Biostatistics

Mithat Gonen

Brenden Bready

Today's Lecture

- Review of The First Two Lectures
- New Material: Type II errors and power
- If we have time → Goodman article

Populations → Samples
Parameters → Estimates

- We need to recognize that our data is a sample from some populations
- It may not be (most likely is not) a random sample
- The population may not be so easy to define but it is there, at least conceptually
- Parameters are population quantities; samples give us estimates of parameters
- Many many concepts in statistics depend on this duality between population and samples

Assuming we agreed on a population

- And we were able to obtain a sample …
- Our conceptual problems have not ended
- There is almost never a random sample
- What we can hope for is a representative sample

Example: Phase II Trial

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Long-Term Follow-up of CD19 CAR Therapy in Acute Lymphoblastic Leukemia

Jae H. Park, M.D., Isabelle Rivière, Ph.D., Mithat Gonen, Ph.D., Xiuyan Wang, Ph.D., Brigitte Sénéchal, Ph.D., Kevin J. Curran, M.D., Craig Sauter, M.D., Yongzeng Wang, Ph.D., Bianca Santomasso, M.D., Ph.D., Elena Mead, M.D., Mikhail Roshal, M.D., Peter Maslak, M.D., Marco Davila, M.D., Ph.D., Renier J. Brentjens, M.D., Ph.D., and Michel Sadelain, M.D., Ph.D.

Population & Parameter

- CD19+ B-cell acute lymphoblastic leukemia (ALL)
- Relapsed of refractory disease
- Primary endpoint: response (complete remission)
- What is our parameter?
- Response rate: P(Response = 1) = r

Things we can do

- Point estimate: produce a single number that represents our best guess at what the parameter value might be
- Interval estimate: produce an interval that is likely to contain the true value of the parameter
- Hypothesis testing: produce a yes/no answer to question about r (such as $r<=r_0$ vs $r>r_0$ where r_0 is a pre-specified number)

Point Estimate

- Most of the time there is a sample analog of the population definition
- r is the proportion of responders in the population; can we use the proportion of responders in the sample to estimate r?
- Yes!
- Sometimes sample analogs are not great estimates, but we will ignore that for now (famous example: standard deviation)

Back to the Example

- 44/53 (= 0.83) responded
- We often say response rate is 83%
- Any time you hear this you should think in your mind “Our point estimate for response rate in this data set is 83%”
- The true response rate in the population is very unlikely to be exactly 83% but we hope it is close
- It will be close if we did our homework: good sampling, good data collection and good statistical analysis

Why is the parameter not 83%

- Imagine we repeated the study, same inclusion/exclusion criteria, same everything but different individuals enrolling.
- It would be possible but unlikely to get 44 responders again.
- Imagine we repeated the study 100 times. Many of these would not have 44 responders.
- So 44 responders and 83% is nothing special. It is somewhere in the vicinity of the right answer but it is not the right answer. Each repeated study will give a slightly different answer.

What then?

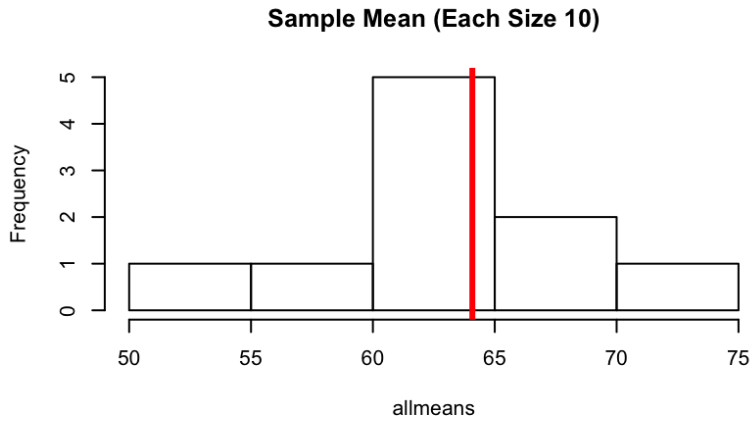
- Interval estimate: Can we produce an interval that is likely to contain the true value?
- Go back to imagining the repeated studies
- What if there is a way to say: here is a formula to produce an interval estimate from a given data set; do it for each of the 100 repeats and obtain 100 interval estimates. 95% of these intervals will contain the true value
- You have gotten yourself a confidence interval

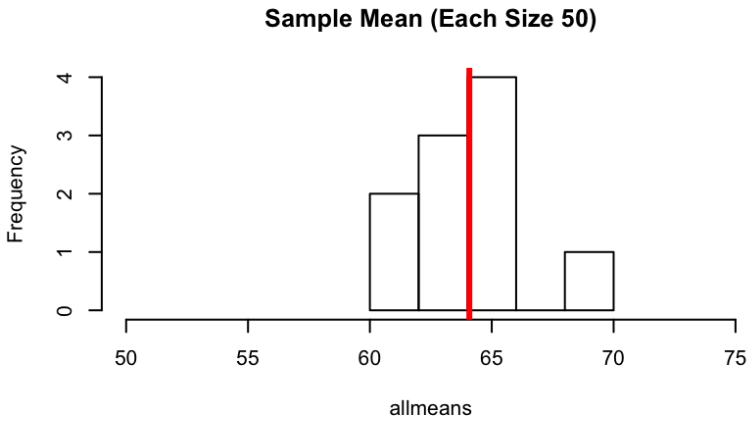
Back to the Example

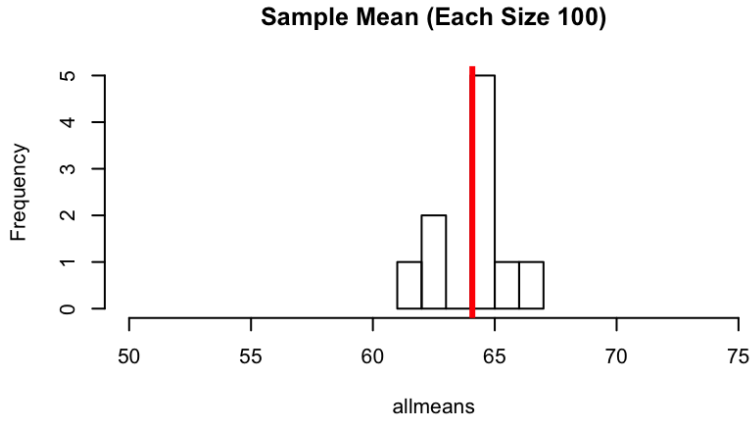
- 44 out of 53 → 95% confidence interval: 70% - 92%
- What is the interpretation?
- There is a 95% chance that the true parameter value is between 70% and 92%?
- 95% of the intervals produced this way will contain the true value of the parameter
- Is this helpful? Maybe.

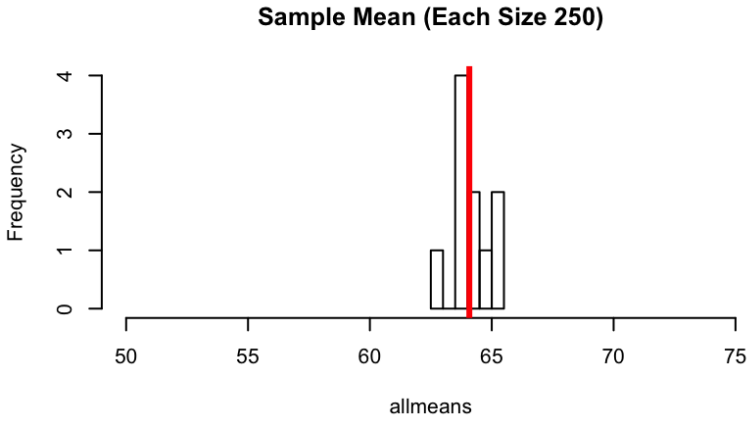
How is it helpful?

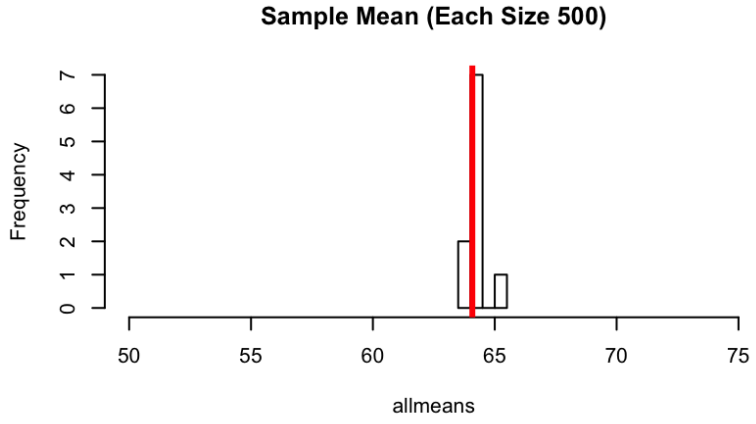
- Precise probabilistic interpretation is cumbersome
- But points out to why this is useful
- If most of the intervals will contain the true value, a single randomly selected one of them is likely to contain the true value
- Confidence intervals are a bridge between point estimation and hypothesis testing
- Single most underused statistical tool

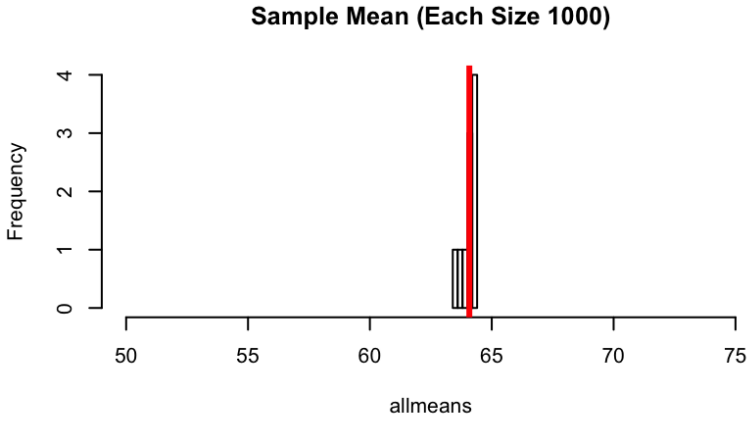

Sample Mean (Each Size 10)
Frequency
0
1
2
3
4
5
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 50)
Frequency
0
1
2
3
4
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 100)
Frequency
0
1
2
3
4
5
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 250)
Frequency
0
1
2
3
4
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 500)
Frequency
0
1
2
3
4
5
6
7
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 1000)
Frequency
0
1
2
3
4
50
55
60
65
70
75
allmeans

Why is this important?

- Look at the previous figure. For all of them the red line is in the middle.
- So regardless of the sample size the distribution of the sample mean is centered around the (true) population mean
- But in real like we will have only one sample, so this is nice but kind of useless

The devil is in the tails

- If I will not give everyone a sample, but instead give only one sample for everyone in the class to use …
- Would you want a size of 10, 50, …, 1000?
- Why?
- If sample size is 10, which sample I get matters quite a lot. My estimate of the mean can be 54 or 70
- If my sample size is 100, it matters less (between 61 and 70) but it still does a little
- If my sample size is 1000, all the samples have a mean between 63.4 and 64.3

Summary of what we did

- We obtained multiple samples from a population
- Calculated the sample mean for each sample
- We looked at the distribution (via a histogram) of the means in a sample
- We saw that each sample size gets larger, the means got concentrated around the population mean
- This is called the law of large numbers

Why is it important?

- It provides a justification for the intuitive thought that large sample sizes are better
- In practice we will be able to observe only one sample. If our sample size is large, sample mean does not vary too much from one sample to the other, hence we can rest assured having observed a single sample is OK. If we had gotten another sample its sample mean would be very close to the first one anyway
- But in a small sample, sample-to-sample variability is substantial. With one sample we can be really off.

Some generalizations

- Instead of sample mean you say “estimate” and this statement will be true
- In fact this is one definition of a good estimate (does it satisfy the law of large numbers?)
- So this idea is not limited to means

Review of Hypothesis Testing

- Suppose we are designing a Phase II study
- Already decided that we will do a single-arm study and use response rate as the primary endpoint
- We think response rate for the standard of care in this population is 50% and our regimen needs to exceed this

Hypothesis Testing

- We can formulate this as follows
  - Hypothesis 0: Response rate is less than or equal to 50%
  - Hypothesis 1: Response rate greater than 50%
- Is the response rate referred to in the hypothesis statement a parameter or an estimate?

- We will use the data to discard one of these and retain the other

Type I and Type II Errors

- There are four things that can happen when we test a hypothesis
  - H0 is true. We discard H1 and retain H0. Correct decision (True Negative).
  - H0 is true. We discard H0 and retain H1. Type I Error (False Positive)
  - H1 is true. We discard H1 and retain H0. Type II Error (False Negative)
  - H1 is true. We discard H0 and retain H1. Correct decision (True Positive).
- It would be good if we have a method to test these hypothesis while controlling the probability of making the errors.

Neyman-Pearson Procedure

- Choose a Type I Error (almost always 0.05)
- Find yourself a test statistic. It will be a function of the data and the sample size. In other words, once I give you a data set you should be able to calculate the value of the test statistic from the data set. Ideally, the values the test statistic takes would be very different when H0 is true from when H1 is true.
  - There will be many test statistics that will be reasonable to use.
- Calculate the value of the test statistic in the data set at hand.

Neyman-Pearson Procedure (cont'd)

- Figure out what values would you have gotten if H0 is true.
  - You can represent these values with a distribution.
- Figure out which value of the test statistic would have given you a 5% Type I Error. This is your reference value.
- If the value of the test statistic exceeds the reference value then your test statistic is unusual with respect to this reference distribution, H0 is unlikely to be true. Reject it, retain H1.
- Else retain H0.

Let’s Work This Out

- What is a good test statistics to use for testing
  - Hypothesis 0: Response rate is less than or equal to 50%
  - Hypothesis 1: Response rate greater than 50%
- There is something called a Wald statistic; widely applicable
  - Numerator = Estimate of the effect size
  - Denominator Uncertainty of the estimate of the effect size
- In our example
  - Numerator = Estimate of response rate – 0.5
  - Denominator = Square root of Data Variance divided by the sample size

In this specific example

- Denominator = sqrt(0.5*(1-0.5)/n)
  - 0.5 is the response rate when H0 is true
  - n is the sample size
- If we have 36 patients and 27 of them respond
  - Numerator = (27/36) - 0.5 = 0.75 – 0.5 = 0.25
  - Denominator = sqrt(0.5*0.5/36) = 0.0833
  - Value of the test statistic = 3

What is the reference distribution?

- It turns out that all Wald test statistics have the same reference distribution, “approximately”
  - Approximately is very critical. The approximation is better when n is large. How large depends on the actual test.
- Normal distribution with mean 0 and variance 1 (standard normal)
  - We did not exactly learn what this is but it is the usual bell shaped curve
- What is the reference value? There are tables for this but you can also use a software or ask Google or chatGPT
  - 1.645

Looks pretty unusual

- 3 > 1.645 → Reject H0.
- What is we had observed 21 responses instead of 27?
- Numerator = 21/36 – 0.5 = 0.0833
- Denominator stays the same at 0.0833
- Test statistic = 1
- 1 < 1.645 → Retain H0.

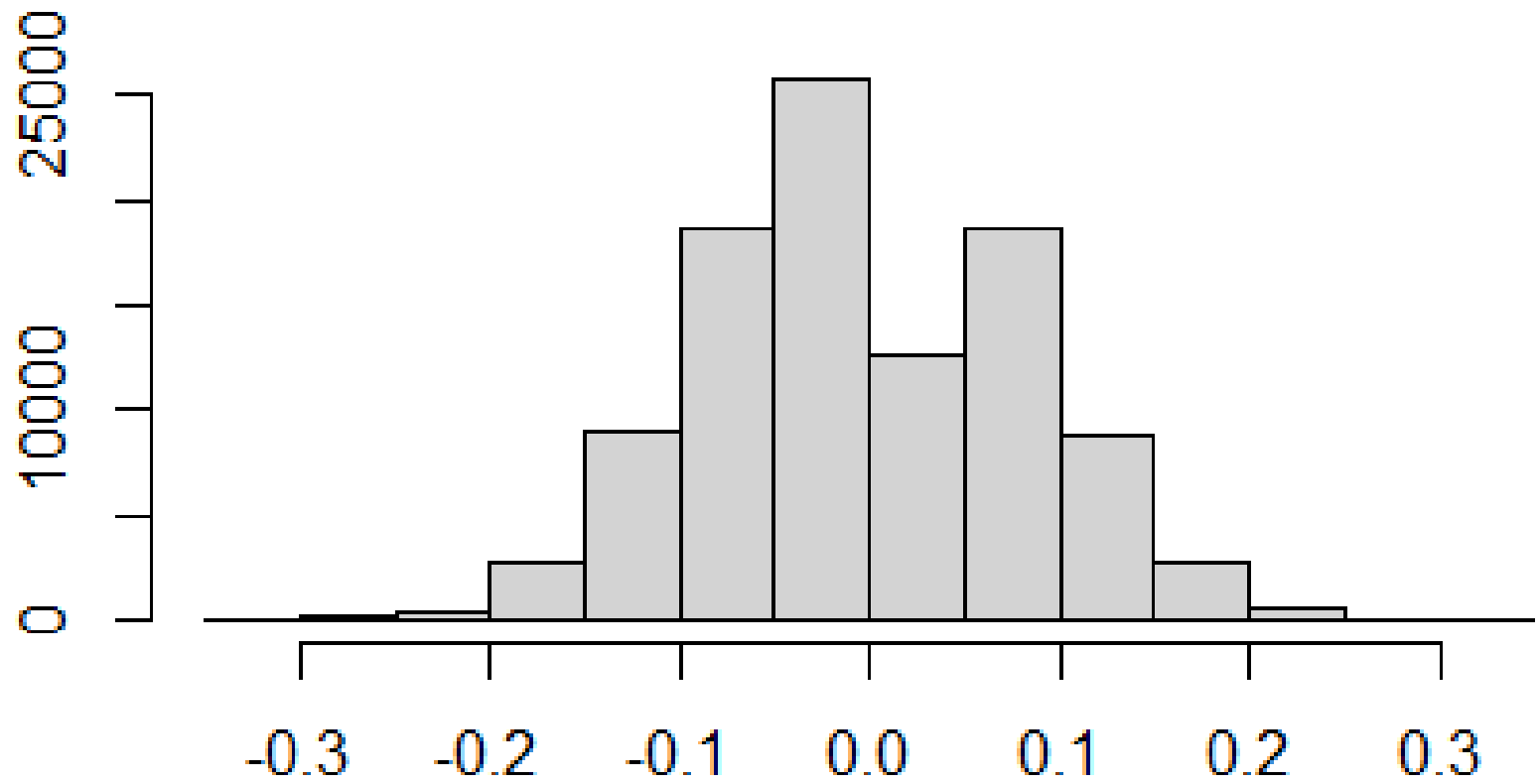
What if?

Number of Responses	21	22	23	24	25	26	27
Test Statistic	1	1.33	1.67	2.00	2.33	2.67	3

- 23 or more responses, reject H0; 22 or fewer retain it
  - This is how we derive the decision rule in a Simon's design
- Test statistic increases linearly with the number of responses
  - This is because the denominator stays the same
  - Only in this rather simple (but quite useful) case of distribution with one parameter
  - In most other cases the denominator will change when the data changes

More on the reference distribution

- What the test statistic would look like if the null distribution is true?
- This has a “parallel universe” interpretation
- Suppose the true response rate is 0.5 and you repeat the trial, get a certain amount of responses and calculate the test statistic.
- Do it over and over a very large number of times
- Do a histogram of the test statistics you get --- and imagine its smooth version
- This is your reference distribution


0
10000
25000
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3

Key Points

- What does the parallel universe ensure?
  - If H0 is true then this is what we expect to see
- What does 1.645 ensure?
  - If H0 is true then the value of the test statistic would exceed 1.645 only 5% of the time
- These two points together means Type I Error is controlled at 5%

Remarks

- What about the Type II error?
  - This procedure has no assurance that we have a small Type II error
  - We will need to deal with it separately
- So the two hypotheses are not treated “symmetrically”
- This is why we put the hypothesis we want to reject in H0
  - If we reject H0, there is no concern about Type II error
  - We could not have a false negative

More Remarks

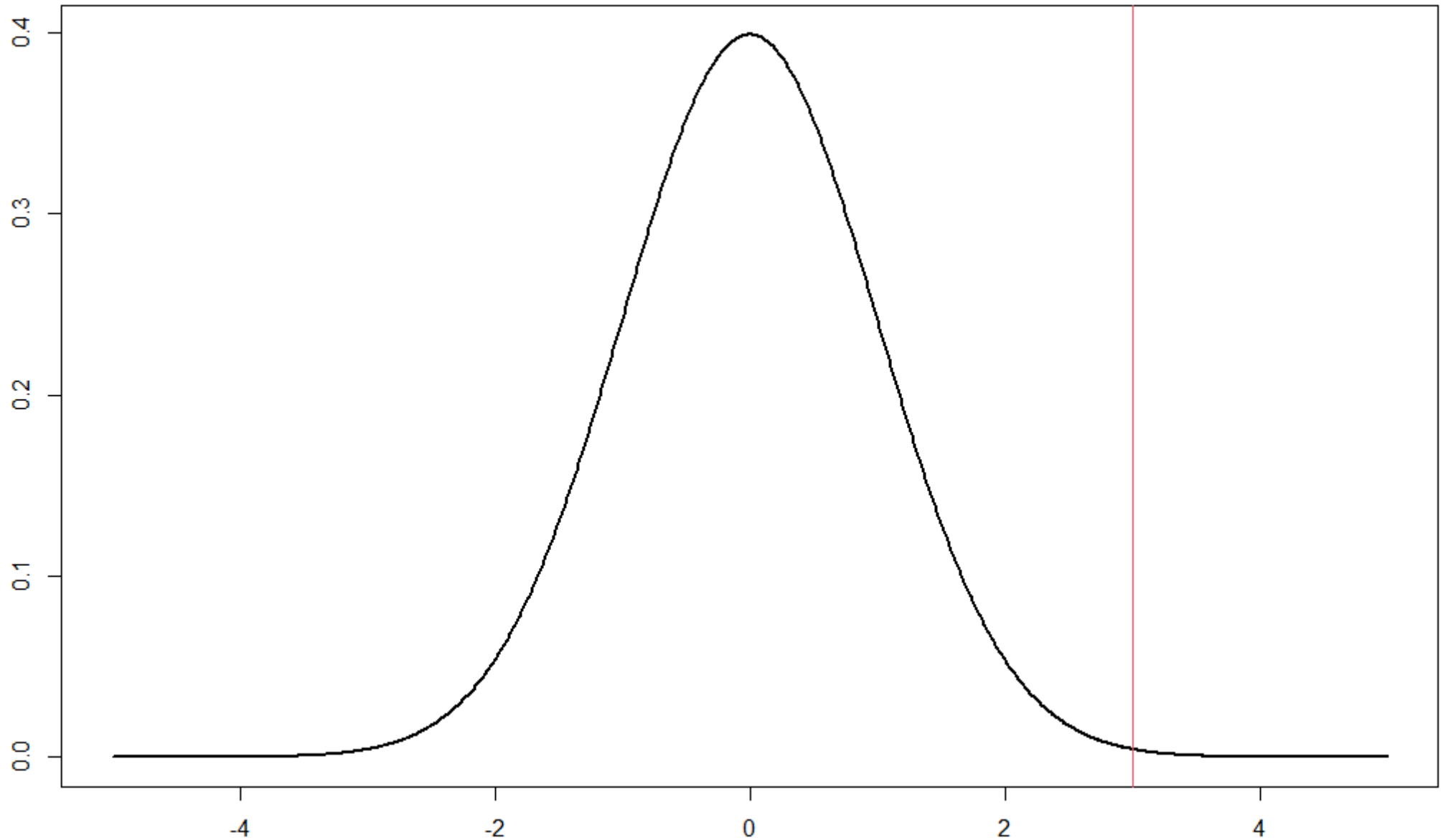
- Remember that the denominator staying the same is not generalizable to all Wald statistics
  - This is the hardest thing to figure out, the correct denominator.
- You do not have to use a Wald statistic; there are other “families” of test statistics
  - Wald, Score, Likelihood ratio
- Standard normal is an approximation. There are other (sometimes better) approximations

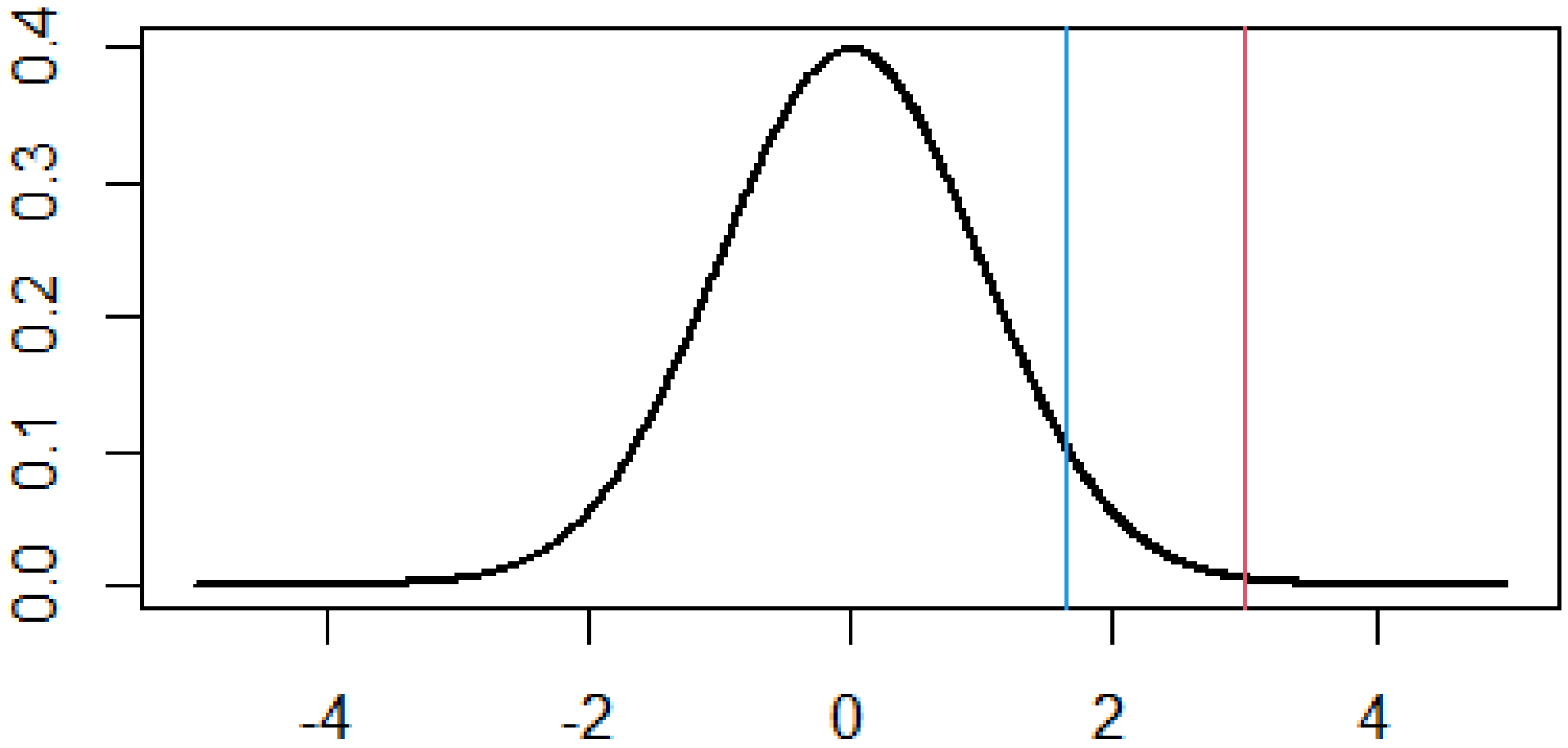
More Remarks

- In rare cases you can also derive the exact reference distribution, i.e. for a given sample size
  - Called an exact test
- In our example, it is possible to get an exact test
  - You might have heard of Fisher's exact test
- You can also derive an approximate reference distribution by computation
  - Like the histogram I showed
  - If you are comparing two or more samples, you would use something called permutation

What is the p-value?

- It is the probability of observing a test statistic higher than the one you observed in your data set
- You can use the p-value to test a hypothesis
  - If p< =0.05 reject H0
  - Else do not reject H0.


0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4


0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

Remarks About the P-Value

- It is not the Type I error !!!
- This cannot be over-emphasized
- P-VALUE IS NOT THE PROBABILITY OF TYPE I ERROR
- Type I error probability is not something you observe, it does not depend on the data. It is something you choose, ahead of time.
- Why do you have to choose it ahead of time?

More Remarks About the P-Value

- Think of the parallel universes
- If you do not choose the Type I error ahead of time, if you allow yourself to see the data first and then choose it ….
- Then these universes will not be parallel
- Parallel means identical procedures, rules and algorithms
- Parallel universes allow you to make a probability out of these parallel universes
- If they are not parallel you cannot calculate probabilities

More about p-values

- They are meant as a standardized way of measuring the strength of evidence against H0
- Smaller p, more evidence
- Test statistics measure that too but they are not necessarily standardized, and you cannot tell what is small or large

More about p-values

- They are strengths of evidence not effect
- You can have a small p-value with a small effect and a large sample size
- Think of the test statistic
- Numerator is the effect size
- It has sqrt(n) in the denominator of the denominator, i.e. n goes up denominator goes down test statistic goes up
- Same effect size but larger n means larger test statistic and smaller p-value

How Does One Test A Hypothesis Using a Confidence Interval?

- Produce a confidence interval and see if it is entirely contained in one of the hypothesized intervals or not.
- If it is then we rule in favor of that hypothesis
- In this example, confidence interval is 0.7 – 0.92, entirely contained within $r>0.5$, hence we conclude $r>0.5$
- What is the interval spanned both intervals (say it was 0.4 - 0.6)?

Asymmetry of hypothesis testing

- As long as our interval estimate contains a shred of the null region we cannot rule in favor of the alternative
  - For example, if the confidence interval here was 0.49-0.69 then we would not reject H0
- You get the same yes/no answer either way
  - Whether you use a p-value of a confidence interval
- Confidence interval has the effect size (center), strength of evidence (width) and the test (null value included or not)
- This is why you hear confidence intervals are better than tests

Why not use it all the time?

- Ignorance
  - Don’t know that it is better
- Habit
  - Editors and reviewers ask for a p-value
- Laziness
  - CI requires more intellectual work