CmpE 343 Lecture Notes
8: Sampling

Ethem Alpaydın

December 4, 2014

1 Population vs. Sample

Given a random experiment, the *population* consists of the whole set of observations. Populations may be very large and even infinite. *Sampling* is the process of randomly choosing an observation from the population; a *sample* is the set of such instances and is generally a small subset of the population. Any value calculated from the sample is a *statistic* and in *statistical inference,* we would like to extract information about the population from the sample.

For example, let us say we are carrying out a healthy weight study of college students in Turkey. In such a case, our population is the set of *all* college students in Turkey, but because we cannot possibly observe and carry measurements on this whole population, we choose a random sample. We calculate statistics on this sample, for example, we can calculate the body mass index (BMI) values[1] on this sample and infer about the BMI values in the whole population from this sample. For example, the average we calculate over the sample gives us information about the mean of the population, and the range of values in the sample gives us information about the variance of the population distribution.

Each observation in the sample is the result of a random selection and as such is represented by a random variable X having the (unknown) population distribution $f(x)$. The sample $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ therefore is a set of random variables and any statistic calculated from the sample, for example, their average, is also a random variable. It is very important that the sample be unbiased and reflects as much as possible the full characteristics of the population. The sample should be a *random sample* where the observations are independent and they are all drawn from the same underlying population; hence we write the joint probability of the sample as

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n) \tag{1}$$

As another example, consider testing. For quality control, a manufacturer cannot possibly test all the items coming out of the production line, because testing is costly and sometimes testing destroys the item. Instead, a small random sample is taken and tested, and the aim is to infer about the quality of the whole population from this small sample.

2 Some Example Statistics

There are certain statistics that we calculate from the sample and they give us a lot of information about the underlying population.

The *sample average* is defined as

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2}$$

which is sometimes denoted as m.

The *sample variance* is defined as

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \tag{3}$$

[1] Body Mass Index is weight in kq/(height in m)^2, and someone with BMI over 25 is considered overweight; see `http://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmi-m.htm`.

and its square root is the *sample standard deviation*. We will later discuss why we divide by $n-1$, and not n.

Generally, when we collect data, for example, after a set of experiments, we summarize the sample and report its sample average and standard deviation, i.e., as $m \pm s$. But this is meaningful only if the sample has a single group symmetric around the average, for example, if it is approximately normal. Otherwise, just these two numbers are not enough and we need to display the whole sample, for example by plotting its *histogram*.

A sample may contain noisy observations. An *outlier* is a value that is very much different from other observations, and may arise as a result of errors in transmission or recording, for example, typing errors, faulty sensors, and so on. Outliers may have a harmful effect on the statistics and the idea in *robust statistics* is to use statistics minimally affected by noise. For example, as a measure of central tendency, the sample median is more robust than the sample average.

The *sample median* is the value halfway when sorted. Let us say we sort the X_i values so that $X_{(1)}$ is the smallest, i.e., $X_{(1)} \equiv \min_{i=1}^{n} X_i$, and $X_{(n)}$ is the largest: $X_{(n)} \equiv \max_{i=1}^{n} X_i$—this is called *order statistics*. Then the sample median is

$$X_{\text{median}} = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } n \text{ is even} \end{cases} \tag{4}$$

For example, given the sample $\{2, 1, 3\}$, both the average and median is 2, but if the sample is $\{2, 1, 30\}$, the median is still 2 but the average is 11. We do not want single instances to have such a large effect on our inferences. Note that the sample variance is not robust to outliers either.

If we want to report the range of possible values, it is not a good idea to report it as the range from the minimum to the maximum, again due to possible outliers in the data. A *quantile* $q(f)$ is defined as the value such that the fraction f of the observations in the sample is less than or equal to $q(f)$. For example, sample median is $q(0.5)$. $q(0.25)$ and $q(0.75)$ are the lower and upper *quartiles* and contains half of the sample between them and such, they may be used instead of $m \pm s$ as a robust range in which the central bulk of the sample lies. Similarly it is better to use $q(0.05)$ instead of the minimum, and $q(0.95)$ instead of the maximum. The *box-and-whisker plot* uses these quantiles.

3 Sampling Distribution

It is important to always keep in mind that from the same population, one can get different random samples. For example, from the population of college students in Turkey, in two different surveys, one can choose two different (but both random) samples. Because the samples are different, the statistics we calculate will also be different. For example we may see a different sample average in each sample. But because they all have the same underlying population distribution, we expect them to be close (and actually close to μ, the population mean). These different sample averages follow, what we call, a *sampling distribution*.

3.1 Sample Average from a Single Population

Let us see how can calculate the sampling distribution for $\overline{X}$. Assume X_i are normal with mean μ and variance σ^2. We see in equation (2) that the sample average is a linear combination of X_i each of which is normal and we know that linear combinations of normals is also normal; hence $\overline{X}$ is also normal. Let us derive its expected value and variance:

$$E[\overline{X}] = E\left[\frac{\sum_i X_i}{n}\right] = \frac{\sum_i E[X_i]}{n} = \frac{n\mu}{n} = \mu \tag{5}$$

$$\text{Var}(\overline{X}) = \text{Var}\left(\frac{\sum_i X_i}{n}\right) = \frac{\sum_i \text{Var}(X_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{6}$$

So $\overline{X}$ has the same mean with X_i, but its variance is divided by n. This makes sense: As n gets larger, the different samples of size n over which we calculate the different $\overline{X}$ will be more and more similar, and hence the calculated $\overline{X}$ will be more and more close to each other, and also to μ.

A very important point is that, even if the underlying population (from which X_i are drawn) is not normal, because of the central limit theorem, $\overline{X}$ will be approximately normal. Note that equations (5) and (6) always hold regardless of the distribution of X_i (as long as they are independent and identically distributed (iid)). Hence in both cases (and therefore without actually caring for the underlying distribution), we can write $\overline{X} \sim N(\mu, \sigma^2/n)$ as the sampling distribution, or equivalently

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim Z \tag{7}$$

Given such a sampling distribution, we can use it for various purposes: For example, given μ and σ^2, we can calculate the probability that $\overline{X}$ is in a certain given range, using this sampling distribution by plugging the values in equation (7) and reading the probabilities from the table/function for the Z distribution. We define z_α such that $P(Z > z_\alpha) = \alpha$. Remember that Z is symmetric around its mean (zero) and therefore, $z_{1-\alpha} = -z_\alpha$; for example, $z_{0.05} = 1.645$ and $z_{0.95} = -1.645$.

Or, more interestingly, if we have a sample of size n, we can calculate $\overline{X}$ and if we know σ^2, we can calculate an interval in which the unknown μ is highly likely to lie—this is called a *confidence interval*. Or let us say somebody makes a claim about the value of μ; we take a sample of size n, calculate $\overline{X}$ and if we know σ^2, we can calculate the *confidence interval* for μ, and then we reject the claim if the claimed value for μ lies outside of this interval—this is called *hypothesis testing*. We will discuss these in more detail in later lectures.

3.2 Difference of Sample Averages from Two Populations

Let us say we have two populations with means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively. From these we draw two samples independently of sizes n_1, n_2 and we calculate the sample averages $\overline{X}_1, \overline{X}_2$. For example, let us say we want to compare the Math 101 grades of CmpE and EE students and $\mu_1 - \mu_2$ is the difference between the grades—for example, we can say that the performances of students from the two departments are comparable if $\mu_1 - \mu_2$ is close to zero. So we sample n_1 CmpE students and n_2 EE students, calculate $\overline{X}_1, \overline{X}_2$ and then look at $\overline{X}_1 - \overline{X}_2$. What can we say about the sampling distribution of $\overline{X}_1 - \overline{X}_2$?

From the previous section, we know that $\overline{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$ and $\overline{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$. Therefore, we see that $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$, or equivalently

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim Z \tag{8}$$

3.3 Sample Variance from a Single Population

Assuming $X_i \sim N(\mu, \sigma^2)$, let us derive the sampling distribution of s^2 (defined in equation (6)):

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

Let us divide both sides by σ^2:

$$\begin{aligned}
\frac{\sum_i (X_i - \mu)^2}{\sigma^2} &= \frac{\sum_i (X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2} \\
\sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2 &= \frac{(n-1)s^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2
\end{aligned}$$

On the left, each $(X_i - \mu)/\sigma$ is Z, its square is chi-squared with one degree of freedom, so the left hand side is chi-squared with n degrees of freedom. On the right, the second term is similarly is chi-squared with one degree of freedom (equation (7)). From Cochran's theorem, the degrees of freedom add up and we say that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \tag{9}$$

The domain of the chi-squared distribution is nonnegative (remember these are variances). For large degrees of freedom, the chi-squared converges to the normal (the sum of chi-squared are chi-squared where we sum also the degrees of freedom) but for small degrees of freedom, it is skewed and not symmetric around its mean.

We see that $\sum_i (X_i - \mu)^2/\sigma^2$ is χ^2_n, but $\sum_i (X_i - \overline{X})^2/\sigma^2$ is χ^2_{n-1}; when we plug $\overline{X}$ instead of μ, we lose one degree of freedom—In the first case, you can pick n numbers however you like, in the second case, you again pick n numbers but their average is fixed, so you lose one degree of freedom.

3.4 Sample Average from a Single Population with Unknown Variance

You may have noticed that equation (7) uses σ^2, but in many applications, when we do not know μ, we do not know σ^2 either. In such a case, we can use the sample standard deviation s instead of the population standard deviation σ in deriving the sampling distribution for $\overline{X}$. In such a case, the statistic is no longer standard normal, but is from another distribution called *student's t* or in short *t distribution*.

If Z is standard normal, X is chi-squared with ν degrees of freedom, and the two are independent,

$$\frac{Z}{\sqrt{X/\nu}} \sim t_\nu \tag{10}$$

where ν is the degrees of freedom of the t distribution, which you can consider as a parameter of the distribution.

We know that $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ is Z and $(n-1)s^2/\sigma^2$ is χ^2_{n-1}, hence

$$\frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{((n-1)s^2/\sigma^2)/(n-1)}} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \tag{11}$$

The t distribution looks very much like the Z; it is centered at zero and decreases on both sides symmetrically as we move away from zero. t has longer tails indicating more spread but this spread (uncertainty) decreases as n increases—for $n \geq 30$, we can use Z instead of t.

We define t_α such that $P(T > t_\alpha) = \alpha$. Just like Z, t is symmetric so $t_{1-\alpha} = -t_\alpha$.

3.5 Proportion of Two Sample Variances from Two Populations

Let us say we have two populations and we want to compare their variances. Instead of looking at their difference as we do with the means, it is easier to look at their proportion. We get two samples from the two populations of sizes n_1, n_2 and calculate the two sample variances s_1^2, s_2^2.

If U and V are two independent chi-squared random variables with degrees of freedom ν_1, ν_2 respectively, then $(U/\nu_1)/(V/(\nu_2))$ is F distributed with n_1 and n_2 degrees of freedom—the F distribution has two parameters.

In our case, we know that $(n_1 - 1)s_1^2/\sigma_1^2 \sim \chi^2_{n_1-1}$ and $(n_2 - 1)s_2^2/\sigma_2^2 \sim \chi^2_{n_2-1}$, so

$$\frac{((n_1 - 1)s_1^2/\sigma_1^2)/(n_1 - 1)}{((n_2 - 1)s_2^2/\sigma_2^2)/(n_2 - 1)} = \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} \sim F_{n_1-1, n_2-1} \tag{12}$$

The domain of F is also nonnegative, and an interesting aside is that $F_{1-\alpha} = 1/F_\alpha$.