation)

- If your data looks normal, it will be fine
- You need to be quite skew to be “too skew”

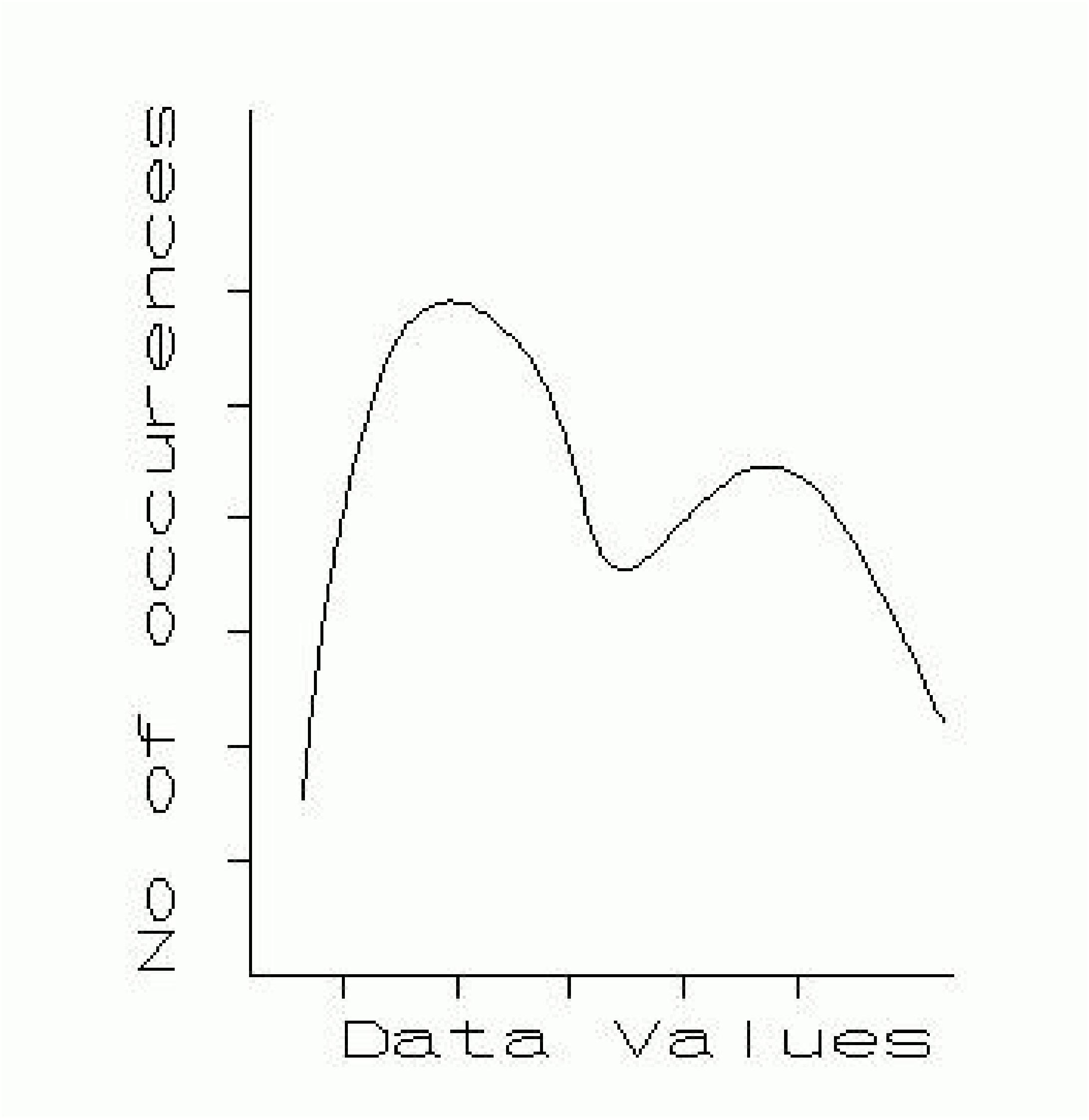
Kurtosis



Kurtosis

- A major issue
- Student’s t distribution is leptokurtic
- But there is data that is too leptokurtic for Student’s t distribution

Bimodal Distribution



Bimodal Distribution

- Causes standard methods to completely explode
- Can be dealt with by fitting the data as a function of two normal curves

Non-parametric tests

- There is a different class of statistics tests you can use, called *non-parametric tests*

- Valid for data with skew, kurtosis, or bimodality issues

Non-parametric tests

- Do not make assumptions about distribution
- Ignore extreme values
- Typically less statistical power, so not preferred when distribution assumptions are not violated

Non-parametric tests

- Some examples include
  - Mann-Whitney U Test (alternative to 2-group t-test)
  - Wilcoxon Signed-Rank Test (alternative to paired t-test)
  - Kruskal-Wallis’s Test (alternative to ANOVA)

Violation of homogeneity of variance

- Recall that for ANOVA we assume σ^2 is the same for each condition

- If this is violated, the model breaks

Violation of homogeneity of variance

- Recall that for ANOVA we assume σ^2 is the same for each condition

- If this is violated, the model breaks

- In this case you can use Kruskal-Wallis's Test (alternative to ANOVA)

Violation of independence assumptions

- Very serious – often leads to high Type I error

Violation of independence assumptions

- Several alternatives
  - – Explicitly model non-independence
    - Student-level terms in linear regression
    - Hierarchical models
  - – Collapse into one data point per student

Questions? Comments?

Upcoming Classes

- 5/6 ***HW 12 due***

- 5/6 3pm REVIEW SESSION, RUSSELL 302
- 5/8 noon REVIEW SESSION, GRACE DODGE 453

- ***5/11 FINAL EXAM***

Review Sheet Posted

- http://www.columbia.edu/~rsb2162/Stats2015/HUDM4122-ReviewSheet-Final.pdf

Practice Problems (and Answers!) Posted

- See Moodle