7. Statistical power and effect size

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

Last time

- Statistical hypothesis testing: tests about μ and π
- Null vs. alternative hypotheses; one-tailed vs. two-tailed tests
- Test statistics and *p*-values
- Significance levels (α)
- Using a $(1-\alpha)$% confidence interval to test a hypothesis with an α significance level
- Using `mean` or `ttest` in Stata for hypotheses about μ
- Using `proportion` or `prtest` in Stata for hypotheses about π
- Stata's `ttesti` and `prtesti` calculators

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 = α)	*Correct decision* (Pr = 1-α)
H_0 is false	Correct decision (Pr = 1-β)	Incorrect decision— ***Type II error*** (Pr = β)

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 man
- **Type II error**: not rejecting H_0 and letting a guilty man go free

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

Type I and Type II errors

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

- H_0: no prostate 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

The probability of committing a Type II error is a bit more difficult to calculate than the probability of committing a Type I error (which is set by the researcher as α). This is because this probability depends on how far away the plausible alternative to μ_0 is.

- 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

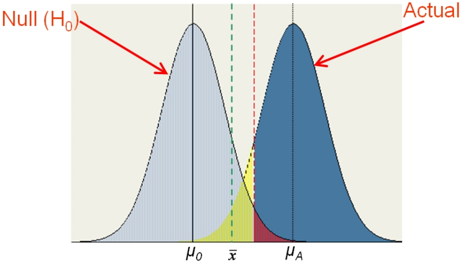


Figure: Distribution of $\bar{x}$ under H_0 and a specific alternative H_A

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.

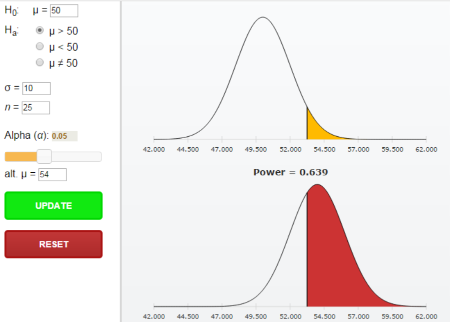
Power of the test

The probability of *correctly* rejecting H_0 when H_0 is false $(1 - \beta)$ is called the **power of the test** (or just **power**). 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 slide.

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`

Also try the following applet (which looks even nicer):
`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.

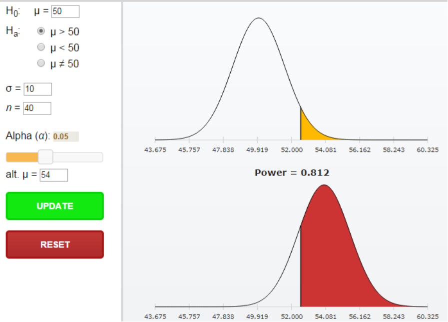
The $\bar{x}$ beyond which H_0 is rejected is: $50 + 1.645 * (10/\sqrt{25}) = 53.29$

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**. `1-normal(-0.355)`

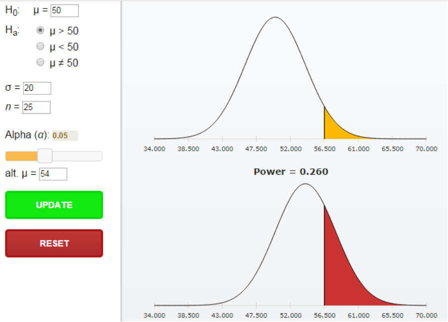
I have assumed normality for simplicity here. If σ were unknown this would affect the multiplier value used above (it would be 1.677 rather than 1.645), and a t distribution would be used in the probability calculation.

Power - 2



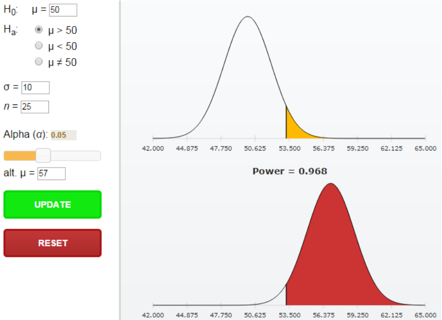
Consider what happens when n increases to 40.

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. $\mu = 57$).

Power of the test

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

- **The effect size of interest**: how far the alternative is away from the null. All else equal, the <u>closer</u> the alternative to the null, the <u>lower</u> the power.
- **Significance level**, which determines when we reject. All else equal, a <u>higher</u> α, the <u>greater</u> the power of the test.
- **The standard error of the sample mean** ($\sigma/\sqrt{n}$). All else equal, the <u>smaller</u> the standard error, the <u>greater</u> the power of the test. Because increasing *sample size* decreases the standard error, a larger n (holding σ constant) will increase the power of the test.
- **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.

Power of the test

Tools for calculating power for tests of μ:

- Power applets like: `http://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html`
- **Stata power command**: Statistics → Power and sample size → Means → One-sample → 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

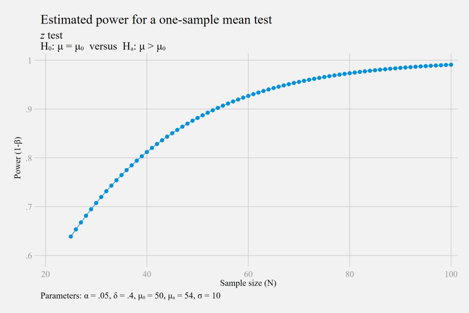
Using "Power - 1" example above. $\mu_0 = 50, \sigma = 10, n = 25, \alpha = 0.05$. Find statistical power when μ is actually 54.

```
. 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 *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!

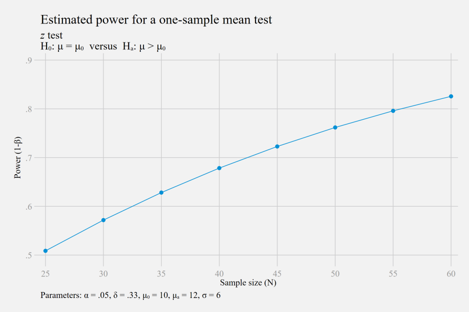
Your team believes the intervention will increase gains by 2 (from 10 to 12)—an effect size of $2/6 = 0.33$. What is the probability of a Type II error (and power) associated with various sample sizes (25-60)?

```
. 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 different than the general population).
- 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 <u>did</u> 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 calculation in Stata - 3

For power calculations for *proportions* use `power oneproportion`. For large samples use the normal approximation for the sampling distribution (the default). For small samples can use optional `binomial` test.

```
. power oneproportion 0.50 0.55,  n(600)

Estimated power for a one-sample proportion test
Score z test
Ho: p = p0  versus  Ha: p != p0

Study parameters:

        alpha =    0.0500
            N =       600
        delta =    0.0500
           p0 =    0.5000
           pa =    0.5500

Estimated power:

        power =    0.6886
```


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 $(\mu_a - \mu_0)$ $(1-\beta)$% 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?

Power analyses can get much more complicated than the one shown here for estimating μ from a single sample. In cases where there aren't ready calculations, simulation may help.

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. For example:

- In our test of mean body temperature $H_0 : \mu = 98.6$, a large sample size could result in $\bar{x} = 98.59$ leading to the rejection of H_0. In other words, the difference could be statistically significant.
- But whether this difference is **practically significant** is a different matter, and depends on the context.

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 in probabilistic terms, i.e., p-values). 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 <u>original scale</u> of x (<u>not</u> standard errors, which always decrease as n gets larger).

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

Effect size

In the body temperature example, suppose we can reject $H_0 : \mu = 98.6$ with $\bar{x} = 98.59$ (and assume $s = 0.60$). The effect size is:

$$d = \frac{\bar{x} - \mu_0}{s} = \frac{98.59 - 98.6}{0.60} = -0.017$$

The standard deviation of body temperature in the sample is 0.60. Our observed sample mean temperature is only 0.017 standard deviations below our hypothesized mean. Even if we can reject H_0, this small difference in temperature may not be practically significant.

Stata esize

Stata offers an `esize` command for calculating Cohen's d and other commonly-used effect sizes (e.g., Hedges's g). It is intended for use when comparing two independent samples (see Lecture 8).

```
. esize twosample achmat08, by(gender)

Effect size based on mean comparison

                               Obs per group:
                                    male =        227
                                  female =        273

      Effect Size   Estimate     [95% Conf. Interval]

        Cohen's d   .2678647     .0908964     .4445664
       Hedges's g   .2674611     .0907595     .4438964
```

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	$225 \pm 1.96 * 0.25$	$201 \pm 1.96 * 0.25$	$225 \pm 2.31 * 33.3$	$201 \pm 1.96 * 0.79$
	(224.51, 225.49)	(200.51, 201.49)	(148.1, 301.9)	(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.

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 of (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**. More on this in Lecture 8.

How to report results: advice

Most papers emphasize two characteristics: statistical significance and practical significance. (Bad ones focus only on statistical significance).

- 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.

Some limitations to this approach:

1. Zero and the point estimate itself are not the only values of interest. Sometimes zero is not an interesting null hypothesis.
2. No information is provided about the *strength* of the evidence against the null.

How to report results: advice

Empirical papers in leading economics journals rarely discuss confidence intervals or the size of standard errors (Romer, 2020).

Table 1—Information Reported in the Text of Empirical Papers in Three Leading Journals in 2019

Discussed prominently	
Confidence intervals	14%
	(3)
Standard errors but not confidence intervals	10%
	(3)
Mentioned in passing	
Confidence intervals	6%
	(2)
Standard errors but not confidence intervals	7%
	(2)
Neither confidence intervals nor standard errors discussed	64%
	(5)

Notes: Standard errors are in parentheses. See text for details.

The upper end of a 2 SE confidence interval for papers that discuss confidence intervals is 20%. ;)

How to report results: advice

Romer (2020): "the tone is often that once it is known that estimates are statistically different from zero, the only aspect of the results that matters is the point estimates—almost as though when an estimate is significantly different from zero, it can be treated as exact."

Romer recommends reporting and discussing confidence intervals.

"Knowing significance is not enough to know what values of the parameter other than zero the data provide strong evidence against, and what values they provide little reason to object to."

Example

Consider two papers estimating the rate of return to an additional year of education (i.e., the % increase in annual earnings) . Both papers estimate $\bar{x} = 9.0$.

- Paper 1 has a standard error of 3.9
- Paper 2 has a standard error of 1.8

Both papers would claim statistical and practical significance. (For most people, the benefit to additional education would outweigh the costs).

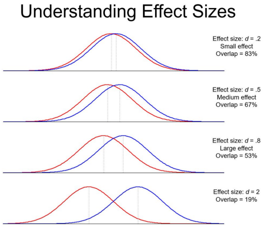
- Paper 1 has a 95% CI of (1.4, 16.6)
- Paper 2 has a 95% CI of (5.5, 12,5)

Paper 1 cannot rule out effects that may be considered practically small. Need to see the standard error and/or confidence interval!

Benchmarking effect sizes

How do we know if an effect size is practically meaningful? Cohen (1969) 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).

In public health: Rutledge & Loh (2004) on effects of health behaviors (smoking, obesity, etc.)

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"
- 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.

Benchmarking effect sizes: typical student growth

Annual growth tends to be greater for younger kids, and varies by subject.

Table 1
Average Annual Gain in Effect Size From Nationally Normed Tests

	Reading tests		Math tests	
Grade transition	Mean	Margin of error	Mean	Margin of error
Grade K–1	1.52	±0.21	1.14	±0.49
Grade 1–2	0.97	±0.10	1.03	±0.14
Grade 2–3	0.60	±0.10	0.89	±0.16
Grade 3–4	0.36	±0.12	0.52	±0.14
Grade 4–5	0.40	±0.06	0.56	±0.11
Grade 5–6	0.32	±0.11	0.41	±0.08
Grade 6–7	0.23	±0.11	0.30	±0.06
Grade 7–8	0.26	±0.03	0.32	±0.05
Grade 8–9	0.24	±0.10	0.22	±0.10
Grade 9–10	0.19	±0.08	0.25	±0.07
Grade 10–11	0.19	±0.17	0.14	±0.16
Grade 11–12	0.06	±0.11	0.01	±0.14

Sources. Annual gain for reading is calculated from seven nationally normed tests: California Achievement Test (CAT)–5th edition, Stanford Achievement Test (SAT)–9th edition, TerraNova–Comprehensive Test of Basic Skills (CTBS), Metropolitan Achievement Test (MAT8), TerraNova–CAT, SAT10, and Gates-MacGinitie. Annual gain for math is calculated from six nationally normed tests: CAT5, SAT9, TerraNova–CTBS, MAT8, Terra Nova–CAT, and SAT10. For further details, contact the authors (Bloom et al. 2006a, 2006b, 2007a, 2007b, in press; Lipsey et al., 2007).

Source: Hill et al. (2008). Note these tests are designed for comparisons across grades. Mean scores in adjacent grades are used for growth; these are converted to effect sizes used pooled std deviations across the two grades.

Benchmarking effect sizes: typical achievement gaps

2002 and 2000 NAEP achievement gaps in reading and math:

Table 2
Demographic Performance Gap in Mean NAEP Scores, by Grade (in Effect Size)

Subject and grade	Black–White	Hispanic–White	Eligible–ineligible for free/reduced-price lunch	Male–Female
Reading				
Grade 4	−0.83	−0.77	−0.74	−0.18
Grade 8	−0.80	−0.76	−0.66	−0.28
Grade 12	−0.67	−0.53	−0.45	−0.44
Math				
Grade 4	−0.99	−0.85	−0.85	0.08
Grade 8	−1.04	−0.82	−0.80	0.04
Grade 12	−0.94	−0.68	−0.72	0.09

Sources. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Reading Assessment, and 2000 Mathematics Assessment (Bloom et al., 2007a, 2007b, in press; Lipsey et al., 2007).

Source: Hill et al. (2008)

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.

More on p-values

- Publication bias / p-Screening
- p-Hacking
- Over-comparing and under-reporting

The lucky coin flipper

Suppose a subject flips a coin 15 times and gets heads 13 of those times. What is the probability this result (or something more extreme) could have occurred by chance?

$$\underbrace{0.5^{13}(1-0.5)^2\frac{15!}{13!(15-13)!}}_{0.0032}+\underbrace{0.5^{14}(1-0.5)^1\frac{15!}{14!(15-14)!}}_{0.00046}+$$

$$\underbrace{0.5^{15}(1-0.5)^0\frac{15!}{15!(15-15)!}}_{0.00003}=0.003693$$

or in Stata: `bitesti 15 13 0.5` yields $Pr(k \geq 13) = 0.003693$. Very unlikely! If this were a hypothesis test, we would reject $H_0 : \pi = 0.5$.

The lucky coin flipper

Now if 100 subjects *independently* flip a coin 15 times, what is the probability that at least 1 of them gets heads 13 or more times?

The probability that <u>none</u> of them meets this benchmark is:

$$(1-0.003693)^{100} = 0.691$$

Thus the probability that <u>at least one</u> of them gets 13 or more heads is:

$$1-0.691 = 0.309$$

The lucky coin flipper

That is, an almost 1 in 3 chance that <u>someone</u> will perform this well!
(Note: `bitesti 100 1 0.003693` confirms $Pr(k \geq 1) = 0.309$)

"If someone flips a coin and gets the same result 9 or 10 times, it is not remarkable in itself, but it will seem remarkable to the person flipping the coin." (From Bueno de Mequita & Fowler, 2022, ch. 7)

The *Journal of Lucky Coin Flippers*

In the real world, we are more likely to hear about the significant cases. "Neither the public nor the scientific community gets to see all of the hypothesis tests that were (or could have been) conducted."

- **Publication bias**: journals tend to favor statistically significant ($p < \alpha$), surprising, and noteworthy results.
- **"File drawer effect"**: studies lacking significant findings are more likely to be shelved by the authors.

These behaviors might collectively be called ***p*-Screening**.

The *Journal of Lucky Coin Flippers*

In the coin flip example, if the 13 heads out of 15 outcome was unusual, it would be unlikely to re-occur in subsequent trials:

- `bitesti 100 1 0.003693` yields $Pr(k \geq 1) = 0.309$
- `bitesti 100 2 0.003693` yields $Pr(k \geq 1) = 0.053$
- `bitesti 100 3 0.003693` yields $Pr(k \geq 1) = 0.006$
- Etc.

The probability that 2 or 3 "studies" obtain the same unusual result—if H_0 is true—is very low. We need to see both the statistically significant and insignificant findings!

Publication bias can inflate perceptions of effect size

- Suppose there are many well-designed, unbiased studies of an effect. The only difference between them is sampling variation.
- Suppose journals only publish *statistically significant* findings (those significantly different from zero).
- To pass the publication test, the estimate must be large relative to its standard error. We only observe estimates that are sufficiently large.
- Ex: suppose the true effect $= 1$ and the studies all have a confidence interval width of ± 2. The studies will only reject H_0 when the estimated effect is 2 or greater.

Over time, the published estimates will tend to *overstate* the true effect.

p-Hacking

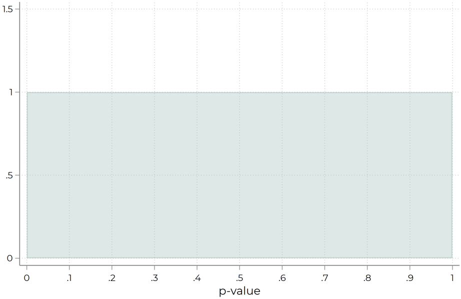
The above example assumed researchers were acting in good faith, and that results only varied due to sampling variation. In practice, analysts may play around with their data or tests until p is below some threshold:

- Lots of subgroup analyses
- Specification searches (e.g., regression analyses)
- Lots of different outcomes

This is called ***p*-Hacking**, and clearly exacerbates the problem.

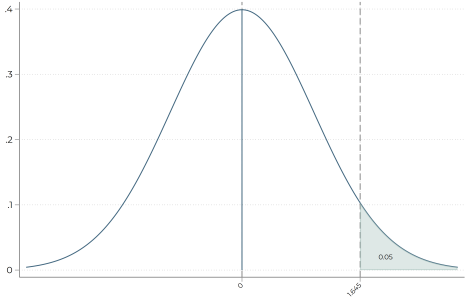
Diagnosing p-Hacking

If H_0 is true and there is no real "effect" (e.g., difference from zero), p-values should be equally likely across studies. That is, they should have a *uniform distribution*:



Diagnosing p-Hacking

This may not feel intuitive, since the distribution of our test statistic (under CLT) follows a *normal distribution*, where values closer to 0 are more likely than values away from 0:



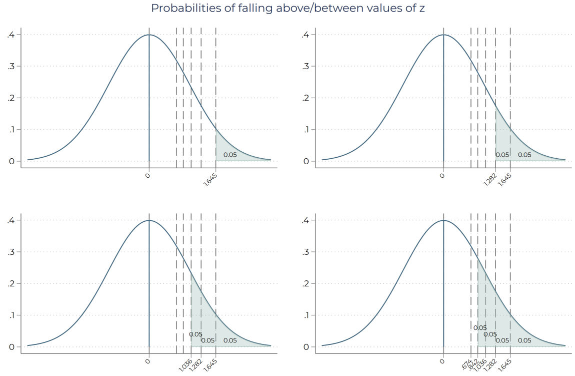
Diagnosing p-Hacking

In fact, the distribution of p-values under H_0 follows directly from how we define p-values (as probabilities). If H_0 is true:

- In 5% of samples, one will obtain $p < 0.05$
- In 10% of samples, one will obtain $p < 0.10$
- In 20% of samples, one will obtain $p < 0.20$
- ...and so on

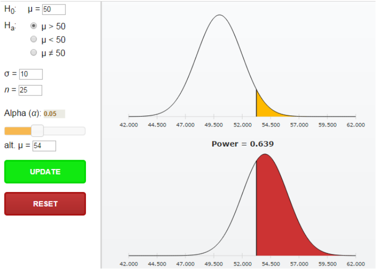
In other words, an even distribution of p-values.

Diagnosing *p*-Hacking



Diagnosing *p*-Hacking

On the other hand, if H_a is true and there is an "effect" (e.g., difference from zero), *p*-values are more likely to be small than large. Example 1 from earlier:



Diagnosing p-Hacking

If researchers are p-Hacking, they may tinker with their analysis until $p < 0.05$. This suggests bunching of p-values just under 0.05.

Diagnosing p-Hacking

Patterns of p-values in three cases (assume journals only publish when $p < 0.05$):

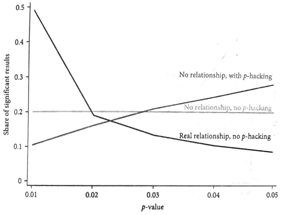


Figure 7.2. p-hacking distorts the distribution of p-values in a literature.

Source: Bueno de Mesquita & Fowler (2022).

Diagnosing p-Hacking

z-values that appear in more than 1 million Medline articles: note the under-representation between -2 and +2.

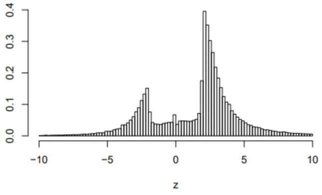


Figure 1: The distribution of more than one million z-values from Medline (1976–2019).

Source: van Zwet & Cator (2021)
`https://onlinelibrary.wiley.com/doi/full/10.1111/stan.12241`

Potential solutions for preventing p-Hacking

1. Reduce significance threshold (e.g., $p < 0.005$). There are pros and cons to this.
2. Adjust p-values for multiple testing (like coin flip example earlier)–see next section. Complicated for non-independent multiple tests.
3. Don't obsess over statistical significance. Use p-values, confidence intervals, practical significance.
4. **Pre-registration**: pre-commit to tests before seeing the data.
5. **Replication**: independent replications of the same study.
6. Focus on important and plausible hypotheses. Is the answer to the question interesting, regardless of what it turns out to be? (Example of the "power pose").

Pre-registration

Examples:

- Open Science Framework: `https://osf.io/`
- Registry of Efficacy and Effectiveness Studies: `https://sreereg.icpsr.umich.edu/sreereg/`
- AEA RCT Registry: `https://www.socialscienceregistry.org/`

See also the excellent book by Glennerster and Takavarsha (2013), *Running Randomized Evaluations*, for an extended discussion of pre-registration in the context of field experiments.

Multiple comparisons problem

A related issue comes up in studies that test **multiple hypotheses** with the same data. This can arise whenever:

- There are many groups being compared against each other. (With g groups, there are $g(g-1)/2$ potential pairwise comparisons).
- You are examining relationships between a large set of variables (e.g., a correlation matrix).
- There are many outcomes of interest. E.g., a school-based reform could affect academic outcomes (in multiple subjects), behavioral outcomes, attendance, engagement, etc.

Key idea: when conducting multiple tests, the probability of making a Type I error on <u>any</u> (1 or more) test is much higher than the probability of a Type I error on any single test.

Family-wise error rates

Suppose α is the significance level used in a single test. You plan to conduct K hypothesis tests. (Sometimes tests are grouped into **families** of related tests).

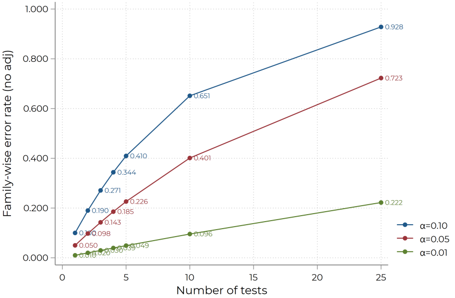
The **family-wise error rate** (α_{FW}) is the probability of making at least one Type I error among your multiple tests. This probability is:

$$\alpha_{FW} = 1 - \left[(1-\alpha)^K\right]$$

aka the **multiple comparison error rate**.

Family-wise error rates

FWERs for $\alpha = 0.10$, $\alpha = 0.05$ and $\alpha = 0.01$:



Family-wise error rates

Easy calculation in Stata with `bitesti`. Suppose you plan 25 tests with $\alpha = 0.05$.

```
. bitesti 25 1 0.05

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
       25          1         1.25       0.05000      0.04000

  Pr(k >= 1)           = 0.722610  (one-sided test)
  Pr(k <= 1)           = 0.642376  (one-sided test)
  Pr(k <= 1 or k >= 2) = 1.000000  (two-sided test)
```

The FWER is 0.723. On average, out of 25 tests, you expect 1.25 to result in rejection even H_0 is true.

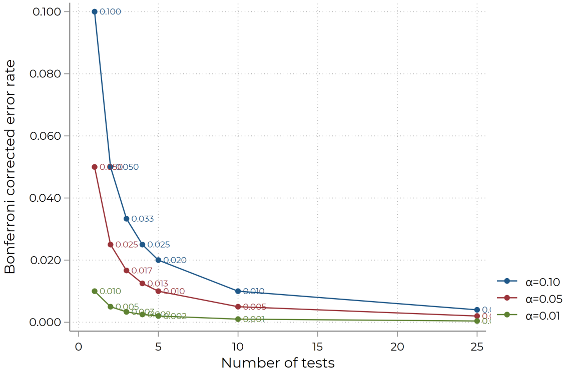
Bonferroni adjustment

The goal when conducting multiple tests is to reduce the FWER to some desired level (e.g., 0.05). Not doing so could be seen as *p*-Hacking.

The **Bonferroni correction** simply divides the FWER by the number of tests to obtain an adjusted significance level for each test: α_{FW}/K.

- For example, if conducting $K = 3$ tests and you want $\alpha_{FW} = 0.05$, use a significance of $0.05/3 = 0.0167$ on each test.
- Note: since the new significance level is α_{FW}/K, you sometimes see the Bonferroni correction described as multiplying your *p*-values by K and then comparing to α_{FW}. Same idea.
- The correction changes your threshold *p*-value for rejection. It also affects your confidence interval width (use the new $\alpha_{FW}/2$).

Bonferroni corrected significance levels



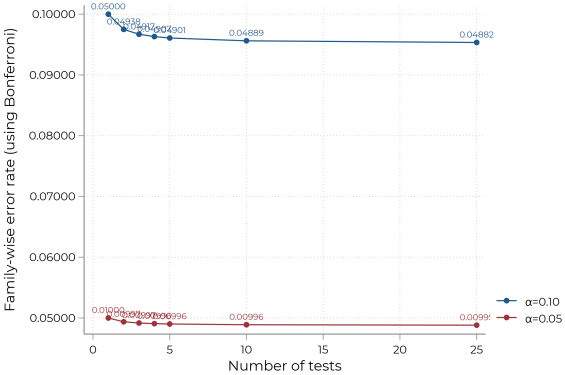
Bonferroni adjustment

The Bonferroni adjustment reduces the number of false positive, but it is a conservative method:

- It is an approximation (see graph next slide).
- It sets a high threshold for rejection (low α_{FW}) with the aim of preventing a false positive on *any one* test. The threshold may become unreasonably high when there is a large number of tests.
- Assumes tests are independent of one another. When the same data are use to conduct multiple hypothesis test, there is likely correlation across tests.

The Bonferroni adjustment *reduces power*. Because you are less likely to reject H_0 on any one test, you are more likely to make Type II errors: failing to reject H_0 when it is false.

FWER using Bonferroni corrected significance levels



Bonferroni adjustment

The Bonferroni adjustment is generally only recommended if you a have a relatively small number of hypothesis tests (comparisons to make), and if the results are not strongly correlated across tests.

There are other methods of adjustment for multiple comparisons available (e.g., Tukey, step-down resampling method)