ISTA 116: Statistical Foundations for the Information Age

Univariate Numeric Data

29 August 2011

Outline

1 Reminders/Announcements

2 Categorical Data Warmup

3 Univariate Numeric Data
- Discrete vs. Continuous Numeric Variables

4 Visualizing Numeric Data
- Stem and Leaf Plots
- Strip Charts
- Histograms

- Lab Assignment 1 Due Friday via d2l dropbox (unless otherwise specified by your lab instructor)
- Wednesday's lecture and lab video available as podcast at `http://itunes.arizona.edu`, or streaming (follow d2l link)
- Some changes to office hours (see updated syllabus on d2l)

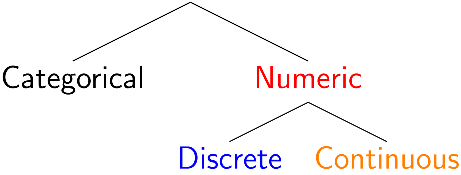
What's your favorite hot beverage?

Coffee	Tea	Mate	Cocoa	Other	None

Meet your neighbor, and draw:

- Relative Frequency Table
- Bar Plot
- Dot Chart
- Pie Chart

Types of Data



Categorical

Numeric

Discrete

Continuous

- Difference betwen discrete and continuous numeric variables?
  - Intuitively, for **discrete** variables, you have consecutive values with nothing in between *possible* (usually, whole numbers)
  - For **continuous** variables, there's always another value possible between any two, no matter how close.
- Math majors: yes, it's a bit more complicated than this (discrete = countable, continuous = uncountable), but above will do for our purposes
- Examples of each?

- Unlike categorical data, can do a lot more than count frequencies
- A numeric scale has a natural spatial arrangement (think “number line”)
- Take advantage of this intuitive notion to visualize the data

- The most detailed graphic is the **stem and leaf plot**
  - Basic idea: group data into “bins”, and “stack” the digits in each bin.
  - Usually, the **stem** is all but the last digit, and the **leaf** is the last digit (though sometimes it makes sense to use larger or smaller “bins”).

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

Is this discrete or continuous? (Careful!)

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

What do you think?

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

Same idea as stem-and-leaf plots, with three differences:

- Instead of stems, create an x-axis
- Rather than bins, identical items stacked vertically
- Display dots rather than digits

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

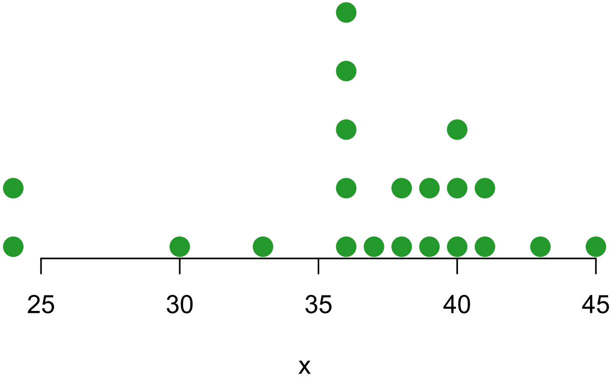


Figure: Heights of two-year-old Children in Inches

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

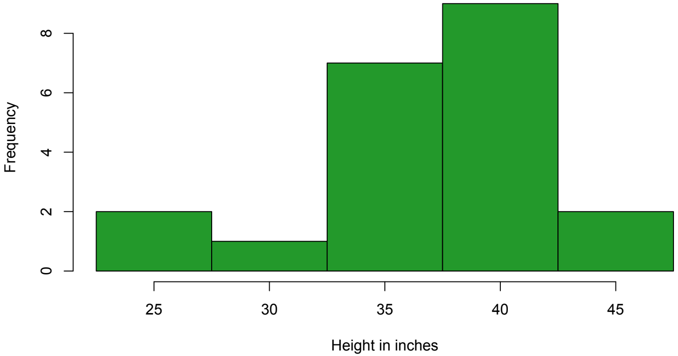
- Both stem-and-leaf plots and strip charts can be cluttered if there’s lots of data

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

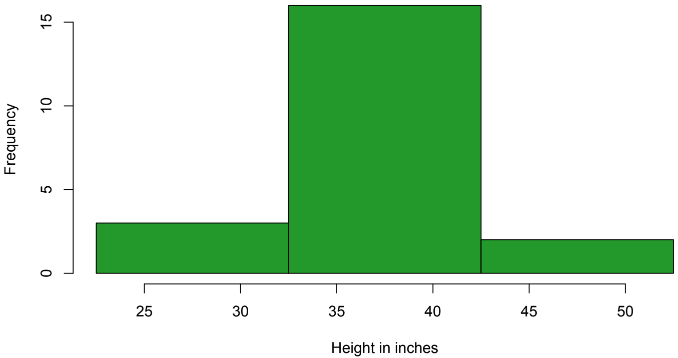
- Basically the same as a strip chart, but with bars instead of stacks of dots
- Back to bins, but can be any range (not just by digit)
- Like a bar chart, but with touching bars, to indicate the underyling numeric scale

Example: Heights of Four-Year-Olds in Inches

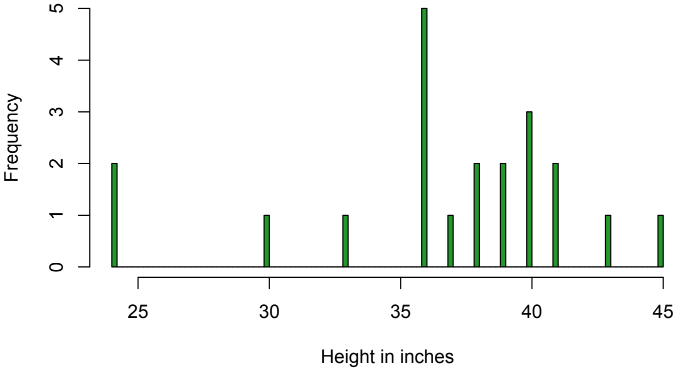
38	24	40	36	36	41	38
24	40	41	45	37	36	36
39	40	36	43	33	39	30


Frequency
0
2
4
6
8
25
30
35
40
45
Height in inches

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

- Next time: Measures of the “Center” of a Data Set
- Reminder: Lab Assignment 1 due in d2l dropbox by Friday by 5 P.M.