Statistical Concepts for Clinical Research

- PJ Heagerty
- Department of Biostatistics
- University of Washington

Overall Focus

- **Ralph Waldo Emerson**
  - "The man who knows how will always have a job. The man who also knows why will always be his boss. As to methods there may be a million and then some, but principles are few. The man who grasps principles can successfully select his own methods."

Session Two Outline

- Hypothesis testing
- Null hypothesis / Alternative hypothesis
- P-value / statistical significance
- Use (overuse) of p-values / testing
- Multiple comparisons
- Inference for means

Our Small Surgical Trial

- Recall that yesterday we talked about a small ($n = 20$) hypothetical trial that compared surgical to non-surgical treatment.
- In our **randomized trial** we observed:
  - ▷ Mean outcome among **surgical**: 3.37
  - ▷ Mean outcome among **non-surgical**: 5.32
- **Q**: how strong is the evidence?
- **Q**: what uncertainty is associated with the observed surgical benefit of **3.37 - 5.32 = -1.95**?
- **Q**: Our trial was small, but was it **underpowered**?

Subject	**Randomized**	Observed		
i	Assignment	$Y_i(0)$	$Y_i(1)$	Difference
1	0	4.5		
2	1		1.0	
3	1		2.0	
4	1		2.2	
5	0	3.3		
6	1		0.8	
7	1		1.5	
8	0	4.9		
9	0	3.8		
10	0	3.6		

11	1		5.1	
12	0	6.7		
13	0	6.0		
14	0	5.6		
15	0	6.5		
16	1		6.0	
17	1		5.1	
18	0	8.3		
19	1		4.6	
20	1		5.3	
Mean		5.32	3.37	**-1.95**

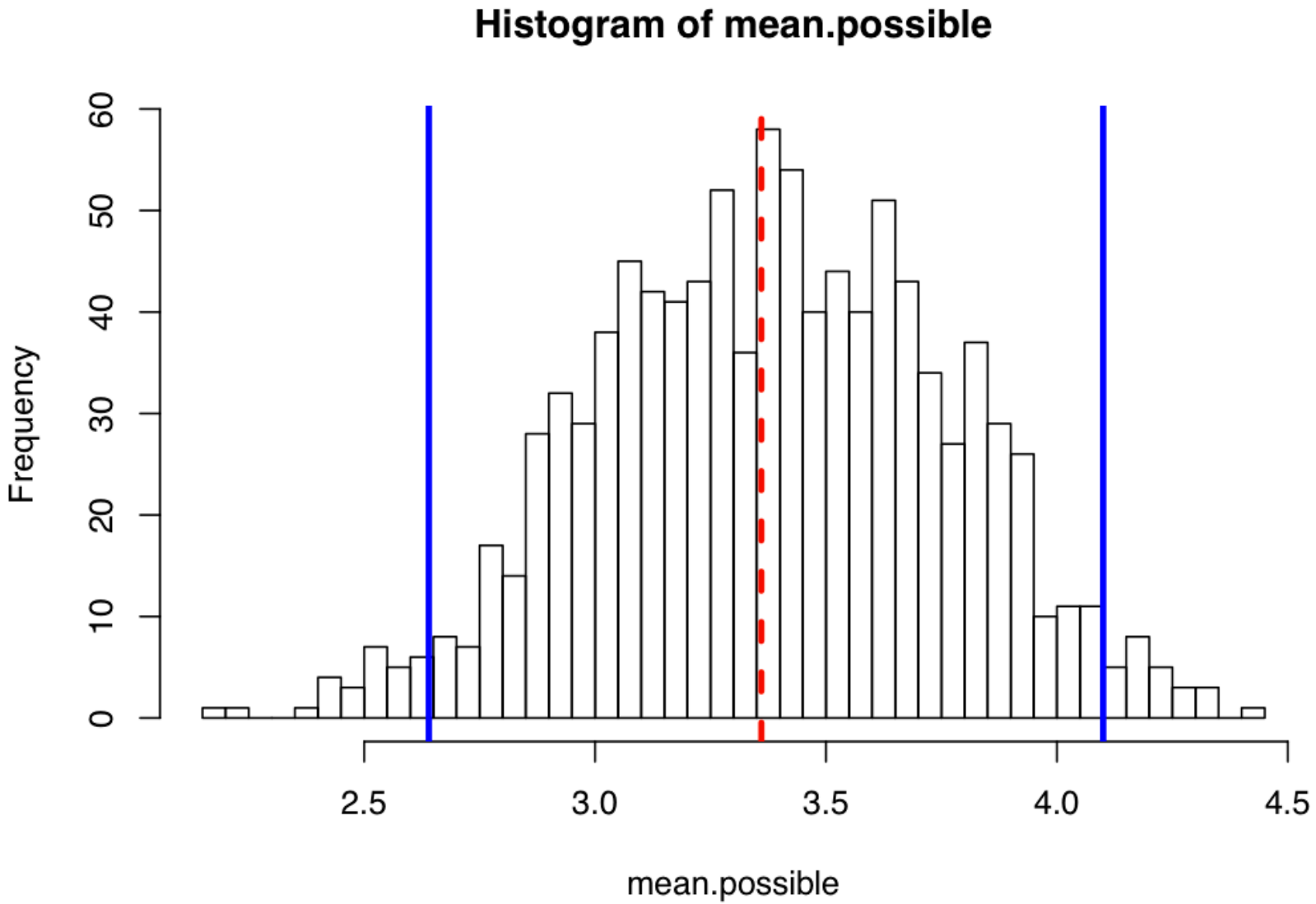
Principle: Variation Exists

- Different studies **will** give different results.
- We also assumed that we could have observed any of the 20 subjects after being treated surgically (and we listed their data).
- **Q**: What might we have seen if we had chosen different individuals for the treatment group (e.g. surgery)?

Subject	Potential Outcomes			
i	$Y_i(0)$	$Y_i(1)$	**Surg Sample 1**	**Surg Sample 2**
1	4.5	2.7	**2.7**	
2	3.1	1.0		**1.0**
3	3.9	2.0	**2.0**	**2.0**
4	4.3	2.2	**2.2**	
5	3.3	1.5		
6	3.3	0.8		**0.8**
7	4.0	1.5	**1.5**	**1.5**
8	4.9	3.2		
9	3.8	2.0	**2.0**	
10	3.6	2.0		**2.0**

Samples...

- Here we could actually just take lots and lots of possible samples and look at the results (next slide!).
- We also can use theoretical results based on probability to predict what the samples should look like.
  - Sample means, $\overline{Y}$ should vary around the **population** mean.
  - The **variability** of these sample means depends on both the intrinsic **person-to-person variability** and the **size of the sample**.


Histogram of mean.possible
Frequency
0
10
20
30
40
50
60
2.5
3.0
3.5
4.0
4.5
mean.possible

Samples...

- To create the previous figure I just fixed these 20 candidate subjects and simulated all possible samples of 10 subjects.
- **Principle**: we have to account for the uncertainty associated with the specific realized study by thinking about what **could** happen.

Samples...

- **Theory**: mathematical statistics can also predict what the sample means should do (assuming knowledge of mean and variance for population).
  - ▷ Sample means are "normally distributed"
  - ▷ Usually sample mean falls in between:

(**pop mean** - 2 · **SD**/$\sqrt{n}$, **pop mean** + 2 · **SD**/$\sqrt{n}$)

  - ▷ For our data the "population mean" is 3.39 (all 20)
  - ▷ For our study sample the "sample mean" is 3.32
  - ▷ **SD** is the "standard deviation", a measure of intrinsic person-to-person variability.

Normal Distribution is Important!

Samples...

- **Variability** exists.
- **Population** $\rightarrow$ **Samples**
  - If we **know** the population characteristics we can predict what samples should look like.
  - **Q**: Why is this helpful?
    - We can compare our DATA to what we'd expect to find if we assume something about the population – e.g. assuming there is no treatment effect. This is the essense of **hypothesis testing**.

Principle: Understand Expected Variation if no association is truth

- **Hypothesis testing**: "Innocent until proven guilty."
- Assume "no association" (null) unless enough evidence.
- **Q**: In my example surgical trial the difference in the means was **-1.95** suggesting a meaningful benefit associated with surgery – have I proven that it works?
- **A**: Not yet. We can try to overturn the assumption of no benefit by asking how likely is it that a study with only 20 patients yields an estimate of **-1.95** or better?

Prediction of Data if No Association

- **Null hypothesis** = no association.
- **Q**: How can we use these data to predict what would be EXPECTED if the null hypothesis were true?
  - ▷ Adopt a **model** that has equal distributions for each treatment group. (t-test)
  - ▷ Use either simulations using the null **model**, or mathematical results to predict the distribution for the difference in means.
  - ▷ Use **permutation** test (easy idea!)

Permutation Test

- **Idea**: if the null is true (e.g. no difference across groups) then there is no real link between the assignment to treatment group and the outcome.
- **Plan**: so permute (scramble, shuffle) the group in which you put an individual's data.
- **Expectation**: by permuting the group you are showing what results you might have gotten if people were otherwise put in the other treatment group, **and** assuming there is no difference across treatment groups.

Subject	**Randomized**	Observed		
i	Assignment	$Y_i(0)$	$Y_i(1)$	Difference
1	0	4.5		
2	1		1.0	
3	1		2.0	
4	1		2.2	
5	0	3.3		
6	1		0.8	
7	1		1.5	
8	0	4.9		
9	0	3.8		
10	0	3.6		

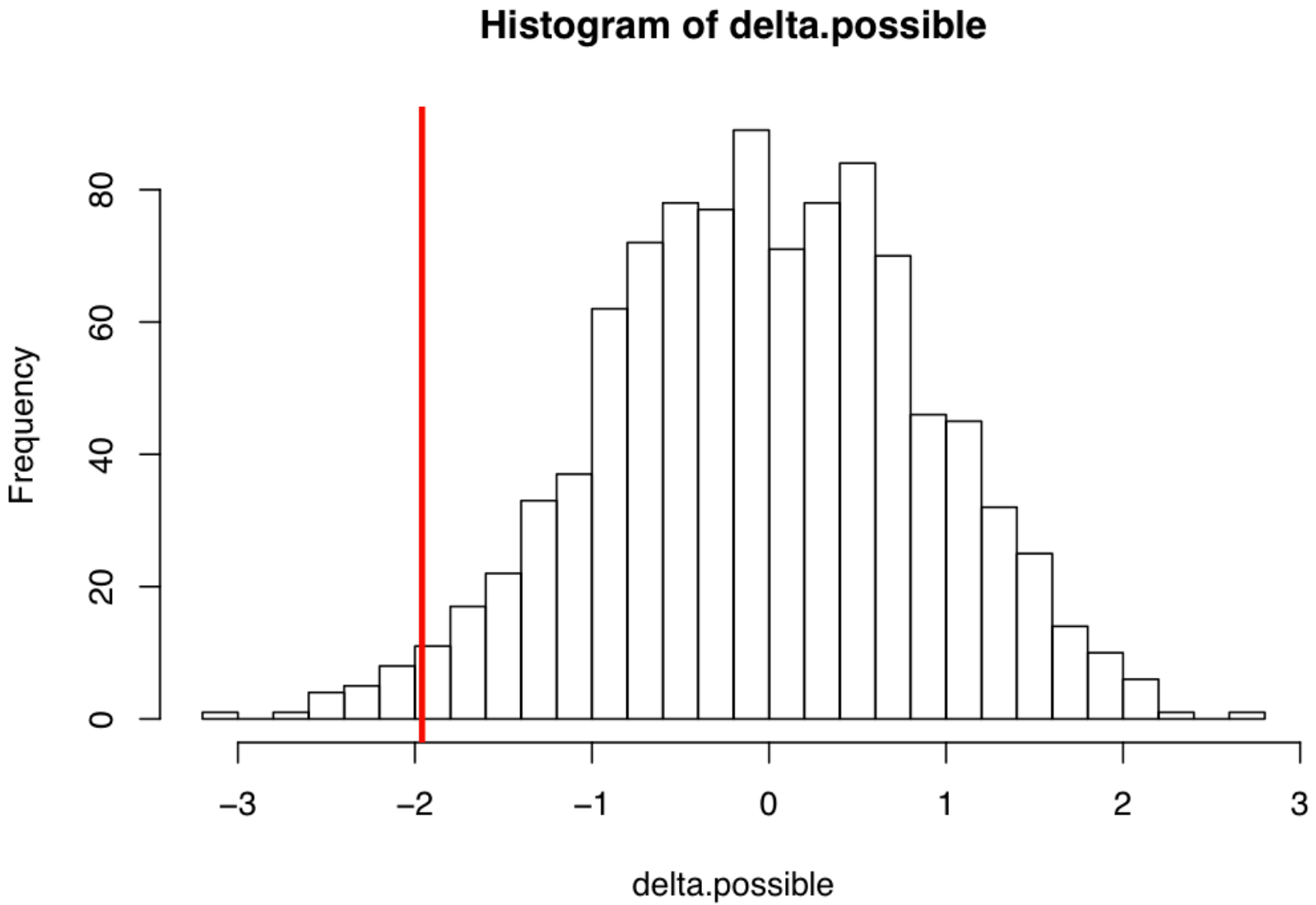
11	1		5.1	
12	0	6.7		
13	0	6.0		
14	0	5.6		
15	0	6.5		
16	1		6.0	
17	1		5.1	
18	0	8.3		
19	1		4.6	
20	1		5.3	
Mean		5.32	3.37	**-1.95**

Subject i	**Randomized** Assignment	Observed Y_i	**Permute 1**	**Permute 2**
1	0	4.5	**1**	**1**
2	1	1.0	**1**	**0**
3	1	2.0	**1**	**1**
4	1	2.2	**1**	**1**
5	0	3.3	**0**	**1**
6	1	0.8	**0**	**0**
7	1	1.5	**0**	**0**
8	0	4.9	**0**	**0**
9	0	3.8	**1**	**1**
10	0	3.6	**1**	**1**

	11	1	5.1	**1**	**1**
	12	0	6.7	**0**	**1**
	13	0	6.0	**0**	**0**
	14	0	5.6	**1**	**1**
	15	0	6.5	**0**	**0**
	16	1	6.0	**0**	**0**
	17	1	5.1	**1**	**1**
	18	0	8.3	**0**	**0**
	19	1	4.6	**0**	**0**
	20	1	5.3	**1**	**0**
Difference			-1.95	**-1.04**	**-0.30**

Permutation Test

- **Calculate**: the results of your experiment if other group assignments were made (but you keep the individual outcomes fixed).
  - ▷ You now have a picture of what results should happen if NO ASSOCIATION.
- **Show**: BUT now look to see where the result you ACTUALLY DID GET falls in relationship to what you would have expected...


Histogram of delta.possible
Frequency
0
20
40
60
80
−3
−2
−1
0
1
2
3
delta.possible

Permutation Test

- **Note**: If NO ASSOCIATION then you'd get results like your observed results (or farther away from zero) only 3% of the time!
- **Conclusion Options**:
  - ▷ There is NO ASSOCIATION and your study results are strange.
  - ▷ You results are sufficiently inconsistent with what would be expected if there is no association that you conclude there IS ASSOCIATION between surgery and the outcome.

Hypothesis Test

- **p-value**: the calculation above – the probability of getting results as or more "strange" than your OBSERVED results if the NULL were true.
- For a single test we usually ask that your p-value is **less than 5%** in order to reject the null hypothesis.
- **Summary**
  - ▷ Method to suggest EXPECTED if NULL is true.
  - ▷ Compare your OBSERVED result to the EXPECTED results above.
  - ▷ REJECT NULL if your results are unlikely.

Permutation Test and t-test

- The **t-test** is a similar idea to the permutation test, but requires specific mathematical **models** be used to compute the p-value.
- For our example:

```
Two Sample t-test

data:  y[tx == 0] and y[tx == 1]
t = 2.4001, df = 17.047, p-value = 0.02809
95 percent confidence interval:
 0.2374206 3.6825794
```

Some Comments on Tests and p-values

- Common interpretations:

$p \leq 0.05$	significant	difference across groups
$p > 0.05$	non-significant	no difference

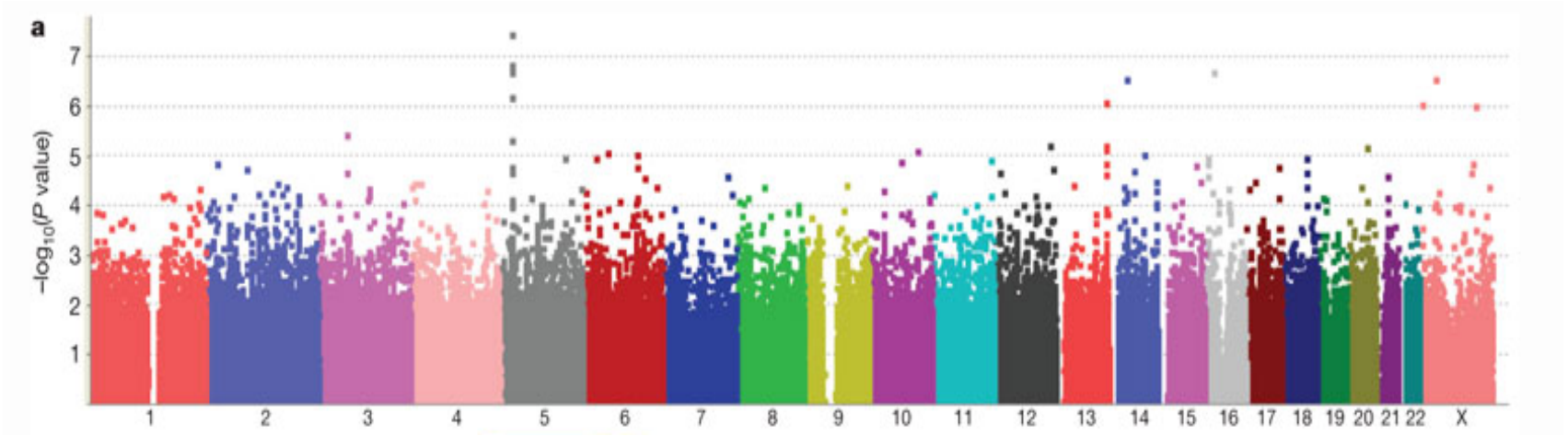
- When p>0.05 we don't have enough evidence to overturn the null hypothesis. And, null hypothesis says "no difference".
- BUT – we usually do have SOME difference observed in the data, and it may be a matter of **too small** of a difference to make us conclude against the null.

Multiple Tests...

- Using a p-value threshold of 0.05 controls the frequency of **false positive** statements. If the null is true then you'll only reject it 5% of the time.
- But with multiple tests...

N Tests	Expected False Positives
1	0.05
10	0.5
100	5
1000	50

Genome-wide Association Studies (Wang 2009) *Nature* – autism



- With $1,000,000$ markers use p$= 0.05/1,000,000$ to control error rate.
- **multiple comparisons** correction.

Some Comments on Tests and p-values

- Use of p-values to interpret your data is a huge simplification of the story. (e.g. YES/NO significant)
- More information is provided by stating the MAGNITUDE of the difference that is observed, interpreting this effect **clinically**, and providing some measure of uncertainty around the estimate.
- **confidence interval**

Example: Inferential Studies

	6 months				12 months			
	Surgical (n=50)	Non-surgical (n=54)	Treatment effect* (95% CI)	p value	Surgical (n=49)	Non-surgical (n=52)	Treatment effect* (95% CI)	p value
Primary outcome								
CTSAQ function (1–5)	1·91(0·88)	2·44 (0·87)	0·46 (0·20 to 0·72)	0·0006	1·74 (0·79)	2·17 (0·96)	0·40 (0·11 to 0·70)	0·0081
Secondary outcomes								
CTSAQ symptom (1–5)	2·02 (1·03)	2·42 (0·80)	0·42 (0·07 to 0·77)	0·0181	1·74 (0·76)	2·07 (0·88)	0·34 (0·02 to 0·65)	0·0357
Days of reduced work or housework (0–28)	4·3 (8·8)	6·3 (9·4)	2·3 (–1·0 to 5·6)	0·1638	2·2 (5·6)	5·2 (8·8)	2·7 (–0·0 to 5·4)	0·0524
Days of lost work (0–28)	0·7 (4·3)	2·8 (8·1)	2·3 (–0·6 to 5·3)	0·1174	0·1 (0·7)	1·6 (5·8)	0·9 (–0·4 to 2·3)	0·1641
Pain intensity (0–10)	4·7 (3·2)	5·7 (3·1)	1·0 (–0·2 to 2·1)	0·0993	3·5 (3·0)	4·3 (3·3)	0·9 (–0·3 to 2·1)	0·1590
Pain interference (0–10)	2·8 (3·0)	3·4 (3·2)	0·1 (–0·8 to 1·1)	0·8068	2·1 (6·9)	3·1 (3·3)	0·6 (–0·3 to 1·6)	0·1957
SF-36 (version 2.0)† PCS	39 (12)	37 (11)	1·5 (–1·7 to 4·7)	0·3608	39 (14)	37 (12)	1·6 (–2·8 to 6·0)	0·4762
SF-36 (version 2.0)† MCS	47 (16)	47 (14)	0·9 (–3·5 to 5·4)	0·6833	45 (15)	47 (15)	–0·5 (–6·0 to 5·0)	0·8520

Data are mean (SD) unless otherwise stated. CTSAQ=Carpal Tunnel Syndrome Assessment Questionnaire. SF-36=short-form-36. PCS=physical component summary. MCS=mental component summary. *Treatment effect indicates the difference between surgical and non-surgical groups on the outcome measure at 6 and 12 months, based on ANCOVA adjusted for the baseline value of the outcome measure and treatment site. A positive effect indicates that patients assigned to surgery had better outcomes than did those assigned to non-surgical care. †SF-36 PCS and MCS scores are norm-based with mean (SD) of 50 (10) in a healthy population. Higher scores indicate better function.

***Table 2:* Primary and secondary outcomes and adjusted treatment effect at 6 and 12 months by randomised treatment assignment (intention-to-treat comparisons)**

Principle: Quantify Uncertainty

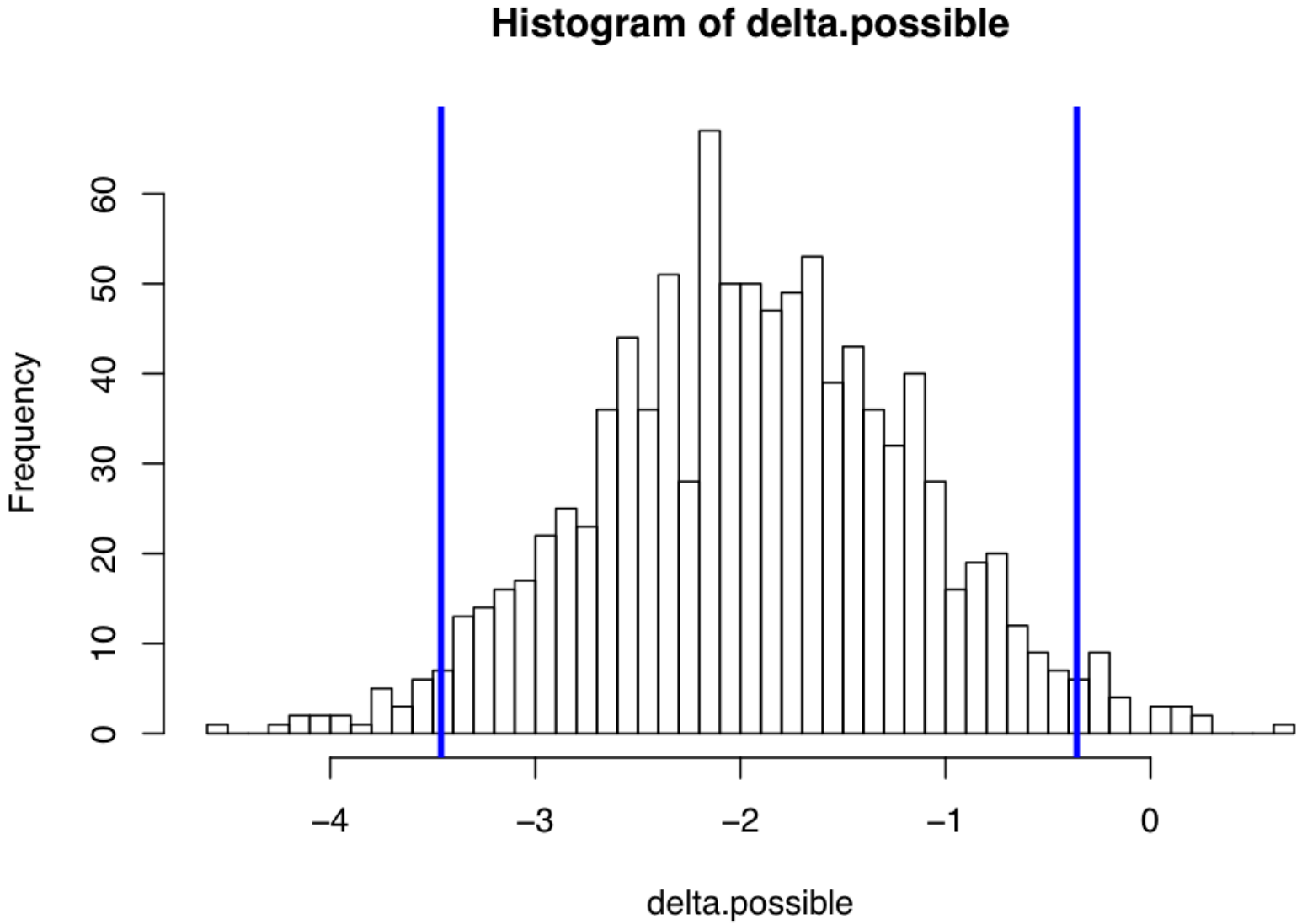
- In our small study we observed a difference of **-1.95** comparing surgical to non-surgical treatment.
- **Q**: What do we think might happen with similar replication studies?
- **Q**: What magnitudes of surgical benefit are supported by our data? Might the benefit actually be as high as -3.0 units?
- **Objective**: provide an INTERVAL of values for the surgical benefit (effect) that we think are plausible based on our data.

Idea: Simulate Trials

- Use our data to anchor some simulations.
- We can use our own data to create a POPULATION and then show the uncertainty associated with a small SAMPLE of only 20 people.
- **Idea**:
  - ▷ **Treated Population**: create a group with 1,000 copies of each treated person we studied.
  - ▷ **Possible Samples**: draw 10 people from this "population" (repeat!)

Idea: Simulate Trials

- **Idea (continued)**:
  - ▷ **Control Population**: create a group with 1,000 copies of each non-surgical person we studied.
  - ▷ **Possible Samples**: draw 10 people from this "population" (repeat!)
- Notice that we are now simulating data where we assume the true relationship to be whatever we've seen in our data – no longer assuming the null hypothesis – like we did with testing.


Histogram of delta.possible
Frequency
0
10
20
30
40
50
60
−4
−3
−2
−1
0
delta.possible

Confidence Interval

- We can then calculate all possible studies that would mirror our actual study and plot the results.
- We can then use the middle 95% of simulated study results to reflect our expectations of what could have happened in a study similar to the one we conducted.
- Additional mathematical results also tell us that this interval represents those values for the true surgical benefit that are plausible based on our data.
- **95% confidence interval**: (-3.39, -0.47)
- **Method** = Bootstrap confidence interval.

Confidence Interval

- **estimate**: -1.95
- **95% confidence interval**: (-3.39, -0.47)
- **Q**: what does this interval tell us?
  - ▷ It's possible that the benefit is as large as 3.0 units or as small as 0.5 units.
  - ▷ We can rule out no benefit since 0.0 isn't supported. That's perfectly consistent with our test and p-value of p=0.03.

Example: Confidence Intervals

	6 months				12 months			
	Surgical (n=50)	Non-surgical (n=54)	Treatment effect* (95% CI)	p value	Surgical (n=49)	Non-surgical (n=52)	Treatment effect* (95% CI)	p value
Primary outcome								
CTSAQ function (1–5)	1·91(0·88)	2·44 (0·87)	0·46 (0·20 to 0·72)	0·0006	1·74 (0·79)	2·17 (0·96)	0·40 (0·11 to 0·70)	0·0081
Secondary outcomes								
CTSAQ symptom (1–5)	2·02 (1·03)	2·42 (0·80)	0·42 (0·07 to 0·77)	0·0181	1·74 (0·76)	2·07 (0·88)	0·34 (0·02 to 0·65)	0·0357
Days of reduced work or housework (0–28)	4·3 (8·8)	6·3 (9·4)	2·3 (–1·0 to 5·6)	0·1638	2·2 (5·6)	5·2 (8·8)	2·7 (–0·0 to 5·4)	0·0524
Days of lost work (0–28)	0·7 (4·3)	2·8 (8·1)	2·3 (–0·6 to 5·3)	0·1174	0·1 (0·7)	1·6 (5·8)	0·9 (–0·4 to 2·3)	0·1641
Pain intensity (0–10)	4·7 (3·2)	5·7 (3·1)	1·0 (–0·2 to 2·1)	0·0993	3·5 (3·0)	4·3 (3·3)	0·9 (–0·3 to 2·1)	0·1590
Pain interference (0–10)	2·8 (3·0)	3·4 (3·2)	0·1 (–0·8 to 1·1)	0·8068	2·1 (6·9)	3·1 (3·3)	0·6 (–0·3 to 1·6)	0·1957
SF-36 (version 2.0)† PCS	39 (12)	37 (11)	1·5 (–1·7 to 4·7)	0·3608	39 (14)	37 (12)	1·6 (–2·8 to 6·0)	0·4762
SF-36 (version 2.0)† MCS	47 (16)	47 (14)	0·9 (–3·5 to 5·4)	0·6833	45 (15)	47 (15)	–0·5 (–6·0 to 5·0)	0·8520

Data are mean (SD) unless otherwise stated. CTSAQ=Carpal Tunnel Syndrome Assessment Questionnaire. SF-36=short-form-36. PCS=physical component summary. MCS=mental component summary. *Treatment effect indicates the difference between surgical and non-surgical groups on the outcome measure at 6 and 12 months, based on ANCOVA adjusted for the baseline value of the outcome measure and treatment site. A positive effect indicates that patients assigned to surgery had better outcomes than did those assigned to non-surgical care. †SF-36 PCS and MCS scores are norm-based with mean (SD) of 50 (10) in a healthy population. Higher scores indicate better function.

***Table 2:* Primary and secondary outcomes and adjusted treatment effect at 6 and 12 months by randomised treatment assignment (intention-to-treat comparisons)**

Summary

- We have seen that with knowledge of a population we can describe how variable we expect samples to be.
- **Population** $\rightarrow$ **Sample**
- We have used this to provide a set of reference results if the null hypothesis were actually true:
- **Null Hypothesis** $\rightarrow$ **Expected Samples**
- A **p-value** reflected where our OBSERVED results stand in relation to results that are expected under the null.

Summary

- We have used sampling knowledge again but then moved away from assuming the null and simply assumed a "truth" that comes from our data.
- **Est. Population** $\rightarrow$ **Expected Samples**
- Using this idea we could create a range of estimates that are consistent with our data – **confidence interval**.
- Simulations require computing but no mathematics!