Confidence Intervals and Hypothesis Testing

# 1 Confidence Intervals

A simpler idea than confidence intervals (though one used much less frequently) is *prediction intervals*. If you knew that women's heights were normally distributed with a mean of 67 and a standard deviation of 2.5 inches, you would say "I am 95% certain that the next woman who walks through that door will have a height between 62 and 72 inches," using the Empirical Rule (more precisely, 95% of the data actually fall between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.). The precise interpretation would be clear: Of all the women who might walk through the door, you believe that 95% of them fall in that height range. Likewise you could say, I am "I am 90% sure that the average height of the 16 women in my stats class is between 65.97 and 68.03" using the fact that 90% of normal data falls within 1.645 standard deviations of the mean and that the standard deviation for the average of 16 women's heights is $\sigma/\sqrt{16} = 0.635in$. Again the interpretation is that if you took a lot of stats classes with 16 women in them, 90% would have average heights in that range.

It is trickier with confidence intervals. When you use a sample to estimate the average airspeed of a North American swallow, you report a 95% confidence interval of say $[26.2, 29.7]$. You are not saying that 95% of the time the average airspeed of all swallows is within that range, because the average airspeed is some particular number, which either fits in that range or it doesn't. In any statement about probability there is something you are imagining repeating over and over again, and in this case it is the process of taking a sample and computing a 95% confidence interval! What you are really saying is "95% of the time when I take a sample and compute a 95% confidence interval, it will be correct in that it includes the correct answer for the average airspeed. I like to think of it like this: when an old statistician looks back over her life and all the 95% confidence intervals she computed, she can be sure that if she did each one perfectly with good data, perfect assumptions and no mistakes, then 1 out of 20 of them were just wrong, they did not include the desired quantity within them.

Calculating them is not hard. If you know the standard deviation $\sigma$, then you know that 95% of all samples of size $n$ will have an average within $1.96\sigma/\sqrt{n}$ of the mean of the population, so to put it another way there is a 95% chance that for a given sample the true mean will fall in the interval

$$\overline{x} \pm 1.96\sigma/\sqrt{n}.$$

If you don't want a 95% confidence interval, use

| 90% | 95% | 99% |
|---|---|---|
| 1.645 | 1.96 | 2.57 |

instead of 1.96. Or more generally for an $\alpha$ confidence interval use the number $Z_{\alpha/2}$ that you get by plugging mean 0, standard deviation 1, and probability $1 - \alpha/2$ into the NORMINV function of Excel.

Of course, you don't usually know the true standard deviation in situations where you don't know the true mean, so this isn't much help. Instead you can use the sample standard deviation $s$ in place of $\sigma$, since it is a pretty good estimate, but then you have introduced a little

fuzziness into the estimate, so to account for that you have to use the fuzzier *t-distribution* instead of the normal distribution. In practice that means using the formula

$$\overline{x} \pm t_{\alpha/2,n-1} s/\sqrt{n},$$

where $t$ is the number you get from plugging $1-\alpha/2$ in for the probability and $n-1$ in for the "degrees of freedom" in Excel's TINV function. for a 95% interval $t$ will be larger than 1.96, a little larger if $n$ is large, and a lot larger if $n$ is small.

# 2 Hypothesis Testing

Here the question is "How strong is this evidence for a particular claim?" A great example to keep in mind is a court trial, where you (say you the jury) are trying to figure out how strong is the evidence that the accused is guilty. The answer is always the same. You assume the claim you are assessing the evidence for is false. In this case, you assume the defendant is guilty, and you ask "what is the probability we would see *evidence this strong* assuming he is innocent. If the probability is extremely small you are convinced that this evidence did not happen "by chance," and that he is in fact guilty. If the probability is not all that small, you "presume innocent until proven guilty" and acquit him. Of course you may still think it is likely that he is guilty, but are saying that the evidence was not strong enough to compel you to the drastic step of convicting him.

First some terminology. The statement you are assessing the evidence in favor of is called the *alternate hypothesis*. In our example this is "he's guilty." The statement you will assume to compute the probability of getting evidence like yours is called the *null hypothesis,* in this case "he's innocent." The null hypothesis is usually close to being the opposite of the alternate, as in this case, but not always (the null hypothesis needs to be very specific in order to compute a probability). If you have trouble picking out your null and alternate hypothesis here is some guidance. The alternate hypothesis is the claim you are assessing the evidence for, the claim you are trying to prove, the claim that involves taking strong or dramatic or new action over, the claim it would be worse to conclude incorrectly (falsely convicted is much worse than falsely exonerated). The null hypothesis is the claim that you will presume in the absence of evidence to the contrary, the claim that involves no action or response, the claim it would be better to assume falsely. When you compute the probability that you would get evidence as strong as what you got assuming the null hypothesis, this is called the $p$-value. It is the measure of how strong your evidence is (lower is stronger).

We also need to distinguish between *formal* hypothesis testing, and *informal* hypothesis testing. This and many other curious features of the process stem from the fact that it is very easy to let the conclusion you want to draw affect how you phrase the question or interpret the answer. Formal hypothesis testing is what is generally done in science, law, and other areas where the objectivity of the conclusions is paramount. Here you choose a cutoff ahead of time and declare if the $p$-value is below that cutoff, you will consider that sufficient evidence (we say then the data is *statistically significant*. Otherwise, you will declare the data insignificant and ignore it. That cutoff is called the significance level, $\alpha$, and each field generally has its

own standard, with common values for the significance level being $1\%, 2\%, 5\%$ and $10\%$. If the question gives a significance level, hey are asking for formal hypothesis testing and your answer should be either "this data is significant evidence for ...(the alternate hypothesis)" or "this data is not significant evidence for ...(the alternate hypothesis)." If they do not give you a significance level, you *should not make one up!* You should report the $p$ value, preferably in a sentence such as " the probability we would of seen evidence at least as strong for ... (the alternate hypothesis ... assuming ...(the null hypothesis).. is ...$p$... You can also give a subjective assessment of the conclusion (this is strong evidence, this is mild evidence, this is not evidence for...).

The first case we will consider is when the alternate hypothesis is "the population mean of some random variable is (different from/greater than/less than) some particular value $\mu_0$" where $\mu_0$ is typically the value that you would otherwise assume for it. In each case the null hypothesis is that the mean is equal to this value $\mu_0$. They differ only in how you interpret *evidence for the alternate hypothesis at least as strong as yours.* We'll cover how to do this next.