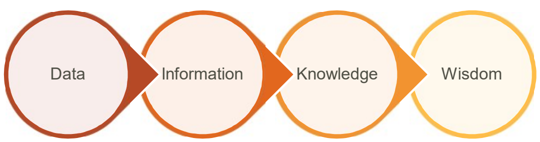
Fundamentals of Analytics

Analytics in Agriculture

The Evolution of Data

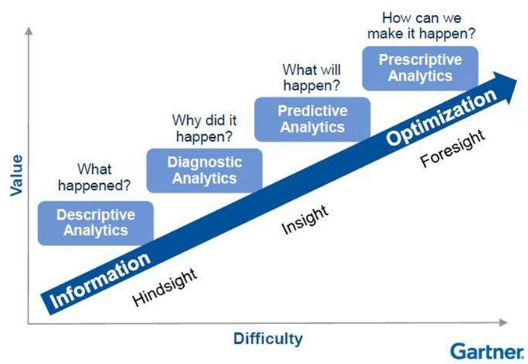
The Evolution of Data



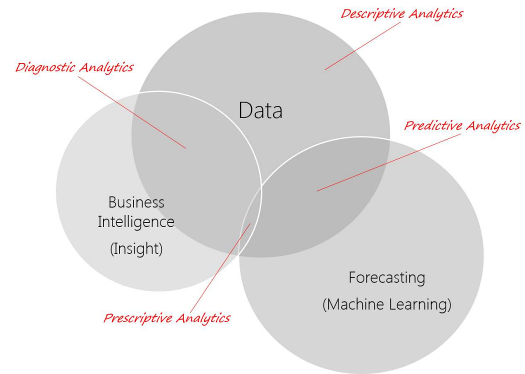
What is analytics?

- The scientific process of transforming data into insight for making better decisions
- Used for data-driven decision making which is often seen as more objective than other alternatives for decision making

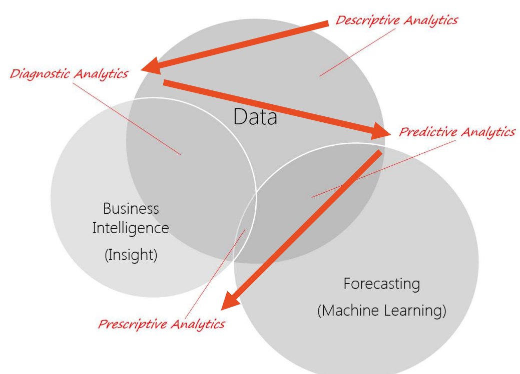
The Analytics Maturity Model



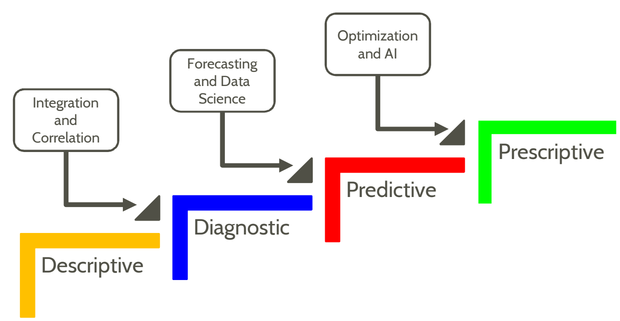
Approach



Approach



How do we get there?



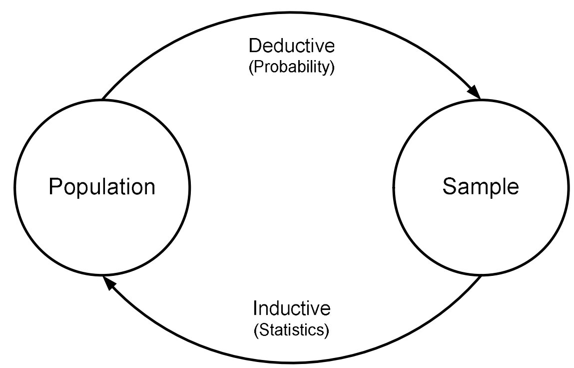
Descriptive Analytics

Agenda

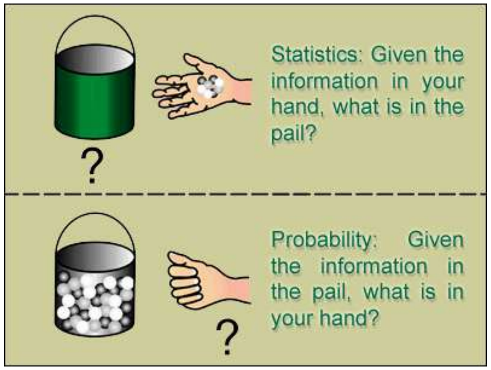
- Descriptive analytics
  - Introduction to analytics, role of probability and statistics
  - Types of data, population and samples
  - Descriptive statistics
    - Measures of location and variation
    - Bivariate relationships
  - Simple charting and visualization
  - Probability distributions
  - Estimation – point and interval estimates
  - Hypothesis testing

What is the difference between probability and statistics?

Probability & Statistics



Probability & Statistics



Probability

- Classical (a priori)
- Relative Frequency
- Subjective

What is a … ?

- Parameter
- Statistic

Types of Statistics

- Descriptive
- Inferential

Data

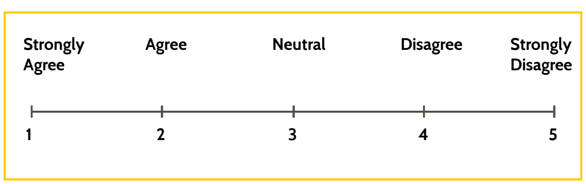
- Non-metric (qualitative)
  - Nominal
  - Ordinal
- Metric (quantitative)
  - Interval
  - Ratio

Nominal

- Lowest level
- Only classify / categorize
  - Examples – Gender, ethnicity, etc.
- Limited opportunities for analysis – (χ^2)

Ordinal

- Can order or rank
- What about differences between ranks?



Interval

- Distances between consecutive numbers have meaning
- Always numerical
- Zero is a matter of convention or convenience
  - Not a natural or fixed zero point
  - Vertical intercept of unit of measure transformation is not zero
  - Examples: Time

Ratio

- Highest level
- There is an absolute zero
  - Represents absence of the characteristic being studies
- Ratio of two numbers is meaningful

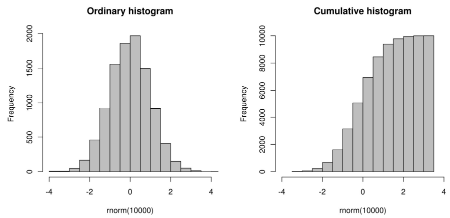
Classify the following...

- Ranking of a company in the Fortune 500
- Number of tickets sold at a movie theater on any given night
- Per capita income
- Amount of rainfall in a particular season
- Time between rainy days
- Socio-economic class
- Whether a farmer used fertilizer in a particular season
- Amount of fertilizer used in a particular season
- Number of units rejected out of an inspected lot

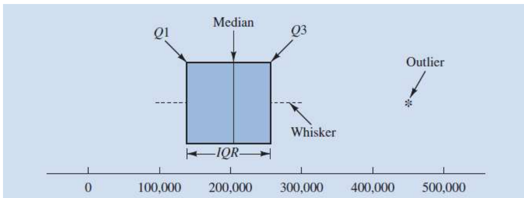
Graphical Representation of the Data

Histogram

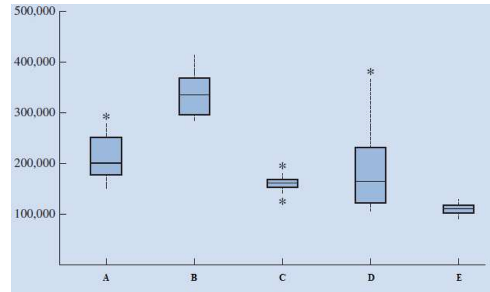
- Representation of the distribution of numerical data
  - Divide the entire range of values into a series of intervals (bins) — count how many values fall into each interval
  - Bins are (but not required to be) often of equal width



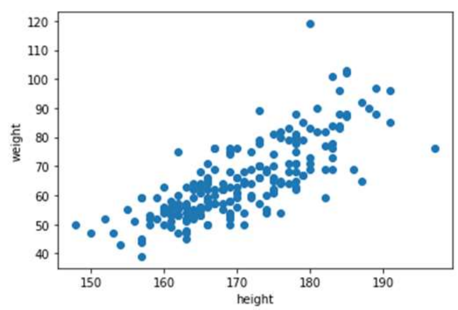
Boxplots (box and whisker plot)



Boxplots



Scatter Plot – Bivariate



Descriptive Statistics

- Parameters and statistics
- Measures of
  - Central tendency
    - Mean (μ or $\bar{x}$), median, mode, percentile, quartiles
  - Dispersion
    - Range, inter-quartile range, MAD, variance (σ^2 or s^2), standard deviation (σ or s)

Mode

- Most frequently occurring value
- What is the mode?

5 8 55 8 7 6 5 4 5 9 11

- Multimodal (when?)

Mean

- (Arithmetic) average
- Compute the mean

$$\boxed{\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

5 8 55 8 7 6 5 4 5 9 11

- What are the advantages and disadvantages?

Median

- The "middle" value
- Ordered array
  - Odd number of items – middle value
  - Even number of items – average of the middle two terms
- Compute the median

5 8 55 8 7 6 5 4 5 9 11

Percentiles

- At least *k%* of the data lie <u>below</u> the k^{th} percentile
  - What is the 50^{th} percentile?
- Organize in ascending order
- Determine location

$$i = \frac{P}{100}(n)$$

  - If i is a whole number, the percentile is the average of the values at the i and (i+1) positions
  - If i is not a whole number, the percentile is at the (i+1) position in the ordered array.

Percentile

- Compute the 30th percentile

5 8 55 8 7 6 5 4 5 9 11

Quartiles

- Divide the data into four groups
  - 25% below the first quartile
  - 50% below the second quartile
  - 75% below the third quartile
- Determine the first quartile

5 8 55 8 7 6 5 4 5 9 11

Excel Functions

- AVERAGE
- MODE
- MEDIAN
- PERCENTILE
- QUARTILE

Population Variance

- Average of the sum of squared deviations from the mean

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

- Units
- Variance?

5 8 55 8 7 6 5 4 5 9 11

Population Standard Deviation

- Square root of the variance

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Sample Variance & SD

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

Interpretation – Empirical

Distance from mean	Percentage of values falling within distance
$\mu \pm 1\sigma$	68
$\mu \pm 2\sigma$	95
$\mu \pm 3\sigma$	99.7

- Assumption?

Chebyshev's Theorem

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

- Applies to all distributions

Chebyshev's Theorem

- Proportion of values falling between
  - ±2σ
  - ±3σ
  - ±4σ

Portfolio Risk

- Portfolio A
  - 57, 68, 64, 71, 62
- Portfolio B
  - 12, 17, 8, 15, 13
- Which portfolio carries a higher risk?

Coefficient of Variation

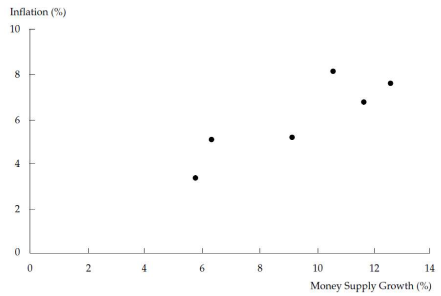
$$CV = \frac{\sigma}{\mu}(100)$$

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion

Software

- COTS
  - Excel, SAS, SPSS, JMP, Minitab
- Languages
  - R, Python

Bivariate Relationships: Scatter Plot



- Are the variables related? What can you say about the relationship?
  - Nature of the relationship
  - Magnitude of the relationship

Covariance

- Sample Variance and SD

$$s_X^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1) \qquad s_X = \sqrt{s_X^2}$$

- Sample Covariance

$$\text{Cov}\,(X,Y) = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/(n-1)$$

- Units?

Correlation Analysis

- The scatter plot leaves a lot to the skill of the interpreter
- Issues with covariance – units and scaling
- Correlation analysis expresses this same relationship using a <u>single</u> number
  - Scaled from -1 to +1
- Measures the direction and extent of <u>linear association</u> between two variables

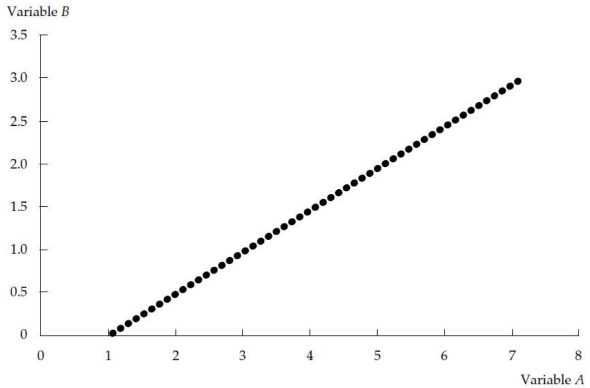
Correlation

- The correlation coefficient (r) is the covariance of two variables (X & Y) divided by the product of their standard deviations

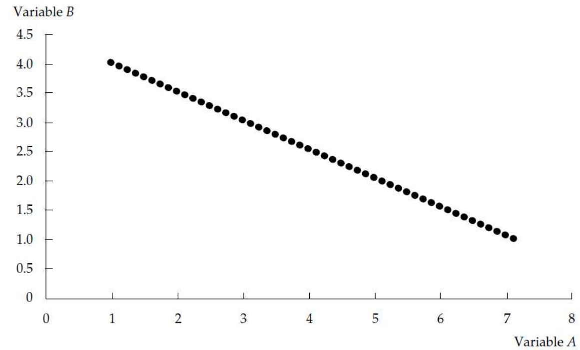
$$r = \frac{\text{Cov}\,(X,Y)}{s_X s_Y}$$

- What is the unit of the correlation coefficient?
- Assumption – mean and variance of X and Y are constant and finite

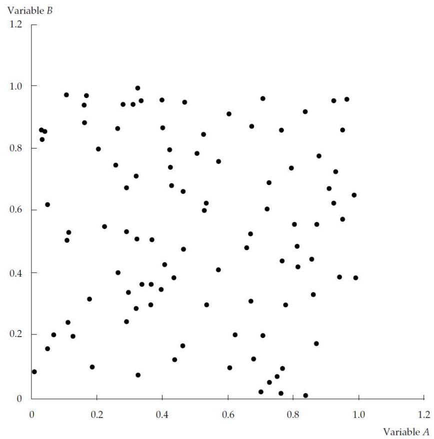
r = 1



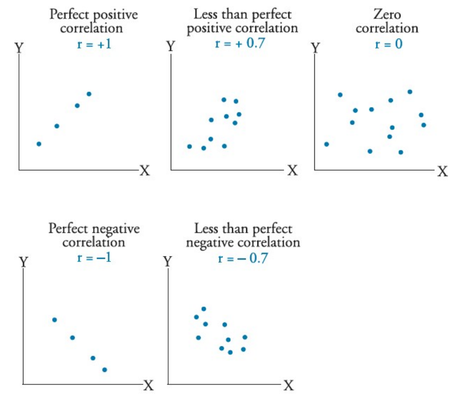
r = -1



r = 0

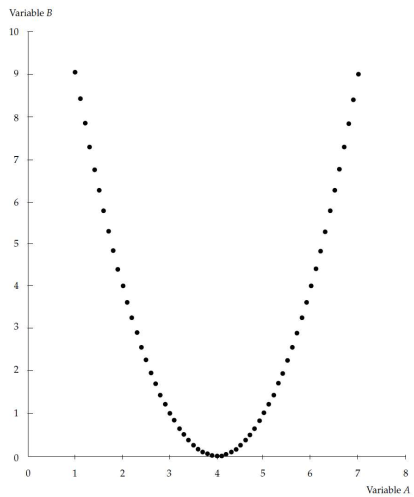


Scatter Plots



Estimate *r*

$$B = (A - 4)^2$$



Outliers

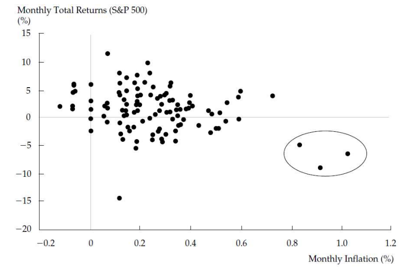
- If outliers are included

  r = -0.2997

- If outliers are excluded

  r = -0.1347

- What should we do about the outliers?



Limitations

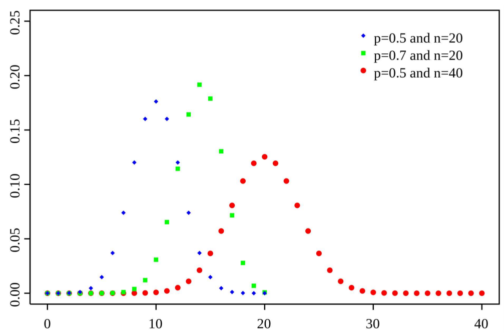
- Reliability – strong non-linear relationship yet low correlation coefficient
- Unreliable when outliers are present
  - What should be done to outliers?
- Correlation does not imply causation
- Spurious correlation
  - Chance relationships
  - No direct relationship but related to a third variable

Distributions

- Discrete
  - Binomial
  - Hypergeometric
  - Geometric
  - Poisson
  - Many others…
- Continuous
  - Normal
  - t
  - F
  - χ^2
  - Many others…

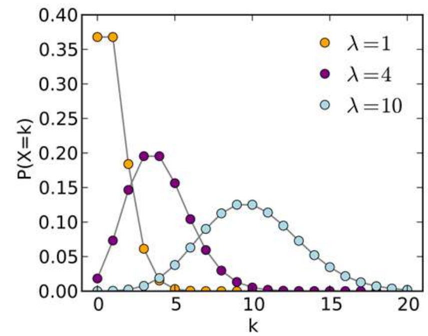
The Binomial Distribution

- Out of n trials, what is the probability of getting x successes?



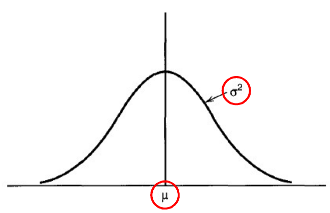
The Poisson Distribution

- Probability of a given number of events occurring in a time interval



The Normal Distribution

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(y-\mu)/\sigma]^2} \qquad -\infty < y < \infty$$

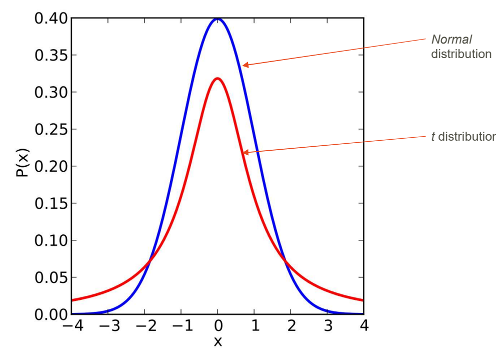


$$z = \frac{(x-\mu)}{\sigma}$$

$$z \sim N(0,1)$$

- Normal and standard normal distributions
- Why is the normal distribution important?
  - CLT

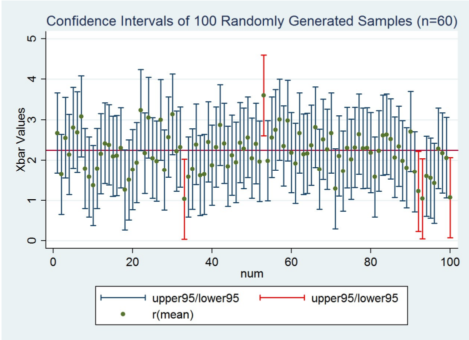
The *t* distribution



Estimation

- Point and interval estimates (why?)
- Confidence interval
  - Interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter
- Confidence level
  - Represents the proportion of possible confidence intervals that contain the true value of the unknown population parameter
- Confidence interval for the mean
  - $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ (for normal distribution and known standard deviation)
  - $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ (for unknown standard deviation)

Confidence Interval



The vertical line segments represent realizations of a confidence interval for the population mean μ. Note that some confidence intervals (shown in red) do not contain the population mean, as expected.

Testing Hypotheses

- Null hypothesis (H_0)
  - Assumes that whatever you are trying to prove did not happen
- Alternate hypothesis (H_a or H_1)
  - What you are trying to prove

- Example

$$H_0: \mu = 45$$
$$H_a: \mu \neq 45$$

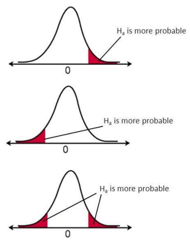
Errors in Hypothesis Testing

- Errors
  - Type I
    - Reject the null hypothesis when it is true
  - Type II
    - Fail to reject the null hypothesis when it is false

		Conclusion about null hypothesis from statistical test	
		Accept Null	Reject Null
Truth about null hypothesis in population	True	Correct	Type I error Observe difference when none exists
	False	Type II error Fail to observe difference when one exists	Correct

Methodology

- Fix the confidence level
- Calculate the observed mean
- Calculate the test statistic (t_{obs})
- Calculate the critical values ($t_{\frac{\alpha}{2}}$)
- If t_{obs} falls outside the region $(-t_{\frac{\alpha}{2}},\ t_{\frac{\alpha}{2}})$ we reject the null hypothesis
- Else, we fail to reject the null hypothesis (we can never prove H_0)
- Alternatively, we can look at p-values. <u>If p-value < α, we reject the null hypothesis</u>.



Right-tail test
H_a: $\mu >$ value

Left-tail test
H_a: $\mu <$ value

Two-tail test
H_a: $\mu \neq$ value

Simple Comparative Experiments

- Compare two conditions (sometimes called treatments)
- Illustration
  - The tension bond strength of Portland cement mortar is an important characteristic of the product. An engineer is interested in comparing the strength of a modified formulation in which polymer latex emulsions have been added during mixing to the strength of the unmodified mortar. The experimenter has collected 10 observations on strength for the modified formulation and another 10 observations for the unmodified formulation. The data are shown in the table.
  - The two different formulations are referred to as two treatments or as two levels of the factor formulations

Tension Bond Strength of Portland Cement

Modified Mortar	Unmodified Mortar
16.85	17.50
16.40	17.63
17.21	18.25
16.35	18.00
16.52	17.86
17.04	17.75
16.96	18.22
17.15	17.90
16.59	17.96
16.57	18.15

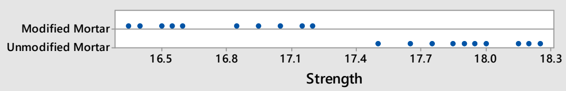
Basic Concepts

- Each observation in the Portland cement experiment would be called a run
- The individual runs differ, so there is fluctuation, or noise, in the results.
- This noise is usually called experimental error or simply error.
- It is a statistical error, meaning that it arises from variation that is uncontrolled and generally unavoidable.
- The presence of error or noise implies that the response variable, tension bond strength, is a random variable.
  - A random variable may be either discrete or continuous.

Graphical Description – Dot Plot

Dotplot of Modified Mortar, Unmodified Mortar

Shows central tendency as well as dispersion.

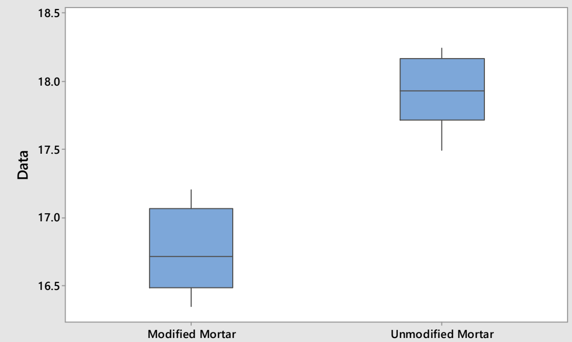


	Mean Strength
Modified Mortar	16.764
Unmodified Mortar	17.922

Do the two samples differ by a **non-trivial** amount?

Graphical Description - Boxplot

Boxplot of Modified Mortar, Unmodified Mortar



The Portland Cement Example – Summary Statistics

Modified Mortar	Unmodified Mortar
$\bar{y}_1 = 16.76$ kgf/cm^2	$\bar{y}_2 = 17.92$ kgf/cm^2
$S_1^2 = 0.100$	$S_2^2 = 0.061$
$S_1 = 0.316$	$S_2 = 0.247$
$n_1 = 10$	$n_2 = 10$

How the Two-Sample t-Test Works:

- Test statistics

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Values of t_0 that are near zero are consistent with the null hypothesis
- Values of t_0 that are very different from zero are consistent with the alternative hypothesis
- t_0 is a "distance" measure-how far apart the averages are expressed in standard deviation units

The Two-Sample (Pooled) t-Test

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{9(0.100)+9(0.061)}{10+10-2} = 0.081$$
$$S_p = 0.284$$

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{16.76 - 17.04}{0.284\sqrt{\frac{1}{10}+\frac{1}{10}}} = -2.20$$

- The two sample means are a little over two standard deviations apart
- Is this a significantly large difference?"

The Two-Sample (Pooled) t-Test

- We need an objective basis for deciding how large the test statistic t_0 really is
- In 1908, W. S. Gosset derived the reference distribution for t_0 called the *t* distribution
  - Tables of the *t* distribution are given in all statistics textbooks
  - Alternatively, we can use Excel to look up *t* values.

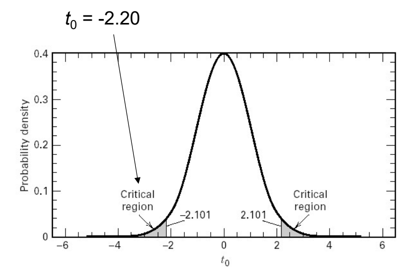


Figure 2-10 The *t* distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$.

The Two-Sample (Pooled) t-Test

- A value of t_0 between -2.101 and +2.101 is consistent with equality of means
- It is possible for the means to be equal and t_0 to exceed either 2.101 or –2.101, but it would be a "rare event" which leads to the conclusion that the means are different
- Could also use the P-value approach

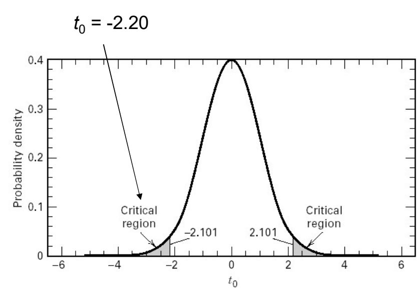


Figure 2-10 The *t* distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$.

The Two-Sample (Pooled) t-Test

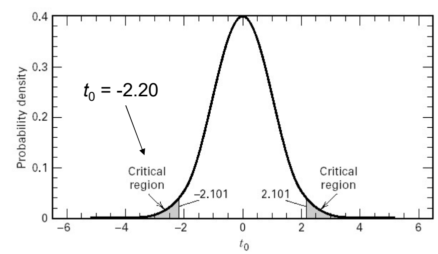


Figure 2-10 The *t* distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$.

- The *p*-value is the risk of wrongly rejecting the null hypothesis of equal means (it measures rareness of the event)
- The *p*-value in our problem is *p* = 0.042