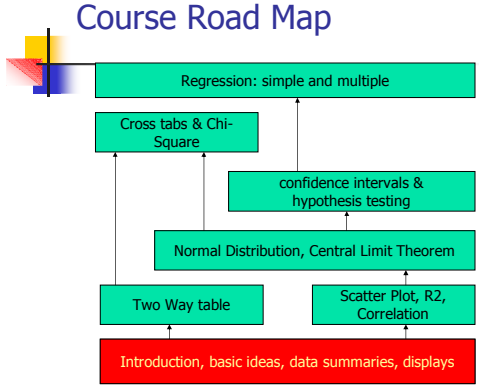
PUB – POS 316

Review

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

Course Road Map

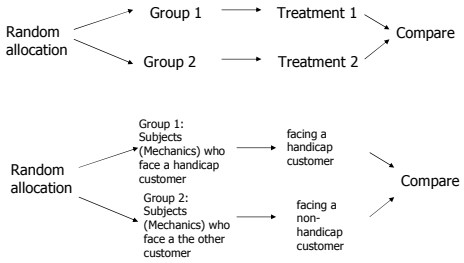


1. Systematic methods of research

- 2.1. Experiment
- 2.2. Survey
- 2.3. Qualitative methods: Interview, Observation

1. Systematic methods of research

- Experiment procedure.



- **(1) Research Design:** A researcher wants to study the effect of more homework on students' performance. He divides the students in two groups and gives one group more homework than the other. At the end of the semester he compares their grades.
- a) Draw a diagram that shows how different treatments are conducted and how different groups are compared in this study.
- b) Offer two ideas that can improve this study?

2. Displaying data with graphs

- 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.

2. Displaying data with graphs

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

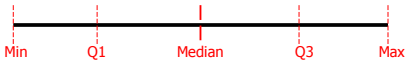
Procedure 1: To make a stemplot

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.

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

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

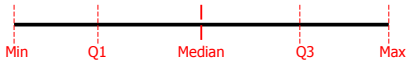


Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79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

2. Displaying data with numbers

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

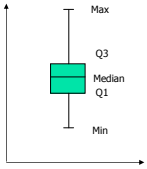
	A	B	C
Mean:	75	80	80
Min:	70	70	75
Q1:	72	72	77
Median:	75	73	81
Q3:	77	75	83
Max:	80	100	90



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

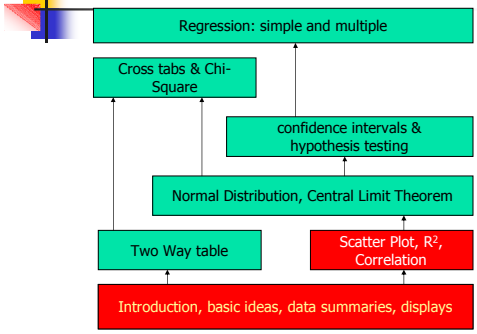


Names	Grades
Person A	72
Person B	67
Person C	77
Person D	78
Person E	89
Person F	94
Person G	65
Person H	55
Person I	88
Person J	91
Person K	89
Person L	70
Person M	71
Person N	68
Person O	66
Person P	75
Person Q	74
Person R	72
Person S	68
Person T	80
Person U	77
Person V	71
Person W	67
Person X	79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

- **(2) Data Summary:** The following data show annual mortality in 20 U.S. cities.
- Sketch 1) a stem plot and 2) a box plot of the data.

City	Mortality	City	Mortality
Akron, OH	921.87	Chattanooga, TN-GA	1017.61
Albany-Schenectady-Troy, NY	997.87	Chicago, IL	1024.89
Allentown, Bethlehem,PA-NJ	962.35	Cincinnati, OH-KY-IN	970.47
Atlanta, GA	982.29	Cleveland, OH	985.95
Baltimore, MD	1071.29	Columbus, OH	958.84
Birmingham, AL	1030.38	Dallas, TX	860.1
Boston, MA	934.7	Dayton-Springfield, OH	936.23
Bridgeport-Milford, CT	899.53	Denver, CO	871.77
Buffalo, NY	1001.9	Detroit, MI	959.22
Canton, OH	912.35	Flint, MI	941.18

Course Road Map



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

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

4. Association

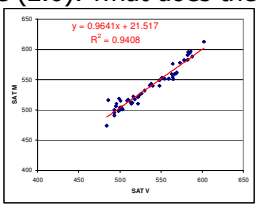
- Association between variables:

Two variables are associated if knowing the value of one of them tells you something about the other one.

- Examples:
  - *Effort and grade*
    - *Positive association*
  - *Price and demand*
    - *Negative association*

4. Association

- Example (2.6): what does the graph say?

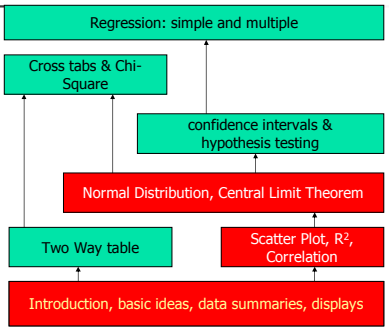


R^2 is the fraction of variance in y account for by the regression line.

Correlation (r) is the sqr root of R^2 with a proper sign.

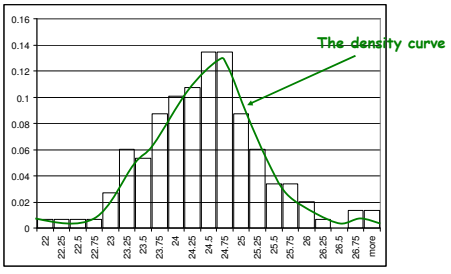
- **(3) Scatter Plots (First test 2010):** Here is a scatter plot. X-axis is the percentage of students who take the SAT test in each state and y-axis is the average SAT verbal in each state. The regression line best fitting these points is also shown.
- (a) The equation of the line is y = -1.0352x + 574.97. (R^2=0.81)
- The slope of this line is _____. What kind of association do you see: _____.
- What is the correlation between the variables?
- (b) What would the regression line predict for a state where the percentage of students who take the test is 50%:
- What about when the percentage is 10%:

Course Road Map



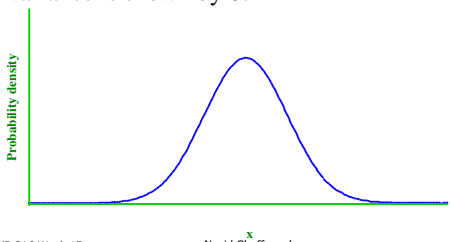
5. Normal Distribution

- Example: Biological clocks



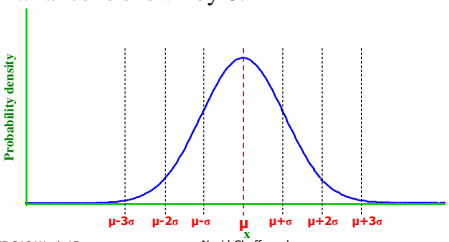
5. Normal Distribution

- Normal distribution is a bell-shaped, symmetric density curve. Average is shown by μ and variance is shown by σ.



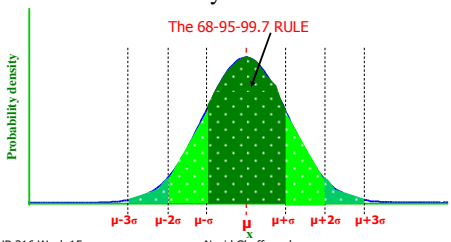
5. Normal Distribution

- Normal distribution is a bell-shaped, symmetric density curve. Average is shown by μ and variance is shown by σ.

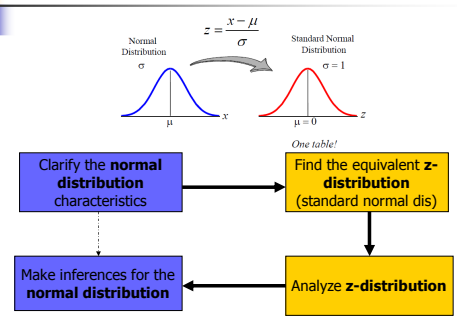


5. Normal Distribution

- Normal distribution is a bell-shaped, symmetric density curve. Average is shown by μ and variance is shown by σ.

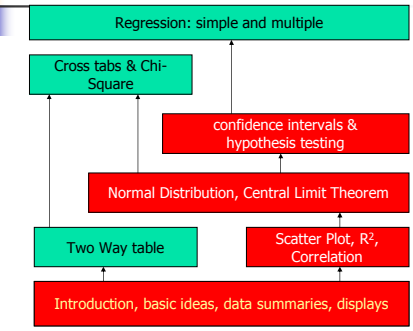


5. Normal Distribution



- **(5) Normal distribution (population):** Assume that milk consumption per capita is 0.78 of a glass a day.
- **a:** If 95% of people drink between 0.98 and 0.58 of a glass every day, what is the standard deviation of this distribution?
- **b:** Assume that doctors suggest every body to drink more than 1 glass of milk of every day. What portion of the population drink more than a glass of milk every day?

Course Road Map



6. Normal Distribution - samples

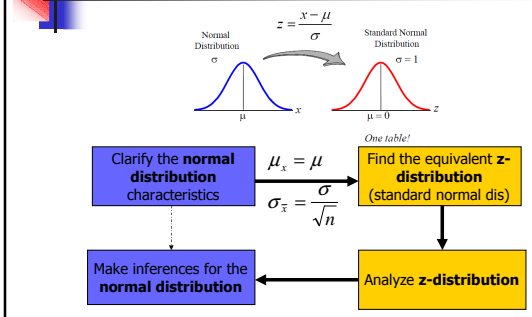
- Almost (always) we study samples. So, we don't have the distribution of population.
  - Two kind of questions:
    - Sample mean (e.g., average income of UAlbany students)
    - Sample proportion (e.g., percentage of people who agree with Obama's health care plan)

6. Normal Distribution - samples

- An example: We would like to know the average income of UAlbany students. We take a sample of 20 students, …..
  - **<u>The central limit theorem</u>**: when n is large, the sampling distribution of the sample mean is approximately normal with mean of μ and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

6. Normal Distribution - samples



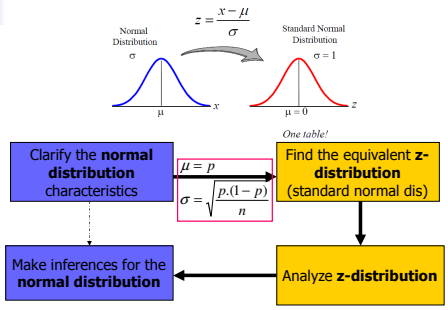
- **(6) Normal distribution (Sample mean):** Assume that milk consumption per capita is 0.78 of a glass a day, with a standard deviation of 0.1.
- a: If I take a sample of 50 people, what is the chance that my results show that milk consumption is more than 0.8 of a glass?
- b: If I take a sample of 50 people, what is the chance that my results show that milk consumption is between 0.79 and 0.8 of a glass?

6. Normal Distribution – samples (proportion)

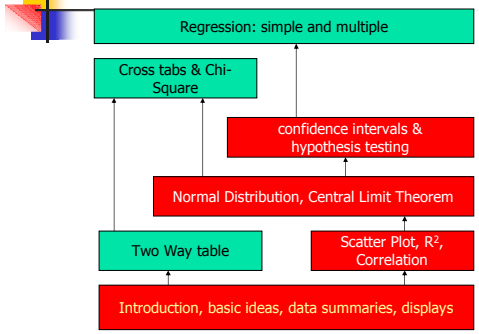
- An example: Suppose 60% of people agree with Obama's health reform. We poll 36 people ….
  - **<u>Every thing is the same BUT the way you calculate the standard deviation</u>**

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ $\qquad$ $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

6. Normal Distribution – samples (proportion)



Course Road Map



7. Confidence interval

- An interval containing the true value of the parameter with some probability.
- Example: Price of a medium size cup of coffee in Albany is:
  - $ 1.85 ±$0.35 (with 95% confidence)
- Margin of error

- X= (the result from the sample) ± z. (proper standard deviation)
- Or
- X= (the result from the sample) ± t. (proper standard deviation)

- Sample mean: $x = \mu \pm z.\frac{\sigma}{\sqrt{n}}$
- Sample proportion: $x = \mu \pm z.\sqrt{\frac{p(1-p)}{n}}$

- (7) From an ad:
- 86% of the UAlbany students are happy! This statement is based on a sample of 1011 students.

- Report the 95% confidence interval for this statement.

8. Hypothesis testing

- Hypothesis testing is a way to make a systematic conclusion from data.
- First you state your hypotheses.
- Then assume the null hypothesis is correct, and draw a distribution curve for it.
- Then, find p-value
  - P-Value: The probability that H_0 is true, based on our data.
- Then compare p-value with a critical number, usually 0.05. If it is lower you can reject the null hypothesis under 0.05 level of significance.

- **(8) Hypothesis testing:** We would like to measure happiness in our university. We ask from a sample of 100 people to report their happiness in the scale of 1-10 (10 very happy). The average of results is 5.5. Can we say that the average of happiness in our university is significantly higher than 5? Assume that the standard deviation in population is 2. State and test proper hypotheses.