Lecture 5: Distributions

Intro Stats with Nathan Favero

February 17, 2023

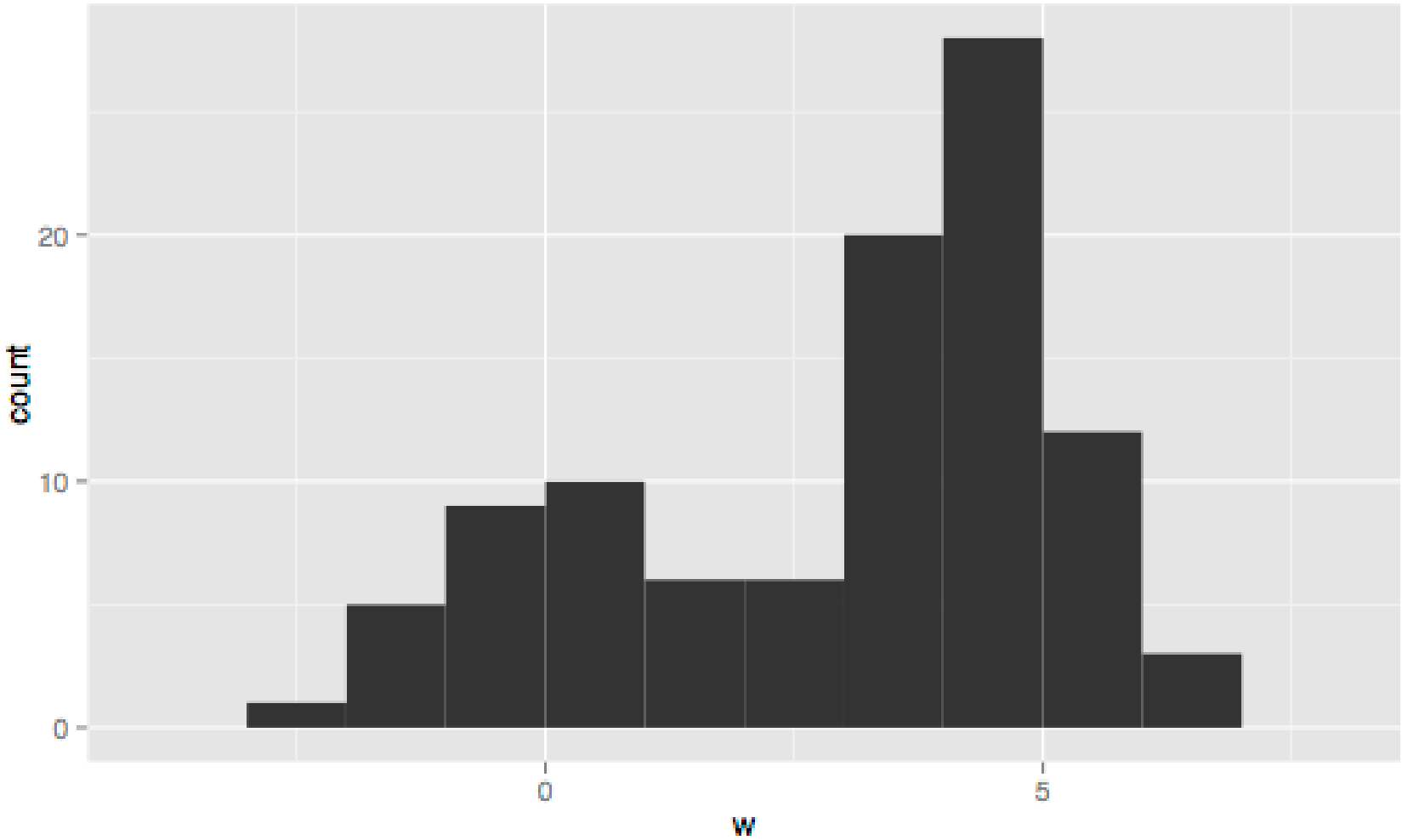
AMERICAN UNIVERSITY
WASHINGTON, DC

Describing distributions

- Distribution: all of the values (and the frequency of each of these values) for a variable
  - Distributions can be depicted in tables or figures
  - Descriptive statistics (e.g., mean, median, variance) can be used to describe a distribution

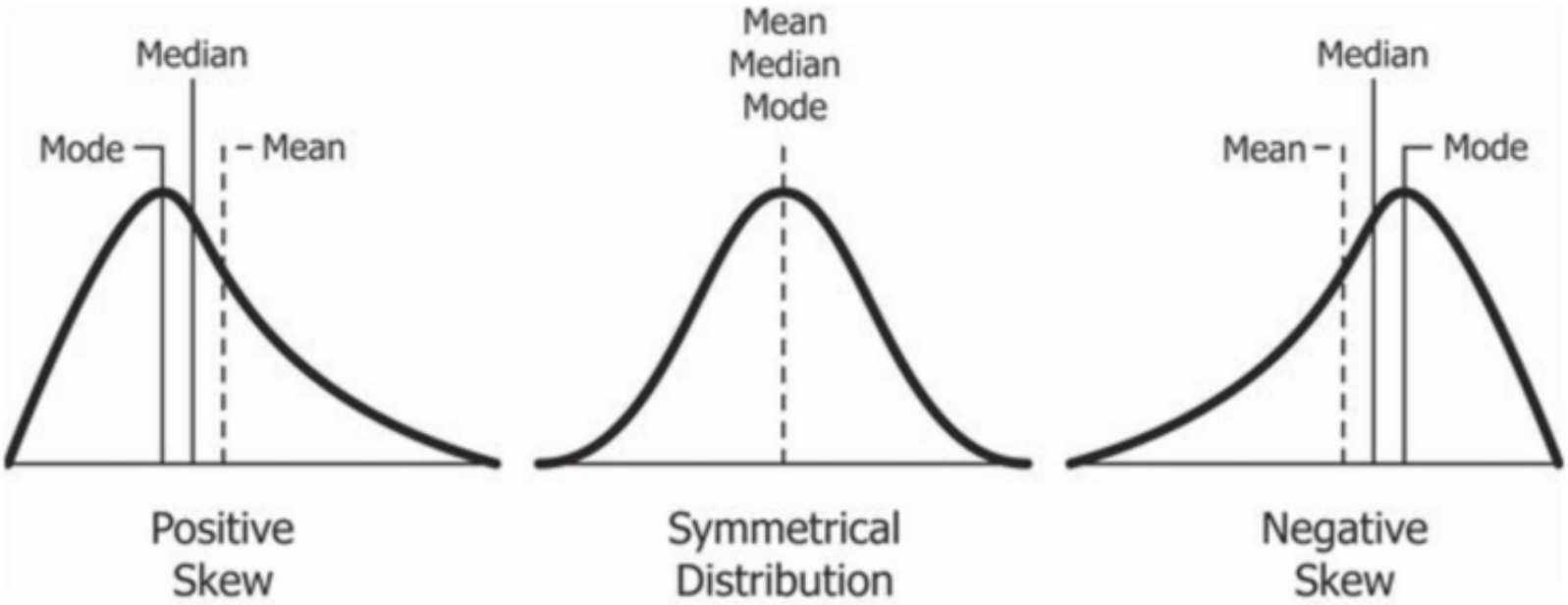
Describing distributions

- Unimodal: has only one (local) mode
- Multimodal: has more than one (local) mode
- Bimodal: has two (local) modes



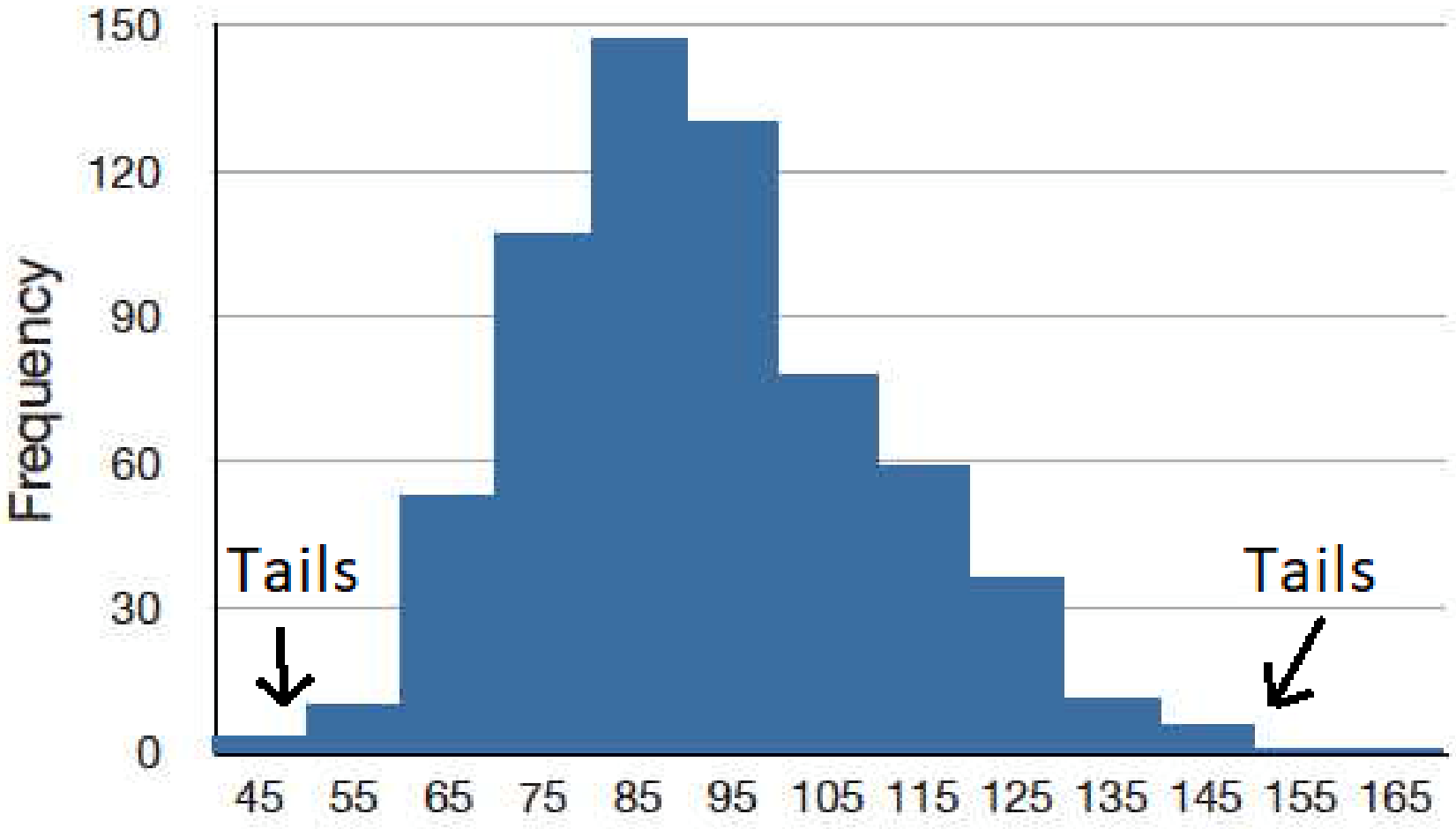
Describing distributions

- Skew: asymmetry in a distribution
- Positively skewed (or right-skewed) : the mean is higher than the median (opposite is negatively skewed)



Describing distributions

- Tails: ends of the distribution (highest and lowest values), typically describing ranges of values where there are few observations (short bars in a histogram)



Probability distributions

- To understand the models we use to do statistical inference, we first need to get familiar with the language and logic of probability distributions
- We learned before that a distribution describes how frequently every possible value for a variable occurs in a dataset (i.e., what you see when you use the `tab` command in Stata)
- *Probability* distributions don't describe data we've collected; instead, they describe a random process and indicate how likely we are to obtain different possible values from this random process
- Before proceeding, you might want to check out the appendix with an example of applying key concpets from this lecture: example: ER visits

Probability distributions

- Probability is going to be central to the way that we talk about data and distributions
- Using probabilities to describe things allows us to acknowledge that there are things we can't know... that we're not looking to predict or explain things with 100% certainty
- Random variable: has a set of possible values that can be described by a probability distribution
- Probability distribution: a list of all possible values and the associated probabilities

Probability distributions

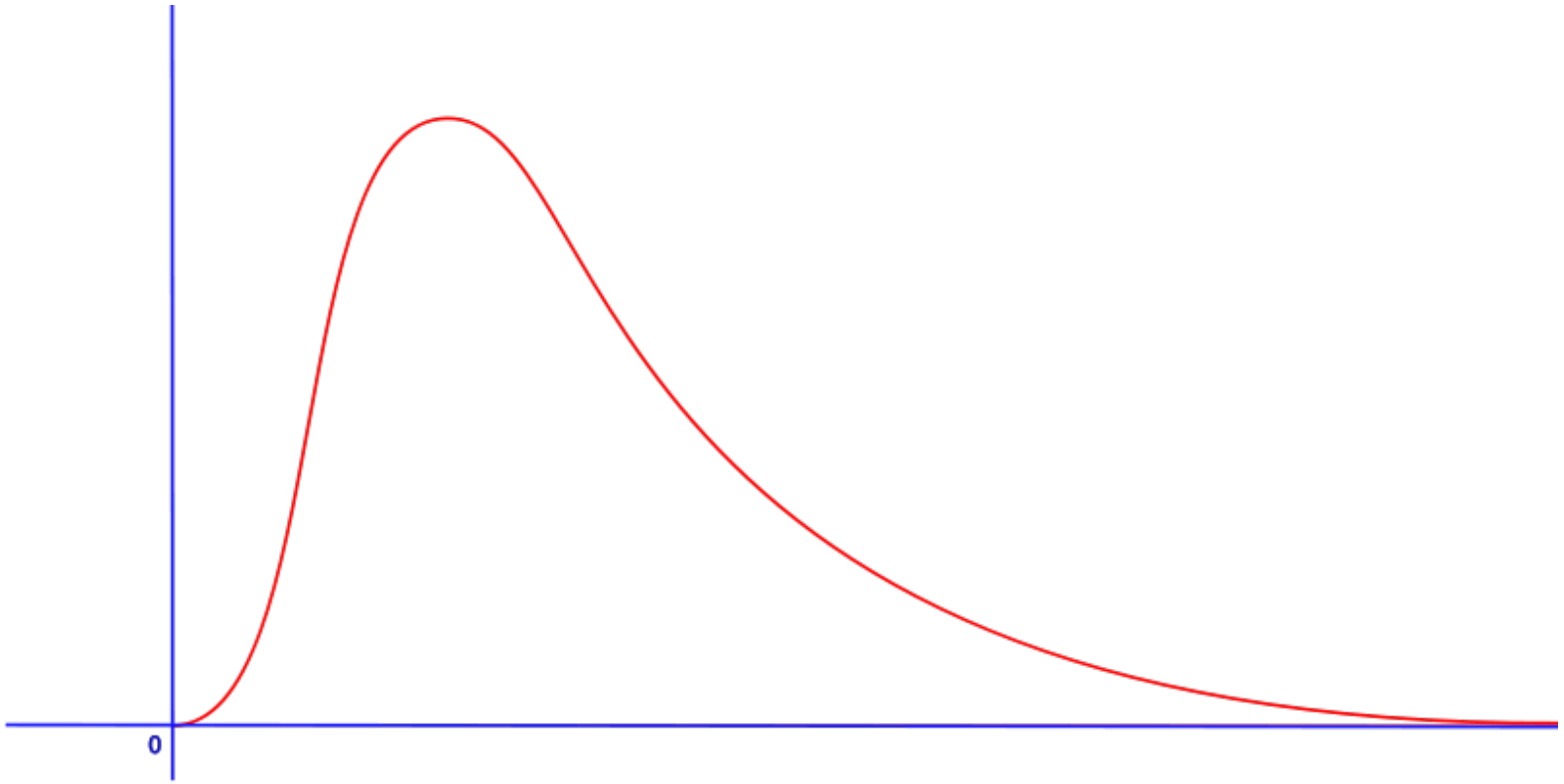
- Probability distributions can be very simple
- Example:
  - Heads: 0.5
  - Tails: 0.5
- Example:
  - 1: 1/6
  - 2: 1/6
  - 3: 1/6
  - 4: 1/6
  - 5: 1/6
  - 6: 1/6

Probability distributions

- Most of the distributions we'll think about in this class will be more complex, and we'll mainly depict them graphically
- We generally depict complex (continuous) distributions using a probability density function (PDF)
- Our statistics software will be doing a lot of fancy math behind the scenes based on these PDFs, but you don't need to worry about such details in this class
- You can basically interpret a graph of a PDF like a kernel density plot (or even a histogram)

Probability distributions

- Statisticians have developed PDFs for distributions with lots of different shapes
- Here's a skewed distribution:

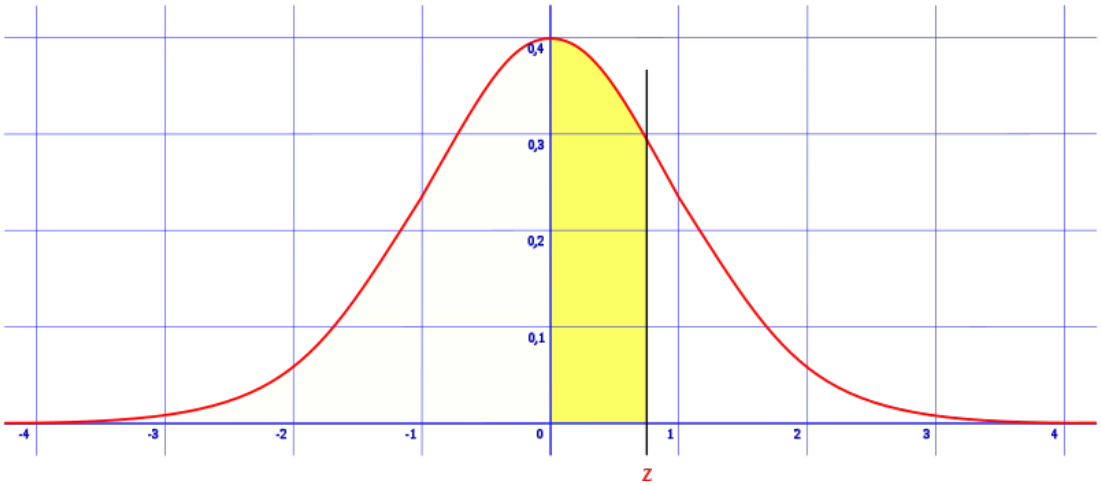


Probability distributions

- With PDF graphs, wherever the height is greater, the values are more likely to be observed (just like a histogram)
- But kernal desnity plots or histograms are generally used for data that we've already collected; PDF graphs depict a probability distribution from which data is "drawn" (the distinction is subtle, and you can't necessarily tell from just looking at the graph)

Probability distributions

- A precise interpretation for PDFs is a bit tricky since any value can be measured to infinite digits (and therefore has infinitely small probability of being selected)
- Total area under line equals 1
- Probability for a range of possible values (e.g., probability of drawing a value between 0 and 0.7) equals the area under the line within that range (requires calculus)



Probability distributions

- When we talked about distributions the last few weeks, we were talking about data that had already been collected
- *Probability* distributions describe a (theoretical) random process from which we get data
- Example:
  - If we write out the (theoretical) probabilities for each possible outcome of a coin flip (heads, tails) or a die roll (1, 2, 3, etc.), we're talking about a *probability distribution*
  - If we flip a coin (rolled a die) 20 times and record the results, we're looking at a distribution of data, which we might depict with a histogram

Probability distributions

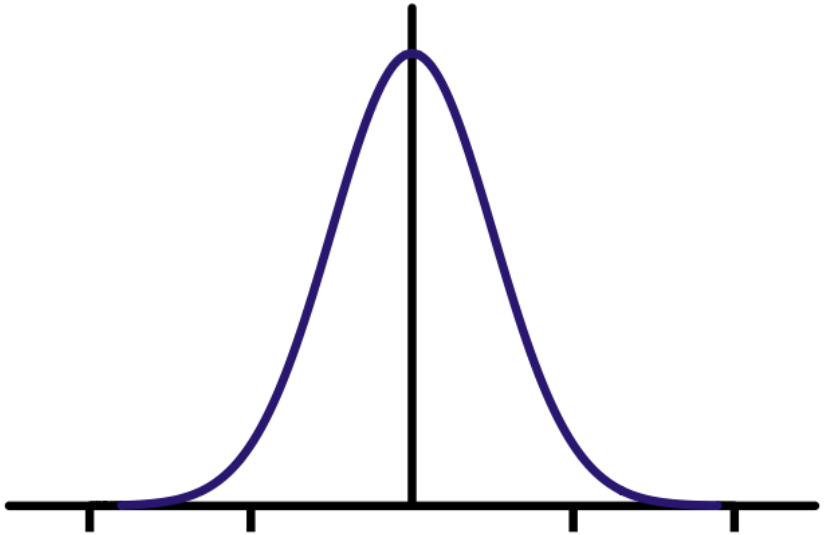
- Another way to think about this distinction is that for a distribution of data we've collected, each observation was (theoretically) drawn from the probability distribution
- Example:
  - If we have a distribution (PDF graph) that describes how likely we think it is that we get various numbers of patients visiting the ER on a given day, we're looking at a *probability distribution*
  - If we're looking at data (e.g., a histogram) of past daily totals for the number of patients who visited the ER, we're looking at a distribution of data

Probability distributions

- Many of the same statistics and words we use to describe data that's been collected can also be used (with a bit of adaptation) to describe probability distributions.
- For example, a probability distribution will (often) have a mean and a variance

Probability distributions

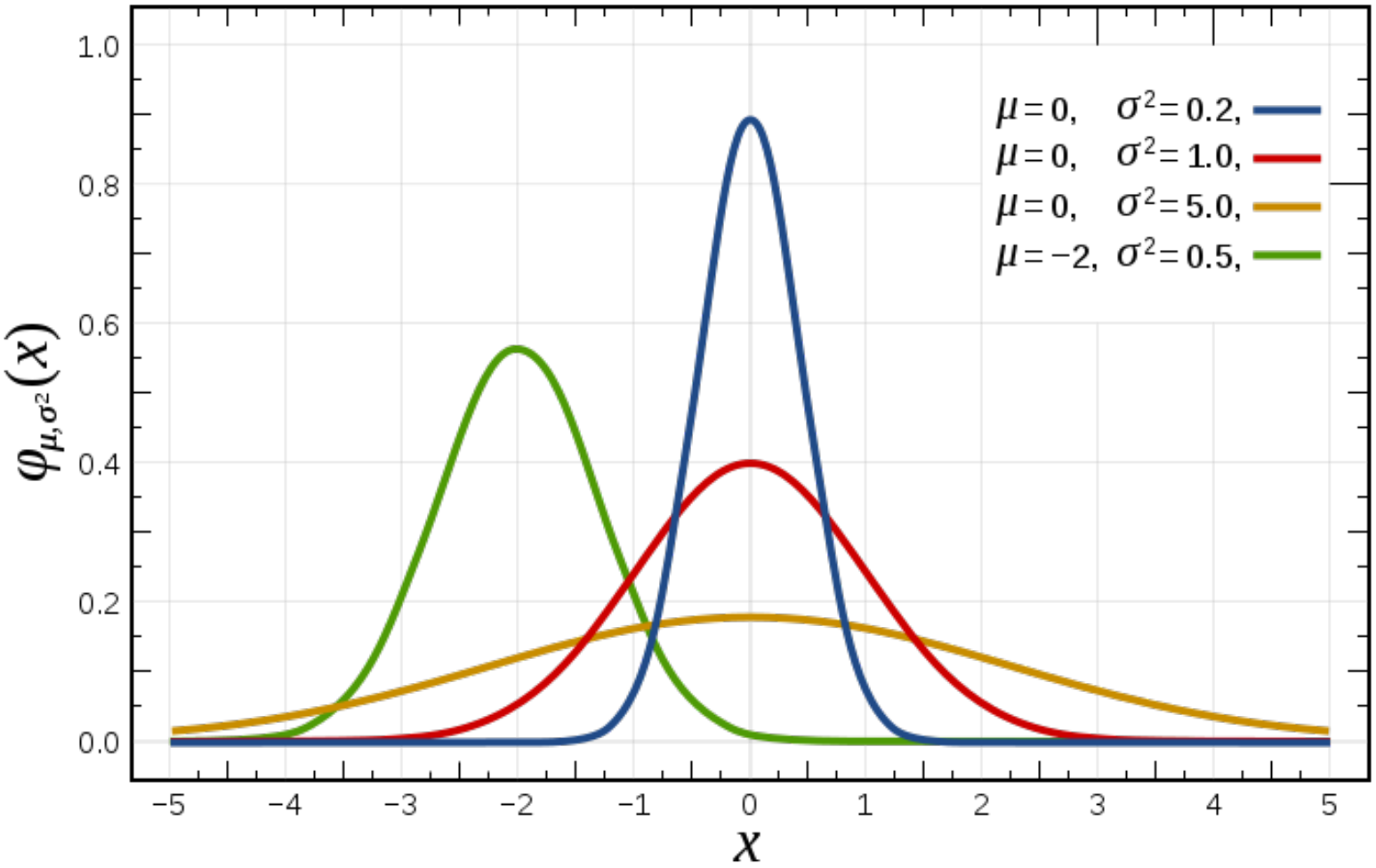
- The normal distribution is one of the most important and commonly-used probability distributions



- All normal distributions have this basic shape, but it can be shifted to the left or right, squished, or expanded

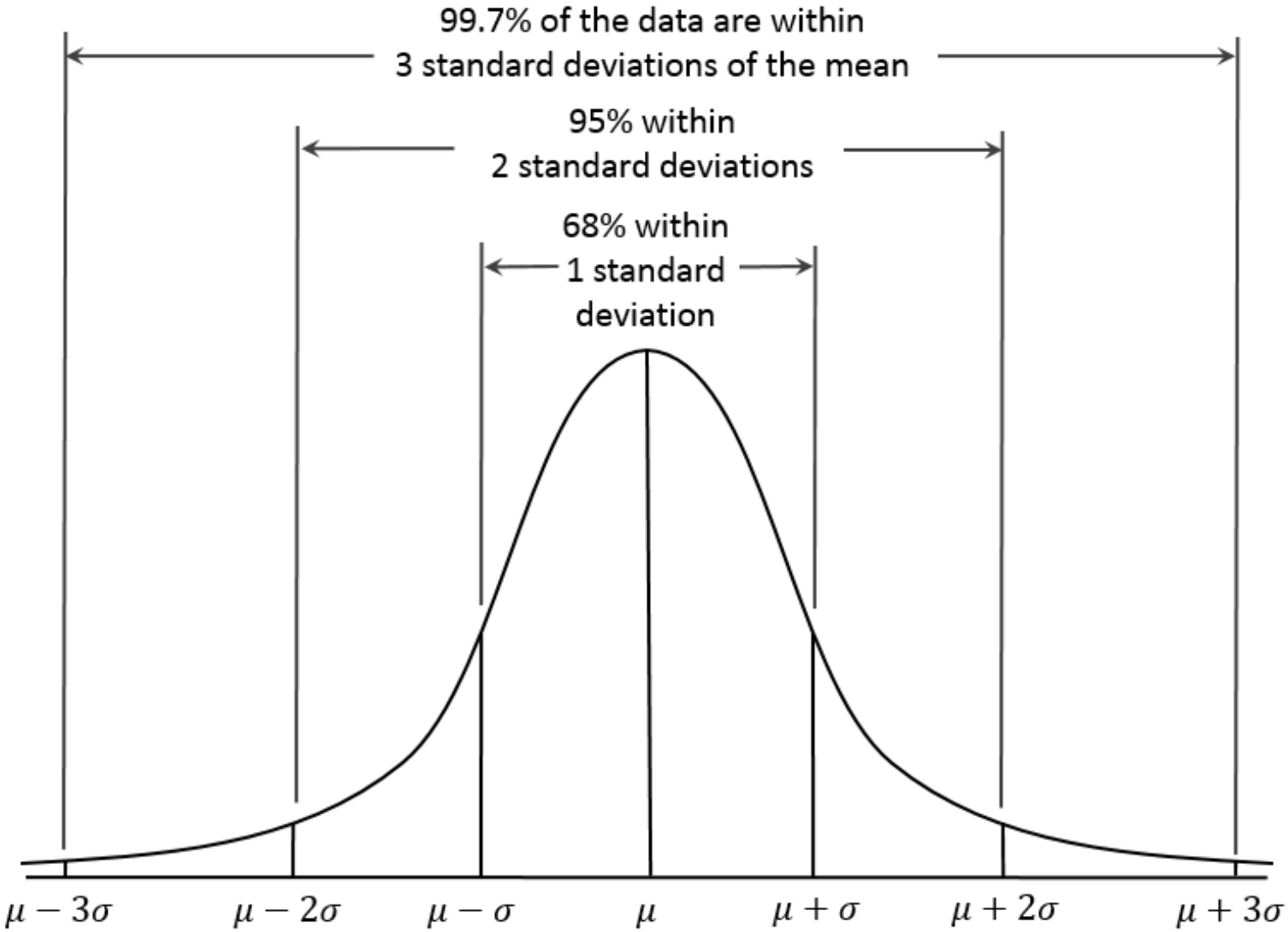
Probability distributions

- For normal distribution, two parameters needed to identify exact distribution: mean (μ) and variance (σ^2)



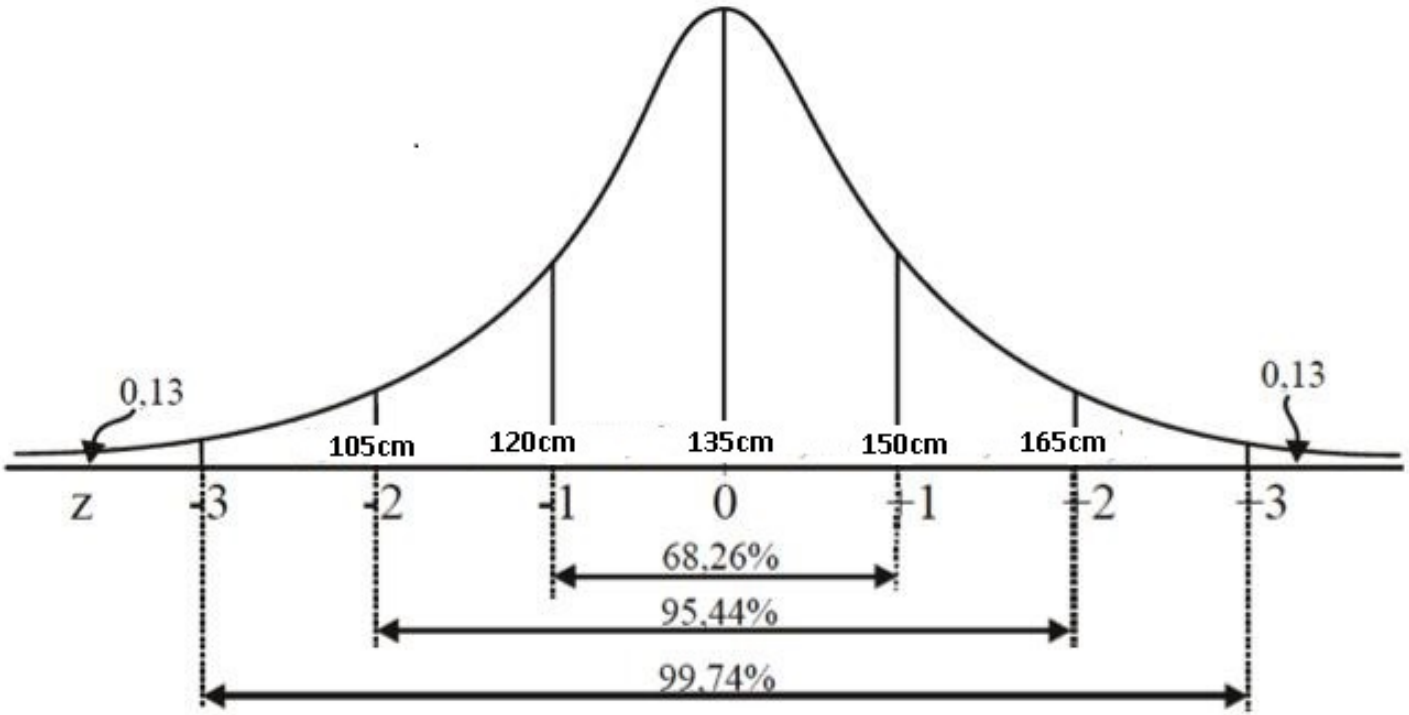
Probability distributions

Important characteristics of the normal PDF



Probability distributions

- Standard normal distribution: normal distribution with mean of 0 and standard deviation of 1 ($\mu = 0; \sigma = 1$)
  - When you hear “standard” normal, think of standardization (Lecture 2); you can transform any normal distribution into a standard normal distribution by standardizing the values in the original distribution



Appendix 1: Example problem: ER visits

Return to main slides

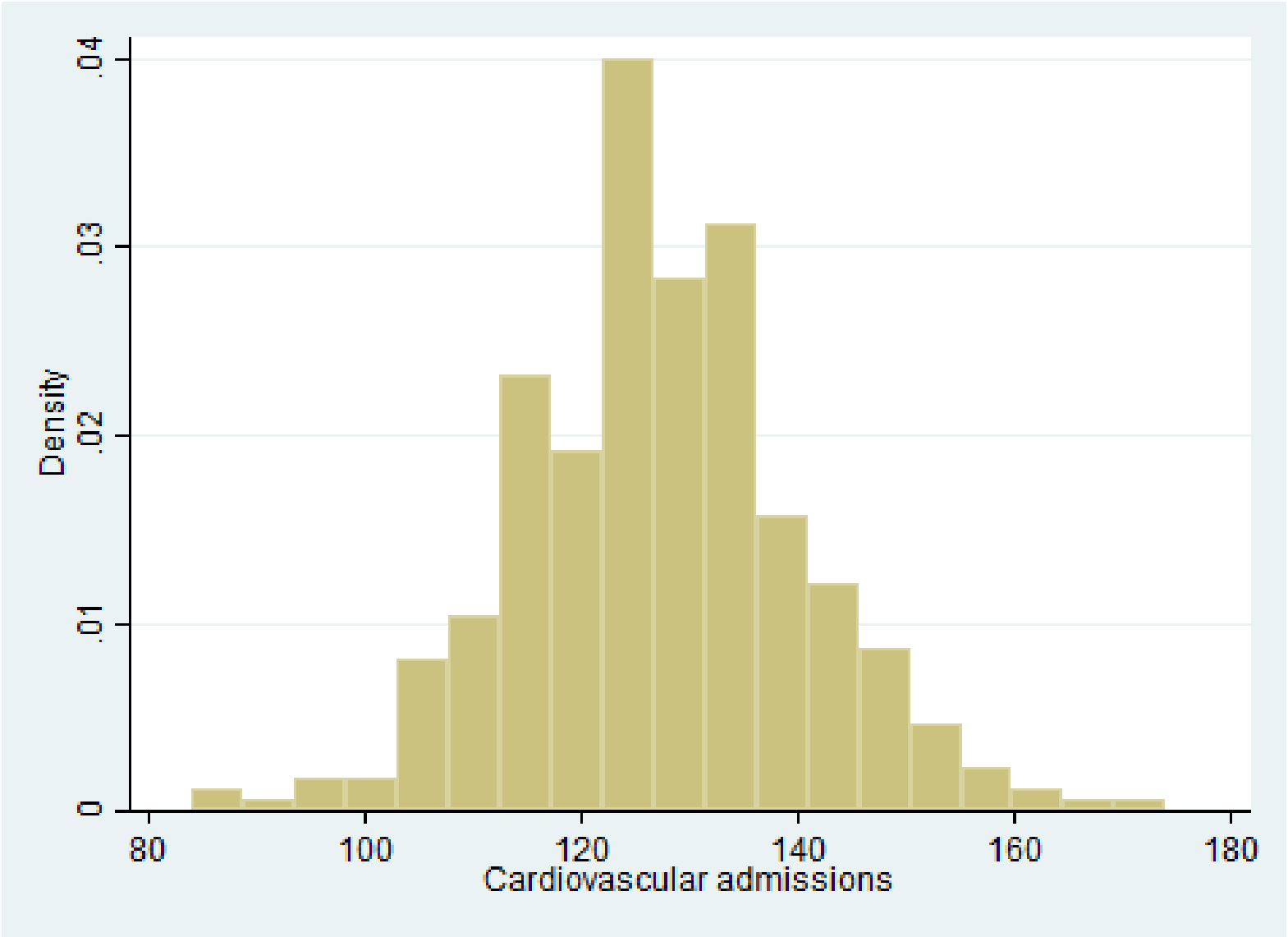
- I'm a hospital administrator, and I want to be able to estimate how many emergency room (ER) visits there will be tomorrow
- I have daily data from the past year, which indicates the number of ER visits each day
- How might I go about creating an estimate of how many ER visits there will be tomorrow?

Example problem: ER visits

- There will be many factors I can't predict that will influence how many people visit the ER tomorrow (air quality, people's physical activities, accidents, etc.)
- I've downloaded some data on ER visits (combined data from 30 hospitals in Seoul; I used the 2011 data only) that were used in a published research project (`https://doi.org/10.1371/journal.pone.0183224`)

```
   |        date   circ |
   |--------------------|
1. | 2011-01-01     134 |
2. | 2011-01-02     126 |
3. | 2011-01-03     149 |
4. | 2011-01-04     115 |
```

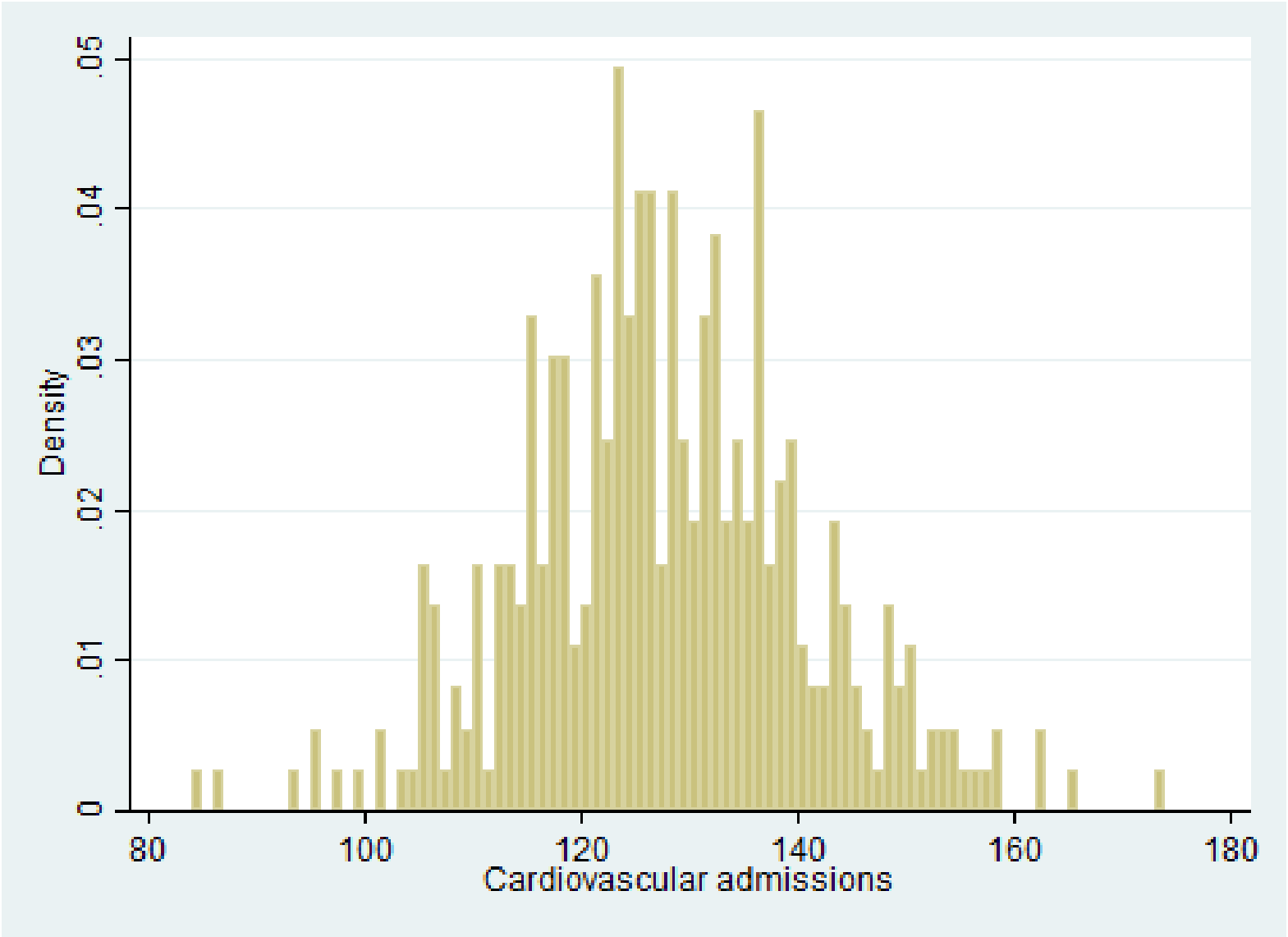
Example problem: ER visits



Example problem: ER visits

- When it comes to making a prediction, we could give our best guess as a single number (maybe 130)
- Or, if we want to give a more sophisticated (and informative) guess, we could list out a bunch of possible values and indicate their relative likelihood
- We could could present this information either graphically or in a table; either way, it's called a *probability distribution*
- Perhaps the simplest probability distribution we could create as an estimate for tomorrow's ER visits would just be to use the exact percentage of times each value was observed in the past year

Example problem: ER visits



Example problem: ER visits

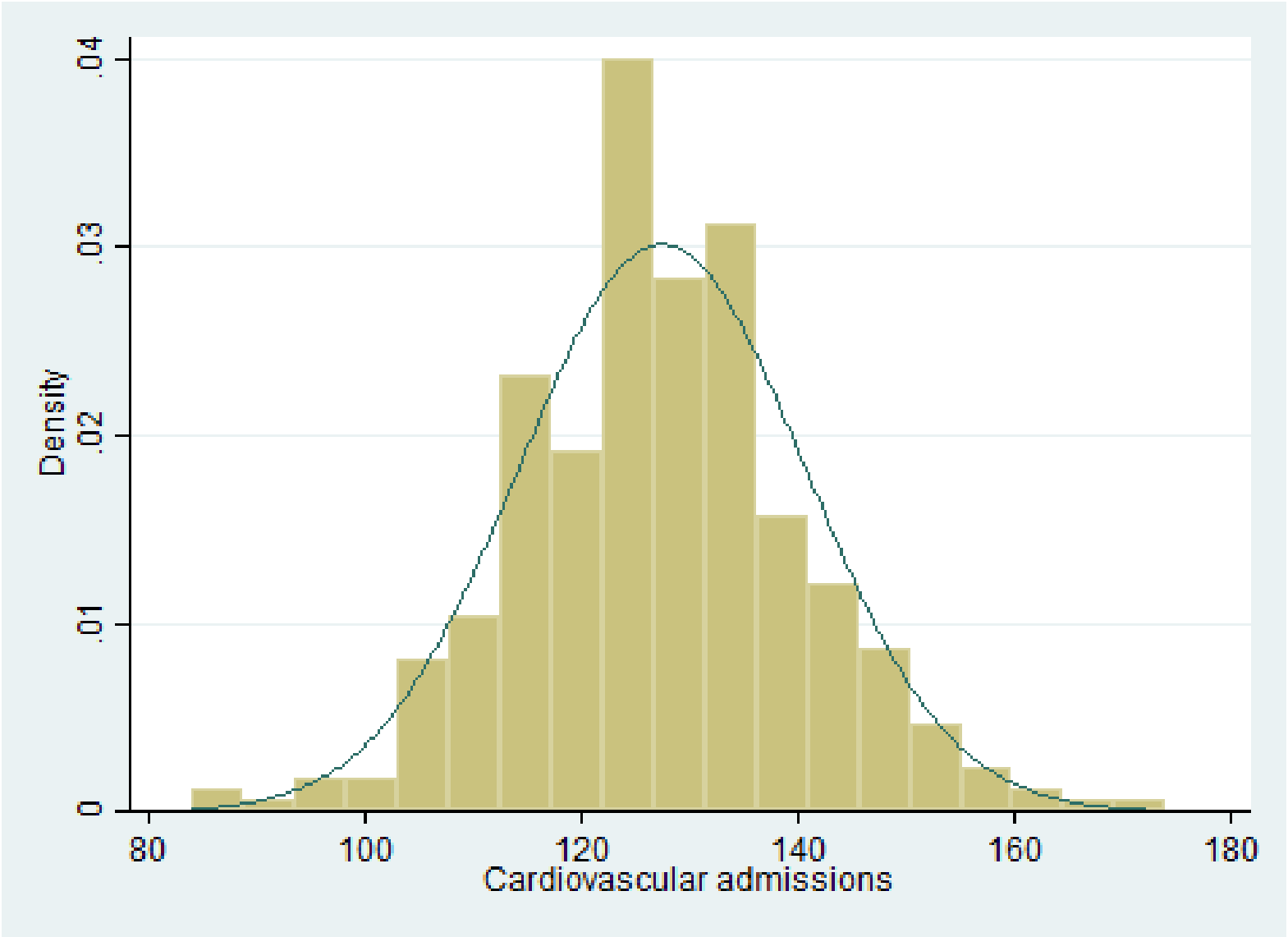
```
. tab circ

Cardiovascu |
        lar |
 admissions |      Freq.     Percent        Cum.
------------+-----------------------------------
         84 |          1        0.27        0.27
         86 |          1        0.27        0.55
         93 |          1        0.27        0.82
...
        114 |          5        1.37       14.25
        115 |         12        3.29       17.53
        116 |          6        1.64       19.18
        117 |         11        3.01       22.19
```

Example problem: ER visits

- There are a couple problems with this simplistic approach:
  - Some values (e.g., 91, 92) were never observed in the last year, but we might not want to rule them out for tomorrow
  - The data is sometimes choppy (e.g., 115: 3.29%, 116: 1.64%, 117: 3.01%), and we might want to smooth it out
- One potential solution: use one of the well-known distributions that statisticians have come up with to model different types of data-generation processes
- Let's try a normal distribution with our data

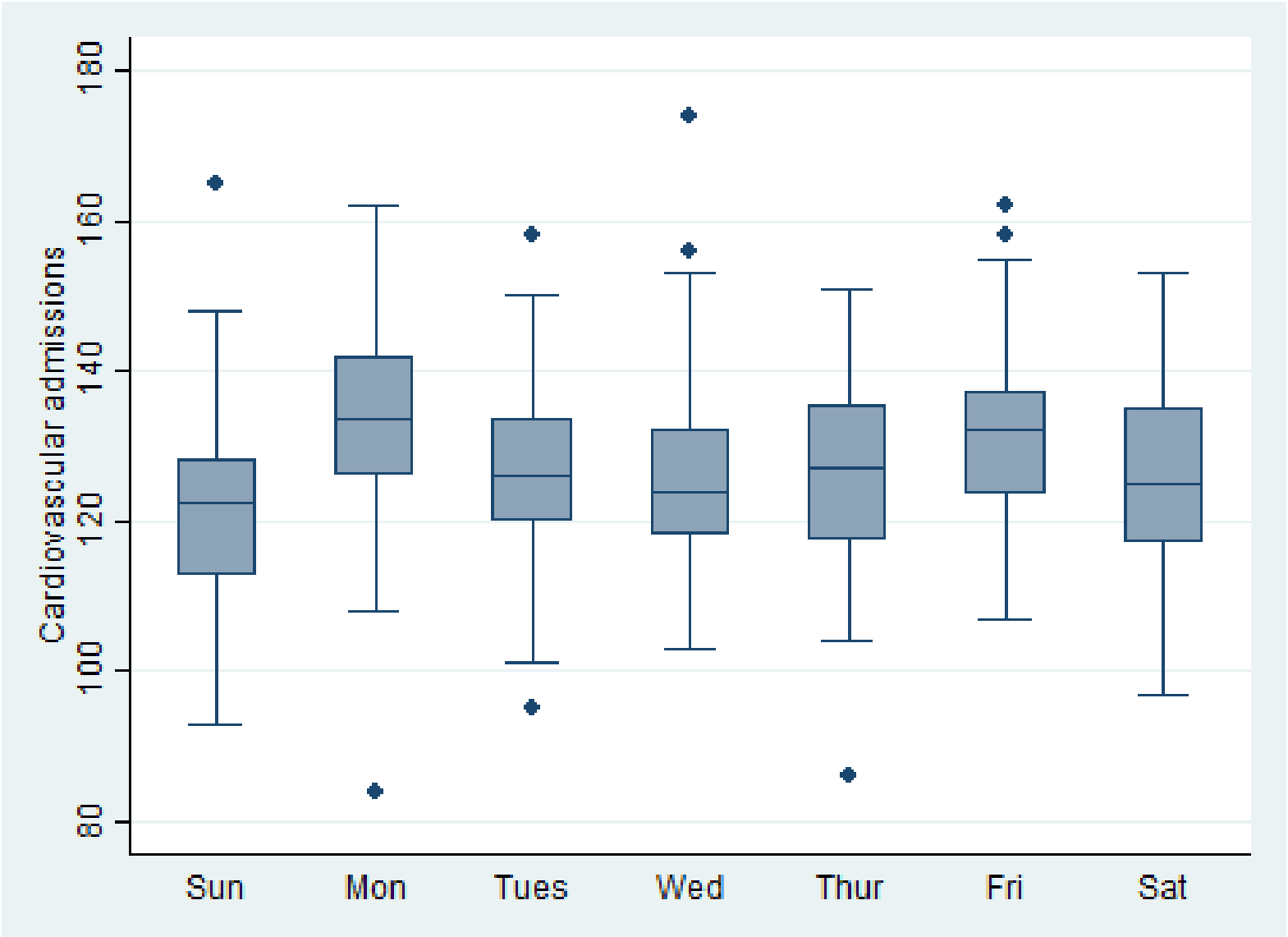
Example problem: ER visits



More on example problem: ER visits

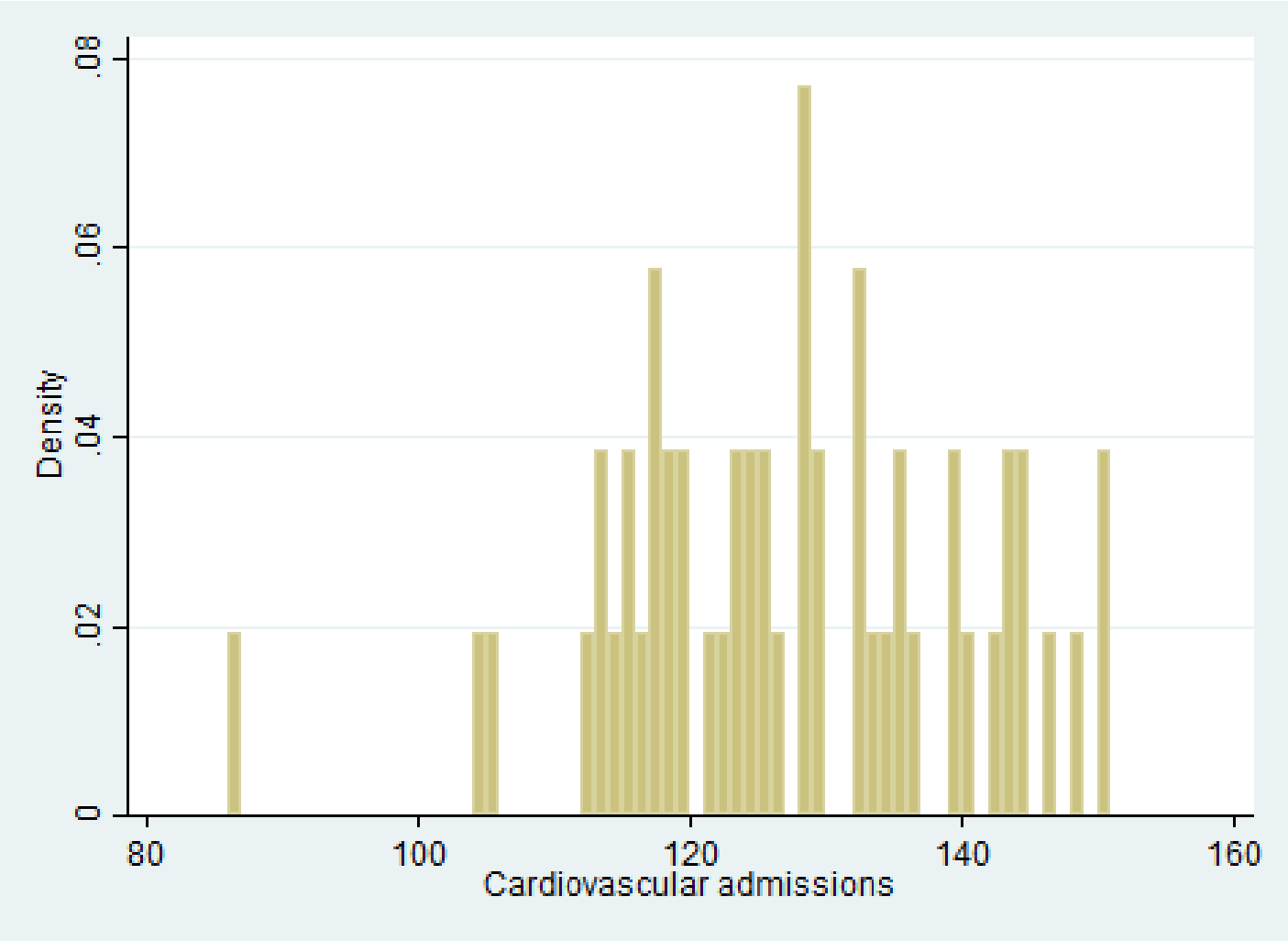
- We can make our analyisis a bit more sophisticated if we think about making different predictions under different conditions (e.g., for different days of the week)
- One way to approach this is to think about there being a different probability distribution for each day of the week

Example problem: ER visits



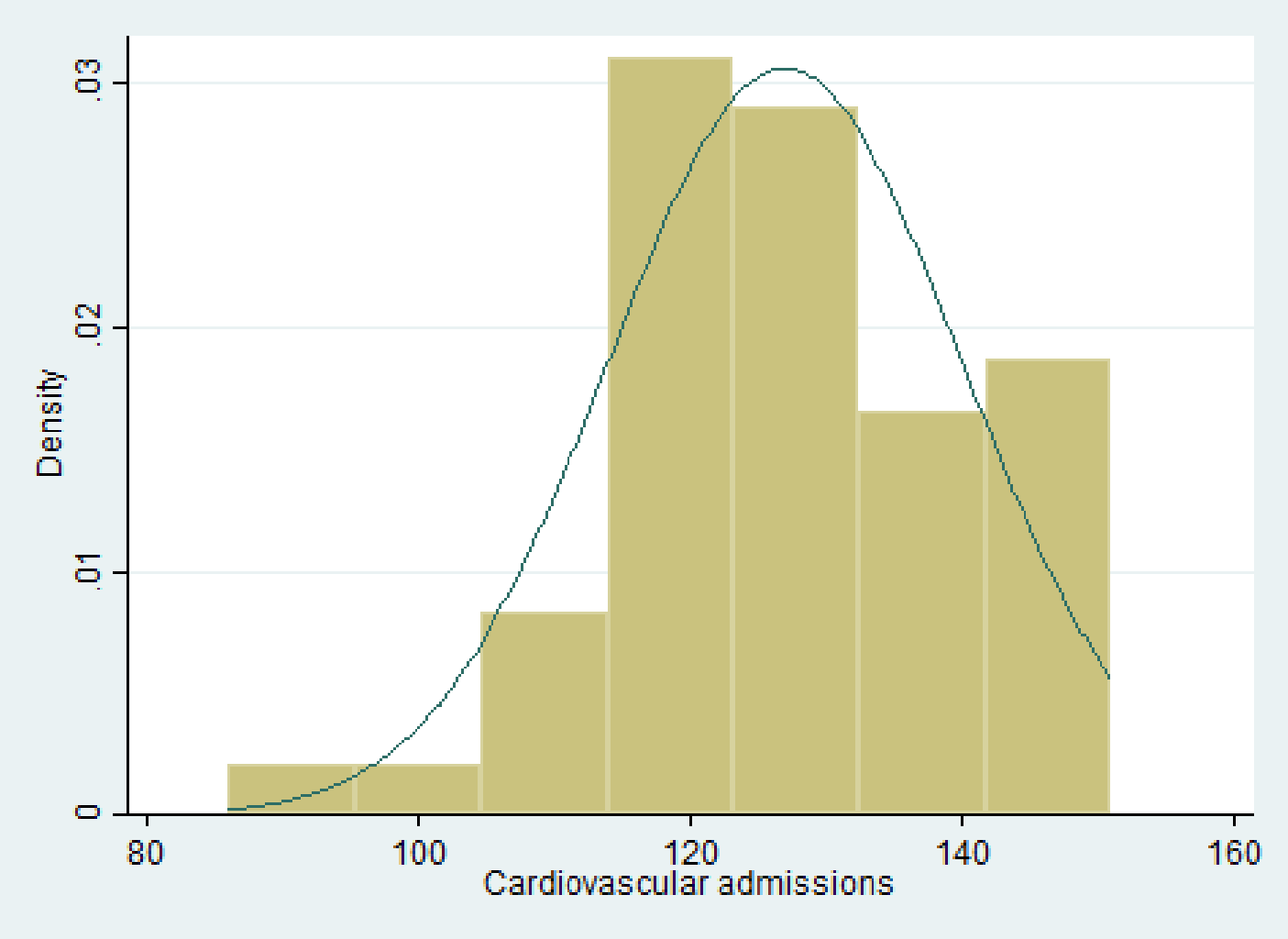
Example problem: ER visits

If we look at data for just Thursdays (from the last year), the data gets even choppier (because there are fewer observations)

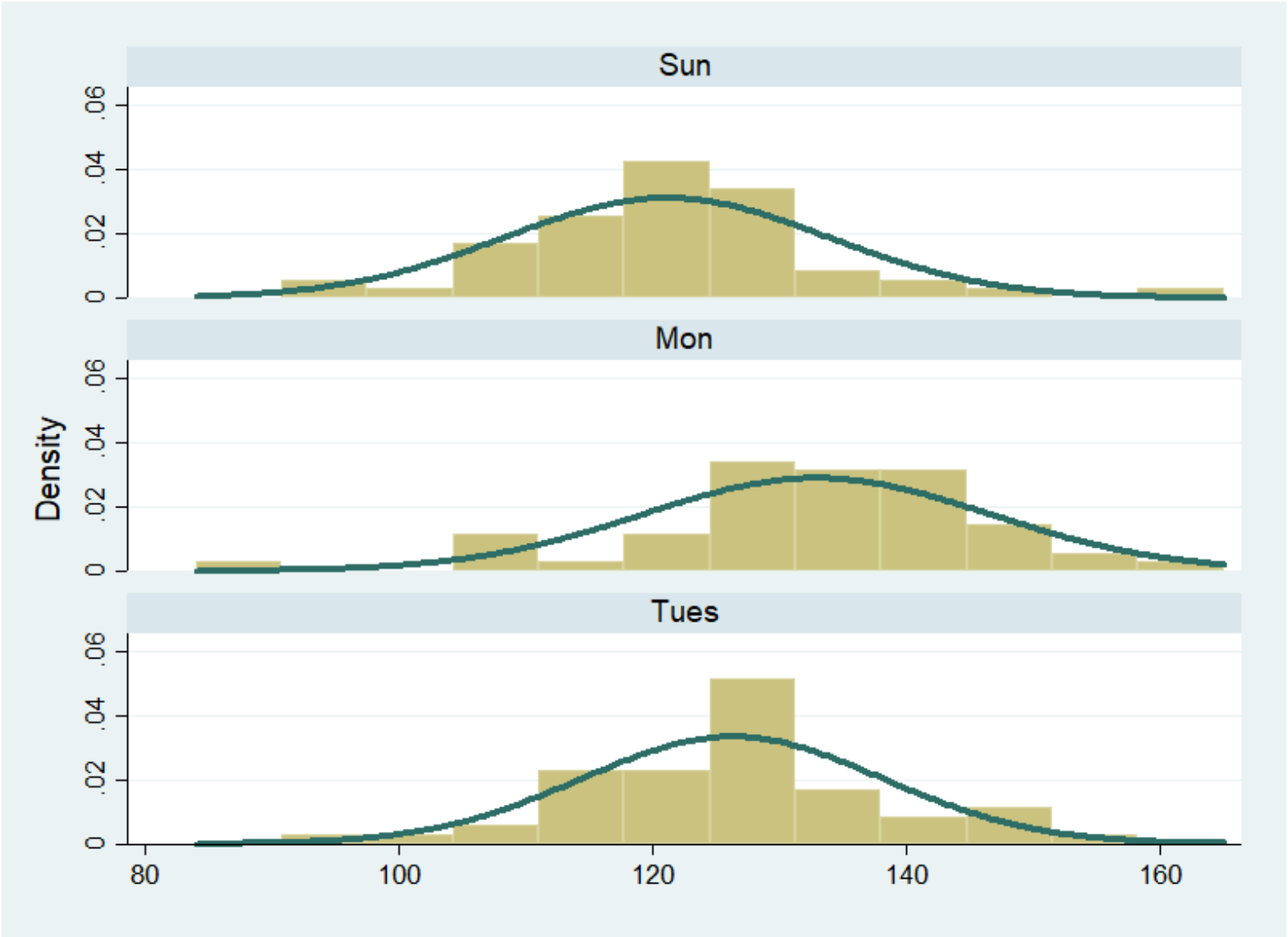


Example problem: ER visits

Thursday gets its own normal distribution



Example problem: ER visits



Example problem: ER visits

Note that not all data from the real world looks like a normal distribution: