M 358K: Applied Statistics

Brady Zhou

Fall 2016

1 Sampling Distribution Models (Ch. 18)

Definitions

i. Sampling distribution - all proportions from all possible samples.

ii. Sampling error/variability - change expected from different samples.

iii. Independence assumption - sampled values must be independent.

iv. Sample size assumption - n must be large enough.

v. Randomization condition - data represents population, sampling unbiased.

vi. 10% condition - n must be no larger than 10% of population, drawn without replacement.

vii. Success/Failure condition - expect at least 10 successes/failures.

viii. Sampling distribution model - sampling distribution is modeled as a normal distribution.

ix. Law of large numbers - as sample size increases, each sample average will be closer to population mean.

x. Central limit theorem - sampling distribution of any mean approaches normal as sample size grows.

xi. Large enough sample condition - CLT requires different sample size depending on true population.

Sampling Distribution Model for a Proportion (Categorical Data)

$$\sigma(\hat{p}) = SD(\hat{p}) = \sqrt{\frac{pq}{n}} \tag{1}$$

Central Limit Theorem

The mean of a random sample is a random variable whose sampling distribution can be approximated by a normal model. The larger the sample, the better the approximation will be.

Sampling Distribution Model for a Mean (Quantitative Data)

$$\sigma(\bar{y}) = SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} \tag{2}$$

2 Confidence Intervals for Proportions (Ch. 19)

i. Standard error - estimation of standard deviation of a sampling distribution.

ii. Confidence interval - range that contains the true population at a given confidence.

iii. One-proportion z-interval - confidence interval dealing with a single population's proportion.

iv. Margin of error - half of the width of the confidence interval.

v. Critical value - number of standard deviations that corresponds with the confidence.

Standard Error

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{3}$$

Confidence Interval (One-Proportion Z-Interval)

The interval given by the following formula *probably* contains the the true population proportion.

$$\hat{p} \pm ME \tag{4}$$

where ME is the margin of error, defined by

$$ME = z^* \ SE(\hat{p}) \tag{5}$$

Note: z^* is critical value that corresponds with the confidence interval.

For example, for 90% confidence, $z^* = 1.645$, for 95% confidence, $z^* = 2$, for 99.7% confidence, $z^* = 3$.

To drop the margin of error, we can either decrease the level of confidence, or increase the sample size.

3 Testing Hypotheses About Proportions (Ch. 20)

i. Null hypothesis - denoted H_0, proposes a value for a population parameter.

ii. Alternative hypothesis - denoted H_A, contains plausible values for a population parameter.

iii. P-value - probability of observing statistic given the null hypothesis is true.

iv. One-proportion z-test - test about proportions.

v. Effect size - difference between true and hypothesized parameters.

vi. Two-sided alternative - $H_A : p \neq p_0$.

vii. One-sided alternative - $H_A : p \leq p_0$ or $H_A : p \geq p_0$.

One-Proportion Z-Test

We test $H_0 : p = p_0$ using the z-statistic, which is defined by

$$z = \frac{(\hat{p} - p_0)}{SD(\hat{p})} \tag{6}$$

where $SD(\hat{p})$ is the standard deviation of the hypothesized population, defined by

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} \tag{7}$$

4 More About Tests and Intervals (Ch. 21)

i. Alpha level (significance level) - if the p-value falls under this threshold, the null hypothesis is rejected.

ii. Plus-four method - confidence interval from padded data so the success/failure condition is satisfied.

iii. Type I error - null hypothesis true, mistakenly reject, a false positive.

iv. Type II error - null hypothesis false, fail to reject, a false negative.

v. Power - probability that the test correctly rejects a false null hypothesis, $1 - \beta$.

vi. Effect Size - distance between hypothesized value and true value.

vii. α - probability of a type I error.

viii. β - probability of a type II error.

Using Confidence Intervals to Test Hypotheses

Establish a confidence interval, and if the null hypothesis does not fall within the range, then we can reject the null hypothesis.

Agresti-Coull "Plus-Four" Interval

If the observed data does not satisfy the success/failure condition, then another confidence interval can be constructed, where

$$\tilde{p} = \frac{y+2}{n+4} \tag{8}$$

Common Alpha and Critical Values

α	1-sided	2-sided
0.05	1.645	1.96
0.01	2.33	2.576
0.001	3.09	3.29

Altering α and β of a Test

We can reduce β by increasing α, which makes it easier to reject the null, regardless if it is true or not. This, however, increases the probability of type I errors.

To reduce both α and β, we can simply collect more data.

5 Comparing Two Proportions (Ch. 22)

i. Independent groups assumption - the two groups being compared must be independent of each other.

ii. Two-proportion z-interval - the confidence interval for the difference of two proportions.

iii. Pooling - combining counts to get an overall proportion.

iv. Two-proportion z-test - hypothesis testing when comparing two different populations.

Standard Deviation of the Difference of Two Proportions

$$SD(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \tag{9}$$

Standard Error of the Difference of Two Proportions

$$SE(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}} \tag{10}$$

Two-Proportion Z-Interval

The confidence interval for the difference of two sample proportions is defined by

$$(\hat{p_1} - \hat{p_2}) \pm z^* SE(\hat{p_1} - \hat{p_2}) \tag{11}$$

Hypothesis Testing for Comparing Two Proportions

$$H_0 : p_1 - p_2 = 0 \tag{12}$$

Pooled Proportion

$$\hat{p}_{pooled} = \frac{n_1 \hat{p_1} + n_2 \hat{p_2}}{n_1 + n_2} \tag{13}$$

Pooled Standard Error

$$SE_{pooled}(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_2}} \tag{14}$$

Two-Proportion Z-Test Statistic

$$z = \frac{(\hat{p_1} - \hat{p_2}) - 0}{SE_{pooled}(\hat{p_1} - \hat{p_2})} \tag{15}$$

6 Inference About Means (Ch. 23)

i. Student's (Gosset's) t-distribution - normal-like distribution pertaining to population mean.

ii. One-sample t-interval - a confidence interval used to bound the mean.

iii. Nearly normal condition - the data comes from a unimodal, symmetric distribution.

iv. One-sample t-test - hypothesis test for the mean.

v. Critical value - denoted t^*_{n-1}, and depends on sample size and confidence.

Sampling Distribution Model For Means

The sample mean can be standardized to a t-statistic, defined by

$$t = \frac{\bar{y} - \mu}{SE(\bar{y})} \tag{16}$$

with $n - 1$ degrees of freedom, and where

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} \tag{17}$$

One-Sample T-Interval for the Mean

We can establish a confidence interval for the mean, defined by

$$\bar{y} \pm t^*_{n-1} SE(\bar{y}) \tag{18}$$

Using this model increases the margin of error, and increases p-values.

One-Sample T-Test for the Mean

We can test $H_0 : \mu = \mu_0$, using the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})} \tag{19}$$

T-Distribution Properties

As the degrees of freedom approaches infinity, the t-distribution approaches normal.

7 Comparing Means (Ch. 24)

i. Two-sample t-interval - difference in means.

ii. Two-sample t-test - hypothesis test for difference in means.

iii. Pooled t-test - if variances assumed to be equal, common variance can be calculated.

iv. Equal variance assumption - variances from two populations are equal.

v. Similar spreads condition - look at boxplots and check spreads.

Standard Deviation of the Difference of Two Means

$$SD(\bar{y_1} - \bar{y_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{20}$$

Standard Error of the Difference of Two Means

$$SE(\bar{y_1} - \bar{y_2}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{21}$$

Two-Sample T-Interval

$$(\bar{y_1} - \bar{y_2}) \pm t_{df}^* \ SE(\bar{y_1} - \bar{y_2}) \tag{22}$$

Two-Sample T-Test

We can test $H_0 : \mu_1 - \mu_2 = \Delta_0$ using the following statistic

$$t = \frac{(\bar{y_1} - \bar{y_2}) - (\mu_1 - \mu_2)}{SE(\bar{y_1} - \bar{y_2})} \tag{23}$$

Note: the degrees of freedom formula is complicated.

Pooled Variance

$$s_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \tag{24}$$

Pooled Standard Error

$$SE_{pooled}(\bar{y_1} - \hat{y_2}) = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} \tag{25}$$

8 Paired Samples and Blocks (Ch. 25)

i. Paired data - data that is not independent, measured before and after.

ii. Blocking - pairs from experiments.

iii. Matching - pairs from observations.

iv. Paired t-test - one-sample t-test for means of pairwise differences.

v. Paired data assumption - the groups must not be independent of each other.

vi. Paired t-interval - confidence interval for the mean of the paired differences.

Paired T-Test

We can test $H_0 : \mu_d = \Delta_0$ with the statistic

$$t_{n-1} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})} \tag{26}$$

where the standard error of the mean of the pairwise differences is defined by

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} \tag{27}$$

Paired T-Interval

The confidence interval is defined by

$$\bar{d} \pm t^*_{n-1}\ SE(\bar{d}) \tag{28}$$

where the standard error of the mean of the pairwise differences is defined by

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} \tag{29}$$

9 Comparing Counts (Ch. 26)

i. Goodness of fit test - measures how closely observed counts fit model, use $(n-1)$ df.

ii. Counted data condition - data must be counts for categorical data.

iii. Expected cell frequency condition - expect to see at least 5 individuals in each cell.

iv. Chi-square statistic - relative magnitude of difference between observed and expected.

v. Two-way table - used to compare pairwise category counts.

vi. Chi-square test of homogeneity - see if data is consistent across all groups, use $(r-1)(c-1)$ df.

vii. Standardized residuals - shows how observed data deviates from hypothesized pattern.

viii. Contingency table - categorize counts to tell whether the two variables are dependent.

ix. Chi-square test of independence - similar to homogeneity, but for a row/col has a yes/no grouping.

Chi-Square Statistic

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{30}$$

Standardized Residual

$$c = \frac{(Observed - Expected)}{\sqrt{Expected}} \tag{31}$$

10 Glossary

i. 68-95-99.7 Rule - The percent of a normal distribution that falls within 1, 2, 3 standard deviations.

ii. Right/left skewed - the right/left tail is longer.

iii. Sample standard deviation - $s = \sqrt{\frac{\Sigma(y-\mu)^2}{n}} = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}$.

iv. Tukey's quick test - count values in high group larger than all of low group, compare to 7, 10, 13.