

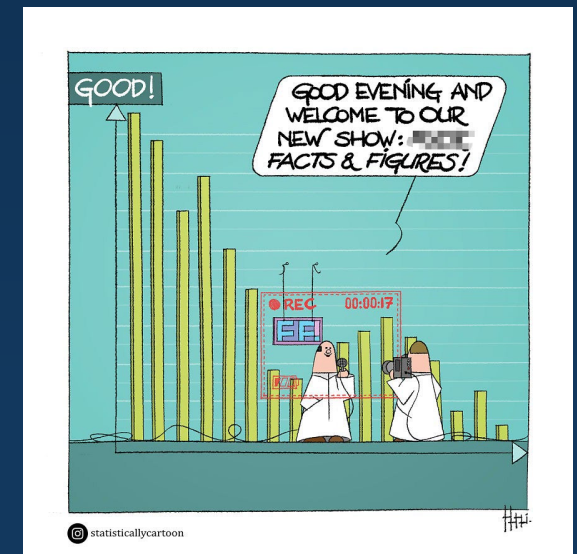
Things Forgotten Since College

Foundational Statistics

Jordan Harlacher and Kyle Davis
Quantech Services, Inc.
May 14, 2024

Table of Contents

- Exploring the World of Statistics
- Data Collection & Organization
- Measures
 - Central Tendency
 - Variability
 - Probability
- Statistical Inference
 - Sampling
 - Correlation & Causation
 - Statistical Distributions
 - Regression Analysis
- Hypothesis Testing
 - Experimental Design
 - Bias and Error
- Data Visualization
- Conclusion
- References



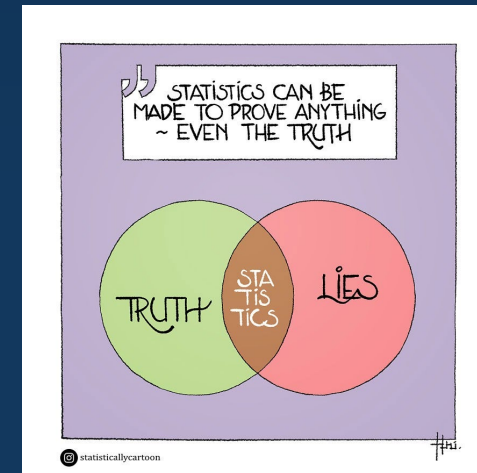
Exploring the World of Statistics: *An Introduction* ¹

Cost estimating requires the analyst to look for trends and patterns in data; Choosing methodologies often involves analysis of key statistics

Definition: A **scientific discipline** that involves the collection, analysis, interpretation, presentation, and organization of **data**. Information is typically gathered through observations, experiments, surveys, or other data collection methods.

Importance: Statistics play a crucial role in informed **decision-making**. The primary goal of statistics is to extract **meaningful** insights and draw valid conclusions from data, enabling informed decision-making and predictions. It encompasses various methods, including descriptive statistics (summarizing and presenting data) and inferential statistics (making predictions or inferences about a population based on a sample).

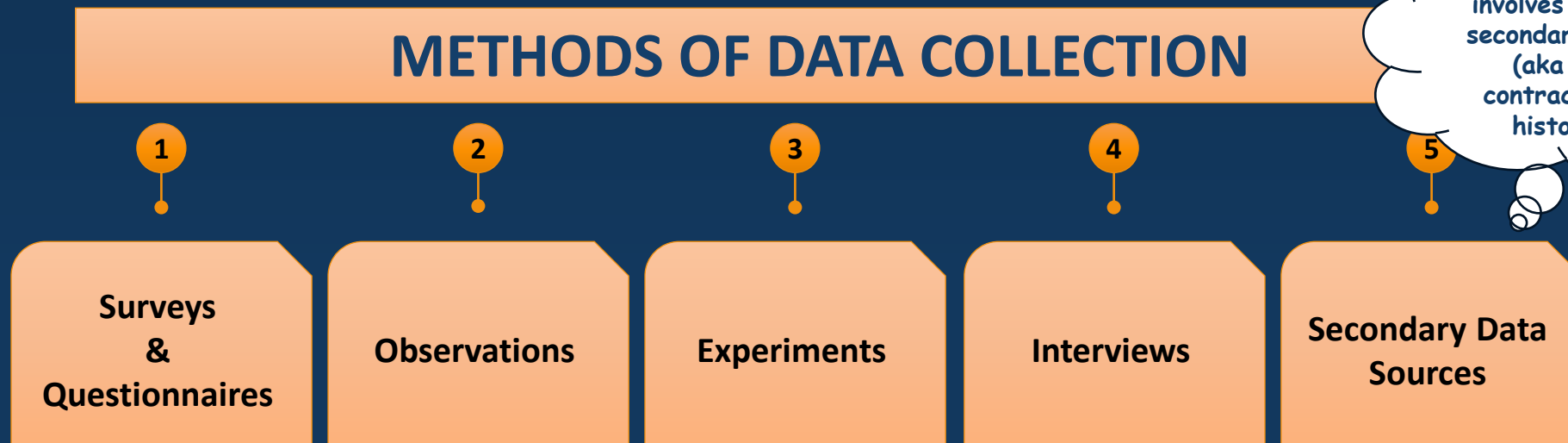
Applicability: Statistics is widely used in **diverse fields** such as economics, psychology, biology, sociology, business, government and many other disciplines to uncover **patterns, relationships, and trends** within datasets.



Data Collection: *Practical Methods* ²

Data collection is a crucial phase in the statistical process, involving the systematic gathering of information to facilitate analysis, interpretation, and decision-making.

The primary purpose of data collection is to provide information that supports the **decision-making** processes. Whether in business, research, or public policy, collecting relevant data helps in making informed and evidence-based choices.



Data collection for cost estimation typically involves interviews and secondary data sources (aka databases, contracts) to look at historical costs.

Data Collection: *Real World Example*

— What is an NBA player's salary?

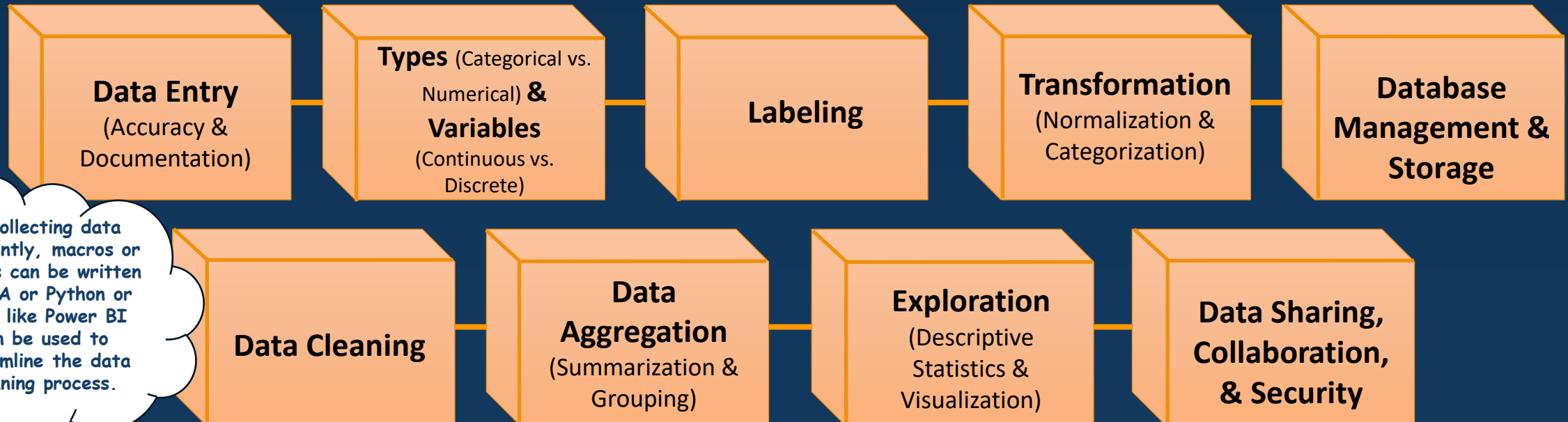
The screenshot shows the ESPN website's NBA section. The main heading is "NBA Player Salaries - 2023-2024". Below this, there is a dropdown menu for the season, currently set to "2023-2024". The main content is a table titled "2023-2024 Player Salaries" with columns for Rank (RK), Name, Team, and Salary. The table lists the top 12 highest-paid players. A red box highlights the "Salaries" option in the "More" dropdown menu on the right side of the page.

RK	NAME	TEAM	SALARY
1	Stephen Curry, PG	Golden State Warriors	\$51,915,615
2	Kevin Durant, PF	Phoenix Suns	\$47,649,433
3	LeBron James, SF	Los Angeles Lakers	\$47,607,350
4	Nikola Jokic, C	Denver Nuggets	\$47,607,350
5	Joel Embiid, C	Philadelphia 76ers	\$46,900,000
6	Bradley Beal, SG	Phoenix Suns	\$46,741,590
7	Giannis Antetokounmpo, PF	Milwaukee Bucks	\$45,640,084
8	Damian Lillard, PG	Milwaukee Bucks	\$45,640,084
9	Kawhi Leonard, SF	LA Clippers	\$45,640,084
10	Paul George, F	LA Clippers	\$45,640,084
11	Jimmy Butler, SF	Miami Heat	\$45,183,960
12	Klay Thompson, SG	Golden State Warriors	\$43,219,440

Data Organization: *Process & Structures* ³

Data organization is a critical step in the statistical process, involving the **structuring and arrangement** of collected data to facilitate analysis, **interpretation, and presentation**. Well-organized data is essential to providing clarity and enables analysts to draw meaningful insights and conclusions.

Sometimes collecting data to fit government formats may seem cumbersome, but having standardization allows for later phases of the data collection process (especially data cleaning) to be easier.

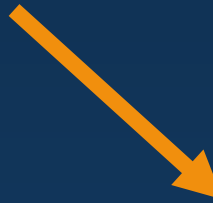


If collecting data frequently, macros or scripts can be written in VBA or Python or tools like Power BI can be used to streamline the data cleaning process.

Data Organization: *Real World Example*

RK	NAME	TEAM	SALARY
1	Stephen Curry, PG	Golden State Warriors	\$51,915,615
2	Kevin Durant, PF	Phoenix Suns	\$47,649,433
3	LeBron James, SF	Los Angeles Lakers	\$47,607,350
4	Nikola Jokic, C	Denver Nuggets	\$47,607,350
5	Joel Embiid, C	Philadelphia 76ers	\$46,900,000
6	Bradley Beal, SG	Phoenix Suns	\$46,741,590
7	Giannis Antetokounmpo, PF	Milwaukee Bucks	\$45,640,084
8	Damian Lillard, PG	Milwaukee Bucks	\$45,640,084
9	Kawhi Leonard, SF	LA Clippers	\$45,640,084
10	Paul George, F	LA Clippers	\$45,640,084
RK	NAME	TEAM	SALARY
11	Jimmy Butler, SF	Miami Heat	\$45,183,960
12	Klay Thompson, SG	Golden State Warriors	\$43,219,440

- Use **Match Destination Format** when pasting data
- Filter for anomalies to quickly clean data
- Functions **Right()**, **Left()** and **Mid()** paired with **Len()** can strip unwanted leading or trailing characters




- Utilize **Text to Columns** to separate data
- Utilize **Find and Replace** to normalize values
- Ensure all cells have intended data types

RANK	NAME	POSITION	TEAM	SALARY
1	Stephen Curry	G	Golden State Warriors	\$51,915,615
2	Kevin Durant	F	Phoenix Suns	\$47,649,433
3	LeBron James	F	Los Angeles Lakers	\$47,607,350
4	Nikola Jokic	C	Denver Nuggets	\$47,607,350
5	Joel Embiid	C	Philadelphia 76ers	\$46,900,000
6	Bradley Beal	G	Phoenix Suns	\$46,741,590
7	Giannis Antetokounmpo	F	Milwaukee Bucks	\$45,640,084
8	Damian Lillard	G	Milwaukee Bucks	\$45,640,084
9	Kawhi Leonard	F	LA Clippers	\$45,640,084
10	Paul George	F	LA Clippers	\$45,640,084
11	Jimmy Butler	F	Miami Heat	\$45,183,960
12	Klay Thompson	G	Golden State Warriors	\$43,219,440

Data Measures: *Overview*

- Measures in statistics refer to various statistical values or metrics that summarize different aspects of a dataset. These measures help in understanding the **central tendency, variability, and distribution of data**.
- These measures are fundamental in statistical analysis and are used to **describe and summarize data, assess relationships, and make meaningful interpretations**. The choice of measures depends on the nature of the data and the specific goals of the analysis.
- This brief will focus on two main areas of measures examining **Central Tendency** (Mean, Median, and Mode) and **Variability** (Range, Variance, and Standard deviation).



When looking at historical cost information, both central tendencies and variability should be considered to select methodologies and risk distributions.

Data Measures: *Central Tendency* 4,5

Mean (Average): The sum of all values divided by the number of observations. It represents the central value of a dataset.

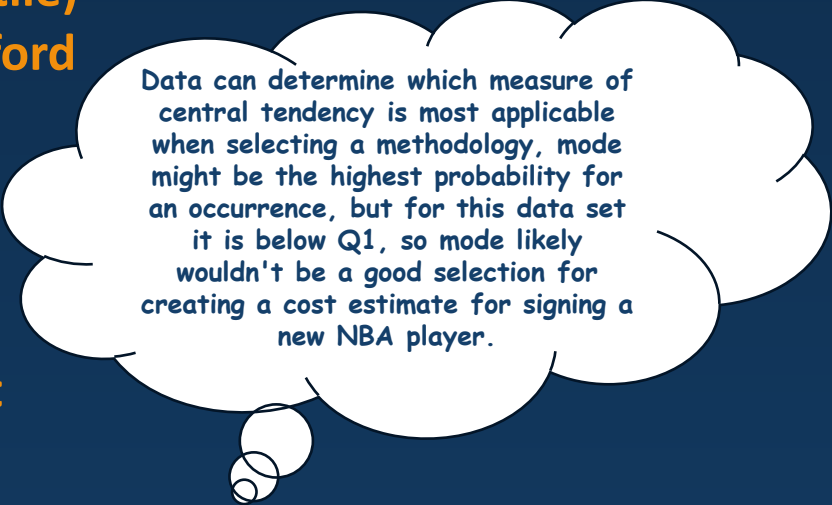
**Mean 2023-2024 Season NBA Salary : \$10,161,384.26 (67th percentile)
Somewhere between Landry Shamet (\$10,250,000.00) and Al Horford (\$10,000,000.00)**

Median: The middle value when the data is arranged in ascending or descending order. It is less sensitive to extreme values than the mean.

**Median 2023-2024 Season NBA Salary : \$5,278,856.50
Midpoint between Cason Wallace (\$5,291,000.00) and Isaiah Stewart / Chuma Okeke (\$5,266,713.00)**

Mode: The value that occurs most frequently in a dataset. A dataset can have one mode, more than one mode (multimodal), or no mode.

**Mode 2023-2024 Season NBA Salary : \$2,019,706.00
Veteran Minimum – 40 Players in 23-24 Season**



Data can determine which measure of central tendency is most applicable when selecting a methodology, mode might be the highest probability for an occurrence, but for this data set it is below Q1, so mode likely wouldn't be a good selection for creating a cost estimate for signing a new NBA player.

Data Measures: *Variability* ^{5,6}

Range: The difference between the maximum and minimum values in a dataset. It provides a simple measure of the spread of data.

Range of 2023-2024 NBA Salaries: \$50,886,132.00

High: Stephen Curry - \$51,915,615.00

Low: Gui Santos - \$1,029,483.00

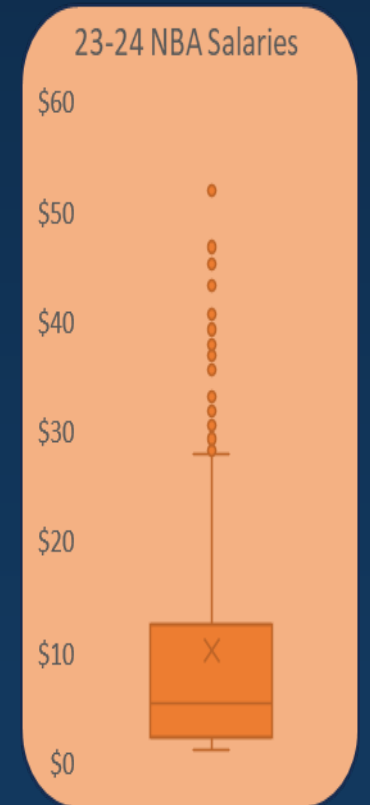
Variance: The average of the squared differences from the mean. It quantifies the degree of variability in a dataset.

Variance of 2023-2024 NBA Salaries: \$²125,846,264,830,440.00

A more common measure of variability in cost estimates is coefficient of variance (Std Dev / Mean), in this example, the CV would be about 1.1 (11.2/10.1), which would be very high for a cost estimate.

Standard Deviation: The square root of the variance. It measures the average distance between each data point and the mean, providing a more interpretable measure of variability.

**Standard Deviation of 2023-2024 NBA Salaries:
\$11,218,122.16**



Data Measures: *Probability* ^{7,8}

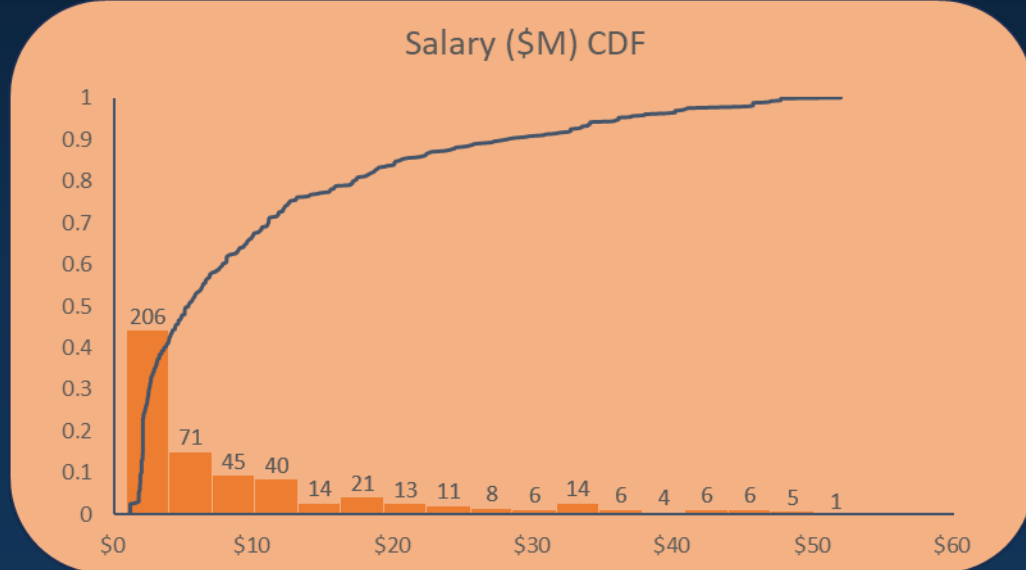
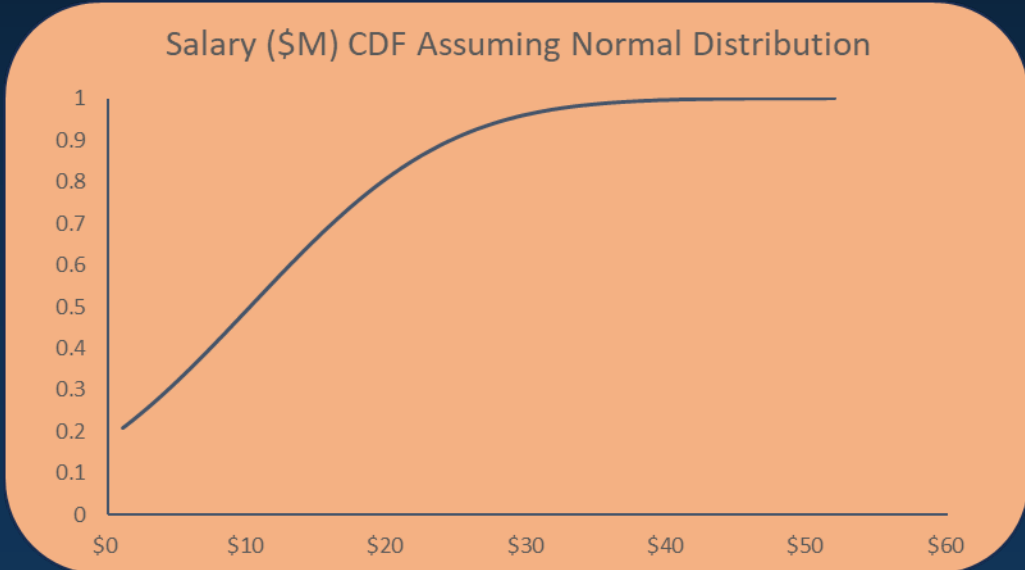
Probability is a principle of mathematics regarding the **likelihood or chance** of a specific event or series of events occurring. It provides a framework for **quantifying uncertainty** and making predictions based on available data and information. Probability is centered around the concept of **experimentation (an uncertain event), outcomes (a possible result), and sample spaces (all possible outcomes)**.

Applications of Probability:

- Risk Assessments: Evaluating the likelihood of undesirable events.
- Data Inference: Forming conclusions about topics based on sample data.
- Game Theory: Analyzing strategic interactions and decision-making.
- Monte Carlo Simulations: Using random sampling to model and analyze complex systems or situations.

Understanding probability is essential in various fields, from science and engineering to finance and decision analysis. It allows for rational decision-making in the face of uncertainty and provides a foundation for statistical inference.

Probability: *Real World Example*



If we were to select a player at random from the NBA, what is the probability that their salary is higher than one of the Minnesota Timberwolves starters salaries?

Player	Salary (Ranking)	Probability of Higher Salary
Mike Conley	\$24,360,000 (58)	12.0%
Anthony Edwards	\$9,219,512 (T171)	35.7%
Jaden McDaniels	\$3,901,399 (277)	58.0%
Karl-Anthony Towns	\$36,016,200 (T24)	4.8%
Rudy Gobert	\$41,000,000 (13)	2.5%

Cumulative distribution functions are known in cost estimates as "S-curves". The CDF on the left utilizes the mean and std dev of the data set and assumes it is normally distributed, here we see a gentle ramp up, like what might be seen in most cost estimates. The CDF on the right shows the right skew in the data set, with a steeper ascent in the curve and a longer right tail.

Statistical Inference: *Sampling*⁹

A **sample** is a subset of the **population** selected for a study or analysis where the population represents the entire set of individuals, items, or data point within a certain criteria.

Sampling in cost estimating is seen when utilizing parametric models. Data is stratified when selecting applicable characteristics for your program, returning a sample data set more appropriate for your methodology development.

Stratified Sampling

Convenience Sampling



Systematic Sampling

Cluster Sampling

Simple Random Sampling

Sampling: *Real World Example*

	Minnesota Timberwolves	Entire NBA
Observations	16	476
Mean	\$10,135,754.19	\$10,161,384.26
Median	\$4,343,750.00	\$5,278,856.50
Mode	\$9,219,512 / \$1,719,864	\$2,019,706.00
Minimum	\$1,719,864.00	\$1,029,483.00
Maximum	\$41,000,000.00	\$51,915,615.00
Interquartile Range	\$7,953,210.00	\$10,338,100.00
Range	\$39,280,136.00	\$50,886,132.00
Standard Deviation	\$12,563,225.51	\$11,218,122.16
Variance	\$ ² 157,834,635,299,288.00	\$ ² 125,846,264,830,440.00

Our salary data has a few categorical variables (position and team) that could be used to create a stratified or clustered sample. Let's look at a clustered sample, selecting a team at random and compare their salary statistics versus the entire league.

Statistical Inference: *Correlation & Causation* ¹⁰

Correlation and causation are fundamental concepts in statistics and research that describe the **relationships** between variables. Understanding the distinction between them is crucial for making **accurate interpretations and informed decisions**.

Correlation is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where: +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear correlation.

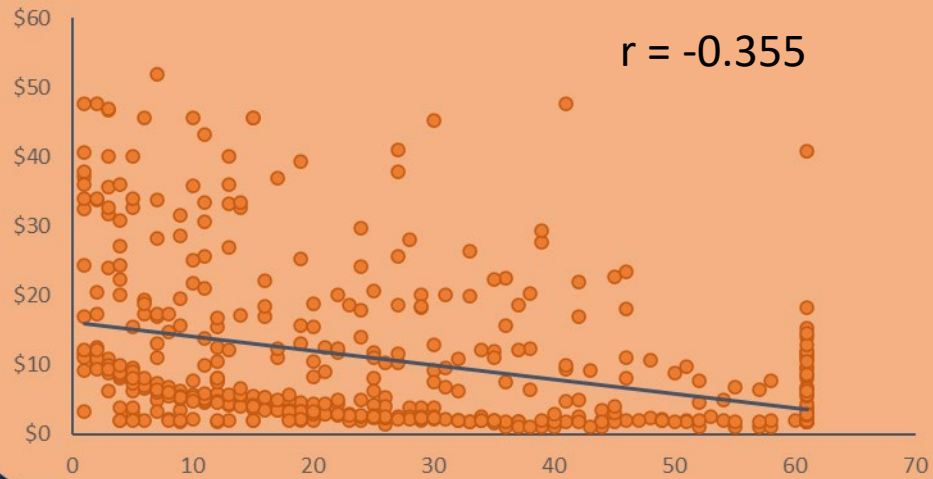
Two variables being correlated does not mean that one causes the other. Remember to not assume causation, as a third unknown variable could be an influencing factor. Understanding the nuances between correlation and causation is essential for drawing valid conclusions from data and avoiding misinterpretations or unwarranted assumptions in research, decision-making and customer interactions.

Causation implies that a change in one variable directly causes a change in another. Establishing causation is more complex than identifying correlation, and additional evidence and research are often required.

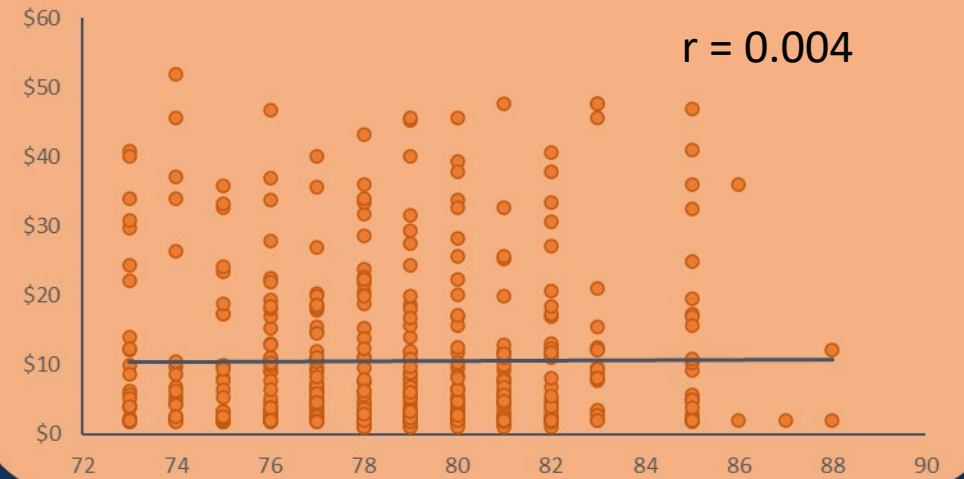


Correlation & Causation: *Real World Example*

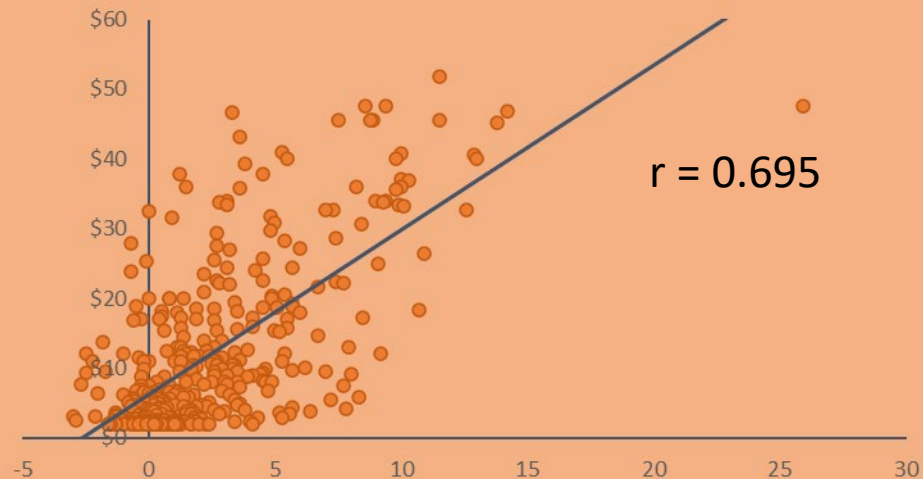
Pick Drafted vs. Salary



Height (inches) vs. Salary



Wins Above Replacement vs Salary



Statistical Inference: *Distributions* ¹¹

Statistical distributions are mathematical functions that describe the **likelihood** of different outcomes in a sample space. These distributions are fundamental in statistics and probability theory and provide a way to **understand the patterns and characteristics of data**.

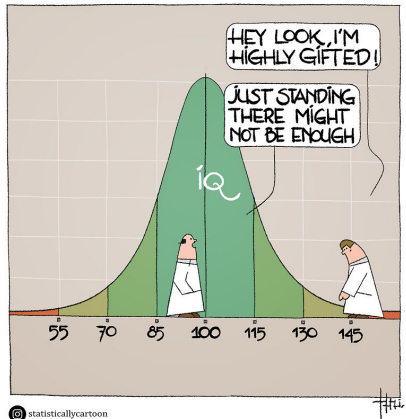
Types of Distributions:

- *Discrete Distributions*: Describes probabilities for individual values (Bernoulli, Binomial, Poisson).
- *Continuous Distributions*: Describes probabilities for a range of values (Normal, Uniform, Exponential).

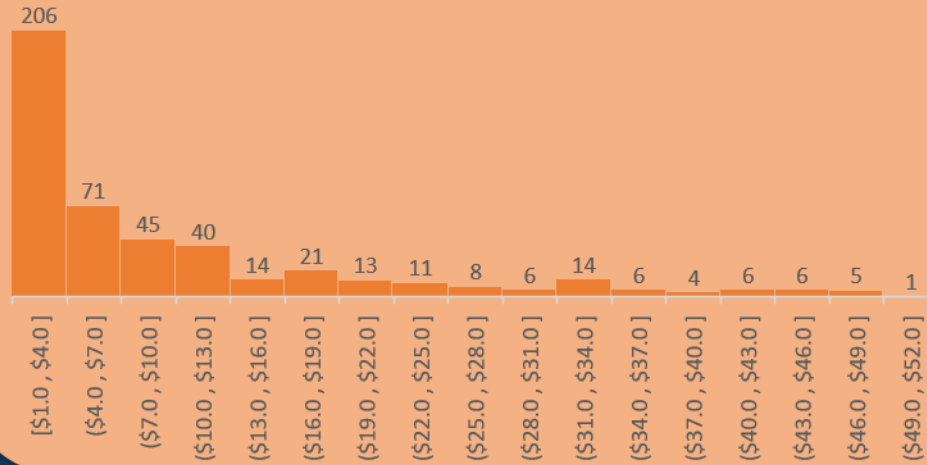
Skewness measures the **asymmetry** of a distribution. Positive skewness indicates a longer right tail, and negative skewness indicates a longer left tail.

Distributions are essential in determining **confidence intervals** that help estimate the **range** within which a parameter is likely to fall. Distributions are also useful when performing **hypothesis testing** helping compare sample statistics against expected values.

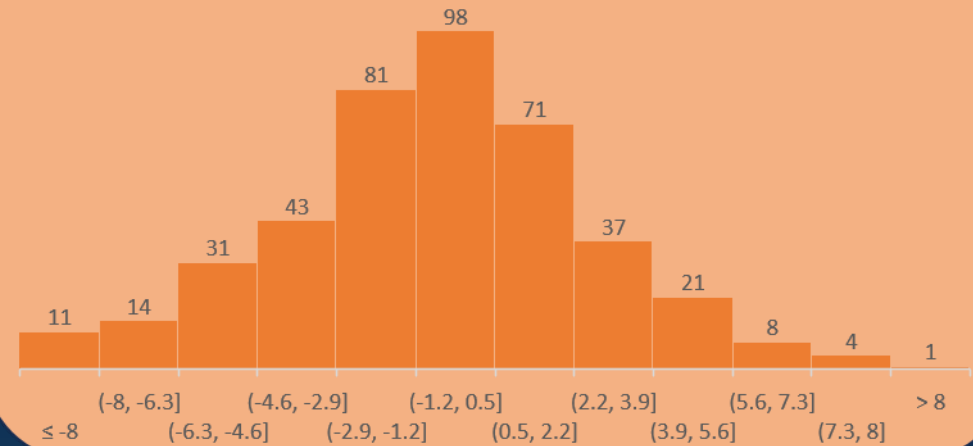
Statistical Distributions: *Real World Example*



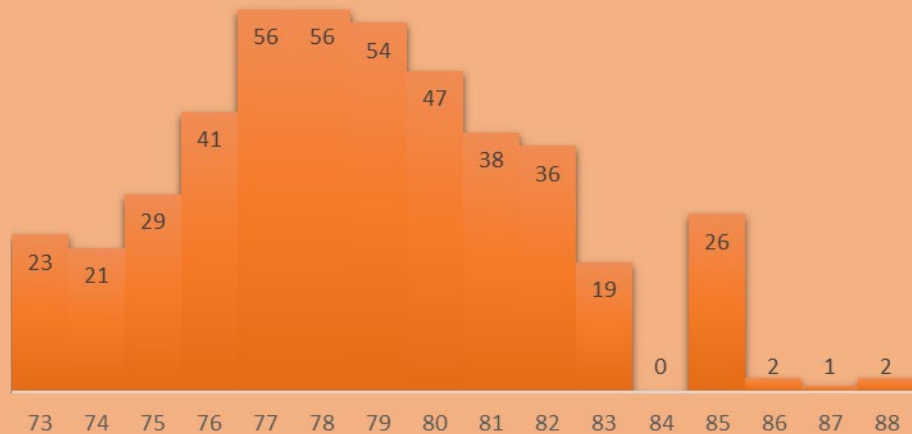
2023-2024 NBA Salaries



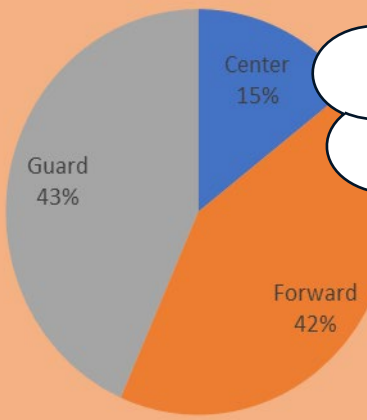
2022- 2023 RAPTOR +/-



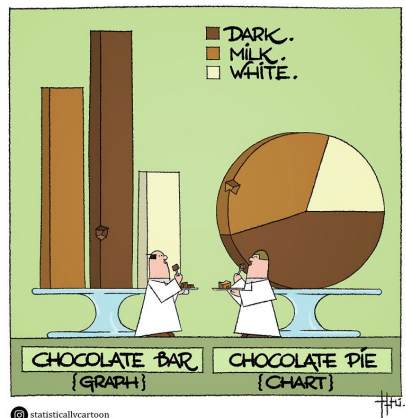
NBA Players' Height (Inches)



Summary of Players by Position



Plotting histograms for large data sets provides insight into the shape of the distribution and can inform selection of appropriate risk distributions for methodologies.



Statistical Inference: *Regression Analysis* ^{11,12}

Regression analysis is a statistical technique used to examine the **relationship** between one or more independent variables and a dependent variable.

TYPES OF REGRESSION ANALYSIS

Simple Linear Regression	Multiple Linear Regression	Polynomial Regression	Logistic Regression	Ridge & Lasso Regression
Examines the relationship between one independent variable and one dependent variable	Involves two or more independent variables to predict a single dependent variable	Extends linear regression by introducing polynomial terms to capture non-linear relationships	Used for binary or categorical dependent variables, estimating the probability of an event occurring	Techniques that add regularization to linear regression to prevent overfitting

Linear Regression Equation: $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_n \cdot X_n + \epsilon$, where Y is the dependent variable, Xi are the independent variables, bi are the coefficients, and ϵ is the error term. Coefficients (b_i) represent the slope of the relationship between each independent variable and the dependent variable. Intercept (b_0) represents the value of the dependent variable when all independent variables are zero.

Statistical Inference: *Regression Analysis Cont.*

Assumptions of Regression Analysis:

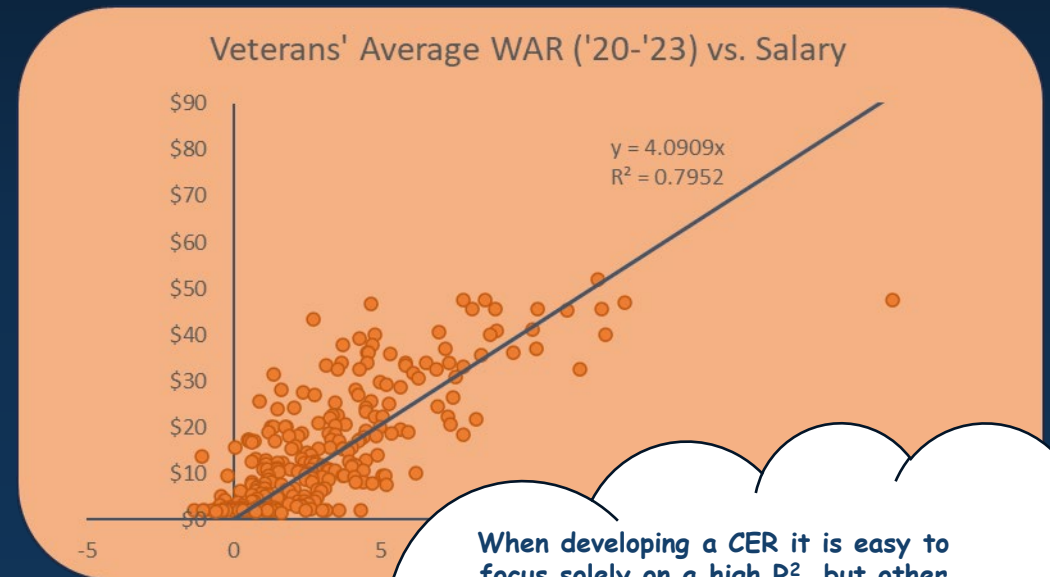
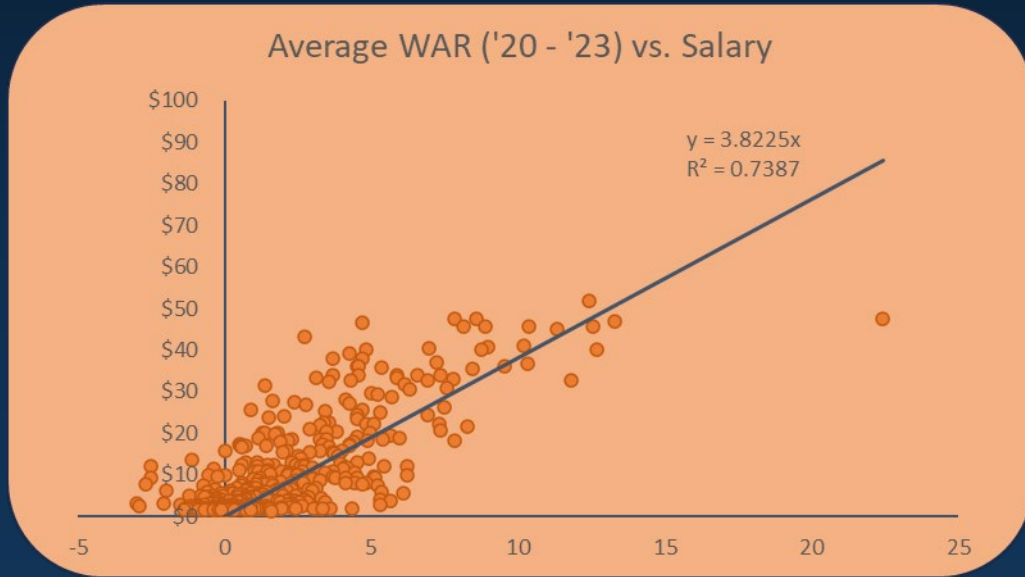
- Linearity: The relationship between independent and dependent variables is linear.
- Independence: Residuals (errors) are independent of each other.
- Homoscedasticity: Residuals have constant variance across all levels of independent variables.
- Normality: Residuals are normally distributed.
- No Perfect Multicollinearity: Independent variables are not perfectly correlated.

Challenges and Limitations:

- Assumption Violations: Violations of model assumptions can lead to biased estimates.
- Overfitting: Including too many variables may result in a model that fits the training data too closely but does not generalize well to new data.

Regression Analysis Error occurs because of **residuals** or differences between observed and predicted values in regression analysis. This error can be measured using the **Mean Squared Error (MSE)** to calculate the average squared difference between observed and predicted values in regression.

Regression Analysis: *Real World Example*



Predicting Salary from Avg WAR	Full Data Set	Veterans Only
Equation	Salary = 3.8225*WAR	Salary = 4.0909*WAR
R-Squared	0.7387	0.7952
Average Actual	\$10.9 M	\$15.0 M
Standard Error of Regression	\$8.2 M	\$8.9 M
RMS of % Errors	140.7%	120.6%
MAD of % Errors	94.9%	73.8%
CV based on Std Error	74.7%	59.7%
CV based on MAD Residuals	53.5%	43.0%

When developing a CER it is easy to focus solely on a high R², but other statistics need to be looked at as well to ensure the fit is good and not overly influenced by larger data points. Root Mean Squared Error is an important statistic to look at for any potential CERs as poor fits for smaller data points would be more visible, like Mean Absolute Deviation checking to make sure there aren't large portions of the data that are all overestimated or underestimated systematically.

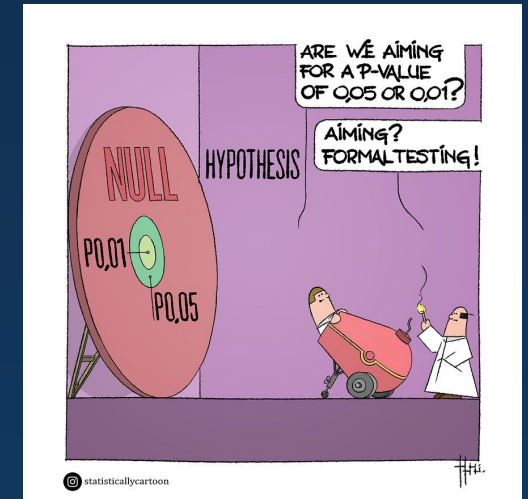
Hypothesis Testing: *Experimental Design* 14,15,16

Hypothesis testing is a statistical method used to make **inferences** about population parameters based on a sample of data. It involves formulating hypotheses, collecting and analyzing data, and drawing **conclusions** about the population from which the sample is drawn.

Basic Structure of Hypothesis Testing

- Null Hypothesis (H_0): Represents a default assumption or status quo about a population parameter. It often states that there is no effect or no difference.
- Alternative Hypothesis (H_1 or H_a): Represents a specific claim or statement about the population parameter. It asserts the presence of an effect or a difference.
- Test Statistic: A numerical summary of the sample data used to make decisions about the null hypothesis.
- Significance Level (α): The probability of rejecting the null hypothesis when it is true. Common choices include 0.05 or 0.01.

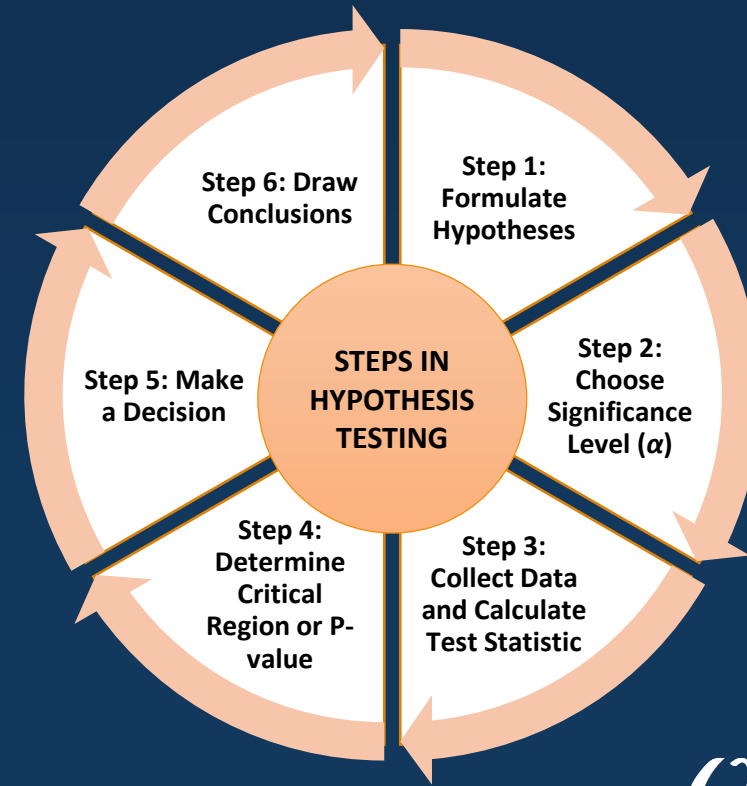
The **probability** of obtaining a test statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true is known as the **P-Value**. If $P\text{-value} \leq \alpha$, reject H_0 . If $P\text{-value} > \alpha$, fail to reject H_0 .



Hypothesis Testing: *Experimental Design (Continued)* ¹⁷

Hypothesis testing is a crucial tool in scientific research, allowing researchers to make evidence-based decisions and draw conclusions about population parameters. Careful consideration of test design, sample size, and effect size enhances the reliability and validity of hypothesis testing results.

<u>TYPES OF HYPOTHESIS TESTS</u>	
One-Sample Test	Compares the mean of a single sample to a known or hypothesized population mean
Two-Sample Test	Compares the means of two independent samples
Paired Sample Test	Compares the means of two related or paired samples
Chi-Squared Test	Tests for independence or goodness-of-fit in categorical data
Analysis of Variance	Tests for differences in means among more than two groups



Hypothesis Testing: *Bias and Error* 18,19

Bias and error are terms commonly used in statistics and research to describe the **deviation** between an estimated or observed value and the true or population value. Understanding these concepts is crucial for assessing the validity and reliability of research findings.

Bias refers to the **systematic error** introduced in the process of data collection, analysis, or interpretation that consistently skews the results in a particular direction. Examples include Selection Bias, Measurement Bias, Recall Bias, Observer Bias, etc.

Error refers to the difference between an observed or estimated value and the true or population value. Errors can be **random or systematic**. Random Error is unpredictable and occurs by chance, it tends to cancel out when averaged over multiple measurements but can affect precision. Systematic error is consistent and occurs in the same direction, it can lead to inaccurate measurements and affect the validity of conclusions.

There are two types of Hypothesis Testing Errors - **Type I** (False Positive) rejecting a true null hypothesis, and **Type II** (False Negative) failing to reject a false null hypothesis. Adjusting the significance level (α) and sample size helps balance Type I and Type II errors.

Bias can lead to overestimation or underestimation of the true values and compromise the internal and external validity of a study. Randomization, blinding, and careful study design are strategies to mitigate bias.

Hypothesis Testing: *Real World Example*

Cost analysts don't often conduct t-tests, but the test results are seen when developing CERs. Coefficients are tested to ensure they are statistically different than zero and p-value results are provided, if those p-value results are less than your significance level you can be confident the independent variable influences the variable that is being predicted.

- Question: Do all NBA teams pay the same average salaries?
 - Set up Two-Sample T Test to test for difference in sample means from different teams
- Formulate Hypothesis
 - Null Hypothesis (H_0): $\mu_{\text{Timberwolves}} = \mu_{\text{Warriors}}$
 - Alternative Hypothesis (H_a): $\mu_{\text{Timberwolves}} \neq \mu_{\text{Warriors}}$
- Choose Significance Level: $\alpha = 0.05$
- Collect Data and Calculate Test Statistic

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} = \frac{\$10.1 - \$13.6}{\sqrt{\frac{\$12.6^2}{16} + \frac{\$16.1^2}{16}}} = \frac{-\$3.4}{\$5.1} = -0.67$$

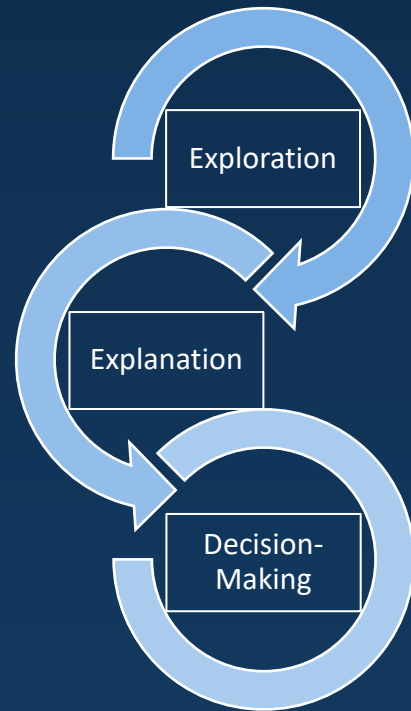
Team	Timberwolves	Warriors
Sample Mean (\bar{x})	\$10,135,754	\$13,583,841
Sample Std Dev (s)	\$12,563,226	\$16,132,306
Sample Size (n)	16	16

- Compare to critical value
 - Critical value for two-tailed t distribution with 15 degrees of freedom is 2.131
- Make Decision
 - T-statistic absolute value is less than critical value, fail to reject Null Hypothesis
 - Fail to reject Null Hypothesis, cannot prove NBA teams pay different average salaries

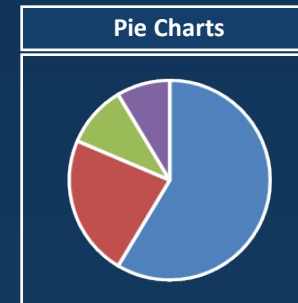
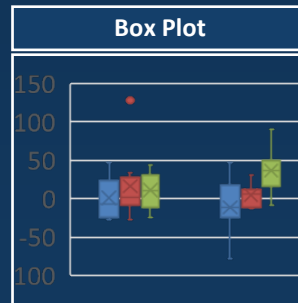
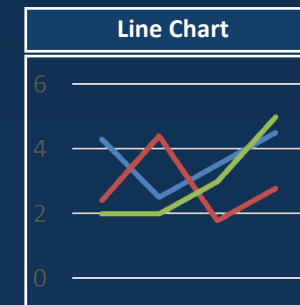
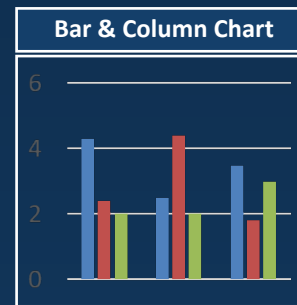
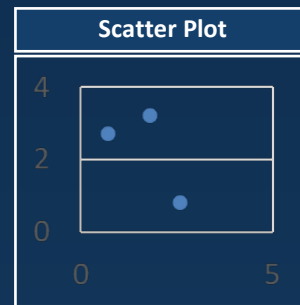


Data Visualization

Data visualization is the presentation of data in a graphical or visual format to facilitate understanding and interpretation. It is a critical component of data analysis, helping to communicate patterns, trends, and insights more effectively than raw data.



Types of Data Visualizations



When presenting information to senior leadership to inform decision making it is often easier for them to visualize the data, selecting the correct visualization for the type of data you have can be a critical part of the presentation.

Additional Chart Types to Consider: Maps, Geospatial Visualizations, Dashboards, Heatmaps, Network Diagrams, and many more!

Conclusion

- Statistics forms the backbone of cost estimation
- Finding a data source is half of the battle; the data still needs to be cleaned and organized before use
- Data analysis starts by examining central tendencies and the variability of the data set
- Selecting appropriate distributions can help assess the probability of events occurring
- Forming a meaningful relationship with regression analysis requires more than just a high R^2 value
- Hypothesis testing assesses the likelihood that results occurred by chance
- Data visualizations help effectively convey results to a wider audience

References

- ¹ Williams, Thomas A. , Anderson, David R. and Sweeney, Dennis J.. "statistics". *Encyclopedia Britannica*, 29 Nov. 2023, <https://www.britannica.com/science/statistics>
- ² Berman H.B., "Data Collection Methods", [online] Available at: <https://stattrek.com/statistics/data-collection-methods> URL
- ³ "Data Cleaning: What It Is, Why It Matters & How to Do It", [online] Available at: <https://www.sigmacomputing.com/resources/learn/what-is-data-cleaning?> URL
- ⁴ Berman H.B., "Mean and Median", [online] Available at: <https://stattrek.com/descriptive-statistics/mean-median> URL
- ⁵ Berman H.B., "What Are Variables in Statistics?", [online] Available at: <https://stattrek.com/descriptive-statistics/variables> URL
- ⁶ Berman H.B., "Mean and Variance of Random Variables", [online] Available at: <https://stattrek.com/random-variable/mean-variance> URL
- ⁷ Berman H.B., "How to Measure Variability in Quantitative Data", [online] Available at: <https://stattrek.com/descriptive-statistics/variability> URL
- ⁸ Berman H.B., "What is Probability?", [online] Available at: <https://stattrek.com/probability/what-is-probability> URL
- ⁹ Berman H.B., "Survey Sampling Methods", [online] Available at: <https://stattrek.com/survey-research/sampling-methods> URL
- ¹⁰ Berman H.B., "Correlation Coefficient", [online] Available at: <https://stattrek.com/statistics/correlation> URL
- ¹¹ Berman H.B., "Data Patterns in Statistics", [online] Available at: <https://stattrek.com/statistics/charts/data-patterns> URL
- ¹² Berman H.B., "Regression Tutorial", [online] Available at: <https://stattrek.com/tutorials/regression-tutorial> URL
- ¹³ Berman H.B., "Residual Analysis in Regression", [online] Available at: <https://stattrek.com/regression/residual-analysis> URL
- ¹⁴ Berman H.B., "What is an Experiment?", [online] Available at: <https://stattrek.com/experiments/what-is-an-experiment> URL
- ¹⁵ Berman H.B., "Experimental Design", [online] Available at: stattrek.com/experiments/experimental-design URL
- ¹⁶ Berman H.B., "What is a Confidence Interval?", [online] Available at: <https://stattrek.com/estimation/confidence-interval> URL
- ¹⁷ Berman H.B., "What is Hypothesis Testing?", [online] Available at: <https://stattrek.com/hypothesis-test/hypothesis-testing> URL
- ¹⁸ Berman H.B., "Bias in Survey Sampling", [online] Available at: <https://stattrek.com/survey-research/survey-bias> URL
- ¹⁹ Stewart, Ken. "mean squared error". *Encyclopedia Britannica*, 30 Mar. 2023, <https://www.britannica.com/science/mean-squared-error>
- "NBA Player Salaries – 2023-2024", [online] Available at: <https://www.espn.com/nba/salaries> URL
- "2023-2024 NBA Players", [online] Available at: <https://basketball.realgm.com/nba/players#> URL
- "2022-23 Hollinger NBA Player Statistics – All Players", [online] Available at: https://www.espn.com/nba/hollinger/statistics/_/year/2023 URL
- "The Best NBA Players, According to RAPTOR", [online] Available at: <https://projects.fivethirtyeight.com/nba-player-ratings/> URL