Chapter 5: Sampling Distributions

Let $x_1, x_2, \cdots, x_n$ denote the outcomes of a random process. A statistic is any function of these $x_i$’s.

**Example 1:**
Suppose you take $n$ samples of a random variable $X$, with values $x_1, x_2, \cdots, x_n$. Examples of statistics include the empirical mean

$$\bar{x} := \frac{\sum_{i=1}^{n} x_i}{n}$$

and the empirical variance

$$s^2 := \frac{1}{n-1} \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2}.$$  

Note that these statistics have interpretations: the mean gives an idea of the “center” of the data, while the variance gives some notion of “spread”. Note too that we are in no way limited to these statistics: since there are infinitely many functions involving the $x_i$, there are infinitely many possible statistics.

**Exercise 1:**
Write your own statistic. Your statistic should be given by an explicit formula involving the $x_1, x_2, \cdots, x_n$, and you should comment on what qualitative feature of the underlying random process it characterizes.

Since the number $x_i$ are random, so are the statistics based on these numbers. On a practical level, the most important question we can ask about any random quantity is what values it can assume and how often it assumes these values, i.e. what is its distribution. The goal of this section is to calculate the distributions (also known as sampling distributions) of several important statistics.

**Important and Potentially Confusing Point:** Sometimes we use the same term to describe both a statistic (i.e. a random quantity) and a parameter (i.e. a fixed number that describes a random process.) For example, the mean of a normal random variable $X \sim N(0, 1)$ is just the number 0. However, if you take $n$ samples from this distribution $x_1, x_2, \cdots, x_n$ and form the quantity $(x_1 + \cdots + x_n)/n$, we also call this the mean. In the first usage, the word mean refers to the theoretical value of the average as the number of samples goes to infinity; in the second usage, the word refers to the actual value of the average for a finite number of samples.

1. **Counts and Proportions**

A binary random variable is one that can assume one of two possible values (think “heads” or “tails” for a coin flip.) Suppose we sample a binary random variable a total of $n$ times, where each sample is independent of every other, and the distribution of the binary variable remains the same from observation to observation. If we set $X$ equal to the **count** of the number of times one particular value appears, then $X$ is called a binomial random variable.
Example 2:
Consider a coin toss. It has two possible outcomes, heads or tails. If we toss a coin multiple times, each toss is independent of every other, and the probabilities of getting heads or tails don’t change from toss to toss. If we decide in advance to toss the coin 10 times and let $X$ denote the number of heads, the $X$ is a binary random variable.

Note that in this coin tossing experiment, the total number of heads is going to be a number between 0 and 10. Each of these numbers occurs with a certain probability. Specifying these probabilities defines the binomial distribution. In general, we use the symbol $B(n, p)$ to denote the distribution for a binomial random variable with $n$ trials, each with probability $p$ of success.

Exercise 2:
Everyone flip a coin 10 times. Record the number of heads. Form an empirical histogram of the outcomes. Convince yourself that this histogram is a good approximation of some distribution $B(n, p)$. Question: what are $n$ and $p$?

A coin toss is an example of a random event that has exactly two possible outcomes. Sometimes we collect data that is more complex, i.e. has more than two possible outcomes. The binomial distribution can still sometimes yield interesting facts about this data, but the analysis requires that we first transform the data (and our questions) into a form that fits the binomial model.

Exercise 3:
Find a partner and play rock-scissors-paper a total of five times. After each game, record who played what symbol, and which symbol won. Note that your outcomes are complex, i.e. they are not simple “success”-“failure”, etc. Find a way to map your data into a new data set that is does fit the binomial model. (One example: for each game, you could ask “did heads win?” The answer will be yes or no. Think of your own question.) Use the transformed data to calculate the value of a binomial statistic.

Suppose $X$ is a random variable distributed binomial $B(n, p)$. Then $X$ can assume values between 0 and $n$. To calculate the probabilities of each of these outcomes, there are two options: either we can use a table, or we can use a formula. The table is in the back of the book, while the formula is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Both methods are rather unpleasant.

One of the most common applications of binomial sampling is to sample a large population, some portion of which has some feature (e.g. has HIV, is a smoker, or likes dogs.) Often, the interest is in knowing what the proportion is. Standard protocol is to do a simple random sample of $n$ individuals, and let $X$ denote the number of individuals that happen to have had the feature. Then $X$ is distributed $B(n, p)$. We don’t know $p$, of course, but we use $\hat{p} = X/n$ as an estimate. We can calculate the mean and the standard deviation of
both $X$ and $\hat{p}$ using standard sum rules:

\[
\begin{align*}
\mu_X &= np \\
\sigma_X &= \sqrt{np(1-p)} \\
\mu_\hat{p} &= p \\
\sigma_\hat{p} &= \sqrt{\frac{p(1-p)}{n}}
\end{align*}
\]

**Example 3:**
Suppose you roll a four-sided die 5 times. Let $X$ denote the number of 1’s you roll. What are the mean and standard deviation of $X$?

**Exercise 4:**
Suppose you wanted to assess whether or not a certain 6-sided die was fair. Design an experiment to estimate $p_6$, the probability of rolling a 6. Suppose you wanted to make sure that the standard deviation of your estimate was less than .01. How could you tailor your experiment to meet this goal?

The above formulas for mean and standard deviation are cumbersome. Luckily, we can approximate binomial variables with normal variables, assuming $n$ is large enough. In particular, if $X$ is distributed $B(n,p)$, and $np \geq 10$, then we have the following approximations:

\[
X \sim N(np, \sqrt{np(1-p)}) \quad \hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})
\]

**Exercise 5:**
Consider the previous exercise. How many times did you need to roll the die? Using the normal approximation, give 95% confidence bounds on your estimate (i.e. find a window $[a, b]$ within which you are 95% sure the true value of $p$ should lie.)

2. **Sample Means**

An estimator is a statistic designed to approximate some “actual” quantity. For example, the sample mean $\overline{x}$ is designed to estimate the true mean $\mu$, and the sample standard deviation $s$ is designed to estimate the true standard deviation $\sigma$.

An estimator is **unbiased** if its mean is identical to the value of the thing it is supposed to be estimating. Otherwise, the estimator is biased.

It is easy to show that the (true) mean and standard deviation of the estimator $\overline{x}$ are given by

\[
\mu_{\overline{x}} = \mu \quad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}
\]

In other words, the sample mean is an unbiased estimator of the mean, and its various decreases the more observations are taken.

Note that, in general, merely knowing the mean and standard deviation does not give the full distribution of a random variable. An exception to this claim is the normal distribution, which is entirely characterized by these quantities. It is an amazing fact that the distribution of $\overline{x}$ gets closer and closer to a normal distribution as $n$ gets larger and larger. This is the content of the **central limit theorem**.

**Example 4:**
Computer demo.
Example 5:
Suppose you keep watch over a heard of 100 sheep. Each sheep needs about 30 minutes of sheering per week, with standard deviation 1 hour. What is the probability that in a given week the total time you need to invest in sheering your sheep is larger than 60 hours?

Exercise 6:
Repeat this exercise with sheep replaced by lions and sheering replaced by taming, with mean 30 minutes and standard deviation 30 minutes.