Statistical Thinking

Vincent Vezza

ASQRS Quality Conference September 25, 2019

Descriptive Statistics

Numerical descriptive measures create a mental picture of a set of data. These measures which are calculated from a sample are numerical descriptive measures called statistics.

Statistical Inference

Statistical Inference is the process of using data analysis to deduce properties of a population for example by testing hypotheses and deriving estimates.

Difference between Descriptive and Inferential statistics

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population. While statistical inference is the process of using data analysis to deduce properties of an underlying population.

Difference between a parameter and a statistic

- Parameter: The true numeric population value.

- Statistic: A numerical data value taken from a sample that may be used to make inference about a population.

Measures of central Tendency

- Mean
- Mode
- Median (Midpoint)

The Mean

The Mean (x-bar or $\bar{x}$)

Formula: $\bar{x} = \frac{\sum x}{n}$

$\bar{x}$ is the mean

$\sum$ sum the x's

x represent the individual numbers

n is the sample size

Advantages of Using the mean

- The mean is the center of gravity
- The mean uses all the data
- There is no sorting when calculating the mean

Disadvantages of using the mean

- Extreme data values may distort the mean
- It can be time consuming
- The mean may not be the actual value of any data points

The Mode

The mode is the most frequently occurring number in a data set

Example: 7 5 9 11 10 7 6 7 10

The mode is 7

Advantages of using the mode

- No sorting the data
- It is not influenced by extreme values
- It is an actual value
- It can be detected visually in distribution plots

Disadvantages of using the mode

- The data may not have a mode
- It may have more than one mode

The Median

The median is the middle value when the data is arranged in ascending or descending order. For an even set of data, the median is the average of the middle two values.

Median of the two data sets

- 10 numbers: 4 4 4 5 6 8 9 9 10 11
- 9 numbers: 4 4 5 6 7 9 10 10 11

The median for both sets is 7

Advantages of the Median

- Idea of where most data is located
- Simple calculation needed
- Little sensitivity to extreme values

Disadvantages of the median

- The data has to be sorted
- Extreme values may be important
- Two medians cannot be averaged to obtain a combined distribution
- The median will have more variation between samples than the average

Measure of Dispersion

- Range (R)
- Variance (σ^2, s^2)
- Standard Deviation (σ, s)
- Coefficient of Variation

Range

The range is the difference between the largest and smallest number.

Example: 15 13 17 19 18 15 14 15 18

range 19 – 13 = 6

Variance

The variance is the measure of dispersion.

Variance (σ^2, s^2)

$$\text{Population, } \sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

$$\text{Sample, } s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Standard Deviation

The standard deviation is the square root of the variance.

$$\text{Population, } \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

$$\text{Sample, } s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Coefficient of Variation COV

The coefficient of variation equals the standard deviation divided by the mean and is expressed as a percentage.

$$\text{COV} = \frac{\sigma}{\mu}(100)\%$$

$$COV = \frac{s}{x-bar}(100)\%$$

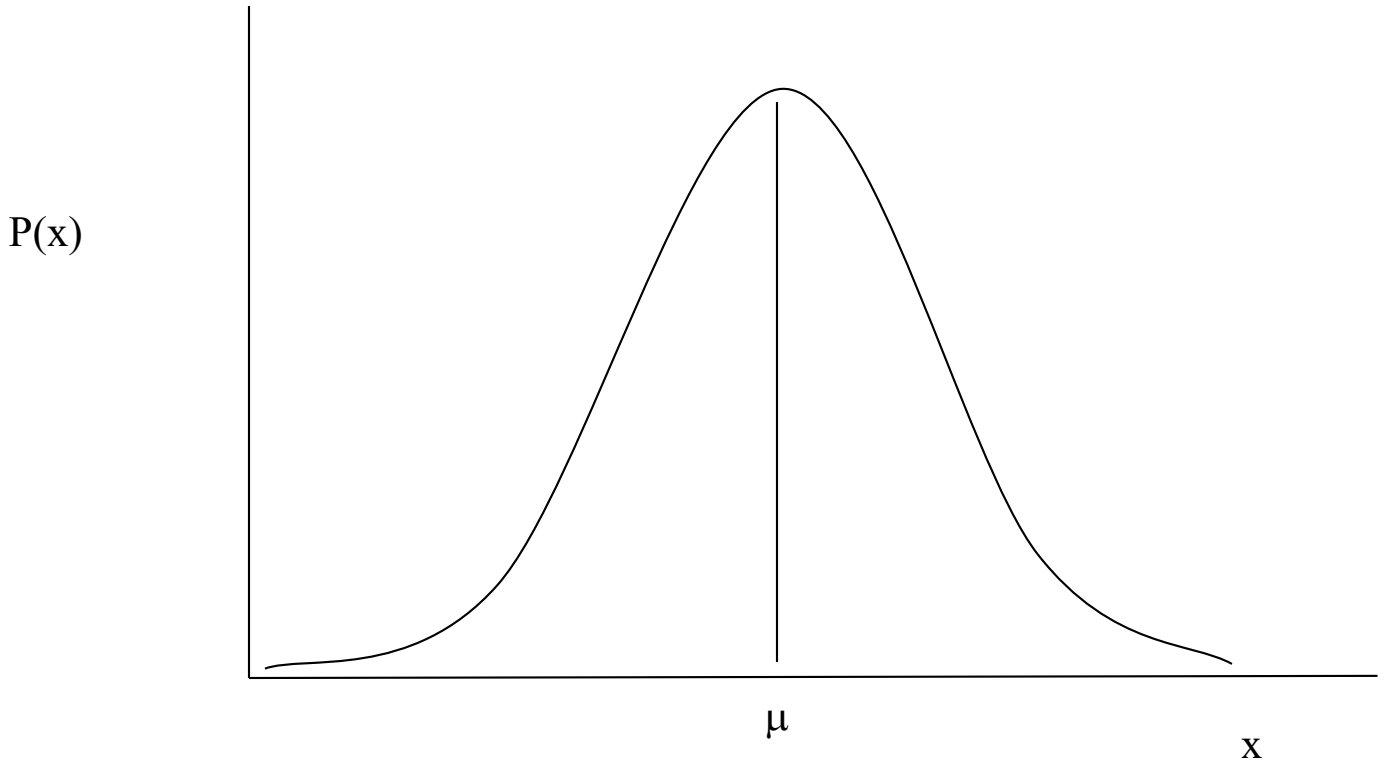
Normal distribution

- Symmetric bell-shaped curve
- μ locates the middle of the distribution
- σ describes how spread out the distribution is.

Equation for Normal Distribution

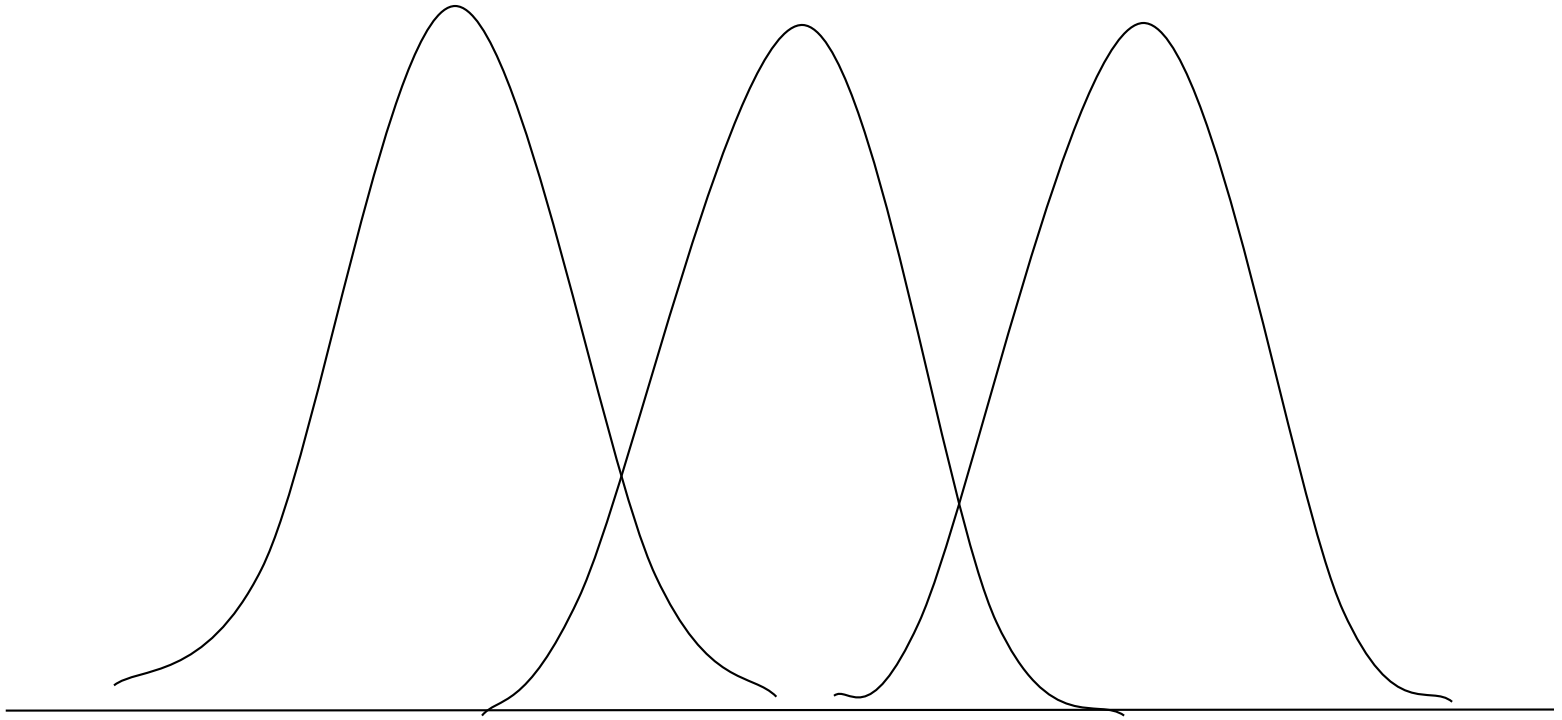
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Normal Probability Distribution



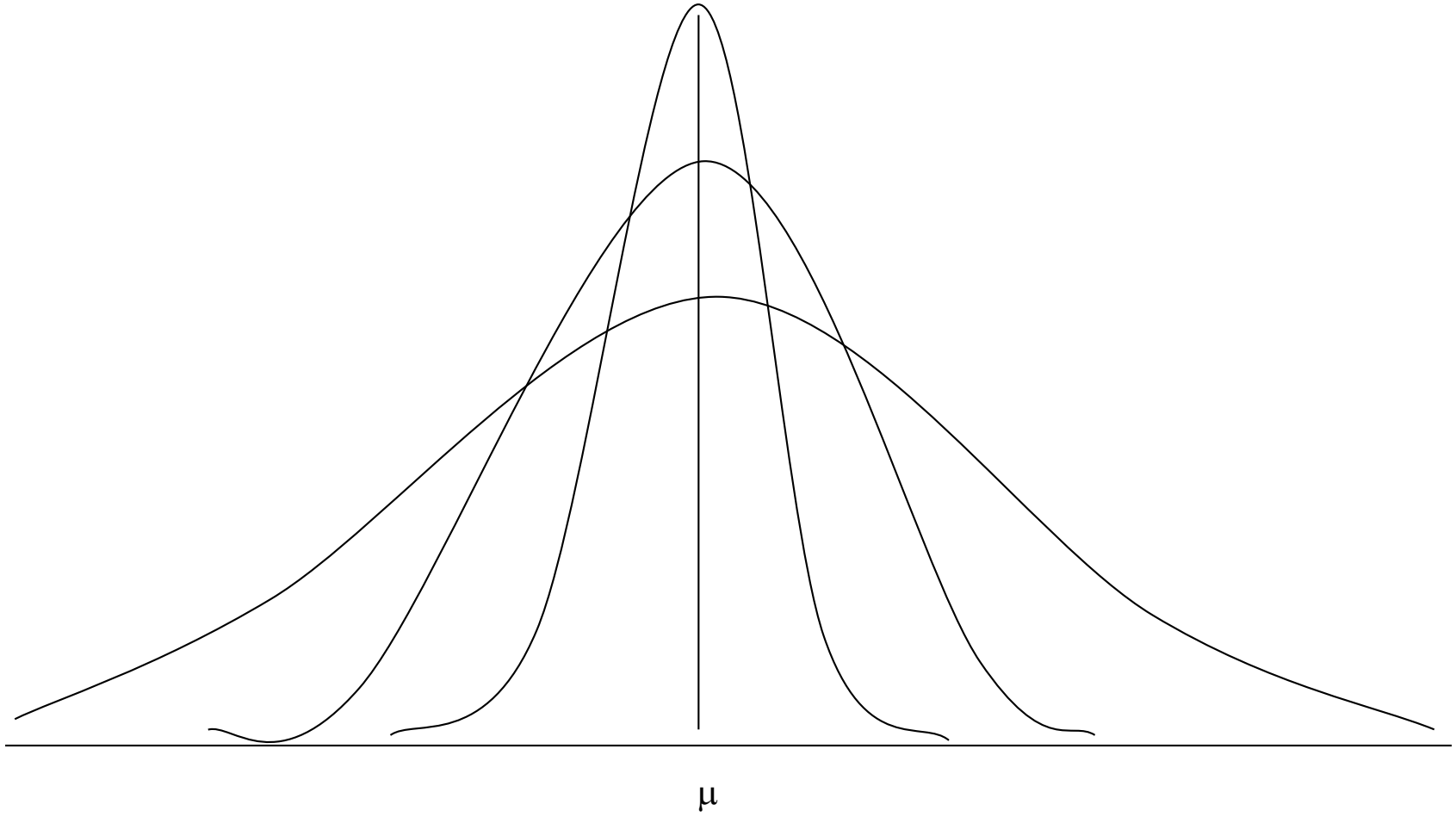
Normal Distributions

- same standard deviations
- Different means

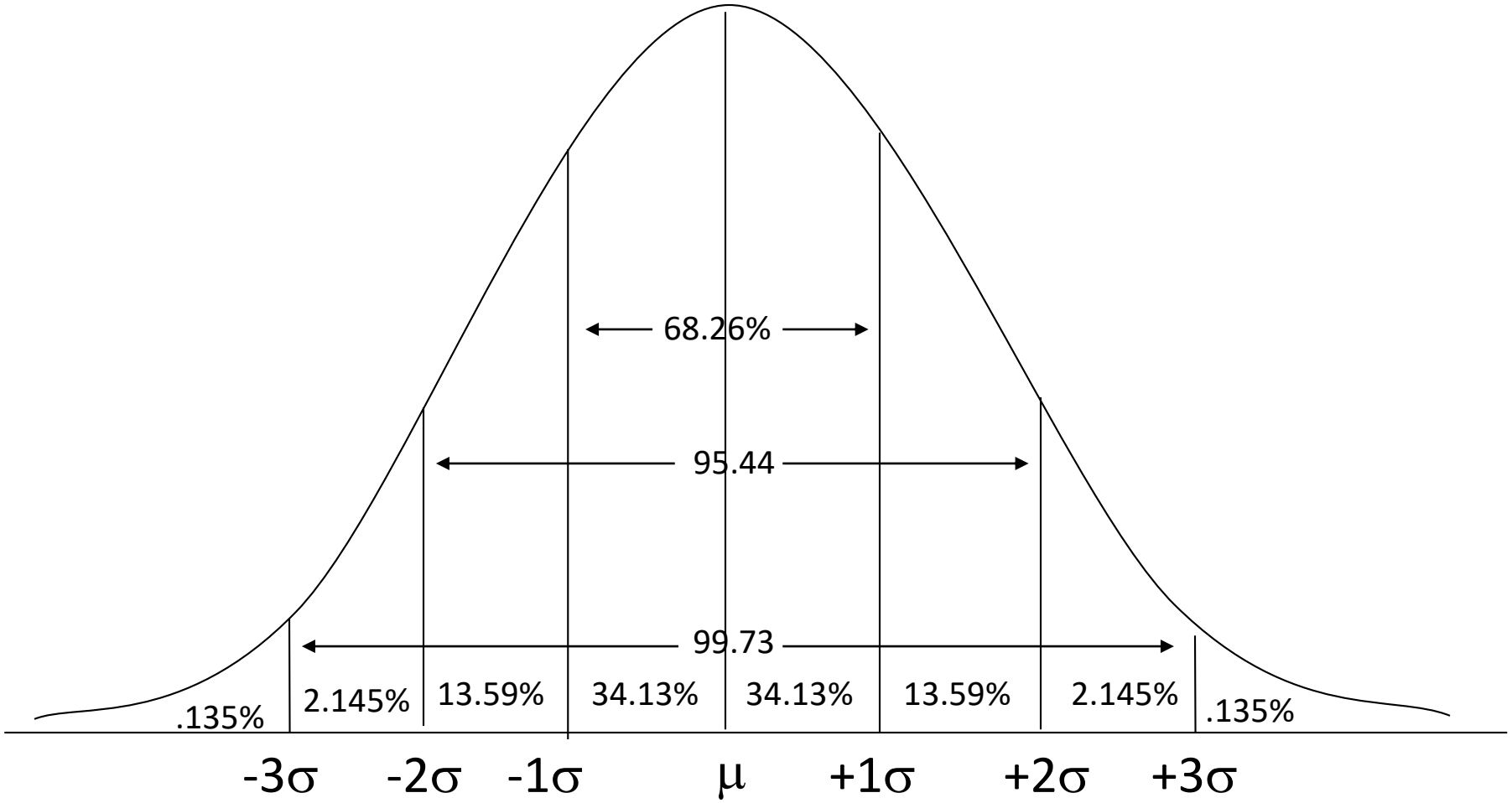


Normal Distributions

- different standard deviations
- Same mean



Percentage of the normal Distribution
(Area under the normal curve)



How to Understand the Normal Curve in a Practical Way

Whenever things of the same kind are measured, a large group of the measurements will tend to cluster around the middle. Most measurements fall close to the middle. In fact, mathematicians can make a fairly accurate prediction of the percentage of measurements in various section sections of the frequency distribution curve. This prediction is shown in the previous slide.

How to Understand the Curve in a Practical Way continued

- If we measure each piece that comes from a machine or operation and make a tally of the measurements, we will eventually have a curve similar to the on in slide 51.
- If we don’t measure each peace, but merely reach into a tote pan, grab a handful pieces, and measure them, the chance are that 68 out of a 100 measurements (34% + 34%) will fall within the two middle sections.
- The chances are 28 out of 100 pieces(14% +14%) will fall into the next two section, one on each side of the middle sections.
- Finally, the chances are that 4 out of 100 of the pieces (2%+2%) will fall into the two outside sections.

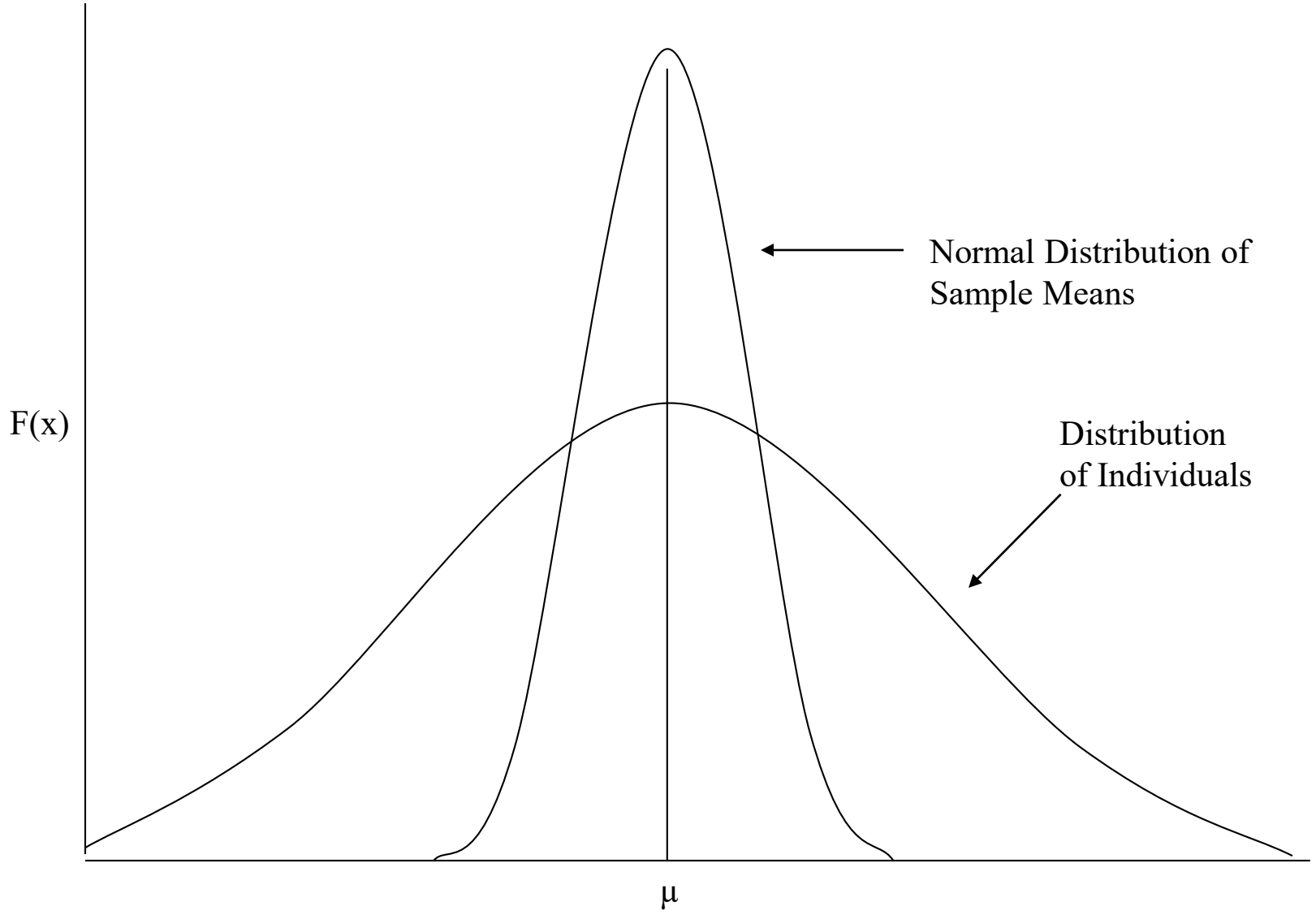
Central Limit Theorem

If a random variable X has a mean μ and finite variance σ^2, as n increases, x-bar approaches a normal distribution with mean μ and variance $\sigma^2_{\text{x-bar}}$. Where:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

and n is the number of observations on which each mean is based.

Distribution of Individuals Vs. Means



Equation for Standard Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

What is Analysis of Variance (ANOVA)

The essence of the technique is where a particular subject of interest is influenced by several different factors, it permits statistical data on the subject be broken down and the influence of each factor to be assessed. The technique does so by measuring the variation (variance) in a set of data and then partitioning, or subdividing, the variance into separate parts, each measures the contribution of a particular factor to the overall variation.

ANOVA

H_o : $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_A : At least two of the means are not equal.

ANOVA table for Three-Factor (k) Factorial, Fixed Effects Model

Source	**Sum of Squares**	**df**	**Mean Square**	**F calculated**
A Treatment	SS_A	a-1	$MS_A = SS_A$ /a-1	$F_0 = MS_A / MS_E$
B Treatment	SS_B	b-1	$MS_B = SS_B$ /b-1	$F_0 = MS_B / MS_E$
C Treatment	SS_C	c-1	$MS_C = SS_C$ /c-1	$F_0 = MS_C / MS_E$
Error	SS_E	$2^k(n-1)$	$MS_E = SS_E / 2^k(n-1)$	
Total	SS_T	$n2^k - 1$		

ANOVA Table

Printed from Minitab©

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	3	1112.50	370.83	7.24	0.043
Linear	3	1112.50	370.83	7.24	0.043
A	1	1058.00	1058.00	20.64	0.010
B	1	50.00	50.00	0.98	0.379
C	1	4.50	4.50	0.09	0.782
Error	4	205.00	51.25		
Total	7	1317.50			

F Test

The F-distribution is applied to the analysis of variance. The F-distribution is called the variance ratio distribution. It was developed by Sir Ronald Fisher who developed the technique for application to agriculture experiments. The F-test compares the variance among treatment means versus the variance of individuals within the specific treatment. High values of F indicates that one or more means are different.

Presentation of Data

- The average human brain is not good at comparing more than a few numbers at a time. Therefore a large amount of data is often difficult to analyze unless it is presented in some easily digested format.

Presentation of Data continued

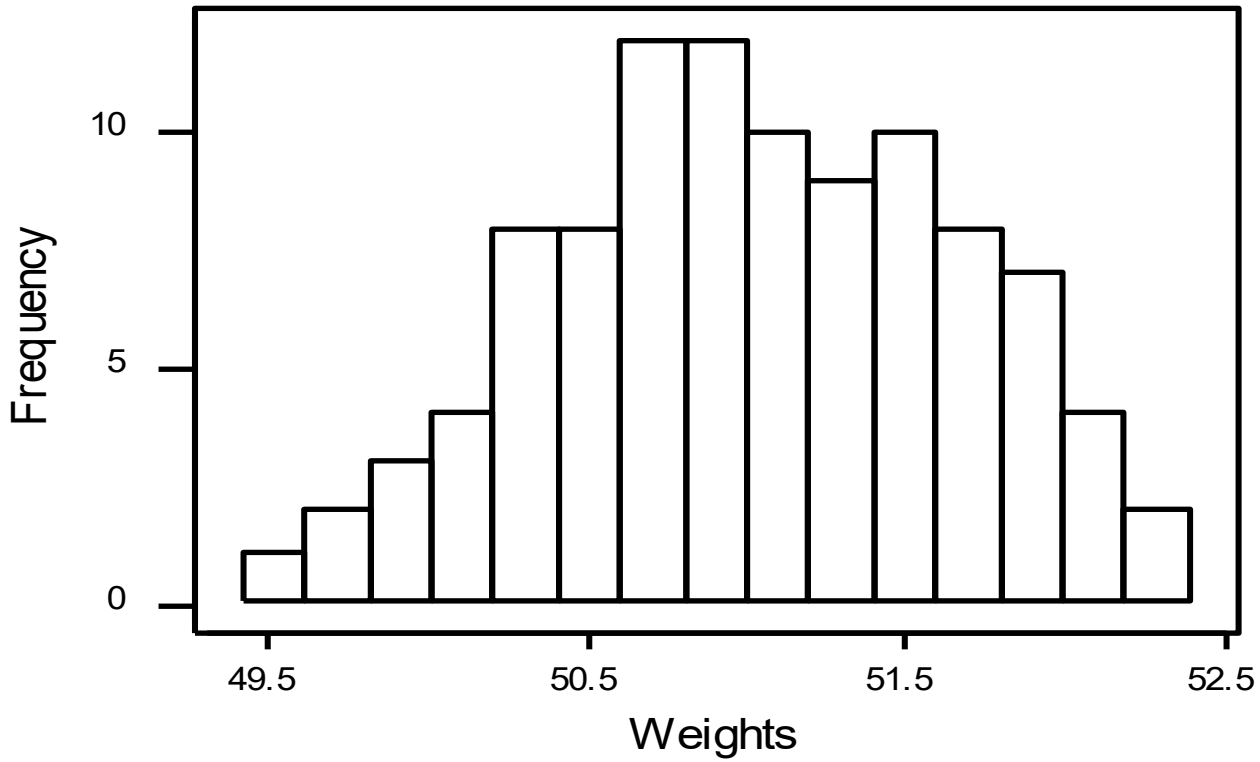
Tools used for the presentation of data.

- Histograms
- Pareto diagram
- Scatter diagram

Histograms

A histogram is a plot of data that has possible values on one axis and frequencies for those values on the other axis. The frequency is traditionally placed on the vertical axis, so that the observations accumulate in a pile at each value.

Histogram



Steps for Histogram

- Count the data points.
- Calculate the total range for the data. Range equal the highest value minus lowest value.
- Select the number of columns (cell intervals) for the histogram. As a rule of thumb, the number of columns should approximate the square root of the number of data points.

Steps in Histogram continued

- Determine class width. Class width equal to the Range divided by the number of classes. Round up.
- Choose the lowest class boundary. Start with a value that is just below the smallest observation and contains one additional significant figure.

Steps in Histogram continued

- Calculate the class boundaries by adding the class width starting at the lowest class boundary until all the data is contained.
- Construct a tally sheet that contains the class boundaries. Complete the histogram.

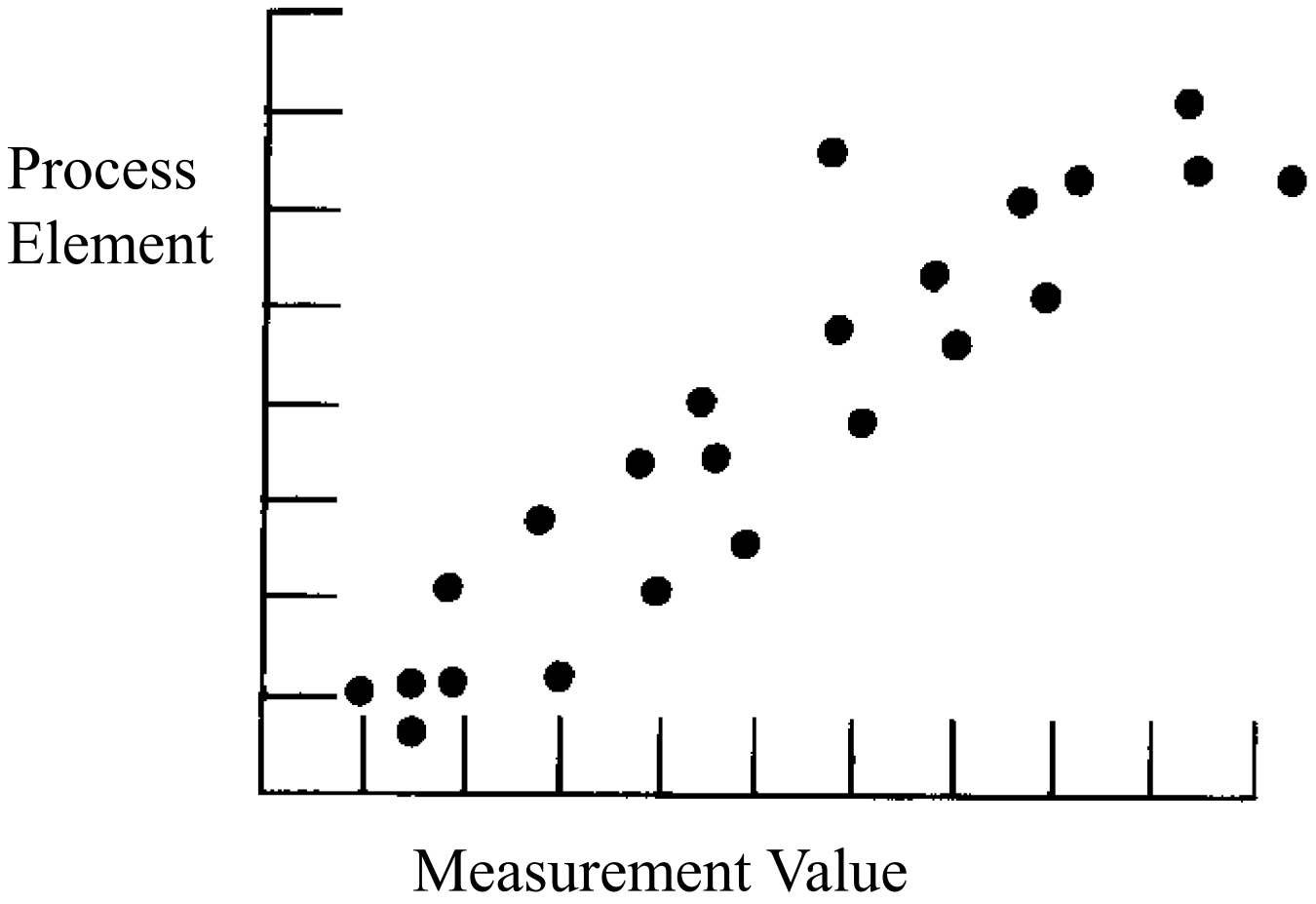
Alternative

N	K
31-50	5 – 7
51- 100	6 - 10
101 – 250	7 – 12
Over 250	10 - 20

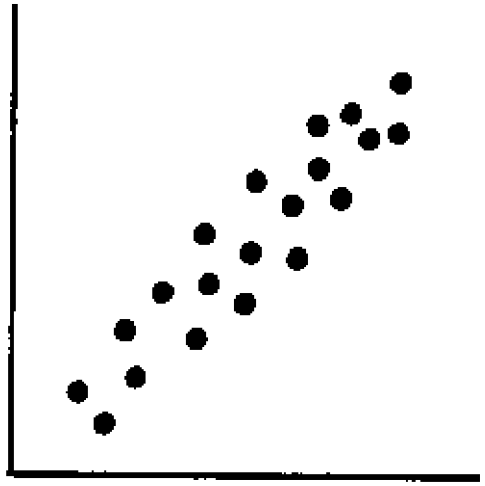
Scatter Diagrams

A scatter diagram is constructed by plotting pairs of sample observation on a two-dimensional plot. If for example, it is known or suspected that two variables move together, a scatter diagram can be used to illustrate that fact. For example peoples height vs their weight.

Scatter Diagram continued

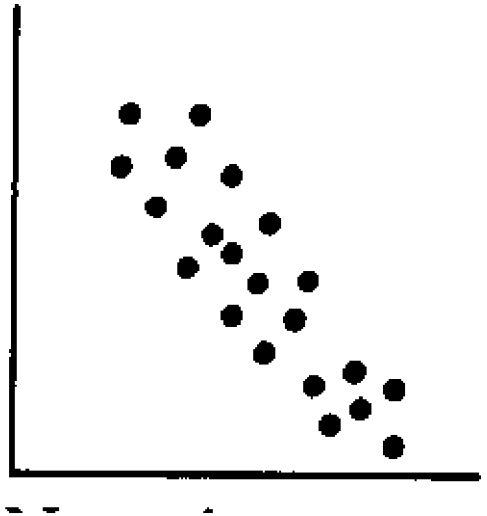


Scatter Diagram Continued



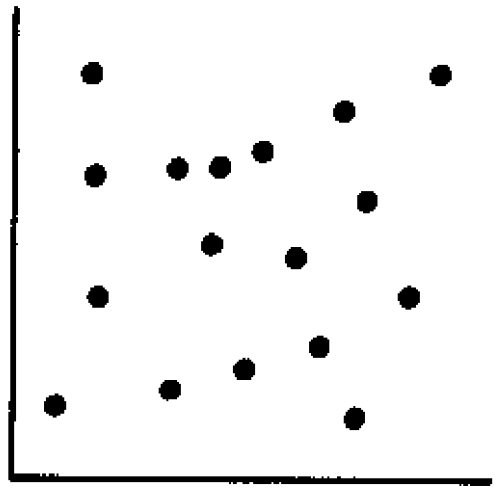
Positive Correlation

Scatter Diagram continued



Negative Correlation

Scatter Diagram Continued



No Correlation

Scatter Diagram Continued

Non-linear Correlation

Hypothesis Testing for the Means and Variance

$H_o : \mu = \mu_o$
$H_a : \mu \neq \mu_o$

$H_o : \mu \leq \mu_o$
$H_a : \mu > \mu_o$

$H_o : \mu \geq \mu_o$
$H_o : \mu < \mu_o$

$H_o : \sigma^2 = \sigma^2_o$
$H_a : \sigma^2 \neq \sigma^2_o$

$H_o : \sigma^2 \leq \sigma^2_o$
$H_a : \sigma^2 > \sigma^2_o$

$H_o : \sigma^2 \geq \sigma^2_o$
$H_a : \sigma^2 < \sigma^2_o$

Hypothesis Testing

- State the null and alternative hypothesis
- Specify the level of significance, α
- Conclude if the null hypothesis is rejected or failed to be rejected

Robustness

All statistical procedures are based on assumptions about their theoretical behavior. When statistics obtain by these procedures are not affected by moderate deviations from theoretical expectation, they are said to be robust or insensitive to these deviations.

A statistical procedure is considered robust when it can be used even when the basic assumptions are violated to a moderate degree.

Process for Data Analysis Process

- Define your Questions
- Set Clear Measurement Priorities
- Collect Data
- Analyze Data
- Interpret Results