

## Math 203 – Review for Exam II

Your second exam covers Chapters 5-8 of your textbook. This review sheet is meant to help you, but is not meant to be a substitute for studying from the book. In particular, I have tried to include all the topics we covered, but it is possible I accidentally left something out. If you think I have left something out, tell me. It is not safe to assume that any omitted topics are not covered on the exam.

Good luck studying!

**Designing Experiments (Chapter 5):** The overall idea from this chapter is that we would like to know some information about a *population*, but because of time, money, or logistical restrictions, we can't get the information from every individual in the population. So we choose a smaller group from the population, a *sample*, and collect the desired information from the individuals in the sample. We would then like to use this information from the sample as an estimate for the information about the whole population. The catch is that we need to choose our sample wisely, i.e. large enough, without bias, etc, and set up our survey or experiment carefully so that we can justifiably make conclusions about our population based on knowledge of our sample. You should be completely comfortable with the following concepts.

- Be able to identify the *population* and the *sample* in a given situation.
- Be able to recognize bad sampling methods and detect sources of *bias*.
- Know and understand the definition of a *simple random sample* of size  $n$ .
- Be able to explain the difference between bias and *variability*. Which one can be avoided by using a simple random sample to perform your experiment?
- Know how to use a table of random digits to select a random sample from a small population.
- Be able to calculate the sample proportion given the sample size and number of positive responses.
- Be able to describe a design for a comparative experiment; for example, design an experiment that tests to see whether listening to classical music before a taking a math exam improves your score. Know why a control group is necessary in a comparative experiment. Understand the following terms and be able to give examples of them:
  - confounded variables
  - placebo effect
  - double-blind experiment

**Describing data (Chapter 6):** This chapter deals with how to describe a list of data. Since lists of numbers don't reveal much information, we need to have a useful way to summarize numerical data to help us see any patterns. The pattern of the values that a variable takes on is called a distribution. We use three characteristics to describe a distribution:

- Shape: Is the distribution symmetric or skewed? Does it have any outliers?
- Center: What is the most typical value in the distribution?
- Spread: What is the range of values that the variable takes on?

We can describe the distribution either visually by means of a graph, or numerically. You should be able to do the following:

- Construct a *histogram* for a small data set.
- Describe in words the shape, center, and spread of a distribution given its histogram.
- Make a *stemplot* for a list of data.
- After sorting a list of numbers from smallest to largest, determine its *median*.
- Calculate the lower and upper *quartiles* of a list of data. (What proportion of all the data points lie between the two quartiles?)
- Give the *five-number summary* for a list of data.
- Display the five-number summary using a *boxplot*.
- Calculate the *mean* of a list of data.
- Know how the shape of the distribution affects the relationship between the mean and the median.
- Be able to make a *scatterplot*, and given a scatterplot, be able to sketch in the *least squares regression line*. Understand the idea and the point behind least squares regression lines. In particular, be able to estimate data points by using both the graph and the equation of the regression line.

**Probability (Chapter 7):** Chapter 7: There are three big ideas in this chapter: *probability*, *normal distributions*, and the *Central Limit Theorem*.

- Probability can be thought of as a mathematical way to assign chances to random phenomena, where individual outcomes are uncertain but in the long run, a regular pattern describes how frequently each outcome occurs.
- A probability model for a random phenomenon is the *sample space*, the set of all possible outcomes, and a way of assigning probabilities to groups of outcomes.
  - Be able to list the sample space for a given random phenomenon.
  - Know the two *laws* of assigning probabilities:
    1. Every probability is a number between 0 and 1.
    2. The sum of the probabilities for all the outcomes in the sample space is exactly 1.

- Be able to use the laws of probability to determine if a probability model is *valid* or not.
  - Given a probability model, be able to compute the *expected value* or *mean of the random variable* of the model.
- The *law of large numbers* says that the the more times you repeat a random phenomenon, the closer the sample mean gets to the mean of the probability model  $\mu$ . For example, we can compute the mean of the probability model for rolling a fair die to be  $\mu = 3.5$ . Suppose we roll a die 100 times and compute the sample mean of our 100 rolls. We expect that number to be pretty close to 3.5. But if we roll the die 10,000 times and compute the sample mean of our 10,000 rolls, we are even more sure that our answer will be close to 3.5. That is the law of large numbers at work.

*Normal curves* are curves that pop up often in statistics because many sets of data are approximately described by normal distributions. All normal cures are symmetric and bell-shaped, with tails that fall off rapidly on both sides. The area under a normal curve is always 1. A normal distribution is totally determined by its *mean*  $\mu$  and *standard deviation*  $\sigma$ . Here are some things you should be able to do with normal distributions:

- Given a graph of a normal curve, be able to find the mean and standard deviation of the distribution. Recall the mean lies at the center of symmetry of the curve and the standard deviation can be found by locating the distance from the mean to the point where the curve changes from curving downward to curving upward.
- Given a mean and standard deviation, be able to draw the normal curve that it describes. Remember to just draw a bell shaped curve and then scale the axis to fit the numbers you are given.
- Be able to apply the “68-95-99.7” rule for a normal distribution. This says that 68% of all observations happen within 1 standard deviation of the mean, 95% of all observations happen within 2 standard deviations of the mean, and 99.7% of all observations lie within 3 standard deviations of the mean. In particular, if you know the mean and standard deviation of a normal distribution, you can be pretty sure that almost all of your observations will lie in the range from  $\mu - 3\sigma$  to  $\mu + 3\sigma$ .

The *Central Limit Theorem* says that the sampling distribution of any random phenomenon is approximately normal, if you average over a large number of independent trials. Suppose  $\mu$  and  $\sigma$  are the mean and standard deviation of the sampling distribution of a single trial of the random phenomenon, and consider the sampling distribution of  $\bar{x}$ , the average of  $n$  trials of the random phenomenon. Then the mean of the sampling distribution of  $\bar{x}$  is still  $\mu$  and the standard deviation of the sampling distribution of  $\bar{x}$  is  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . Intuitively, this says that if you increase the number of trials (or increase the sample size), then the center of the new sampling distribution is the same but the spread is smaller.

**Statistical Inference (Chapter 8):** This chapter describes two different statistics whose sampling distributions are given by normal curves (or almost normal curves): *sample proportion*  $\hat{p}$  and *sample mean*  $\bar{x}$ . You should:

- Know the difference between a *parameter* and a *statistic*.
- This is a repeat from Chapter 5, but be sure you can calculate a sample proportion.

Suppose we chose a simple random sample of size  $n$  from a population with an unknown population proportion  $p$ . Then we know the sampling distribution of the sample proportion  $\hat{p}$  is almost normal with  $p$  as the mean of the sampling distribution and  $\sigma_{\hat{p}} = \sqrt{\frac{p(100-p)}{n}}$  as the standard deviation. Since we don't know what  $p$  is, but we do know an approximation ( $\hat{p}$ ) for  $p$ , we approximate  $\sigma_{\hat{p}}$  with  $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(100-\hat{p})}{n}}$ .

Suppose we take one sample of size  $n$  and compute  $\hat{p}$  as our best guess for  $p$ . How close to the true  $p$  is our guess likely to be? To answer this question, we compute a 95% confidence interval for the population proportion  $p$ . This is a range of values computed in such a way that for each different sample proportion we use, 95% of the time the confidence interval about that sample proportion will actually contain the true value of the population proportion. Recall that the method for calculating this interval comes from using the "95" part of the "68-95-99.7" rule.

- Given a sample size  $n$  and a value for the sample proportion  $\hat{p}$ , be able to compute the 95% confidence interval for  $p$ .

We can use the same ideas to estimate a population mean using a sample mean. Suppose we calculate  $\bar{x}$ , the sample mean, from a simple random sample of size  $n$  from a large population with unknown mean  $\mu$  and known standard deviation  $\sigma$ . Then the sampling distribution of  $\bar{x}$  is normal (as long as the distribution of the population is normal). The mean of the sampling distribution of  $\bar{x}$  is  $\mu$  and the standard deviation is  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ .

- Given a sample mean  $\bar{x}$  calculated from a random sample of size  $n$  from a normal population with unknown mean  $\mu$  and known standard deviation  $\sigma$ , be able to compute the 95% confidence interval for  $\mu$ .