Quantitative Skills Workshop

Statistics

Basic Concepts

Frequency Distribution

⇨Definition: a way to display the frequency of measurements (observations) by dividing them into class intervals

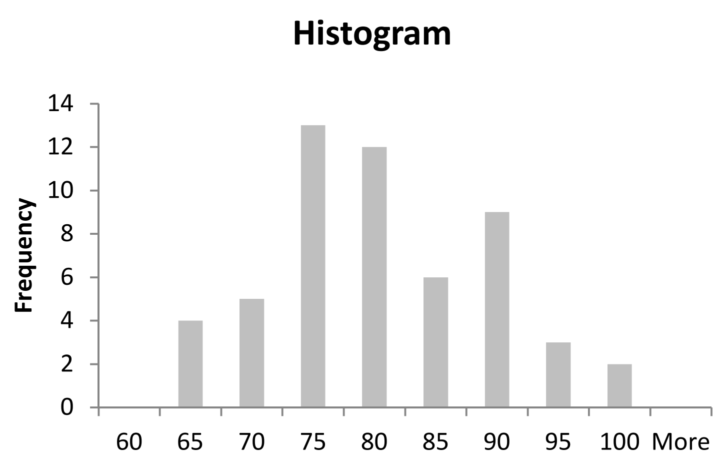
Example: Frequency distribution of test scores

Scores
86.53
84.85
78.55
73.92
71.03
78.57
75.07
96.43
89.00
88.31
71.78
73.15
81.65
70.82
75.47
79.38
87.97
86.62
69.98
70.73
91.63
78.63
88.42
99.02
73.84
67.98
69.16
88.76
78.51
70.82
79.35
81.35
69.59
89.43
92.47
61.89
71.35
76.80
73.61
80.69
73.92
71.45
75.18
66.30
75.95
77.07
70.20
91.75
64.52
61.22
62.97
89.02
83.94
82.50

Frequency Distribution

Classes	Observations	Frequency	% Frequency
Scores Between 60-65	4	4/54	7.41%
Scores Between 65-70	5	5/54	9.26%
Scores Between 70-75	13	13/54	24.07%
Scores Between 75-80	12	12/54	22.22%
Scores Between 85-90	6	6/54	11.11%
Scores Between 90-95	9	9/54	16.67%
Scores Between 95-100	3	3/54	5.56%
Scores Between 100-105	2	2/54	3.70%
Total Observations:	54		

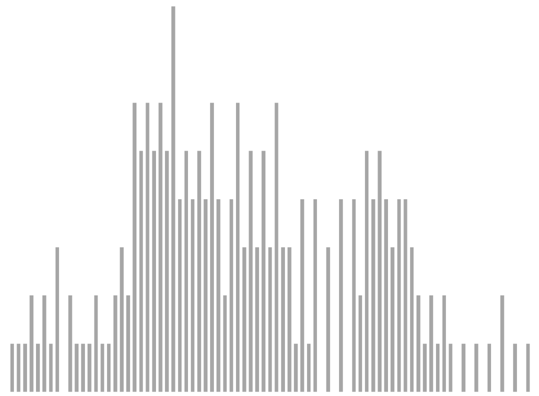
Histogram: a graph of a frequency distribution



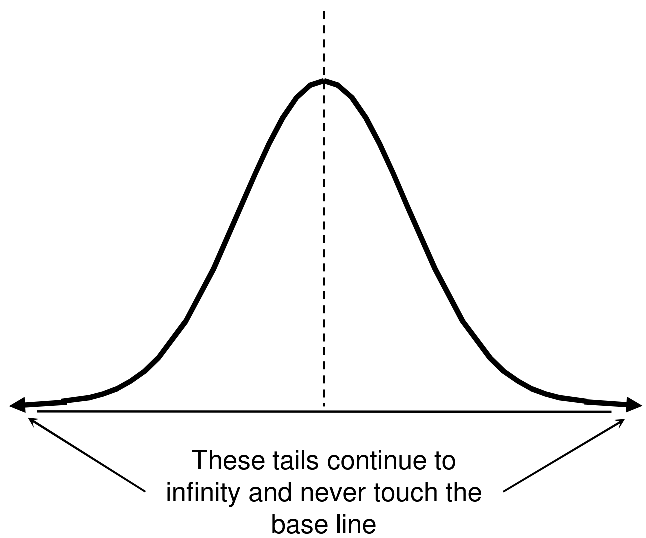
Basic Concepts

Histograms (continued)

If there are a large number of measurements, the histogram looks more like this:



A histogram with an infinite number of random observations that are balance around the mean would look like this:

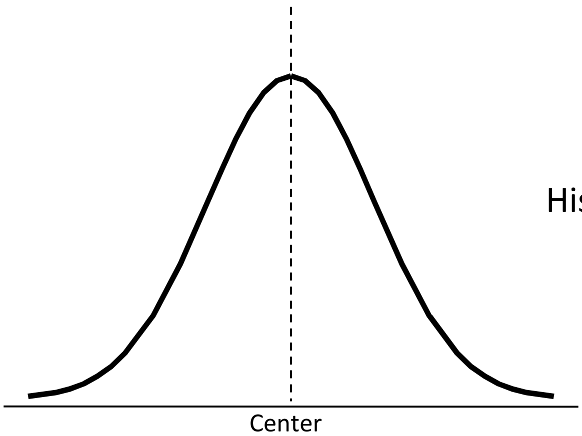


Basic Concepts

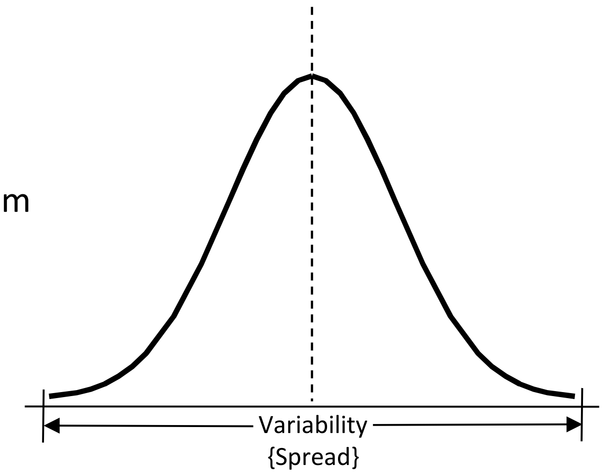
Numerical Descriptive Measures of Historical Data

⇨Central Tendency
⇨Variability/Dispersion

Visual Representation:



Histogram



Premise:

1} Examine a sample of a population
2} Describe the sample (central tendency and variability)
3} Based on the description of the sample, infer something about the population

Population Size: Symbol N
Sample Size: Symbol n
Reliability of Inference: as n approaches N (as n/N approaches 1), the more reliable the inference

Basic Concepts

Mean of a Sample:

⇨Definition: The sum of measurements (observations) divided by the number of measurements in a data set

⇨The average value

⇨The most probable future value (more on this later)

⇨Symbol: $\bar{x}$ or x-bar

⇨Formula: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

⇨**The mean is used to describe the central tendency of a sample**

Example B1: Find the mean of these measurements: 4,7,2,3,8

x-bar = (4 + 7 + 2 + 3 + 8) / 5 = 4.8

Interpretation:

1} The average value of this data is 4.8

2} The next event {outcome, occurrence} is likely to be 4.8

Data Set:

⇨Definition: the collection of data points (measurements or observations) in a sample

⇨In the example above, the data set is: {4,7,2,3,8}

Excel Example B1:

Basic Concepts

Mean of a Population:
⇨Definition: The sum of measurements divided by the number of measurements in an entire population
⇨Symbol: μ (mu)
⇨x-bar is used to infer μ

→it is often impractical if not impossible to compute the mean of an entire population
→it is relatively easy to compute x-bar of a sample from a population
→x-bar $\neq \mu$, x-bar is only used to estimate μ
→the larger the sample, the more accurate x-bar will be as an estimate of μ

Median of a Sample:
⇨Definition: The middle number of a set of measurements when the data is arranged in ascending order.
⇨Symbol: m
⇨If n (the total number of measurement) is an odd number, m is the middle number of the set
⇨If n is an even number, m is the average of the middle two numbers

Why care about the median?
⇨m is some times a better measure of central tendency than x-bar. This is especially true when some measurements in a data set are extremely small and some are very large

Basic Concepts

Median of a Sample: {continued}

Example: Find the median of these measurements: 4,7,2,3,8

Sort the data in ascending order: 2, 3, 4, 7, 8; m = 4

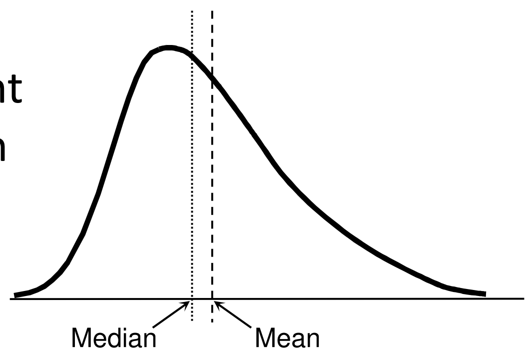
Skewness:

⇨A data set is said to be skewed if one tail of the distribution has more extreme observations than the other

⇨Comparing the median to the mean reveals whether the data set is skewed

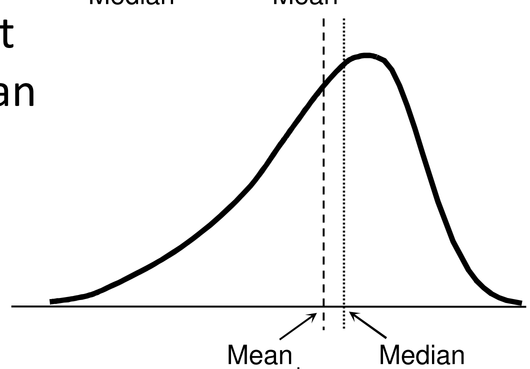
→Skewed to the right

Mean > Median



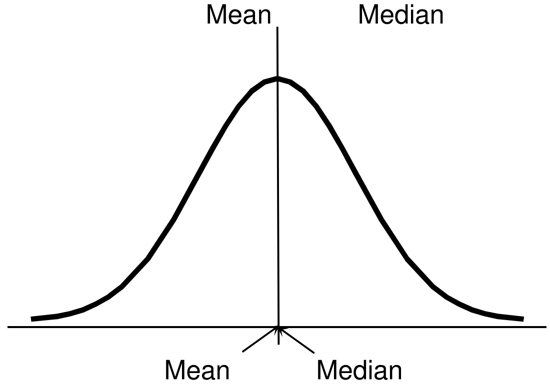
→Skewed to the left

Mean < Median



→Not Skewed

Mean = Median



Point: When data is skewed, the median may be a more accurate descriptor of the central tendency than the mean

Basic Concepts

Mode: The measurement that occurs most frequently in the data set

Range of a Sample: The largest measurement minus the smallest measurement
⇨Easy to compute
⇨Not much use in expressing variation; two data set can have the same range but vastly different variance

Variability of a Sample:
⇨Indicates how much the observations vary from the mean
⇨Used to express relative uncertainty concerning the next event or outcome (more on this later)
⇨Symbol: s^2 (sigma squared)

⇨Formula: Variance =

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

Variability of a Population:

⇨Formula:

$$\sigma^2 = \sum_{i=1}^{N} (X_i - \mu)^2 f(X_i)$$

Point: $s^2 \neq \sigma^2$. There's a slight but important difference between the variance of a sample and the variance of a population. You must use the correct formula for the appropriate situation to avoid inaccuracies.

Basic Concepts

Standard Deviation of a Sample:

⇨Symbol: s (sigma)

⇨Standard Deviation = Square Root of Variance

⇨Formula: $s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

⇨Unlike variance, standard deviation is expressed in the original units of the data being observed

Standard Deviation of a Population:

⇨Formula: σ = square root of σ^2

⇨s ≠ σ

⇨The difference between s and σ is small and therefore they are often used interchangeably; this should be done only if the difference is negligible.

Example B2 : the data set is: {4,7,2,3,8}

Excel Example B2

Probability of Discrete Random Variables

Variables: the observations/data points

Discrete: This means that the observations/data points are specific, separate and distinct and from a population that has a finite number of elements.

Random: the data points are the result of random sampling and there were no biases in the selection process

Random Sampling: if *n* elements are selected from a population in such a way that every set of *n* in a population has an equal probability of being selected, the *n* elements are said to be a random sample. (This is the definition of a simple random sample which is the most common technique)

Examples:

⇨Examine every 50th invoice produced by a sales staff in a year

⇨Select for inspection every tenth machined part coming off an assembly line

⇨Choose every 100th name in a telephone book to participate in a survey

⇨Select the closing stock price at the end of every week over the last 52 weeks to determine the average stock price over a year

Probability of Discrete Random Variables

Premise:
⇨Conduct an experiment a large number of times
⇨Arrange the observations in a relative frequency distribution (graphically, this would be a histogram)
⇨This relative frequency distribution is theoretical model of the entire population, ***past, present and future***
⇨The frequency distribution of observations that have already occurred is, in theory, a probability distribution of future events
⇨Thus past performance infers future performance

Experiment: is an act or process of observation that leads to a single outcome that cannot be predicted with certainty.
Sample Point: the outcome of a single experiment.
Examples:
⇨flipping a coin
⇨rolling two 6-sided dice
⇨checking the dimensions of a machined part
⇨examining the monthly closing price of a particular stock
Sample Space of an Experiment: is the collection of all the sample points
⇨Sample Space Size: Symbol: n
Examples:
⇨flipping a coin 10 times; {H, H, T, H, T, T, H, T, H, T}; n = 10
⇨rolling two 6-sided dice 50 times; {6, 11, 5, 7, etc.}; n = 50
⇨checking the dimensions of 500 machine parts; {5.4567”, 5.4582”, 5.4548”, etc.}; n =500
⇨examining the monthly closing prices of a particular stock for the last five years; {$45.67, $51.25, $48.76, etc.}; n = 60

Probability of Discrete Random Variables

Probability of a sample point occurring:
⇨Symbol: p
⇨Must be between 0 and 1 (i.e. $0 \leq p \leq 1$ or $0\% \leq p \leq 100\%$)
⇨For a sample space, the probabilities of each sample point must add up to 1 (or 100%)

Event:
1) a specific collection of sample points
2) a specific outcome

Probability of an Event:
⇨The probability (chance) of an event occurring
⇨Designated as P(x)
⇨Determined by summing the probabilities of the sample points in the sample space
⇨Theory: the frequency of an event occurring in the past is the probability that it will occur in the future

Example : Consider a 6-sided die. One roll of this die has six possible outcomes: 1, 2, 3, 4, 5 or 6 thus the sample space is {1, 2, 3, 4, 5, 6}. If the die is balanced and there are no external influences on the roll of the die, then there is a 1 in 6 chance of any one particular sample point occurring (i.e. p = 1/6). The "event" is the next roll of the die and the outcome will be one element of the sample space (i.e. 1 through 6) with a 1/6 probability of any of these elements occurring.

Example : What is the probability of rolling a 5? (i.e. P(5) = ?). The sample space is: {5}. The probability of rolling the die with an outcome of 5 is p(5) which is equal to 1/6. Thus P(5) = p(5) = 1/6 = 16.666666%

Probability of Discrete Random Variables

Example: Again consider a 6-sided die. What is the probability of rolling an even number. The event is rolling the die in which the outcome is an even number, designated as P(even number). The sample space is: {2, 4, 6}. The probability of the event is determined by summing the probabilities of each of the sample points: P(even number) = p(2) + p(4) + p(6) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2 = 50%. Thus P(even number) = 50%

Determining the Probability of an event (Process Summary)

⇨Define the experiment
⇨List the sample points
⇨Assign probabilities to the sample points
⇨Determine the collection of sample points contained in the event of interest (i.e. define a sample space)
⇨Sum the sample point probabilities to determine the event probability

Probability of Discrete Random Variables

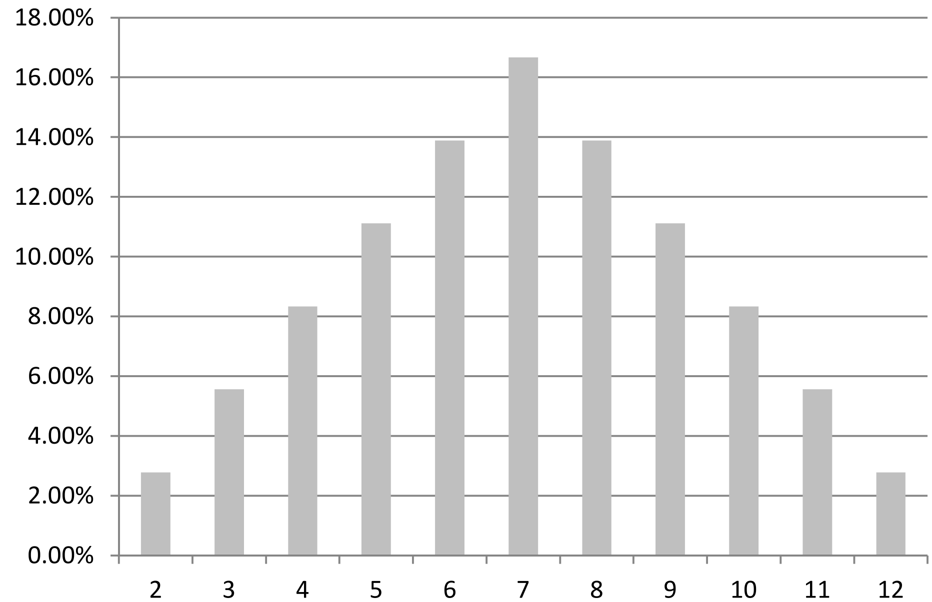
Probability Distribution: the possible values of outcomes associated with the probability of their occurrence

Example: Probability distribution for the role of two 6-sided dice

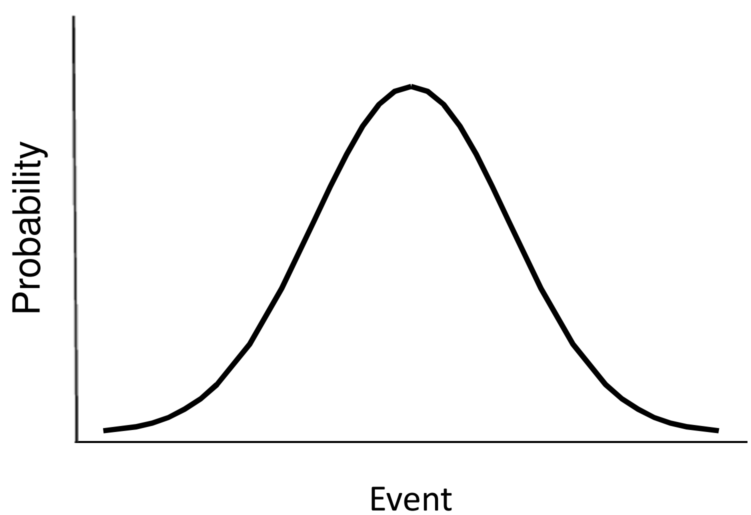
Chart Format

Event	Probability (P)
2	2.78%
3	5.56%
4	8.33%
5	11.11%
6	13.89%
7	16.67%
8	13.89%
9	11.11%
10	8.33%
11	5.56%
12	2.78%
Sum:	100.00%

Graph



If there are a large number of discrete random events, the probability distribution looks more like this:



Probability of Discrete Random Variables

Determining the Expected Value:

⇨$E(x) = \Sigma xP(x)$ = the sum of every event times the probability of that outcome

⇨Symbols: $E(x)$, $\hat{x}$

⇨This is an estimate of the value of a future event

Example: Determine the expected value of a bushel of wheat grown in Kansas

Environmental Conditions	P	Price	P x Price	
Extreme Drought	3.4%	$3.08	$0.10	
Moderate Drought	21.4%	$2.78	$0.59	
Normal Rainfall	42.0%	$2.56	$1.08	
Above Normal Rainfall	23.5%	$2.13	$0.50	
Excessive Rainfall	9.7%	$2.26	$0.22	
			$2.49	**E(Price of Wheat)**

Example: Determine the expected value of the roll of two 6-sided dice

Roll	P(x)	Roll x P(X)	
2	2.78%	0.06	
3	5.56%	0.17	
4	8.33%	0.33	
5	11.11%	0.56	
6	13.89%	0.83	
7	16.67%	1.17	
8	13.89%	1.11	
9	11.11%	1.00	
10	8.33%	0.83	
11	5.56%	0.61	
12	2.78%	0.33	
	Sum	**7**	**E(Roll)**

Probability of Discrete Random Variables

Example: Ten newly manufactured bolts are randomly selected from an assembly line for inspection. The length of each of the bolts is measured. What would you expect to be the length of the next bolt selected from the assembly line?

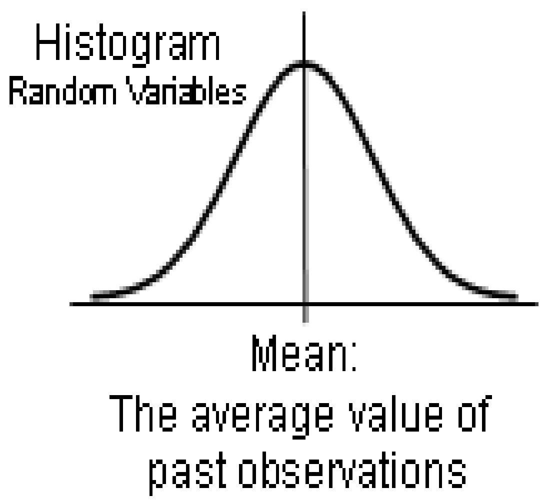
Bolt	P(x)	Length (mm)	x times P(x)
1	10.00%	6.030	0.603
2	10.00%	6.010	0.601
3	10.00%	5.980	0.598
4	10.00%	5.970	0.597
5	10.00%	6.020	0.602
6	10.00%	6.010	0.601
7	10.00%	6.040	0.604
8	10.00%	5.990	0.599
9	10.00%	6.020	0.602
10	10.00%	6.030	0.603
			6.010

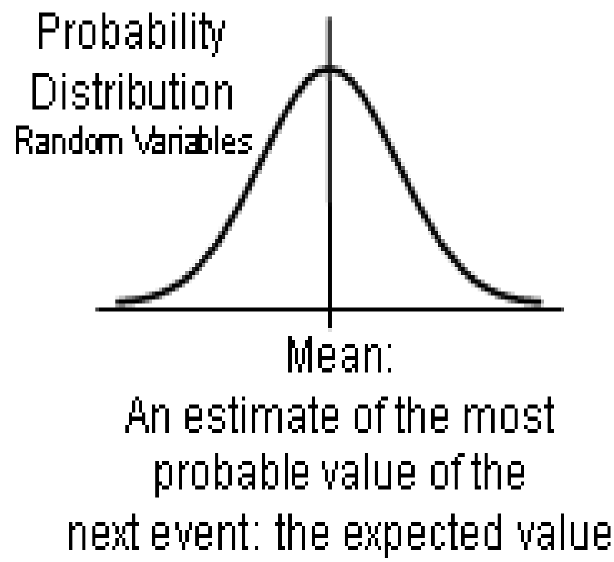
	Length (mm)
	6.030
	6.010
	5.980
	5.970
	6.020
	6.010
	6.040
	5.990
	6.020
	6.030
Sum:	60.100
Average:	**6.010**

Random vs. Non-Random Variables: In the above example the bolts were randomly selected and each event (length of a particular bolt) had the same probability. Thus the expected value can be determined by simple finding the average length.

Probability Distribution Graphs of Discrete Random Variables :

Graphically, it's the same thing as a histogram but the interpretation is different. Instead of showing the distribution of the frequencies of observations that occurred in the past, we interpret it as the distribution of probability of future events.





Probability of Discrete Random Variables

Excel Example P1:

Standard Deviation of Discrete Random Variables:

⇨Formula: $\sigma = \sqrt{\sum_{i=1}^{n}(x_i - \mu)^2 p(x)}$
⇨xbar is substituted for μ
⇨Note: we use σ and not s;   **Why?**

Example: Compute σ of the probability distribution of the bolt example from the previous page.

Excel Example P2:

Bolt	P(x)	Length (mm)	x- xbar	(x- xbar)^2	(x- xbar)^{2}p(x)
1	10.00%	6.030	0.0200	0.000400	0.000040
2	10.00%	6.010	0.0000	0.000000	0.000000
3	10.00%	5.980	-0.0300	0.000900	0.000090
4	10.00%	5.970	-0.0400	0.001600	0.000160
5	10.00%	6.020	0.0100	0.000100	0.000010
6	10.00%	6.010	0.0000	0.000000	0.000000
7	10.00%	6.040	0.0300	0.000900	0.000090
8	10.00%	5.990	-0.0200	0.000400	0.000040
9	10.00%	6.020	0.0100	0.000100	0.000010
10	10.00%	6.030	0.0200	0.000400	0.000040
				Variance(mm^2)	0.000480
				Std. Dev.(mm)	**0.021909**

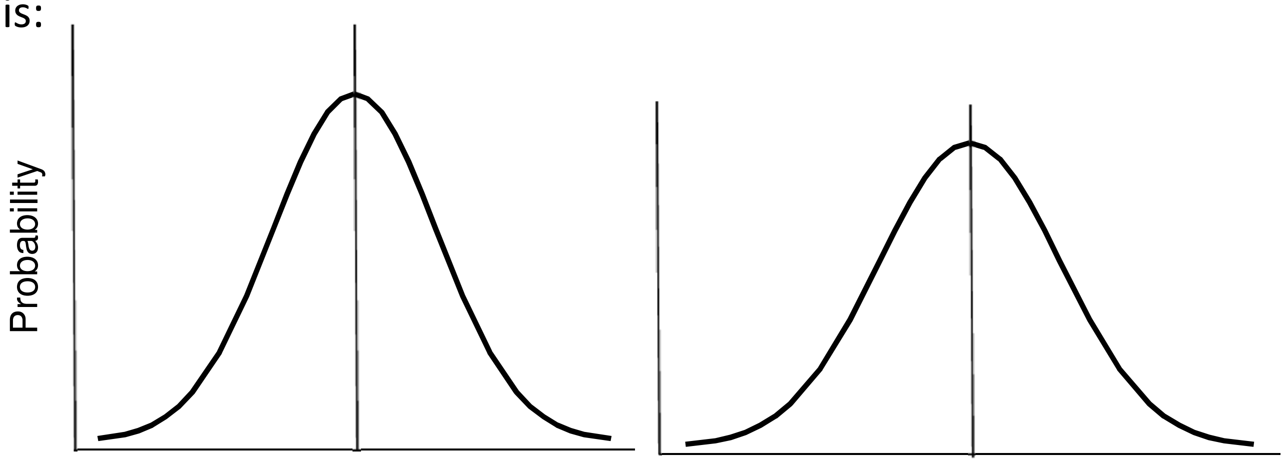
Probability of Discrete Random Variables

Interpreting Standard Deviation:

⇨σ represents the variability of an estimated expected value
⇨It provides some indication of the reliability of the estimate
⇨The smaller the variability, the more reliable the estimate

Example: Consider a company that manufactures fasteners (nuts & bolts, screws, etc.). Statistical analysis is conducted the diameter on 20mm length bolts and the hole diameter of the matching nuts. For the bolts, E(diameter) = xbar = 3.0215mm and s = 0.0073mm. For the nuts, E(diameter) = xbar = 3.0211mm and s = 0.0097mm. Which expected value is a more reliable predictor of the future bolt/nut diameters?

Answer: The expected value of bolt length is a more reliable predictor of future bolt lengths since the variability of diameter as expressed by s is smaller than that of the nuts. Graphically, the situation looks like this:



Bolts
E(diameter) = 3.0215mm
s = 0.0073mm

Nuts
E(diameter) = 3.0211mm
s = 0.0097mm

Probability of Discrete Random Variables

<u>Normal Distribution of Data</u>: A set of data is said to be normally distributed if:

⇨The data are randomly selected

⇨The value of each data point is independent of (not influenced by) any other data point

⇨The probability distribution curve is bell shaped

Why care?

⇨Computation of numerical descriptive measures (i.e. mean, σ, etc.) for normally distributed data is easy compared to doing the same for data that are not normally distributed

⇨Many situations can be approximated as normally distributed therefore the statistics can be computed easily

<u>Business Related Examples</u>:

→security prices

→machined parts dimensions

→responses from survey participants

→production output

None of these examples are normally distributed but they are close enough.

Bivariate Relationships

⇨Examining the relationship of two variables

⇨One variable is the “independent variable” and the other is the “dependent variable”

⇨The data are assumed to be normally distributed and the variables are random (simple case)

<u>Examples</u>:

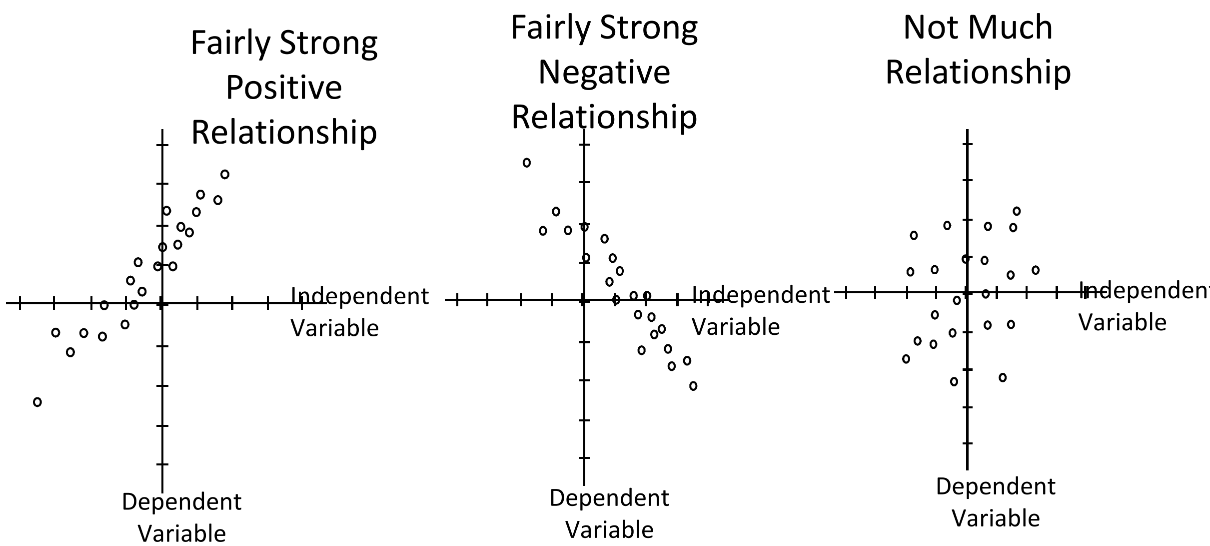
⇨How does the temperature (independent variable) of a plastic forming process affect the stiffness of plastic water bottles (dependent variable)?

⇨How does a consumer’s income level (independent variable) influence where he shops (dependent variable)?

⇨How does a consumer’s gender (independent variable) affect his/her choice of automobile (dependent variable)?

⇨Is there a relationship between the number of times a stock is traded during a single day (independent variable) and the stock’s closing price at the end of that day (dependent variable)?

<u>Depicting the Relationship</u>: If you make a graph that plots the dependent variables against the independent variables, you have a “Scatter Diagram” or “Scatter Plot”. Each point is an x-y coordinate.



Bivariate Relationships

Quantifying the the Relationship:

Covariance (of a Sample):

⇨Definition: the measure of how much two random variables (x, y) vary together (as distinct from variance, which measures how much a single variable varies).

⇨If two variables tend to vary together (that is, when one of them is above its expected value, then the other variable tends to be above *its* expected value too), then the covariance between the two variables will be positive.

⇨Symbol: cov(x,y) or $\sigma_{x,y}$

⇨Formula: covariance of a sample: $\sigma_{x,y}$ = $\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})/(N-1)$

Coefficient of Correlation (of a Sample):

⇨Definition: a measure of the strength of **the linear relationship** between two variables, *x* and *y*

⇨Usually, x is the independent variable and y is the dependent variable

⇨Symbol: r(x,y) or $r_{x,y}$; (for a population, the symbol is $\rho_{x,y}$)

⇨Formula: $r_{x,y}$ of a sample = $\sigma_{x,y} / (\sigma_x \sigma_y)$

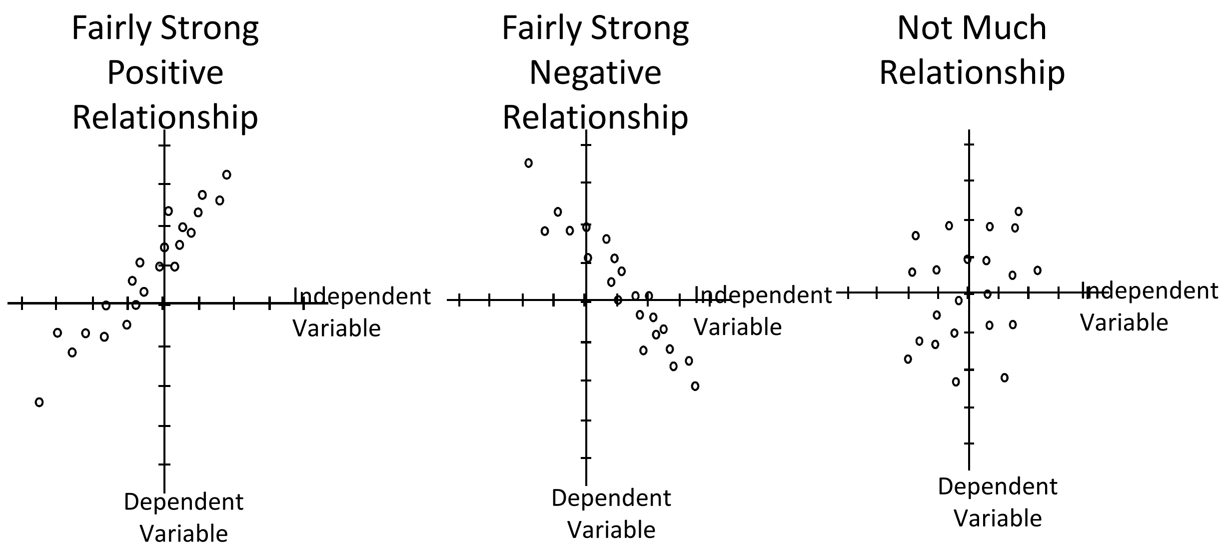
⇨$1 \geq r \geq -1$

- →If y=ax+b, a>0, then r =1; interpretation: the relationship is perfectly positive
- →If y=ax+b, a<0, then r = -1; interpretation: the relationship is perfectly negative

Bivariate Relationships

<u>Quantifying the the Relationship</u>: (continued)

<u>Example</u>: Coefficient of Correlation interpretation (Scatter plots from previous page)



$r_{x,y}$ is close to 1 $r_{x,y}$ is close to -1 $r_{x,y}$ is close to 0

Point: The Coefficient of Correlation can be used to assess the amount of diversification in a portfolio of stocks

Excel Example BV1: Find Covariance of a sample of stock returns

Excel Example BV2: Find Coefficient of Correlation of a sample of stock returns

Arithmetic Mean vs. Geometric Mean

<u>Arithmetic Mean</u>:

⇨The descriptive and probabilistic means that we have so far discussed are arithmetic means; they were computed by summing values

⇨The arithmetic mean is quite useful for a large variety of purposes

⇨However, the arithmetic mean provides less than fully accurate estimates for situations in which data is changing over time; for example: security prices and security returns

<u>Geometric Mean</u>:

⇨This method captures and accounts for the period-to-period increases or decreases of a value over time

⇨General Formula: $GM = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = [(x_1)(x_2)(x_3)\ldots(x_n)]^{1/n}$

⇨For financial returns: $[(1+k_1)(1+k_2)(1+k_3)\ldots(1+k_n)]^{1/n}$

⇨The multiplication of the data is a form of ***geometric progression***

<u>Example</u>: Consider an investment that has had the following annual RORs for the last 4 years. The profit realized each year is reinvested.
Annual RORs: {1.5%,1.3%,0.5%,1%}

a. Find the geometric mean of these returns:

$1+R = [(1+0.015)(1+0.013)(1+0.005)(1+0.01)]^{1/4}$

$R = (1.04367)^{1/4} - 1$

$= 1.01074 – 1 = 0.01074 = 1.074\%$

b. Compute the Arithmetic Mean:

x-bar = (1.5% + 1.3% + 0.5% + 1%) / 4 = 4.3% / 4 = 1.075%

Excel Example (Geometric Mean) :

Probability Equations (Formal Expression):

Probability Distribution: $\int_a^b f(y)dy \rightarrow f(y) = F'(y)$

Mean: $E(Y) = \int yf(y)dy$

Variance: $V(Y) = E([Y - E(Y)]^2)$

Normal Distribution Probability Density Function:

$$\Phi(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{\frac{(y-\mu)^2}{2\sigma^2}}$$

Some Formula

- X and Y are random variables; a and b are constants
- $E(aX+b) = aE(X)+b$
- $E(X+Y) = E(X)+E(Y)$
- $Var(X) = E(X^2)-[E(X)]^2$
- $Var(aX+b) = a^2Var(X)$
- $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$
- $Cov(X,Y) = E(XY)-E(X)E(Y)$
- $Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y)$
- $Cov(aX,bY) = abCov(X,Y)$
- $Cov(X+a,Y+b) = Cov(X,Y)$
- $Cov(X, aX+b) = aVar(X)$
- If $\sigma_{XY} = Cov(X,Y)$ $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$