Categorical variables inference

Inference for p

The **point estimate** for a sample of size n from a population with a true proportion p, is

$$\hat{p} = \frac{\#\text{ of ``successes''}}{\#\text{ of cases}}$$

Inference for p

What is the sampling distribution of $\hat{p}$?

Inference for p

Conditions for the sampling distribution of $\hat{p}$ being nearly normal

1. the sample observations are independent and
2. successes-failure condition
   - $np \geq 10$,
   - $n(1 - p) \geq 10$.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Inference for p

Confidence interval:

$$\hat{p} \;\pm\; z^{\star} SE$$

Hypothesis testing

H_0: $p = p_0$.

$$H_A: \begin{cases} p > p_0 & \text{(upper-tail alternative)} \\ p \neq p_0 & \text{(two-tailed alternative)} \\ p < p_0 & \text{(lower-tail alternative)} \end{cases}$$

Test statistic: $z = \frac{\hat{p} - p_0}{SE}$
We reject H_0 when:

- $P(Z > z) < \alpha$ (upper-tail alternative)
- $P(|Z| > z) < \alpha$ (two-tailed alternative)
- $P(Z < -z) < \alpha$ (lower-tail alternative)

Inference for groups

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal

- each proportion separately follows a normal model
- the two samples are independent of each other.

The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where p_1 and p_2 represent the population proportions, and n_1 and n_2 represent the sample sizes.

Inference for groups

Confidence interval:

$$\hat{p}_1 - \hat{p}_2 \pm \; z^{\star} SE$$

Inference for groups

Hypothesis testing

$$H_0 : p_1 = p_2$$

Which value should we select for the SE?

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Inference for groups

Pooled estimate of a proportion For the null hypothesis $p_1 = p_2$, use the **pooled estimate** of the shared proportion:

$$\hat{p} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2},$$

where $\hat{p}_1 n_1$ represents the number of successes in sample 1

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

Inference for groups

Hypothesis testing

H_0: $p_1 = p_2$.

$$H_A: \begin{cases} p_1 > p_2 & \text{(upper-tail alternative)} \\ p_1 \neq p_2 & \text{(two-tailed alternative)} \\ p_1 < p_1 & \text{(lower-tail alternative)} \end{cases}$$

Test statistic: $z = \frac{\hat{p}}{SE}$,
where

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

Chi-Square Tests

Number of juror-Racial bias

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Q: Are the jurors racially representative of the population?

Chi-Square Tests

Testing for goodness of fit using chi-square

- H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.
- H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

Chi-Square Tests

Hypothesis testing (up to this point):

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

Chi-Square Tests

1. **Evaluate the Expected counts**

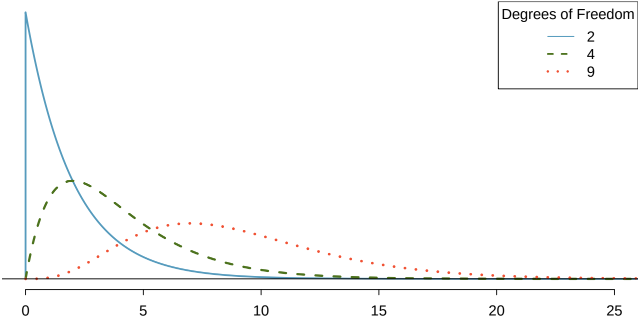
Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

2. Calculate the X^2-statistic

$$\begin{aligned} X^2 &= \frac{(\text{observed count}_1 - \text{expected count}_1)^2}{\text{expected count}_1} + \ldots \\ &+ \frac{(\text{observed count}_4 - \text{expected count}_4)^2}{\text{expected count}_4} \\ &= 5.89 \end{aligned}$$

Chi-Square Tests

$$X^2 \sim \text{Chi-square distribution}$$



Degrees of freedom: $df = k - 1$, where k is the number of bins.

Chi-Square Tests

- $X^2 = 5.89$
- $df = 3 - 1 = 2$
- $p - value =$
- The data do not provide convincing evidence of racial bias in the juror selection.

Chi-Square distributions

Chi-square test

- k categories
- Observed counts O_1, O_2, ..., O_k
- Expected counts E_1, E_2, ..., E_k

Test statistic

$$x^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k},$$

where X^2 follows the Chi-square distribution with $df = k - 1$.
$p - value = P(X^2 > x^2)$

Chi-square Test

Conditions

- Independence.
- Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.
- Degrees of freedom: $df \geq 2$

Chi-Square tests

Google Experiment

- Test three algorithms using a sample of 10,000 google.com search queries.
- Breakdown of test subjects into three search groups.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

Chi-Square tests

Q: Do the results align with the user's interest?

Quantifying the result

1. The user clicked one of the links provided and did not try a new search or
2. The user performed a related search.

Chi-Square tests

Results of the Google search algorithm experiment.

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

Chi-Square tests

Testing for independence

H_0: The algorithms each perform equally well.

H_A: The algorithms do not perform equally well.

Chi-Square tests

- $r \times c$ categories, r and c are number of rows and columns respectively
- Observed counts O_1, O_2, ..., $O_{r \cdot c}$
- Expected counts E_1, E_2, ..., $E_{r \cdot c}$

Test statistic

$$x^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_{r \cdot c} - E_{r \cdot c})^2}{E_{r \cdot c}},$$

where X^2 follows the Chi-square distribution with $df = (r-1)(c-1)$.

$p - value = P(X^2 > x^2)$

Chi-Square tests

Computing expected counts in a two-way table

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

Chi-Square tests

The observed counts and the (expected counts).

Search algorithm	current		test 1		test 2		Total
No new search	3511	(3539)	1749	(1769.5)	1818	(1769.5)	7078
New search	1489	(1461)	751	(730.5)	682	(730.5)	2922
Total	5000		2500		2500		10000

- $X^2 = 6.120$
- $df = (2-1) \times (3-1) = 2$
- $p - value < 0.05$
- The data provide convincing evidence that there is some difference in performance among the algorithms.