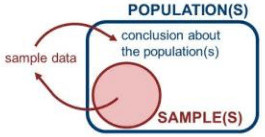
SE POWER ANALYSIS

It's about α & $(1-\beta)$

Quick review



We often use sample data to test hypotheses about population data.

But sampling error (i.e., noise) means that our samples are sometimes misleading.

This leads to statistical **errors** (due to randomness).

These are not **mistakes** (that's doing the math wrong).

Mistakes can be avoided, error cannot.

Classifying statistical errors

Type I error: Rejecting a true null hypothesis

Type II error: Failing to reject a false null hypothesis

Reality	Conclusion: Accept H_0	Conclusion: Reject H_0
H_0 true	Correct	**Type I error** α
H_0 false	**Type II error** β	Correct

α: the probability of rejecting a true null hypothesis
β: the probability of failing to reject a false null hypothesis

Understanding α and β

α: the probability of rejecting a true null hypothesis.

- The risk of deciding there is a real pattern or difference if there isn't one.
- Easily reduced by specifying smaller p value for test.

β: the probability of failing to reject a false null hypothesis

- The risk of not seeing a real pattern or difference.
- Depends on many factors.
- The value $1-\beta$ is the **power** of a statistical test.

The variance is small

Consider two homoscedastic populations differing by 3.0 and two samples of n=10. For the t-test, $t_{crit} = t_{18,0.025} = 2.101$

If sample var is 8: $t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{3.0}{\sqrt{8\left(\frac{1}{10}+\frac{1}{10}\right)}} = \frac{3.0}{1.2649} = 2.372 > 2.101$

If sample var is 11: $t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{3.0}{\sqrt{11\left(\frac{1}{10}+\frac{1}{10}\right)}} = \frac{3.0}{1.483} = 2.023 \ngtr 2.101$

Large sample variances obscure real differences.

The pattern/difference is strong/large

Consider two homoscedastic populations with Var=8.0 and two samples of n=10. For the t-test, $t_{crit} = t_{18,0.025} = 2.101$

If means differ by 3: $t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{3.0}{\sqrt{8\left(\frac{1}{10}+\frac{1}{10}\right)}} = \frac{3.0}{1.2649} = 2.372 > 2.101$

If means differ by 2: $t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{2.0}{\sqrt{8\left(\frac{1}{10}+\frac{1}{10}\right)}} = \frac{2.0}{1.2649} = 1.581 \ngtr 2.101$

Small magnitudes of the difference obscure real differences.

What influences the power, $1-\beta$?

In general, the power of a test will be higher when:

- The pattern/difference is strong/large. It's easier to detect big differences (Beavis Effect).
- The variance is small.
- The sample size is large.
- The test is parametric vs nonparametric. Parametric tests use more information from the sample.

Understanding α and β

Reality	Conclusion: Accept H_0	Conclusion: Reject H_0
H_0 true	Correct	**Type I error**
H_0 false	**Type II error**	Correct

Choice of α used to minimize this

We estimate β to determine this

$1-\beta$ is the power of our test
What we can or can't see when we look

The sample size is large

Consider two homoscedastic populations differing by 3.0 with Var=8, two samples of size n=10 or n=8.

If n=10: $t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{3.0}{\sqrt{8\left(\frac{1}{10}+\frac{1}{10}\right)}} = \frac{3.0}{1.2649} = 2.372 > 2.101$

If n=8: $t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{3.0}{\sqrt{8\left(\frac{1}{8}+\frac{1}{8}\right)}} = \frac{3.0}{1.4142} = 2.121 \ngtr 2.145$

Smaller sample sizes obscure real differences.

The test is parametric vs nonparametric

Nonparametric tests discard some of the information so have less power. The power comparisons are more complicated.

Mann-Whitney U test vs unpaired t-test (with large samples)

$$\frac{(1-\beta)_{MWU}}{(1-\beta)_{t-test}} \rightarrow \frac{3}{\pi} = 0.955$$

Wilcoxon signed-rank test vs paired t-test (with large samples)

$$\frac{(1-\beta)_{WSR}}{(1-\beta)_{t-test}} \rightarrow \frac{2}{\pi} = 0.6437$$

The main use of power analysis

Using an α=0.05 for the risk of type I error is standard. Planning for a $(1-\beta)$=0.80 is common, but not as orthodox.

When planning a study, calculations of power are common:

- How big does the study need to be to detect a relevant pattern? (how many mice need die, how much $).
- Can we detect a relevant pattern with the data available? (when data sets have fixed sizes)

The second use of power analysis

We can also calculate the power after the fact.

If we do a statistical test and fail to reject the null hypothesis, we can calculate *what we could have detected* based on our sample size and observed variance.

Amateur conclusion: "We fail to reject the null hypothesis and therefore conclude that there's no difference in the population means."

Professional conclusion: "We fail to reject the null hypothesis, which suggests that any difference is less than approximately ... "

Recap

Power of a test is $1-\beta$ (β is the probability of a type II error).
Power increases when:

- The pattern/difference is stronger/larger.
- The variance is smaller.
- The sample size is larger.
- The test is parametric vs nonparametric.

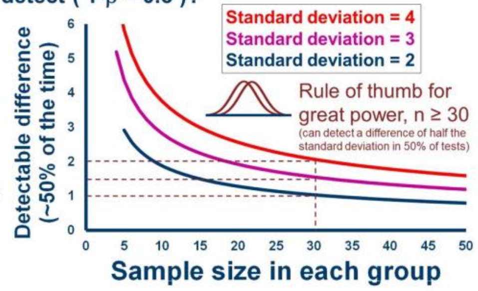
Power is calculated:

- Before an experiment to determine the design and cost needed to detect a relevant pattern/difference.
- After an experiment to describe range of patterns/differences that could have been detected.

What can we detect ($1-\beta = 0.5$)?

Larger sample size is better, but there are *diminishing returns*.

I.e., doubling the sample size doesn't halve the detectable difference.



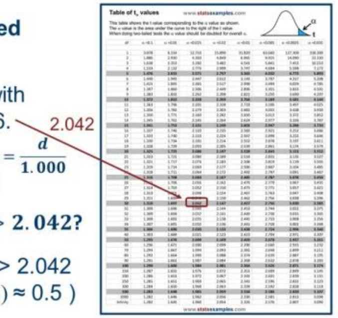
What we *could have* detected

Consider a t-test of two homoscedastic populations with Var = 8.0 and two samples 16.

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{d}{\sqrt{8\left(\frac{1}{16}+\frac{1}{16}\right)}} = \frac{d}{1.000}$$

$$\frac{d}{1.000} > 2.042?$$

This study can only detect d > 2.042 (i.e., about half the time, $(1-\beta) \approx 0.5$)



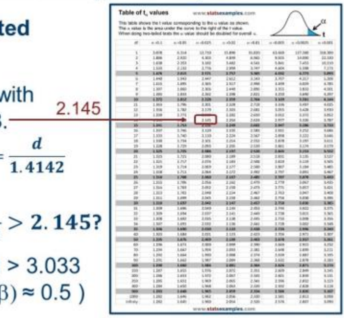
What we *could have* detected

Consider a t-test of two homoscedastic populations with Var = 8.0 and two samples 8.

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{d}{\sqrt{8\left(\frac{1}{8}+\frac{1}{8}\right)}} = \frac{d}{1.4142}$$

$$\frac{d}{1.4142} > 2.145?$$

This study can only detect d > 3.033 (i.e., about half the time, $(1-\beta) \approx 0.5$)



StatsExamples.com