Section 1.7 – Misrepresentations of Data

MDM4U
David Chen

The Media

--- The media are major users of data. In addressing issues and presenting points of view, the media rely on information based on data

--- One of the main purposes of the media is to inform the general public about world events in as an objective manner as possible

--- However, the media may sometimes provide misleading or false impressions to sway the public

--- An important reason to study statistics is to understand how information is represented or misrepresented

"I can prove it or disprove it! What do you want me to do?"

Part 1: Warm-up

Democrats say that they have won 60% of recent elections, however, Republicans say that they have won 62.5% of the most recent elections.

What is going on? Who do you think is lying?

Lets examine the real statistics.

2008 --- Obama --- Democrat
2004 --- Bush --- Republican
2000 --- Bush --- Republican
1996 --- Clinton --- Democrat
1992 --- Clinton --- Democrat
1988 --- Bush --- Republican
1984 --- Reagan --- Republican
1980 --- Reagan --- Republican

So, who was lying?

Neither

Democrats have won 3 of the last 5 elections = 60%
Republicans have won 5 of the last 8 elections = 62.5%

Deception of Sample Sizes is a common way data is misrepresented.

Other common ways data can be misrepresented:

1. **Data not displayed properly**
   a. Truncated y-axis
   b. Area principle violated
   c. Missing axis labels
   d. Improper scale

2. Sample size is too small

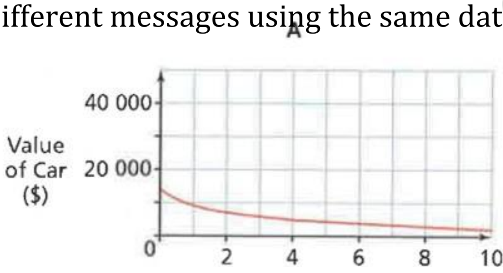
3. Insufficient information

4. Sample is not representative of the population

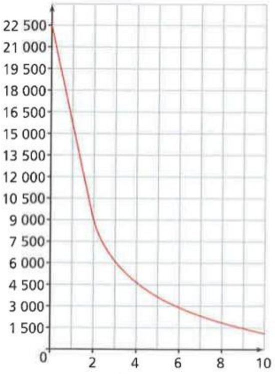
Part 2: Data not Displayed Properly

Example 1: When you purchase a new vehicle, its value drops dramatically the moment it is driven off the car dealer's lot, and then continues to drop each year thereafter. A graph is used to show this change in value over time. It is possible to communicate different messages using the same data by changing the vertical scale.

Graph A: This graph shows the car's value go from $9000 after 2 years to $1000 after 10 years.



Graph B: This graph also shows the car's value go from $9000 after 2 years to $1000 after 10 years.



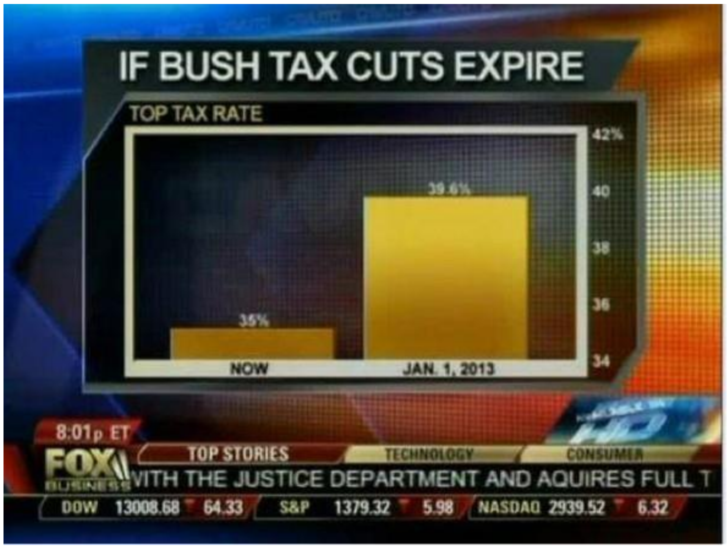
a) Look quickly at each graph. What impression does graph A give you about the change in value of the car compared to graph B?

The value of the car in graph A seems to be decreasing but at a much slower rate than the value of the car in graph B.

b) Once you look more careful at both graphs, how does your impression change? What information changed your first impression of the graphs?

The change in the value of the car is actually the same. However, your impression likely changed when you looked at the scale provided for the two graphs. Scales that go up by small differences exaggerate trends in the data.

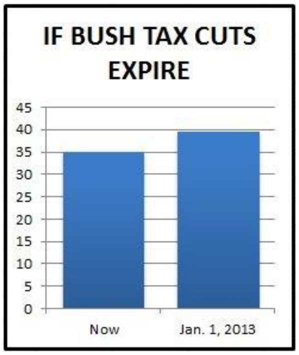
Example 2: How did FOX news misrepresent this data?



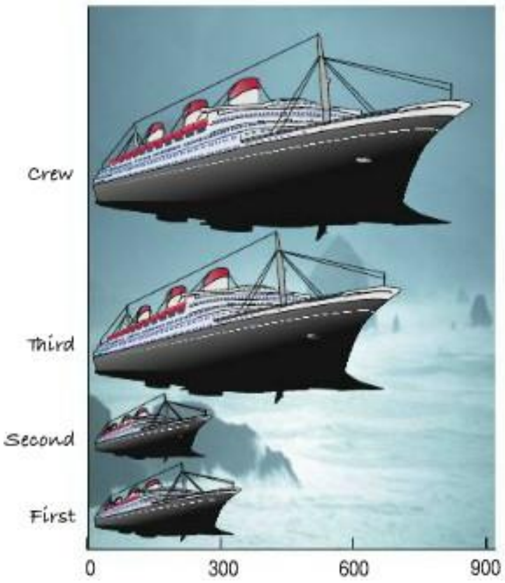
This is an example of a truncated y-axis.

Looks like the percentage changed a lot from "now" to Jan 1, 2013. But examining closely, you can see that ***the minimum point on the vertical axis is 34% instead of 0***. That's what made it misleading. Fox News exaggerated the percentage just to serve the purpose of pushing Bush's tax cut renewal. This is called "truncating the y-axis".

This is what the real percentage looks like:



Example 3: The following graph shows the number of people on board the Titanic for each class. How does this graph misrepresent the data?



Although the lengths of the ships are accurate, our eyes respond to the area of the pictures. There are about three times as many crew members on the ship as first class passengers but the picture of the ship for crew members has an area about 9 times larger than the first class ship.

The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents.

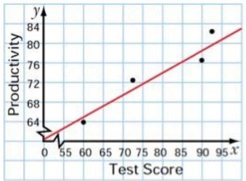
Part 3: Sample Size is Too Small

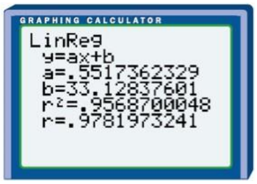
Example 4: A manager wants to know if a new aptitude test accurately predicts employee productivity. The manager has all 30 current employees write the test and then compares their scores to their productivities as measured in the most recent performance reviews. The data is ordered alphabetically by employee surname. In order to simplify the calculations, the manager selects a systematic sample using every seventh employee. Based on this sample, the manager concludes that the company should hire only applicants who do well on the aptitude test. Determine whether the manager's analysis is valid.

"After closer investigation, it's become clear that we need to enter more than one value."

Test Score	Productivity
98	78
57	81
82	83
76	44
65	62
72	89
91	85
87	71
81	76
39	71
50	66
75	90
71	48
89	80
82	83
95	72
56	72
71	90
68	74
77	51
59	65
83	47
75	91
66	77
48	63
61	58
78	55
70	73
68	75
64	69

Based on the linear regression of the systematics sample, what would you conclude?

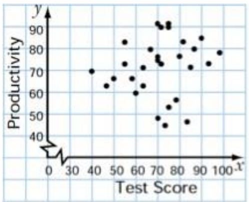


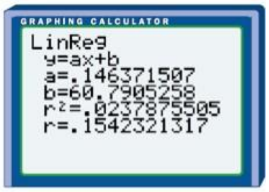


```
LinReg
 y=ax+b
 a=.5517362329
 b=33.12837601
 r²=.9568700048
 r=.9781973241
```

There is a strong positive linear correlation between test score and productivity. Therefore the aptitude test is a great indicator of employee productivity.

Based on the linear regression of the raw data, do you think the sample is a good representation of the population?





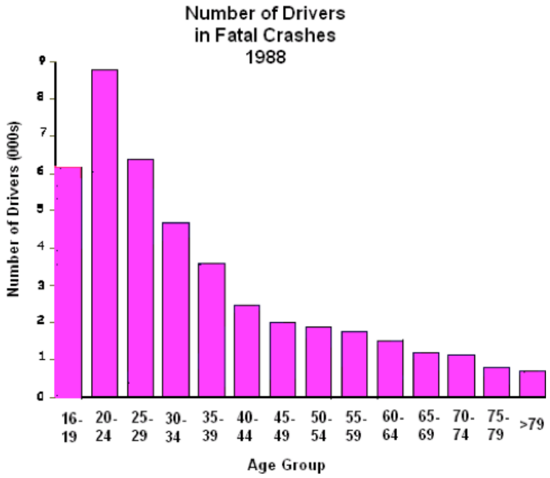
```
LinReg
 y=ax+b
 a=.146371507
 b=60.7905258
 r²=.0237875505
 r=.1542321317
```

No, there appears to be a very weak correlation between test scores and productivity. Therefore the aptitude test is not a good predictor of employee productivity.

Part 4: Insufficient Information

Example 5: What does this graph tell you about the ability of drivers as they age?

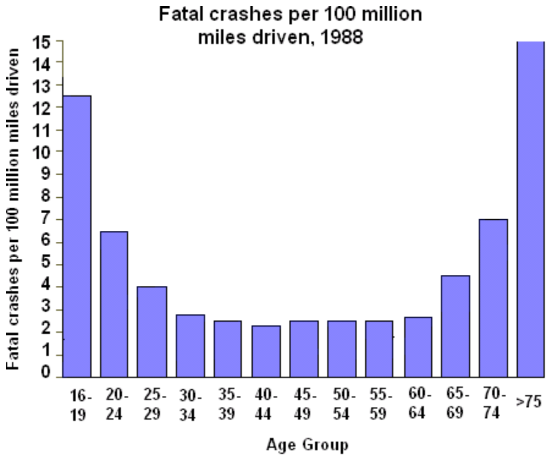
This graph indicates that drivers get better with age because they are involved in fewer fatal crashes.



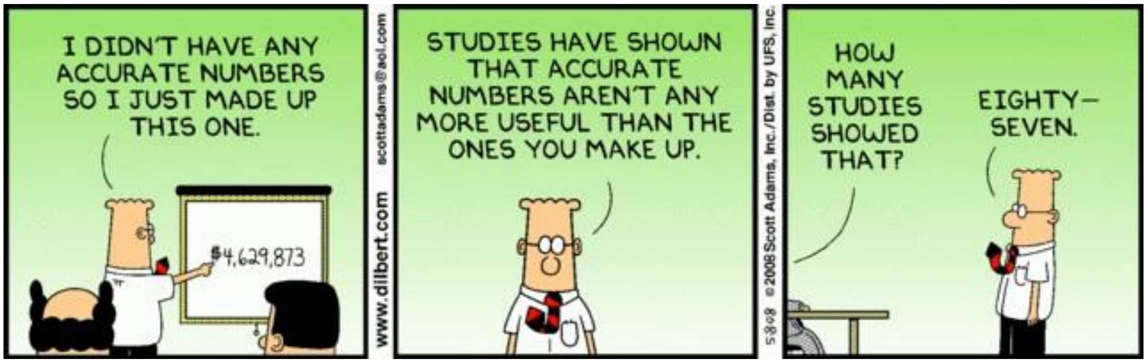
Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

What does this graph tell you about the ability of drivers as they age?

This graph indicates that drivers are at the highest risk of a fatal crash when they are >75 years old. The previous graph was misleading because it didn't take in to account how many miles each age group drives.

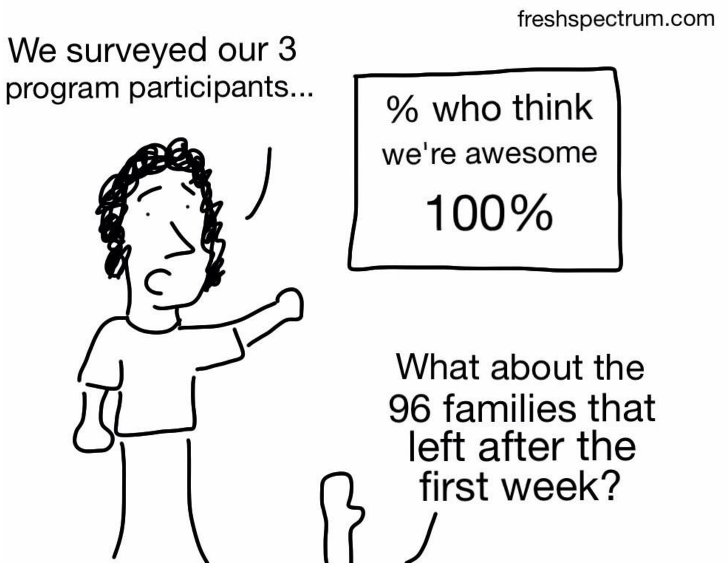


Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
scottadams@aol.com
www.dilbert.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
©2008 Scott Adams, Inc./Dist. by UFS, Inc.
5-8-08
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

Part 4: Sample is not Representative of the Population

When reading statistics, look carefully for an indication of how the sample was chosen. Often, companies will carefully select a sample so that they can inflate their results.


freshspectrum.com
We surveyed our 3 program participants...
% who think we're awesome
100%
What about the 96 families that left after the first week?