6 Estimation

6.1 Introduction

The purpose of statistics is to make inferences about populations based on data from a small sample of that population. Populations are characterized by probability distributions which can be described by numerical parameters. The population mean μ and population variance σ^2 are parameters common to all probability distributions. In addition, some populations can be described by other natural parameters. One example is the population of all registered voters in Rhode Island. If we are interested in the yes/no question "Are you going to vote for Clinton?", a natural parameter is p, the proportion of the population who plans on voting for Clinton.

The setup is exactly the same as with sampling. Suppose we are studying a population whose distribution is characterized by a parameter of interest which we will denote θ (this could be the population mean, population variance, proportion of voters supporting Clinton, or some other parameter). We will take n independent, identically distributed samples $Y_1, \ldots, Y_n$ from our population. An *estimator* is a function of our n samples which is designed to give us information about the population parameter. There are two types of estimators we will discuss:

1. A *point estimator* produces a single number which we think is close to the parameter of interest. We will learn several ways to quantitatively evaluate the "goodness" of a point estimator

2. An *interval estimate* produces an interval (often called a confidence interval) in which we believe our parameter of interest lies. We will learn how to construct confidence intervals which include the parameter of interest with a given probability.

6.2 Point Estimators

We start our discussion with point estimators. We have already met one point estimator, the sample mean $\bar{Y}$. The sample mean is an estimator since it is a function of our n samples. It is an estimator for the population mean. For another example, suppose we are polling n voters out of a population of registered voters and asking them if they are voting for Clinton. The parameter of interest is p, the proportion of registered voters who will vote for Clinton. Let Y be the number of voters in our sample who are voting for Clinton. Then the sample proportion $\hat{p} = Y/n$ is an estimator for the population proportion p. (The "hat" over the p indicates that it is an estimator for the parameter p).

Since we are applied mathematicians, we need quantitative tools to tell whether an estimator is any good. Let $\hat{\theta}$ be an estimator for the parameter θ. The first criterion we can use to evaluate an estimator is *bias*. We would like the expected value of our estimator to be the actual value of the parameter we are trying to estimate, i.e. $\mathbb{E}(\hat{\theta}) = \theta$. An estimator which has this property is called *unbiased*

Let $\hat{\theta}$ be an estimator for a parameter θ. The estimator $\hat{\theta}$ is *unbiased* if $\mathbb{E}(\hat{\theta}) = \theta$. Otherwise, the estimator $\hat{\theta}$ is *biased*. The *bias* of $\hat{\theta}$ is given by

$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Let's look at the estimators we have seen so far. The sample mean $\bar{Y}$ is an estimator for the population mean μ (here we abandon our "hat"-convention, and call this $\bar{Y}$ rather than $\hat{\mu}$). We showed in the last section that $\mathbb{E}(\bar{Y}) = \mu$, thus the sample mean is an unbiased estimator for the population mean.

What about the sample proportion estimator we use in polling? Suppose we poll n voters in a population, and let Y be the number of voters in our sample who are voting for Clinton. As long as the sample is small enough (less than 1/20 of the population size), we can take Y to be a binomial random variable with parameters n and p. For our estimator $\hat{p} = Y/n$,

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{Y}{n}\right) = \frac{1}{n}\mathbb{E}(Y) = \frac{np}{n} = p$$

where we have used the expected value of a binomial random variable. Since $\mathbb{E}(\hat{p}) = p$, this estimator is unbiased as well. We will see an example of a biased estimator later.

A perhaps better measure of the "goodness" of an estimator is its *mean square error*, the average of the square distance of the estimator from the parameter of interest.

Mean Square Error

Let $\hat{\theta}$ be an estimator for a parameter θ. The *mean square error* of $\hat{\theta}$ is defined by

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

If $Bias(\hat{\theta})$ is the bias of $\hat{\theta}$ and $Var(\hat{\theta})$ is the variance of $\hat{\theta}$, then

$$MSE(\hat{\theta}) = [Bias(\hat{\theta})]^2 + Var(\hat{\theta})$$

Note that the MSE can be divided into two components: variance and bias squared. Both of these are positive. We have discussed above how low bias is a good quality for an estimator. Low variance is a desired quality for an estimator as well since ideally we would like the distribution of the estimator to cluster tightly about the parameter of interest. In general, if we keep the sample size fixed, there is a tradeoff between bias and variance. For a given MSE, if we wish our estimator to have lower bias, then we must accept a higher variance

and vice versa.

To show the relationship between MSE, bias, and variance, we take the definition of MSE and add and subtract $\mathbb{E}(\hat{\theta})$ inside the parentheses.

$$
\begin{aligned}
MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta})) + (\mathbb{E}(\hat{\theta}) - \theta)]^2 \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))\underbrace{(\mathbb{E}(\hat{\theta}) - \theta)}_{\text{constant}}] + \mathbb{E}[\underbrace{(\mathbb{E}(\hat{\theta}) - \theta)^2}_{\text{constant}}] \\
&= Var(\hat{\theta}) + 2(\mathbb{E}(\hat{\theta}) - \theta)\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))] + (\mathbb{E}(\hat{\theta}) - \theta)^2 \\
&= Var(\hat{\theta}) + 2(\mathbb{E}(\hat{\theta}) - \theta)\underbrace{(\mathbb{E}(\hat{\theta}) - \mathbb{E}(\hat{\theta}))}_{=0} + [Bias(\hat{\theta})]^2 \\
&= Var(\hat{\theta}) + [Bias(\hat{\theta})]^2
\end{aligned}
$$

Let's look at the MSE of the two estimators we discussed above.

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

   We showed that this estimator is unbiased, so the MSE is equal to the variance. We showed in a previous section that the variance of the sample mean is σ^2/n, where σ^2 is the population variance. Thus we have

   $$
   MSE(\bar{Y}) = \frac{sigma^2}{n}
   $$

   Note that the MSE goes to 0 as $n \to \infty$, i.e. the error of our estimator decreases as our sample gets larger. This makes intuitive sense that a larger sample provides a better estimator for the population mean.

2. Sample proportion. $\hat{p} = \frac{Y}{n}$.

   Recall that we are assuming that $Y \sim$ Binomial(n, p). We showed above that this estimator is also unbiased, thus once again the MSE is equal to the variance. Recalling that the variance of a binomial random variable is $np(1-p)$,

   $$
   \begin{aligned}
   MSE(\hat{p}) &= Var\left(\frac{Y}{n}\right) \\
   &= \frac{1}{n^2}Var(Y) \\
   &= \frac{np(1-p)}{n^2} \\
   &= \frac{p(1-p)}{n}
   \end{aligned}
   $$

Sometimes we are interested in studying the difference between two populations. Here are some examples of that.

1. Suppose we are interested in whether Brown University first-year students or seniors get more sleep. In this case, the parameter of interest is the *difference* in the mean amount of sleep between first-years and seniors. If μ_1 and σ_1^2 are the mean and variance of the amount of sleep of first-years and μ_2 and σ_2^2 are the same for seniors, then mathematically our parameter of interest is $\mu_1 - \mu_2$. If we take a sample of n_1 first-year students and n_2 seniors and compute the sample means $\bar{Y}_1$ and $\bar{Y}_2$, then $\bar{Y}_1 - \bar{Y}_2$ is an estimator for $\mu_1 - \mu_2$. By linearity of expectation and our result for the expected value of the sample mean, the expected value of the sample mean is $\mu_1 - \mu_2$, so this estimator is unbiased. For the variance of this estimator, we use the formula for the variance of a sum (assuming the samples are independent), and recall that constants are squared when they are pulled out of the variance:

$$
\begin{aligned}
Var(\bar{Y}_1 - \bar{Y}_2) &= Var(\bar{Y}_1) + Var(-\bar{Y}_2) \\
&= Var(\bar{Y}_1) + (-1)^2 Var(\bar{Y}_2) \\
&= Var(\bar{Y}_1) + Var(\bar{Y}_2) \\
&= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
\end{aligned}
$$

2. Suppose we are interested in the preference for Clinton in rural versus urban voters in Pennsylvania. The parameter of interest here is the difference in the proportion of Clinton supporters between rural and urban areas. First, we have to define "rural" and "urban". Suppose we take "urban" to mean living in a city with population of 250,000 or more (in Pennsylvania, this would only include Philadelphia and Pittsburgh). What are advantages or drawbacks to this definition? If p_1 is the proportion of Clinton supports in rural areas and p_2 is the proportion of Clinton supporters in urban areas, then the parameter of interest is $p_1 - p_2$. Suppose we sample n_1 voters from rural areas and n_2 voters from urban areas. Let Y_1 and Y_2 be the proportion of rural and urban voters (respectively) who support Clinton. Then

$$
\hat{p}_1 - \hat{p}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}
$$

is an estimator for $p_1 - p_2$. By linearity of expectation and the result for a single population, the expected value of this estimator is $p_1 - p_2$, so this estimator is unbiased. What is the variance of this estimator? As above, using the formula for the variance of a sum of two independent random variables,

$$
\begin{aligned}
Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(-\hat{p}_2) \\
&= Var(\hat{p}_1) + (-1)^2 Var(\hat{p}_2) \\
&= Var(\hat{p}_1) + Var(\hat{p}_2) \\
&= \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}
\end{aligned}
$$

As another example, let's look at our estimators for population variance. Recall from the previous section that the sample variance is given by:

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

where $\bar{Y}$ is the sample mean. The factor of $(n-1)$ in the denominator seems peculiar. It seems more natural to divide by n and use the following estimator for the population variance:

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

We will show that while they are both estimators for the sample variance, S'^2 is biased whereas S^2 is unbiased. Since S^2 is unbiased, we call it the sample variance.

First we show S'^2 is biased by computing its expected value. This is done in several steps.

1. First we find a nice formula for $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

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

$$= \sum_{i=1}^{n} Y_i^2 - 2\bar{Y} \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} \bar{Y}^2$$

$$= \sum_{i=1}^{n} Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2$$

$$= \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

2. Next we take the expected value of this. By linearity of expectation,

$$\mathbb{E}\left[\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right] = \mathbb{E}\left(\sum_{i=1}^{n} Y_i^2\right) - n\mathbb{E}(\bar{Y}^2)$$

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

3. Next we use the Magic Variance Formula "in reverse" to compute $\mathbb{E}(Y_i^2)$ and $\mathbb{E}(\bar{Y}^2)$.

$$Var(Y_i) = \mathbb{E}(Y_i^2) - [\mathbb{E}(Y_i)]^2$$
$$\mathbb{E}(Y_i^2) = Var(Y_i) + [\mathbb{E}(Y_i)]^2$$
$$= \sigma^2 + \mu^2$$

Similarly,

$$\mathbb{E}(\bar{Y}^2) = Var(\bar{Y}) + [\mathbb{E}(\bar{Y})]^2$$
$$= \frac{\sigma^2}{n} + \mu^2$$

5

4. We then plug these into the expression from step 3.

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] = \sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2$$
$$= (n-1)\sigma^2$$

5. Finally we use this result to compute the expected value of S'^2.

$$\mathbb{E}(S'^2) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]$$
$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]$$
$$= \frac{1}{n}(n-1)\sigma^2$$
$$= \frac{n-1}{n}\sigma^2$$

Since $\mathbb{E}(S'^2) \neq \sigma^2$, this estimator is biased. For large n, however, $(n-1)/n \approx 1$, so the bias is minimal. We can convert this to an unbiased estimator by multiplying by $n/(n-1)$. This is legitimate since n is a known constant (the sample size) and not one of the parameters we are trying to estimate:

$$\mathbb{E}\left(\frac{n}{n-1}S'^2\right) = \frac{n}{n-1}\mathbb{E}(S'^2)$$
$$= \frac{n}{n-1}\frac{n-1}{n}\sigma^2$$
$$= \sigma^2$$

So we have found an unbiased estimator for σ^2. But we also have:

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

Thus S^2 is an unbiased estimator for the population variance σ^2.