Statistical Inference

Inference for Categorical Variables

Behnam Bahrak
Spring 2020

Jury Selection

- In a county where jury selection is supposed to be random, a civil rights group sues the county, claiming racial disparities in jury selection.
- Distribution of ethnicities of the people in the county who are eligible for jury duty (based on census results):

ethnicity	white	black	nat. amer.	asian & PI	other
% in population	80.29%	12.06%	0.79%	2.92%	3.94%

- Distribution of 2500 people who were selected for jury duty the previous year:

ethnicity	white	black	nat. amer.	asian & PI	other
% in population	1920	347	19	84	130

Jury Selection

- The court retains you as an independent expert to assess the statistical evidence that there was discrimination. You propose to formulate the issue as an hypothesis test.

H_0 (nothing going on): People selected for jury duty are a simple random sample from the population of potential jurors. The observed counts of jurors from various race/ ethnicities follow the same ethnicity distribution in the population.

H_A (something going on): People selected for jury duty are not a simple random sample from the population of potential jurors. The observed counts of jurors from various ethnicities do not follow the same race/ethnicity distribution in the population.

Evaluating the Hypotheses

- Quantify how different the observed counts are from the expected counts
- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- Called a goodness of fit test since we're evaluating how well the observed data **fit** the expected distribution

Example

➢ Calculate expected number of jurors from each ethnicity if in fact the jury selection is random.

ethnicity	white	black	nat. amer.	asian & PI	other	total
% in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007 +	302 +	20 +	73 +	98 $\stackrel{?}{=}$	2500 ✓

2500×0.8029

2500×0.1206

Example

ethnicity	white	black	nat. amer.	asian & PI	other	total
% in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007	302	20	73	98	2500
observed #	1920	347	19	84	130	

observed < expected

observed > expected

Conditions for the Chi-square Test

1. ***Independence:*** Sampled observations must be independent.
    - random sample/assignment
    - if sampling without replacement, $n < 10\%$ of population
    - each case only contributes to one cell in the table

2. ***Sample size:*** Each particular scenario (i.e. cell) must have at least 5 expected cases.

Anatomy of a Test Statistic

General form of a test statistic $\frac{point\ estimate - null\ value}{SE\ of\ point\ estimate}$

1. Identifying the difference between a point estimate and an expected value if the null hypothesis were true

2. Standardizing that difference using the standard error of the point estimate

Chi-square Statistic

- When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the chi-square (χ^2) statistic.

χ^2 **statistic:** $$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

O : observed
E : expected
k : number of cells

Why square?

- Positive standardized difference
- Highly unusual differences between observed and expected will appear even more unusual
- Ease of mathematical calculations

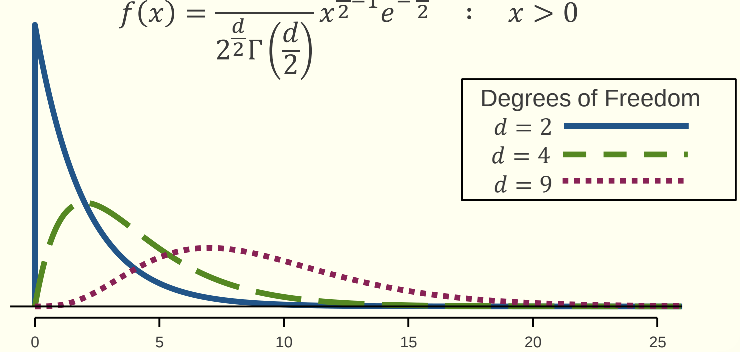
Degrees of Freedom

- To determine if the calculated χ^2 statistic is considered unusually high or not we need to first describe its distribution
- Chi-square distribution has just one parameter:
  - Degrees of freedom (df): influences the shape, center, and spread

χ^2 degrees of freedom for a goodness of fit test:	$df = k - 1$ k : number of cells

Chi-square distribution & degrees of freedom

$$f(x) = \frac{1}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)} x^{\frac{d}{2}-1} e^{-\frac{x}{2}} \quad : \quad x > 0$$



Example

ethnicity	white	black	nat. amer.	asian & PI	other	total
% in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007	302	20	73	98	2500
observed #	1920	347	19	84	130	2500

H_0: The observed counts of jurors from various race/ ethnicities follow the same ethnicity distribution in the population.

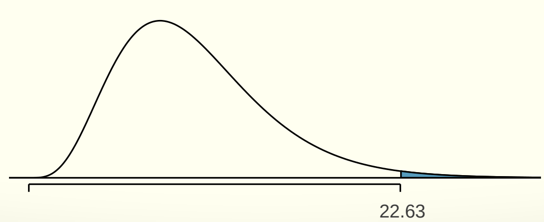
H_A: The observed counts of jurors from various ethnicities do not follow the same race/ethnicity distribution in the population.

$$\chi^2 = \frac{(1920-2007)^2}{2007} + \frac{(347-302)^2}{302} + \frac{(19-20)^2}{20} + \frac{(84-73)^2}{73} + \frac{(130-98)^2}{98} = 22.63$$

$$df = k - 1 = 5 - 1 = 4$$

p-value

- P-value for a chi-square test is defined as the tail area above the calculated test statistic
- Because the test statistic is always positive, and a higher test statistic means a higher deviation from the null hypothesis

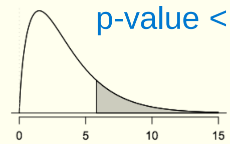


p-value

$$\chi^2 = 22.63$$
$$df = 4$$

Chi-square probability table

p-value < 0.001



Using the table:

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Using R:

```
R
> pchisq(22.63, 4, lower.tail = FALSE)
[1] 0.0002
```


Example: Game of Thrones!

- There may not be a situation more perilous than being a character on Game of Thrones!
- So what do all these gruesome deaths have to do with statistics?
- They are data that come from a Poisson distribution.

Number of deaths	0	1	2	3	4	5	6	≥7	total
episodes	5	10	9	8	10	8	2	5	57

Expected number of death

- We model the number of deaths in each episode of GoT with a Poisson distribution with $\lambda = 3.22807$:

$P(X=0) = e^{-\lambda}\frac{\lambda^0}{0!} = 0.039634$ → Expected = $57 \times P(X=0) = 2.2591$

$P(X=1) = e^{-\lambda}\frac{\lambda^1}{1!} = 0.127941$ → Expected = $57 \times P(X=1) = 7.2926$

⋮

$P(X=6) = e^{-\lambda}\frac{\lambda^6}{6!} = 0.062286$ → Expected = $57 \times P(X=6) = 3.5503$

$P(X \geq 7) = 1 - P(X=0) - \cdots - P(X=6) = 0.046346$
→ Expected = $57 \times P(X \geq 7) = 2.6417$

Goodness of Fit

Number of deaths	0	1	2	3	4	5	6	≥7	total
observed	5	10	9	8	10	8	2	5	57
expected	2.259	7.293	11.77	12.67	10.22	6.6	3.55	2.64	57

H_0: The observed counts of deaths follow a Poisson distribution with λ =3.22807

H_A: The observed counts of deaths **do not** follow a Poisson distribution with λ =3.22807

$$\chi^2 = \frac{(5-2.259)^2}{2.259} + \frac{(10-7.293)^2}{7.293} + \cdots + \frac{(2-3.55)^2}{3.55} + \frac{(5-2.64)^2}{2.64} = 9.792$$

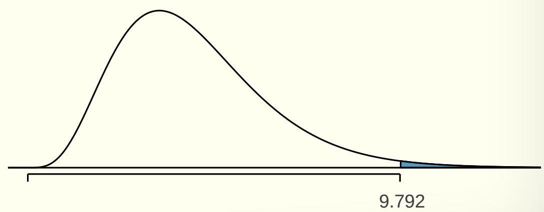
$$df = k - 1 = 8 - 1 = 7$$

Goodness of Fit

```
R
> pchisq(9.792, 7, lower.tail = FALSE)
[1] 0.2006703
```

p-value = 0.2 > 0.05

Thus we fail to reject H_0



Chi-square Independence Test

- In a study, students in grades 4-6 (age 10-12) were asked whether good grades, athletic ability, or popularity was most important to them.
- A two-way table separating the students by age and by choice of most important factor is shown below. Do these data provide evidence to suggest that goals vary by age?

	Grades	Popularity	Sports	Total
10 yrs old	63	31	25	**119**
11 yrs old	88	55	33	**176**
12 yrs old	96	55	32	**183**
Total	**247**	**141**	**90**	**478**

Hypotheses

H_0 (nothing going on):
- Age and goals are independent.
- Goals do not vary by age.

H_A (something going on):
- Age and goals are dependent.
- Goals vary by age.

Evaluating the Hypotheses

- Quantify how different the observed counts are from the expected counts
- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- Called an independence test since we're evaluating the relationship between two categorical variables

Chi-square test of independence

χ^2 **test of independence:** $$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

$$df = (R-1) \times (C-1)$$

O : Observed

E : Expected

R : number of rows

C : number of columns

k : number of cells

- The p-value is the area under the χ^2_{df} curve, above the calculated test statistic.

Conditions for the chi-square test

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, $n < 10\%$ of population
   - each case only contributes to one cell in the table

2. ***Sample size:*** Each particular scenario (i.e. cell) must have at least 5 expected cases.

Expected Counts

	Grades	Popularity	Sports	Total
10 yrs old	63	31	25	**119**
11 yrs old	88	55	33	**176**
12 yrs old	96	55	32	**183**
Total	**247**	**141**	**90**	**478**

Expected counts in two-way tables:

$$Expected\ Count = \frac{(row\ total) \times (column\ total)}{table\ total}$$

$$E_{row\ 1,\ col\ 1} = \frac{119 \times 247}{478} = 61$$

Calculating the test statistic in two-way tables

	Grades	Popularity	Sports	Total
10 yrs old	63 (61)	31 (35)	25 (23)	**119**
11 yrs old	88 (91)	55 (52)	33 (33)	**176**
12 yrs old	96 (95)	55 (54)	32 (34)	**183**
Total	**247**	**141**	**90**	**478**

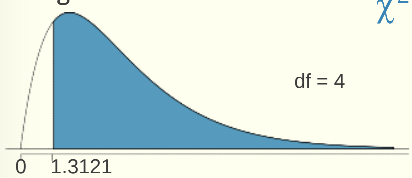
$$\chi^2 = \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \ldots + \frac{(32-34)^2}{34} = 1.3121$$

$$df = (R-1) \times (C-1) = 2 \times 2 = 4$$

Decision Making

- Test the hypothesis that age and goals are associated at the 5% significance level.

$$\chi^2 = 1.3121 \qquad df = 4$$



p-value $>$ 0.3

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52

Decision Making

R

```
> pchisq(1.312, 4, lower.tail = FALSE)
[1] 0.8593193
```

- Since p-value is high, we fail to reject H_0.
- The data do not provide convincing evidence that age and goals are dependent.
- It doesn't appear that goals vary by age.

Chi-square Tests

- goodness of fit: comparing the distribution of one categorical variable (with more than 2 levels) to a hypothesized distribution
- independence: evaluating the relationship between two categorical variables (at least one with more than 2 levels)

Behnam Bahrak
bahrak@ut.ac.ir