BTRY 6010: Statistical Methods I

PRELIM 2 REVIEW SESSION: 4:55 PM – 6:10 PM

TA: KIM HOCHSTEDLER (SHE/HER)
NOVEMBER 17, 2020

WELCOME!

TODAY'S TOPIC: CATEGORICAL DATA ANALYSIS (CHI-SQUARED TESTS)

CALENDAR REMINDER

Tuesday, 11/17 @ 8:00 am EST: Inference on proportions with Dave

Tuesday, 11/17 @ 3:00 pm EST: Inference on means with Indra

Tuesday, 11/17 @ 4:55 pm EST: Categorical data analysis with Kim

Wednesday, 11/18 @ 3:00 pm EST: ANOVA with Steve

Wednesday, 11/18 @ 5:00 pm EST: OH with Dave

Thursday, 11/19 @ 8:00 am EST: Non-parametric inference and regression with Sumanta

Thursday, 11/19 @ 5:00 pm EST: OH with Sumanta

Friday, 11/20 @ 5:00 pm EST: OH with Kim

Sunday, 11/22 @ 7:30 pm EST: Prelim 2

TODAY'S SESSION

1. Overview and example of chi-square test for goodness-of-fit
2. Overview and example of chi-square test for independence
3. Question and answer time!

Please turn your camera on if you can and you are comfortable with it

Be ready to chat in your questions when we reach the Q&A time

Chi-squared tests of goodness-of-fit

OVERVIEW OF CHI-SQUARED TEST OF GOODNESS-OF-FIT

Ex. jury problems

When is this test appropriate?

We want to know if the **distribution of a variable in our sample** matches what we would expect from the **distribution of the population**.

given

General hypotheses

Null hypothesis: $p_1 = x, p_2 = y, \ldots, p_k = z$

Alternative hypothesis: At least one of the probabilities differs from those listed in the null hypothesis.

p_i = proportion of some group in our sample

$x, y, \ldots z$ = known population proportions

OVERVIEW OF CHI-SQUARED TEST OF GOODNESS-OF-FIT

Test Statistic

observed in our sample

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{H_0}{\sim} \chi^2_{df}$$

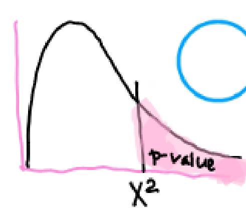
Always list the degrees of freedom

Expected number

Degrees of freedom = $k-1$, number of groups -1

P-value determined by:

pchisq(X^2, df, lowertail = F)



OVERVIEW OF CHI-SQUARED TEST OF GOODNESS-OF-FIT

Assumptions

1. Independent observations

Check: Simple random sample

2. Expected cell counts of at least 5

Check: $n \cdot p_i$

Group A	B	C
%	%	%

Observed $n=100$

Population: 33%, 33%, 34%
33, 33, 34

PROBLEM #1: SET-UP

In 2018, the city of Ithaca took a census where they asked residents which ice cream shop was their favorite: Purity, Sweet Melissa's, or Cornell Dairy Bar. The results of the census were as follows.

- 56% of Ithacans preferred Purity.
- 25% of Ithacans preferred Cornell Dairy Bar.
- 19% of Ithacans preferred Sweet Melissa's.

In 2019, Cornell asked a simple random sample of 250 graduate students to answer the same question. They found the following results.

- 102 Cornell graduate students preferred Purity.
- 70 Cornell graduate students preferred Cornell Diary Bar.
- 78 Cornell graduate students preferred Sweet Melissa's.

Question: Are the ice cream shop preferences of Cornell Graduate Students different from the population of Ithacans?

PROBLEM #1: HYPOTHESES

In 2018, the city of Ithaca took a census where they asked residents which ice cream shop was their favorite: Purity, Sweet Melissa's, or Cornell Dairy Bar. The results of the census were as follows.

- 56% of Ithacans preferred Purity.
- 25% of Ithacans preferred Cornell Dairy Bar.
- 19% of Ithacans preferred Sweet Melissa's.

p_1 = probability that a Cornell grad student prefers Purity

Question: Are the ice cream shop preferences of Cornell Graduate Students different from the population of Ithacans?

p_2 = Cornell Dairy Bar p_3 = Sweet Melissa's

Null hypothesis:

$p_1 = .56$, $p_2 = .25$, $p^3 = .19$

Alternative hypothesis:

Any of the probabilities differ from those listed in the null hypothesis.

PROBLEM #1: TEST

Question: Are the ice cream shop preferences of Cornell Graduate Students different from the population of Ithacans?

Conclusion: $\alpha = 0.05$

```
# Run a chi-square goodness-of-fit test.
cornell_students_icecream <- c(102, 70, 78)
ithacans_icecream <- c(.56, .25, .19)

chisq.test(x = cornell_students_icecream,
           p = ithacans_icecream)
```

sample data (#'s)

population proportions

```
##
##  Chi-squared test for given probabilities
##
## data:  cornell_students_icecream
## X-squared = 30.798, df = 2, p-value = 2.052e-07
```

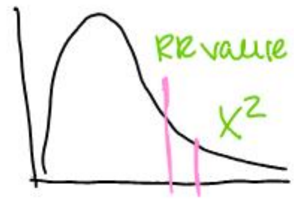
p-value < .05, reject the null

X^2 $k-1$

PROBLEM #1: TEST

Question: Are the ice cream shop preferences of Cornell Graduate Students different from the population of Ithacans?

Rejection region




```
##  Chi-squared test for given probabilities
##
## data:  cornell_students_icecream
## X-squared = 30.798, df = 2, p-value = 2.052e-07
```

Interpretation:

Since the p-value was $< .05 = \alpha$, at the 5% level of significance we have sufficient evidence to conclude that Cornell grad students have a different distribution of ice cream preferences than that given in the 2018 census of all residents of Ithaca.

PROBLEM #1: ASSUMPTIONS

Question: Are the ice cream shop preferences of Cornell Graduate Students different from the population of Ithacans?

1. Independent observations

Check: SRS of Cornell grad students ✓

2. Expected cell count of at least 5

Check:

$n = 250$, sample size

$p_1 = .56 \quad p_2 = .25 \quad p_3 = .19$

$\times 250 \quad \times 250 \quad \times 250$

$140 \geq 5 \quad 62.5 \geq 5 \quad 47.5 \geq 5$ ✓

Chi-squared tests of independence

OVERVIEW OF CHI-SQUARED TEST OF INDEPENDENCE

When is this test appropriate?

We want to know if **two responses/measures/variables** are **independent** of one another in a **single population**.

- Statistical independence occurs when an observed response of one variable does not depend on the response of another variable.

$P(A|B) = P(A)$

General hypotheses

Null hypothesis: There is no association between the variables (they are *independent*).

Alternative hypothesis: The variables are associated.

OVERVIEW OF CHI-SQUARED TEST OF INDEPENDENCE

Test Statistic

$$X^2 = \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{c}}_{\text{row and column}} \frac{(\overset{n_{ij}}{\overbrace{n_{ij}}}-E_{ij})^2}{E_{ij}} \overset{H_0}{\sim} \chi^2_{df}$$

Degrees of freedom = $(r-1)(c-1)$, r = rows, c = columns

P-value determined by:

same as χ^2 GOF

OVERVIEW OF CHI-SQUARED TEST OF INDEPENDENCE

Assumptions

1. independent observations

Check: SRS

2. expected cell counts ≥ 5

Check:

	Europe	Asia	Africa	
Often	*	·	·	A
Sometimes	·			
Never	·			
B				N

$$\frac{\text{rowsum} \times \text{column sum}}{\text{total}}$$

$$\frac{A \times B}{N} = * \geq 5$$

PROBLEM #2: SET-UP

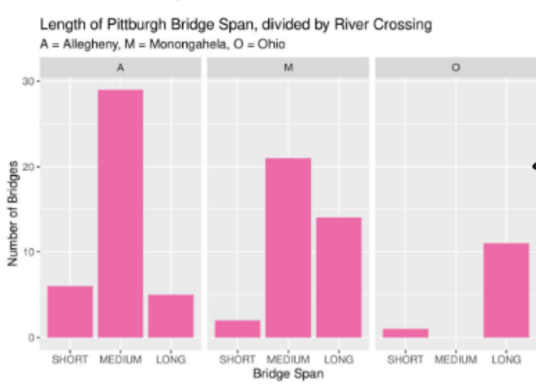
We have data on a simple random sample of 89 bridges in Pittsburgh. The following columns are present in the dataset, with one observation per bridge.

- **River:** which of 3 rivers the bridge crosses (M = Monongahela, A = Allegheny, or O = Ohio)
- **Purpose:** bridge purpose, including 3 categories ("Highway", "Aquduct", or "RR" = Railroad)
- **Material:** primary type of material the bridge is made of ("wood", "iron", or "steel")
- **Span:** categorical length of bridge span ("short", "medium", or "long")

Question: Are the span of the bridge and the river over which the bridge crosses independent? Provide statistical evidence.

PROBLEM #2: SET-UP

Question: Are the span of the bridge and the river over which the bridge crosses independent? Provide statistical evidence.



River

SPAN	Allegheny	Monongahela	Ohio
SHORT	6	2	1
MEDIUM	29	21	0
LONG	5	14	11

PROBLEM #2: HYPOTHESES

Question: Are the span of the bridge and the river over which the bridge crosses independent? Provide statistical evidence.

Null hypothesis:

There is no association between bridge span and river

Alternative hypothesis:

The bridge span and the river are associated.

PROBLEM #2: TEST

Question: Are the span of the bridge and the river over which the bridge crosses independent? Provide statistical evidence.

M, L, S

M, A, O

```
# Run a chi-square test of independence.
chisq.test(x = pitt_bridges$span, y = pitt_bridges$river)
```

```
##  Pearson's Chi-squared test
##
## data:  pitt_bridges$span and pitt_bridges$river
## X-squared = 27.917, df = 4, p-value = 1.297e-05
```

Conclusion: p-value < .05

Reject the null hypothesis

PROBLEM #2: TEST

Question: Are the span of the bridge and the river over which the bridge crosses independent? Provide statistical evidence.

```
##  Pearson's Chi-squared test
##
## data:  pitt_bridges$span and pitt_bridges$river
## X-squared = 27.917, df = 4, p-value = 1.297e-05
```

Interpretation: Since the p-value < .05, at the 5% significance level, we have sufficient evidence of an association between

variable 1 variable 2

bridge span and the river it was built over for bridges in Pittsburgh

setting

PROBLEM #2: ASSUMPTIONS

Question: Are the span of the bridge and the river over which the bridge crosses independent? Provide statistical evidence.

1. Independent observations

Check: SRS

2. Expected cell count of at least 5

Check:

	Short	Medium	Long	Total
Allegheny	4	22.5	13.5	40
Monongahela	3.7	20.8	12.5	37
Ohio	1.2	6.7	4.0	12
Total	9	50	30	89

$$\frac{40 \cdot 9}{89} = 4$$