Biostatistics

Mithat Gonen

Brendon Bready

Today’s Lecture

- Data Types and Visualization
- Review of Last Week

Types of Statistical Variables

- Outcomes vs predictors (correlates, causes, exposures)
- Outcome
  - The things we want to know
  - Variables that we do not want to “control” even if we can
- Predictor
  - The things that we think have an effect on or move with the outcome
  - Variables that we want to “control” if we can

Outcomes

- Most clinical research is conducted to find about outcome variables
- Example: Vemurafenib in BRAF-mutant melanoma
  - Serial tumor biopsies and look at changes in tumor cell proliferation (cyclin D1, ki-67) and phosphorylated extracellular signal-related kinase (p-ERK)
  - This is an outcome, we treat cells, or animals or humans with the drug and measure these markers at multiple time points.
  - This is an outcome because it is why we do the study, what we want to know

Predictors

- Same example: Vemurafenib in BRAF-mutant melanoma
  - Dose and schedule of treatment
  - Tumor characteristics (size, location etc)
  - Timing of biopsies
- These are all predictors (or correlates, or some of them may actually carry a causal relationship with the outcome) because we do not want to know about them but we think their presence (or amount) is related to the outcome

One person’s outcome is an another’s predictor

- p-ERK levels can be a predictor in another study
- Is the amount of change in p-ERK associated with survival?
- Now change in p-ERK is a predictor?
- What is the outcome in this case?

Types of Outcomes

- Binary
- Continuous
- Censored
- There are many others but these cover 90%+ of all clinical cancer research studies

Binary Outcomes

- Yes/No, Present/Absent, Up/Down, Good/Bad
- Represent the many dichotomies that we deal with in clinical care and research
- Easy and a strict definition: each case must be classified as one of exactly two possibilities.

Binary variables follow a Bernoulli distribution

- Sometimes Bernoulli is used exchangeably for binary
- I will assume binary variable is coded as 0/1.
- P(V = 1) = p (read this is as “probability of the variable V being equal to 1 is p”)
- There is a nice symmetry to Bernoulli distribution. We only need to know p to understand how it behaves.
- Why? Because P(V = 0) = 1 – p
  - This happens because there are exactly two possibilities.

History Corner

- Bernoulli family originated in Antwerp, Belgium and settled in Basel, Switzerland in the 17^{th} century
- Eight members of the family turned out to be prominent mathematicians
- Jacob Bernoulli (1654-1704) formalized the Bernoulli distribution
- In addition to the Bernoulli distribution: Bernoulli differential equations, Bernoulli triangle, Bernoulli inequality and Bernoulli polynomials

Cancer example: Response

- Did a cancer patient given a certain treatment respond?
- We need a precise definition of response
  - Such as RECIST, or MRD
- Response rate (clinical language)
- Probability of response (statistical language)
- Both refer to P(Response = 1) = r

Continuous Outcomes

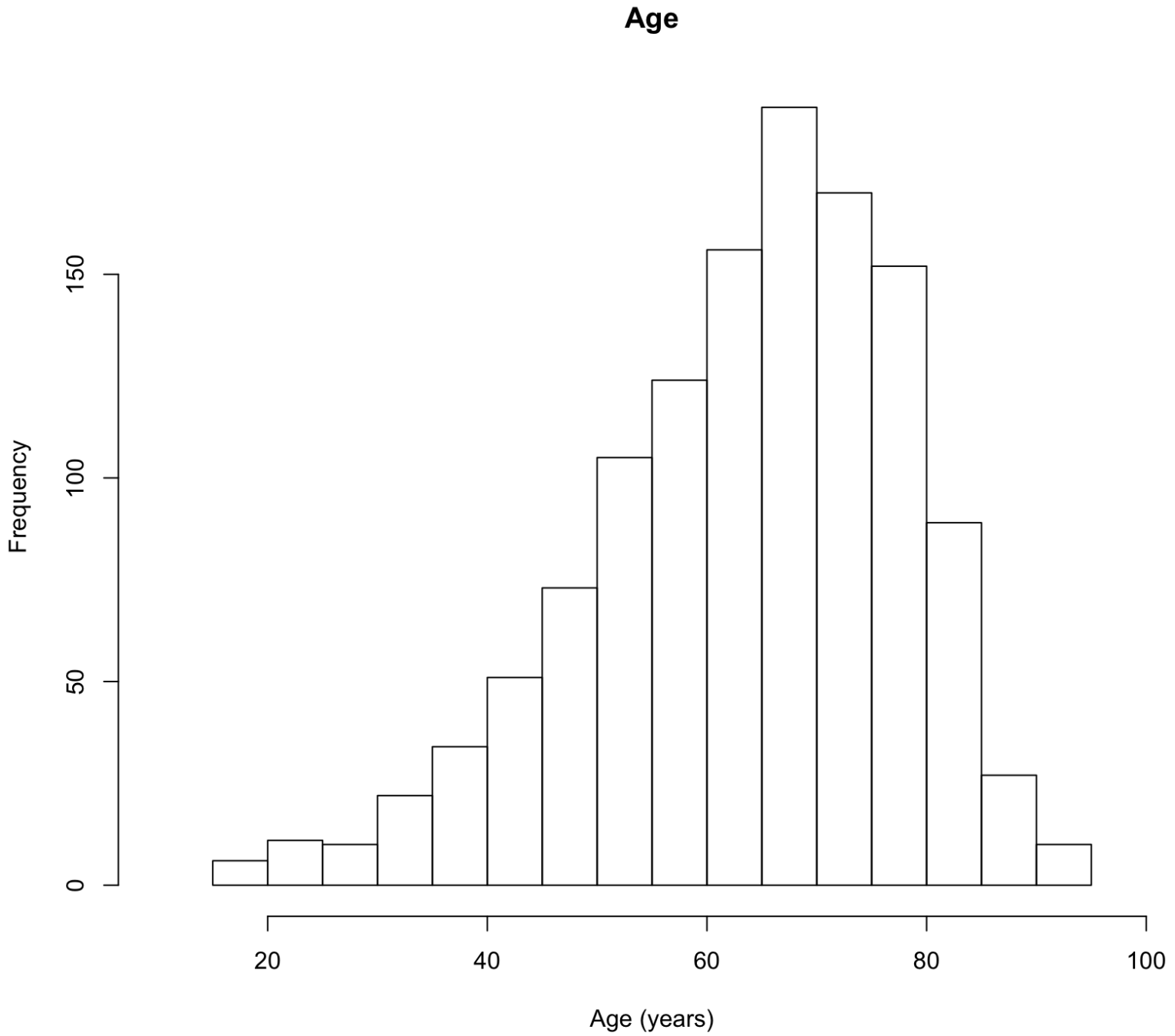
- A variable that takes on numeric values
- Age, bilirubin, tumor size
- Technically it refers to a variable that can take on infinitely many values
- You can talk about the mean, median, standard deviation etc for continuous variables

Surgical Data Example (Sex and Age)

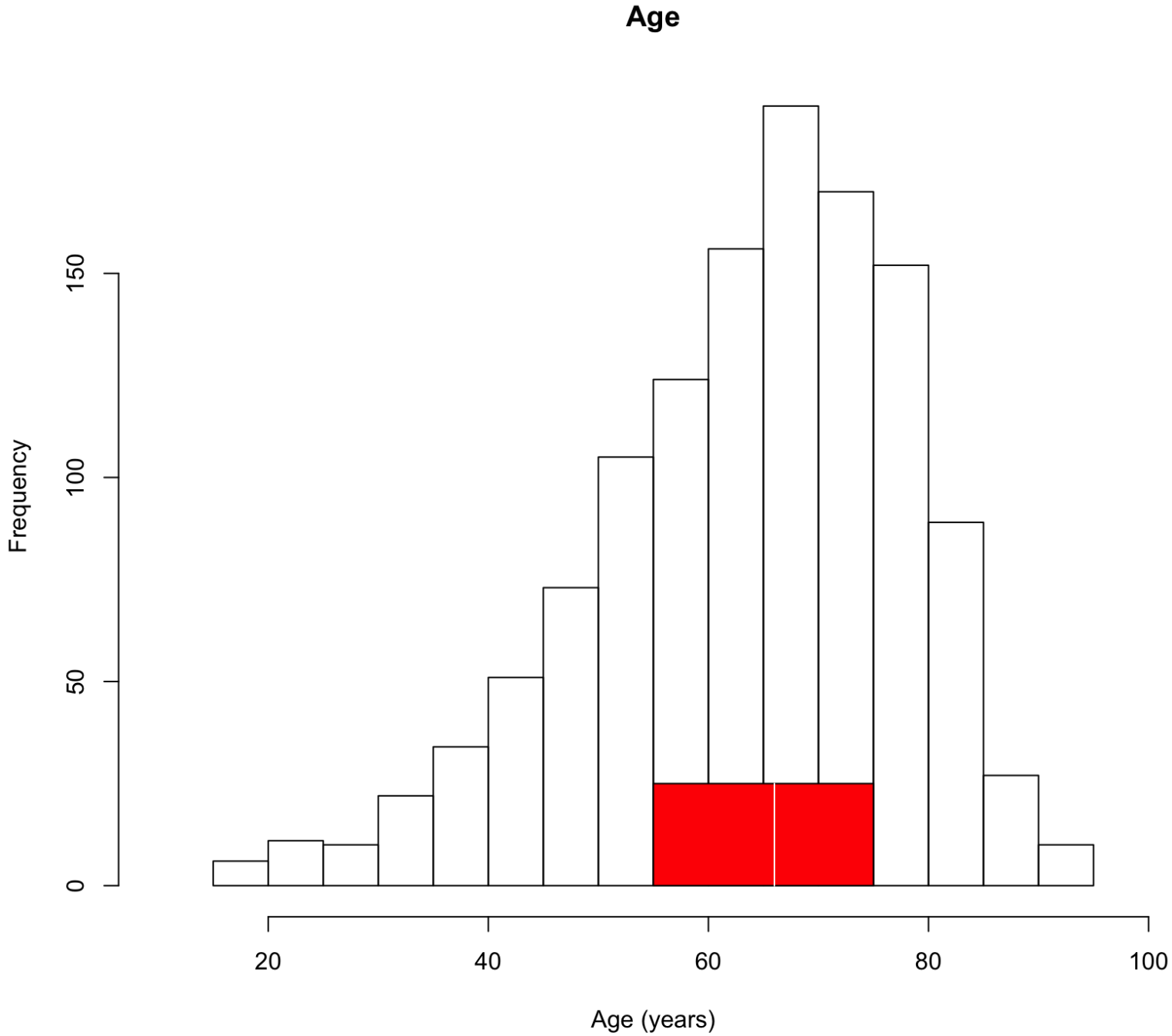
- 1231 patients undergoing a certain type of GI resection for cancer at MSK
- Example variables: Sex and Age
- Out of 1231, 552 were female (45%)
- Pretty much the only thing you can report is this, an estimate of P(Sex = Female)
- Equivalent to reporting 1 – P(Sex = Female) = P(Sex = Male) = 55%

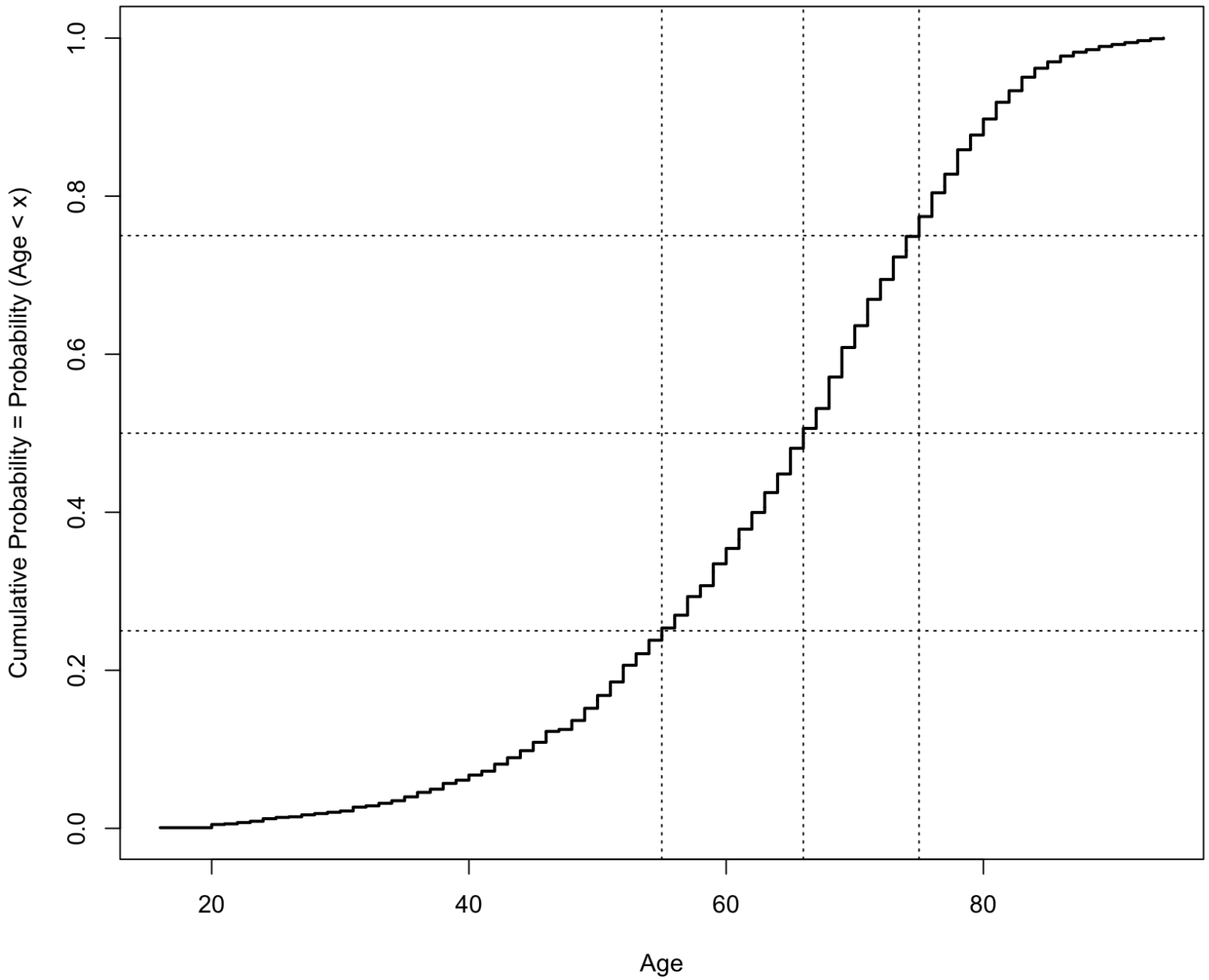
Surgical Data Example (Sex and Age)

- Age: continuous variables --- so many things to report
  - Mean: 64.1
  - Median: 66.0
  - Range: 16 – 94
  - Interquartile Range: 55 – 75
  - Five Number Summary: 16, 55, 66, 75, 94 (min, quartiles, max)
- What are these numbers? What do they mean?
  - Histogram
  - Cumulative density function (CDF)



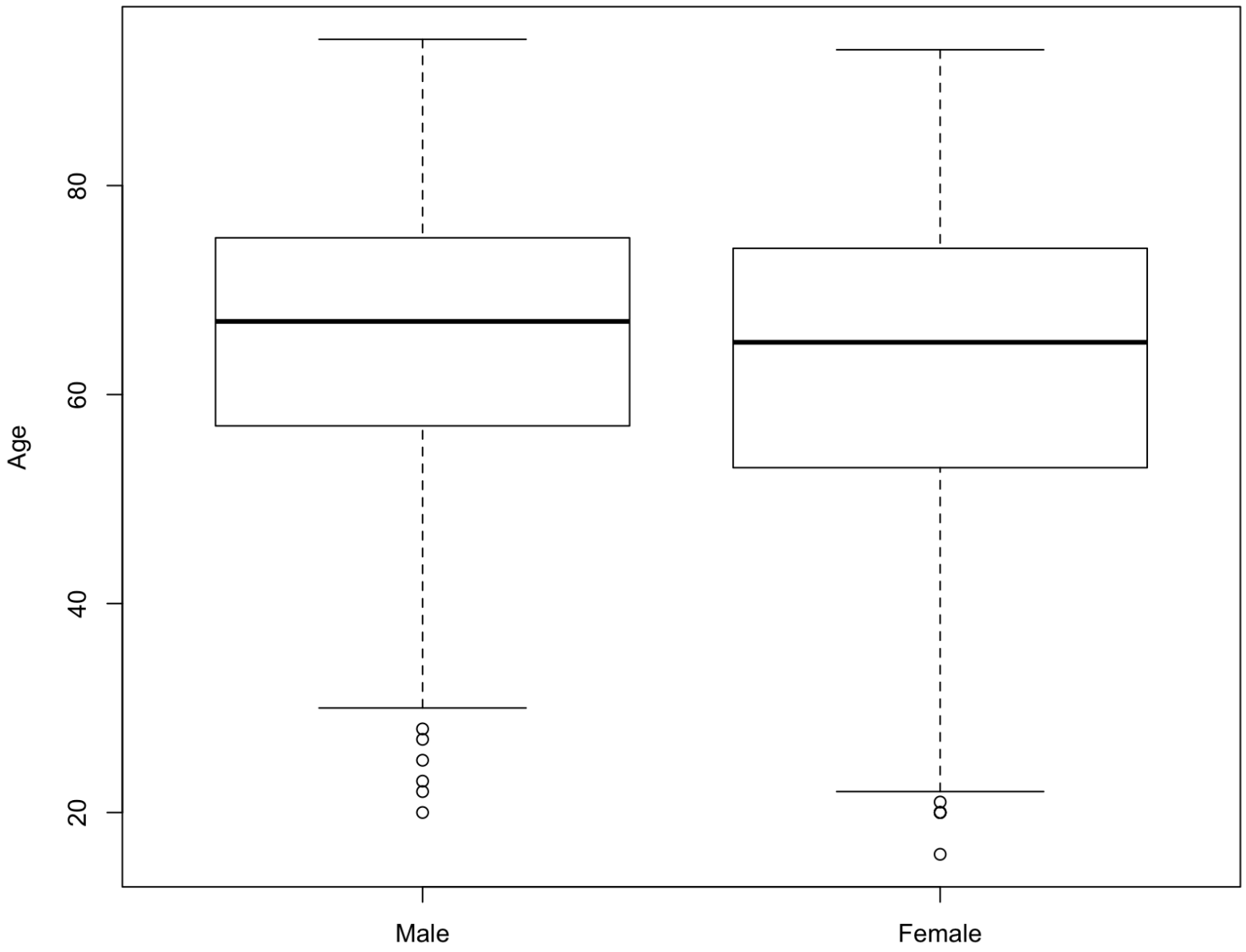
Histogram


Age
Frequency
0
50
100
150
20
40
60
80
100
Age (years)



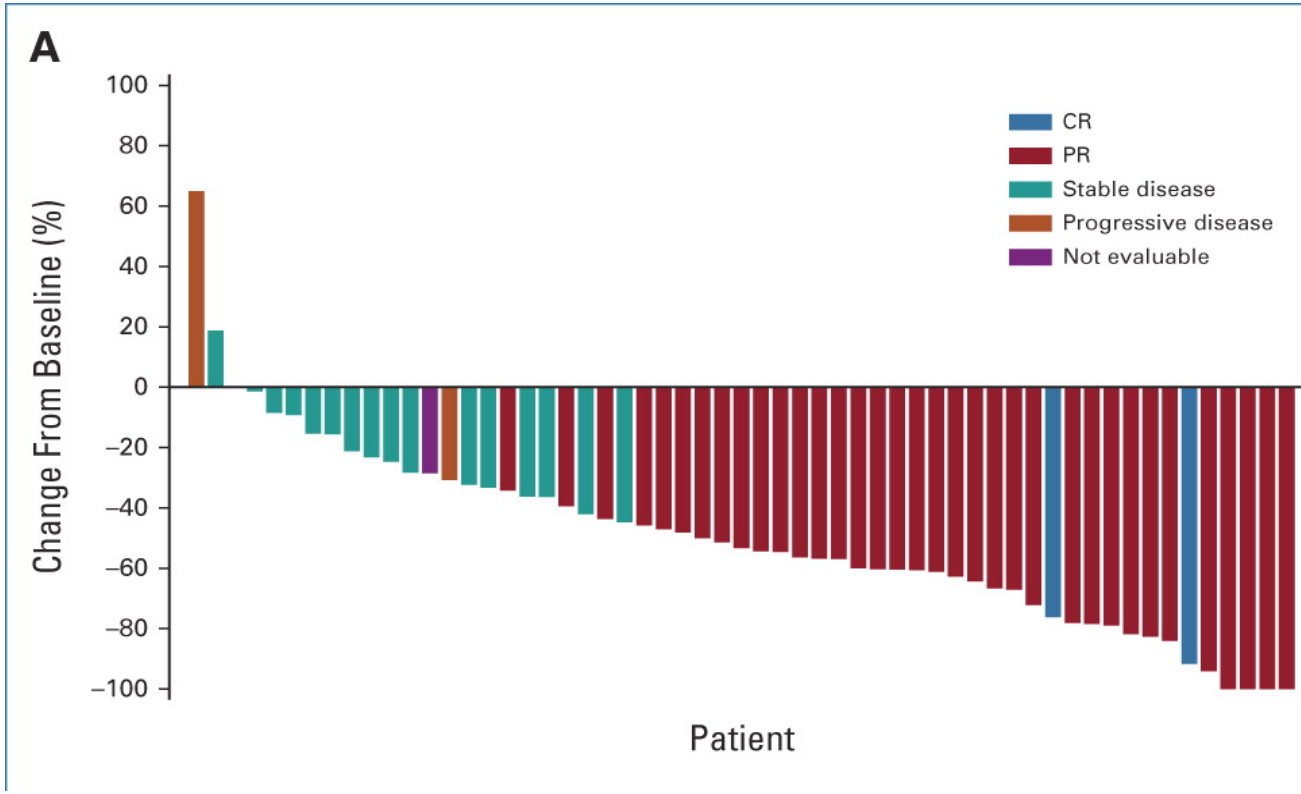
Cumulative Distribution

Box Plots



Waterfall Plots

- Not every bar chart is a histogram
- Waterfall plot: A popular way of displaying response rates
- https://ascopubs.org/doi/full/10.1200/JCO.23.00774



Surgical Data Example (Survival Time)

- Dead/Alive
  - A binary variable
- If we only use dead/alive (very common misleading practice) time element is completely ignored
  - Someone who is alive with 6 months of follow-up is treated the same way with someone who is alive with 10 years of follow-up; and differently than someone who died at 5 years

Surgical Data Example (Survival Time)

- Time Between Resection and Death (always define in papers)
  - A continuous variable with a time unit
- How to deal with those who are alive?
- Use time from resection to last follow-up?
  - We need to distinguish which times are survival times and which are follow-up times
- Hence we need two columns to represent survival data
  - Time to death or last follow-up
  - Status (dead/alive), equivalently an indicator of whether the time variable contains survival or follow-up

Censored variables

- This kind of variable is called censored in statistical literature
- Many people mistakenly believe censored means excluded, i.e. only survival times are analyzed
- No observations are excluded in censored data analysis, the word censoring refers to the situation that we have not observed death time for some patients (hence their death times are censored)
- The binary component of a survival variable is sometimes called survival indicator

Examples of Censored Variables

- Almost always they are time to an event
  - Death
  - Progression
  - Treatment
- Censored variables require special methods. They cannot be analyzed as if they are binary (just the censoring indicator column) or continuous (just the time column)
- Most common outcomes in cancer studies are censored
- Most incorrect statistical analysis in the literature pertain to censored variables

Three Musketeers of Statistics

- Point Estimation
- Interval Estimation
- Hypothesis Testing

- Prediction

- REVIEW

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 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
- Famous failures of sampling …

History Corner: 1948 Presidential Elections

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, most of the time
- Some parameters (like slope, hazard ratio) does not have so easily defined sample analogs
- Sometimes sample analogs are not great estimates, but we will ignore that now (famous example: standard deviation)

Maximum Likelihood Estimates

- We use something called maximum likelihood to produce estimates for them
- 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.

Confidence Interval

- An interval that is likely to contain the true value of the parameter
- More precisely, a 95% confidence interval means
  - If the data were to be collected again under “identical” conditions
  - And a confidence interval is formed for every one of these data sets
  - Then 95% of these intervals will contain the true value
- You observe only one data set and one confidence interval
- One interval out of many, 95% of which would contain the true value, is likely to contain the true value

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

Confidence Interval Brain Teaser

- For a given data set and a variable I give you two confidence intervals
- One is wider than the other
- Which one has higher confidence level? Narrower or wider?
- Example: two intervals for age
  - 63.3 – 64.9 (wider)
  - 63.5 – 64.7 (narrower)

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?

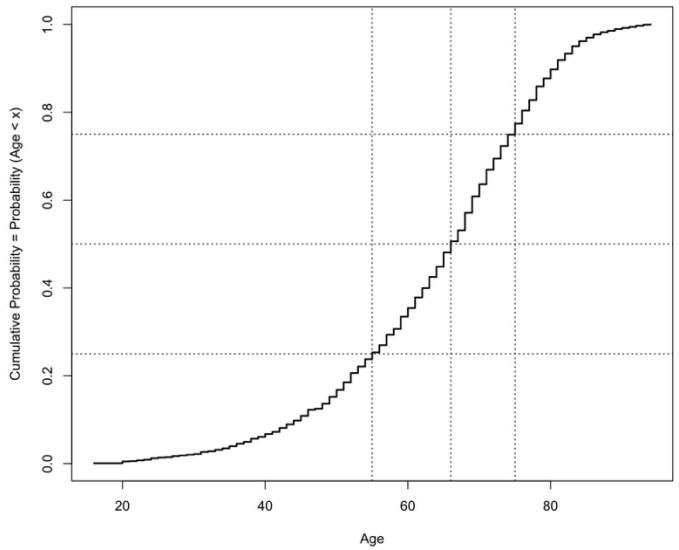
- Produce a confidence interval and see if it is entirely contained in one of the hypothesized intervals or not.
- If it is then we rule in favor of that hypothesis
- In this example, confidence interval is 0.7 – 0.92, entirely contained within $r>0.5$, hence we conclude $r>0.5$
- What is the interval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

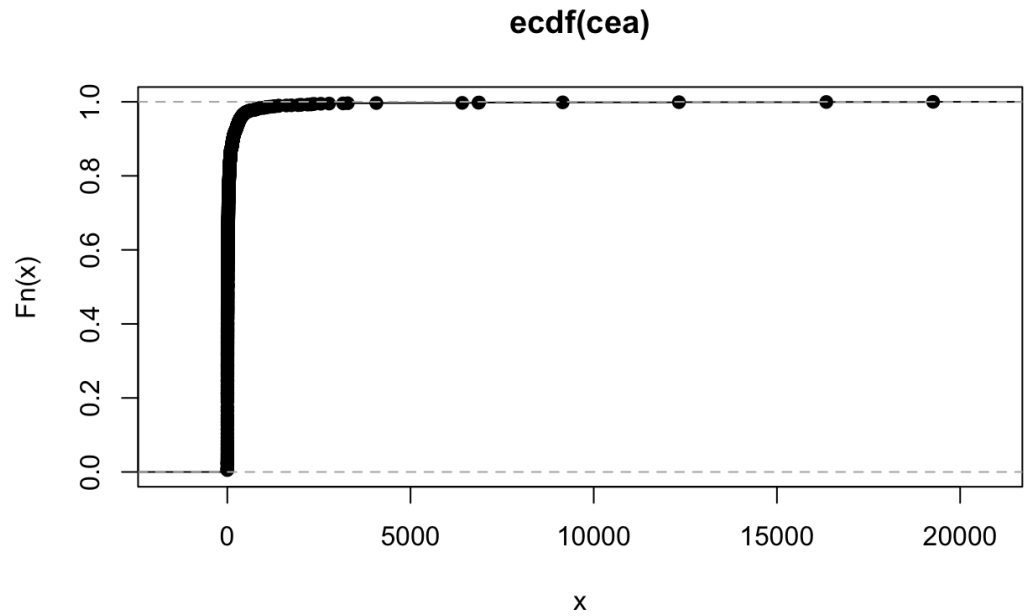
Asymmetry of hypothesis testing

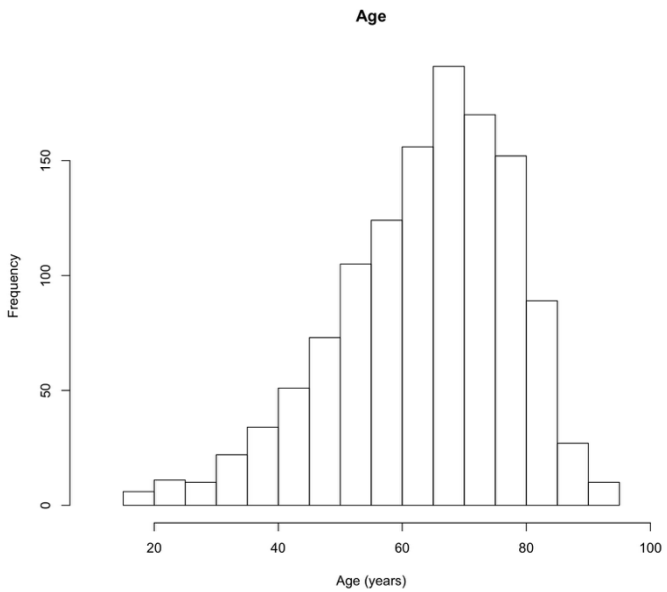
- r<=0.50 vs r>0.50 – each is a hypothesis. One of them we want to disprove (to be called the null hypothesis, or H_0) and the other we want to prove (alternative hypothesis, H_1).
- They are not symmetrical for reasons we will discuss over and over in this class
- As long as our interval estimate contains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

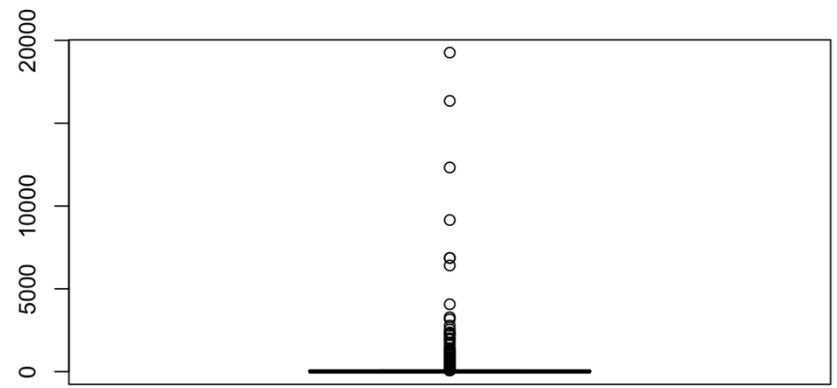
Another way of testing a hypothesis

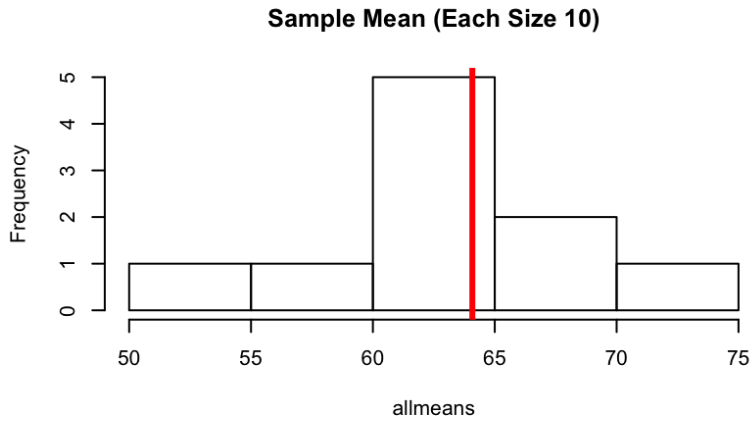
- Generate a p-value (to be defined next week) from the data
- If $p < 0.05$ the conclude alternative hypothesis is consistent with the data, otherwise conclude null hypothesis is still the best thing we have
- We have many discussions coming on this alternative, very very popular and infamous method

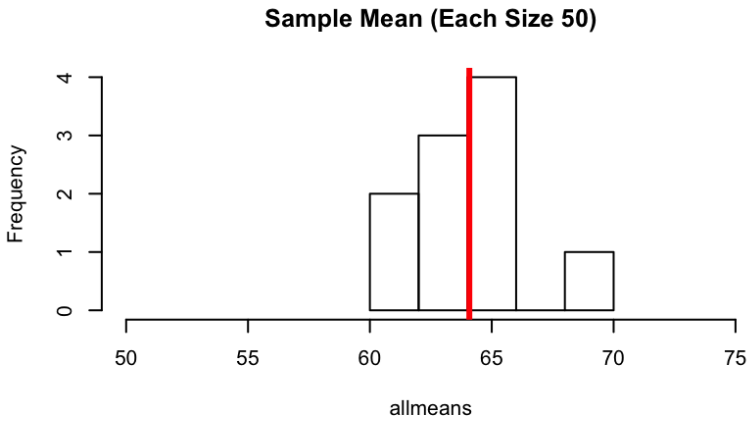

Cumulative Probability = Probability (Age < x)
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
Age

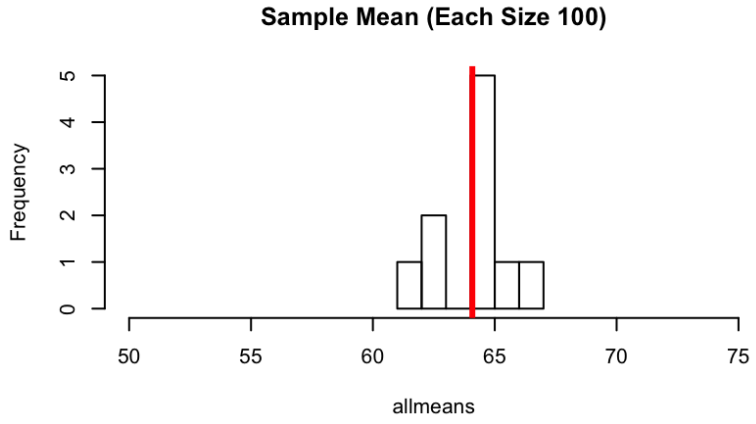

ecdf(cea)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
x

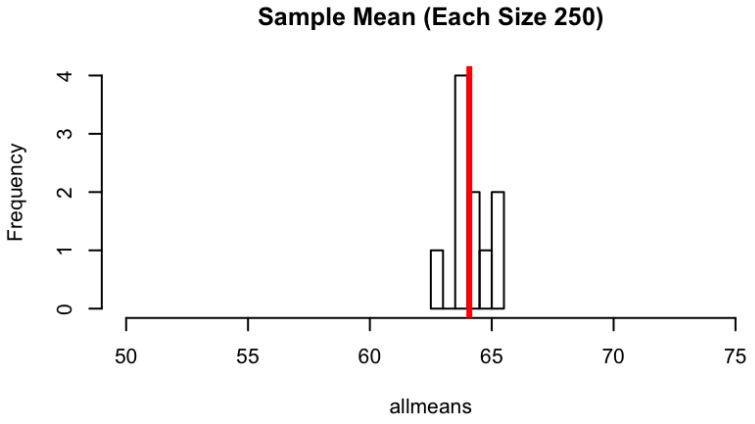

Age
Frequency
0
50
100
150
20
40
60
80
100
Age (years)

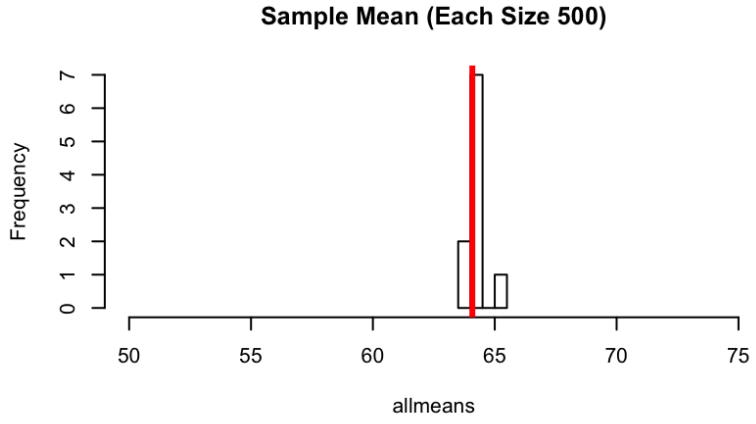

0
5000
10000
20000

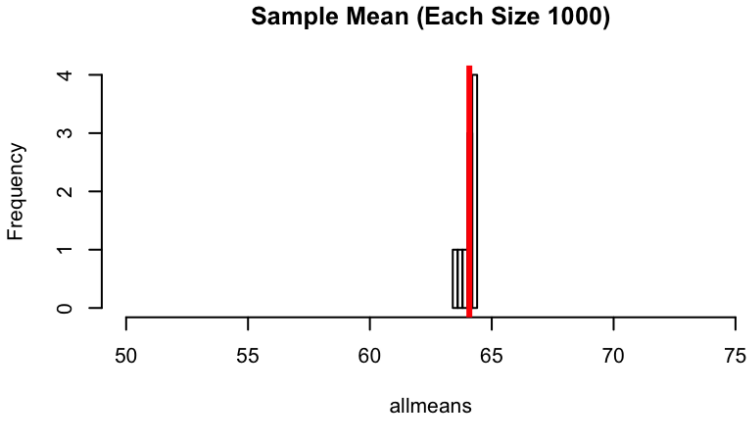

Sample Mean (Each Size 10)
Frequency
0
1
2
3
4
5
50
55
60
65
70
75
allmeans

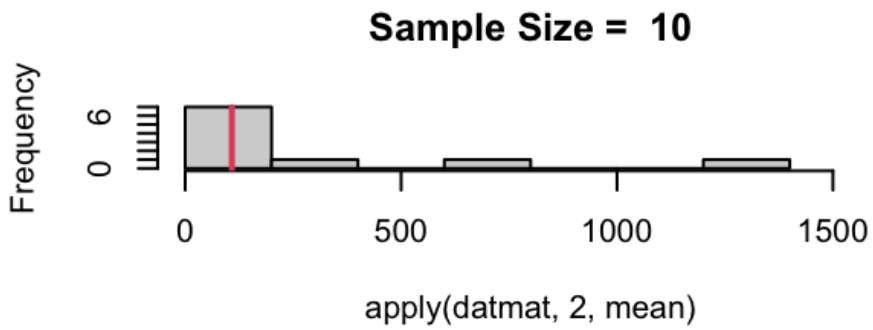

Sample Mean (Each Size 50)
Frequency
0
1
2
3
4
50
55
60
65
70
75
allmeans

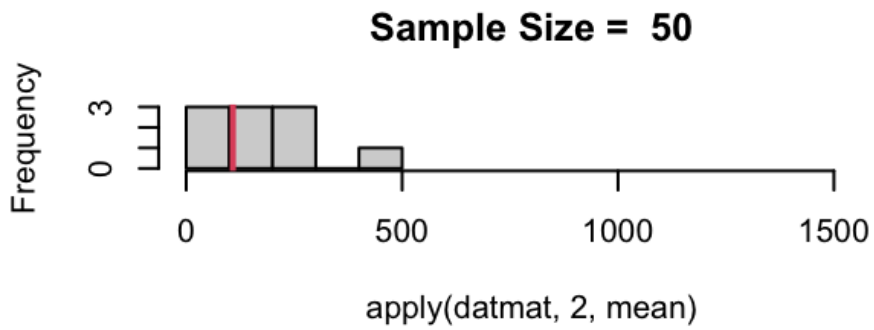

Sample Mean (Each Size 100)
Frequency
0
1
2
3
4
5
50
55
60
65
70
75
allmeans

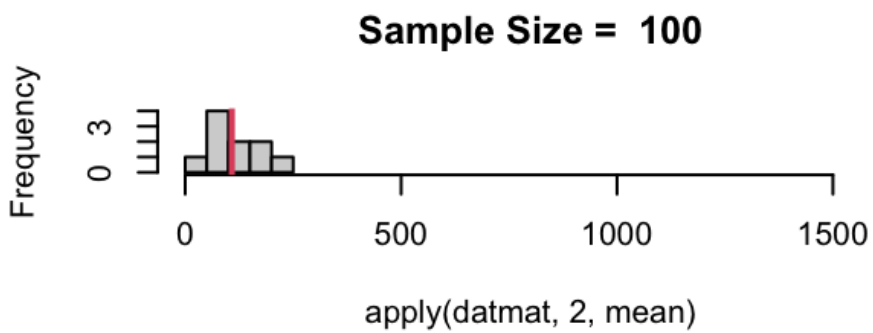

Sample Mean (Each Size 250)
Frequency
0
1
2
3
4
50
55
60
65
70
75
allmeans

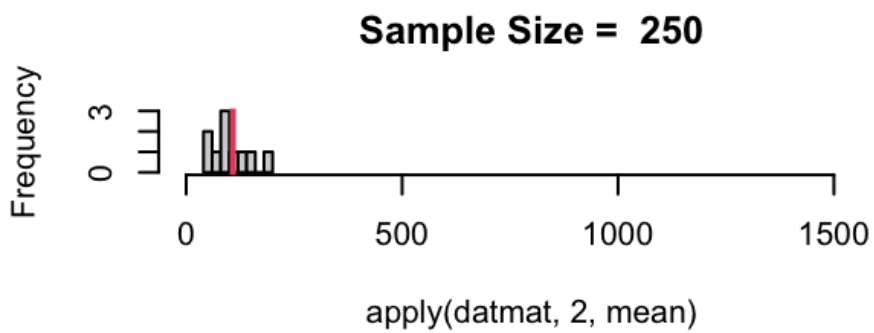

Sample Mean (Each Size 500)
Frequency
0
1
2
3
4
5
6
7
50
55
60
65
70
75
allmeans

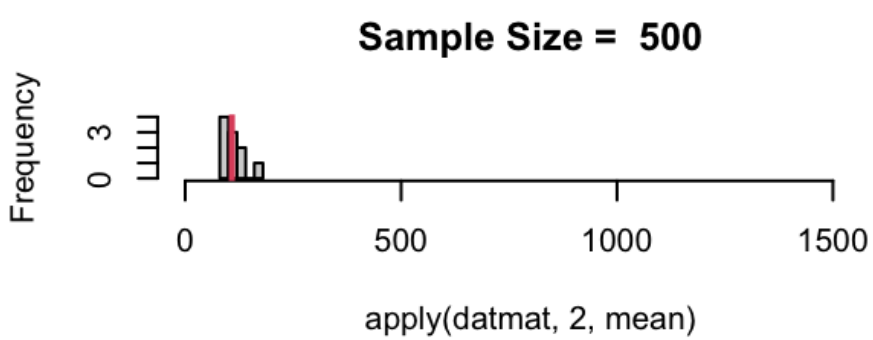

Sample Mean (Each Size 1000)
Frequency
0
1
2
3
4
50
55
60
65
70
75
allmeans

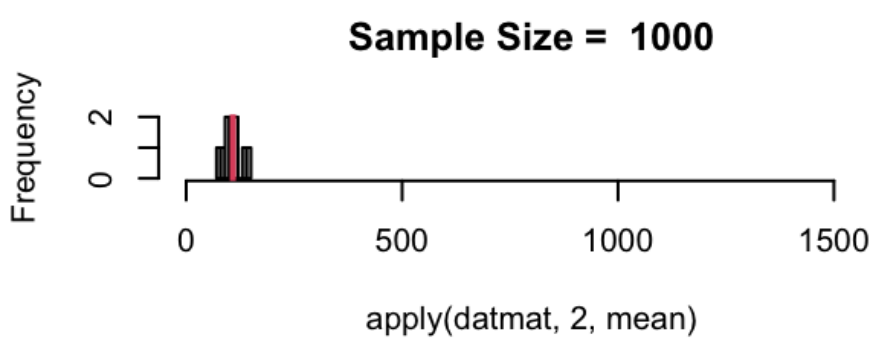

Sample Size = 10
Frequency
0
6
0
500
1000
1500
apply(datmat, 2, mean)

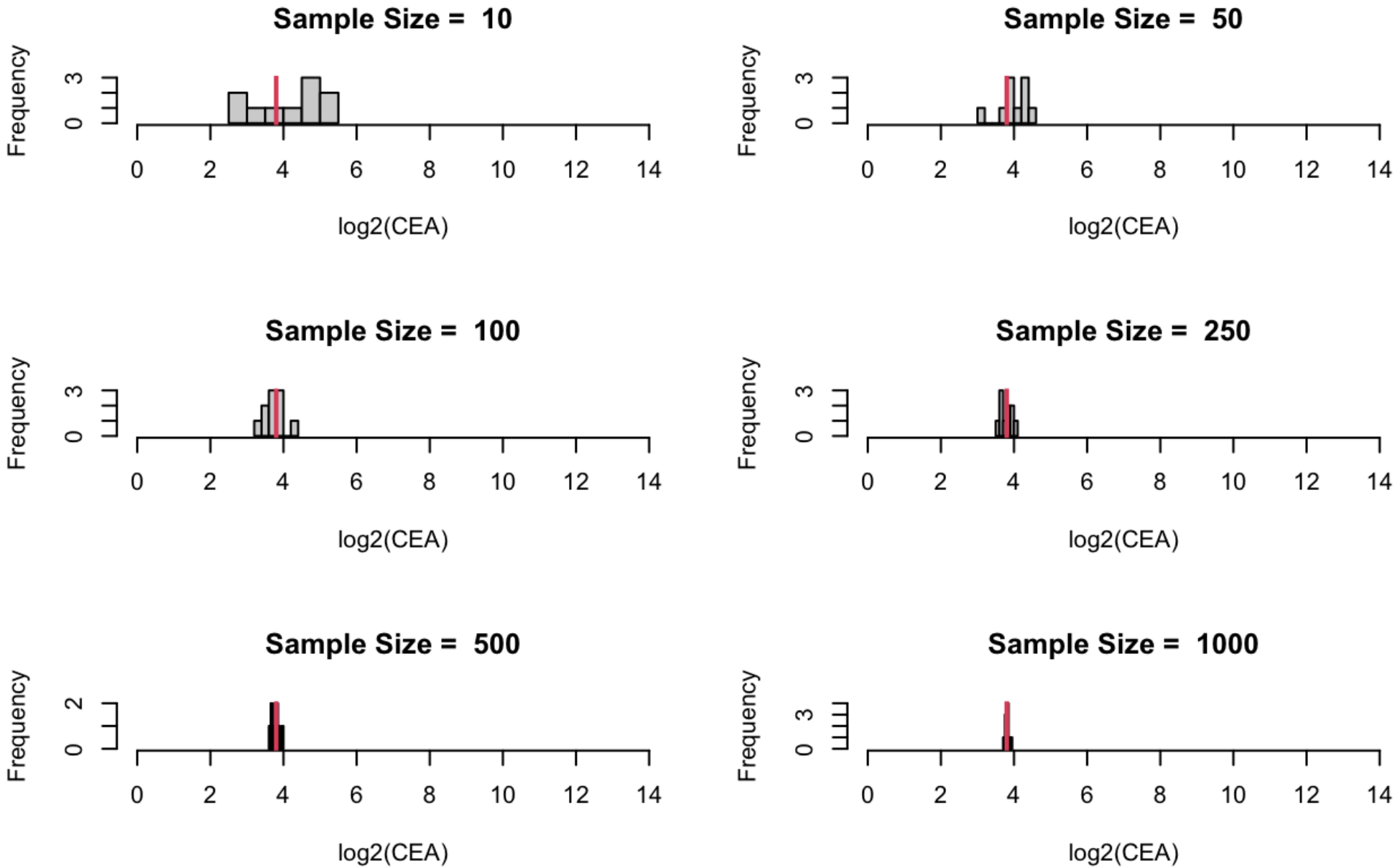

Sample Size = 50
Frequency
0
3
0
500
1000
1500
apply(datmat, 2, mean)


Sample Size = 100
Frequency
0
3
0
500
1000
1500
apply(datmat, 2, mean)


Sample Size = 250
Frequency
0
3
0
500
1000
1500
apply(datmat, 2, mean)


Sample Size = 500
Frequency
0
3
0
500
1000
1500
apply(datmat, 2, mean)


Sample Size = 1000
Frequency
0
2
0
500
1000
1500
apply(datmat, 2, mean)


Sample Size = 10
Frequency
0 3
0 2 4 6 8 10 12 14
log2(CEA)
Sample Size = 50
Frequency
0 3
0 2 4 6 8 10 12 14
log2(CEA)
Sample Size = 100
Frequency
0 3
0 2 4 6 8 10 12 14
log2(CEA)
Sample Size = 250
Frequency
0 3
0 2 4 6 8 10 12 14
log2(CEA)
Sample Size = 500
Frequency
0 2
0 2 4 6 8 10 12 14
log2(CEA)
Sample Size = 1000
Frequency
0 3
0 2 4 6 8 10 12 14
log2(CEA)