JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Describing Data: Part I

John McGready
Johns Hopkins University

Lecture Topics

- What role does statistics have in public health?
- Types of data: continuous, binary, categorical, time-to event
- Continuous data: numerical summary measures
- Continuous data: visual summary measure
- Sample data versus population (process) level data

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

What Role Does Biostatistics Play in Public Health (Why Do I Need this Stuff?)

Data Is Everywhere!

- Data is utilized and summarized frequently in research literature

- From *Archives of Surgery* article, August 2000:
  - "**Hypothesis**: Surgeon-directed institutional peer review, associated with positive physician feedback, can decrease the morbidity and mortality rates associated with carotid endarterectomy."
  - "**Results**: Stroke rate decreased from 3.8% (1993-1994) to 0% (1997-1998). The mortality rate decreased from 2.8% (1993-1994) to 0% (1997-1998). (Average) length of stay decreased from 4.7 days (1993-1994) to 2.6 days (1997-1998). The (average) total cost decreased from $13,344 (1993-1994) to $9,548 (1997-1998)."

Source: Olcott IV, C., et al. (2004). Institutional peer review can reduce the risk and cost of carotid endarterectomy *Arch Surg*, 135: 939-942.

Data Is Everywhere!

- Data is utilized and summarized with statistics frequently in popular media

- From cnn.com, Monday July 8th, 2008:
  - "'For the first time, an influential doctors group is recommending that some children as young as eight be given cholesterol-fighting drugs to ward off future heart problems . . . With one-third of U.S. children overweight and about 17 percent obese, the new recommendations are important,' said Dr. Jennifer Li, a Duke University children's heart specialist."

Data Is Everywhere!

- Data is utilized and summarized with statistics frequently in popular media
- From *Washington Post*, June 27th, 2008:
  - “The number of young homosexual men being newly diagnosed with HIV infection is rising by 12 percent a year, with the steepest upward trend in young black men, according to a new report.”

Data Provides Information

- Good data can be analyzed and summarized to provide useful information

- Bad data can be analyzed and summarized to provide incorrect/ harmful/non-informative information

Steps in a Research Project

- Planning/design of study
- Data collection
- Data analysis
- Presentation
- Interpretation
- Biostatistics CAN play a role in each of these steps! (but sometimes is only called upon for the data analysis part)

Biostatistics Issues

- Planning/design of studies
  - Primary question(s) of interest:
    - Quantifying information about a single group?
    - Comparing multiple groups?
  - Sample size
    - How many subjects needed total?
    - How many in each of the groups to be compared?
  - Selecting study participants
    - Randomly chosen from “master list?”
    - Selected from a pool of interested persons?
    - Take whoever shows up?
  - If group comparison of interest, how to assign to groups?

Biostatistics Issues

- Data collection

- Data analysis
  - What statistical methods are appropriate given the data collected?
  - Dealing with variability (both natural and sampling related):
    - Important patterns in data are obscured by variability
    - Distinguish real patterns from random variation
  - Inference: using information from the single study coupled with information about variability to make statement about the larger population/process of interest

Biostatistics Issues

- Presentation
  - What summary measures will best convey the "main messages" in the data about the primary (and secondary) research questions of interest
  - How to convey/ rectify uncertainty in estimates based on the data

- Interpretation
  - What do the results mean in terms of practice, the program, the population etc.?

1954 Salk Polio Vaccine Trial

Polio Cases:	Vaccine	82
	Placebo	162

Source: Meier, P. (1972), "The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine," In J. Tanur (Editor), *Statistics: A Guide to the Unknown*. Holden-Day.

Design: Features of the Polio Trial

- Comparison group
- Randomized
- Placebo controls
- Double blind
- Objective—the groups should be equivalent except for the factor (vaccine) being investigated

Analysis Question

- Question
  - There were almost twice as many polio cases in the placebo compared to the vaccine group
  - Could the results be due to chance?

Such Great Imbalance by Chance?

- Polio cases
  - Vaccine—82
  - Placebo—162

- Statistical methods tell us how to make these probability calculations

Section B

Types of Data

Binary Data

- Binary (dichotomous) data
  - Yes/no
  - Polio: Yes/no
  - Cure: Yes/no
  - Sex: Male/female (or as yes/no, “is subject male?”)

Categorical Data

- Categorical data (*place individuals in categories*)

- *Nominal categorical data: no inherent order to categories*
  - Race/ethnicity
  - Country of birth
  - Religious affiliation

- *Ordinal categorical data: order to categories*
  - Income level categorized into four categories, least to greatest
  - Degree of agreement, five categories from strongly disagree to strongly agree

Continuous Data

- Continuous data (*finer measurements*)
  - Blood pressure, mmHg
  - Weight, pounds (kilograms, ounces, etc.)
  - Height, feet (centimeters, inches, etc.)
  - Age, years (months)
  - Income level, dollars/year (Euro by year, etc.)

Time to Event Data

- Data that is a hybrid of continuous data and binary data
  - Whether an event occurs and time to the occurrence (or time to last follow-up without occurrence)

Different Methods for Different Data Types

- To compare the number of polio cases in the two treatment arms of the Salk Polio vaccine, you could use . . .
  - Fisher's Exact Test
  - Chi-Square Test

- To compare blood pressures in a clinical trial evaluating two blood pressure-lowering medications, you could use . . .
  - 2-Sample t-Test
  - Wilcoxon Rank Sum Test

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section C

Continuous Data: Numerical Summary Measures; Sample Estimates versus Population Measures

Summarizing and Describing Continuous Data

- Measures of the center of data
  - Mean
  - Median

- Measure of data variability
  - Standard deviation (variance)
  - Range

Sample Mean: The Average or Arithmetic Mean

- Add up data, then divide by sample size (*n*)

- The sample size *n* is the number of observations (pieces of data)

Mean, Example

- Five systolic blood pressures (mmHg) (n = 5)
  - 120, 80, 90, 110, 95
- Can be represented with math type notation:
  - $x_1 = 120$, $x_2 = 80$, . . . $x_5 = 95$
- The sample mean is easily computed by adding up the five values and dividing by five—in statistical notation the sample mean is frequently represented by a letter with a line over it
  - For example (pronounced "x bar")
  - $\bar{x}$

Mean, Example

- Five systolic blood pressures (mmHg) (n = 5)
  - 120, 80, 90, 110, 95

$$\bar{x} = \frac{120 + 80 + 90 + 110 + 95}{5} = 99mmHg$$

Notes on Sample Mean

- Generic formula representation

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

- In the formula to find the mean, we use the "summation sign"—$\sum$
  - This is just mathematical shorthand for "add up all of the observations"

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots\ldots + x_n$$

Notes on Sample Mean

- Also called *sample average* or *arithmetic mean*

- Sensitive to extreme values
  - One data point could make a great change in sample mean

- Why is it called the *sample* mean?
  - To distinguish it from population mean (will discuss at end of this section)

Sample Median

- The median is the middle number (also called the 50th percentile)
  - Other percentiles can be computed as well, but are not measures of center

80 90 95 110 120

Sample Median

- The sample median is not sensitive to extreme values
  - For example, if 120 became 200, the median would remain the same, but the mean would change to 115

80 90 95 110 200

Sample Median

- If the sample size is an even number

80   90   95   110   120   125

Median

$$\frac{95+110}{2}=102.5mmHg$$

Describing Variability

- Sample variance (s^2)
- Sample standard deviation (*s* or *SD*)
- The sample variance is the average of the square of the deviations about the sample mean

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

Describing Variability

- The sample standard deviation is the square root of s^2

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

Describing Variability

- Recall, the five systolic blood pressures (mm Hg) with sample mean ($\bar{x}$) of 99 mmHg

- Five systolic blood pressures (mmHg) (n = 5)
  - 120, 80, 90, 110, 95

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (120 - 99)^2 + (80 - 99)^2 + (90 - 99)^2$$

$$+(110 - 99)^2 + (95 - 99)^2$$

Describing Variability

- Example: n = 5 systolic blood pressures (mm Hg)

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (21)^2 + (-19)^2 + (-9)^2 + (11)^2 + (-4)^2$$

$$= (441) + (361) + (81) + (121) + (16)$$

$$= 1020 \text{mmHg}^2$$

Describing Variability

- Sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1020}{4} = 255$$

- Sample standard deviation (*s*)

$$\sqrt{s^2} = \sqrt{255}$$

s = 15.97 (mmHg)

Notes on *s*

- The bigger s is, the more variability there is

- s measures the spread about the mean

- s can equal 0 only if there is no spread
  - All n observations have the same value

- The units of s are the same as the units of the data (for example, mm Hg)

- Often abbreviated *SD* or *sd*

- s^2 is the best estimate from the sample of the population variance σ^2; s is the best estimate of the population standard deviation σ

Population Versus Sample

- *Sample*: a subset (part) of a larger group (population) from which information is collected to learn about the larger group
  - For example, sample of blood pressures *n* = five 18-year-old male college students in the United States

- *Population*: the entire group for which information is wanted
  - For example, the blood pressure of all 18-year-old male college students in the United States

Random Sampling

- For studies it is optimal if the sample which provides the data is representative of the population under study
  - Certainly not always possible!

- For this term, we will make this assumption unless otherwise specified

- One way of getting a representative sample: simple random sampling
  - A sampling scheme in which every possible sub-sample of size *n* from a population is equally likely to be selected
  - How to do it? More detail in second half of term, but think of the “names in a hat” idea

Population Versus Sample

- The sample summary measures (mean, median, sd) are called statistics, and are just estimates of their population (process) counterparts

- Assuming the sample is representative of the population from which it is taken (for example, a randomly drawn sample) these sample estimates should be "good" estimates of true quantities

Population

Population (true) mean: μ

Population (true) SD: σ

Sample

Sample mean: $\bar{x}$

Sample SD: s

Population Versus Sample

- For example, we will never know the population mean μ but would like to know it

- We draw a sample from the population

- We calculate the sample mean $\overline{x}$

- How close is $\overline{x}$ to μ?

- Statistical theory allow us to estimate how close $\overline{x}$ is to μ using information computed from the same single sample we use to estimate $\overline{x}$

The Role of Sample Size on Sample Estimates

- Increasing sample size, increases "Goodness" of sample statistics as estimates for their population counterparts
  - Sample mean based on random sample of 1,000 observations is "better estimate" of true (population) mean than sample mean than sample mean based on random sample of 100 from same population
  - Same logic applies to sample standard deviation estimates
  - We'll define "better estimate" in the third lecture

The Role of Sample Size on Sample Estimates

- Increasing sample size does not dictate how sample estimates from two different representative samples of difference size will compare in value!

- Researcher can not systematically decrease (or increase) value of sample estimates such as mean and standard deviation by taking larger samples!

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

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

The Role of Sample Size on Sample Estimates

- Extreme values, both larger and smaller, are actually more likely in larger samples
  - The smaller and larger extremes in larger samples they "balance each other out"
  - This "balancing" act tends to keep the mean in a "steady state" as sample size increases—it tends to be about the same

- In addition, "non-extreme" values (values closer to mean) are also more likely in larger samples
  - Hence, sample SD also stays "balanced," i.e., does not systematically increase/decrease with larger samples

SD: Why Do We Divide by n-1 Instead of n?

- We really want to replace $\bar{x}$ with μ in the formula for s^2

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- Since we don't know μ, we use $\bar{x}$

- But generally, $\sum_{i=1}^{n} (x_i - \bar{x})^2$ tends to be smaller than $\sum_{i=1}^{n} (x_i - \mu)^2$

  - To compensate, we divide by a smaller number: *n-1* instead of *n*

- This will be explored further in an optional component of the third lecture

n-1

- *n-1* is called the *degrees of freedom of the variance* or *SD*
- Why?
  - The sum of the deviations is zero
  - The last deviation can be found once we know the other *n-1*
  - Only *n-1* of the squared deviations can vary freely
- The term *degrees of freedom* arises in other areas of statistics
- It is not always *n-1*, but it is in this case

Why SD as Measure of Variation

- Why note use the range of the data for example?
  - Range = maximum – minimum

- What happens to the sample maximum and minimum as sample size increases?
  - As it turns out, as sample size increases, the maximum tends to increase, and the minimum tends to decrease: Extreme values are more likely with larger samples!
  - This will tend to increase the range systematically with increased sample size

Section D

Visually Displaying Continuous Data: Histograms

Pictures of Data: Continuous Variables

- Histograms
  - Means and medians and standard deviations do not tell the whole story
  - Differences in shape of the distribution
  - Histograms are a way of displaying the distribution of a set of data by charting the number (or percentage) of observations whose values fall within pre-defined numerical ranges

How to Make a Histogram

- Consider the following data collected from the 1995 Statistical Abstracts of the United States
  - For each of the 50 United States, the proportion of individuals over 65 years of age has been recorded

How to Make a Histogram

State	%	State	%	State	%	State	%
AL	12.9	IN	12.8	NE	14.1	SC	11.9
AK	4.6	IA	15.4	NV	11.3	SD	14.7
AZ	13.4	KS	13.9	NH	11.9	TN	12.7
AR	14.8	KY	12.7	NJ	13.2	TX	10.2
CA	10.6	LA	11.4	NM	11.0	UT	8.8
CO	10.1	ME	13.9	NY	13.2	VT	12.1
CT	14.2	MD	11.2	NC	12.5	VA	11.1
DE	12.7	MA	14.1	ND	14.7	WA	11.6
FL	18.4	MI	12.4	OH	13.4	WV	15.4
GA	10.1	MN	12.5	OK	13.6	WI	13.4
HI	12.1	MI	12.5	OR	13.7	WY	11.1
ID	11.6	MO	14.1	PA	15.9		
IL	12.6	MT	13.3	RI	15.6		

Source: Statistical Abstracts of the United States (1995).

How to Make a Histogram

State	%	State	%	State	%	State	%
AL	12.9	IN	12.8	NE	14.1	SC	11.9
AK	4.6	IA	15.4	NV	11.3	SD	14.7
AZ	13.4	KS	13.9	NH	11.9	TN	12.7
AR	14.8	KY	12.7	NJ	13.2	TX	10.2
CA	10.6	LA	11.4	NM	11.0	UT	8.8
CO	10.1	ME	13.9	NY	13.2	VT	12.1
CT	14.2	MD	11.2	NC	12.5	VA	11.1
DE	12.7	MA	14.1	ND	14.7	WA	11.6
FL	18.4	MI	12.4	OH	13.4	WV	15.4
GA	10.1	MN	12.5	OK	13.6	WI	13.4
HI	12.1	MI	12.5	OR	13.7	WY	11.1
ID	11.6	MO	14.1	PA	15.9		
IL	12.6	MT	13.3	RI	15.6		

Source: Statistical Abstracts of the United States (1995).

How to Make a Histogram

- Break the data range into mutually exclusive, equally sized "bins": here each is 1% wide

- Count the number of observations in each bin

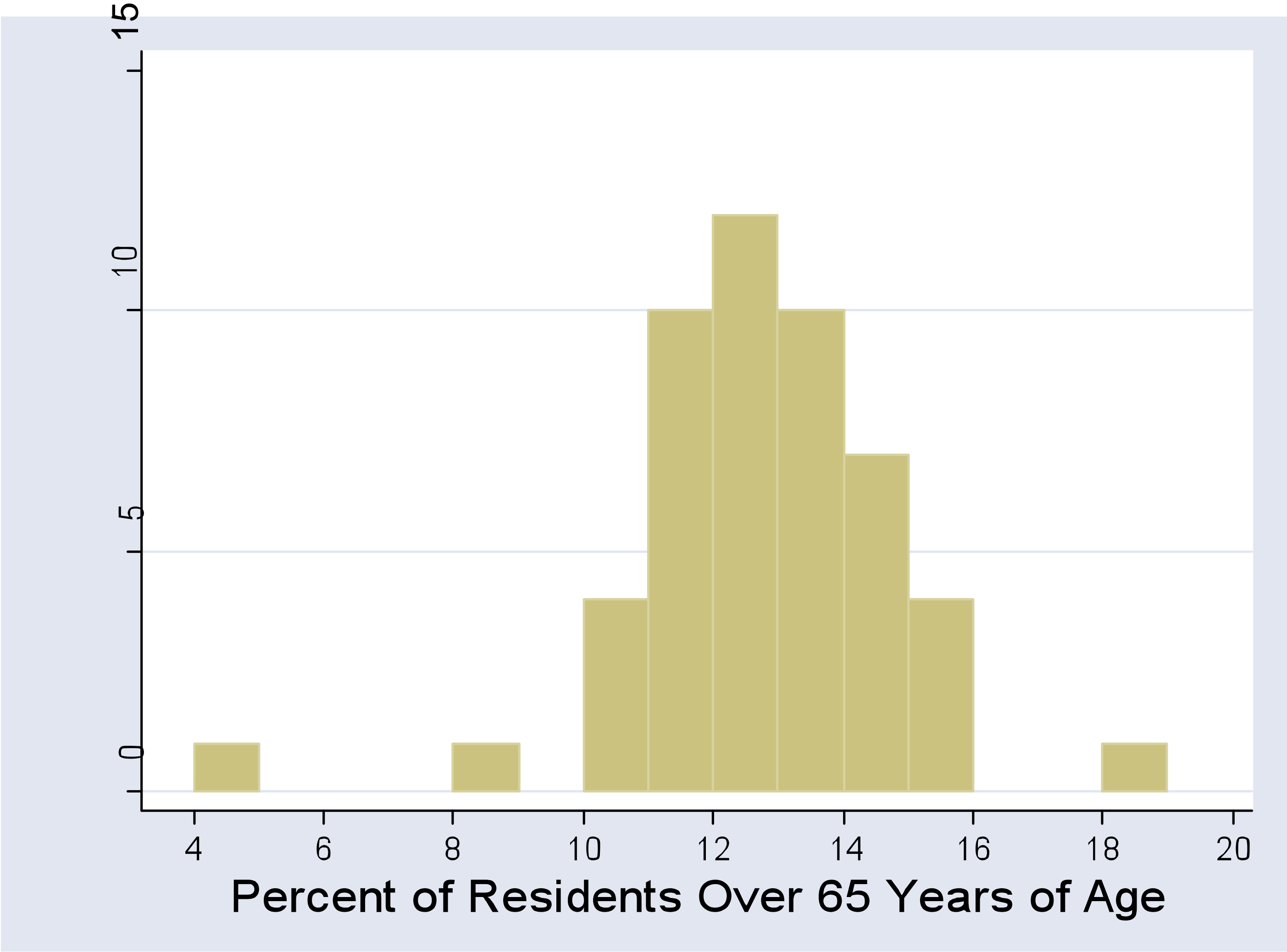
Class	Count	Class	Count	Class	Count
[4.0-5.0]	1	[9.0-10.0]	0	[14.0-15.0]	7
[5.0-6.0]	0	[10.0-11.0]	5	[15.0-16.0]	4
[6.0-7.0]	0	[11.0-12.0]	9	[16.0-17.0]	0
[7.0-8.0]	0	[12.0-13.0]	12	[17.0-18.0]	0
[8.0-9.0]	1	[13.0-14.0]	10	[18.0-19.0]	1

$$\bar{x} = 12.7\%;\ s = 2.1\%$$

How to Make a Histogram

- Draw the histogram

- Label scales



Pictures of Data: Histograms

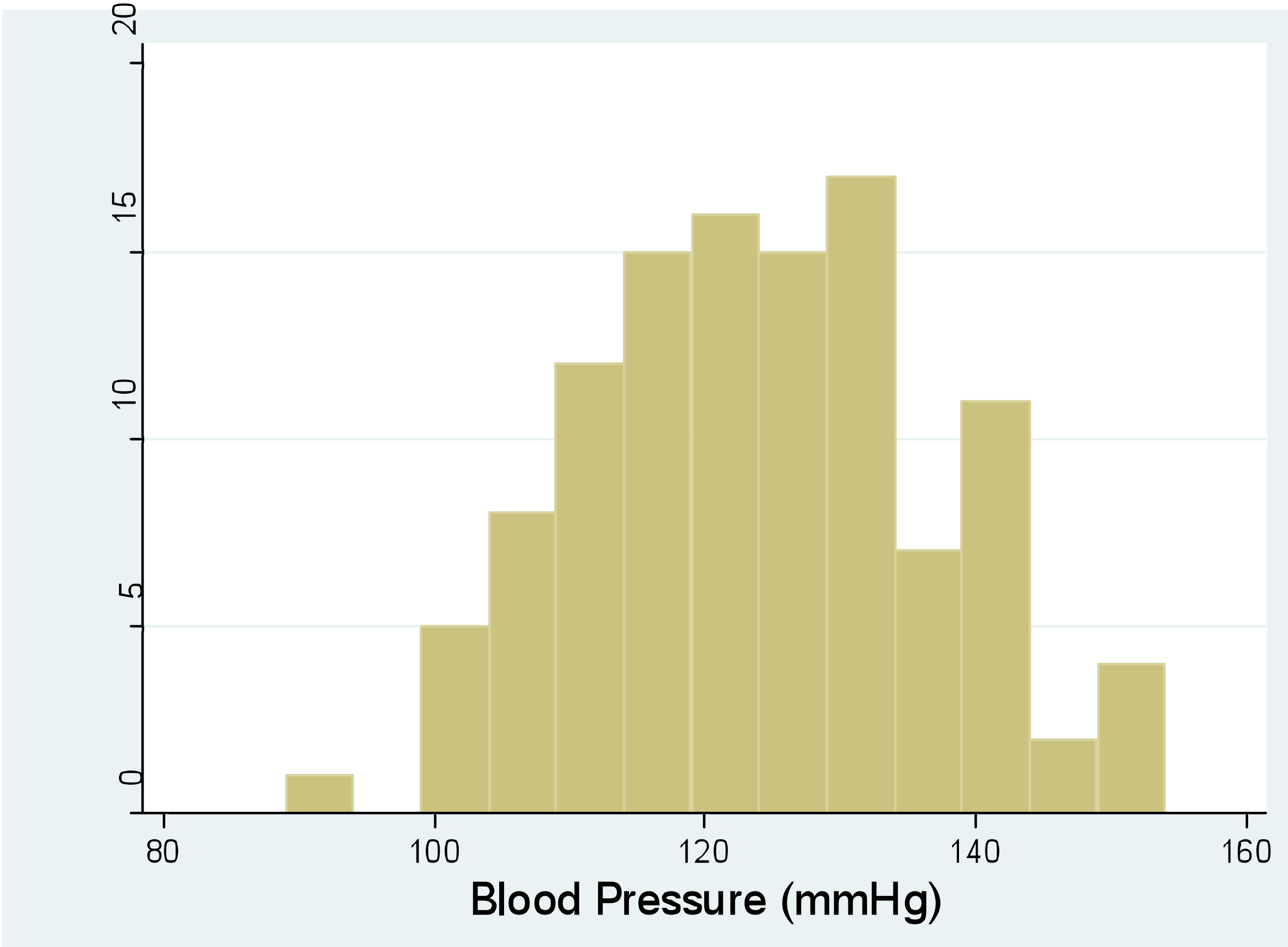
- Suppose we have a sample of blood pressure data on a sample of 113 men

- Sample mean ($\bar{x}$) : 123.6 mmHg

- Sample Median (*m*): 123.0 mmHg

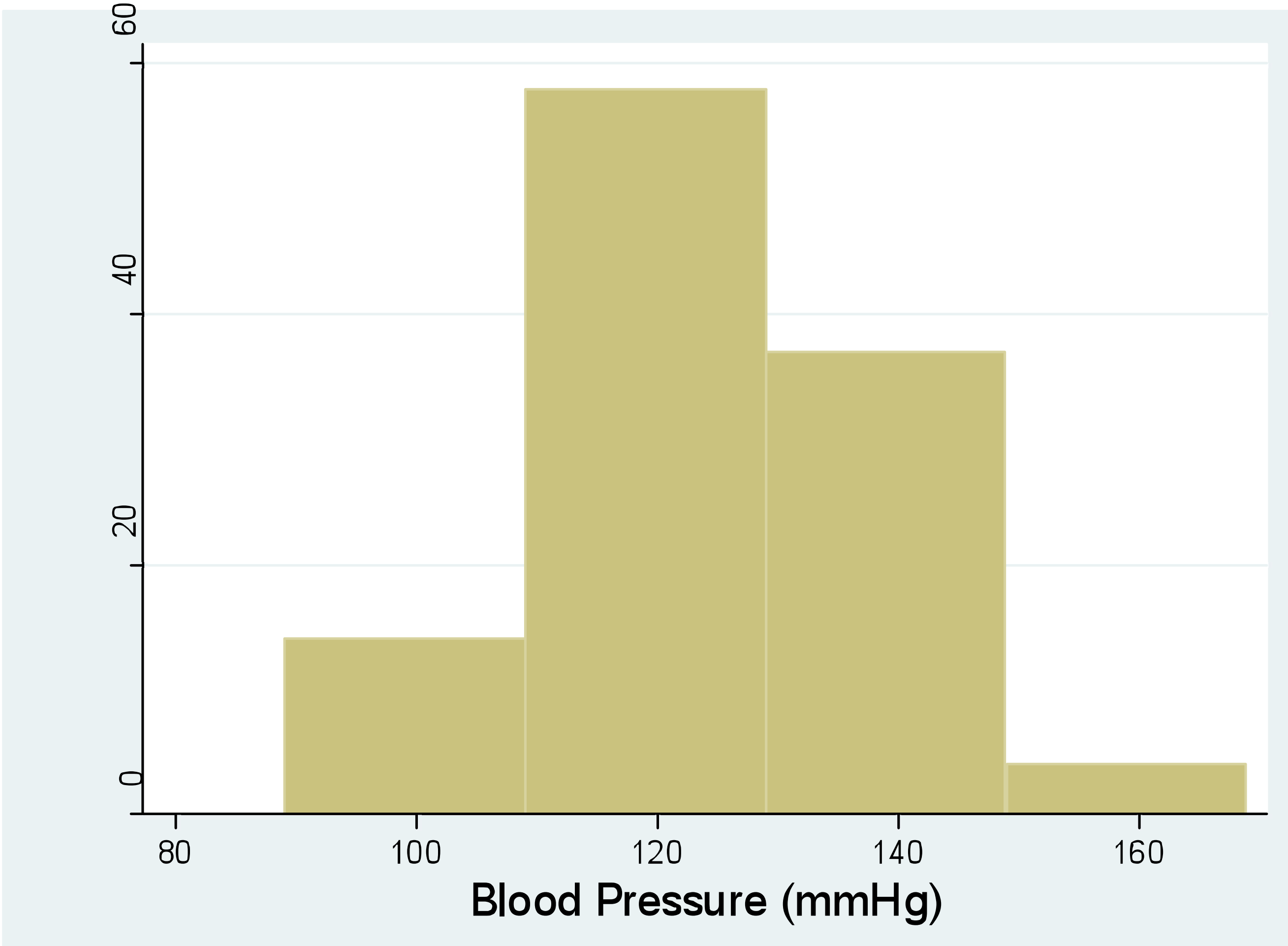
- Sample sd: (*s*): 12.9 mmHg

Pictures of Data: Histograms



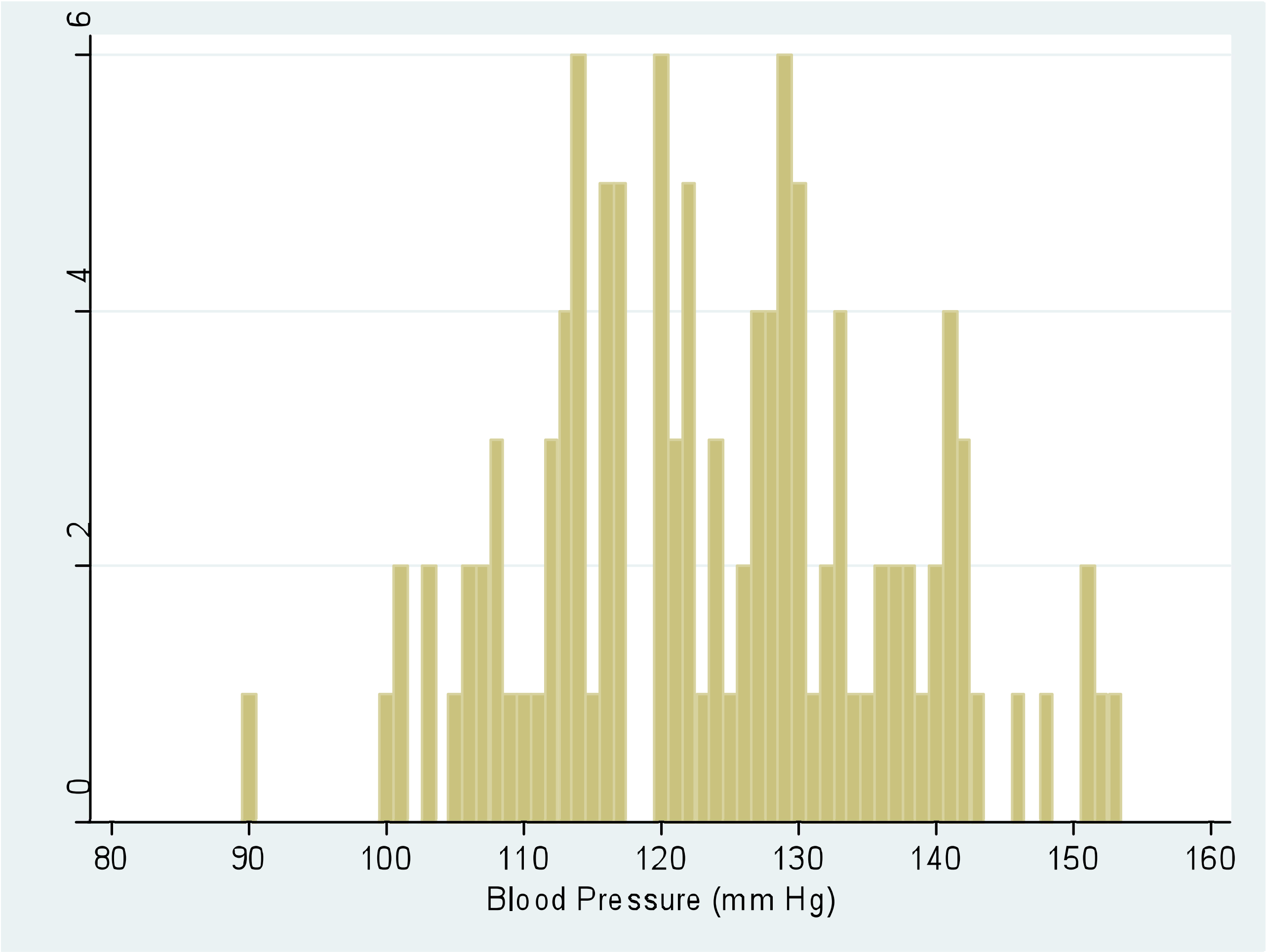
Histogram of the Systolic Blood Pressure for sample of 113 men. Each bar spans a width of five mmHg on the horizontal axis. The height of each bar represents the number of individuals with SBP in that range.

Pictures of Data: Histograms



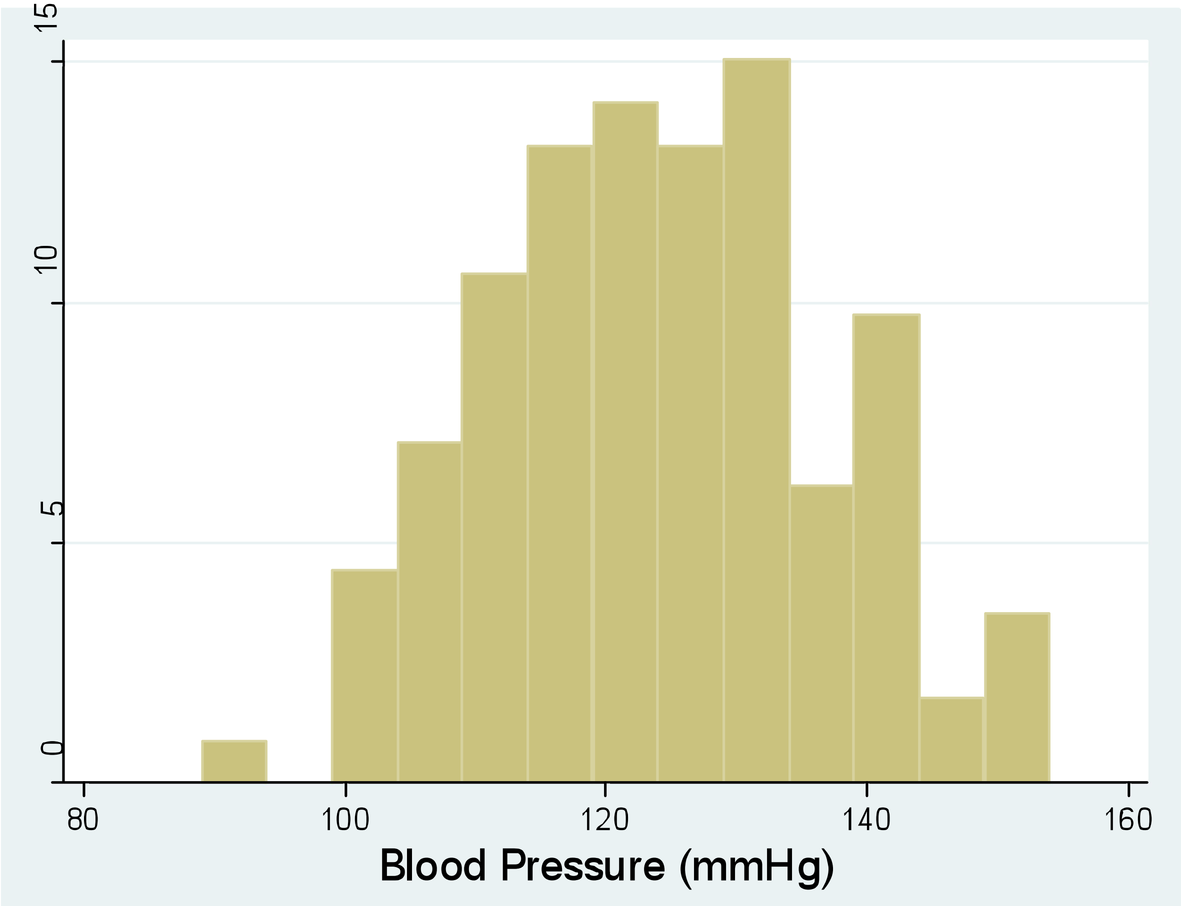
Another histogram of the blood pressure of 113 men. In this graph, each bar has a width of 20 mmHg and there are a total of only four bars making it hard to characterize the distribution of blood pressures in the sample.

Pictures of Data: Histograms



Yet another histogram of the same BP information on 113 men. Here, the bin width is one mmHg, perhaps giving more detail than is necessary.

Other Examples



Another way to present the data in a histogram is to label the y-axis with relative frequencies as opposed to counts. The height of each bar represents the percentage of individuals in the sample with BP in that range. The bar heights should add to one.

Intervals

- How many intervals (bins) should you have in a histogram?
  - There is no perfect answer to this
  - Depends on sample size *n*
  - Rough rule of thumb: # Intervals $\approx \sqrt{n}$

n	Number of Intervals
10	About 3
50	About 7
100	About 10

Section E

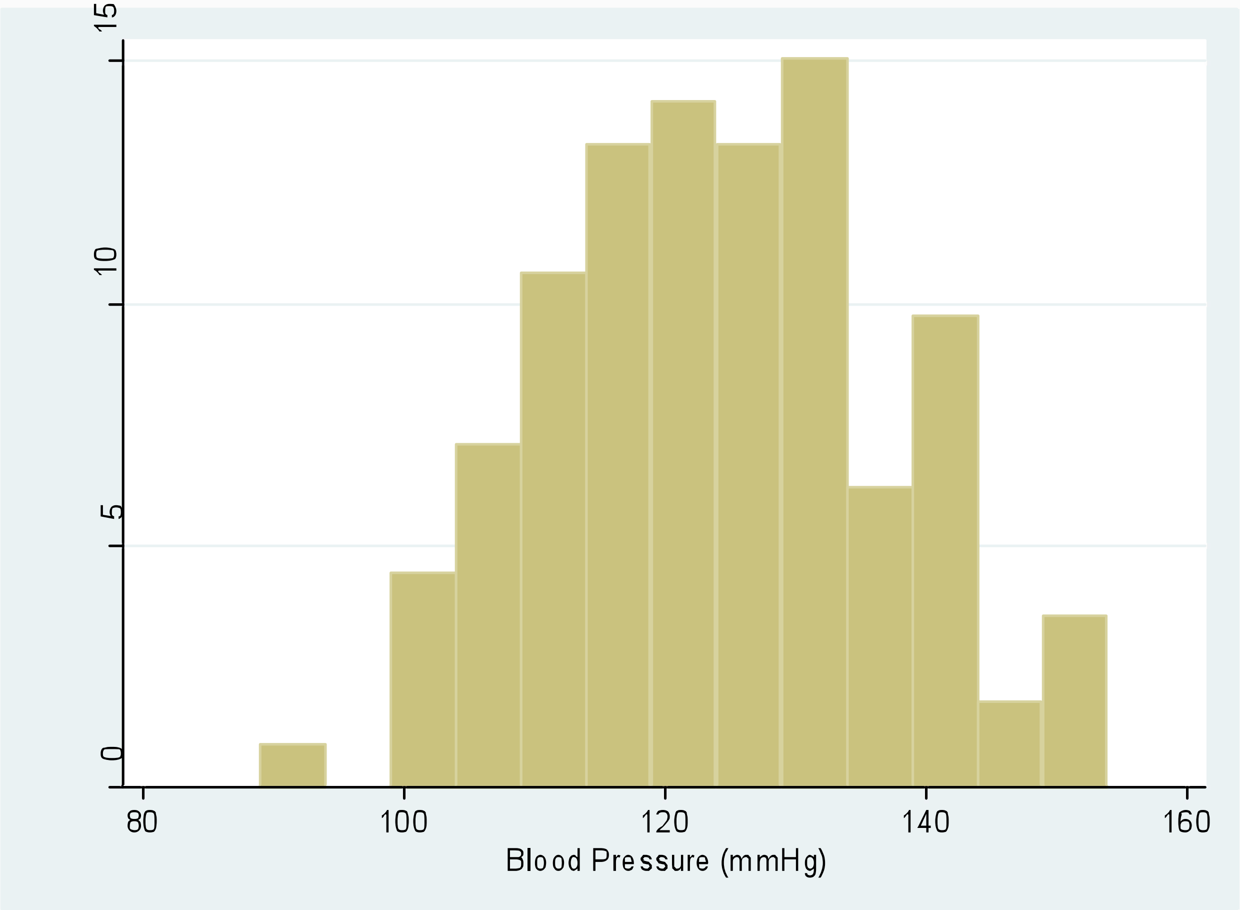
Stem and Leaf Plots, Box Plots

Sample 113 Men

- Suppose we took another look at our random sample of 113 men and their blood pressure measurements

- One tool for “visualizing” the data is the histogram

Histogram: BP for 113 males



Sample 113 Men: Stem and Leaf

- Another common tool for visually displaying continuous data is the "stem and leaf" plot

- Very similar to a histogram
  - Like a "histogram on its side"
  - Allows for easier identification of individual values in the sample

Stem and Leaf: BP for 113 Males

```
  8. | 9
  9* |
  9. | 9
 10* | 11334
 10. | 566777899
 11* | 111223333344444
 11. | 55666667779
 12* | 00000000111223344
 12. | 5566677778888999999
 13* | 000112222334
 13. | 5677789
 14* | 0000112222
 14. | 67
 15* | 0122
```

Stem and Leaf: BP for 113 Males

"Stems"

```
 8. | 9
 9* |
 9. | 9
10* | 11334
10. | 566777899
11* | 11122333344444
11. | 55666667779
12* | 00000000111223344
12. | 5566677778888999999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
```

Stem and Leaf: BP for 113 Males

```
 8. | 9
 9* |
 9. | 9
10* | 11334
10. | 566777899
11* | 1112233333444444
11. | 55666667779
12* | 0000000011122334 4
12. | 5566677778888999999
13* | 00011222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
```

"Leaves"

Stem and Leaf: BP for 113 Males

```
 8. | 9
 9* |
 9. | 9
10* | 11334
10. | 566777899
11* | 111223333344444
11. | 55666667779
12* | 00000000111223344
12. | 5566677778888999999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
```

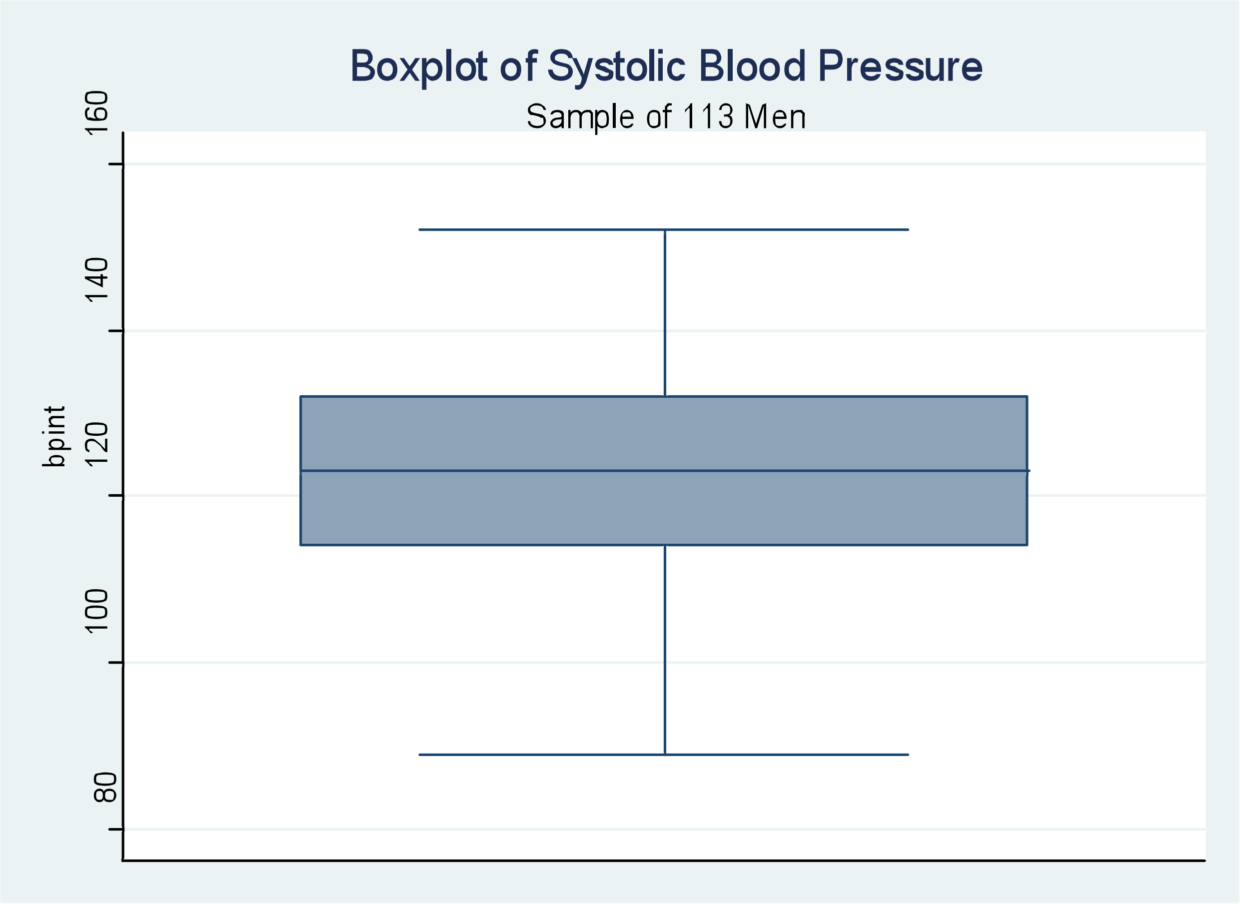
Stem and Leaf: BP for 113 Males

```
 8. | 9
 9* |
 9. | 9
10* | 11334
10. | 566777899
11* | 11122333334444
11. | 5566666677779
12* | 000000001112233 44
12. | 5566677778888999999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
```

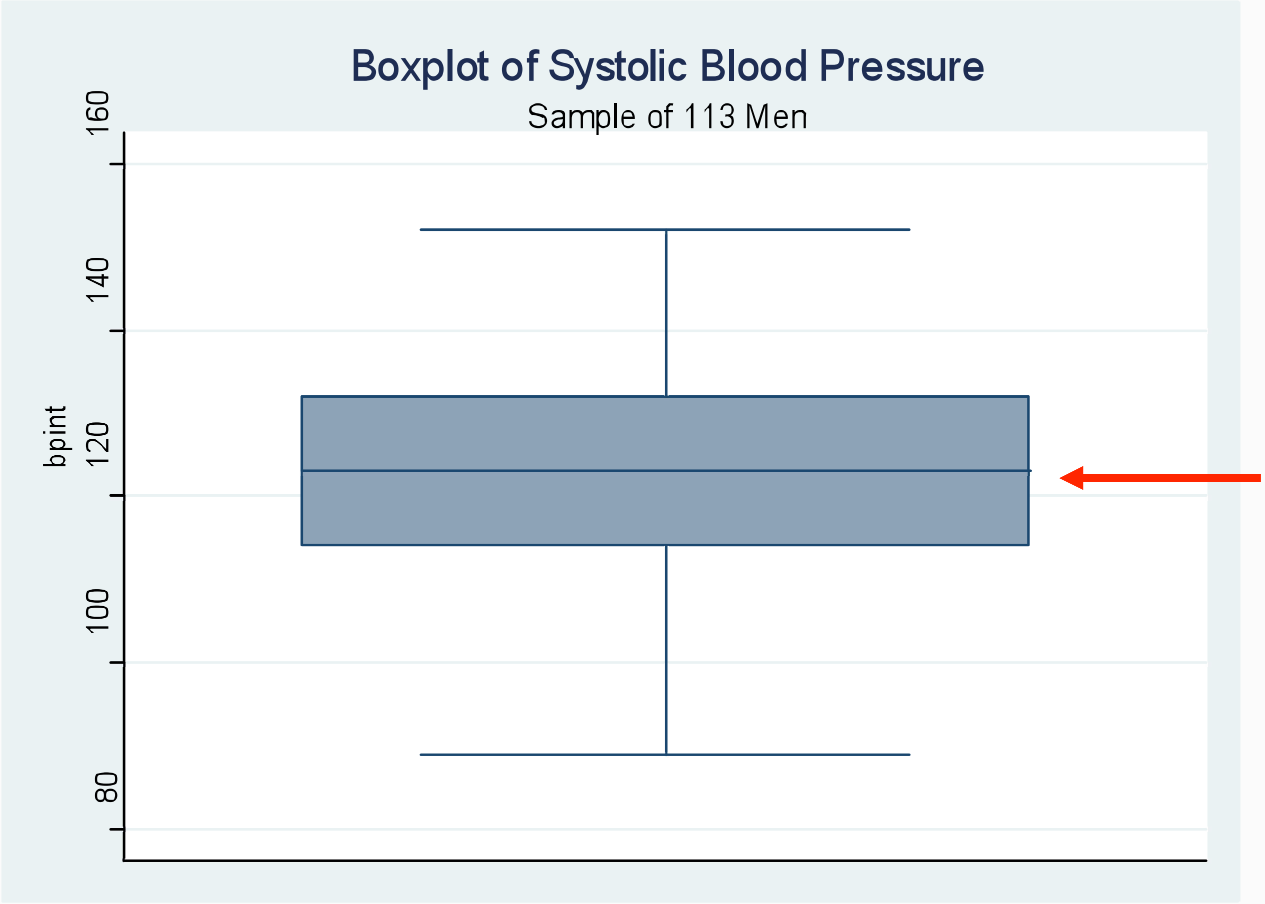
Sample 113 Men: Stem and Boxplot

- Another common visual display tool is the boxplot
  - Gives good insight into distribution shape in terms of skewness and outlying values (extremes: values different than “most” of the rest of the data)
  - Very nice tool for easily comparing distribution of continuous data in multiple groups—can be plotted side by side

Boxplot: BP for 113 Males

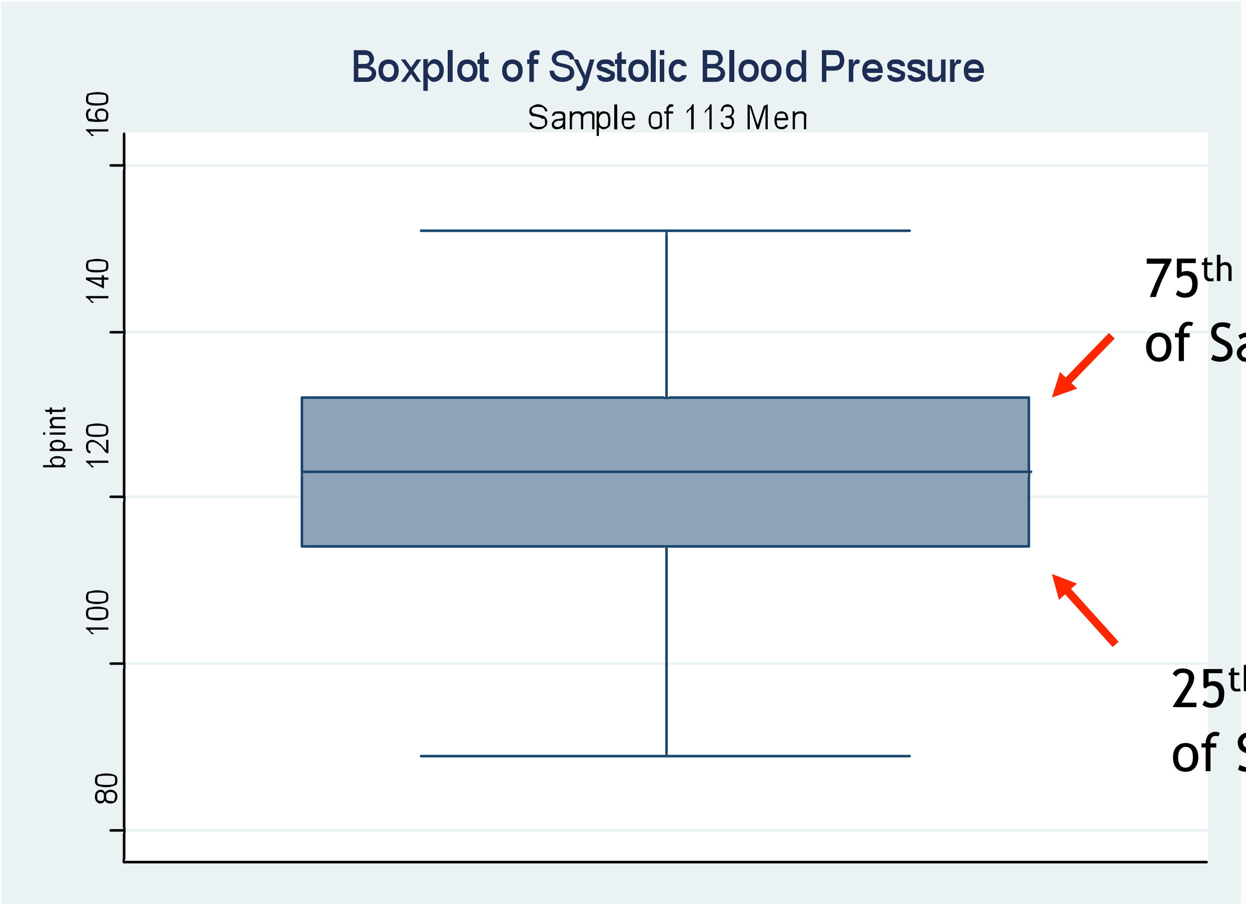


Boxplot: BP for 113 Males



Sample Median Blood Pressure

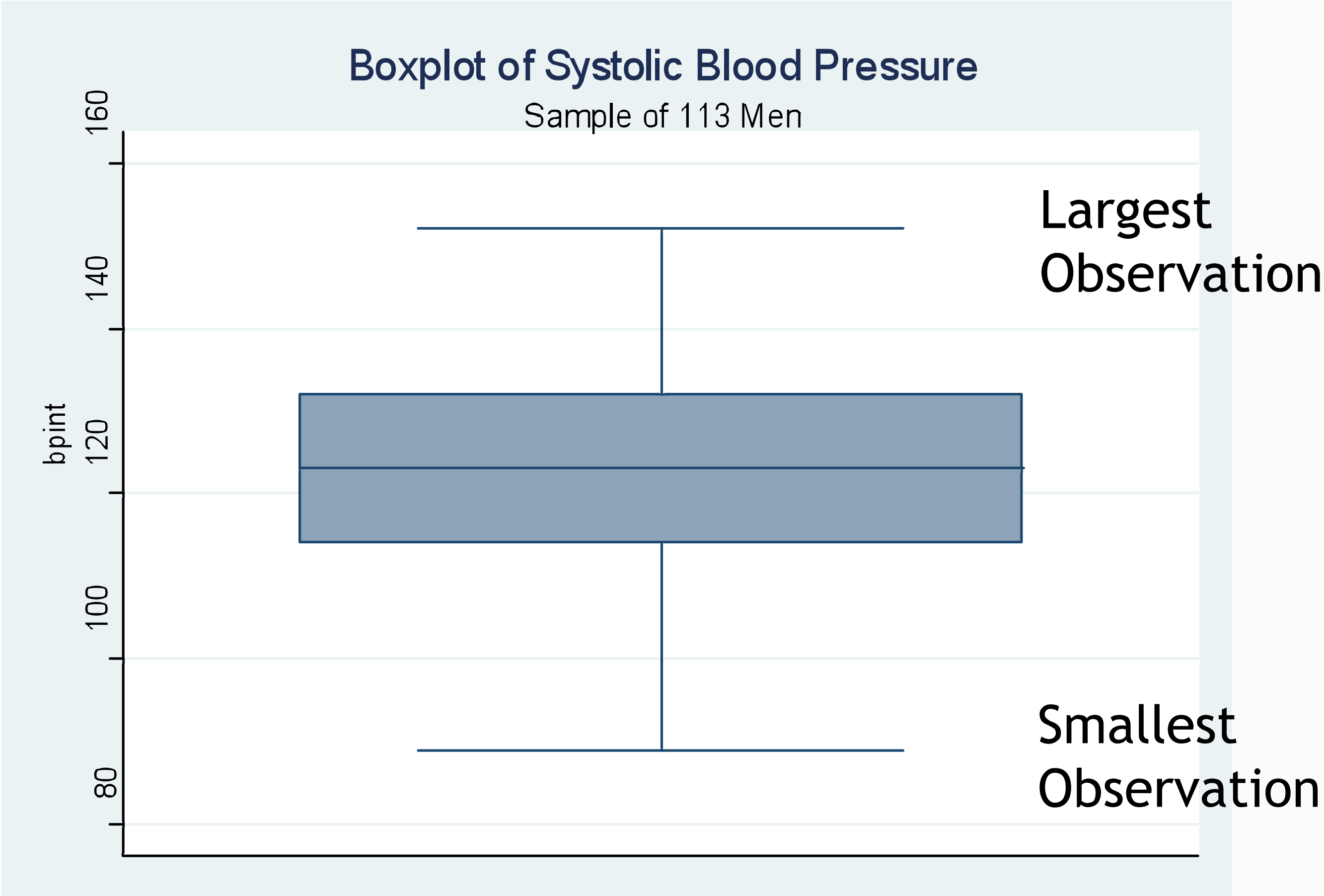
Boxplot: BP for 113 Males



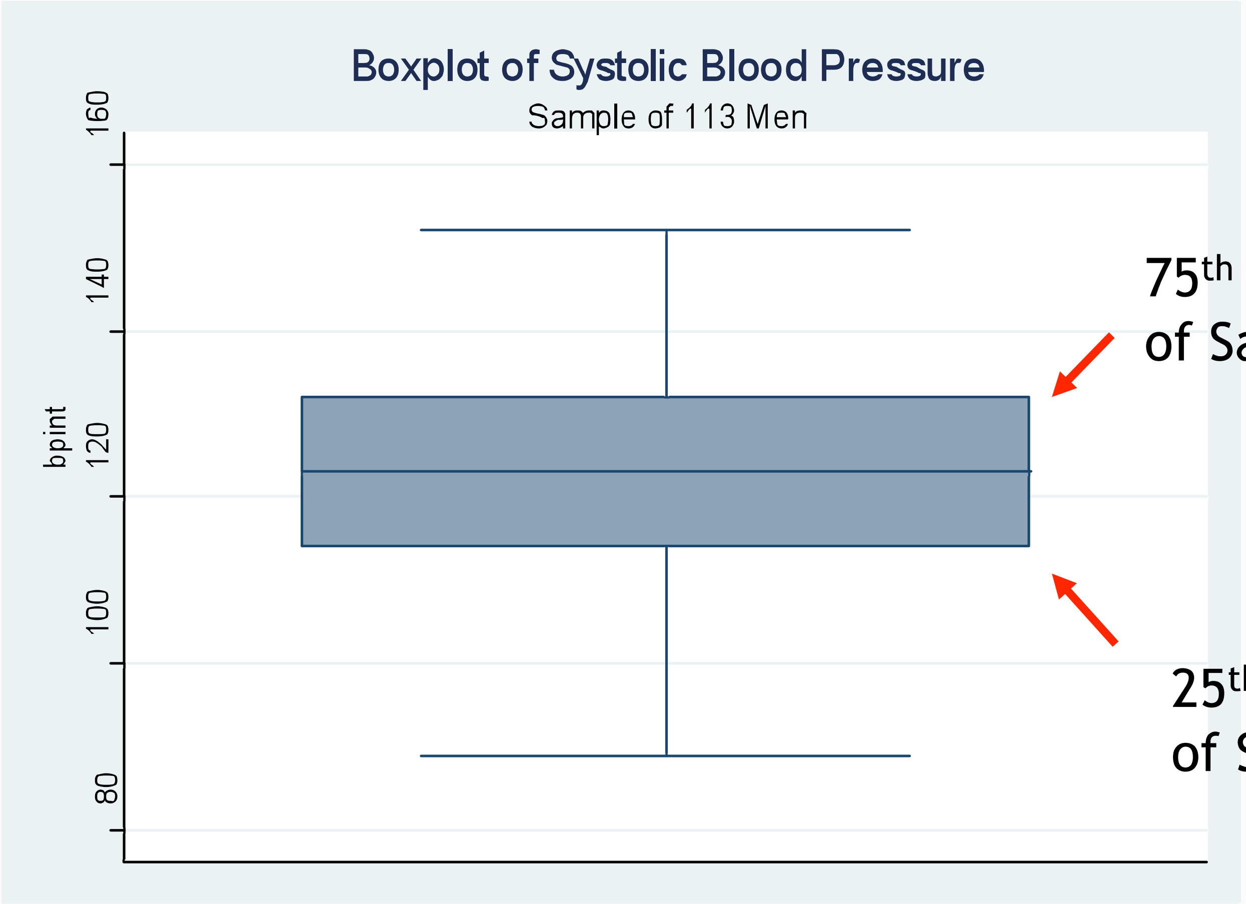
75th Percentile of Sample

25th Percentile of Sample

Boxplot: BP for 113 Males



Boxplot: BP for 113 Males



75th Percentile of Sample

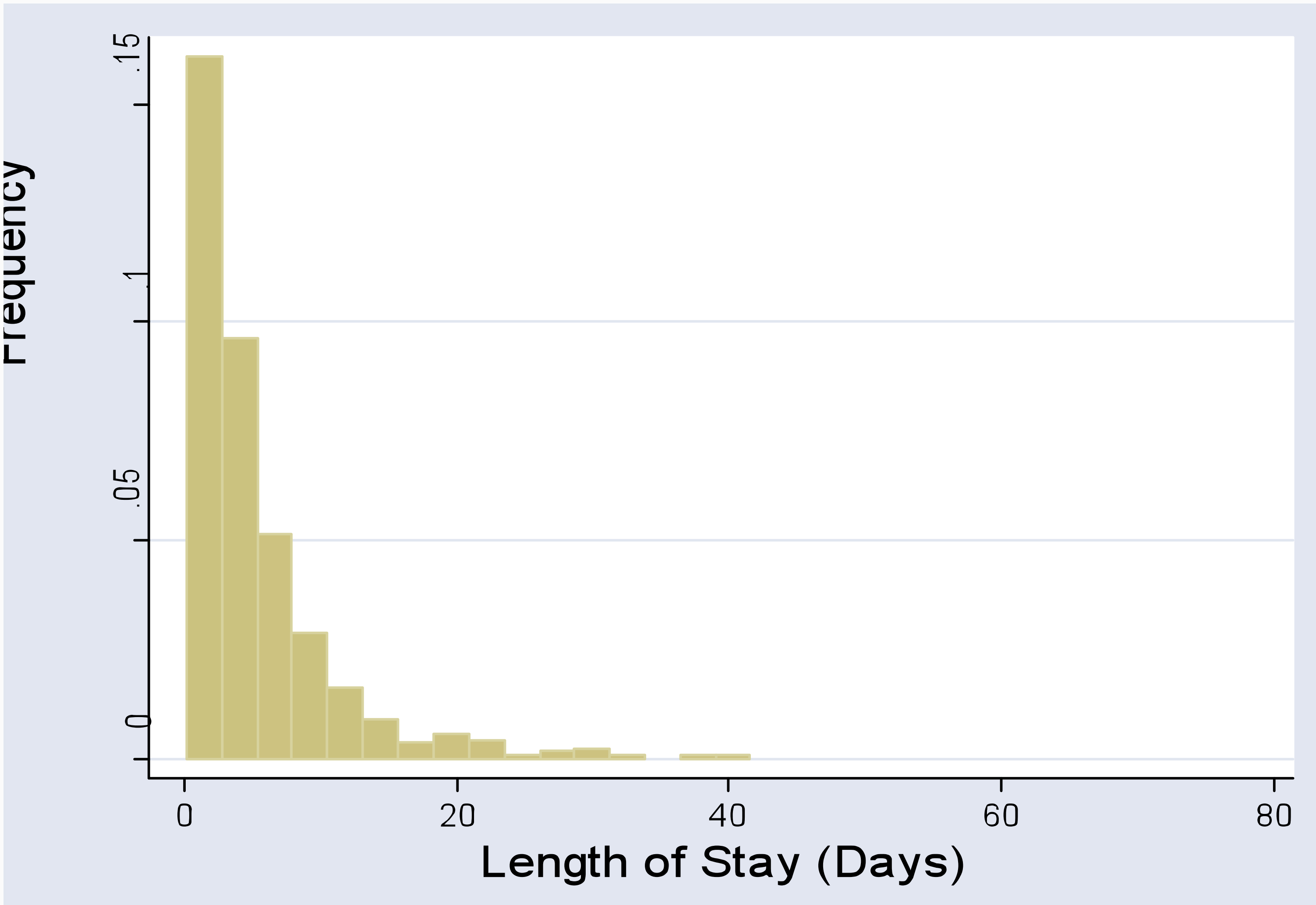
25th Percentile of Sample

Hospital Length of Stay for 1,000 Patients

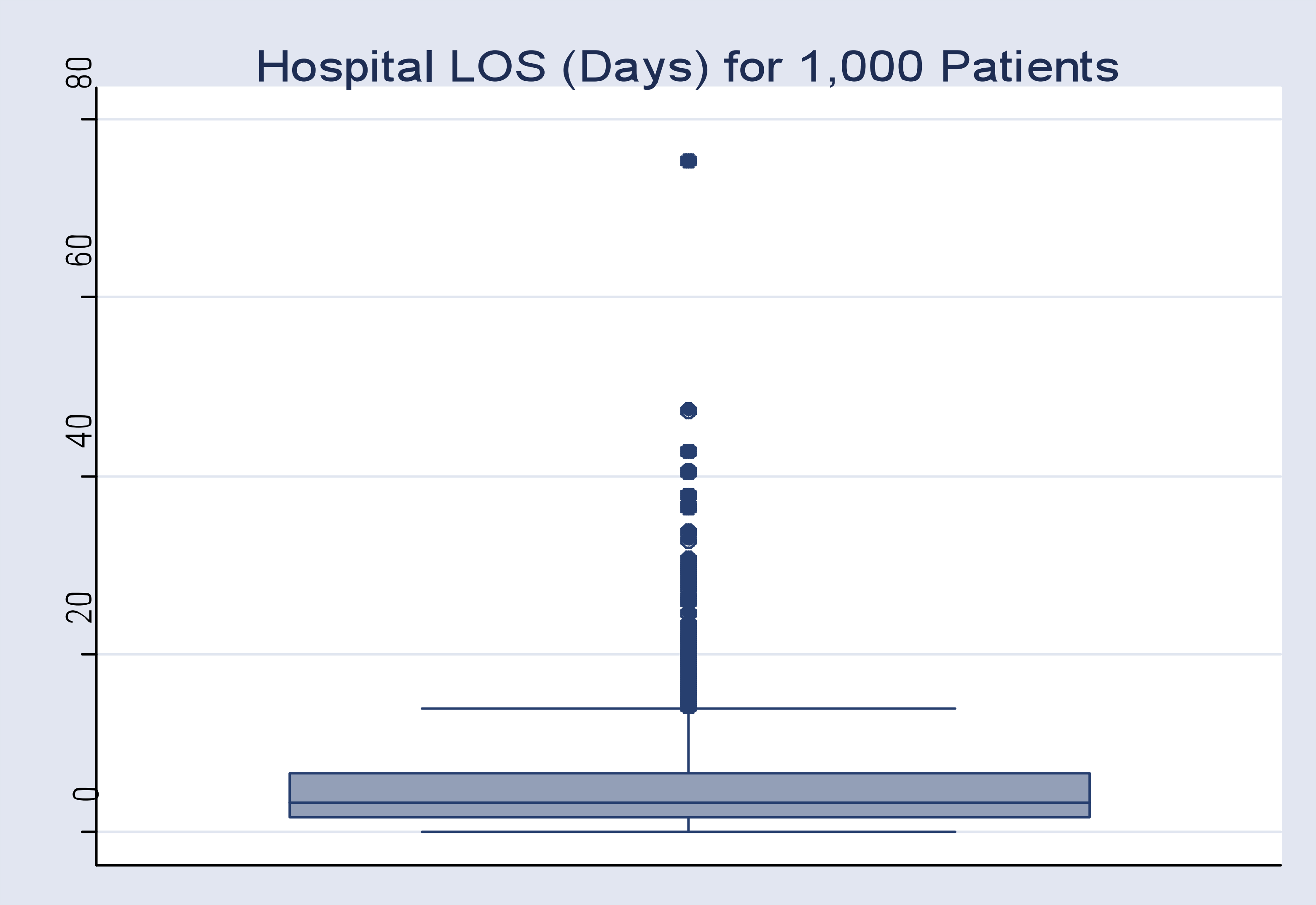
- Suppose we took a representative sample of discharge records from 1,000 patients discharged from a large teaching hospital in a single year

- How could we visualize this data?

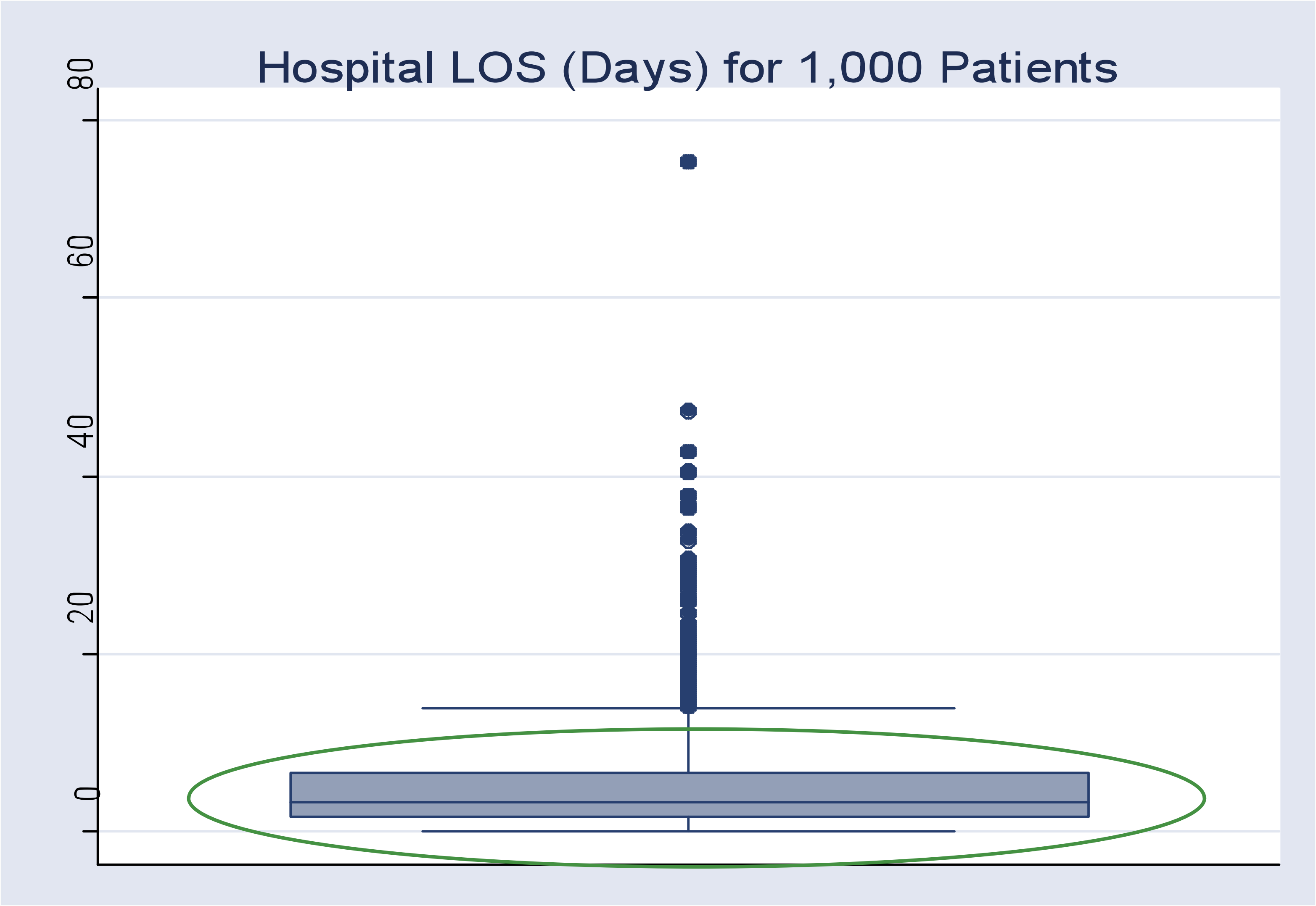
Histogram: Length of Stay



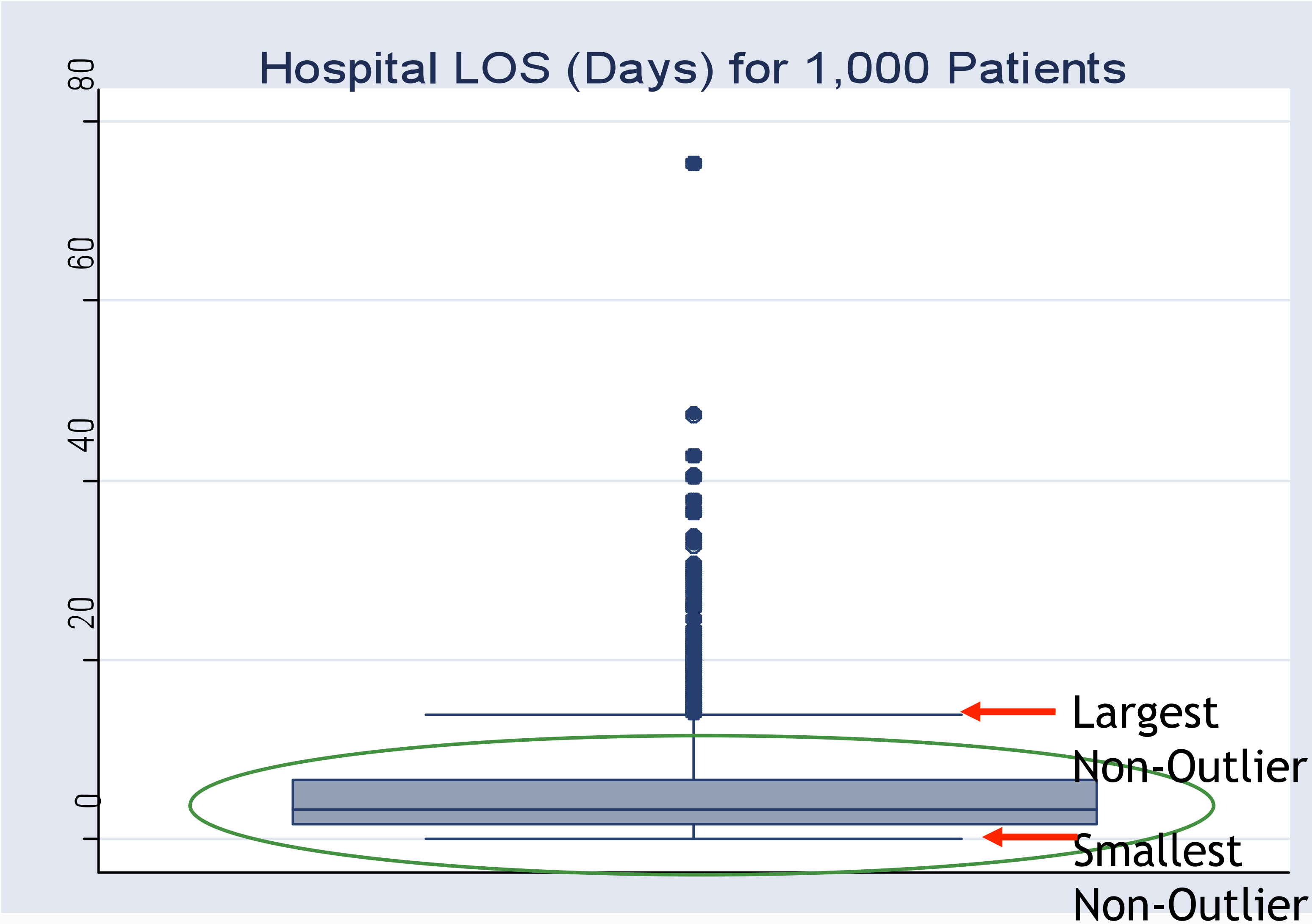
Boxplot: Length of Stay



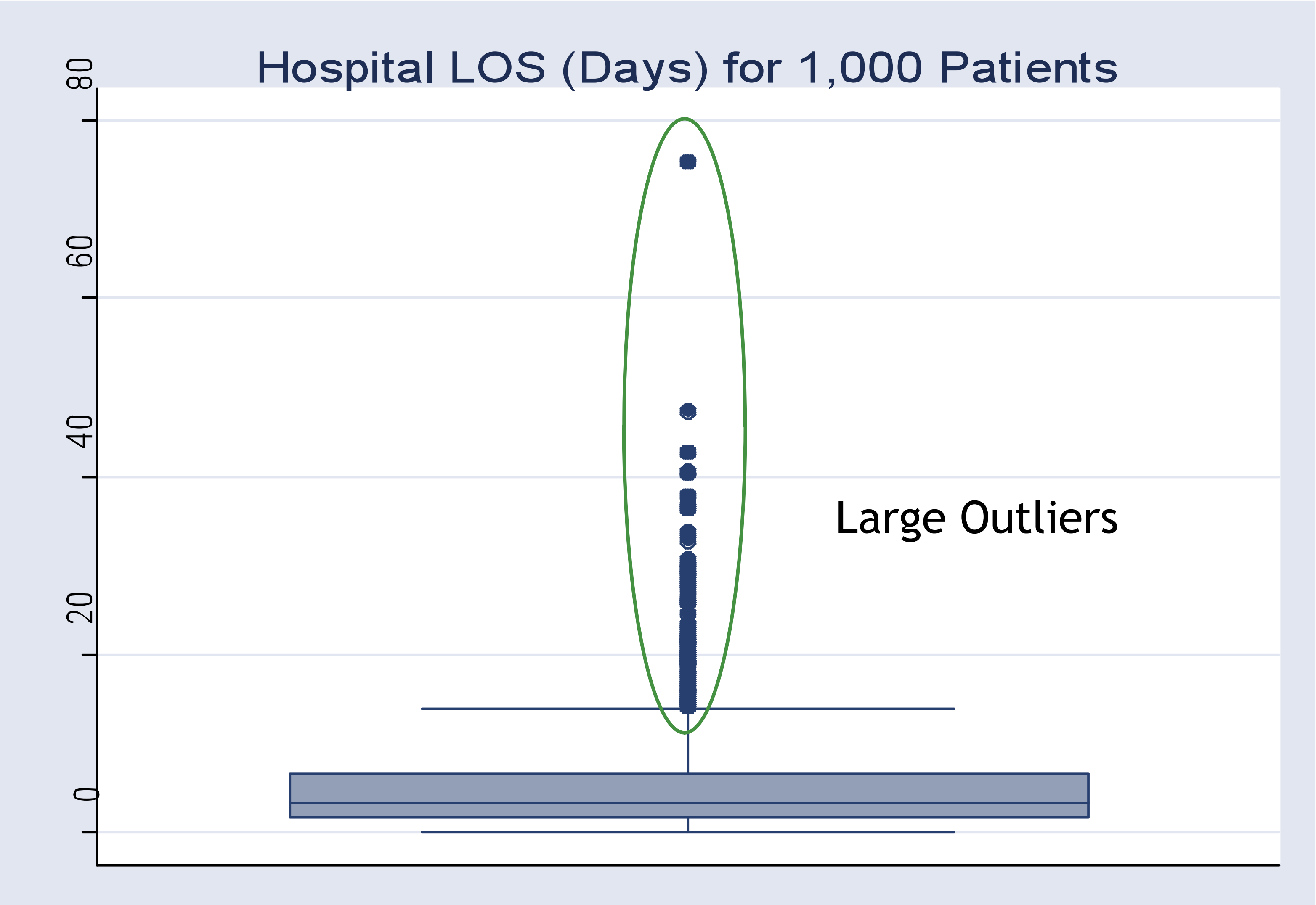
Boxplot: Length of Stay



Boxplot: Length of Stay



Boxplot: Length of Stay

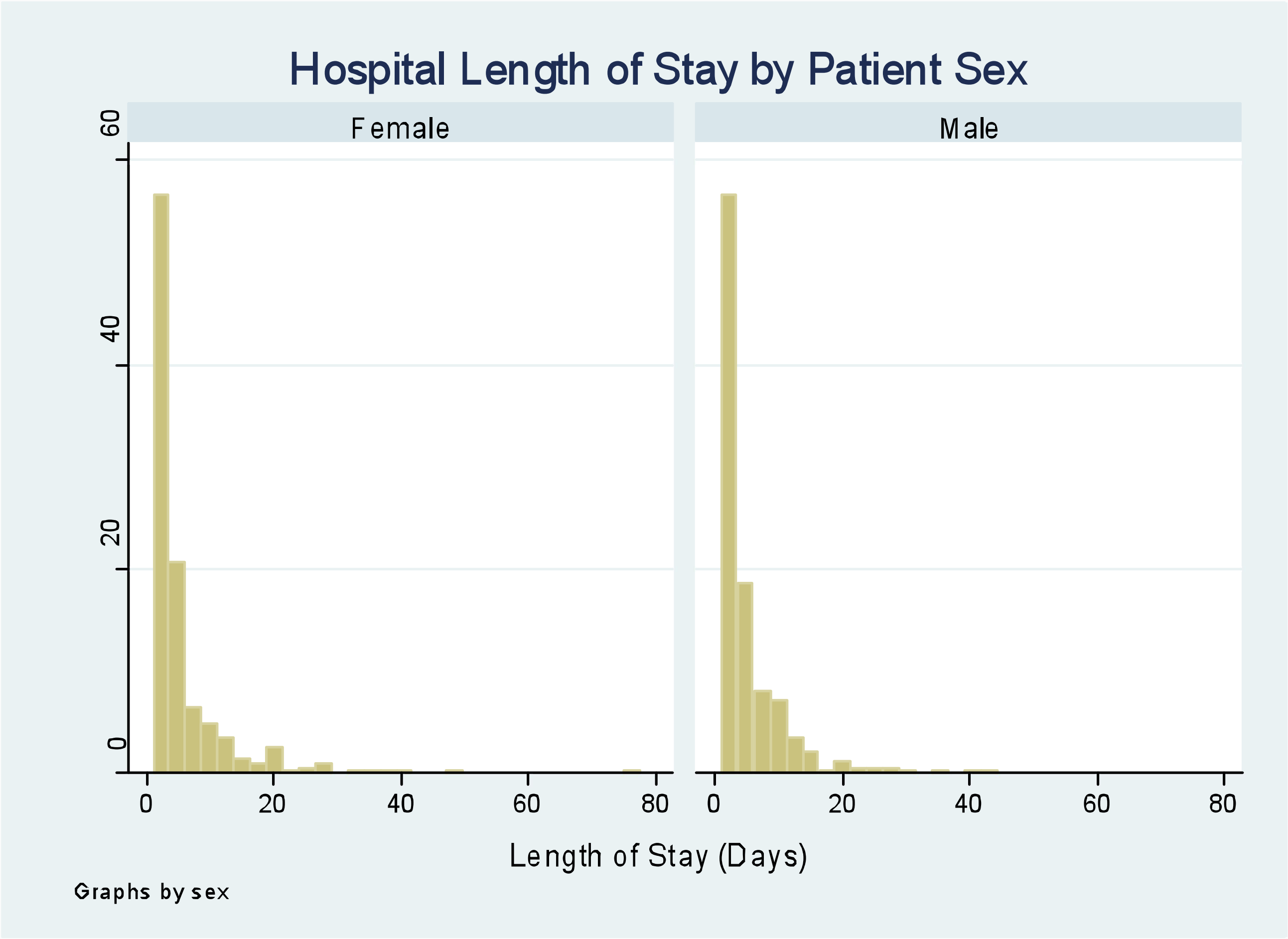


Stem and Leaf: Length of Stay

```
0* | 11111111111111111111111111111111111111111111111111111111111111 ... (298)
0t | 22222222222222222222222222222222222222222222222222222222222222 ... (269)
0f | 44444444444444444444444444444444444444444444444444444444444444 ... (140)
0s | 666666666666666666666666666666666666666666666666666666667777777 ... (95)
0. | 888888888888888888888888888888888999999999999999999999999999999
1* | 000000000000000000011111111111
1t | 22222222222222222222222333333333333
1f | 444444445555
1s | 66666777
1. | 8889999999
2* | 000001111111
2t | 2233
2f | 455
2s | 6677
2. | 889999
3* | 0
3t | 23
3f |
3s | 667
3. |
4* | 00
4t | 2
4f |
4s | 7
4. |
5* |
5t |
5f |
5s |
5. |
6* |
6t |
6f |
6s |
6. |
7* |
7t |
7f | 5
```

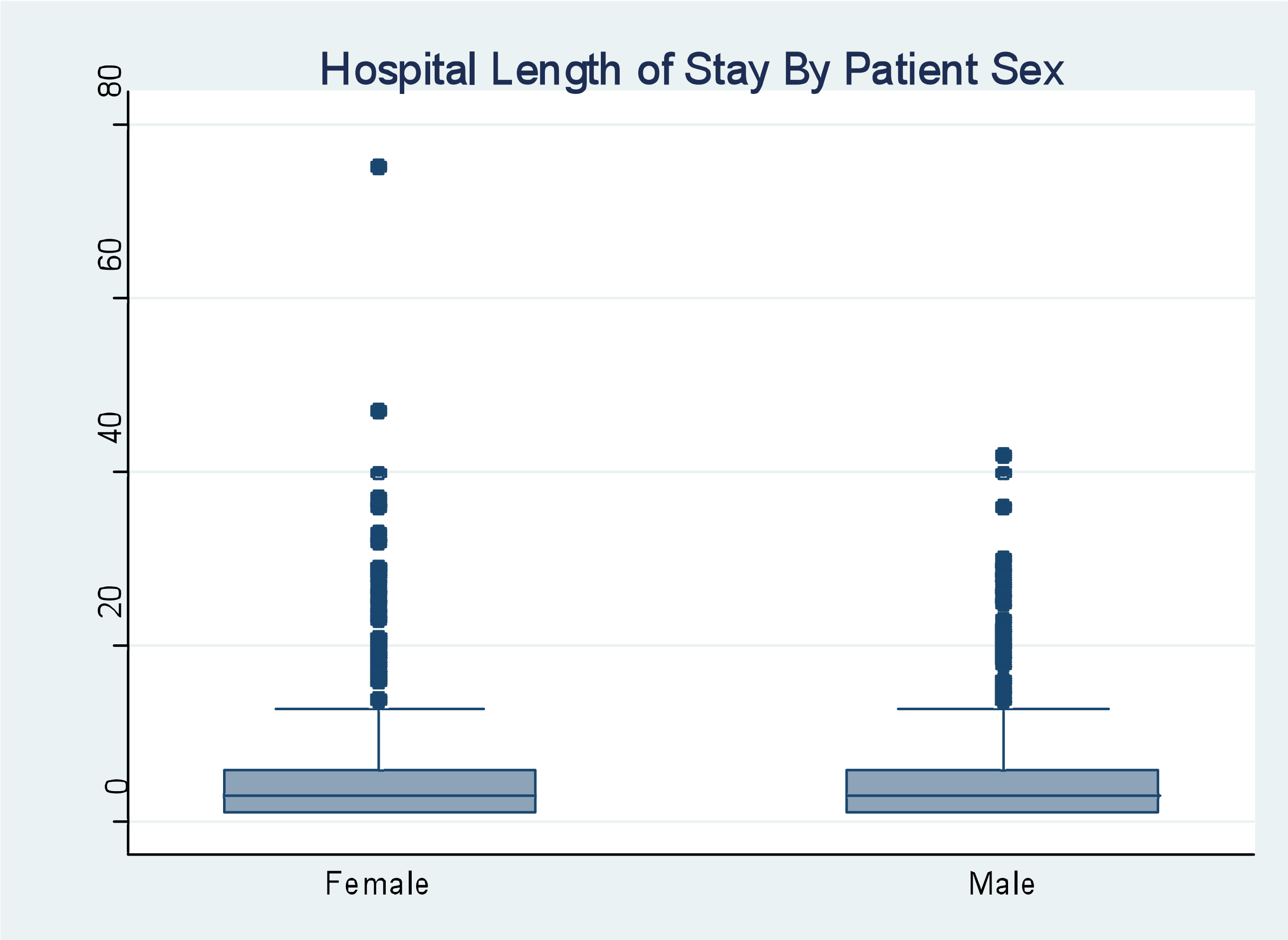
Side by Side Distribution Comparison

- Side by side histograms of length of stay for female and male patients in sample



Side by Side Distribution Comparison

- Side by side boxplots of length of stay for female and male patients in sample



JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

Section F

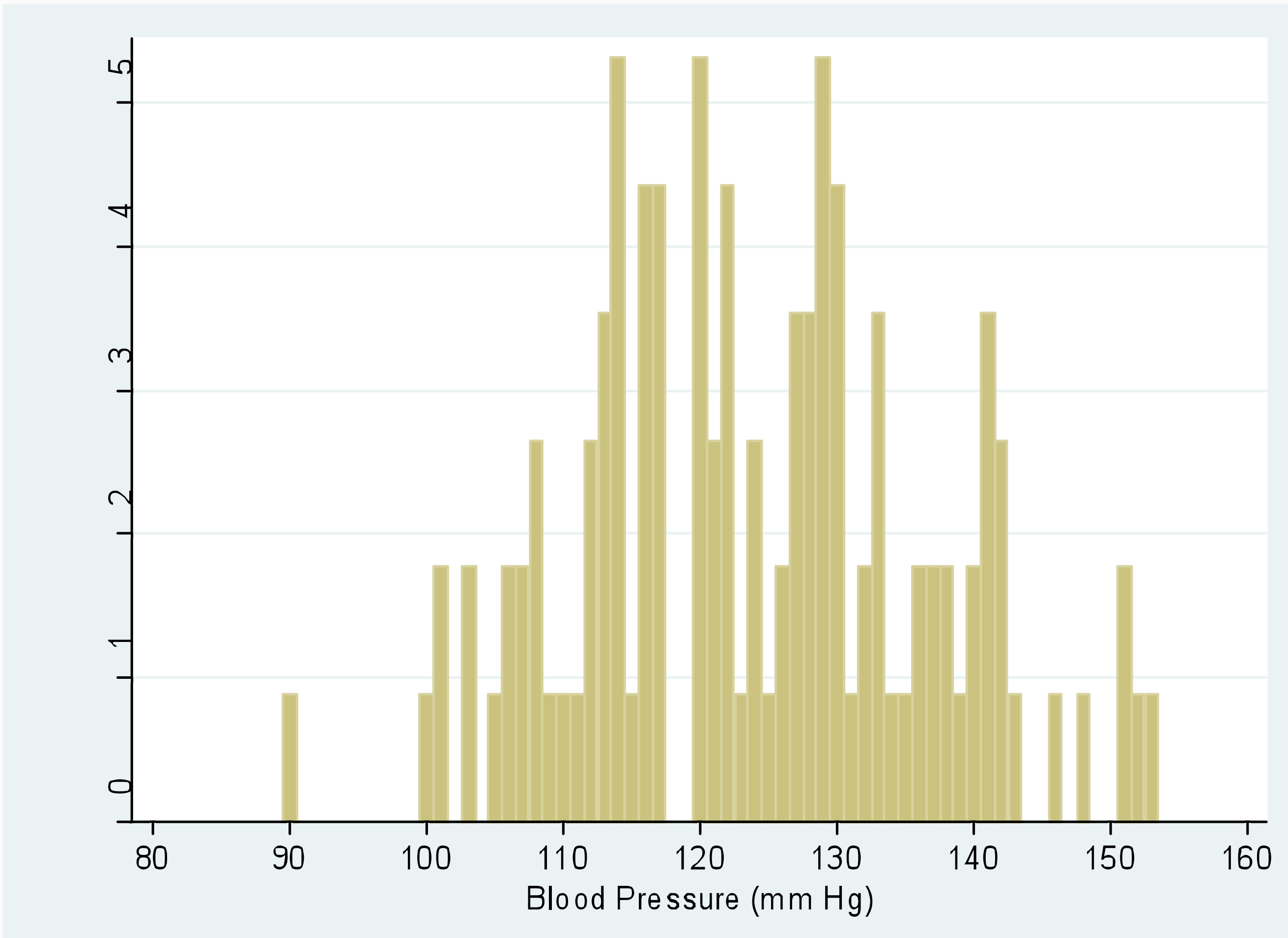
Samples versus Populations, Part 2: Sample Distribution versus Underlying “Population Distribution”

Sample Distribution

- In research, samples are taken from larger population
- If the sample is taken randomly, the sample characteristics will imperfectly mimic the population characteristics
- The characteristics include the mean, median and sd (but also the distribution of individual values)

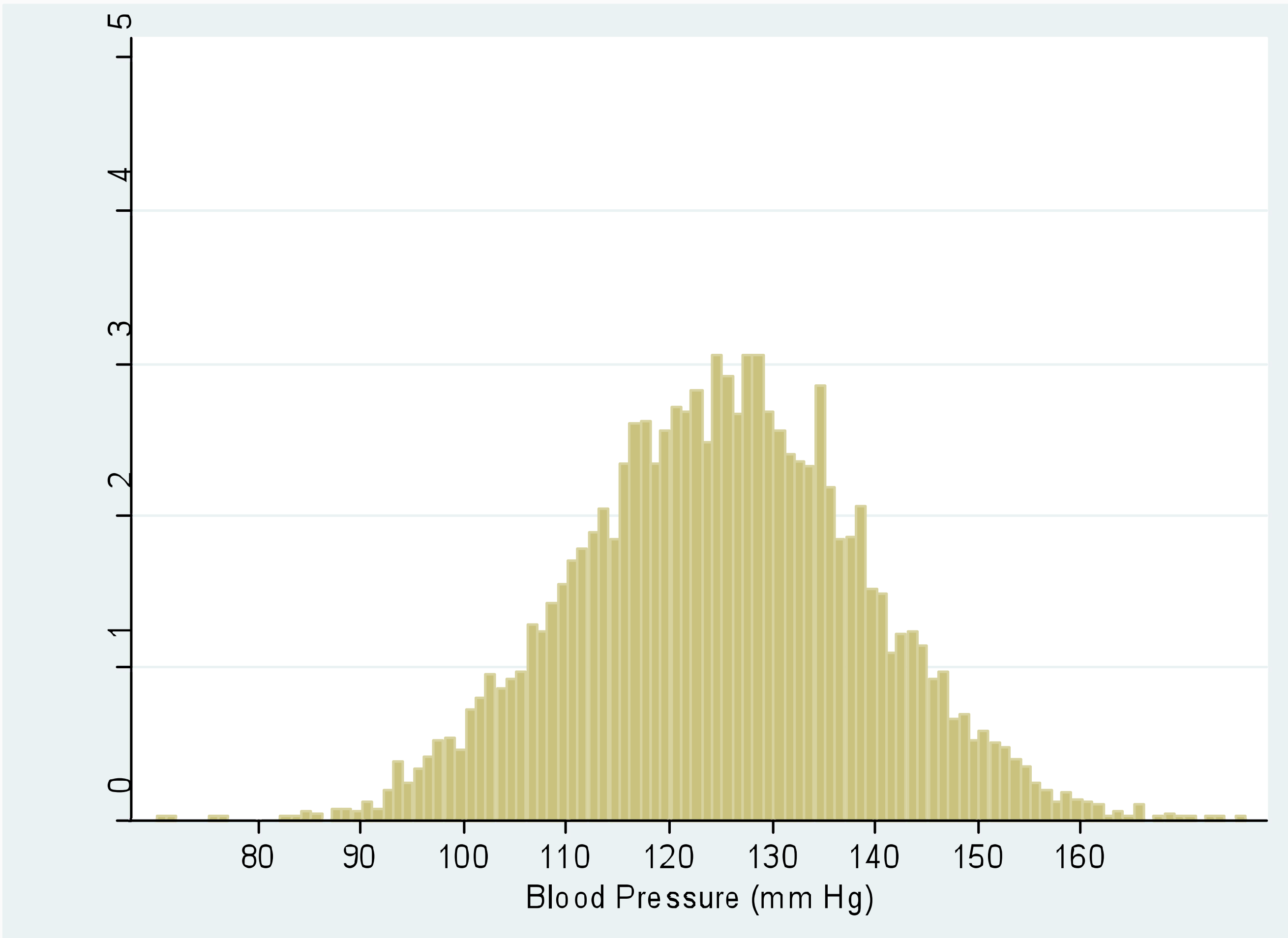
Example 1: Blood Pressure in Males

- Histogram of BP values for random sample of 113 men



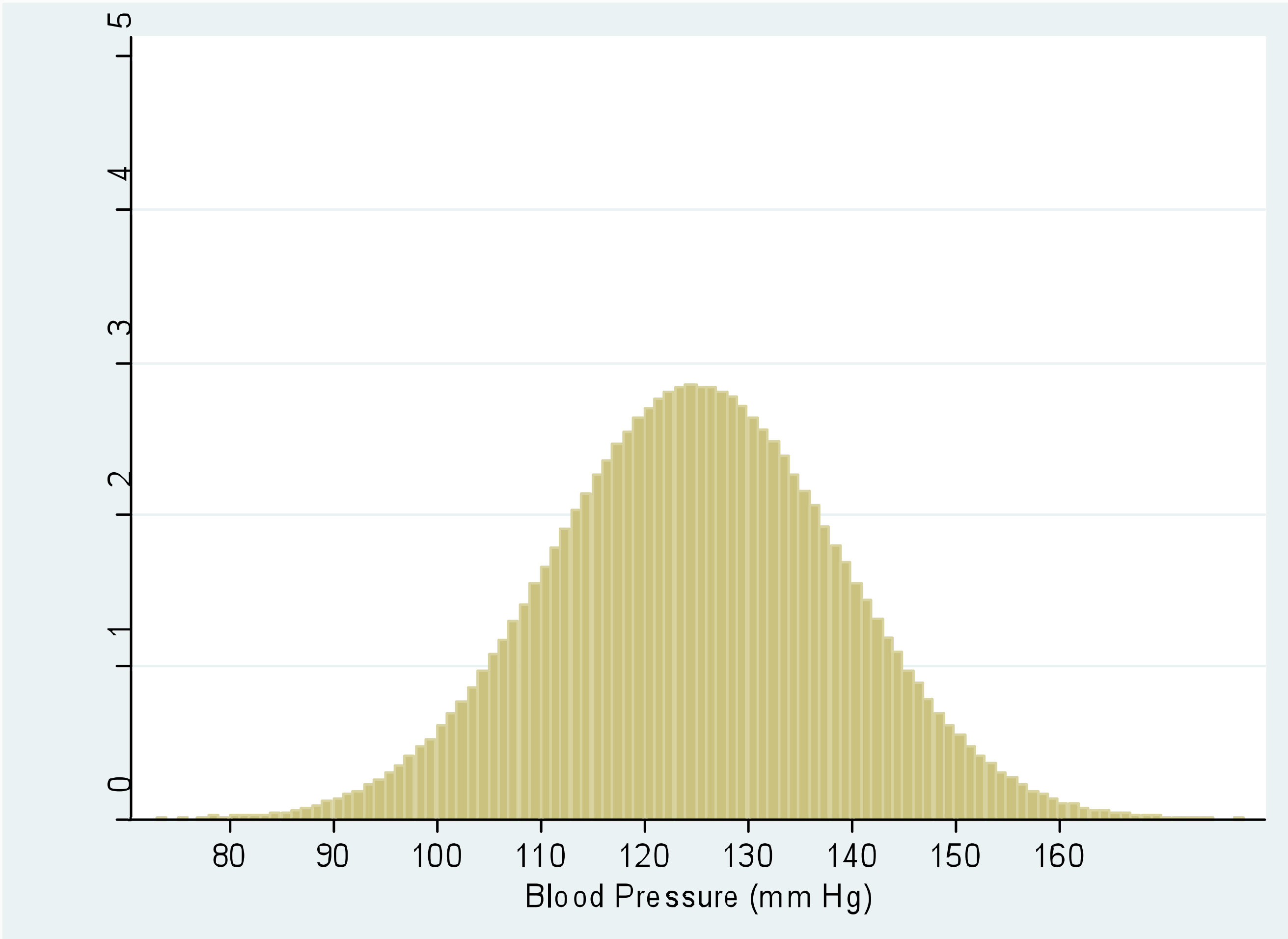
Example 1: Blood Pressure in Males

- Histogram of BP values for random sample of 500 men



Example 1: Blood Pressure in Males

- Histogram of BP values for male population

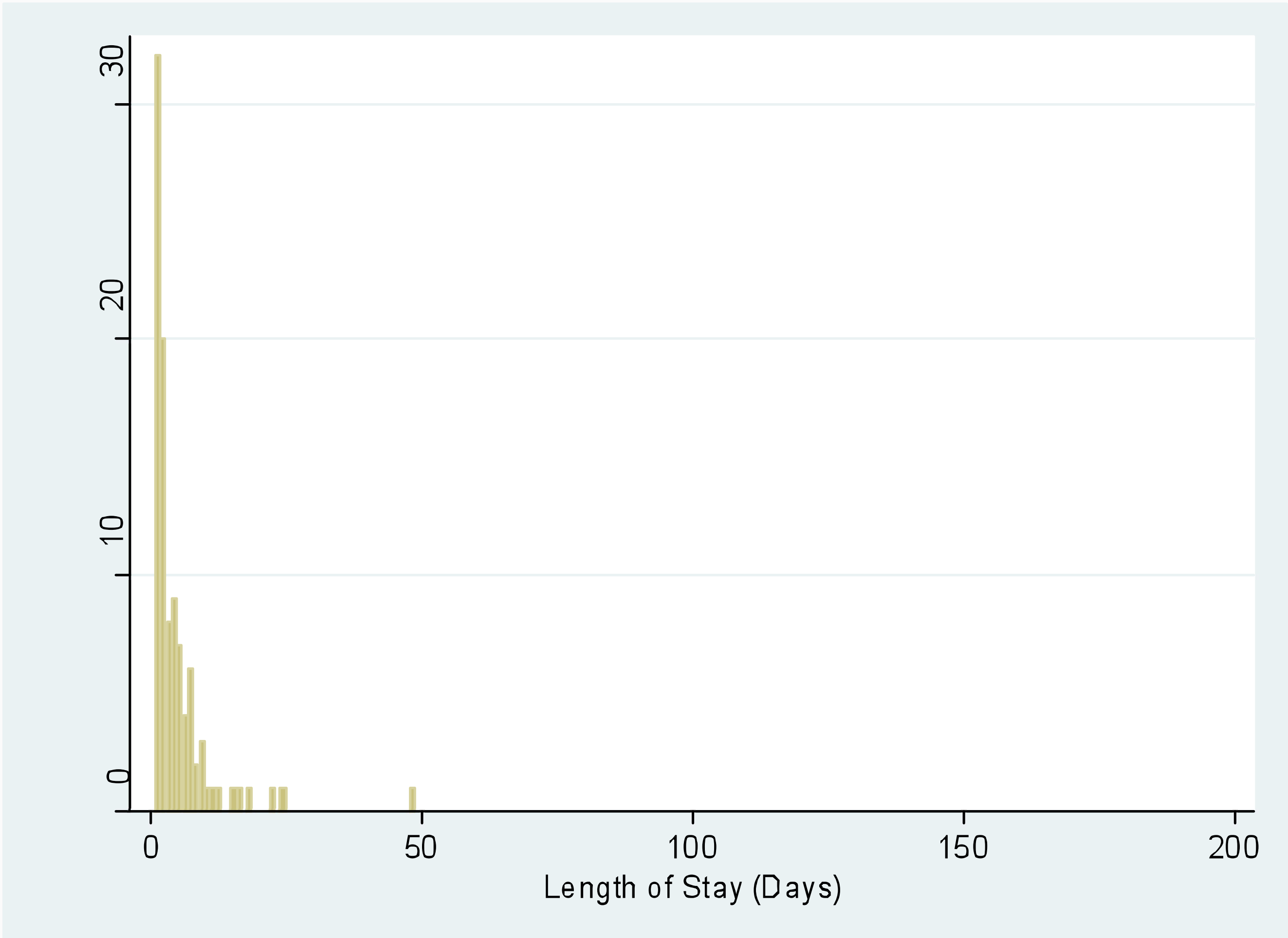


The Histogram and the Probability Density

- The *probability density* is a smooth idealized curve that shows the shape of the distribution in the population
- This is generally a theoretical distribution that we can never see: we can only estimate it from the distribution presented by a representative (random) sample from the population
- Areas in an interval under the curve represent the percentage of the population in the interval
- The distributions shown are indicative of a symmetric, bell shaped distribution for blood pressure measurements in men

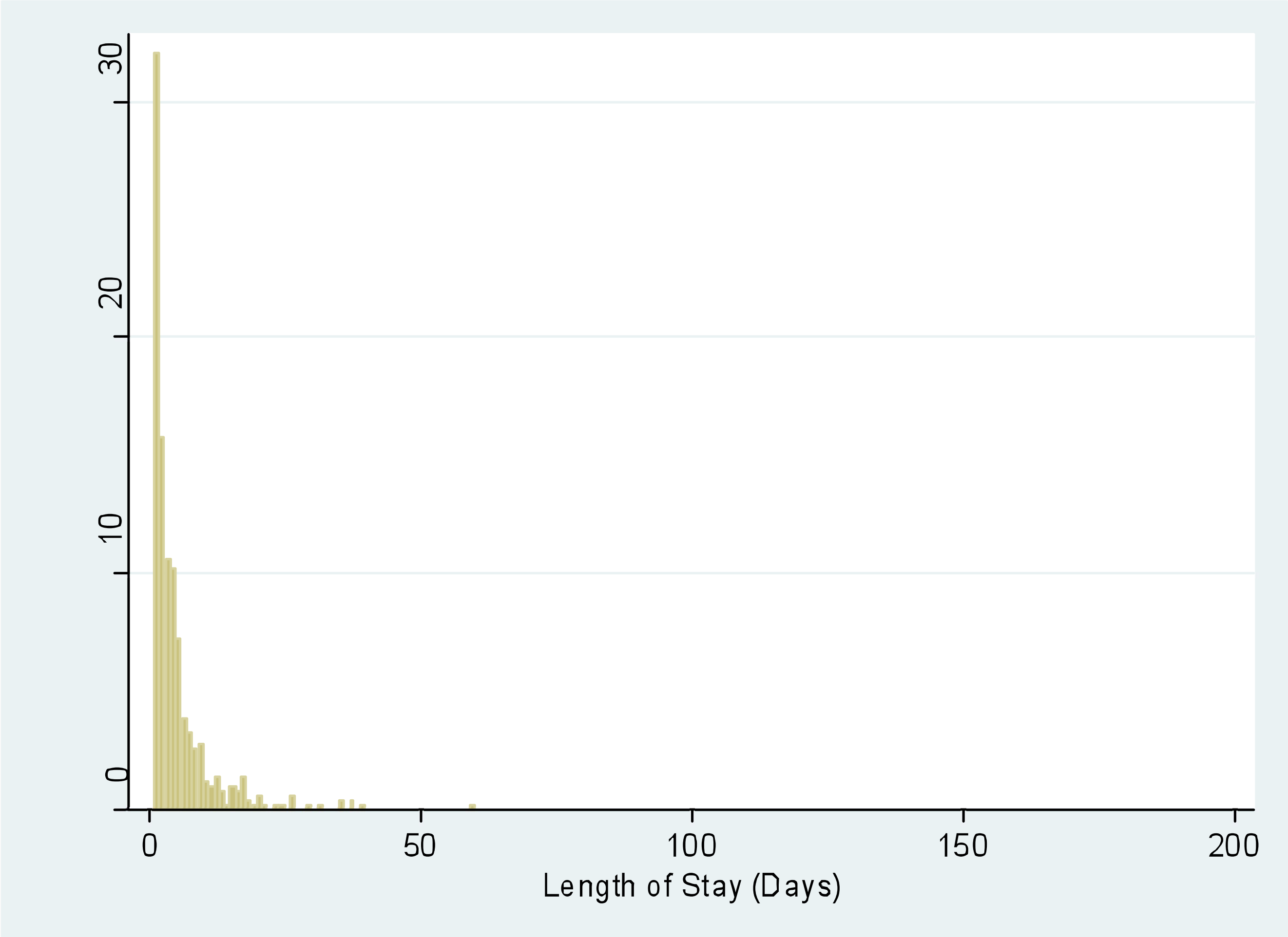
Example 2: Hospital Length of Stay

- Histogram of LOS values for 100 patients



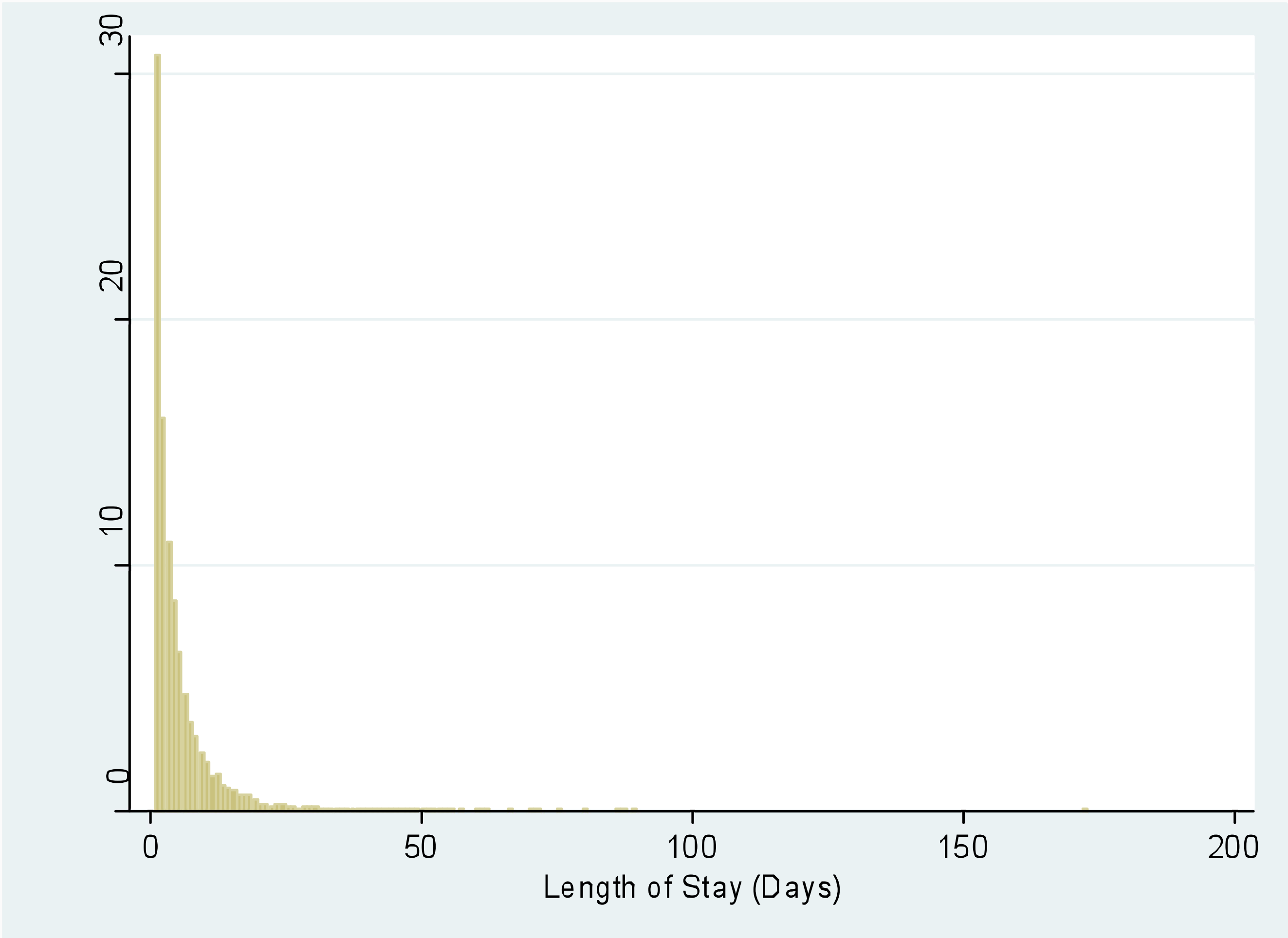
Example 2: Hospital Length of Stay

- Histogram of LOS values for 500 patients



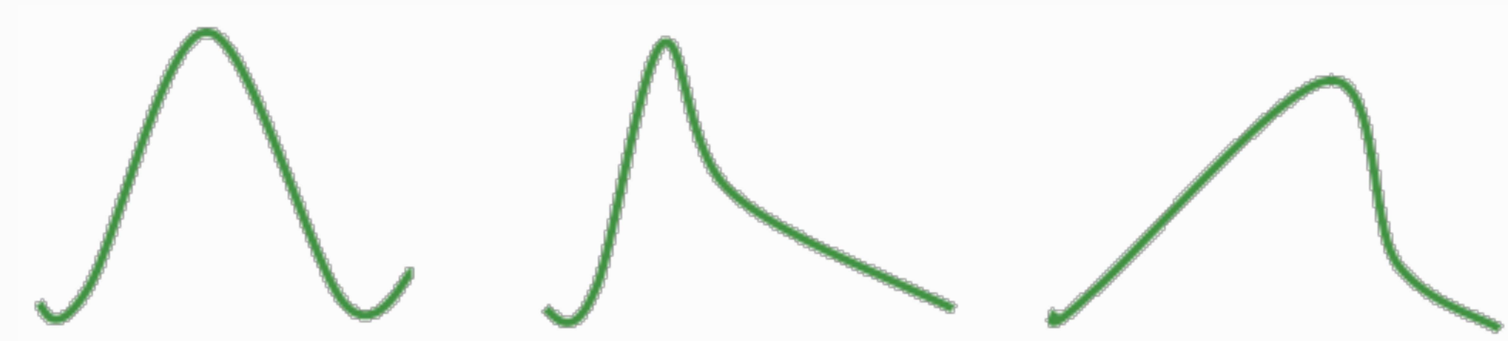
Example 2: Hospital Length of Stay

- Histogram of LOS values for all patients



Common Shapes of the Distribution

- Some shapes of data distributions



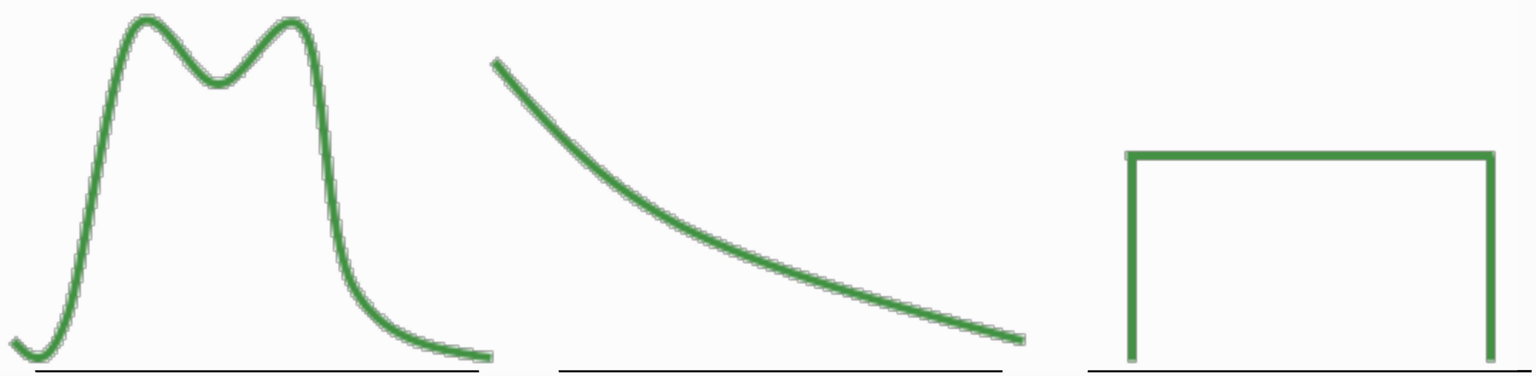
A Symmetrical and bell shaped

B Positively skewed or skewed to the right

C Negatively skewed or skewed to the left

Shapes of the Distribution

- Some possible shapes for frequency distributions



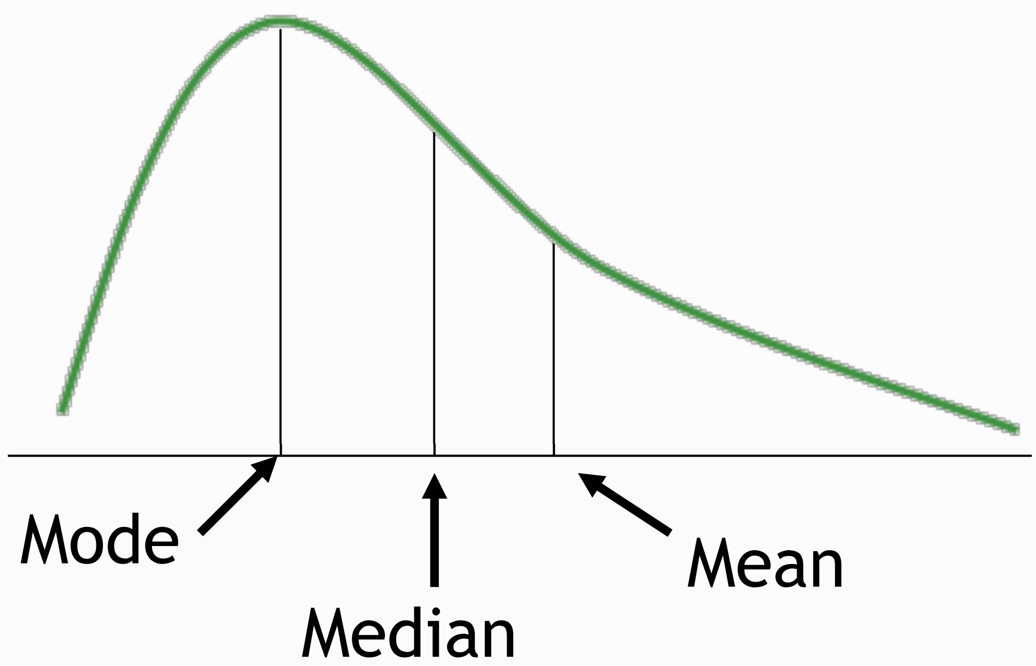
A Bimodal

B Reverse J-shaped

C Uniform

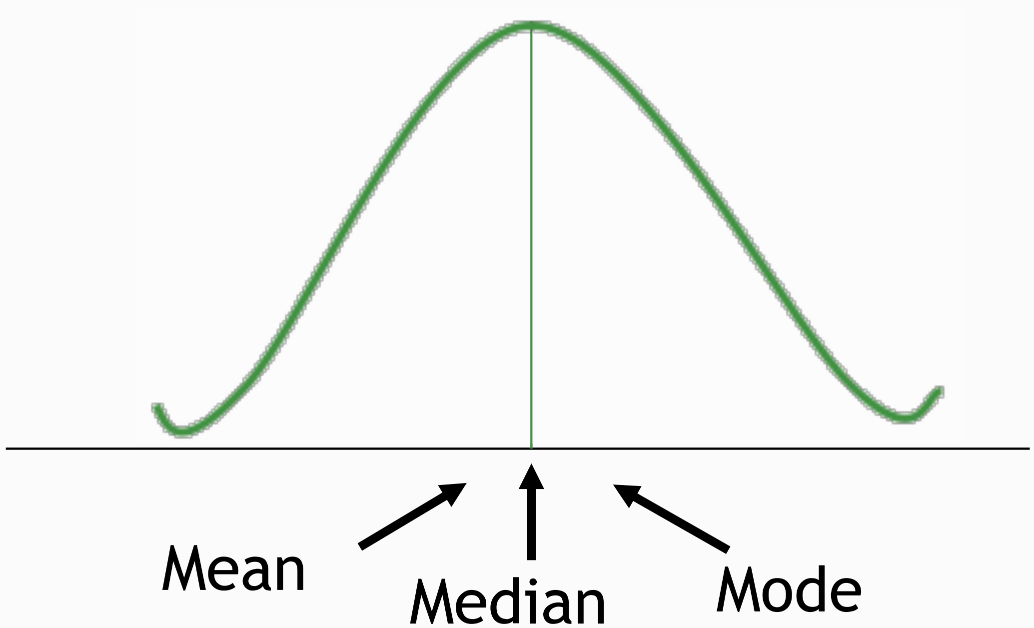
Distribution Characteristics

- Mode: Peak(s)
- Median: Equal areas point
- Mean: Balancing point



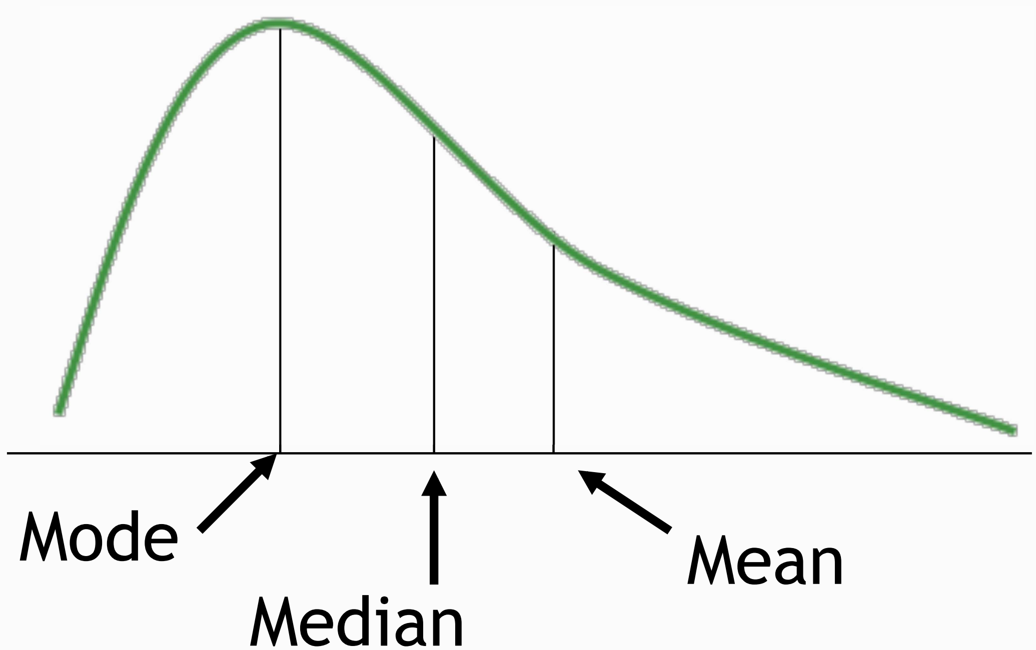
Shapes of Distributions

- *Symmetric* (right and left sides are mirror images)
  - Left tail looks like right tail
  - Mean = Median = Mode



Shapes of Distributions

- *Right skewed* (positively skewed)
  - Long right tail
  - Mean > Median



Shapes of Distributions

- *Left skewed* (negatively skewed)
  - Long left tail
  - Mean < Median