Review for Exam I
Stat 205: Statistics for the Life Sciences

Tim Hanson, Ph.D.

University of South Carolina

Logistics...

* Thursday, February 17, 3:30–4:45.
* Open book. Bring a calculator.
* Problems will be patterned after homework problems.
* *Be on time*.
* Note that scanned pages from the textbook's solutions manual are posted on the course website. Also quiz solutions.
* Each of exams I, II, and III are worth 20%. Your quizzes are worth 40%.

2.2, 2.3: Histograms, distributions, skew and modality

* Have data $y_1, y_2, \ldots, y_n$; want to describe it with pictures and numerical summaries.
* If data categorical, can make a bar chart (like a histogram) or a pie chart. Can table 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 shape: unimodal, bimodal, multimodal.
* Histogram skew: left skew, right skew, symmetry.
* HW 2.4, 2.7, 2.8, 2.10 (also describe the shape of the histograms in terms of skew and modality).

2.4, 2.5, 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.
* $y_{(1)}, Q_1, Q_2, Q_3, y_{(n)}$ is 5 number summary, used to make 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."
* HW 2.18, 2.19, 2.31, 2.32 (for 2.31 and 2.32 construct boxplots as in class, but also determine if there are any outliers by computing the upper and lower fence), 2.46(a,b,d), 2.47, 2.48.

3.3, 3.5: Probability

* 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. (LTP) $\Pr\{A\} = \Pr\{A|B\}\Pr\{B\} + \Pr\{A|B^C\}\Pr\{B^C\}$.
  5. (Bayes') $\Pr\{A|B\} = \Pr\{B|A\}\Pr\{A\}/\Pr\{B\}$.
  6. (compliment rules) $\Pr\{A^C\} = 1 - \Pr\{A\}$ *and* $\Pr\{A^C|B\} = 1 - \Pr\{A|B\}$.
  7. (independence) A and B are independent if $\Pr\{A\} = \Pr\{A|B\}$. This happens when $\Pr\{A \text{ and } B\} = \Pr\{A\}\Pr\{B\}$.
* HW 3.7, 3.8, 3.11, 3.12, 3.13, 3.14.

3.6: Continuous random variables, densities

* 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.
* Weird stuff: $\Pr\{Y = a\} = 0$, $\Pr\{Y < a\} = \Pr\{Y \leq a\}$, etc. Only with *continuous* random variables.
* HW 3.16, 3.17, 3.46.

3.7: Discrete 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 $\Pr\{Y \text{ in } A\}$ by summing probabilities in table for y in A.
* *Now* $\Pr\{Y < a\}$ will be different than $\Pr\{Y \leq a\}$.

3.7: Mean and variance of a discrete random variable

* Mean is now *weighted average*

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

* Variance is *weighted average* squared deviation about mean

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

* Standard deviation is σ_Y.
* HW 3.18, 3.20, 3.21, 3.22, 3.23.

3.8: Binomial distribution

* "BInS"
* Notation $Y \sim \text{binomial}(n, p)$. Y counts number of "success" trials out of n. Y can be $0, 1, 2, \ldots, n$.
* PMF: $\Pr\{Y = j\} = {}_nC_j \; p^j (1-p)^{n-j}$ for $j = 0, 1, \ldots, n$.
* Table 2 (p. 674) has ${}_nC_j$.
* Use PMF to get probabilities $\Pr\{Y \text{ in } A\}$.
* $\mu_Y = E(Y) = n\,p$, $\sigma_Y^2 = n\,p\,(1-p)$.
* HW 3.27, 3.28, 3.29, 3.30, 3.32, 3.33.

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 by *standardizing* Y and using table on inside cover of textbook.

Normal distribution

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

$$\begin{aligned}\Pr\{a < Y < b\} &= \Pr\left\{\frac{a-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right\} \\ &= \Pr\left\{\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right\} \\ &= \Pr\left\{Z < \frac{b-\mu}{\sigma}\right\} - \Pr\left\{Z < \frac{a-\mu}{\sigma}\right\}.\end{aligned}$$

*

$$\Pr\{Y < b\} = \Pr\left\{Z < \frac{b-\mu}{\sigma}\right\}.$$

*

$$\Pr\{Y > a\} = 1 - \Pr\left\{Z < \frac{a-\mu}{\sigma}\right\}.$$

Normal distribution

* HW 4.3, 4.4 (use 4.3), 4.9, 4.10 (use 4.9), 4.13.

---

* Let $Y \sim N(\mu, \sigma^2)$. Say we want y^* such that $\Pr\{Y < y^*\} = p$ where p is given.
* Use table inside front cover "in reverse" to find z^* such that $\Pr\{Z < z^*\} = p$.
* Then $y^* = \mu + z^*\sigma$.
* y^* is called pth percentile of Y.
* HW 4.11, 4.14, 4.15.