Stats 95

Effect Size
Statistical Significance
Statistical Power
Confidence Intervals

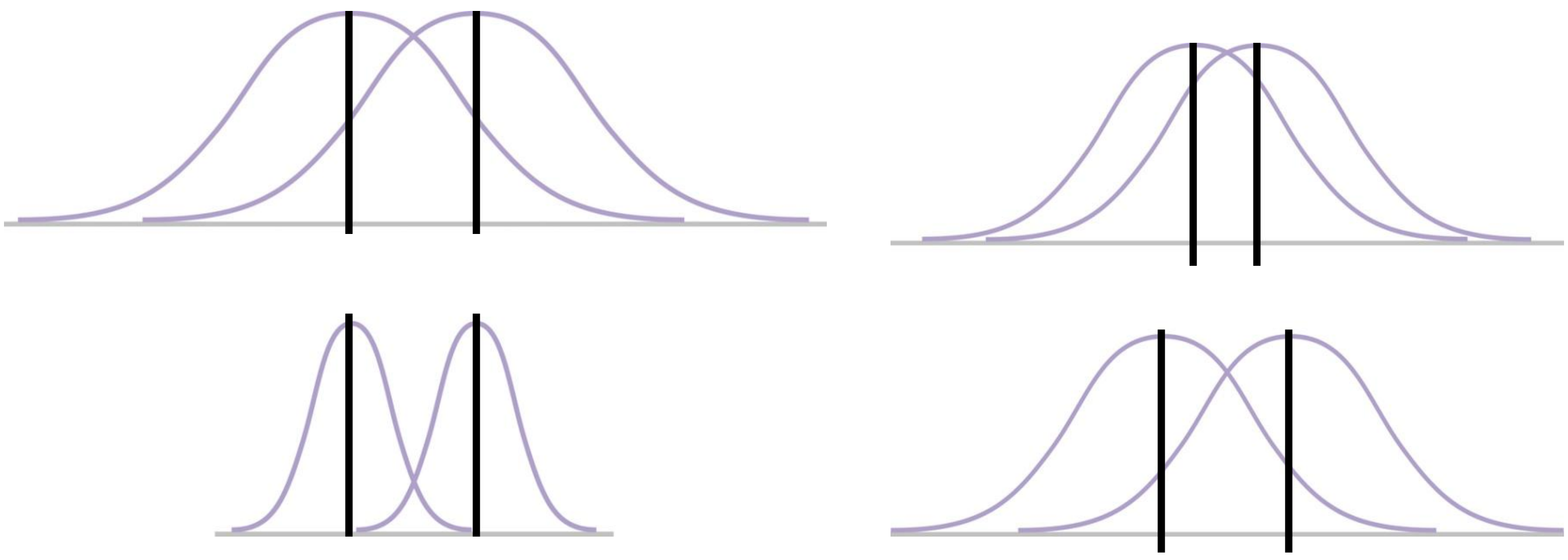
Effect Size

- Ratio of the Distance and Spread
  - – distance between means of two distributions to the standard deviation of the pop.
    - Affected by distance and standard deviation
- Expressed in z-scores
- Unaffected by sample size

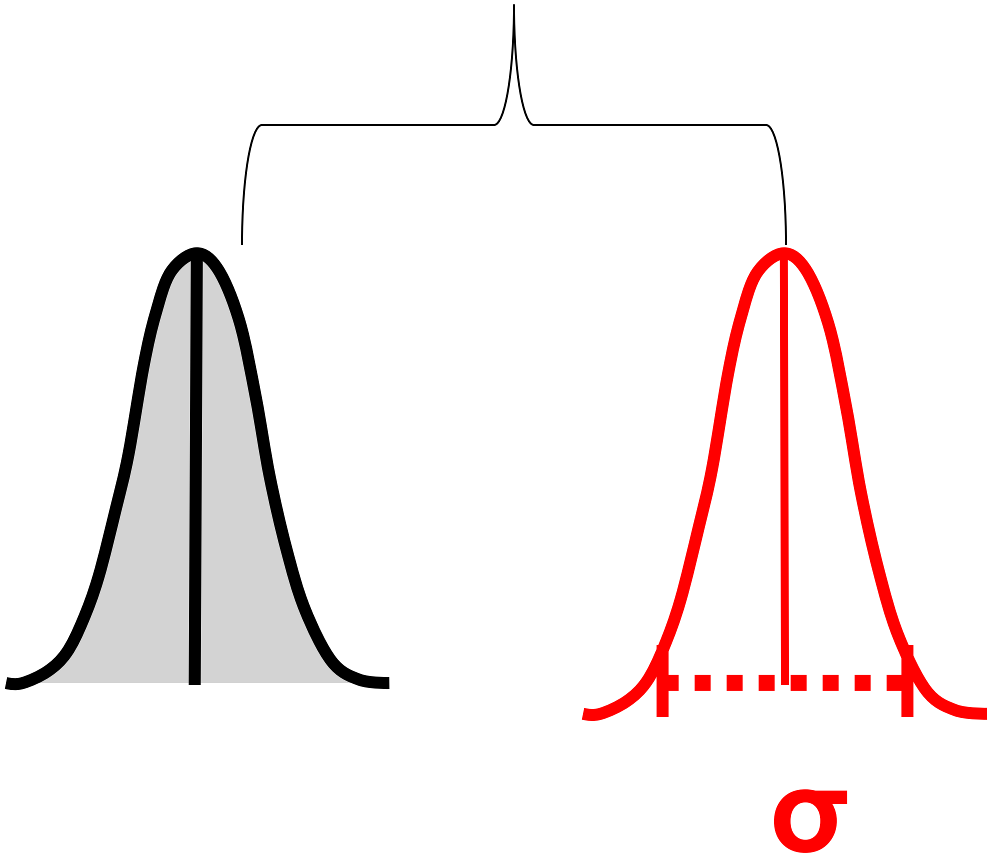
Effect Size and Mean Differences

Imagine both represent significant effects

Note the Spreads and Distances: Which effect is bigger?



Playing with Effect Size



Effect Size

- Cohen's *d*: effect size estimate
  - Effects size for *z statistic*

$$d = \frac{(M - \mu_M)}{\boxed{\sigma}}$$

TABLE 12-1. COHEN'S CONVENTIONS FOR EFFECT SIZES: *d*

Jacob Cohen has published guidelines (or conventions), based on the overlap between two distributions, to help researchers determine whether an effect is small, medium, or large. These numbers are not cutoffs, merely rough guidelines to aid researchers in their interpretation of results.

EFFECT SIZE	CONVENTION	OVERLAP
Small	0.2	85%
Medium	0.5	67%
Large	0.8	53%

CAUTION: *The formula for the z-stat and d, though similar, differ importantly at the denominator -- size matters for z but not d*

$$z = \frac{(M - \mu_M)}{\boxed{\sigma_M}}$$

Statistical Significance

- A finding is statistically significant if the data differ from what we would expect <u>from chance alone with a probability of 5% or less</u>.

- They may not be significant in the sense of big, important differences, but they occurred with a probability below the critical cutoff value, usually Alpha = .05 (z-stat = ±1.96 two-tailed or 1.64 one-tailed)
- **<u>The risk we are willing to accept of rejecting the Null Ho when it is CORRECT.</u>**

Significance

Pg 99. How did Fisher conceive of p-value.

The p value of .05 is an arbitrary but convenient level of significance for the practical investigator, but it does not mean that he allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained (because they failed to understand all the factors they needed to control). He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result.

(In a way, it means the experimenter knows how to design the experiment so that it will fail to give him a significant result only .05)

Statistical Power

- Important because it informs sample size

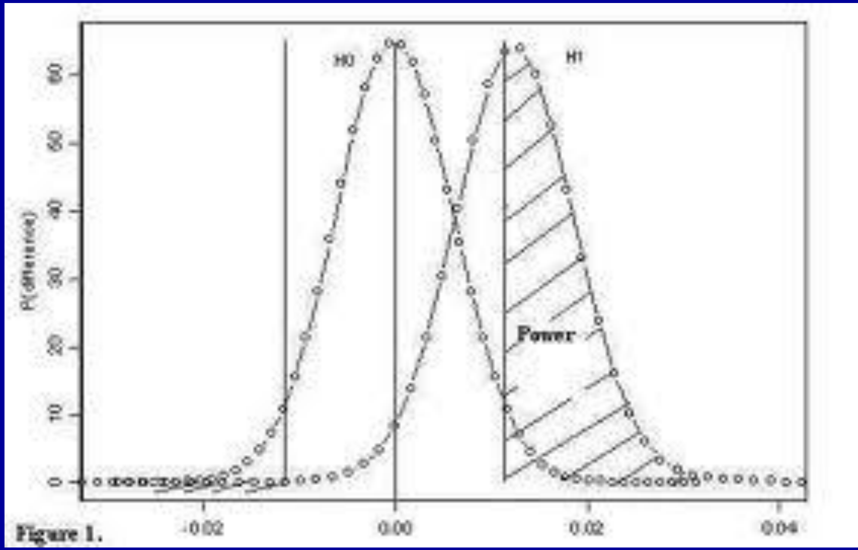
“Before the middle of the eighteenth century there is little indication of a willingness of astronomers to combine observations; indeed there was sometimes an outright refusal to combine them. The idea that accuracy could be increased by combining measurements made under different conditions was slow to come. They feared that errors in one observation would contaminate others, that errors would multiply, not compensate.” – Stigler, 1986

As long as you

“…randomize, randomize, randomize.” -- Fisher

Statistical Power

- Defining statistical power:
  - % Area of Hits – % Area of Misses
  - ideally 80% minimum



- Statistical power is used to estimate the required sample size.

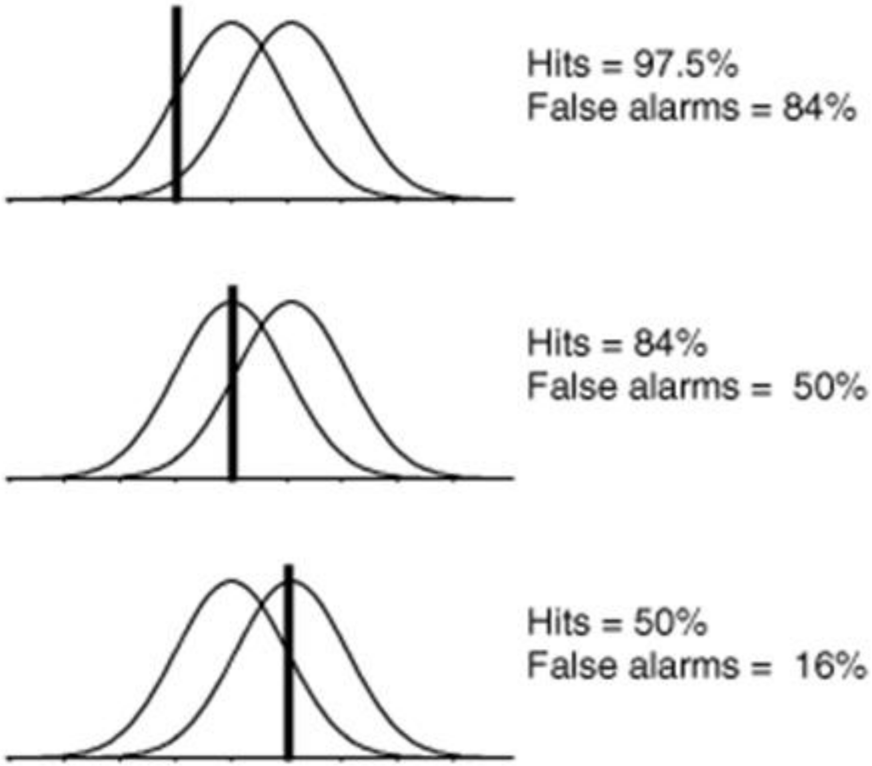


Figure 3: Effect of shifting the criterion

Lady Tasting Tea

Chapter 11, pg 109

What did Newman and Pearson call the two forms of hypotheses, and what did statistical power refer to? *Newman and Pearson called the hypothesis being tested the "null hypothesis" and the other hypothesis as the "alternative."*

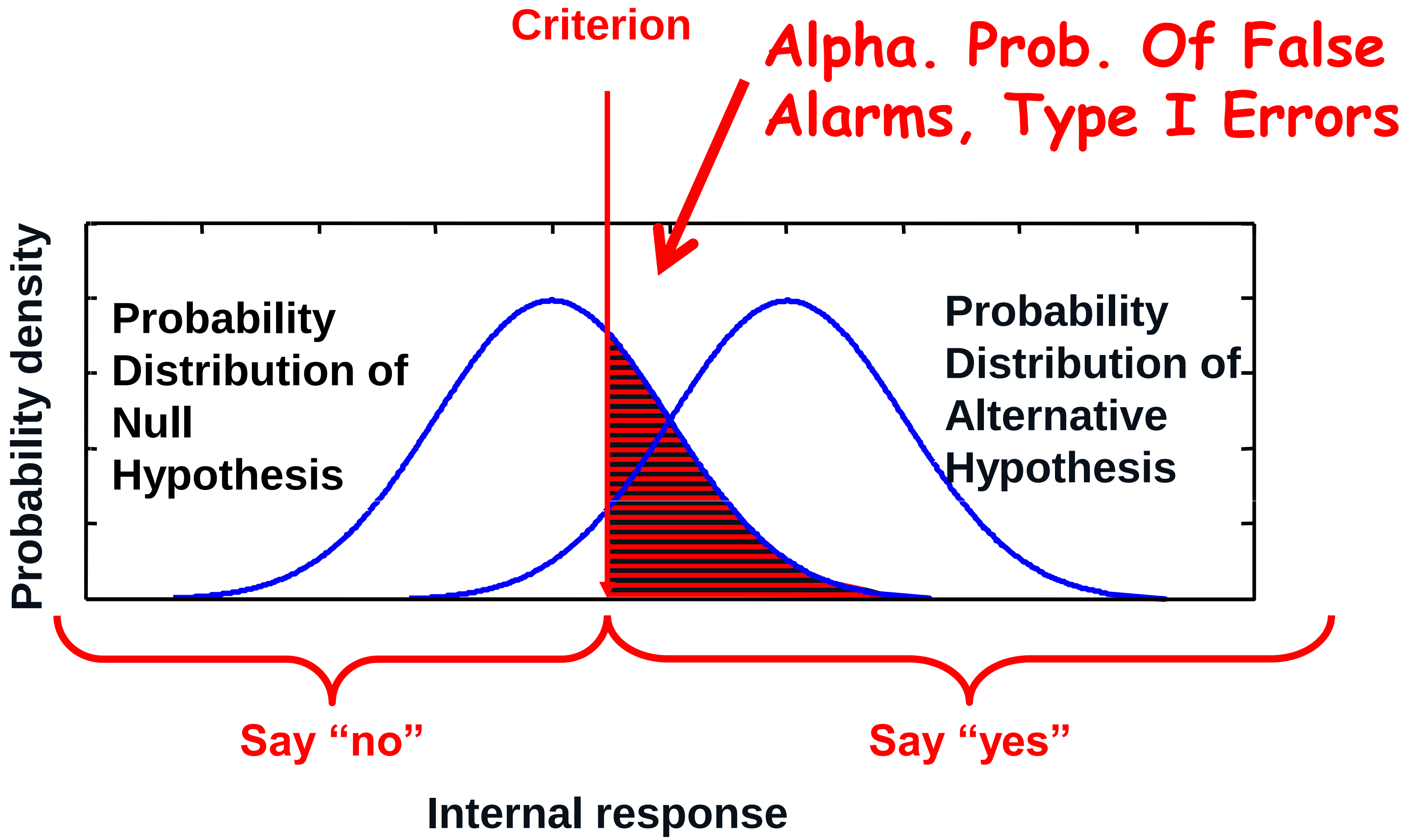
In their formulation, the p value is calculated for testing the null hypothesis but the power refers to how the p-value will behave if the alternative is in fact true.

What was power according to Newman? *The probability of detecting that alternative, if it is true.*

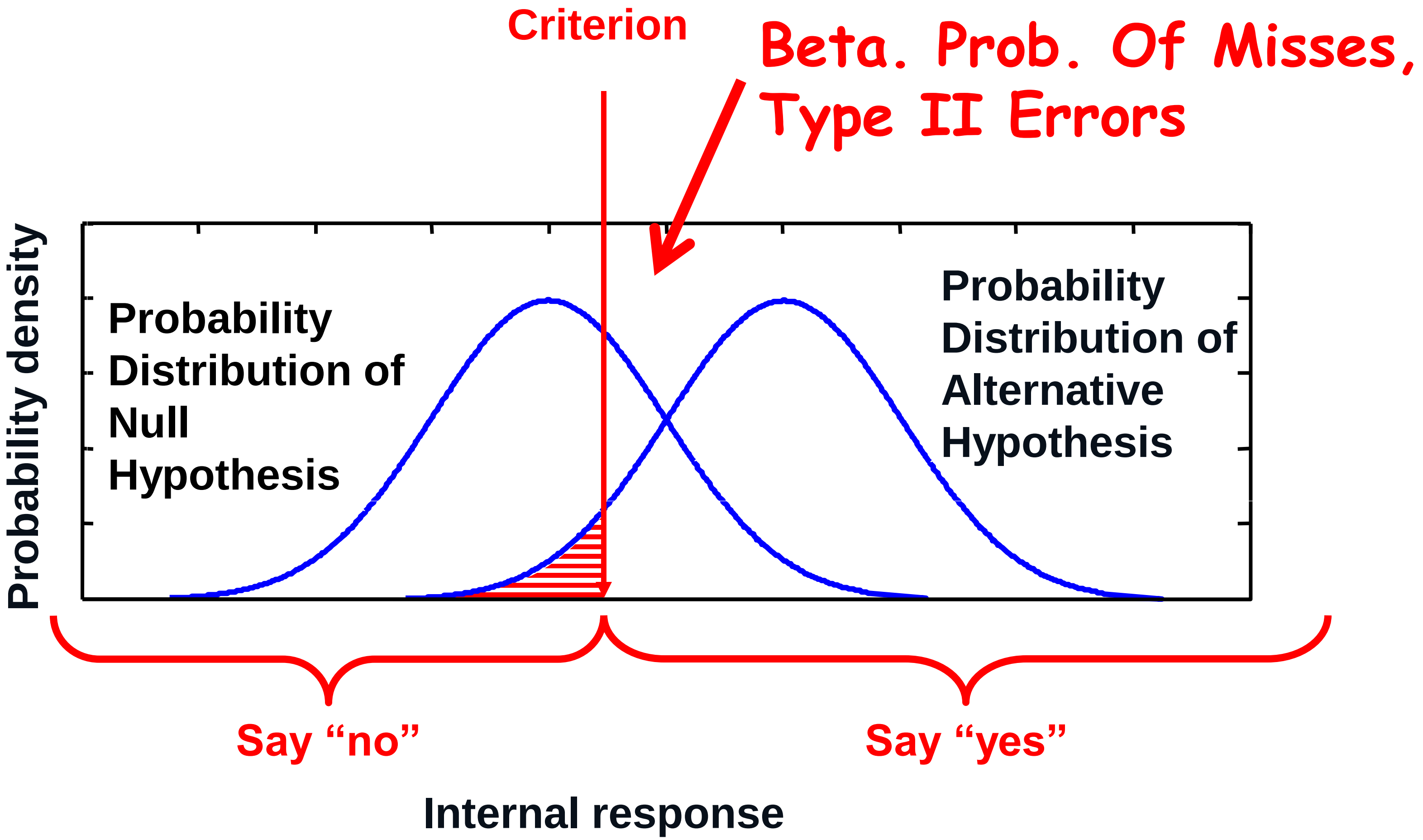
Power & Significance

- Power is the probability of accepting the Alternative Hypothesis when it is True
  - the likelihood that a study will detect an effect when there is an effect there to be detected.
- Significance is the probability of rejecting the Null Hypothesis when it is True

False Alarms (Type I Error, *Alpha*)
(response "yes" on no-signal trial)

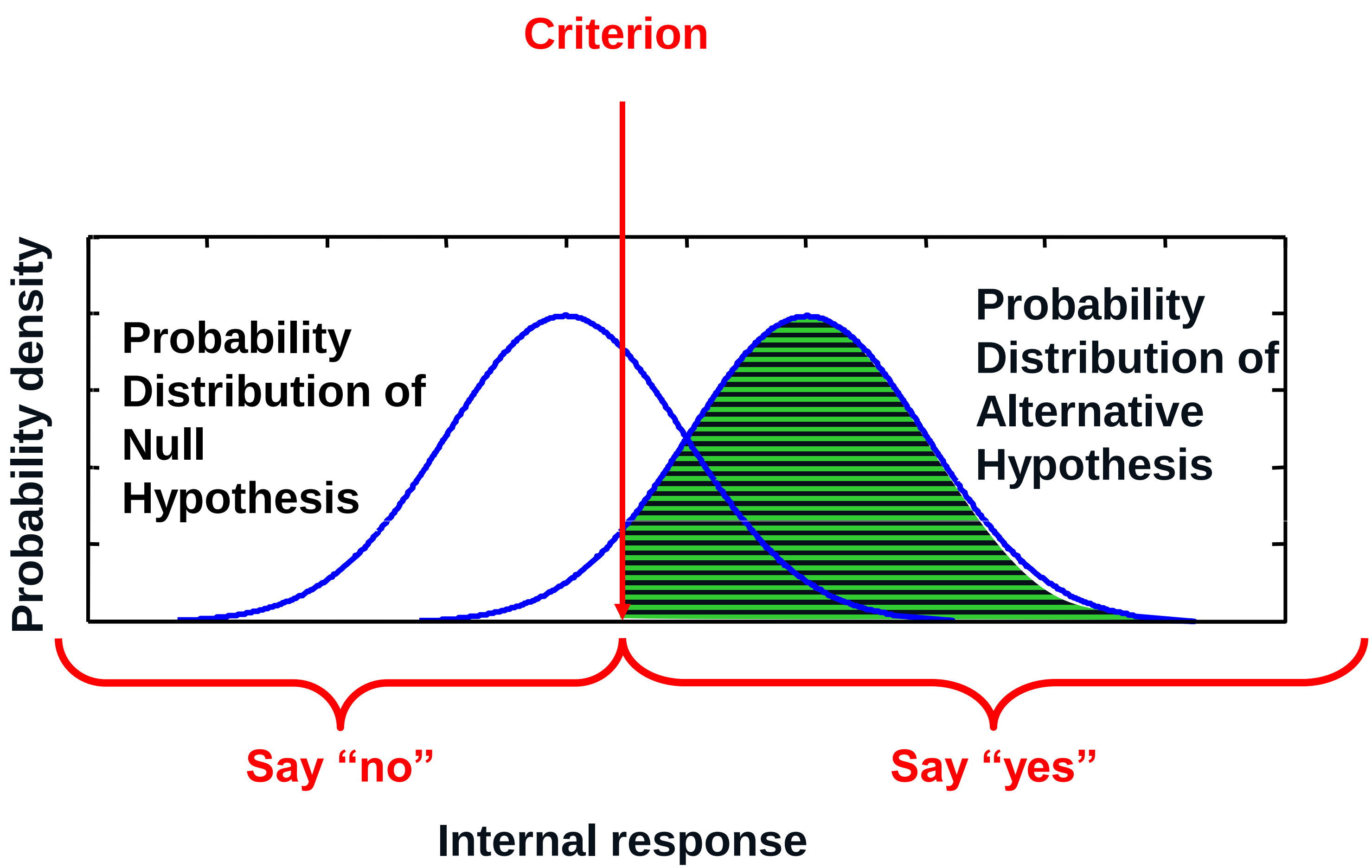


Misses (Type II Error, βeta)
(response “no” on signal trial)



Hits (1-beta)

(response “yes” on signal trial)



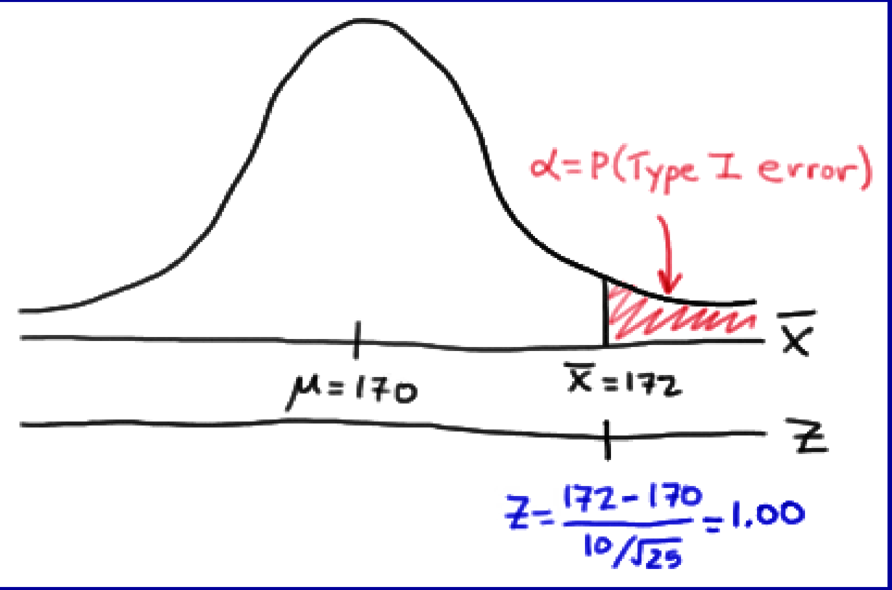
- **Example**
- The Brinell hardness scale is one of several definitions used in the field of materials science to quantify the hardness of a piece of metal. The Brinell hardness measurement of a certain type of rebar used for reinforcing concrete and masonry structures was assumed to be normally distributed with a standard deviation of 10 kilograms of force per square millimeter. Using a random sample of n = 25 bars, an engineer is interested in performing the following hypothesis test:
- the null hypothesis H_0: $\mu = 170$
- against the alternative hypothesis H_A: $\mu > 170$
- If the engineer decides to reject the null hypothesis if the sample mean is 172 or greater, that is, if $\bar{X} \geq 172$, what is the probability that the engineer commits a Type I error?

https://onlinecourses.science.psu.edu/stat414/node/304

Calculating Alpha: Calculating the probability of getting a sample mean of 172 or more when the true (unknown) population is 170

- Now, we can calculate the engineer's value of α by making the transformation from a normal distribution with a mean of 170 and a standard deviation of 10, N = 25, to that of Z, the standard normal distribution using:

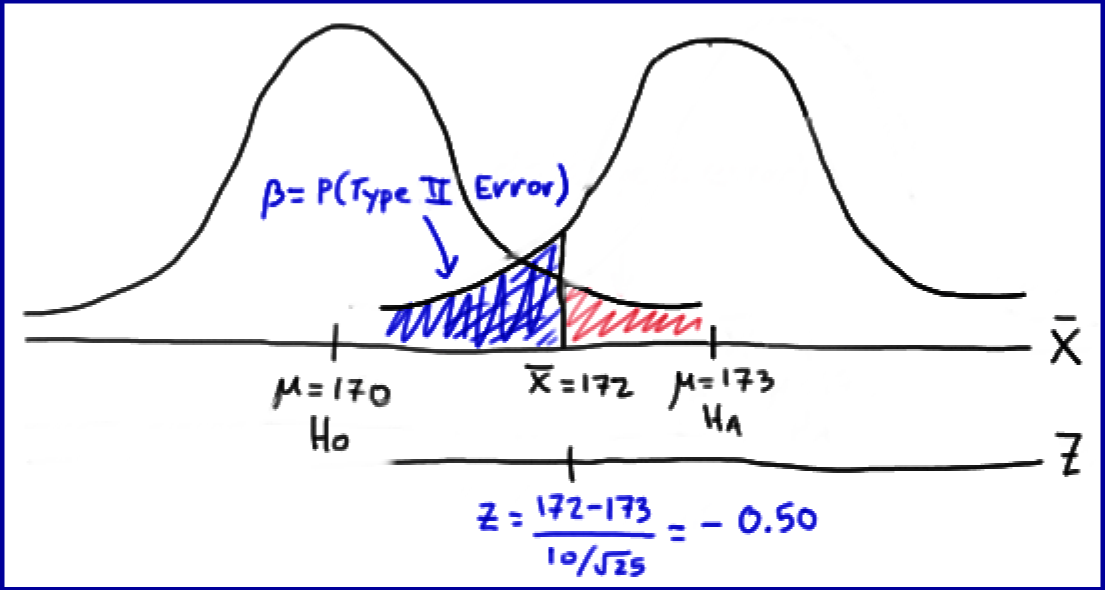
$$z = \frac{M - \mu}{\sigma / \sqrt{N}}$$



Calculating Beta: Calculating the probability of getting a sample mean less than 172 when the true (unknown) population mean is 173

- Now, we can calculate the engineer's value of β by making the transformation from a normal distribution with a mean of 173 and a standard deviation of 10, N = 25, to that of Z, the standard normal distribution using:

$$z = \frac{M - \mu}{\sigma / \sqrt{N}}$$



Calculating Power: Calculating the probability of getting a sample mean of 172 or more when the true (unknown) population mean is 173.

- Now, we can calculate the engineer's value for Power, 1-β , given a normal distribution with a mean of 173 and a standard deviation of 10, N = 25:

$$\text{Power} = 1 - \beta = 1 - 0.3085 = 0.6915$$

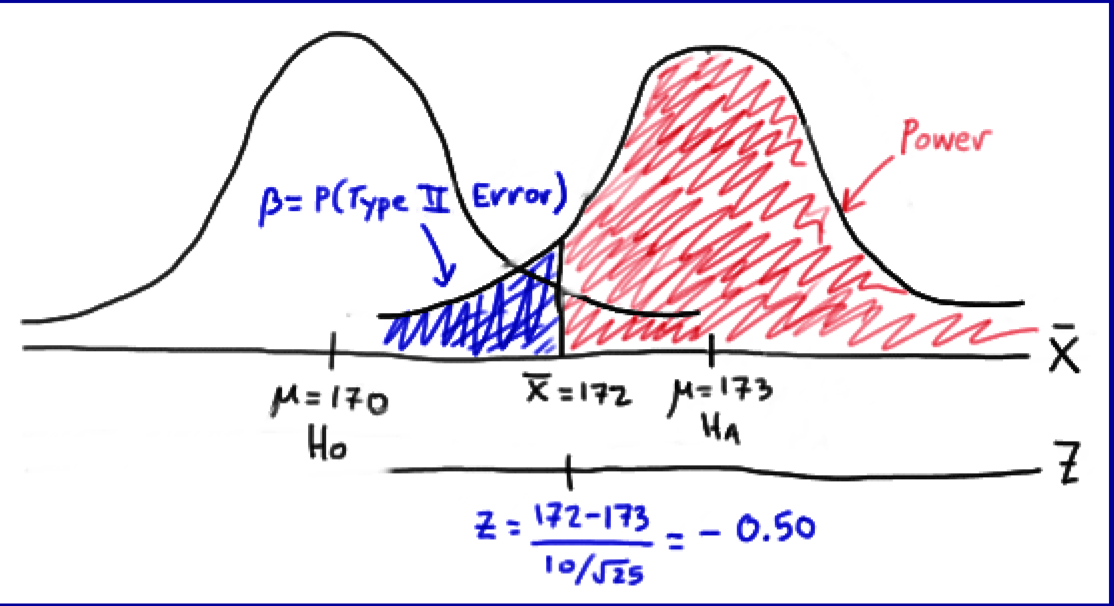


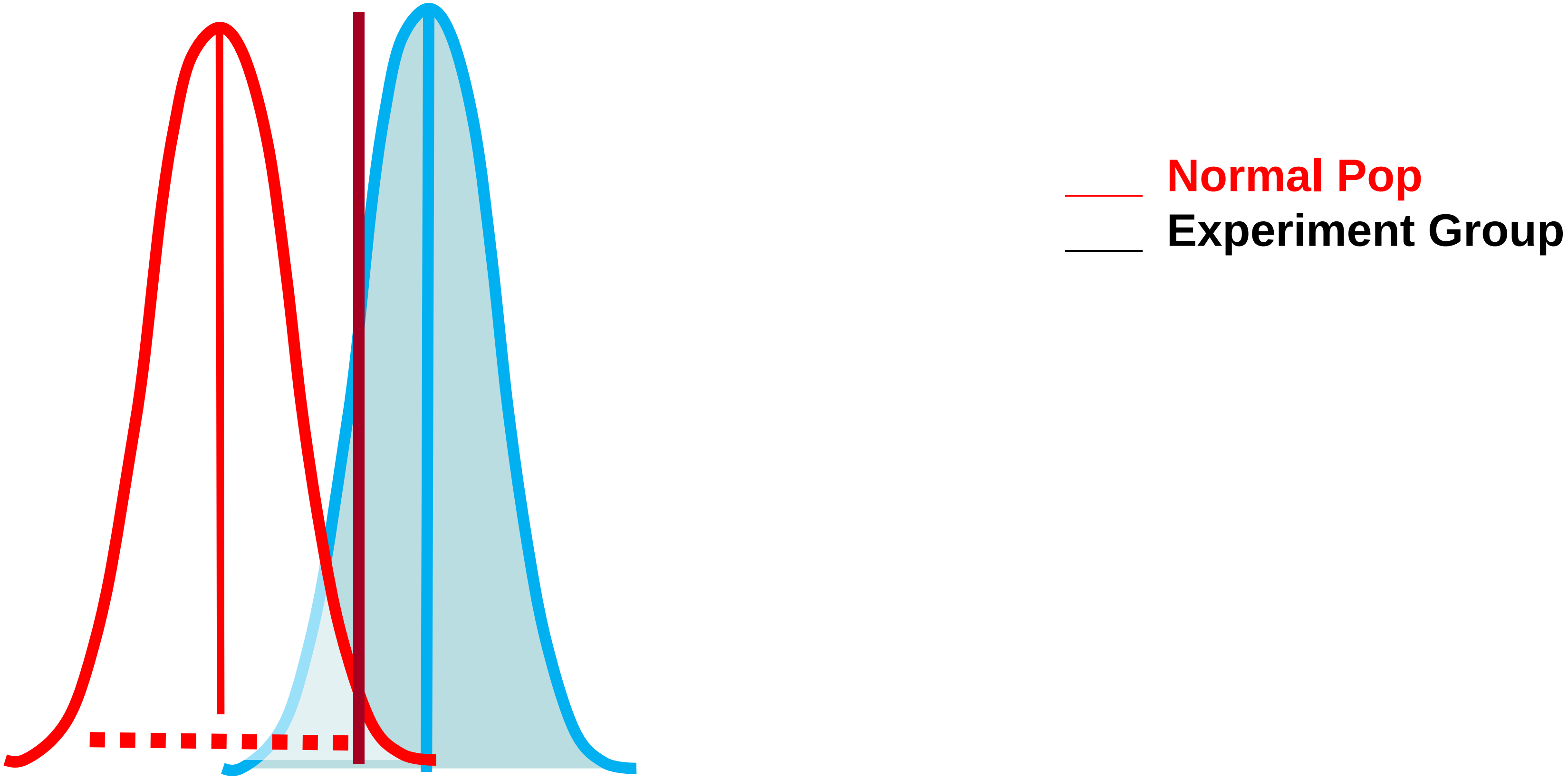
Table 2 Power of one-sample two-tailed t-test at .05 level

n	d														
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50
2	0.05	0.05	0.05	0.06	0.06	0.07	0.07	0.08	0.09	0.09	0.10	0.11	0.12	0.13	0.13
3	0.05	0.06	0.06	0.07	0.08	0.10	0.12	0.14	0.16	0.18	0.20	0.23	0.26	0.29	0.32
4	0.05	0.06	0.07	0.09	0.11	0.14	0.17	0.21	0.25	0.29	0.34	0.38	0.43	0.48	0.53
5	0.05	0.06	0.08	0.11	0.14	0.18	0.23	0.28	0.34	0.40	0.47	0.53	0.59	0.65	0.71
6	0.06	0.07	0.09	0.13	0.17	0.22	0.29	0.36	0.43	0.51	0.58	0.66	0.72	0.78	0.83
7	0.06	0.07	0.10	0.15	0.20	0.27	0.35	0.43	0.52	0.60	0.68	0.75	0.82	0.87	0.91
8	0.06	0.08	0.11	0.17	0.23	0.31	0.40	0.50	0.59	0.68	0.76	0.83	0.88	0.92	0.95
9	0.06	0.08	0.13	0.19	0.26	0.35	0.46	0.56	0.66	0.75	0.82	0.88	0.93	0.96	0.98
10	0.06	0.09	0.14	0.21	0.29	0.40	0.51	0.62	0.72	0.80	0.87	0.92	0.95	0.98	0.99
11	0.06	0.09	0.15	0.23	0.32	0.44	0.55	0.67	0.77	0.85	0.91	0.95	0.97	0.99	0.99
12	0.06	0.10	0.16	0.25	0.35	0.48	0.60	0.71	0.81	0.88	0.93	0.97	0.98	0.99	1.00
13	0.06	0.10	0.17	0.26	0.38	0.51	0.64	0.75	0.85	0.91	0.95	0.98	0.99	1.00	1.00
14	0.06	0.11	0.18	0.28	0.41	0.55	0.68	0.79	0.88	0.93	0.97	0.99	0.99	1.00	1.00
15	0.07	0.11	0.19	0.30	0.44	0.58	0.71	0.82	0.90	0.95	0.98	0.99	1.00	1.00	1.00
16	0.07	0.12	0.20	0.32	0.47	0.61	0.75	0.85	0.92	0.96	0.98	0.99	1.00	1.00	1.00
17	0.07	0.12	0.21	0.34	0.49	0.64	0.77	0.87	0.94	0.97	0.99	1.00	1.00	1.00	1.00
18	0.07	0.13	0.23	0.36	0.52	0.67	0.80	0.89	0.95	0.98	0.99	1.00	1.00	1.00	1.00
19	0.07	0.13	0.24	0.38	0.54	0.70	0.82	0.91	0.96	0.98	1.00	1.00	1.00	1.00	1.00
20	0.07	0.14	0.25	0.40	0.57	0.72	0.84	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00
21	0.07	0.14	0.26	0.42	0.59	0.74	0.86	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00
22	0.07	0.15	0.27	0.43	0.61	0.77	0.88	0.95	0.98	0.99	1.00	1.00	1.00	1.00	1.00
23	0.07	0.15	0.28	0.45	0.63	0.79	0.89	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00
24	0.08	0.16	0.29	0.47	0.65	0.80	0.91	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00
25	0.08	0.16	0.30	0.48	0.67	0.82	0.92	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00
26	0.08	0.17	0.31	0.50	0.69	0.84	0.93	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
27	0.08	0.17	0.32	0.52	0.71	0.85	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
28	0.08	0.18	0.33	0.53	0.72	0.86	0.95	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
29	0.08	0.18	0.35	0.55	0.74	0.88	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
30	0.08	0.19	0.36	0.56	0.75	0.89	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Five Factors that Influence Power

- Increase Alpha. Move the criteria of .05 towards the middle of the Normal Distribution.
  - Bad choice: you increase power but you also increase Type I errors (False Alarms)
- Turn Two-Tailed Test into One-Tailed test
  - Same as above
- Increase *N*. Increase the sample size
  - By increasing sample size you increase the numerator of the test statistics, increasing the likelihood of rejecting the Null Hyp.
- (Run a Good Experiment) maximize the distance between the means
- (Run a Good Experiment) Decrease the standard deviation

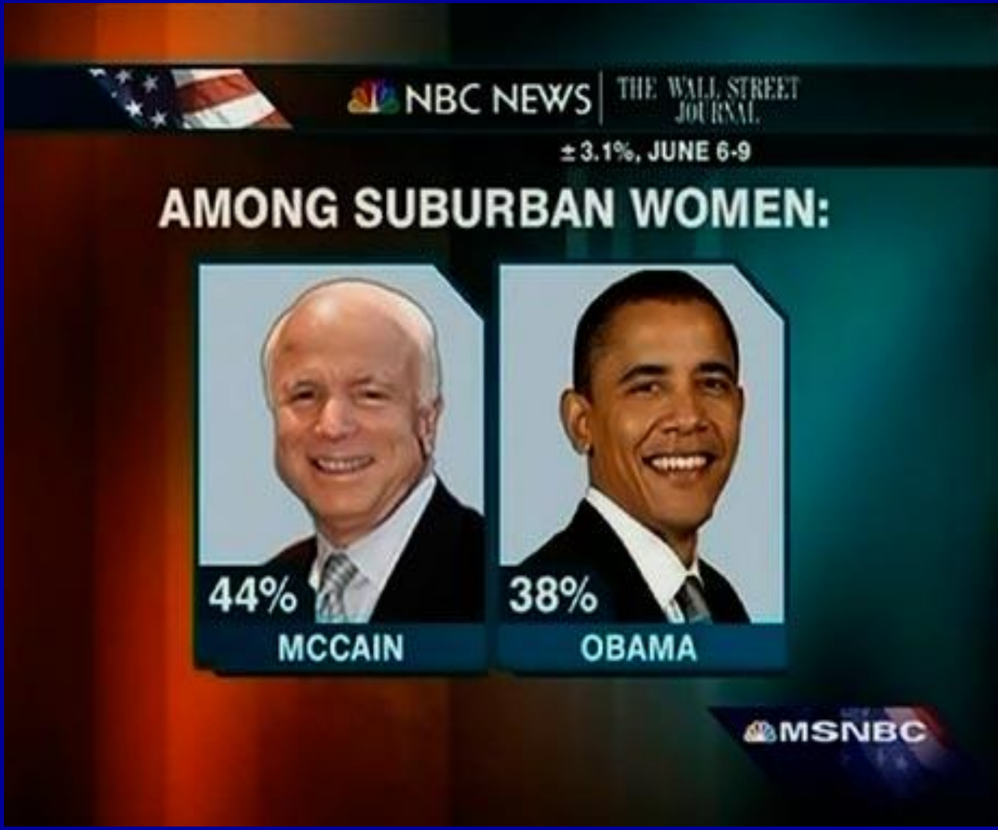
Playing with Statistical Power



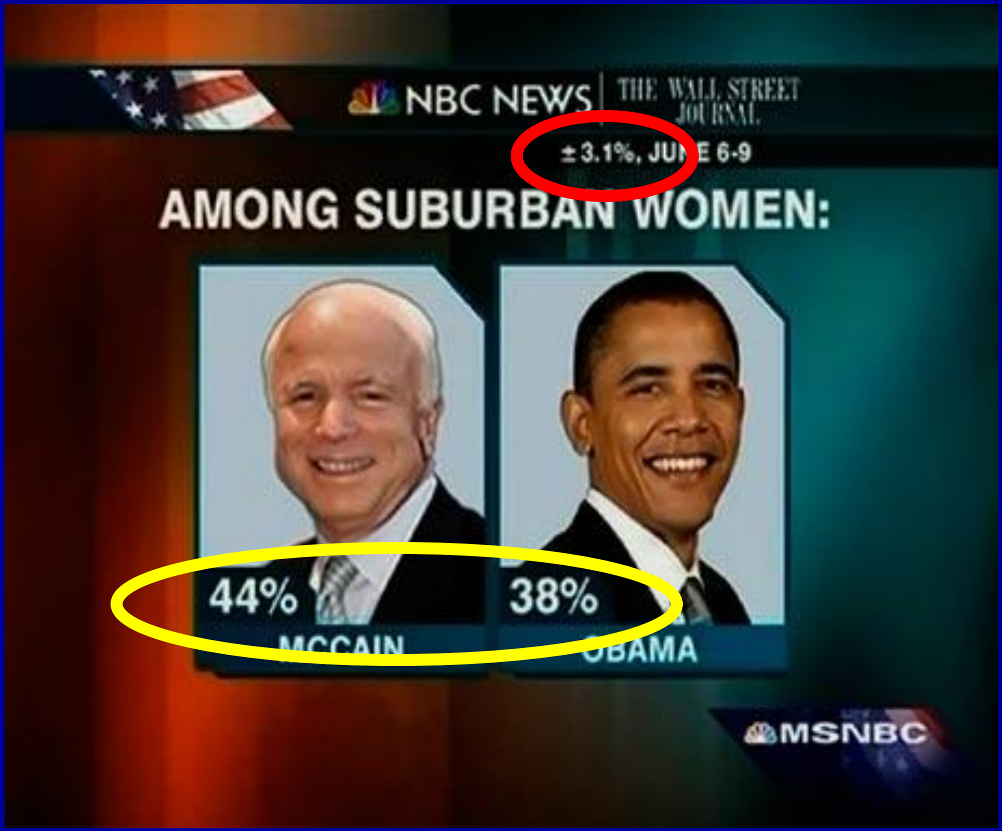
Confidence Intervals

An Alternative to Hypothesis Testing

- Confidence Interval answers: How close? How Confident?
- Gives more information than just hypothesis testing, and in some cases used instead.
- Used in other forms of statistical test (*z*, *t*, *f*)



Confidence Intervals: An Alternative to Hypothesis Testing



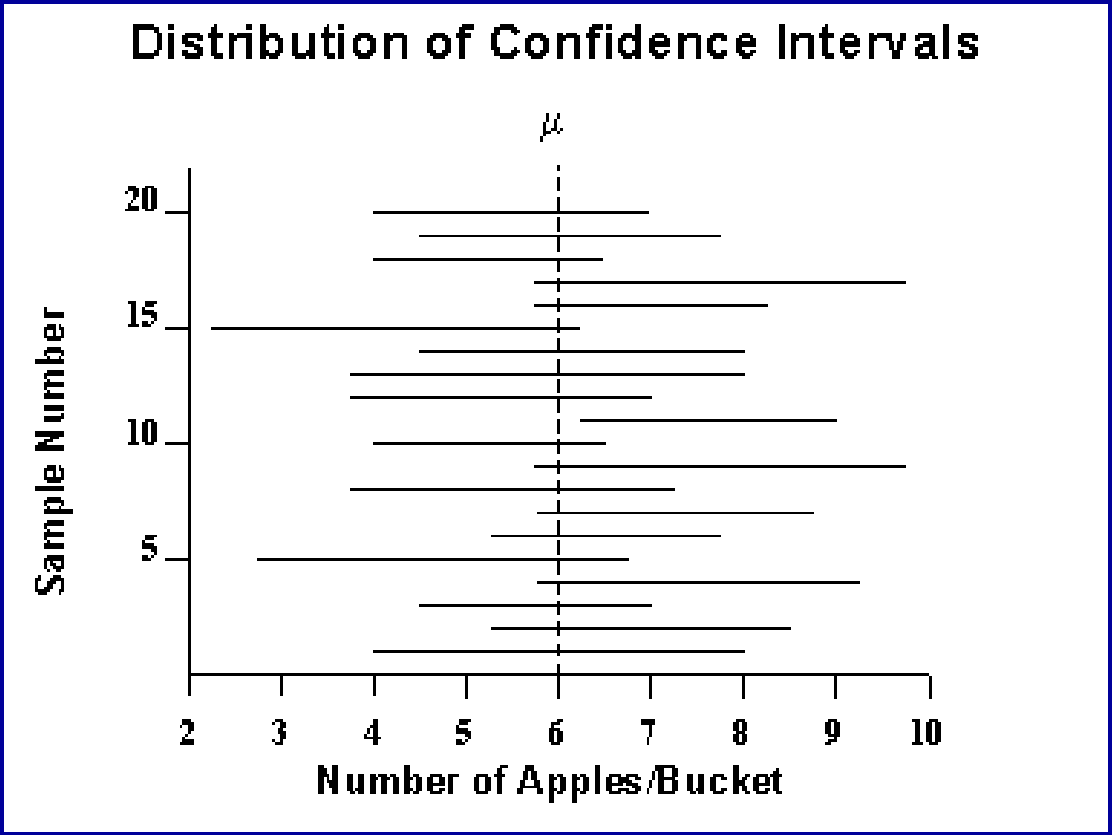
- Point estimate: summary statistic – one number as an estimate of the population
  – E.g., 32 point difference b/w Boys and Girls
- Interval estimate: based on our sample statistic, range of sample statistics we would expect if we repeatedly sampled from the same population
- Confidence interval
  – Interval estimate, the range of values that will include the value of the population (from which our sample is drawn) a percentage of the time were we to sample repeatedly.
  – Typically set at 95%, the Confidence Level

Confidence Intervals

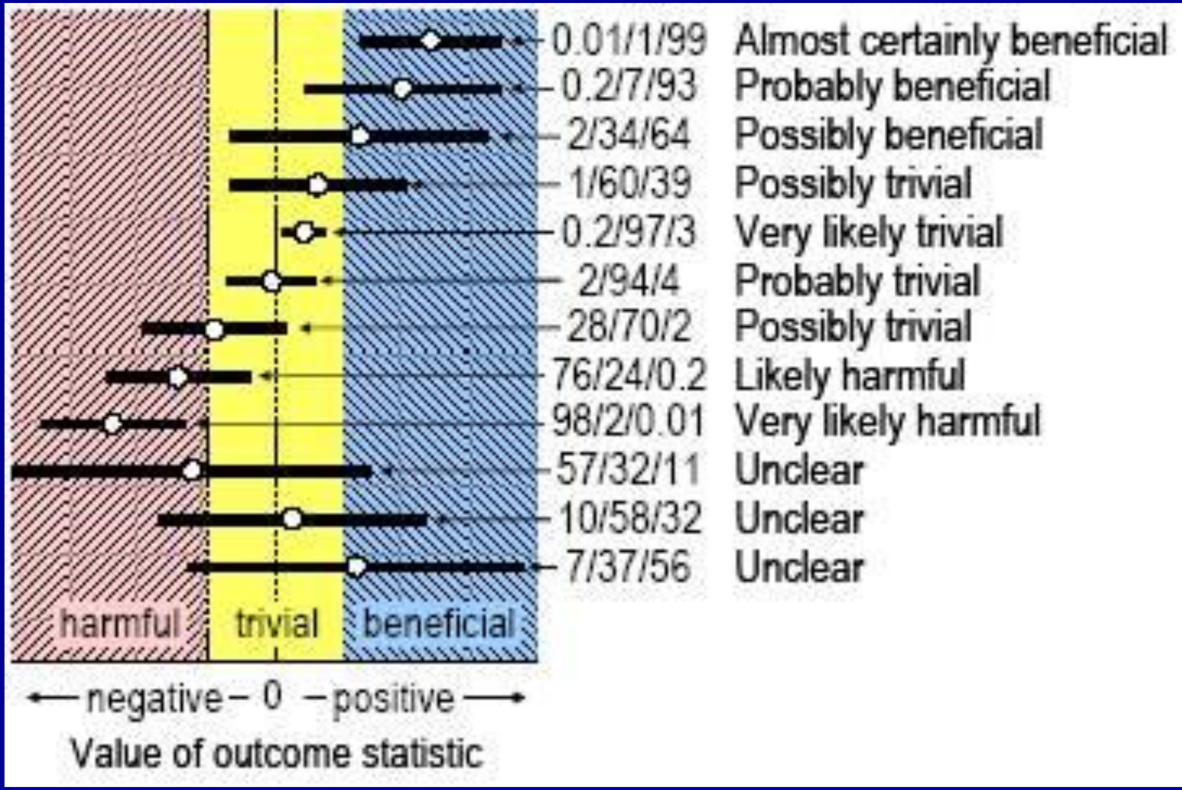
- The range (Interval Estimate) that includes the mean of the population of interest were we to sample from the the same population repeatedly
  - Typically set at 95%
- The range around the mean when we add and subtract a margin of error
- Confirms findings of hypothesis testing and adds more detail

Confidence Interval

- Use the Confidence **LEVEL** (e.g. 95% ➔ z-score of ±1.96), to calculate the Confidence **INTERVAL** (range, e.g., 31,268.33 – 36.731.67) .
- If C. Interval includes the mean of the population of interest, then sample is not statistically significant
- Confidence intervals combines statistical significance and effect size.
- As sample size increases, Confidence Interval narrows.

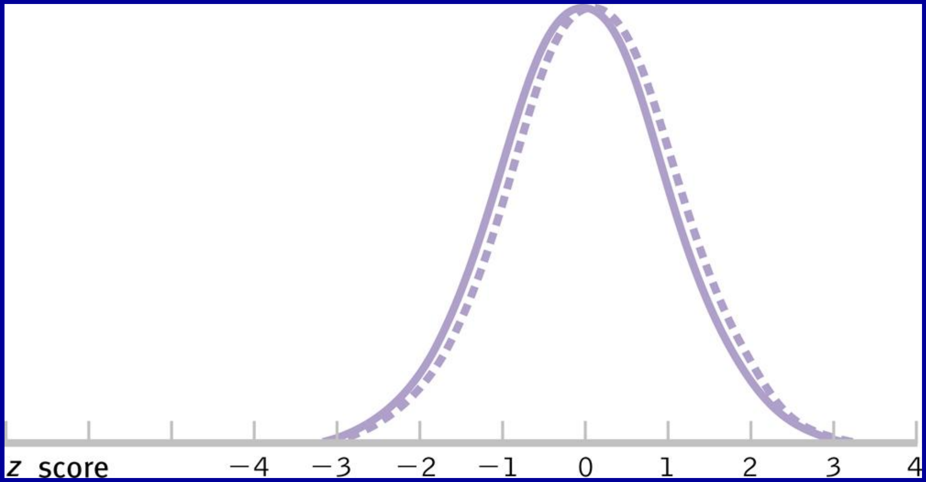


- Confidence interval
  - Interval estimate that includes the mean of the population a certain percentage of the time were we to sample from the same population repeatedly
  - Typically set at 95%, the Confidence Level


0.01/1/99 Almost certainly beneficial
0.2/7/93 Probably beneficial
2/34/64 Possibly beneficial
1/60/39 Possibly trivial
0.2/97/3 Very likely trivial
2/94/4 Probably trivial
28/70/2 Possibly trivial
76/24/0.2 Likely harmful
98/2/0.01 Very likely harmful
57/32/11 Unclear
10/58/32 Unclear
7/37/56 Unclear
harmful
trivial
beneficial
negative – 0 – positive
Value of outcome statistic

Confidence Intervals and Effect Sizes

Figure 12-1: A Gender Difference in Mathematics Performance – amount of overlap as reported by Hyde (1990)



- Teen Talk Barbie: A sample of 10,000 boys and 10,000 females in grades 7-10, who were the top 2-3% of on standardized math tests,
- Average for the boys was 32 points higher than the average for the girls. So, obviously, boys are better at math than girls, correct?
- Finding a difference between the boys group and the girls group does not mean that ALL boys score above ALL girls,
- *Statistically significant* does not mean *quantitatively meaningful.*
- The distance between the means, Effect Size, counts!

Confidence Interval for *Z*-test

$$M_{lower} = -\,z(\sigma_M) + M_{sample}$$

$$M_{upper} = z(\sigma_M) + M_{sample}$$

The length of the confidence *<u>interval</u>* *is influenced by sample size of the sample mean.*

The larger the sample, the narrower the interval.

But that does not influence the confidence *<u>level,</u>* *because the standard error also decreases as the sample size increases,*

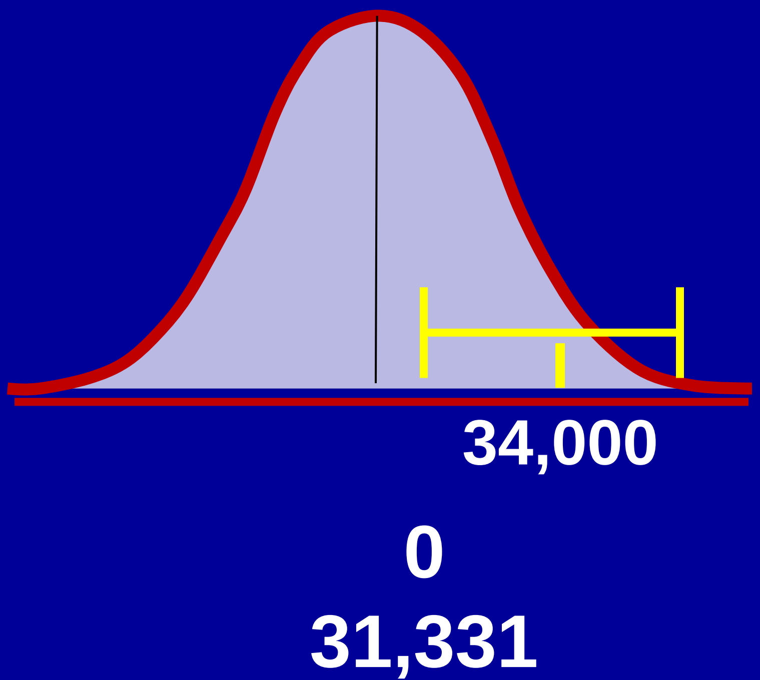
Confidence Intervals: Example

- According to the 2003-2004 annual report of the Association of Medical and Graduate Departments of Biochemistry, the **average stipend** for a postdoctoral trainee in biochemistry was **$31,331** with **a standard deviation of $3,942**. Treating this as the population, assume that you asked the **8** biochemistry postdoctoral trainees at your institution what their annual stipend was and that it **averaged $34,000**.
- a. Construct an 80% confidence interval for this sample mean.
- b. Construct a 95% confidence interval for this sample mean.
- c. Based on these two confidence intervals, if you had performed a two-tailed hypothesis test with a *p level of 0.20, would you have found that the trainees at your school earn more, on average, than the population of trainees? If you had performed the same test with a two-tailed p level of 0.05, would you have made another decision regarding the null hypothesis?*

Confidence Intervals: Example

1. Draw a Graph
2. Summarize Parameters

$$\mu = 31{,}331$$
$$\sigma = 3{,}942$$
$$N = 8$$
$$M = 34{,}000$$



3. Choose Boundaries

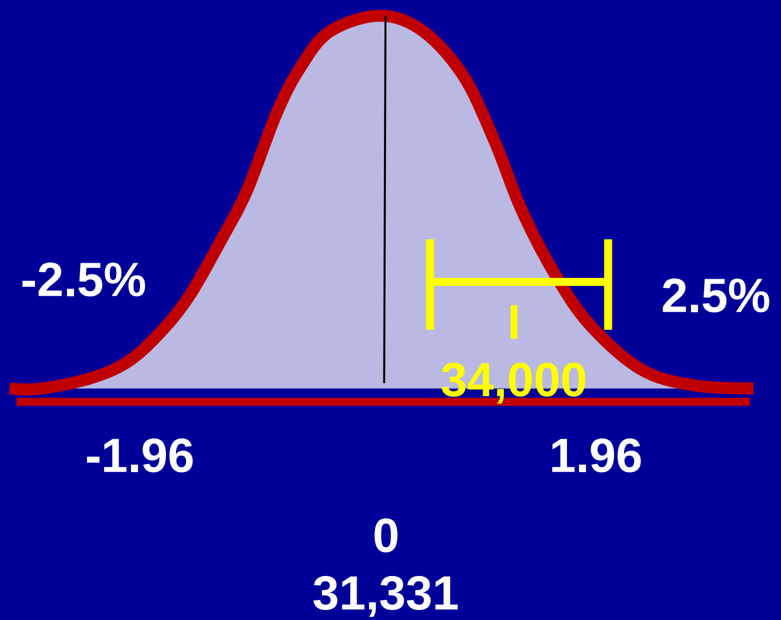
4. Determine z Statistics

Confidence Intervals: Example

- Draw a Graph
- Summarize parameters

$$\mu = 31{,}331$$
$$\sigma = 3{,}942$$
$$N = 8$$
$$M = 34{,}000$$

- Choose Boundaries
  - Two-Tailed 80%
  - Two-Tailed 95%



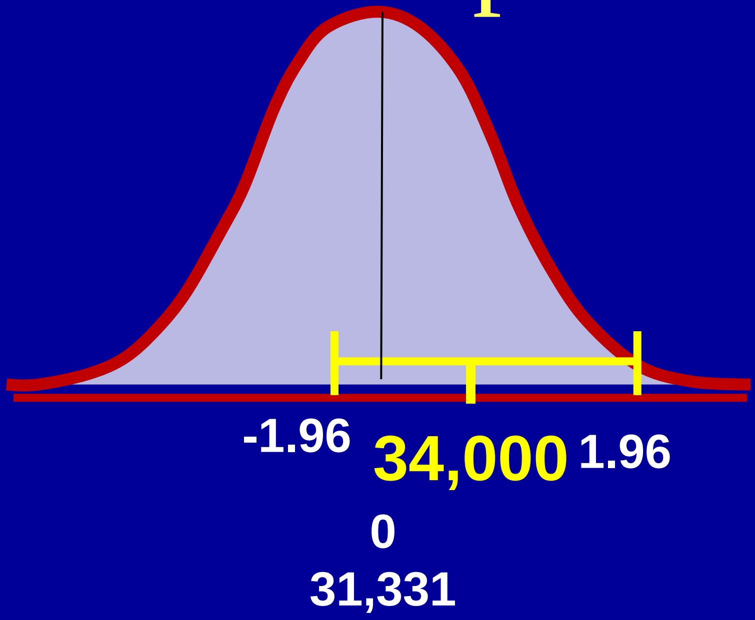
- Determine z Statistics. Choosing the bounds of confidence of 5%, in a two-tailed test means we divide by two, for 2.5% and find the z-statistic associated with that percentage.

Confidence Intervals: Example

- Turn z Statistics into raw scores
  - *Strategy: start with the final formula and work backwards*

$M_{lower} = - z(\sigma_M) + M_{sample}$

$M_{upper} = z(\sigma_M) + M_{sample}$



$$\sigma_m = \frac{\sigma}{\sqrt{N}} = \frac{3{,}942}{\sqrt{8}} = 1393.707$$

$$M_{Lower} = -1.29(1393.707) + 34{,}000 = 32{,}327.55$$
$$M_{Upper} = 1.29(1393.707) + 34{,}000 = 35{,}797.88$$

$$M_{Lower} = -1.96(1393.707) + 34{,}000 = 31{,}268.33$$
$$M_{Upper} = 1.96(1393.707) + 34{,}000 = 36{,}731.67$$

The End

The Definition Of Statistical Power

- Statistical power is the probability of not missing an effect, due to sampling error, when there really is an effect there to be found.

- Power is the probability (prob = 1 - β) of correctly rejecting H_o when it really is false.

When H_o is False And You Fail To Reject It, You Make A Type II Error

- *When, in the population, there really is an effect, but your statistical test comes out non-significant, due to inadequate power and/or bad luck with sampling error, you make a Type II error.*

- When H_o is false, (so that there really is an effect there waiting to be found) the probability of making a Type II error is called ***beta*** (β).

When H_o Is True And You Reject It, You Make A Type I Error

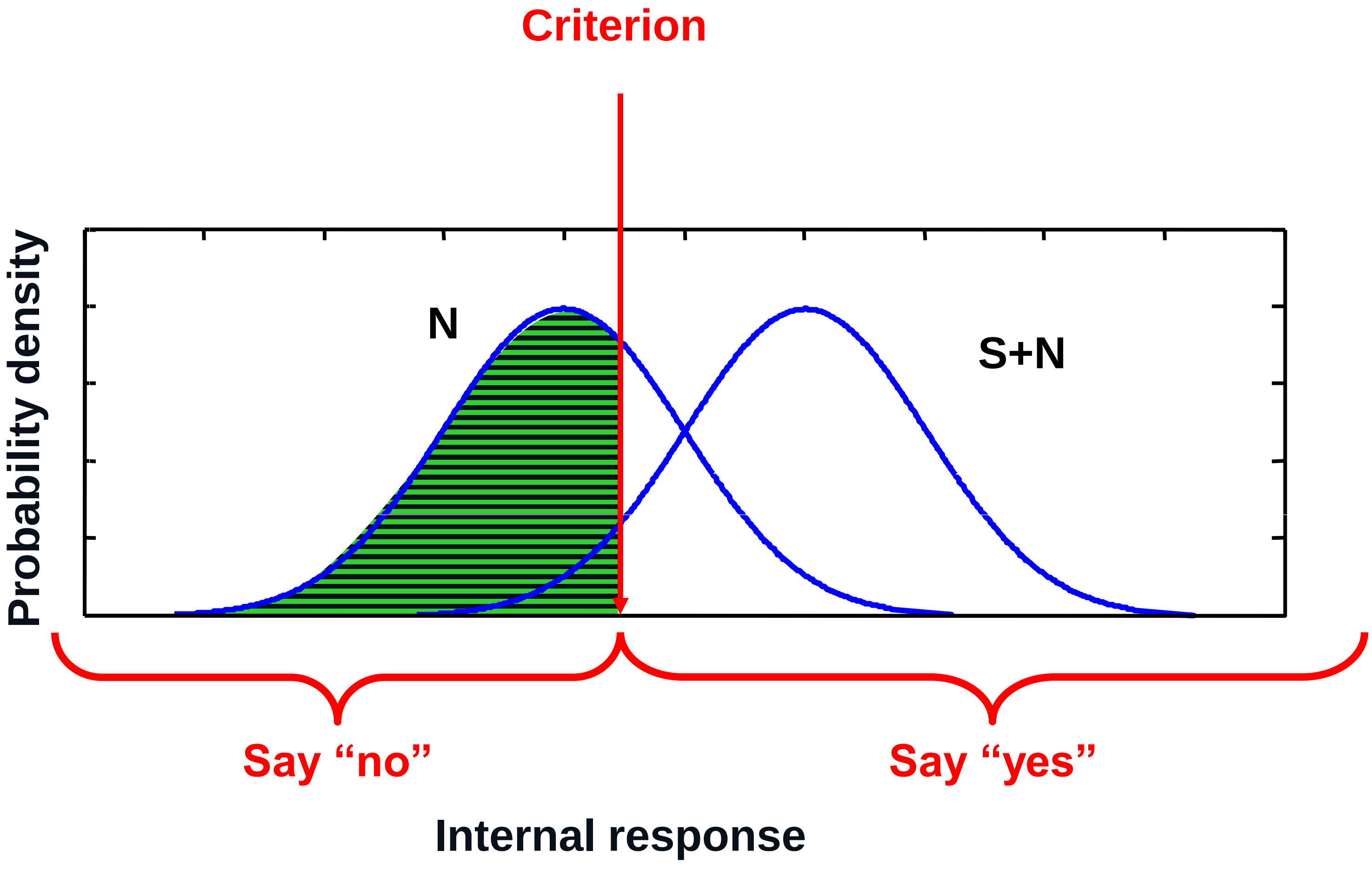
- *When there really is no effect, but the statistical test comes out significant by chance, you make a Type I error.*

- When H_o is true, the probability of making a Type I error is called ***alpha (α)***. This probability is the *significance level* associated with your statistical test.

"What Cold Possibly Go Worng?": Type I and Type II Errors

	Perception	
Reality	Yes	No
Yes / Exists (True)	Hit	Miss (Type II) *False Negative*
No / Does Not Exist (False)	False Alarm (Type I) *False Positive*	Correct Rejection

Correct rejects (1-beta)
(response “no” on no-signal trial)



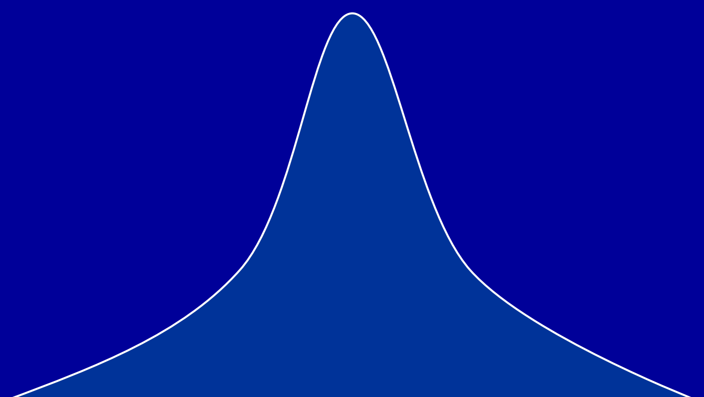


Figure 12-2: A 95% Confidence Interval, Part I

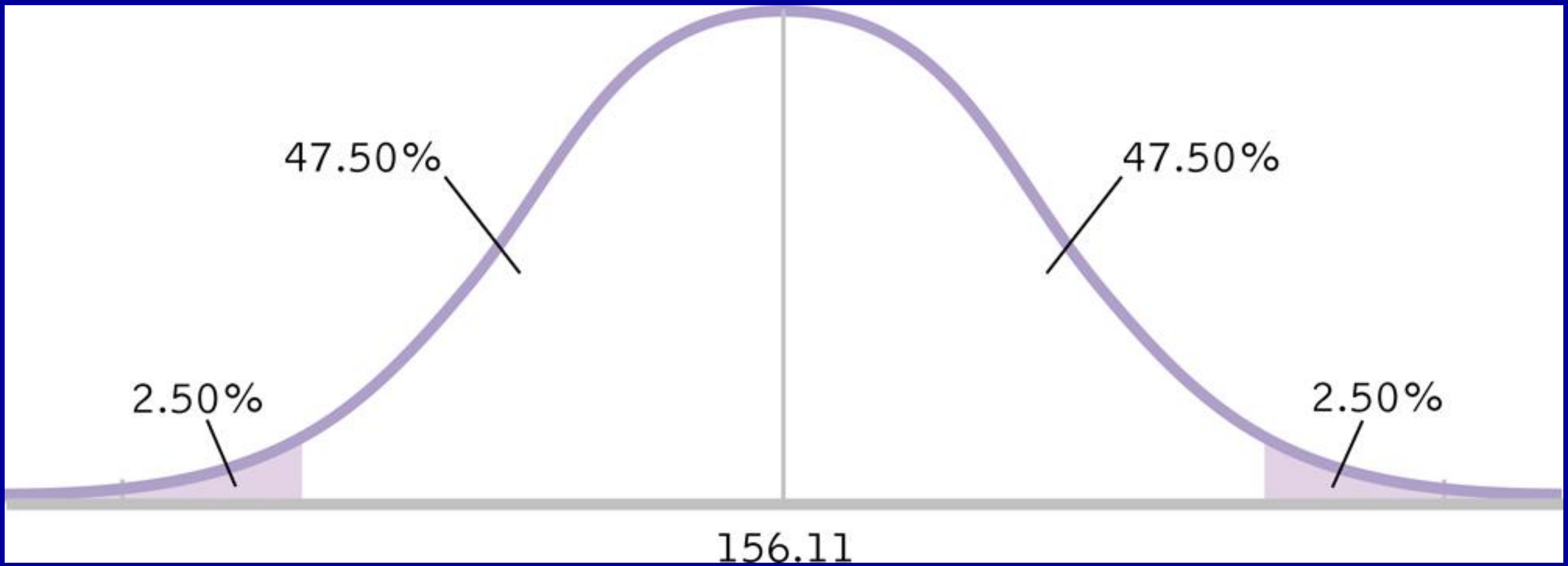


Figure 12-3: A 95% Confidence Interval, Part II

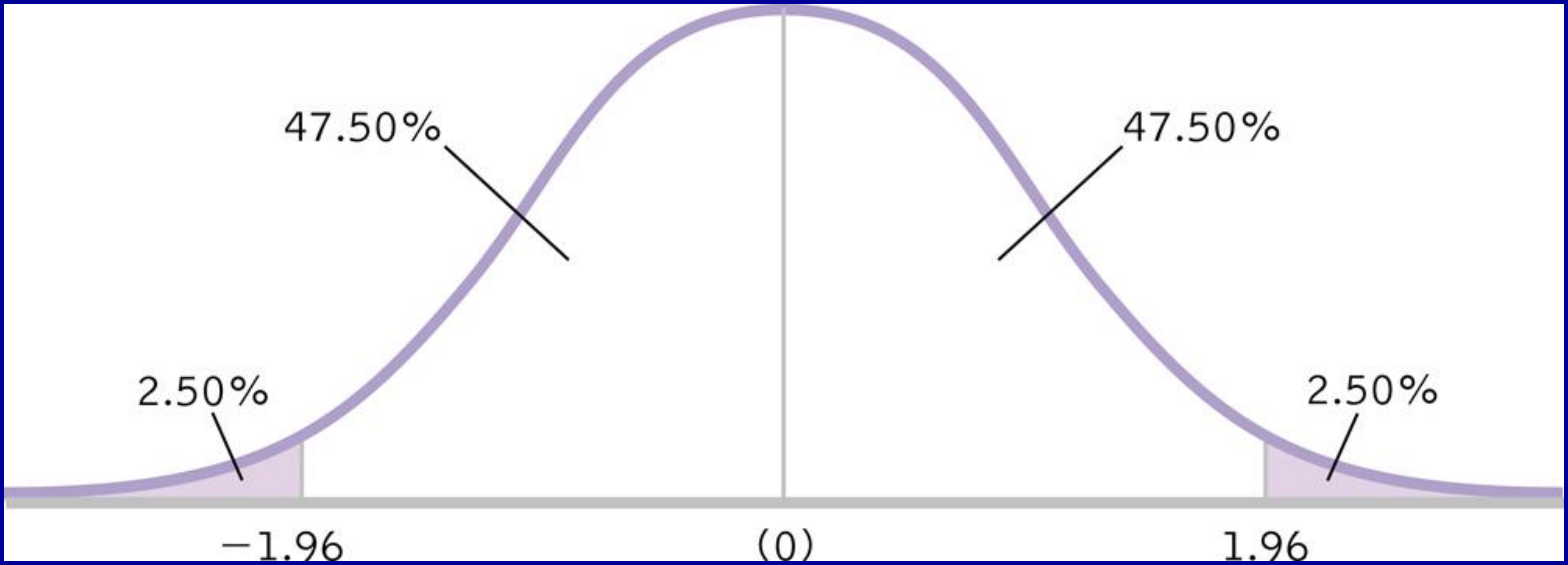
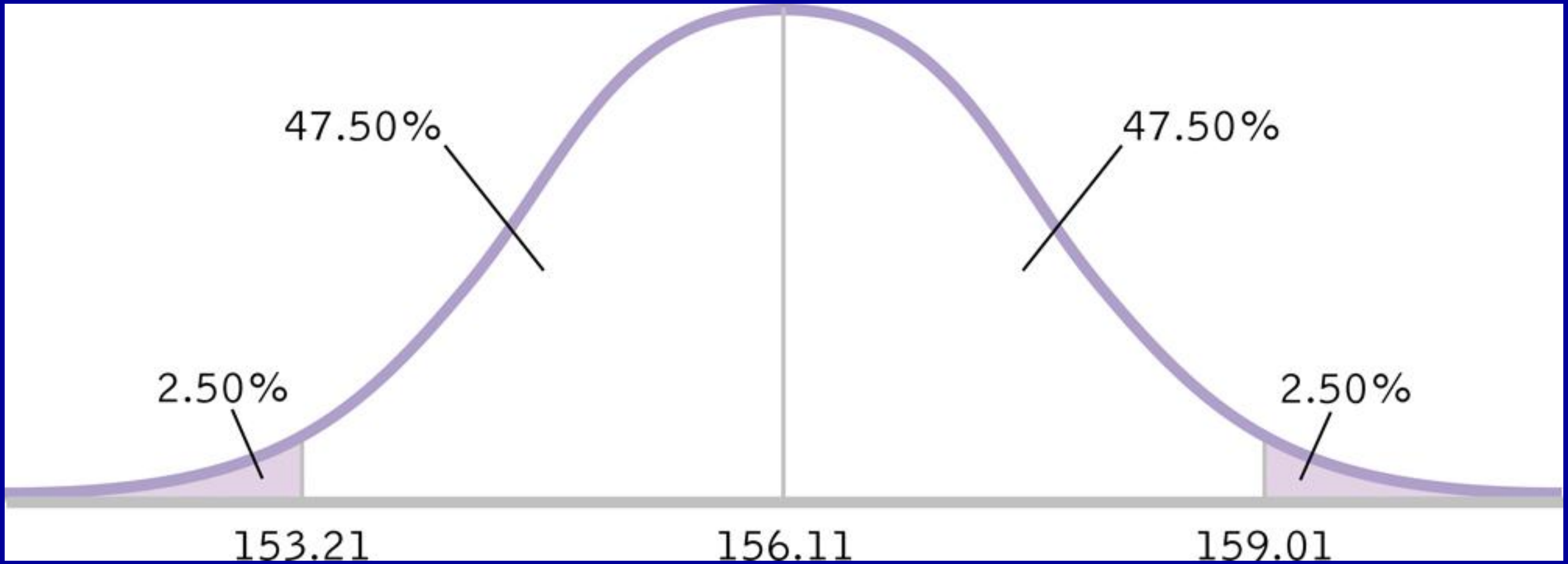


Figure 12-4: A 95% Confidence Interval, Part III



- **Kepler tried to record the paths of planets in the sky, Harvey to measure the flow of blood in the circulatory system, and chemists tried to produce pure knowing it was an element, though they failed at it. What was there approach to science, and what was the new perspective Pearson made, and what was the one crucial distinction he failed to make.** *Before Pearson, the things that science dealt with were real and palpable. Pearson proposed that the observable phenomena were only random reflections, what was real was the probability distribution. The real things of science were not things that we could observe but mathematical functions that described the randomness of what was observed. The four parameters of the distribution were what we were really interested in, but in the end we can never really determine them, we can only estimate them, and Pearson failed to understand that they could only be estimated.*

What Terms Calculate Statistical Power?

- The power of any test of statistical significance will be affected by four main parameters:
- the effect size
- the sample size (N)
- the alpha significance criterion (α)
- statistical power, or the chosen or implied beta (β)