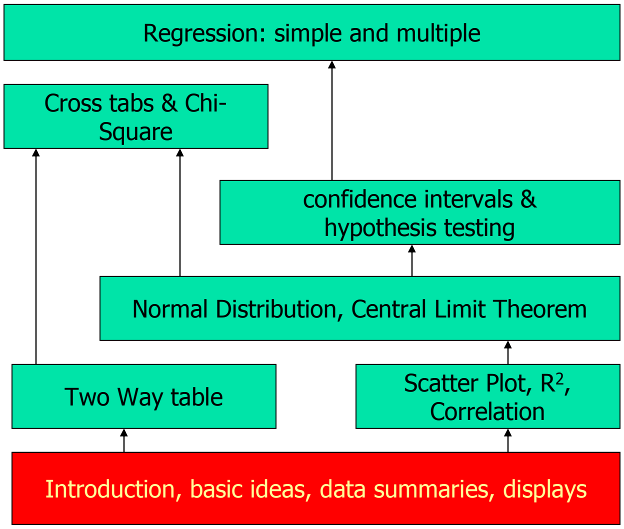
PUB – POS 316
Week 3

Data Summaries

Navid Ghaffarzadegan
navidg@gmail.com
Last updated – Jan 1, 10

Course Road Map



Regression: simple and multiple

Cross tabs & Chi-Square

confidence intervals & hypothesis testing

Normal Distribution, Central Limit Theorem

Two Way table

Scatter Plot, R^2, Correlation

Introduction, basic ideas, data summaries, displays

Agenda

- 1. Displaying data with graphs
  - 1.1. Graphs for categorical variables
  - 1.2. Stemplot
  - 1.3. Histograms
  - 1.4. Time plots
- 2. Displaying data with numbers
  - 2.1. Mean, Median, Quartiles, Box plot
  - 2.2. Standard Deviation

1. Displaying data with graphs
1.1. bar chart and Pie chart

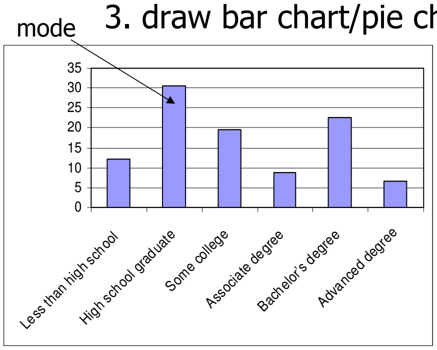
- Graphs for categorical variables
  - The distribution of the highest level of education for people aged 25 to 34 years:

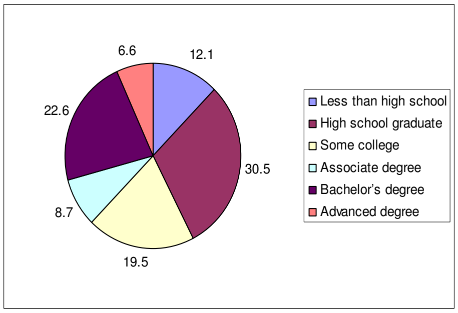
education	Count (millions)	Percent
Less than high school	4.6	
High school graduate	11.6	
Some college	7.4	
Associate degree	3.3	
Bachelor’s degree	8.6	
Advanced degree	2.5	

1. Displaying data with graphs
1.1. bar chart and Pie chart

- Graphs for categorical variables
  - Why is this called categorical data?
  - How should we calculate the percentage?
  - 1. calculate total, 2. calculate the ratio for each row

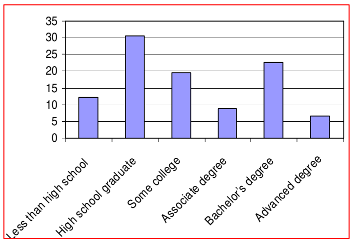
  3. draw bar chart/pie chart in excel.

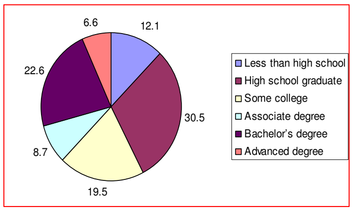




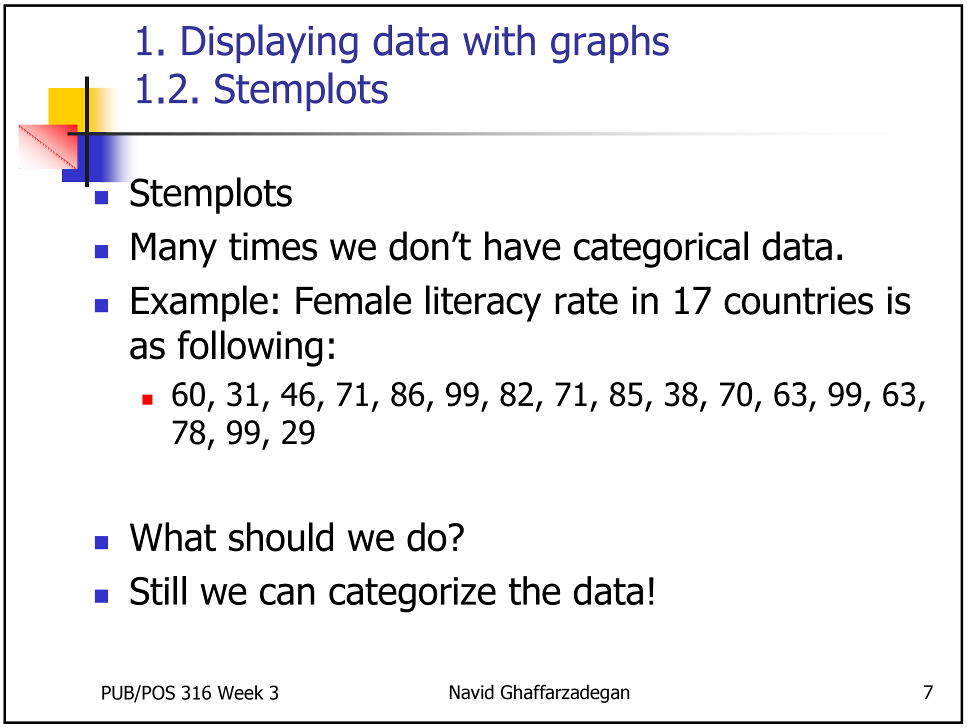
1. Displaying data with graphs
1.1. bar chart and Pie chart

- Graphs for categorical variables
  - Compare: which is better?



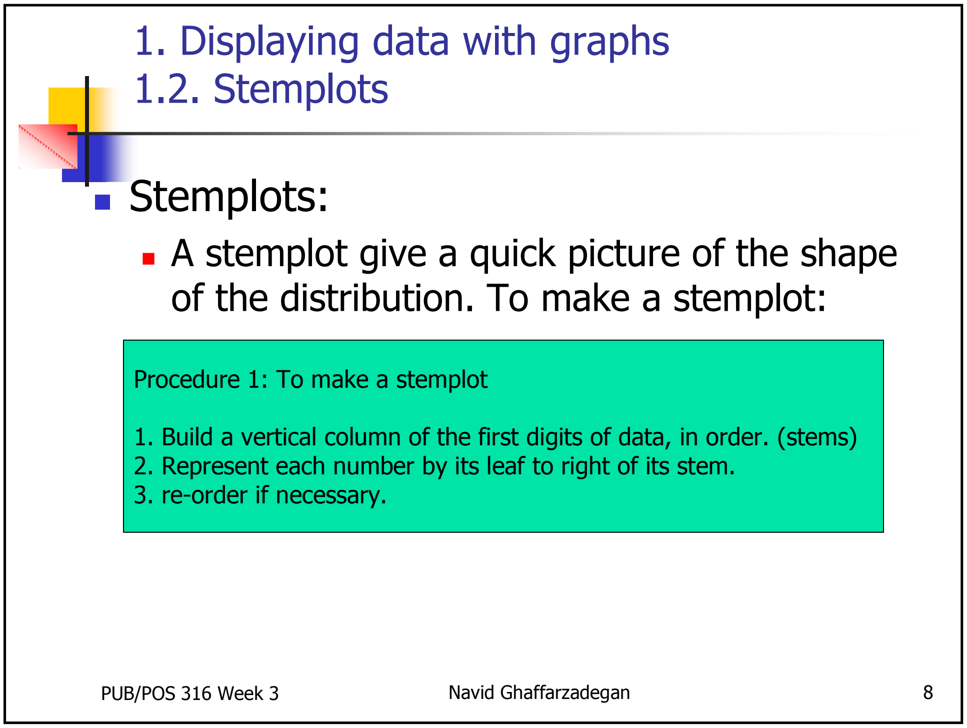


education	Count (millions)	Percent
Less than high school	4.6	12.1
High school graduate	11.6	30.5
Some college	7.4	19.5
Associate degree	3.3	8.7
Bachelor’s degree	8.6	22.6
Advanced degree	2.5	6.6
Total	38	



1. Displaying data with graphs
1.2. Stemplots

- Stemplots
- Many times we don’t have categorical data.
- Example: Female literacy rate in 17 countries is as following:
  - 60, 31, 46, 71, 86, 99, 82, 71, 85, 38, 70, 63, 99, 63, 78, 99, 29
- What should we do?
- Still we can categorize the data!



1. Displaying data with graphs
1.2. Stemplots

- Stemplots:
  - A stemplot give a quick picture of the shape of the distribution. To make a stemplot:

Procedure 1: To make a stemplot

1. Build a vertical column of the first digits of data, in order. (stems)
2. Represent each number by its leaf to right of its stem.
3. re-order if necessary.

1. Displaying data with graphs
1.2. Stemplots

- Example:
  - 60, 31, 46, 71, 86, 99, 82, 71, 85, 38, 70, 63, 99, 63, 78, 99, 29

```
2 | 9
3 | 1 8
4 | 6
5 |
6 | 0 3 3
7 | 0 1 1 8
8 | 2 5 6
9 | 9 9 9
```

- What can we learn from this representation?
  - Overall pattern? Max, min, median?

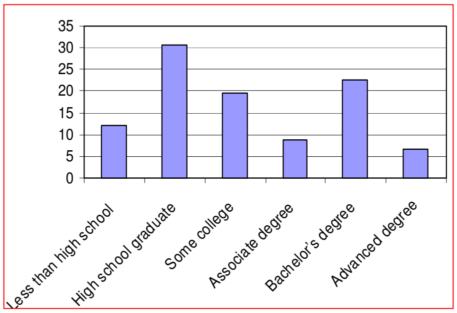
1. Displaying data with graphs
1.3. Histograms

- The book differentiates bar chart and histograms. They don’t differ that much!
- Y-axis in bar chart is percentage, but in histogram is frequency.

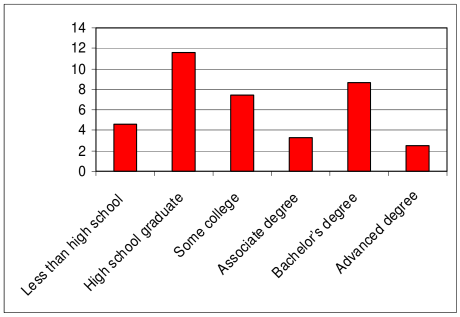
education	Count (millions)	Percent
Less than high school	4.6	12.1
High school graduate	11.6	30.5
Some college	7.4	19.5
Associate degree	3.3	8.7
Bachelor’s degree	8.6	22.6
Advanced degree	2.5	6.6
Total	38	

1. Displaying data with graphs
1.3. Histograms

- The book differentiates bar chart and histograms. They don’t differ that much!



Bar chart



histogram

1. Displaying data with graphs
1.3. Histograms

- Example: How can we make of a sense of performance of a class?

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

1. Displaying data with graphs
1.3. Histograms

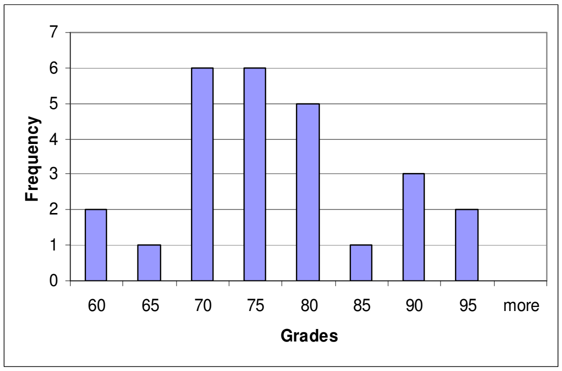
- Example: How can we make of a sense of performance of a class?

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

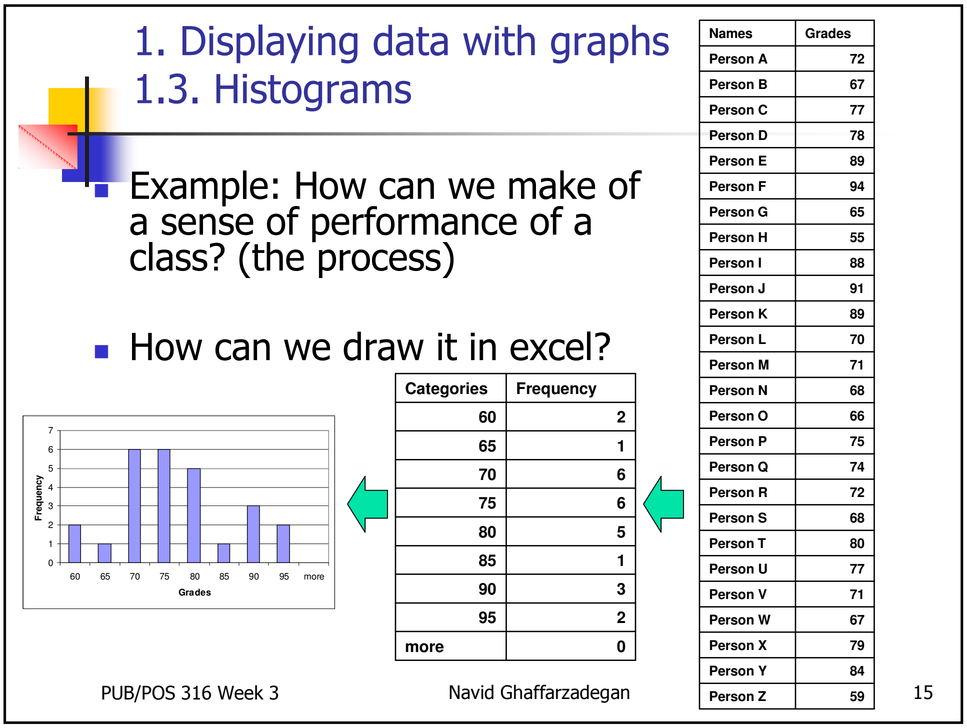
Categories	Frequency
60	2
65	1
70	6
75	6
80	5
85	1
90	3
95	2
more	0

1. Displaying data with graphs
1.3. Histograms

- Example: How can we make of a sense of performance of a class?



Categories	Frequency
60	2
65	1
70	6
75	6
80	5
85	1
90	3
95	2
more	0

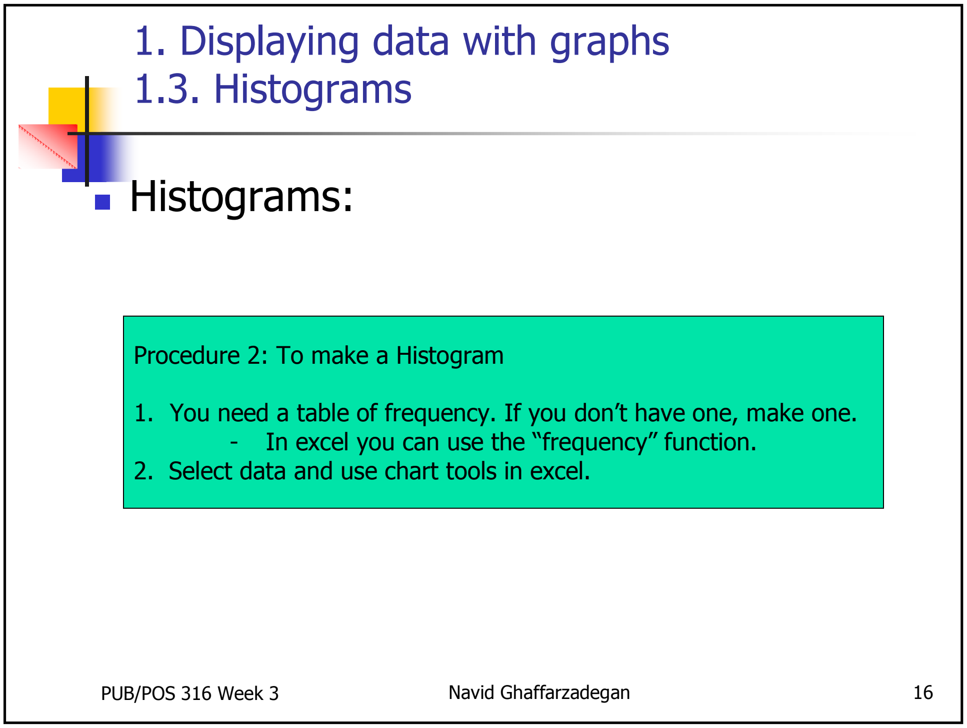


1. Displaying data with graphs
1.3. Histograms

- Example: How can we make of a sense of performance of a class? (the process)
- How can we draw it in excel?

Categories	Frequency
60	2
65	1
70	6
75	6
80	5
85	1
90	3
95	2
more	0

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59



1. Displaying data with graphs
1.3. Histograms

- Histograms:

Procedure 2: To make a Histogram

1. You need a table of frequency. If you don't have one, make one.
   - In excel you can use the "frequency" function.
2. Select data and use chart tools in excel.

1. Displaying data with graphs
1.3. Histograms

- Histograms:

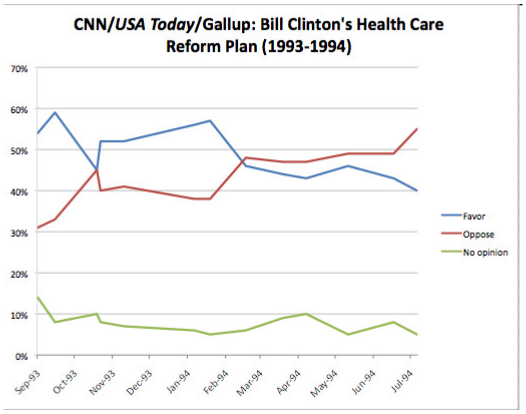
Procedure 2: To make a Histogram

1. You need a table of frequency. If you don’t have one, make one.
   - In excel you can use the “frequency” function.
2. Select data and use chart tools in excel.

- For more details see “Dynamic Do-It-Yourself Histograms“ at: http://peltiertech.com/Excel/Charts/Histograms.html skip the beginning of the page, it may confuse you. Only read Dynamic Do-It-Yourself Histograms
- Note: you don’t necessarily need to install analysis tool pack here (we will need it later)

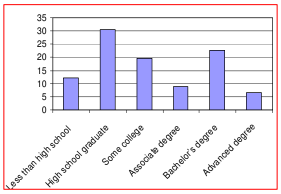
1. Displaying data with graphs
1.4. Time Series

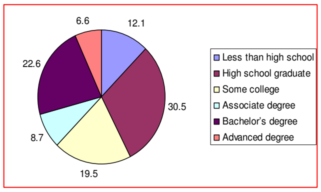
- In time series the X-axis is time.
- Graph over time.



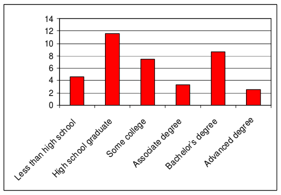
1. Displaying data with graphs Summary

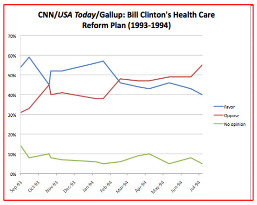
- Bar chart, Pie chart, stemplot, histograms, time series, [scatter plot]





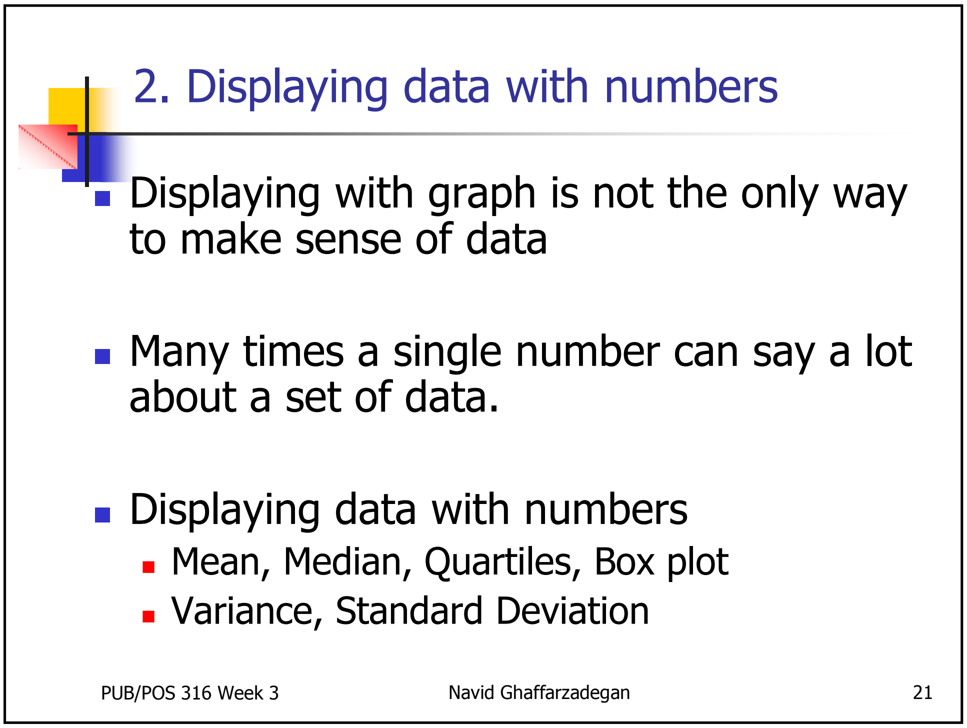
```
2 | 9
3 | 1 8
4 | 6
5 |
6 | 0 3 3
7 | 0 1 1
8 |
8 | 2 5 6
9 | 9 9 9
```





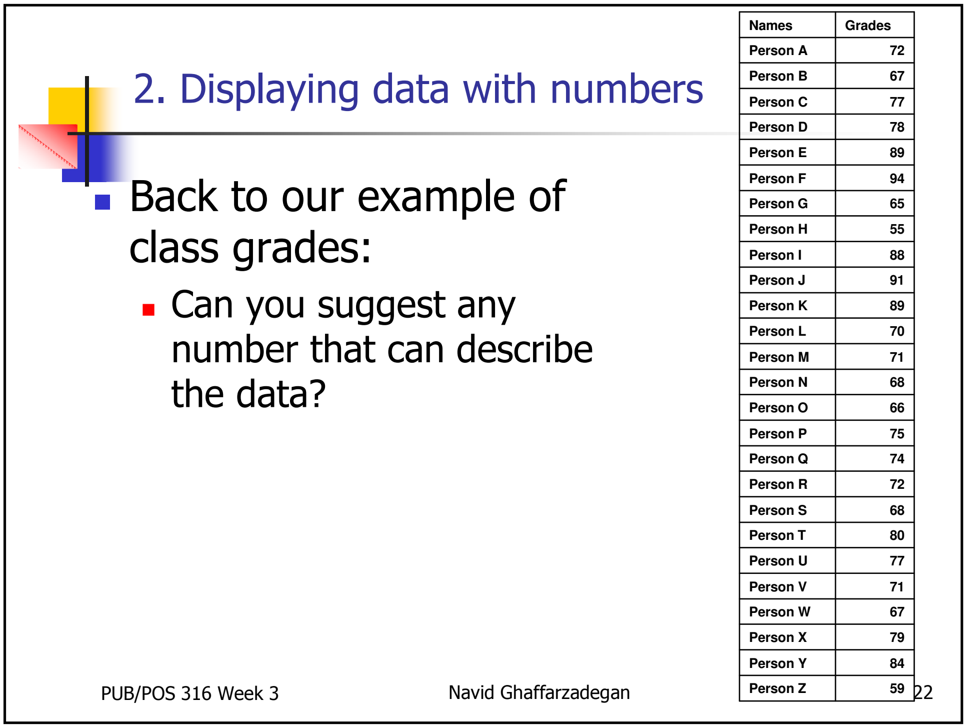
Agenda

- 1. Displaying data with graphs
  - 1.1. Graphs for categorical variables
  - 1.2. Stemplot
  - 1.3. Histograms
  - 1.4. Time plots
- 2. Displaying data with numbers
  - 2.1. Mean, Median, Quartiles, Box plot
  - 2.2. Standard Deviation



2. Displaying data with numbers

- Displaying with graph is not the only way to make sense of data
- Many times a single number can say a lot about a set of data.
- Displaying data with numbers
  - Mean, Median, Quartiles, Box plot
  - Variance, Standard Deviation



2. Displaying data with numbers

- Back to our example of class grades:
  - Can you suggest any number that can describe the data?

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

2. Displaying data with numbers

- Back to our example of class grades:
  - Can you suggest any number that can describe the data?
    - Mean
      - Add values divide by number of observation.→ 74.8

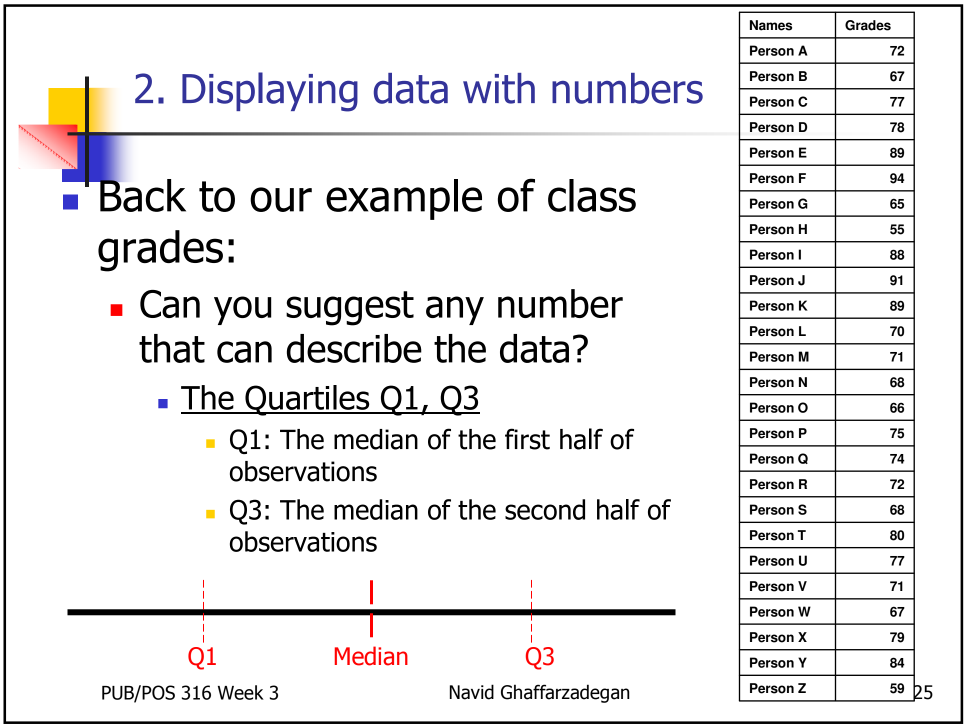
$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum x_i$$

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

2. Displaying data with numbers

- Back to our example of class grades:
  - Can you suggest any number that can describe the data?
    - Median
      - The midpoint of a distribution
      - 1. Arrange all observations in order
      - 2. 1. If you have odd number of observation, pick the middle one.
      - 2.2. If you have even number, report the average of two center observations
      - In our example: median = 73

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

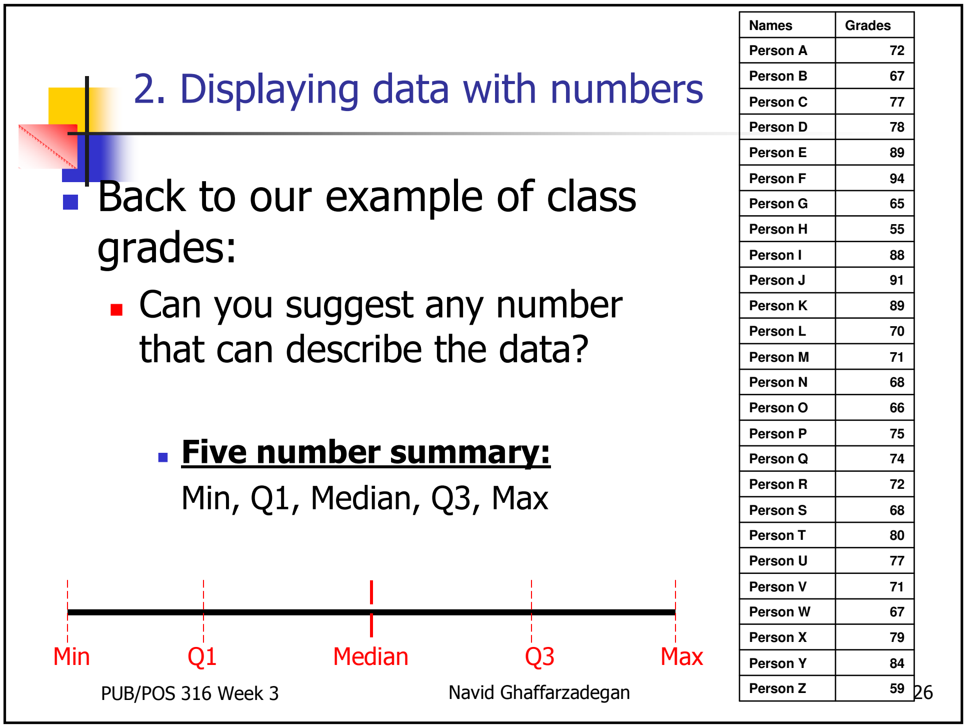


2. Displaying data with numbers

- Back to our example of class grades:
  - Can you suggest any number that can describe the data?
    - The Quartiles Q1, Q3
      - Q1: The median of the first half of observations
      - Q3: The median of the second half of observations

Q1
Median
Q3

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59



2. Displaying data with numbers

- Back to our example of class grades:
  - Can you suggest any number that can describe the data?
    - **Five number summary:**
      Min, Q1, Median, Q3, Max

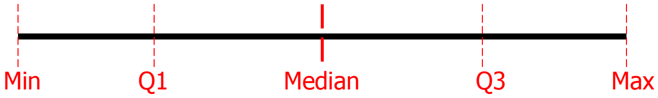
Min
Q1
Median
Q3
Max

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

2. Displaying data with numbers

- Example: How do you evaluate the difference in grading across these three classes?

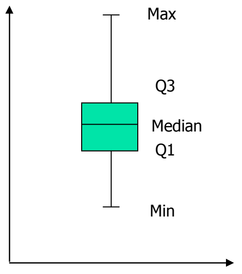
	A	B	C
Mean:	75	80	80
Min:	70	70	75
Q1:	72	72	77
Median:	75	73	81
Q3:	77	75	83
Max:	80	100	90



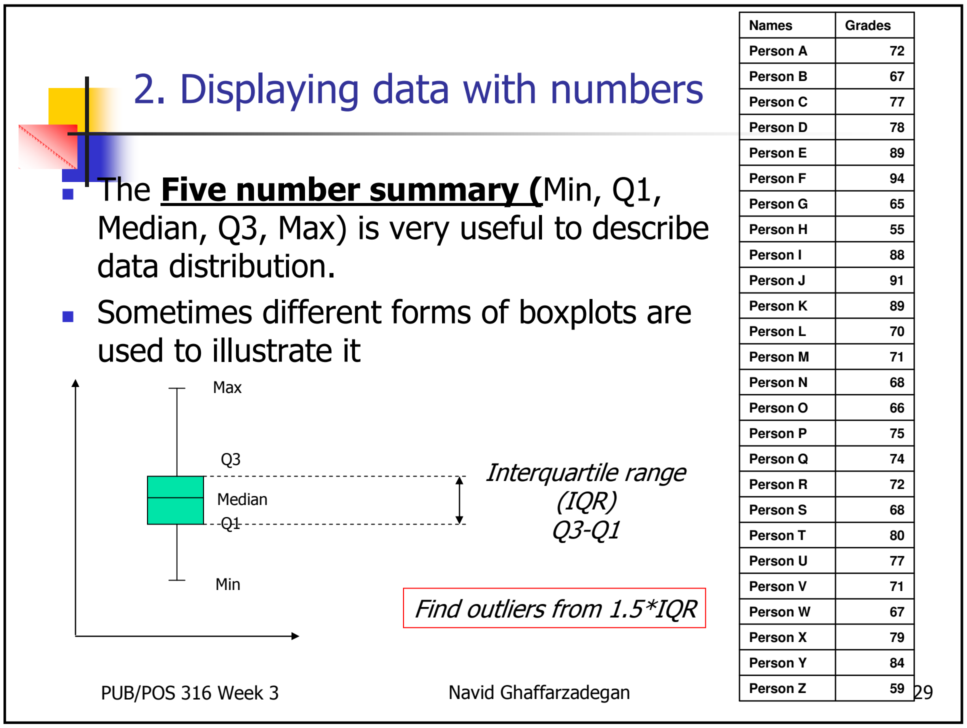
Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

2. Displaying data with numbers

- The **Five number summary (**Min, Q1, Median, Q3, Max) is very useful to describe data distribution.
- Sometimes different forms of boxplots are used to illustrate it



Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59



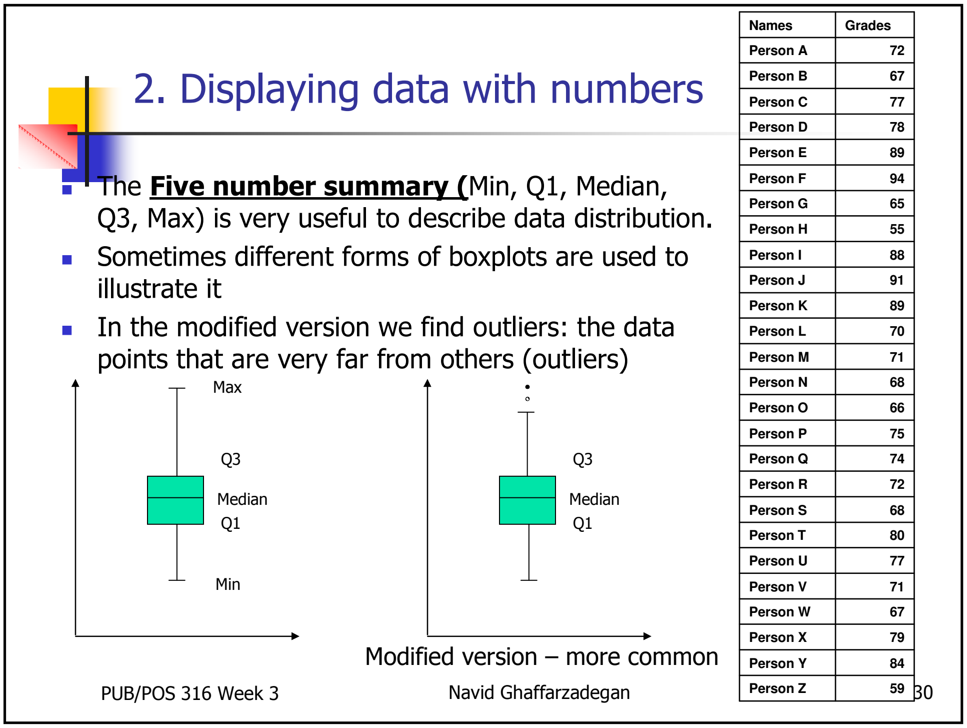
2. Displaying data with numbers

- The **Five number summary (**Min, Q1, Median, Q3, Max) is very useful to describe data distribution.
- Sometimes different forms of boxplots are used to illustrate it

Interquartile range (IQR) Q3-Q1

*Find outliers from 1.5*IQR*

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

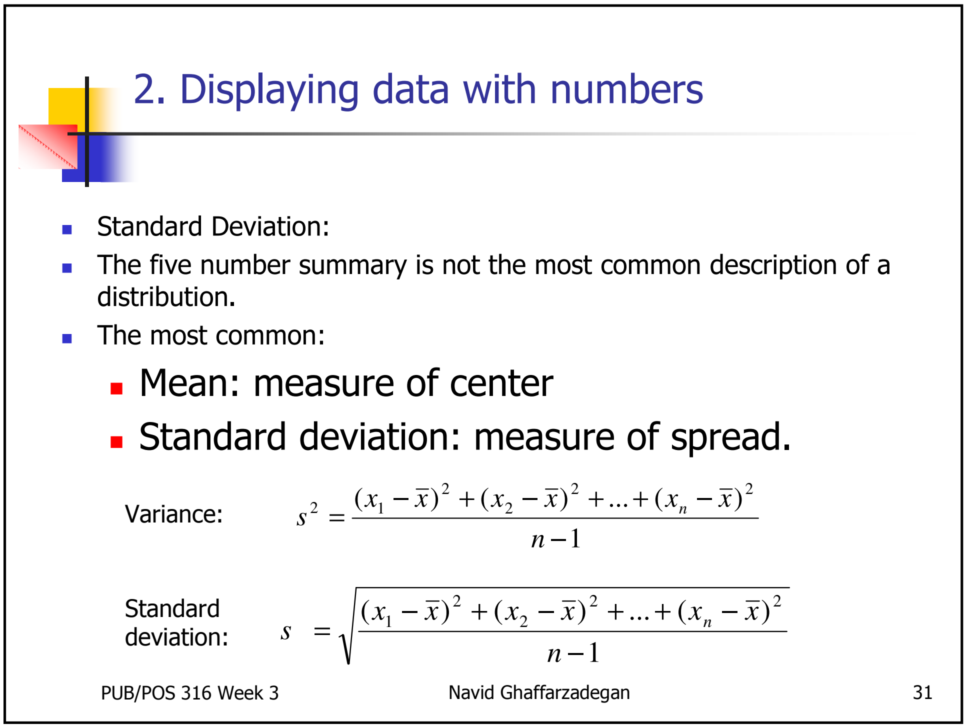


2. Displaying data with numbers

- The **Five number summary (**Min, Q1, Median, Q3, Max) is very useful to describe data distribution.
- Sometimes different forms of boxplots are used to illustrate it
- In the modified version we find outliers: the data points that are very far from others (outliers)

Modified version – more common

Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79
Person Y	84
Person Z	59

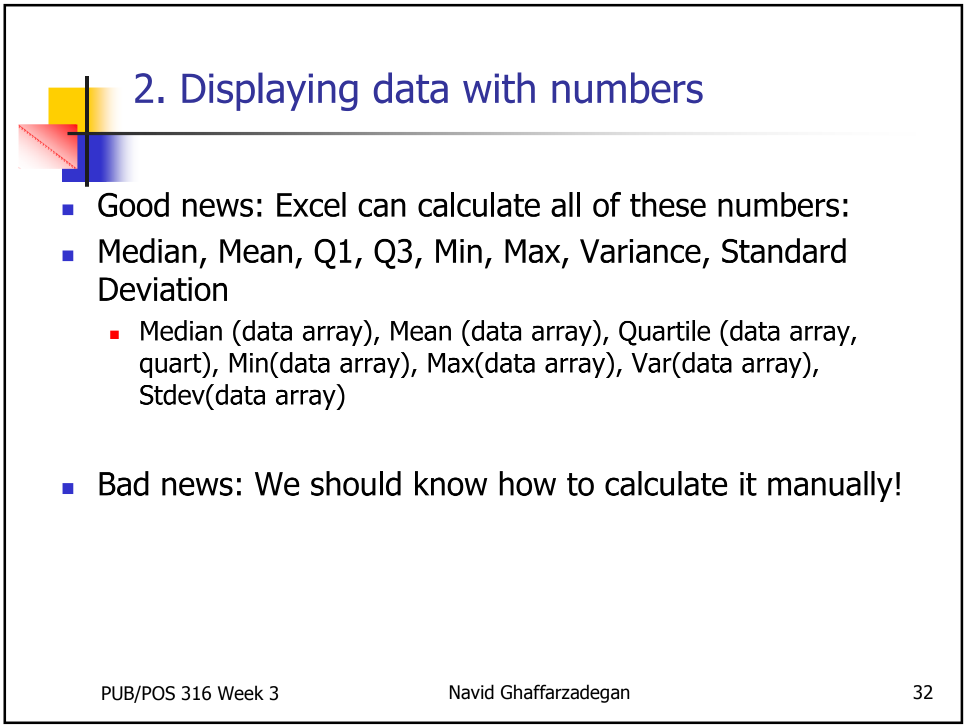


2. Displaying data with numbers

- Standard Deviation:
- The five number summary is not the most common description of a distribution.
- The most common:
  - Mean: measure of center
  - Standard deviation: measure of spread.

Variance: $$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}$$

Standard deviation: $$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}}$$



2. Displaying data with numbers

- Good news: Excel can calculate all of these numbers:
- Median, Mean, Q1, Q3, Min, Max, Variance, Standard Deviation
  - Median (data array), Mean (data array), Quartile (data array, quart), Min(data array), Max(data array), Var(data array), Stdev(data array)
- Bad news: We should know how to calculate it manually!