Math 3339

Section 27204
MWF 10-11:00am AAAud 2

Bekki George
bekki@math.uh.edu
639 PGH

Office Hours:
M & Th noon – 1:00 pm & T 1:00 – 2:00 pm
and by appointment

Inferences Concerning a Difference Between Population Proportions

Confidence Intervals:

$$\left(\widehat{p_1}-\widehat{p_2}\right) \pm z^* \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

Tests of two population proportions:

The rejection region for a hypothesis testing of two population proportions:

$$H_0: p_1 = p_2 \quad (or\ p_1 - p_2 = \delta)$$

$$H_a: p_1 \neq p_2\ or\ p_1 < p_2\ or\ p_2 > p_2$$

$$z = \frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad where\ \hat{p} = \frac{X+Y}{n_1+n_2}$$

Assumptions:

1. $n_1p_1 \geq 10, n_1(1-p_1) \geq 10, n_2p_2 \geq 10, n_2(1-p_2) \geq 10$

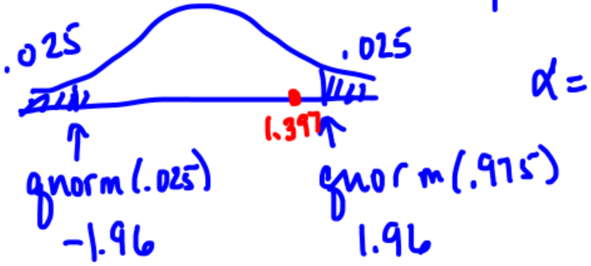
2. two independent samples

Ex: A private ① and a public ② university are located in the same city. For the private university, 1046 ^{n_1} alumni were surveyed and 653 said that they attended at least one class reunion. For the public university, 791 out of 1327 ^{n_2} sampled alumni claimed they have attended at least one class reunion. Is the difference in the sample proportions statistically significant?

$\hat{p}_1 = \frac{653}{1046}$ $\hat{p}_2 = \frac{791}{1327}$

$H_0: p_1 = p_2$

$H_a: p_1 \neq p_2$



$\alpha = .05$

$\hat{p} = \frac{653+791}{1046+1327}$

$$z = \frac{653/1046 - 791/1327}{\sqrt{\frac{1444}{2373}\left(\frac{979}{2373}\right)\left(\frac{1}{1046} + \frac{1}{1327}\right)}} = 1.397$$

pvalue $[1 - \text{pnorm}(1.397)] * 2 = .162 > \alpha$

Fail to reject H_0 + conclude there is no diff in prop. of attendees for class reunion in public vs private univ.

Goodness of Fit Test

Suppose we want to make an inference about a group of data (instead of just one or two). Or maybe we want to test counts of categorical data. **Chi-square** (or χ^2) testing allows us to make such inferences.

There are several types of Chi-square tests but in this section we will focus on the goodness-of-fit test. **Goodness-of-fit** test is used to test how well one sample proportions of categories "match-up" with the known population proportions stated in the null hypothesis statement. The Chi-square goodness-of-fit test extends inference on proportions to more than two proportions by enabling us to determine if a particular population distribution has changed from a specified form.

The null and alternative hypotheses do not lend themselves to symbols, so we will define them with words.

H_o: _____ is the same as _____
H_a: _____ is different from _____

For each problem you will make a table with the following headings:

total "items"

$\% \cdot N$

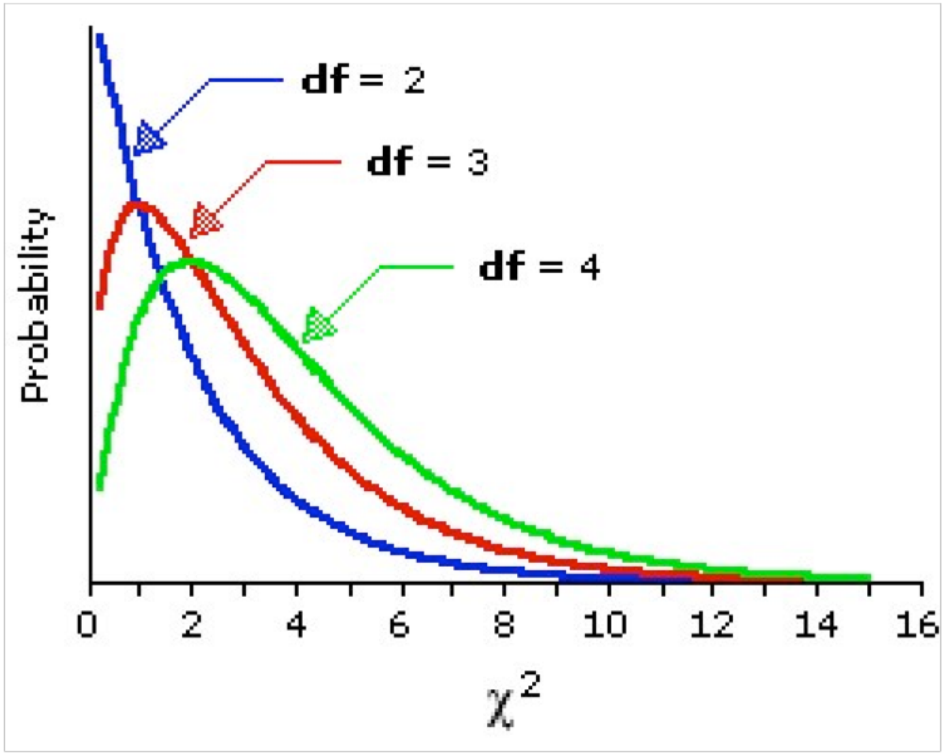
Observed Counts (O)	Expected Counts (E)	$\frac{(O-E)^2}{E}$

The sum of the third column is called the Chi-square test statistic.

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

Our p-values for χ^2 use $\underset{\substack{\uparrow \\ \text{groups}}}{n} - 1$ degrees of freedom.

Chi-square distributions have only positive values and are skewed right. As the degrees of freedom increase it becomes more normal. The total area under the χ^2 curve is 1.



Pvalues:

$P(\chi^2 > \text{test stat})$

The assumptions for a Chi-square goodness-of-fit test are:

2. The sample must be an SRS from the populations of interest.
3. The population size is at least ten times the size of the sample.
4. All expected counts must be at least 5.

To find probabilities for χ^2 distributions:

R-Studio command is: $\underbrace{1 - \text{pchisq}(\textit{test statistic}, \textit{df})}_{\text{pvalue}}$

Examples:

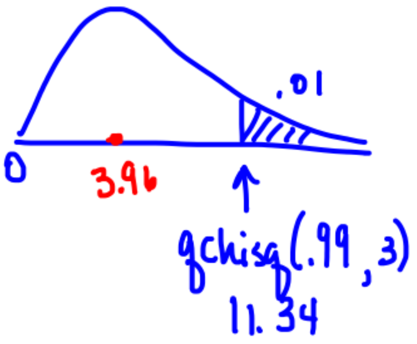
1. The Mixed-Up Nut Company advertises that their nut mix contains (by weight) 40% cashews, 15% Brazil nuts, 20% almonds and only 25% peanuts. The truth-in-advertising investigators took a random sample (of size 50 lbs) of the nut mix and found the distribution to be as follows:

Cashews	*Brazil Nuts*	*Almonds*	*Peanuts*	
15 lb	11 lb	13 lb	11 lb	← OBSERVED
20 lb	7.5 lb.	10 lb.	12.5 lb.	← EXPECTED

At the 1% level of significance, is the claim made by Mixed-Up Nuts true?

H_0: the nuts fit company's stated percentages

H_a: the distribution is different from claim



$$\chi^2 = \frac{(15-20)^2}{20} + \frac{(11-7.5)^2}{7.5} + \frac{(13-10)^2}{10} + \frac{(11-12.5)^2}{12.5} = 3.96$$

pvalue: $P(\chi^2 > 3.96) = 1 - \text{pchisq}(3.96, 3) = .2658 > \alpha$ FRH$_0$

Distribution of nuts fit the company's claim.

2. The community hospital is studying its distribution of patients. A random sample of 321 patients presently in the hospital gave the following information:

df = 5-1 = 4

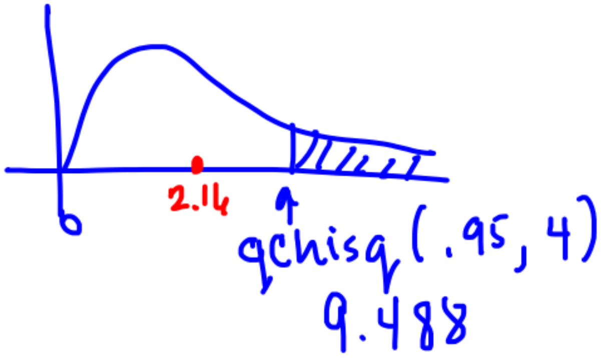
Type of Patient	Old Rate of Occurrences Exp	Present Number of Occurrences OBS.
Maternity Ward	321 · 20% = 64.2	73
Cardiac Ward	321 · 32% = 102.72	94
Burn Ward	321 · 10% = 32.1	30
Children's Ward	321 · 15% = 48.15	50
All Other Wards	321 · 23% = 73.83	74
		321

Test the claim at the 5% significance level that the distribution of patients in these wards has not changed.

α

H_0: patient occurance same as old rate

H_a: patient distribution is different from old rate



$$\chi^2 = \frac{(73-64.2)^2}{64.2} + \frac{(94-102.72)^2}{102.72} + \frac{(30-32.1)^2}{32.1} + \frac{(50-48.15)^2}{48.15} + \frac{(74-73.83)^2}{73.83} = 2.16$$

pvalue: 1-pchisq(2.16, 4) = .7 > α FRH_0

Inference for Two-way Tables

We can also use the Chi-Square method to make inferences for data in two-way tables.

The formula to find the expected count in a two-way table is:

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

Where *n* is the grand total of all values.

When conducting a Chi-square test of independence in a two-way table, the null and alternate hypothesis will be:

$H_{0:}$ There is no association between the row and column variables.
H_a: There is an association between the two variables.

The test statistic is:

$$\chi^2 = \sum_{allcells} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The *p*-value: $P(\chi_k^2 > \chi^2)$, where χ_k^2 represents a Chi-square distribution with $df = (r - 1)(c - 1)$ degrees of freedom.

The assumptions necessary for the test to be valid are:

1. The observations constitutes a simple random sample from the population of interest, and
2. The expected counts are at least 5 for each cell of the table.

By itself, the chi-square test determines only whether the data provide evidence of a relationship between the two variables. If the result is significant, one can go on to identify the source of that relationship by finding the cells of the table that contribute most to the χ^2 value (i.e. those cells with the biggest discrepancy between the observed and expected counts) and by noting whether the observed count falls above or below the observed count in those cells.

Example:

3. Use the data below to determine if there is sufficient evidence to conclude that an association exists between car color and the likelihood of being in an accident.

2 rows
3 cols

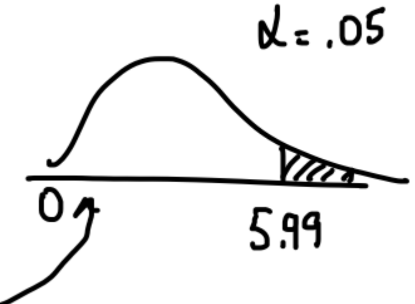
	Red	Blue	White	
Car has been in accident	28	33	36	97
Car has not been in accident	23	22	30	75
	51	55	66	$\boxed{172}$

$df = (2-1)(3-1) = 2$

	Red	Blue	White
acc	$\frac{97 \cdot 51}{172} = 28.8$	$\frac{97 \cdot 55}{172} = 31.0$	$\frac{97 \cdot 66}{172} = 37.2$
no acc	$\frac{75 \cdot 51}{172} = 22.2$	$\frac{75 \cdot 55}{172} = 24.0$	$\frac{75 \cdot 66}{172} = 28.8$

Exp

$$\chi^2 = \frac{(28-28.8)^2}{28.8} + \cdots + \frac{(30-28.8)^2}{28.8} \approx .43$$



pvalue $P(\chi^2 > .43) = 1 - \text{pchisq}(.43, 2) \approx .81 > .05$ (FRH_0)

Fail to reject the null hyp. + conclude there is no association between car color and if in accident.

Popper 29

There are 4 TV sets in the student center of a large university. At a particular time each day, four different soap operas (1, 2, 3 and 4) are viewed on these TV sets. It is believed that the percentages of the audience captured by these shows are 25%, 30%, 25%, and 20%, respectively. 300 students are surveyed and the summary of their response follows:

	1	2	3	4
Observed	80	88	79	53

	1	2	3	4
EXP				

1. What are the expected counts for this distribution?
   a. 25, 30, 25, 20
   [✓ b.] 75, 90, 75, 60
   c. 50, 60, 50, 40
   d. cannot be determined

2. What is the degrees of freedom for this problem?
   a. 2
   [✓ b.] 3
   c. 4
   d. none of these

3. What is the value of the test statistic?
   a. 94
   b. 0.3133
   [✓ c.] 1.408
   d. 1.485

```
> qchisq(.95,3)
[1] 7.814728
```

4. Using $\alpha = .05$
   [✓ a.] FRHo
   b. RHo

5. E

4. Do these observed data fit the belief of percentages who watch each show ($\alpha = 0.05$)?

   a) Yes
   b) No

5. Choose E.