STAT 110: Review lecture

University of Otago

Outline

- Big picture review of the course
- Connect key elements to (practice) exam questions
- I have not included the context in many cases

Data

- We looked at data, summaries, and R
  - The R object `penguin` contains information on a random sample of chinstrap penguins from the Palmer archipelago. There are two variables: `bill`, the bill length (mm), and `flipper`, the flipper length (mm). We consider the R code below:

```
mean(penguin$bill)
sd(penguin$flipper)
```

    - What is being evaluated in the first line of R code: `mean(penguin$bill)`?
  - We observe data y = (39.7,41.3,44.4,39.0,45.5)
    - The sample mean $\bar{y}$ is closest to

Probability and random variables

- We want to fit statistical models
- We need knowledge of probability[1]
  - What is the best interpretation of $\Pr(B|V^{\complement})$?
  - The probability $Pr(V|B)$ is closest to
  - Find the quantity $E[Y]$
  - What is the best description of a random variable?
  - What is the expected nutrient score per serving, $E[2X - 3Y]$?

[1] I haven't included the context for these questions.

The normal distribution

- Looked in detail at the normal distribution
  - Working memory span refers to the amount of information a person can temporarily hold and manipulate in their mind while performing a cognitive task. A score of working memory span has been developed that is normally distributed with mean $\mu = 40$ and standard deviation $\sigma = 8$ for healthy adults in the population.
    - A randomly selected healthy adult has a working memory score that is 1.5 standard deviations below the mean ($z = -1.5$). Their working memory score is closest to
    - Which of the following options calculates the probability that a randomly selected healthy adult has a score above 48?
- We found the sampling distribution for $\bar{y}$
  - If we were to collect a sample of $n = 64$ healthy adults and calculate their working memory score, select the option below that best describes the sampling distribution of the sample mean $\bar{y}$

Normal models

- One sample & paired data
  - The R code below carries out the hypothesis test:

$$\mathrm{H}_0 : \mu_d = 0; \qquad \mathrm{H}_A : \mu_d \neq 0,$$

    where μ_d is the mean difference in the nitrogen levels (after-before). If $\alpha = 0.05$, select the best interpretation:
- Two independent samples
  - The sample mean reduction for drug A is $\bar{y}_1 = 19.00$ with sample standard deviation $s_1 = 13.579$. The sample mean reduction for drug B is $\bar{y}_2 = 15.95$ with sample standard deviation $s_2 = 9.054$. The estimated standard error for $\bar{y}_1 - \bar{y}_2$ is closest to
  - Which of the following options should we use to find a 95% confidence interval for $\mu_1 - \mu_2$
  - The R output of a suitable model is below. Select the best interpretation

Normal models

- ANOVA
  - Select the hypotheses that are being tested with ANOVA
  - The F-value for the appropriate test is closest to
  - Select the option that is not correct with respect to ANOVA
  - Select the best interpretation of the p-value from the ANOVA test

Linear regression

- Understanding the linear regression model
  - What is the best interpretation of β_1?
  - Which of the following is correct for the subpopulation of mammals that have body mass of 20kg?
  - Does it make sense to interpret $\hat{\beta}_0$ in this application?
- Estimating / fitting a linear regression model
  - What is the best description of the method used to estimate the parameters in the linear regression model below?
  - Select the correct expression for the fitted regression model based on the R output
- Assumptions
  - Suppose that we fit a linear regression model with outcome y and predictor variable x. Based on the plot below, select the option that best describes which regression assumptions, if any, appear to be violated

Linear regression

- Prediction
  - The researchers want to use the model to predict the aptitude of a child who first speaks at 60 months. This quantity is closest to
  - The code below finds two intervals. The type of interval is hidden (we have replaced the type of interval by A and B). Select the best description of these intervals
- Multiple linear regression
  - Which of the following statements about multiple linear regression is correct?
  - Researchers fit a model that includes both temperature and activity. Select the option that gives $\hat{\beta}_2$ and the standard error for $\hat{\beta}_2$.
- Categorical predictors: see Assignment 8
- Model fit
  - The R^2 is 84.8%. Which of the statements below is not correct.

Binary/binomial models

- Assumptions
  - Researchers are studying how frogs respond to a predator cue. They expose individual frogs to the cue and record whether each frog jumps away (yes/no). They continue collecting data until they observe 20 frogs that run away. Which of the binomial assumptions, if any, are violated?
- Model fitting and interpretation
  - The sample proportion of field goals made from less than 50 yards is closest to
  - A confidence interval can be found using `prop.test` as below. Select the best description of the parameter being estimated by the confidence interval shown in the output
  - What hypothesis test is being carried out when using `prop.test`

Contingency table

- χ^2-test
  - If we assume independence between diet and cancer, the expected count of those with a moderate diet of fish and no cancer is closest to
  - What is the appropriate hypotheses for the χ^2-test for contingency tables.
  - What are the degrees of freedom for the χ^2-test?
  - Select the best interpretation from the χ^2-test below if $\alpha = 0.05$

Other methods

- Nonparametric methods
  - Select the option that best describes how the Mann-Whitney test statistic is found
  - Which of the following is a benefit of using a non-parametric test such as the Mann-Whitney test?
  - Interpret the test carried out below if $\alpha = 0.05$
- Central limit theorem
  - As a summer research project we develop a new working memory score that is not normally distributed but still has mean $\mu = 40$ and standard deviation $\sigma = 8$ (we can assume it is not excessively skewed). If we were to collect a sample of $n = 64$ healthy adults and calculate their working memory score, select the option below that best describes the sampling distribution of $\bar{y}$

Where is the data from?

- Sampling
  - Which of the following best describes stratified sampling?
  - ...which of the following is likely to be the largest source of bias, and why?
- Culturally informed design and governance
  - Which of the following is not a characteristic of co-designed research studies?
  - In Indigenous data soverignty, the CARE acronym refers to which of the following
- Experiments and observational data
  - What is the best description of a placebo group?
  - What is a confounding variable?

Other topics

- Replication crisis
  - What is the replication crisis in science primarily about?
  - What is the best description of HARKing?
  - What is the main danger of performing many statistical tests without adjustment?
- Estimation of statistical models
  - Which of the following best describes maximum likelihood estimation (MLE)?
  - Which of the following is a feature of Bayesian inference?

Summary

- The final exam is comprehensive
  - Questions cover the entire course
  - Questions cover all of the learning outcomes for the course
    - Be able to describe the information contained in a data set
    - Be able to carry out common statistical data analyses in R
    - Be able to interpret the results of common statistical analyses in the context of the scientific study
    - Be aware of the appropriate use of study designs
    - Be able the understand advantages and disadvantages of various statistical procedures
- Keep an eye out for exam help sessions closer to the exam