Introduction to Machine Learning in Geosciences
Introduction

GEO371T/GEO39D.1

Mrinal K. Sen
Geosciences
UT Austin

September 19, 2023

Content

- What is statistics?
- Population and Sample
- Data types
- Probability
- Probability Distribution
- Measures of dispersion
- Bayesian Inference
- Hypothesis Testing

- I see blank faces when I start to discuss statistical methods in my Inverse Theory class.
- "Statistics was my worst class in school!"
- "There are three kinds of lies: *Lies, damned lies and statistics*" - Mark Twain (?)
- Statistics Never Lie But Liars Use Statistics!

The need to be competent in statistics is fast becoming a necessity in many fields of work. It is also becoming a requirement to be a thoughtful participant in modern society, as we are bombarded daily by statistical information and arguments, many of questionable merit.[1]

[1] *Statistics in a Nutshell* by Boslaugh ad Watters

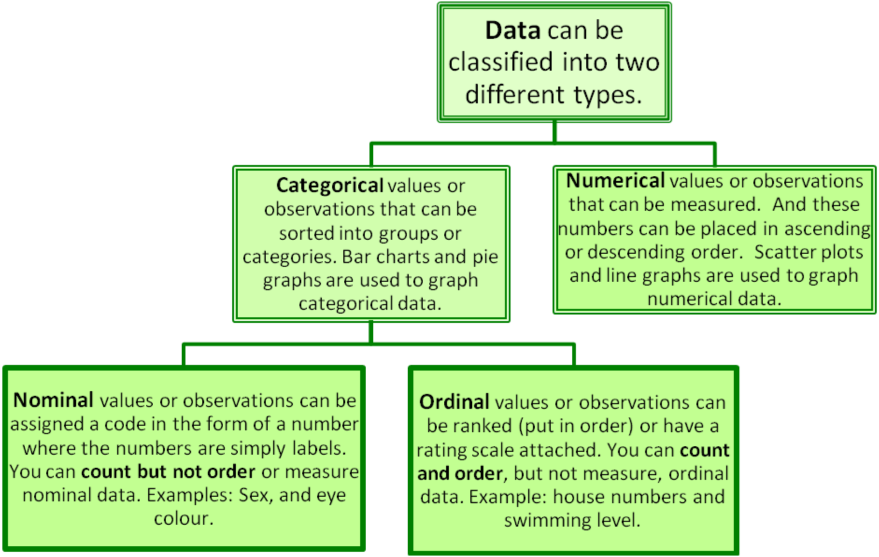
"I believe that it would be worth trying to learn something about the world even if in trying to do so we should merely learn that we do not know much. This state of learned ignorance might be a help in many of our troubles. It might be well for all of us to remember that, while differing widely in the various little bits we know, in our infinite ignorance we are all equal."[a]

[a] *Knowledge Without Authority*, 1960 - Karl Popper

What is Statistics?

- Statistics: “a branch of mathematics used to summarize, analyze, and interpret a group of numbers or observations (data). It is the science of how data is collected, analyzed, and interpreted.
- Data is information.
- Typically, data takes the form of observed measurements (e.g. height, temperature) or descriptions (e.g. dog, cat, blue, female)

Data Types



Data can be classified into two different types.

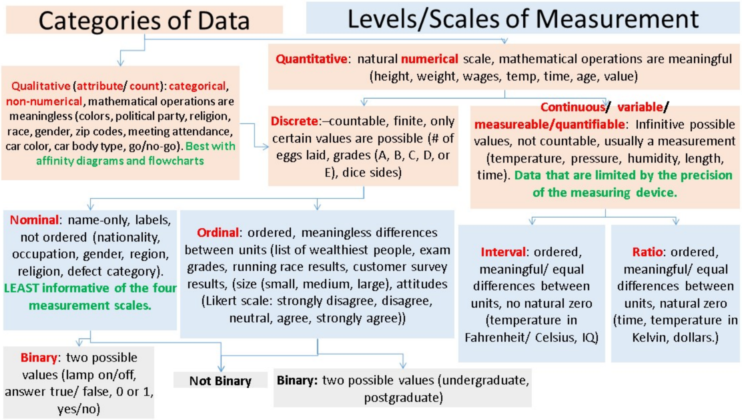
Categorical values or observations that can be sorted into groups or categories. Bar charts and pie graphs are used to graph categorical data.

Numerical values or observations that can be measured. And these numbers can be placed in ascending or descending order. Scatter plots and line graphs are used to graph numerical data.

Nominal values or observations can be assigned a code in the form of a number where the numbers are simply labels. You can **count but not order** or measure nominal data. Examples: Sex, and eye colour.

Ordinal values or observations can be ranked (put in order) or have a rating scale attached. You can **count and order**, but not measure, ordinal data. Example: house numbers and swimming level.

Data Types



Categories of Data

Levels/Scales of Measurement

Quantitative: natural **numerical** scale, mathematical operations are meaningful (height, weight, wages, temp, time, age, value)

Qualitative (**attribute**/ **count**): **categorical**, **non-numerical**, mathematical operations are meaningless (colors, political party, religion, race, gender, zip codes, meeting attendance, car color, car body type, go/no-go). **Best with affinity diagrams and flowcharts**

Discrete:–countable, finite, only certain values are possible (# of eggs laid, grades (A, B, C, D, or E), dice sides)

Continuous/ variable/ measureable/quantifiable: Infinitive possible values, not countable, usually a measurement (temperature, pressure, humidity, length, time). **Data that are limited by the precision of the measuring device.**

Nominal: name-only, labels, not ordered (nationality, occupation, gender, region, religion, defect category). **LEAST informative of the four measurement scales.**

Ordinal: ordered, meaningless differences between units (list of wealthiest people, exam grades, running race results, customer survey results, (size (small, medium, large), attitudes (Likert scale: strongly disagree, disagree, neutral, agree, strongly agree))

Interval: ordered, meaningful/ equal differences between units, no natural zero (temperature in Fahrenheit/ Celsius, IQ)

Ratio: ordered, meaningful/ equal differences between units, natural zero (time, temperature in Kelvin, dollars.)

Binary: two possible values (lamp on/off, answer true/ false, 0 or 1, yes/no)

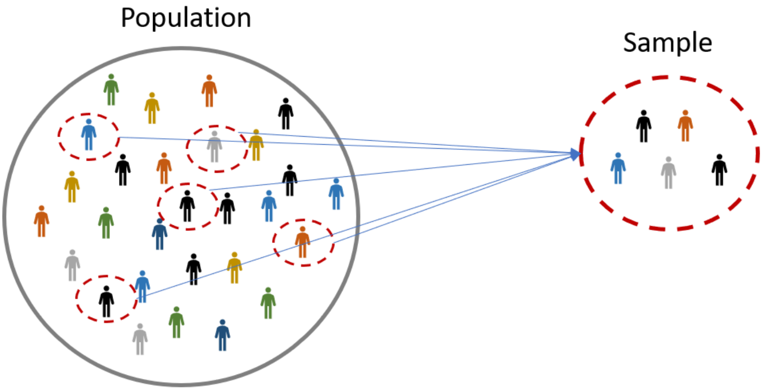
Not Binary

Binary: two possible values (undergraduate, postgraduate)

Population and Sample

- A **population** is the entire overall group we are interested in.
- A **sample** is a subset of the entire population that we collect data on. The variable(s) of interest is/are measured on each member of the sample. single member of a sample is called an **observation**.

We work with samples because the entire population of interest is not available to us!



https://www.sigmamagic.com/blogs/online-sample-size-calculators/

Random Numbers

- A computer's output is always predictable - computer generated random numbers are used extensively and are called **pseudo-random numbers**.
- **Congruential method** to generate random samples from a uniform distribution.
- A sequence of pseudo random integers $I_1, I_2, I_3, \ldots$ between 0 and N by the recurrence relation:

$$I_{j+1} = \mathrm{mod}(aI_j + c, N),$$

$$\mathrm{mod}(a_1, a_2) = a_1 - \mathrm{int}\left(\frac{a_1}{a_2}\right) a_2.$$

- a and c : positive integers called the multiplier and increment.
- he number thus generated will repeat itself with a period no greater than N. If N, a and c are properly chosen, then the period will be of maximum length.
- The highest possible value of N is controlled by the precision of the machine used.

Random Variables

- A random variable is the outcome of a random experiment (the tossing of a die and flipping of a coin).
- The outcome of a random experiment is usually represented by a point, called a sample point s
- The set which consists of all possible sample points is called the sample space S
- Subsets of S represent certain events such
- that an event A consists of a certain collection of possible outcomes s.
- subsets contain no points s in common, they are said to be disjoint, and the corresponding events are said to be *mutually exclusive*.
- Any single valued numerical function $X(s)$ defined on a sample space S is called a random variable and a unique real number is associated with each point s

Probability

Axioms

$$P\{A\} \geq 0 \quad \text{for any event} A,$$

i.e.,the probabilities of all the events are non-negative and

$$P(S) = 1,$$

i.e., the probability of the whole sample space is unity.
Thus, by convention or axiom, the probability that one of a sequence of mutually exclusive events $\{A_i\}$ will occur is given by

$$P\{A_1 \ or \ A_2 \ or \ A_3, ...\} = \sum_i P(A_i).$$

Probability

The Classical Interpretation

,

Let N be the total number of possible outcomes of an experiment. If in $N_A$of all these outcomes, the event A occurs, then $P(A)$ is given by

$$P(A) = \frac{N_A}{N},$$

provided that the occurrence of all the events are *equally likely*. The main criticism of this definition is that it is circular since *equally likely* also means *equally probable*.

Probability

Relative frequency interpretation [von Mises, 1957]

the probability of an event A is the following limit of the relative frequency

$$P(A) = \lim_{N \to \infty} \frac{N_A}{N},$$

where N_A is the number of occurrences of A in N number of trials.

- The concept of equally likely events is completely avoided.
- A trial is used to mean **repetition of an experiment under identical circumstances**.
- The limit can only be assumed to exist.
- The number of trials is always finite.
- The definition gives no meaning to the probability of a hypothesis [Jeffreys, 1939]

Probability:Bayesian interpretation (Bayes 1763)

- It is a degree of belief!
- The probability theory can be viewed as an extension of deductive logic and is called inductive logic.
- In deductive logic, a proposition can either be true or false.
- In inductive logic, the probability of a proposition constitutes a degree of belief, with proof or disproof as extremes.
- In the subjective interpretation, the degree of belief is a personal degree of belief such that the axioms of probability theory are not violated
- Most statisticians are proponents of one interpretation or another but more than one interpretation may be helpful since different situations may simply ask for different interpretations.A Bayesian interpretation of probability appears to be conceptually very clear.

Probability distribution and Distribution function

A discrete random variable $X(s)$ assumes a finite number of values. For each value x_i, there is a unique probability that the random variable assumes the value x_i::

$$P\left\{X(s) = x_i\right\} = p_i, i = 0, 1, 2, ...$$

The sequence $\{p_i\}$: the **probability cumulative probability**

$$P\left\{X(s) \le x\right\} = \sum_{x_i \le x} p_i = F(x), -\infty < x < \infty.$$

is called the **distribution function** of $X(s)$.

Probability Density Function

Distribution function:$F(-\infty) = 0, F(\infty) = 1;$, A non-decreasing function of $x : F(X_1) \leq F(x_2),$ for $x_1 < x_2$. If $F(x)$ is differentiable, **the probability density function (pdf)**, $p(x)$, is given by

$$p(x) = \frac{dF(x)}{dx}.$$

If $F(\mathbf{m})$ denotes the distribution function of a vector of variables m, with $\mathbf{m} = [m_1, m_2, ...m_n]^T$, then if $F(\mathbf{m})$ is differentiable, the probability density function (pdf), $p(\mathbf{m})$ of m, is given by

$$p(\mathbf{m}) = \frac{\partial^n F(\mathbf{m})}{\partial m_1, \partial m_2, ..., \partial m_n}.$$

The probability of $\mathbf{m}$ being in a certain region or volume A is given by

$$P(\mathbf{m} \in A) = \int_A p(\mathbf{m}) \, dm_1 dm_2.....dm_n.$$

Probability Distribution & Distribution function

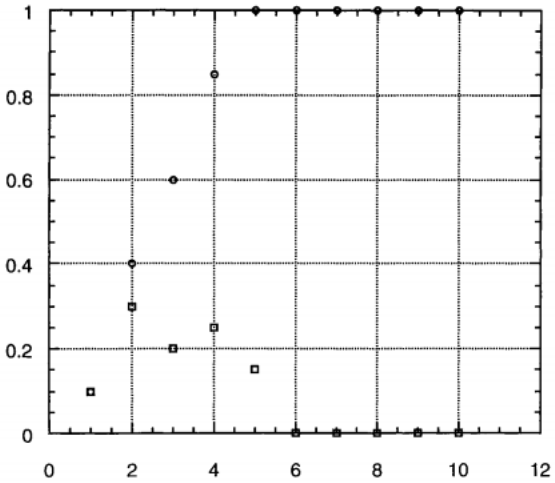


Figure 1.1. An example of a probability distribution (open squares) and its corresponding distribution function (open circles).

Continuous Distribution: Normal or Gaussian Distribution

A continuous random variable x is said to be normally distributed if its density function $p(x)$ is given as

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x- < x >)^2}{2\sigma^2} \right],$$

where $< x >$ and σ^2 are the mean and variance respectively. Note that

$$\int_{-\infty}^{\infty} dx\ p(x) = 1.$$

Properties of the normal distribution:

- The normal distribution curve is symmetric around its mean $< x >$.
- Approximately 68% of the area below the normal curve is covered by the interval of plus minus one standard deviation around its mean: $< x > \pm\sigma$.
- Approximately 95% of the area below the normal curve is covered by the interval of plus minus two standard deviations around its mean: $< x > \pm 2\sigma$.
- Approximately 99.7% of the area below the normal curve is covered by the interval of plus minus three standard deviations around its mean: $< x > \pm 3\sigma$.
- A linear combination of two or more normal random variables is also normal.

Continuous Distribution: Normal or Gaussian Distribution

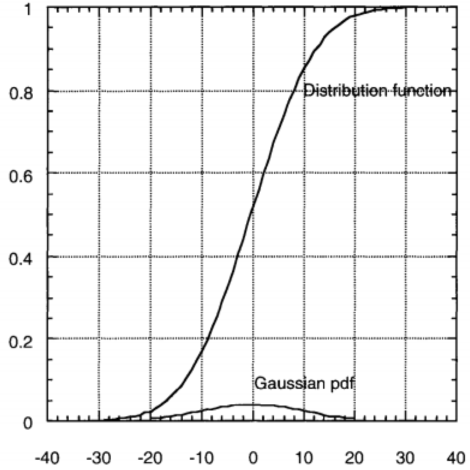


Figure 1.2. A Gaussian probability density function for a mean = 0.0 and a standard deviation = 10.0. Its corresponding distribution function is also shown.

Sampling Distribution & Sample Mean

- A lot of data drawn and used by academicians, statisticians, researchers, marketers, analysts, etc. are actually **samples**, not **populations**. A sample is a subset of a population.
- Now suppose that instead of taking just one sample, we take repeated random samples from the general population, and compute the sample mean for each sample group.
- The mean computed for each sample set is the sampling distribution of the mean.
- Other statistics, such as the standard deviation, variance, proportion, and range can be calculated from sample data. The standard deviation and variance measure the variability of the sampling distribution.

Sampling Distribution & Sample Mean

- Consider a normal population with mean μ and variance σ^2.
- We repeatedly take samples of a given size from this population and calculate the arithmetic mean $\bar{x}$ for each sample – this statistic is called **the sample mean**.
- The distribution of these means, or averages, is called the **sampling distribution of the sample mean**.
- This distribution is normal $N(\mu, \sigma^2/n)$ (n is the sample size) since the underlying population is normal, although sampling distributions may also often be close to normal even when the population distribution is not!

Central Limit Theorem

- **The Central Limit Theorem states that the sampling distribution of the sample mean approximates the normal distribution, regardless of the distribution of the population from which the samples are drawn, if the sample size is sufficiently large.**
- The Central Limit Theorem states that the sampling distribution of the sample mean approximates the normal distribution, regardless of the distribution of the population from which the samples are drawn, if the sample size is sufficiently large.
- The distribution of these means, or averages, is called the **sampling distribution of the sample mean**.
- This distribution is normal $N(\mu,\ \sigma^2/n)$ (n is the sample size) since the underlying population is normal, although sampling distributions may also often be close to normal even when the population distribution is not!
- Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2, then for large n,

  $$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

  even if the underlying distribution of individual observations in the population is not normal.
- The application of the Central Limit Theorem in practice can be seen through computer simulations that repeatedly draw samples of specified size from a nonnormal population.

Central Limit Theorem

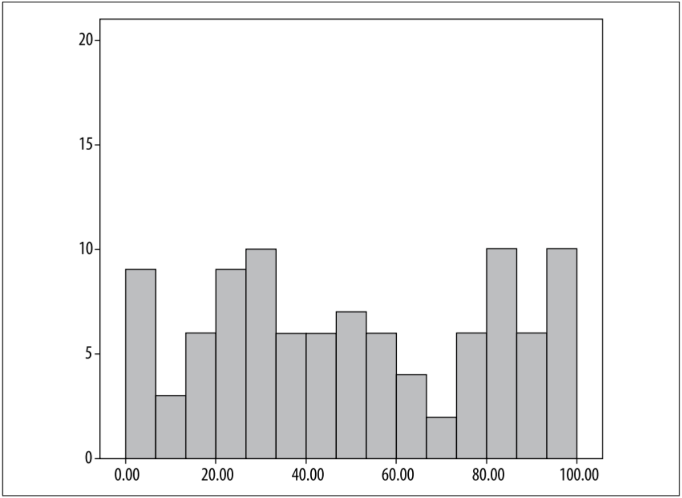


Figure 7-4. Histogram of a uniformly distributed population (N = 100) with range 0–100

Central Limit Theorem

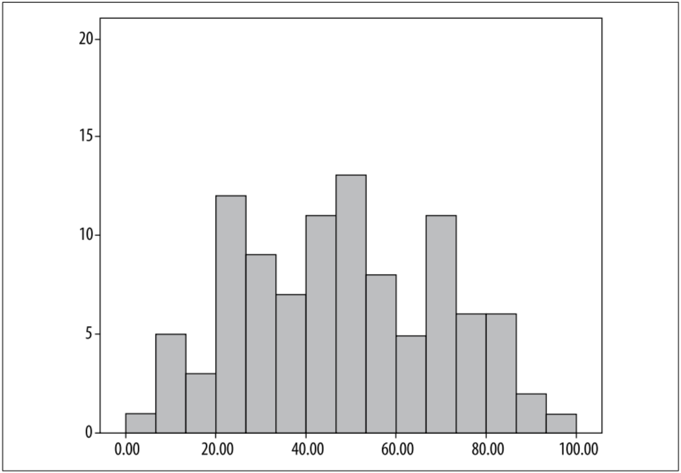


Figure 7-5. Distribution of the means of 100 samples of size n = 2, drawn from a uniform distribution

Central Limit Theorem

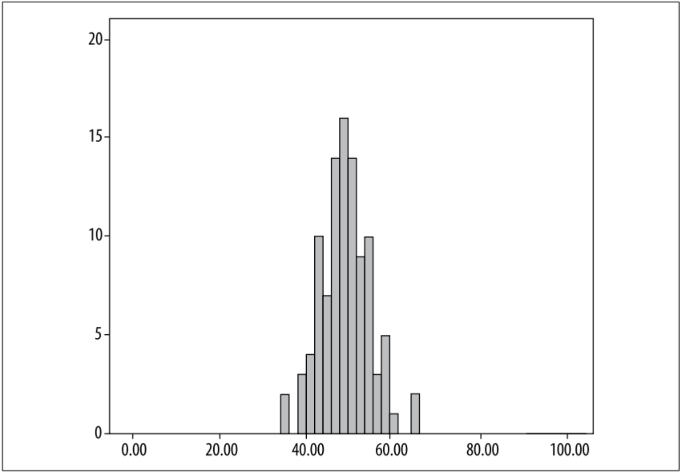
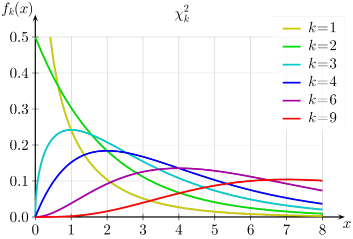


Figure 7-6. Distribution of means of 100 samples of size n = 25, drawn from a uniform distribution

The Chi-square distribution

- if $Z_1, ..., Z_k$ are independent, standard normal variables, then the sum of their squares, $Q = \sum_{i=1}^{k} Z_i^2$ is the **chi-square distribution with k degrees of freedom**, denoted by $Q \sim \chi^2(k)$ or $Q \sim \chi_k^2$.
- The χ^2-distribution is used to model errors measured as sum of squares or the distribution of the sample variance.

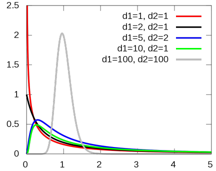


The F-distribution or Fisher's F-distribution

- A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates:

$$X = \frac{U_1/d_1}{U_2/d_2}.$$

- The F-distribution plays a central role in hypothesis testing answering the question: Are two variances equals? Is the ratio or two errors significantly large ?



Student's t Distribution

- Note that the Gaussian distribution is sensitive to ouliers! *Student's t distribution* is more robust, given by

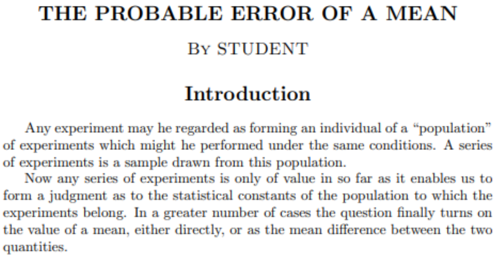
$$\tau(x| <x>, \sigma^2, \nu) \propto \left[1 + \frac{1}{\nu} \left(\frac{x- <x>}{\sigma} \right)^2 \right]^{-\left(\frac{\nu+1}{2}\right)},$$

  where $<x>$ is the mean, σ^2 is the scale parameter and $\nu > 0$ is called the **degrees of freedom**.

- if $\nu = 1$, the distribution is called the **Cauchy or Lorentz** distribution. For $\nu > 5$, it approaches a Gaussian distribution!

Student's t Distribution

- Let $M \sim N(0,1)$, and $V \sim \chi^2_n$. The t distribution, T_n, with n degrees of freedom is the ratio: $T_n = \frac{M}{\sqrt{V/n}}$.
- The distribution of the difference between an estimated parameter and its true (or assumed) value divided by the standard deviation of the estimated parameter (standard error) follow a t-distribution. Is this parameters different from a given value?



THE PROBABLE ERROR OF A MEAN

By STUDENT

Introduction

Any experiment may he regarded as forming an individual of a "population" of experiments which might he performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a greater number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

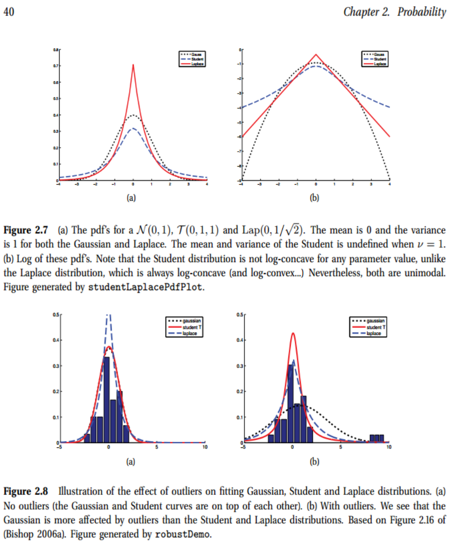
Laplace Distribution

- The Laplace distribution has heavy tails (also known as **double-sided exponential distribution**).
- This is given by

$$\text{Lap}(x| < x >, \sigma) = \frac{1}{2\sigma} \exp\left(- \frac{|x - < x > |}{b} \right),$$

  where $< x >$ is the location parameter and $b > 0$ is a scale parameter.
- **This distribution is robust to outliers and is a very useful way to encourage sparsity in a model.**

Distributions



40 *Chapter 2. Probability*

Figure 2.7 (a) The pdf's for a $\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $\text{Lap}(0,1/\sqrt{2})$. The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when $\nu = 1$. (b) Log of these pdf's. Note that the Student distribution is not log-concave for any parameter value, unlike the Laplace distribution, which is always log-concave (and log-convex...) Nevertheless, both are unimodal. Figure generated by `studentLaplacePdfPlot`.

Figure 2.8 Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions. Based on Figure 2.16 of (Bishop 2006a). Figure generated by `robustDemo`.

Discrete Distribution: Binomial or Bernoulli Distribution

- Suppose we toss a coin n times. Let $X \in \{0, ..., n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial distribution**, written as $X \sim \text{Bin}(n, \theta)$. The pmf[2] is given by

$$\text{Bin}(k|n, \theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

- Mean $= \theta$; var$=n\theta(1-\theta)$

[2] a probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.

Discrete Distribution: Multinomial and multinoulli Distribution

- The binomial distribution can be used to model the outcomes of coin tosses. To model the outcomes of tossing a K-sided die, we can use the **multinomial distribution**. This is defined as follows: let $x = (x_1, ..., x_K)$ be a random vector, where x_j is the number of times side j of the die occurs. Then x has the following pmf:

$$\mathrm{Mu}(\mathbf{x}|n, k) \triangleq \binom{n}{x_1, ..., x_K} \prod_{j=1}^{K} \theta_j^{x_j},$$

$$\binom{n}{x_1, ..., x_K} \triangleq \frac{n!}{x_1!x_2!.....x_K!}.$$

Discrete Distribution: Poisson Distribution

- We say that $X \in \{0, 1, 2, ...\}$ has a Poisson distribution with parameter $\lambda > 0$, written as $X \sim \text{Poi}(\lambda)$, if its pmf is

$$Poi(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}.$$

- The first term is just the normalization constant, required to ensure the distribution sums to 1.

Multivariate Normal Distribution

- **Covariance** :The covariance between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

$$\mathrm{cov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

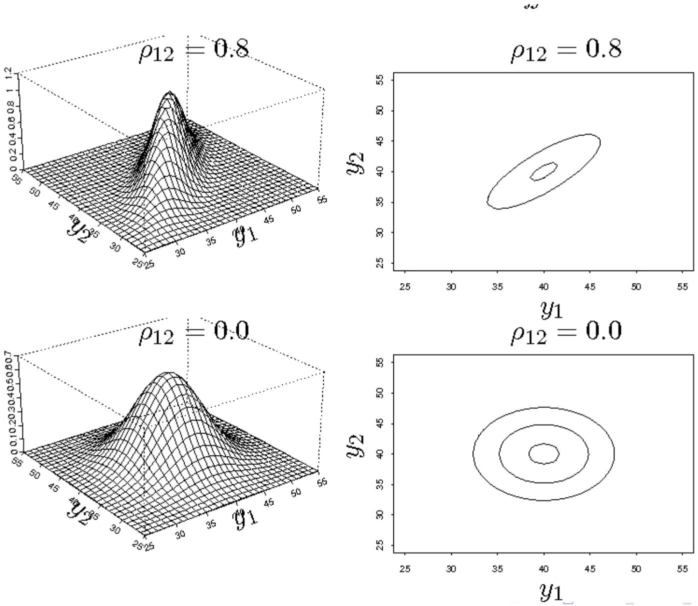
- **The (Pearson) correlation coefficient** between X and Y is defined as

$$\mathrm{corr}[X, Y] \triangleq \frac{\mathrm{cov}[X, Y]}{\sqrt{var[X]var[Y]}}.$$

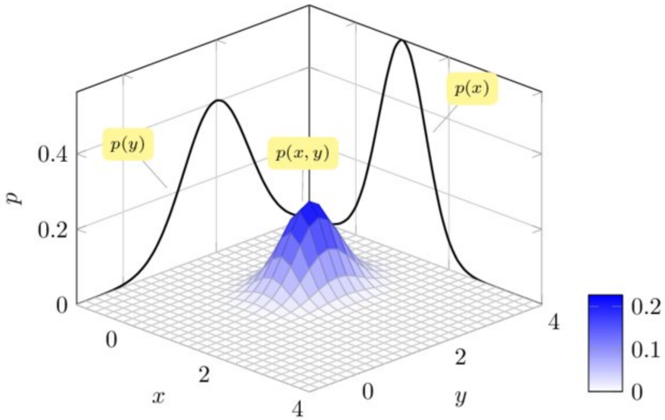
- The pdf of the **MVN** in D dimensions is defined by the following:

$$N(\mathbf{x}|\mu, \Sigma) \triangleq \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu)\right].$$

Multi-variate Normal Distribution



Multi-variate Normal Distribution



Multivariate Student *t* Distribution

- **Covariance and correlation**:The covariance between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

$$(\mathbf{x}|\mu,\Sigma,\nu)=\frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)}\frac{|\Sigma|^{-1/2}|}{\nu^{D/2}\pi^{D/2}}\left[1+\frac{1}{\nu}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]^{\frac{\nu+D}{2}}$$

- Σ : Scale Matrix; The smaller ν is, the flatter the tails. As $\nu\rightarrow\infty$, the distribution tends toward Normal!

Conditional Probabilty and Bayes' Rule

Let us assume that the events A_1 and A_2 are among the possible outcomes of an experiment and that $P\{A_1\}$ and $P\{A_2\}$ are the probabilities of the two events respectively. According to the classical definition of probability,

$$P\{A_i\} = \frac{\text{sum of probabilities of sample points } \textit{in } A_j}{\text{sum of probabilities of all sample points in the sample space}}.$$

We also use the symbol $P\{A1, A2\}$ to represent joint probability of events A_1 and A_2.

Conditional Probabilty and Bayes' Rule

Next we assume that if A_2 has occurred, then the symbol $P\{A_1|A_2\}$ stands for the conditional probability of event A_1 given that A_2 has occurred. In such a situation, the outcome of the experiment must be in a more restricted region than the entire sample space. Based on the axioms of probability, we can write

$$P\{A_1|A_2\} = \frac{P\{A_1, A_2\}}{P\{A_2\}}, \text{ assuming that } P\{A_2\} \neq= 0.$$

Thus

$$P\{A_1, A_2\} = P\{A_2\}P\{A_1|A_2\}.$$

Again, assuming that $P\{A_1|A_2\}$ stands for the conditional probability of event A_2 given that A_1 has occurred, we can write

$$P\{A_2|A_1\} = \frac{P\{A_1, A_2\}}{P\{A_1\}}, \text{ assuming that } P\{A_1\} \neq= 0.$$

Conditional Probabilty and Bayes' Rule

The conditional probability distribution and density functions can now easily be defined. The conditional distribution function $F(x|x_0)$ of a random variable $X(s)$, assuming x_0, is defined as the conditional probability of the event $\{X(s) \leq x|x_0\}$,

$$F(\mathbf{x}|\mathbf{x}_0) = P\{X(s) \leq x|x_0\} = \frac{P\{X(s) \leq x, x_0\}}{P\{x_0\}},$$

where $\{X(s) \leq x|x_0\}$ is the event consisting of all outcomes ϵ such that $X(\epsilon) \leq x$ and $\epsilon \in x_0$.

If we consider two vectors of random variables $\mathbf{x}$ and $\mathbf{y}$ such that their joint probability density functions is denoted by $p(\mathbf{x}, \mathbf{y})$, the marginal pdf of $\mathbf{x}$ will be given as

$$p(\mathbf{x}) = \int d\mathbf{y}\; p(\mathbf{x}, \mathbf{y}),$$

which can be regarded as the pdf of $\mathbf{x}$ ignoring or averaging over

Conditional Probabilty and Bayes' Rule

The conditional pdf $p(\mathbf{x}|\mathbf{y})$ called the pdf of $\mathbf{x}$ for given values of $\mathbf{y}$ is defined as

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})}.$$

Thus, we have

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}).$$

Also,

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})},$$

or,

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}).$$

Therefore, we have

$$\boxed{p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}.}$$

Conditional Probabilty and Bayes' Rule

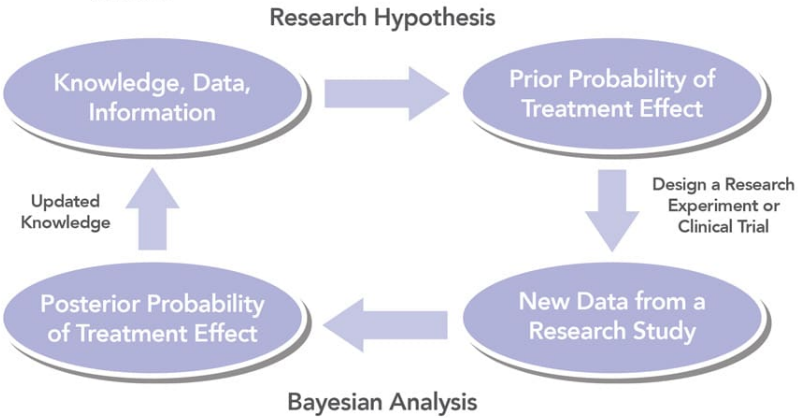
$$\boxed{p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}.}$$

The above relation is called *Bayes' rule* in statistics and is widely used in problems of model parameter estimation by data fitting. If $\mathbf{x}$) is identified with the model and *mathbfy* with the data vector, $p(\mathbf{x}|\mathbf{y})$, the probability of $\mathbf{x}$) given measurements of $\mathbf{y}$ is expressed as a function of $p(\mathbf{x})$, the marginal pdf of the model independent of data and $p(\mathbf{y}|\mathbf{x}$ the conditional probability of $\mathbf{y}$ given $\mathbf{x}$). $p(\mathbf{x}|\mathbf{y})$ is called the posterior or *a posteriori* pdf while $p(\mathbf{x})$ is called the prior or *a priori* pdf and contains the prior information about the model independent of measurements. The marginal density function $p(\mathbf{y})$ is usually assumed to be a constant and the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is usually called a *likelihood function*.

Conditional Probabilty and Bayes' Rule

The interpretation of Bayes' theorem is as follows: The state of information on $\mathbf{x}$ and $\mathbf{y}$ is described by the joint pdf $p(\mathbf{x}, \mathbf{y})$. Information becomes available as values of $\mathbf{y}$) are obtained. So the question we ask is "how should the pdf of $\mathbf{x}$ be calculated in this situation?" According to the definitions of conditional probability, this pdf should be proportional to $p(\mathbf{y}|\mathbf{x})$ with the obtained values for $\mathbf{y}$ substituted. The final expression for the conditional pdf $p(\mathbf{x}|\mathbf{y})$ is thus given by Eq. (45). Bayes' rule is especially appealing because it provides a mathematical formulation of how the current knowledge can be updated when new information becomes available.

Bayesian Analysis



Research Hypothesis

Knowledge, Data, Information

Prior Probability of Treatment Effect

Design a Research Experiment or Clinical Trial

New Data from a Research Study

Posterior Probability of Treatment Effect

Updated Knowledge

Bayesian Analysis

Statistical Analysis

- **Descriptive Statistics**: Procedures used to summarize, organize, and make sense of a set of scores or observations.
- **Inferential statistics**: procedures used that allow researchers to infer or generalize observations made with samples to the larger population from which they were selected.

Descriptive Statistics

- **Measures of Central tendency**:
  - Mean
  - Median
  - Mode
- **Range**: Difference between the largest value and smallest values (informative for data without ouliers)
- **Variance**: It measures the average squared distance that scores deviate from their mean [sample variance: s^2, population variance σ^2.]
- **Standard Deviation**: square root of variance [average distance that by which the data values deviate from their mean. How do we measure distance?]

Choosing a proper measure of central tendencey depends on: (1) the type of distribution, and (2) the scale of measurement!

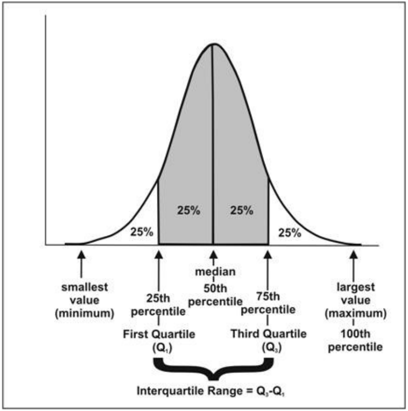
Descriptive Statistics

- **Measures of Position**:
  - **Percentile**: The p-th percentile of a data set is the data value such that p percent of the values in the data set are **at or below this value**.
  - **Percentile rank**: The percentile rank of a data value equals the percentage of values in the data set that are at or below that value.
  - **Interquartile Range (IQR)** : The first quartile (Q1) is the 25th percentile of a data set; the second quartile (Q2) is the 50th percentile (median); and the third quartile (Q3) is the 75th percentile. The IQR measures the difference between 75th and 25th observation using the formula: IQR = Q3 Q1.
  - **Z-score**: The Z-score for a particular data value represents how many standard deviations the data value lies above or below the mean.

  $$\text{Z-score} = \frac{x - \bar{x}}{\sigma}.$$

  So, If Z is positive, it means that the value is above the average.

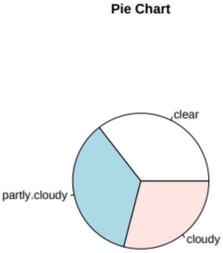
RANGE

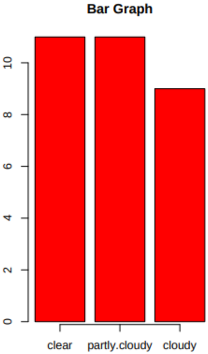


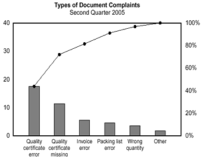
25%
25%
25%
25%
smallest value (minimum)
25th percentile
First Quartile (Q_1)
median
50th percentile
75th percentile
Third Quartile (Q_3)
largest value (maximum)
100th percentile
Interquartile Range = Q_3-Q_1

DISPLAYS: CATEGORICAL

Example: Categorical: Weather in Central Park

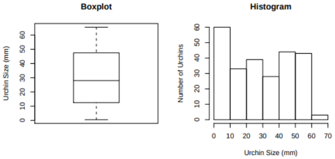






- Pie charts are harder to read.
- Pareto chart combines the properties of a bar chart and a line chart. The bars display frequency and relative frequency, whereas the line display cumulative frequency!

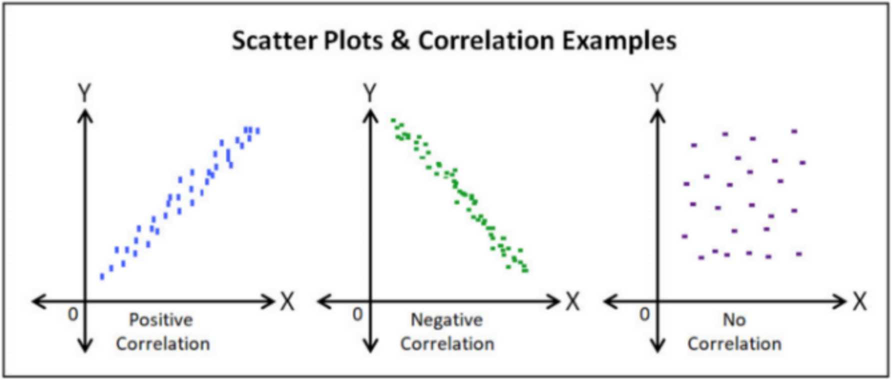
DISPLAYS: CATEGORICAL



- Boxplot shows five point summary: minimum, first quartile, median, third quartile, maximum. Especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.

Bi-variate Descriptive Statistics

- Scatterplots
- Coerrelation



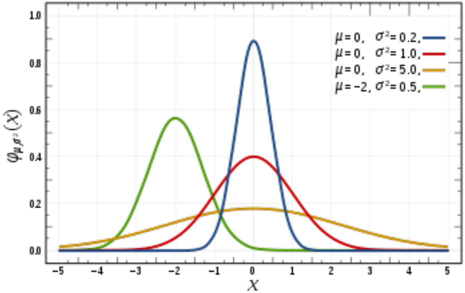
Positive correlation ($r > 0$), Negative correlation ($r < 0$), No correlation ($r = 0$)

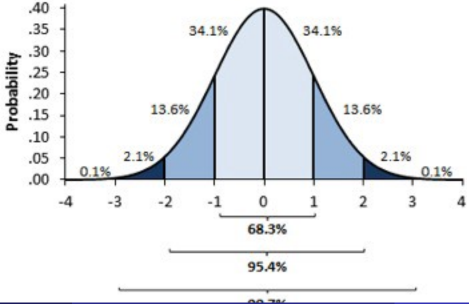
Inferential Statistics: Hypothesis Testing

- Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.
- Example : you say avg student in class is 40 or a boy is taller than girls.
- all those examples we assume need some statistic way to prove those.
- A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- Standardized normal distribution (mean 0, standard deviation=1); $x_{new} = \frac{x-\mu}{\sigma}$

https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

Inferential Statistics: Hypothesis Testing





Inferential Statistics: Hypothesis Testing

- We use inferential statistics because it allows us to measure behavior in samples to learn more about the behavior in populations that are often too large or inaccessible.

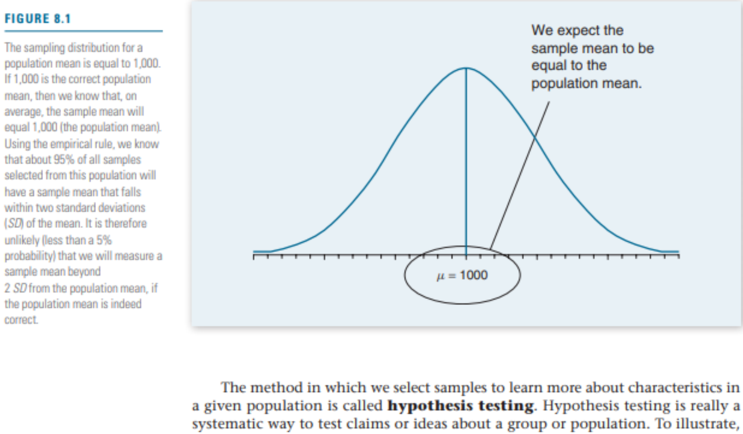


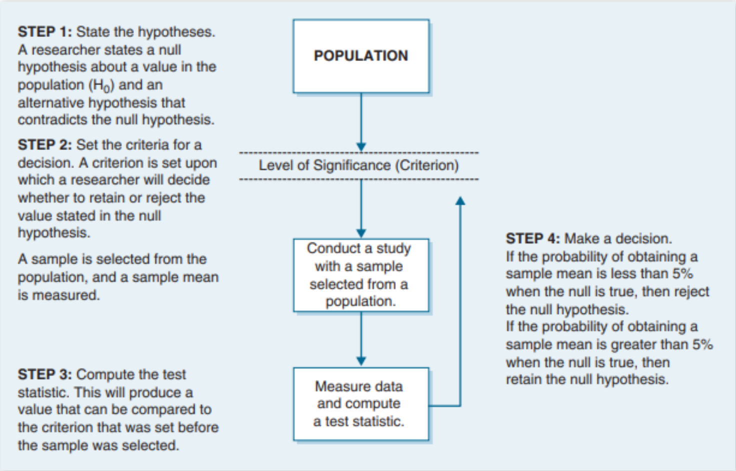
FIGURE 8.1

The sampling distribution for a population mean is equal to 1,000. If 1,000 is the correct population mean, then we know that, on average, the sample mean will equal 1,000 (the population mean). Using the empirical rule, we know that about 95% of all samples selected from this population will have a sample mean that falls within two standard deviations (*SD*) of the mean. It is therefore unlikely (less than a 5% probability) that we will measure a sample mean beyond 2 *SD* from the population mean, if the population mean is indeed correct.

We expect the sample mean to be equal to the population mean.

The method in which we select samples to learn more about characteristics in a given population is called **hypothesis testing**. Hypothesis testing is really a systematic way to test claims or ideas about a group or population. To illustrate,

Inferential Statistics: Hypothesis Testing



STEP 1: State the hypotheses. A researcher states a null hypothesis about a value in the population (H_0) and an alternative hypothesis that contradicts the null hypothesis.

STEP 2: Set the criteria for a decision. A criterion is set upon which a researcher will decide whether to retain or reject the value stated in the null hypothesis.

A sample is selected from the population, and a sample mean is measured.

STEP 3: Compute the test statistic. This will produce a value that can be compared to the criterion that was set before the sample was selected.

POPULATION

Level of Significance (Criterion)

Conduct a study with a sample selected from a population.

Measure data and compute a test statistic.

STEP 4: Make a decision. If the probability of obtaining a sample mean is less than 5% when the null is true, then reject the null hypothesis. If the probability of obtaining a sample mean is greater than 5% when the null is true, then retain the null hypothesis.

FIGURE 8.3

A summary of hypothesis testing.

https://www.sagepub.com/sites/default/files/upm-binaries/

Inferential Statistics: Hypothesis Testing - STEP I

H0: Null Hypothesis; H1: Alternate Hypothesis

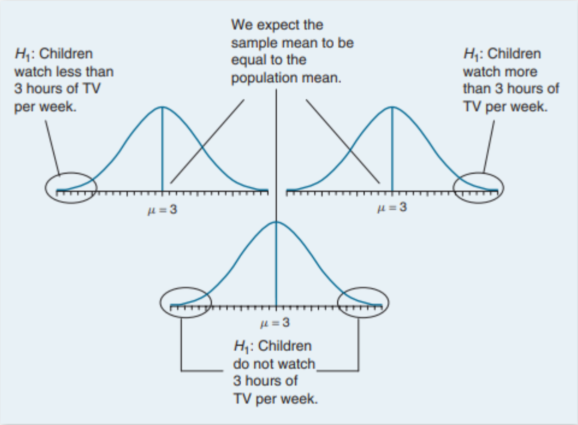

H_1: Children watch less than 3 hours of TV per week.
We expect the sample mean to be equal to the population mean.
H_1: Children watch more than 3 hours of TV per week.
$\mu = 3$
$\mu = 3$
$\mu = 3$
H_1: Children do not watch 3 hours of TV per week.

FIGURE 8.2

The alternative hypothesis determines whether to place the level of significance in one or both tails of a sampling distribution. Sample means that fall in the tails are unlikely to occur (less than a 5% probability) if the value stated for a population mean in the null hypothesis is true.

https://www.sagepub.com/sites/default/files/upm-binaries/40007_Chapter8.pdf

Inferential Statistics: Hypothesis Testing - STEP II: Decision Criteria

- **Level of significance**: A criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of obtaining a statistic measured in a sample if the value stated in the null hypothesis were true.
- The alternative hypothesis establishes where to place the level of significance.

Inferential Statistics: Hypothesis Testing - STEP III: Compute Test Statistic

- Suppose we measure a sample mean!
- To make a decision, we need to evaluate how likely this sample outcome is, if the population mean stated by the null hypothesis is true.
- We use a test statistic to determine this likelihood.
- Specifically, a test statistic tells us how far, or how many standard deviations, a sample mean is from the population mean. The larger the value of the test statistic, the further the distance, or number of standard deviations, a sample mean is from the population mean stated in the null hypothesis.
- The value of the test statistic is used to make a decision in Step 4.

Inferential Statistics: Hypothesis Testing - STEP IV: Make a Decision

- We use the value of the test statistic to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean, given that the value stated in the null hypothesis is true.
- 1. Reject the null hypothesis. The sample mean is associated with a low probability of occurrence when the null hypothesis is true
- 2. Retain the null hypothesis. The sample mean is associated with a high probability of occurrence when the null hypothesis is true.

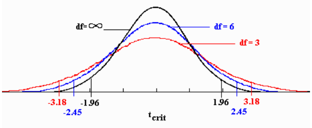
Inferential Statistics: Hypothesis Testing - Z Score

- What Is a Z-Test?
- A z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large. The test statistic is assumed to have a normal distribution, and nuisance parameters such as standard deviation should be known in order for an accurate z-test to be performed.
- z-statistic, or z-score, is a number representing how many standard deviations above or below the mean population a score derived from a z-test is.
- The Z-score allows you to decide if your sample is different from the population mean. In order to use z, you must know four things:
  - The population mean.
  - The population standard deviation.
  - The sample mean.
  - The sample size.

  Z score is a conversion of raw data to a standard score, when the conversion is based on the population mean and population standard deviation. When a full data set is available with us, we can compute the Z score. Z score is the subtraction of the population mean from the raw score and then divides the result with population standard deviation. T score is a conversion of raw data to the standard score when the conversion is based on the sample mean and sample standard deviation. When the population data set is not available, then we have to pick up some sample data to compute the sample mean and population standard deviation.

Inferential Statistics: Hypothesis Testing - T statistic

- The T Statistic is used in a T test when you are deciding if you should support or reject the null hypothesis.
- It's very similar to a Z-score and you use it in the same way: find a cut off point, find your t score, and compare the two. You use the t statistic when you have a small sample size, or if you don't know the population standard deviation.
- When you run a hypothesis test, you use the T statistic with a p value. The p-value tells you what the odds are that your results could have happened by chance.



Inferential Statistics: Hypothesis Testing - *p* value

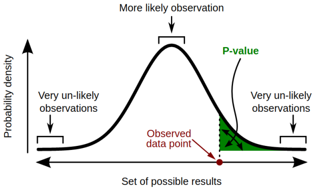
A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

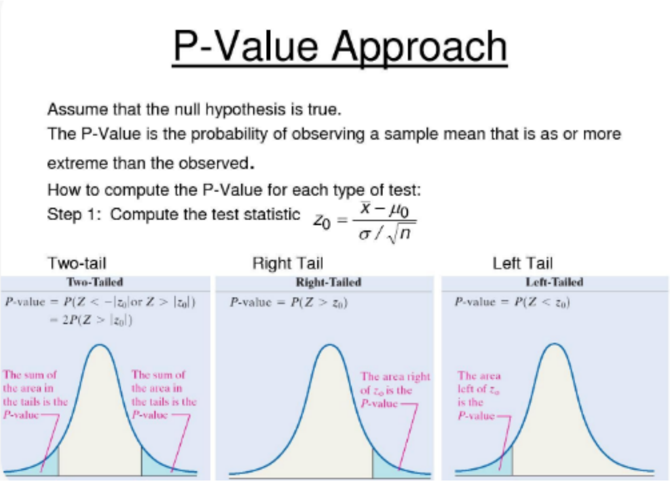
Using the p-value as a "score" is committing an egregious logical error:
the transposed conditional fallacy.



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

Inferential Statistics: Hypothesis Testing - *p* value

$$H_0 : \mu = \mu_0, H_1 : \mu > \mu_0$$



P-Value Approach

Assume that the null hypothesis is true.

The P-Value is the probability of observing a sample mean that is as or more extreme than the observed.

How to compute the P-Value for each type of test:

Step 1: Compute the test statistic $z_0 = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$

Two-tail	Right Tail	Left Tail
Two-Tailed	**Right-Tailed**	**Left-Tailed**
$P\text{-value} = P(Z < -\lvert z_0 \rvert \text{ or } Z > \lvert z_0 \rvert)$ $= 2P(Z > \lvert z_0 \rvert)$	$P\text{-value} = P(Z > z_0)$	$P\text{-value} = P(Z < z_0)$

The sum of the area in the tails is the P-value
The sum of the area in the tails is the P-value
The area right of z_0 is the P-value
The area left of z_o is the P-value

https://www.youtube.com/watch?v=HTZ8YKgDOMI

Inferential Statistics: Hypothesis Testing - *p* value

Standard Error

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

Z=(Sample mean - population mean)/(standard error)

$$Z_c = \frac{< X > - \mu}{\sigma_x}$$

$$P - value = 1 - Probability(Z_score)$$

If the P-value ¿ Significance Level, then we Fail To Reject The Null Hypothesis. Or else, if the P-value ¡ Significance Level, we Reject the Null Hypothesis.

`https://towardsdatascience.com/p-value-method-for-hypothesis-testing-c967c0f78d74`