Math 263 Section 005: Statistics and Bio-Statistics, Spring 2014

Deborah Hughes Hallett
Gould Simpson 829, 621-6886
dhh@math.arizona.edu

Course Policies and Information at http://math.arizona.edu/~dhh/263-14.html. You will need a webassign account for the course. Go to www.webassign.net. The class key is **arizona 7989 8719**

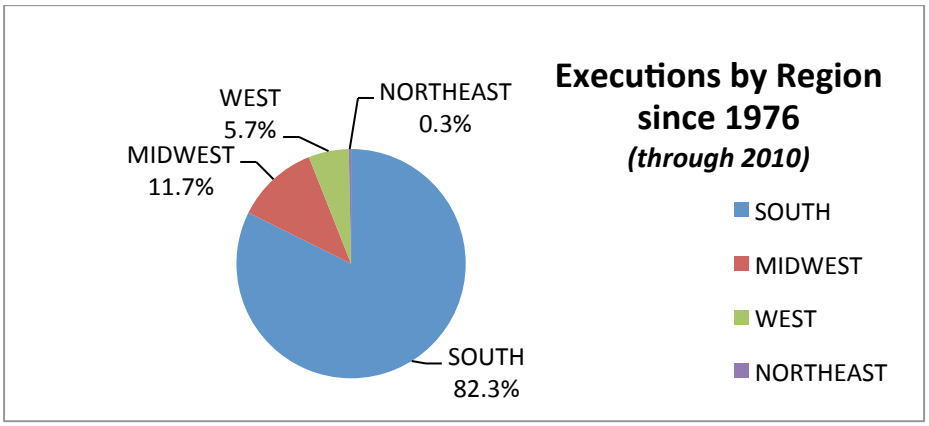
Today's class will cover graphs (bar graphs, pie charts, histograms) and numerical descriptors (mean, median, standard deviation, IQR). This is Sections 1.1, 1.2 in Introduction to the Practice of Statistics, 7-th edition, by D. Moore, G. McCabe, B. Craig. (W.H. Freeman, 2012)

GRAPHS

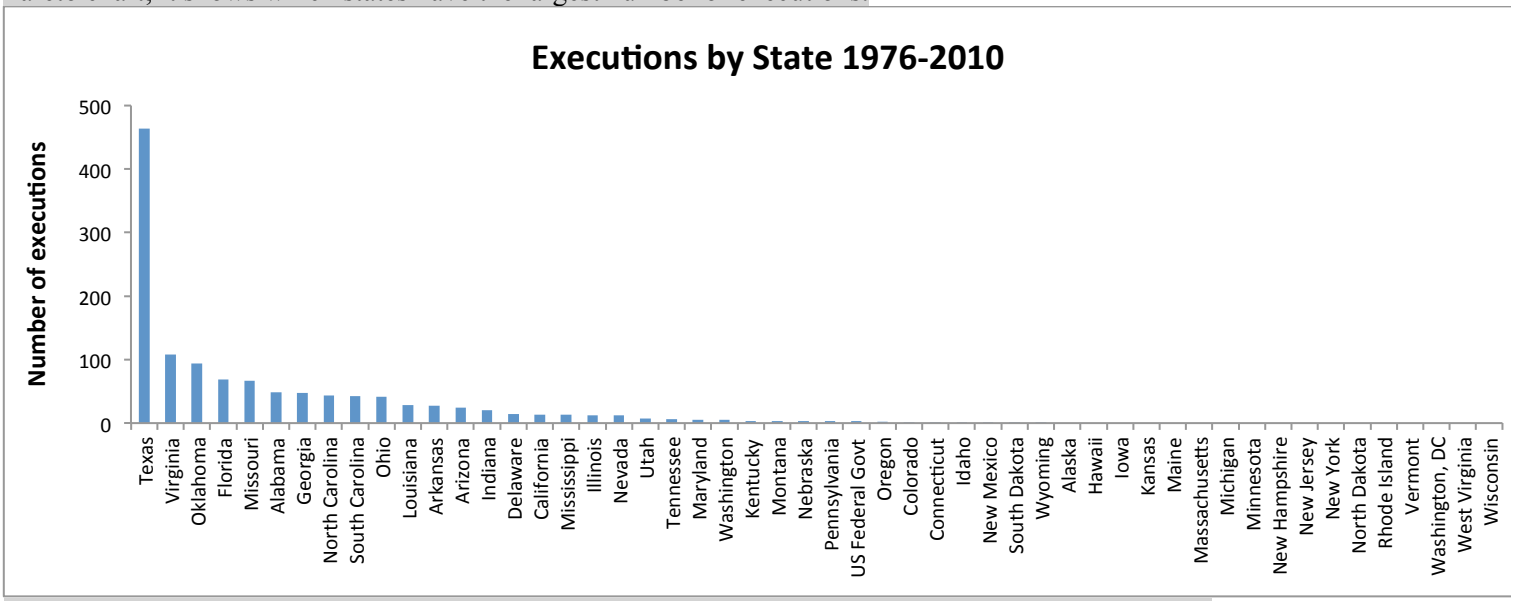
Executions by Region

REGION	TOTAL	
SOUTH	1016	0.823
MIDWEST	144	0.117
WEST	70	0.057
NORTHEAST	4	0.003
Total	1234	1.000

How would you make a pie chart?



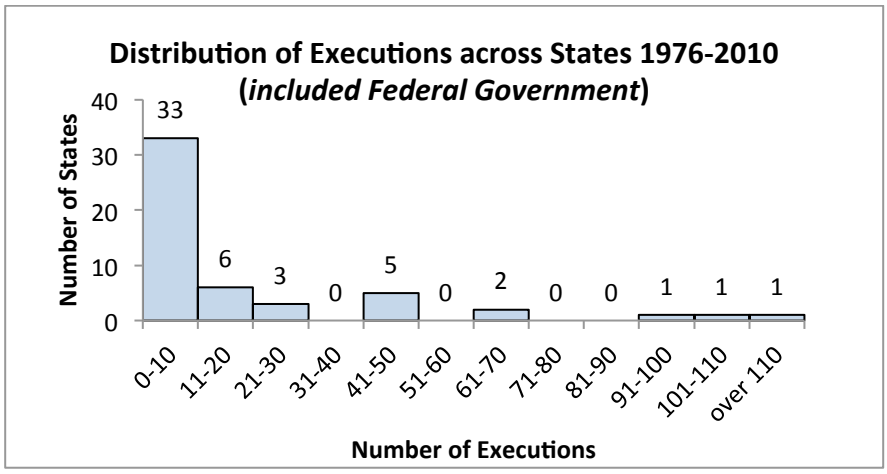
Executions by State What sort of graph is this? What is it useful for?
Pareto chart; it shows which states have the largest number of executions.



We may also be interested in number of executions per capita, to take the size of the state into account.

What is a histogram? For the executions by state, make a count using the bins below, then make a bar chart:

Bin	*Frequency*
10 or less	33
20	6
30	3
40	0
50	5
60	0
70	2
80	0
90	0
100	1
110	1
More	1



From Death Penalty Information Center (http://www.deathpenaltyinfo.org)

STATE	TOTAL EXECUTIONS
Texas	464
Virginia	108
Oklahoma	94
Florida	69
Missouri	67
Alabama	49
Georgia	48
North Carolina	43
South Carolina	42
Ohio	41
Louisiana	28

Arkansas	27
Arizona	24
Indiana	20
Delaware	14
California	13
Mississippi	13
Illinois	12
Nevada	12
Utah	7
Tennessee	6
Maryland	5
Washington	5
Kentucky	3

Montana	3
Nebraska	3
Pennsylvania	3
US Federal Govt	3
Oregon	2
Colorado	1
Connecticut	1
Idaho	1
New Mexico	1
South Dakota	1
Wyoming	1
Total	**1234**

Also 17 states with no executions.

Numbers that Describe the Distribution of the Data: "Sample Statistics"

Statistics *including* Texas
Mean = 24.2
Median = 3
Standard Deviation = 67.6
Quartiles: Q1 = 0
Quartiles: Q2 = 3
Quartiles: Q3 = 22
IQR = 22 – 0 = 22

Outliers *with* Texas
1.5×IQR = 33 above Q3 and below Q1.
So above 55. That is, Texas, Virginia, Oklahoma, Florida, Missouri

Descriptors *without* Texas
Mean = 15.4
Median = 3
Standard Deviation = 25.3
Quartiles: Q1 = 0
Quartiles: Q2 = 3
Quartiles: Q3 = 18.5
IQR = 18.5 – 0 = 18.5

Outliers *without* Texas
1.5×IQR = 27.25 above Q3 and below Q1.
So above 45.75. That is, Virginia, Oklahoma, Florida, Missouri, Alabama, Georgia

Statistics: Mean, Median, Standard Deviation

To communicate about a distribution without showing a histogram, ***sample statistics*** are useful.

Suppose the data values are $x_1, x_2, x_3, \ldots, x_n$, so n is the size of the data set. Then x_i is a typical data value for $i = 1,2, \ldots, n$.

- **Mean**, or average value, $\bar{x}$, obtained by adding the values and dividing by the number of values.

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **Median** is the middle value. If there is an even number of values, the median is the average of the two middle values.
- **Standard Deviation** measures spread of the data:

$$s_x = \sqrt{\frac{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}{n-1}}$$

Why take the square root? To get same units as x.
Why is it $(n-1)$ and not n? This is subtle (not necessary for Math 263). It is because the formula is for a sample, not a population.

On a histogram:
Mean is "balance point"
Median has half the data to the left, half to the right.
Standard deviation is "average distance from mean"

In Excel,
Mean is =AVERAGE(array);
Median is =MEDIAN(array);
Standard deviation is =STDEV(array) or STEV.S(array) in Excel 2010 and 2011

Ex: If data set is 1, 2, 9, 20, 5 then $n = 5$ and the statistics are:

$\text{Mean} = \bar{x} = \frac{1+2+9+20+5}{5} = 7.4$

$\text{Median} = \text{Middle value when arranged in order} = 5$

$\text{Standard deviation} = \sqrt{\frac{1}{4}((1-7.4)^2 + (2-7.4)^2 + (9-7.4)^2 + (20-7.4)^2 + (5-7.4)^2)} = 7.7.$

Why is the mean larger than the median?
The 20 "pulls up" the mean, but doesn't affect the median. The 20 can be considered an **outlier**.

What does the standard deviation represent?
It measures the spread and can be thought of as the "average distance from the average". The value is large because the 20 is far from the mean and the difference is squared.

What are can we learn from Sample Statistics ?

1. Look at the relationship between the mean and the median: How does this relate to the shape of the graph? This histogram is called **skewed right** because the tail goes to the right.

   Mean is larger because histogram skewed to right.

2. Which of the statistics change a lot when Texas is removed? Why? These statistics are called **sensitive**, or said to be **not robust**.

   The mean and standard deviation are not robust. The value for Texas enters directly into the formula and so changes the mean/SD when Texas is included.

3. Which statistics are robust?
   The median and IQR.

4. What generalization can you make about outliers and the shape of the graph?

   If the histogram is skewed right, the mean will be larger than the median. If the histogram is skewed left, the mean will be less than the median.

Box Plot