Biostatistics
Lecture 1

Mithat Gonen

Brendon Bready

Preliminaries

- Once a week, in person
- We might have one week over zoom when I am traveling
- Asking Questions During Lectures
  - Please do. The more the better. There are no bad questions in this class

Preliminaries

- All documents on Moodle
- Email for announcements/updates
- No textboooks
- These slides and your class notes will function like a textbook

Homework Assignments

- Assigned on Tuesday, due back on next Tuesday
- Do not wait until Sunday evening to start the homework
- A mix of problem sets and reading
  - Problems sets will be graded
  - Reading assignments will be discussed in class. We expect everyone will participate
- Suggest reading materials from your literature
  - Send to us and we will decide when it is the right time

What This Course Is and Is Not

- What This Course Will NOT Do
  - Give you skills of a data analyst or a statistician
  - Make you proficient in statistical software
- We hope you do not become
  - Someone who is overconfident in their knowledge and understanding of statistics
- What This Course Will Try to DO
  - Provide Statistical Literacy
  - Teach Critical Thinking from a Quantitative Perspective in Clinical Research
- We Hope You Will
  - Know what you know
  - Know what you don’t know

How Is This Course Different Than Other Statistics Classes You Might Have Taken?

- Students are different
  - Completed advanced medical training
  - Already engaged in research
  - (Some ?) respect for statistics as the language of empirical research
  - Highly motivated to learn, not just to graduate

How Is This Course Different Than Other Statistics Classes You Might Have Taken?

- Teachers are different
  - Combined experience of ~ 25 years in medical research
  - First-hand experience with hundreds of clinical researchers
  - Highly motivated to teach, not just go through the motions

How Is This Course Different Than Other Statistics Classes You Might Have Taken?

- Unique focus on cancer research
- Statistics is a field widely used in many disciplines
- Fundamental principles are the same, but details differ widely
- What is relevant to behavioral science or experimental physics, or even infectious diseases can be quite different than what is relevant to cancer

Potential Pitfalls

- Diverse baseline knowledge of statistics
- Do not doze off thinking “I know this”
  - You either do not know it
  - Or you know it wrong
  - Or at best, your knowledge is incomplete

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

What is our population?

- Ideally, we should begin all clinical research studies with a definition of the population
- Clinical trials try to do this
  - Inclusion/exclusion criteria in the protocols is an attempt to define the population
- In observational studies we too often start with the data (sample) and try to figure out the population from the data
  - Exactly the opposite of what we should do
  - We will come back to this over and over again in this course

But what is it?

- It is difficult to define your population
- Suppose we have a single-institution Phase II clinical trial
- All patients with stage IV colorectal cancer scheduled for liver resection and candidates for adjuvant chemotherapy
- This is our population, give or take some details such as sufficient liver function, no chronic GI disease etc etc
- So, the sample (patients who will enroll) is from this population

Not so quick

- All patients will be at MSK
- Does this mean our population is such MSK patients?
- Or do we mean such patients everywhere, but we think sampling MSK patients is enough
  - Worth thinking
- Not entirely a statistical issue but has statistical consequences

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

Sample and Estimate

- 53 patients
- Table 1 of the paper describes the sample
- It is a subjective evaluation whether this is representative of the population
- 44 of 53 patients responded
- What can we do this with this information?

Table 1. Characteristics of the 53 Patients at Baseline.*

Characteristic	Value
Age	
Median (range) — yr	44 (23–74)
Distribution — no. (%)	
18–30 yr	14 (26)
31–60 yr	31 (58)
>60 yr	8 (15)
No. of previous therapies — no. (%)	
2	21 (40)
3	13 (25)
≥4	19 (36)
Primary refractory disease — no. (%)	
Yes	12 (23)
No	41 (77)
Previous allogeneic HSCT — no. (%)	
Yes	19 (36)
No	34 (64)
Previous treatment with blinatumomab — no. (%)	
Yes	13 (25)
No	40 (75)
Pretreatment disease burden†	
Median bone marrow blasts (range) — %	63 (5–97)
Bone marrow blasts — no. (%)	
≥5%	27 (51)
<5% with extramedullary disease	5 (9)
≥0.01% and <5%	15 (28)
<0.01%	6 (11)
Philadelphia chromosome–positive — no. (%)	
Yes	16 (30)
No	37 (70)

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)

Maximum Likelihood Estimates

- This is a general method to generate point estimates
- It turns out that sample analogs are also maximum likelihood estimates
- We will not discuss what maximum likelihood estimates are in this class, but you should know that it is a generically good way of obtaining estimates to pretty much any parameter

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

Hypothesis Testing

- Suppose at the time of study design we thought 50% of patients in this population would respond to standard of care
- Then a reasonable hypothesis to test is $r \leq 0.50$ vs $r > 0.50$
- This is the inverse of interval estimation
- We start with pre-defined intervals and ask which interval is more likely to contain the true value

How Does One Test A Hypothesis?

- 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

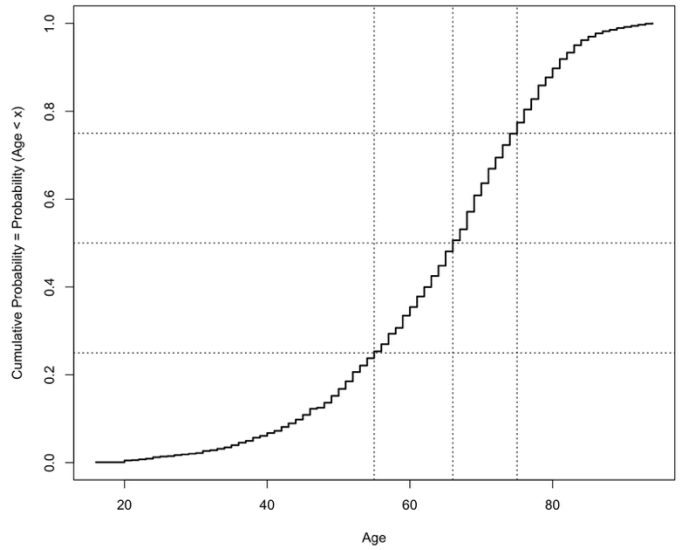
- r<=0.50 vs r>0.50 – each is a hypothesis. One of them we want to disprove (to be called the null hypothesis, or H_0) and the other we want to prove (alternative hypothesis, H_1).
- They are not symmetrical for reasons we will discuss later in this class
- As long as our interval estimate contains a 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

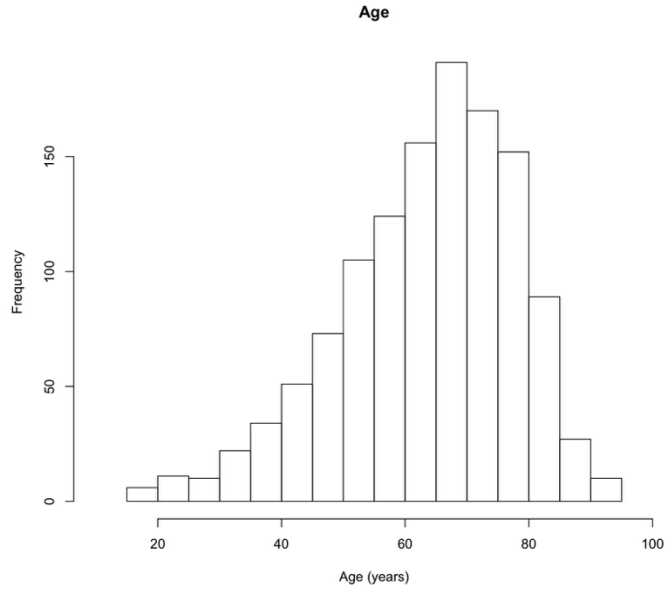
Another way of testing a hypothesis

- Generate a p-value (to be defined in the coming weeks) from the data
- If $p < 0.05$ the conclude alternative hypothesis is consistent with the data, otherwise conclude null hypothesis is still the best thing we have
- We have many discussions coming on this very popular and infamous method

More on Point Estimation

- Let’s use a continuous variable for illustration
- We have a (real) data set of 1231 patients undergoing resection at MSK





Entire data as population

- Only for pedagogic purposes here
- Assume that the 1231 numbers we have as age is the entire population
- I have access to the entire population (i.e. all 1231 records); you don’t
- You want to estimate the mean age of the population and I agreed to give you a very small sample from the population, for the time being only 10 samples.

Sample

- Here is a randomly chosen sample of 10 ages from this population
  - 68 73 80 48 67 74 68 55 52 63
- How do you estimate the population mean from this sample?

Sample Mean

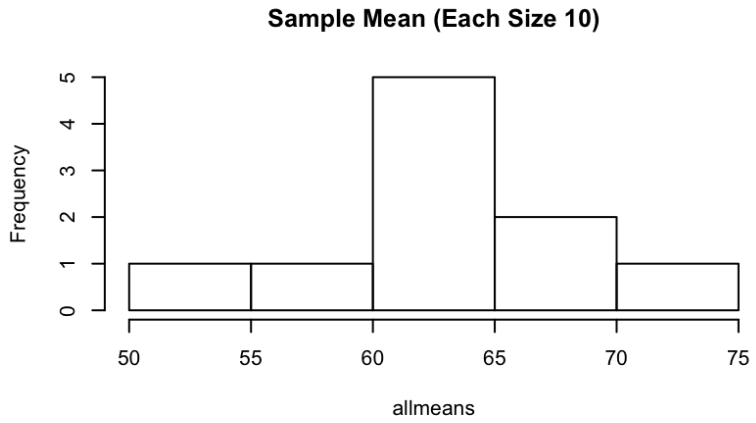
- 89 59 60 44 65 82 83 72 56 81
- Sample Mean: 64.8, our estimate of the population mean

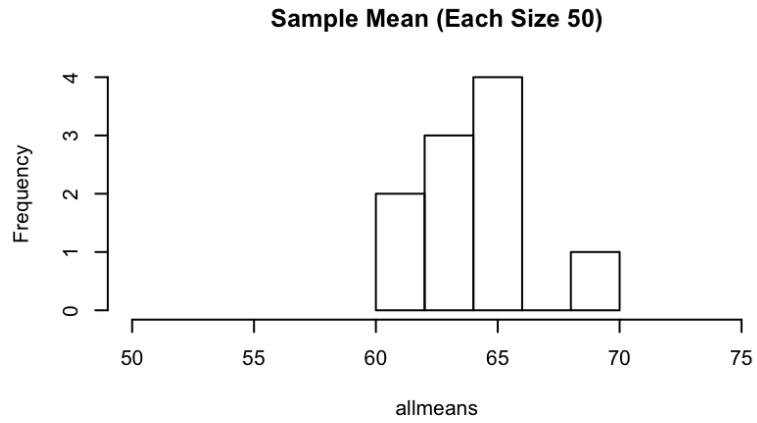
A sample for everyone in this class

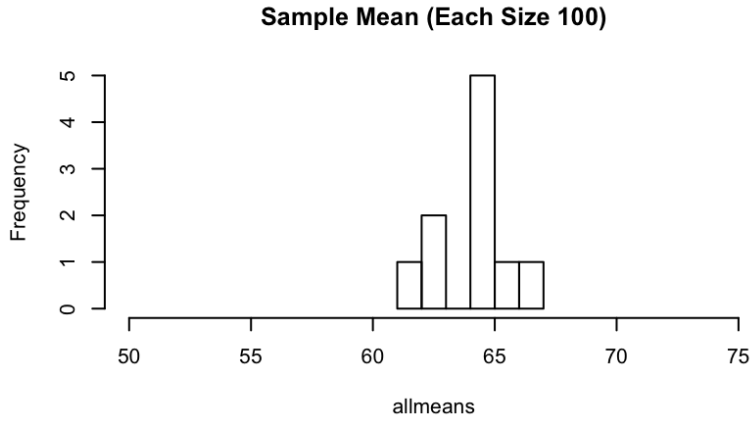
- 89 59 60 44 65 82 83 72 56 81
  - 69.1
- 68 73 80 48 67 74 68 55 52 63
  - 64.8
- 62 22 85 69 66 83 55 66 60 53
  - 62.1
- 38 54 61 48 78 31 75 89 76 75
  - 62.5
- 69 73 81 66 72 28 63 57 79 86
  - 67.4
- 68 51 57 40 52 58 69 67 41 42
  - 54.5
- 57 36 81 73 44 29 64 70 60 77
  - 59.1
- 60 50 44 58 75 63 80 74 77 64
  - 64.5
- 83 66 62 67 70 68 65 68 74 80
  - 70.3
- 44 48 67 50 82 75 70 75 57 65
  - 63.3

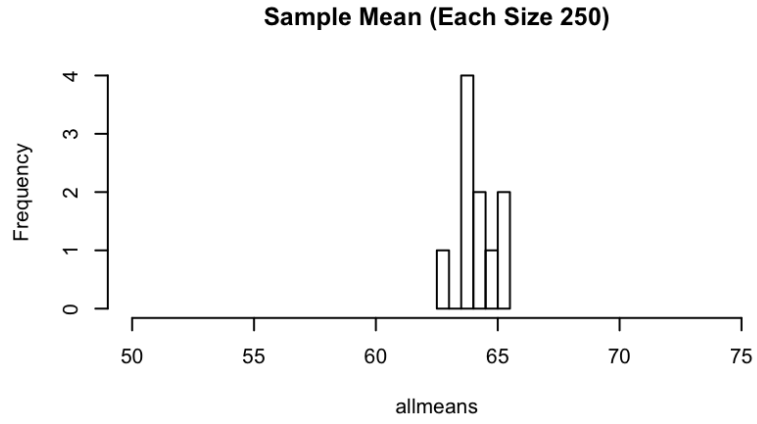
Distribution of the Sample Mean

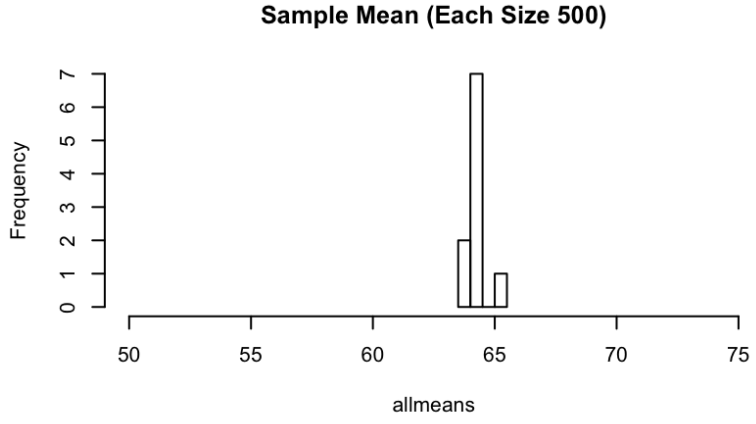
- I want to understand how the sample mean behave
- I have 10 samples, I also have 10 sample means
- Can I use these 10 samples to see how sample mean is distributed?
- What happens to this distribution if sample size increases?
  - I will continue to give everyone in the class additional rounds of samples
  - At each round sample size will increase
  - I will plot histograms of each round

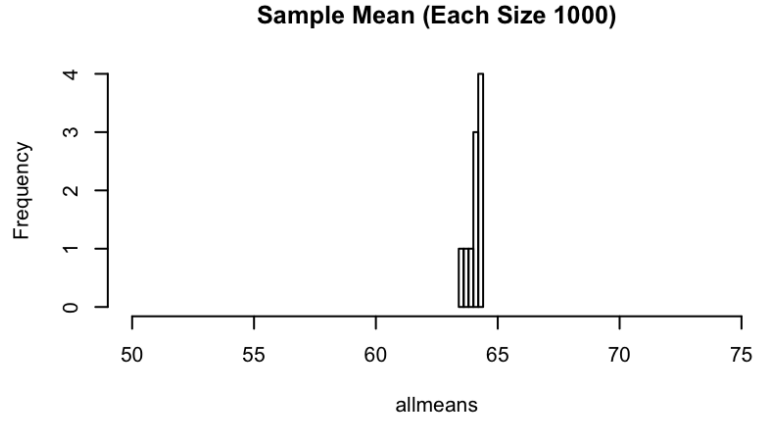

Sample Mean (Each Size 10)
Frequency
0 1 2 3 4 5
50 55 60 65 70 75
allmeans


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

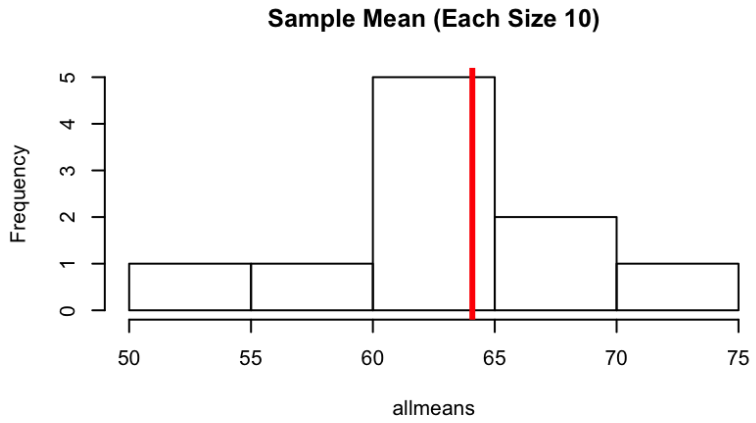

Sample Mean (Each Size 250)
Frequency
0 1 2 3 4
50 55 60 65 70 75
allmeans

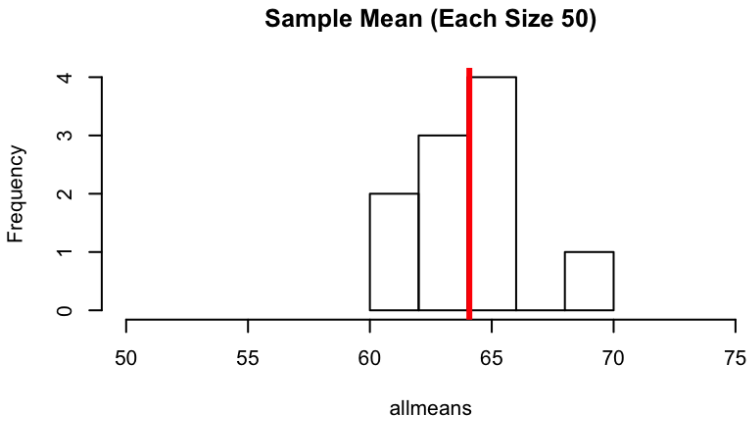

Sample Mean (Each Size 500)
Frequency
0 1 2 3 4 5 6 7
50 55 60 65 70 75
allmeans

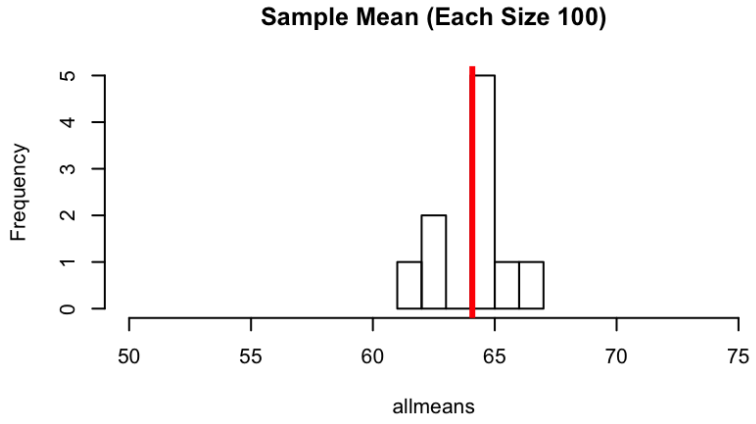

Sample Mean (Each Size 1000)
Frequency
0 1 2 3 4
50 55 60 65 70 75
allmeans

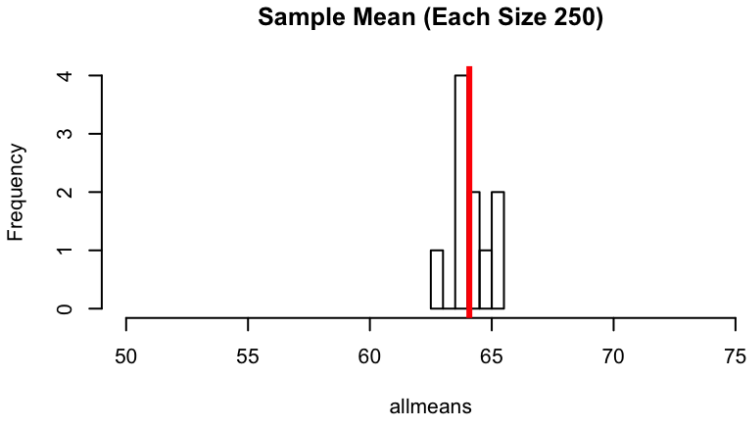
Law of Large Numbers

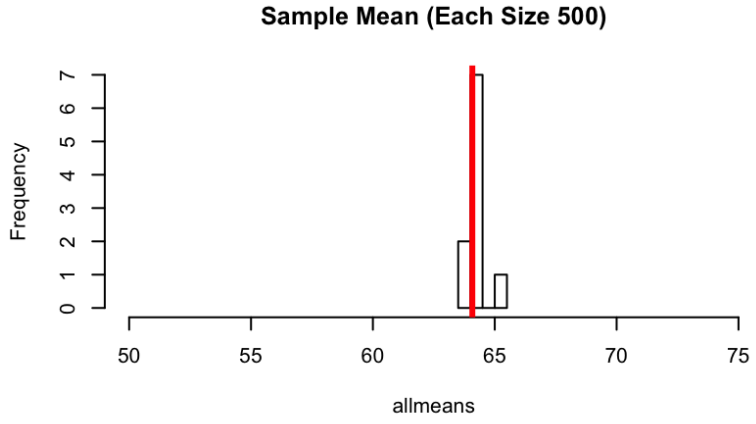
- As the sample size increases, the distribution of the sample mean gets more and more concentrated
- This is called the law of large numbers
- Can we figure out the value around which the sample means gets more concentrated?

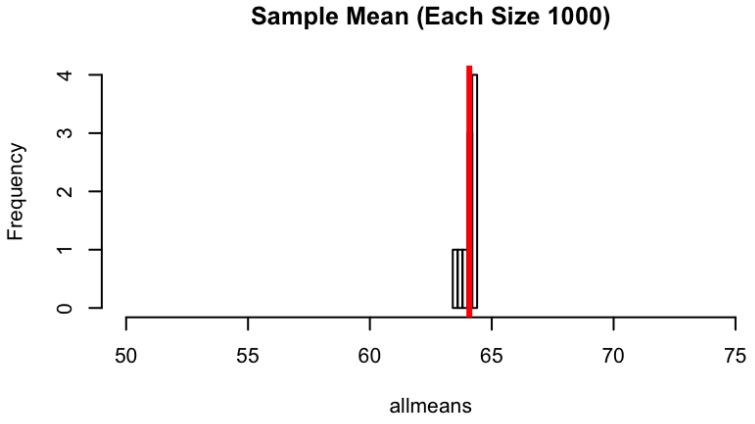

Sample Mean (Each Size 10)
Frequency
0
1
2
3
4
5
50
55
60
65
70
75
allmeans


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.

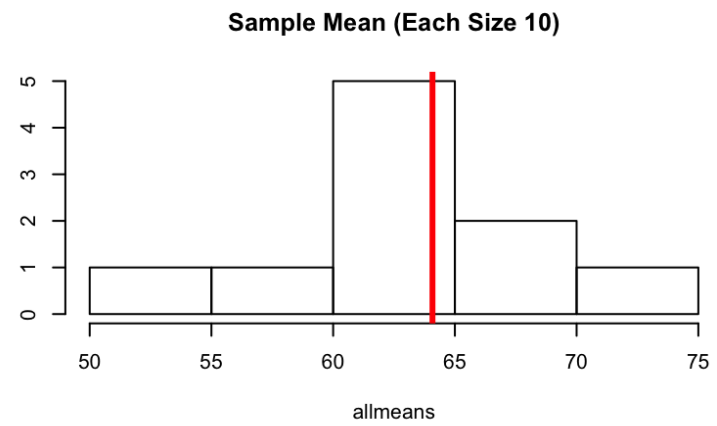
Some generalizations

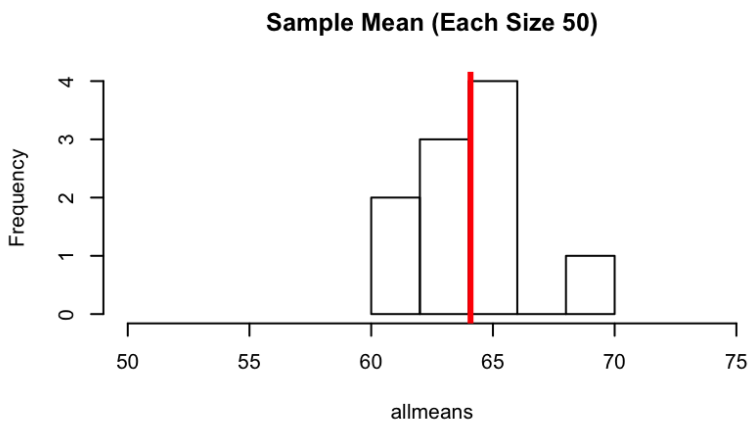
- Instead of sample mean you say “estimate” and this statement will be true
- In fact this is one definition of a good estimate (does it satisfy the law of large numbers?)
- So this idea is not limited to means

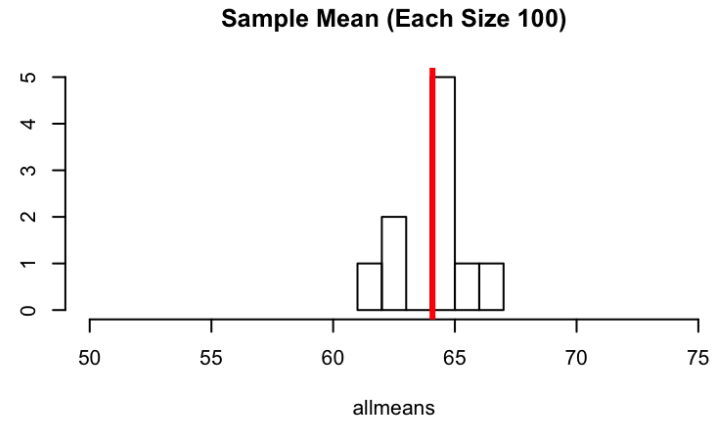
Works for binary data too!

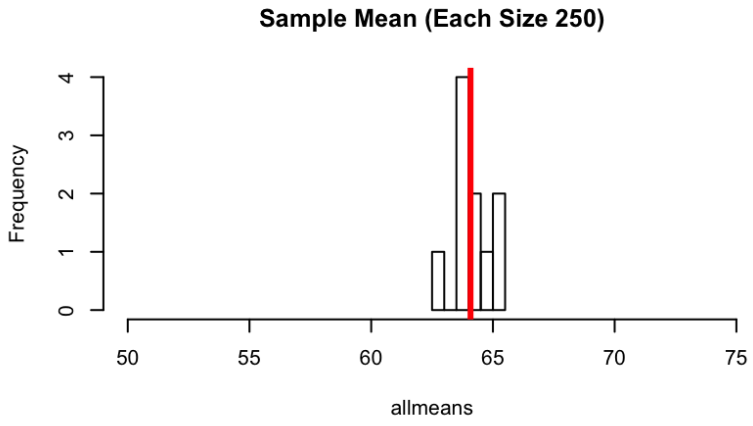
- This idea is not limited to continuous variables either
- Sample proportion is like a sample mean
- If we repeat what we just did for a proportion we will observe a similar finding.
- In fact, “works” for all kinds of data

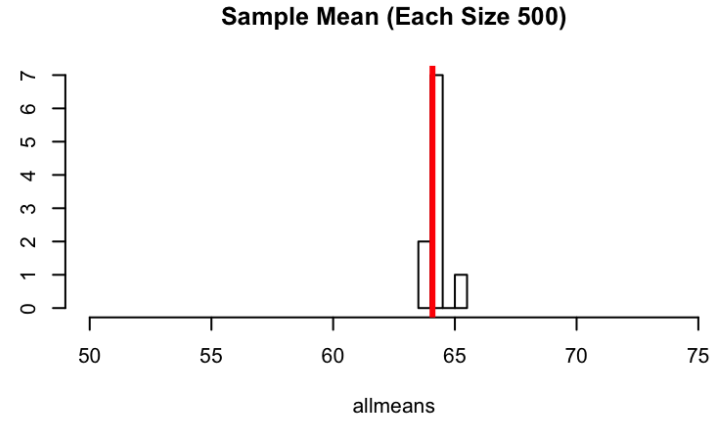
Can we say more about the distribution of these repeated samples?

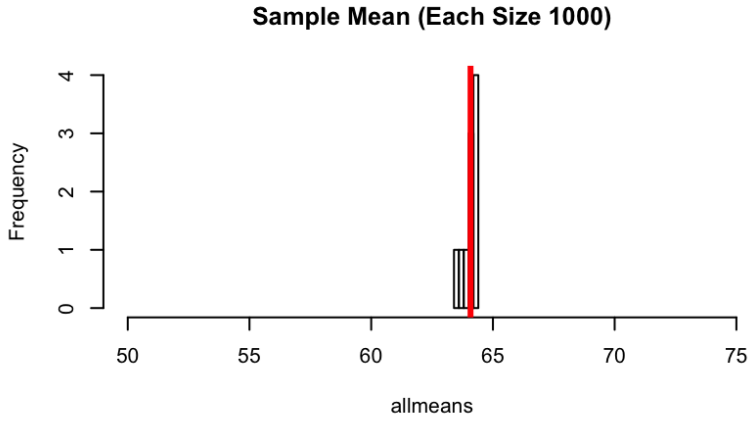




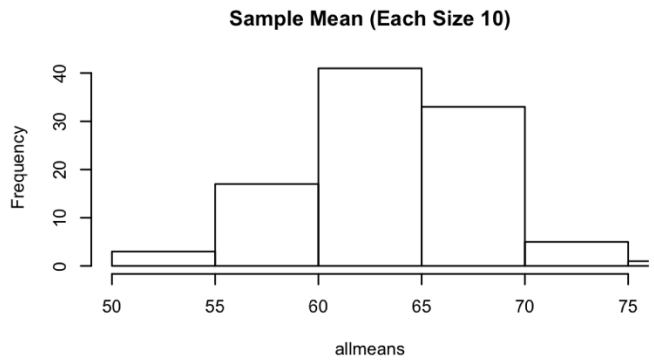


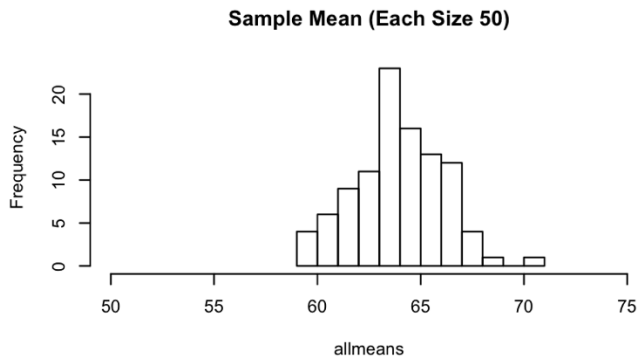


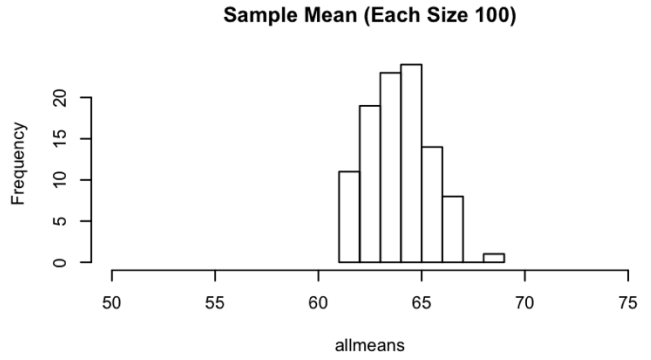


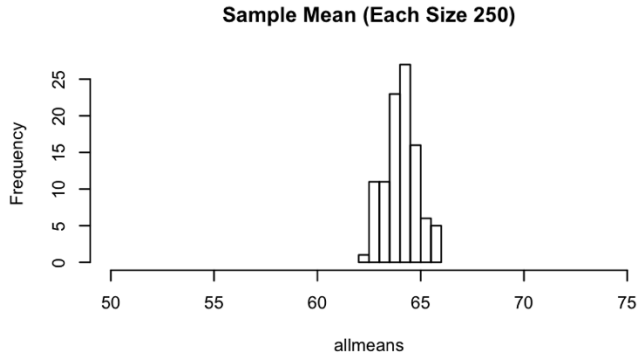


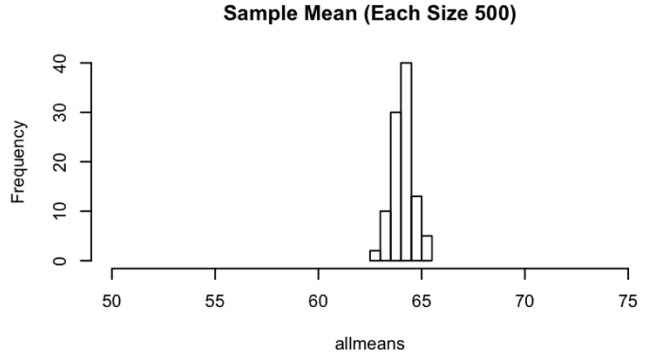
100 Samples Instead of 10

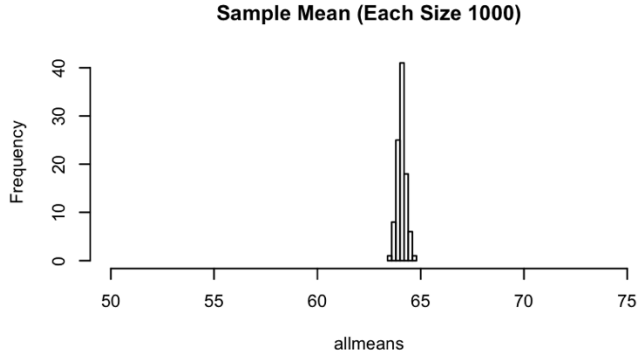


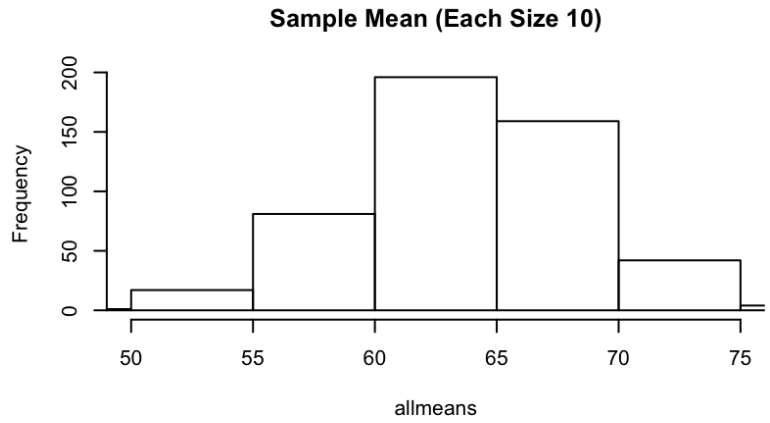


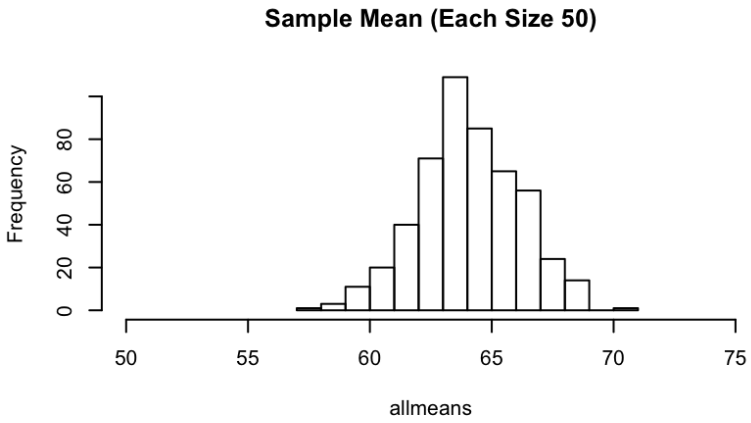


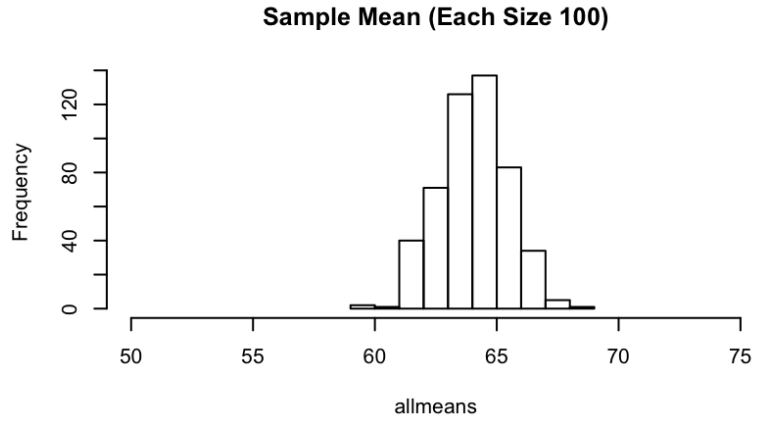


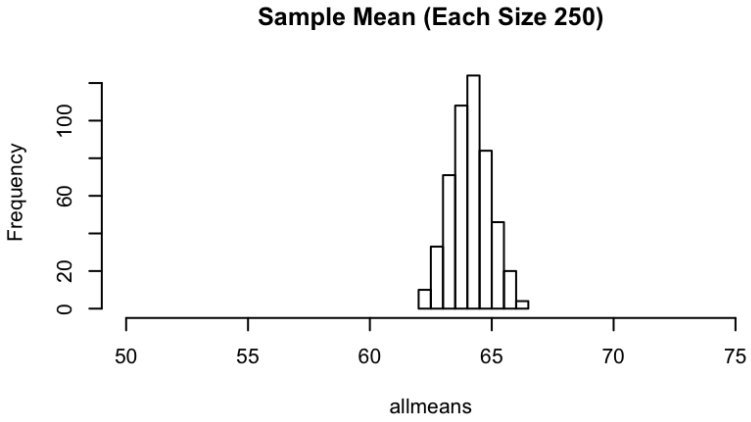


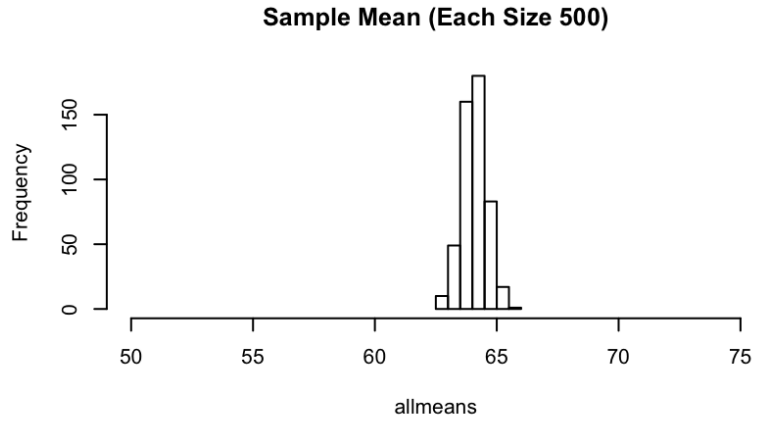


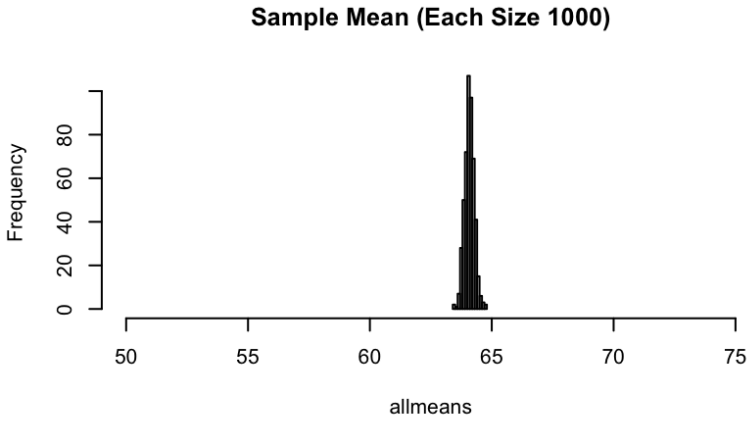

Sample Mean (Each Size 10)
Frequency
0
50
100
150
200
50
55
60
65
70
75
allmeans

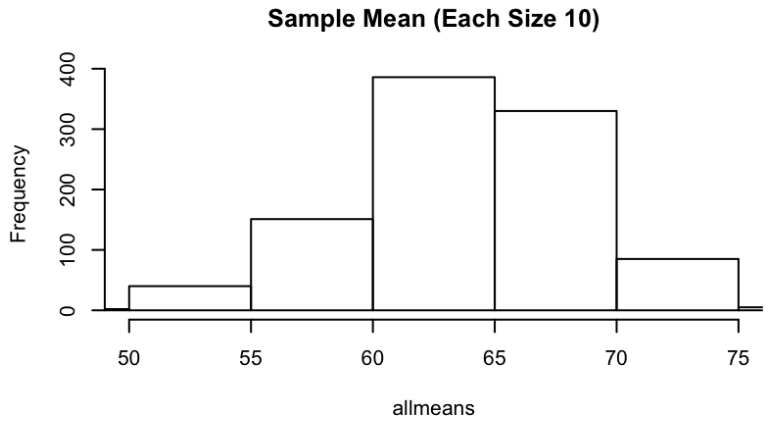

Sample Mean (Each Size 50)
Frequency
0
20
40
60
80
50
55
60
65
70
75
allmeans

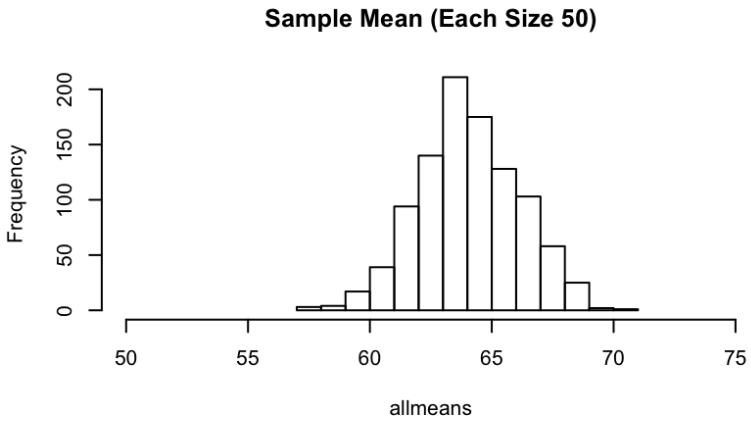

Sample Mean (Each Size 100)
Frequency
0
40
80
120
50
55
60
65
70
75
allmeans

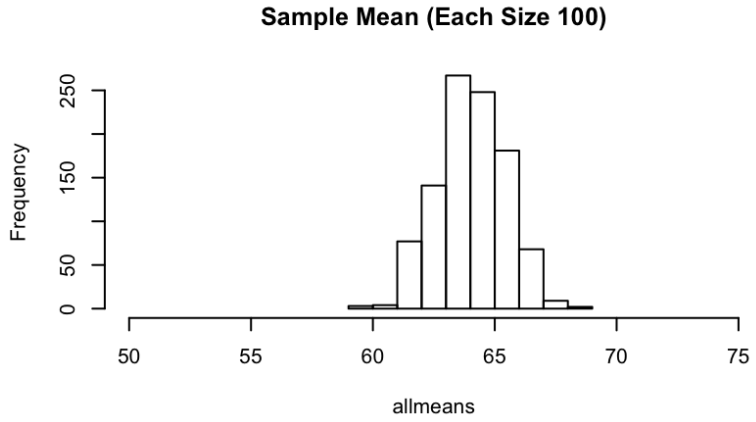

Sample Mean (Each Size 250)
Frequency
0
20
60
100
50
55
60
65
70
75
allmeans

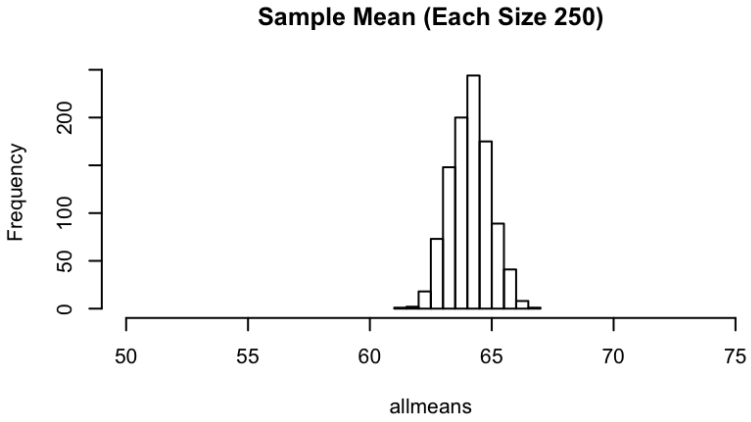

Sample Mean (Each Size 500)
Frequency
0
50
100
150
50
55
60
65
70
75
allmeans

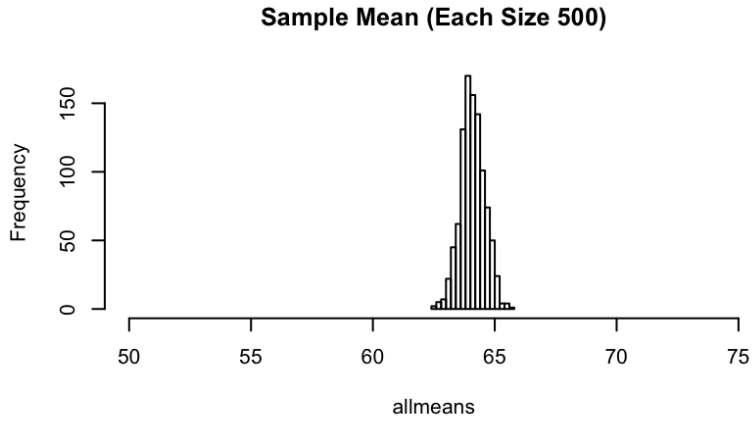

Sample Mean (Each Size 1000)
Frequency
0
20
40
60
80
50
55
60
65
70
75
allmeans

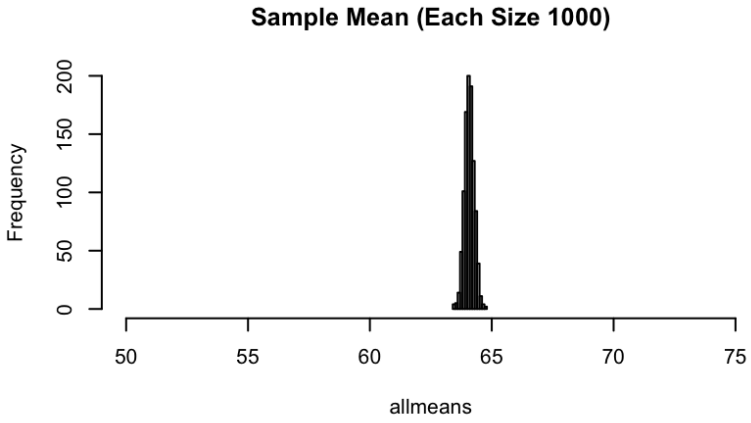

Sample Mean (Each Size 10)
Frequency
0
100
200
300
400
50
55
60
65
70
75
allmeans

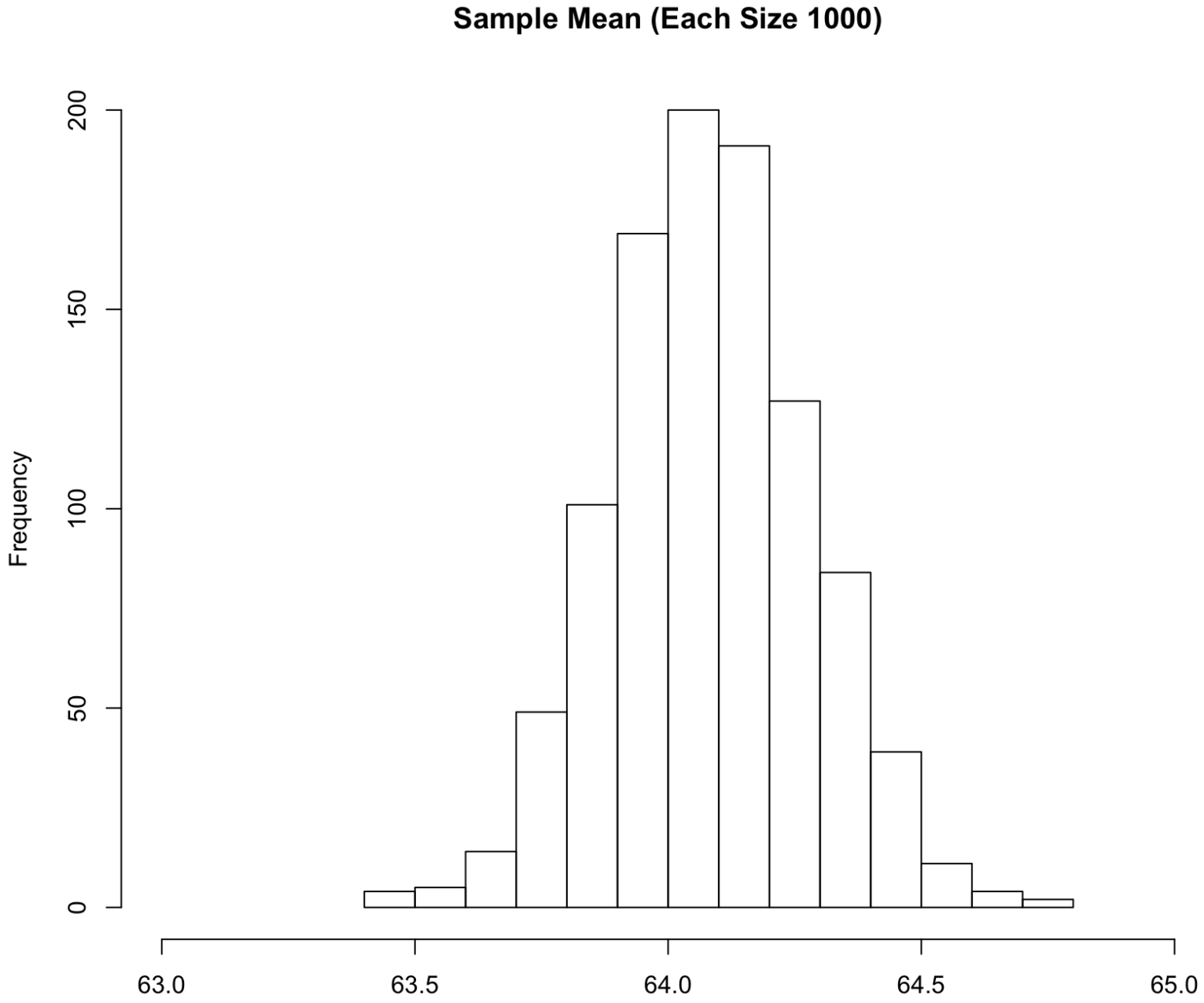

Sample Mean (Each Size 50)
Frequency
0
50
100
150
200
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 100)
Frequency
0
50
150
250
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 250)
Frequency
0
50
100
200
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 500)
Frequency
0
50
100
150
50
55
60
65
70
75
allmeans


Sample Mean (Each Size 1000)
Frequency
0
50
100
150
200
50
55
60
65
70
75
allmeans

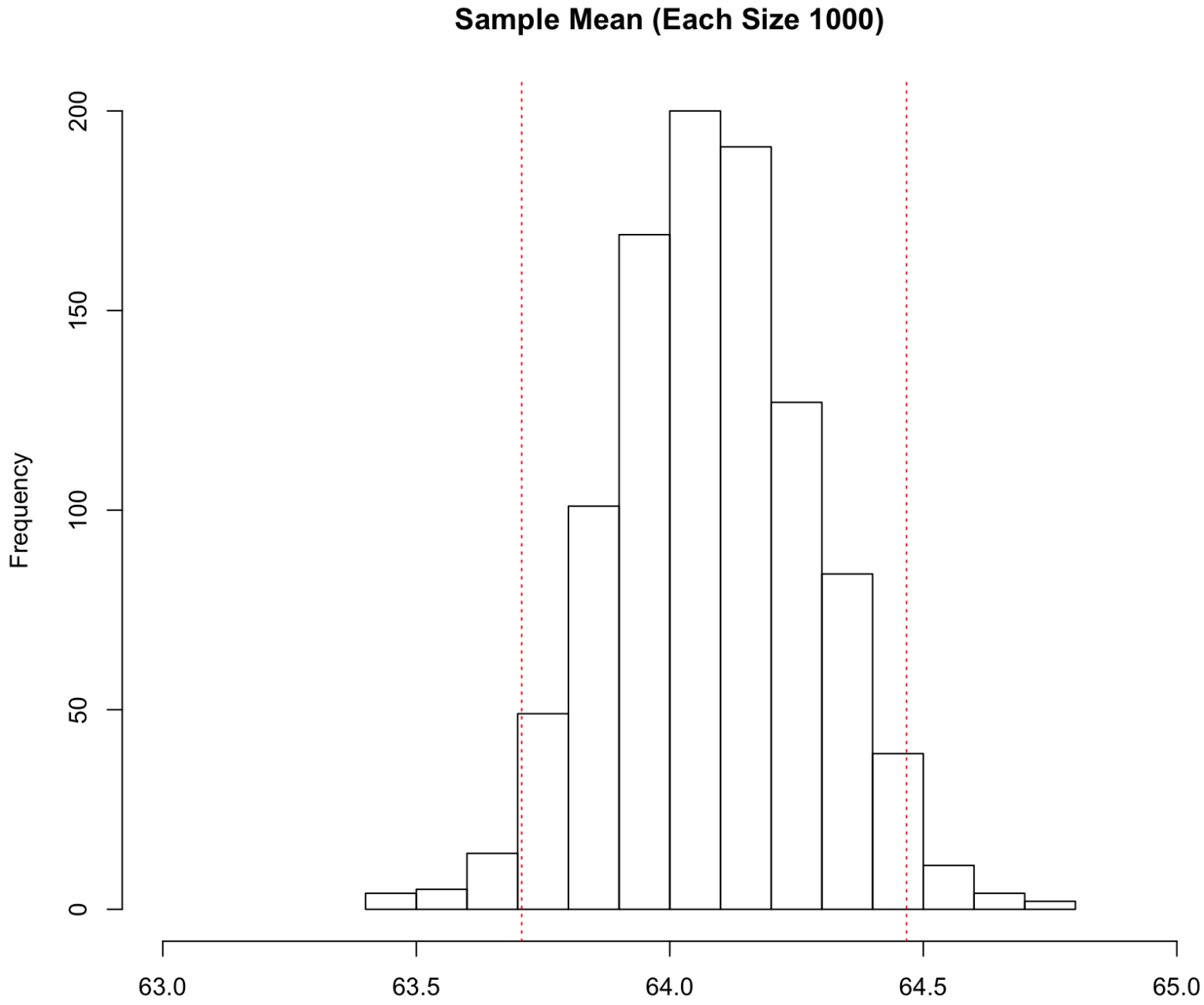

Sample Mean (Each Size 1000)
Frequency
0
50
100
150
200
63.0
63.5
64.0
64.5
65.0

Summary of what we did

- We obtained a large number of 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
- We also saw that the distribution looks more and more like a bell curve
- This is called the Central Limit Theorem

Why is it important?

- This bell-shaped distribution turns out to be the normal distribution
- It is not exactly normal, only approximately
- But as the size of each sample increases the approximation gets better better
- We can make “approximate” probability calculations using this fact



The interval between the red lines contains the sample means of 950 of the 1000 samples, i.e. 95% probability

But we only have one sample in real life

- Very true
- We cannot generate any of these histograms during a data analysis
- But we can make these calculations using mathematical formulae even though we do not have repeated samples
- The learning goal for this class is not that, it is this concept of repeated sampling that underlies almost all statistics
- Anytime you are looking at data, close your eyes and imagine these histograms.