STAT:2010/4200
Statistical Methods and Computing

Contingency Tables and the Chi-Square Test
Introduction to ANOVA

Lecture 21
Apr. 5, 2017

Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

The Chi-square test for differences among more than 2 proportions

We are interested in the *independent samples* case.

Example:

- A study investigated the accuracy of death certificates by comparing the results of 575 autopies to the causes of death listed on the certificates.
- Two hospitals participated in the study.
  - community hospital, labeled A
  - university hospital, labeled B
- Three possible cases
  - death certificate confirmed accurate
  - death certificate contained inaccuracies but did not require correction of underlying cause of death

  - death certificate incorrect and required recoding of underlying cause of death
- Question of interest: Are there differences between the two hospitals with respect to practices in completing death certificates
- One way to address the question: Test null hypothesis that, within each category of death certificate status, the proportions of death certificates coming from Hosptial A are the same.

A multiple comparisons problem!

$$H_0: \quad p_c = p_i = p_r$$
$$H_a: \quad p_c \neq p_i \; or \; p_c \neq p_r \; or \; p_i \neq p_r$$

- We will *first* test whether there are *any* significant differences.
- Only if we reject H_0 in the overall test will we do pairwise tests to find out *which* population proportions are different.

Results

	Hospital A	Hospital B	Total
Confirmed accurate	157	268	425
Inacc, no recoding	18	44	62
Incorrect, recoding	54	34	88
Total	229	346	575

The overall sample proportion of death certificates from hospital A is

$$\frac{229}{575} = 0.398$$

If H_0 is true, we would expect this same proportion of hospital A certificates in all three categories.

Observed and expected counts

	Hospital A	Hospital B	Hospital A	Hospital
Accurate	157	268	169.3	255.7
Incorrect	18	44	24.7	37.3
Recode	54	34	35.0	53.0

The Chi-square statistic is

$$X^2 = 21.62$$

- $r = 3$ rows
- $c = 2$ columns
- So the degrees of freedom is $(r-1)(c-1) = 2(1) = 2$

According to Table E, the .05 cutoff under a Chi-square distribution with 2 d.f. is 5.99.

We can reject H_0 because $21.62 > 5.99$. The p-value < 0.001.

We conclude that the proportions of death certificates from Hospital A are not the same for the three different categories of certificate status.

This Chi-square test in SAS

```
options linesize = 72 ;

data dthcert ;
input hosp $ status $ count ;
datalines ;
A C 157
A I  18
A R  54
B C 268
B I  44
B R  34
;

proc freq data = dthcert ;
tables status * hosp / expected ;
weight count ;
run ;

proc freq data = dthcert ;
tables status * hosp / chisq ;
weight count ;
run ;
```

```
      TABLE OF STATUS BY HOSP

STATUS     HOSP

Frequency|
Expected |
Percent  |
Row Pct  |
Col Pct  |A       |B       |  Total
---------+--------+--------+
C        |    157 |    268 |    425
         | 169.26 | 255.74 |
         |  27.30 |  46.61 |  73.91
         |  36.94 |  63.06 |
         |  68.56 |  77.46 |
---------+--------+--------+
I        |     18 |     44 |     62
         | 24.692 | 37.308 |
         |   3.13 |   7.65 |  10.78
         |  29.03 |  70.97 |
         |   7.86 |  12.72 |
---------+--------+--------+
R        |     54 |     34 |     88
         | 35.047 | 52.953 |
         |   9.39 |   5.91 |  15.30
         |  61.36 |  38.64 |
         |  23.58 |   9.83 |
---------+--------+--------+
```



```
Total          229      346      575
             39.83    60.17   100.00
```



```
      TABLE OF STATUS BY HOSP

STATUS     HOSP

Frequency|
Percent  |
Row Pct  |
Col Pct  |A       |B       |  Total
---------+--------+--------+
C        |    157 |    268 |    425
         |  27.30 |  46.61 |  73.91
         |  36.94 |  63.06 |
         |  68.56 |  77.46 |
---------+--------+--------+
I        |     18 |     44 |     62
         |   3.13 |   7.65 |  10.78
         |  29.03 |  70.97 |
         |   7.86 |  12.72 |
---------+--------+--------+
R        |     54 |     34 |     88
         |   9.39 |   5.91 |  15.30
         |  61.36 |  38.64 |
         |  23.58 |   9.83 |
---------+--------+--------+
Total          229      346      575
             39.83    60.17   100.00
```


STATISTICS FOR TABLE OF STATUS BY HOSP

Statistic	DF	Value	Prob
Chi-Square	2	21.523	0.001
Likelihood Ratio Chi-Square	2	21.189	0.001
Mantel-Haenszel Chi-Square	1	12.864	0.001
Phi Coefficient		0.193	
Contingency Coefficient		0.190	
Cramer's V		0.193	

Sample Size = 575

The sample proportions are

	Hospital A	Hospital B	Total
Confirmed accurate	157	268	0.369
Inacc, no recoding	18	44	0.409
Incorrect, recoding	54	34	0.614
Total	229	346	575

More advanced methods provide tests and confidence intervals to formalize analysis of which population proportions are significantly different.

Comparing more than two population means

Example: Does the presence of pets or friends affect responses to stress?

- Allen, Blascovich, Tomaka, and Kelsey, 1988, *Journal of Personality and Social Psychology*
- subjects: 45 women who described themselves as dog lovers
- randomly assigned to three groups: to do a stressful task
  1. alone
  2. with a good friend present
  3. with their dog present
- Subjects' mean heart rate during the task was one measure of the effect of stress.

Goal: to compare population means under three different "treatments"

- a *three*-independent-sample problem
- Call the population mean heart rates μ_1 for when pets are present, μ_2 for when friends are present, and μ_3 for when women perform task alone: then
  - $H_0 : \mu_1 = \mu_2 = \mu_3$
  - $H_a : \mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or $\mu_2 \neq \mu_3$
    - not one-sided or 2-sided

SAS descriptive statistics:

```
          Analysis Variable : BEATS

--------------- GROUP=C ---------------------------------

 N          Mean        Std Dev       Minimum       Maximum
-----------------------------------------------------------
15    82.5240667     9.2415747    62.6460000    99.0460000
-----------------------------------------------------------

--------------- GROUP=F ---------------------------------

 N          Mean        Std Dev       Minimum       Maximum
-----------------------------------------------------------
15    91.3251333     8.3411341    76.9080000   102.1540000
-----------------------------------------------------------

--------------- GROUP=P ---------------------------------

N          Mean        Std Dev       Minimum       Maximum

-----------------------------------------------------------
15    73.4830667     9.9698202    58.6920000    97.5380000
-----------------------------------------------------------
```

To infer about the three population means, we *might* use the two-independent-sample t test 3 times:

- Test $H_0 : \mu_1 = \mu_2$ to see if mean heart rate when pet is present differs from mean when friend is present.
- Test $H_0 : \mu_1 = \mu_3$ to see if mean heart rate when pet is present differs from mean when alone.
- Test $H_0 : \mu_2 = \mu_3$ to see if mean heart rate when friend is present differs from mean when alone.

Problem with this approach:

- 3 p-values for 3 different tests don't tell us how likely it is that *three* sample means are spread apart as far as these are.
- might be that $\bar{x}_1 = 73.48$ and $\bar{x}_2 = 91.32$ are significantly different if we look at just 2 groups but *not* significantly different if we know they are the smallest and largest means in 3 groups
  - As more and more groups are considered, we expect gap between smallest and largest sample mean to get larger.
  - (Imagine comparing heights of shortest and tallest person in larger and larger groups of people.)
- the probability of Type I error for the whole set of t-tests will be much bigger than the α level set for each one

Multiple comparisons procedures in statistics

- issue: how to do many comparisons at once with some overall measure of confidence in all our conclusions
- two steps
  - overall test of whether there is good evidence of *any* differences among parameters we wish to compare
  - follow-up analysis to decide which of parameters differ and to estimate size of differences

Step one: One-Way Analysis of Variance (ANOVA)

- step one (overall test) for *some* difference among 3 or more population means
- uses an F *test* to compute a p-value

Dogs, friends, and stress example:

```
                     Analysis of Variance Procedure

                        Class    Levels    Values

                        GROUP         3    C F P

                 Number of observations in data set = 45

                     Analysis of Variance Procedure

Dependent Variable: BEATS
                                Sum of            Mean
Source           DF            Squares          Square  F Value    Pr > F

Model             2          2387.6889920    1193.8444960    14.08    0.0001
Error            42          3561.2994916      84.7928450
Corrected Total  44          5948.9884836

           R-Square              C.V.         Root MSE          BEATS Mean
           0.401360          11.16915        9.2083030           82.444089


Source           DF          Anova SS     Mean Square  F Value    Pr > F

GROUP             2          2387.6889920    1193.8444960    14.08    0.0001
```


Main idea of ANOVA

What matters is how far apart sample means are *relative to variability of individual observations.*

- F statistic

$$F = \frac{\textit{variation among the sample means}}{\textit{variation among individuals in the same sample}}$$

- compare to a cutoff value in an **F distribution**

Notation:

- I = number of different populations whose means we are studying
- n_i = number of observations in sample from ith population
- N = total number of observations in all samples combined

F distributions

- many different F distributions, identified by two parameters
  - numerator degrees of freedom = I - 1
  - denominator degrees of freedom = N - I

Example

Do four varieties of tomato plant differ in mean yield? Agronomists grew 10 plants of each variety and recorded the yield of each plant in pounds of tomatoes.

What are

- the populations of interest
- the variable of interest
- I
- each n_i
- the degrees of freedom for the ANOVA F statistic

Assumptions for One-Way ANOVA

- We have I independent simple random samples, one from each of I populations.
- Each population i has a normal distribution with unknown mean μ_i.
  - As with t-tests, if sample sizes are large enough in each sample, Central Limit Theorem says inference based on sample means is OK even if population distributions are not exactly normal.

- All of the populations have the same standard deviation σ (unknown)
  - unlike t-tests, there is no general procedure when population standard deviations are not assumed to be equal
  - rough rule of thumb: if largest sample standard deviation is no more than twice the smallest sample standard deviation, then population standard deviations probably are close enough to equal that ANOVA procedure is OK

Step two: individual t-tests with correction for multiple comparisons

This is the *follow-up* test.

- should be carried out *only* if the F test from one-way ANOVA is significant at the chosen significance leve.

Goal: to set the *overall* probability of committing a type I error at α when doing pairwise comparisons of k different means

- we will perform $\binom{k}{2}$ two-independent-sample t-tests
- we will conduct each one at the significance level

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$

- This is called the *Bonferroni correction*

  - very conservative

Dogs, friends, and stress example

- There are $k = 3$ samples, so there are $\binom{k}{2} =$ 3 different pairs to compare.
- To get an overall significance level $\alpha = .05$ on all 3 tests considered together, we conduct each one at

$$\alpha^* = \frac{.05}{3} = .0167$$

  - That is, we would consider the difference between two population means to be significantly different from zero at the .05 level only if the p-value for the the t-test for that pair was less than .0167.

  - Equivalently, we could multiply the p-value from each t-test by 3.
    - If the result was less than .05, we would consider the difference between two population means to be significantly different from zero at the .05 level

SAS does the adjusting and prints a grouped list of the classes. Means with the same letter are not significantly different at the specified alpha level.

```
                Analysis of Variance Procedure

          Bonferroni (Dunn) T tests for variable: BEATS

   NOTE: This test controls the type I experimentwise error rat
         generally has a higher type II error rate than REGWQ.

               Alpha= 0.05  df= 42  MSE= 84.79285
                    Critical Value of T= 2.49
             Minimum Significant Difference= 8.3847

   Means with the same letter are not significantly different.

          Bon Grouping              Mean      N  GROUP

                     A            91.325     15  F

                     B            82.524     15  C

                     C            73.483     15  P
```


One-way ANOVA in SAS

```
options linesize = 79 ;

data pet ;
infile '/temp/pet.dat' ;
input group $ beats ;
run ;

proc sort data = pet ;
by group ;
run ;

proc means data = pet ;
by group ;
var beats ;
run ;

proc anova data = pet ;
class group ;
model beats = group ;
run ;

proc anova data = pet ;
class group ;
model beats = group  ;
means group / bon alpha = .05 ;
run ;
```