ISTA 116: Statistical Foundations for the Information Age

Measures of Central Tendency

31 August 2011

Outline

- Lab Assignment 1 Due Friday via d2l dropbox (unless otherwise specified by your lab instructor)
- No class Monday (Labor Day)

Example: Heights of Four-Year-Olds in Inches

2	4 4
3	0 3 6 6 6 6 6 7 8 8 9 9
4	0 0 0 1 1 3 5

Ratio of data to bins is pretty high. Maybe try subdividing.

2	4 4
2	
3	0 3
3	6 6 6 6 6 7 8 8 9 9
4	0 0 0 1 1 3
4	5

Important tips:

- Make sure all bins have the same # of possible values
- Keep digits vertically aligned so that horizontal space corresponds to # of observations

Example: Heights of Four-Year-Olds in Inches

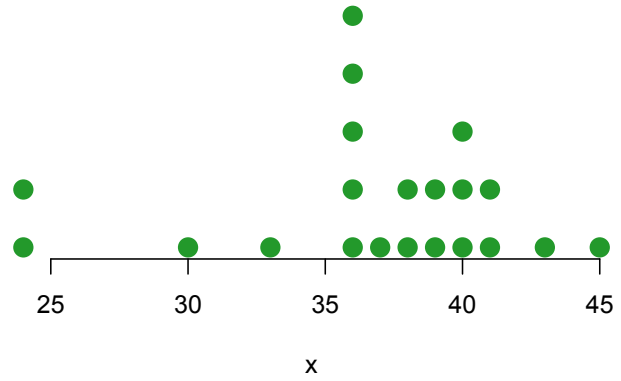


Figure: Height of 4-year-old Children in Inches

Question: When might you prefer a stem-and-leaf plot to a strip chart, or vice-versa?

Stem-and-Leaf	**Strip Chart**
Can bin nearby values	Less vertical space
More easily read exact values	A bit "cleaner"
	Easier to see exactly repeated values

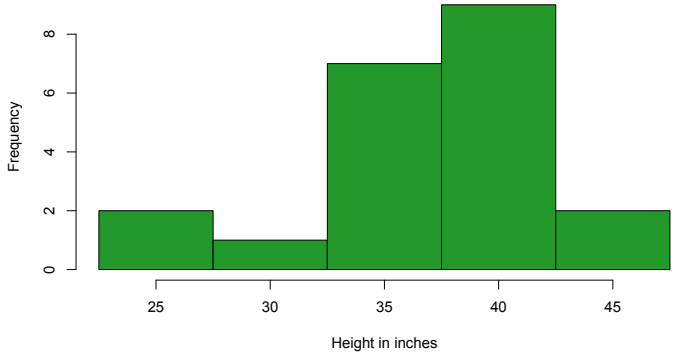
- Both stem-and-leaf plots and strip charts can be cluttered if there's lots of data

Another closely-related graphic is the **histogram**.

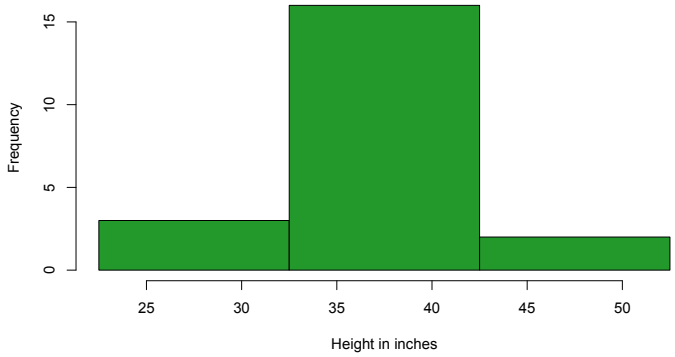
- Basically the same as a strip chart, but with bars instead of stacks of dots
- Back to bins, but can be any range (not just by digit)
- Like a bar chart, but with touching bars, to indicate the underyling numeric scale

Example: Heights of Four-Year-Olds in Inches

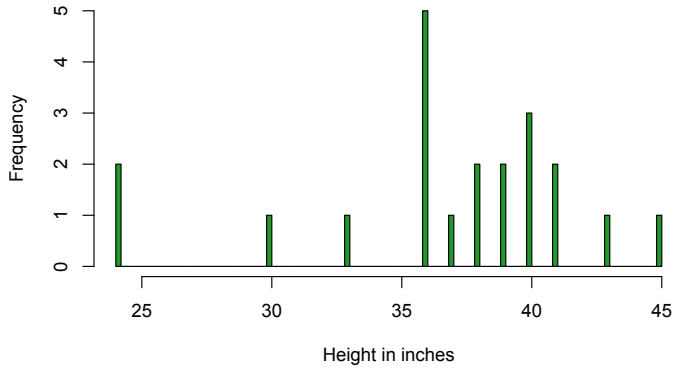
38 24 40 36 36 41 38
24 40 41 45 37 36 36
39 40 36 43 33 39 30



Notice that different bins can give very different impressions:



Here, the bins are smaller than the precision of our data:



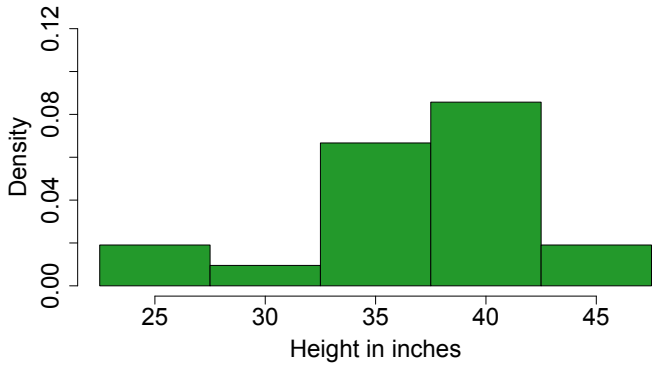
Some rules and rules of thumb for histograms:

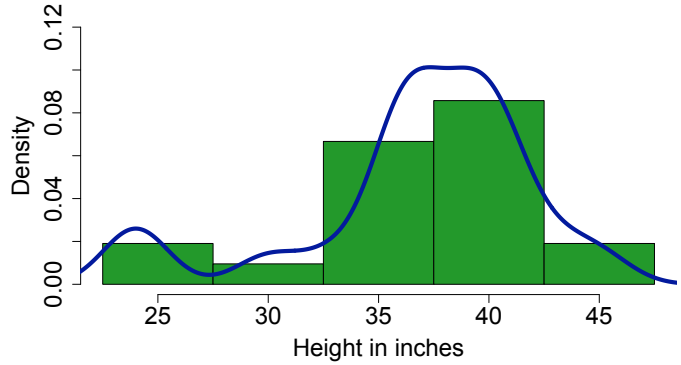
- Always select equal-width bins.
- Bin width should be no smaller than data precision.
- Convention: for discrete data, data on a boundary goes to the left (why is this not an issue for continuous data?)
- Guiding principle 1: Use wide enough bins to avoid "gaps", unless there's a good reason to think a gap is meaningful.
- Guiding principle 2: Use narrow enough bins that there's not much observable "structure" within bins (data is pretty evenly spread out within bins).

When would you prefer a histogram vs. one of the other graphics?

- Large data sets
- Continuous variables

- With a continuous variable, the "edges" between bins are artificial.
- If we kept collecting data, would expect the histogram to smooth out.
- Can capture what might happen with more data using a **density estimate**: smooth curve showing the shape of the data distribution.

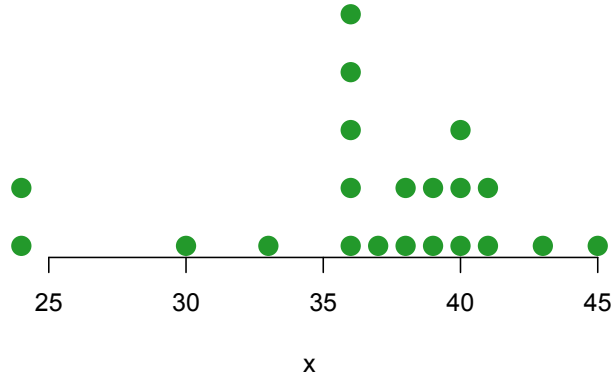




Notice the use of densities on the y-axis. Extrapolation to "infinite" data.

- Graphics are subjectively informative, but often we want to summarize data with a single number
- Usually representing a "typical", or "middle" value.
- But how do we define "typical"?
- Depends on the data and the question.

Where is the "center" / what is a "typical" value?



Some different intuitions:

- Most common value (the **mode**)
- Value separating the data into halves (the **median**)
- Value at the "balance point" (the **mean**)
- Halfway between highest and lowest values (the **midrange**)

- Intuitively, the mean is the "balance point" of the data.
- Computationally, it's the usual average:

$$\bar{x} = (\sum_{i=1}^{n} x_i)/n \qquad (1)$$

  - x_i is the i^{th} observation
  - n is the **sample size** (number of observations)

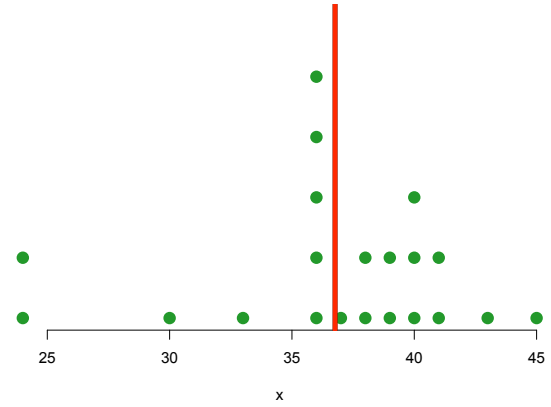


Figure: Height of 4-year-old Children in Inches

$\bar{x} = 36.76$

> "If an individual at any given epoch in society possessed all the qualities of the average man, he would represent all that is great, good, or beautiful."
>
> — Adolphe Quetelet, 19th Century French Statistician

- The ideal individual is 67.3 inches tall, comes from a family of 2.7 children, and is 0.49 male.
- Mean still defined, even for discrete variables, but does not represent a possible value for an individual.

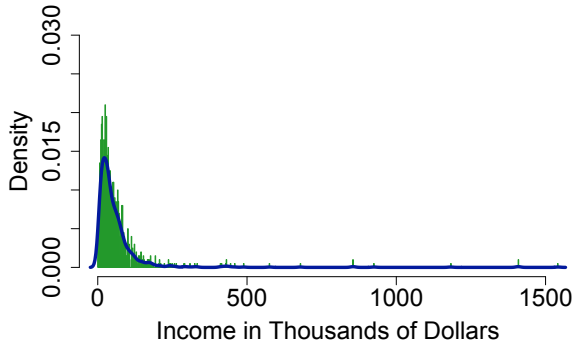


Figure: Annual Incomes of U.S. consumers in 2001

Where is the mean?

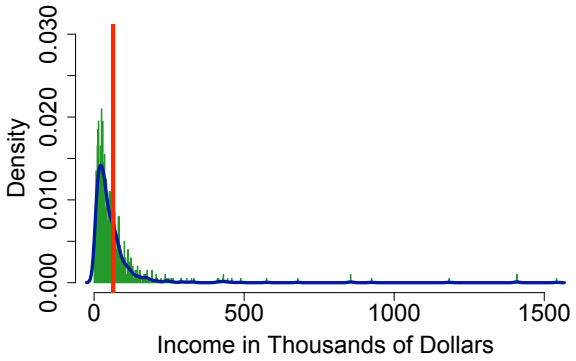


Figure: Annual Incomes of U.S. consumers in 2001

$$\bar{x} = \$63,400$$

- The mean is heavily influenced by extreme values
- For strongly asymmetric distributions, it will be pulled far from the "center" of the data.
- In cases like these, may be better to rely on a more **robust** (insensitive to extreme values) measure, such as the **median**.
- Intuitively, the median is _____

- To define the median, introduce some notation.
- Saw before, we use x_i to denote the i^{th} observation
- This is in the order that the data is collected.
- With parentheses around the index, denotes the i^{th} *smallest value* in the data set. Called the i^{th} **order statistic**.

$$\begin{aligned} x_{(1)} &= \text{minimum value} \\ x_{(2)} &= \text{next lowest (may be same)} \\ &\dots \\ x_{(n)} &= \text{maximum value} \end{aligned}$$

The Median

The **median** is written Q_2, and defined as:

$$Q_2 = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ Mean(\{x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}\}) & \text{if } n \text{ is even} \end{cases}$$

- What if $n = 1$?
- If $n = 2$?
- If $n = 400$?

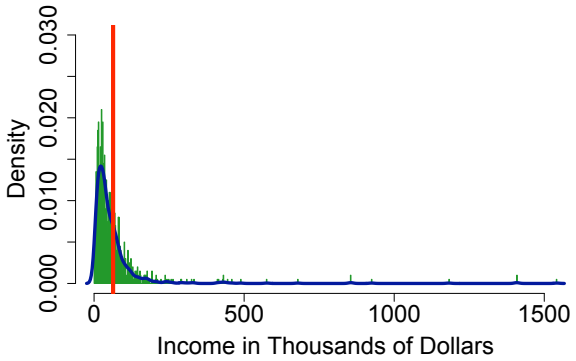


Figure: Annual Incomes of U.S. consumers in 2001

$\bar{x} = \$63\,400$

ISTA 116: Statistical Foundations for the Information Age
Measures of Central Tendency
The Median

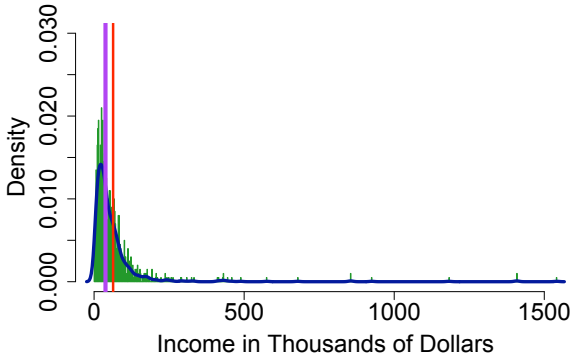


Figure: Annual Incomes of U.S. consumers in 2001

$\bar{x} = \$63\,400$ $Q_2 = \$38\,000$

ISTA 116: Statistical Foundations for the Information Age
Measures of Central Tendency
The Median

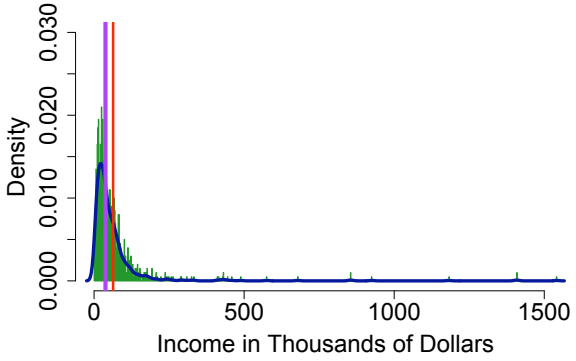


Figure: Annual Incomes of U.S. consumers in 2001

The mean is actually higher than 70.5% of the data!

ISTA 116: Statistical Foundations for the Information Age
Measures of Central Tendency
The Median

- The median is more representative of a "typical" observation when the data is asymmetric
- Whereas $\bar{x}$ is heavily influenced by extreme values, can change data around within two halves in any way without affecting Q_2

- Other measures are less commonly used:
  - The **mode** is the most frequent value.
    - What kinds of variables does this make sense for?
    - Can you think of a way to generalize to continuous variables?
  - The **midrange** is half way between the smallest and largest values:

$$\text{Midrange} = \frac{x_{(1)} + x_{(n)}}{2}$$

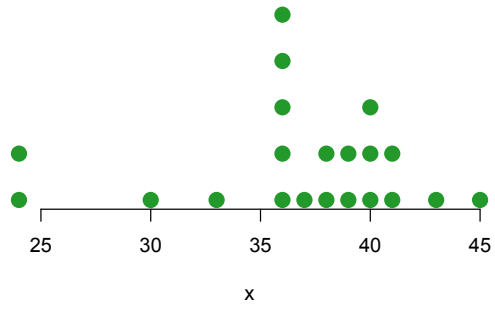


Figure: Heights of 4-year-old Children

Where are the mode and midrange?

- The mode may be nowhere near the middle of the data.
- If the mean is too influenced by extreme values, the midrange is *only* influenced by extreme values.

- Next time: Characterizing variability in data
- Reminder: Lab 1 due in d2l dropbox by Friday, 5 P.M.
- See you next Wednesday!