Wednesday, April 24, 2019

Days of Week	Frequency
Sunday	14
Monday	13
Tuesday	12
Wednesday	15
Thursday	14
Friday	17
Saturday	15

- Warm-up
  - Each observation in a random sample of 100 fatal bicycle accidents was classified according to the day of the week on which the accident occurred. Data consistent with the information on the web site: www.highwaysafety.com are given in the table. Based on these data, is it reasonable to conclude that the proportion of accidents is not the same for all days of the week? Use a significance level of 0.05.

- More uses for the chi-square test

Each observation in a random sample of 100 fatal bicycle accidents was classified according to the day of the week on which the accident occurred. Data consistent with the information on the web site: www.highwaysafety.com are given in the table. Based on these data, is it reasonable to conclude that the proportion of accidents is not the same for all days of the week? Use a significance level of 0.05.

Days of Week	Frequency
Sunday	14 / 14.28
Monday	13 / 14.28
Tuesday	12 / 14.28
Wednesday	15 / 14.28
Thursday	14 / 14.28
Friday	17 / 14.28
Saturday	15 / 14.28
total	100

Counted Data
Random Sample Stated
Expected cell count = 14.28 > 5
100 < 10% of "fatal bicycle accidents" (independent)

χ^2 GOF

H_0: Proportion of accidents are the same for days of the week

H_A: Proportion of accidents is NOT the same for days of the week.

A] Obs	B] Expe
14	14.28
13	14.28
12	14.28
15	⋮
⋮	⋮

df = 6

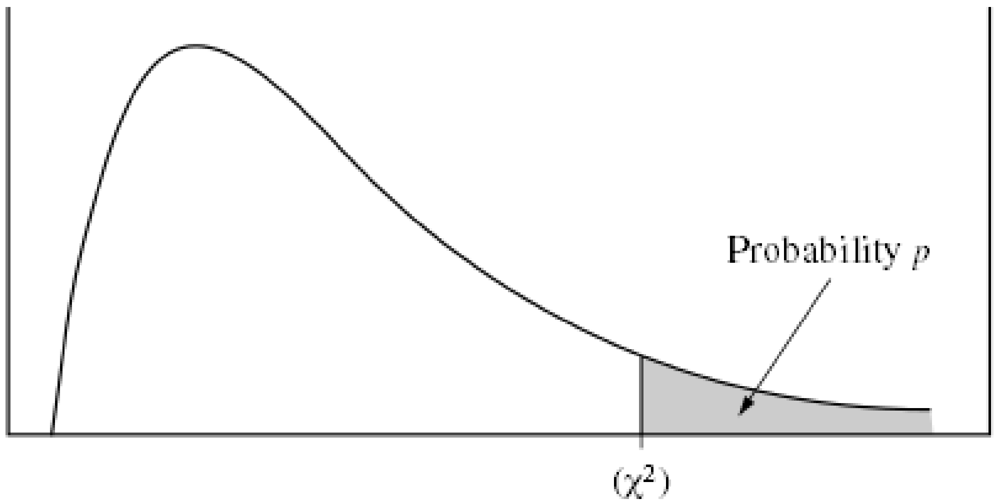
$\chi^2 = 1.08$
df = 6
p-val = 0.98

Due to a p-value of 0.98, which is higher than $\alpha = 0.05$, we fail to reject the null. There is not sufficient statistical evidence that the proportion of accidents is not the same for all days of the week.

p-value

p-value > 0.25

Table entry for p is the point (χ^2) with probability p lying above it.



able C χ^2 critical values

df	Tail probability p .25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87

Objectives

Content Objective: I will use chi-square models to test hypothesis.

Language Objective: I will take clear notes to complement my reading so that I can use chi-square models correctly.

- Social Objective: I will stay focused on the lesson and not distract myself and others.

Comparing Observed Distributions

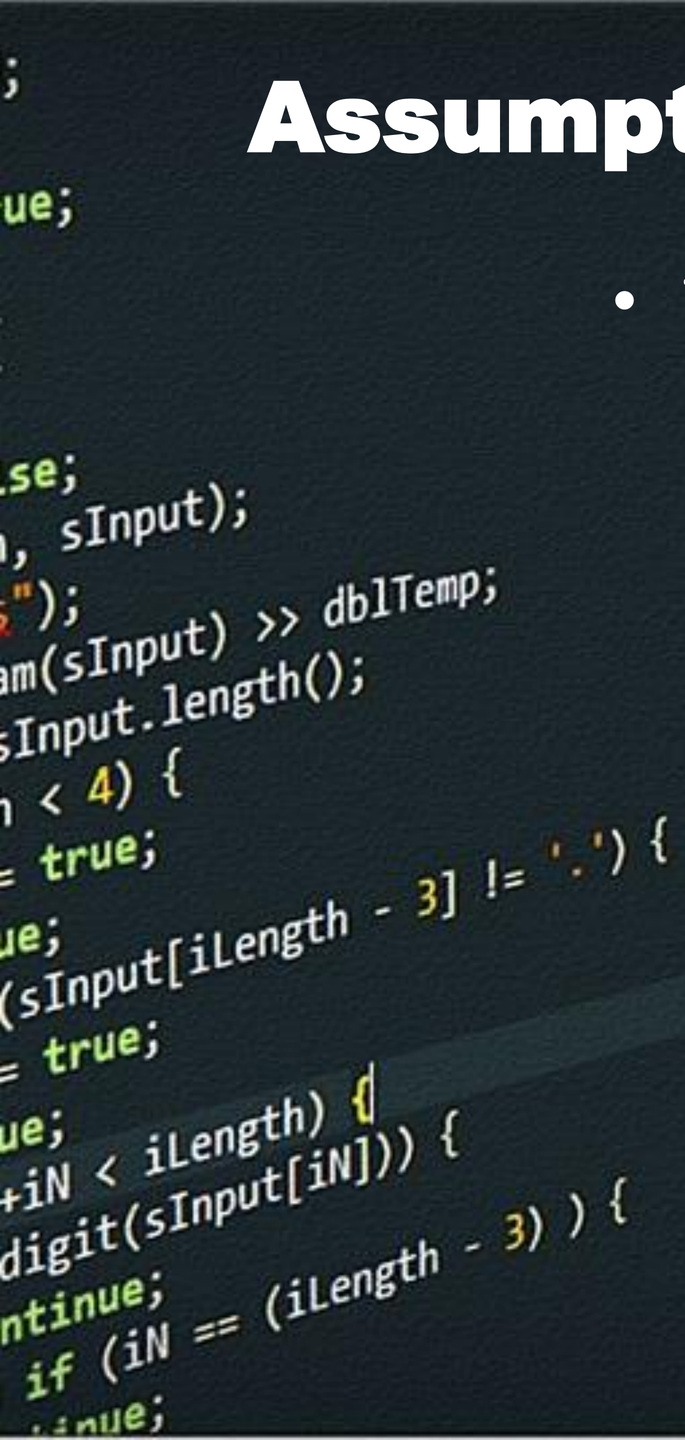
A test comparing the distribution of counts for two or more groups on the same categorical variable is called a **chi-square test of homogeneity**.

A test of homogeneity is actually the generalization of the two-proportion *z*-test.

Comparing Observed Distributions

- The statistic that we calculate for this test is *identical* to the chi-square statistic for goodness-of-fit.
- In this test, however, we ask whether choices are the same among different groups (i.e., there is no model).
- The expected counts are found directly from the data and we have different degrees of freedom.

Assumptions and Conditions



- The assumptions and conditions are the same as for the chi-square goodness-of-fit test:
  - Counted Data Condition: The data must be counts.
  - Randomization Condition and 10% Condition: As long as we don't want to generalize, we don't have to check these conditions.
  - Expected Cell Frequency Condition: The expected count in each cell must be at least 5.

Mechanics

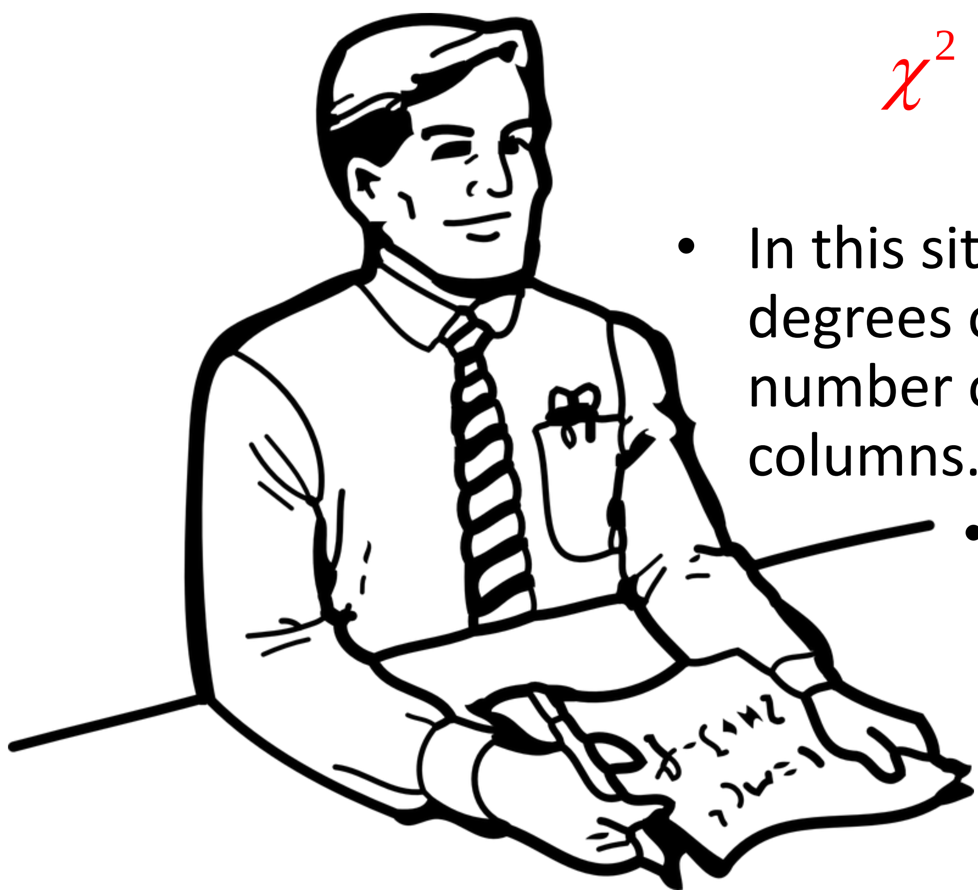
(row total)×(column total)/grand total

- To find the expected counts, we multiply the row total by the column total and divide by the grand total.
- We calculated the chi-square statistic as we did in the goodness-of-fit test:

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}$$

df

- In this situation we have $(R-1)(C-1)$ degrees of freedom, where R is the number of rows and C is the number of columns.
  - We'll need the degrees of freedom to find a P-value for the chi-square statistic.



Practice

Calculator -> Scratch Pad
Menu -> Matrix: Create rows = 5 column = 3

The non profit group Public Agenda conducted telephone interviews with three randomly selected groups of parents of high school children. There were 202 black parents, 202 Hispanic parents, and 201 white parents. One question asked was "Are the high schools in your state doing an excellent, good, fair, or poor job, or don't you know enough to say?" Here are the survey results. What conclusion would you draw about the different groups?

	Black Parents	Hispanic Parents	White Parents	Total
Excellent	12	34	22	68
Good	69	55	81	205
Fair	75	61	60	196
Poor	24	24	24	72
Don't know	22	28	14	64
Total	202	202	201	605

expected $68 \times 202/605$

H_0: The responses are equally distributed among the groups of parents.

H_A: The responses are not equally distributed among groups of parents.

[ ] -> a (ctrl var STO, ENTER)

Menu -> Stat Tests -> χ^2 2 way

$64 \times 201/605$ = 21.26

Counted Data

"Randomly Selected"

605 < 10% of all parents in the state

Lowest expected value > 5

χ^2 Homogeneity

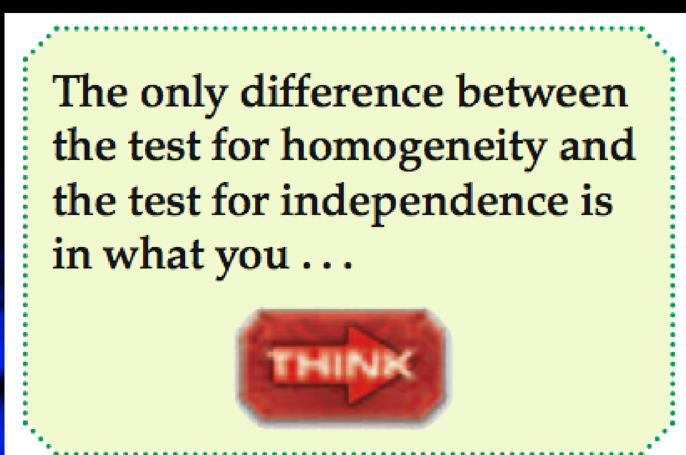
$df = 8$

$\chi^2 = 22.426$

p-value: 0.004

Independence

- **Contingency tables** categorize counts on two (or more) variables so that we can see whether the distribution of counts on one variable is contingent on the other.
- A test of whether the two categorical variables are **independent** examines the distribution of counts for one group of individuals classified according to both variables in a contingency table.



- **chi-square test of independence** uses the same calculation as a test of homogeneity.

Assumptions and Conditions

- We still need counts and enough data so that the expected values are at least 5 in each cell.
- If we're interested in the independence of variables, we usually want to generalize from the data to some population.
  - In that case, we'll need to check that the data are a representative random sample from that population.

Another Example $$$$$$$

The following table was constructed using data from the article "Influence of Socioeconomic Status on Mortality After Stroke" (*Stroke* [2005]: 310-314). One of the questions of interest to the author was whether there was an association between survival after a stroke and level of education. Medical records for a sample of 2333 residents of Vienna, Austria, who had suffered a stroke were used to classify each individual according to two variables – survival (survived, died) and level of education (no basic education, secondary school graduation, technical training/apprenticed, higher secondary school degree, university graduate). Expected cell counts (computed under the assumption of no association between survival and level of education) appear in parenthesis in the table.

	No Basic Education	Secondary School Graduation	Technical Training/ Apprenticed	Higher Secondary School Degree	University Graduate
Died	13 (17.40)	91 (77.18)	196 (182.68)	33 (41.91)	36 (49.82)
Survived	97 (92.60)	397 (410.82)	959 (972.32)	232 (223.09)	279 (265.18)

H_0: There is no association between education & survival.

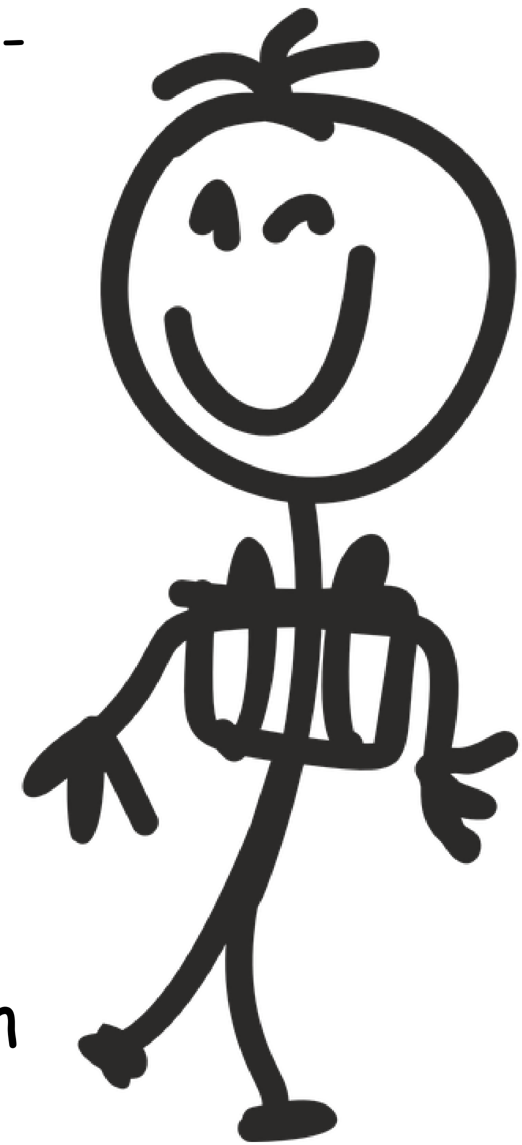
H_A: There is an association between education & survival.

χ^2 test of association (independence)

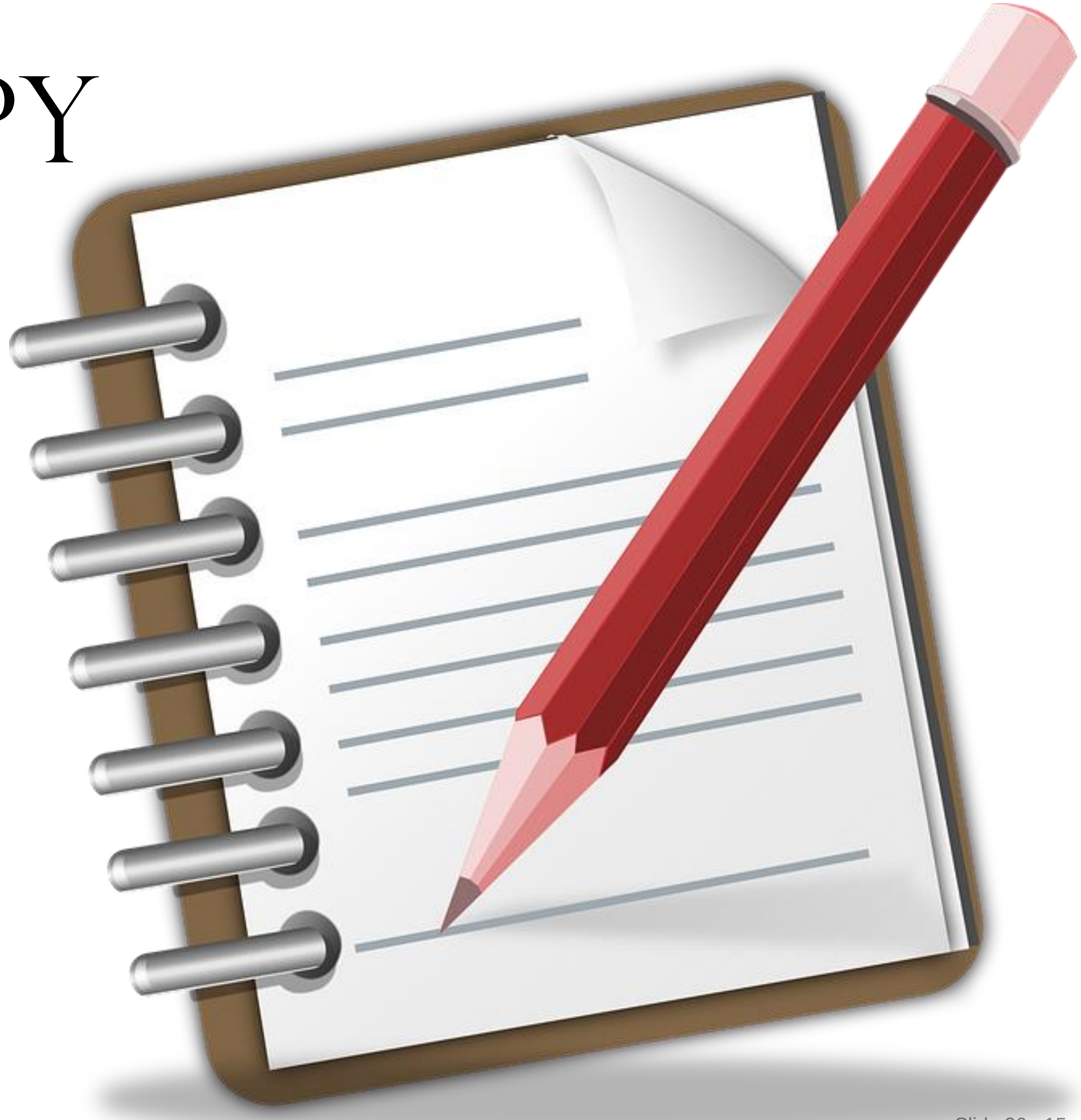
What Can Go Wrong?

- Don’t use chi-square methods unless you have counts.
  - Just because numbers are in a two-way table doesn’t make them suitable for chi-square analysis.
- Beware large samples.
  - With a sufficiently large sample size, a chi-square test can always reject the null hypothesis.
- Don’t say that one variable “depends” on the other just because they’re not independent.
  - Association is not causation.

- We've learned how to test hypotheses about categorical variables.
- All three methods we examined look at counts of data in categories and rely on chi-square models.
  - **Goodness-of-fit tests** compare the observed distribution of a single categorical variable to an expected distribution based on theory or model.
  - **Tests of homogeneity** compare the distribution of several groups for the same categorical variable.
  - **Tests of independence** examine counts from a single group for evidence of an association between two categorical variables.



FRAPPY



Homework:

p 647 (33,34)