1 Introduction

We define a set of random variables $X_1, \ldots, X_n$ to be a random sample from a population if all the $X_i$ are independent and each $X_i$ has the same distribution. Our goal in this handout is to describe the distribution of the sample sum and the sample mean. The assumption that $X_1, \ldots, X_n$ is a random sample will be used in two ways. First, since all the $X_i$ have the same distribution, they will have the same mean and variance. For simplicity, we will call $E[X_i]=\mu$ and $V[X_i]=\sigma^2$. Second, since all the $X_i$ are independent, we have a formula for the mean and variance of a linear combination. Recall

$$E[a_1X_1 + \cdots + a_nX_n + b] = a_1E[X_1] + \cdots + a_nE[X_n] + b$$
$$V[a_1X_1 + \cdots + a_nX_n + b] = a_1^2V[X_1] + \cdots + a_n^2V[X_n]$$

2 Expectation and Variance of $\overline{X}$

The sample sum $\sum_{i=1}^{n}X_i$ is a particularly simple linear combination of the $X_i$ values. The $a_i$ values are all 1 and $b$ is 0. Therefore (recall all the $X_i$ have the same mean and same variance)

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] + 0 = n\mu$$
$$V[X_1 + \cdots + X_n] = V[X_1] + \cdots + V[X_n] = n\sigma^2$$

In addition, observe that $\overline{X}$ is a linear transformation of $\sum_{i=1}^{n}X_i$ (specifically, $\overline{X} = (1/n) \sum_{i=1}^{n}X_i + 0$), so

$$E[\overline{X}] = \frac{1}{n} E\left[ \sum_{i=1}^{n}X_i \right] + 0 = \frac{n\mu}{n} = \mu$$
$$V[\overline{X}] = \frac{1}{n^2} V\left[ \sum_{i=1}^{n}X_i \right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

The expectation and variance of $\overline{X}$ are fundamental to statistical inference. Because $E[\overline{X}] = \mu$, the sample mean provides an unbiased estimator of the population mean. In other words, on average the sample mean will be equal to the population mean. That is NOT to say the sample mean is exactly equal to the population mean, but that the center of its distribution is the population mean $\mu$. Because $V[\overline{X}] = \frac{\sigma^2}{n}$, the variance of $\overline{X}$ decreases as the sample size increases.

These two facts together imply something called the Law of Large Numbers. We just established that, on average, the sample mean is equal to the population mean AND the variance decreases (to 0) as the sample size $n$ increases. Together, these facts imply that as $n$ increases, the sample mean $\overline{X}$ will get closer and closer to the population mean $\mu$. In large samples, the sample mean $\overline{X}$ is extremely likely to be almost exactly equal to $\mu$.

3 Central Limit Theorem

Recall from a previous handout that a linear combination of independent normal random variables is normally distributed. Thus, if $X_1, \ldots, X \sim N(\mu, \sigma^2)$, then the sample mean will be normally distributed, with the mean and variance we computed in the previous section.
THEOREM If \( X_1, \ldots, X \sim N(\mu, \sigma^2) \), then
\[
\bar{X} \sim N(\mu, \sigma^2/n)
\]
The **Central Limit Theorem** states that, for large samples, this result holds MUCH more generally. Suppose that the sample size \( n \) is large (the rule of thumb is \( n \geq 30 \)). Then the sample mean is approximately normally distributed **no matter how** the individual \( X \) are distributed.

**THEOREM (Central Limit Theorem)** Suppose \( X_1, \ldots, X \) are independent, each with \( E[X_i]=\mu \) and \( V[X_i]=\sigma^2 \). Then for large \( n \),
\[
\bar{X} \approx N(\mu, \sigma^2/n)
\]

### 4 Two Populations

Suppose we observe two sets of individuals. We observe a random sample \( X_1, \ldots, X_{n_X} \) from one population (with \( E[X]=\mu_X \) and \( V[X]=\sigma_X^2 \)) and a random sample from \( Y_1, \ldots, Y_{n_Y} \) from another population (with \( E[Y]=\mu_Y \) and \( V[Y]=\sigma_Y^2 \)). Assume the \( X \) and \( Y \) are independent.

In inference, we are often interested in comparing the means \( \mu_X \) and \( \mu_Y \). One possible situation is the comparison of two teaching methods, where \( \mu_X \) is the mean student score using method \( X \) and \( \mu_Y \) is the mean student score using method \( Y \). Assuming we want students to do well, then \( \mu_X > \mu_Y \) indicates method \( X \) is better, \( \mu_X < \mu_Y \) indicates method \( Y \) is better, and \( \mu_X = \mu_Y \) indicates the methods are equal. Another way to make the same conclusion is to consider the quantity \( \mu_X - \mu_Y \). If \( \mu_X - \mu_Y > 0 \) then \( X \) is better, if \( \mu_X - \mu_Y < 0 \) then \( Y \) is better, and if \( \mu_X - \mu_Y = 0 \) then the methods are equal.

Usually the quantity \( \mu_X - \mu_Y \) is estimated with \( \bar{X} - \bar{Y} \). This is a linear combination of \( \bar{X} \) and \( \bar{Y} \). Using the rules for linear combinations,
\[
E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_X - \mu_Y
\]
\[
V[\bar{X} - \bar{Y}] = V[\bar{X}] + V[\bar{Y}] = \sigma_X^2/n_X + \sigma_Y^2/n_Y
\]
If \( n_X \) and \( n_Y \) are both large, or both populations are normally distributed, then \( \bar{X} \) and \( \bar{Y} \) will both be normally distributed and thus \( \bar{X} - \bar{Y} \) will be normally distributed, and thus
\[
(\bar{X} - \bar{Y}) \sim N(\mu_X - \mu_Y, \sigma_X^2/n_X + \sigma_Y^2/n_Y)
\]

### 5 Bernoulli Distributions

Suppose we observe \( X_1, \ldots, X \sim \text{Bern}(p) \) (so all the \( X \) are independent and have the same probability \( p \)). We have previously investigated this situation, called a Binomial Experiment. Recall that if we summed the number of successes, we found \( Y = \sum X_i \sim \text{Bin}(n, p) \).

However, when \( n \geq 30 \), the Central Limit Theorem also applies, so let’s see what the Central Limit Theorem says about this situation. The individual observations (the \( \text{Bern}(p) \) random variables) each have mean \( E[X] = p \) and \( V[X] = p(1-p) \). Using the Central Limit Theorem
\[
\bar{X} = \frac{\sum X_i}{n} \approx N(p, \frac{p(1-p)}{n})
\]
Thus, we have an approximation for the Binomial distribution when \( n \geq 30 \). The mean and variance of the approximating normal are just the mean and variance of the exact Binomial distribution.

Usually we do not use this approximation to compute probabilities such as \( P(X=30) \). After all, the exact distribution is easy enough to compute. We typically use the normal approximation when computing the probabilities of a range, such as \( P(30 \leq X < 80) \). Here is an example. Suppose that 40% of the vehicles on a highway are trucks, and that the vehicles are independent. You observe 400 vehicles and record whether or not each vehicle is a truck. Let \( X \) be the number of trucks. This is a Binomial experiment, so \( X \sim \text{Bin}(n=400, p=0.40) \). In addition, since \( n \geq 30 \), we may approximate the distribution of \( X \) by \( X \sim \text{N}(\mu=np=160, \sigma^2=np(1-p)=96) \). Suppose we wanted \( P(X<140) \). We could use

\[
P(X < 140) = \sum_{k=0}^{139} \binom{400}{k} 0.4^k 0.6^{400-k}
\]

Considering this is a summation of 140 terms, this may be difficult to evaluate except on a computer (the actual answer is 0.01762126). For interval probabilities, the normal approximation is usually quite effective. To use the normal approximation, find \( P(X<140) = 0.02061342 \). (Side Note: The approximation is improved using something called the continuity correction, which is not required for this course. The continuity correction would take \( P(X<139.5)=0.01820707 \) instead of \( P(X<140) \)).

The approximation works best in the range around the mean. Suppose we wanted \( P(150<X \leq 190) \). The exact answer is

\[
\sum_{k=151}^{190} \binom{400}{k} 0.4^k 0.6^{400-k} = 0.832821
\]

while the approximate answer is

\( P(150.5<X<190.5)=0.8329482 \).

These answers are almost identical. Here is a summary of some guidelines for using the normal approximation to the binomial (otherwise known as the Central Limit Theorem for Bernoullis):

1. The normal approximation should work for \( n \geq 30 \). However, if the Poisson approximation works well (so large \( n \) but \( p \) is small), use the Poisson approximation. When \( p \) is small, the Poisson approximation is more accurate.
2. There is no point in approximating the probability of a single point with the normal approximation, just use the Binomial probabilities.
3. The approximation works best when the range is near the mean of the distribution.

6 Problems

1) Suppose that each weekend an airplane rental company attempts to rent its stock of planes. The expected rental fees each week are $910 with variance $294,900. What is the probability they make over $55,000 in rent in a year (52 weekends)? Assume the weekends are independent.

2) A lake contains 10 red fish and 30 blue fish and is restocked daily. Each day a fisherman arrives and catches 5 fish at random (without replacement). Over the course of a year (365 days), find the probability that the fisherman catches at least 450 fish.
3) Two teaching methods are being compared in a study. The methods are assessed by giving each student in the study a test which is graded pass/fail. The school board decides to adopt whichever method has the highest proportion of passing grades. Suppose that (unknown to the school board) method A results in 80% of the students passing the test, while method B results in 75% of the students passing the test.

a) Suppose that 50 students are taught by method A and 50 students are taught by method B. Let \( \hat{p}_A \) be the proportion of students that were taught by method A that pass, and \( \hat{p}_B \) is the corresponding proportion for the students taught with method B. What are the approximate distributions of \( \hat{p}_A \) and \( \hat{p}_B \)? What is the approximate distribution of \( \hat{p}_A - \hat{p}_B \)?

b) Note that the school board adopts method B if \( \hat{p}_A - \hat{p}_B < 0 \). What is the approximate probability that the school board mistakenly adopts method B?

c) Suppose the sample sizes were increased to 500 students in each method. Now what is the approximate probability the school board mistakenly adopts method B?

4) A small store has an average of 15 customers per day with a standard deviation of 3 customers per day.

a) Over the course of 50 days, what is the distribution of the total number of customers in the store?

b) What is the probability that, over 50 days, there are over 775 customers in the store?

5) Suppose scores on an exam may be approximated by a continuous distribution that has density 
\[
f(x) = 63574540(x/100)^9(1 - (x/100))^9 \text{ on the range 0 to 100.}
\]
Using this density, it is possible to find the expectation and variance for any particular score is \( E[X]=75 \) and \( V[X]=45.73171 \). Suppose 60 students take the exam. Let \( \bar{X} \) be the average of the 60 scores. What is the approximate probability \( P(\bar{X}<77) \)?

6) A class with 500 students is held in a large lecture hall. Each day, there is a 0.75 chance any particular student attends class. What is the approximate probability at least 390 students attend class? You may assume the students attend class independently.

7) A student takes a 500 question, multiple choice exam, without knowing anything. By guessing, the student has a 20% chance of getting any individual question correct. You may assume all questions are independent. Let \( X \) be the number of questions answered correctly.

a) What are the exact and approximate distributions of \( X \)?

b) Approximate \( P(95<X<120) \).

8) The number of calls to an emergency number in any given minute has a Poisson distribution with the parameter \( \lambda=1.5 \). Let \( T \) be the number of calls in any particular hour (assume minutes are independent). Find the approximate distribution of \( T \) and find \( P(80<T<140) \).