Statistics to English Translation, Part 2a: 'Significant' Doesn't Always Mean 'Important'

Nina Zumel*

December, 2009

In this installment of our ongoing Statistics to English Translation series[1], we will look at the technical meaning of the term "significant". As you might expect, what it means in statistics is not exactly what it means in everyday language.

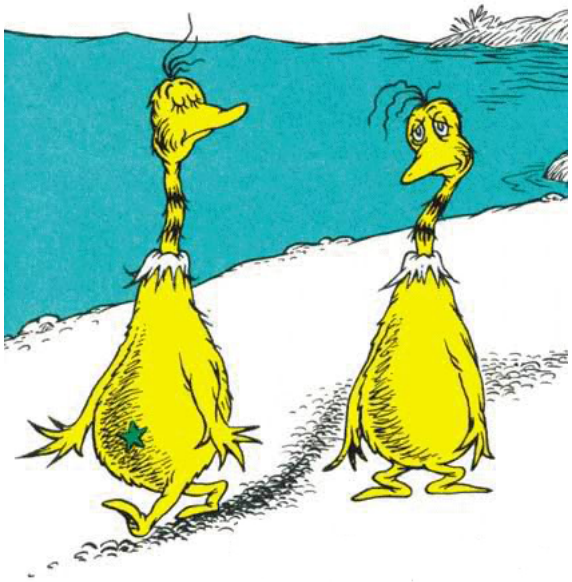
> Does too much salt cause high blood pressure, or doesn't it? That debate has raged for decades, with a slew of studies finding "yes" and a slew of others finding "no." Two new studies out today in the journal *Hypertension* tip the scales in favor of reducing sodium – particularly for those 1 in 4 Americans who have high blood pressure. One study found that reducing salt intake from 9,700 milligrams a day to 6,500 milligrams decreased blood pressure significantly in blacks, Asians, and whites who had untreated mild hypertension. Another study found that switching to a lower-salt diet helped lower blood pressure in folks with treatment-resistant hypertension.
>
> – "10 salt shockers that could make hypertension worse," *U.S. News & World Report* [Kot09]

"Great!" you think. "Who needs to spend money on high-blood pressure meds? I can just cut down my salt!" Well, maybe so, maybe not. To come to that conclusion, you need more information than you were given in that paragraph. What was the "significant" decrease in blood pressure? What was the "before" and the "after"? Does "significant" mean important, or useful? And why has there been so much controversy over this?

Let's discuss the important points with an example.

*http://www.win-vector.com/

[1] http://www.win-vector.com/blog/category/statistics-to-english-translation/



Suppose that we wanted to test for a difference in intelligence between two groups, say Star-Bellied Sneetches and Plain-Bellied Sneetches[2]. We take a group of 50 Star-Bellies and a group of 40 Plain-Bellies, and give them both a series of tests designed to measure their mathematical, linguistic, and problem-solving abilities. After evaluating the data, we conclude that there is "a significant difference in mathematical performance ($t(88) = 2.499$, $p = 0.014$) between the two groups". The mean mathematics score of the Star-Bellies is 78, with a standard deviation of 7, and the mean mathematics score of the Plain-Bellies is 74, with a standard deviation of 8, for a difference of 4 points[3].

Should we interpret this result to mean that Star-Bellied Sneetches are better than Plain-Bellied ones at math? It depends.

How Hypothesis Tests Work

The Sneetch example above and the blood-pressure study cited earlier are both examples of *hypothesis tests*. In hypothesis testing, researchers set their proposed hypothesis (that there is an effect or a relationship) against the *null hypothesis* that there is no effect or relationship. In this article, we consider proposed relationships of the form

> The mean value of X measured for group A is different from the mean value of X measured for group B.

[2] "The Sneetchs," from *The Sneetches and Other Stories* by Dr. Seuss. http://www.youtube.com/watch?v=Ln3V0HgW4eM (link) and http://www.youtube.com/watch?v=s0LgMpfLD1Y (link)

[3] This example is based on Exercise 10.17 in [SS99]; the original exercise did not, unfortunately, involve Sneetches.

In this case, the null hypothesis is

> The mean value of X is the same for groups A and B, and any difference observed in the data is only by observational chance.

In fact, we are actually testing the stricter null hypothesis:

> The distribution of X is the same for groups A and B, and any difference observed is only by observational chance.

A and B are sometimes called *treatment groups*; this terminology comes from the original applications of hypothesis testing procedures, in agriculture and medicine. In the blood pressure study above, the treatment is daily salt intake. One group ingests about 9,700 milligrams of sodium a day, the other group about 6,500 milligrams a day. The question of interest is: does the difference in sodium intake make a difference in the average blood pressure of the two groups? The null hypothesis is "No."

Significance

We call an observed difference *significant* – meaning that a difference as large as we observed is probably not by chance – if the the value $1 - p$ is "high enough." In the Sneetch example, $p = 0.014$ is the *significance level* of the result. To interpret the p-value, suppose the null hypothesis is true: there is truly no difference between Star-Bellied math scores and Plain-Bellied math scores. If this is so, then there is only a 0.014 (1.4%) chance that the difference in the average scores of the two groups will be 4 points or larger. In other words, if the null hypothesis is true, and we administer this same test to different groups of 50 Star-Bellies and 40 Plain-Bellies a hundred times, then the difference in scores will be 4 points or more only about once or twice.

We interpret the fact that we have seen a difference that should be rare to be evidence that the null hypothesis *isn't* true. So we *reject the null hypothesis* and say that there is a "significant difference" in the performance of the two groups. Alternatively, we could say that Star-Bellied Sneetches performed "significantly better" than Plain-Bellied Sneetches on the math test.

Effect Size

Four points (or about a 5% difference) is the *effect size* of the comparison. The effect size represents what might be called the "practical significance" of the result. In general, the larger the effect size, the better. In this example, Star-Bellies might truly outperform Plain-Bellies by about four points on average, but if we were to examine the relationship between math scores and real-life math performance (say, how well college-attending Sneetches do in their math and science courses), we might discover that it takes a test score difference of ten points or more to reliably predict which Sneetches will do better. In that case, a four point average difference would not be a practical difference.

Evaluating a Result

When evaluating a result, you should look both for its significance and its effect size. In practice, researchers usually consider a finding to be significant if $p \leq 0.05$. This is actually a pretty large p; it means even if the null hypothesis is true, you would still observe a difference as large as the one that you observed about five times out of every one hundred trials. In fact, Sachs noted that $p < 0.0027$ used to be the commonly used threshold for significance ([Sac84, p. 114]).

Sometimes results are reported using an asterisk convention: (*) means $p \leq 0.05$, (**) means $p \leq 0.01$, and (***) means $p \leq 0.001$. Hopefully, the actual significance level is reported (it isn't always), as well as the actual effect size (it isn't always).

The effect size in medical studies is often reported in the popular press with statements like "those who abstained from coffee had triple the risk of contracting colon cancer compared to those who drank three or more cups a day." Does that mean that all confirmed Lapsang Souchong drinkers and the uncaffeinated should run out and learn to embrace Starbucks? Well, no. First of all, ask yourself: what is the baseline risk of colon cancer? If abstaining from coffee triples the risk from 0.01% to 0.03%, well, it probably isn't worth worrying about. On the other hand, if the risk triples from 5% to 15%, perhaps that is a reason to take up espressos. You should also see who were the subjects of the study, and how similar they are to you. Suppose the study was done on Caucasian males in the U.S., ages 55-65, with no family history of colon cancer. If you are a young white American male, it's possible that this study says something about your future health. If you are female or non-Caucasian or not living in the U.S., the finding may or may not be relevant to you. It depends on the mechanism that drives the relationship, and whether or not it applies to you as well as to the subjects of the study.

"Significant" is not the same as "Important"

> With a large sample, even a small difference can be "statistically significant"... . This doesn't necessarily make it important. Conversely,

> an important difference may not be statistically significant if the sample size is too small.
>
> – Freedman, Pisani and Purves, *Statistics* [FPP07, p. 550]

The ability of a study to detect a significant difference depends almost entirely on its size. When a researcher designs a study, she has to decide how much risk of error – and what type of error – she is willing to tolerate.

> How big a risk [of inventing a difference] between two indistinguishable treatments are we willing to put up with? This risk is known as the significance level α. [Sac84, p. 214]

α is the probability of rejecting a null hypothesis that should be accepted. This is a Type I error (a false positive). α enters the design of the study as the threshold for p-values that the researcher will accept as significant.

> How big a risk do we allow of missing a substantial difference between two treatments? ... This risk is called β. [Sac84, p. 214]

β is the probability of accepting a null hypothesis that should have been rejected. This is a Type II error (a false negative). The quantity $1 - \beta$ is known as the *power* of the test: the probability that the test will correctly reject the null hypothesis when the alternative hypothesis is true.

> How small a difference should still be recognized as significant? This difference is called δ. [Sac84, p. 214]

δ is the minimum effect size that we are willing to consider "practically significant."

It is important to consider *all three* of α, β, and δ when determining an appropriate sample size for a trial. The power of a test and the significance of a result both increase as the sample size n increases. So if δ is not specified, **any difference can appear significant, with a large enough** n, even if the difference is really by chance.

The Central Limit Theorem

To see why the above statement is true, we need a few more facts about estimating the mean. Suppose we have a random variable X that is normally (or nearly normally) distributed, with a true mean μ and (unknown) variance σ^2. You want to estimate μ by drawing n samples; the sample mean $\bar{x}$ gives you an estimate of μ. According to the *Central Limit Theorem*, if you were to repeat this experiment over and over again, you would see that the estimated $\bar{x}$ has a normal distribution, with mean μ and variance σ^2/n. So $\bar{x}$ is a good estimate of μ, one that improves with a larger sample size n.

Another fact about normal distributions is that a little over 95% of the probability mass is within ± 2 standard deviations of the mean. So, for a single

experiment, we can reason that the true mean μ is in the interval $\bar{x} \pm 2\sigma/\sqrt{n}$ with 95% probability[4].

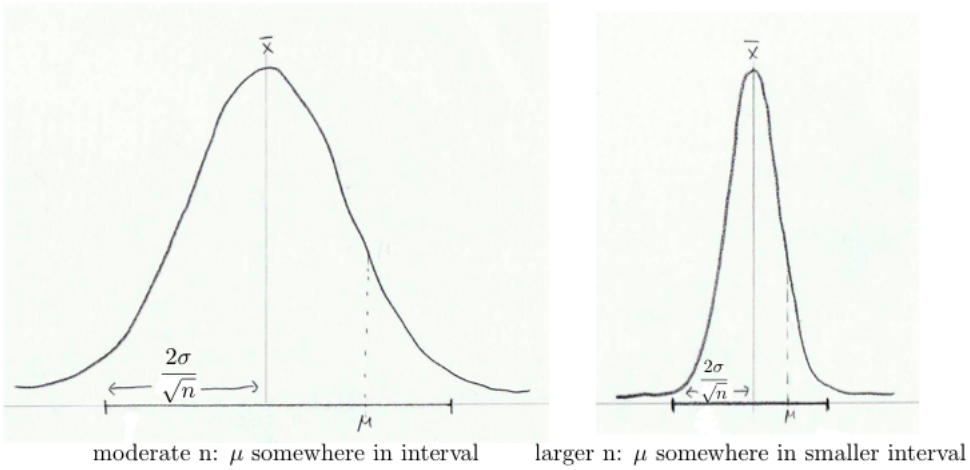


Figure 1: Confidence bounds on the estimate of μ for different values of n

So, as n gets larger, we zoom in on μ[5].

Now, back to the problem of checking for the difference of means. We'll take n samples from population A and n from population B. Let's assume for now that the variances are equal.

[4]The correct way to state this is that for a given (unknown) μ, the estimate $\bar{x}$ falls in the interval $\mu \pm 2\sigma/\sqrt{n}$ just over 95% of the time. This gets awkward to reason about. Luckily, symmetry arguments let us center the appropriate confidence interval around $\bar{x}$ instead.

[5]Of course, we don't actually know σ, so we don't know exactly how fast we zoom in. That doesn't affect our argument, though, since only n changes

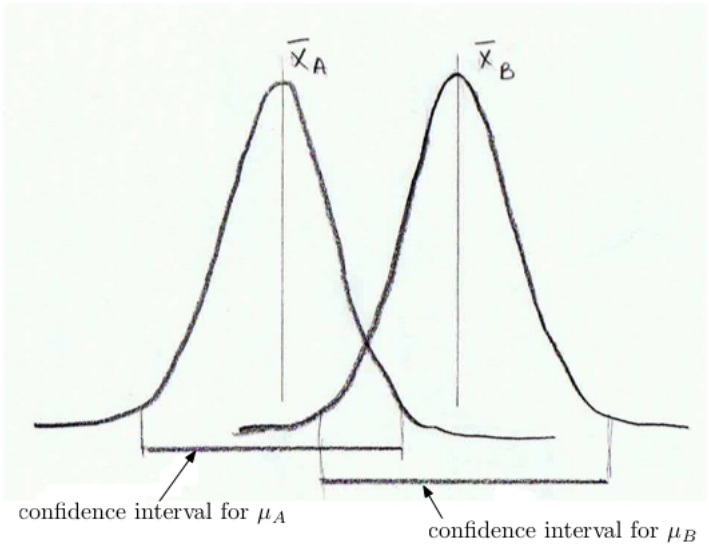


Figure 2: Confidence bounds overlap; means may not be truly different

With 95% probability, $\mu_A \in \bar{x}_A \pm 2\sigma/\sqrt{n}$, and $\mu_B \in \bar{x}_B \pm 2\sigma/\sqrt{n}$. If $|\bar{x}_A - \bar{x}_B|$ is small compared to $4\sigma/\sqrt{n}$, then the two confidence intervals overlap substantially, and we cannot reject the null hypothesis that $\mu_A = \mu_B$.

If, on the other hand, $|\bar{x}_A - \bar{x}_B|$ is wide compared to $4\sigma/\sqrt{n}$:

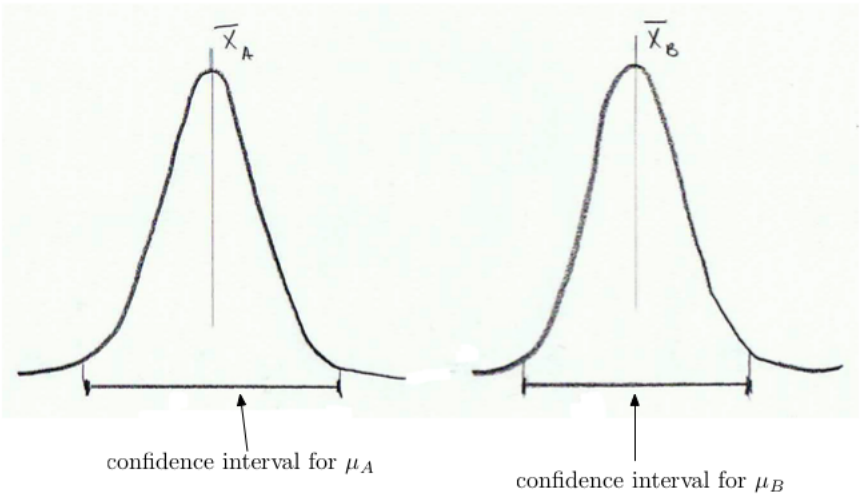


Figure 3: Confidence bounds don't overlap; means are significantly different

then the confidence intervals are well separated, and we can reject the null hypothesis.

So δ, the minimum significant distance – the "resolution" of the experiment – is about the distance when the two confidence intervals touch: $4\sigma/\sqrt{n}$, if our desired significance level is 0.05.

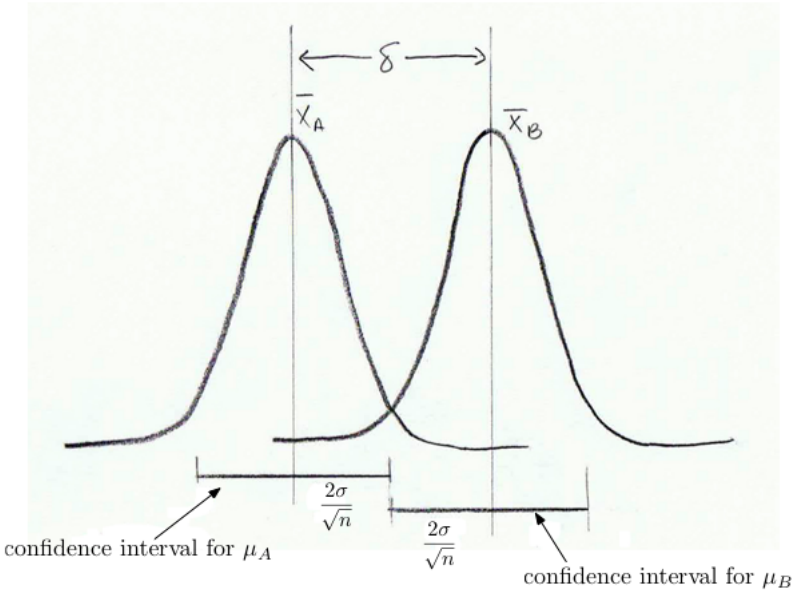


Figure 4: Minimum significant distance for a given sample size n

If δ is too large, the experiment may be unable to detect important differences because the confidence intervals overlap too soon. This means that the sample size was too small (the test didn't have enough power), and the experiment should be repeated with a larger test population.

If δ is too small, then the experiment will potentially detect statistically significant differences that are, for all practical intents and purposes, meaningless. To go back to the Sneetch example, if the math exam has one hundred questions, then an effect size of two points would correspond to one group answering two additional questions correctly, on average. Practically speaking, that's probably not a very big difference. But if we made the experiment big enough, about 250 Sneetches in each group, it would be a *statistically* significant difference, to the 0.05 level. In theory, we could even make a difference of less than one point statistically significant! That is why knowing the effect size of a significant result is important.

"Significant" is not the same as "True"

The power and significance level of a test play similar roles to the sensitivity and specificity of a diagnostic test. You'll remember from Part 1 of this series[6] that sensitivity and specificity are properties of the test, *not* how the test performs in a given population. To know the practical accuracy of a screening test, you must know the underlying prevalence of the condition that it is screening for. If it is crucial that the screening not miss any positive cases, then the test will be designed to be highly sensitive, possibly at the cost of specificity. In that case, the test will

[6]http://www.win-vector.com/blog/2009/11/i-dont-think-that-means-what-you-think-it-means-statistics-to-english-translation-part-1-accuracy-measures/ (link)

tend to have a high false positive rate if the condition is relatively rare. And yet, this same screening test will have a lower overall false positive rate when used in a population where the condition is more prevalent.

The same is true for hypothesis tests. The probability that a statistically significant result is actually *true* depends on the underlying probability that results "of that type" tend to be true in the domain of study. It also depends on whether the researcher was trying to minimize the chance of a false positive error, or a false negative error.

You should also be careful interpreting the results of exploratory work, where the researchers have run a series of several different studies, but only highlight the "significant" ones. Running twenty experiments and having one of them return a significant result to the $p = 0.05$ level is actually not significant at all.

John Ioannides discusses these points (and a few others) in his 2005 essay "Why Most Published Research Findings are False"[Ioa05]. The essay made a few waves at the time of its publication, and it is still available online. We recommend that you read it, along with the 2007 followup article by Moonesinghe, et.al [MKJ07]. Now that you've read the first two installments of the Statistics to English translation, both essays should be a breeze!

Some Points to Remember

- "Significant" is a statistical statement that an observed relationship is unlikely to be by chance. It is not an necessarily a statement about the magnitude or the importance (or the truth!) of the relationship.
- Knowing the effect size of a significant result will help you decide if the relationship is "practically significant."
- With a large enough sample size, any difference in means can appear significant, even when it is by chance.

You now have a general idea what a "statistically significant result" is. The next installment will go into a little more technical detail of how significance is calculated. You should read that installment if you want to decipher statements in research papers like "$(F(2, 864) = 6.6, p = 0.0014)$" – or if you are simply curious.

References

[FPP07] David Freedman, Robert Pisani, and Roger Purves, *Statistics*, 4th ed., W. W. Norton & Company, New York, 2007.

[Ioa05] John P. A. Ioannidis, *Why most published research findings are false*, PLoS Med **2** (2005), no. 8, e124, Available as http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124 (link).

[Kot09] Deborah Kotz, *10 salt shockers that could make hypertension worse*, U.S. News & World Report (2009), Online as http://health.usnews.com/articles/health/heart/2009/07/20/10-salt-shockers-that-could-make-hypertension-worse.html (link).

[MKJ07] Ramal Moonesinghe, Muin J Khoury, and A. Cecile J. W Janssens, *Most published research findings are false - but a little replication goes a long way*, PLoS Med **4** (2007), no. 2, e28, Available as http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0040028 (link).

[Sac84] Lothar Sachs, *Applied statistics: A handbook of techniques*, 2nd ed., Springer-Verlag, New York, 1984.

[SS99] Murray R. Spiegel and Larry J. Stephens, *Schaum's outline of statistics*, 4th ed., McGraw-Hill, New York, 1999.