

# ESSENTIAL STATISTICS AND PROBABILITY WITH PYTHON PROGRAM

---

Vijay Dwivedi

# ELEMENTS OF STRUCTURED DATA

---

- Data comes from many sources! Much of this data is unstructured.
- There are two basic types of structured data: **numeric** and **categorical**.

Numeric		
Continuous	Any value or interval	Interval, float, numeric
Discrete	Only integer values	Integer, count
Categorical		
Categorical	Categories	Enums, factors, nominal
Binary	Two categories	Dichotomous, logical, Boolean
Ordinal	Categorical that has explicit order	Ordered factor

# RECTANGULAR DATA

---

- The typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table
- Data frame:
  - Rectangular data is the basic data structure for statistical and machine learning models
- Feature:
  - A column in the table is commonly referred to as a feature. (attribute, input, predictor, independent variable)
- Outcome:
  - The features are sometimes used to predict the outcome. (dependent variable, response, target, output)
- Records:
  - A row in the table is commonly referred to as a record. (case, example, instance, observation, pattern, sample)

# DATA FRAMES AND INDEXES

---

## R

- `data.frame` object
- An automatic index is created for a `data.frame` based on the order of the rows.
- Doesn't support multilevel indexes. To overcome this use `data.table` or `dplyr` libraries.

## PYTHON

- `DataFrame` object (pandas).
- An automatic index is created for a `DataFrame` based on the order of the rows.
- Pandas can handle multiple indexes.

# NONRECTANGULAR DATA STRUCTURES

---

- **Time series** records successive measurements of the same variable.
- **Spatial data structures**, used in mapping and location analytics.
- **Graph (network)** used to represent physical, social and abstract relationships.



# ESTIMATES

---

- **Estimates:** values calculated from the data at hand, to draw a distinction between what we see from the data and the theoretical or true value. Data scientist and business analyst are more likely to refer to such values as **metric**.
- Bias: difference between the expected value of an estimator and the true value.

$$Bias_{\theta} = E[\hat{\theta}] - \theta$$

- If the bias of an estimator is 0, we have an unbiased estimator.

# ESTIMATES OF LOCATION

---

- A basic step in exploring data is getting a “**typical value**” for each feature an estimate of where most of the data is located.
- **Mean** (average): sum of all values divided by the number of value.

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

\*probe that the mean is an unbiased estimator.

# ESTIMATES OF LOCATION

---

- Weighted mean: sum of all values times a weight divided by the sum of weights.
  - When some values are intrinsically more variable than others, and highly variable observations are given a lower weight.

$$\bar{x}_w = \frac{\sum_i^n w_i x_i}{\sum_i^n w_i}$$

- Trimmed mean: the average of all values after dropping a fixed number of extreme values.
  - Eliminates the influence of extreme values.

$$\bar{x} = \frac{\sum_{i=p}^{n-p} x_{(i)}}{n - 2p}$$



# ROBUST ESTIMATES

---

- **Robust:** Not sensitive to extreme values
- **Outlier:** A data value that is very different from most of the data
- **Median (50<sup>th</sup> percentile):** The value such that one half of the data lies above and below.
- **Weighted median:** The value such that one half of the sum of the weights lies above and below the sorted data.

# ESTIMATES OF VARIABILITY

---

- **Variability**, also referred to as dispersion, measures whether the data values are tightly clustered or spread out.
- **Deviations**: The difference between the observed values and the estimate of location (errors, residuals)

$$e_i = x_i - \bar{x}$$

- **Variance**: the sum of squared deviations from the mean divided by  $n-1$  where  $n$  is the number of data values.

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

- **\*probe that the variance is an unbiased estimator.**

# ESTIMATES OF VARIABILITY

---

- **Standard deviation:** The square root of the variance
  - Is much easier to interpret than the variance since it is in on the same scale as the original data.
- **Mean absolute deviation:** The mean of the absolute value of the deviation from the mean.

$$\text{mean absolute deviation} = \frac{\sum_i^n |x_i - \bar{x}|}{n}$$

- **Median absolute deviation (MAD):** The median of the absolute value of the deviation from the median.

$$MAD = \text{Median}(|x_1 - m|, |x_2 - m|, \dots, |x_N - m|)$$

# ESTIMATED BASED ON PERCENTILES

---

- A different approach to estimating dispersion is based on looking at the spread of the sorted data. Also known as **order statistics**.
- **Range:** The difference between the largest and the smallest value in a data set. (ranks)
- **Percentile:** The value such that P percent of the values take on this value or less and (100-P) percent take on this value or more. (quantile)
- **Interquartile range:** The difference between the 75<sup>th</sup> percentile and the 25<sup>th</sup> percentile. (IQR)

# EXPLORING THE DATA DISTRIBUTION

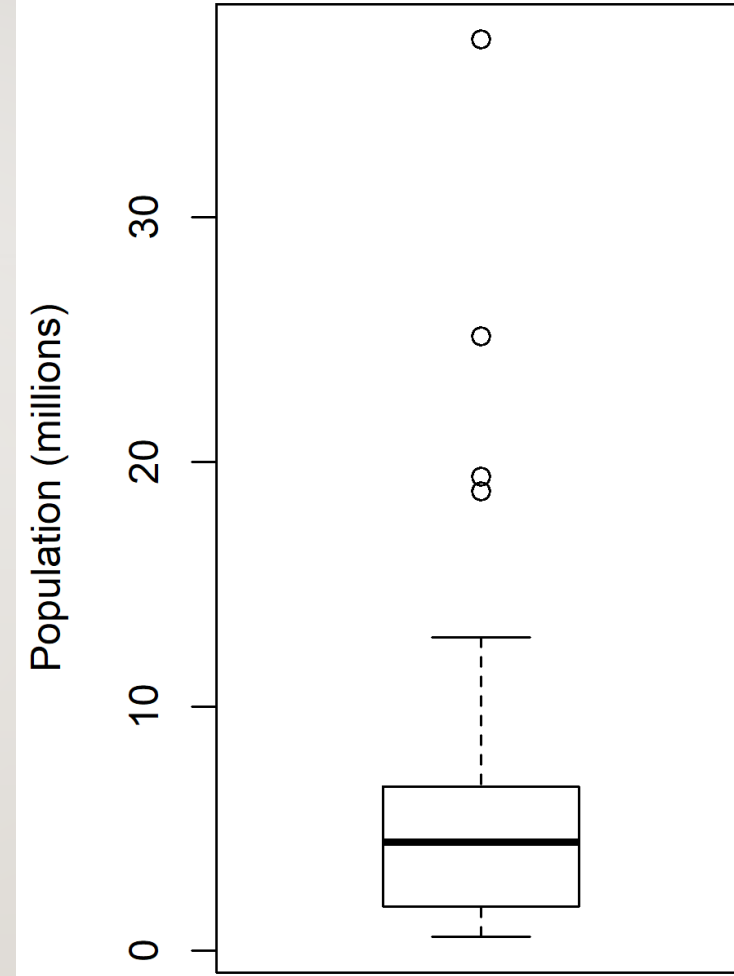
---

- Estimators are useful to explore how data is distributed, overall.
- There are a groups of tools that provide us insights about the data distribution:
  - Boxplot (box and whiskers plot)
  - Frequency table
  - Histogram
  - Density plot



# PERCENTILES AND BOXPLOTS

---



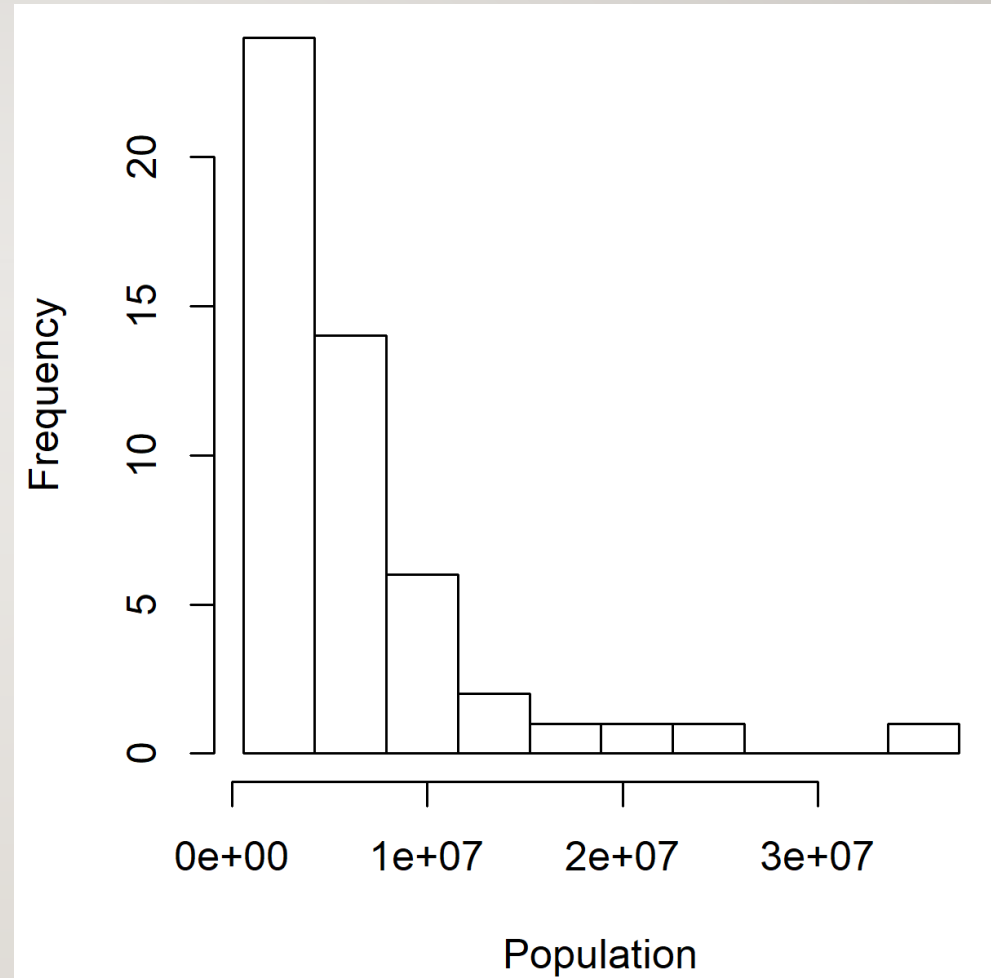
# FREQUENCY TABLES

---

BinNumber	BinRange	Count	States
1	563,626-4,232,658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,NV,UT,K...
2	4,232,659-7,901,691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7,901,692-11,570,724	6	VA,NJ,NC,GA,MI,OH
4	11,570,725-15,239,757	2	PA,IL
5	15,239,758-18,908,790	1	FL
6	18,908,791-22,577,823	1	NY
7	22,577,824-26,246,856	1	TX
8	26,246,857-29,915,889	0	
9	29,915,890-33,584,922	0	
10	33,584,923-37,253,956	1	CA

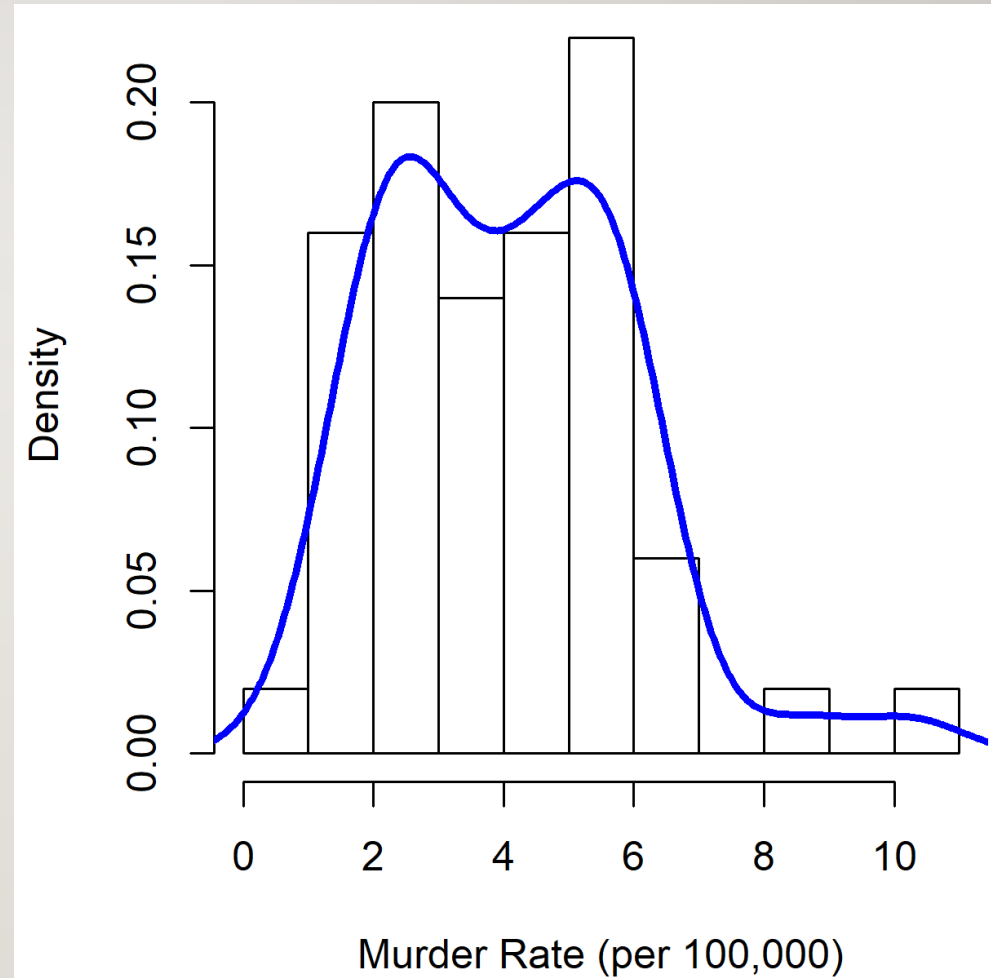
# HISTOGRAMS

---



# DENSITY ESTIMATES

---



# EXPLORING BINARY AND CATEGORICAL DATA

---

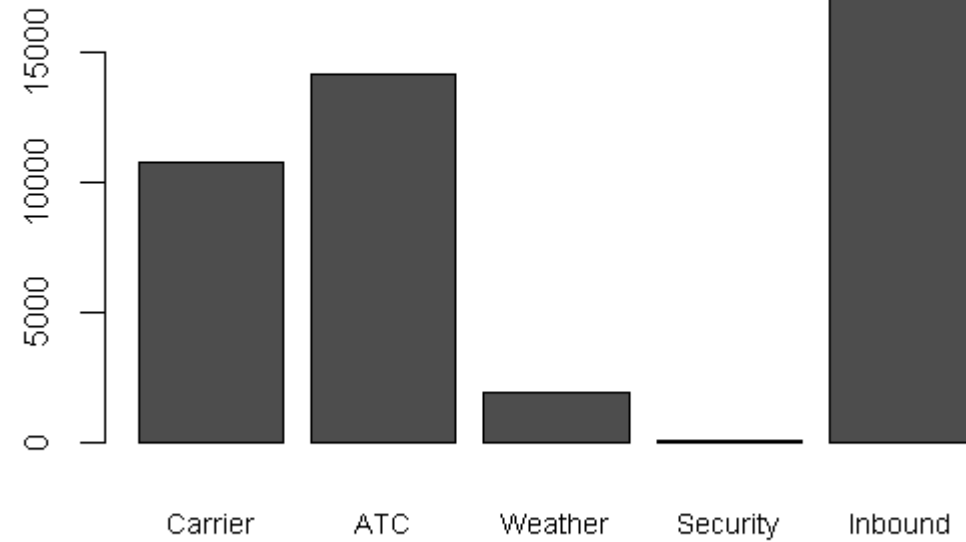
- For categorical data, simple proportions or percentages tell the story of the data.
- **Mode:** The most commonly occurring category or value in a data set.
- **Expected value:** When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.
- **Bar charts:** The frequency or proportion for each category plotted as bars.
- **Pie charts:** The frequency or proportion for each category plotted as wedges in a pie.



## BAR CHART

---

Bar chart resembles a histogram' in a bar chart the x-axis represents different categories of a factor variable, while in a histogram the x-axis represents values of a single variable on a numeric scale.



# EXPECTED VALUE

---

- Expected value is the sum of values times their probability of occurrence, often used to sum up factor variable levels.
- Example: A new cloud technology offers two levels of service. Service A is priced at \$300/month and service B at \$50/month. 5% of webinar attendees will sign up for the \$300 service, 15% for the \$50 service and 80% will not sign up for anything.

$$EV = (0.05)(300) + (0.15)(50) + (0.80)(0) = 22.5$$

# CORRELATION

---

- Correlation coefficient: A metric that measures the extent to which numeric variables are associated with one another (ranges from -1 to 1)

$$r = \frac{\sum_1^N (x_i - \bar{x})(y_i - \bar{y})}{(N - 1)s_x s_y}$$

\* Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric.

\*\* Correlation coefficient is sensitive to outliers in the data.

\*\*\* The definition above corresponds to Pearson's correlation definition.

- Correlation matrix: A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

## TWO CATEGORICAL VARIABLES

**Contingency tables:** A tally of counts between two or more categorical variables.

Grade	Charged Off	Current	Fully Paid	Late	Total
A	1562	50051	20408	469	72490
	0.022	0.690	0.282	0.006	0.161
B	5302	93852	31160	2056	132370
	0.040	0.709	0.235	0.016	0.294
C	6023	88928	23147	2777	120875
	0.050	0.736	0.191	0.023	0.268
D	5007	53281	13681	2308	74277
	0.067	0.717	0.184	0.031	0.165
E	2842	24639	5949	1374	34804
	0.082	0.708	0.171	0.039	0.077
F	1526	8444	2328	606	12904
	0.118	0.654	0.180	0.047	0.029
G	409	1990	643	199	3241
	0.126	0.614	0.198	0.061	0.007
Total	22671	321185	97316	9789	450961

# CATEGORICAL AND NUMERIC DATA

---

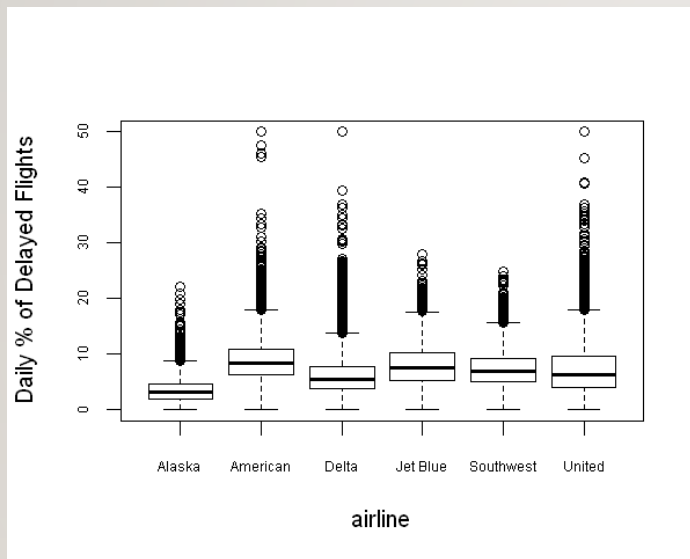
- Boxplots are a simple way to visually compare the distribution of a numeric variable grouped according to a categorical variable.
- A violin plot is an enhancement to the boxplot and plots the density estimate with the density on the y-axis. The advantage of a violin plot is that it can show nuances in the distribution that aren't receptive in a boxplot.



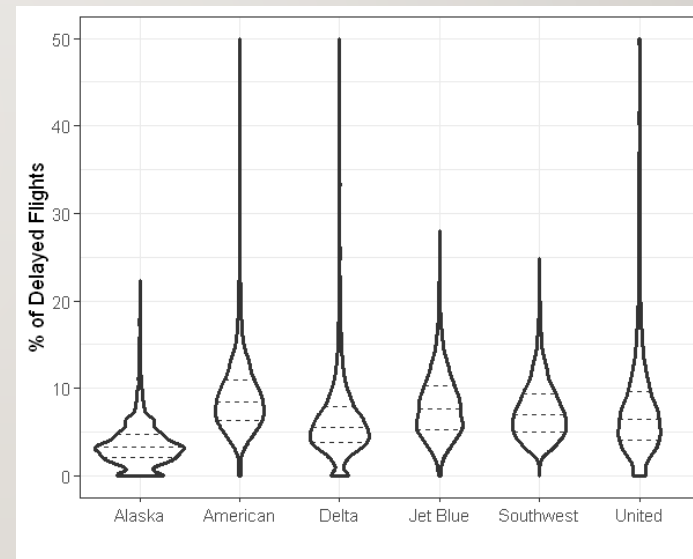
# CATEGORICAL AND NUMERIC DATA

---

## BOXPLOT

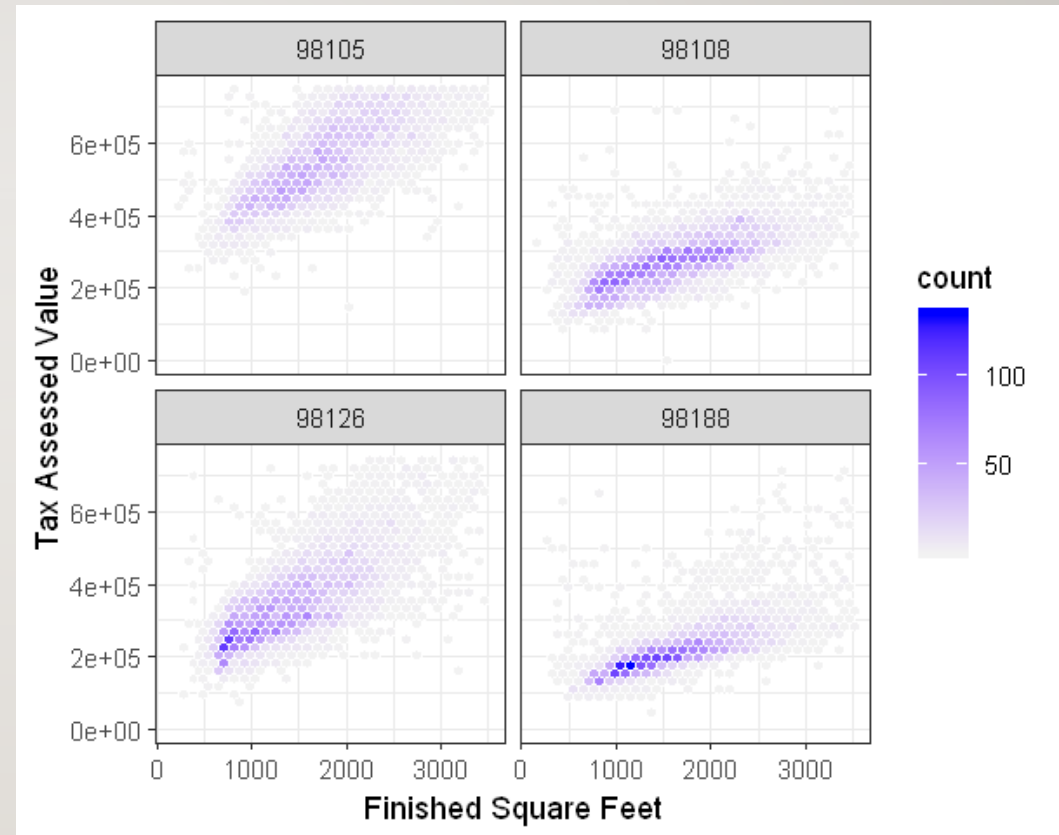


## VIOLIN PLOT



# VISUALIZING MULTIPLE VARIABLES

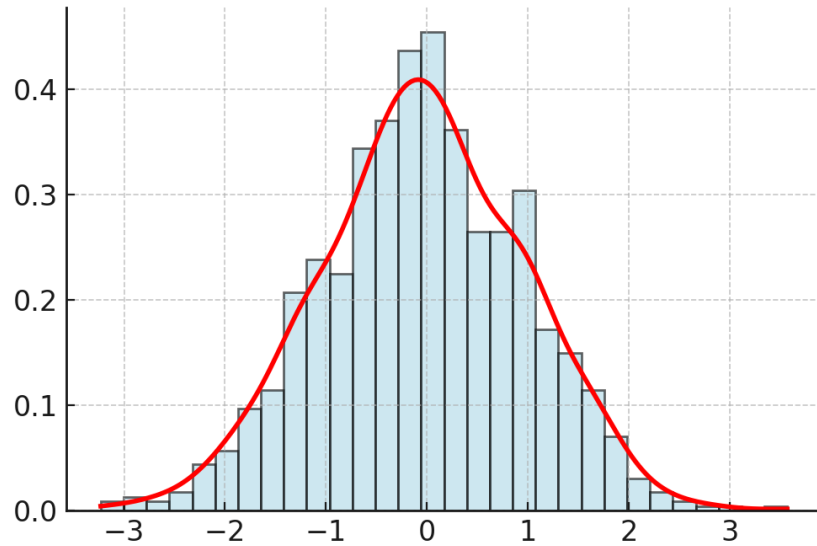
---



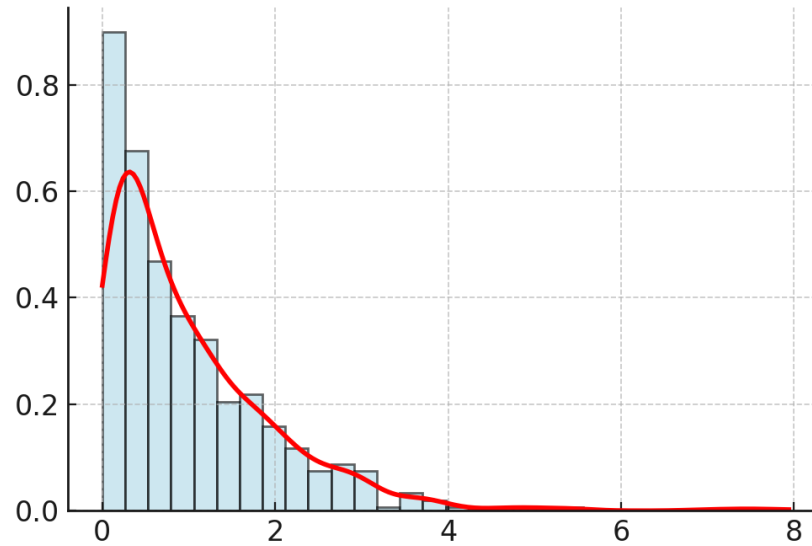
# SKEWNESS & KURTOSIS PLOT

---

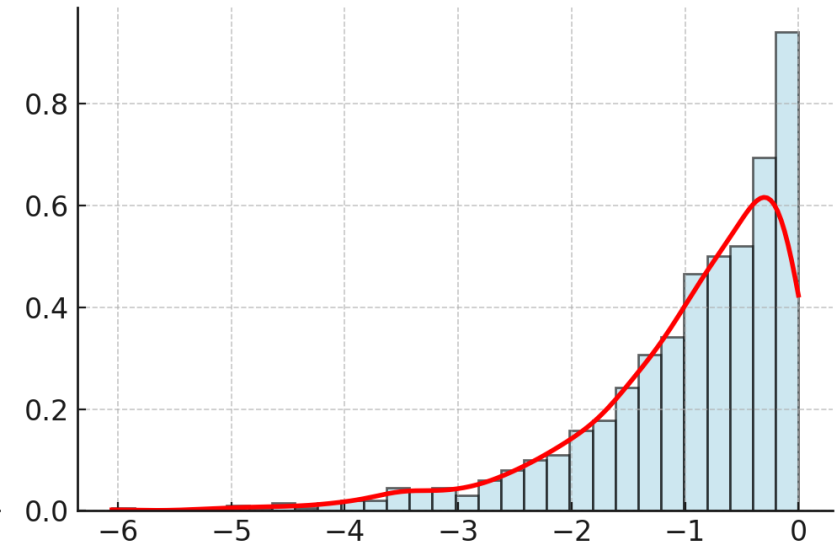
Normal  
Skew=-0.03, Kurt=2.97



Right Skewed  
Skew=2.12, Kurt=10.58



Left Skewed  
Skew=-1.61, Kurt=6.11



# SKEWNESS

---

- Skewness measures the asymmetry of a distribution.
- Positive Skew (Right-skewed): Tail on the right side is longer.
- Negative Skew (Left-skewed): Tail on the left side is longer.
- Zero Skew: Data is symmetric (like a normal distribution).
- Examples:
  - Test scores (low outliers) → Left Skew
  - Incomes (few high earners) → Right Skew

# KURTOSIS

---

- Kurtosis measures the peakedness or flatness of a distribution.
- Mesokurtic ( $\approx 3$ ): Normal distribution.
- Leptokurtic ( $> 3$ ): Sharp peak, heavy tails (more outliers).
- Platykurtic ( $< 3$ ): Flat peak, light tails (fewer outliers).



# WHY SKEWNESS & KURTOSIS MATTER?

---

- Helps understand the shape of data distribution.
- Shows if mean, median, and mode are aligned or shifted.
- Identifies outliers and data spread.
- Important in statistical modeling and hypothesis testing.

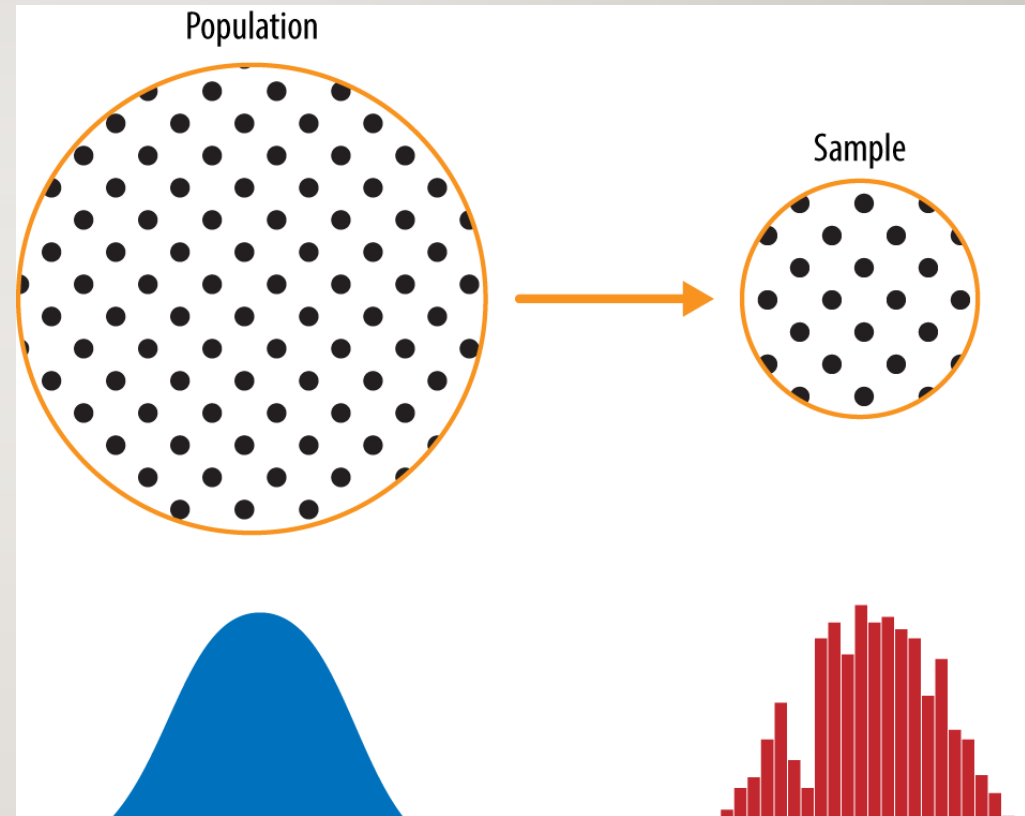
# PROBABILITY, PROBABILITY & SAMPLING DISTRIBUTION

---



# KEY CONCEPTS

- **Population:** The larger dataset or idea of a dataset.
  - **Sample:** A subset from a larger dataset.
  - **$N(n)$ :** The size of the population sample.
- 



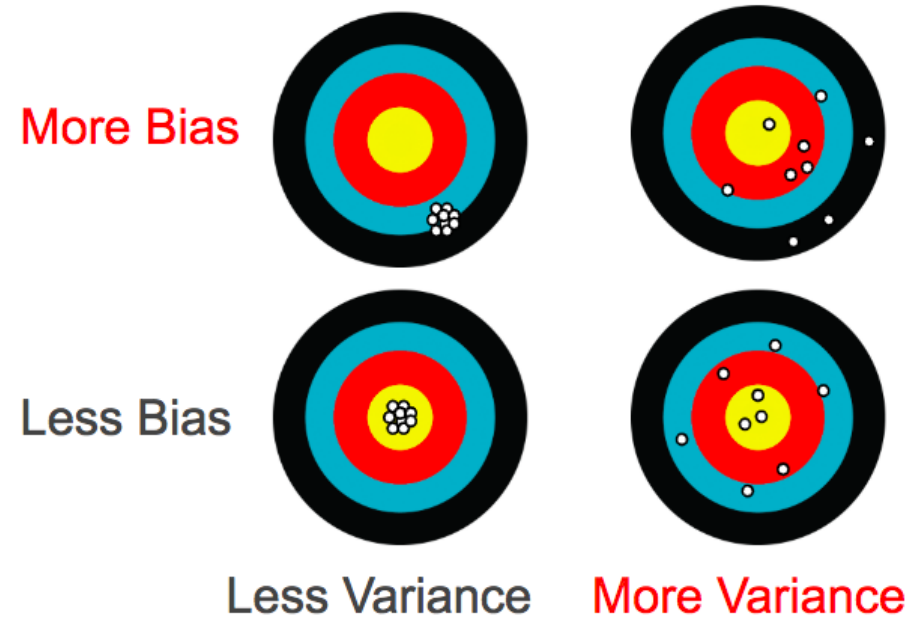
# RANDOM SAMPLING AND BIAS

---

- **Random sampling:** Drawing elements into a sample at random. Each available member of the population being sampled has an equal chance of being chosen for the sample at each draw.
  - With Replacement: observations are put back in the population after each draw for possible future reselection.
  - Without replacement: observations, once selected, are unavailable for future draws.
- **Sample Bias:** A sample that misrepresents the population. (Poll Example)

# BIAS

- Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process.





# SAMPLING DISTRIBUTION OF A STATISTIC

---

- The **sampling distribution** of a statistic refers to the distribution of some sample statistic, over many samples drawn from the same population.
- Typically, a sample is drawn with the goal of measuring something or modeling something. We are interested on ***sampling variability***.
- The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of data itself.

# LAW OF LARGE NUMBERS

---

- The average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.

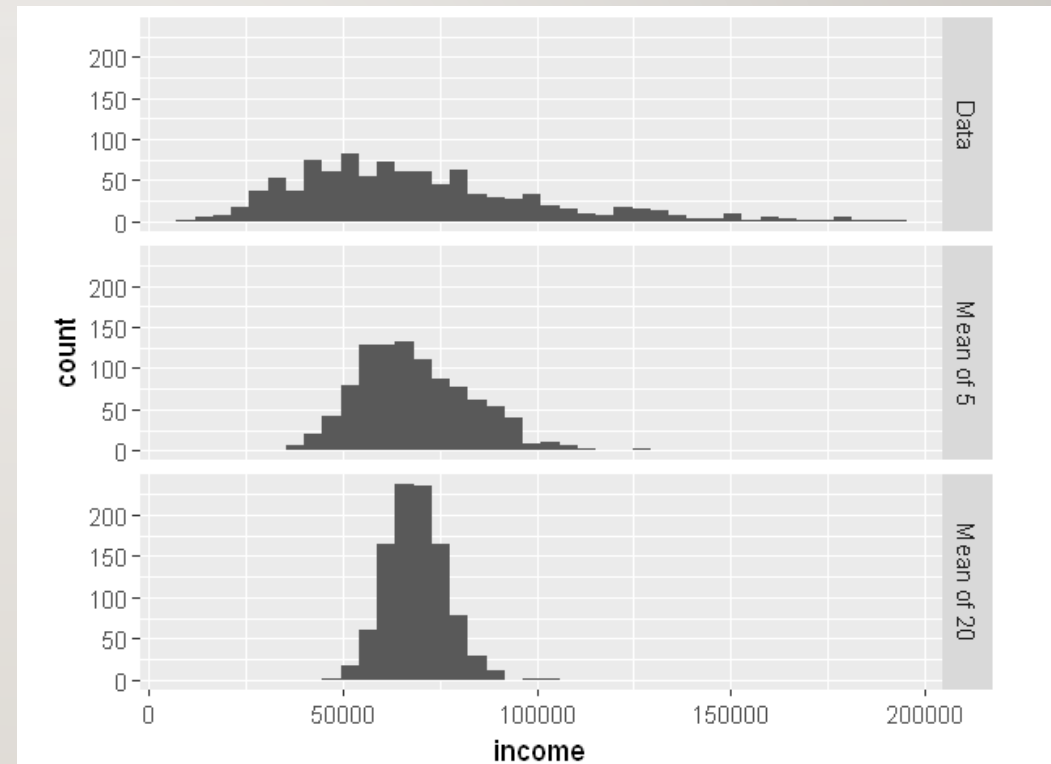
$$\overline{X}_n \xrightarrow{P} \mu \quad \text{when } n \rightarrow \infty.$$

That is, for any positive number  $\varepsilon$ ,

$$\lim_{n \rightarrow \infty} \Pr\left(|\overline{X}_n - \mu| > \varepsilon\right) = 0.$$

# LAW OF LARGE NUMBERS

- The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of data itself.



# CENTRAL LIMIT THEOREM

---

- The sampling distribution of the mean approaches a normal distribution, as the sample size increases.

$$z = \sqrt{n} \frac{\bar{x} - \mu}{\sigma} \rightarrow_d \mathcal{N}(0, 1) \text{ as } n \rightarrow \infty.$$

# STANDARD ERROR

---

- Single metric that sums up the variability in the sampling distribution for a statistic. The standard error can be estimated using a statistic based on the standard deviation  $s$  of the sample values, and the sample size  $n$ :

$$SE = \frac{s}{\sqrt{n}}$$

NOTE: standard deviation measures the variability of individual data points and standard error measures the variability of a sample metric

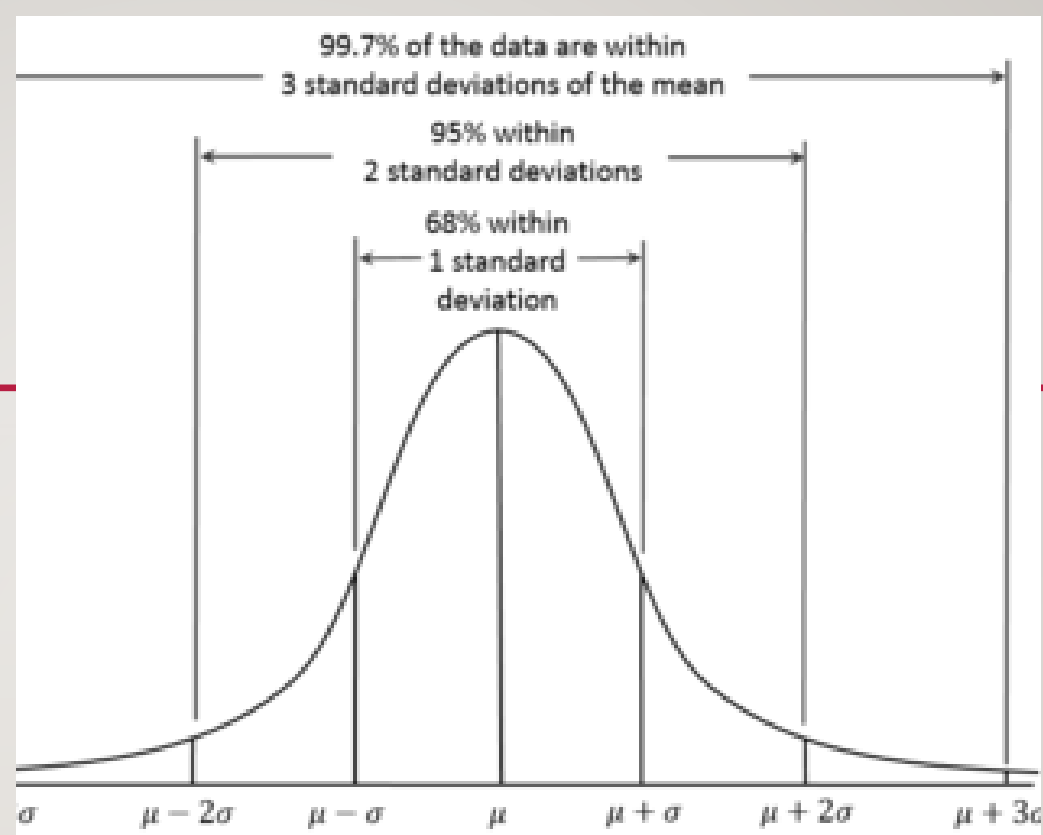


# CONFIDENCE INTERVALS

---

- Confidence intervals is an alternative to point estimation. It is a good way to deal with uncertainty. Confidence interval are the typical way to present estimates as an interval range.
- The more data you have, the less variable a sample estimate will be.
- The lower the level of confidence you can tolerate, the narrower the confidence interval will be.
- Bootstrap is an effective way to construct confidence intervals.

# NORMAL DISTRIBUTION



- Bell-shaped distribution, Gaussian distribution.

$$\phi_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

# STANDARD NORMAL

---

- Let  $\theta \sim N(\mu, \sigma^2)$  normal distributed.
- $z = \frac{\theta - \mu}{\sigma} \sim N(0, 1)$  This transformation is commonly called standardization or z-scores.
- Note: Converting data to z-scores does not make the data normally distributed. It just puts the data on the same scale as the standard normal distribution.

# STUDENT'S T-DISTRIBUTION

---

- The t-distribution is a normally shaped distribution, but a bit thicker and longer on the tails. Often called Student's t.

$$f(t) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\nu\pi} \Gamma(\nu/2)} (1 + t^2/\nu)^{-(\nu+1)/2}$$

- $\nu \equiv \text{Degrees of freedom}$
- Degrees of freedom: A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and number of groups.
- It is widely used as a reference basis for the distribution of sample means, differences between two sample means, regression parameters, and more.

# BINOMIAL DISTRIBUTION

---

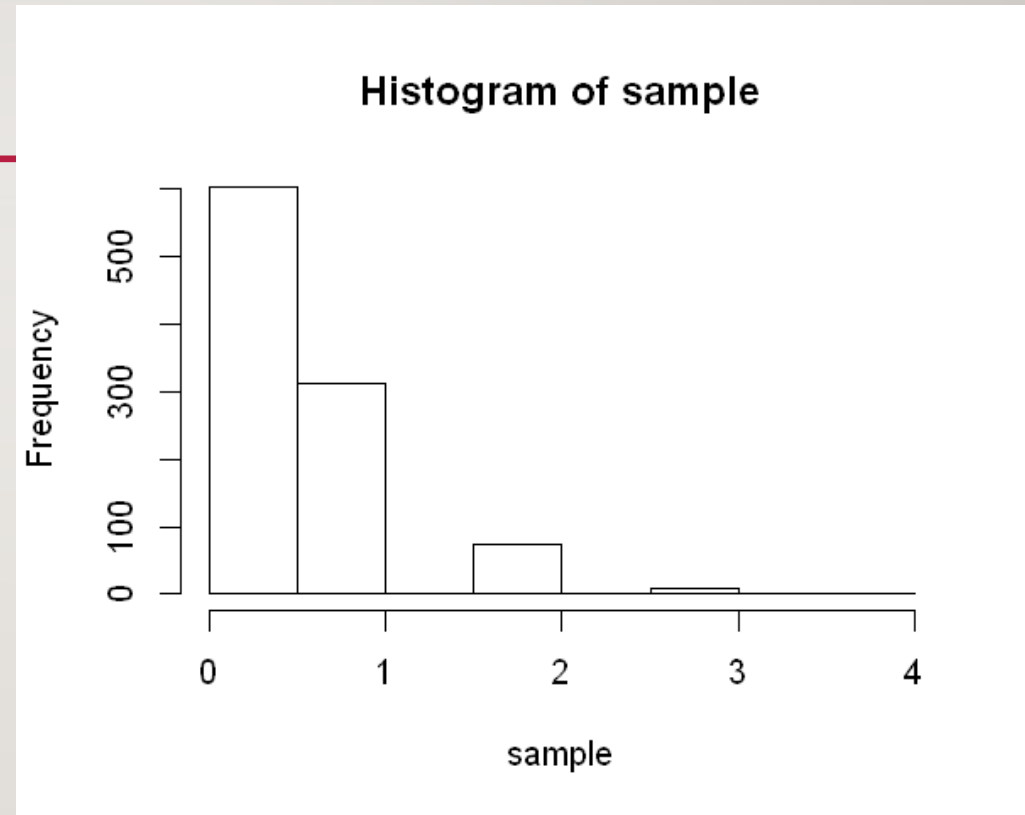
- Yes/No (binomial) outcomes lie at the heart of analytics since they are often the culmination of a decision or other process.
  - **Trial:** An event with a discrete outcome.
- The binomial distribution is the frequency distribution of the number of successes ( $x$ ) in a given number of trials ( $n$ ) with specified probability ( $p$ ) of success in each trial.
- With large  $n$ , and provided  $p$  is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution. (**proof\***)



# BINOMIAL DISTRIBUTION

---

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 \leq p \leq 1$$



# POISSON DISTRIBUTION

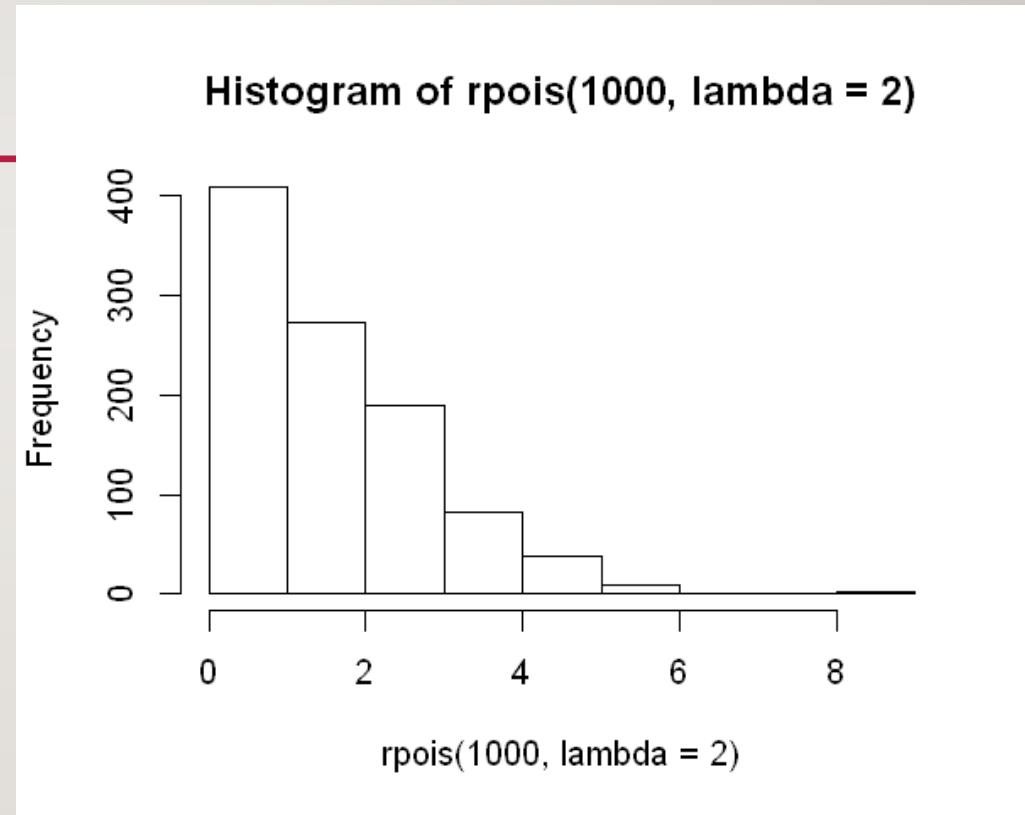
---

- The Poisson distribution tell us the distribution of events per unit of time or space when we sample many such units.
- “Internet traffic that arrives on a server in any 5-second period”
- “Number of car that cross a bump in any 5-minutes period”

# POISSON DISTRIBUTION

---

$$f(k, \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$



# CHI-SQUARE DISTRIBUTION

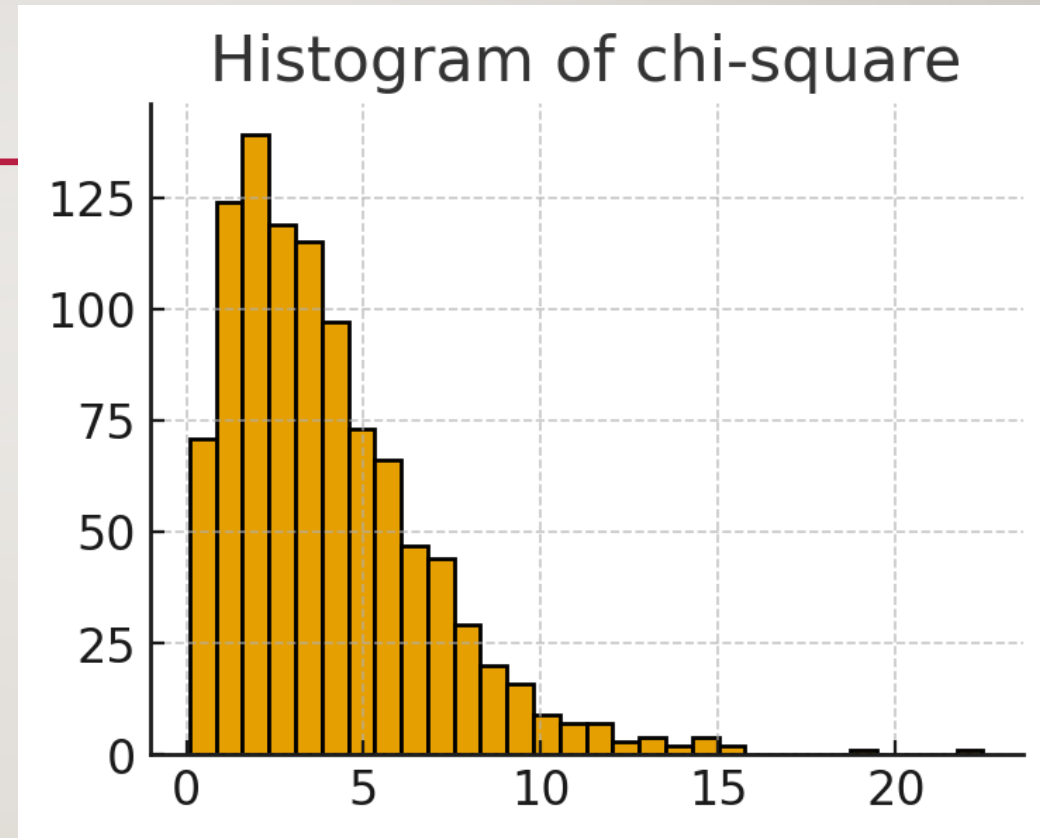
---

- The Chi-Square distribution is widely used in hypothesis testing and confidence interval estimation for variance.
- It arises from the sum of the squares of independent standard normal variables.
- Applications:
  - Goodness of fit test for categorical data
  - Test of independence in contingency tables

# CHI-SQUARE DISTRIBUTION

---

$$f(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2 - 1} e^{-x/2}, \quad x > 0$$





# F DISTRIBUTION

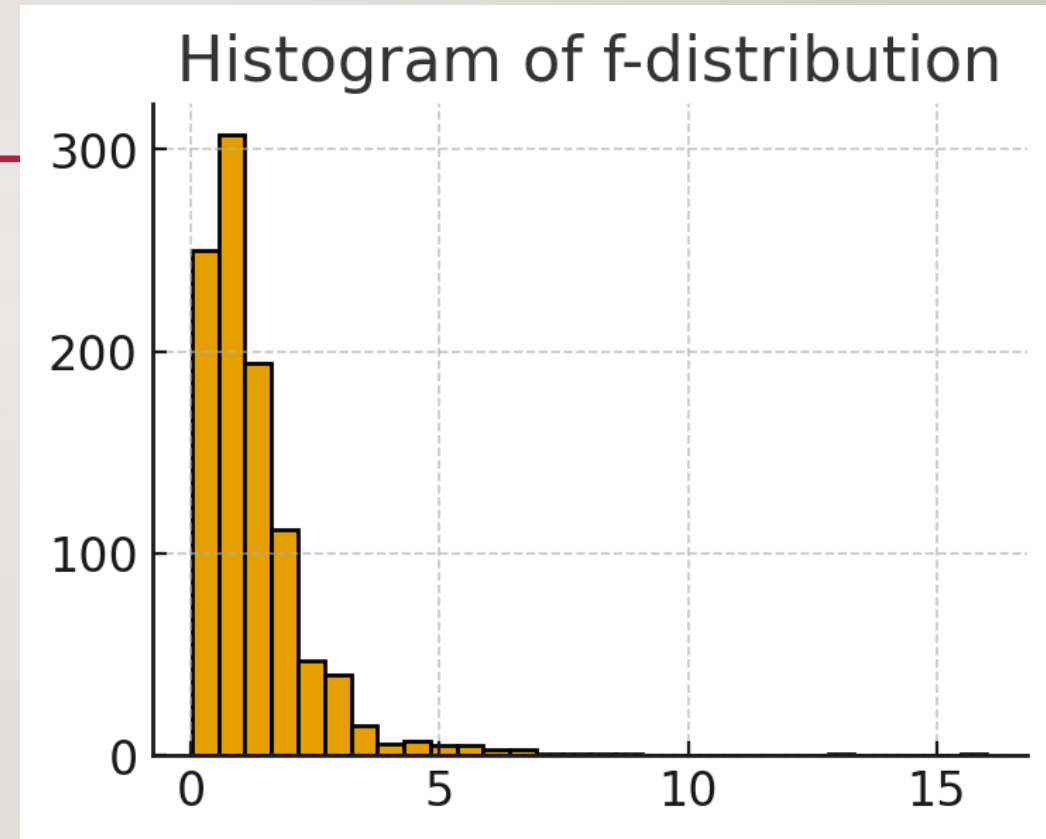
---

- The F distribution is the ratio of two scaled chi-square distributions.
- It is commonly used to compare variances and in analysis of variance (ANOVA).
- Applications:
  - Testing if two populations have equal variances
  - ANOVA for comparing multiple group means

# F DISTRIBUTION

---

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B(d_1/2, d_2/2)}, \quad x > 0$$



# STATISTICAL EXPERIMENTS AND SIGNIFICANCE TESTING

---



# INTRODUCTION

---

- Data scientist are faced with the need to conduct continual experiments in order to confirm or reject a hypothesis.
- In this chapter we will review traditional experimental design and discuss some common challenges in data science.
- Classical Statistical inference Pipeline:



# HYPOTHESIS TEST

## KEY TERMS

---

- Also called significance test. Their purpose is to help you learn whether random chance might be responsible for an observed effect.
- **Null hypothesis.** The hypothesis that chance is to blame.
- **Alternative hypothesis:** counterpoint to the null (what you hope to prove).
- **One-way test:** Hypothesis test that counts chance results only in one direction
- **Two-way test:** Hypothesis test that counts chance results in two directions.



# HYPOTHESIS TEST EXAMPLE

---

- Is it true that vitamin C has the ability to cure or prevent the common cold? Or is it just a myth? There's nothing like an in-depth experiment to get to the bottom of it all.



# HYPOTHESIS TEST EXAMPLE

---

- **Null hypothesis** - Children who take vitamin C are no less likely to become ill during flu season.
- **Alternative hypothesis** - Children who take vitamin C are less likely to become ill during flu season.



# ONE-WAY, TWO-WAY HYPOTHESIS TEST

---

Often, we want to test a new option B, against an established default option A and the presumption is that we will stick with the default option unless the new option proves itself definitively better.

Null : “no difference between the means of group A and group B”, alternative: “A is different from B”

$$H_0: A = B \text{ vs } H_1: A \neq B$$

Null : “group B is equal or better than group A”, alternative: “A is better than B”

$$H_0: A \leq B \text{ vs } H_1: A > B$$

# PERMUTATION TEST



- Procedure of combining two or more samples together, and randomly (or exhaustively) reallocating the observations to resamples.
  1. Combine results from different groups in a single data set.
  2. Shuffle and draw without replacement a resample of group A size.
  3. From the remaining, draw without replacement a resample of group B size.
  4. Do the same for groups C,D (if necessary)
  5. Calculate test statistic for the resamples, and record (one iteration)
  6. Repeat R time to yield a permutation distribution of the test statistic

Go back to the observed difference between group and compare it to the set of permuted differences.

# PERMUTATION TEST

---

- Exhaustive permutation test: instead of just randomly shuffling and dividing the data, we actually figure out all possible ways it could be divided. With a large number of repeated shufflings, the random permutation test results approximate those of the exhaustive permutation test.
- Bootstrap permutation test: The  $B$  draws in steps 2 and 3 are made with replacement.

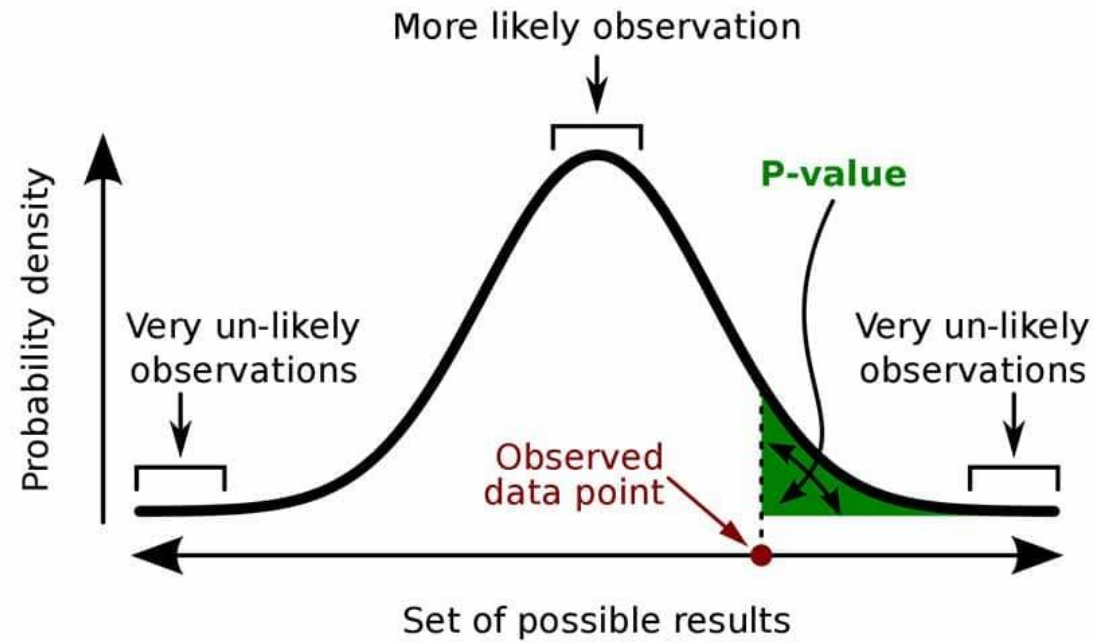


# STATISTICAL SIGNIFICANCE AND P-VALUES

---

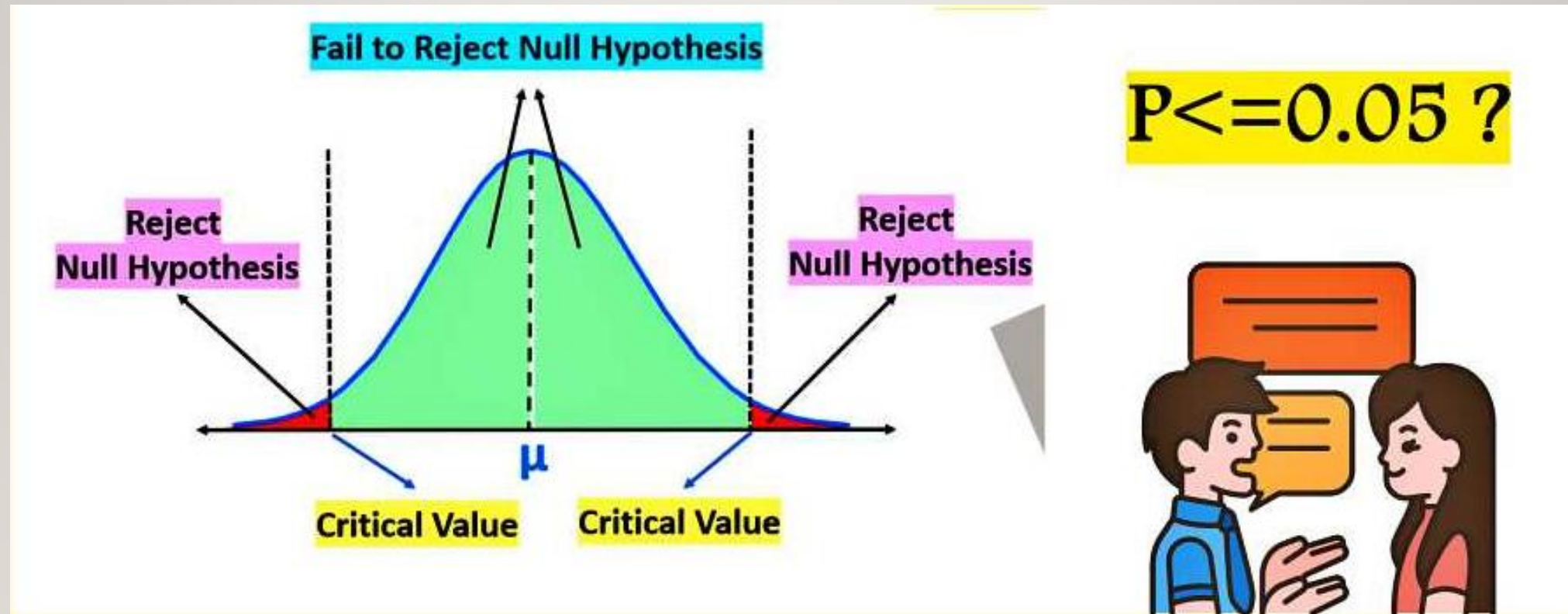
- Statistical significance is how statisticians measure whether an experiment (or even a study of existing data) yields a results more extreme than what chance might produce.
- If the result is beyond the realm of chance variation , it is said to be statistically significant.
- **P-value:** Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.
- **Alpha:**The probability threshold of unusualness that chance results must surpass, for actual outcomes to be deemed statically significant.

# EXAMPLE



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

# EXAMPLE



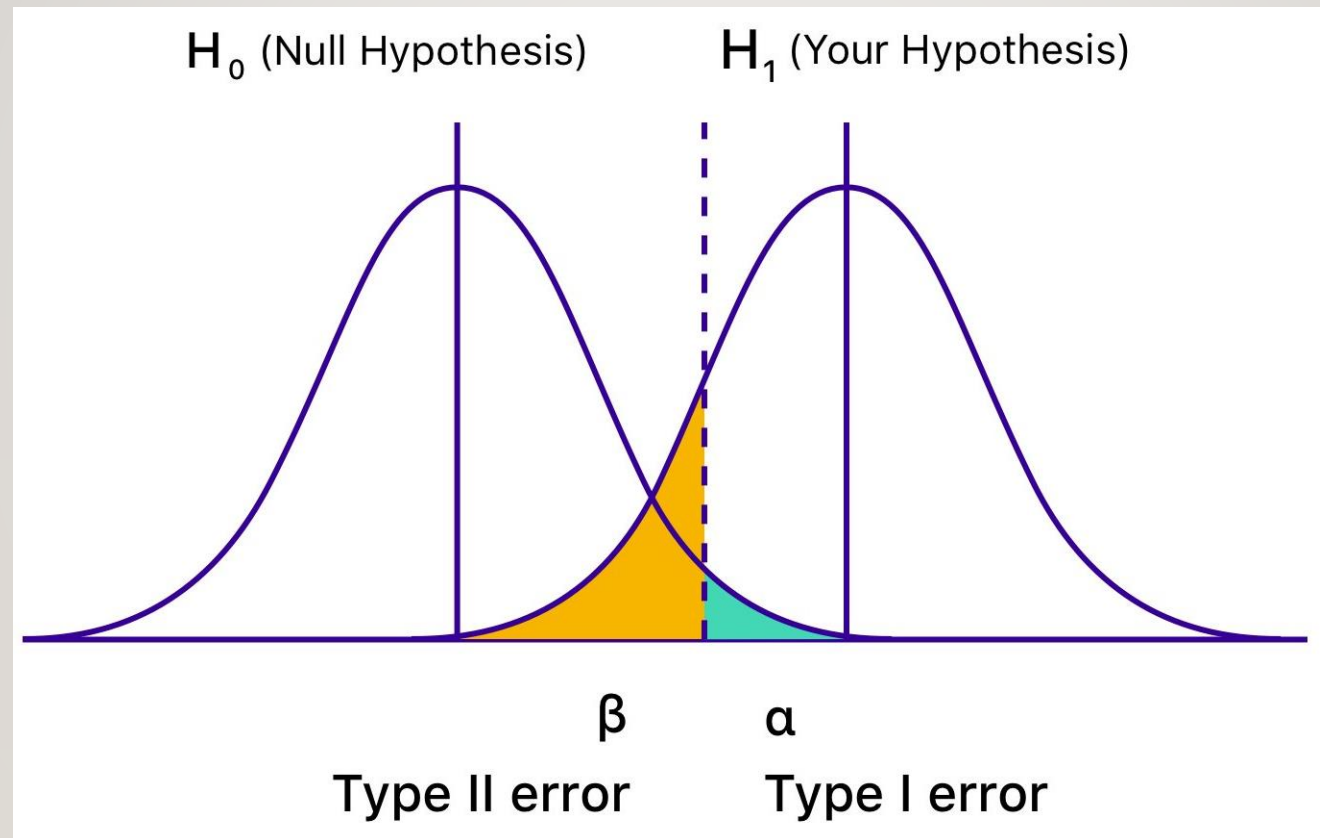
# ERROR TYPES

---

- Type I Error (False Positive):
  - Rejecting the null hypothesis when it is actually true.
  - Example: Concluding a new drug works when it actually does not.
- Type II Error (False Negative):
  - Failing to reject the null hypothesis when it is actually false.
  - Example: Concluding a drug does not work when it actually does.





# ERROR TYPES





# ERROR TYPES

---

		Reality	
		True	False
Measured or Perceived	True	Correct 	<b>Type 1 error</b> False Positive
	False	<b>Type 2 error</b> False Negative	Correct 

---

Any Questions