3. Statistical Power

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Describing data

- Quantitative vs. categorical variables; discrete vs. continuous
- Histograms and densities for continuous variables
- Measures of central tendency (mean, median), location (percentiles), variability (variance, standard deviation)

Inferential statistics

- The importance of *quantifying uncertainty*: confidence intervals and significance testing
- Sampling distributions: what you would expect to see from an estimator (like $\bar{x}$) over *repeated sampling*.
- *Standard error*: a measure of variability in the sampling distribution.

Tonight

Statistical power: given we are using a sample, do we have enough information to detect "real" effects?

- Relatedly: how large of a sample size do we need to detect an effect of a given size?
- How can we improve our ability to detect real effects?

Murnane & Willett discuss their review of the What Works Clearinghouse (`https://ies.ed.gov/ncee/wwc/`). Most studies of educational policies and programs cannot support causal inference due to:

- Omitted variables bias (most)
- Lack of statistical significance: positive impact but *under-powered*

Tonight's sample datasets

We will refer to one dataset tonight (on Brightspace):

1. `nyvoucher.dta`: pre and post-test scores from the New York Scholarship Program, a randomized experiment of private school vouchers in NYC (see M&W ch. 4). Also used in Lecture 2.

Motivating example

Suppose we are evaluating a randomized intervention in rural India designed to improve literacy among primary school children.

$$H_0 : \mu_T - \mu_C = 0$$
$$H_1 : \mu_T - \mu_C \neq 0$$

We estimate $\bar{x}_T - \bar{x}_C = 6$ with a 95% confidence interval of (-2, 14). Can we conclude that the program had a causal effect on literacy? What if this estimate had been based on a larger sample?

Murnane & Willet: research designs are a "magnifying glass" for detecting effects, and there are things we can do to make our magnifying glass more powerful.

Type I and Type II errors

Type I and Type II errors

A hypothesis test can result in one of two types of incorrect decisions:

- **Type I error**: rejecting H_0 when it is actually true
- **Type II error**: not rejecting H_0 when it is false

Type I and Type II errors

	Reject H_0	**Do not reject H_0**
H_0 is true	Incorrect decision: ***Type I error* ($Pr = \alpha$)**	Correct decision: $(Pr = 1 - \alpha)$
H_0 is false	Correct decision: $(Pr = 1 - \beta)$	Incorrect decision: ***Type II error* ($Pr = \beta$)**

Note: the β symbol used here is not the same as a regression slope coefficient! Statistics often uses the same symbols for multiple purposes...

Type I and Type II errors

Example: criminal court

- H_0: not guilty (presumption of innocence)
- **Type I error**: rejecting H_0 and convicting an innocent person
- **Type II error**: not rejecting H_0 and letting a guilty person go free

There are costly consequences for both types of errors. Guilt "beyond a reasonable doubt" implies that a very low p-value is required for conviction (i.e., a low threshold for a Type I error α).

Type I and Type II errors

Example: PSA (prostate specific antigen) screening for prostate cancer

- H_0: no cancer
- **Type I error**: false positive—finding elevated levels of PSA and inferring a cancer growth when it does not exist
- **Type II error**: false negative—failing to detect an actual cancerous growth when it does exist

There are costly consequences for both types of errors:

- If a Type II error is made, growth exists and is untreated
- If a Type I error is made, detect tumor and perform unnecessary surgery

Probability of a Type II error

Consider a one-sided hypothesis test for a population mean μ:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

You take a random sample from the population of size n and calculate the sample mean $\bar{x}$ (and s) in order to test this hypothesis.

Probability of a Type II error

The probability of a Type II error is a bit more difficult to calculate than the probability of a Type I error (which is set by the researcher in advance as α). This is because the probability of a Type II error depends on how far away the effect we failed to detect is (often denoted δ).

- All else equal, we will be *more* likely to make a Type II error if the true μ is close—but not equal to—μ_0.
- All else equal, we will be *less* likely to make a Type II error if the true μ is far away from μ_0.

Probability of a Type II error

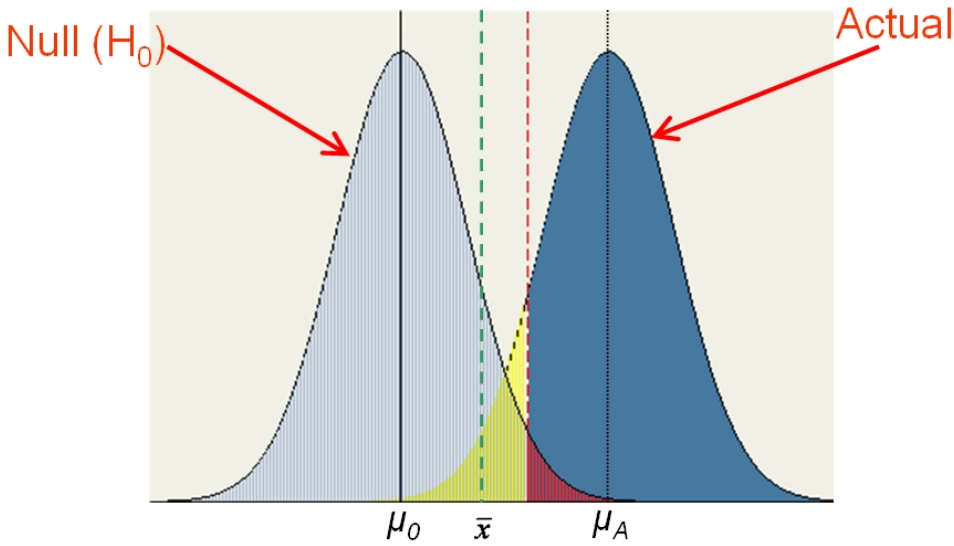


Figure: Distribution of $\bar{x}$ under H_0 and a specific alternative H_A

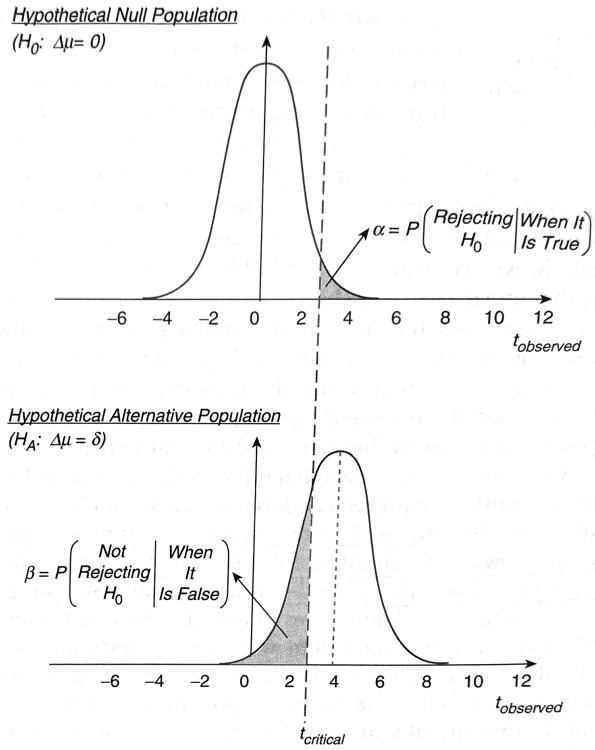
Note: δ is the difference between μ_0 and μ_A

Probability of a Type II error

A few notes on the previous figure:

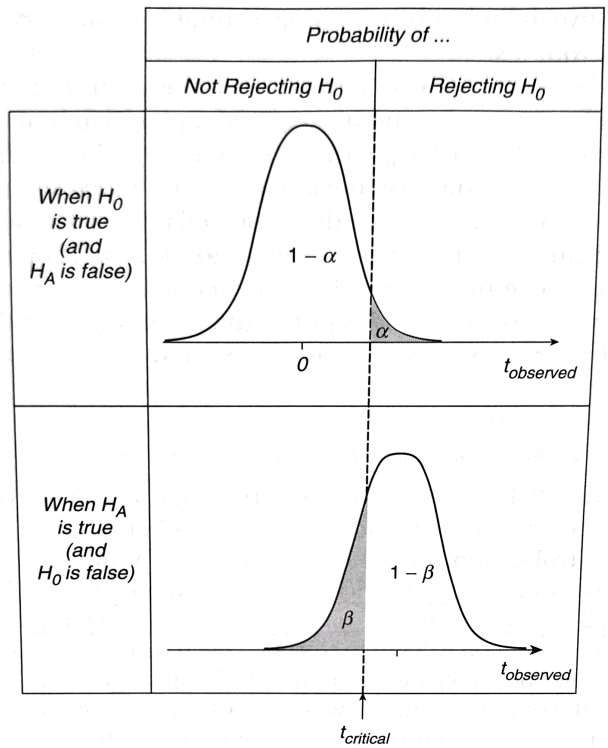
- The sampling distribution of $\bar{x}$ has the same shape and standard deviation under the null and alternative; the only difference is where the distribution is centered.
- The red area is the traditional rejection region. If $\bar{x}$ falls within this region we reject H_0. If $\alpha = 0.05$ we will reject H_0 when it is true (a Type I error) 5% of the time.
- If the true mean is μ_A, the yellow area represents a region in which we would make a Type II error.

Probability of a Type II error



Murnane & Willett Figure 6.1

Probability of a Type II error



Murnane & Willett Figure 6.2

Statistical power

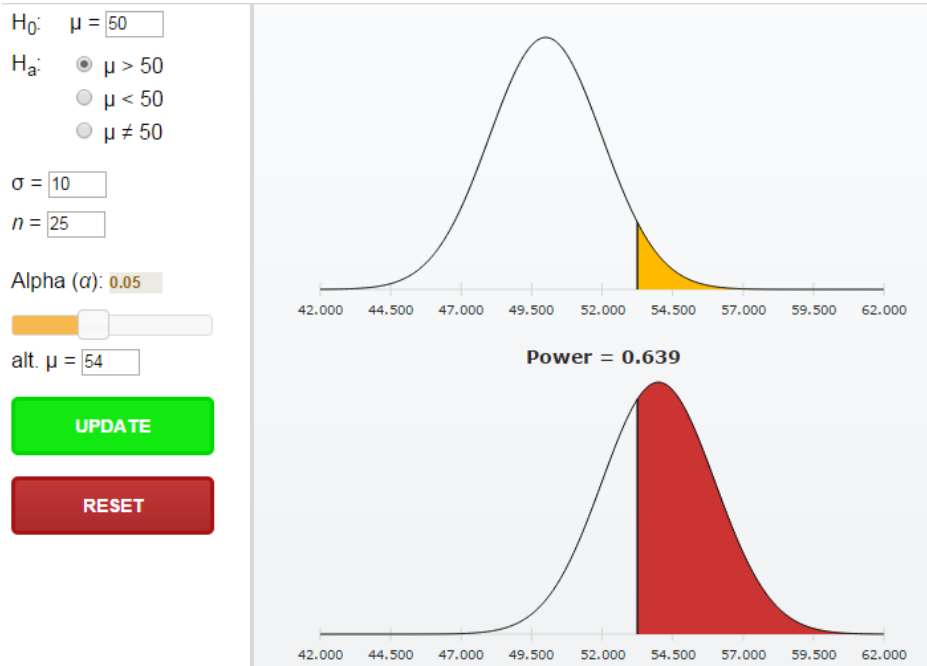
Statistical power

Statistical power is the probability of *correctly* rejecting H_0 when H_0 is false $(1 - \beta)$. Power represents our ability to detect a difference between the null hypothesis and a particular alternative hypothesis. This is the dark blue region in the previous figure.

Note: the following figures were taken from an online applet linked on the class website:
`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html`

The following applet is similar:
`https://istats.shinyapps.io/power/`

Power - 1



One-sided hypothesis test: $\mu_0 = 50$, $\sigma = 10$, $n = 25$, $\alpha = 0.05$. Find statistical power $(1 - \beta)$ when μ is actually 54.

Power - 1

If you were doing this manually, you would need to determine the value of $\bar{x}$ beyond which H_0 will be rejected (i.e., the yellow region above), find its z (or t) score in the *alternative* sampling distribution, and determine the probability of obtaining that score or something greater if H_a were true.

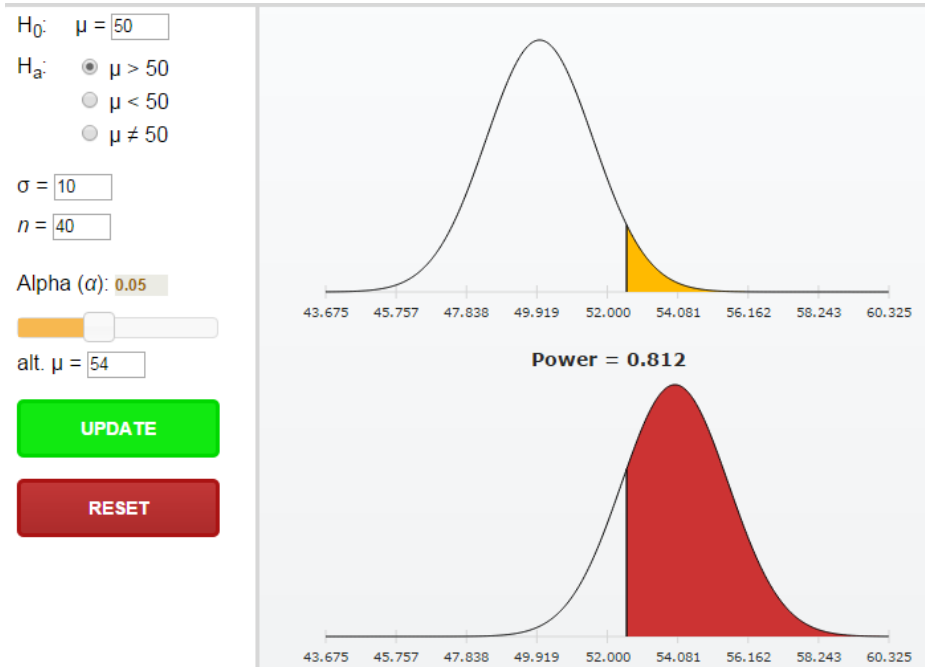
The $\bar{x}$ beyond which H_0 is rejected is: $50 + 1.645 * (10/\sqrt{25}) = 53.29$

The z-value in the *alternative* is: $(53.29 - 54)/(10/\sqrt{25}) = -0.355$

The probability of obtaining a $z > -0.355$ is **0.639**.

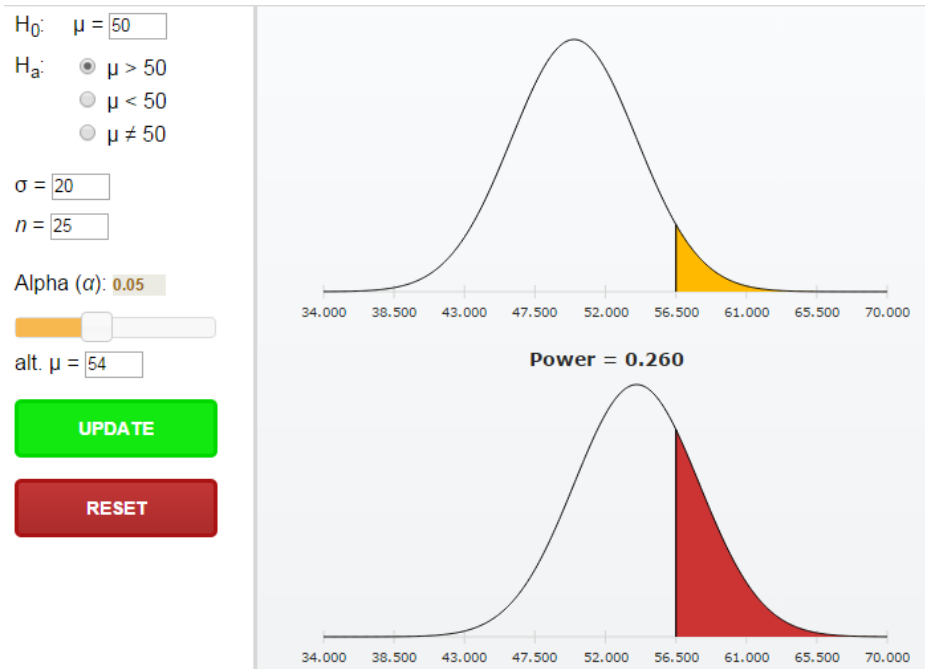
I have assumed normality for simplicity here. If σ were unknown this would affect the multiplier value used above (it would be 1.711 rather than 1.645), and a t distribution would be used in the probability calculation.

Power - 2



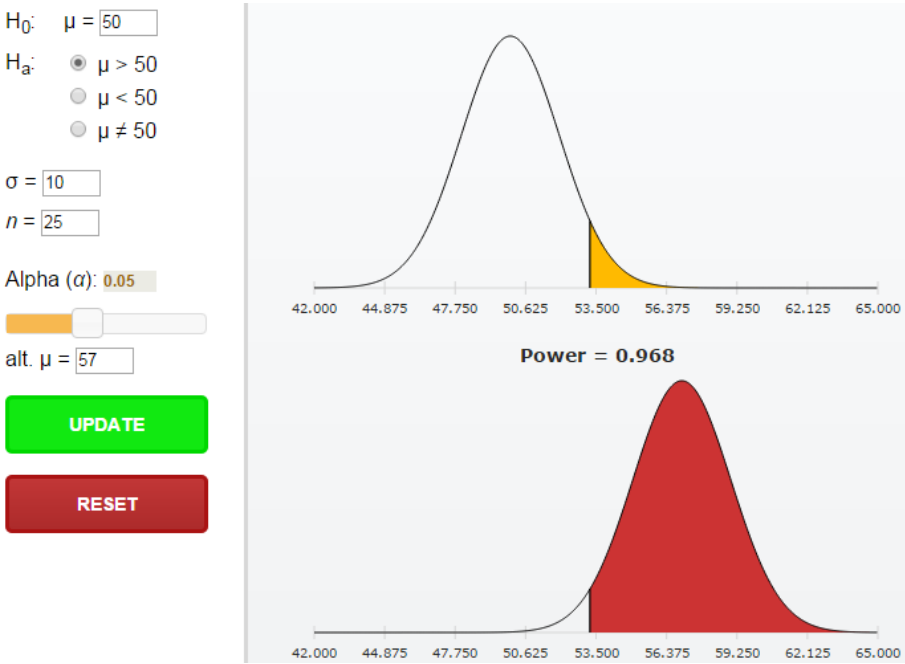
Consider what happens when n increases to 40.

Power - 3



Consider what happens when σ increases to 20 (keep $n = 25$).

Power - 4



Consider what happens when the alternative is further away (e.g. $\mu = 57$).

Statistical power

Things that affect statistical power (our ability to discern the null hypothesis from an alternative):

- **The effect size of interest** (δ): how far the alternative is away from the null. All else equal, the closer the alternative to the null, the lower the power.
- **Significance level**, which determines when we reject. All else equal, a higher α, the greater the power.
- **The standard error of the sample mean** ($\sigma/\sqrt{n}$ or $s/\sqrt{n}$). All else equal, the smaller the standard error, the greater the power. Because increasing *sample size* decreases the standard error, a larger n (holding σ constant) will increase power.
- **1- vs. 2-sided test**. 1-tailed tests have more power to detect an effect in one direction vs. a 2-tailed test with the same α.

Statistical power

Tools for calculating power for tests of μ:

- Online power applets like: `http://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html`
- Special software like G*Power and PowerUp!
- **Stata power command**: Statistics $\rightarrow$ Power and sample size $\rightarrow$ Means $\rightarrow$ One-sample $\rightarrow$ Test comparing one mean to a reference value. Select Compute: Power. Can calculate:
  - Power $(1 - \beta)$
  - Sample size requirements
- Stata can accept ranges of values (e.g., sample sizes, alternative hypotheses) and plot the results

Power calculation in Stata - 1

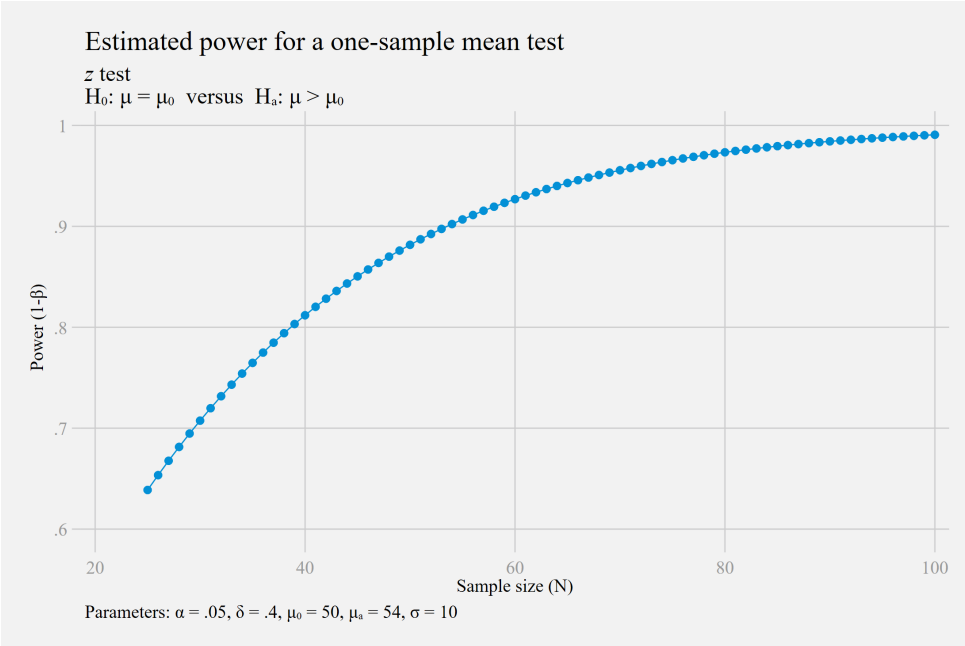
Using "Power - 1" example above. $\mu_0 = 50, \sigma = 10, n = 25, \alpha = 0.05$. Find statistical power when μ is actually 54.

```
. power onemean 50 54, n(25) sd(10) knownsd onesided

Estimated power for a one-sample mean test
z test
Ho: m = m0  versus  Ha: m > m0

Study parameters:

        alpha =    0.0500
            N =        25
        delta =    0.4000
           m0 =   50.0000
           ma =   54.0000
           sd =   10.0000

Estimated power:

        power =    0.6388
```

Note: *delta* (δ) is the standardized *effect size*: (54-50)/10. More on this later.

Power calculation in Stata - 1

Using "Power - 1" example. $\mu_0 = 50, \mu_a = 54, \sigma = 10, \alpha = 0.05$. Find power for sample sizes ranging from $n = 25$ to $n = 100$.



Power calculation in Stata - 1

What if the test were *two-sided* and $\mu_a = 54$? We will reject the null less often than if the test were one-sided (lower power).

```
. power onemean 50 54, n(25) sd(10) knownsd

Estimated power for a one-sample mean test
z test
Ho: m = m0  versus  Ha: m != m0

Study parameters:

        alpha =    0.0500
            N =        25
        delta =    0.4000
           m0 =   50.0000
           ma =   54.0000
           sd =   10.0000

Estimated power:

        power =    0.5160
```

Power calculation in Stata - 2

Your research group has developed an intervention designed to improve reading comprehension in 3rd grade. The typical (mean) gain on the 3rd grade reading test is 10 points, with a standard deviation of 6. Your intervention intends to improve on this. You randomly select n students to receive the intervention and calculate their mean gains ($\bar{x}$).

A standard significance test would be set up as:

$$H_0 : \mu = 10$$

$$H_1 : \mu > 10$$

The test statistic is: $t = (\bar{x} - 10)/(6/\sqrt{n})$, and you will reject if the probability of obtaining a t at least that large is < 0.05 (α).

Power calculation in Stata - 2

In some circumstances, your test will fail to reject H_0 even when it is false (a Type II error). If your intervention has a positive effect, you'd like your test to reject H_0!

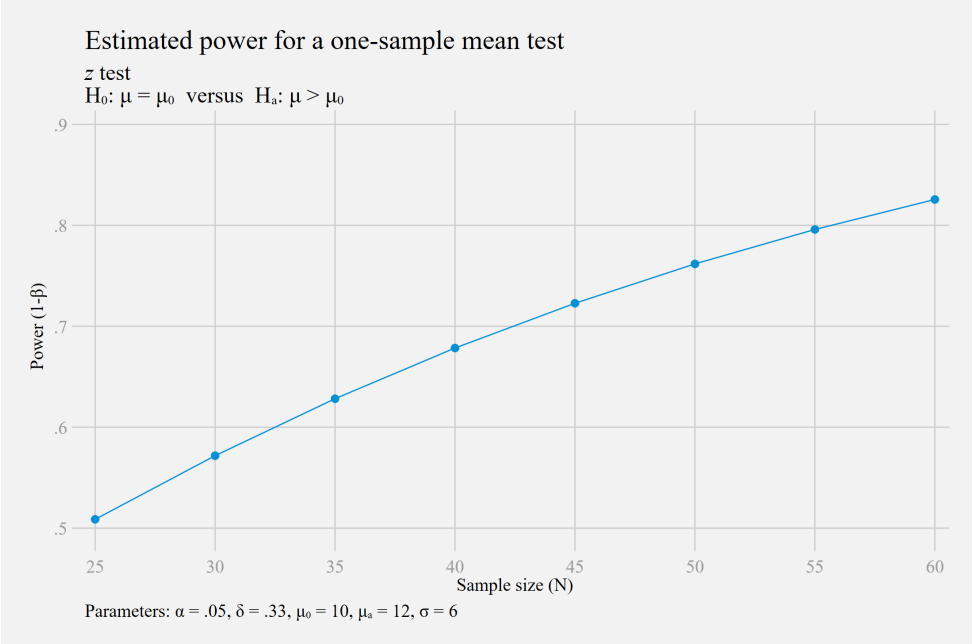
Your team believes the intervention will increase gains by 2 (from 10 to 12)—an effect size of $2/6 = 0.33$. What is the probability of a Type II error (and power) associated with various sample sizes (25-60)?

```
. power onemean 10 12, n(25(5)60) sd(6) knownsd onesided table(power beta N) graph

Estimated power for a one-sample mean test
z test
Ho: m = m0  versus  Ha: m > m0
```

power	beta	N
.5087	.4913	25
.5718	.4282	30
.6282	.3718	35
.6784	.3216	40
.7228	.2772	45
.7618	.2382	50
.7959	.2041	55
.8257	.1743	60

Power calculation in Stata - 2

Graphically:



Power - interpretation

Sticking with the previous example and $n = 45$:

- In 95% of random samples, this test will not reject H_0 if it is true (i.e., the mean gain in the study sample is 10, no impact of the intervention).
- In 5% of random samples, this test *will* reject H_0 when it is true—a Type I error. This is by design, since $\alpha = 0.05$.
- Suppose $\mu = 12$ in the study population: the study did have a $+2$ point effect. If $n = 45$, we fail to reject H_0 in 27.7% of random samples—a Type II error. We don't detect the effect in these cases.
- In **72.3%** of random samples, we properly reject H_0 (*power*, or $1 - \beta$)

What is a desirable power? A generally accepted value is **80%**.

Power - uses

The most common reasons for power analysis are:

- Determining the **minimum required sample size**. How large of a sample n do I need in order to detect a given effect size δ $(1-\beta)\%$ of the time?
- Determining the **minimum detectable effect size** (MDES). Given a sample size n, what is the smallest effect I will detect $(1-\beta)\%$ of the time?

Other factors affecting statistical power

- Type of statistical analysis: some types of analyses have more power than others
  - Techniques with stronger assumptions have more power (e.g., *parametric* vs. non-parametric tests)
- Controlling for covariates: reduces residual variance and therefore standard errors.
  - Light, Singer, & Willett (1990) note that if you include covariates that predict 50% of the variation in y together, you can get the same power with half the sample size.
- Non-random sampling designs (e.g., clustered)
- Reliability of the outcome measure: y is sometimes measured with error, which increases the standard error of test statistics

Reliability

Reliability in a measure can be thought of as the ratio of the variance in the "true" score divided by the variance in the observed score (σ_x/σ_{x^*}).

- Educational test score measures may have a reliability of 80-90%
- Survey items may have a lower reliability (e.g., 60%).
- Power analyses should account for reliability.

Murnane & Willett recommend multiplying your desired effect size by $\sqrt{reliability}$. For example if your desired effect is 0.2 and your reliability is 0.95, conduct a power analysis based on a minimum effect size of $0.2 * \sqrt{0.95} = 0.195$.

Practical significance and effect size

Practical significance

A *statistically significant* effect or difference is not necessarily a *practically important* one. In fact, with a large enough n, one can find statistically significant differences between the observed and hypothesized mean, even when the absolute difference between the two is quite small.

Practical significance is sometimes referred to as a "meaningfully large" effect, an "educationally significant" effect, "economically significant effect," or "clinically significant effect," depending on the context.

Effect size

An **effect size** is a measure of the degree to which the null hypothesis is false, in some meaningful unit (rather than p-values, say). One measure of effect size is the number of *standard deviations* in the original distribution the observed sample mean is from the hypothesized one. This measure is sometimes called **Cohen's d**:

$$d = \frac{\bar{x} - \mu_0}{s}$$

Note we are using s from the original scale of x. Put another way, we are comparing the effect to a measure of the overall variability in x.

Note: in the Stata `power` output this is called *delta* (δ).

Practical vs. statistical significance

	A	B	C	D
Sample size	10,000	10,000	9	1,000
Mean test score under H_0	200	200	200	200
Sample mean ($\bar{x}$)	225	201	225	201
Sample std deviation (s)	25	25	100	25
Δ = Difference from H_0	**25**	**1**	**25**	**1**
Standard error ($s/\sqrt{n}$)	0.25	0.25	33.3	0.79
t-statistic (Δ/se)	100	4	0.75	1.26
p-value	$p<0.0001$	$p<0.001$	$p>0.40$	$p>0.20$
Statistically significant?	Yes	Yes	No	No
Confidence interval for μ	$225 \pm 1.96 * 0.25$ (224.51, 225.49)	$201 \pm 1.96 * 0.25$ (200.51, 201.49)	$225 \pm 2.31 * 33.3$ (148.1, 301.9)	$201 \pm 1.96 * 0.79$ (199.5, 202.5)
Effect size δ (Δ/s)	1	0.04	0.25	0.04
Practically significant?	Yes	No	Yes (if true)	No

Source: based on Remler & Van Ryzin ch. 8. Effect size δ is Cohen's d using μ under H_0 as a benchmark ($H_0 : \mu = 200$)

Practical vs. statistical significance

In column D, the difference from H_0 is neither practically nor statistically significant. But the results still provide valuable information. Note the 95% confidence interval is (199.5, 202.5). If we don't consider the *bounds* of this interval to be meaningful differences from H_0, then we can rule out practically meaningful effects. (In column C we can't rule out practically meaningful effects).

When the confidence interval (for a difference) includes zero and rules out meaningful effects, it is sometimes called a **precise zero**.

How to report results: advice

Good studies provide information about both:

- **Statistical significance**: tells us that the point estimate is statistically different from H_0 (which is often zero).
- **Practical significance**: assesses whether the point estimate is meaningful in size, given the context.

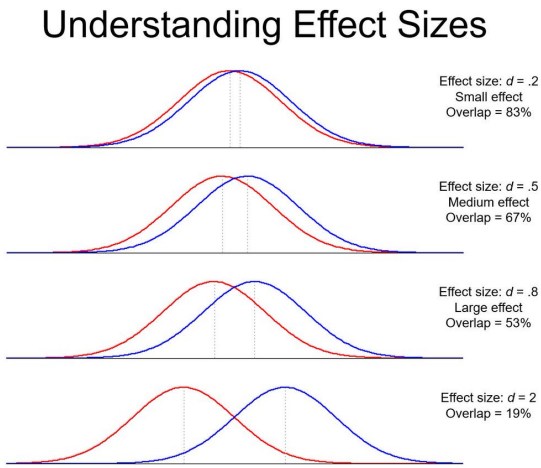
Even better studies include information about:

- The range of population parameters that cannot be rejected (i.e., a confidence interval).
- The *strength* of the evidence against the null (i.e., p-values)

Benchmarking effect sizes

How do we know if an effect size is practically meaningful? Cohen (1969, 1988) proposed the following guidelines for interpreting d:

- 0.2 = small effect
- 0.5 = medium effect
- 0.8 = large effect



Benchmarking effect sizes

These benchmarks do not work well in all contexts, however, and the evidence suggests they are much too large for educational interventions (Kraft, 2020; Hill et al., 2008).

For benchmarks in international education and development, see Evans (2022).

Benchmarking effect sizes

A better approach to interpreting effect size is to look to **empirical benchmarks**—that is, looking to *existing evidence* to tell us whether an effect is meaningful or not. Approaches in Hill et al. (2008):

- Normative expectations for growth in student achievement: "typical yearly growth" for a given age and subject.
- Policy-relevant gaps in student achievement by demographic group or school performance (e.g., commonly-observed gaps by income, race, gender)
- Effect sizes from past research for similar interventions and similar target populations.

Note: annual growth tends to be greater for younger kids, and in math

Benchmarking effect sizes

Based on a review of 747 randomized controlled trials in education, Kraft (2020) proposes the following benchmarks:

- < 0.05 = small effect size
- 0.05 - 0.2 = medium effect size
- > 0.20 = large effect size

In determining practical significance in your context, ask "how large is the effect relative to other studies with broadly comparable features?"

Benchmarking effect sizes

Typical effect sizes vary by test subject (math or reading), scope of test, and sample size.

Table 1
Empirical Distributions of Effect Sizes From Randomized Control Trials of Education Interventions With Standardized Achievement Outcomes

		Subject		Sample Size					Scope of Test		DoE
	Overall	Math	Reading	≤100	101–250	251–500	501–2,000	>2,000	Broad	Narrow	Studies
Mean	0.16	0.11	0.17	0.30	0.16	0.16	0.10	0.05	0.14	0.25	0.03
Standard deviation	0.28	0.22	0.29	0.41	0.29	0.22	0.15	0.11	0.24	0.44	0.16
Mean (weighted)	0.04	0.03	0.05	0.29	0.15	0.16	0.10	0.02	0.04	0.08	0.02
P1	–0.38	–0.34	–0.38	–0.56	–0.42	–0.29	–0.23	–0.22	–0.38	–0.78	–0.38
P10	–0.08	–0.08	–0.08	–0.10	–0.14	–0.07	–0.05	–0.06	–0.08	–0.12	–0.14
P20	–0.01	–0.03	–0.01	0.02	–0.04	0.00	–0.01	–0.03	–0.03	0.00	–0.07
P30	0.02	0.01	0.03	0.10	0.02	0.06	0.03	0.00	0.02	0.05	–0.04
P40	0.06	0.04	0.08	0.16	0.07	0.10	0.06	0.01	0.06	0.11	–0.01
P50	0.10	0.07	0.12	0.24	0.12	0.15	0.09	0.03	0.10	0.17	0.03
P60	0.15	0.11	0.17	0.32	0.17	0.18	0.12	0.05	0.14	0.22	0.05
P70	0.21	0.16	0.23	0.43	0.25	0.22	0.15	0.08	0.20	0.34	0.09
P80	0.30	0.22	0.33	0.55	0.35	0.29	0.19	0.11	0.29	0.47	0.14
P90	0.47	0.37	0.50	0.77	0.49	0.40	0.27	0.17	0.43	0.70	0.23
P99	1.08	0.91	1.14	1.58	0.93	0.91	0.61	0.48	0.93	2.12	0.50
k (number of effect sizes)	1,942	588	1,260	408	452	328	395	327	1,352	243	139
n (number of studies)	747	314	495	202	169	173	181	124	527	91	49

Note. A majority of the standardized achievement outcomes (95%) are based on math and English language art test scores, with the remaining based on science, social studies, or general achievement. Weights are based on sample size for weighted mean estimates. For details about data sources, see Appendix A, available on the journal website. DoE = U.S. Department of Education.

Benchmarking effect sizes

How should we think about effect sizes? (advice from Kraft, 2020)

- Effect sizes can be *descriptive* (correlational) or *causal*. Descriptive "effect sizes" are often much larger than causal ones.
- Effects on short-run outcomes are often larger than effects on long-run outcomes.
- Effects on specialized and researcher-designed instruments are often larger than those on broader instruments.
- Effect sizes are smaller when more measurement error is expected.

Benchmarking effect sizes

How should we think about effect sizes? (advice from Kraft, cont.)

- Studies with targeted samples tend to have bigger effects than those with more inclusive samples.
- Effect sizes for an intervention tend to be larger when there is a greater treatment-control contrast.
- Treatment effects are larger if they are based on actual treatment, rather than a treatment *offer*.
- Cost matters: effects from lower-cost interventions are arguably more impressive than effects from higher-cost interventions.
- Effects of interventions are generally smaller when they are taken to scale.

Benchmarking effect sizes

Example of a new tool that provides benchmarks for effect sizes (on non-academic outcomes like antisocial behavior and self-regulation):

`https://ebcontextualizer.shinyapps.io/EmpBench/`

Source: Wilson, Freeman, and Hedberg (forthcoming)

Next time

Describing relationships: correlation and regression

- Read: Huntington-Klein chapter 4, Stock & Watson chapter 4