Review for Final M1100 Section 3

1. **Observational study**
   - population, population parameter
   - sample, sample statistic
   - samples: random, systematic, convenience, cluster, stratified, representative versus biased
   - case-control study
2. **Experiment**
   - treatment group, control group, double-blind
   - confounding factors
3. A **variable** is something that one measures or observes in different individuals or in different objects or on various occasions.
   - Be able to give and to recognize examples.
   - The class of all those things for which a variable is measured is the *reference class* of the variable. The members of the reference class are called *observational units*.
   - When we make a measurement or observation, we determine a *value* of whatever variable it is that we are observing. A variable takes different values at different times or for different individuals.
   - measurement procedure
   - qualitative (continuous, discrete), quantitative, boolean
   - In describing a variable, one needs to consider the kind of variable, the reference class and the procedure used to make the measurement. If we are clear and explicit in describing these things, then we have a better chance of avoiding error or misinterpretation.
   - bins
4. The **distribution** of a variable is a record of the frequency with which each of the possible values occurs within the reference class.
   - frequencies and relative frequencies, tables of frequencies and relative frequencies
   - bar graph, dot chart, histogram, stem-and-leaf plots
   - pie charts
5. **Probability basics**
   - random, random processes and examples (coin flips, dice)
   - relative frequency (as an interpretation of what probability means); Law of Large Numbers
   - probability distribution (page 238; related to relative frequency distribution)
6. **Probability models**
   - outcome/event models
     - probabilistic experiment (Example: flip 6 coins and a roll a die)
     - outcome (Example: $HHTTTH4$)
     - sample space ($\{TTTTTT1, TTTTTT2, ..., HHHHHH6\}$. There are 384 outcomes.)
     - event ("flipped more heads than tails" or "there are more heads than the number showing on the die")
     - the condition of "equal likelihood" (all 384 outcomes are equally likely) and how it's used
     - the complement of an event (page 235-236)
   - tree models (for multi-stage processes) (Be able to draw the tree diagram for the experiment of taking beads (of some specified number and color) from a jar, with or without replacement.)
   - reference class models (In a given town, 1/4 of the people are democrat and 3/4 are republican. One in 4 is black and the remaining citizens are white. If 1 in 8 is a black republican, then what proportions are the other possibilities? Is race associated with political party?)
7. **The binomial distribution**
   - Binomial coefficients: "7 choose 3" = $\binom{7}{3}$ = the number of H-T-sequences of length 7 in which 3 Hs appear (or the number of 0-1-sequences of length 7 in which 3 1s appear, or *etc.*).

- If the probability is 1/3 that a random individual from Weston (a city) is Republican, then what is the probability that in random sample of 8 there will be 5 or more Republicans. *Answer.* The probability that there will be exactly 5 Republicans in the sample is

$$\binom{8}{5}(1/3)^5(2/3)^3 = 56 \cdot (1/243) \cdot (8/27) \cong 0.068.$$

Similarly, the probability of exactly 6 is $\binom{8}{6}(1/3)^6(2/3)^2 = 28 \cdot (1/729) \cdot (4/9) \cong 0.017$, the probability of exactly 7 is $\binom{8}{7}(1/3)^7(2/3) = 8 \cdot (1/2187) \cdot (2/3) \cong 0.0024$, and the probability of exactly 8 is $(1/3)^8 \cong 0.00015$. We add these numbers and get 0.088 (rounded to the nearest thousandth) as the probability of 5 or more. (That's 8.8%.)

8. Hypothesis testing.

- **Text.** Chapter 9, especially pages 360–365.
- **Vocabulary.** Null hypothesis, alternative hypothesis, reject the null hypothesis, statistical significance, P-value (probability of data at least as extreme, assuming the null hypothesis).
- **Concepts.** The null hypothesis is a set of assumptions about the factors that influence our data. We reject the null hypothesis when the data that we have obtained would be unlikely if the null hypothesis were true. The statistician quantifies how unlikely, and decides a level at which to reject.

9. Determining Population Proportions.

- **Text.** Chapter 8, especially pages pages 333–335 and pages 347–350.
- **Vocabulary.** Population proportion, sample proportion, distribution of sample proportions (p. 335), margin of error, level of confidence.
- **Concepts.** "Population proportion" refers to the fraction of a population with some characteristic. One estimates population proportion from sample proportion, recognizing that samples seldom have exactly the same proportion as the population from which they are drawn, but also recognizing that the sample proportions will be close. The range in which sample proportions are likely to fall depends upon the sample size. This is described by the sampling distribution. From the sampling distribution, we can compute margin of error and level of confidence. Level of confidence really means how often we expect our sample proportion to differ from the population proportion by no more than the margin of error, if we select random samples of the same size over and over.

10. Correlation.

- **Text.** Chapter 7, especially pages pages 274–285.
- **Vocabulary.** Scatter diagram, correlation, positive/negative correlation, strength of correlation, r-value.
- **Concepts.** A correlation between two quantitative variables exists the values of one variable enable us to predict the values of the other. Correlations are often apparent to the eye when data is displayed in a scatter diagram. When both variables rise together, we say there is a positive correlation; when one rises as the other falls, we have a negative correlation. When data points are scattered widely, the correlation is weak. When all data points lie very close to a line, the correlation is strong. For any given data set, statisticians compute an r-value, which is a measure of the strength of the correlation. An r-value close to 1 indicates a strong positive correlation; r-values close to 1 are seldom seen in random data sets, especially when the data set contains many points. Statisticians can calculate the probability that a random data set (of a given size) will have an r-value exceeding a specified value. This is the basis for evaluating the statistical significance of a given r-value.