Lecture 21: more 11.3 (ANOVA) Categorical & Quantitative Variable Begin Ch.12 Inf. for 2 Categorical Vars.

- ANOVA: Table, Test Stat, *P*-value
- 1st Step in Practice: Displays, Summaries
- ANOVA Output
- Guidelines for Use of ANOVA
- Formulating Hypotheses about 2 Cat. Vars.
- Test Based on Proportions or Counts: *z* or ChiSq

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing (Lectures 3-8)
  - Probability (discussed in Lectures 9-14)
  - Statistical Inference
    - 1 categorical (discussed in Lectures 14-16)
    - 1 quantitative (discussed in Lectures 16-18)
    - cat and quan: paired, 2-sample, several-sample
    - 2 categorical
    - 2 quantitative

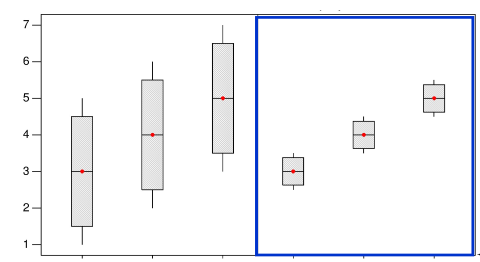
The *F* Statistic *(Review)*

$$F=\frac{\left[n_1(\bar{x}_1-\bar{x})^2+n_2(\bar{x}_2-\bar{x})^2+\cdots+n_I(\bar{x}_I-\bar{x})^2\right]/(I-1)}{\left[(n_1-1)s_1^2+(n_2-1)s_2^2+\cdots+(n_I-1)s_I^2\right]/(N-I)}$$

- **Numerator:** variation among groups
  - How different are $\bar{x}_1,\cdots,\bar{x}_I$ from one another?
- **Denominator:** variation within groups
  - How spread out are samples? (sds $s_1,\cdots,s_I$)

Role of Variations on Conclusion *(Review)*

Boxplots with same variation *among* groups (3, 4, 5) but different variation *within*: sds large (left) or small (right)



Scenario on right: smaller s.d.s → larger $F=\frac{\text{var among}}{\text{var within}}$ →smaller *P*-value→ likelier to reject H_0 →conclude pop means differ

ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F	P
Factor	$DFG = I - 1$	SSG	$MSG = SSG/DFG$	$F = \frac{MSG}{MSE}$	p-value
Error	$DFE = N - I$	SSE	$MSE = SSE/DFE$		
Total	$N - 1$	SST			

- Organizes calculations
  - "Source" refers to source of variation:
    - "Factor" refers to variation among groups (expl var)

    This variation is from the numerator.
    - "Error" refers to individuals differing within groups

    This variation is from the denominator.

ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F	P
Factor	$DFG = I - 1$	SSG	$MSG = SSG/DFG$	$F = \frac{MSG}{MSE}$	p-value
Error	$DFE = N - I$	SSE	$MSE = SSE/DFE$		
Total	$N - 1$	SST			

- Organizes calculations
  - "Source" refers to source of variation
  - DF: use I = no. of groups, N = total sample size
    - DFG = $I - 1$
    - DFE = $N - I$

ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F	P
Factor	$DFG = I - 1$	SSG	$MSG = SSG/DFG$	$F = \frac{MSG}{MSE}$	p-value
Error	$DFE = N - I$	SSE	$MSE = SSE/DFE$		
Total	$N - 1$	SST			

- Organizes calculations
  - "Source" refers to source of variation
  - DF: use I = no. of groups, N = total sample size
  - SSG measures overall variation among groups
  - SSE measures overall variation within groups

SSG and SSE tedious to calculate; other table entries straightforward, except for P-value

ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F	P
Factor	$DFG = I - 1$	SSG	$MSG = SSG/DFG$	$F = \frac{MSG}{MSE}$	p-value
Error	$DFE = N - I$	SSE	$MSE = SSE/DFE$		
Total	$N - 1$	SST			

- Organizes calculations
  - "Source" refers to source of variation
  - DF: use I = no. of groups, N = total sample size
  - SSG measures overall variation among groups
  - SSE measures overall variation within groups
  - Mean Sums: Divide Sums by DFs
  - F: Take quotient of MSG and MSE
  - P-value: Found with software or tables

Example: *Key ANOVA Values*

- **Background**: Compare mileages for 8 sedans, 8 minivans, 12 SUVs; find SSG=42.0, SSE=181.4.
- **Question:** What are the following values for table:
  - **DFG? DFE? MSG? MSE? *F*?**
- **Response:**
  - **DFG** = 3 – 1 =___
  - **DFE** = *N* – *I* = (8+8+12) – 3 =____
  - **MSG** = SSG/DFG = 42/2 =____
  - **MSE**= SSE/DFE = 181.4/25 =_____
  - ***F***= MSG/MSE = 21/7.256 =_____

 Practice: 11.29e-g p.561

Example: *Completing ANOVA Table*

- **Background**: Found these values for ANOVA:
  - **DFG**=3-1= 2
  - **DFE**=*N*-*I*=(8+8+12)-3= 25
  - **MSG**=SSG/DFG=42/2= 21
  - **MSE**=SSE/DFE=181.4/25= 7.256
  - ***F***=MSG/MSE=21/7.256= 2.89
- **Question:** Complete ANOVA table?
- **Response:** Software→*P*-val=0.0743→ ___________

Source	DF	SS	MS	F	P
Factor					
Error					

 Practice: 11.32 p.563

ANOVA *F* Statistic and *P*-Value

- Sample means very different→
  F large→
  P-value small→
  Reject claim of equal population means.
- Sample means relatively close→
  F *not* large→
  P-value *not* small→
  Believe claim of equal population means.

How Large is "Large" *F*

Particular *F* distribution determined by DFG, DFE

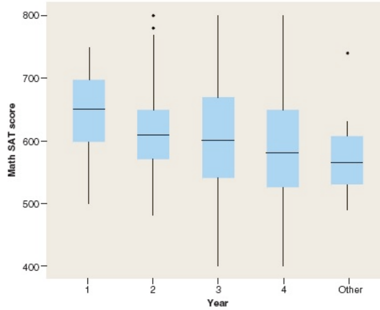
(these determined by sample size, number of groups)

P-value in software output lets us know if *F* is large.

Note: P-value is "bottom line" of test; "top line" is examination of display and summaries.

Example: *Examining Boxplots*

- **Background**: For all students at a university, are Math SATs related to what year they're in?



- **Question:** What do the boxplots suggest?
- **Response:** As year goes up, mean ______________ (Suggests ___________ students scored better in Math.)

Example: *Examining Summaries*

- **Background**: For all students at a university, are Math SATs related to what year they're in?

Level	N	Mean	StDev
1	32	643.75	63.69
2	233	613.91	61.00
3	87	601.84	89.79
4	28	581.79	89.73
other	10	578.00	72.08

- **Question:** What do the summaries suggest?
- **Response:** Means decrease by about ______ points for each successive year 1 to 4. Standard deviations are around ____, and sample sizes are ____________.

Example: *ANOVA Output*

- **Background**: For all students at a university, are Math SATs related to what year they're in?

Analysis of Variance for Math

Source	DF	SS	MS	F	P
Year	4	78254	19563	3.87	0.004
Error	385	1946372	5056		
Total	389	2024626			

- **Question:** What does the output suggest?
- **Response:** Test H_o :

  P-value=0.004. Small? ____ Reject H_0?

  ____ Conclude all 5 population means may be equal?

  ____Year and Math SAT related in population?____

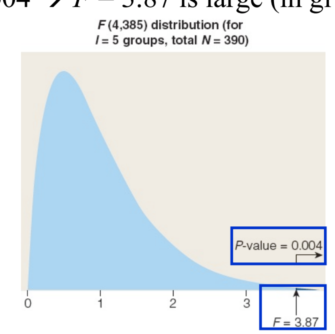
How Large is "Large" *F (Review)*

Particular *F* dist determined by DFG, DFE

(these determined by sample size, number of groups)

P-value in software output lets us know if *F* is large.

P-value = 0.004 → *F* = 3.87 is large (in given situation)



Example: *ANOVA Output*

- **Background**: A test for a relationship between Math SAT and year of study, based on data from a large sample of intro stats students at a university, produced a large *F* and a small *P*-value.
- **Question:** What issues should be considered before we use these results to draw conclusions about the relationship between year of study and Math SAT for all students at that university?
- **Response:**

Guidelines for Use of ANOVA Procedure

- Need random samples taken independently from several populations.
- Confounding variables should be separated out.
- Sample sizes must be large enough to offset non-normality of distributions.
- Need populations at least 10 times sample sizes.
- Population variances must be equal.

Pooled Two-Sample *t* Procedure *(Review)*

If we can assume $\sigma_1 = \sigma_2$, standardized difference between sample means follows a pooled *t* distribution.

- Some apply Rule of Thumb: use pooled *t* if larger sample s.d. not more than twice smaller.

The F distribution is in a sense "pooled": our standardized statistic follows the F distribution only if population variances are equal (same as equal s.d.s)

Example: *Checking Standard Deviations*

- **Background**: For all students at a university, are Math SATs related to what year they're in?

Level	N	Mean	StDev
1	32	643.75	63.69
2	233	613.91	61.00
3	87	601.84	89.79
4	28	581.79	89.73
other	10	578.00	72.08

- **Question:** Is it safe to assume equal population variances?
- **Response:**

Largest s.d.=_____ > 2(smallest s.d.)___________?

____ Assumption of equal variances OK? ____

Example: *Reviewing ANOVA*

- **Background**: For all students at a university, are Verbal SATs related to what year they're in?

```
Level         N      Mean     StDev
1            32    596.25     86.91
2           234    592.76     65.87
3            86    596.51     77.26
4            29    579.83     79.47
other        10    551.00    124.32
Source     DF        SS        MS        F        P
Year        4     23559      5890     1.10    0.357
```

- **Questions:** Are conditions met? Do the data provide evidence of a relationship?
- **Response:** n_i large and 124.32 not > 2(65.87)→___ P-val=0.357 small?___ Evidence of a relationship?___

 Practice: 11.77c-d p.582

Guidelines for Use of ANOVA *(Review)*

- Need random samples taken independently from several populations
- Confounding variables should be separated out
- Sample sizes must be large enough to offset non-normality of distributions
- Need populations at least 10 times sample sizes
- Population variances must be equal.

Example: *Considering Data Production*

- **Background**: F test found evidence of relationship between Math SAT and year (P-value 0.004), but not Verbal SAT and year (P-value 0.357).
- **Question:** Keeping in mind that the sample consisted of students in various years taking an introductory statistics class, are there concerns about bias/confounding variables?
- **Response:** For Math, ____. For Verbal, ____

 Practice: 11.31 p.562

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing (Lectures 3-8)
  - Probability (discussed in Lectures 9-14)
  - Statistical Inference
    - 1 categorical (discussed in Lectures 14-16)
    - 1 quantitative (discussed in Lectures 16-18)
    - cat and quan: paired, 2-sample, several-sample (Lectures 19-21)
    - 2 categorical
    - 2 quantitative

Inference for Relationship *(Review)*

- H_0 and H_a about variables: not related or related
  - Applies to all three C→Q, C→C, Q→Q
- H_0 and H_a about parameters: equality or not
  - C→Q: pop means equal?
  - C→C: pop proportions equal?
  - Q→Q: pop slope equals zero?

Example: *2 Categorical Variables: Hypotheses*

- **Background**: We are interested in whether or not smoking plays a role in alcoholism.
- **Question:** How would H_0 and H_a be written
  - in terms of variables?
  - in terms of parameters?
- **Response:**
  - in terms of variables
    - H_0: smoking and alcoholism ________ related
    - H_a: smoking and alcoholism________related
  - in terms of parameters
    - H_0: Pop proportions alcoholic________for smokers, non-smokers
    - H_a: Pop. proportions alcoholic ________ for smokers, non-smokers

The word "not" appears in Ho about variables, in Ha about parameters.

 Practice: 12.24 p.616

Example: *Summarizing with Proportions*

- **Background**: Research Question: Does smoking play a role in alcoholism?
- **Question:** What statistics from this table should we examine to answer the research question?
- **Response:** Compare proportions__________(response) for ________________________________ (explanatory).

	Alcoholic	Not Alcoholic	Total
Smoker	30	200	230
Nonsmoker	10	760	770
Total	40	960	1,000

 Practice: 12.12b-c p.609

Example: *Test Statistic for Proportions*

- **Background**: One approach to the question of whether smoking and alcoholism are related is to compare proportions.

	Alcoholic	Not Alcoholic	Total
Smoker	30	200	230
Nonsmoker	10	760	770
Total	40	960	1,000

$$\hat{p}_1 = \frac{30}{230} = 0.130$$
$$\hat{p}_2 = \frac{10}{770} = 0.013$$

- **Question:** What would be the next step, if we've summarized the situation with the difference between sample proportions 0.130-0.013?
- **Response:** _____________the difference between sample proportions 0.130-0.013.

Stan. diff. is normal for large *n*: ______________________

 Practice: 12.2a-b p.595

z Inference for 2 Proportions: Pros & Cons

Advantage:

Can test against *one-sided* alternative.

Disadvantage:

2-by-2 table: comparing proportions straightforward

Larger table: comparing proportions complicated, can't just standardize one difference $\hat{p}_1 - \hat{p}_2$

Another Comparison in Considering Categorical Relationships *(Review)*

- Instead of considering how different are the *proportions* in a two-way table, we may consider how different the *counts* are from what we'd expect if the "explanatory" and "response" variables were in fact unrelated.
- Compared observed, expected counts in wasp study:

Obs	A	NA	T
B	16	15	31
U	24	7	31
T	40	22	62

Exp	A	NA	T
B	20	11	31
U	20	11	31
T	40	22	62

Inference Based on Counts

To test hypotheses about relationship in r-by-c table, compare counts observed to counts expected if H_0(equal proportions in response of interest) were true.

Example: *Table of Expected Counts*

- **Background**: Data on smoking and alcoholism:

	Alcoholic	Not Alcoholic	Total
Smoker	30	200	230
Nonsmoker	10	760	770
Total	40	960	1,000

- **Question:** What counts are expected if H_0 is true?
- **Response:** Overall proportion alcoholic is______________

If proportions alcoholic were same for S and NS, expect

- (40/1,000)(230)=____ smokers to be alcoholic
- (40/1,000)(770)=____ non-smokers to be alcoholic; also
- (960/1,000)(230)=____ smokers not alcoholic
- (960/1,000)(770)=____ non-smokers not alcoholic

Practice: 12.12d p.609

Example: *Table of Expected Counts*

- **Background**: If proportions alcoholic were same for S and NS, expect
  - (40/1,000)(230)= 9.2 smokers to be alcoholic
  - (40/1,000)(770)= 30.8 non-smokers to be alcoholic; also
  - (960/1,000)(230)= 220.8 smokers not alcoholic
  - (960/1,000)(770)= 739.2 non-smokers not alcoholic
- **Question:** Where do they appear in table of expected counts?
- **Response:**

	Alcoholic	Not Alcoholic	Total
Smoker			230
Nonsmoker			770
Total	40	960	1,000

Note:
9.2/230 =
30.8/770 =
40/1,000

 Practice: 12.12e p.609

Example: *Table of Expected Counts*

	Alcoholic	Not Alcoholic	Total
Smoker	9.2	220.8	230
Non-smoker	30.8	739.2	770
Total	40	960	1000

- **Note:** Each expected count is $\frac{\text{Column total} \times \text{Row total}}{\text{Table total}}$

  Expect:
  - (40)(230)/1,000 = 9.2 smokers to be alcoholic
  - (40)(770)/1,000 = 30.8 non-smokers to be alcoholic; also
  - (960)(230)/1,000 = 220.8 smokers not alcoholic
  - (960)(770)/1,000 = 739.2 non-smokers not alcoholic

 Practice: 12.12e p.609

Chi-Square Statistic

- Components to compare observed and expected counts, one table cell at a time:

$$\text{component} = \frac{(\text{observed - expected})^2}{\text{expected}}$$

Components are **individual** standardized squared differences.

- **Chi-square** test statistic χ^2 combines all components by summing them up:

$$\text{chi-square} = \text{sum of } \frac{(\text{observed - expected})^2}{\text{expected}}$$

Chi-square is **sum** of standardized squared differences.

Example: *Chi-Square Statistic*

- **Background**: Observed and Expected Tables:

Obs	A	NA	Total
S	30	200	230
NS	10	760	770
Total	40	960	1000

Exp	A	NA	Total
S	9.2	220.8	230
NS	30.8	739.2	770
Total	40	960	1000

- **Question:** What is the chi-square statistic?
- **Response:** Find $\text{chi-square} = \text{sum of } \frac{(\text{observed - expected})^2}{\text{expected}}$

  =

 Practice: 12.12g,i p.609

Example: *Assessing Chi-Square Statistic*

- **Background**: We found chi-square = 64.
- **Question:** Is the chi-square statistic (64) large?
- **Response:**

Chi-Square Distribution

chi-square = sum of $\frac{(\text{observed - expected})^2}{\text{expected}}$ follows a predictable pattern (assuming H_0 is true) known as **chi-square distribution** with df = (r-1) × (c-1)

- r = number of rows (possible explanatory values)
- c= number of columns (possible response values)

Properties of chi-square:

- Non-negative (based on squares)
- Mean=df [=1 for smallest (2×2) table]
- Spread depends on df
- Skewed right

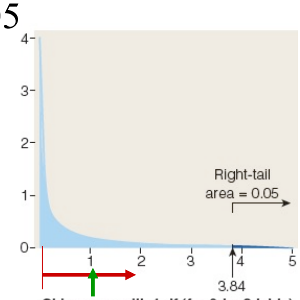
Chi-Square Density Curve

For chi-square with 1 df, $P(\chi^2 \geq 3.84) = 0.05$

→ If $\chi^2 > 3.84$, P-value < 0.05

Properties of chi-square:

- Non-negative
- Mean = df
  df=1 for smallest [2×2] table
- Spread depends on df
- Skewed right



Chi-square with 1 df (for 2-by-2 table)

Example: *Assessing Chi-Square (Continued)*

- **Background**: In testing for relationship between smoking and alcoholism in 2×2 table, found $\chi^2 = 64$
- **Question:** Is there evidence of a relationship in general between smoking and alcoholism (not just in the sample)?
- **Response:** For df=(2-1)×(2-1)=1, chi-square considered "large" if greater than 3.84
  →chi-square=64 large? ____ P-value small?____
  Evidence of a relationship between smoking and alcoholism? ____

 Practice: 12.12j-l p.609

Inference for 2 Categorical Variables; z or χ^2

For 2×2 table, $z^2 = \chi^2$

- z statistic (comparing proportions)→ combined tail probability=0.05 for z=1.96
- chi-square statistic (comparing counts)→ right-tail prob=0.05 for $\chi^2 = 1.96^2 = 3.84$

Example: *Relating Chi-Square & z*

- **Background**: We found chi-square = 64 for the 2-by-2 table relating smoking and alcoholism.
- **Question:** What would be the z statistic for a test comparing proportions alcoholic for smokers vs. non-smokers?
- **Response:**

Practice: 12.10 p.608

Assessing Size of Test Statistics *(Summary)*

When test statistic is "large":

- z: greater than 1.96 (about 2)
- t: depends on df; greater than about 2 or 3
- F: depends on DFG, DFE
- χ^2 depends on df=(r-1)×(c-1); greater than 3.84 (about 4) if df=1

Lecture Summary
(Inference for Cat→Quan; More About ANOVA)

- ANOVA for several-sample inference
  - ANOVA table
  - F statistic and P-value
- 1st step in practice: displays and summaries
  - Side-by-side boxplots
  - Compare means, look at sds and sample sizes
- ANOVA output
- Guidelines for use of ANOVA

Lecture Summary

(Inference for Cat→Cat; Chi-Square)

- Hypotheses in terms of variables or parameters
- Inference based on proportions or counts
- Chi-square test
  - Table of expected counts
  - Chi-square statistic, chi-square distribution
  - Relating z and chi-square for 2×2 table
  - Relative size of chi-square statistic