1 Supplementary notes on probability

1.1 Random samples

A **simple random sample** of n items is one in which each all possible subsets of size n are equally likely. This can arise from drawing names from a hat one at a time, where at each draw all of the remaining names in the hat are equally likely and there is no dependence on previous draws from the hat.

Samples that are not random are prone to bias. (Recall the classroom exercise where the class distribution of average rectangle size from judgment samples was centered to the right of the center of the distribution of random sample means.

1.2 Probability

Probability measures the likelihood of events on a scale from 0 to 1.

A **random variable** is a variable whose value is randomly determined.

The probability distribution of a **discrete random variable** places discrete "chunks" of probability at specific locations. The distribution can be described with a table that lists the possible values of the random variable and the probability associated with each. Sometimes, we use a formula to specify the probability of each possible outcome. It is always the case with discrete random varaibles that the sum the probabilities of each possible outcome is one.

Probability distributions of **continuous random variables** describes how thick a probability "dust" is spread over the line. A **probability density curve** is a nonnegative function where the total area under the curve is one that has the property that the area under the curve between two points a and b is the probability that the random variable is between a and b.

1.3 The binomial distribution

The binomial distribution arises from counting the number of heads in a prespecified number of coin tosses. This is a model for the way that data is produced for a vast number of examples in statistics. In particular, we will use this model when examining the proportion of a random sample that belongs to a particular category.

Every binomial random variable is described by two parameters: n is the number of trials and p is the probability that an individual trial is a success.

The binomial setting: You may recognize a setting in which the binomial distribution is appropriate with the acronym BINS: binary outcomes, independent trials, n is fixed in advance, same value of p for all trials.

A trial has one of two possible values. One is called a "success" and the other is called a "failure". We want to count the number of successes.

The binomial distribution is appropriate when we have this setting:

1. there are a fixed number of trials;
2. there are two possible outcomes for each trial;
3. the trials are independent of one another;
4. there is the same chance of success for each trial;
5. we count the number of successes

The binomial probability formula for exactly j successes (and $n-j$ failures) in n independent trials with success probability p is

$$\Pr\{Y = j\} = {}_nC_j p^j (1-p)^{n-j} \qquad \text{for } j = 0, 1, \ldots, n \text{ where } {}_nC_j = \frac{n!}{j!(n-j)!}$$

There is no simple formula to sum binomial probabilities: to calculate the probability that a binomial random variable is one of several outcomes, you need to compute the outcomes individually and sum them.

1.4 The normal distribution

Many naturally occuring variables have distributions that are well-approximated by a "bell-shaped curve", or a normal distribution. These variables have histograms which are approximately symmetric, have a single mode in the center, and tail away to both sides. Two parameters, the mean μ and the standard deviation σ describe a normal distribution completely and allow one to approximate the proportions of observations in any interval by finding corresponding areas under the appropriate normal curve.

In addition, the sampling distributions of important statistics such as the sample mean are approximately normal for moderately large samples for many populations.

Characteristics of all normal curves:

- Each bell-shaped normal curve is symmetric and centered at its mean μ.
- The total area under the curve is 1.
- About 68% of the area is within one standard deviation of the mean, about 95% of the area is within two standard deviations of the mean, and almost all (99.7%) of the area is within three standard deviations of the mean.
- The places where the normal curve is steepest are a standard deviation below and above the mean ($\mu - \sigma$ and $\mu + \sigma$).

Standardization: In working with normal curves, the first step in a calculation is invariably to standardize.

$$z = \frac{x - \mu}{\sigma}$$

This z-score tells how many standard deviations an observation x is from the mean. Positive z-scores are greater than the mean, and negative z-scores are below the mean.

If the z-score is known and the value of x is needed, solving the previous equation for x gives

$$x = \mu + z \times \sigma$$

Reading the algebra, this simply states that x is z standard deviations above the mean.

The standard normal distribution: Areas under all normal curves are related. For example, the area to the right of 1.76 standard deviations above the mean is identical for all normal curves. Because of this, we can find an area over an interval for any normal curve by finding the corresponding area under a standard normal curve which has mean $\mu = 0$ and standard deviation $\sigma = 1$.

Using the normal table: The standard normal table is located in the inside cover of your textbook. It tells you the area to the left of z. Because the normal curve is symmetric and the total area under the curve is 1, this is sufficient to find the area under the curve over any interval.

You will need to be able to use the table to find areas when the numbers on the axis are known, and to be able to use the table to find numbers on the axis when areas are known.

It is helpful to draw a sketch of a normal curve in working out problems. Draw one axis with the units of the problem. Draw a second axis with standard units.

The normal approximation to the binomial distribution: (This is useful background information, but we will not test it. You have access to software to compute sums of binomial probabilities exactly, so there is no need to use a normal approximation except as a check.)

Consider this example: find the probability that there are 410 or more successes in 500 independent trials when the probability of success on a single trial is 0.8. An exact expression of this probability is

$$P(X \geq 410) = \sum_{x=410}^{500} \frac{500!}{x!(500-x)!}(0.75)^x(0.25)^{500-x}$$

Even with a calculator, this is an intimidating computation. Statistical software gives the answer numerically as 0.1437.

When the number of trials is sufficiently large, a graph of the binomial distribution resembles the familiar shape of the normal distribution. When the number of expected successes and failures in the trials are each sufficiently large, an area under the normal curve will be a good numerical approximation to the exact binomial computation. A good rule of thumb is that if $np \geq 5$ and $n(1-p) \geq 5$, then an approximation will be good (usually to about two or three digits).

There are many normal distributions, each described completely by μ and σ. To find the one which matches a binomial distribution best, let $\mu = np$ be the mean (or expected value) and $\sigma = \sqrt{np(1-p)}$ be the standard deviation. This μ will be the balancing point of the binomial distribution. You expect the random variable to be close to μ, but it may not be μ exactly. In fact, μ may not even be an integer and may not be a possible value of the random variable. Although the random variable could take on any integer value between 0 and n, it is highly unlikely to be more than a two or three standard deviations σ from μ. You may interpret σ as the rough size of a typical deviation from the mean.

Follow these steps to apply the approximation.

1. Check to see if n is large enough for the given p.

2. Write the outcome in the form $a \leq X \leq b$ where X is the random number of successes in the n trials.

3. Find μ and σ.

4. Find the area between $a - 0.5$ and $b + 0.5$ under a normal curve with mean μ and standard deviation σ.

For the problem above, $500(.8) \geq 5$ and $500(.2) \geq 5$ so the sample size is easily large enough to get a good approximation. We want to find $P(410 \leq X \leq 500)$ with $n = 500$ and $p = 0.8$. $\mu = 500(.8) = 400$ and $\sigma = \sqrt{500(.8)(.2)} = 8.94$. The z-score of 409.5 is $(409.5 - 400)/8.94 = 1.06$. The z-score of 500.5 is $(500.5 - 400)/8.94 = 11.2$. The area between 1.06 and 11.2 under the standard normal curve is $.5000 - .3554 = .1446$, which is close to the exact .1437.

1.5 The Poisson distribution

(You may see the Poisson distribution in biological problems. We did not cover it in class.)

We will use the Poisson distribution to describe the probability of counting the number of rare events. It depends on a single parameter λ and satisfies

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

for x a nonnegative integer. When n is very large and p is very small, the binomial probabilities are numerically very close to the Poisson probabilities with $\lambda = np$.

For example, say that over the past several years, there have been an average of 13 new cases of cancer of the esophogus diagnosed in Pittsburgh. The probability that any individual gets a this type of cancer is rare. The population of Pittsburgh is fairly stable so it is not too outrageous to make the approximation that it stays constant. The population of the Pittsburgh metropolitan area is the large number n and the probability an individual is diagnosed with this form of cancer is the small number p. We may expect that the number of diagnoses in the upcoming year is a random Poisson variable with $\lambda = 13$. If so, the probability that there are exactly 10 new cases would be

$$P(X = 10) = \frac{e^{-13}(13)^{10}}{10!} = 0.086$$

This is far more reasonable to compute than plugging into the binomial distribution formula with $n = 2,500,000$ and $p = 13/n$.

1.6 Sampling distributions

Sampling distributions are probability distributions of statistics that may be calculated from random samples. I find it extremely helpful to keep in mind this metaphor. A population is like big box full of balls, where the balls may be colored (in the case of categorical variables) or have a number written on them (for a quantitative variable). From a random sample of size n we calculate a sample statistic, such as a sample proportion or a sample mean. Write this statistic on a ticket and place it into another box. Repeat this process indefinitely. The distribution of values in the new box are the sampling distribution of the statistic.

For sample proportions, the sampling distribution is just the binomial distribution except that the possible values are $0, 1/n, 2/n, \ldots, 1$ instead of $0, 1, 2, \ldots, n$. For the sampling distribution of the sample mean, these three important facts hold.

1. The mean of the sampling distribution $\mu_{\bar{Y}}$ equals the population mean μ.
2. The standard deviation of the sampling distribution $\sigma_{\bar{Y}}$ is $\sigma/\sqrt{n}$, so it decreases with the square root of the sample size.
3. For (almost) any population, the shape of the sampling distribution is approximately normal if the sample size n is big enough. (Big enough depends on the specific population.)