CSCI446/946 Big Data Analytics

Week 3 – Lecture: Data Exploration

School of Computing and Information Technology
University of Wollongong Australia
Spring 2024

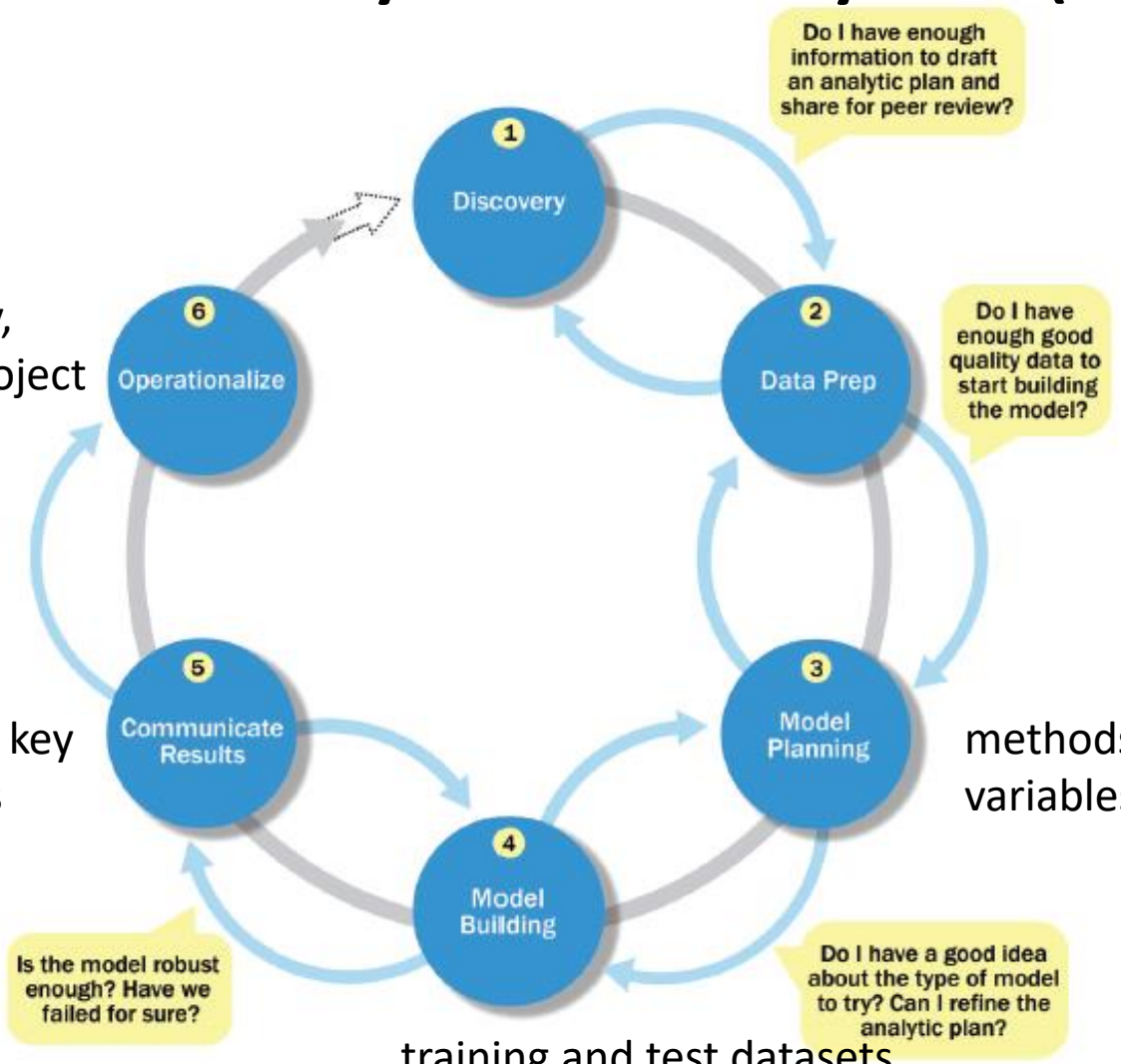
Content

- Brief Recap
  - Big Data Analytics Lifecycle
  - An Example
- Data Exploration
  - Exploratory Data Analysis
    - Visualization before analysis (in lab)
    - Visualizing single and multiple variables (in lab)
  - Statistical Methods for Evaluation
    - Hypothesis, Hypothesis Testing, t-test, ANOVA

Content

- Brief Recap
  - Big Data Analytics Lifecycle
  - An example
- Data Exploration
  - Exploratory Data Analysis
    - Visualization before analysis (in lab)
    - Visualizing single and multiple variables (in lab)
  - Statistical Methods for Evaluation
    - Hypothesis, Hypothesis Testing, t-test, ANOVA

Data Analytics Lifecycle (recap)



delivery, pilot project

identify key findings

methods, techniques, workflow, variables, relationships, models

training and test datasets, software, and hardware

FIGURE 2-2 *Overview of Data Analytics Lifecycle*

Phase 1: Discovery (recap)

- Learning the Business Domain
- Interviewing the Analytical Sponsor
- Identifying Key Stakeholders
- Resources & Goals
- Identifying Potential Data Sources
- Framing the Problem
- Developing Initial Hypotheses (IH)

Phase 2: Data Preparation (recap)

- Explore, pre-process, and condition data prior to modelling and analysis
  - Prepare an analytics sandbox
  - Perform ETLT
  - Understanding the data in detail is critical
  - Data conditioning
  - Get the data into a format to facilitate analysis
  - Perform data visualisation
- The **most labour-intensive step** in the lifecycle

Phase 3: Model Planning (recap)

- Select variables
  - Understand the relationships of variables
  - Use domain knowledge
- Identify candidate models
  - Refer to the hypothesis in Phase I
  - Translate it into machine leaning problems
    - Clustering, classification, association rules…
  - Critical literature review for latest models
  - Document the modelling assumptions
  - Model selection

Phase 4: Model Building (recap)

- Create datasets for training, validation and testing
- Perform training and testing
- Evaluation of the trained model(s)
  - Model appear valid and accurate on test/validation data?
  - Tweak training parameters as needed.
  - Model appear valid and accurate on test data?
  - Output/behaviour make sense to domain expert?
  - Model parameters make sense?
  - Model is sufficiently accurate to meet the goal?
  - Model supports run-time requirements?
  - Is a different form of the model required?

Phase 5: Communicate Results (recap)

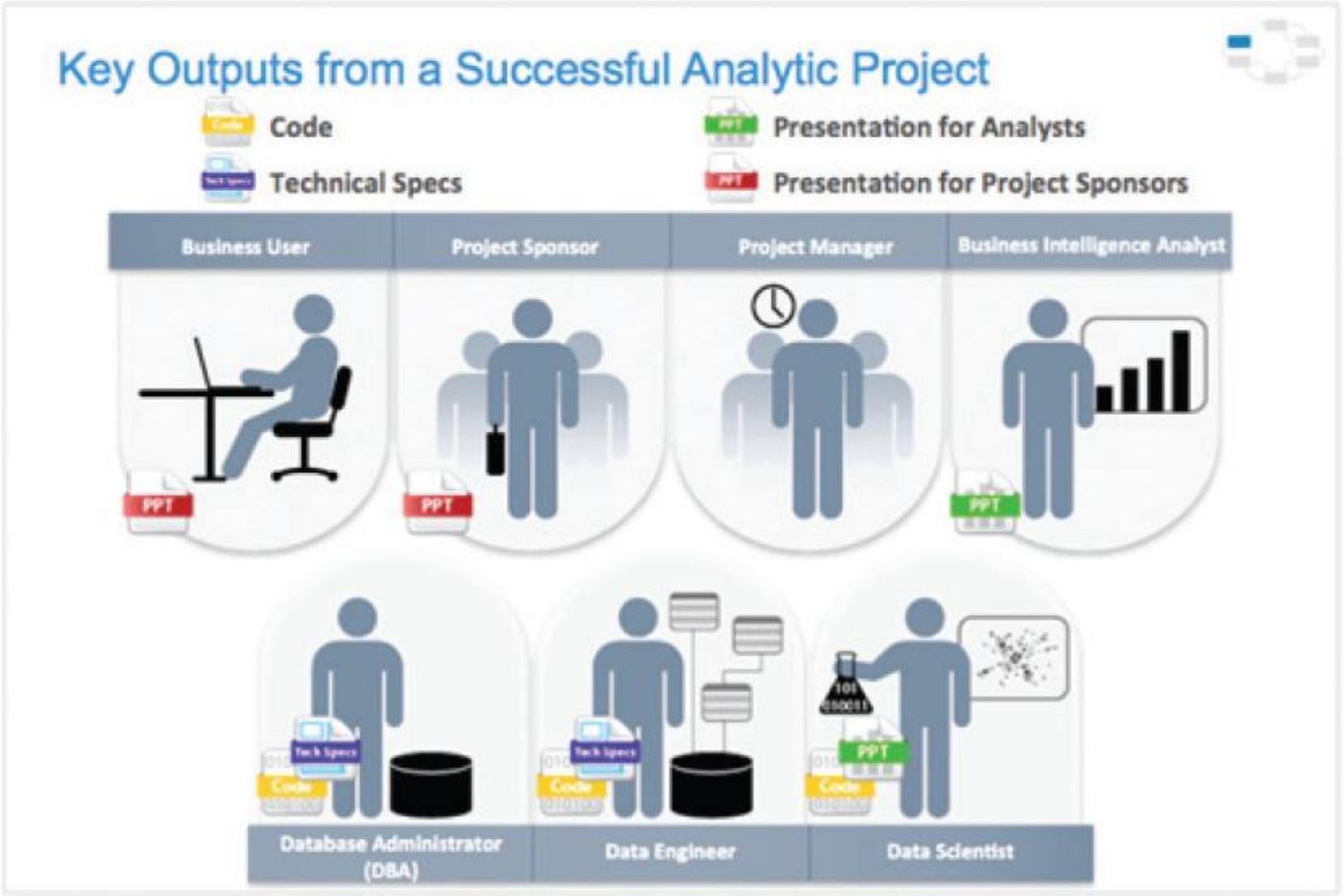
- Compare the outcomes of the modelling to the criteria established for success and failure
- Articulate the findings and outcomes to team members and stakeholders
- Take into account caveats, assumptions, and any limitations of the results
- Make recommendations for future work or improvements

- The deliverable of this phase will be the most visible portion to stakeholders and sponsors

Phase 6: Operationalize

- In the final phase, communicate the benefits of the project more broadly - Deliverables
- Set up a pilot project to deploy the work in a controlled way, before deploying the work to a full enterprise or ecosystem of users
  - Risk can be managed more effectively
- Learn the performance and constraints of the model and make adjustments as needed before a full deployment
- Train a new set of people (e.g., engineers) responsible for the production environment
- Create a mechanism for performing ongoing monitoring and improvement of model accuracy.

Phase 6: Operationalize



Phase 6: Operationalize

- Business users: benefits and implications
- Project sponsor: business impact, risk, ROI
- Project manager: completion on time, within budget, goals are met?
- BI analyst: reports and dashboards impacted?
- DE and DBA: code and documents
- Data scientist: code, model, and explanation

Phase 6: Operationalize

- Four main types of deliverables
  - Presentation for project sponsors
  - Presentation for analysts
  - Code for technical people
  - Technical specifications of implementing the code
- A general rule: the more executive the audience, the more succinct the presentation needs to be

A Simplified Example

Example: A hospital wants to reduce patient readmission rates.

- **Discovery**
  - **Example**: A healthcare organization wants to reduce patient readmission rates.
  - **Identify Problem**: High readmission rates lead to increased costs and lower patient satisfaction.
  - **Stakeholder Engagement**: Discuss with doctors, nurses, and administrators to understand the factors contributing to readmissions.
  - **Data Sources**: Identify relevant data sources such as patient records, treatment histories, and demographic data.

A Simplified Example

- **Data Preparation -** Preparing the healthcare data for analysis.
  - **Data Collection**: Extract patient records, treatment histories, and demographic data from electronic health records (EHR) systems.
  - **Data Cleaning**: Handle missing values, correct errors in patient records, and standardize formats (e.g., date formats).
  - **Data Integration**: Combine data from different sources into a unified dataset, ensuring consistency across variables.

A Simplified Example

- **Model Planning -** Planning a model to predict patient readmissions.
  - **Feature Selection**: Identify key features such as age, gender, diagnosis, treatment type, and length of stay that might influence readmissions.
  - **Algorithm Selection**: Choose suitable algorithms for the prediction task, such as logistic regression, decision trees, or random forests.
  - **Evaluation Criteria**: Define metrics for model evaluation, such as accuracy, precision, recall, and the F1-score.

A Simplified Example

- **Model Building-** Building and training the predictive model.
  - **Data Splitting**: Split the data into training and testing sets (e.g., 80% training, 20% testing).
  - **Training**: Train the chosen algorithms on the training data to learn patterns and relationships.
  - **Validation**: Use cross-validation techniques to validate the model and tune hyperparameters for optimal performance.

A Simplified Example

- **Communicate Results -** Sharing the analysis and model results with stakeholders.
  - **Visualization**: Create visualizations such as confusion matrices, ROC curves, and feature importance charts to explain model performance.
  - **Reporting**: Generate a comprehensive report detailing the analysis process, model outcomes, and actionable insights.
  - **Stakeholder Presentation**: Present the findings to healthcare providers, highlighting how the model can predict high-risk patients and recommending preventive measures.

A Simplified Example

- **Operationalize -** Implementing the predictive model in the healthcare setting.
  - **Integration**: Integrate the predictive model into the hospital's EHR system to provide real-time readmission risk scores for patients.
  - **Monitoring**: Continuously monitor model performance and retrain it periodically with new data to maintain accuracy.
  - **Actionable Use**: Develop workflows where high-risk patients identified by the model receive targeted interventions such as follow-up calls, personalized care plans, and closer monitoring. Explain the factors

Brief Recap

Issue & Question & Answer

Content

- Brief Recap
  - Big Data Analytics Lifecycle
  - An Example
- Data Exploration
  - Exploratory Data Analysis
    - Visualization before analysis (in lab)
    - Visualizing single and multiple variables (in lab)
  - Statistical Methods for Evaluation
    - Hypothesis, Hypothesis Testing, t-test, ANOVA

Key Objectives of Data Exploration

- Understanding Data Structure
  - Types of Data; Distribution; Summary Statistics:
- Assessing Data Quality
  - Missing Values; Outliers; Duplicates
- Identifying Patterns and Relationships
  - Correlation Analysis; Cross-Tabulation (relationships between categorical variables); Visualizations
- Formulating Hypotheses
  - Hypotheses or ideas about potential patterns, trends, or relationships

Exploratory Data Analysis

- Consider two situations:

  1. A company collects data about customer satisfaction levels, and wishes to investigate whether a change in the product design would improve the costumer satisfaction. How can the company ascertain whether the change had the desired effect?

  2. A data scientist produces a set of results by deploying a machine learning model. The data scientist wishes to investigate whether a variation of the model architecture would improve results. How can the data scientist be certain that the modified model yields an improvement in results?

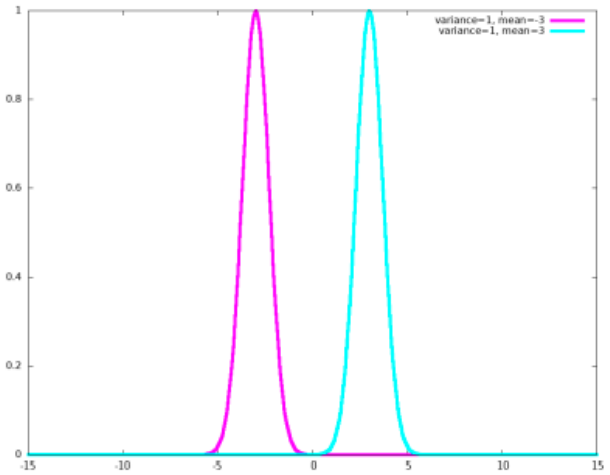
Exploratory Data Analysis

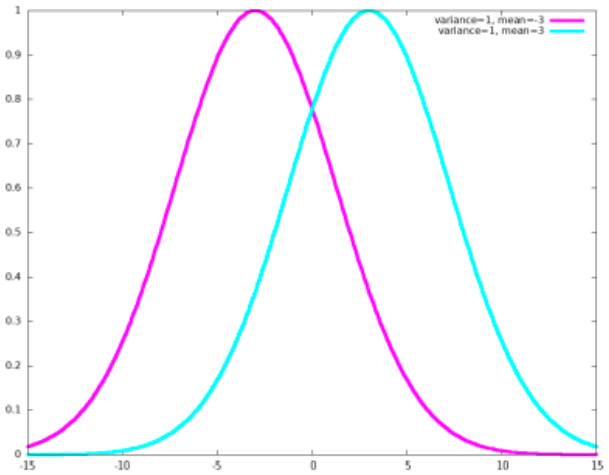
- For each of the two situations we could compute the average (mean) value of the data prior to the change, and compute the mean value of the data after the change.
  - Analyse the difference of Means

- Would it be correct to state that if the new mean is larger than the old mean value then the new product or model is better than the old?

Statistical Methods for Evaluation

- Analysis of the difference of Means
  - Very common hypothesis test.
  - But simple comparison is often not sufficient.
  - Example: Assume we have two populations, one with mean=-3 and the other with mean=3
    - By comparing the means can we say that the difference between the two populations is significant?
    - Answer depends on variance.





- Descriptive Statistics in R
  - `Summary()` function: mean, median, min, max
  - R functions include descriptive statistics

```
# to simplify the function calls, assign
x <- sales$sales_total
y <- sales$num_of_orders

cor(x,y)          # returns 0.7508015 (correlation)
cov(x,y)          # returns 345.2111 (covariance)
IQR(x)            # returns 215.21 (interquartile range)
mean(x)           # returns 249.4557 (mean)
median(x)         # returns 151.65 (median)
range(x)          # returns 30.02 7606.09 (min max)
sd(x)             # returns 319.0508 (std. dev.)
var(x)            # returns 101793.4 (variance)
```

- Descriptive Statistics in Python: NumPy
- https://numpy.org/doc/stable/reference/routines.statistics.html

Statistical Methods for Evaluation

- Statistics is crucial because it may exist throughout the entire Data Analytics Lifecycle
  - Initial data exploration and data preparation
  - Model building and planning
    - Best input variables, predictability
  - Evaluation of the final models
    - Accuracy, better than guess or another one?
  - Assessment of the new models when deployed
    - Sound prediction? Have desired effect?

Statistical Methods for Evaluation

- Hypothesis Testing
  - Form an assertion and test it with data
  - Common assumption (there is no difference)
  - Null hypothesis (H_0)
  - Alternative hypothesis (H_A)
- Example: identify the effect of drug A compared to drug B on patients
  - What are the H_0 and H_A ?
- A hypothesis is formed before validation
  - It can define expectations.

Definition (Hypothesis): a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

A Question

- But why Hypothesis testing in BDA lifecycle? In Phase 2 Data Preparation?
  - Hint: *Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases.*

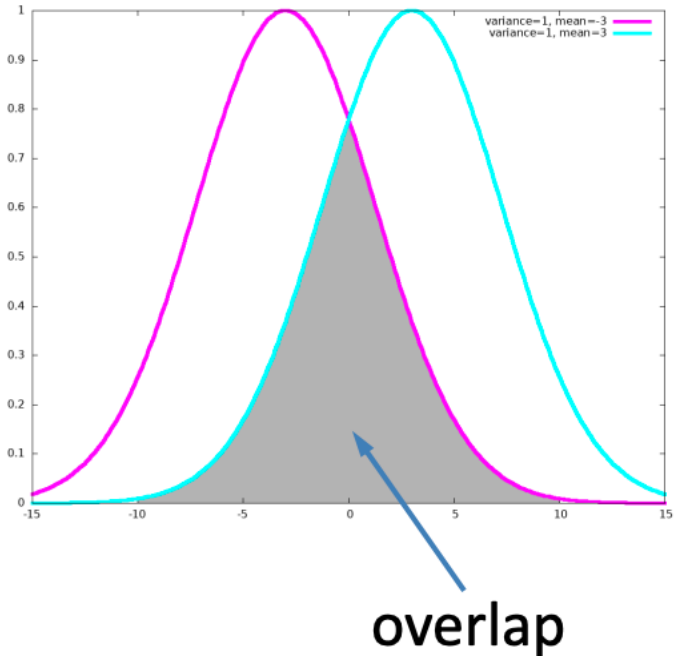
Statistical Methods for Evaluation

- Hypothesis Testing
  - Clearly state Null and Alternative hypotheses
  - **Either** reject the null hypothesis in favour of the alternative **or** not reject the null hypothesis

Application	Null Hypothesis	Alternative Hypothesis
Medical drug development	Drug A is not more effective than drug B	Drug A is more effective than drug B.
Accuracy Forecast	Model X does not predict better than the existing model.	Model X predicts better than the existing model.
Regression Modelling	This variable does not affect the outcome because its coefficient is zero.	This variable affects the outcome because its coefficient is not zero.

Exploratory Data Analysis

- The overlap between two populations is larger, the closer the means (x axe) and the larger the variances (y axe).
- The greater the overlap, the less the significance of the differences between the population.
- Thus, if the overlap is large we would accept the null hypothesis, otherwise we would reject it.
- This can be tested using student's t-test.



Statistical Methods for Evaluation

- Student's *t*-test
  - Assume that distributions of the two populations have equal but unknown variance
  - If each population is normally distributed with the same mean and with the same variance, then

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Signal

Noise

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Pooled variance

T (the *t-statistic*) follows a *t-distribution* with (n_1+n_2-2) degree of freedom

$$S_k^2 = \frac{\sum_i^{n_k}(x_i - \bar{X}_k)^2}{n_k - 1}$$

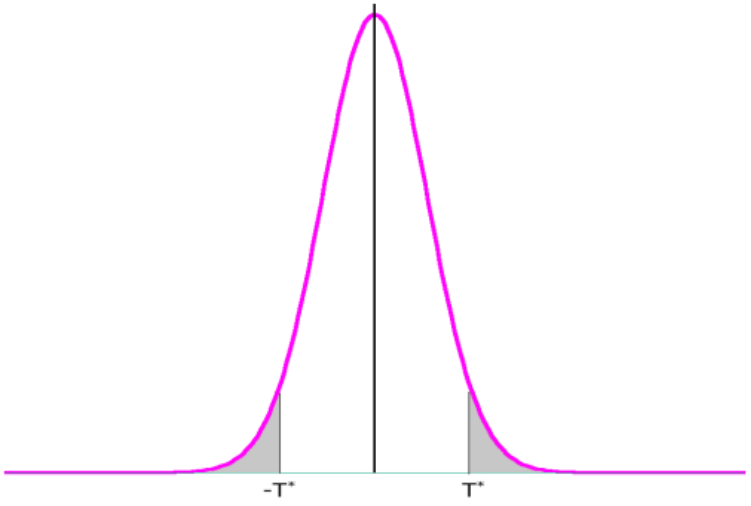
Variance

Statistical Methods for Evaluation

- Student's *t*-test
  - The further T is from zero the more significant the difference between the populations. If |T| is large then one would reject the null hypothesis

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

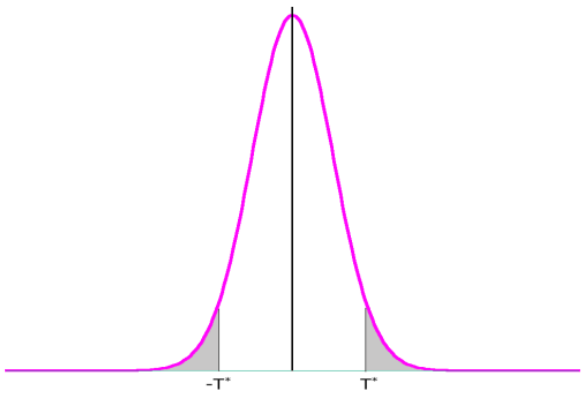


Statistical Methods for Evaluation

- Student's *t*-test
  - Significance level of the test (α): the probability of rejecting the null hypothesis, when the null hypothesis is actually TRUE
  - So, what does it mean by setting $\alpha = 0.05$?
  - Find T* such that $P(|T| \geq T^*) = \alpha$
  - Reject H_0 if $|T| \geq T^*$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$



Statistical Methods for Evaluation

- Student's *t*-test (an example)

```
# generate random observations from the two populations
x <- rnorm(10, mean=100, sd=5)     # normal distribution centered at 100
y <- rnorm(20, mean=105, sd=5)     # normal distribution centered at 105

t.test(x, y, var.equal=TRUE)       # run the Student's t-test
Two Sample t-test

data:  x and y
t = -1.7828, df = 28, p-value = 0.08547
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -6.1611557  0.4271893
sample estimates:
  mean of x mean of y
102.2136  105.0806
```

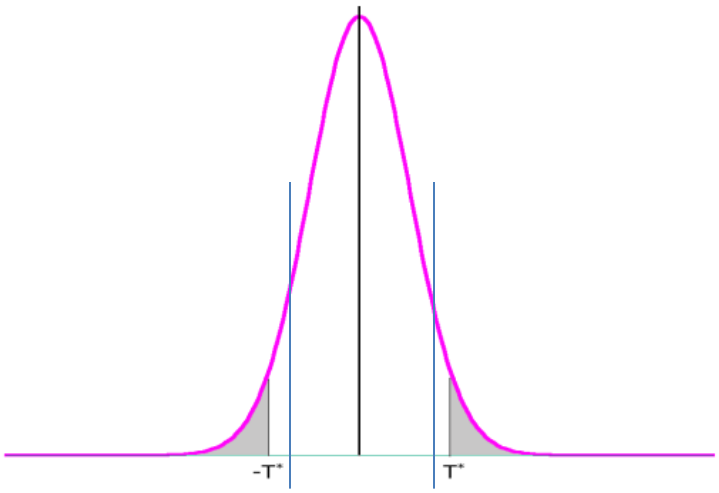
Statistical Methods for Evaluation

- Shall we reject or accept the null hypothesis?
- What does the "two-sided test" mean?

```
# obtain t value for a two-sided test at a 0.05 significance level
qt(p=0.05/2, df=28, lower.tail= FALSE)
2.048407
```

T*

- Perform T-test:
  - Is |T| greater or equal to T*?
  - Let check: |-1.7828|>=2.048407
  - Answer is "No"
    - Insufficient evidence to reject
    - H_0 is accepted.



Statistical Methods for Evaluation

- Student's *t*-test (an example)
  - What does the "p-value" mean? – two-sided test

```
t = -1.7828, df = 28, p-value = 0.08547
```

  - The significance level α corresponds to the sum of P(T ≤ -t) and P(T ≥ t)
    - The two "tails" in the distribution function
    - The significance level for each of these two tails is thus α/2
  - p-value offers the probability of observing |T| ≥ t given the null hypothesis is TRUE : 0.08547>0.05
    - α=0.05 is very common in statistics.

Statistical Methods for Evaluation

- Student's *t*-test (an example)
  - What is the "95 percent confidence interval"?

```
95 percent confidence interval:
  -6.1611557  0.4271893
```

  - A confidence level is an interval estimate of a population parameter based on sample data, indicating the uncertainty of a point estimate
    - If $\bar{x}$ is the estimate of an unknown population mean μ, the confidence interval provides an idea of how close $\bar{x}$ is to the unknown μ.
    - The above "95 percent confidence interval" straddles the TRUE mean value of the difference of the population means 95% of the time

Statistical Methods for Evaluation

- Student's *t*-test (an example)
  - Why is var.equal = TRUE?

```
t.test(x, y, var.equal=TRUE)
```

  - Student's t-test requires that variances of the two is equal!
  - If such assumption is not appropriate then the Welch's t-test should be used.

Statistical Methods for Evaluation

- One-sided t-Test Example - A pharmaceutical company has developed a new drug, and they claim that it lowers blood pressure more than the current standard drug. They conduct an experiment to compare the average reduction in blood pressure between the new drug and the standard drug.
  - **Null Hypothesis (H_0)**: The new drug does not lower blood pressure more than the standard drug, i.e., $\mu_{new} \leq \mu_{stardarad}$
  - **Alternative Hypothesis (H_1)**: The new drug lowers blood pressure more than the standard drug, i.e., $\mu_{new} > \mu_{stardarad}$
- This is a **one-sided t-test** because the alternative hypothesis is directional, focusing only on whether the new drug performs better.

Statistical Methods for Evaluation

- Two-sided t-Test Example - A company wants to know if a new packaging design leads to a difference in the average sales of a product, compared to the current design. They are not sure whether the new design will increase or decrease sales, just that it might cause a difference.
  - **Null Hypothesis (H_0)**: There is no difference in the average sales between the new and old packaging designs, i.e., $\mu_{new} = \mu_{stardarad}$
  - **Alternative Hypothesis (H_1)**: There is a difference in the average sales between the new and old packaging designs, i.e., $\mu_{new} \neq \mu_{stardarad}$
- This is a **two-sided t-test** because the alternative hypothesis is non-directional, considering the possibility of a difference in either direction.

Statistical Methods for Evaluation

- Welch's *t*-test
  - Shall be used when the equal population variance assumption is NOT justified
  - It uses the sample variance for each population instead of the pooled sample variance
  - Still assume two populations are normal with the same mean

$$T_{welch} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Statistical Methods for Evaluation

- Welch's *t*-test

```
t.test(x, y, var.equal=FALSE)       # run the Welch's t-test

Welch Two Sample t-test

data:  x and y
t = -1.6596, df = 15.118, p-value = 0.1176
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -6.546629  0.812663
sample estimates:
  mean of x mean of y
102.2136  105.0806
```

Statistical Methods for Evaluation

- Wilcoxon Rank-Sum Test
  - What if the two populations are not normal?
- Parametric test (i.e. student's t-test)
  - Makes assumptions about the population distributions from which the samples are drawn
- Nonparametric test (i.e. Wilcoxon rank-sum test)
  - Shall be used if the populations cannot be assumed (or transformed) to be normal

Statistical Methods for Evaluation

- Wilcoxon Rank-Sum Test
  - A nonparametric test to check whether two populations are identically distributed
  - It uses "ranks" instead of numerical outcomes to avoid specific assumption about the distribution
- How to conduct the test
  - Rank two samples as if they are from one group
  - Sum assigned ranks for one population's sample
  - Determine the significance of the rank-sums

Statistical Methods for Evaluation

- Wilcoxon Rank-Sum Test

```
wilcox.test(x, y, conf.int = TRUE)

Wilcoxon rank sum test

data:  x and y
W = 55, p-value = 0.04903
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  -6.2596774 -0.1240618
sample estimates:
  difference in location
-3.417658
```

p-value: the probability of the rank-sums of this magnitude being observed assuming that the population distributions are identical

Statistical Methods for Evaluation

- Wilcoxon Rank-Sum Test – Suppose we have the following data:
  - Group A: [85, 80, 78, 90, 95]; Group B: [88, 82, 85, 87, 92]
- Step 1: Combine and Rank the Data
  - Combine: [85, 80, 78, 90, 95, 88, 82, 85, 87, 92]
  - Rank: [4.5, 2, 1, 8, 10, 6, 3, 4.5, 5, 9]
- Step 2: Sum the Ranks for Each Group
  - Group A: [4.5, 2, 1, 8, 10]; Sum of ranks W_1=4.5+2+1+8+10=25.5
  - Group B: [6, 3, 4.5, 5, 9]; Sum of ranks W_2=6+3+4.5+5+9=27.5
- Step 3: Choose the Test Statistic
  - W can be either 25.5 or 27.5 depending on the test design, but usually, the smaller sum is used if conducting a one-sided test.
- Step 4: Determine Significance
  - Compare the test statistic W to a critical value from the Wilcoxon rank-sum distribution or use a p-value from statistical software.

Statistical Methods for Evaluation

- Type I and Type II Errors
  - Type I error: the rejection of the null hypothesis when the null hypothesis is TRUE
    - The probability of type I error is denoted by α
    - Fix by select an appropriate significance level
  - Type II error: the acceptance of the null hypothesis when the null hypothesis is FALSE
    - The probability of type II error is denoted by β
    - Fix by increase sample size
- Power (statistical power):
  - determine necessary sample size
  - The probability of correcting rejecting the null hypothesis (1- β)

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)
  - What if there are more than two populations?
  - Multiple *t*-test may not perform well now
- A generalization of the hypothesis testing
  - ANOVA tests if any of the population means differ from the other population means
  - Each population is assumed to be normal and have the same variance

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)

$$H_o: \mu_1 = \mu_2 = \ldots = \mu_n$$

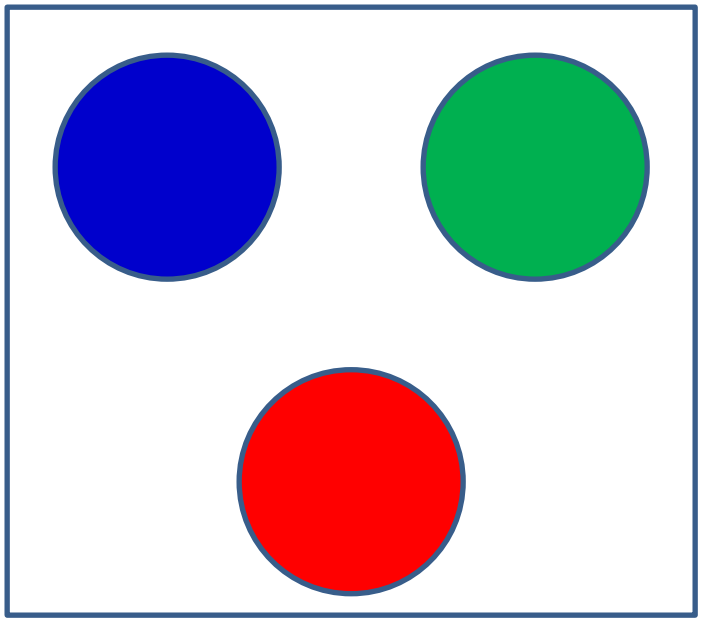
$$H_A: \mu_i \neq \mu_j \text{ for at least one pair of } i, j$$

- Compute *F*-test statistic
  - Between-groups mean sum of squares
  - Within-groups mean sum of squares

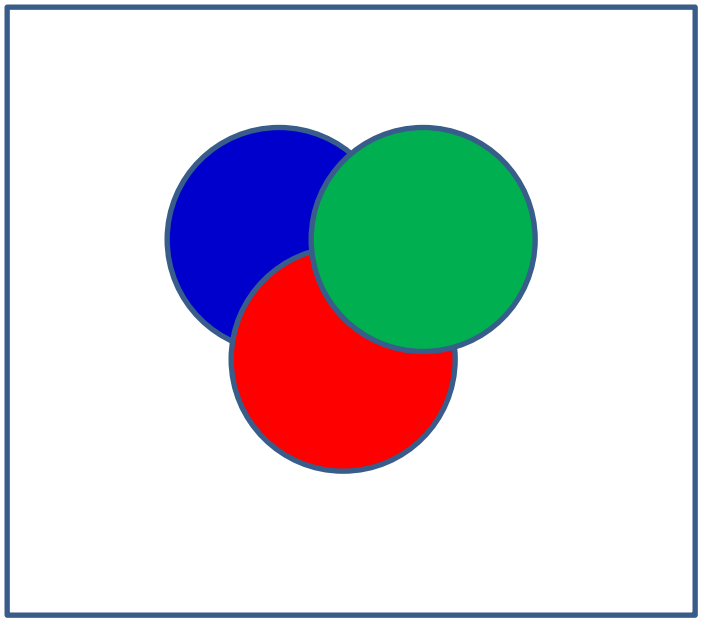
$$S_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i \cdot (\bar{x}_i - \bar{x}_0)^2 \qquad S_W^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)



$$F = \frac{S_B^2}{S_W^2}$$



$$S_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i \cdot (\bar{x}_i - \bar{x}_0)^2$$

$$S_W^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)
  - Measures how different the means are relative to the variability within each group
  - The larger the *F*-test statistic, the greater the likelihood that the difference of means are due to something other than chance alone
  - The *F*-test statistic follows an *F*-distribution

$$F = \frac{S_B^2}{S_W^2}$$

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)

```
# fit ANOVA test
model <- aov(purchase_amt ~ offers, data=offertest)

summary(model)
             Df Sum Sq Mean Sq F value Pr(>F)
offers        2 225222  112611   130.6 <2e-16 ***
Residuals   497 428470     862
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Shall we accept or reject the null hypothesis?

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)
  - One-Way ANOVA
    - Compares means across different groups based on a single independent variable (factor).
    - e.g. Comparing the mean test scores of students across different teaching methods (Method A, Method B, Method C).
  - Two-Way ANOVA:
    - Compares means across groups based on two independent variables (factors), and can also evaluate the interaction between the factors.
    - e.g. Comparing test scores based on teaching methods (Factor 1) and study times (Factor 2).

Statistical Methods for Evaluation

- Limitations of ANOVA (Analysis of Variance)
  - Assumptions
    - Normality: Data should be approximately normally distributed.
    - Homogeneity of Variances: Variances within each group should be equal (tested using Levene's test).
    - Independence: Observations should be independent of each other.
  - Limitations:
    - Sensitivity to Outliers: Outliers can affect the F-statistic and lead to misleading results.
    - Assumes Equal Variances: Violations of this assumption can impact the validity of the results.
    - Identifies Differences but Not Specifics: ANOVA indicates whether a difference exists but does not specify which groups are different without further tests (post-hoc).

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)
  - Additional tests for each pair of groups
  - Tukey's Honest Significant Difference (HSD)

```
TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = purchase_amt ~ offers, data = offertest)

$offers
                     diff        lwr      upr     p adj
offer1-nopromo 40.961437 33.4638483 48.45903 0.0000000
offer2-nopromo 48.120286 40.5189446 55.72163 0.0000000
offer2-offer1   7.158849 -0.4315769 14.74928 0.0692895
```

Statistical Methods for Evaluation

- Tukey's Honest Significant Difference (HSD)
  - Assumptions
    - Norm + equal variance + sample sizes are approximately equal (though it can still be used if they are not).
  - Perform ANOVA test
    - establish whether there is a significant difference between the means of the groups
  - Calculation of the HSD
    - Critical value from studentized range distribution, Mean square within groups(from ANOVA), number of groups
  - Decision Rule
    - For each pair of means, calculate the absolute difference.
    - Compare the absolute difference to the HSD value.
    - If the absolute difference is greater than the HSD, the pair of means is considered significantly different.

Q&A

Images Courtesy of Google Image