BASIC STATISTICS (Cheat Sheet)

codybaldwin.com

THE FUNDAMENTALS

Data Types:

- *Attribute Data – Qualitative:*
  * Text Data – e.g. yes/no, pass/fail, approve/reject…
- *Variable Data – Quantitative:*
  * Discrete – counted numbers – e.g. # of defects (74), # of customer returns (13)
  * Continuous – decimal numbers – e.g. time (12:24:59), money ($17.4354), pressure (25.44534 lbs.)

Types of Statistics:

- Descriptive Stats – Used to describe and summarize data.
- Inferential Stats – Drawing conclusions about a population, when sample data is used.
  * As we gather data, we work with samples.
  * We need confidence that our sample represents the population.

Measures of Central Tendency:

- Mean – The average.
- Median – The middle value.
- Mode – The most frequently occurring value.
- Trimmed Mean – A compromise between the mean and median, removes some outliers then averages.

Measures of Variation:

- Range – Difference between the largest and smallest value.
- Interquartile Range – Difference between the 75th and 25th percentile.
- Standard Deviation – Average deviation of values from the mean.
- Variance – Average squared deviation of values from the mean.

Basic Graphs:

- Histogram – shows central tendency and variation within a *single* distribution.
- Dotplot – similar to a histogram, but shows each value as an individual point.
- Boxplot – shows central tendency and a variation within *several* distributions, not just one.
- Time-Series Plot – shows critical quality measurements over time.
- Scatterplot – shows the relationship between two variables.

Data Measurement Scales:

- Nominal – Cannot be ordered; no arithmetic can be performed. e.g. city (Detroit, Cleveland, Seattle).
- Ordinal – Can be ordered; differences between values meaningless. e.g. taste (bad, okay, good).
- Interval – Can be ordered; differences between values meaningful (not ratios). e.g. temp (0^o, 10^o , 20^o).
- Ratio – Can be ordered; ratios meaningful; zero indicates an absence. e.g. weight (0kg, 25kg, 50kg).

Types of Sampling & Measurement Errors:

- Sampling Error – Differences among samples drawn at random ("luck of the draw").
- Sampling Bias – A lack of random samples (e.g. height of basketball players only).
- Measurement Error – Issues with our measurement systems.
- Measurement Invalidity – Not measuring what it is intended (e.g. temperature near a furnace).

HYPOTHESIS TESTING

Helps answer: *"Is the sample a fair representation of the population?"*

Hypotheses:

- Null Hypothesis (Ho) – assumes NO differences (the same), *p-value* > 0.05
- Alternative Hypothesis (Ha) – states there is a difference, *p-value* < 0.05

Tests for Normal Data ("t-tests"):

- 1-Sample t-Test – study one sample's mean against a target.
- 2-Sample t-Test – study means from two different samples.
- ANOVA Test – study means from more than two samples.
- Paired t-Test – study paired data (e.g. same part before/after improvement).

Normal vs Non-Normal Data

- Hypothesis tests with NORMAL data use the **mean** for central tendency
- Hypothesis tests with NON-NORMAL data use the **median** for central tendency

DESIGN OF EXPERIMENTS

- Shows the cause and effect relationship between X and Y.
- Helps determine the proper settings (levels) for our inputs (X) in order to optimize our output (Y).

Key Terminology:

- Factors (x) – The independent variables being used (e.g. temperature).
- Levels – The various settings for the factors (e.g. 300°, 500°).
- Run – A set of experimental conditions. (Experiments have multiple.)
- Response (y) – The result from an experimental run (e.g. material strength).
- Replication – The repetition of experimental runs. (Challenges the result.)

Common Types of Experiments:

- Full Factorials – use 2-5 input variables with all combinations of levels (or settings).
- Fractional Factorials – use 4-15 input variables and a fraction of combinations.

General Notation for Full Factorial Design (2k):

- k = # of input variables
- 2 = # of levels used for each factor

Principles of Good Experimental Design:

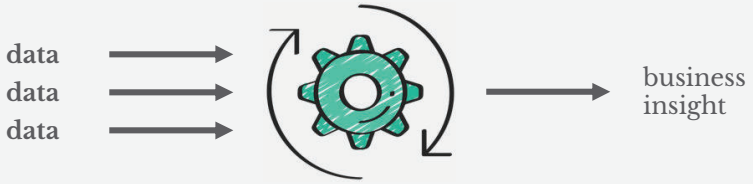
- Randomization of runs to remove bias and spread noise
- Replication of the experiment to challenge or strengthen the validity of results.
- Monitoring of noise.
- Holding other factors constant. (Those that are not a focus on the experiment.)

BUSINESS ANALYTICS (Cheat Sheet)

codybaldwin.com

BASIC CONCEPTS

What is Analytics?



Types of Analytics

- **Descriptive Analytics** — What happened?
- **Predictive Analytics** — What might happen?
- **Prescriptive Analytics** — What should we do?

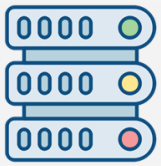
Lifecycle of Analytics *("CRISP-DM")*

- **Business Understanding** — Define the business problem.
- **Data Understanding** — Identify available data and gaps in data.
- **Data Preparation** — Clean and prepare the data.
- **Modeling** — Build predictive models.
- **Evaluation** — Evaluate how the models perform.
- **Deployment** — Start using the chosen model.

BIG DATA

Big data is so large that "it requires the use of new technical architectures ... to enable insights that unlock new sources of business value." (McKinsey)

3 V's of Big Data *(Defining Characteristics)*



Volume

Velocity

Variety

POPULAR TOOLS

Microsoft Excel
Allows you to explore/analyze smaller data sets

Tableau Desktop *(or Power BI)*
Allows you to visualize your data with dashboards

Python Language *(or R)*
Allows you to build models to make predictions

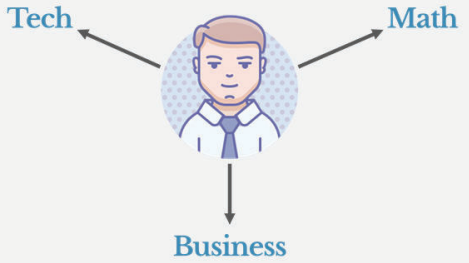
Structured Query Language (SQL)
Allows you to communicate and interact with databases

CAREERS IN ANALYTICS

Common Job Titles

- Business Analyst
- Business Intel. Analyst
- Analytics Manger
- Data Analyst
- Data Scientist *
- ...

** Most people feel this job is more technical.*



Most job postings ask about software, so:

- Select a tool from above
- Download a free trial.
- Get a pizza!
- Spend a weekend to learn.
- State you have "Experience with..." on your resume.

SQL (Cheat Sheet)

codybaldwin.com

CATEGORIES OF SQL COMMANDS

Category	Objective	Examples
Data Definition Language (DDL)	Define the database.	CREATE, DROP, ALTER
Data Manipulation Language (DML)	Manipulate the data.	SELECT, INSERT, UPDATE
Data Control Language (DCL)	Manage permissions.	GRANT, REVOKE
Transaction Control Language (TCL)	Manage transactions.	COMMIT, ROLLBACK

BASIC SQL COMMANDS

Command	Description
SELECT	Query information in a database.
INSERT	Add records to a database.
UPDATE	Update information in a database.
DELETE	Delete information in a database.

SQL STYLE GUIDE

Capitalize SQL commands for readability – SELECT, FROM, etc.
Use new lines and indenting for readability – avoid long statements on one line
If possible, use only standard SQL functions instead of vendor SQL functions for portability
Consider incorporating comments to provide more context
In general, use single quotes for values and double quotes for tables and columns.
Remember, there might be more than one way to accomplish a query

GENERAL OPTIMIZATION BEST PRACTICES

Select only the columns you need. Querying unnecessary data consumes resources.
Use LIMIT to sample query results. You may find the query needs refinement.
Use wildcards at the end of a phrase only. Wildcards generate wide searches.
Avoid SELECT DISTINCT if possible. This can require lots of processing power.
Run large queries during off-peak hours. This minimizes impact to others.

BASIC SQL STATEMENT

```
SELECT column1, column2...
FROM table;
```

ORDER BY – Sort your results.

```
SELECT column1, column2, ...
FROM table
ORDER BY column1, column2 ... ASC | DESC ;
```

GROUP BY – Summarizes your results into groups.

```
SELECT column1, SUM(column2)
FROM table
GROUP BY column1;
```

INSERT INTO – Inserts new records into an existing table.

```
INSERT INTO table_name ( column1, column2, column3, ... )
VALUES ( value1, value2, value3, ... );
```

UPDATE – Modify existing records in a table.

```
UPDATE table_name
SET column1 = value1, column2 = value2 ...
WHERE condition;
```