Review for Midterm

Timothy Hanson

Department of Statistics, University of South Carolina

STAT 205H: Elementary Statistics for the Biological and Life Sciences

Midterm logistics...

- Tuesday, October 10 during class. Penny Wang will proctor.
- Open book, open notes. Bring your laptop (w/ internet access & R) and pencil/pen.
- Problems will be patterned after homework problems; a few multiple choice.
- *Be on time.*
- You will answer the questions on the actual test; just handwrite R code you used. Keep midterm R code in a file in case I've got questions while grading. Should be straightforward though.
- Note that textbook problem solutions are posted on the course website.
- The Midterm and Final are worth 40% of your final grade, 20% each.
- Chapter 1: what is a simple random sample?

2.1 Types of variables

- Categorical
  - Ordinal (e.g. "low, medium, high", "infant, toddler, child, teen, adult")
  - Nominal (e.g. eye color, car type)
- Numeric
  - Continuous (e.g. height, cholesterol, tree diameter)
  - Discrete (e.g. number of cracked eggs in a carton, die roll)

- Have data $y_1, y_2, \ldots, y_n$; want to describe it with pictures and tables.
- If data categorical, can make a bar chart. Can record frequency of data value occurrences in a table.
- Continuous data can be displayed in a histogram defined by bins. Again, need a table of frequency values for occurrences within each bin.
- Histogram/density shape: unimodal, bimodal, multimodal.
- Histogram/density skew: left skew, right skew, symmetry.
- R code: `hist`, `boxplot`, `barplot`, `plot`, `density`.

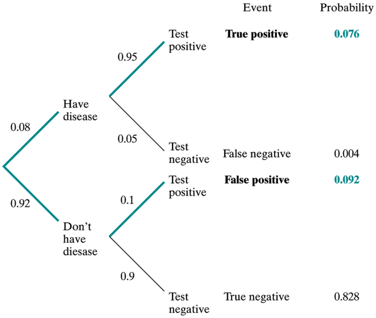
2.3, 2.4, 2.6: Descriptive statistics: mean, median, quartiles, 5 number summary, IQR, boxplots, outliers.

- Mean $\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$ is "balance point" of data.
- Median Q_2 cuts ordered data into halves of equal size.
- First quartile Q_1 is median of lower half; Third quartile Q_3 is median of upper half.
- min, $Q_1, \tilde{y}, Q_3,$ max is 5 number summary, used to make boxplot. Be able to intepret R's boxplot!
- $IQR = Q_3 - Q_1$, length of interval containing middle 50% of data. Sample variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, standard deviation is s.
- $UF = Q_3 + 1.5 \times IQR$, $LF = Q_1 - 1.5 \times IQR$. Any of $y_1, \ldots, y_n$ **larger than** UF or **smaller than** LF are "outliers."
- R code: `mean`, `quantile`, `median`, `summary`, `var`, `sd`, `IQR`, `boxplot` (gives outliers!).

- Let A and B two events. A and B is that both occur. A or B is either occurs. A^C is that A does not occur. *Always*: $0 \leq \Pr\{A\} \leq 1$.
- A and B are *disjoint* if they have no outcomes in common.
- Formulas:
  1. If $E_1, E_2, \ldots, E_k$ disjoint, then $\Pr\{E_1 \text{ or } E_2 \text{ or } \cdots \text{ or } E_k\} = \Pr\{E_1\} + \Pr\{E_2\} + \cdots + \Pr\{E_k\}$.
  2. $\Pr\{A \text{ or } B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \text{ and } B\}$.
  3. (conditional probability) $\Pr\{A|B\} = \Pr\{A \text{ and } B\}/\Pr\{B\}$.
  4. (compliment rules) $\Pr\{A^C\} = 1 - \Pr\{A\}$ *and* $\Pr\{A^C|B\} = 1 - \Pr\{A|B\}$.
  5. (independence) A and B are independent if $\Pr\{A\} = \Pr\{A|B\}$.

Recall computing probabilities from a table of counts!

3.2 Probability trees



$$\Pr\{\text{Disease, Test positive}\} = 0.08(0.95) = 0.076 \qquad \Pr\{\text{Disease, Test negative}\} = 0.92(0.05) = 0.004$$
$$\Pr\{\text{No disease, Test positive}\} = 0.08(0.10) = 0.092 \qquad \Pr\{\text{No disease, Test negative}\} = 0.92(0.90) = 0.828$$

What is the probability of testing positive? What is $\Pr\{\text{disease}|\text{test positive}\}$?

- A continuous random variable Y has a *density* $f(y)$. Examples: cholesterol, height, GPA, blood pressure.
- $\Pr\{a < Y < b\}$ is the area under the density curve $f(y)$ between a and b. Total area equals one.
- Note that $\Pr\{Y < a\} = \Pr\{Y \leq a\}$. Only with *continuous* random variables.

- A *discrete* random variable can only take on a countable number of values. Examples: number of broken eggs in a carton, number of earthquakes in a day.
- Finite discrete random variables have probability mass functions, e.g.

No. vertebrae y	$\Pr\{Y = y\}$
20	0.03
21	0.51
22	0.40
23	0.06

- Get probabilities $\Pr\{a \le Y \le b\}$ by summing probabilities in table for $a \le y \le b$.
- For discrete $\Pr\{Y < a\}$ will be different than $\Pr\{Y \le a\}$.

- Mean is now *weighted average*

$$\mu_Y = E(Y) = \sum y_i \text{Pr}\{Y = y_i\}.$$

- Variance is *weighted average* squared deviation about mean

$$\sigma_Y^2 = \sum (y_i - \mu_Y)^2 \text{Pr}\{Y = y_i\}.$$

- Standard deviation is σ_Y.

- Notation $Y \sim \text{binomial}(n, p)$. Y counts number of "success" trials out of n. Y can be $0, 1, 2, \ldots, n$.
- $\Pr\{Y = j\} = {}_nC_j \; p^j(1-p)^{n-j}$ for $j = 0, 1, \ldots, n$.
- $\mu_Y = E(Y) = n\ p$, $\sigma^2_Y = n\ p\ (1-p)$. How about σ_Y?
- R code: `dbinom` for $\Pr\{Y = j\}$, `pbinom` for $\Pr\{Y \leq j\}$.
- Use R!

4.2, 4.3: Normal distribution

- Used to model *many, many* different kinds of continuous data: cholesterol, eggshell thickness, creatinine clearance, $T_{1\rho}$ measurements from MRI, health care expenditures, etc.
- Notation: $Y \sim N(\mu, \sigma^2)$.
- μ is mean and σ^2 is variance of Y (requires calculus to show this). σ is standard deviation.
- Y is *continuous* random variable that can be any number $-\infty < Y < \infty$.
- Get probabilities from R using `pnorm`(y,μ,σ).

Normal distribution

μ and σ *are given to you* in Chapter 4.

$$
\begin{aligned}
\Pr\{a < Y < b\} &= \texttt{pnorm}(b,\mu,\sigma)\text{-}\texttt{pnorm}(a,\mu,\sigma) \\
\Pr\{Y < b\} &= \texttt{pnorm}(b,\mu,\sigma) \\
\Pr\{Y > a\} &= 1\text{-}\texttt{pnorm}(a,\mu,\sigma)
\end{aligned}
$$

How to get $\Pr\{Y < a \;\text{ or }\; Y > b\}$?

Normal distribution

- Let $Y \sim N(\mu, \sigma^2)$. Say we want y^* such that $\Pr\{Y < y^*\} = p$ where p is given.
- `qnorm`(p,μ,σ) gives y^*.
- y^* is called $p(100)$th percentile of Y.
- e.g. If $Pr\{Y \leq 10\} = 0.7$ then the $70th$ percentile of Y is 10.

6.3: Confidence interval for μ

- Data are generated $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$.
- Use $Y_1, \ldots, Y_n$ to come up with plausible range for μ, called a confidence interval.
- A 95% confidence interval for μ is given by

$$\bar{y} \pm t_{0.025} SE_{\bar{Y}} \text{ where } SE_{\bar{Y}} = \frac{s}{\sqrt{n}}.$$

- If $n < 30$ then data need to be $\boxed{\ ?\ }$ Can check this with a $\boxed{\ ?\ }$.
- A 99% confidence interval is $\boxed{\ ?\ }$ than a 95% confidence interval.
- True or False: a confidence interval always contains the unknown μ.
- W.S. Gossett invented the t-distribution doing quality control for the $\boxed{\ ?\ }$ brewery.

6.3: Confidence interval for μ

- t-distribution is used because we estimate σ by s in $SE_{\bar{y}}$; t has fatter tails than normal.
- Probability of confidence interval covering μ is 95% before we conduct experiment. After experiment the interval either covers μ or not, we don't know which.
- After we conduct experiment and compute $\bar{Y} \pm t_{0.025} SE_{\bar{y}}$, we call refer to "confidence" instead of "probability."
- R code: `t.test` to get the CI. `qqnorm` to test assess normality. Can also use `shapiro.test` to formally test data are normal.

6.7: Confidence interval for $\mu_1 - \mu_2$

- Now have *two random samples* from *two populations*:

$$\begin{array}{ll} \text{Population 1:} & \mu_1 \text{ and } \sigma_1 \\ \text{Population 2:} & \mu_2 \text{ and } \sigma_2 \end{array}$$

- Have sample statistics:

$$\begin{array}{ll} \text{Sample 1:} & \bar{y}_1 \text{ and } s_1 \text{ and } n_1 \\ \text{Sample 1:} & \bar{y}_2 \text{ and } s_2 \text{ and } n_2 \end{array}$$

- (p. 201) Standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

- 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{y}_1 - \bar{y}_2 \pm t_{0.025} SE_{\bar{y}_1 - \bar{y}_2}.$$

- The degrees of freedom for the t-distribution is (p. 206)

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

  R will do the work for us.

- `t.test`(sample1,sample2) or `t.test`(response$\sim$group).

μ_1 & μ_2 versus $\bar{y}_1$ & $\bar{y}_2$

- We do not know μ_1 or μ_2. These are unknown population means.
- We *do know* the sample means $\bar{y}_1$ and $\bar{y}_2$.
- Don't write something like $\mu_1 = 142$ miles per hour.
- Write: $\mu_1 =$ population mean tennis ball serve speed using the new composite racquet, $\mu_2 =$ population mean tennis ball serve speed using the old-type racquet.
- Perhaps "$\bar{y}_1 = 142$*mph estimates* μ_1*, the true typical serve speed using the new composite racquet.*"
- Can also use 'average' or 'mean' instead of 'typical'.

7.2: The t-test for $H_0 : \mu_1 = \mu_2$

- Initially consider $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$.
- $t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}}$ is the *test statistic*.
- The p-value is $\Pr\{|T| \geq |t_s|\}$, where T is a student t random variable with degrees of freedon df given by the Welch-Satterthwaite formula on slide 18.
- The P-value will be computed for you. Recall that the P-value is the probability of seeing two sample means $\bar{Y}_1$ and $\bar{Y}_2$ *even further apart than what we saw* given that $H_0 : \mu_1 = \mu_2$ is true.
- Reject $H_0 : \mu_1 = \mu_2$ in favor of $H_A : \mu_1 \neq \mu_2$ if P-value $< \alpha$ (otherwise accept H_0). α is called the *significance level* of the test, usually $\alpha = 0.05$.

- Pages 234–235 explains the following *important* rule:
- **Reject** $H_0 : \mu_1 = \mu_2$ **in favor of** $H_A : \mu_1 \neq \mu_2$ **at the 5% level whenever a 95% confidence interval for** $\mu_1 - \mu_2$ **does not contain zero.**

- Type I error is rejecting $H_0 : \mu_1 = \mu_2$ when H_0 **is true**.
- Type II error is *accepting* $H_0 : \mu_1 = \mu_2$ when H_0 **is false**.
- α is the probability of making a Type I error, usually 5%. This is called the significance level of the test.
- β is the probability of a Type II error. This number depends on the *true, unknown value of* $\mu_1 - \mu_2$. It also depends on σ_1, σ_2, n_1, and n_2.
- The Power of a hypothesis test is $\boxed{\ ?\ }$.

7.4: Association vs. causation

- When can we ascribe causality?
- A carefully controlled experiment creates two populations that are essentially identical except for an experimental manipulation (treatment vs. control). If we're careful, we can ascribe causality.
- An observational study simply collects some data and looks for association. Here, lurking variables, or unmeasured *confounders* may be *may be* driving any association that we see.

- P-value $\boxed{\quad ? \quad}$ the probability that H_0 is true.
- P-value $\boxed{\quad ? \quad}$ the probability of seeing a test statistic as extreme or more extreme than what we saw.

Good luck!

- Go over HW problems and solutions.
- 7 problems total; one is some true/false Q's. Several places where R will make quick, light work of your solution.
- 4 pages, problems on both side.
- Just handwrite your R code next to your answer. Maybe keep R code in a file in case I've got questions while grading.