Section 2.2

One Quantitative Variable: Shape and Center

Hollywood Movies2011

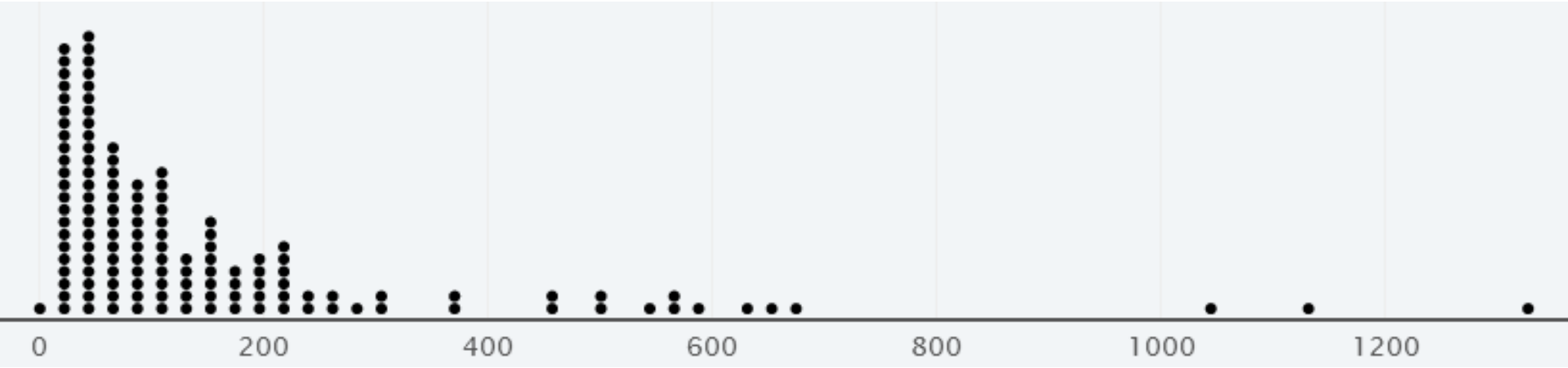
	Movie	LeadStudio	RottenTomatoes	AudienceScore	Story	Genre	TheatersOp
1	Insidious	Sony	67	65	Monster Force	Horror	2408
2	Paranormal Activity 3	Independent	68	58	Monster Force	Horror	3321
3	Bad Teacher	Independent	44	38	Comedy	Comedy	3049
4	Harry Potter and the Deathly Hallows Part 2	Warner Bros	96	92	Rivalry	Fantasy	4375
5	Bridesmaids	Relativity Media	90	77	Rivalry	Comedy	2918
6	Midnight in Paris	Sony	93	84	Love	Romance	944
7	The Help	DreamWorks Pictures	75	91	Maturation	Drama	2534
8	The Hangover Part II	Legendary Pictures	35	58	Comedy	Comedy	3615
9	Another Earth	Independent	63	74	Temptation	Fantasy	NA
10	Limitless	Virgin	69	73	Wretched Excess	Thriller	2756
11	Horrible Bosses	Warner Bros	69	72	Revenge	Comedy	3040
12	No Strings Attached	Spyglass Entertainment	49	57	Comedy	Comedy	3018
13	Twilight: Breaking Dawn	Independent	26	68	Love	Romance	4061
14	Transformers: Dark of the Moon	DreamWorks Pictures	35	67	Quest	Action	4088
15	Gnomeo and Juliet	Disney	56	52	Love	Animation	2994
16	Rio	20th Century Fox	71	73	Quest	Animation	3826
17	Super 8	Paramount	82	78	Monster Force	Horror	3379
18	Rise of the Planet of the Apes	20th Century Fox	83	87	Revenge	Action	3648
19	Apollo 18	Weinstein Company	23	31	Monster Force	Horror	3328
20	The Smurfs	Sony Pictures Animation	23	50	Fish Out Of Water	Animation	3395
21	Fast Five	Universal	78	83	Escape	Action	3644
22	Our Idiot Brother	The Weinstein Company	68	79	Comedy	Comedy	2555
23	50/50	Independent	93	93	Discovery	Comedy	2458
24	Drive	Independent	93	79	Rivalry	Thriller	2886
25	Beginners	Independent	84	80	Love	Comedy	NA
26	Kung Fu Panda 2	DreamWorks Animation	82	80	Rivalry	Animation	3925
27	Unknown	Independent	55	57	The Riddle	Thriller	3043
28	The Ides of March	Columbia	85	76	Transformation	Thriller	2199

One Quantitative Variable

- We need summary statistics and visualizations that show the **center**, **spread**, and **shape** of the quantitative data.

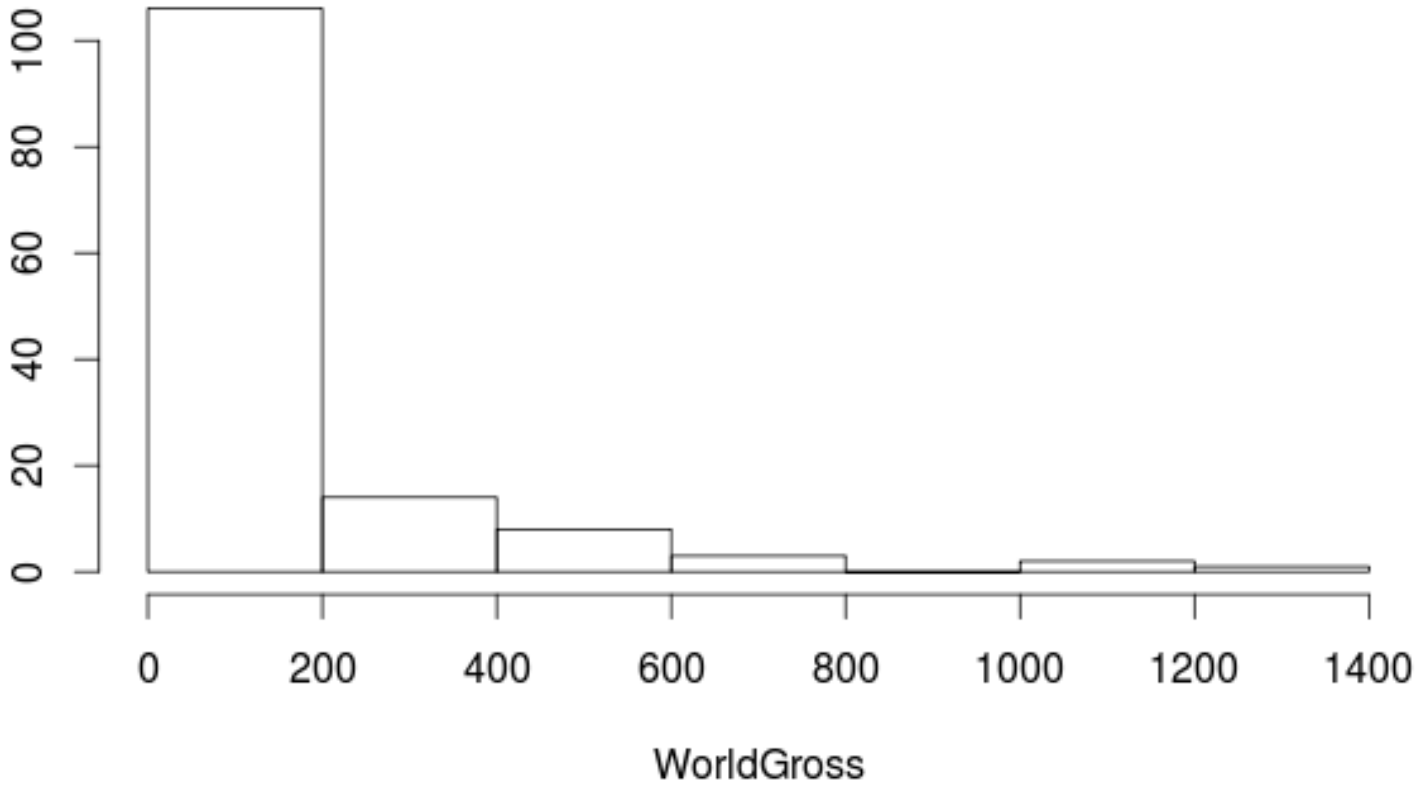
Dotplot

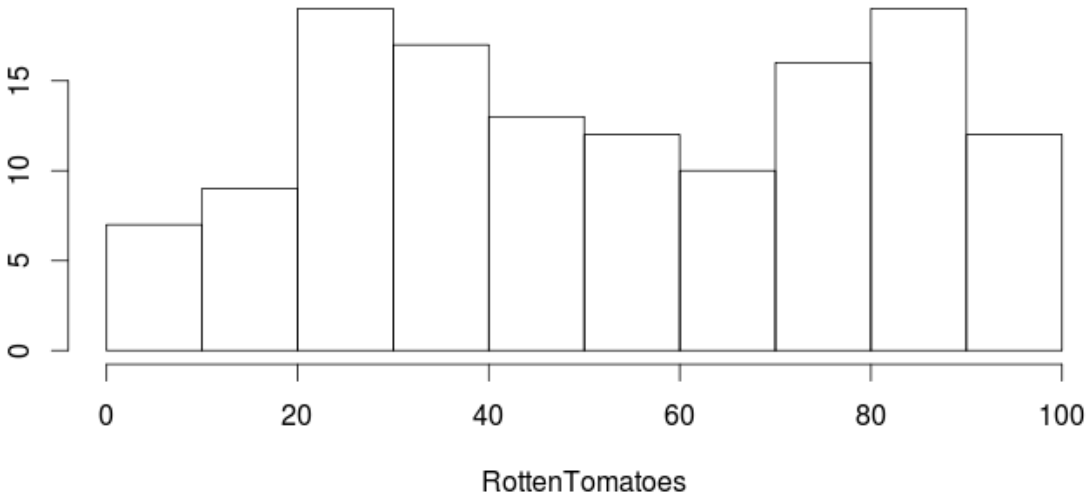
- In a **dotplot**, each case is represented by a dot and dots are stacked.
- Easy way to see each case



Histogram

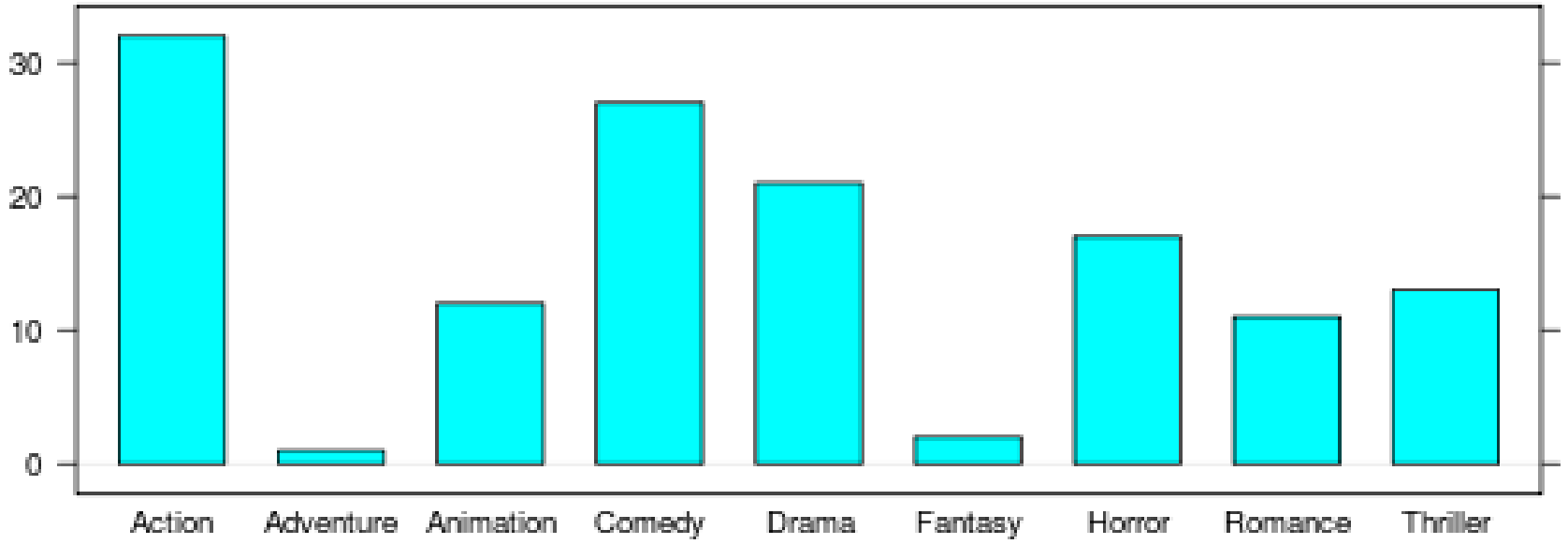
- The height of the each bar corresponds to the number of cases within that range of the variable





This is a

1. Histogram
2. Bar chart
3. Other
4. I have no idea



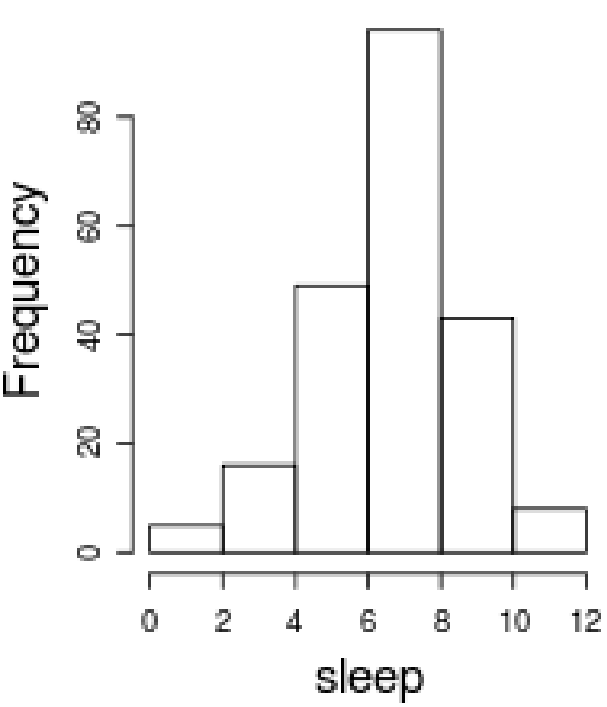
This is a

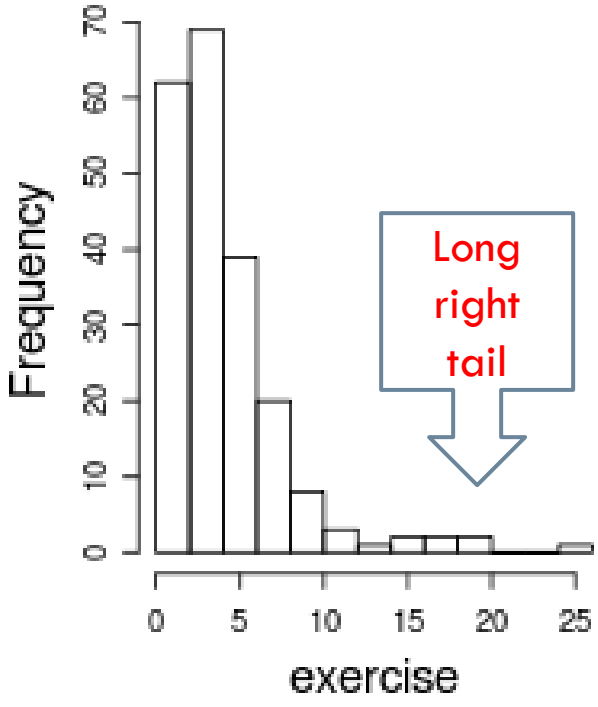
1. Histogram
2. Bar chart
3. Other
4. I have no idea

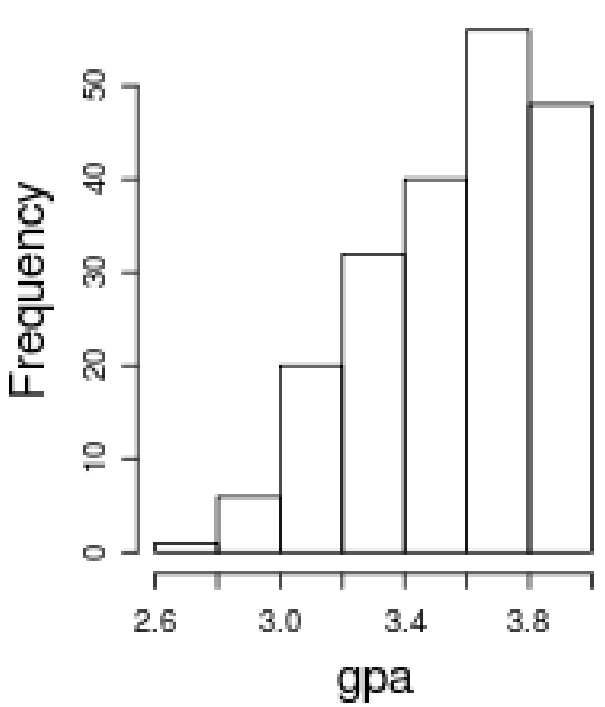
Histogram vs. Bar Chart

- A bar chart is for categorical data, and the x-axis has no numeric scale
- A histogram is for quantitative data, and the x-axis is numeric
- For a categorical variable, the number of bars equals the number of categories, and the number in each category is fixed
- For a quantitative variable, the number of bars in a histogram is up to you (or your software), and the appearance can differ with different number of bars

Shape







Symmetric

Right-Skewed

Left-Skewed

Demonstration: Measures of Center

- We have 5 stacks of blocks with heights 1, 3, 3, 6, and 7.
- We would like to describe the typical value (or center) of the height data. Describe multiple methods that we could use to do so.

Notation

- The **sample size**, the number of cases in the sample, is denoted by ***n***
- We often let *x* or *y* stand for any variable, and

$x_1, x_2, \ldots, x_n$ represent the *n* values of the variable *x*

- Ex) $x_1 = 97.009$, $x_2 = 201.897$, $x_3 = 216.196$, …

Movie	WorldGross
Insidious	97.009
Paranormal Activity 3	201.897
Bad Teacher	216.196
Harry Potter and the Deathly Hallows Part 2	1328.111
Bridesmaids	288.382
Midnight in Paris	139.177
The Help	199.324
The Hangover Part II	581.464
Another Earth	1.321

Mean

The **mean** or **average** of the data values is

$$\textbf{mean} = \frac{\textbf{sum of all data values}}{\textbf{number of data values}}$$

$$mean = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

- Sample mean: $\bar{x}$
- Population mean: μ ("mu")

Median

The **median**, ***m***, is the middle value when the data are ordered.

If there are an even number of values, the median is the average of the two middle values.

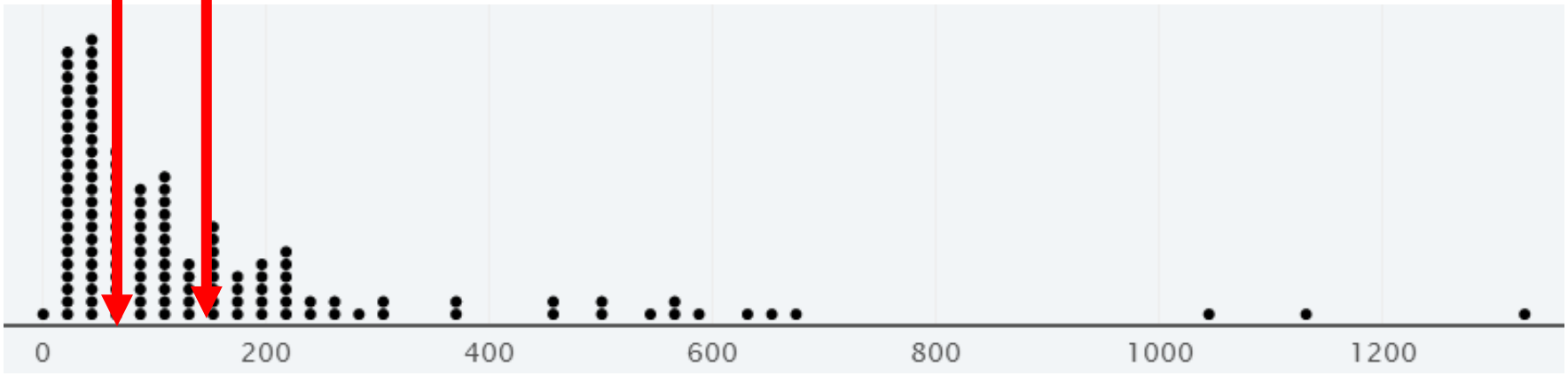
- The median splits the data in half.

Measures of Center

$m = 76.66$

$\mu = 150.74$

Mean is "pulled" in the direction of skewness

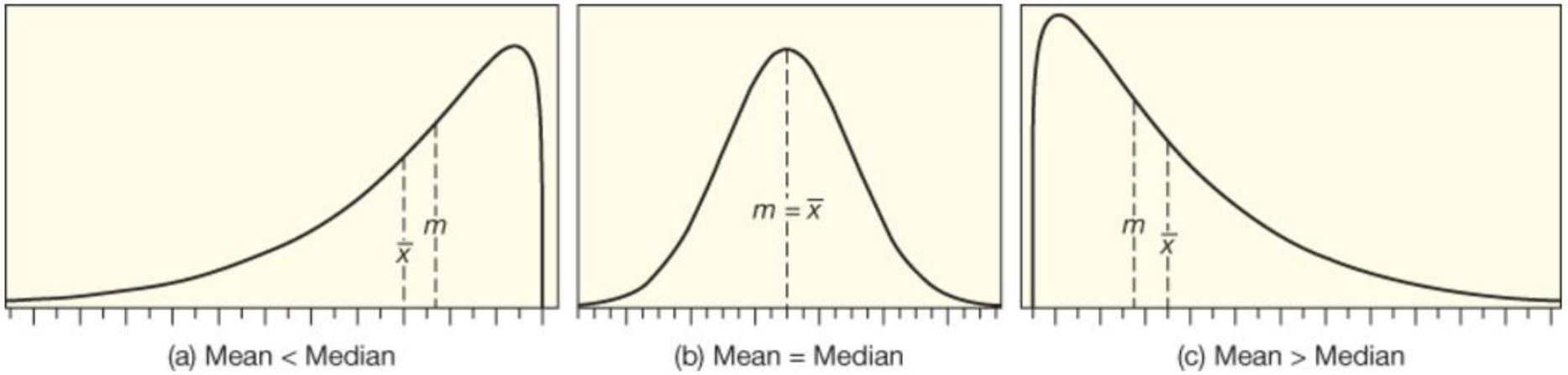


World Gross (in millions)

A distribution is left-skewed. Which measure of center would you expect to be higher?

1. Mean
2. Median

Skewness



(a) Mean < Median

(b) Mean = Median

(c) Mean > Median

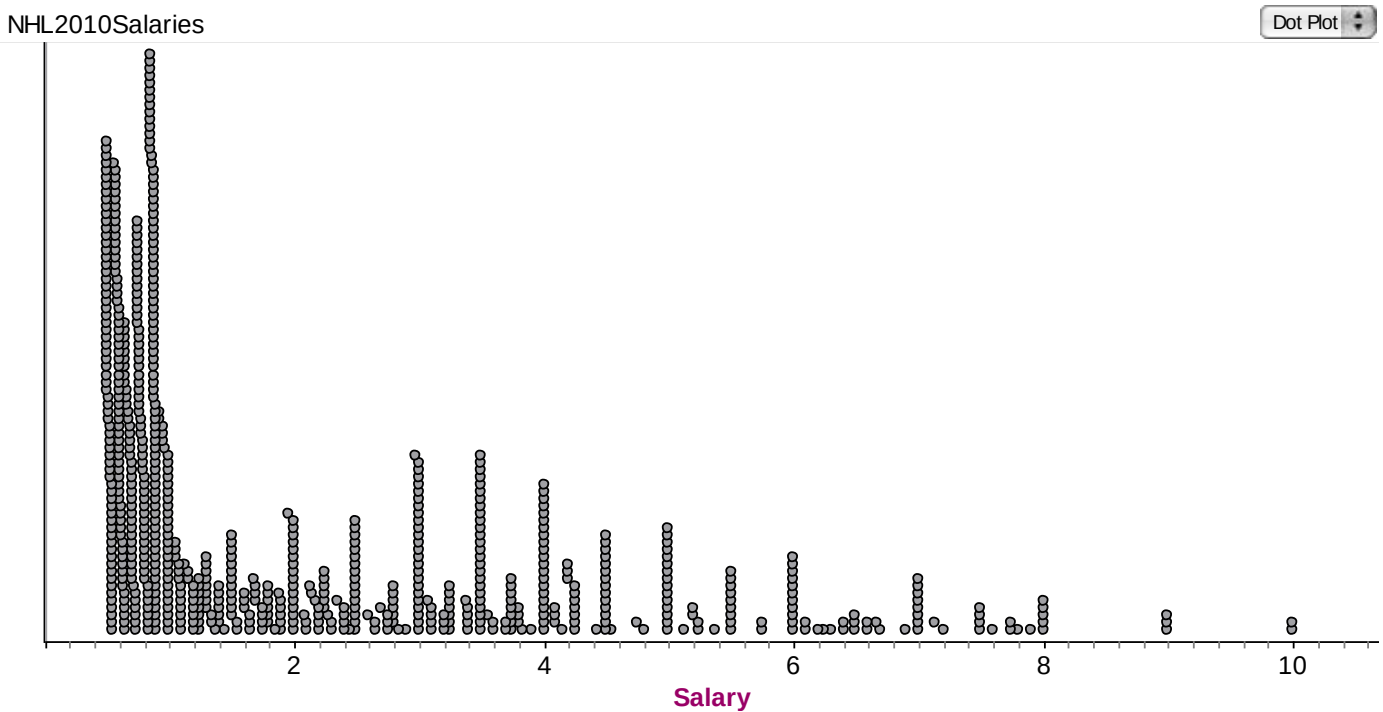
Skewed Left

Symmetric and Bell-Shaped

Skewed Right

The distribution of 2010-11 NHL Salaries is shown, in millions of dollars. The distribution is:

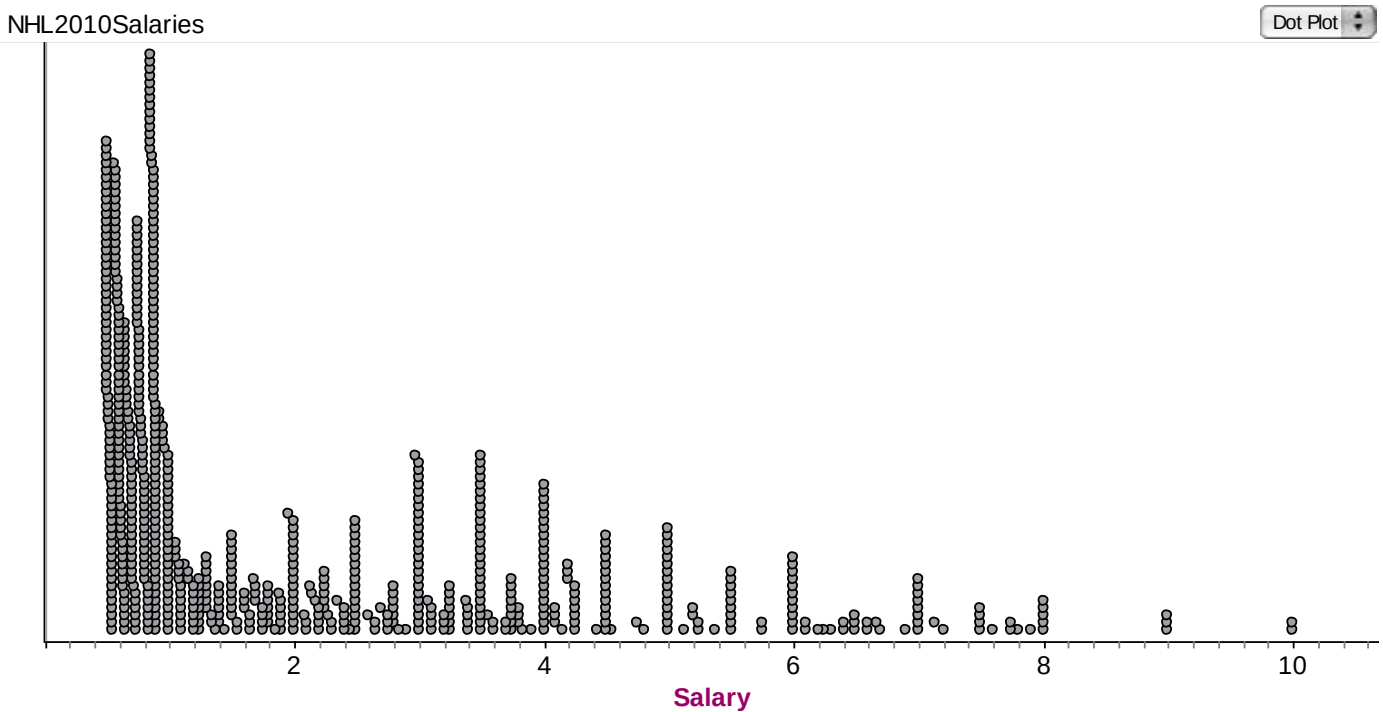
A. Symmetric
B. Skewed right
C. Skewed left



The distribution of 2010-11 NHL Salaries is shown. Which is larger, the mean or the median?

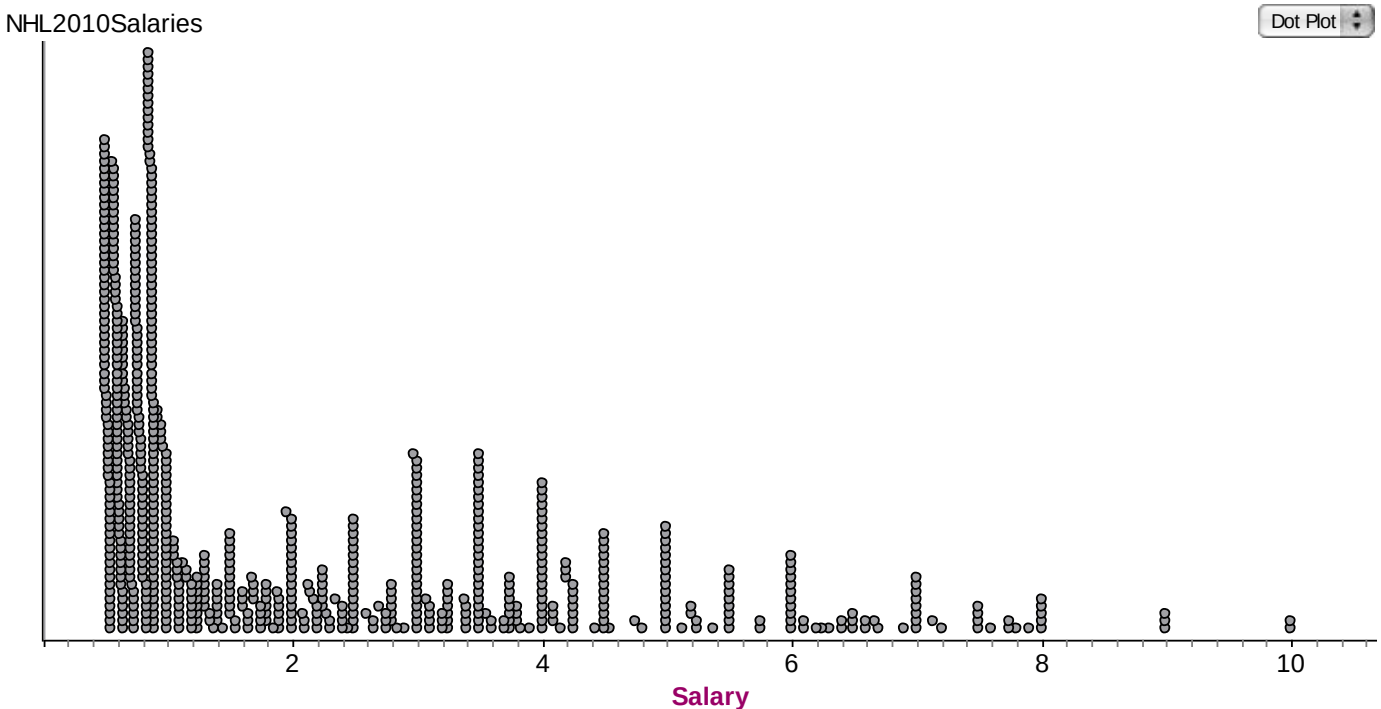
A. The mean

B. The median



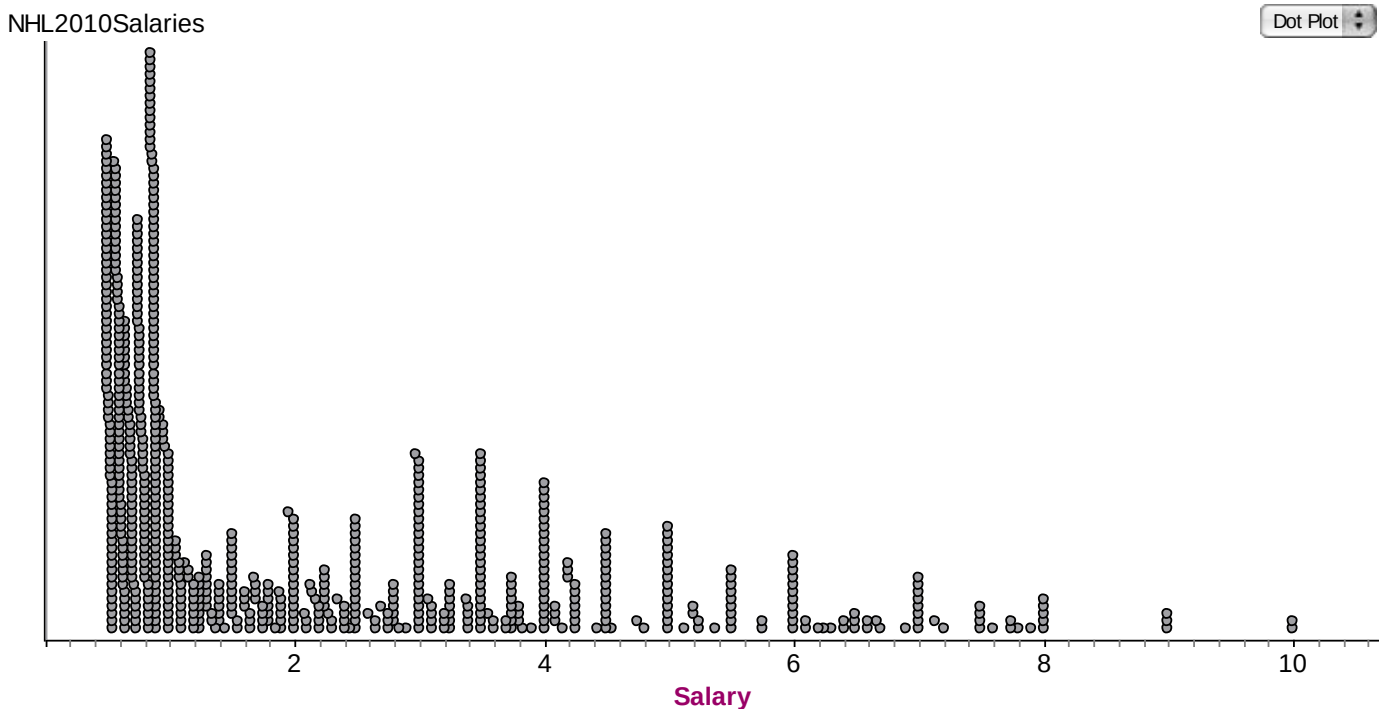
The mean is approximately (in millions of dollars)

A. 0.46
B. 1.25
C. 2.21
D. 4.35
E. 5.0



The median is approximately (in millions of dollars)

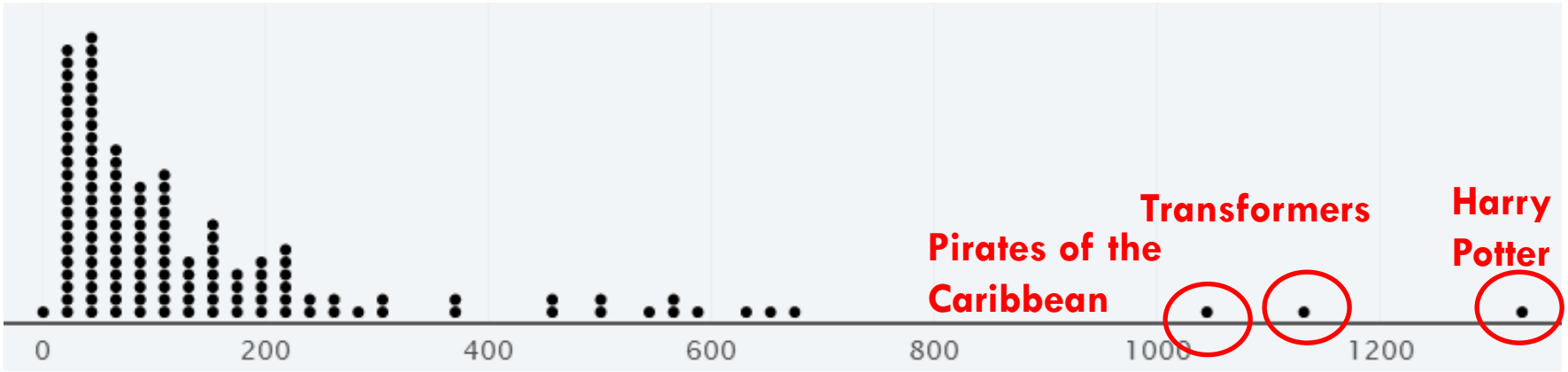
A. 0.46
B. 1.25
C. 2.21
D. 4.35
E. 5.0



Outlier

An **outlier** is an observed value that is notably distinct from the other values in a dataset. Values that are considered outliers will be *much larger* or *much smaller* than the rest of the data values.

Outliers



World Gross (in millions)

Resistance

	1			1
	2			2
	3			3
	7			7
	8			8000000
Mean	4.2		Mean	1600003
Median	3		Median	3

Resistance

A statistic is **resistant** if it is relatively unaffected by extreme values.

The **median** is resistant while the **mean** is not.

	Mean	Median
With Harry Potter	$150,742,300	$76,658,500
Without Harry Potter	$141,889,900	$75,009,000

Outliers

- When using statistics that are not resistant to outliers, stop and think about whether the outlier is a mistake
- If not, you have to decide whether the outlier is part of your population of interest or not
- Usually, for outliers that are not a mistake, it's best to run the analysis twice, once with the outlier(s) and once without, to see how much the outlier(s) are affecting the results

Example: Normal Body Temperature

- It is commonly believed that "normal" human body temperature is 98.6°F (or 37°C). In fact, "normal" temperature can vary from person to person, and for a given person it can vary over the course of a day.
- The table on the next slide gives a set of temperature readings of a healthy woman taken over a two-day period.
- (cont'd…)

Example: Normal Body Temperature

97.2	**97.6**	**98.4**	**98.5**	**98.3**	**97.7**
97.3	**97.7**	**98.5**	**98.5**	**98.4**	**97.9**

a) Make a dotplot of the data.
b) Compute the mean of the data and locate it on the dotplot as the balance point.
c) Compute the median of the data and locate it on the dotplot as the midway point.

http://lock5stat.com/statkey/

Summary

- Visualizing one quantitative variable:
  - Dotplot
  - Histogram
- Shape:
  - Symmetric
  - Skewed
- Measures of center:
  - Mean (not resistant to outliers)
  - Median (resistant to outliers)