Basic Biostatistics for Clinicians: How to Use and Interpret Statistics (for the boards)

Elizabeth Garrett-Mayer, PhD
Associate Professor
Director of Biostatistics
Hollings Cancer Center

Outline for today's talk

1. Experimental design
2. Motivating example
3. Types of variables
4. Descriptive statistics
5. Population vs. sample
6. Confidence intervals
7. Hypothesis testing
8. Type I and II errors

Experimental Design

- How do we set up the study to answer the question?
- Two main situations
  - Controlled designs
    - The experimenter has control
    - “exposure” or “treatments”
    - Randomized clinical trials
  - Observational designs
    - Cohort studies
    - Case-control studies

Controlled Designs

- Not necessarily randomized
- E.g. Cancer research
  - Phase I: dose finding
  - Phase II: single arm efficacy
  - Phase III: randomized design
- The “experimenter” dictates
- Gold-standard: RCT
  - Controls biases
  - “balances” treatment arms

Observational studies: Cohort

- Process:
  - Identify a cohort
  - Measure exposure
  - Follow for a long time
  - See who gets disease
  - Analyze to see if disease is associated with exposure
- Pros
  - Measurement is not biased and usually measured precisely
  - Can estimate prevalence and associations, and relative risks
- Cons
  - Very expensive
  - Very very expensive if outcome of interest is rare
  - Sometimes we don't know all of the exposures to measure

Observational Studies: Case-Control

- Process:
  - Identify a set of patients with disease, and corresponding set of controls without disease
  - Find out retrospectively about exposure
  - Analyze data to see if associations exist
- Pros
  - Relatively inexpensive
  - Takes a short time
  - Works well even for rare disease
- Cons
  - Measurement is often biased and imprecise (‘recall bias’)
  - Cannot estimate prevalence due to sampling

Observational Studies: Why they leave us with questions

- Confounders
- Biases
  - Self-selection
  - Recall bias
  - Survival bias
  - Etc.

Motivating example

- The **primary goal** of this study is to determine whether epsilon aminocaproic acid (EACA) is an effective strategy to reduce the morbidity and costs associated with allogeneic blood transfusion in adult patients undergoing spine surgery. (Berenholtz)

- Comparative study with EACA arm and placebo arm.
- Randomized
- N=182 (91 patients per arm)
- Investigators would be interested in regularly using EACA if it could reduce the number of units transfused by 30% comparing placebo to EACA

Study Endpoints

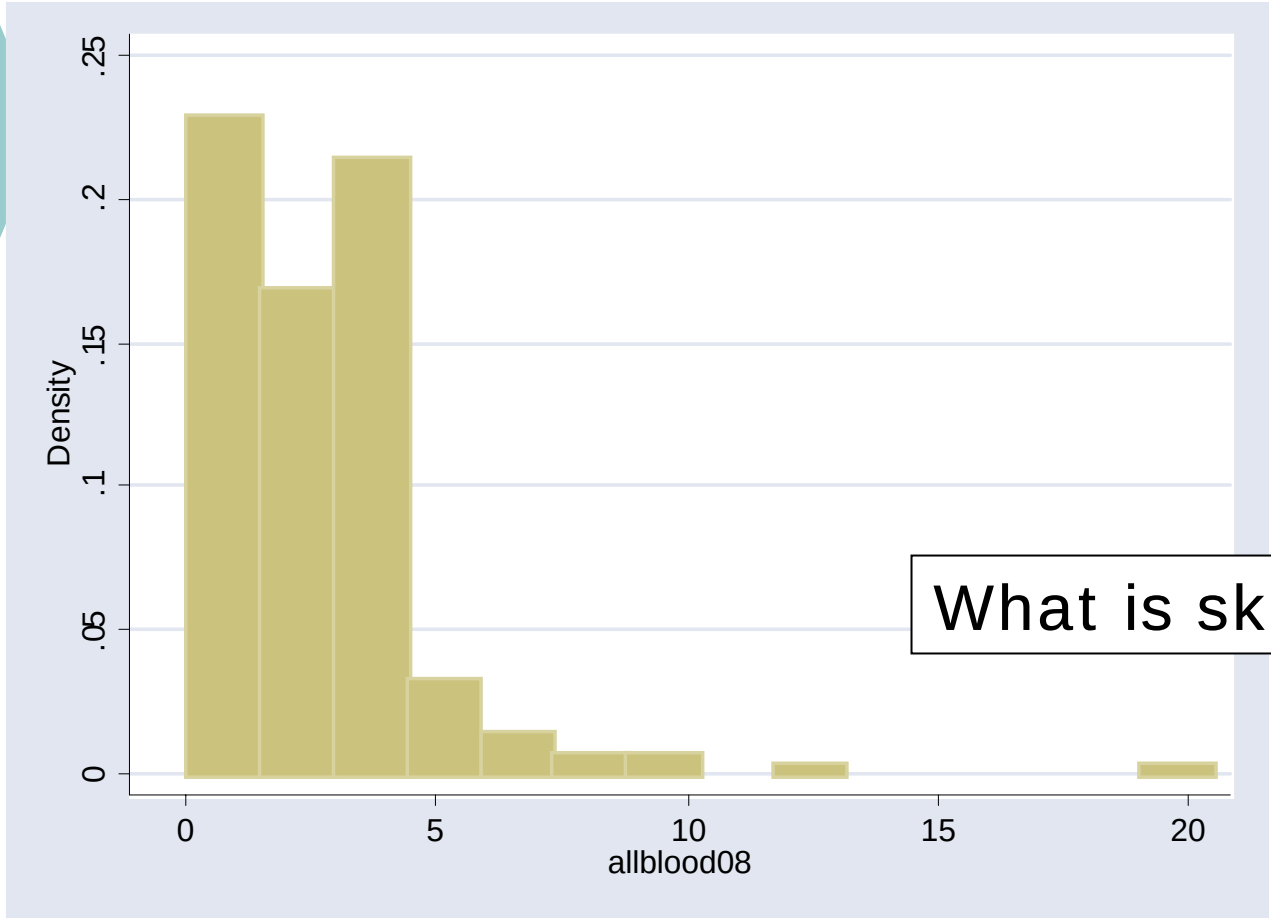
	Intraoperative	Post-surgery to 48 hours	>48 hours through day 8	Total
Allo	s	s	s	**P**
Auto	s	s	s	s
Allo + Auto	s	**P**		s

FFP	s	s	s	s
Platelets	s	s	s	s
All products	s	s	s	s

Three Primary Types of Variables in Medical Research

- continuous:
  - blood pressure
  - cholesterol
  - quality of life
  - units of blood
- categorical
  - blood type
  - transfused/not transfused
  - cured/not cured
- time-to-event
  - time to death
  - time to progression
  - time to immune reconstitution
  - time to discharge(?)

Descriptive Statistics (and Graphical Displays)

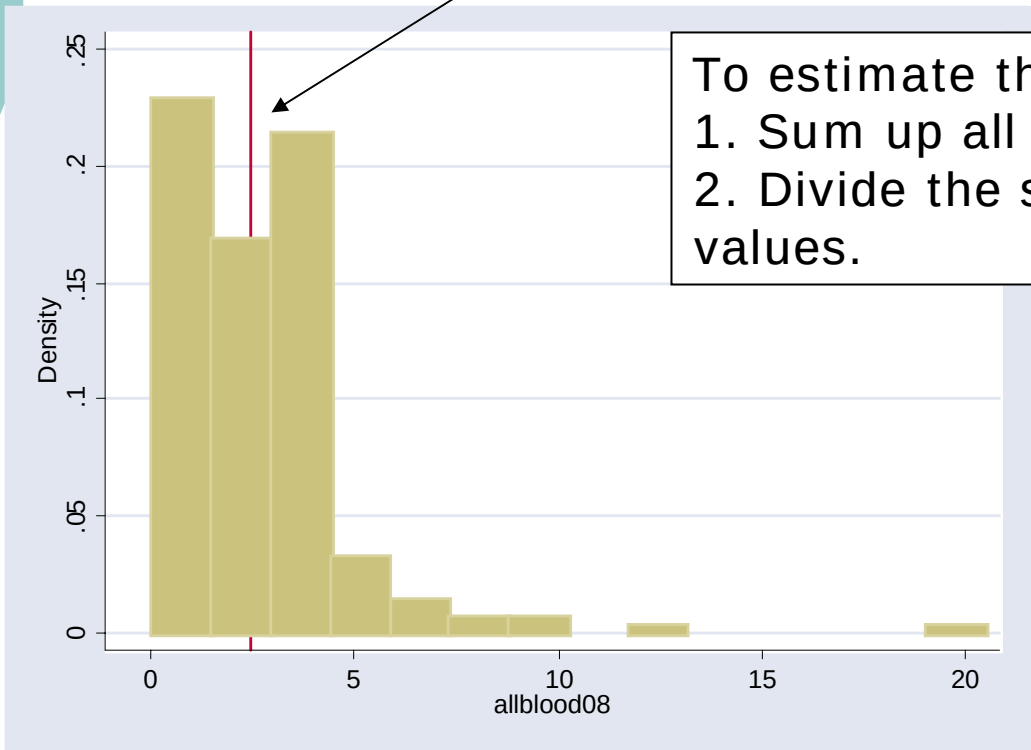


What is skewness?

Allo+Auto Units, post-op through Day 8

The Mean: The statistical average.

2.42 units

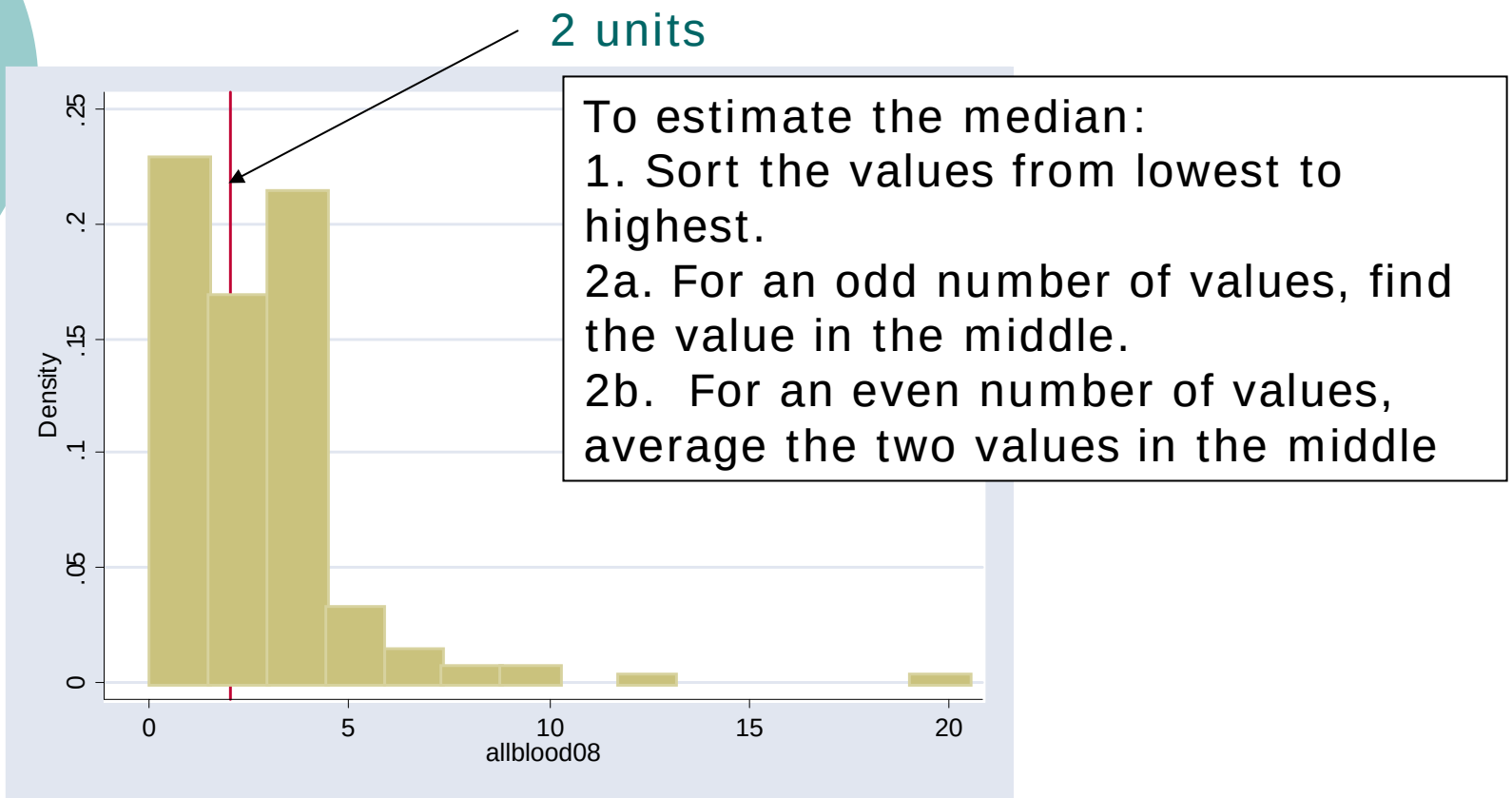


To estimate the mean:
1. Sum up all the values.
2. Divide the sum by the number of values.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Allo+Auto Units, post-op through Day 8

The Median: The “middle” value



To estimate the median:
1. Sort the values from lowest to highest.
2a. For an odd number of values, find the value in the middle.
2b. For an even number of values, average the two values in the middle

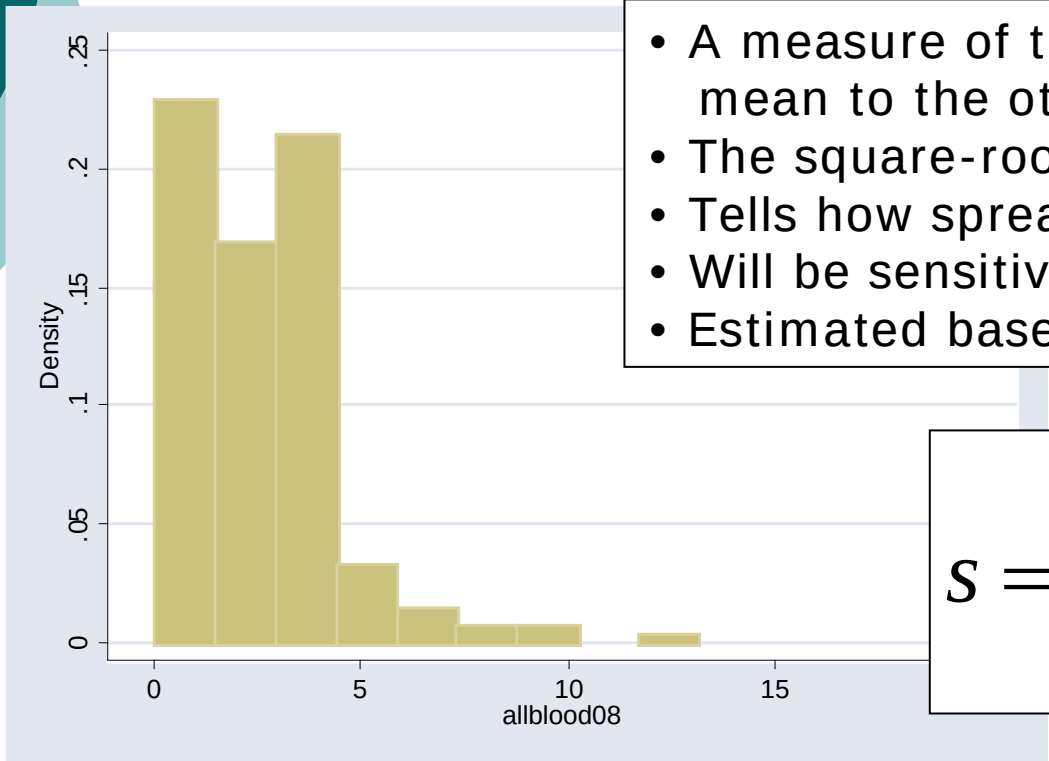
Allo+Auto Units, post-op through Day 8

The mean versus the median

- The mean is sensitive to “outliers”
- The median is not

- When the data are highly skewed, the median is usually preferred
- When the data are not skewed, the median and the mean will be very close

The standard deviation: s = 2.3



- A measure of the distance from the mean to the other values.
- The square-root of the variance
- Tells how spread out the data are
- Will be sensitive to skewness
- Estimated based on a sample of data

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Allo+Auto Units, post-op through Day 8

Others

- Range
- Interquartile range
- Mode
- Skewness

What about categorical outcomes?

- Focus on binary:

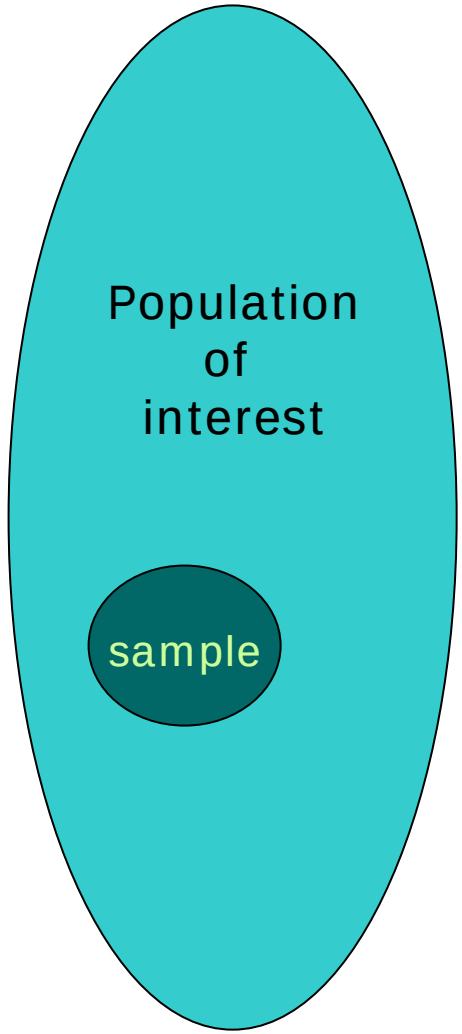
$$y = \begin{cases} 1 & \text{if units} > 5 \\ 0 & \text{if units} \leq 5 \end{cases}$$

- How do we summarize that?
- Usually just a proportion will do:

```
         y |      Freq.     Percent
------------+-----------------------------
         0 |       106       58.24
         1 |        76      41.76
------------+-----------------------------
     Total |       182      100.00
```

A key distinction: Population versus Sample

- We collect data from a population
- “sample”
- We use the data on the sample to make INFERENCES about the population
- We have a “sample” mean
- It is NOT the true mean, but it might be pretty close
- **How close depends on the size of the sample**



Parameters versus Statistics

- A parameter is a population characteristic
- A statistic is a sample characteristic
- Example: we estimate the sample mean to tell us about the true population mean
  - the sample mean is a ‘statistic’
  - the population mean is a ‘parameter’

Statistical Inference

- Use the data from the sample to inform us about the population
- "Generalize"
- Two common approaches
  - <u>confidence intervals</u>: tell us likely values for the true population value based on our sample data
  - <u>hypothesis testing</u>: find evidence for or against hypotheses about the population based on sample data

Confidence Intervals

- We want to know the true mean
- All we have is the sample mean.
- How close is the sample mean to the true mean?
- A confidence interval can tell us
- It is based on
  - the sample mean ($\bar{x}$)
  - the sample standard deviation (s)
  - the sample size (N)
  - (& the level of confidence)
- We usually focus on **95%** confidence intervals

Confidence Intervals

- What does it mean?
- **It is an interval which contains the TRUE population parameter with 95% certainty**
- How do we calculate it?
- First, we need to learn about the standard error

Standard Error

- A measure of the precision of the **sample statistic**
- For the sample mean:

$$se_{\bar{x}} = \frac{s}{\sqrt{N}}$$

- **Standard error ≠ standard deviation!**
- What is the difference?
  - The ***standard deviation*** is a measure of precision of the <u>population distribution</u>. Tells us what we could expect about individuals in the population.
  - The ***standard error*** is a measure of precision of a <u>sample statistic</u>. Tells us how precise our estimate of the parameter is.
- By increasing N, what happens to our estimate of
  - The standard error?
  - The standard deviation?

Confidence Intervals

- We use the standard error to calculate confidence intervals
- 95% confidence interval:

$$\bar{x} \pm \boxed{1.96} se_{\bar{x}}$$

multiplier

- Or,

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{N}}$$

What does it mean?

- **It is an interval that we are 95% sure contains the true population mean.**
- **It provides a "reasonable range" for the true population parameter**
- Example: EACA and placebo

$$Placebo:$$

$$\bar{x} = 2.81, s = 2.81, N = 91$$

$$2.81 \pm 1.96 \frac{2.81}{\sqrt{91}} = (2.23, 3.40)$$

$$EACA:$$

$$\bar{x} = 2.04, s = 1.83, N = 91$$

$$2.04 \pm 1.96 \frac{1.83}{\sqrt{91}} = (1.67, 2.41)$$

What about other levels of confidence?

- Might see 99% or 90%.
- Use a different multiplier
- For 99%, replace 1.96 with **2.58**
- For 90%, replace 1.96 with **1.645**

- **More confident: wider interval**
- **Less confident: narrower interval**

Caveats

- Validity of CI requires either
  - A relatively large sample size (>30-ish)
  - A normally distributed variable
  - (or both)
- EACA example:
  - Very skewed
  - But, N=91 per group
  - If N=25 instead, confidence interval would not be valid

Caveats

- For sample sizes <100, use "t-correction"
- Adjusts for imprecision in estimate of standard deviation
- Examples: for 95% CI
  - For N=20: multiplier = 2.09
  - For N=50: multiplier = 2.01
  - For N=100: multiplier = 1.98
  - For N=1000: multiplier = 1.96

Confidence Intervals

- **We can make confidence intervals for any parameter**
- We just need:
  - Sample estimate
  - Standard error of estimate
  - (a little theory)
- Example: proportion
- Width ALWAYS depends on sample size!!!

$$Placebo:$$

$$\hat{p} = \frac{46}{91} = 0.51$$

$$95\%\ \mathrm{CI} = (0.40, 0.61)$$

$$EACA:$$

$$\hat{p} = \frac{30}{91} = 0.33$$

$$95\%\ \mathrm{CI} = (0.23, 0.44)$$

Hypothesis Testing

- Helps us to choose between two conclusions:
  - The treatment did work versus did not work
  - There is an association versus there is not an association
- Setup usually looks formal (and involves Greek letters):
  - $\boxed{H_0: \ \mu_1 = \mu_2}$
  - $H_1: \ \mu_1 \neq \mu_2$

Null distribution

- In words:
  - The population mean in group 1 (placebo) is **<u>the same</u>** as in group 2 (EACA)
  - The population mean in group 1 (placebo) is **<u>different</u>** in group 2 (EACA)

“Null” distribution

- The “it didn’t work” distribution
- “there is no association”
- “the means are the same”
- “there is no difference”
- **We generally try to disprove the null**

Continuous outcomes: t-test

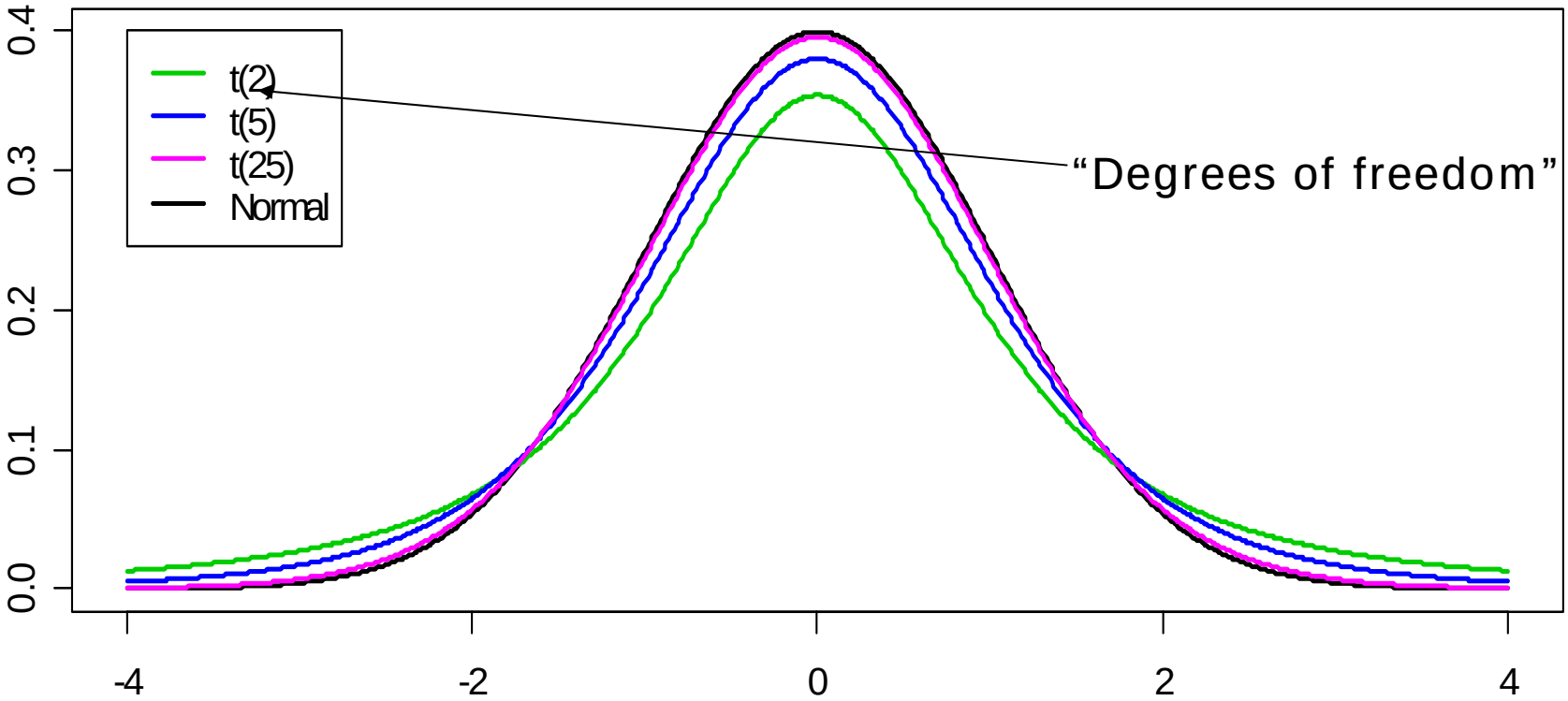
- Several kinds of t-tests
  - Two sample
  - one sample
  - Paired
- EACA Example: two independent groups $\rightarrow$ two sample t-test
- Construction of test statistic depends on this

Two-sample t-test

- No mathematics here
- Just know that the following are included:
  - The means of both groups
  - The standard deviations of both groups
  - The sample sizes of both groups
- Plug it all in the formula and….
- Out pops a number: **the t-statistic**
- We then compare the t-statistic to the appropriate t-distribution

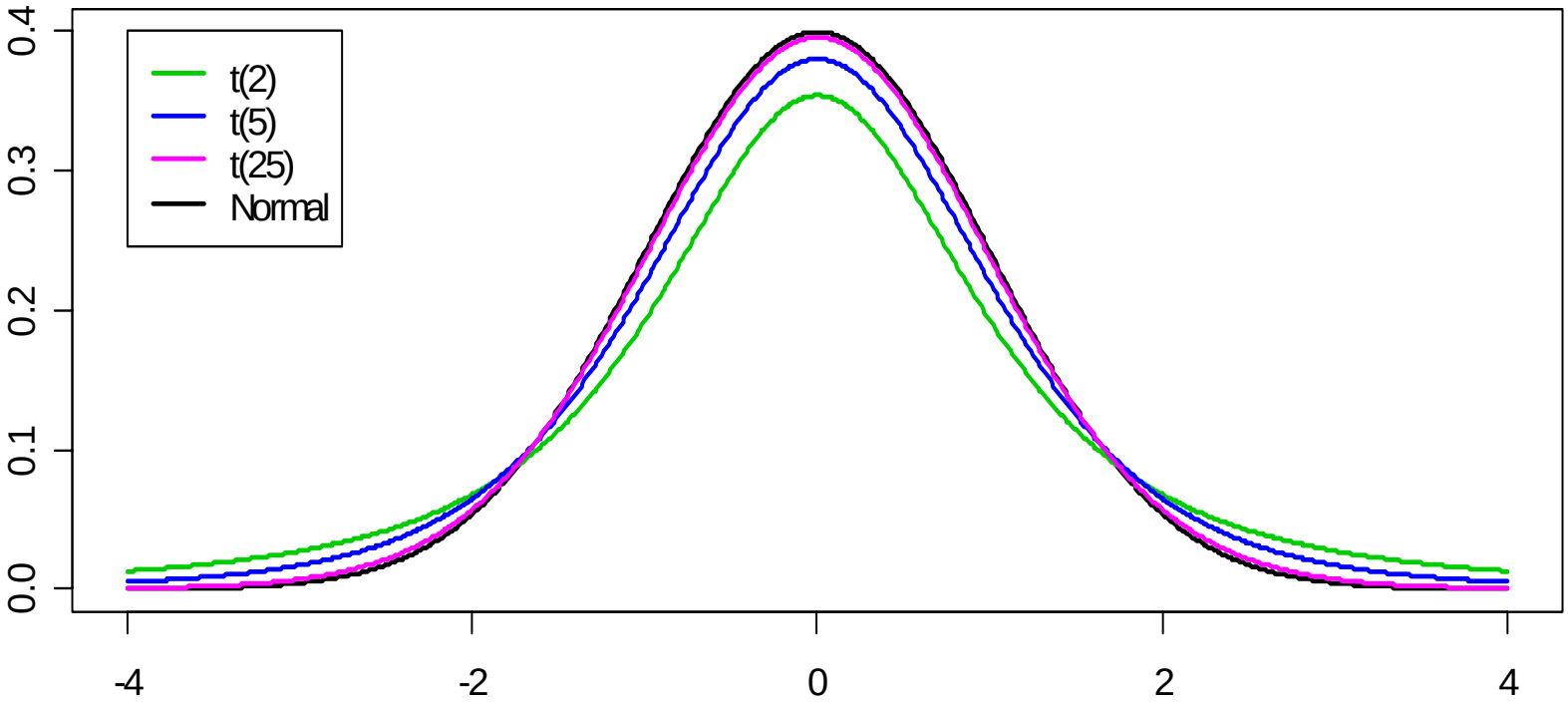
T-distribution

- Looks like a standard normal distribution (mean=0, s=1)
- Remember the t-correction?
- The larger the sample size, the narrower the t-distribution



T-distribution

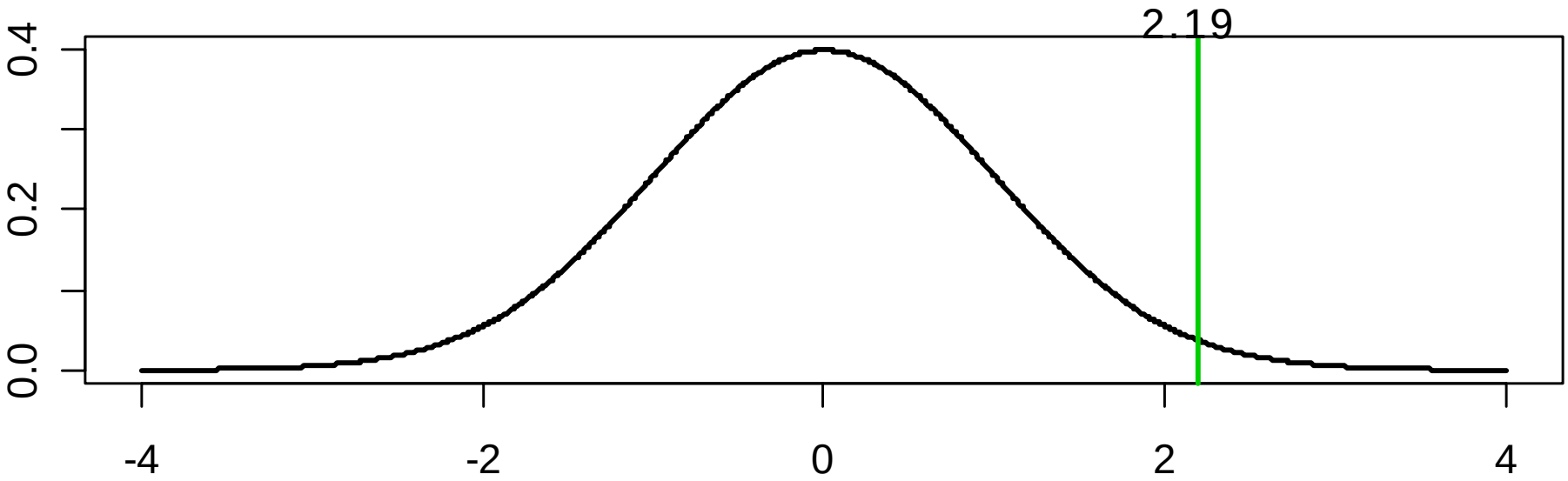
- Represents the “null” distribution
- Observations in the ‘bulk’ of the curve are things that would be common if the null were true
- “extreme” observations are rare under the null



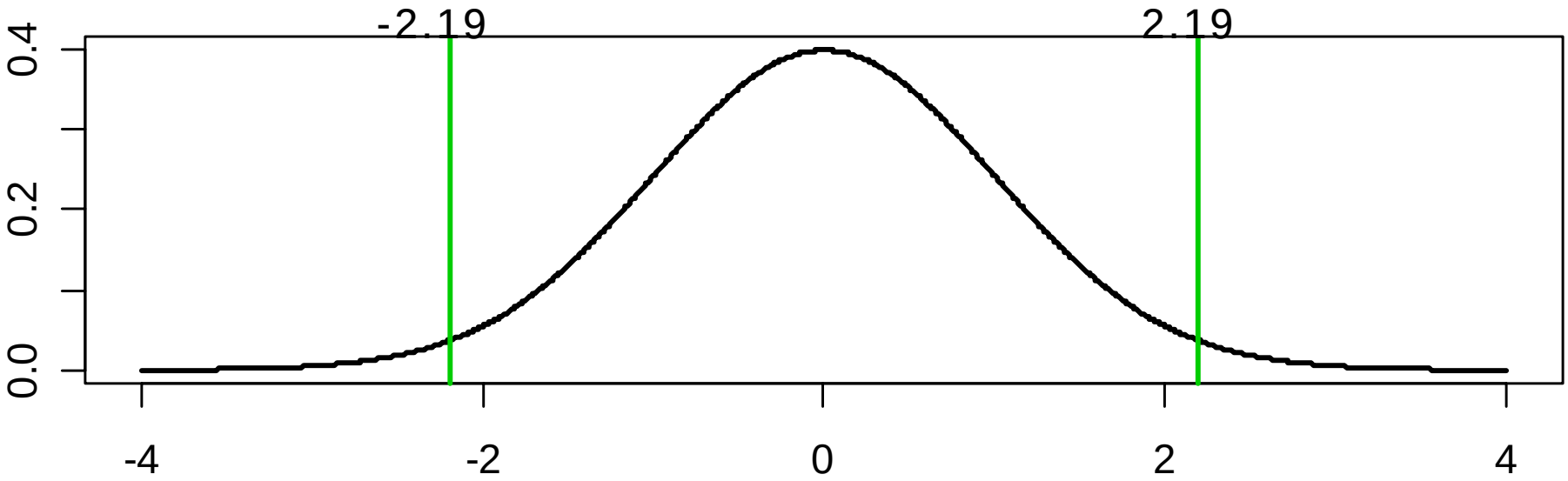
EACA and placebo

- Two-sample t-test
- t-statistic = 2.19
- Total N=182
- Use N to determine “degrees of freedom”
- Relationship between N and degrees of freedom depends on type of t-test
  - One sample: N-1
  - Two sample: N-2 or other…

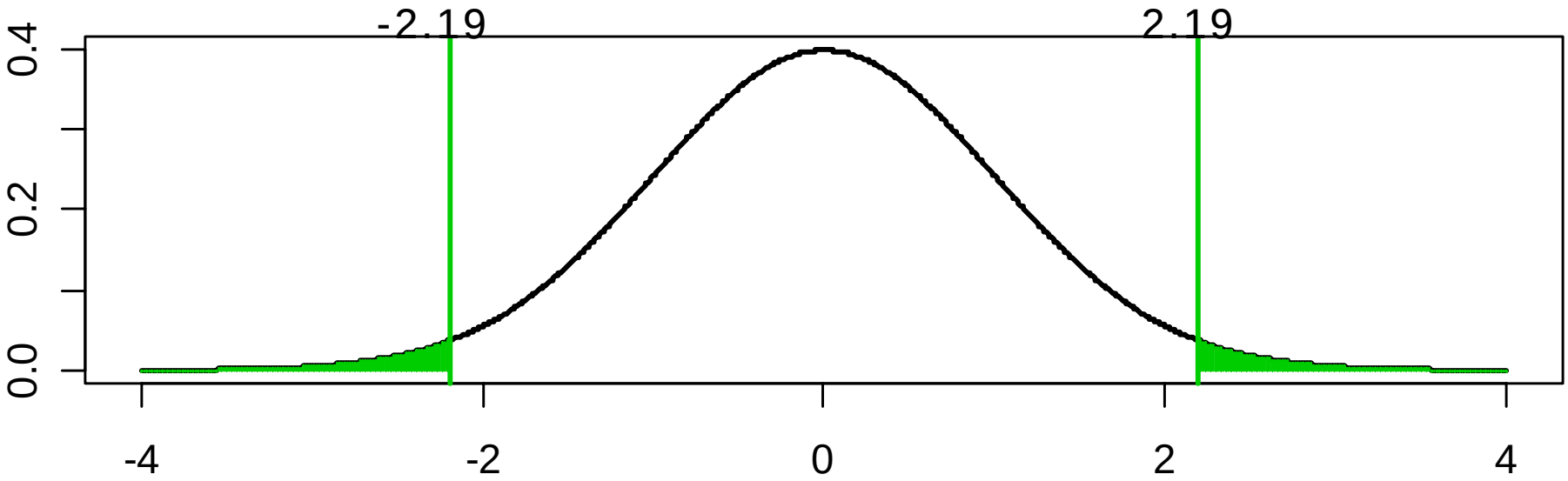
1. Choose the appropriate t-distribution
2. Locate t-statistic on x-axis



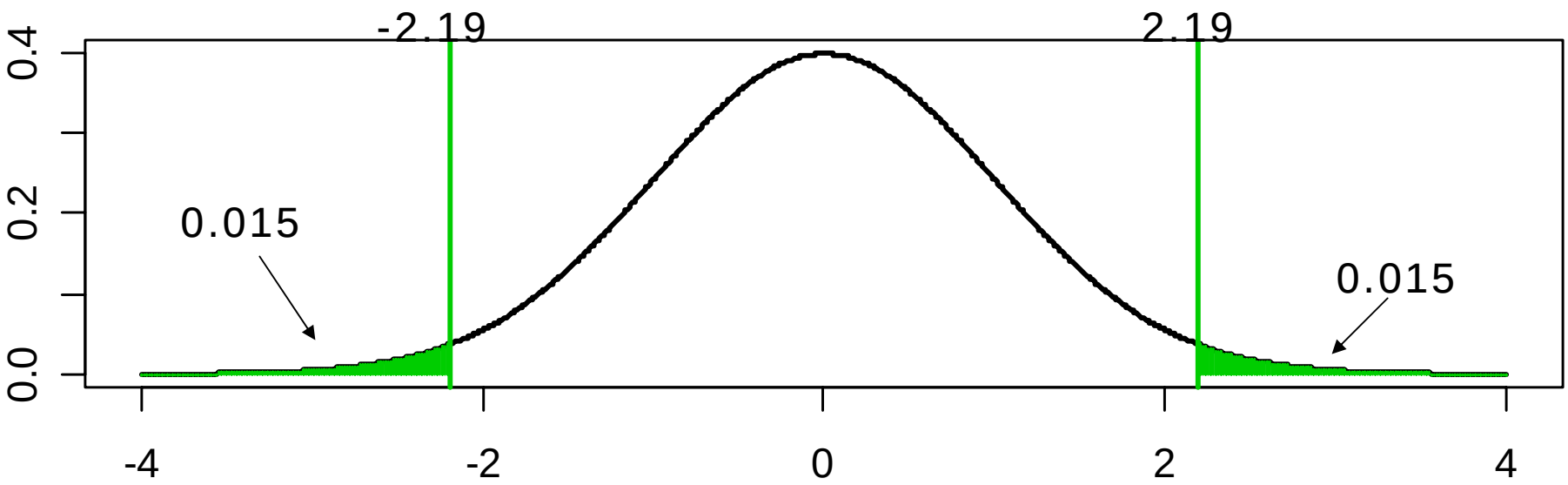
1. Choose the appropriate t-distribution
2. Locate t-statistic on x-axis
3. Locate -1*t-statistic on x-axis



1. Choose the appropriate t-distribution
2. Locate t-statistic on x-axis
3. Locate -1*t-statistic on x-axis
4. Identify area that is ‘more extreme’ in the tails of the t-dist’n



1. Choose the appropriate t-distribution
2. Locate t-statistic on x-axis
3. Locate -1*t-statistic on x-axis
4. Identify area that is ‘more extreme’ in the tails of the t-dist’n
5. Calculate green area



The p-value

- sum of the green area = P-VALUE
- EACA vs. Placebo: p-value=0.03
- What does that mean?
  - Version 1: “If the null hypothesis were true, the probability of seeing something as or more extreme than we saw is 0.03”
  - Version 2: “There is a less than 3% chance of seeing something this or more extreme if the two groups truly have the same means.”

The p-value **IS NOT**

- The probability that the null is true
- The probability of seeing the data we saw
- Key issues to remember:
  - "...as or more extreme..."
  - "..if the null is true..."
  - Statistic is calculated based on the null distribution!

What about proportions?

- T-tests are ONLY for continuous variables
- There are other tests for proportions:
  - Fisher's exact test
  - Chi-square tests
- P-values always mean the same thing regardless of test: the probability of a result as or more extreme under the null hypothesis
- Example: comparison of proportions
  - 0.50 and 0.33 in placebo and EACA
  - p-value = 0.02

Now what?

- What do we do with the p-value?
- We need to decide if 0.03 is low enough to ‘reject’ the null
- General practice:
  - Reject the null if $p<0.05$
  - “fail to reject” the null if $p>0.05$
- AD HOC cutoff
- **DEPENDS HEAVILY ON SAMPLE SIZE!!!!!!!!!!!!**

Type I error (alpha)

- The “significance” cutoff
- General practice: alpha = 0.05
- Sometimes:
  - alpha = 0.10
  - Alpha = 0.01
- Why might it differ?
  - Phase of study
  - How many hypotheses you are testing

Interpretation of Type I error

- The probability of FALSELY rejecting the null hypothesis
- Recall, 5% of the time, you will get an “extreme” result if the null is true
- People worry a lot about making a type I error
- That is why they set it pretty low (5%)

Type II error

- The opposite of type I error
- “the probability of failing to reject the null when it is true”
- People don’t worry about this so much
- Happens all the time
- Why?
- Because sample size is too small: not enough evidence to reject the null
- How can we ensure that our sample size is large enough? Power calculations

QUESTIONS???

- Contact me:

Elizabeth Garrett-Mayer
garrettm@musc.edu

- Other resources
  - Glaser: High-Yield Biostatistics
  - Norman & Streiner: PDQ Statistics
  - Dawson-Saunders & Trapp: Basic Biostatistics