3. Making estimates

The most basic kind of inference about a population is an estimate of the location (or shape) of a distribution. The central limit theorem says that the sample mean is an unbiased estimator of the population mean and can be used to make a single point inference of the population mean. While making this kind of inference will give you the correct estimate on average, it seldom gives you exactly the correct estimate. As an alternative, statisticians have found out how to estimate an interval that almost certainly contains the population mean. In the next few pages, you will learn how to make three different inferences about a population from a sample. You will learn how to make interval estimates of the mean, the proportion of members with a certain characteristic, and the variance. Each of these procedures follows the same outline, yet each uses a different sampling distribution to link the sample you have chosen with the population you are trying to learn about.

Estimating the population mean

Though the sample mean is an unbiased estimator of the population mean, very few samples have a mean exactly equal to the population mean. Though few samples have a mean, exactly equal to the population mean, m, the central limit theorem tells us that most samples have a mean that is close to the population mean. As a result, if you use the central limit theorem to estimate μ, you will seldom be exactly right, but you will seldom be far wrong. Statisticians have learned how often a point estimate will be how wrong. Using this knowledge you can find an interval, a range of values, which probably contains the population mean. You even get to choose how great a probability you want to have, though to raise the probability, the interval must be wider.

Most of the time, estimates are interval estimates. When you make an interval estimate, you can say "I am z per cent sure that the mean of this population is between x and y". Quite often, you will hear someone say that they have estimated that the mean is some number "± so much". What they have done is quoted the midpoint of the interval for the "some number", so that the interval between x and y can then be split in half with + "so much" above the midpoint and - "so much" below. They usually do not tell you that they are only "z per cent sure". Making such an estimate is not hard— it is what Kevin Schmidt did at the end of the last chapter. It is worth your while to go through the steps carefully now, because the same basic steps are followed for making any interval estimate.

In making any interval estimate, you need to use a sampling distribution. In making an interval estimate of the population mean, the sampling distribution you use is the t-distribution.

The basic method is to pick a sample and then find the range of population means that would put your sample's t-score in the central part of the t-distribution. To make this a little clearer, look at the formula for t:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

n is your sample's size and $\bar{x}$ and s are computed from your sample. μ is what you are trying to estimate. From the t-table, you can find the range of t-scores that include the middle 80 per cent, or 90 per cent, or whatever per cent, for n-1 degrees of freedom. Choose the percentage you want and use the table. You now have the lowest and highest t-scores, $\bar{x}$, s and n. You can then substitute the lowest t-score into the equation and solve for μ to find one of the limits for μ if your sample's t-score is in the middle of the distribution. Then substitute the highest t-score into the equation, and find the other limit. Remember that you want two μ's because you want to be able to say that the population mean is between two numbers.

The two t-scores are almost always ± the same number. The only heroic thing you have done is to assume that your sample has a t-score that is "in the middle" of the distribution. As long as your sample meets that assumption, the population mean will be within the limits of your interval. The probability part of your interval estimate, "I am z per cent sure that the mean is between...", or "with z confidence, the mean is between...", comes from how much of the t-distribution you want to include as "in the middle". If you have a sample of 25 (so there are 24df), looking at the table you will see that .95 of all samples of 25 will have a t-score between ±2.064; that also means that for any sample of 25, the probability that its t is between ±2.064 is .95.

As the probability goes up, the range of t-scores necessary to cover the larger proportion of the sample gets larger. This makes sense. If you want to improve the chance that your interval contains the population mean, you could simply choose a wider interval. For example, if your sample mean was 15, sample standard deviation was 10, and sample size was 25, to be .95 sure you were correct, you would need to base your mean on t-scores of ±2.064. Working through the arithmetic gives you an interval from 10.872 to 19.128. To have .99 confidence, you would need to base your interval on t-scores of ±2.797. Using these larger t-scores gives you a wider interval, one from 9.416 to 20.584. This trade-off between precision (a narrower interval is more precise) and confidence (probability of being correct), occurs in any interval estimation situation. There is also a trade-off with sample size. Looking at the t-table, note that the t-scores for any level of confidence are smaller when there are more degrees of freedom. Because sample size determines degrees of freedom, you can make an interval estimate for any level of confidence more precise if you have a larger sample. Larger samples are more expensive to collect, however, and one of the main reasons we want to learn statistics is to save money. There is a three-way trade-off in interval estimation between precision, confidence, and cost.

At Foothill Hosiery, John McGrath has become concerned that the hiring practices discriminate against older workers. He asks Kevin to look into the age at which new workers are hired, and Kevin decides to find the average age at hiring. He goes to the personnel office, and finds out that over 2,500 different people have worked at Foothill in the past fifteen years. In order to save time and money, Kevin decides to make an interval estimate of the mean age at date of hire. He decides that he wants to make this estimate with .95 confidence. Going into the personnel files, Kevin chooses 30 folders, and records the birth date and date of hiring from each. He finds the age at hiring for each person, and computes the sample mean and standard deviation, finding $\bar{x}$ = 24.71 years and s = 2.13 years. Going to the t-table, he finds that .95 of t-scores with 29df are between ±2.045. He solves two equations:

$$\pm 2.045 = (24.71 - \mu)/\ (2.13/\sqrt{30})$$

and finds that the limits to his interval are 23.91 and 25.51. Kevin tells Mr McGrath: "With .95 confidence, the mean age at date of hire is between 23.91 years and 25.51 years."

Estimating the population proportion

There are many times when you, or your boss, will want to estimate the proportion of a population that has a certain characteristic. The best known examples are political polls when the proportion of voters who would vote for a certain candidate is estimated. This is a little trickier than estimating a population mean. It should only be done with large samples and there are adjustments that should be made under various conditions. We will cover the simplest case here, assuming that the population is very large, the sample is large, and that once a member of the population is chosen to be in the sample, it is replaced so that it might be chosen again. Statisticians have found that, when all of the assumptions are met, there is a sample statistic that follows the standard normal distribution. If all of the possible samples of a certain size are chosen, and for each sample, p, the proportion of the sample with a certain characteristic, is found, and for each sample a z-statistic is computed with the formula:

$$z = \frac{p - \pi}{\sqrt{\frac{(p)(1-p)}{n}}}$$

where π = proportion of population with the characteristic these will be distributed normally. Looking at the bottom line of the t-table, .90 of these z's will be between ±1.645, .99 will be between ±2.326, etc.

Because statisticians know that the z-scores found from sample will be distributed normally, you can make an interval estimate of the proportion of the population with the characteristic. This is simple to do, and the method is parallel to that used to make an interval estimate of the population mean: (1) choose the sample, (2) find the sample p, (3) assume that your sample has a z-score that is not in the tails of the sampling distribution, (4) using the sample p as an estimate of the population π in the denominator and the table z-values for the desired level of confidence, solve twice to find the limits of the interval that you believe contains the population proportion, p.

At Foothill Hosiery, Ann Howard is also asked by John McGrath to look into the age at hiring at the plant. Ann takes a different approach than Kevin, and decides to investigate what proportion of new hires were at least 35. She looks at the personnel records and, like Kevin, decides to make an inference from a sample after finding that over 2,500 different people have worked at Foothill at some time in the last fifteen years. She chooses 100 personnel files, replacing each file after she has recorded the age of the person at hiring. She finds 17 who were 35 or older when they first worked at Foothill. She decides to make her inference with .95 confidence, and from the last line of the t-table finds that .95 of z-scores lie between ±1.96. She finds her upper and lower bounds:

$$+1.96 = \frac{.17 - \pi}{\sqrt{\frac{(.17)(1-.17)}{100}}}$$

π = .17 -(.038)(1.96) = .095

and, she finds the other boundary:

$$-1.96 = \frac{.17 - p}{\sqrt{\frac{(.17)(1-.17)}{100}}}$$

π = .17 - (.038)(1.96) = .245

and concludes, that with .95 confidence, the proportion of people who have worked at Foothills Hosiery who were over 35 when hired is between .095 and .245. This is a fairly wide interval. Looking at the equation for constructing the interval, you should be able to see that a larger sample size will result in a narrower interval, just as it did when estimating the population mean.

Estimating population variance

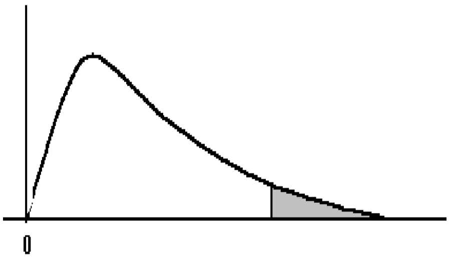
Another common interval estimation task is to estimate the variance of a population. High quality products not only need to have the proper mean dimension, the variance should be small. The estimation of population variance follows the same strategy as the other estimations. By choosing a sample and assuming that it is from the middle of the population, you can use a known sampling distribution to find a range of values that you are confident contains the population variance. Once again, we will use a sampling distribution that statisticians have discovered forms a link between samples and populations.

Take a sample of size n from a normal population with known variance, and compute a statistic called " χ^2 " (pronounced "chi square") for that sample using the following formula:

$$\chi^2 = \frac{\Sigma(x - \bar{x})^2}{\sigma^2}$$

You can see that χ^2 will always be positive, because both the numerator and denominator will always be positive. Thinking it through a little, you can also see that as n gets larger, χ^2 ,will generally be larger since the numerator will tend to be larger as more and more $(x-\bar{x})^2$ are summed together. It should not be too surprising by now to find out that if all of the possible samples of a size n are taken from any normal population, that when χ^2 is computed for each sample and those χ^2 are arranged into a relative frequency distribution, the distribution is always the same.

Because the size of the sample obviously affects χ^2 , there is a different distribution for each different sample size. There are other sample statistics that are distributed like χ^2 , so, like the t-distribution, tables of the χ^2 distribution are arranged by degrees of freedom so that they can be used in any procedure where appropriate. As you might expect, in this procedure, df = n-1. A portion of a χ^2 table is reproduced below.



The χ^2 distribution

	p	**.95**	**.90**	**.10**	**.05**
n	**df**				
2	1	0.004	0.02	2.706	3.841
10	9	3.33	4.17	14.68	19.92
15	14	6.57	7.79	21.1	23.7
20	19	10.12	11.65	27.2	30.1
30	19	17.71	19.77	39.1	42.6

Exhibit 4: The χ^2 distribution

Variance is important in quality control because you want your product to be consistently the same. John McGrath has just returned from a seminar called "Quality Socks, Quality Profits". He learned something about variance, and has asked Kevin to measure the variance of the weight of Foothill's socks. Kevin decides that he can fulfill this request by using the data he collected when Mr McGrath asked about the average weight of size 11 men's dress socks. Kevin knows that the sample variance is an unbiased estimator of the population variance, but he decides to produce an interval estimate of the variance of the weight of pairs of size 11 men's socks. He also decides that .90 confidence will be good until he finds out more about what Mr McGrath wants.

Kevin goes and finds the data for the size 11 socks, and gets ready to use the χ^2 distribution to make a .90 confidence interval estimate of the variance of the weights of socks. His sample has 15 pairs in it, so he will have 14 df. From the χ^2 table he sees that .95 of χ^2 are greater than 6.57 and only .05 are greater than 23.7 when there are 14df. This means that .90 are between 6.57 and 23.7. Assuming that his sample has a χ^2 that is in the middle .90, Kevin gets ready to compute the limits of his interval. He notices that he will have to find $\sum(x-\bar{x})^2$ and decides to use his spreadsheet program rather than find $(x-\bar{x})^2$ fifteen times. He puts the original sample values in the first column, and has the program compute the mean. Then he has the program find $(x-\bar{x})^2$ fifteen times. Finally, he has the spreadsheet sum up the squared differences and finds 0.062.

Kevin then takes the χ^2 formula, and solves it twice, once by setting χ^2 equal to 6.57:

χ2 = 6.57 = .062/σ2

Solving for σ^2, he finds one limit for his interval is .0094. He solves the second time by setting $x^2=23.6$:

23.6 = .062/σ2 a

and find that the other limit is .0026. Armed with his data, Kevin reports to Mr McGrath that "with .90 confidence, the variance of weights of size 11 men's socks is between .0026 and .0094."

What is this confidence stuff mean anyway?

In the example we just did, Ann found "that with .95 confidence..." What exactly does "with .95 confidence" mean? The easiest way to understand this is to think about the assumption that Ann had made that she had a sample with a z-score that was not in the tails of the sampling distribution. More specifically, she assumed that her sample had a z-score between ±1.96; that it was in the middle 95 per cent of z-scores. Her assumption is true 95% of the time because 95% of z-scores *are* between ±1.96. If Ann did this same estimate, including drawing a new sample, over and over, in .95 of those repetitions, the population proportion *would* be within the interval because in .95 of the samples the z-score would be between ±1.96. In .95 of the repetitions, her estimate would be right.