Statistics and Quantitative Analysis U4320

Segment 4:

Probability Distributions: Univariate & Bivariate

URL: http://www.columbia.edu/itc/sipa/U4320y-003/

Prof. Sharyn O'Halloran

Copyright Sharyn O'Halloran 2001

Probability Distributions

- Outline
  - Focus on the distribution of a single event
    - Discrete random variables
      - Probability Tables
    - Continuous random variables
      - Probability distributions
      - Normal density curve
  - Distribution of two variables
    - Joint probability
    - Covariance
    - Correlation

Copyright Sharyn O'Halloran 2001

Probability Distributions: Discrete Random Variable

- Definition
  - A discrete random variable X has a finite number of possible values.
  - The probability distribution of X lists the values and their probabilities:

Value of X	X_1	X_2	X_3	...	x_k
Probability	P_1	P_2	P_3	...	P_k

- The probabilities p_i must satisfy:
  - Every probability p_i lies between 0 and 1.
  - $p_1 + p_2 + \ldots + p_k = 1$

Copyright Sharyn O'Halloran 2001

Probability Distributions: Discrete Random Variable (con t)

- Example: Distribution of grades in a large class
  - 15% of the students get A $ and D $, 30% receive B $ and C $, and 10% F $.
  - How can we put this into a table? (convert the grade to a 4 pt. scale)

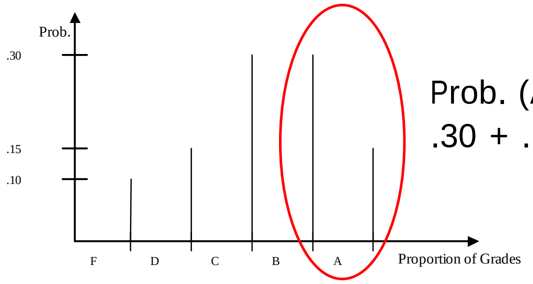
Grade	F=0	D=1	C=2	B=3	A=4
Probability	.10	.15	.30	.30	.15

- The probability of getting a B or better is:
  - P(A or B) = P(A) + P(B) = .15 + .30 = .45

Copyright Sharyn O'Halloran 2001

Probability Distributions: Discrete Random Variable (con t)

- How would we represent the grades as a distribution?

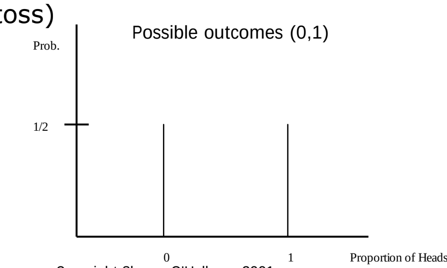


Prob. (A OR B)=
.30 + .15 = .45

Copyright Sharyn O'Halloran 2001

Probability Distributions: Discrete Random Variable (con t)

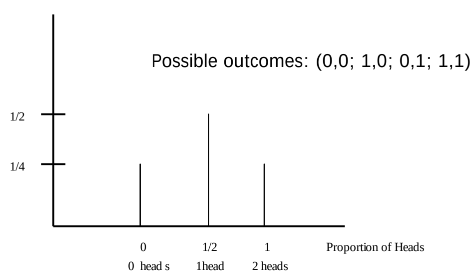
- How do discrete probability tables relate to continuous distributions?
  - What is the probability of getting a head? (1 coin toss)



Copyright Sharyn O'Halloran 2001

Probability Distributions(cont.)

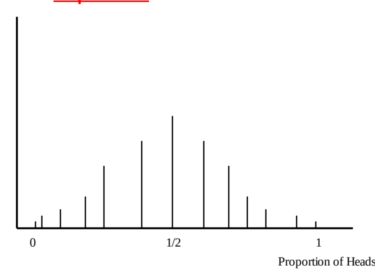
- Now say we flip the coin twice.
  - The picture now looks like:



Copyright Sharyn O'Halloran 2001

Probability Distributions(cont.)

- As number of coin tosses increases,
  - The distribution looks like a bell-shaped curve: flips.xls



Copyright Sharyn O'Halloran 2001

Probability Distributions:
Continuous Random Variable

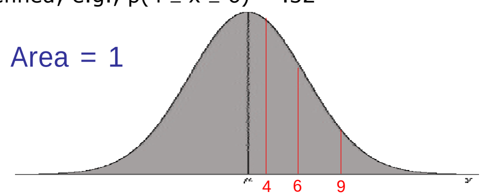
- Definition:
  - A continuous random variable is a variable which can take on an infinite (uncountable) number of values
- Example:
  - Suppose the time taken by all workers to commute from home to work falls between 5 minutes (min) and 130 minutes (max).
    - Then x can assume any value in the interval 5 to 130 minutes.
    - The interval contains an infinite number of possible values
    - The probability of any one (exact) value is zero.
- Solution:
  - Need to define a probability distribution over continuous range
  - Need to be able to calculate probabilities of possible events

Copyright Sharyn O'Halloran 2001

Probability Distributions(cont.)

- Probability distributions are idealized bar graphs or histograms.
  - The curve is drawn so that the total area underneath it is equal to 1
  - The probability of any one value is 0; e.g., $p(x=9)=0$
  - But the probability of a range of values is well-defined; e.g., $p(4 \le x \le 6) = .32$

Total Area = 1



Copyright Sharyn O'Halloran 2001

Probability Distributions:
Continuous Random Variable

- Definition:
  - A continuous random variable X takes all values in an interval of numbers
  - The probability distribution of X is described by a density curve.
  - The probability of any event is the area under the density curve and above the values of X that make up the event.
    - Any density curve gives the distribution of a continuous random variable. (e.g. normal curve)
    - Normal distributions as an idealized description of data are closely related to normal probability distributions.

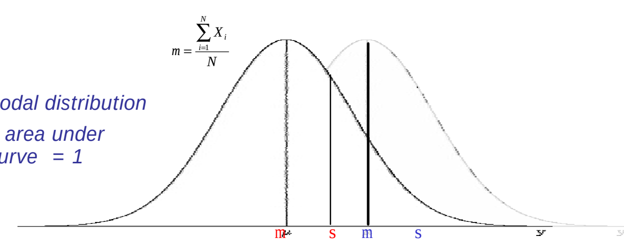
Copyright Sharyn O'Halloran 2001

Probability Distributions:
Normal Distribution

- Probabilities are areas under the curve.
  - Mean μ describes the central tendency
    - Changing μ w/o changing σ moves the curve horizontally

Unimodal distribution

Total area under the curve = 1



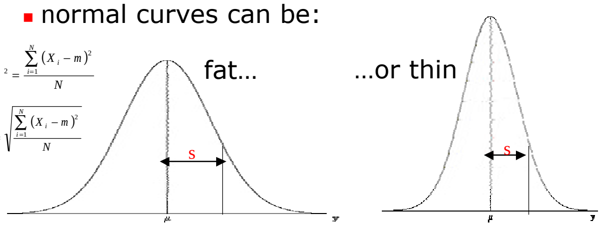
Copyright Sharyn O'Halloran 2001

Probability Distributions:
Normal Distribution

- Standard deviation σ controls the spread of the curve
  - normal curves can be:

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - m)^2}{N}$$

$$s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - m)^2}{N}}$$

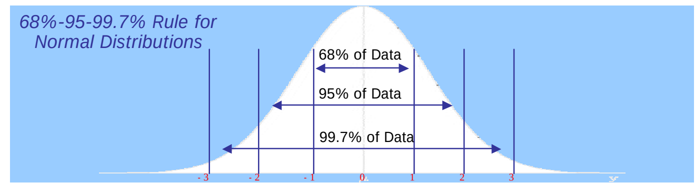


- As σ changes, so does the shape of the curve.
  - As σ increases the curve becomes more spread out.
  - As σ decreases the curve becomes less spread out.

Copyright Sharyn O'Halloran 2001

Probability Distributions:
Normal Distribution

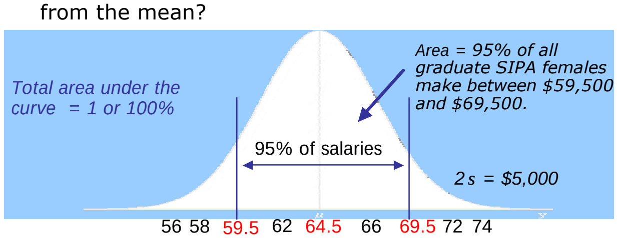
- But they all follow the 68-95-99.7 Rule:
  - 68% of the observations fall within σ of the mean μ.
  - 95% of the observations fall within 2σ of μ.
  - 99.7% of the observations fall within 3σ of μ.



Copyright Sharyn O'Halloran 2001

Probability Distributions:
Example: SIPA Salaries

- Distribution of salaries for female SIPA graduates
  - Mean μ = $64,500; Standard Deviation σ = $2,500
  - What is the area representing 2 standard deviations from the mean?

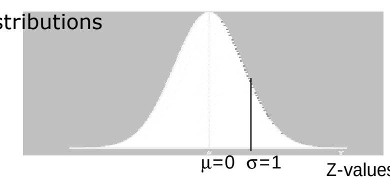


- What if σ = $2,000?

Copyright Sharyn O'Halloran 2001

Probability Distributions:
Standard Normal Distribution

- Definition:
  - Standard Normal Curve is a normal distribution with mean 0 and standard deviation 1.
- Characteristics
  - continuous distributions
  - symmetric
  - Unimodal

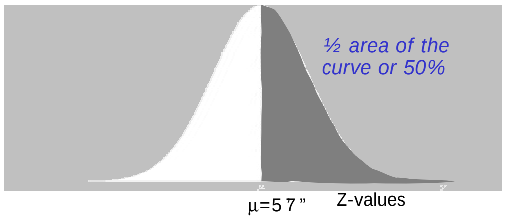


- Z-values
  - points on the x-axis that show how many standard deviations the observation is away from the mean μ.

Copyright Sharyn O'Halloran 2001

Probability Distributions:
Standard Normal Distribution (cont.)

- Example: Height of people are normally distributed with mean 5'7"
  - What is the proportion of people taller than 5'7"?



Copyright Sharyn O'Halloran 2001

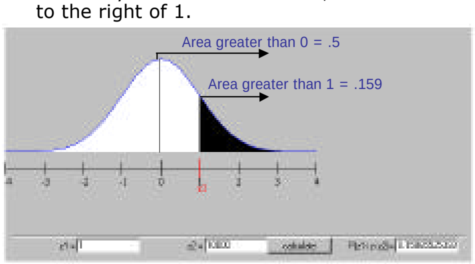
Probability Distributions:
How to Calculate a Z-score

- Z-Score
  - Z-value is the number of standard deviations away from the mean
  - Z-tables give the probability (score) of observing a particular z-value or greater
    - You can calculate any area under the standard normal curve using a combination of z-scores.
- Finding a z-Score
  - Use z-Tables in back of book
    - First two digits come from the left column
    - Third digit comes from the top right
  - Or, use online sources!

Copyright Sharyn O'Halloran 2001

Probability Distributions:
How to Calculate a Z-score (cont.)

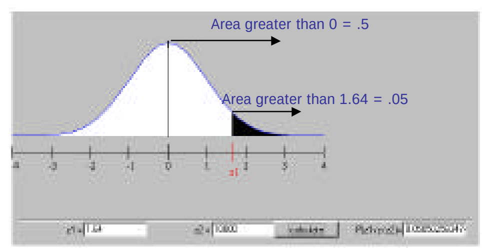
- What is the area under the curve that is greater than 1 ?
  - Prob (Z>1)
    - The entry in the table is 0.159, which is the total area to the right of 1.



Copyright Sharyn O'Halloran 2001

Probability Distributions:
How to Calculate a Z-score (cont.)

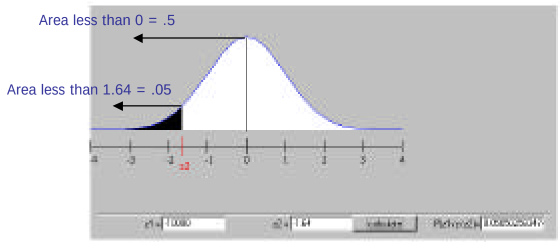
- What is the area to the right of 1.64?
  - Prob (Z > 1.64)
    - The table gives 0.051, or about 5%.



Copyright Sharyn O'Halloran 2001

Probability Distributions:
How to Calculate a Z-score (cont.)

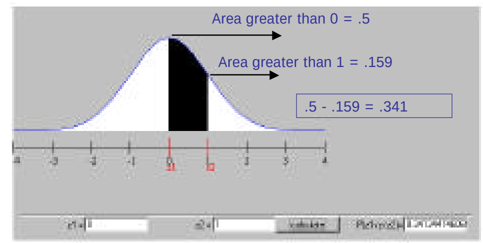
- What is the area to the left of -1.64?
  - Prob (Z < -1.64)



Copyright Sharyn O'Halloran 2001

Probability Distributions:
How to Calculate a Z-score (cont.)

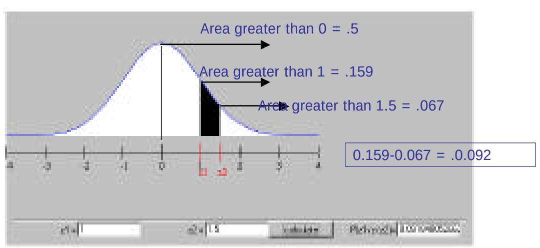
- What is the probability that an observation lies between 0 and 1?
  - Prob (0 < Z < 1)



Copyright Sharyn O'Halloran 2001

Probability Distributions:
How to Calculate a Z-score (cont.)

- How would you figure out the area between 1 and 1.5 on the graph?
  - Prob (1 < Z < 1.5)

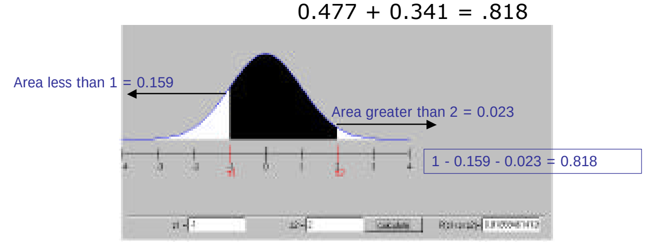


Copyright Sharyn O'Halloran 2001

Probability Distributions:
How to Calculate a Z-score (cont.)

- What is the area between -1 and 2?
  - Prob (-1 < Z < 2)
    - P (-1<Z<0) = .341
    - P (0<z<2) = 0.50-.023 =.477

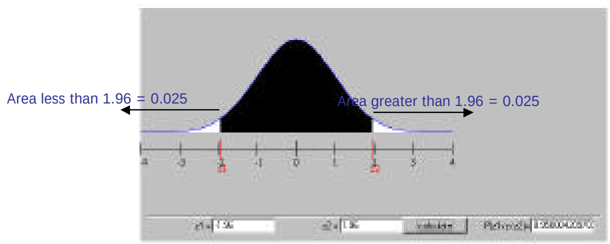
      0.477 + 0.341 = .818



Copyright Sharyn O'Halloran 2001

Probability Distributions: How to Calculate a Z-score (cont.)

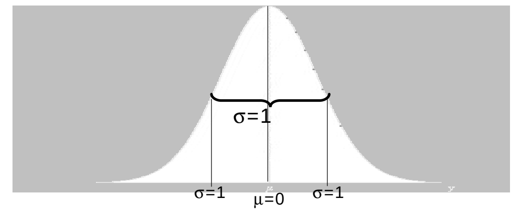
- What is the area between −1.96 and 1.96?
  - Prob(-1.96<Z< 1.96)
    - 1- Prob (Z< -1.96) - Prob (Z>1.96)
    - = 1 - .025 - .025 = .9500



Copyright Sharyn O'Halloran 2001

Probability Distributions: Standardization

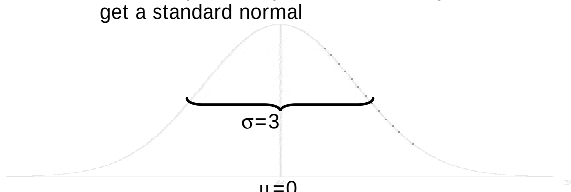
- Standard Normal Distribution
  - SND is a special case where
    - the mean of distribution equals 0 and
    - the standard deviation equals 1.
  - But you can convert any normal curve to a SND



Copyright Sharyn O'Halloran 2001

Probability Distributions: Standardization (cont.)

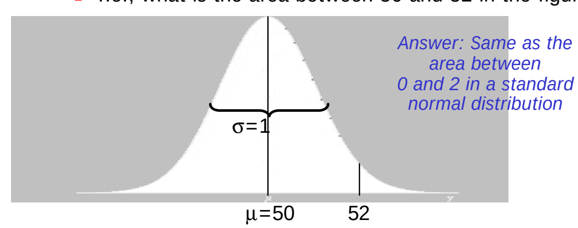
- Case 1: standard deviation differs from 1
  - Take a normal distribution with mean 0 and some standard deviation σ.
  - Convert any point x to the standard normal distribution by changing it to x/σ.
    - For example, change each x in the figure below to x/3 to get a standard normal



Copyright Sharyn O'Halloran 2001

Probability Distributions: Standardization (cont.)

- Case 2: Mean differs from 0
  - Take a normal distribution with mean μ and standard deviation 1
  - You can convert any point x to the standard normal distribution by changing it to $x- \mu$
  - I.e., what is the area between 50 and 52 in the figure?



Copyright Sharyn O'Halloran 2001

Probability Distributions:
Standardization (cont.)

- General Case:
  - Mean not equal to 0 and Standard Deviation not equal to 1
    - Say you have a normal distribution with mean μ & standard deviation σ.
    - You can convert any point x in that distribution to the same point in the standard normal by computing:

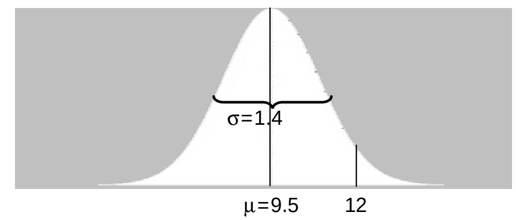
$$Z = \frac{x - m}{s}$$

  - This is called standardization
    - The Z-value is the equivalent of the original x value
    - You can then look up the Z-value in the normal table

Copyright Sharyn O'Halloran 2001

Probability Distributions:
Standardization (cont.)

- Trout Example:
  - The lengths of trout caught in a lake are normally distributed with mean 9.5" and standard deviation 1.4".
  - There is a law that you can't keep any fish below 12". What percent of the trout is this?



Copyright Sharyn O'Halloran 2001

Probability Distributions:
Standardization (cont.)

- Trout Example (continued)
  - Step 1: Standardize
    - Find the Z-score of 12: Prob (x>12)

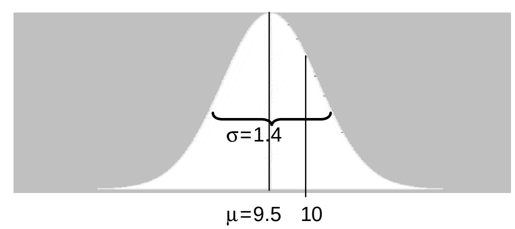
$$Z = \frac{x - m}{s} = \frac{12 - 9.5}{1.4} = 1.79$$

  - Step 2: Find Prob (Z>1.79)
    - Look up 1.79 in Standard Normal table.
    - Only .037, or about 4% of the fish could be kept.

Copyright Sharyn O'Halloran 2001

Probability Distributions:
Standardization (cont.)

- Trout Example (part 2):
  - Now they're thinking of changing the standard to 10" instead of 12"
  - What proportion of fish could be kept under the new limit?



Copyright Sharyn O'Halloran 2001

Probability Distributions: Standardization (cont.)

- Trout Example (Part 2)
  - Step 1: Standardize
    - Find the Z-score of 10: Prob (x>10)

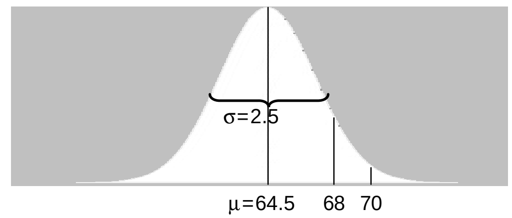
$$Z = \frac{x-m}{s} = \frac{10-9.5}{1.4} = 0.36$$

  - Step 2: Find Prob (Z>0.36)
    - Look up 0.36 in your table
    - Now .359, or almost 36% of the fish could be kept under the new law.

Copyright Sharyn O'Halloran 2001

Probability Distributions: Standardization (cont.)

- Salary Example
  - Recall that SIPA grads make an average of \$64.5K per year, with a standard dev. of \$2.5K
  - Let's solve the original problem: what percentage make between 68K and 70K?



Copyright Sharyn O'Halloran 2001

Probability Distributions: Standardization (cont.)

- Salary Example
  - Step 1: Standardize
    - Find the Z-scores of 68 and 70:

$$Z_{68} = \frac{x-m}{s} = \frac{68-64.5}{2.5} = 1.40$$

$$Z_{70} = \frac{x-m}{s} = \frac{70-64.5}{2.5} = 2.20$$

  - Step 2: Find Prob (1.40<Z<2.20)
    - Look up 1.40 and 2.20 in your table;
    - .0808 - .0139 = .067, or 6.7% of grads make between \$68K and \$70K per year

Copyright Sharyn O'Halloran 2001

Joint Distributions

- Probability Tables
  - Example: Toss a coin 3 times.
    - How many heads and how many runs do we observe?
      - Def: A <u>run</u> is a sequence of one or more of the same event in a row
      - Possible outcomes look like:

Toss	Probability	Heads x	Runs y
TTT	1/8	0	1
TTH	1/8	1	2
THT	1/8	1	3
THH	1/8	2	2
HTT	1/8	1	2
HTH	1/8	2	3
HHT	1/8	2	2
HHH	1/8	3	1

Copyright Sharyn O'Halloran 2001

Joint Distributions (cont.)

- The joint probability of x and y is the probability that both x and y occur.
  - p(x,y) = Pr(X and Y)
- Joint Distribution Table

Heads x \ Runs y	1	2	3	
0	1/8	0	0	1/8
1	0	1/4 (2/8)	1/8	3/8
2	0	1/4 (2/8)	1/8	3/8
3	1/8	0	0	1/8
	1/4	1/2	1/4	1

  - p(0, 1) = 1/8,
  - p(1, 2) = 1/4,
  - p(3, 3) = 0.

Copyright Sharyn O'Halloran 2001

Joint Distributions
Marginal Probabilities (cont.)

- Marginal probability
  - The sum of the rows and columns.
    - The overall probability of an event occurring.

$$p(x) = \sum_y p(x,y).$$

  - The probability of just 1 head is the probability of 1 head and 1 runs + 1 head and 2 runs + 1 head and 3 runs
    = 0 + 1/4 + 1/8 = 3/8

Copyright Sharyn O'Halloran 2001

Joint Distributions (cont.)

- Independence
  - A and B are independent if P(A|B) = P(A).

$$\Rightarrow P(A|B) = \frac{P(A\&B)}{P(B)};$$

$$\Rightarrow P(A) = P(A|B),$$

$$\Rightarrow P(A,B) = P(A)P(B).$$

- The joint probability
  - is the product of the marginal probabilities of 2 independent events

Copyright Sharyn O'Halloran 2001

Joint Distributions (cont.)

- Are the number of heads and the number of runs independent?
  - If so, the table would look like this:

# heads x \ # Runs y	1	2	3	marg dist
0	1/32	1/16	1/32	1/8
1	3/32	3/16	3/32	3/8
2	3/32	3/16	3/32	3/8
3	1/32	1/16	1/32	1/8
marg dist	1/4	1/2	1/4	1

NO, because the # of heads observed is related to the number of runs

Copyright Sharyn O'Halloran 2001

Correlation and Covariance

- Definition of Covariance
  - The expected value of the product of the differences from the means.

$$s_{x,y} = E(X - m_x)(Y - m_Y)$$

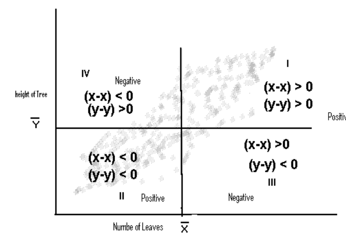
$$= \frac{\sum_{i=1}^{N}(X_i - m_x)(Y_i - m_Y)}{N}$$

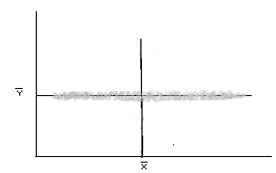
$$= \sum (X_i - m_x)(Y_i - m_Y)\,p(x,y).$$

Copyright Sharyn O'Halloran 2001

Correlation & Covariance (con't)

- Graph





Copyright Sharyn O'Halloran 2001

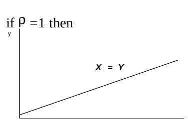
Correlation & Covariance (con't)

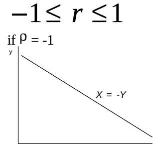
- Definition of Correlation

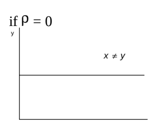
$$r = \frac{s_{x,y}}{s_x s_y} = \frac{\text{Covariance}}{\text{SD}_x * SD_y}$$

- Characteristics of Correlation

$$-1 \le r \le 1$$







Copyright Sharyn O'Halloran 2001

Correlation & Covariance (con't)

- Why? $\quad -1 \le r \le 1$

$$r = \frac{s_{x,y}}{s_x s_y} = \frac{\frac{\sum (x_i - m_x)(y_i - m_y)}{N}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - m_x)^2}{N}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - m_y)^2}{N}}}$$

Copyright Sharyn O'Halloran 2001