CmpE 343 Lecture Notes
9: Estimation

Ethem Alpaydın

December 30, 2014

LET us say we have a population drawn from some unknown probability distribution $f(x)$ with some parameter θ. When we do not know θ, we can *estimate* it using a random sample. We discuss two types of estimation, namely, point and interval estimation.

1 Point Estimation

In *point estimation*, we estimate a single value that we denote by $\hat{\theta}$—in statistics, the hat indicates that the value is an estimate. We collect a sample $\mathcal{X} = \{X_i\}_{i=1}^n$, and a point estimator $d(\mathcal{X})$ is a function that takes the sample as its argument and returns a value. For example, if μ is unknown, $\overline{X}$ is a point estimator and on a sample, the sample average is one specific value.

For the same population parameter, there can be different point estimators; for example, for μ, one point estimator is the sample average, another may be the sample median. When there are multiple possible estimators or when we are proposing a new one, we need a way of quantifying its goodness.

Let us say θ is the unknown population parameter and $d(\mathcal{X})$ (we write d in short) is the estimator for θ. The *mean square error* of d as estimator for θ is defined as

mean square error

$$r(d, \theta) = E[(d - \theta)^2] \tag{1}$$

The estimate d can be larger or smaller than θ and we square the difference so that it ia always nonnegative (square is easier to manipulate than the absolute value of the difference), and we want to look at the average performance in general, and not on just one specific sample, so we take the expected value over all possible samples of size n (but of course, all should be drawn from the same population with the same θ).

Let us rewrite equation (1):

$$\begin{aligned} r(d, \theta) &= E[(d - \theta)^2] \\ &= E[(d - E[d] + E[d] - \theta)^2] \\ &= E[(d - E[d])^2] + (E[d] - \theta)^2 + 2(E[d] - \theta)E[d - E[d]] \end{aligned}$$

Remember that θ is a constant; d is a random variable but $E[d]$ is a constant, and so we have $E[E[d]] = E[d]$. Hence $E[d - E[d]] = 0$ and the cross-term disappears, and we are left with

bias and variance

$$r(d, \theta) = \underbrace{E[(d - E[d])^2]}_{\text{variance of } d} + \underbrace{(E[d] - \theta)^2}_{\text{bias of } d \text{ squared}} \tag{2}$$

The first term is the variance of d, that is, how much the different d calculated on different samples vary around their expected value $E[d]$. Variance is a measure of uncertainty and we want to decrease it. The second term is the bias of d, that is, how much the expected value of d differs from the parameter it is estimating. If $E[d] = \theta$, d is an *unbiased estimator*, that is, though on any sample, the calculated d may be different from θ, we know that over all, it is correct. We also want the bias to be as small as possible, and if possible we want our estimator to be unbiased.

Let us see some examples: $\overline{X}$ is a point estimator for μ. We know (see Lecture 7) that $E[\overline{X}] = \mu$, so $\overline{X}$ is an unbiased estimator for μ. We also know that $\text{Var}(\overline{X}) = \sigma^2/n$, so the mean square error is

$$r(\overline{X}, \mu) = \sigma^2/n \tag{3}$$

Let us consider another estimator for μ as X_1, that is, the first instance in my sample (Remember that the sample is unordered, so X_1 is not the minimum, it is one random instance from the sample). In this case $E[X_1] = \mu$, so this is also unbiased; but $\text{Var}(X_1) = \sigma^2$ and hence the mean square error is σ^2. That is why the sample average is a better estimator than a single instance, because it has smaller variance (because it uses the whole sample and not just a single instance).

One point estimator for σ^2 is the sample variance s^2 defined as

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Let us see if it is unbiased. We start by

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_i (X_i - \mu + \mu - \overline{X})^2 \\ &= \sum_i (X_i - \mu)^2 - n(\overline{X} - \mu)^2\end{aligned}$$

Then

$$\begin{aligned}E[s^2] &= E\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}\right] \\ &= \frac{1}{n-1}\left(\sum_i E[(X_i - \mu)^2] - nE[(\overline{X} - \mu)^2]\right) \\ &= \frac{1}{n-1}\left(n\sigma^2 - n(\sigma^2/n)\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2\end{aligned}$$

where we used the fact that $E[(X_i - \mu)^2] = \text{Var}(X_i) = \sigma^2$ and $E[(\overline{X} - \mu)^2] = \text{Var}(\overline{X}) = \sigma^2/n$. The fact that $E[s^2] = \sigma^2$ shows that s^2 is an unbiased estimator for σ^2, and also explains why we divide by $n-1$ and not n, if we divided by n, it would be a biased estimator—actually it is an *asymptotically unbiased* estimator because as n goes to infinity, $(n-1)/n$ converges to 1.

2 Interval Estimation

The point estimate returns a single value but we know that from the same population, if we draw another sample, there will be a different point estimate value (as given by the sampling distribution of the point estimating statistic—see Lecture 8). In *interval estimation*, we estimate an interval $[\hat{\theta}_L, \hat{\theta}_U]$ that includes the unknown θ with a high probability as specified by a parameter α. The length of this interval defines the uncertainty we have in estimating the unknown parameter.

2.1 Mean of a Single Population

Let us start with the case of a single population whose mean μ is unknown. To get the interval estimator, we use the sampling distribution of the point estimator. For μ, a point estimator is $\overline{X}$ and assuming σ^2 is known, we have from Lecture 8 that

$$\frac{\overline{X} - \mu}{\sigma/n} \sim Z$$

Given α, in defining the $(1-\alpha)100\%$ *confidence interval*, we make use of the sampling distribution and α:

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha \tag{4}$$

For example when $\alpha = 0.05$, 95% of Z lies between $-z_{0.025} = -1.96$ and $z_{0.025} = 1.96$. Then we leave the population parameter we are interested in alone and move all the things whose values we know outside and get

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{5}$$

Hence, $\left(\overline{X} - z_{\alpha/2}(\sigma/\sqrt{n}), \overline{X} + z_{\alpha/2}(\sigma/\sqrt{n})\right)$ is the $(1-\alpha)100\%$ *confidence interval* for μ.

Remember that $\overline{X}$ is the point estimator; the confidence interval can be viewed as indicating our uncertainty regarding our point estimate. We know that our point estimate will always be wrong, but how much it can be off is given by the confidence interval. The confidence interval states that if we draw samples of size n from the same population and calculate intervals like that for all, in $(1-\alpha)100\%$ of the time, the actual (unknown) μ will fall in the interval.

Because it is a measure of uncertainty, we want intervals to be as small as possible while having $1-\alpha$ as large as possible. We can view $z_{\alpha/2}(\sigma/\sqrt{n})$ as the *error term* and we see that this term increases with σ (as the variance in the original population increases so does the variance of $\overline{X}$) and decreases with $\sqrt{n}$ (as the sample size increases, the different samples become more alike and statistics calculated from them get similar). Actually if we have a bound b as to how large the error term should be, we can calculate how large n should be:

$$b \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow n \geq \lceil \left(z_{\alpha/2}\sigma/b\right)^2 \rceil \tag{6}$$

Above we assume that σ is known which is not very likely; if we do not know μ, we do not know σ either. When we do not know σ, we plug the sample standard deviation s in its stead and we know from Lecture 8 that $(\overline{X} - \mu)/(s/\sqrt{n})$ is t distributed with $n-1$ degrees of freedom. In such a case, we have

$$\begin{aligned} P\left(-t_{\alpha/2,n-1} < \frac{\overline{X} - \mu}{s/\sqrt{n}} < t_{\alpha/2,n-1}\right) &= 1 - \alpha \\ P\left(\overline{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right) &= 1 - \alpha \end{aligned} \tag{7}$$

Let us now consider a different setting. We draw s sample of size n from a population whose mean is unknown (assume σ is known) and then using this sample, we would like to make a prediction about the next, $n+1$st observation X_0. The point estimator would be $\overline{X}$, that is, the sample average over the n observations. The confidence interval for X_0 is called the *prediction interval*. We define a new random variable $X' = X_0 - \overline{X}$ where $E[X'] = \mu - \mu = 0$ and $\text{Var}(X') = \text{Var}(X_0) + \text{Var}(\overline{X}) = \sigma^2 + \sigma^2/n$ (X_0 and $\overline{X}$ are independent). Hence

$$\frac{(X_0 - \overline{X}) - 0}{\sqrt{\sigma^2 + \sigma^2/n}} \sim Z$$

which we use to define the $(1-\alpha)100\%$ confidence interval for the next observation X_0: *prediction interval*

$$P\left(\overline{X} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < X_0 < \overline{X} + z_{\alpha/2}\sigma\sqrt{1 + 1/n}\right) = 1 - \alpha \tag{8}$$

If we do not know, we use s instead of σ, and t_{n-1} instead of Z. Prediction interval can be used for outlier detection. An *outlier* is an observation that is very much different from the other observations and generally is a result of faults or errors; we would like to detect such outliers and discard them as otherwise they can corrupt the statistics we calculate over the sample. Given the n previous observations (for large enough n) if the $n+1$st do not lie in the prediction interval, we can consider it as outlier and discard. *outlier detection*

2.2 Difference of Means of Two Populations

Let us say we have two populations with unknown means μ_1 and μ_2 and we want to compare them. The variances may be known or unknown as we will see shortly. In comparing two means, we look at their difference $\mu_1 - \mu_2$, which is what we want to estimate.

We collect two independent random samples of sizes n_1 and n_2 using which we calculate $\overline{X}_1$ and $\overline{X}_2$ respectively, and the point estimator to $\mu_1 - \mu_2$ is $\overline{X}_1 - \overline{X}_2$. To get the interval estimator, we need the sampling distribution of the point estimator.

We know that $\overline{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$ and $\overline{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$, then $E[\overline{X}_1 - \overline{X}_2] = \mu_1 - \mu_2$ and $\text{Var}(\overline{X}_1 - \overline{X}_2) = \sigma_1^2/n_1 + \sigma^2/n_2$ and therefore

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma^2/n_2}} \sim Z$$

So the $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

$$P\left(\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma^2/n_2} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma^2/n_2}\right) = 1 - \alpha \tag{9}$$

This assumes the variances are known; if they are not, they are estimated and plugged in and we use the t distribution instead of Z.

For example, let us say we draw two random samples from two populations whose means are equal, that is, $\mu_1 = \mu_2$. In such a case, we will not have $\overline{X}_1 - \overline{X}_2 = 0$, but we expect the interval $(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma^2/n_2}$ to contain zero.

2.3 Paired Difference of Means of Two Populations

Let us say we want to compare the success of students in two courses Phys101 and Math101. We can do this as above, by first randomly choosing n_1 students and recording their grades for Phys101 and then randomly choosing another n_2 students and recording their Math101 grades, and then looking at the difference between the two average grades.

However we know that the grade of student in a course is not only influenced by the course but by all sorts of factors that have an effect on the student or on the environment, so in checking for the difference between the courses, if possible, we would like to set equal all other factors that may have an effect. If the Phys101 grades are by a different set of students, any difference we detect may not be because of the difference of the courses but may be because of the students. So then a better strategy would be to choose n students that take both courses and for each student, look at the difference at the observation level and then check for the average of these differences, rather than averaging samples separately and looking at the difference of the averages. This is called *pairing*.

We collect $i = 1, \ldots, n$ observations from two populations and in each observation, we use $d_i = X_{1i} - X_{2i}$. A $(1-\alpha)100\%$ confidence interval for $\mu_d = \mu_1 - \mu_2$ is

$$P\left(\overline{d} - t_{n-1,\alpha/2}s_d/\sqrt{n} < \mu_d < \overline{d} + t_{n-1,\alpha/2}s_d/\sqrt{n}\right) = 1 - \alpha \tag{10}$$

where $\overline{d}$ and s_d are the average and standard deviation of d_i.

Let us consider d_i. If X_{1i} and X_{2i} are independent, then $\text{Var}(d_i) = \text{Var}(X_{1i}) + \text{Var}(X_{2i})$, but in pairing, because they come from the same source (e.g., student), they are dependent, and actually they are positively correlated: If a student is smart or lives in conditions that are suitable for studying, his/her grades will be high for both courses and if not, his/her grades will be low for both courses, that is, $\text{Cov}(X_{1i}, X_{2i}) > 0$. Hence, $\text{Var}(d_i) = \text{Var}(X_{1i}) + \text{Var}(X_{2i}) - 2\text{Cov}(X_{1i}, X_{2i}) < \text{Var}(X_{1i}) + \text{Var}(X_{2i})$. This is the advantage of pairing.

Note that pairing is not always possible and should be used with care; we need to make sure that $\text{Cov}(X_{1i}, X_{2i}) > 0$ holds. In particular, note that from two samples of total size of $n_1 + n_2$ observations, we get a sample of size n, which implies a decrease in sample size and hence in the degrees of freedom.

2.4 Proportions as Means

Remember that even if X_i are not normal, unless n is very small ($n \geq 30$), we can still write $(\overline{X} - \mu)/(\sigma/n) \sim Z$ due to the central limit theorem. We know from earlier lectures that this is for example true for the binomial distribution which is the sum of 0/1 Bernoullis.

Let us say p_0 is the unknown probability of "success" for Bernoulli and we want to estimate it, for example, it is the probability of heads for tossing a particular coin. We toss the coin n times and see X heads. The point estimator for p_0 is $\hat{p}_0 = X/n$. To get the interval estimator, we need the sampling distribution of $\hat{p}_0$.

$X/n = X_1/n + X_2/n + \cdots + X_n/n$ where $X_i \in \{0, 1\}$ ($E[X_i] = p_0$ and Var$(X_i) = p_0(1-p_0)$) and from the central limit theorem, X/n is approximately normal. $E[\hat{p}_0] = np_0/n = p_0$ ($\hat{p}_o$ is an unbiased estimator) and Var$(\hat{p}_0) = np_0(1-p_0)/n^2 = p_0(1-p_0)/n$. Therefore $\hat{p}_0 \sim N(p_0, p_0(1-p_0)/n)$ and

$$\frac{\hat{p}_0 - p_0}{\sqrt{p_0(1-p_0)/n}} \sim Z \tag{11}$$

and we can write a $(1-\alpha)100\%$ confidence interval for p_0 as

$$P\left(\hat{p}_0 - z_{\alpha/2}\sqrt{\hat{p}_0(1-\hat{p}_0)/n} < p_0 < \hat{p}_0 + z_{\alpha/2}\sqrt{\hat{p}_0(1-\hat{p}_0)/n}\right) = 1-\alpha \tag{12}$$

Note how we used $\hat{p}_0$ instead of p_0 in the variance term—this is not ideal but inevitable, because the unknown parameter of Bernoulli defines both the mean and the variance.

Similarly one can derive the point and interval estimators for the difference of two proportions.

2.5 Variance of a Single Population

Assume we have a normal population whose variance σ^2 is unknown. We collect a sample of size n and the point estimator is the sample variance s^2. To get the confidence interval, we need the sampling distribution of the point estimator which is $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$, which we use to define a $(1-\alpha)100\%$ confidence interval for σ^2:

$$\begin{aligned} P\left(\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right) &= 1-\alpha \\ P\left(\frac{(n-1)}{\chi^2_{n-1,\alpha/2}}s^2 < \sigma^2 < \frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}}s^2\right) &= 1-\alpha \end{aligned} \tag{13}$$

Note that unlike for the case of means (which uses symmetric Z or t) where the interval is calculated by adding two error terms (one less than, one greater than 0) to the point estimate, here with the χ^2 distribution, the interval is calculated by multiplying the point estimate by two factors (one smaller than, one larger than 1).

2.6 Ratio of Variances of Two Populations

When we have two populations and want to compare their variances, we look at their ratios (rather than differences as we do with the means). We collect two independent samples of sizes n_1 and n_2 and the point estimator for σ_1^2/σ_2^2 is s_1^2/s_2^2. To get the interval estimate, we need the sampling distribution, and we know from Lecture 8 that

$$\frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} \sim F_{n_1-1,n_2-1}$$

which we use to define a $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2:

$$\begin{aligned} P\left(F_{n_1-1,n_2-1,1-\alpha/2} < \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} < F_{n_1-1,n_2-1,\alpha/2}\right) &= 1-\alpha \\ P\left(\frac{1}{F_{n_1-1,n_2-1,\alpha/2}}\frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < F_{n_1-1,n_2-1,\alpha/2}\frac{s_1^2}{s_2^2}\right) &= 1-\alpha \end{aligned} \tag{14}$$

where we used the fact that $F_{n_1-1,n_2-1,1-\alpha/2} = 1/F_{n_1-1,n_2-1,\alpha/2}$. Note that as with the single population case, the two bounds of the interval is found by multiplying the point estimate with two factors. For example, if we collect two samples from two populations where the first population has twice the variance of the second one, s_1^2/s_2^2 we calculate may not be equal to two, but with probability $1-\alpha$, the confidence interval above will contain two.