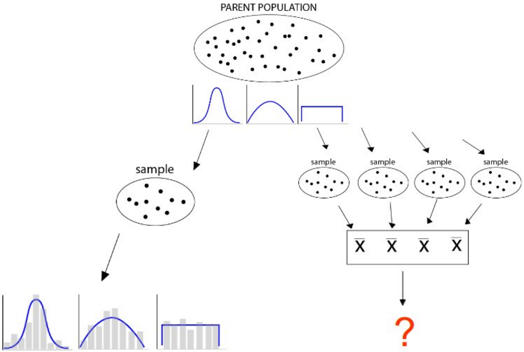
Likelihood of Sample Mean

Central Limit Theorem

Regardless of the shape of the parent population, the sampling distribution of the mean will be *normally distributed* with a mean of μ and a standard deviation (standard error) of $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$.



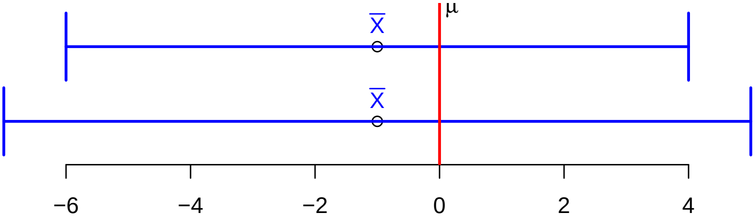
Stating the probability of a sample statistic when σ is known

You take a sample of size N from the parent population (μ, σ are known) and obtain an mean of $\bar{X}$. What is the probability of obtaining a sample mean at least that extreme?

1. Calculate the standard error $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$
2. Calculate a z-score $z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$
3. Look up probability from SND table
4. Multiply probability by two ("two-tailed" probability)

Confidence Intervals (CIs)

- Used when the population mean μ is unknown
- Colloquially referred to as "margin of error"
- Specify a range of values that "captures" the population mean with some error rate (5%, 1%)
  - "The population mean is between LL and UL, with an error rate of X%."
- CIs are "centered" at the sample mean $\bar{X}$
  - because that is the best guess at the population mean



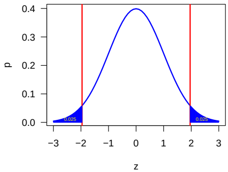
Calculating confidence intervals σ known, μ unknown

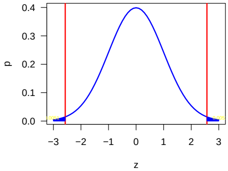
(NB: This is a highly unusual case!)

1. Calculate the standard error of the mean $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$.
2. Find the **critical value** of z such that $P(z_{obs} \geq z_{crit}) = \frac{1-pCI}{2}$.
   - For 95% CI, use $z_{crit} = 1.96$
   - For 99% CI, use $z_{crit} = 2.57$
3. Calculate:

$$LL = \bar{X} - z_{crit}\sigma_{\bar{X}}$$

$$UL = \bar{X} + z_{crit}\sigma_{\bar{X}}$$





Calculating confidence intervals
σ unknown, μ unknown

(NB: This is the typical case!)

1. Estimate the standard error of the mean $\hat{s}_{\bar{X}} = \frac{\hat{s}}{\sqrt{N}}$.
2. Determine the degrees of freedom (df). For these problems, $df = N - 1$.
3. Find the **critical value** t_{crit} such that $P(t_{obs} \geq t_{crit}) = \frac{1-pCI}{2}$.
   - You will need to look this up in a table based on df and desired CI (95%, 99%).
4. Calculate:

$$LL = \bar{X} - t_{crit} s_{\bar{X}}$$
$$UL = \bar{X} + t_{crit} s_{\bar{X}}$$

Logic of Hypothesis Testing

- Logic of “null hypothesis significance testing” (NHST)
- The one-sample t-test
- Effect size for one-sample designs

Logic of NHST

1. State a *null hypothesis* about a population parameter that you wish to *disprove*, and a mutually exclusive *alternative hypothesis*.
   - $H_0 : \mu = 500$
   - $H_1 : \mu \neq 500$
2. Obtain a sample, and identify the appropriate test statistic
3. Compare the observed test statsitic to a sampling distribution to obtain the probability of your sample mean $\bar{X}$ *assuming H_0 is true*
4. If the probability is "sufficiently" small, REJECT H_0, otherwise RETAIN H_0

A Legal Analogy

In a criminal court...

- Presumption of innocence
  - H_0: Defendent is innocent
  - H_1: Defendent is guilty
- Evidence is presented to jury
- Jury decides to "reject" or "retain" the presumption of innocence
- Evidence for rejecting the assumption must be "beyond the shadow of a doubt"

Statistical Significance

We say that an observed sample mean (or difference in sample means) is "statistically significant" when the mean (or difference) is "sufficiently unlikely" if the null hypothesis is true.

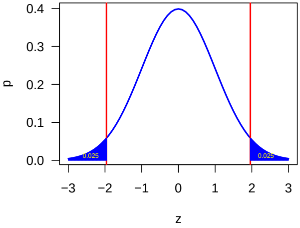
- Let $P(D|H_0)$ stand for the probability of the data (or some statistic representing the data) if the null hypothesis is true
- To reject the null hypothesis, $P(D|H_0)$ should be less than or equal to some threshold value, which we designate as α
- Conventional level of α: .05, .01
- $P(D|H_0)$ can obtained directly by looking up the probability of some observed value (obs) of a test statistic for a sampling distribution (SND, t)
- Another way is to find the **critical value** for the test statistic, beyond which the $P(D|H_0) \leq \alpha$

Assume $\alpha = .05$.

Two-tailed test

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$



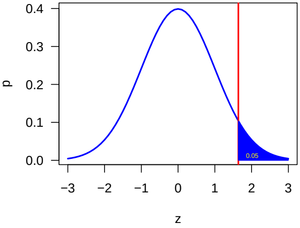
$$z_{crit} = 1.96$$

if $|z_{obs}| \geq z_{crit}$ then REJECT H_0

One-tailed test

$$H_0 : \mu < \mu_0$$

$$H_1 : \mu \geq \mu_0$$



$$z_{crit} = 1.64$$

if $z_{obs} \geq z_{crit}$ then REJECT H_0

The one-sample t-test

Purpose: Test the hypothesis that $\mu = \mu_0$, where μ_0 is some hypothesized population mean, and the population standard deviation σ is unknown

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$
$$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}} \qquad df = N - 1$$

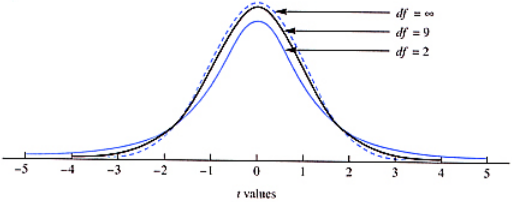


FIGURE 7.5 Three different t distributions

TABLE D The t distribution*

	Confidence interval percents (two-tailed)					
	80%	90%	95%	98%	99%	99.9%
	α level for two-tailed test					
	.20	.10	.05	.02	.01	.001
	α level for one-tailed test					
df	.10	.05	.025	.01	.005	.0005
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.474	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Mistakes in significance testing

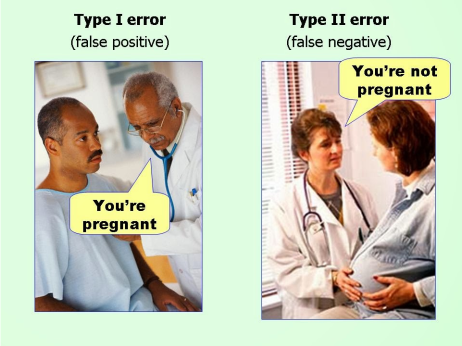
Research decisions

	State of the world	
	H_0 true	*H_0 false*
Reject H_0	Type I error	Correct decision
Retain H_0	Correct decision	Type II error

Mistakes in significance testing

Research decisions

	State of the world	
	H_0 *true*	H_0 *false*
Reject H_0	Type I error	Correct decision
Retain H_0	Correct decision	Type II error



Alpha level and Power

P(Type I error) = α
P(Type II error) = β

Power: the probability of rejecting the null hypothesis when it is, in fact, false. That is, the probability that your test will detect an effect of a given size.

- Power = 1 - β
- Power is inversely related to α

Effect size

- NHST determines the *statistical significance* of an effect. However, this doesn't say anything about its *practical significance*.
- The effect size index (d) "standardizes" an effect relative to population variability (like a z-score). This standardization makes it possible to compare effect sizes across studies with different sample sizes.

Calculating and Interpreting Effect Size

$d = \frac{|\bar{X} - \mu_0|}{\sigma}$ when σ is known, or:

$d = \frac{|\bar{X} - \mu_0|}{\hat{s}}$ when σ is unknown

d = .20	small
d = .50	medium
d = .80	large

Bringing it all together: Reporting results.

Example

You wish to test a company's claim in an advertisement that its batteries last, on average, 500 hours. You purchase 100 batteries and find that the mean battery life is 493 hours, with a standard deviation ($\hat{s}$) of 33. Is the company's claim accurate?

"In this study, we measured the battery life of each of 100 batteries. A one-sample t-test (two-tailed) was conducted to test the manufacturer's claim that the mean battery life is 500 hours. For the test, α was set to .05. The observed mean battery life for the sample was 493 hours ($SD = 33$). This was significantly lower than the manufacturer's claim of 500 hours, $t(99) = 2.12$, $p < .05$, although the effect size was small ($d = .21$). The 95% confidence interval for the population mean was [486.47,499.53]."

Low reproducibility of findings

Open Science Collaboration (2015), "Estimating the reproducibility of psychological science"

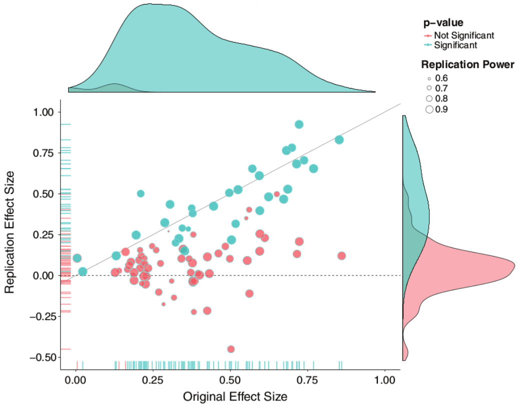


Fig. 3. Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

- attempted replications of 100 studies published in 2008 in three journals (JEP:LMC, PS, JPSP)
  - 97% of original $p < .05$
  - 36% of replications had $p < .05$

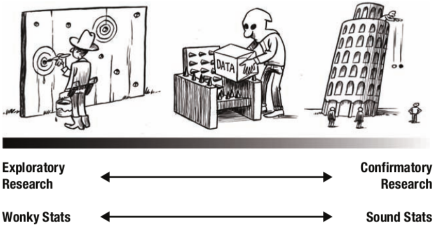
Publication bias

Sterling (1959); Rosenthal (1979)

- Journals typically uninterested in publishing “negative” findings
- Not all studies that have been conducted on a topic are available for meta-analysis
- significant results far more likely to be published than papers with a null result
- this gives a biased picture of the evidence on a phenomenon

Misrepresenting “exploratory” as “confirmatory”

Wagenmakers, Wetzels, Borsboom, van der Mass, & Kievit (2012)
Simmons, Nelson, & Simonsohn (2011)



(Figure downloaded from Flickr, courtesy of Dirk-Jan Hoek, reprinted in Wagenmakers et al., 2012)

- Hypothesizing After Results are Known (HARKing)
- Cherry picking DVs
- Failing to report nonsignificant experiments
- “Data massaging”
- “Data peeking”

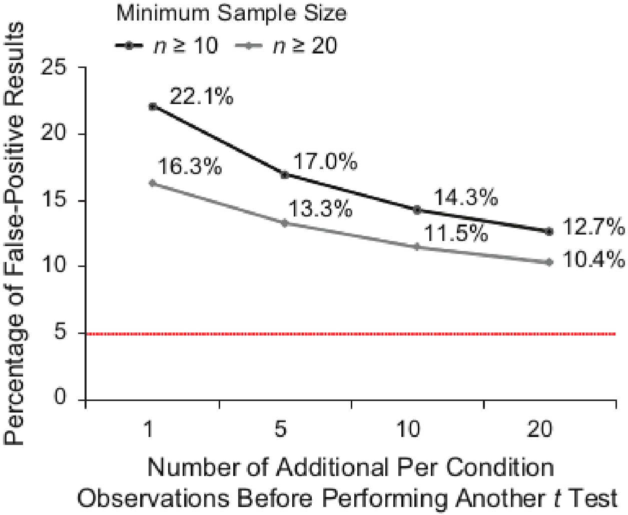
p-hacking

Simmons, Nelson, & Simonsohn (2010)

Table 1. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Data peeking



- researcher has already collected 10 or 20 and continues until $p < .05$ or $N = 50$

How widespread are such practices?

John, Loewenstein, & Prelec (2012)