

Introduction to Statistics I

Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition, by Rossman and Chance).

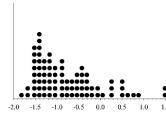
Previous Lecture

- ◆ Proportion
- ◆ Bar Charts & Distributions (distr)

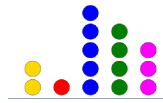
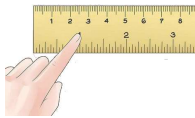


How many states have you visited?

§2.3: Graphical Summaries for Quantitative Data

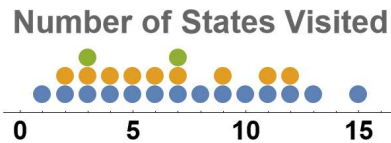


Dot plots



Dot plots visualize quantitative vars.

They're useful w/small amounts of data having repeated values.



All dot plots should contain:

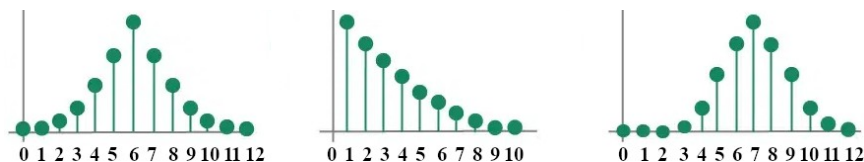
- ▶ A title
- ▶ Labeled axis showing the possible values of the var
- ▶ One dot per obv-unit at its value

Distr of Quantitative Vars

You'll be asked to describe a dot plot's pattern of variation.

- ◆ What values does the var take on? (min/max)
- ◆ Most frequent values?
- ◆ Approximate center?
- ◆ What trends or patterns do you notice?

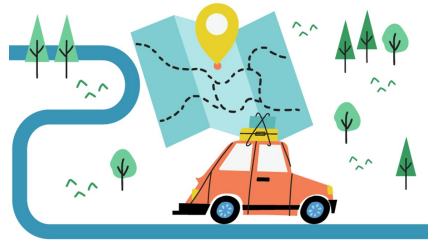
That is, try to describe the key aspects of a dot plot to someone who isn't looking at it.



How would you describe these?

Example: How many states have you been to?

What kind of var is this?



bit.ly/introstatsdata

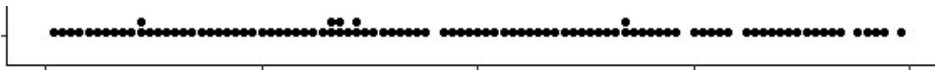
Polls: States Visited

Activity: 2.3a

Guessing Game

Is the dot plot below:

- ◆ Point values of letters in the game Scrabble?
- ◆ Blood pressure measurements of healthy adults?
- ◆ Jersey numbers of Cal Poly State University football players in 2006?



A ₁	B ₃	C ₃	D ₂		
E ₁	F ₄	G ₂	H ₄	I ₁	J ₈
K ₅	L ₁	M ₃	N ₁	O ₁	P ₃
Q ₁₀	R ₁	S ₁	T ₁	U ₁	V ₄
W ₄	X ₈	Y ₄	Z ₁₀		

Scrabble

Is the dot plot below:

- ◆ Annual inches of snowfall for some cities around the US?
- ◆ Prices of properties on the Monopoly game board?
- ◆ Blood pressure measurements of healthy adults?



Name of the Property	Cost
Mediterranean Ave.	60
Baltic Ave.	60
Oriental Ave.	100
Vermont Ave.	100
Connecticut Ave.	120
St. Charles Place	140
States Ave.	140
Virginia Ave.	160
St. James Place	180
Tennessee Ave.	180
New York Ave.	200
Kentucky Ave.	220
Indiana Ave.	220
Illinois Ave.	240
Atlantic Ave.	260
Ventnor Ave.	260
Marvin Gardens	280
Pacific Ave.	300
North Carolina Ave.	300
Pennsylvania Ave.	320
Park Place	350
Boardwalk	400

Monopoly

Is the dot plot below:

- ◆ Quiz percentages for a class of stats students (where quizzes went well for most students)?
- ◆ Prices of properties on the Monopoly game board?
- ◆ Weights (lbs) of rowers on the 2008 US Men's Olympic team?



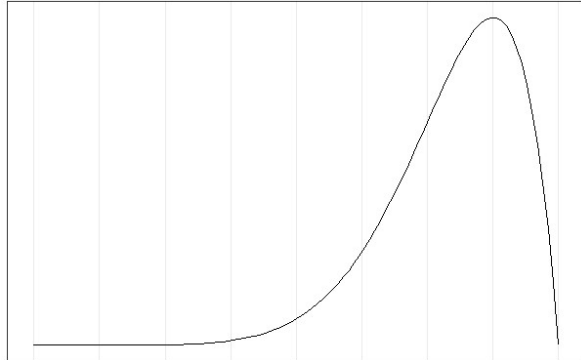
Describing Distributions

Without numbers, we're still able to figure it out. The distr contains useful info about a var.

For a quantitative var's distr, we focus on the **three "S"s**:

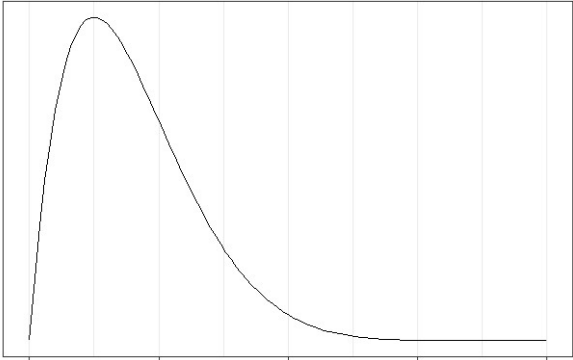
- ◆ **Shape** of distr (today's focus)
- ◆ **Center** of distr
- ◆ **Spread** of distr

Shape



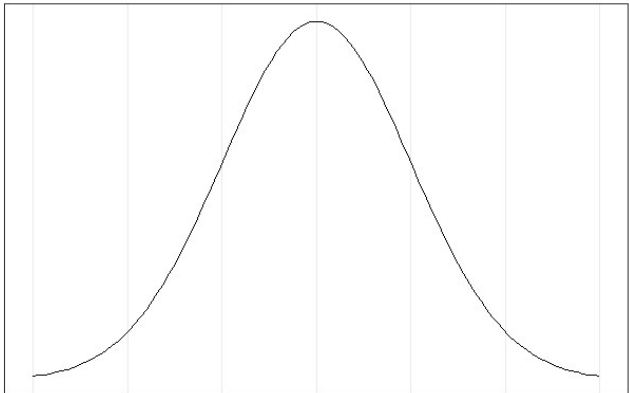
Skewed Left

e.g. retirement ages



Skewed Right

e.g. income distribution

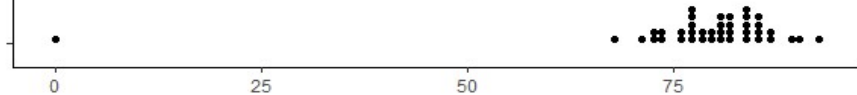


Symmetric, Mound Shaped

e.g., height

Outliers: data pts that are significantly above or below other values of the var.

When describing a distr, mention the outliers, then report the shape of the distr, excluding the outliers.



An outlier at zero, but otherwise mound shaped

Other Ways to Graph Quantitative Data

1, 5, 12, 7, 22, 4, 5, 27, 3, 13, 19

Stem	Leaf
0	1 3 4 5 5 7
1	2 3 9
2	2 7

Stem-and-Leaf Plot

It quickly graphs small data sets
 Most effective to have $\approx 4 - 15$ stems.

! In a stem-and-leaf plot, the right-most digit in your data is often the leaf, and the other digits are the stem.

Example: Age at Which Poets Wrote Their Famous Poems

Poet	Poem	Age
Eliot	<i>Prufrock</i>	23
Lowell	<i>Silent Hour</i>	41
Frost	<i>Stopping by the Woods on a Snowy Evening</i>	48
Williams	<i>Red Wheelbarrow</i>	40
Bishop	<i>The Fish</i>	29
Pound	<i>The River Merchant's Wife</i>	30
Plath	<i>Daddy</i>	30
Pound	<i>In a Station of the Metro</i>	28
Frost	<i>Mending Wall</i>	38
Stevens	<i>The Snow Man</i>	42
Williams	<i>The Dance</i>	59



Tens Place - Stem

Age
23
41
48
40
29
30
30
28
38
42
59

2
3
4
5

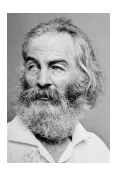
Ones Place - Leaf

Arranged in numerical order.

```

2 | 3 8 9
3 | 0 0 8
4 | 0 1 2 8
5 | 9
    
```

- ◆ How many poems were written by poets who were in their 20s?
- ◆ Are great poems typically written by younger poets (younger than 30)?

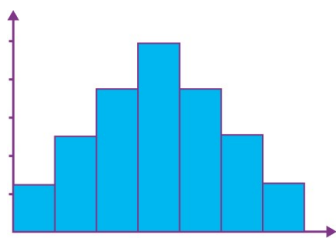


W. Whitman

❗ Stem doesn't have to be the tens digit. When measuring incomes, stem could be 10,000s digit, and leaf could be 1,000s digit.

What if you have a lot of quantitative data?

Histograms



Best for large data sets.

Example: Age at which patients receive their first diabetes diagnosis (2004 survey).

Sample size: 548 subjects (this would be difficult to dot plot!)

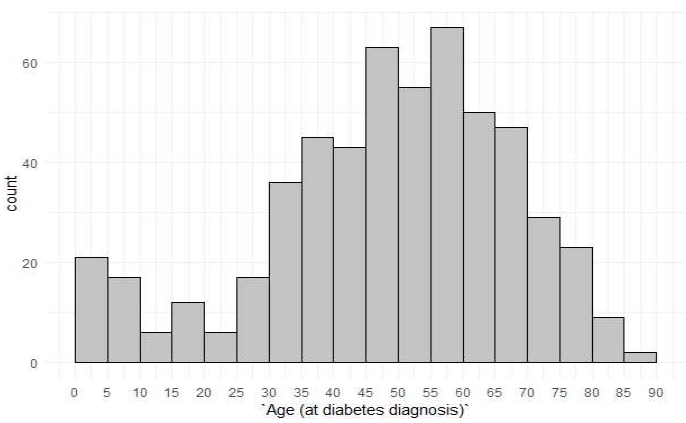


Group the data using **bins**, created by breaking the range of the data (ages from 0-90) into equal sized sub-intervals.

(you decide how big the bins are)

Then, count how many fall into each bin (called **frequencies**).

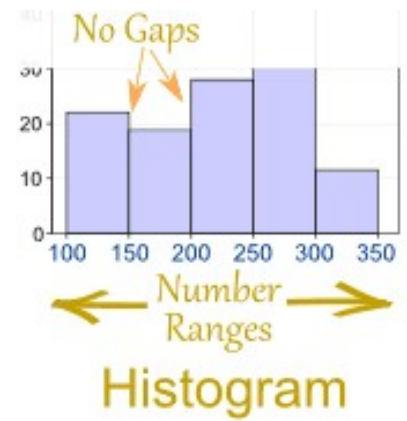
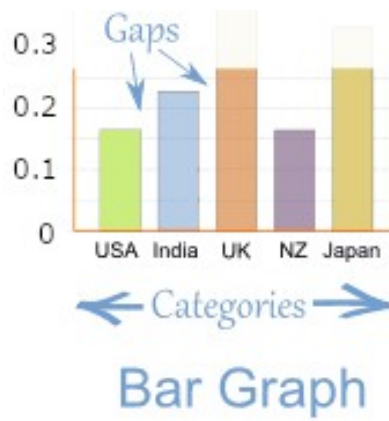
Bins	Frequency
0 – 4	21
5 – 9	17
10 – 14	6
15 – 19	12
20 – 24	6
etc,



Diabetes Age Histogram

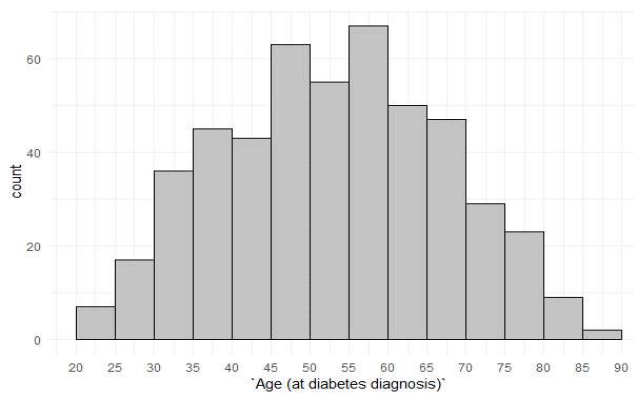
Histograms are different than bar graphs:

- ◆ Histograms are used only for quantitative data.
- ◆ Bars in a histogram touch each other (why?).
- ◆ Counts, not proportions, are on vertical-axis.



Other Histogram Characteristics

- ◆ Easy to see patterns in the shape of a hist distr.
- ◆ Unlike stem-and-leaf, hists destroy data (data loss).



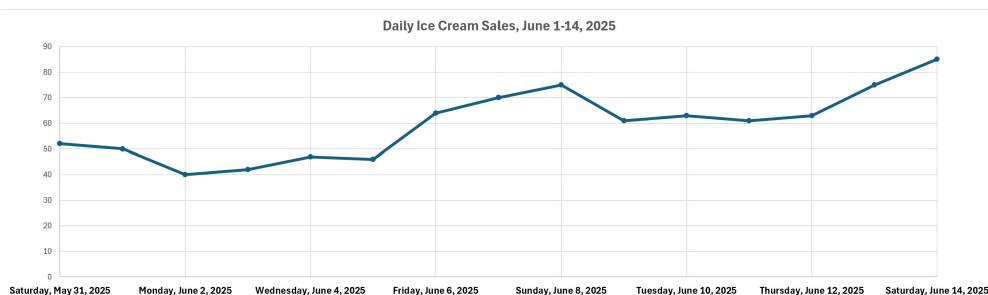
Diabetes Histogram (20 or Older)

Time Series Plot

A seaside ice cream shop tracks the number of cones sold each day from June 1st thru 14th, 2025. Here's the data:

Date	Cones Sold	Date	Cones Sold
Sat, 6/1	52	Sat, 6/8	70
Sun, 6/2	50	Sun, 6/9	75
Mon, 6/3	40	Mon, 6/10	61
Tue, 6/4	42	Tue, 6/11	63
Wed, 6/5	47	Wed, 6/12	61
Thu, 6/6	46	Thu, 6/13	63
Fri, 6/7	64	Fri, 6/14	75

Plot these in Excel. What are the trends?



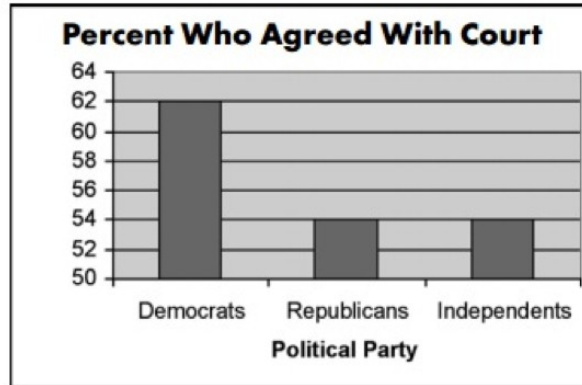
The owner notices sales tend to climb as the weather warms, and also as more tourists arrive on weekends.

Graphs Can Be Misleading

Positioning the Vertical Scale

Common way of generating a misleading graph is to set the vertical baseline at some number other than zero.

Example: when asked whether people agreed with a particular court decision, the following graph was generated:

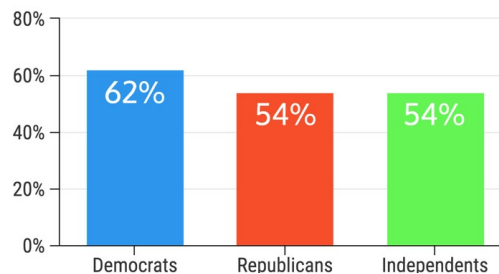


Observe the baseline number on the graph (the lowest number on the vertical axis) is 50, not zero.

If you weren't paying close attention, you might think Democrats agreed 3 times as much as the Republicans or Independents.

When the graph is redone without this misleading choice, you get the following graph.

Percent Who Agreed With Court



This more accurately portrays the actual difference between Democrats and the other respondents.

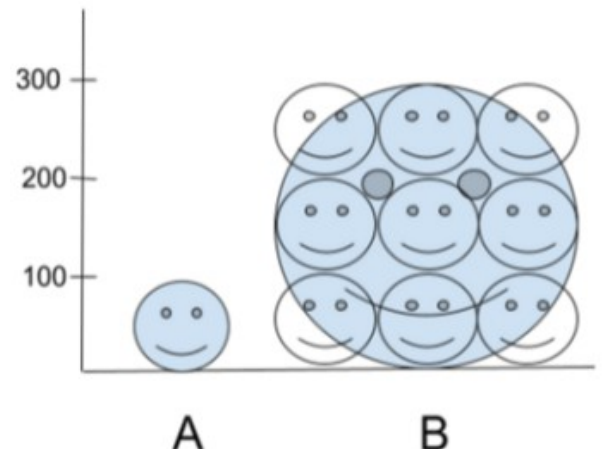
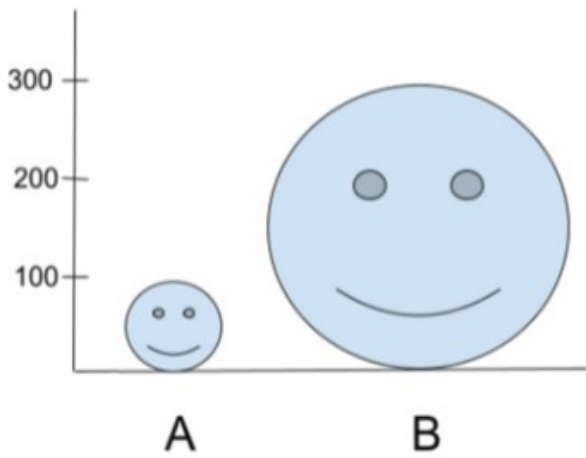
Area Principle

If you increase the length of a 2D object in your graph which you are using to indicate some quantity, you must be careful.

For instance, in the graph below, the suggestion is that B is three times as happy as A.

But, when you make the height of one object 3 times the height of another, you cause the area to increase by 9 times!

The 2nd image gives some indication of this.



So this gives the unconscious impression that B is 9 times as large, instead of only 3 times as large.

Activities: 2.3b



2.3c

Applets: Histogram Bin Width
 (then choose Histogram under the graph)
bit.ly/introstatsdata



Data: Diabetes
bit.ly/introstatsdata

Bin width (in yrs) and # of Bins

- ◆ The given graph in this activity has: 18 bins, with width of 5 yrs.
- ◆ Unfortunately, the app default is: 20 bins, width 4.5 yrs.
- ◆ "Bin width" of 20 yrs is not ideal (why?). In the applet, instead try 2 bins, or a "bin width" of 45 yrs.
- ◆ Then try 90 bins, which is a "bin width" of 1 yr.

What did we learn?

- ◆ Dot Plots & Distr's
- ◆ Distributions: shape/center/spread, skew, symmetry, outliers
- ◆ Stem-and-Leaf Plots
- ◆ Histogram: bins/frequencies
- ◆ Time Series

