Chapter 6. Inference for categorical data

6.1 Inference for a single proportion

Z-score calculators
1) https://thepercentagecalculator.net/Zscore/Table/plus/Z-score-0.html
2) https://www.calculator.net/z-score-calculator

Normality assumption

- Sample proportion p̂ can be assumed to be nearly normal when conditions are met:
    - Sample observations are independent (e.g., simple random sample)
    - Sample size is sufficiently large (at least 10 successes and 10 failures in the sample following, again, the '≥ 10 rule'; i.e., **success-failure condition**)
- When these are true, sample distribution of p̂ is nearly normal with:
    - Mean of p
    - Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}.$
    - For hypothesis tests, typically the null value (proportion claimed in H0 is used in place of p

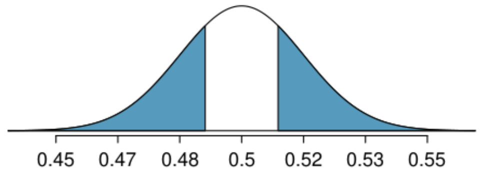
Confidence interval

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

- Z-score for a 95% confidence interval is 1.96 (look up using z-score calculator)
- 4 steps: prepare, check, calculate (compute SE using p̂, find z*, generate CI interval), conclude (interpret confidence interval in context of problem).
- E.g. interpretation: "We are 95% confident that the true proportion of … in context of …. was between … and …"

1-proportion hypothesis test

- 4 steps: prepare, check, calculate (compute SE using p0, Z-score and identify p-value using z-score calculator), conclude (compare p-value to α, interpret in context of problem).
- Diagram like this is helpful for computing p-value



Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.51 - 0.50}{0.017} = 0.59$$

- E.g. interpretation: “The poll does not provide convincing evidence that a proportion of …. support or oppose …”

When normality condition(s) aren’t met:

- When success-failure condition isn’t met:
  - For a hypothesis test: simulate null distribution of $\hat{p}$ using the null value, p_0 (see strategy in Section 2.3)
  - For a confidence interval: use Clopper-Pearson interval (beyond scope)
- When independence condition isn’t met:
  - Important to understand how and why but special methods (beyond scope) may need to be used for cluster samples, for example
  - Inherent biases of data from a convenience sample may never be correctable

Sample size calculation

- Large enough n so that margin of error is sufficiently small and the sample is useful
- E.g., How big of a sample is required to ensure the margin of error is smaller than 0.04 of the actual proportion using a 95% confidence interval?
- Remember the confidence interval formula: $\hat{p} \pm z^{\star} \times SE$
- Margin of error: $z^{\star}\sqrt{\frac{p(1-p)}{n}} < 0.04$
- For 95% CI, $z^* = 1.96$ (from a lookup table)
- If an estimate of p is available, use it
- If not, use worst case value of $p = 0.5$ (margin of error is largest)
- Solve for n (always round up for sample size calculations!)
- Also make sure the success-failure condition is checked in the final sample to ensure normal approximation is reasonable

6.2 Difference of two proportions

Normality assumption

- Difference of two sample proportions $\hat{p}_1 - \hat{p}_2$ can be modeled using normal distribution when:
  - Data are independent within and between 2 groups (e.g., 2 independent random samples)
  - Success-failure condition holds for both groups

When these conditions are satisfied, the standard error of $\hat{p}_1 - \hat{p}_2$ is

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

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

Confidence intervals

$$\hat{p}_1 - \hat{p}_2 \ \pm \ z^\star \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Same 4 steps as before: prepare, check, calculate, conclude

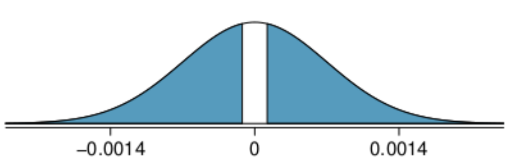
2-proportion hypothesis tests

- When H0 is p1 − p2 = 0 (or, p1 = p2): you can use **pooled proportion** to to get the best estimate of both proportions and to check success-failure condition

$$\begin{aligned}\hat{p}_{pooled} &= \frac{\#\text{ of patients who died from breast cancer in the entire study}}{\#\text{ of patients in the entire study}} \\ &= \frac{500 + 505}{500 + 44{,}425 + 505 + 44{,}405} \\ &= 0.0112\end{aligned}$$

- Use pooled proportion to calculate SE
- Use to calculate Z-score and draw a picture

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{-0.00012 - 0}{0.00070} = -0.17$$



- Compute p-value and make interpretations/conclusions
- Where H0: p1 − p2 is not 0 but some particular number: don't use pooled proportion

6.3 Testing for goodness of fit using chi-square

- Assessing a model when data are binned (e.g., dealing with count data)

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

Figure 6.6: Actual and expected make-up of the jurors.

- Hypotheses:
  - H0: sample proportions are randomly chosen (no bias); observed counts reflect natural sampling fluctuations
  - HA: sample proportions are not randomly chosen (there is a bias in selection)
  - To evaluate, we quantify how different observed counts are from expected counts

Chi-square test statistic

- So far, we dealt with test statistics like this (z-score): $\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$
- This was based on 2 ideas:
  - Calculating the difference between point estimate and expected value if H0 was true
  - Standardizing that difference using a standard error
- We need a single test statistic to determine if several standardized differences are irregularly far from zero: combine them (first square them and then combine)!

$$X^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \cdots + \frac{(\text{observed count}_4 - \text{null count}_4)^2}{\text{null count}_4}$$

- Chi-squared: is the sum of squared Z values, summarizes how strongly observed counts tend to deviate from the null counts
- If H0 is true, then X^2 follows a new distribution called a chi-square distribution, which we can use to compute a p-value to evaluate the hypotheses

Chi-square distribution

- Sometimes used to characterize data and statistics that are always positive and typically right skewed
- Degrees of freedom (df): just 1 parameter that influences shape, center and spread
  - As df increases, distribution becomes more symmetric and center moves to the right and variability increases
- Need to know the upper tail area to calculate the p-value

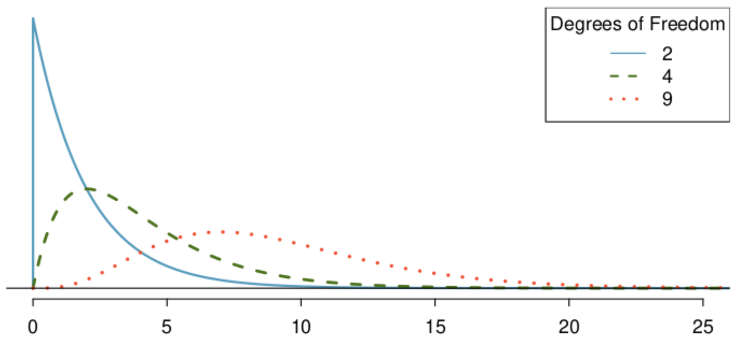


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

- Chi-square calculator
  https://stattrek.com/online-calculator/chi-square.aspx
- Finding a p-value
  - Large X^2 values would suggest strong evidence for HA
  - If H0 was true and there was no bias, X^2 would follow a chi-square distribution, with k - 1 degrees of freedom (k = number of bins)
  - Conditions
    - Independence: each case/count must be independent of all the other cases/counts in the table
    - Sample size: each cell count must be at least 5!
  - If conditions are met, chi-square model can be applied
  - In the juror hypothesis, df = 3, X^2 = 5.89 -> p = 0.12 (cannot reject H0)

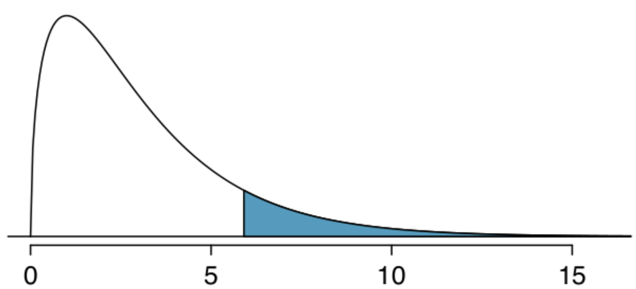


Figure 6.9: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

- If data with only 2 bins, pick a single bin and use the 1-proportion hypothesis test
- Evaluating goodness of fit
  - E.g., waiting time until a positive trading day for S&P500; we can test whether the observed counts follow geometric distribution which is expected if stock market up/down status was independent from all other days

6.4 Testing for independence in two-way tables

Two-way table

- One-way table describes counts for each outcome in a single variable
- Two-way table describes counts for combinations of outcomes for 2 variables
- Often want to know if variables are related in any way (they are dependent?)
- Start with computing expected counts based on row totals, column totals and table total

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2 **(20.33)**	23 **(20.33)**	36 **(20.33)**	61
Hide Problem	71 **(52.67)**	50 **(52.67)**	37 **(52.67)**	158
Total	73	73	73	219

Figure 6.15: The observed counts and the **(expected counts)**.

Chi-square test

General formula	$\frac{(\text{observed count } - \text{expected count})^2}{\text{expected count}}$
Row 1, Col 1	$\frac{(2-20.33)^2}{20.33} = 16.53$
Row 1, Col 2	$\frac{(23-20.33)^2}{20.33} = 0.35$
$\vdots$	$\vdots$
Row 2, Col 3	$\frac{(37-52.67)^2}{52.67} = 4.66$

- X^2 is computed by adding the value for each cell
- Degree of freedom (df) = (# rows -1) x (# columns -1)
- When analyzing 2-by-2 contingency tables, use 2-proportion methods