10. Sampling and Confidence Intervals

Assuming that we know the distribution of a random variable, how do we determine its "parameters" (namely, the mean and variance)? For example, assume that we know that a Poisson process generates arrivals to a system. How do we determine the parameter λ (rate of arrivals) of that distribution? The only way to "estimate" this parameter is to observe the system for a period of time and to obtain a set of measurements (or samples) from the underlying distribution (Poisson in our example). However, how do we go from a set of samples to the distribution parameters? In this chapter we show how we can estimate the distribution parameters from observations obtained through sampling.

10.1 Sample Mean and Variance

Let $\mathbf{X} = \{X_1, \ldots, X_N\}$ be the set of N samples selected at random from a population that follows some unknown probability mass function $f(x)$. We denote the *sample mean* by $\bar{X}$ and the sample variance by S^2. The sample mean and sample variance are given by:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{10.1}$$

and

$$S^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1} \tag{10.2}$$

Obviously, $\bar{X}$ and S^2 are themselves random variables. To understand why this is the case, one only needs to observe that if we obtain another set of samples, the new values of $\bar{X}$ and S^2 will likely be different from what we obtained before.

Thus, the sample mean and variance are **not** the same as the mean and variance of the "population." But, while the sample mean and variance are "random variables" that are not equal to the mean and variance of the underlying distribution $f(x)$, they are related to that distribution.

Box 10.1.1 An Example of Sample Mean and Standard Deviation

To give a concrete example, think about the response time for web server requests. Let's say that you send 5 requests to a web server and got response times: 30, 40, 20, 50, and 10 msec. You can think of these five values as a sample of the random response times of that web server. For this specific sample, the mean is 30 msec (150/5). Clearly if you send five other requests to that same server, you are likely to get five different values, say 35, 25, 30, 40, and 15 msec, whose mean is different from the first sample, namely 29 msec (145/5). So, every time we "sample" from a distribution, the sample mean (and any other function of the sampled values) is a random variable.

Specifically, the sample mean and variance are **estimates** of the distribution mean and variance. These estimates are likely to become more accurate as the size of the sample becomes larger. Indeed, at the limit, as the sample size approaches infinity, we would be able to characterize (with absolute confidence) the distribution's mean and variance.

However, obtaining infinite-size samples (if the population is infinite in size) is impractical. Thus, we have to be content with some level of sampling error (or inaccuracy). Luckily, this sampling error could be bounded (as we will see next). In other words, by knowing the size of a sample, we can estimate the confidence we have in that sample's "characterization" of the distribution.

10.2 Building Confidence Intervals

As we explained above, sample mean and sample variance are estimators of the mean and variance of a probability distribution (that is unknown to us). Building confidence intervals around these estimates is the process of "bounding" the distribution's mean and variance given the sample mean and variance.

The idea of confidence intervals is similar to that of bounding errors in numerical analysis. Thus, we motivate the concept of a confidence interval around an estimate using an analogy.

Consider the process of measuring a quantity x. If we can assert that the error in the measurement of x is less than ε then we can state that the "accurate" value of x lies anywhere between $x-\varepsilon$ and $x+\varepsilon$. For example, if x is a distance measured with a ruler that has a resolution of $1/16^{th}$ of an inch, then the actual (accurate) value for whatever we are measuring is anywhere between $x-0.0625''$ and $x+0.0625''$.

In the above example, establishing bounds around the estimate was easy, because we know the "resolution" of our measuring device (namely ε). The process is more involved when we use sample mean and variance to estimate distribution mean and variance. In particular, the question is *What is the resolution of a sampling process?* To answer this question we rely on the Central-Limit Theorem (see Section 8.2).

Let $X_1, X_2, X_3, \ldots, X_N$ be the values of a random sampling from a population with a finite mean μ and a finite variance σ^2. Obviously, these samples are independent random variables, each of which follows the same distribution.

Recall that the Central-Limit Theorem states that if $X_1, X_2, X_3, \ldots, X_N$ are random variables with identical probability distributions, each with a finite mean μ and a finite standard deviation σ^2 then the sum of these random variables approaches a **Normal distribution** with mean $N\mu$ and variance $N\sigma^2$. Specifically, the central limit theorem states that the random variable Z follows a *standard normal distribution* $N(0,1)$, where Z is given by:

$$Z = \frac{\left(\sum_{i=1}^{N} X_i\right) - N\mu}{\sigma\sqrt{N}} \tag{10.3}$$

We can then rewrite this expression as:

$$Z = \frac{N \cdot \left(\frac{\sum_{i=1}^{N} X_i}{N} - \mu\right)}{\sigma\sqrt{N}} = \frac{\frac{\sum_{i=1}^{N} X_i}{N} - \mu}{\frac{\sigma}{\sqrt{N}}} \tag{10.4}$$

But recall that $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$. With this in mind, we can write:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \tag{10.5}$$

where Z is a random variable that follows a standard normal distribution.

We wish to define an "interval" around in such a way that the probability that the actual mean of the distribution falls within the interval is larger than a given bound, as illustrated by the diagram in Figure 10.1.

We define the confidence interval to be $\bar{X} \pm E$. Thus, we would like to find the value of E such that:

$$P(\bar{X} + E > \mu > \bar{X} - E) > (1 - \alpha) \tag{10.6}$$

We can then write:

$$P(\bar{X} + E > \mu > \bar{X} - E) = P(E > \bar{X} - \mu > -E) = P\left(\frac{E}{\frac{\sigma}{\sqrt{N}}} > \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} > \frac{-E}{\frac{\sigma}{\sqrt{N}}}\right) > (1 - \alpha) \tag{10.7}$$

Substituting Z from Equation 10.5, we get:

$$P\left(\frac{-E}{\frac{\sigma}{\sqrt{N}}} < Z < \frac{E}{\frac{\sigma}{\sqrt{N}}}\right) > (1 - \alpha) \tag{10.8}$$

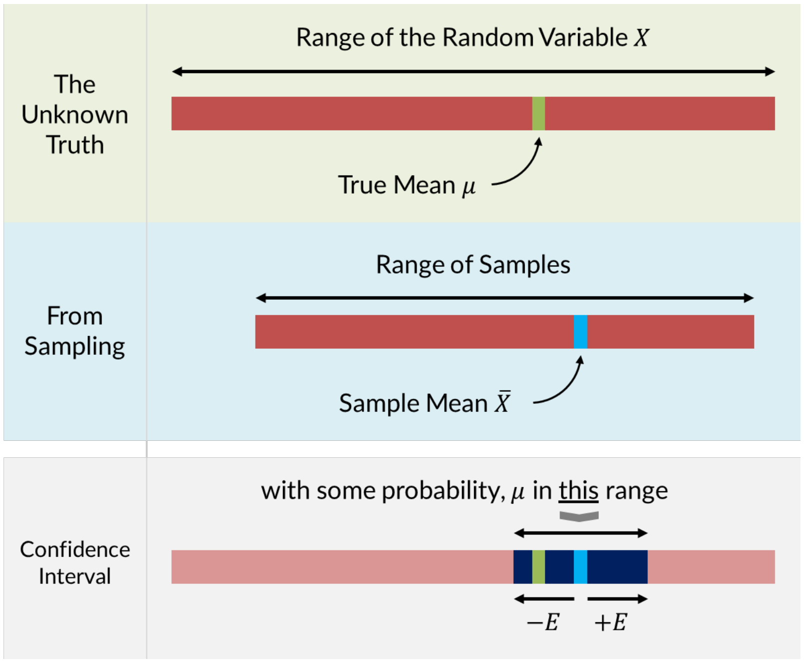

Figure 10.1: A confidence interval is used to expresses the probability that the true but unknown mean of the sampled distribution μ is within the range $[\bar{X}-E, \bar{X}+E$.

The standard normal random variable Z is symmetric. Thus, we can write the above inequality as follows:

$$P\left(Z < \frac{E}{\frac{\sigma}{\sqrt{N}}}\right) > \left(1 - \frac{\alpha}{2}\right) \tag{10.9}$$

By definition of cumulative distribution function, we can then write:

$$F\left(\frac{E}{\frac{\sigma}{\sqrt{N}}}\right) > \left(1 - \frac{\alpha}{2}\right) \tag{10.10}$$

I order to re-write the equation above to find E, one would need to reverse the $F(x)$ function for the normal distribution. This is no trivial endeavor. For this reason, a look-up table is used instead to find the value, say $Z_{\alpha/2}$, such that $F(Z_{\alpha/2}) = \left(1 - \frac{\alpha}{2}\right)$.

At this point, we imposed that E because we know that:

$$\frac{E}{\frac{\sigma}{\sqrt{N}}} = Z_{\alpha/2} \tag{10.11}$$

and thus:

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \tag{10.12}$$

The above equation allows us to define an interval around $\bar{X}$, namely $\bar{X} \pm E$ in such a way that the probability that the "actual" mean is outside this interval is bounded by $(1-\alpha)$.

There is still one caveat though! Equation 10.12 enables us to give the confidence interval around the mean, but to do so we need to know the distribution's standard deviation (i.e. σ)! The good news is that for large enough sample sizes, we may approximate σ by the standard deviation of the sample (i.e. $S = \sqrt{S^2}$ – see Equation 10.2). This leads to the following formula for computing E, and hence the confidence interval $\bar{X} \pm E$:

$$E = Z_{\alpha/2} \frac{S}{\sqrt{N}} \tag{10.13}$$

The above relationship shows that the width of the confidence interval is related to S and N (for a given level of confidence). In particular, if the standard deviation of our measurement is small, then we can give a tighter range for μ Similarly, we can tighten the range for μ by increasing the number of samples (or observations) N. In particular, the interval width is inversely related to the square root of the sample size. Thus, we can cut the confidence interval in half by quadrupling the sample size. Similarly, we can reduce the confidence interval ten-fold by increasing the size of our sample 100 times!

Obviously, to be "right on the money" in our estimation of the mean (i.e. to reduce E to zero), then one of three conditions must hold. Namely: (1) the standard deviation is zero (i.e. we detect no variability) in our measurements; (2) the sample size becomes infinite (i.e. we sample the whole population); or (3) we can be content with a confidence of 0%! This is why it is impossible for any sampling to give us a deterministic "answer."

There are many uses of Equation 10.13. For example, if we want our confidence interval to be tighter than some given bounds, then we can use Equation 10.13 to figure out how many samples we need for the calculation of . Alternatively, if we are given a number of measurements (or samples) from an experiment, then we can calculate the probability that the "actual" mean will be within a given interval.

Example: To exemplify the use of confidence intervals, assume that we need to calculate a 95% confidence interval for the mean of some distribution. Furthermore, assume that we have 100 samples, and that the sample mean $\bar{X} = 5$ and that the sample variance $S^2 = 9$. Since we want a 95% confidence interval, it follows that $\alpha = 1 - 0.95 = 0.05$. It follows that $1 - \alpha/2 = 0.975$. Now we look at Table 8.2 and find the value $Z_{\alpha/2}$ for which $F(Z_{\alpha/2}) = 0.975$. This results in a value of $Z_{\alpha/2} = 1.96$ (row at 1.9, column at 0.06). Using Equation 10.13, we can calculate $E = Z_{\alpha/2} \cdot \sqrt{\frac{S^2}{N}} = 1.96 \cdot 3/10 = 0.588$, which leads to a confidence interval around the mean given by $\bar{X} \pm E$, i.e. the range $[4.412, 5.588]$.

10.3 Correlation Metrics

Imagine that we are performing two experiments concurrently. For example, we are counting the number of people that walk into a bank per minute and at the same time (at some other neighboring location) we are counting the number of people walking into a restaurant. Let the sequence of measurements for the bank be X_i, where $i = 1,2,3,\dots,N$ and let the sequence of measurements for the restaurant be Y_i, where $i = 1,2,3,\dots N$. Intuitively, we say that the two sequences are "correlated" if whenever the measurements for the bank go up/down, we notice a similar trend in the measurements for the restaurant (and vice versa). Obviously, correlation can be strong or weak. Thus, we want to find a quantitative measure of *how correlated two sequences are*. How do we do that?

Well, let $\bar{X}$ denote the mean of the sequence X and $\bar{Y}$ denote the mean of the sequence Y. If the two sequences are correlated, then we expect that whenever X_i is larger than $\bar{X}$, then similarly Y_i would be larger than $\bar{Y}$. Also, whenever X_i is smaller than $\bar{X}$, then similarly Y_i would be larger than $\bar{Y}$.

More formally, we define the **covariance** for two sequences X and Y to be:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}) \cdot (Y_i - \bar{Y}) \tag{10.14}$$

Notice that the quantity within the summation is positive only when both X and Y are on the same side of their respective means. Thus, the overall quantity $Cov(X,Y)$ will be larger if the two sequences are correlated. Under such a condition, we say that the two sequences are **positively correlated**. It is also possible that the two sequences are correlated in a different way. This happens when X and Y are on the opposite sides of their respective means. Under such a condition, we say that the two sequences are **negatively correlated**.

However, covariance can produce values of an arbitrary magnitude, and hence it is not possible to directly compare two covariance values to each other. Therefore we also define a scaled version of that measure, which we call the **correlation** of the two sequences:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{1}{N}\frac{\sum_{i=1}^{N}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sigma_X \cdot \sigma_Y} \tag{10.15}$$

where σ_X is the standard deviation of the samples of X, and σ_Y is the standard deviation for the samples of Y.

One can easily show that the above quantity ranges between -1 and 1.

Correlation is a scaled version of covariance; note that the two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be positively correlated; when the sign is negative, the variables are said to be negatively correlated; and when the sign is 0, the variables are said to be **uncorrelated**. As these terms suggest, covariance and correlation measure the level of dependence between the variables. Sequences of values that are independent (e.g., "temperature in Arizona" and "pressure in Alaska") would produce a correlation that would approach zero.

Example: In a computing system, the number of processes in the "ready" queue and the number of processes in the "blocked" queue at 20 time instants were given by the following sequences: Ready Queue: (20, 15, 17, 6, 2, 3, 3, 5, 8, 9, 8, 12, 16, 22, 38, 29, 32, 35, 28, 33) and Blocked

Queue: (7, 9, 10, 8, 8, 4, 2, 2, 1, 4, 2, 6, 4, 7, 7, 6, 9, 12, 14, 11). Calculate the correlation coefficient between these two sequences.

Sometimes we are also interested in finding out if there are correlations within the same sequence of measurements.

Consider the sequence of measurements of the number of people walking into a bank per minute. One question that we may want to answer is whether subsequent samples are correlated. For example, if a sample is "large" then the next sample is large (and vice versa). This correlation is not necessarily between a sample and the next sample. For example, it may be the case that a sample is correlated with the sample after the next one, etc. It may be the case that the samples are totally uncorrelated. In general, we define the auto-correlation function for a sequence to be:

$$AutoCorr(X,d) = \frac{1}{N}\frac{\sum_{i=1}^{N}(X_{(i-d)} - \bar{X}) \cdot (X_i - \bar{X})}{\sigma_X^2} \tag{10.16}$$

where d is a parameter that controls the distance between samples (which we are interested in checking for correlation). Notice that $AutoCor(0) = 1$, which is to say that each sample is maximally correlated with itself!

In the previous example with a sequence of twenty observations of the number of processes in the Ready Queue, one may be interested in finding out if the number of processes in the ready queue at $t = i$ are correlated with those at $t = i + 1$, etc. Clearly, if the time interval between the two measurements is small enough, then in a typical computing systems one would expect them to be correlated. If a system is overloaded in one instant, it is likely to continue to be overloaded in the next instant. One could also be interested in figuring out if a system which is overloaded at $t = i$ would be under-loaded at $t = i + 12$ hours (which would be the case if the auto correlation computed with $d = 12$ hours would be negative).

Example: Compute the auto-correlation of the number of processes in the Ready Queue using $d = 1$ and using $d = 10$. What can you say about that sequence?

Recall the arrival independence concept that we discussed earlier in the course when presenting the Poisson arrival model!? Well, to say that arrivals are independent is the same as saying that the auto correlation function for any value of d is *zero* for any **infinite** sequence of values that represent Poisson arrivals. Notice that the word "infinite" in the previous sentence is crucial! In particular, for any finite sequence the auto correlation function is unlikely to be "exactly" zero, but rather a small value that approaches zero (as the length of the sequence increases).