Week 3 Lab

Tuesday, January 30/Thursday, February 1

Plan for today

1. Changes of lab session format
2. Bernoulli distribution
3. Binomial Coefficient
4. Binomial Distribution
5. Confidence Intervals
6. One sample Student's t-test
7. One sample Wilcoxon Signed-Rank test
8. Power analysis for one-sample t-test

Changes of lab session format

Changes of lab session administration

- Lab handout:
  - Include ”To Think About” questions
- Lab format:
  - More similar to benchwork lab, e.g. chemistry
  - Less lecturing
  - More reading and practicing

If you were confused by the previous lab format...

- Sorry. We (TAs) did not clearly explain the design.
- The previous lab sessions were designed to prepare you for reading the write-up independently:
  - Review important theoretical points in the lecture
  - Explore programming concepts interactively (more time for questions)
  - Make sure that the code can be executed on you computer
  - Generally explain the design of code
  - Give personal suggestions on controversial topics
  - Leave the reading after the lab session at your own pace
  - Leave trouble-shooting to Canvas or office hour
  - Leave practicing to homework assignments

If you were confused by the previous homework...

- Grading:
  - We are not expecting you to give us a single ”correct” answer.
  - We are expecting you to explain your reasoning in your answer.
  - We will accept all reasonable answers with important reasoning steps.
  - However, do not think too far away. There is always a solution using the lecture and lab materials.
  - If you are thinking about methods not covered in the lecture or lab, discuss with the TAs or instructors.

If you were confused by the previous homework...

- You can:
  - Start early.
  - Get help during office hour or through Canvas.
- We will:
  - The instructors will try to put the homework quesetions in a more controlled context.
  - The instructors will try to ask more natural questions.
  - We TAs will try to discuss possible confusions about the homework at the end of the lab.

Recap

Recap

- Different interpretations of random variable.
- The normal distribution.
- Testing for normality:
  - Plotting: Histogram, density plot, and Q-Q plot.
  - Statistical testing: Shapiro-Wilk normality test and Anderson-Darling normality test
- The central limit theorem.
- One-sample tests of proportion
  - z-test and χ^2-test
  - One-sided versus two-sided
  - Power analysis

Bernoulli distribution

Bernoulli distribution

Notation: $X \sim \text{Bern}(p)$

Parameter: $p \in [0, 1]$ – success probability of a Bernoulli trial.

Support: $x \in \{0, 1\}$

Probability mass function:

$$f(x; p) = \begin{cases} p, & \text{for } x = 1. \\ q = (1 - p), & \text{for x = 0.} \end{cases}$$

Bernoulli distribution is a special case of binomial distribution.

R has `rbinom()` but not `rbern()`.

Binomial Coefficient

Binomial Coefficient

Theorem (Binomial theorem)

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

Definition (Binomial coefficient)

Any of the positive integers that occurs as a coefficient in the binomial theorem is a binomial coefficient.

Notation:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Source: https://en.wikipedia.org/wiki/Binomial_theorem
Source: https://en.wikipedia.org/wiki/Binomial_coefficient

Binomial Distribution

Binomial Distribution

Intuition: the distribution of the sum of n random variables *i.i.d.* Bern(p).

$$X \sim B(n, p)$$

- Parameters:
  - $n \in N_0$ – number of trials
  - $p \in [0, 1]$ – success probability in each trial
- Support: $x \in \{0, 1, 2, ..., n\}$
- Probability mass function:

$$f(x; n, p) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

Source: `https://en.wikipedia.org/wiki/Binomial_distribution`

Confidence Intervals

Confidence interval

- Intuition: express the confidence of estimating population parameter using a sample of observations as an interval of estimates.
- Straightforward in algebraical definitions:
  - Closely related to hypothesis testing.
  - Multiple methods to calculate the confidence intervals (CIs) of the same population parameter.

Controversy of interpreting confidence interval

- Controversial interpretations of confidence level (described without mathematical rigor):
  - Apply algebraical equivalence: probability that the population parameter lies in the interval.
  - Apply frequentist claimed definition: the proportion of intervals containing the population parametres in a large number of samples.
- The controversy lies in the philosophical question of whether the population parameter is random (algebraical interpretation) or constant (frequentist interpretation).
- The algebraical interpretation is usually referred as "misinterpretation" or "incorrect".

Source: `https://onlinecourses.science.psu.edu/stat414/node/197`
Source: whuber at StatsExchange

Example: constant or random population parameter

- Scenario: estimate the mean effect of a drug on a disease measured as continuous value with arbitrary unit in $\mathbb{R}$.
- Population parameter: the mean effect of the drug, μ.
- Categorization (determined using domain specific knowledge): for the same disease, a drug is good if its $\mu \geq 100$, otherwise bad.
- Sample: 500 patients. Mean 110. 95% confidence interval of population mean: $[100, 120]$.
- Which claim on the drug do you prefer?
  - Significantly good, $p \leq 0.05$. (constant)
  - 97.5% chance to be good. (random)

One sample Student’s t-test

One sample Student's t-test

Intuition from CI perspective: calculate CI for population mean without requiring large sample size or known population standard deviation.

Theorem

If $X_1, X_2, ..., X_n$ are normally distributed random variables with mean μ and variance σ^2, then a confidence interval of confidence level $1 - \alpha$ for the population mean μ is:

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Source: `https://onlinecourses.science.psu.edu/stat414/node/199`

One sample Wilcoxon Signed-Rank test

One sample Wilcoxon Signed-Rank test

- Null hypothesis (H_0): the distribution of a sample is symmetric about μ_0.
- Non-parametric test procedure: no known model for data points.
- Intuition:
  - Use rank to ignore the magnitude of differences between different data points.
  - Use sign, whether greater than or less than μ_0, to describe the ”direction” of the data point relative to μ_0
  - If H_0 is true, the sum of signed rank follows certain distribution.

Power analysis for one-sample t-test

General procedure of power analysis

- Calculation of power: usually before having the data, if the alternative hypothesis H_1 is true, how likely can we reject the null hypothesis H_0 with significance level α.
- Parameters:
  - Arbitrary constants: H_1, H_0, and α.
  - Constant depend on H_1 and H_0: *effectsize*.
  - Variables: power and sample size n.
  - A priori power analysis: set power and calculate n.
  - Post hoc power analysis: set n and calculate power.
- By convention from Wikipedia, power is set to 0.8 when calculating sample size.

Intuition of power analysis

- Good intuition of power analysis is unknown to me.
- Best I can get is power $= 1 - \beta$.
- Why? No generalized mathematical framework for calculating power comparing to hypothesis testing.
- Better to reason through the calculation of each specific case using the definition:

$$\text{power} = \mathsf{Pr}\left(\text{reject } H_0 \mid H_1 \text{ is true}\right)$$

Power analysis for one-sample t-test

- H_0: population mean $\mu = \mu_0$.
- $\alpha = 0.05$.
- Set n or power.
- Very complicated in deriving the mathematical formula of power $= \Pr\big(\text{reject } H_0 \mid H_1 \text{ is true}\big)$.
  - Various H_1s: $\mu = \mu_1$, or $\mu > \mu_0$, or $\mu < \mu_0$, or $\mu \neq \mu_0$.
  - Different H_1s may result in different formulas.
- Very complicated interpretation of the effect size.
  - Assumes population variance takes certain value.
  - Possibly various calculation methods.
- Call `pwr.t.test` to avoid complexities, and check the mathematics when necessary (usually not).

Questions?