Math 316

Section 7.6 More on Interpretation of Statistical Significance

Significant: The difference between the samples is large enough for us to conclude their respective populations are different. Important: So we've shown that there really is a difference in population means, but is this difference big enough to really matter? In **Class Example 1**, we compare male and female GPAs for college students nationwide (with made up numbers). Let's use $\alpha = 0.05$.

Sample	1	2	1	2	1	2
n	100	90	1000	900	10000	9000
$\bar{y}$	2.71	2.73	2.71	2.73	2.71	2.73
s	.41	.45	.41	.45	.41	.45
t_s	$-.319$		-1.009		-3.19	
P Table 4	$P > .40$		$.20 < P < .40$		$.001 < P < .010$	
P Exact	.750		.313		.0014	
Reject H_0?	No		No		Yes	

In all three cases the sample statistics (mean, standard deviation) are the same. But larger samples result in a larger test static and a smaller P value. In the first two cases, the difference between the two samples is *not* statistically significant (using $\alpha = 0.05$). That is, the samples are not different enough for us to conclude that there really is a difference nationwide in male and female GPAs. In the third case, the difference *is* statistically significant, in which case we would conclude that the two populations (male and female) really are different in average GPA. And so what? Is it important? Not necessarily. (Probably grad schools and employers wouldn't care about a 0.02 difference in GPA.) Let's look at **Example 7.6.7 with Table 7.6.3**.

Just like being *significant* is somewhat subjective (for example, do we want to be 80% confident or 90% or 95%?), being *important* is also subjective. Like significance, to me it's not like the difference between two means is either important or not. It's *to what degree* it is important. You might measure this by relative difference:

$$\frac{|\mu_1 - \mu_2|}{\mu_1} \text{ or } \frac{|\mu_1 - \mu_2|}{\mu_2} \text{ or } \frac{|\bar{y}_1 - \bar{y}_2|}{\bar{y}_1} \text{ or } \frac{|\bar{y}_1 - \bar{y}_2|}{\bar{y}_2}$$

Let's work **Class Example 2.**

We can also look at effect size, where σ is the larger population standard deviation,

$$effect\ size = \frac{|\mu_1 - \mu_2|}{\sigma} \text{ or (more commonly) } effect\ size = \frac{|\bar{y}_1 - \bar{y}_2|}{s}$$

where s the larger sample standard deviation. Often both populations, thus the samples, have similar standard deviations. BTW, this feels similar to a z-value: difference relative to (i.e. divided by) standard deviation.

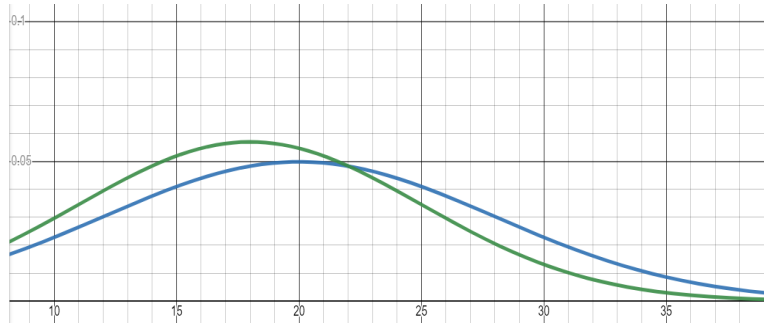
With effect size, we are computing how different the one sample mean is from the other, relative to how spread out the data is. Larger effect ⇔ the samples are more different. Effect size is basically the *opposite* of the amount of *overlap* between the two samples:

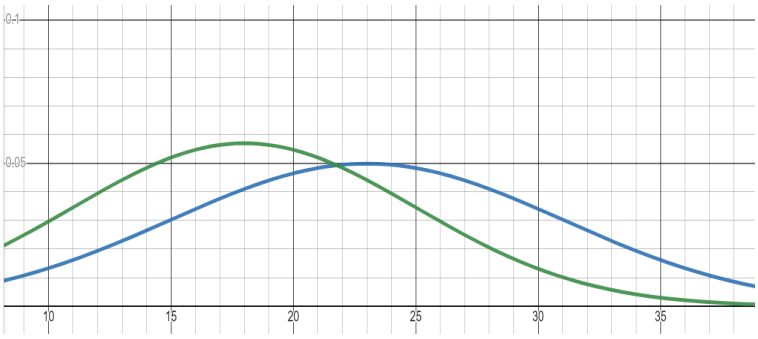
Samples are not too different ⇔ Smaller effect size (more overlap)

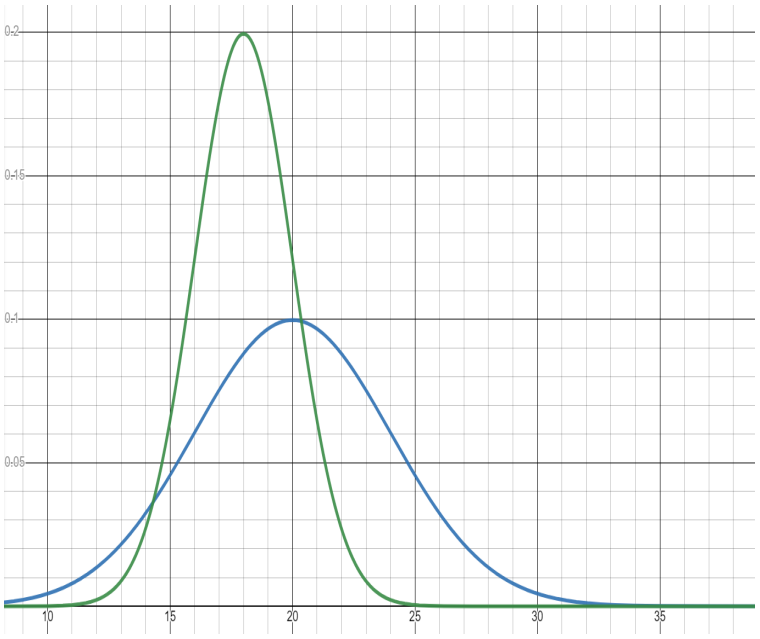
Samples are more different ⇔ Larger effect size (less overlap).

Let's discuss the Example below. $H_0: \mu_1 = \mu_2$, $H_A: \mu_1 \neq \mu_2$. We'll use $\alpha = 0.10$.

Case	Original		Larger $\bar{y}_1 - \bar{y}_2$		Smaller s_1 and/or s_2		Larger samples	
Sample	1	2	1	2	1	2	1	2
n	16	25	16	25	16	25	110	100
$\bar{y}$	20	18	23	18	20	18	20	18
s	8	7	8	7	4	2	8	7
Effect size	2/8 = .250		5/8 = .625		2/4 = .500		2/8 = .250	
t_s	0.819		2.048		1.857		1.932	
P	0.420		0.050		0.079		0.055	
Conf. int. for $\mu_1 - \mu_2$	(-2.153, 6.153)		(0.847, 9.153)		(0.138, 3.862)		(0.289, 3.711)	
$\mu_1 \neq \mu_2$, $i.e.\ \mu_1 - \mu_2 \neq 0$?	Not sure		Yes		Yes		Yes	
Reject H_0, accept H_A?	No		Yes		Yes		Yes	







Observations (all of which we've seen before, but now described along with effect size):

- Two things that will result in a narrower confidence interval: smaller standard deviations in the samples and/or larger sample sizes. Both of these make us happy.
- Larger effect size occurs when t_s is larger, so larger effect size means we're more likely to end up concluding that the populations (from which the samples come) have different means.
- Even with the same effect size, if the samples are large enough, we can still detect (determine, decide) that the population means really are different. It's like our test is more sensitive to (and trusts more) even small difference in samples.

See **Examples 7.6.5 – 7.6.7** for examples of what difference is and is not important.

So why the term "effect"? We're often interested in the effect of something on some group of things, e.g. the effect of a drug on treating an illness or the effect of a certain diet on losing weight. If there is more effect, the two samples (one control, one being treated) will be more different, and there will be less overlap between the samples:

Samples are more different $\Leftrightarrow$ Larger effect size (less overlap of samples).

Section 7.7 Planning for Adequate Power

Situation	What we should do	What we actually do	Did we err?
Population means are approximately the same	Don't reject H_0	Don't reject H_0	Nope
		Reject H_0, accept H_A	Type I error
Population means are different	Reject H_0, accept H_A	Don't reject H_0	Type II error
		Reject H_0, accept H_A	Nope

Confidence is $1-\alpha$, where α is the max chance of making a Type I error.
Power is $1-\beta$, where β is the max chance of making a Type II error.

Then <u>confidence level</u> is the likelihood that we will:

- Not make a Type I error.
- Not reject H_0 when we should not.
- Not decide the populations are different if in fact they are not different.

Then <u>power</u> is the likelihood that we will

- Not make a Type II error.
- Reject H_0 when we should.
- Decide the populations are different if in fact they are different.
- For example, if $\beta = .10$, then there is (at most) at 10% chance we might make a Type II error, and (at least) a 90% chance that we will not. More formally, Power = Pr{significant evidence for H_A} if H_A is true.

The trade-off between confidence and power: if α is smaller, we are *less likely to make a Type I error*, since we will be less likely to reject H_0 and accept H_A (including in the case

that we should reject H_0), but we are *more likely to make a Type II error*. Conversely, decreasing the likelihood of a Type II error increases the likelihood of a Type I error.

There are certain things we can control and others we cannot control regarding power. Things that will *increase* power:

- Larger difference between sample means, i.e. larger $\bar{y}_1 - \bar{y}_2$. The more different the samples are, the more likely it is we'll conclude the populations are different, if that is actually the case.
- Smaller sample standard deviations s_1 and/or s_2 (since more variation in the sample makes us trust the sample mean less, which makes us less certain about deciding that the two population means really are different just because the sample means are).
- Larger sample sizes: in general, whatever we are deciding or concluding about the populations based on the samples, we will believe it even more if we have more data (in this case, larger samples) to support that claim—it's a trade-off: we will spend more time and money for larger samples, but we'll get more reliable results.
- Larger α, that is, greater tolerance for Type I errors (since the less tolerance we have for Type I errors, the more likely it is that we will make Type II errors)—it's a trade-off, as mentioned above.

Given a particular Type I error tolerance (such as $\alpha = .05$), the one thing we can control is sample size. Table 5 helps us know the necessary minimum sample size to attain a certain power level given a certain effect size $\frac{|\mu_1 - \mu_2|}{\sigma}$ which we approximate with $\frac{|\bar{y}_1 - \bar{y}_2|}{s}$.

Even though Table 5 is a little strange at first and you will likely wonder where they came up with all of these sample size values (you can see Appendix 7.1 on page 614 for all of the beautiful details), it is basically telling us that larger sample sizes result in:

- Higher confidence: larger $1 - \alpha$ (smaller α), *less chance of making a Type I error*.
- More power: larger $1 - \beta$ (smaller β), *less chance of making a Type II error*,
- Hypothesis testing that is more sensitive to differences in the samples, that is, effect size.

As with Tables 3 and 4, we could use technology for a more complete set of values for α and power and effect size $\frac{|\mu_1 - \mu_2|}{\sigma}$, rather than just those in Table 5. Also, in general, it turns out that the power will be maximized if the sample sizes are the same, so that is usually what we (approximately) try to do.

If time, let's work **HW 7.7.3**. Also see **Example 7.7.3** and the thoughts at the end of that example.