Chapter 1A Summary

Paul Luo, Tina Wang, Jackson Cong

2022-11-08

Investigation 1.1

- Rossman/Chance Applet Collection (https://www.rossmanchance.com/applets/2021/oneprop/OneProp.htm)
- We summary the observed data by making a bar graph, one bar for each outcome, with heights representing the number of observations in each category, separating the bars to indicate distinct categories.

Investigation 1.2

- Null hypothesis (H_0)
  - A type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations.
- Alternative hypothesis (H_a)
  - Contradictory to the null hypothesis and a statement which a researcher is testing.
- Binary variable
  - A categorical variable with only two possible outcomes (e.g., heads, tails)
- Categorical variable
  - A variable that places observational units into categories (e.g., small, medium, or large), rather than measuring a numerical value.
- Quantitative variable
  - A variable that takes on numerical characteristics (where it makes sense to average the values of the outcomes)
- Observational units
  - The people or objects about which data are recorded.
- Parameter
  - A numerical summary describing the larger process than generated the data or to the population from which the sample was selected.
- P-value
  - Strength of evidence against the null hypothesis (see 1.8).
- Sample size
  - The number of observational units in the study (for which data have been recorded). Typically denoted by n.
- Variable
  - Any characteristic that varies from observational unit to observational unit.
- Binomial probability
  - k(success), n(independent trials), pi(probability of success)
  - $P(X = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$ where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
  - Independent trial means the success of one trial does not affect the result of next trial

Investigation 1.3

- Statistics

    - A number describing a sample.
- Parameter
    - A number describing a whole population.
- Standardization
    - The process of putting different variables on the same scale.
- Mean
    - The average of a data set.
- Standard Deviation
    - A statistic that measures the dispersion of a dataset relative to its mean.

Investigation 1.4

- Quantitative variable
    - Any variables where the data represent amounts(numbers).
- Categorical variable
    - Any variables where the data represent groups(names).
- Two sided(p-value)
    - A significance test for which no particular direction is specified in the alternative hypothesis, using “not equal to” in the alternative hypothesis.

Investigation 1.5

- Confidence interval
    - The probability that parameter will fall between a set of values (see 1.9)
- Confidence level
    - The long-run proportion of intervals that capture the parameter value. If the procedure is valid the observed coverage rate under repeated random sampling will match the stated confidence level.
- Duality
    - The correspondence between a two-sided test of significance and a confidence interval.
- Level of significance
    - The cut-off for the p-value that leads us to reject the null hypothesis. The probability of a type I error.

Investigation 1.6

- Power
    - The probability of rejecting the null hypothesis at a particular alternative value of the parameter
- Rejection region
    - The values of the statistic that lead us to reject the null hypothesis for a particular level of significance
- Type-I Error (See 1.11)
- Type-II Error (See 1.11)

Investigation 1.8

- Central Limit Theorem (CLT)
    - If sample is large enough, to prove binomial random varible is normal, we use $n \times \pi \geq 10$ and $n \times (1 - \pi) \geq 10$.
- One proportion Z-test
    - Get Z score, which means a numerical measurement that describes a value’s relationship to the mean of a group of values. Eg. Z-score = 1 means one standard deviation away from the π_0

- $Z_0 = \frac{\hat{p}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
- P-value
  - The P-value is based on a normal Z-distribution. The P-value of a test is the probability that we would get this sample result or one more extreme if the null-hypothesis is true. The smaller the P-value is, the stronger the evidence against the null-hypothesis provided by the data.
  - We use interval $(\mu - 1, \mu + 1)$ to capture approximately 68% of the distribution.
  - We use interval $(\mu - 2, \mu + 2)$ to capture approximately 95% of the distribution.
  - We use interval $(\mu - 3, \mu + 3)$ to capture approximately 99.7% of the distribution.

Investigation 1.9

- Statistical Significance
  - If the P-value is as small as or smaller that α, we say that the data are statistically significant at level α. In general, use $\alpha = 0.05$ unless otherwise noted.
- Standard Error
  - The standard error of the sample proportion, $SE(\hat{p})$, is an estimate for the standard deviation of $\hat{p}$.
  - $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- Critical Value Z* -The multiplier of the standard error in a confidence interval corresponding to the nominal confidence level.
- Margin of Error
  - The half-width of a confidence interval.
  - $MOE = Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- Confidence Interval
  - $statistic \pm margin\ of\ error$
  - A confidence interval gives an interval of plausible values for a parameter based on sample data.
  - $\hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where Z^* correspond to the confidence level.

Confidence Level	90%	95%	99%	99.9%
Critical Value Z*	1.645	1.960	2.576	3.291

Investigation 1.11

- Type-I Error
  - Rejecting the H_0 when it is actually **true** (a false positive).
- Type-II Error
  - Accepting the H_0 when it is actually **false** (a false negative).
  - The probablity of a **Type I error** is the same as alpha, the significant level.
- Plus Four procedures
  - Adding two successes and two failures to the sample before computing a one-sample z-interval to improve the long-run coverage rate of the procedure.