STA101 Midterm Review

Fan Bu

May 27, 2021

Summer Session I 2021

Department of Statistical Science, Duke University

Logistics of the Midterm

About the Midterm

- Chapters 1-5 will be covered
- “Tests & Quizzes” on Sakai
- Open at 00:00am May 28 and close at 11:59pm May 30
- **2-hour** time window (you can start any time, but once started you must finish within 2 hours)
- 3 parts: (1) True or false, (2) Fill in the blank, and (3) Multiple choices
- You can try out a sample exam (with the same operations and structures, but dummy questions) on Sakai

About the Midterm (cont’d)

- **Individual** exam - you should **NOT** consult other people or the Internet
- Open-book (sort of) - you can consult the textbook, your notes, lecture slides, etc.
- Find a place with a steady Internet connection
- If you encounter **technical** issues (or if you believe certain questions are mis-worded or wrong), you should email Fan (`fan.bu1@duke.edu`); emails will be replied 9am-6pm EDT, May 28-30
- I’ll also be available on Zoom during lecture time (10-11:15am) and my office hour time (7-8pm) on May 28 (usual Zoom links on Sakai)

A brief review of the topics

Chapter 1

- **Data types**: numerical (continuous, discrete) & categorical (regular, ordinal)
- **Association** $\neq$ **Causation** between explanatory & response variables
- **Sample v.s. Population**: size & representative
- **Observational studies**: sampling strategies
- **Experiment design**: control, treatment, random assignment, etc.

Chapter 1 - example question

Lisa conducted a survey on 100 randomly selected college students and asked them about their sleep time and GPA.

(a) Observational study, or experiment?

(b) Can she make a causal statement about relationship between sleep time and GPA?

(c) Type of variable for GPA (4.0 scale)?

(d) If she wants to make sure that students in all 4 years are included in her sample, what sampling strategy should she use?

Chapter 2

- **Numerical data**:
  - Dot plots, histograms, boxplots, etc.
  - Mean, SD (non-robust); median, IQR (robust)
- **Categorical data**: contingency table, (all kinds of) bar plots.
- **Randomization test**: simulations for hypothesis testing

Chapter 2 - example question

For a right-skewed distribution, median < mean or median > mean?

Which plot would be the best choice for visualizing heights of male and female adults (relationship between height and gender)?

Which plot would be the best choice for visualizing the relationship between whether or not a voter has a college degree and their attitude towards vaccination?

Chapter 3

- **Probability theory basics**: disjoint & non-disjoint events, probability distribution, sample space, complementary events, independence.
- **Conditional probability**: definition and calculation, connection with independence, inverse probability, Bayes theorem
- **Sampling from small population**: with and without replacement
- **Random variables**: mean, variance (standard deviation), linear combination of random variables

Chapter 3 - example questions

Consider events A and B:

- A = supports FC Barcelona, B = supports Real Madrid, then $P(A \text{ or } B) = P(A) + P(B)$;
- A = chicken sandwich for lunch, B = sweet tea for lunch, then $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$;
- A = heads up in 1st coin toss, B = heads up in 2nd coin toss, and consecutive coin tosses are independent, then $P(A \text{ and } B) = P(A)P(B)$;
- A = get A+ in 1st math course, B = get A+ in 2nd math course, then $P(A \text{ and } B) = P(A)P(B \mid A)$.

Chapter 3 - example questions

Some syndrome X has prevalence $x\%$ in the population, and some testing method Y is $y\%$ accurate for those without syndrome X and $z\%$ accurate for those with syndrome X. Person W tests positive. What is the probability that W actually has X?

Chapter 3 - example questions

Some syndrome X has prevalence $x\%$ in the population, and some testing method Y is $y\%$ accurate for those without syndrome X and $z\%$ accurate for those with syndrome X. Person W tests positive. What is the probability that W actually has X?

Random variables X and Y have standard deviations σ_X and σ_Y respectively. X and Y are independent. What is the standard deviation of $X + Y$?

Chapter 4

- **Normal**: shape of distribution, z-score, percentiles, quantiles
- **Geometric**: Bernoulli trials, probability formula, mean & SD
- **Binomial**: choose function, probability formula, mean& SD, normal approximation of Binomial
- **Poisson**: probability formula, mean& SD

Chapter 4 - example question

$$X \sim N(\mu = 50, \sigma = 3)$$

(a) $Pr(X > 59) =$?

(b) If $Pr(X < x) = 0.35$, what is x?

(c) What is the Z-score for 45?

Chapter 5

- **Point estimates and sampling variability**: point estimates, sampling distribution, central limit theorem (independence & success-failure conditions)
- **Confidence intervals**: formula for constructing CIs for a proportion, margins of error, widths of CIs, interpretation of CIs (“we’re xx% confident that ...”)
- **Hypothesis testing** (for single proportions): how to set up H_0 and H_A, Type 1 and 2 errors, significant level, p-values (and interpretations)

Chapter 5 - example question

How large should be expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

Chapter 5 - example question

How large should be expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

For a fixed sample and a specific parameter of interest, is the 90% CI or 99% CI wider?

Chapter 5 - example question

How large should be expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

For a fixed sample and a specific parameter of interest, is the 90% CI or 99% CI wider?

If $\alpha = 0.01$ and p-value $= 0.005$, should we reject H_0? Does this mean that H_0 must be true/false?

Chapter 5 - example question

How large should be expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

For a fixed sample and a specific parameter of interest, is the 90% CI or 99% CI wider?

If $\alpha = 0.01$ and p-value $= 0.005$, should we reject H_0? Does this mean that H_0 must be true/false?

The margin of error for a $n = 50$ sample is 0.1. What will the margin of error be if the sample size increases to $n = 200$?

Chapter 5 - example question

How large should be expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

For a fixed sample and a specific parameter of interest, is the 90% CI or 99% CI wider?

If $\alpha = 0.01$ and p-value $= 0.005$, should we reject H_0? Does this mean that H_0 must be true/false?

The margin of error for a $n = 50$ sample is 0.1. What will the margin of error be if the sample size increases to $n = 200$? (*Note: margin of error* $\sim \frac{1}{\sqrt{n}}$)

Some `R` stuff

- How to check out data: `str` or `dim`
- Normal percentiles & quantiles: `pnorm` and `qnorm`
- Check out the help document: “?” before a function name
- Have your `RStudio` container (or local environment) ready for the exam (e.g., you’ll be asked to load a dataset in `R` and check it out)