5 Sampling Distributions

5.1 Introduction

In this section, we transition from the world of probability to the world of statistics. The field of probability is concerned with making predictions about probability distributions which are completely known. The contents of a deck of cards, for example, is known, and thus we can derive probabilistic statements about the likelihood of certain cards being flipped in a game of Texas Hold'em poker. The field of statistics addresses the converse problem. It allows us to make inferences about the distribution of a population based on a small sample drawn from that population. We then use the tools from probability to make probabilistic statements about the accuracy of our inferences. For example, we can model the distribution of Rhode Island voter preferences in the upcoming presidential election with a binomial distribution. The parameters of the distribution are $n = 754,224$ (the number of registered voters as of election day, 2014) and p, the proportion of voters who prefer Clinton. The parameter p is unknown, and unless we accurately survey every single registered voter, we will not know its exact value. Such a survey is logistically and financially unfeasible, thus we turn to statistics. We poll a smaller sample of voters (say, 1000) and find the proportion of voters in that smaller sample who prefer Clinton. This sample proportion is designated $\hat{p}$, where the "hat" indicates that it is an estimator for the true value p based off of a sample drawn from the larger population. Using the tools of statistics, we will be able to quantitatively evaluate how close our estimate $\hat{p}$ is to the true value p.

5.2 Statistics

We will use the following setup for our discussion.

1. We have a large population and are interested in studying a particular quantitative feature of that population. For example, the population could be the registered voters in Rhode Island, and we are interested in the yes/no question "Are you going to vote for Clinton?". Another population could be the ball bearings produced by the factory, and we are interested in their diameter.

2. The population feature can be characterized by a pmf $p(y)$ (if it is discrete, as in the case of the number of voters who prefer Clinton) or a density function $f(y)$ (if it is continuous, as in the case of the ball bearing diameters).

3. The population feature will have certain parameters. All populations have a mean and a variance, which will be designated by μ and σ^2. We will refer to these as the population mean and the population variance. Populations may also have other parameters. As an example, if we model a population with a uniform distribution on an unknown interval $[a, b]$, the endpoints a and b are parameters of the population. Much of statistics is concerned with estimating parameters of populations.

4. We take a small sample from the population. Let n be the sample size, and let the random variables $Y_1, \ldots, Y_n$ be the samples we take from the population.

5. We will assume that the sample size is small enough relative to the population size that the samples $Y_1, \ldots, Y_n$ are independent and have the same distribution $p(y)$ or $f(y)$ as the population. As an example, recall that we mentioned before that sampling without replacement (as in political polling) was approximately binomial if the sample size was less than $1/20$ of the population size.

6. A *statistic* is a function of our samples $Y_1, \ldots, Y_n$. The function can only involve the samples themselves and known constants such as the sample size n. Since a statistic is a function of random variables, it is itself a random variable, thus we can characterize its distribution using the tools of probability.

In this first section we will be concerned with the probability distribution of various statistics. Before we continue, we will give a table of the most common statistics we will encounter. Let $Y_1, \ldots, Y_n$ be independent and identically distributed (iid) samples drawn from a population which is distributed according to a pmf $p(y)$ or a density $f(y)$. Let μ and σ^2 be the population mean and variance.

Statistic	Definition
sample mean	$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$
sample variance	$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
smallest order statistic (sample minimum)	$Y_1 = \min(Y_1, \ldots, Y_n)$
largest order statistic (sample maximum)	$Y_n = \max(Y_1, \ldots, Y_n)$
sample range	$R = Y_n - Y_1$

Note that you have to compute $\bar{Y}$ before you can compute S^2. The $n-1$ in the denominator of the sample variance may seem a bit mysterious, but we will see in a few classes why that makes sense. For now, we will look at the sample mean. To get the sample mean, we add together the samples and divide by the number of samples, so this is the empirical mean we have discussed before. The sample mean is a random variable, so we can find its mean and variance. Using linearity of expectation and the fact that all the Y_i have expected value of μ,

$$
\begin{aligned}
\mathbb{E}(\bar{Y}) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu \\
&= \mu
\end{aligned}
$$

Thus the expected value of the sample mean is the population mean μ.

For the variance, we recall that $Var(aY) = a^2 Var(Y)$ and that for independent random variables, $Var(Y_1 + \cdots + Y_n) = Var(Y_1) + \cdots + Var(Y_n)$. Using this plus the fact the all the Y_i have variance of σ^2,

$$
\begin{aligned}
Var(\bar{Y}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2}n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

The variance of the sample mean is thus the population variance divided by n. As n gets larger, this variance gets smaller, which makes intuitive sense since there should be less spread if we average more samples together.

Note that while we have characterized the mean and variance of the sample mean, we have said nothing about its distribution. In general, there is not much additional we can say. In the special case where the population has a normal distribution, however, we can say a lot more.

5.3   Sampling Distributions for Normally Distributed Populations

5.3.1   Distribution of Sample Mean

Suppose we have a population which is normally distributed. Then the sample mean is also normally distributed.

Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and variance σ^2. Then the sample mean

$$
\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i
$$

has a normal distribution with mean $\mathbb{E}(\bar{Y}) = \mu$ and variance $Var(\bar{Y}) = \sigma^2/n$.

The mean and variance of the sample mean was shown above. To show the sample mean has a normal distribution, we can use the method of moment generating functions or characteristic functions, which will not be done in this course for sake of time.

Since the sample mean $\bar{Y}$ is normally distributed with mean μ and variance σ^2/n,

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right)$$

is a standard normal random variable. (Recall that to get a standard normal, we subtract the mean and divide by the standard deviation, which is the square root of the variance.)

Let's use this in an example.

Example. A ball bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm. We unfortunately have lost the manual for the machine, so we do not know the value of μ. We call the company to get more information, but all they can tell us is that $\sigma = 0.1$.

1. We take a sample of 16 ball bearings from the machine and compute the sample mean $\bar{Y}$. Find the probability that $\bar{Y}$ is within 0.02 mm of the true mean μ.

   By what we learned above, the sample mean $\bar{Y}$ has a normal distribution with mean μ and variance $\sigma^2/n = 0.1^2/16$. The standard deviation of $\bar{Y}$ is $\sigma = 0.1/\sqrt{16}$. Thus we can find the probability by converting to the standard normal distribution.

$$\begin{aligned} \mathbb{P}(|\bar{Y} - \mu| \leq 0.02) &= \mathbb{P}(-0.02 \leq (\bar{Y} - \mu) \leq 0.02) \\ &= \mathbb{P}\left(\frac{-0.02}{\sigma/\sqrt{n}} \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.02}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}\left(\frac{-0.02}{0.1/\sqrt{16}} \leq Z \leq \frac{0.02}{0.1/\sqrt{16}}\right) \\ &= \mathbb{P}(-0.8 \leq Z \leq 0.8) \\ &= 0.7881 - 0.2119 \\ &= 0.5762 \end{aligned}$$

2. How many ball bearings should we sample if we wish the sample mean to be within 0.02 mm of μ with probability 0.95?

   We can use the 68-95-99.7 rule here. We want the sample mean to be with 0.02 mm of the mean with probability 0.95. Another way of writing this is that we want the population mean μ to lie within two sample standard deviations of the sample mean. Thus we want the sample standard deviation to be $0.02/2 = 0.01$. Since the sample standard deviation is $\sigma/\sqrt{n} = 0.1/\sqrt{n}$, all we have to is set these equal to each other and solve for n.

$$\begin{aligned} \frac{0.1}{\sqrt{n}} &= 0.01 \\ \sqrt{n} &= \frac{0.1}{0.01} = 10 \\ n &= 100 \end{aligned}$$

3. How many ball bearings should we sample if we wish the sample mean to be within 0.02 mm of μ with probability 0.99?

Here we can't rely on the 68-95-99.7 rule and have to actually do the math. We want:

$$\mathbb{P}(|\bar{Y} - \mu| \leq 0.02) = \mathbb{P}(-0.02 \leq (\bar{Y} - \mu) \leq 0.02) = 0.99$$

Dividing this by the sample standard deviation $\sigma/\sqrt{n}$ to convert to the standard normal, we get

$$
\begin{aligned}
0.99 = \mathbb{P}\left(\frac{-0.02}{\sigma/\sqrt{n}} \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.02}{\sigma/\sqrt{n}}\right) \\
= \mathbb{P}\left(\frac{-0.02}{0.1/\sqrt{n}} \leq Z \leq \frac{0.02}{0.1/\sqrt{n}}\right) \\
= \mathbb{P}(-0.2\sqrt{n} \leq Z \leq 0.2\sqrt{n})
\end{aligned}
$$

At this point, we take complements and use the symmetry of the normal distribution.

$$\mathbb{P}(Z \leq -0.2\sqrt{n} \text{ or } Z \geq 0.2\sqrt{n}) = 1 - 0.99 = 0.01$$
$$\mathbb{P}(Z \leq -0.2\sqrt{n}) = 0.005$$

Consulting a Z table, we find that $\mathbb{P}(Z \leq -2.57) = 0.0051$ and $\mathbb{P}(Z \leq -2.58) = 0.0049$. Choosing the first one (you can choose either one) and setting that value of z equal to $-0.2\sqrt{n}$, we get:

$$
\begin{aligned}
-0.2\sqrt{n} &= -2.57 \\
\sqrt{n} &= 12.85 \\
n &= 165.123
\end{aligned}
$$

Rounding up to the nearest integer, we need a sample size of 166 or more to have 99% confidence that our sample mean is within 0.02 of the true population mean.

5.4 Other distributions

We will now briefly discuss two other useful distributions, the chi-square and t distribution. Since computations with these distributions is done via tables and not by using the densities directly, we will not give the densities here. If you are curious what these densities look like, you can consult Wikipedia or a statistics textbook.

5.4.1 Chi-square distribution

This section was not covered in class but is included for completeness; unless I cover it in the future, I am presenting this purely for reference. It will not appear on any problem set

or exam.

The chi-square distribution is the distribution of the sum of the squares of independent, standard normal random variables. Since the variance of a random variable is computed by first finding the expected value of the square of the random variable, the chi-square distribution is used when we make inferences about the variance of a population.

> Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and variance σ^2. Then the random variables $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, and
>
> $$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$
>
> has a χ^2 (chi-square) distribution with n degrees of freedom (df).

To use the chi-square distribution, we use a chi-square table, which is provided on the course website. Chi-square tables are a little weird. Here's how we use them. Suppose Y has a chi-square distribution with 10 df. We will be looking at the row labeled 10 df. Now look across the top row. You see $\chi^2_{.995}, \chi^2_{.990}, \chi^2_{.975}$, etc. The decimals there are probabilities. What do they represent? Let's look at the row for 10 df. The first entry in that row, in the column labeled $\chi^2_{.995}$, is the value of y for which $\mathbb{P}(Y > y) = 0.995$. The entry in that box is 2.156, so $\mathbb{P}(Y > 2.156) = 0.995$. If we want the probability that Y is less than or equal to 2.156, then we just subtract from 1 to get $\mathbb{P}(Y \leq 2.156) = 1 - 0.995 = 0.005$. Similarly, looking at the 6th column, we have $\mathbb{P}(Y > 15.987) = 0.100$, so $\mathbb{P}(Y \leq 15.987) = 1 - 0.100 = 0.90$. Note that the probabilities across the top are "extreme" probabilities, i.e. very high or very low. In general, those are the probabilities we will be interested in. For other values, you can use a statistical package such as R.

First, let's practice using the chi-square table, then we will use the chi-square distribution in an example.

Example. Let $Z_1, \ldots, Z_6$ be independent samples from a standard normal distribution. Find a value a such that

$$\mathbb{P}\left(\sum_{i=1}^{6} Z_i^2 \leq a \right) = 0.95$$

By the above, $\sum_{i=1}^{6} Z_i^2$ has a chi-square distribution with 6 df. We want the probability that this is less than a to be 0.95, which is equivalent to

$$\mathbb{P}\left(\sum_{i=1}^{6} Z_i^2 > a \right) = 1 - 0.95 = 0.05$$

Thus we look at the chi-square table in the 6 df row and find the entry in the column labeled $\chi^2_{.050}$ to get 12.592. Thus we conclude that $a = 12.592$.

How do we use this practically. Recall from above that the sample variance is given by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Then we can say the following about the distribution of S^2:

Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and variance σ^2. Let S^2 be the sample variance as given above. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

has a χ^2 (chi-square) distribution with $n-1$ degrees of freedom (df).

Again, the proof of this is omitted for sake of time. Let's go back to our ball bearing example.

Example. A ball bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm. Once again, all we know is that $\sigma = 0.1$. Suppose we select 10 samples and compute the sample variance S^2. Find an interval $[a, b]$ such that the probability that S^2 lies in the interval is 0.90, i.e.

$$\mathbb{P}(a \leq S^2 \leq b) = 0.90$$

Multiplying by $(n-1)$ and dividing by σ^2, we get:

$$0.90 = \mathbb{P}(a \leq S^2 \leq b) = \mathbb{P}\left(\frac{(n-1)a}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b}{\sigma^2} \right)$$

From above, we know that $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $(n-1) = 10 - 1 = 9$ df. Let's call this X. Then we have:

$$0.90 = \mathbb{P}(a \leq S^2 \leq b) = \mathbb{P}\left(\frac{(n-1)a}{\sigma^2} \leq X \leq \frac{(n-1)b}{\sigma^2} \right)$$

where X has a chi-square distribution with 9 df. There are many ways we can choose a and b to make this happen, but the easiest way is to "split the difference". Taking complements, we can write this as:

$$\mathbb{P}\left(X \leq \frac{(n-1)a}{\sigma^2} \text{ or } X \geq \frac{(n-1)b}{\sigma^2} \right) = 1 - 0.90 = 0.10$$

Splitting the probability of 0.10 in half, we will find a and b such that:

$$\mathbb{P}\left(X \leq \frac{(n-1)a}{\sigma^2} \right) = 0.05 \text{ and } \mathbb{P}\left(X \geq \frac{(n-1)b}{\sigma^2} \right) = 0.05$$

Since the chi-square table gives us probabilities of being greater than a value, this is equivalent to:
$$\mathbb{P}\left(X > \frac{(n-1)a}{\sigma^2}\right) = 0.95 \text{ and } \mathbb{P}\left(X > \frac{(n-1)b}{\sigma^2}\right) = 0.05$$
(these are continuous random variables, so the strictness of the inequalities does not matter). Now we look at the chi-square table. We need the values for 0.95 and 0.05 in the 9 df row, which are 3.325 and 16.919. Then we can solve for a and b, since we know $n = 10$ and $\sigma = 0.1$.

$$3.325 = \frac{(n-1)a}{\sigma^2} = \frac{9a}{0.1^2}$$
$$a = \frac{3.325 \cdot 0.1^2}{9} = 0.0037$$

$$16.919 = \frac{(n-1)b}{\sigma^2} = \frac{9b}{0.1^2}$$
$$a = \frac{16.919 \cdot 0.1^2}{9} = 0.0187$$

Thus our interval $[a, b]$ is $[0.0037, 0.0187]$, which has 90% probability of including our sample variance S^2. Note that the true population variance $\sigma^2 = 0.1^2 = 0.01$ lies in that interval. If our sample variance is not in that interval, perhaps we should be suspicious of either our sampling technique or that something is going wrong with the machine.

5.4.2   t distribution

What happens when we do not know the population standard deviation. Recall that if we have normal population whose mean and standard deviation is known, the quantity
$$\sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right)$$
is a standard normal random variable (essentially a normalized version of the sample mean). If the population standard deviation is not known, then we can replace σ by our sample statistic S (the square root of the population variance S^2) to get
$$\sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right)$$
The probability distribution of this quantity is known as the t-distribution.

Let $Y_1, \ldots, Y_n$ be a sample of size n drawn from a population which has a normal distribution with mean μ and unknown variance. Let S^2 be the sample variance as given above. Then
$$T = \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right)$$
has a t-distribution with $n - 1$ degrees of freedom (df).

Like the chi-square distribution, we use tables to work with the t-distribution. Let's do an example of this.

Example. Once again, we have a bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm. We not only have lost the manual for the machine, but the company has gone bankrupt so they cannot tell us anything about the machine. We take a sample of 16 ball bearings from the machine and compute the sample mean $\bar{Y}$. We also compute the sample standard deviation S and find that $S = 0.1$. Find a range $[a, b]$ such that the probability that the difference between the sample mean and the population mean μ falls within the interval with a probability of 0.90.

We are interested in finding a and b such that $\mathbb{P}(a \leq (\bar{Y} - \mu) \leq b) = 0.90$. Multiplying by $\sqrt{n}$ and dividing by S, we get:

$$\mathbb{P}\left(\frac{\sqrt{n}a}{S} \leq \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right) \leq \frac{\sqrt{n}a}{S}\right) = 0.90$$

$$\mathbb{P}\left(\frac{\sqrt{16}a}{0.1} \leq T \leq \frac{\sqrt{16}a}{0.1}\right) = 0.90$$

$$\mathbb{P}\left(40a \leq T \leq 40b\right) = 0.90$$

Now we look at our t-table. Note that the t-table gives probabilities for the upper tail of the t-distribution and that the t-distribution is symmetric about the mean. We have a sample of 16, so we want 15 df. We will take a symmetric interval about the mean, so we want $\mathbb{P}(T \leq 40a) = 0.05$ and $\mathbb{P}(T \leq 40b) = 0.05$. For the upper tail, we look at the t-table in the row for 15 df and go across until we get to 0.05, which gives us a t-value of 1.753. By symmetry, the lower t-value is -1.753. Thus we have $40b = 1.753$, which implies $b = 0.0438$. By symmetry, $a = -0.0438$. Thus we are 90% confident that the difference between the sample and population means falls within the interval $[-0.0438, 0.0438]$.

What if we actually know the population standard deviation, as in the example above. Then we can use the standard normal distribution instead of the t-distribution. How does the interval we get compare to that where the population standard deviation is not known. We expect to get a narrower interval in this case, since we have more information. Let's see if that is indeed the case.

$$\mathbb{P}\left(\frac{\sqrt{n}a}{\sigma} \leq \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right) \leq \frac{\sqrt{n}a}{\sigma}\right) = 0.90$$

$$\mathbb{P}\left(\frac{\sqrt{16}a}{0.1} \leq Z \leq \frac{\sqrt{16}a}{0.1}\right) = 0.90$$

$$\mathbb{P}\left(40a \leq Z \leq 40b\right) = 0.90$$

So far, this is exactly the same except we have Z in place of T. Again, we will look for a symmetric interval about the mean, so we want $\mathbb{P}(Z \leq 40a) = 0.05$ and $\mathbb{P}(Z \leq 40b) = 0.05$. For the lower tail, we look at the Z-table and can choose either $z = -1.65$ (corresponding to

a probability of 0.0495) or $z = -1.64$ (corresponding to a probability of 0.0505). Choosing $z = -1.64$ we get $40a = -1.64$, so $a = -0.041$. By symmetry, $b = 0.041$. Thus we are 90% confident that the difference between the sample and population means falls within the interval $[-0.041, 0.041]$. As predicted, this is a narrower interval than the case where we used the t-distribution (where the population standard deviation was unknown).