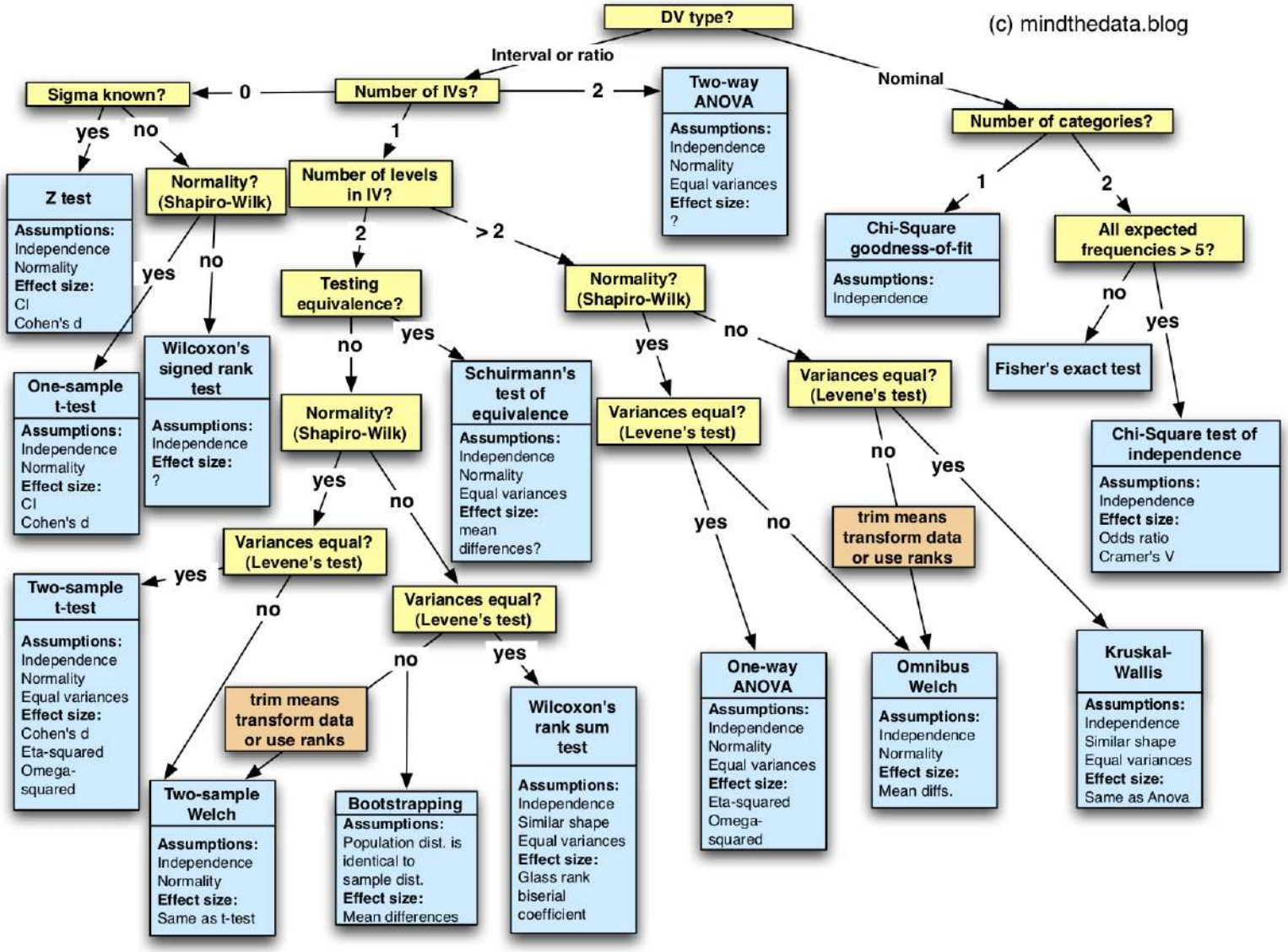
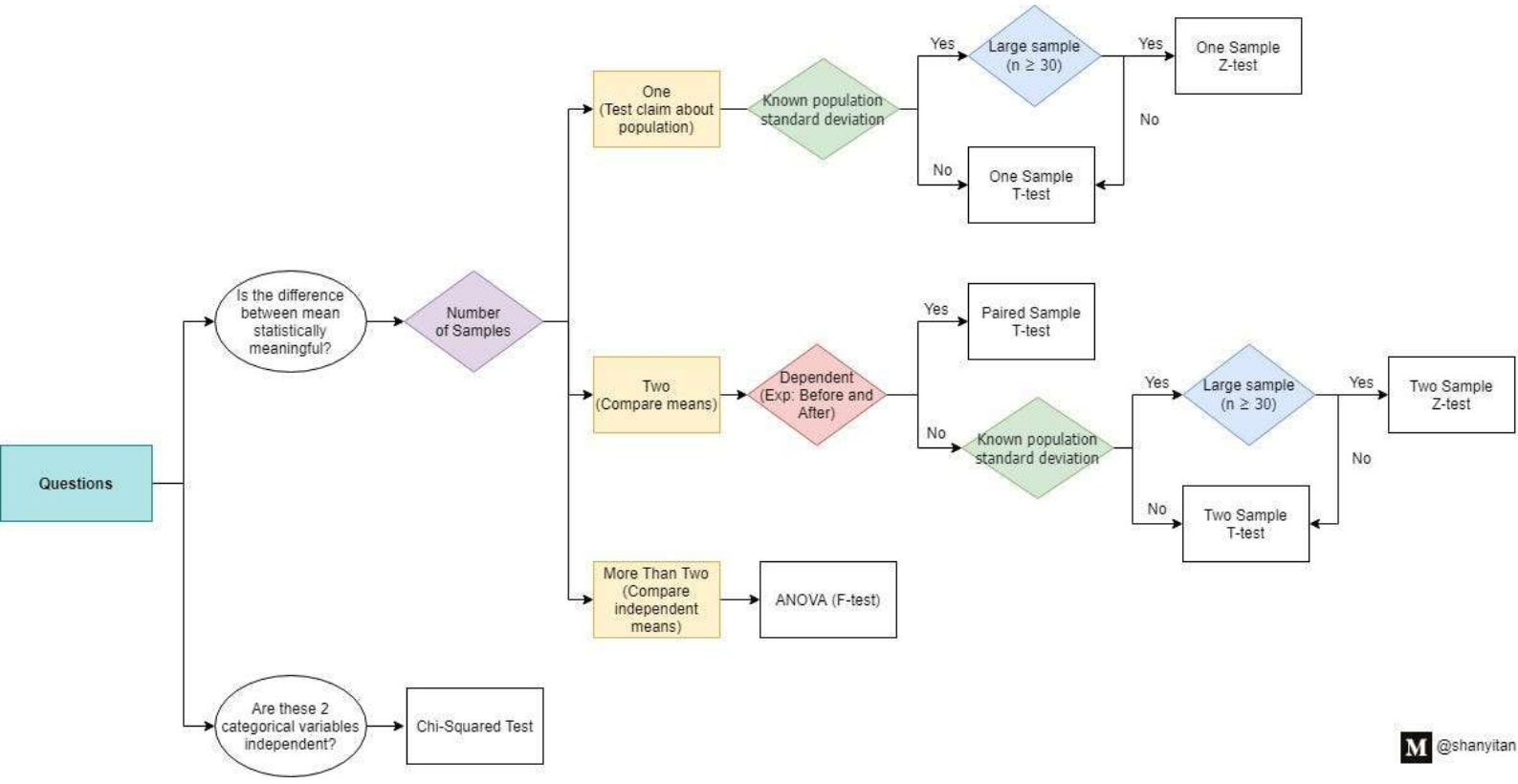
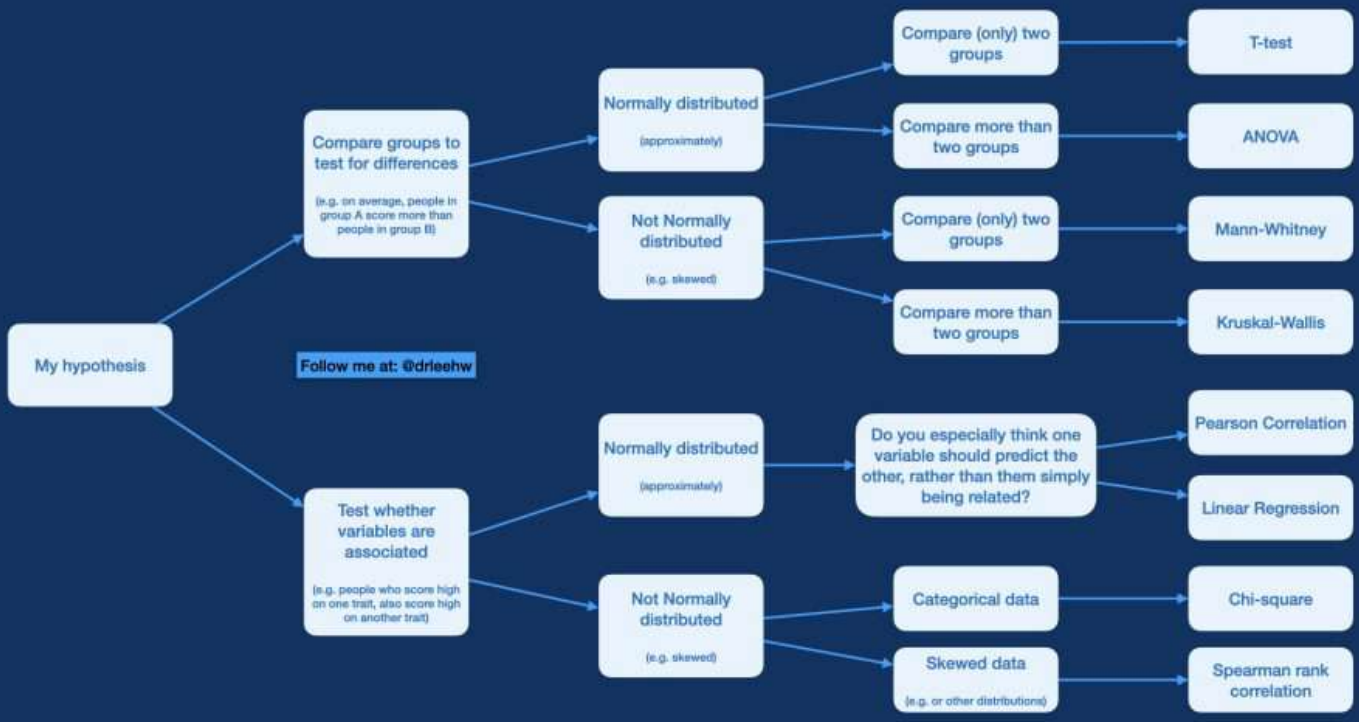

DV type?
(c) mindthedata.blog
Interval or ratio
Nominal
Sigma known?
0
Number of IVs?
2
Two-way ANOVA
Assumptions:
Independence
Normality
Equal variances
Effect size:
?
Number of categories?
yes
no
1
Z test
Assumptions:
Independence
Normality
Effect size:
CI
Cohen's d
Normality? (Shapiro-Wilk)
Number of levels in IV?
1
2
Chi-Square goodness-of-fit
Assumptions:
Independence
All expected frequencies > 5?
yes
no
> 2
Testing equivalence?
Normality? (Shapiro-Wilk)
Fisher's exact test
Chi-Square test of independence
Assumptions:
Independence
Effect size:
Odds ratio
Cramer's V
One-sample t-test
Assumptions:
Independence
Normality
Effect size:
CI
Cohen's d
Wilcoxon's signed rank test
Assumptions:
Independence
Effect size:
?
Schuirmann's test of equivalence
Assumptions:
Independence
Normality
Equal variances
Effect size:
mean differences?
Variances equal? (Levene's test)
Variances equal? (Levene's test)
trim means transform data or use ranks
Kruskal-Wallis
Assumptions:
Independence
Similar shape
Equal variances
Effect size:
Same as Anova
Two-sample t-test
Assumptions:
Independence
Normality
Equal variances
Effect size:
Cohen's d
Eta-squared
Omega-squared
Variances equal? (Levene's test)
One-way ANOVA
Assumptions:
Independence
Normality
Equal variances
Effect size:
Eta-squared
Omega-squared
Omnibus Welch
Assumptions:
Independence
Normality
Effect size:
Mean diffs.
trim means transform data or use ranks
Wilcoxon's rank sum test
Assumptions:
Independence
Similar shape
Equal variances
Effect size:
Glass rank biserial coefficient
Two-sample Welch
Assumptions:
Independence
Normality
Effect size:
Same as t-test
Bootstrapping
Assumptions:
Population dist. is identical to sample dist.
Effect size:
Mean differences


Questions
Is the difference between mean statistically meaningful?
Number of Samples
One (Test claim about population)
Known population standard deviation
Yes
Large sample (n ≥ 30)
Yes
One Sample Z-test
No
No
One Sample T-test
Two (Compare means)
Dependent (Exp: Before and After)
Yes
Paired Sample T-test
No
Known population standard deviation
Yes
Large sample (n ≥ 30)
Yes
Two Sample Z-test
No
No
Two Sample T-test
More Than Two (Compare independent means)
ANOVA (F-test)
Are these 2 categorical variables independent?
Chi-Squared Test
@shanyitan

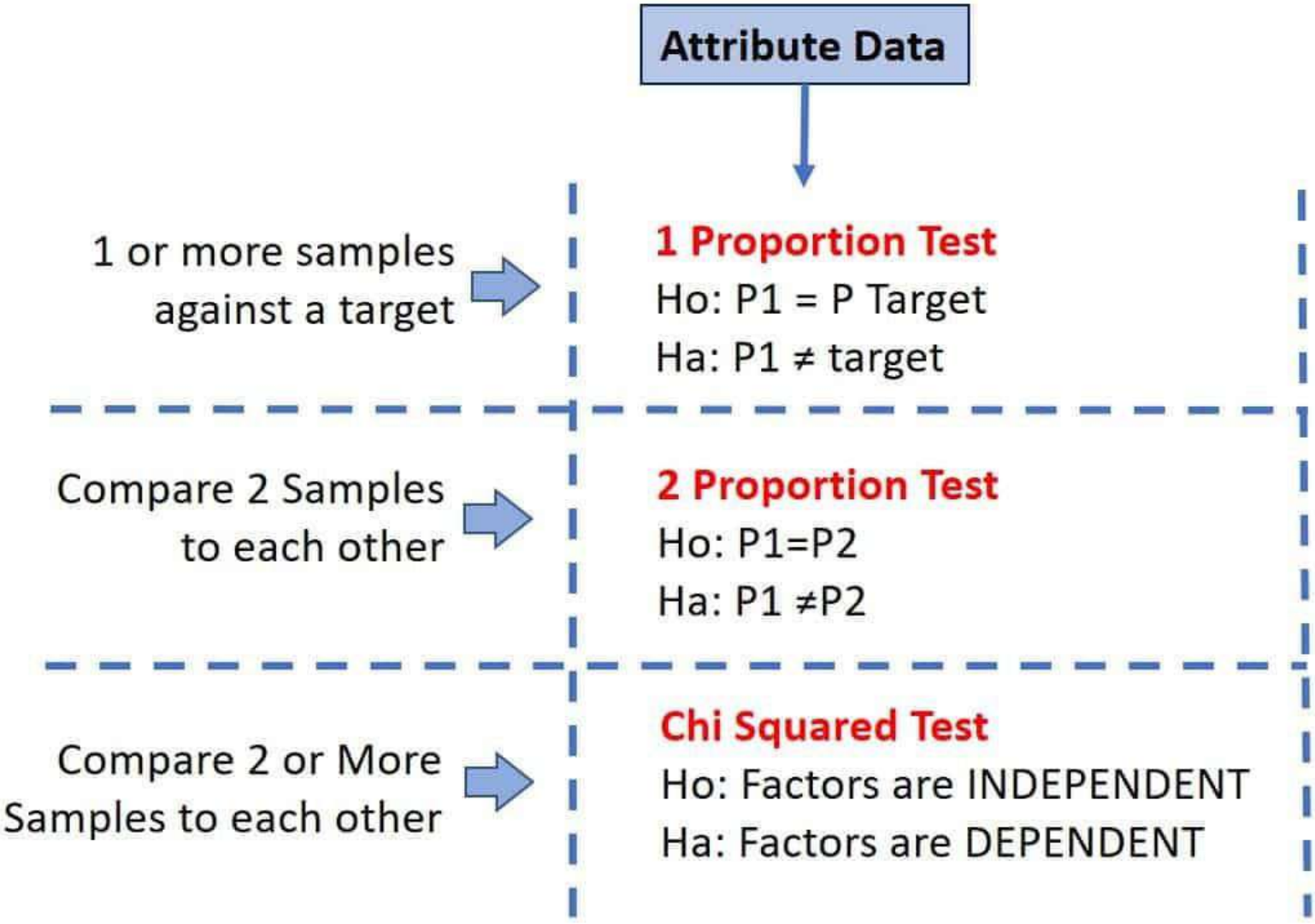
Choosing a statistical test

Remember always to check the assumptions of the test you choose.



Follow me at: @drleehw

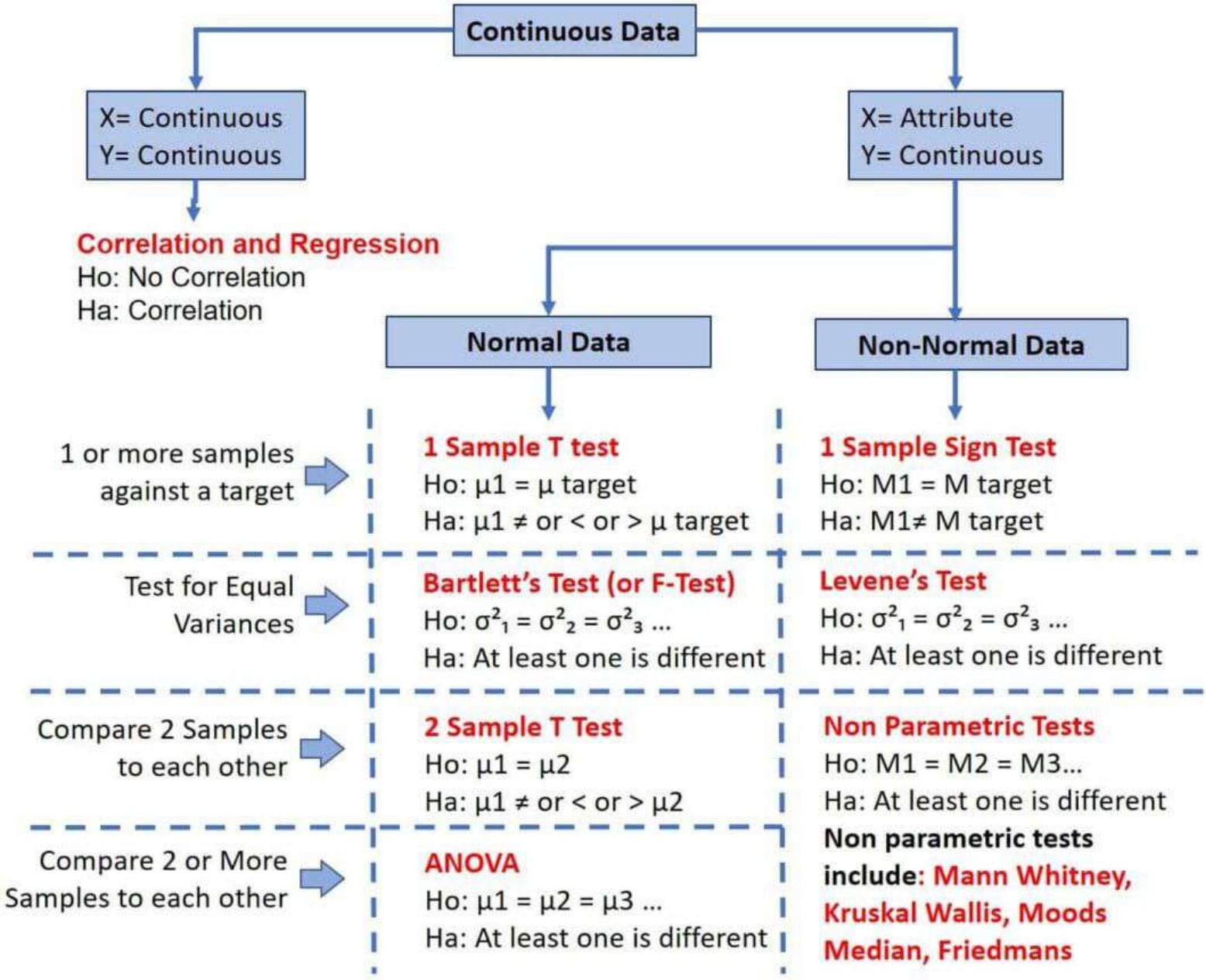
Hypothesis Testing Cheat Sheet



For all tests, choose appropriate α (or alpha). The default α (or alpha) is 0.05

- P > 0.05 Fail to Reject Ho (this is the Null Hypothesis)
- P < 0.05 Reject Ho (This is the Alternative Hypothesis)

Hypothesis Testing Cheat Sheet



For all tests, choose appropriate α (or alpha). The default α (or alpha) is 0.05

- P > 0.05 Fail to Reject Ho (this is the Null Hypothesis)
- P < 0.05 Reject Ho (This is the Alternative Hypothesis)

Hypothesis Testing

www.FairlyNerdy.com

Type Of Test	Purpose	Example	Equation	Comment	Excel Function
Z Test	Test if the average of a single population is equal to a target value	Do babies born at this hospital weigh more than the city average	$Z = \frac{\bar{x} - u_0}{\frac{\sigma}{\sqrt{n}}}$	Z test does not need df σ = population standard deviation	=Ztest(array,x,sigma)
1 Sample T-Test	Test if the average of a single population is equal to a target value	Is the average height of male college students greater than 6.0 feet?	$t = \frac{\bar{x} - u_0}{\frac{s}{\sqrt{n}}}$ $df = n - 1$	s = sample standard deviation	no built in equation use =STDEVA for standard deviation use =AVERAGE for mean use =T.DIST.RT to get 1 tailed confidence use =T.DIST.2T to get 2 tailed confidence
Paired T-Test	Test if the average of the differences between paired or dependent samples is equal to a target value	Weigh a set of people. Put them on a diet plan. Weigh them after. Is the average weight loss significant enough to conclude the diet works?	$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$ $df = n - 1$	d bar = average difference between samples s = sample deviation of the difference n = count of one set of the pairs (don't double count)	=TTEST(Array1,Array2,*,1) * -> 1 for 1 tailed, 2 for 2 tailed
2 Sample T-Test **Equal Variance**	Test if the difference between the averages of two independent populations is equal to a target value	Do cats eat more of type A food than type B food	$df = n_1 + n_2 - 2$ $t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	n1, n2 = count of sample 1, 2	=TTEST(Array1,Array2,*,2)
2 Sample T-Test **Unequal Variance**	Test if the difference between the averages of two independent populations is equal to a target value	Is the average speed of cyclists during rush hour greater than the average speed of drivers	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$		=TTEST(Array1,Array2,*,3)

VIP Refresher: Probabilities and Statistics

Afshine Amidi and Shervine Amidi

August 6, 2018

Introduction to Probability and Combinatorics

❒ **Sample space** – The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by S.

❒ **Event** – Any subset E of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in E, then we say that E has occurred.

❒ **Axioms of probability** – For each event E, we denote $P(E)$ as the probability of event E occuring. By noting E_1,...,E_n mutually exclusive events, we have the 3 following axioms:

$$(1) \quad \boxed{0 \leqslant P(E) \leqslant 1} \quad (2) \quad \boxed{P(S) = 1} \quad (3) \quad \boxed{P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)}$$

❒ **Permutation** – A permutation is an arrangement of r objects from a pool of n objects, in a given order. The number of such arrangements is given by $P(n,r)$, defined as:

$$\boxed{P(n,r) = \frac{n!}{(n-r)!}}$$

❒ **Combination** – A combination is an arrangement of r objects from a pool of n objects, where the order does not matter. The number of such arrangements is given by $C(n,r)$, defined as:

$$\boxed{C(n,r) = \frac{P(n,r)}{r!} = \frac{n!}{r!(n-r)!}}$$

Remark: we note that for $0 \leqslant r \leqslant n$, we have $P(n,r) \geqslant C(n,r)$.

Conditional Probability

❒ **Bayes' rule** – For events A and B such that $P(B) > 0$, we have:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Remark: we have $P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$.

❒ **Partition** – Let $\{A_i, i \in [\![1,n]\!]\}$ be such that for all i, $A_i \neq \varnothing$. We say that $\{A_i\}$ is a partition if we have:

$$\boxed{\forall i \neq j, A_i \cap A_j = \emptyset \quad \text{and} \quad \bigcup_{i=1}^{n} A_i = S}$$

Remark: for any event B in the sample space, we have $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$.

❒ **Extended form of Bayes' rule** – Let $\{A_i, i \in [\![1,n]\!]\}$ be a partition of the sample space. We have:

$$\boxed{P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}}$$

❒ **Independence** – Two events A and B are independent if and only if we have:

$$\boxed{P(A \cap B) = P(A)P(B)}$$

Random Variables

❒ **Random variable** – A random variable, often noted X, is a function that maps every element in a sample space to a real line.

❒ **Cumulative distribution function (CDF)** – The cumulative distribution function F, which is monotonically non-decreasing and is such that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$, is defined as:

$$\boxed{F(x) = P(X \leqslant x)}$$

Remark: we have $P(a < X \leqslant B) = F(b) - F(a)$.

❒ **Probability density function (PDF)** – The probability density function f is the probability that X takes on values between two adjacent realizations of the random variable.

❒ **Relationships involving the PDF and CDF** – Here are the important properties to know in the discrete (D) and the continuous (C) cases.

Case	CDF F	PDF f	Properties of PDF
(D)	$F(x) = \sum_{x_i \leqslant x} P(X = x_i)$	$f(x_j) = P(X = x_j)$	$0 \leqslant f(x_j) \leqslant 1$ and $\sum_j f(x_j) = 1$
(C)	$F(x) = \int_{-\infty}^{x} f(y)dy$	$f(x) = \frac{dF}{dx}$	$f(x) \geqslant 0$ and $\int_{-\infty}^{+\infty} f(x)dx = 1$

❒ **Variance** – The variance of a random variable, often noted $\text{Var}(X)$ or σ^2, is a measure of the spread of its distribution function. It is determined as follows:

$$\boxed{\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2}$$

❒ **Standard deviation** – The standard deviation of a random variable, often noted σ, is a measure of the spread of its distribution function which is compatible with the units of the actual random variable. It is determined as follows:

$$\boxed{\sigma = \sqrt{\text{Var}(X)}}$$

❒ **Expectation and Moments of the Distribution** – Here are the expressions of the expected value $E[X]$, generalized expected value $E[g(X)]$, k^{th} moment $E[X^k]$ and characteristic function $\psi(\omega)$ for the discrete and continuous cases:

Case	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(D)	$\sum_{i=1}^{n} x_i f(x_i)$	$\sum_{i=1}^{n} g(x_i) f(x_i)$	$\sum_{i=1}^{n} x_i^k f(x_i)$	$\sum_{i=1}^{n} f(x_i) e^{i\omega x_i}$
(C)	$\int_{-\infty}^{+\infty} x f(x) dx$	$\int_{-\infty}^{+\infty} g(x) f(x) dx$	$\int_{-\infty}^{+\infty} x^k f(x) dx$	$\int_{-\infty}^{+\infty} f(x) e^{i\omega x} dx$

Remark: we have $e^{i\omega x} = \cos(\omega x) + i \sin(\omega x)$.

❒ **Revisiting the k^{th} moment** – The k^{th} moment can also be computed with the characteristic function as follows:

$$\boxed{E[X^k] = \frac{1}{i^k} \left[\frac{\partial^k \psi}{\partial \omega^k} \right]_{\omega=0}}$$

❒ **Transformation of random variables** – Let the variables X and Y be linked by some function. By noting f_X and f_Y the distribution function of X and Y respectively, we have:

$$\boxed{f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|}$$

❒ **Leibniz integral rule** – Let g be a function of x and potentially c, and a, b boundaries that may depend on c. We have:

$$\boxed{\frac{\partial}{\partial c} \left(\int_a^b g(x) dx \right) = \frac{\partial b}{\partial c} \cdot g(b) - \frac{\partial a}{\partial c} \cdot g(a) + \int_a^b \frac{\partial g}{\partial c}(x) dx}$$

❒ **Chebyshev's inequality** – Let X be a random variable with expected value μ and standard deviation σ. For $k, \sigma > 0$, we have the following inequality:

$$\boxed{P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}}$$

Jointly Distributed Random Variables

❒ **Conditional density** – The conditional density of X with respect to Y, often noted $f_{X|Y}$, is defined as follows:

$$\boxed{f_{X|Y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}}$$

❒ **Independence** – Two random variables X and Y are said to be independent if we have:

$$\boxed{f_{XY}(x,y) = f_X(x) f_Y(y)}$$

❒ **Marginal density and cumulative distribution** – From the joint density probability function f_{XY}, we have:

Case	Marginal density	Cumulative function
(D)	$f_X(x_i) = \sum_j f_{XY}(x_i, y_j)$	$F_{XY}(x,y) = \sum_{x_i \leqslant x} \sum_{y_j \leqslant y} f_{XY}(x_i, y_j)$
(C)	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y) dy$	$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y') dx' dy'$

❒ **Distribution of a sum of independent random variables** – Let $Y = X_1 + ... + X_n$ with $X_1, ..., X_n$ independent. We have:

$$\boxed{\psi_Y(\omega) = \prod_{k=1}^{n} \psi_{X_k}(\omega)}$$

❒ **Covariance** – We define the covariance of two random variables X and Y, that we note σ_{XY}^2 or more commonly $\text{Cov}(X,Y)$, as follows:

$$\boxed{\text{Cov}(X,Y) \triangleq \sigma_{XY}^2 = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y}$$

❒ **Correlation** – By noting σ_X, σ_Y the standard deviations of X and Y, we define the correlation between the random variables X and Y, noted ρ_{XY}, as follows:

$$\boxed{\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}}$$

Remarks: For any X, Y, we have $\rho_{XY} \in [-1,1]$. If X and Y are independent, then $\rho_{XY} = 0$.

❒ **Main distributions** – Here are the main distributions to have in mind:

Type	Distribution	PDF	$\psi(\omega)$	$E[X]$	$\text{Var}(X)$
(D)	$X \sim \mathcal{B}(n,p)$ Binomial	$P(X=x) = \binom{n}{x} p^x q^{n-x}$ $x \in [\![0,n]\!]$	$(pe^{i\omega} + q)^n$	np	npq
	$X \sim \text{Po}(\mu)$ Poisson	$P(X=x) = \frac{\mu^x}{x!} e^{-\mu}$ $x \in \mathbb{N}$	$e^{\mu(e^{i\omega}-1)}$	μ	μ
(C)	$X \sim \mathcal{U}(a,b)$ Uniform	$f(x) = \frac{1}{b-a}$ $x \in [a,b]$	$\frac{e^{i\omega b} - e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$X \sim \mathcal{N}(\mu,\sigma)$ Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $x \in \mathbb{R}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	μ	σ^2
	$X \sim \text{Exp}(\lambda)$ Exponential	$f(x) = \lambda e^{-\lambda x}$ $x \in \mathbb{R}_+$	$\frac{1}{1 - \frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Parameter estimation

❒ **Random sample** – A random sample is a collection of n random variables $X_1, ..., X_n$ that are independent and identically distributed with X.

❒ **Estimator** – An estimator $\hat{\theta}$ is a function of the data that is used to infer the value of an unknown parameter θ in a statistical model.

❒ **Bias** – The bias of an estimator $\hat{\theta}$ is defined as being the difference between the expected value of the distribution of $\hat{\theta}$ and the true value, i.e.:

$$\boxed{\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta}$$

Remark: an estimator is said to be unbiased when we have $E[\hat{\theta}] = \theta$.

❒ **Sample mean and variance** – The sample mean and the sample variance of a random sample are used to estimate the true mean μ and the true variance σ^2 of a distribution, are noted $\overline{X}$ and s^2 respectively, and are such that:

$$\boxed{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i} \quad \text{and} \quad \boxed{s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

❒ **Central Limit Theorem** – Let us have a random sample $X_1, ..., X_n$ following a given distribution with mean μ and variance σ^2, then we have:

$$\boxed{\overline{X} \underset{n\to+\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$$

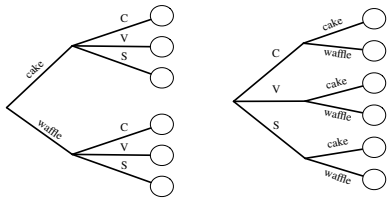
Probability Cheatsheet v2.0

Compiled by William Chen (http://wzchen.com) and Joe Blitzstein, with contributions from Sebastian Chiu, Yuan Jiang, Yuqi Hou, and Jessy Hwang. Material based on Joe Blitzstein's (@stat110) lectures (http://stat110.net) and Blitzstein/Hwang's Introduction to Probability textbook (http://bit.ly/introprobability). Please share comments, suggestions, and errors at http://github.com/wzchen/probability_cheatsheet.

Last Updated September 4, 2015

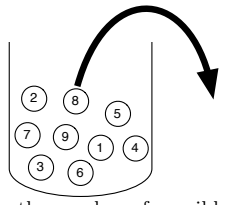
Counting

Multiplication Rule



Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, ..., and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \ldots n_r$ possibilities for the whole experiment.

Sampling Table



The sampling table gives the number of possible samples of size k out of a population of size n, under various assumptions about how the sample is collected.

	Order Matters	Not Matter
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Naive Definition of Probability

If all outcomes are equally likely, the probability of an event A happening is:

$$P_{\text{naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}$$

Thinking Conditionally

Independence

Independent Events A and B are independent if knowing whether A occurred gives no information about whether B occurred. More formally, A and B (which have nonzero probability) are independent if and only if one of the following equivalent statements holds:

$$P(A \cap B) = P(A)P(B)$$
$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Conditional Independence A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$. Conditional independence does not imply independence, and independence does not imply conditional independence.

Unions, Intersections, and Complements

De Morgan's Laws A useful identity that can make calculating probabilities of unions easier by relating them to intersections, and vice versa. Analogous results hold with more than two sets.

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Joint, Marginal, and Conditional

Joint Probability $P(A \cap B)$ or $P(A, B)$ – Probability of A and B.

Marginal (Unconditional) Probability $P(A)$ – Probability of A.

Conditional Probability $P(A|B) = P(A, B)/P(B)$ – Probability of A, given that B occurred.

Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B. Any theorem that holds for probability also holds for conditional probability.

Probability of an Intersection or Union

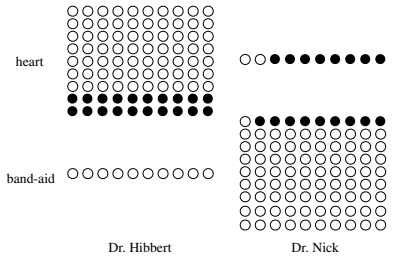
Intersections via Conditioning

$$P(A, B) = P(A)P(B|A)$$
$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

Unions via Inclusion-Exclusion

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Simpson's Paradox



It is possible to have

$$P(A \mid B, C) < P(A \mid B^c, C) \text{ and } P(A \mid B, C^c) < P(A \mid B^c, C^c)$$

yet also $P(A \mid B) > P(A \mid B^c)$.

Law of Total Probability (LOTP)

Let $B_1, B_2, B_3, \ldots B_n$ be a *partition* of the sample space (i.e., they are disjoint and their union is the entire sample space).

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$$
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

For **LOTP with extra conditioning**, just add in another event C!

$$P(A|C) = P(A|B_1, C)P(B_1|C) + \cdots + P(A|B_n, C)P(B_n|C)$$
$$P(A|C) = P(A \cap B_1|C) + P(A \cap B_2|C) + \cdots + P(A \cap B_n|C)$$

Special case of LOTP with B and B^c as partition:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$
$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Bayes' Rule

Bayes' Rule, and with extra conditioning (just add in C!)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)}$$

We can also write

$$P(A|B, C) = \frac{P(A, B, C)}{P(B, C)} = \frac{P(B, C|A)P(A)}{P(B, C)}$$

Odds Form of Bayes' Rule

$$\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)} \frac{P(A)}{P(A^c)}$$

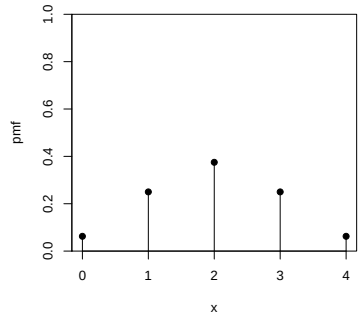
The *posterior odds* of A are the *likelihood ratio* times the *prior odds*.

Random Variables and their Distributions

PMF, CDF, and Independence

Probability Mass Function (PMF) Gives the probability that a *discrete* random variable takes on the value x.

$$p_X(x) = P(X = x)$$

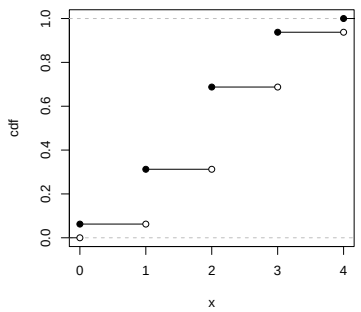


The PMF satisfies

$$p_X(x) \geq 0 \text{ and } \sum_x p_X(x) = 1$$

Cumulative Distribution Function (CDF) Gives the probability that a random variable is less than or equal to x.

$$F_X(x) = P(X \leq x)$$



The CDF is an increasing, right-continuous function with

$$F_X(x) \to 0 \text{ as } x \to -\infty \text{ and } F_X(x) \to 1 \text{ as } x \to \infty$$

Independence Intuitively, two random variables are independent if knowing the value of one gives no information about the other. Discrete r.v.s X and Y are independent if for *all* values of x and y

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Expected Value and Indicators

Expected Value and Linearity

Expected Value (a.k.a. *mean*, *expectation*, or *average*) is a weighted average of the possible outcomes of our random variable. Mathematically, if $x_1, x_2, x_3, \ldots$ are all of the distinct possible values that X can take, the expected value of X is

$$E(X) = \sum_i x_i P(X = x_i)$$

X	Y	$X+Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	−3	−2
1	0	1
5	9	14
4	1	5
...	...	...
$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\sum_{i=1}^{n} y_i$	$\frac{1}{n}\sum_{i=1}^{n}(x_i+y_i)$
$E(X)$	$E(Y)$	$E(X+Y)$

Linearity For any r.v.s X and Y, and constants a, b, c,

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

Same distribution implies same mean If X and Y have the same distribution, then $E(X) = E(Y)$ and, more generally,

$$E(g(X)) = E(g(Y))$$

Conditional Expected Value is defined like expectation, only conditioned on any event A.

$$E(X|A) = \sum_x xP(X = x|A)$$

Indicator Random Variables

Indicator Random Variable is a random variable that takes on the value 1 or 0. It is always an indicator of some event: if the event occurs, the indicator is 1; otherwise it is 0. They are useful for many problems about counting how many events of some kind occur. Write

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Note that $I_A^2 = I_A, I_A I_B = I_{A\cap B}$, and $I_{A\cup B} = I_A + I_B - I_A I_B$.

Distribution $I_A \sim \text{Bern}(p)$ where $p = P(A)$.

Fundamental Bridge The expectation of the indicator for event A is the probability of event A: $E(I_A) = P(A)$.

Variance and Standard Deviation

$$\text{Var}(X) = E\left(X - E(X)\right)^2 = E(X^2) - (E(X))^2$$

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

Continuous RVs, LOTUS, UoU

Continuous Random Variables (CRVs)

What's the probability that a CRV is in an interval? Take the difference in CDF values (or use the PDF as described later).

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, this becomes

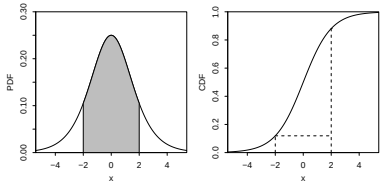
$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

What is the Probability Density Function (PDF)? The PDF f is the derivative of the CDF F.

$$F'(x) = f(x)$$

A PDF is nonnegative and integrates to 1. By the fundamental theorem of calculus, to get from PDF back to CDF we can integrate:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



To find the probability that a CRV takes on a value in an interval, integrate the PDF over that interval.

$$F(b) - F(a) = \int_a^b f(x)dx$$

How do I find the expected value of a CRV? Analogous to the discrete case, where you sum x times the PMF, for CRVs you integrate x times the PDF.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

LOTUS

Expected value of a function of an r.v. The expected value of X is defined this way:

$$E(X) = \sum_x xP(X = x) \text{ (for discrete } X)$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \text{ (for continuous } X)$$

The **Law of the Unconscious Statistician (LOTUS)** states that you can find the expected value of a *function of a random variable*, $g(X)$, in a similar way, by replacing the x in front of the PMF/PDF by $g(x)$ but still working with the PMF/PDF of X:

$$E(g(X)) = \sum_x g(x)P(X = x) \text{ (for discrete } X)$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \text{ (for continuous } X)$$

What's a function of a random variable? A function of a random variable is also a random variable. For example, if X is the number of bikes you see in an hour, then $g(X) = 2X$ is the number of bike wheels you see in that hour and $h(X) = \binom{X}{2} = \frac{X(X-1)}{2}$ is the number of *pairs* of bikes such that you see both of those bikes in that hour.

What's the point? You don't need to know the PMF/PDF of $g(X)$ to find its expected value. All you need is the PMF/PDF of X.

Universality of Uniform (UoU)

When you plug any CRV into its own CDF, you get a Uniform(0,1) random variable. When you plug a Uniform(0,1) r.v. into an inverse CDF, you get an r.v. with that CDF. For example, let's say that a random variable X has CDF

$$F(x) = 1 - e^{-x}, \text{ for } x > 0$$

By UoU, if we plug X into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim \text{Unif}(0, 1)$$

Similarly, if $U \sim \text{Unif}(0, 1)$ then $F^{-1}(U)$ has CDF F. The key point is that for any continuous random variable X, we can transform it into a Uniform random variable and back by using its CDF.

Moments and MGFs

Moments

Moments describe the shape of a distribution. Let X have mean μ and standard deviation σ, and $Z = (X - \mu)/\sigma$ be the *standardized* version of X. The kth moment of X is $\mu_k = E(X^k)$ and the kth standardized moment of X is $m_k = E(Z^k)$. The mean, variance, skewness, and kurtosis are important summaries of the shape of a distribution.

Mean $E(X) = \mu_1$

Variance $\text{Var}(X) = \mu_2 - \mu_1^2$

Skewness $\text{Skew}(X) = m_3$

Kurtosis $\text{Kurt}(X) = m_4 - 3$

Moment Generating Functions

MGF For any random variable X, the function

$$M_X(t) = E(e^{tX})$$

is the **moment generating function (MGF)** of X, if it exists for all t in some open interval containing 0. The variable t could just as well have been called u or v. It's a bookkeeping device that lets us work with the *function* M_X rather than the *sequence* of moments.

Why is it called the Moment Generating Function? Because the kth derivative of the moment generating function, evaluated at 0, is the kth moment of X.

$$\mu_k = E(X^k) = M_X^{(k)}(0)$$

This is true by Taylor expansion of e^{tX} since

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} \frac{E(X^k)t^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!}$$

MGF of linear functions If we have $Y = aX + b$, then

$$M_Y(t) = E(e^{t(aX+b)}) = e^{bt}E(e^{(at)X}) = e^{bt}M_X(at)$$

Uniqueness *If it exists, the MGF uniquely determines the distribution.* This means that for any two random variables X and Y, they are distributed the same (their PMFs/PDFs are equal) if and only if their MGFs are equal.

Summing Independent RVs by Multiplying MGFs. If X and Y are independent, then

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X(t) \cdot M_Y(t)$$

The MGF of the sum of two random variables is the product of the MGFs of those two random variables.

Joint PDFs and CDFs

Joint Distributions

The **joint CDF** of X and Y is

$$F(x, y) = P(X \leq x, Y \leq y)$$

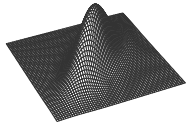
In the discrete case, X and Y have a **joint PMF**

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

In the continuous case, they have a **joint PDF**

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint PMF/PDF must be nonnegative and sum/integrate to 1.



Conditional Distributions

Conditioning and Bayes' rule for discrete r.v.s

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Conditioning and Bayes' rule for continuous r.v.s

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

Hybrid Bayes' rule

$$f_X(x|A) = \frac{P(A|X = x)f_X(x)}{P(A)}$$

Marginal Distributions

To find the distribution of one (or more) random variables from a joint PMF/PDF, sum/integrate over the unwanted random variables.

Marginal PMF from joint PMF

$$P(X = x) = \sum_y P(X = x, Y = y)$$

Marginal PDF from joint PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Independence of Random Variables

Random variables X and Y are independent if and only if any of the following conditions holds:

- Joint CDF is the product of the marginal CDFs
- Joint PMF/PDF is the product of the marginal PMFs/PDFs
- Conditional distribution of Y given X is the marginal distribution of Y

Write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent.

Multivariate LOTUS

LOTUS in more than one dimension is analogous to the 1D LOTUS. For discrete random variables:

$$E(g(X, Y)) = \sum_x \sum_y g(x, y)P(X = x, Y = y)$$

For continuous random variables:

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$$

Covariance and Transformations

Covariance and Correlation

Covariance is the analog of variance for two random variables.

$$\text{Cov}(X, Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X, X) = E(X^2) - (E(X))^2 = \text{Var}(X)$$

Correlation is a standardized version of covariance that is always between -1 and 1.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Covariance and Independence If two random variables are independent, then they are uncorrelated. The converse is not necessarily true (e.g., consider $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$).

$$X \perp\!\!\!\perp Y \longrightarrow \text{Cov}(X, Y) = 0 \longrightarrow E(XY) = E(X)E(Y)$$

Covariance and Variance The variance of a sum can be found by

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)$$

If X and Y are independent then they have covariance 0, so

$$X \perp\!\!\!\perp Y \Longrightarrow \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

If $X_1, X_2, \ldots, X_n$ are identically distributed and have the same covariance relationships (often by **symmetry**), then

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = n\text{Var}(X_1) + 2\binom{n}{2}\text{Cov}(X_1, X_2)$$

Covariance Properties For random variables W, X, Y, Z and constants a, b:

$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(Y, X) \\
\text{Cov}(X + a, Y + b) &= \text{Cov}(X, Y) \\
\text{Cov}(aX, bY) &= ab\text{Cov}(X, Y) \\
\text{Cov}(W + X, Y + Z) &= \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) \\
&\quad + \text{Cov}(X, Z)
\end{aligned}$$

Correlation is location-invariant and scale-invariant For any constants a, b, c, d with a and c nonzero,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

Transformations

One Variable Transformations Let's say that we have a random variable X with PDF $f_X(x)$, but we are also interested in some function of X. We call this function $Y = g(X)$. Also let $y = g(x)$. If g is differentiable and strictly increasing (or strictly decreasing), then the PDF of Y is

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

The derivative of the inverse transformation is called the **Jacobian**.

Two Variable Transformations Similarly, let's say we know the joint PDF of U and V but are also interested in the random vector (X, Y) defined by $(X, Y) = g(U, V)$. Let

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}$$

be the **Jacobian matrix**. If the entries in this matrix exist and are continuous, and the determinant of the matrix is never 0, then

$$f_{X,Y}(x, y) = f_{U,V}(u, v)\left|\left|\frac{\partial(u, v)}{\partial(x, y)}\right|\right|$$

The inner bars tells us to take the matrix's determinant, and the outer bars tell us to take the absolute value. In a 2×2 matrix,

$$\left|\left| \begin{matrix} a & b \\ c & d \end{matrix} \right|\right| = |ad - bc|$$

Convolutions

Convolution Integral If you want to find the PDF of the sum of two independent CRVs X and Y, you can do the following integral:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x)dx$$

Example Let $X, Y \sim \mathcal{N}(0, 1)$ be i.i.d. Then for each fixed t,

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-(t-x)^2/2}dx$$

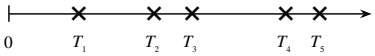
By completing the square and using the fact that a Normal PDF integrates to 1, this works out to $f_{X+Y}(t)$ being the $\mathcal{N}(0, 2)$ PDF.

Poisson Process

Definition We have a **Poisson process** of rate λ arrivals per unit time if the following conditions hold:

1. The number of arrivals in a time interval of length t is Pois(λt).
2. Numbers of arrivals in disjoint time intervals are independent.

For example, the numbers of arrivals in the time intervals $[0, 5]$, $(5, 12)$, and $[13, 23)$ are independent with Pois(5λ), Pois(7λ), Pois(10λ) distributions, respectively.



Count-Time Duality Consider a Poisson process of emails arriving in an inbox at rate λ emails per hour. Let T_n be the time of arrival of the nth email (relative to some starting time 0) and N_t be the number of emails that arrive in $[0, t]$. Let's find the distribution of T_1. The event $T_1 > t$, the event that you have to wait more than t hours to get the first email, is the same as the event $N_t = 0$, which is the event that there are no emails in the first t hours. So

$$P(T_1 > t) = P(N_t = 0) = e^{-\lambda t} \longrightarrow P(T_1 \leq t) = 1 - e^{-\lambda t}$$

Thus we have $T_1 \sim$ Expo(λ). By the memoryless property and similar reasoning, the interarrival times between emails are i.i.d. Expo(λ), i.e., the differences $T_n - T_{n-1}$ are i.i.d. Expo(λ).

Order Statistics

Definition Let's say you have n i.i.d. r.v.s $X_1, X_2, \ldots, X_n$. If you arrange them from smallest to largest, the ith element in that list is the ith order statistic, denoted $X_{(i)}$. So $X_{(1)}$ is the smallest in the list and $X_{(n)}$ is the largest in the list.

Note that the order statistics are *dependent*, e.g., learning $X_{(4)} = 42$ gives us the information that $X_{(1)}, X_{(2)}, X_{(3)}$ are ≤ 42 and $X_{(5)}, X_{(6)}, \ldots, X_{(n)}$ are ≥ 42.

Distribution Taking n i.i.d. random variables $X_1, X_2, \ldots, X_n$ with CDF $F(x)$ and PDF $f(x)$, the CDF and PDF of $X_{(i)}$ are:

$$F_{X_{(i)}}(x) = P(X_{(i)} \leq x) = \sum_{k=i}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

$$f_{X_{(i)}}(x) = n\binom{n-1}{i-1} F(x)^{i-1} (1 - F(x))^{n-i} f(x)$$

Uniform Order Statistics The jth order statistic of i.i.d. $U_1, \ldots, U_n \sim$ Unif$(0, 1)$ is $U_{(j)} \sim$ Beta$(j, n - j + 1)$.

Conditional Expectation

Conditioning on an Event We can find $E(Y|A)$, the expected value of Y given that event A occurred. A very important case is when A is the event $X = x$. Note that $E(Y|A)$ is a *number*. For example:

- The expected value of a fair die roll, given that it is prime, is $\frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 5 = \frac{10}{3}$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success. Let A be the event that the first 3 trials are all successes. Then

  $$E(Y|A) = 3 + 7p$$

  since the number of successes among the last 7 trials is Bin$(7, p)$.
- Let $T \sim$ Expo$(1/10)$ be how long you have to wait until the shuttle comes. Given that you have already waited t minutes, the expected additional waiting time is 10 more minutes, by the memoryless property. That is, $E(T|T > t) = t + 10$.

Discrete Y	Continuous Y
$E(Y) = \sum_y yP(Y = y)$	$E(Y) = \int_{-\infty}^{\infty} yf_Y(y)dy$
$E(Y\|A) = \sum_y yP(Y = y\|A)$	$E(Y\|A) = \int_{-\infty}^{\infty} yf(y\|A)dy$

Conditioning on a Random Variable We can also find $E(Y|X)$, the expected value of Y given the random variable X. This is *a function of the random variable* X. It is *not* a number except in certain special cases such as if $X \perp\!\!\!\perp Y$. To find $E(Y|X)$, find $E(Y|X = x)$ and then plug in X for x. For example:

- If $E(Y|X = x) = x^3 + 5x$, then $E(Y|X) = X^3 + 5X$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success and X be the number of successes among the first 3 trials. Then $E(Y|X) = X + 7p$.
- Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Then $E(Y|X = x) = x^2$ since if we know $X = x$ then we know $Y = x^2$. And $E(X|Y = y) = 0$ since if we know $Y = y$ then we know $X = \pm\sqrt{y}$, with equal probabilities (by symmetry). So $E(Y|X) = X^2, E(X|Y) = 0$.

Properties of Conditional Expectation

1. $E(Y|X) = E(Y)$ if $X \perp\!\!\!\perp Y$
2. $E(h(X)W|X) = h(X)E(W|X)$ (**taking out what's known**) In particular, $E(h(X)|X) = h(X)$.
3. $E(E(Y|X)) = E(Y)$ (**Adam's Law**, a.k.a. Law of Total Expectation)

Adam's Law (a.k.a. Law of Total Expectation) can also be written in a way that looks analogous to LOTP. For any events $A_1, A_2, \ldots, A_n$ that partition the sample space,

$$E(Y) = E(Y|A_1)P(A_1) + \cdots + E(Y|A_n)P(A_n)$$

For the special case where the partition is A, A^c, this says

$$E(Y) = E(Y|A)P(A) + E(Y|A^c)P(A^c)$$

Eve's Law (a.k.a. Law of Total Variance)

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$$

MVN, LLN, CLT

Law of Large Numbers (LLN)

Let $X_1, X_2, X_3 \ldots$ be i.i.d. with mean μ. The **sample mean** is

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

The **Law of Large Numbers** states that as $n \to \infty$, $\bar{X}_n \to \mu$ with probability 1. For example, in flips of a coin with probability p of Heads, let X_j be the indicator of the jth flip being Heads. Then LLN says the proportion of Heads converges to p (with probability 1).

Central Limit Theorem (CLT)

Approximation using CLT

We use $\dot{\sim}$ to denote *is approximately distributed*. We can use the **Central Limit Theorem** to approximate the distribution of a random variable $Y = X_1 + X_2 + \cdots + X_n$ that is a sum of n i.i.d. random variables X_i. Let $E(Y) = \mu_Y$ and $\text{Var}(Y) = \sigma_Y^2$. The CLT says

$$Y \dot{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$$

If the X_i are i.i.d. with mean μ_X and variance σ_X^2, then $\mu_Y = n\mu_X$ and $\sigma_Y^2 = n\sigma_X^2$. For the sample mean $\bar{X}_n$, the CLT says

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \dot{\sim} \mathcal{N}(\mu_X, \sigma_X^2/n)$$

Asymptotic Distributions using CLT

We use $\xrightarrow{D}$ to denote *converges in distribution to* as $n \to \infty$. The CLT says that if we standardize the sum $X_1 + \cdots + X_n$ then the distribution of the sum converges to $\mathcal{N}(0, 1)$ as $n \to \infty$:

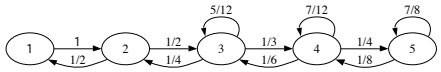
$$\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu_X) \xrightarrow{D} \mathcal{N}(0, 1)$$

In other words, the CDF of the left-hand side goes to the standard Normal CDF, Φ. In terms of the sample mean, the CLT says

$$\frac{\sqrt{n}(\bar{X}_n - \mu_X)}{\sigma_X} \xrightarrow{D} \mathcal{N}(0, 1)$$

Markov Chains

Definition



A Markov chain is a random walk in a **state space**, which we will assume is finite, say $\{1, 2, \ldots, M\}$. We let X_t denote which element of the state space the walk is visiting at time t. The Markov chain is the sequence of random variables tracking where the walk is at all points in time, $X_0, X_1, X_2, \ldots$. By definition, a Markov chain must satisfy the **Markov property**, which says that if you want to predict where the chain will be at a future time, if we know the present state then the entire past history is irrelevant. *Given the present, the past and future are conditionally independent.* In symbols,

$$P(X_{n+1} = j|X_0 = i_0, X_1 = i_1, \ldots, X_n = i) = P(X_{n+1} = j|X_n = i)$$

State Properties

A state is either recurrent or transient.

- If you start at a **recurrent state**, then you will always return back to that state at some point in the future. ♪*You can check-out any time you like, but you can never leave.* ♪
- Otherwise you are at a **transient state**. There is some positive probability that once you leave you will never return. ♪*You don't have to go home, but you can't stay here.* ♪

A state is either periodic or aperiodic.

- If you start at a **periodic state** of period k, then the GCD of the possible numbers of steps it would take to return back is $k > 1$.
- Otherwise you are at an **aperiodic state**. The GCD of the possible numbers of steps it would take to return back is 1.

Transition Matrix

Let the state space be $\{1, 2, \ldots, M\}$. The transition matrix Q is the $M \times M$ matrix where element q_{ij} is the probability that the chain goes from state i to state j in one step:

$$q_{ij} = P(X_{n+1} = j | X_n = i)$$

To find the probability that the chain goes from state i to state j in exactly m steps, take the (i, j) element of Q^m.

$$q_{ij}^{(m)} = P(X_{n+m} = j | X_n = i)$$

If X_0 is distributed according to the row vector PMF $\vec{p}$, i.e., $p_j = P(X_0 = j)$, then the PMF of X_n is $\vec{p}Q^n$.

Chain Properties

A chain is **irreducible** if you can get from anywhere to anywhere. If a chain (on a finite state space) is irreducible, then all of its states are recurrent. A chain is **periodic** if any of its states are periodic, and is **aperiodic** if none of its states are periodic. In an irreducible chain, all states have the same period.

A chain is **reversible** with respect to $\vec{s}$ if $s_i q_{ij} = s_j q_{ji}$ for all i, j. Examples of reversible chains include any chain with $q_{ij} = q_{ji}$, with $\vec{s} = (\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M})$, and random walk on an undirected network.

Stationary Distribution

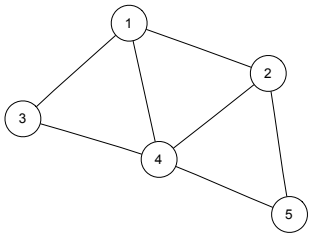
Let us say that the vector $\vec{s} = (s_1, s_2, \ldots, s_M)$ be a PMF (written as a row vector). We will call $\vec{s}$ the **stationary distribution** for the chain if $\vec{s}Q = \vec{s}$. As a consequence, if X_t has the stationary distribution, then all future $X_{t+1}, X_{t+2}, \ldots$ also have the stationary distribution.

For irreducible, aperiodic chains, the stationary distribution exists, is unique, and s_i is the long-run probability of a chain being at state i. The expected number of steps to return to i starting from i is $1/s_i$.

To find the stationary distribution, you can solve the matrix equation $(Q' - I)\vec{s}' = 0$. The stationary distribution is uniform if the columns of Q sum to 1.

Reversibility Condition Implies Stationarity If you have a PMF $\vec{s}$ and a Markov chain with transition matrix Q, then $s_i q_{ij} = s_j q_{ji}$ for all states i, j implies that $\vec{s}$ is stationary.

Random Walk on an Undirected Network



If you have a collection of **nodes**, pairs of which can be connected by undirected **edges**, and a Markov chain is run by going from the current node to a uniformly random node that is connected to it by an edge, then this is a random walk on an undirected network. The stationary distribution of this chain is proportional to the **degree sequence** (this is the sequence of degrees, where the degree of a node is how many edges are attached to it). For example, the stationary distribution of random walk on the network shown above is proportional to $(3, 3, 2, 4, 2)$, so it's $(\frac{3}{14}, \frac{3}{14}, \frac{3}{14}, \frac{4}{14}, \frac{2}{14})$.

Continuous Distributions

Uniform Distribution

Let us say that U is distributed Unif(a, b). We know the following:

Properties of the Uniform For a Uniform distribution, the probability of a draw from any interval within the support is proportional to the length of the interval. See *Universality of Uniform* and *Order Statistics* for other properties.

Example William throws darts really badly, so his darts are uniform over the whole room because they're equally likely to appear anywhere. William's darts have a Uniform distribution on the surface of the room. The Uniform is the only distribution where the probability of hitting in any specific region is proportional to the length/area/volume of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

Normal Distribution

Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

Central Limit Theorem The Normal distribution is ubiquitous because of the Central Limit Theorem, which states that the sample mean of i.i.d. r.v.s will approach a Normal distribution as the sample size grows, regardless of the initial distribution.

Location-Scale Transformation Every time we shift a Normal r.v. (by adding a constant) or rescale a Normal (by multiplying by a constant), we change it to another Normal r.v. For any Normal $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0, 1)$ by the following transformation:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Standard Normal The Standard Normal, $Z \sim \mathcal{N}(0, 1)$, has mean 0 and variance 1. Its CDF is denoted by Φ.

Exponential Distribution

Let us say that X is distributed Expo(λ). We know the following:

Story You're sitting on an open meadow right before the break of dawn, wishing that airplanes in the night sky were shooting stars, because you could really use a wish right now. You know that shooting stars come on average every 15 minutes, but a shooting star is not "due" to come just because you've waited so long. Your waiting time is memoryless; the additional time until the next shooting star comes does not depend on how long you've waited already.

Example The waiting time until the next shooting star is distributed Expo(4) hours. Here $\lambda = 4$ is the **rate parameter**, since shooting stars arrive at a rate of 1 per 1/4 hour on average. The expected time until the next shooting star is $1/\lambda = 1/4$ hour.

Expos as a rescaled Expo(1)

$$Y \sim \text{Expo}(\lambda) \to X = \lambda Y \sim \text{Expo}(1)$$

Memorylessness The Exponential Distribution is the only continuous memoryless distribution. The memoryless property says that for $X \sim \text{Expo}(\lambda)$ and any positive numbers s and t,

$$P(X > s + t | X > s) = P(X > t)$$

Equivalently,

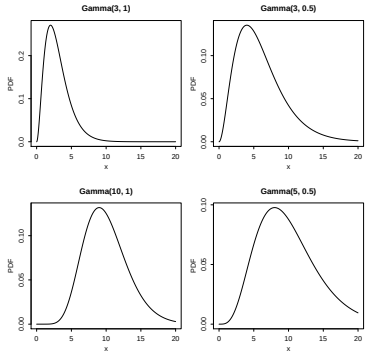
$$X - a | (X > a) \sim \text{Expo}(\lambda)$$

For example, a product with an Expo(λ) lifetime is always "as good as new" (it doesn't experience wear and tear). Given that the product has survived a years, the additional time that it will last is still Expo(λ).

Min of Expos If we have independent $X_i \sim \text{Expo}(\lambda_i)$, then $\min(X_1, \ldots, X_k) \sim \text{Expo}(\lambda_1 + \lambda_2 + \cdots + \lambda_k)$.

Max of Expos If we have i.i.d. $X_i \sim \text{Expo}(\lambda)$, then $\max(X_1, \ldots, X_k)$ has the same distribution as $Y_1 + Y_2 + \cdots + Y_k$, where $Y_j \sim \text{Expo}(j\lambda)$ and the Y_j are independent.

Gamma Distribution

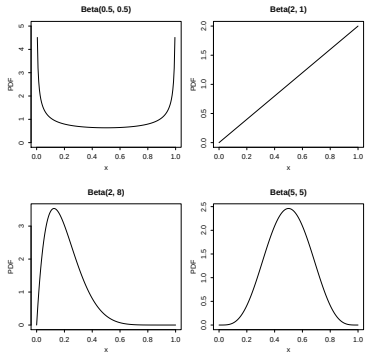


Let us say that X is distributed Gamma(a, λ). We know the following:

Story You sit waiting for shooting stars, where the waiting time for a star is distributed Expo(λ). You want to see n shooting stars before you go home. The total waiting time for the nth shooting star is Gamma(n, λ).

Example You are at a bank, and there are 3 people ahead of you. The serving time for each person is Exponential with mean 2 minutes. Only one person at a time can be served. The distribution of your waiting time until it's your turn to be served is Gamma$(3, \frac{1}{2})$.

Beta Distribution



Conjugate Prior of the Binomial In the Bayesian approach to statistics, parameters are viewed as random variables, to reflect our uncertainty. The *prior* for a parameter is its distribution before observing data. The *posterior* is the distribution for the parameter after observing data. Beta is the *conjugate* prior of the Binomial because if you have a Beta-distributed prior on p in a Binomial, then the posterior distribution on p given the Binomial data is also Beta-distributed. Consider the following two-level model:

$$X | p \sim \text{Bin}(n, p)$$
$$p \sim \text{Beta}(a, b)$$

Then after observing $X = x$, we get the posterior distribution

$$p | (X = x) \sim \text{Beta}(a + x, b + n - x)$$

Order statistics of the Uniform See *Order Statistics*.

Beta-Gamma relationship If $X \sim \text{Gamma}(a, \lambda)$, $Y \sim \text{Gamma}(b, \lambda)$, with $X \perp\!\!\!\perp Y$ then

- $\frac{X}{X+Y} \sim \text{Beta}(a, b)$
- $X + Y \perp\!\!\!\perp \frac{X}{X+Y}$

This is known as the **bank–post office result**.

χ^2 (Chi-Square) Distribution

Let us say that X is distributed χ^2_n. We know the following:

Story A Chi-Square(n) is the sum of the squares of n independent standard Normal r.v.s.

Properties and Representations

$$X \text{ is distributed as } Z_1^2 + Z_2^2 + \cdots + Z_n^2 \text{ for i.i.d. } Z_i \sim \mathcal{N}(0,1)$$

$$X \sim \text{Gamma}(n/2, 1/2)$$

Discrete Distributions

Distributions for four sampling schemes

	Replace	**No Replace**
Fixed # trials (n)	Binomial (Bern if $n = 1$)	HGeom
Draw until r success	NBin (Geom if $r = 1$)	NHGeom

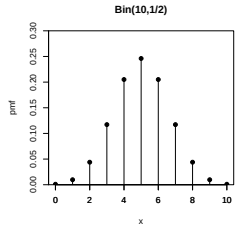
Bernoulli Distribution

The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial ($n = 1$). Let us say that X is distributed Bern(p). We know the following:

Story A trial is performed with probability p of "success", and X is the indicator of success: 1 means success, 0 means failure.

Example Let X be the indicator of Heads for a fair coin toss. Then $X \sim \text{Bern}(\frac{1}{2})$. Also, $1 - X \sim \text{Bern}(\frac{1}{2})$ is the indicator of Tails.

Binomial Distribution



Let us say that X is distributed Bin(n, p). We know the following:

Story X is the number of "successes" that we will achieve in n independent trials, where each trial is either a success or a failure, each with the same probability p of success. We can also write X as a sum of multiple independent Bern(p) random variables. Let $X \sim \text{Bin}(n, p)$ and $X_j \sim \text{Bern}(p)$, where all of the Bernoullis are independent. Then

$$X = X_1 + X_2 + X_3 + \cdots + X_n$$

Example If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed Bin(10, $\frac{3}{4}$).

Properties Let $X \sim \text{Bin}(n, p), Y \sim \text{Bin}(m, p)$ with $X \perp\!\!\!\perp Y$.

- **Redefine success** $n - X \sim \text{Bin}(n, 1 - p)$
- **Sum** $X + Y \sim \text{Bin}(n + m, p)$
- **Conditional** $X|(X + Y = r) \sim \text{HGeom}(n, m, r)$
- **Binomial-Poisson Relationship** Bin(n, p) is approximately Pois(λ) if p is small.
- **Binomial-Normal Relationship** Bin(n, p) is approximately $\mathcal{N}(np, np(1-p))$ if n is large and p is not near 0 or 1.

Geometric Distribution

Let us say that X is distributed Geom(p). We know the following:

Story X is the number of "failures" that we will achieve before we achieve our first success. Our successes have probability p.

Example If each pokeball we throw has probability $\frac{1}{10}$ to catch Mew, the number of failed pokeballs will be distributed Geom($\frac{1}{10}$).

First Success Distribution

Equivalent to the Geometric distribution, except that it includes the first success in the count. This is 1 more than the number of failures. If $X \sim \text{FS}(p)$ then $E(X) = 1/p$.

Negative Binomial Distribution

Let us say that X is distributed NBin(r, p). We know the following:

Story X is the number of "failures" that we will have before we achieve our rth success. Our successes have probability p.

Example Thundershock has 60% accuracy and can faint a wild Raticate in 3 hits. The number of misses before Pikachu faints Raticate with Thundershock is distributed NBin(3, 0.6).

Hypergeometric Distribution

Let us say that X is distributed HGeom(w, b, n). We know the following:

Story In a population of w desired objects and b undesired objects, X is the number of "successes" we will have in a draw of n objects, without replacement. The draw of n objects is assumed to be a **simple random sample** (all sets of n objects are equally likely).

Examples Here are some HGeom examples.

- Let's say that we have only b Weedles (failure) and w Pikachus (success) in Viridian Forest. We encounter n Pokemon in the forest, and X is the number of Pikachus in our encounters.
- The number of Aces in a 5 card hand.
- You have w white balls and b black balls, and you draw n balls. You will draw X white balls.
- You have w white balls and b black balls, and you draw n balls without replacement. The number of white balls in your sample is HGeom(w, b, n); the number of black balls is HGeom(b, w, n).
- **Capture-recapture** A forest has N elk, you capture n of them, tag them, and release them. Then you recapture a new sample of size m. How many tagged elk are now in the new sample? HGeom($n, N - n, m$)

Poisson Distribution

Let us say that X is distributed Pois(λ). We know the following:

Story There are rare events (low probability events) that occur many different ways (high possibilities of occurences) at an average rate of λ occurrences per unit space or time. The number of events that occur in that unit of space or time is X.

Example A certain busy intersection has an average of 2 accidents per month. Since an accident is a low probability event that can happen many different ways, it is reasonable to model the number of accidents in a month at that intersection as Pois(2). Then the number of accidents that happen in two months at that intersection is distributed Pois(4).

Properties Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$.

1. **Sum** $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. **Conditional** $X|(X + Y = n) \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$
3. **Chicken-egg** If there are $Z \sim \text{Pois}(\lambda)$ items and we randomly and independently "accept" each item with probability p, then the number of accepted items $Z_1 \sim \text{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \text{Pois}(\lambda(1-p))$, and $Z_1 \perp\!\!\!\perp Z_2$.

Multivariate Distributions

Multinomial Distribution

Let us say that the vector $\vec{X} = (X_1, X_2, X_3, \ldots, X_k) \sim \text{Mult}_k(n, \vec{p})$ where $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Story We have n items, which can fall into any one of the k buckets independently with the probabilities $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Example Let us assume that every year, 100 students in the Harry Potter Universe are randomly and independently sorted into one of four houses with equal probability. The number of people in each of the houses is distributed $\text{Mult}_4(100, \vec{p})$, where $\vec{p} = (0.25, 0.25, 0.25, 0.25)$. Note that $X_1 + X_2 + \cdots + X_4 = 100$, and they are dependent.

Joint PMF For $n = n_1 + n_2 + \cdots + n_k$,

$$P(\vec{X} = \vec{n}) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Marginal PMF, Lumping, and Conditionals Marginally, $X_i \sim \text{Bin}(n, p_i)$ since we can define "success" to mean category i. If you lump together multiple categories in a Multinomial, then it is still Multinomial. For example, $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$ for $i \neq j$ since we can define "success" to mean being in category i or j. Similarly, if $k = 6$ and we lump categories 1-2 and lump categories 3-5, then

$$(X_1 + X_2, X_3 + X_4 + X_5, X_6) \sim \text{Mult}_3(n, (p_1 + p_2, p_3 + p_4 + p_5, p_6))$$

Conditioning on some X_j also still gives a Multinomial:

$$X_1, \ldots, X_{k-1} | X_k = n_k \sim \text{Mult}_{k-1}\left(n - n_k, \left(\frac{p_1}{1-p_k}, \ldots, \frac{p_{k-1}}{1-p_k}\right)\right)$$

Variances and Covariances We have $X_i \sim \text{Bin}(n, p_i)$ marginally, so $\text{Var}(X_i) = np_i(1-p_i)$. Also, $\text{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$.

Multivariate Uniform Distribution

See the univariate Uniform for stories and examples. For the 2D Uniform on some region, probability is proportional to area. Every point in the support has equal density, of value $\frac{1}{\text{area of region}}$. For the 3D Uniform, probability is proportional to volume.

Multivariate Normal (MVN) Distribution

A vector $\vec{X} = (X_1, X_2, \ldots, X_k)$ is Multivariate Normal if every linear combination is Normally distributed, i.e., $t_1X_1 + t_2X_2 + \cdots + t_kX_k$ is Normal for any constants $t_1, t_2, \ldots, t_k$. The parameters of the Multivariate Normal are the **mean vector** $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ and the **covariance matrix** where the (i, j) entry is $\text{Cov}(X_i, X_j)$.

Properties The Multivariate Normal has the following properties.

- Any subvector is also MVN.
- If any two elements within an MVN are uncorrelated, then they are independent.
- The joint PDF of a Bivariate Normal (X, Y) with $\mathcal{N}(0, 1)$ marginal distributions and correlation $\rho \in (-1, 1)$ is

$$f_{X,Y}(x, y) = \frac{1}{2\pi\tau} \exp\left(-\frac{1}{2\tau^2}(x^2 + y^2 - 2\rho xy)\right),$$

with $\tau = \sqrt{1 - \rho^2}$.

Distribution Properties

Important CDFs

Standard Normal Φ

Exponential(λ) $F(x) = 1 - e^{-\lambda x}$, for $x \in (0, \infty)$

Uniform(0,1) $F(x) = x$, for $x \in (0, 1)$

Convolutions of Random Variables

A convolution of n random variables is simply their sum. For the following results, let X and Y be *independent*.

1. $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \longrightarrow X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X \sim \text{Bin}(n_1, p)$, $Y \sim \text{Bin}(n_2, p) \longrightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$. $\text{Bin}(n, p)$ can be thought of as a sum of i.i.d. $\text{Bern}(p)$ r.v.s.
3. $X \sim \text{Gamma}(a_1, \lambda)$, $Y \sim \text{Gamma}(a_2, \lambda)$ $\longrightarrow X + Y \sim \text{Gamma}(a_1 + a_2, \lambda)$. $\text{Gamma}(n, \lambda)$ with n an integer can be thought of as a sum of i.i.d. $\text{Expo}(\lambda)$ r.v.s.
4. $X \sim \text{NBin}(r_1, p)$, $Y \sim \text{NBin}(r_2, p)$ $\longrightarrow X + Y \sim \text{NBin}(r_1 + r_2, p)$. $\text{NBin}(r, p)$ can be thought of as a sum of i.i.d. $\text{Geom}(p)$ r.v.s.
5. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $\longrightarrow X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Special Cases of Distributions

1. $\text{Bin}(1, p) \sim \text{Bern}(p)$
2. $\text{Beta}(1, 1) \sim \text{Unif}(0, 1)$
3. $\text{Gamma}(1, \lambda) \sim \text{Expo}(\lambda)$
4. $\chi^2_n \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$
5. $\text{NBin}(1, p) \sim \text{Geom}(p)$

Inequalities

1. **Cauchy-Schwarz** $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$
2. **Markov** $P(X \geq a) \leq \frac{E|X|}{a}$ for $a > 0$
3. **Chebyshev** $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$ for $E(X) = \mu, \text{Var}(X) = \sigma^2$
4. **Jensen** $E(g(X)) \geq g(E(X))$ for g convex; reverse if g is concave

Formulas

Geometric Series

$$1 + r + r^2 + \cdots + r^{n-1} = \sum_{k=0}^{n-1} r^k = \frac{1 - r^n}{1 - r}$$

$$1 + r + r^2 + \cdots = \frac{1}{1 - r} \text{ if } |r| < 1$$

Exponential Function (e^x)

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

Gamma and Beta Integrals

You can sometimes solve complicated-looking integrals by pattern-matching to a gamma or beta integral:

$$\int_0^\infty x^{t-1} e^{-x}\, dx = \Gamma(t) \qquad \int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Also, $\Gamma(a + 1) = a\Gamma(a)$, and $\Gamma(n) = (n - 1)!$ if n is a positive integer.

Euler's Approximation for Harmonic Sums

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \log n + 0.577\ldots$$

Stirling's Approximation for Factorials

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Miscellaneous Definitions

Medians and Quantiles Let X have CDF F. Then X has median m if $F(m) \geq 0.5$ and $P(X \geq m) \geq 0.5$. For X continuous, m satisfies $F(m) = 1/2$. In general, the ath quantile of X is $\min\{x : F(x) \geq a\}$; the median is the case $a = 1/2$.

log Statisticians generally use log to refer to natural log (i.e., base e).

i.i.d r.v.s Independent, identically-distributed random variables.

Example Problems

Contributions from Sebastian Chiu

Calculating Probability

A textbook has n typos, which are randomly scattered amongst its n pages, independently. You pick a random page. What is the probability that it has no typos? **Answer:** There is a $\left(1 - \frac{1}{n}\right)$ probability that any specific typo isn't on your page, and thus a $\boxed{\left(1 - \frac{1}{n}\right)^n}$ probability that there are no typos on your page. For n large, this is approximately $e^{-1} = 1/e$.

Linearity and Indicators (1)

In a group of n people, what is the expected number of distinct birthdays (month and day)? What is the expected number of birthday matches? **Answer:** Let X be the number of distinct birthdays and I_j be the indicator for the jth day being represented.

$$E(I_j) = 1 - P(\text{no one born on day } j) = 1 - (364/365)^n$$

By linearity, $\boxed{E(X) = 365\left(1 - (364/365)^n\right)}$. Now let Y be the number of birthday matches and J_i be the indicator that the ith pair of people have the same birthday. The probability that any two specific people share a birthday is $1/365$, so $\boxed{E(Y) = \binom{n}{2}/365}$.

Linearity and Indicators (2)

This problem is commonly known as the ***hat-matching problem****.* There are n people at a party, each with hat. At the end of the party, they each leave with a random hat. What is the expected number of people who leave with the right hat? **Answer:** Each hat has a $1/n$ chance of going to the right person. By linearity, the average number of hats that go to their owners is $\boxed{n(1/n) = 1}$.

Linearity and First Success

This problem is commonly known as the ***coupon collector problem****.* There are n coupon types. At each draw, you get a uniformly random coupon type. What is the expected number of coupons needed until you have a complete set? **Answer:** Let N be the number of coupons needed; we want $E(N)$. Let $N = N_1 + \cdots + N_n$, where N_1 is the draws to get our first new coupon, N_2 is the *additional* draws needed to draw our second new coupon and so on. By the story of the First Success, $N_2 \sim \text{FS}((n-1)/n)$ (after collecting first coupon type, there's $(n-1)/n$ chance you'll get something new). Similarly, $N_3 \sim \text{FS}((n-2)/n)$, and $N_j \sim \text{FS}((n-j+1)/n)$. By linearity,

$$E(N) = E(N_1) + \cdots + E(N_n) = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = \boxed{n\sum_{j=1}^{n} \frac{1}{j}}$$

This is approximately $n(\log(n) + 0.577)$ by Euler's approximation.

Orderings of i.i.d. random variables

I call 2 UberX's and 3 Lyfts at the same time. If the time it takes for the rides to reach me are i.i.d., what is the probability that all the Lyfts will arrive first? **Answer:** Since the arrival times of the five cars are i.i.d., all 5! orderings of the arrivals are equally likely. There are 3!2! orderings that involve the Lyfts arriving first, so the probability that the Lyfts arrive first is $\boxed{\frac{3!2!}{5!} = 1/10}$. Alternatively, there are $\binom{5}{3}$ ways to choose 3 of the 5 slots for the Lyfts to occupy, where each of the choices are equally likely. One of these choices has all 3 of the Lyfts arriving first, so the probability is $\boxed{1/\binom{5}{3} = 1/10}$.

Expectation of Negative Hypergeometric

What is the expected number of cards that you draw before you pick your first Ace in a shuffled deck (not counting the Ace)? **Answer:** Consider a non-Ace. Denote this to be card j. Let I_j be the indicator that card j will be drawn before the first Ace. Note that $I_j = 1$ says that j is before all 4 of the Aces in the deck. The probability that this occurs is $1/5$ by symmetry. Let X be the number of cards drawn before the first Ace. Then $X = I_1 + I_2 + \ldots + I_{48}$, where each indicator corresponds to one of the 48 non-Aces. Thus,

$$E(X) = E(I_1) + E(I_2) + \ldots + E(I_{48}) = 48/5 = \boxed{9.6}.$$

Minimum and Maximum of RVs

What is the CDF of the maximum of n independent Unif(0,1) random variables? **Answer:** Note that for r.v.s $X_1, X_2, \ldots, X_n$,

$$P(\min(X_1, X_2, \ldots, X_n) \geq a) = P(X_1 \geq a, X_2 \geq a, \ldots, X_n \geq a)$$

Similarly,

$$P(\max(X_1, X_2, \ldots, X_n) \leq a) = P(X_1 \leq a, X_2 \leq a, \ldots, X_n \leq a)$$

We will use this principle to find the CDF of $U_{(n)}$, where $U_{(n)} = \max(U_1, U_2, \ldots, U_n)$ and $U_i \sim \text{Unif}(0, 1)$ are i.i.d.

$$\begin{aligned} P(\max(U_1, U_2, \ldots, U_n) \leq a) &= P(U_1 \leq a, U_2 \leq a, \ldots, U_n \leq a) \\ &= P(U_1 \leq a)P(U_2 \leq a)\ldots P(U_n \leq a) \\ &= \boxed{a^n} \end{aligned}$$

for $0 < a < 1$ (and the CDF is 0 for $a \leq 0$ and 1 for $a \geq 1$).

Pattern-matching with e^x Taylor series

For $X \sim \text{Pois}(\lambda)$, find $E\left(\frac{1}{X+1}\right)$. **Answer:** By LOTUS,

$$E\left(\frac{1}{X+1}\right) = \sum_{k=0}^{\infty} \frac{1}{k+1} \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-\lambda}}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!} = \boxed{\frac{e^{-\lambda}}{\lambda}(e^\lambda - 1)}$$

Adam's Law and Eve's Law

William really likes speedsolving Rubik's Cubes. But he's pretty bad at it, so sometimes he fails. On any given day, William will attempt $N \sim \text{Geom}(s)$ Rubik's Cubes. Suppose each time, he has probability p of solving the cube, independently. Let T be the number of Rubik's Cubes he solves during a day. Find the mean and variance of T.
Answer: Note that $T|N \sim \text{Bin}(N,p)$. So by Adam's Law,

$$E(T) = E(E(T|N)) = E(Np) = \boxed{\frac{p(1-s)}{s}}$$

Similarly, by Eve's Law, we have that

$$\text{Var}(T) = E(\text{Var}(T|N)) + \text{Var}(E(T|N)) = E(Np(1-p)) + \text{Var}(Np)$$
$$= \frac{p(1-p)(1-s)}{s} + \frac{p^2(1-s)}{s^2} = \boxed{\frac{p(1-s)(p+s(1-p))}{s^2}}$$

MGF – Finding Moments

Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$ using the MGF of X. **Answer:** The MGF of an $\text{Expo}(\lambda)$ is $M(t) = \frac{\lambda}{\lambda - t}$. To get the third moment, we can take the third derivative of the MGF and evaluate at $t = 0$:

$$\boxed{E(X^3) = \frac{6}{\lambda^3}}$$

But a much nicer way to use the MGF here is via pattern recognition: note that $M(t)$ looks like it came from a geometric series:

$$\frac{1}{1 - \frac{t}{\lambda}} = \sum_{n=0}^{\infty} \left(\frac{t}{\lambda}\right)^n = \sum_{n=0}^{\infty} \frac{n!}{\lambda^n} \frac{t^n}{n!}$$

The coefficient of $\frac{t^n}{n!}$ here is the nth moment of X, so we have $E(X^n) = \frac{n!}{\lambda^n}$ for all nonnegative integers n.

Markov chains (1)

Suppose X_n is a two-state Markov chain with transition matrix

$$Q = \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & \quad 1 \end{matrix} \\ \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \end{matrix}$$

Find the stationary distribution $\vec{s} = (s_0, s_1)$ of X_n by solving $\vec{s}Q = \vec{s}$, and show that the chain is reversible with respect to $\vec{s}$. **Answer:** The equation $\vec{s}Q = \vec{s}$ says that

$$s_0 = s_0(1-\alpha) + s_1\beta \text{ and } s_1 = s_0(\alpha) + s_0(1-\beta)$$

By solving this system of linear equations, we have

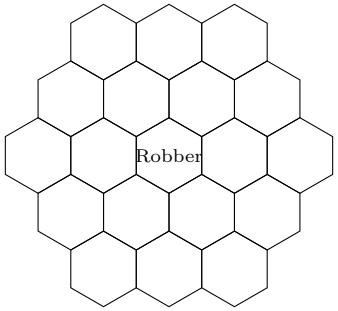
$$\boxed{\vec{s} = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)}$$

To show that the chain is reversible with respect to $\vec{s}$, we must show $s_i q_{ij} = s_j q_{ji}$ for all i, j. This is done if we can show $s_0 q_{01} = s_1 q_{10}$. And indeed,

$$s_0 q_{01} = \frac{\alpha\beta}{\alpha+\beta} = s_1 q_{10}$$

Markov chains (2)

William and Sebastian play a modified game of Settlers of Catan, where every turn they randomly move the robber (which starts on the center tile) to one of the adjacent hexagons.



(a) Is this Markov chain irreducible? Is it aperiodic? **Answer:** $\boxed{\text{Yes to both.}}$ The Markov chain is irreducible because it can get from anywhere to anywhere else. The Markov chain is aperiodic because the robber can return back to a square in $2, 3, 4, 5, \ldots$ moves, and the GCD of those numbers is 1.

(b) What is the stationary distribution of this Markov chain? **Answer:** Since this is a random walk on an undirected graph, the stationary distribution is proportional to the degree sequence. The degree for the corner pieces is 3, the degree for the edge pieces is 4, and the degree for the center pieces is 6. To normalize this degree sequence, we divide by its sum. The sum of the degrees is $6(3) + 6(4) + 7(6) = 84$. Thus the stationary probability of being on a corner is $3/84 = 1/28$, on an edge is $4/84 = 1/21$, and in the center is $6/84 = 1/14$.

(c) What fraction of the time will the robber be in the center tile in this game, in the long run? **Answer:** By the above, $\boxed{1/14}$.

(d) What is the expected amount of moves it will take for the robber to return to the center tile? **Answer:** Since this chain is irreducible and aperiodic, to get the expected time to return we can just invert the stationary probability. Thus on average it will take $\boxed{14}$ turns for the robber to return to the center tile.

Problem-Solving Strategies

Contributions from Jessy Hwang, Yuan Jiang, Yuqi Hou

1. **Getting started.** Start by *defining relevant events and random variables*. ("Let A be the event that I pick the fair coin"; "Let X be the number of successes.") Clear notion is important for clear thinking! Then decide what it is that you're supposed to be finding, in terms of your notation ("I want to find $P(X = 3|A)$"). Think about what type of object your answer should be (a number? A random variable? A PMF? A PDF?) and what it should be in terms of.
   Try simple and extreme cases. To make an abstract experiment more concrete, try *drawing a picture* or making up numbers that could have happened. Pattern recognition: does the structure of the problem resemble something we've seen before?
2. **Calculating probability of an event.** Use counting principles if the naive definition of probability applies. Is the probability of the complement easier to find? Look for symmetries. Look for something to condition on, then apply Bayes' Rule or the Law of Total Probability.
3. **Finding the distribution of a random variable.** First make sure you need the full distribution not just the mean (see next item). Check the *support* of the random variable: what values can it take on? Use this to rule out distributions that don't fit. Is there a *story* for one of the named distributions that fits the problem at hand? Can you write the random variable as a function of an r.v. with a known distribution, say $Y = g(X)$?
4. **Calculating expectation.** If it has a named distribution, check out the table of distributions. If it's a function of an r.v. with a named distribution, try LOTUS. If it's a count of something, try breaking it up into indicator r.v.s. If you can condition on something natural, consider using Adam's law.
5. **Calculating variance.** Consider independence, named distributions, and LOTUS. If it's a count of something, break it up into a sum of indicator r.v.s. If it's a sum, use properties of covariance. If you can condition on something natural, consider using Eve's Law.
6. **Calculating $E(X^2)$.** Do you already know $E(X)$ or $\text{Var}(X)$? Recall that $\text{Var}(X) = E(X^2) - (E(X))^2$. Otherwise try LOTUS.
7. **Calculating covariance.** Use the properties of covariance. If you're trying to find the covariance between two components of a Multinomial distribution, X_i, X_j, then the covariance is $-np_ip_j$ for $i \neq j$.
8. **Symmetry.** If $X_1, \ldots, X_n$ are i.i.d., consider using symmetry.
9. **Calculating probabilities of orderings.** Remember that all $n!$ ordering of i.i.d. continuous random variables $X_1, \ldots, X_n$ are equally likely.
10. **Determining independence.** There are several equivalent definitions. Think about simple and extreme cases to see if you can find a counterexample.
11. **Do a painful integral.** If your integral looks painful, see if you can write your integral in terms of a known PDF (like Gamma or Beta), and use the fact that PDFs integrate to 1?
12. **Before moving on.** Check some simple and extreme cases, check whether the answer seems plausible, check for biohazards.

Biohazards

Contributions from Jessy Hwang

1. **Don't misuse the naive definition of probability.** When answering "What is the probability that in a group of 3 people, no two have the same birth month?", it is *not* correct to treat the people as indistinguishable balls being placed into 12 boxes, since that assumes the list of birth months {January, January, January} is just as likely as the list {January, April, June}, even though the latter is six times more likely.
2. **Don't confuse unconditional, conditional, and joint probabilities.** In applying $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, it is *not* correct to say "$P(B) = 1$ because we know B happened"; $P(B)$ is the *prior* probability of B. Don't confuse $P(A|B)$ with $P(A, B)$.
3. **Don't assume independence without justification.** In the matching problem, the probability that card 1 is a match and card 2 is a match is not $1/n^2$. Binomial and Hypergeometric are often confused; the trials are independent in the Binomial story and dependent in the Hypergeometric story.
4. **Don't forget to do sanity checks.** Probabilities must be between 0 and 1. Variances must be ≥ 0. Supports must make sense. PMFs must sum to 1. PDFs must integrate to 1.
5. **Don't confuse random variables, numbers, and events.** Let X be an r.v. Then $g(X)$ is an r.v. for any function g. In particular, X^2, $|X|$, $F(X)$, and $I_{X>3}$ are r.v.s. $P(X^2 < X|X \geq 0)$, $E(X)$, $\text{Var}(X)$, and $g(E(X))$ are numbers. $X = 2$ and $F(X) \geq -1$ are events. It does not make sense to write $\int_{-\infty}^{\infty} F(X)dx$, because $F(X)$ is a random variable. It does not make sense to write $P(X)$, because X is not an event.

6. **Don't confuse a random variable with its distribution.** To get the PDF of X^2, you can't just square the PDF of X. The right way is to use transformations. To get the PDF of $X+Y$, you can't just add the PDF of X and the PDF of Y. The right way is to compute the convolution.

7. **Don't pull non-linear functions out of expectations.** $E(g(X))$ does not equal $g(E(X))$ in general. The St. Petersburg paradox is an extreme example. See also Jensen's inequality. The right way to find $E(g(X))$ is with LOTUS.

Recommended Resources

- Introduction to Probability Book (`http://bit.ly/introprobability`)
- Stat 110 Online (`http://stat110.net`)
- Stat 110 Quora Blog (`https://stat110.quora.com/`)
- Quora Probability FAQ (`http://bit.ly/probabilityfaq`)
- R Studio (`https://www.rstudio.com`)
- LaTeX File (`github.com/wzchen/probability_cheatsheet`)

Please share this cheatsheet with friends!
`http://wzchen.com/probability-cheatsheet`

Distributions in R

Command	What it does
`help(distributions)`	shows documentation on distributions
`dbinom(k,n,p)`	PMF $P(X=k)$ for $X \sim \text{Bin}(n,p)$
`pbinom(x,n,p)`	CDF $P(X \leq x)$ for $X \sim \text{Bin}(n,p)$
`qbinom(a,n,p)`	ath quantile for $X \sim \text{Bin}(n,p)$
`rbinom(r,n,p)`	vector of r i.i.d. $\text{Bin}(n,p)$ r.v.s
`dgeom(k,p)`	PMF $P(X=k)$ for $X \sim \text{Geom}(p)$
`dhyper(k,w,b,n)`	PMF $P(X=k)$ for $X \sim \text{HGeom}(w,b,n)$
`dnbinom(k,r,p)`	PMF $P(X=k)$ for $X \sim \text{NBin}(r,p)$
`dpois(k,r)`	PMF $P(X=k)$ for $X \sim \text{Pois}(r)$
`dbeta(x,a,b)`	PDF $f(x)$ for $X \sim \text{Beta}(a,b)$
`dchisq(x,n)`	PDF $f(x)$ for $X \sim \chi^2_n$
`dexp(x,b)`	PDF $f(x)$ for $X \sim \text{Expo}(b)$
`dgamma(x,a,r)`	PDF $f(x)$ for $X \sim \text{Gamma}(a,r)$
`dlnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{LN}(m,s^2)$
`dnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{N}(m,s^2)$
`dt(x,n)`	PDF $f(x)$ for $X \sim t_n$
`dunif(x,a,b)`	PDF $f(x)$ for $X \sim \text{Unif}(a,b)$

The table above gives R commands for working with various named distributions. Commands analogous to `pbinom`, `qbinom`, and `rbinom` work for the other distributions in the table. For example, `pnorm`, `qnorm`, and `rnorm` can be used to get the CDF, quantiles, and random generation for the Normal. For the Multinomial, `dmultinom` can be used for calculating the joint PMF and `rmultinom` can be used for generating random vectors. For the Multivariate Normal, after installing and loading the `mvtnorm` package `dmvnorm` can be used for calculating the joint PDF and `rmvnorm` can be used for generating random vectors.

Table of Distributions

Distribution	PMF/PDF and Support	Expected Value	Variance	MGF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q=1-p$	p	pq	$q+pe^t$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k q^{n-k}$ $k\in\{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^k p$ $k\in\{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t}$, $qe^t<1$
Negative Binomial NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^r q^n$ $n\in\{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r$, $qe^t<1$
Hypergeometric HGeom(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k\in\{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\left(\frac{w+b-n}{w+b-1}\right)n\frac{\mu}{n}(1-\frac{\mu}{n})$	messy
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k\in\{0,\ 1,\ 2,\ \dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x\in(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x\ -\ \mu)^2/(2\sigma^2)}$ $x\in(-\infty,\infty)$	μ	σ^2	$e^{t\mu+\frac{\sigma^2 t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$, $t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x\in(0,\infty)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^a$, $t<\lambda$
Beta Beta(a,b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x\in(0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	messy
Log-Normal $\mathcal{LN}(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}$ $x\in(0,\infty)$	$\theta=e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$	doesn't exist
Chi-Square χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x\in(0,\infty)$	n	$2n$	$(1-2t)^{-n/2}$, $t<1/2$
Student-t t_n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$ $x\in(-\infty,\infty)$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$	doesn't exist

Statistics Cheat Sheet

Ch 1: Overview & Descriptive Stats

Populations, Samples and Processes

`Population:` well-defined collection of objects
`Sample:` a subset of the population
`Descriptive Stats:` summarize & describe features of data
`Inferential Stats:` generalizing from sample to population
`Probability:` bridge btwn descriptive & inferential techniques.
In probability, properties of the population are assumed known & questions regarding a sample taken from the population are posed and answered.
`Discrete and Continuous Variables:` A numerical variable is *discrete* if its set of possible values is at most countable.
A numerical value is *continuous* if its set of possible values is an uncountable set.
Probability: pop → sample
Stats: sample → pop

Measures of Location

For observations $x_1, x_2, \cdots, x_n$

`Sample Mean`	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
`Sample Median`	$\tilde{x} = (\frac{n+1}{2})^{\text{nth}}$ observation
`Trimmed Mean`	btwn $\tilde{x}$ and $\bar{x}$, compute by removing smallest and largest observations

Measures of Variability

`Range`	=lgst-smllst observation
`Sample Variance,` σ^2	$= \frac{\sum(x_i - \bar{x}))^2}{n-1} = \frac{S_{xx}}{n-1}$
S_{xx}	$= \sum x_i^2 - \frac{(\sum x_i)^2}{n}$
`Sample Standard Deviation,` σ	$= \sqrt{\sigma^2}$

Box Plots

Order the n observations from small to large. Separate the smallest half from the largest (If n is odd then $\tilde{x}$ is in both halves). The lower fourth is the median of the smallest half (upper fourth..largest..). A measure of the spread that is resistant to outliers is the *fourth spread* f_s given by $f_s =$ upper fourth- lower fourth. Box from lower to upper fourth with line at median. Whiskers from smallest to largest x_i.

Ch 2: Probability

Sample Space and Events

`Experiment`	activity with uncertain outcome
`Sample Space`($\mathcal{S}$)	the set of all possible outcomes
`Event`	any collection of outcomes in $\mathcal{S}$

Axioms, Interpretations and Properties of Probability

Given an experiment and a sample space $\mathcal{S}$, the objective probability is to assign to each event A a number $P(A)$, called the probability of event A, which will give a precise measure of the chance that A will occur. Behaves very much like norm.

Axioms & Properties of Probability:

1. $\forall A \in \mathcal{S}, 0 \le P(A) \le 1$
2. $P(\mathcal{S}) = 1$
3. If $A_1, A_2, \ldots$ is an infinite collection of disjoint events, $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$
4. $P(\emptyset) = 0$
5. $\forall A, P(A) + P(A') = 1$ from which $P(A) = 1 - P(A')$
6. For any two events $A, B \in \mathcal{S}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
7. For any three events $A, B, C \in \mathcal{S}, P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Equally Likely Outcomes : $P(A) = \frac{N(A)}{N}$

Counting Techniques

`Product Rule for Ordered k-Tuples:` If the first element can be selected in n_1 ways, the second in n_2 ways and so on, then there are $n_1 n_2 \cdots n_k$ possible k-tuples.
`Permutations:` An ordered subset. The number of permutations of size k that can be formed from a set of n elements is $P_{k,n}$

$$P_{k,n} = (n)(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

`Combinations:` An unordered subset.

$$\binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

Conditional Probability

$P(A|B)$ is the conditional probability of A given that the event B has occurred. B is the conditioning event.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplication Rule: $P(A \cap B) = P(A|B) \cdot P(B)$

Baye's Theorem

Let $A_1, A_2, \ldots, A_k$ be disjoint and exhaustive events (that partition the sample space). Then for any other event B
$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)$
$\quad = \sum_{i=1}^{k} P(B|A_i)P(A_i)$

Independence

Two events A and B are `independent` if $P(A|B) = P(A)$ and are `dependent` otherwise.
A and B are `independent` iff $P(A \cap B) = P(A) \cdot P(B)$ and this can be generalized to the case of n mutually independent events.

Random Variables

`Random Variable`: any function $X : \Omega \to \mathbb{R}$

`Prob Dist.`:	describes how the probability of Ω is distributed along the range of X
`Discrete rv`:	rv whose domain is at most countable
`Continuous rv`:	rv whose domain is uncountable and where $\forall c \in \mathbb{R}, P(X = c) = 0$
`Bernoulli rv`:	discrete rv whose range is $\{0, 1\}$

The *probability distribution of X* says how the total probability of 1 is distributed among the various possible X values.

1. Distributions

Discrete RVs

Probabilities assigned to various outcomes in $\mathcal{S}$ in turn determine probabilities associated with the values of any particular rv X.

`Probability Mass Fxn/Probability Distribution,(pmf)`:

$$p(x) = P(X = x) = P(\forall w \in \mathcal{W} : X(w) = x)$$

Gives the probability of observing $w \in \mathcal{W} : X(w) = x$
The conditions $p(x) \ge 0$ and $\sum_{\text{all possible x}} p(x) = 1$ are required for any pmf.
`parameter:` Suppose p(x) depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a parameter of distribution. The collection of all probability distributions for different values of the parameter is called a family of probability distributions.

Cumulative Distribution Function

(To compute the probability that the observed value of X will be at most some given x)
`Cumulative Distribution Function(cdf):` $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \le x) = \sum_{y : y \le x} p(y)$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.
For discrete rv, the graph of $F(x)$ will be a step function- jump at every possible value of X and flat btwn possible values.
For any two number a and b with $a \le b$:
$P(a \le X \le b) = F(b) - F(a^-)$
$P(a < X \le b) = F(b) - F(a)$
$P(a \le X \le a) = F(a) - F(a^-) = p(a)$
$P(a < X < b) = F(b^-) - F(a)$
(where a^- is the largest possible X value strictly less than a)
Taking $a = b$ yields $P(X = a) = F(a) - F(a-1)$ as desired.
`Expected value or Mean Value`

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

Describes where the probability distribution is centered and is just a weighted average of the possible values of X given their distribution. However, the sample average of a sequence of X values may not settle down to some finite number (harmonic series) but will tend to grow without bound. Then the distribution is said to have a *heavy tail*. Can make if difficult to make inferences about μ.
`The Expected Value of a Function`: Sometimes interest will focus on the expected value of some function $h(x)$ rather than on just $E(x)$.
If the RV X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(x)$, denoted by $E[h(X)]$ or $\mu_{h(X)}$ is computed by

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

`Properties of Expected Value`:

$$E(aX+b) = a \cdot E(X) + b$$

`Variance of X`: Let X have pmf $p(x)$ and expected value μ. Then the $V(X)$ or σ_X^2 is

$$V(X) = \sum_D (x-\mu)^2 \cdot p(x) = E[(X-\mu)^2]$$

The standard deviation (SD) of X is $\sigma = \sqrt{\sigma}$
Alternatively,

$$V(X) = \sigma^2 = \big[\sum_D x^2 \cdot p(x)\big] - \mu^2 = E(X^2) - [E(X)]^2$$

`Properties of Variance`

1. $V(aX+b) = a^2 \cdot \sigma^2$
2. In particular, $\sigma_{aX} = |a| \cdot \sigma_x$
3. $\sigma_{X+b} = \sigma_X$

Continuous RVs

Probabilities assigned to various outcomes in $\mathcal{S}$ in turn determine probabilities associated with the values of any particular rv X. Recall: an rv X is continuous if its set of possible values is uncountable and if $P(X=c) = 0 \quad \forall c.//$

`Probability Density Fxn/Probability Distribution,(pdf)`: $\forall a, b \in \mathbb{R}, a \leq b$

$$P(\forall w \in \mathcal{W} : a \leq X(w) \leq b) = \int_a^b f(x)dx$$

Gives the probability that X takes values between a and b. The conditions $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x) = 1$ are required for any pdf.

`Cumulative Distribution Function(cdf):`

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

For any number $x, F(x)$ is the probability that the observed value of X will be at most x.
By the continuity arguments for continuous RVs we have that

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a < X < b)$$

Other probabilities can be computed from the cdf $F(x)$:

$$P(X > a) = 1 - F(a)$$

$$P(a \leq X \leq b) = F(b) - F(a)$$

Furthermore, if X is a cont rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which $F'(x)$ exists, $F'(x) = f(x)$.
`Median`($\tilde{\mu}$): is the 50th percentile st $F(\tilde{\mu}) = .5$. That is half the area under the density curve. For a symmetric curve, this is the point of symmetry.

`Expected/Mean Value`(μ or $E(X)$): of cont rv with pdf $f(x)$

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

If X is a cont rv with pdf $f(x)$ and $h(X)$ is any function of X then

$$E[h(X)] = \mu = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$$

`Variance:` of a cont rv X with pdf $f(x)$ and mean value μ is

$$\sigma_x^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2]$$

Alternatively,

$$V(X) = E(X^2) - [E(X)]^2$$

Discrete Distributions

The Binomial Probability Distribution

1) The experiment consists of n *trials* where n is fixed
2) Each trial can result in either success (S) or failure (F)
3) The trials are independent
4) The probability of success $P(S)$ is constant for all trials
Note that in general if the sampling is without replacement, the experiment will not yield independent trials. However, if the sample size (number of trials) n is at most 5% of the population, then the experiment can be analyzed as though it were exactly a binomial experiment.
`Binomial rv X:` = no of S's among the n trials
`pmf of a Binomial RV:`,

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x} \quad : x = 0, 1, 2, \ldots$$

`cdf for Binomal RV:` Values in Tble A.1

$$B(x;n,p) = P(X \leq x) = \sum_{y=0}^{x} b(y;n,p)$$

`Mean & Variance of X` If $X \sim Bin(n,p)$ then

$$E(X) = np \quad V(X) = npq$$

Negative Binomial Distribution

1) The experiment consists of independent trials
2) Each trial can result in either Success(S) or Failure(F)
3) The probability of success is constant from trial to trial
4) The experiment continues until a total of r successes have been observed, where r is a specified integer.
`RV Y:` = the no of trials before the rth success.
`Negative Binomial rv:` $X = Y - r$ the number of failures that precede the rth success. In contrast to the binomial rv, the number of successes is fixed while the number of trials is random.
`pmf of the negative binomial rv` : with parameters r = number of S's and $p = P(S)$ is

$$nb(x;r,p) = \binom{x+r-1}{r-1} p^r (1-p)^x \qquad x = 0, 1, 2, \ldots$$

`Mean & Variance of negative binomial rv X:` with pmf $nb(x;r,p)$

$$E(X) = \frac{r(1-p)}{p} \quad V(X) = \frac{r(1-p)}{p^2}$$

Geometric Distribution

`RV X:` = the no of trials before the 1st success.
`pmf of the geometric rv` :

$$p(x) = q^{x-1}p$$

$$E(X) = \sum xq^{x-1}p = 1/p$$

The Poisson Probability Distribution

Useful for modeling rare events
1) independent: no of events in an interval is independent of no of events in another interval
2) Rare: no 2 events at once
3) Constant Rate: average events/unit time is constant ($\mu > 0$)
`RV X=` no of occurrence in unit time interval
`Possion distribution/ Poisson pmf`: of a random variable X with parameter $\mu > 0$ where

$$p(x;\mu) = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad x = 0, 1, 2, \ldots$$

`Binomial Approximation`: Suppose that in the binomial pmf $b(x;n,p)$, we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. Then $b(x;n,p) \to p(x;\mu)$.
That is to say that in any binomial experiment in which n(the number of trials) is large and p(the probability of success) is small, then $b(x;n,p) \approx p(x;\mu)$, where $\mu = np$.
`Mean and Variance of X:` If X has probability distribution with parameter μ, then $E(X) = V(X) = \mu$

Continuous Distributions

The Normal Distribution, $X \sim N(\mu, \sigma^2)$

`PDF:` with parameters μ and σ where $-\infty < \mu < \infty$ and $0 < \sigma$

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty$$

We can then easily show that $E(X) = \mu$ and $V(X) = \sigma^2$.
`Standard Normal Distribution:` The specific case where $\mu = 0$ and $\sigma = 1$. Then

$$\text{pdf}: \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{cdf}: \quad \Phi(z) = \int_{-\infty}^{z} \phi(u)du$$

`Standardization:` Suppose that $X \sim N(\mu, \sigma^2)$. Then

$$Z = (X-\mu)/\sigma$$

transforms X into standard units. Indeed $Z \sim N(0,1)$.

$$P(a \leq X \leq b) = P\Big(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\Big) = \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big)$$

`Independence:` If $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ and X and Y are independent, then $X \pm Y \sim N(\mu_x \pm \mu_y, \sigma_x^2 + \sigma_y^2)$
NOTE: By symmetry of the standard normal distribution, it follows that $\Phi(-z) = 1 - \Phi(z) \quad \forall z \in \mathbb{R}$

`Normal Approx to Binomial Dist:` Let $X \sim Bin(n,p)$. As long as a binomial histogram is not too skewed, Binomial probabilities can be well approximated by normal curve areas.

$$P(X \le x) = B(x;n,p) \approx \Phi\left(\frac{x+0.5-np}{\sqrt{np(1-p)}}\right)$$

As a rule, the approx is adequate provided that both $np \ge 10$ and $n(1-p) \ge 10$.

The Exponential Distribution, $X \sim Exp(\lambda)$

Model for lifetime of firms/products/humans

`Exponential Distribution:` A cont rv X has exp distribution if its pdf is given by

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \ge 0 \quad \lambda > 0$$

$F(x,\lambda)=$ $\quad P(X \le x) = 1 - e^{\lambda x} \quad x \ge 0$
$E(X)=$ $\quad 1/\lambda$
$V(X)=$ $\quad 1/\lambda^2$
`Memoryless Prop`: $P(X > a+x | X > a) = P(X > x)$ for $x \in D, a > 0$

Note: If Y is an rv distributed as a Poisson $p(y;\lambda)$, then the time between consecutive Poisson events is distributed as an exponential rv with parameter λ

Joint Probability Dist

`Joint Range:` Let $X : \mathcal{S} \to \mathbb{D}_1$ and $Y : \mathcal{S} \to \mathbb{D}_2$ be 2 rvs with a common sample space. We define the joint range of the vector (X,Y) of the form

$$\mathbb{D} = \mathbb{D}_1 \times \mathbb{D}_2 = \{(x,y) : x \in \mathbb{D}_1, y \in \mathbb{D}_2\}$$

`Random Vector:` A 2-D random vector(X,Y) is a function from $\mathcal{S} \to \mathbb{R}^2$. It is defined $\forall \omega \in \mathcal{S}$ such that

$$(X,Y)(\omega) = (X(\omega), Y(\omega)) = (x,y) \in \mathbb{D}$$

`Joint Probability Mass Fxn:` For two discrete rv's X and Y. The joint pmf of (X,Y) is defined $\forall (x,y) \in \mathbb{D}$

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

It must be that $p(x,y) \ge 0$ and $\sum_i \sum_j p(x_i,y_j) = 1$.

`Marginal Prob Mass Fxn:` of X and of Y, denoted $p_X(x)$ and $p_Y(y)$ respectively,

$$p_X(x) = \sum_{y : p(x,y) > 0} p(x,y) \quad \forall x \in \mathbb{D}_1$$

`Joint Probability Density Fxn:` For two continuous rv's X and Y. The joint pdf of (X,Y) is defined $\forall A \subseteq \mathbb{R}^2$

$$P((X,Y) \in A) = \iint_A f(x,y)dxdy$$

It must be that $f(x,y) \ge 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$.
Note also that this integration is commutative.

`Marginal Prob Density Fxn:` of X and of Y, denoted $f_X(x)$ and $f_Y(y)$ respectively,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \forall x \in \mathbb{D}_1$$

Note that if $f(x,y)$ is the joint density of the random vector (X,Y) and $A \in \mathbb{R}^2$ is of the form $A = [a,b] \times [c,d]$ we have that

$$P((X,Y) \in A) = \int_c^d \int_a^b f(x,y)dxdy = \int_a^b \int_c^d f(x,y)dxdy$$

`Independence:` Two rvs are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad f(x,y) = f_X(x)f_Y(y)$$

`Conditional Distribution(discrete):` For two discrete rv's X and Y with joint pmf $p(x_i,y_j)$ and marginal X pmf $p_X(x)$, then for any realized value x in the range of X, the conditional mass function of Y, given that $X = x$ is

$$p_{Y|X}(y|x) = \frac{p(x_i,y_j)}{p_X(x)}$$

`Conditional Distribution(cont):` For two continuous rv's X and Y with joint pdf $f(x,y)$ and marginal X pdf $f_X(x)$, then for any realized value x in the range of X, the conditional density function of Y, given that $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

Expected Values, Covariance & Correlation

`Expected value`: The expected value of a function $h(X,Y)$ of two jointly distributed random variables is

$$E(g(X,Y)) = \sum_{x \in \mathbb{D}_1} \sum_{y \in \mathbb{D}_2} g(x,y)p(x,y)$$

and can be generalized to the continuous case with integrations.//

`Covariance:` Measures the strength of the relation btwn 2 RVs, however very

$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

Shortcut Formula:

$$Cov(X,Y) = E(XY) - \mu_x\mu_y$$

The defect of the covariance however is that its value depends critically on the units of measurement.

`Correlation:` Cov after standardization. Helps interpret Cov.

$$\rho = \rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

Has the property that $Corr(aX+b, cY+d) = Corr(X,Y)$ and that for any rvs X, Y $-1 \le \rho \le 1$.
Note also that ρ is independent of units, the larger $|\rho|$ the stronger the linear association, considered strong linear relationship if $|\rho| \ge 0.8$.
Caution though: if X and Y are independent then $\rho = 0$ but $\rho = 0$ does not imply that X,Y are independent.
Also that $\rho = 1$ or -1 iff $Y = aX + b$ for some a, b with $a \ne 0$.

`Statistic:` Any quantity whose value can be calculated with sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

`Sampling Distribution:` probability distribution of a statistic, it describes how the statis- tic varies in value across all samples that might be selected

Stats & Their Distributions

Fxns of Observed Sample Observ

`Obs Sample Mean` $\bar{x} = \frac{1}{n}\sum x_i$
`Obs Sample Var` $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$
`Obs Sample Max` $x_{(n)} = max(x_i)$

A statistic is a random variable and the most common are listed above.

`Simple Random Samples:` The random variables $X_1, \ldots, X_n$ are said to form a simple random sample of size n if each X_i is an independent random variable, every X_i has the same probability distribution.

`Sampling Distribs:` Every statistic has a probability distribution (a pmt or pdf) which we call its sampling distribution. To determine its distrib can be hard but we use simulations and the CLT to do so.

`Simulation Experiments:` we must specify the statistic of interest, the population distribution, the sample size(n) and the number of samples (k). Use a computer to simulate each different simple random sample, construct a histogram which will give approx sampling distribution of the statistic.

The Dist % Sample Mean

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then $E(\bar{X}) = \mu_{\bar{X}} = \mu$ and $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$. Also if $S_n = X_1 + \ldots + X_n$ then $E(S_n) = n\mu$ and $V(S_n) = n\sigma^2$.

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from a normal distribution with mean μ and variance σ^2. Then for any n, $\bar{X}$ is normal distributed with mean μ and variance σ^2/n. Also S_n is normal distributed with mean $n\mu$ and variance $n\sigma^2$.

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from Bernoulli(p), then $S_n \sim$ Binomial(n,p).

Distribution of The Sample Mean $\bar{X}$

Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then $E(\bar{X}) = \mu_{\bar{X}} = \mu$ and $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$

The standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is often called the *standard error of the mean.*

For a NORMAL random sample with the same mean and std as above, then for any n, $\bar{X}$ is normally distributed with the same mean and std.

`Central Limit Theorem:` Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\bar{X}$ has approximately a normal dis with mean μ and variance σ^2/n. Also S_n is normal distributed with mean $n\mu$ and variance $n\sigma^2$. No matter which population we sample from, the probability histogram of the sample mean follow closely a normal curve when n is sufficiently large. *Rule of thumb: if $n \ge 30$ CLT can be used.* It follows from CLT that is $X \sim Bin(n,p)$ and n is large, then n can be distributed by a $N(np, npq)$.

Dist of a Linear Combination

`Linear Comb:` Let $X_1, \ldots, X_n$ be a collxn of n random variables and let $a_1 \ldots a_n$ be n numerical constants. Then the random variable $Y = a_1X_1 + \ldots a_nX_n$ is a linear comb of the $X_i's$.

1. Regardless of whether the $X_i's$ are independent or not
$$E(Y) = a_1E(X_1) + \ldots + a_nE(X_n) = a_1\mu_1 + \ldots + a_n\mu_n$$
2. If $X_1, \ldots, X_n$ are independent
$$V(Y) = V(a_1X_1 + \ldots a_nX_n) + a_1^2\sigma_1^2 + \ldots$$
3. For any $X_1, \ldots, X_n$,
$$V(Y) = \sum_{i=1}\sum_{j=1} a_ia_jCov(X_i, X_j)$$
4. If $X_1, \ldots, X_n$ are independent, normally distributed rvs, then any linear combination of the rvs also has a normal distribution- as does their difference.

$E(X_1 - X_2) = E(X_1) - E(X_2), \forall X, Y$ while
$V(X_1 - X_2) = V(X_1) + V(X_2)$ if X1, X2 independent,

2. Estimators

`Parameter of Interest` (θ) true yet unknown pop parameter
`Point Estimate:` ($\hat{\theta}$) Our guess for θ based on sample data
`Point Estimator:` ($\hat{\theta}$) statistic selected to get a sensible pt est
A sensible way to quantity the idea of $\hat{\theta}$ being close to θ is to consider the least squared error $(\hat{\theta} - \theta)^2$. A good measure of the accuracy is the expected or mean square error MSE = $E[(\hat{\theta} - \theta)^2]$. It is often not possible to find the estimator with the smallest MSE so we often restrict our attention to *unbiased* estimators and find the best estimator of this group.
`Unbiased:` Pt Est $\hat{\theta}$ if $E(\hat{\theta}) = \theta$ for all θ.
Then $\hat{\theta}$ has a prob distribution that is always "centered" at the true θ value.
When choosing estimators, select the unbiased and the one that has the minimum variance.

Estimators

-When $X \sim Bin(n, p)$, the sample proportion $\hat{p} = X/n$ is an unbiased est of p.
- Let $X_1, \ldots, X_n$ be a SRS from a distribution with mean μ and variance σ^2.Then $\hat{\sigma}^2 = S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ is unbiased for σ^2.
-Let $X_1, \ldots, X_n$ be a SRS from a distribution with mean μ, then $\bar{X}$ is MVUE for μ.
`Standard Error:` of an estimator is its standard deviation $\sigma_{\hat{\theta}} - \sqrt{V(\hat{\theta})}$
`Estimated Standard Error:` If the standard error itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields $\sigma\hat{\theta} = s_{\hat{\theta}}$.

Method of Moments

Let $X_1, \ldots, X_n$ be a SRS from a pdf $f(x)$. For $k = 1, 2, \ldots$ the kth population moment, or kth moment of the distribution $f(x)$, is $E(X^k)$. The kth sample moment is $(1/n)\sum_{i=1}^n X_i^k$.
Let $X_1, \ldots, X_n$ be a SRS from a distribution with pdf $f(x; \theta_1 \ldots \theta_m)$ where θ_i's are unknown. Then the moment estimators $\hat{\theta}_i$'s are obtained from the first m sample moments to the corresponding first m population moments and solving for the θ_i's.

Maximum Likelihood Estimator

Works best when the sample size is large!
Let $X_1, \ldots, X_n$ have joint pmf or pdf
$$f(x_1, \ldots, x_n; \theta_1 \ldots \theta_m)$$
where the θ_i's have unknown values.
When $x_1, \ldots, x_n$ are observed sample values, the above is considered a fxn of the θ_i's and is called the **likelihood function**.
The maximum likelihood estimates (mles) $\hat{\theta}_i$'s are those θ_i's that maximize the likelihood function such that
$$f(x_1, \ldots, x_n; \hat{\theta_1} \ldots \hat{\theta_m}) \geq f(x_1, \ldots, x_n; \theta_1 \ldots \theta_m) \quad \forall \theta_1 \ldots \theta_m$$
When $X_1, \ldots, X_n$ substituted in, the **maximum likelihood estimators** result.

3. Confidence Intervals

Tests in a single sample

When measuring n random variables $Y_i \sim i.i.d.$

Hypotheses about the population mean $E[Y_i]$
`Z-test` (when $n > 40$ or if normality with known variances could be assumed)
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
`CI for Normal Population:` A $100(1-\alpha)\%$ CI for the mean μ of a population when σ is known is
$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$
`T -test` (normality must be assured; for large n this is the same as the z-test). When $\bar{X}$ is the sample mean of a SRS of size n from a Normal(μ, σ^2) population then the RV
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$
has a probability distribution-t with n-1 degrees of freedom.
Note: the density of t_ν is symmetric around 0. t_ν is more spread out than a normal, indeed the few dof the more spread. When dof is large (< 40), the t and normal curve are close. In addition we have that
$$P(|\frac{\bar{X} - \mu}{S/\sqrt{n}}| \leq t_{\alpha/2, n-1}) = 1 - \alpha$$
As a result, the $(1-\alpha)100\%$ CI for the population mean μ under the normal model is
$$\bar{X} \pm t_{\alpha/2, n-1}\frac{S}{\sqrt{n}}$$
Note that here we make the assumption that the observations are realizations of a SRS from a Normal distribution with unknown mean and variance.// `Large Sample Test` for the population proportion (proportions are just means; only valid for $np_0 \geq 10$ and $n(1-p_0) \geq 10$. The $(1-\alpha)$ confidence interval for a population mean μ is
$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$
For a population proportion
$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \qquad \hat{p} = \bar{X}$$
Hypotheses about the population variance $V[X_i]$
The $(1-\alpha)100\%$ CI for the variance σ^2 of a normal population has a lower limit:
$$(n-1)s^2/\chi^2_{\alpha/2, n-1}$$
and Upper limit:
$$(n-1)s^2/\chi^2_{1-\alpha/2, n-1}$$
A confidence interval for σ has lower and upper limits that are the square roots of the corresponding limits in the interval for σ^2. An upper or a lower confidence bound results from replacing $\alpha/2$ with α in the corresponding limit of the CI.

When measuring two variables for each unit $(X_i, Y_i) \sim i.i.d.$

`Paired t-test` about the difference of population means:
`Test about parameters` β_1 and β_0

Tests in two non-paired, independent samples

4. Hypothesis Testing

In it hard to example the evidence of such a strong count as a lucky draw. The p-value or observed significance level determines whether or not a hypothesis will be rejected- the smaller it is, the stronger evidence against he null hypothesis. The plausibility of statistical models determined by the null hypothesis is based on the sample data and their distributions. The idea is that the null is not rejected unless it is testified implausible overwhelmingly by data.

`Possible Errors:` Type I: reject the null hypothesis when it is true; Type II: fail to reject the null even though it is false.

Power Function

For a given test with critical or rejection region $\{x : T(x) \geq c\}$, the power function is defined as

$$\phi(\theta) = P(T(X_1, \ldots, X_n) \geq c|\theta) = P(T \geq c|\theta)$$

In other words, $\phi(\theta)$ represents the *probability of rejection* H_0 if a particular θ were the true value of parameter of the pmt or pdf $f(x;\theta)$.
In other words, if H_0 is true, $\phi(\theta)$ = Probability of type 1 error. If H_0 is false, $\phi(\theta)$ = 1- Probability of type 2 error.
A court trial, where the null hypothesis is "not guilty" unless there is convincing evidence against it. The aim or purpose of court hearings (collecting data) is to establish the assertion of "guilty" rather than to prove "innocence."
P-value (or observed signicance level) is the probability, calculated assuming that H0 is true, of obtaining a value of the test statistic at least as contradictory to H0 as the value calculated from the available sample. It is also the smallest significance level at which one can reject H0.
In other words, suppose we have observed a realization $x_{obs} = (x1, \ldots, xn)$ of our random sample $X_1, \cdots, X_n \sim f(x, \theta)$. We wish to investigate the compatibility of the null hypothesis, with the observed data. We do so by comparing the probability distribution of the test statistic $T(X1, \cdots, Xn)$ with its observed value $t_{obs} = T(x_{obs})$, assuming H0 to be true. As a measure of compatibility, we calculate

$$p(xobs) = \text{p-value} = P(T(X1, \ldots, X_n) \geq t_{obs}|H0)$$

In general, report the p-value. When it is less than 5% or 1 %, the result is statistically significant.

Hypotheses and Test Procedures

Statistical hypothesis(hypothesis) is a claim or assertion about the value of a single parameter, about the values of several parameters, or about the form of an entire population distribution.
In any hypothesis-testing problem, there are two contradictory hypotheses under consideration.
The **null hypothesis**, denoted H_0 is the claim that is initially assumed to be true (the "prior belief" claim). Often called the hypothesis of no change (from current opinion) and will generally be stated as an equality claim, equal to the *null value*. The **alternative hypothesis** or researcher's hypothesis, denoted by H_a is the assertion that is contradictory to H_0. The alt hypothesis is often the claim that the researcher would really like to validate.
The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis. The two possible conclusions from a hypothesis-testing analysis are then reject H_0 or fail to reject H_0.
A **test of hypotheses** is a method for using sample data to decide whether the null hypothesis should be rejected.
A **test procedure** is a rule based on sample data, for deciding whether to reject H_0. A procedure has 2 consitutents: 1) a test static, or function of the sample data used to make a decision and 2) a rejection region consisting of those x values for which H_0 will be rejected in favor of H_a.
A test procedure is specified by the following:

1. A **test statistic**, a function of the sample data on which the decision (reject H_0 or do not reject H_0) is to be based
2. A **rejection region**, the set of all test statistic values for which H_0 will be rejected. The basis for choosing a rejection region lies in consideration of the errors that one might be faced with in drawing a conclusion.

The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.
A **type I error** consists of rejecting the null hypothesis H_0 when it is true- a false negative. A **type II error** involves not rejecting H_0 when H_0 is false- a false positive.
In the best of all possible worlds, test procedures for which neither type of error is possible could be developed. However, this ideal can be achieved only by basing a decision on an examination of the entire population. The difficulty with using a procedure based on sample data is that because of sampling variability, an unrepresentative sample may result, e.g., a value of $\bar{X}$ that is far from μ or a value of $\hat{p}$ that differs considerably from p.
Suppose an experiment and a sample size are fixed and a test statistic is chosen. Then decreasing the size of the rejection region to obtain a smaller value of α results in a larger value of β for any particular parameter value consistent with H_a. In other words, once the test statistic and n are fixed, there is no rejection region that will simultaneously make both α and all β's small. A region must be chosen to effect a compromise between α and β.

Tests About a Population Mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

α = P(H_0 is rejected when H_0 is true) = false negative = $P(\bar{X} \leq 70.8$ when $\bar{X} \sim$ normal with $\mu_{\bar{X}} = 75, \sigma_{\bar{X}} = 1.8) = P(Z \geq c$ null when $Z \sim N(0,1))$
β = P(H_0 is accepted when H_0 is false) = false positive = $P(\bar{X} > 70.8$ when $\bar{X} \sim$ normal with $\mu_{\bar{X}} = 72, \sigma_{\bar{X}} = 1.8)$

Tests about a Population Mean

Case1: A Normal Population with a Known σ

Assuming that the sample mean $\bar{X}$ has a normal distribution with $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. When H_0 is true, $\mu_{\bar{X}} = \mu_0$. Consider now the statistic Z obtained by standardizing $\bar{X}$ under the assumption that H_0 is true:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

5. Simple Linear Regression and Correlation

Common theme: to study the relationships among variables.

Model and Summary Statistics

`Bivariate Data:` $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$
`Generic Pair` (X, Y) X- predictor, independent variable, covariate
`Simple Linear Regression:` $Y = \beta_0 + \beta_1 x + \varepsilon$
`Betas` regression coeffs, ε measurement error, cannot be explained by x
The ith observation is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and we further assume that ε_i are iid $N(0, \sigma^2)$
`Conditional Expected Value:` For the linear model we have that $E(Y|x) = E(\beta_0 + \beta_1 x + \varepsilon_0) = \beta_0 + \beta_1 x$ which is the average for the group with covariate $\sim x$
`Conditional Standard Deviation:` Similarly we have that $V(Y|x) = \sigma^2$ which is the variance for the group with covariate $\sim x$
`Summary Stats x:` $\bar{x}$ and $SD_x = \sqrt{\frac{S_{xx}}{n-1}}$ or $S_{xx} = \sum(x_i - \bar{x})^2$
`Sum Stats y:` $\bar{y}$ and $SD_y = \sqrt{\frac{S_{yy}}{n-1}}$ or $S_{yy} = \sum(y_i - \bar{y})^2$
`Strength of Linear Assoc:` $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ the sample correlation coeff.

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

`Purpose of the Regression:` To quantify the contribution of the predictors $X_1 \ldots X_p$ on the outcome of Y, given $(x_1, \ldots, x_p)$ predict the mean response, quantify the uncertainty in this prediction (with standard error/confidence interval), extrapolate

Estimation of Model Parameters

Data are modeled as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1 \ldots n \quad \varepsilon_i \sim N(0, \sigma^2)$$

How to find good estimates for β_0 & β_1?
- The error between y_i and $\beta_0 + \beta_1 x$ is ε_i and we want to minimize the total "loss"
-In the case of squared-error loss functions, the total loss is $\sum \varepsilon_i^2$
-To minimize, take partial derivatives of SSE wrt each β and set each to zero. Then solve the system of linear equations for each β. In this case

$$\hat{\beta_1} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} = r\frac{SD_x}{SD_y}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

`Fitted Values:` $\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$ The value y_i predicted based on x_i
`Residuals:` $\hat{\varepsilon_i} = y_i - \hat{y_i}$ Difference between predicted and actual y
`Resdal Sum of Squares:` SSE = SSE $(\hat{\beta_0}, \hat{\beta_1}) = \sum \hat{\varepsilon_i}^2$
`Regression Line:` $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$ Used to predict the mean response $\hat{y}$ for a given x

Estimating σ^2

$$\sigma^2 = \frac{1}{n-2}\sum \hat{\varepsilon}^2 = SSE/2$$

It can be shown that $SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy}(1-r^2)$ and hence

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{n-1}{n-2}} SD_y \sqrt{1-r^2}$$

which is smaller that SD_y- the regression has decreased uncertainty about y.

Goodness of fit

`Sum of squares due to regression (SSR)`

$$SS_{reg} = S_{yy} - SSE$$

`Coeff of Determination` R^2: Percentage of variability of Y explained by the regression on X. The larger it is, the better the fit.

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$$

Inference for Model Parameters

$$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta_0} = \bar{y} - \hat{\beta_1} x$$

$$SE(hat)(\hat{\beta_1}) = \hat{\sigma}/\sqrt{S_{xx}} \quad SE(hat)(\hat{\beta_0}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

where the T-statistic is:

$$T = \frac{\hat{\beta_1} - \beta_1}{\hat{SE}(\hat{\beta_1})}$$

`Standard Errors:` SInce the estimators are linear in Y
`Confidence Intervals:` $\hat{\beta_0} \pm t_{\alpha/2, n-2}$

6. Goodness of Fit

Condition: for each cell, the expected count is greater than five
Multinomial dist: probability weights on discrete, unordered possible outcomes
`Homogeneity:` Along the rows we have diff populations and columns are difference categories.
H0: proportion of individuals in category j is the same for each population and that is true for every category. $p_{1j} = \ldots = p_{Ij}$ for $j = 1 \ldots J$
Estimated expected: $\hat{e_{ij}} = \frac{\text{(ith row total)(jth column total)}}{n}$ Test Statistic:

$$\chi^2 = \sum \frac{(ob - estex)^2}{estex} = \sum\sum \frac{(n_{ij} - \hat{e_{ij}})^2}{\hat{e_{ij}}}$$

Rejection Region: $\chi^2 \geq \chi^2_{\alpha(I-1)(J-1)}$
`Independence:` Only one population but looking at the relationship btwn 2 different factors. Each individual in one category associated with first factor and one category associated with second factor.
H0: The null hypothesis here says that an individuals category with respect to factor 1 is independent of the category with respect to factor 2. In symbols, this becomes $p_{ij} = p_i p_j \forall (i,j)$.
Test Statistic, RR and Condition: Same as above
State the uncertainty in a particular estimate of ours.

Basics

The actual sample observations $x1, \ldots, xn$ are assumed to be the result of a random sample $X1, \ldots, Xn$ from a normal distribution with mean value μ and standard deviation σ. We know then(from Ch5) that $\bar{X} \sim N(\mu, \sigma^2/n)$. Standardizing yields

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Obtain an inequality such as

$$P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

and we manipulate the inequality so that it apprears in the form $l \leq \mu \leq u$ where l,u involve factors save μ.
This interval we now describe is random since the endpoints involve a random variable and centered at $\bar{X}$. It says the probability is .95 that the random interval includes or covers the true value of μ. The confidence level 95% is not so much a statement about any particular interval, instead it pertains to what would happen if a very large number of like intervals were to be constructed using the same CI formula.
`CI for Normal Population:` A $100(1-\alpha)$% CI for the mean μ of a population when σ is known is

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

or equivalently,

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

`Necc Sample Size:` for a CI to have width w is $n = (2z_{\alpha/2} \cdot \frac{\sigma}{w})^2$
Note that for sufficiently large n, σ is replaced by S, the sample variance.
`General Large-Sample CI`: Suppose that $\hat{\theta}$ is an estimator approx normal, unbiased, and has an expression for $\sigma_{\hat{\theta}}$. Then standardizing yields

$$P(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}) \approx 1 - \alpha$$

Population Mean (if variance unknown)

With 95% chance the random interval covers μ, population mean.
`Interpretation:` When the estimator is replaced by an estimate, the random interval becomes a realized interval. The word confidence refers to the procedure. If we repeat the experiment many times and construct 95% confidence intervals int he same manner, about 95% of them cover the unknown, but fixed, μ. We don't know whether the current interval covers μ or not but we know that of all the intervals ever constructed 95% will cover.

General Confidence Intervals

When the sample size is large (> 40), the $(1-\alpha)$ confidence interval for a population mean μ is

$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

For a population proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \qquad \hat{p} = \bar{X}$$

Steps for calculating Confidence Intervals

1) Find an RV having an (approximately) known distribution
2) Cut off tails, that is, select a confidence level $(1-\alpha)$
3) Solve the equation to obtain confidence intervals- isolate the population mean in an approbate string of inequalities.

Intervals Based on a Normal Population

When the sample size is small, we can no longer use the CLT. But maybe we can assume that the data comes from a normal population. In that case we need to account for the uncertainty in estimating σ but by how much?
`T-Statistic:` When $\bar{X}$ is the sample mean of a SRS of size n from a Normal(μ, σ^2) population then the RV

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a probability distribution-t with n-1 degrees of freedom.
Note: the density of t_ν is symmetric around 0. t_ν is more spread out than a normal, indeed the few dof the more spread. When dof is large (< 40), the t and normal curve are close. In addition we have that

$$P(|\frac{\bar{X} - \mu}{S/\sqrt{n}}| \leq t_{\alpha/2, n-1}) = 1 - \alpha$$

As a result, the $(1-\alpha)$% CI for the population mean μ under the normal model is

$$\bar{X} \pm t_{\alpha/2, n-1}\frac{S}{\sqrt{n}}$$

Note that here we make the assumption that the observations are realizations of a SRS from a Normal distribution with unknown mean and variance.

One-Sided Confidence Bounds

`Lower Confidence Bound:` When n is large, then

$$P\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z_\alpha\right) = 1 - \alpha$$

and solve to find the $(1-\alpha)$ confidence bound $\bar{X} - z_\alpha \frac{s}{\sqrt{n}}$.
`Upper Confidence Bound:` With $(1-\alpha)$ confidence, μ is bounded by $\bar{X} + z_\alpha \frac{s}{\sqrt{n}}$
Note that when n is small, replace z_α by $t_{\alpha, n-1}$.

CI for the Variance of a Normal Population

`Theorem:` Let $X_1, ..., X_n$ be a SRS from a Normal(μ, σ^2) population, where both parameters are unknown. The RV

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum^n (X_i - X)^2}{\sigma^2}$$

has a probability distribution called the χ^2 distribution with n-1 dof.
The density of chi is always positive and has long upper tails. As n increases, the densities become more symmetric.
Furthermore, we have that

$$P\left(\chi_{1-\alpha/2, n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2, n-1}\right) = 1 - \alpha$$

Hence, the $(1-\alpha)$ CI for the population variance σ^2 under the normal model is

$$\left[\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}}, \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}}\right]$$

Hypothesis Testing Cheat Sheet

Key Concepts

◊ Hypothesis testing uses statistical tests to determine if a hypothesis is true.

◊ **The null hypothesis, H_0,** is the statement that there **IS NO** statistically significant difference or relationship between variables.

◊ Any differences that are observed are due to chance. It is a statement of "no effect" or "no difference." It is the hypothesis which a researcher tries to disprove, reject or nullify.

◊ **The alternative hypothesis, H_1 or H_a,** is the statement created by researchers when they speculate upon the outcome of a research or experiment.

◊ The alternative hypothesis states that there **IS** a statistically significant difference or relationship between variables.

◊ The alternative hypothesis is what the researcher really thinks is the cause of a difference. For example, they may be testing the effects of a new drug.

How Does Hypothesis Testing Improve Products & Processes?

Hypothesis testing can be used in business operations as well. Tests can help identify differences between machines, formulas, raw materials, medications, etc. Without such testing, employees may change the product or process causing more variation. Hypothesis tests enable data driven decisions.

Three Hypothesis Testing Methods

1. **Classical:** Compare a test statistic to a critical value.
2. **p value:** Probability of a test statistic being contrary to the null hypothesis.
3. **Confidence Interval:** Is the test statistic between or outside of the confidence interval.

Type I and Type II Errors

Type I error - Reject a null hypothesis that is true (Producer's Risk)

Type II error - Not reject a null hypothesis that is false (Consumer's Risk)

How to Conduct a Hypothesis Test

Steps to Follow

1. Define the null and alternative hypothesis.
2. Conduct the test.
3. Using data from the test:
   - Calculate the test statistic (i.e. F) and the critical value (i.e. F crit).
   - Calculate a p value and compare it to a significance level (α) or confidence level (1-α). For example, if the significance level = 5%, then the confidence level = 95%.
4. Interpret the results to accept or reject the null hypothesis.

Interpreting the Results

Test Method	Compare	Result
Classical	test statistic > critical value (i.e. F > F crit)	Reject the null hypothesis
Classical	test statistic < critical value (i.e. F < F crit)	Cannot reject the null hypothesis (Accept the null hypothesis)
p value	p value < α	Reject the null hypothesis
p value	p value > α	Cannot reject the null hypothesis (Accept the null hypothesis)

Translating Stat Speak to English

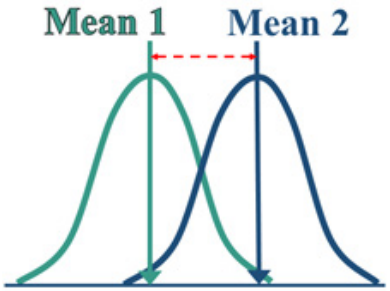
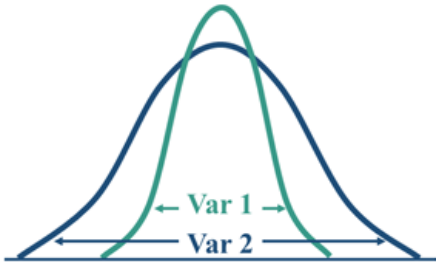
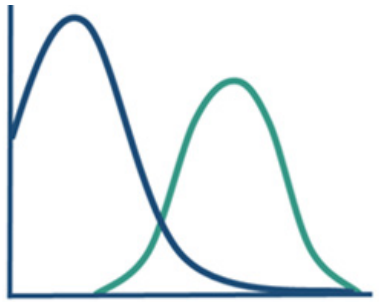
Null Hypothesis: means or variances are **not significantly different.**

Reject the Null Hypothesis	**Cannot Reject the Null Hypothesis (Accept the Null Hypothesis)**
p value < α	p value > α
Means or Variances are different Means or Variances are not the same	Means or Variances are the same Means or Variances are not different

rev 2-2022

Hypothesis Testing Cheat Sheet

Examples of Statistical Tests included in QI Macros for Excel

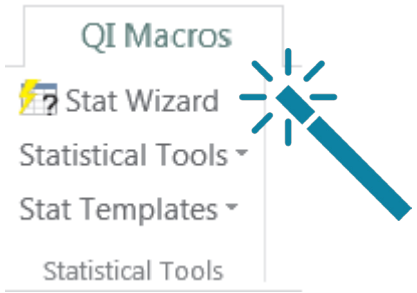
Test of Means	Test of Variances	Test of Relationships	And More
			
ANOVA t tests z test	f test Levenes test	Chi-square Descriptive Statistics Multiple Regression Analysis	AQL Sampling Tables Normality Test Sample Size Calculator

Advantages of Using QI Macros

- Just select your data and then the test you want from QI Macros menu.
- Performs all of the calculations and interprets the results for you. (i.e. Reject null hypothesis because p < α, means are different)
- Draws a box plot or other chart to better explain the results.
- Compatible with PC and Mac. Excel 2013-2021 and Office 365.
- Reduce risk of manual calculations or your own Excel formulas.
- Save Time!

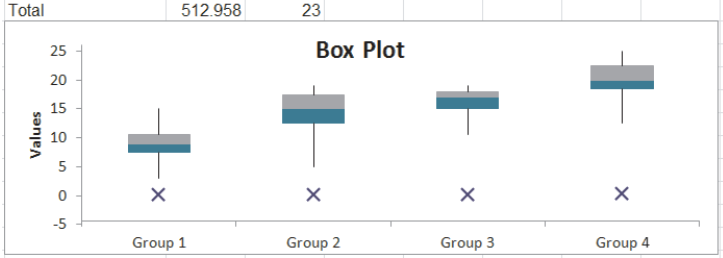
Not Sure Which Statistical Test to Run?

QI Macros Stat Wizard will analyze your data and run the correct tests for you.



Example of QI Macros Results

Anova: Single Factor		α	0.05							
							LSD	3.07242		
SUMMARY							HSD	4.12226		
Groups	*Count*	*Sum*	*Average*	*Variance*			Scheffe	5.79732		
5%	6	60	10	8			Post Hoc	5%	10%	15%
10%	6	94	15.6667	7.86667			10%	5.66667		
15%	6	102	17	3.2			15%	7	1.33333	
20%	6	127	21.1667	6.96667			20%	11.1667	5.5	4.16667
							Colored cells have signficant mean differences			
ANOVA		Reject Null Hypothesis because p < 0.05 (Means are Different)								
Source of Variation	*SS*	*df*	*MS*	*F*	*P-Value*	*F crit*				
Between Groups	382.792	3	127.597	19.6052	0.000	3.09839				
Within Groups	130.167	20	6.50833							
Total	512.958	23								

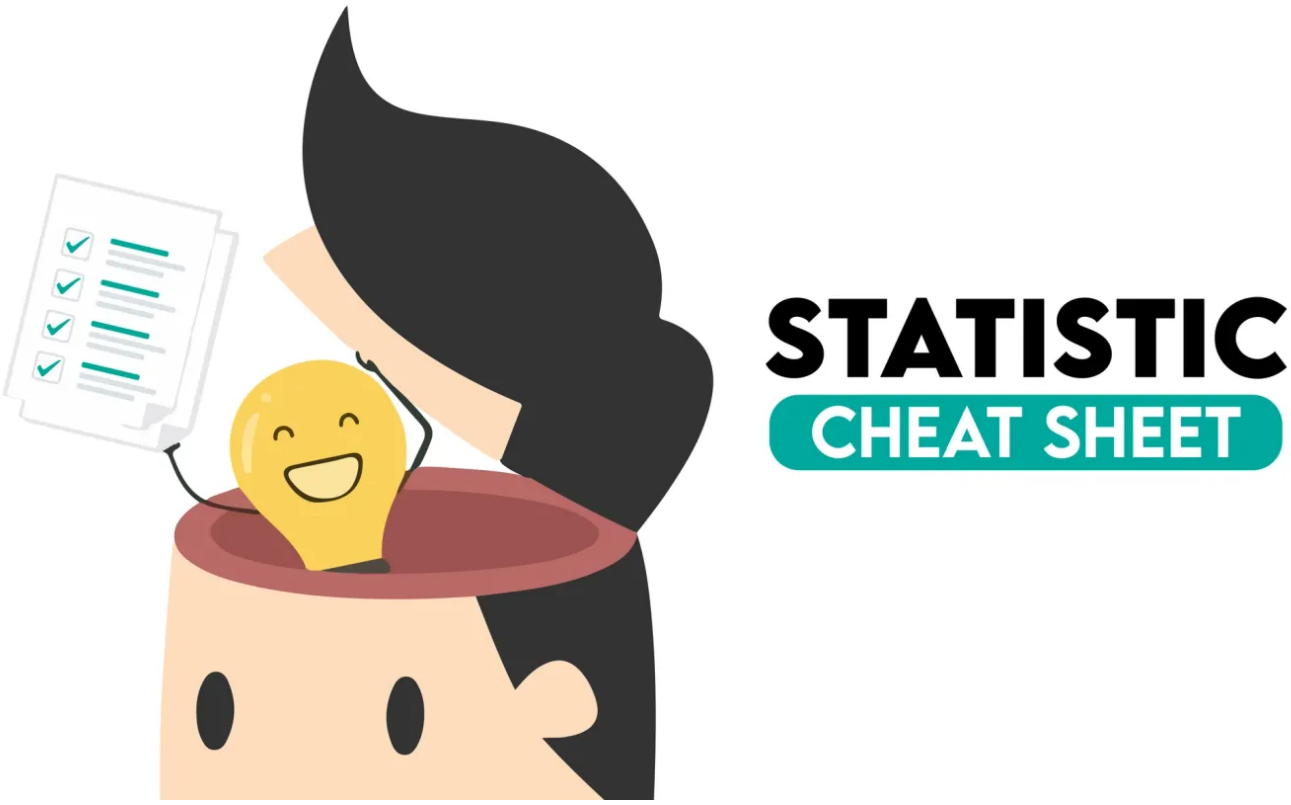


QIMacros®
rev 2-2022

Download 30-Day Trial at: www.qimacros.com

2696 S. Colorado Blvd., Ste. 555
Denver, CO 80222
888-468-1537 or 303-756-9144

A Comprehensive Statistics Cheat Sheet for Data Science Interviews



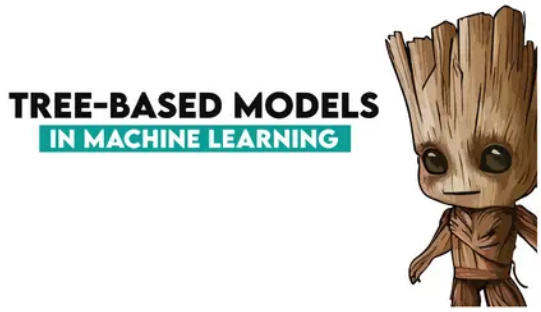
Categories Interviews Statistics

Written by: Nathan Rosidi

Author Bio

October 24th, 2023

Latest Posts:



Modeling Guides

Tree-Based Models in Machine Learning

Interviews

Top Data Warehouse Interview Questions with Answers

Guides

Data Cleaning 101: Avoid These 5 Traps in Your Data

Main Topics

The statistics cheat sheet overviews the most important terms and equations in statistics and probability. You'll need all of them in your data science career.

About This Statistics Cheat Sheet

When I was applying to Data Science jobs, I noticed that there was a need for a comprehensive statistics and probability cheat sheet that goes beyond the very fundamentals of statistics (like mean/median/mode).

And so, I'm going to cover the most important topics that commonly show up in data science interviews. These topics focus more on statistical methods rather than fundamental properties and concepts, meaning it covers topics that are more practical and applicable in real-life situations.

With that said, I hope you enjoy it!

Mean

The mean is a way to find the middle value of numbers. You add up all the numbers and then divide by the total number. It can significantly be swayed by really high or really low values in the set.

Example: Let's say we have the heights of 5 people: 60, 62, 65, 68, and 72 inches. To find the mean height, we add up all the heights and divide by 5:

$$Mean\ height = (60 + 62 + 65 + 68 + 72)/5 = 65.4\ inches$$

This means that the average height of the 5 people is 65.4 inches.

Median

The median is an alternative method for finding a group of numbers' middle value.

You put all the numbers in order from smallest to largest and then find the middle value. If there are an odd number of values, there will be one middle value.

If there are an even number of values, then you find the average of the two middle values.

Example: If we have the same set of heights {60, 62, 65, 68, 72}, we put them in order from smallest to largest: {60, 62, 65, 68, 72}. Since there are an odd number of values, the middle value is 65. So the median height of the 5 people is 65 inches.

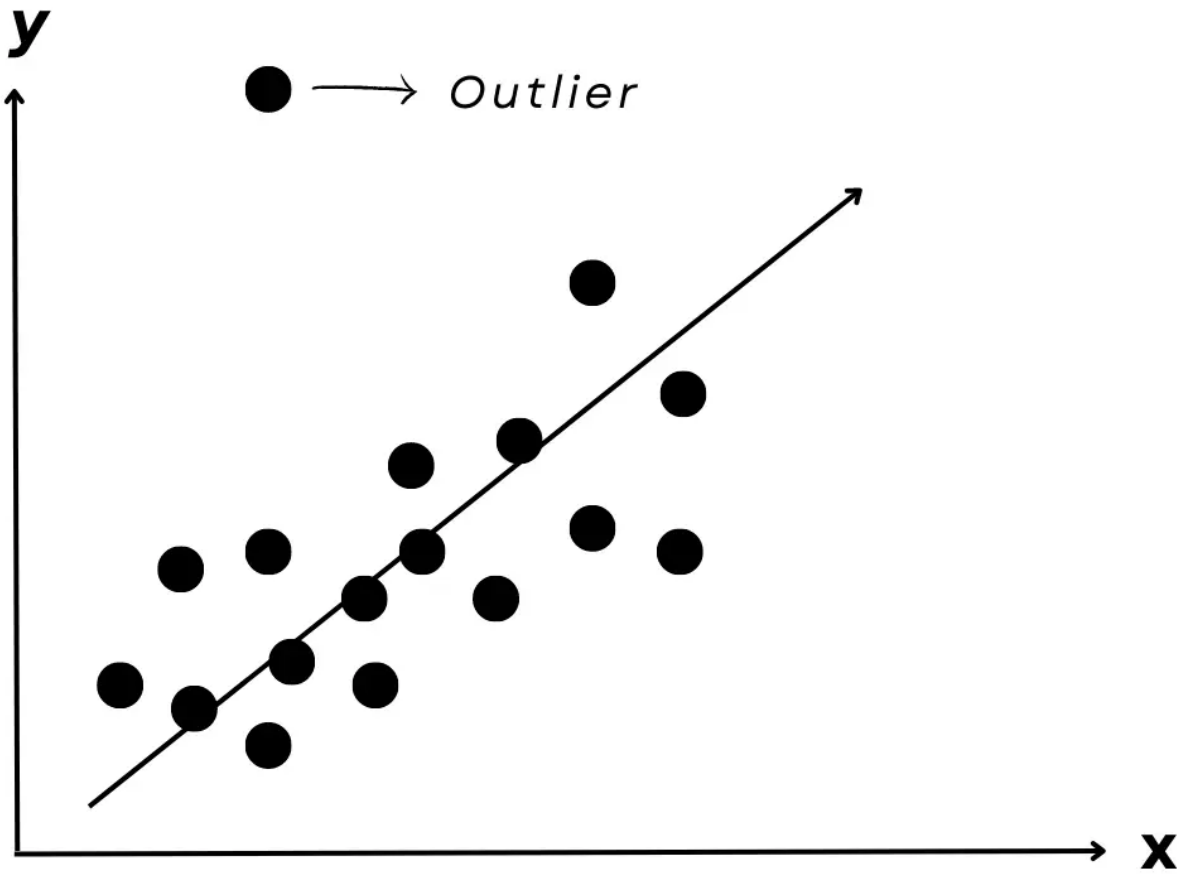
Mode

The mode is another way to find the most common number in a group of numbers.

You look for the value that appears the most often. There won't be a mode if no value shows more than once.

Example: Let's say we have some test scores: {75, 80, 70, 75, 85, 80, 90}. To find the mode, we count how many times each number appears. In this case, 75 and 80 each appear twice, while 70, 85, and 90 each appear once. So the modes of the data set are 75 and 80 because they are the numbers that appear the most often.

Outliers

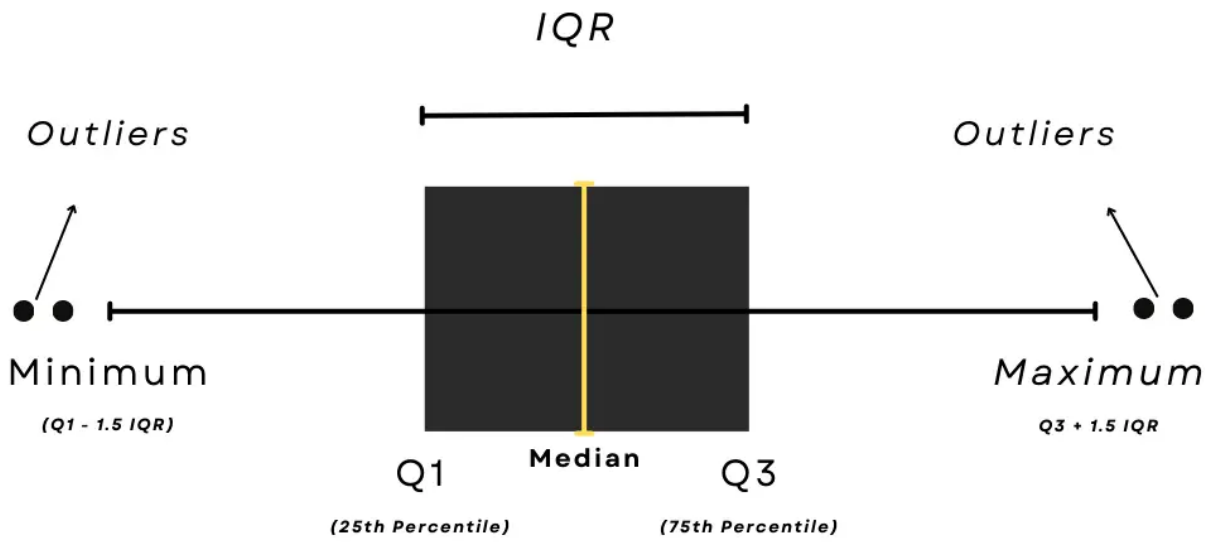


Outliers are extreme values that can skew statistical results.

Measurement errors, data entry errors, or other factors can cause them. Identifying and handling outliers appropriately in statistical analysis is vital to avoid inaccurate results.

In a normal distribution, outliers are typically defined as values that are more than 1.5 times the interquartile range (IQR) above the third quartile or below the first quartile.

Interquartile Range (IQR)



The interquartile range or IQR is the difference between the third and first quartiles, which contains the middle 50% of the data.

Example: Assume a data set follows a normal distribution with a mean of 50 and a standard deviation of 10. If the third quartile is 60 and the first quartile is 40, then the IQR is 20.

Any value that is more than 1.5 times this IQR above the third quartile (i.e., more than 90) or below the first quartile (i.e., less than 10) would be considered an outlier.

Sampling

Sampling is the process of selecting a subset of a population to be analyzed in order to make inferences about the entire population.

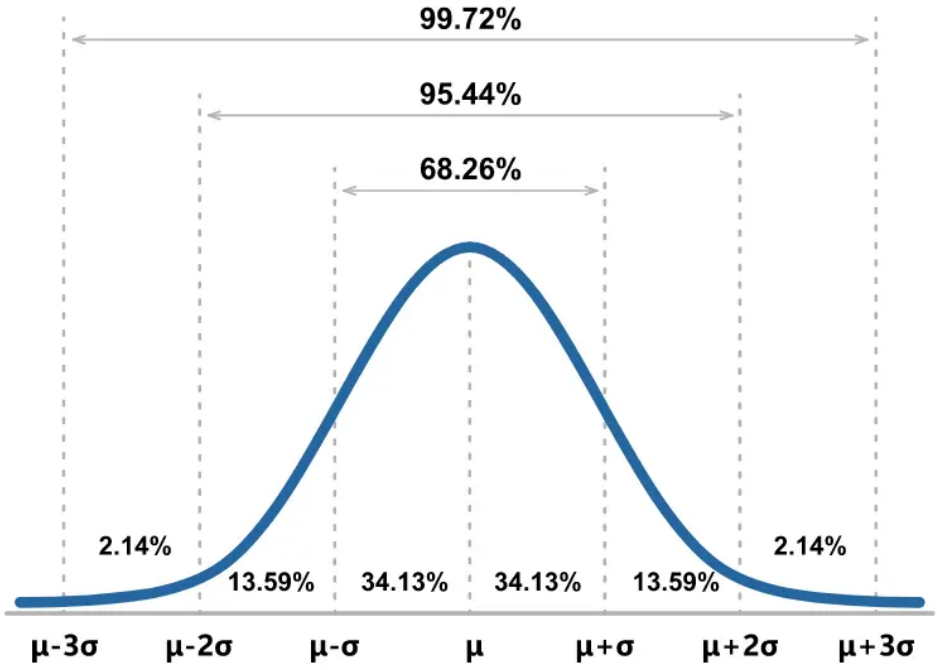
Sampling methods can be probability-based or non-probability based.

Probability-based sampling refers to methods where each member of the population has a known probability of being selected.

The non-probability method is where the selection of individuals is based on convenience, judgment, or quota.

Example: Suppose a market research firm wants to estimate the average income of households in a city. They could select a random sample of households and collect income data from them. By analyzing the income data from the sample, they can make inferences about the average income of the entire population.

Normal Distribution



μ=mean, σ=standard deviation

Normal distribution is a type of probability distribution that is commonly observed in real-world phenomena such as heights, weights, or IQ scores.

Due to its shape, it is also known as a **Gaussian distribution** or a **bell curve**.

A normal distribution's bell curve is defined by the mean and the standard deviation.

The mean is simply the average of the data points.

The standard deviation represents the degree of variability of the data.

The 68-95-99.7 rule is a useful tool for understanding the proportion of data that falls within different ranges of a normal distribution. According to this rule, 68% of the data are within one standard deviation of the mean, 95% are within two standard deviations of the mean, and 99.7% are within three standard deviations of the mean.

The 68-95-99.7 rule is used in many fields, including finance, engineering, and medicine, to help understand the variability and distribution of data.

By using this rule, we can predict the probability of a given event happening within a range of values and make informed decisions based on this.

Example: Imagine the mean height of a group of people is 5 feet 8 inches, and the standard deviation is 2 inches. The 68-95-99.7 rule estimates that approximately 68% of the group will have heights between 5 feet 6 inches and 5 feet 10 inches,

approximately 95% will have heights between 5 feet 4 inches and 6 feet, and approximately 99.7% will have heights between 5 feet 2 inches and 6 feet 2 inches.

Confidence Intervals

A **confidence interval** suggests a range of values that is highly likely to contain a parameter of interest.

For example, suppose you sampled 5 customers who rated your product a mean of 3.5 out of 5 stars. You can use confidence intervals to determine what the population mean (the average rating of all customers) is based on this sample statistic.

Confidence Interval for means (n ≥ 30)

$$(\overline{x} \pm z\frac{\sigma}{\sqrt{n}})$$

Confidence Interval for means (n < 30)

$$(\overline{x} \pm t\frac{s}{\sqrt{n}})$$

Confidence Interval for proportions

$$(\hat{p} \pm z\sqrt{\frac{\hat{p} - (1 - \hat{p})}{n}})$$

Hypothesis Testing

Hypothesis testing is used to determine how likely or unlikely a hypothesis is for a given sample of data. Technically, hypothesis testing is a method in which a sample dataset is compared against the population data.

Here are the steps to performing a hypothesis test:

1. State your null and alternative hypotheses. To reiterate, the null hypothesis typically states that everything is as normally was—that nothing has changed.
2. Set your significance level, the alpha. This is typically set at 5% but can be set at other levels depending on the situation and how severe it is to commit a type 1 and/or 2 error.
3. Collect sample data and calculate sample statistics (z-statistic or t-statistic)
4. Calculate the p-value given sample statistics. Once you get the sample statistics, you can determine the p-value through different methods. The most common methods are the T-score and Z-score for normal distributions.
5. Reject or do not reject the null hypothesis.

Example: Promotional Campaign

Here is an example scenario of hypothesis testing.

A marketing team at a retail store is interested in determining whether a new promotional campaign has had a significant impact on sales.

They randomly selected two groups of customers: one group who saw the promotional ads and another who did not. They then record the sales from each group over the course of a week.

The hypothesis test can be performed as follows:

1. **State the null and alternative hypotheses.** The null hypothesis is that there is no difference in sales between the two groups of customers (i.e., the promotional campaign had no impact on sales). The alternative hypothesis is that there is a difference in sales between the two groups (i.e., the promotional campaign had a significant impact on sales).

$$H0 : \mu1 = \mu2, H1 : \mu1 \neq \mu2$$

2. **Set the significance level, alpha.** Let's assume that the significance level is set at 0.05.

3. **Collect the sample data and calculate the sample statistics.** Let's say that the mean sales for the group who saw the promotional ads were $500, and the mean sales for those who did not were $450.

4. **Calculate the p-value given the sample statistics.** The p-value is the probability of obtaining a sample statistic as extreme or more extreme than the observed one, assuming that the null hypothesis is true. We can use a t-test or z-test depending on the sample size and distribution. Let's assume that we use a t-test with a two-tailed test.
The calculated t-value is:

$$t = \frac{500 - 450}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{500 - 450}{\sqrt{\frac{50^2}{50} + \frac{30^2}{50}}}$$

where:
s1=50
n1=50
s2=30
n2=50

This yields a t-value of 6.06, with a corresponding p-value of 0.0001.

5. **Reject or do not reject the null hypothesis.** Since the p-value is less than the significance level (0.000000004 < 0.05), we reject the null hypothesis and conclude that there is enough evidence to suggest that the promotional campaign had a significant impact on sales.

This hypothesis test can help the marketing team make informed decisions about the effectiveness of their promotional campaign and potentially make adjustments to improve future campaigns.

Z Statistic vs T Statistic

Z Statistics and T Statistics are important to know because they are required for step 3 in the steps to performing a hypothesis test (see above).

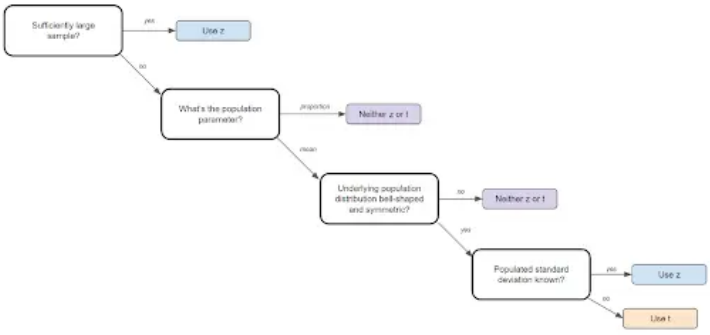
A **Z-test** is a hypothesis test with a normal distribution that uses a **z-statistic**. A z-test is used when you know the population variance or if you don't know the population variance but have a large sample size.

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

A **T-test** is a hypothesis test with a t-distribution that uses a **t-statistic**. You would use a t-test when you don't know the population variance and have a small sample size. You also need the degrees of freedom to convert a t-statistic to a p-value.

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$df = n - 1$$



Example: Blood Pressure

Now let's see how these statistics can be used in a real-life scenario.

A medical researcher wants to test whether a new drug is effective in reducing blood pressure in patients with hypertension.

They recruit a random sample of 30 patients with hypertension and measure their blood pressure before and after taking the drug.

To perform a hypothesis test on the drug's effectiveness, the researcher must determine whether the observed difference in blood pressure before and after taking medicine is statistically significant.

This involves calculating a test statistic, either a z-statistic or a t-statistic, and comparing it to a critical value or calculating a p-value.

If the population standard deviation is known, you can use a z-test to compare the mean blood pressure before and after taking the drug. The null hypothesis is that there is no difference in mean blood pressure before and after taking the medication, while the alternative hypothesis is that there is a difference.

If the population standard deviation is unknown, you must use a t-test. This is because the standard error of the mean must be estimated using the sample standard deviation, which will introduce additional variability into the test statistic. The t-test is more appropriate for small sample sizes, where the sample standard deviation is likely to be a poor estimate of the population standard deviation.

For example, let's say that the sample mean blood pressure before taking the drug is 150 mmHg, and the sample mean blood pressure after taking the drug is 140 mmHg, with a sample standard deviation of 10 mmHg.

The null hypothesis is that the mean difference is zero, and the alternative hypothesis is that the mean difference is less than zero (i.e., the drug effectively reduces blood pressure).

Z-Test

If the population standard deviation is known to be 10 mmHg, the researcher can use a z-test. The z-statistic is:

$$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{(140 - 150)}{\frac{10}{\sqrt{30}}} = -5.48$$

If the significance level is set at 0.05, the critical z-value for a one-tailed test is 1.645 (if it's a right-tailed test) or -1.645 (if it's a left-tailed test). Since it's a drug effectiveness study, we are likely interested in a decrease in blood pressure, which would make it a left-tailed test.

Therefore, the critical z-value is -1.645.

Since the calculated z-statistic is smaller (more negative) than the critical z-value, we reject the null hypothesis and conclude that the drug is effective in reducing blood pressure.

T-Test

If the population standard deviation is unknown, the researcher must use a t-test. The t-statistic is:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{(140 - 150)}{\frac{10}{\sqrt{30}}} = -5.48$$

The result is the same as the z-value, as the population and sample standard deviation is the same in this example.

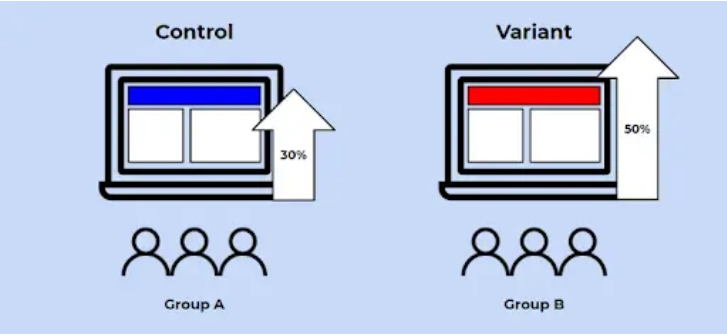
Using the t-distribution with 29 degrees of freedom and a significance level of 0.05, the critical t-value for a one-tailed test is 1.699. Since it's a drug effectiveness study, we are likely interested in a decrease in blood pressure, which would make it a left-tailed test.

Therefore, the critical t-value is -1.699.

Since the calculated t-statistic is smaller (more negative) than the critical t-value, we reject the null hypothesis and conclude that the drug is effective in reducing blood pressure.

This example shows how the choice between a z-test and a t-test depends on the known or estimated population standard deviation and the sample size. It also shows how these tests can be used in hypothesis testing to draw conclusions about the effectiveness of treatment.

A/B Testing



A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Technically speaking, A/B testing is a form of two-sample hypothesis testing, which is a method in determining whether the differences between two samples are statistically significant or not.

The steps to conducting an A/B test are exactly the same as a hypothesis test, except that the p-value is calculated differently depending on the type of A/B test.

An important aspect of A/B testing is choosing the right metric to measure the performance of the two variants. This metric should be relevant to the test goal, such as conversion rate, click-through rate, or revenue.

It's also essential to define the hypothesis before conducting the test, which involves specifying the null and alternative hypotheses and setting the significance level.

Another factor to consider in A/B testing is the sample size. The sample size should be large enough to detect a meaningful difference between the two variants while being small enough to keep the costs and time of the experiment manageable. The sample size calculation should take into account the expected effect size, the significance level, and the statistical power of the test.

A/B testing can be applied in many real-life situations.

Example: A software company may want to test two new feature versions to see which leads to more user engagement.

A fashion retailer may want to test two different product descriptions to see which one leads to more sales.

An e-commerce platform may want to test two different checkout processes to see which one leads to more completed orders.

In all of these cases, A/B testing can help determine which variant performs better and provide valuable insights for making data-driven decisions. By following the steps of hypothesis testing and using appropriate statistical tests, A/B testing can help businesses optimize their products, services, and marketing campaigns and ultimately improve their bottom line.

The type of A/B test that is conducted depends on a number of factors, which I'll go over below:

Note: I won't be covering the math behind these tests but feel free to check out Francesco's article on A/B testing here.

Fisher's Exact Test

The test was developed by Sir Ronald A. Fisher in the early 1900s and is used in various fields, such as genetics, social sciences, and marketing.

The Fisher's test is used when testing against a discrete metric, like clickthrough rates (1 for yes, 0 for no). With a Fisher's test, you can compute the exact p-value, but it is computationally expensive for large sample sizes.

Example: Advertising Campaign

Let's say you are a marketing researcher, and you want to know if a new advertising campaign effectively increases clickthrough rates on a website.

You randomly select two groups of website visitors. One group is shown the new ads, and the other is shown the old ones. You record the clickthrough rates for each group, i.e., the number of visitors who clicked on an ad divided by the total number of visitors in the group.

To test whether the new ads are more effective than the old ads, you can use Fisher's test.

This test will help you determine whether the difference in clickthrough rates between the two groups is statistically significant.

If the p-value is low enough (usually less than 0.05), you can conclude that the new ads are more effective in increasing clickthrough rates than the old ads.

Pearson's Chi-squared Test

Chi-squared tests are an alternative to Fisher's test when the sample size is too large. It is also used to test discrete metrics.

Example: Marketing Research

One real-life example of using Pearson's Chi-squared test is in the field of marketing research.

A company wants to determine if there is an association between a customer's age group (young, middle-aged, or elderly) and their preferred product category (food, clothing, or electronics).

They collect a random sample of 500 customers and record their age group and preferred product category.

To test for association, they use Pearson's Chi-squared test.

They create a contingency table with age group as the rows and product category as the columns. Then they calculate the expected frequencies assuming no association between the two variables.

They then compare the observed frequencies with the expected frequencies using the Chi-squared test statistic and calculate the p-value.

If the p-value is less than the significance level (e.g., 0.05), they reject the null hypothesis of no association and conclude that there is a significant association between a customer's age group and their preferred product category.

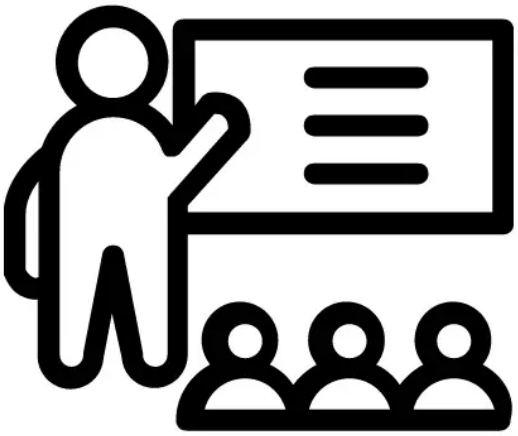
This information can be used by the company to tailor their marketing strategies to specific age groups and product categories.

Student's t-test

I included the t-test and not the z-test because the z-test is typically impractical in reality since the population standard deviations are typically unknown. However, since we can get the sample standard deviation, a t-test is suitable.

It can be used under the conditions that the sample size is large (or the observations are normally distributed), and if the two samples have similar variances.

Example: Teaching Method Testing



An example of using Student's t-test is a study where you want to compare the mean test scores of two groups of students who received different teaching methods.

The first group received traditional teaching methods. The second group received a new teaching method that the researchers hypothesized would lead to better test scores.

To conduct Student's t-test, the researchers randomly assigned students to the two groups and administered the same test to both groups. They then calculated the sample mean and sample standard deviation for each group.

Since the population standard deviation was unknown and the sample size was small, they used Student's t-test. Student's t-test is a specific type of t-test that is commonly used for small sample sizes when the population variance is unknown. It helps you to find out whether the difference in means between the two groups was statistically significant or not, assuming the data are normally distributed and the two groups have similar variances.

By conducting the t-test, the researchers were able to determine whether the new teaching method was more effective than the traditional method in improving test scores.

Welch's Test

Welch's t-test is essentially the same thing as Student's t-test except that it is used when the two samples **do not** have similar variances. In that case, Welch's test can be used.

Example: Salary Testing

An example of using Welch's t-test is when a study wants to compare the salaries of two groups of employees in a company. One group is made up of managers and the other is made up of entry-level employees.

The researchers wanted to see if there was a significant difference in the average salary between the two groups.

However, the salaries of managers are typically more variable than entry-level employees, so the assumption of equal variances required for the Student's t-test was not met.

To address this, the researchers used Welch's t-test, which is appropriate when the two groups being compared have unequal variances.

This allowed them to determine whether there was a statistically significant difference in the average salary between the two groups, despite the unequal variances.

By conducting Welch's t-test, the researchers were able to determine if there was a significant salary gap between the two groups and take appropriate actions based on the findings.

Mann-Whitney U test

The Mann-Whitney test is a non-parametric test and should only be used when all assumptions for all previous tests are violated. For example, if you have a small sample size and the distributions are not normal, a Mann-Whitney test might be suitable.

Example: Medication Testing

An example of using the Mann-Whitney U test is in a study comparing the effectiveness of two different pain medications.

Suppose a small sample size is available, and the data collected is not normally distributed. In this case, the assumption for a t-test may not be met, and the Mann-Whitney U test could be used instead.

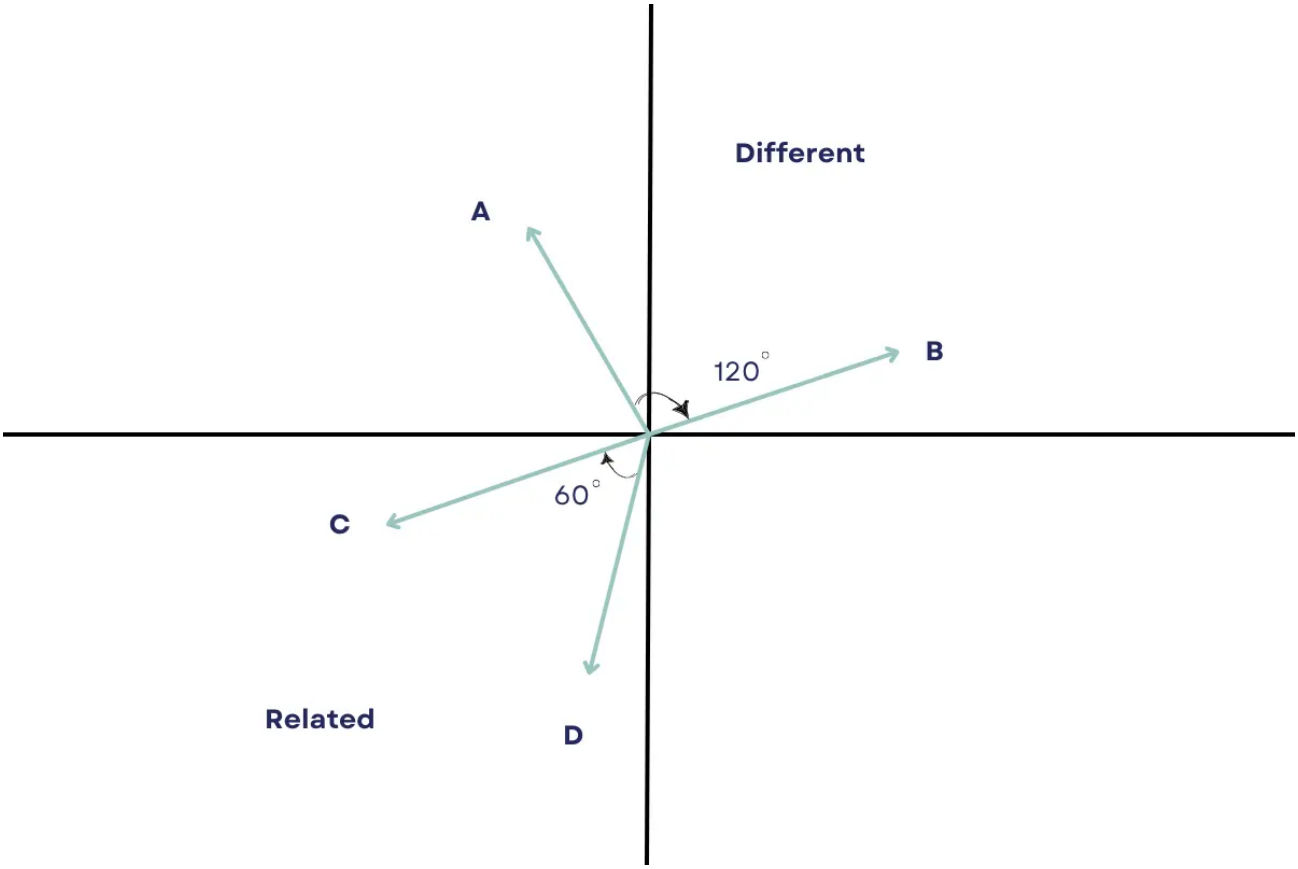
The researchers randomly assign patients to one of two groups to perform the test. The first group receives medication A, while the second group receives medication B.

The researchers then ask the patients to rate their pain on a scale of 1 to 10 after taking the medication. The scores are ranked from lowest to highest, and the test is conducted to determine if there is a significant difference in pain relief between the two medications.

The Mann-Whitney U test allows researchers to compare the effectiveness of two treatments without making assumptions about the normality of the data or the equality of variances between the groups.

To see how to conduct these tests in Python, check out this repository.

Cosine Similarity



Cosine similarity is a way to measure how similar two sets of things are.

The more similar the two sets of things are, the closer they are to each other in this space.

The cosine similarity is calculated by looking at the angle between the two sets of things. If the angle is 0 degrees, that means they are exactly the same. If the angle is 90 degrees, that means they are completely different.

Example: Collaborative Filtering in Netflix

Collaborative filtering often uses cosine similarity to measure the similarity between items or users in a recommendation system.

By using cosine similarity to compare users or items, collaborative filtering can make recommendations based on the preferences of similar users or items.

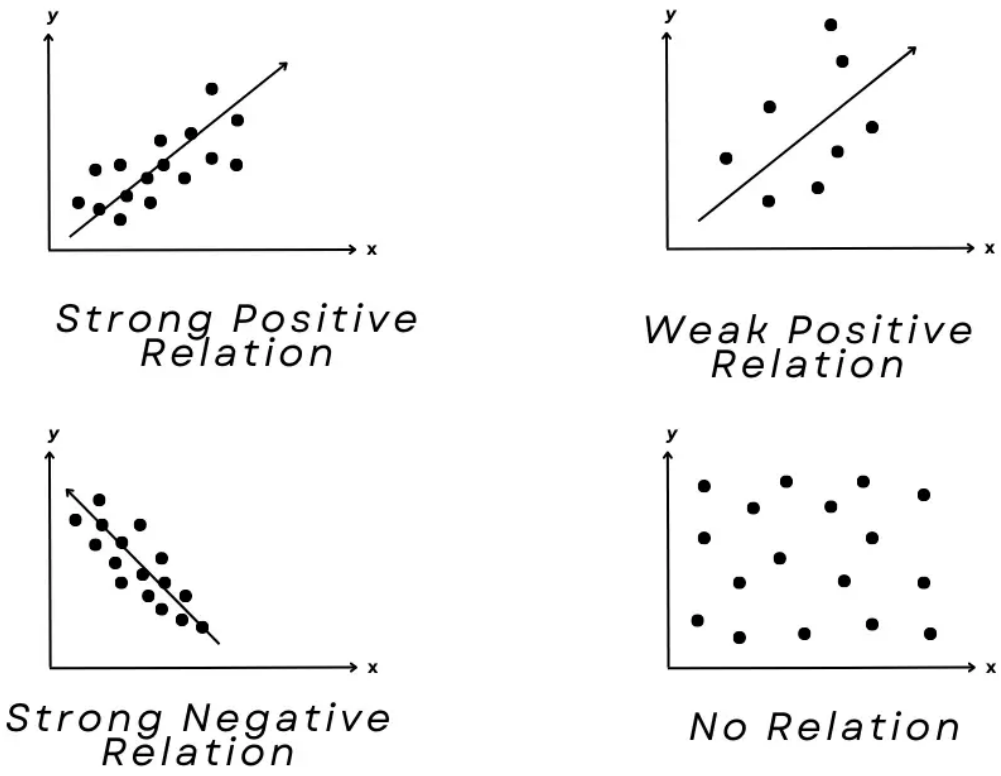
A real-life example of collaborative filtering is the recommendation system used by Netflix. Netflix uses collaborative filtering to recommend movies and TV shows to its users based on their viewing history and the viewing history of similar users.

When a user watches a movie or TV show on Netflix, the system analyzes the user's viewing history. It compares it to the viewing histories of other users who have watched similar content. Based on the similarities between the viewing histories of different users, Netflix recommends other movies and TV shows that the user might enjoy.

For example, if a user has watched several action movies and TV shows, the collaborative filtering algorithm might recommend other action movies and TV shows to the user. Similarly, if several other users with similar viewing histories have enjoyed a particular movie or TV show, the algorithm might recommend that movie or TV show to the user.

By using collaborative filtering to make recommendations, Netflix can personalize its content for each individual user and provide them with a more enjoyable viewing experience.

Correlation



A statistical measure known as correlation analyzes the strength of the relationship or connection between two variables.

It ranges from -1 to 1.

A value of -1 indicates a perfect negative correlation.
A value of 0 indicates no correlation.

A value of 1 indicates a perfect positive correlation.

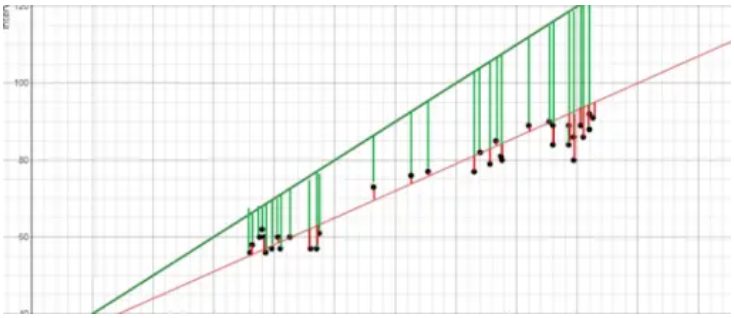
Correlation does not imply causation, but it can be used to identify potential relationships between variables.

Example: Suppose a researcher wants to investigate the relationship between the amount of exercise people do each week and their overall level of physical fitness. They could use a correlation coefficient to determine whether there is a significant relationship between these two variables.

Linear Regression

What is regression?

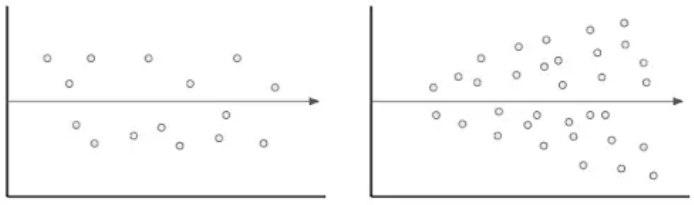
Regression is simply a statistical method for estimating the relationship between one or more independent variables (x) and a dependent variable (y). In simpler terms, it involves finding the 'line of best fit' that represents two or more variables.



The line of best fit is found by minimizing the squared distances between the points and the line of best fit—this is known as **least squares regression**. A **residual** is simply equal to the predicted value minus the actual value.

Residual analysis

A residual analysis can be conducted to assess the quality of a model, and also to identify outliers. A good model should have a homoscedastic residual plot, meaning that the error values are consistent overall.



Homoscedastic plot (left) vs heteroscedastic plot (right)

Variable Selection

Two very simple and common approaches to variables selection are **backward elimination** (removing one variable at a time) or **forward selection** (adding one variable at a time).

You can assess whether a variable is significant in a model by calculating its p-value. Generally speaking, a good variable has a p-value of less than or equal to 0.05.

Model Evaluation

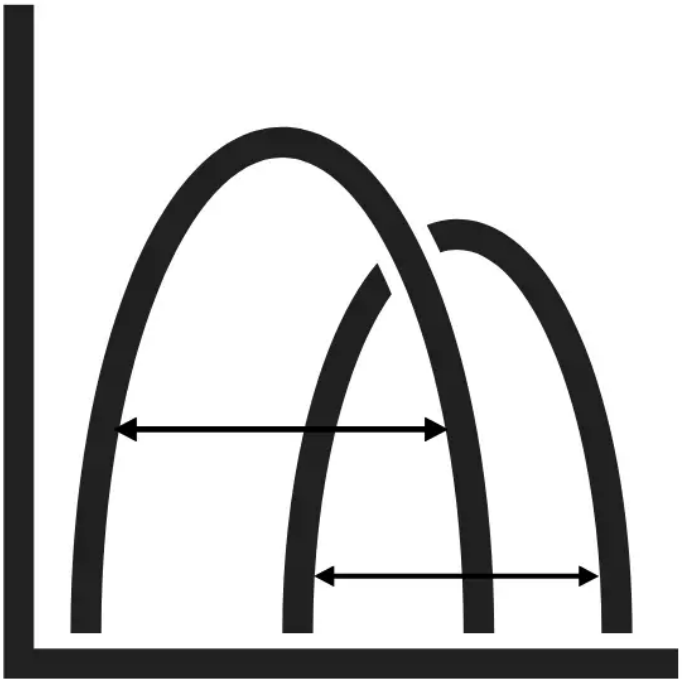
To evaluate a regression model, you can calculate its **R-squared**, which tells us how much of the variability in the data that the model accounts for. For example, if a model has an R-squared of 80%, then 80% of the variation in the data can be explained by the model.

The **adjusted R-squared** is a modified version of r-squared that adjusts for the number of predictors in the model; it increases if the new term improves the model more than would be expected by chance and vice versa.

3 Common pitfalls to avoid

1. **Overfitting**: Overfitting is an error where the model 'fits' the data too well, resulting in a model with high variance and low bias. As a consequence, an overfit model will inaccurately predict new data points even though it has a high accuracy on the training data. This typically happens when there are too many independent variables in the model.
2. **Collinearity**: This is when two independent variables in a model are correlated, which ultimately reduces the accuracy of the model.
3. **Confounding variables**: a confounding variable is a variable that isn't included in the model but affects both the independent and dependent variables.

ANOVA (Analysis of Variance)



ANOVA is a statistical test used to analyze the differences between two or more groups of data.

It helps to determine whether the means of the groups are significantly different from each other or not.

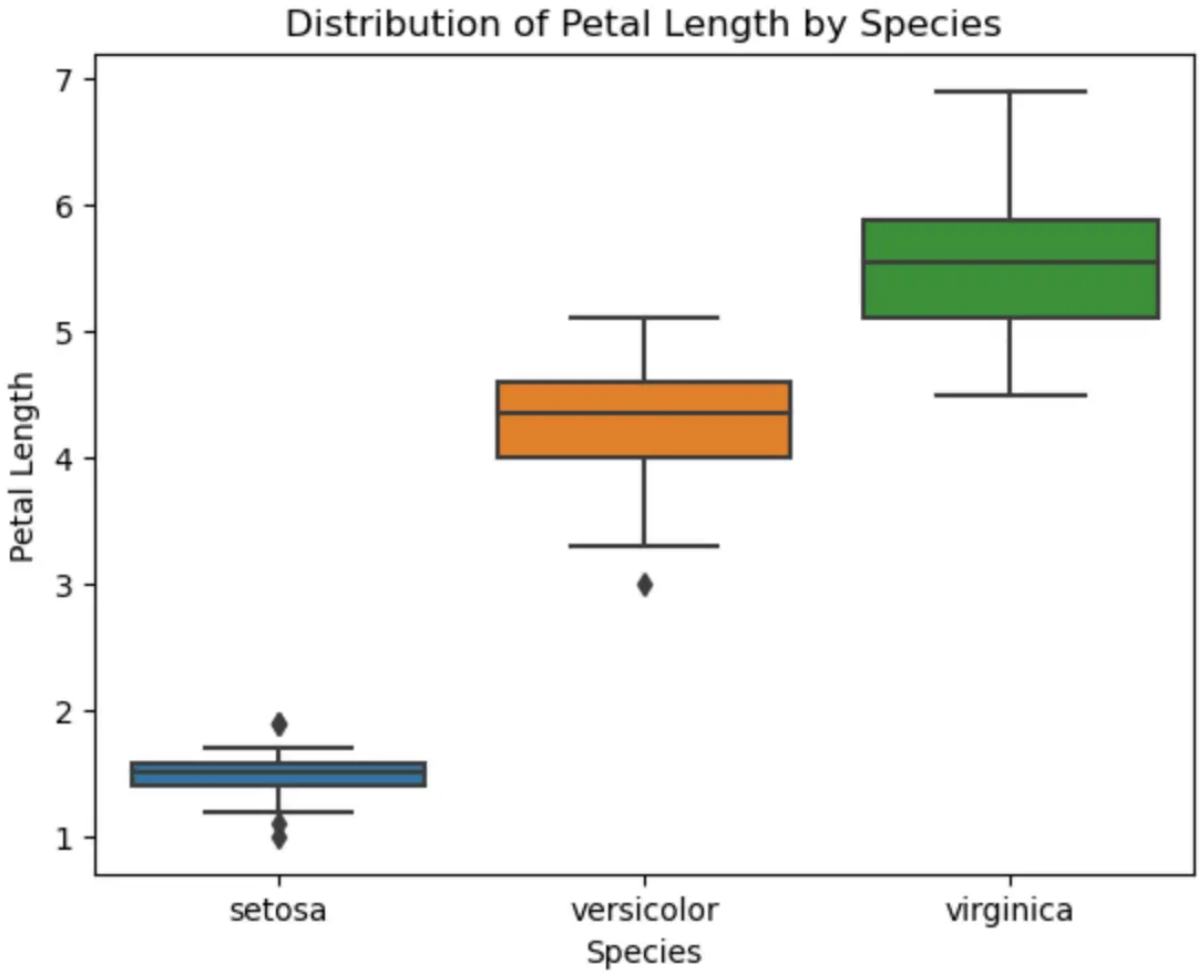
The test compares the variance between the groups to the variance within the groups, and a significant result indicates that at least one group differs significantly from the others.

The formula for ANOVA is:

$$F = \frac{variance\ between\ groups}{variance\ within\ groups}$$

To evaluate the ANOVA score, you need to calculate the F-statistic and compare it to the critical value from an F-distribution table with a certain level of significance (alpha). Suppose the calculated F-statistic is greater than the critical value. In that case, you can reject the null hypothesis and conclude that there is a significant difference between at least two of the groups.

Example: Species Comparison



In this example, we performed an ANOVA test in Python to analyze the differences in the petal length of three different species of iris: setosa, versicolor, and virginica.

We first loaded the iris dataset using the seaborn library, which contains information about the petal length, width, and other variables for each species of iris.

We then separated each species' petal length data into three sets: setosa_petal_length, versicolor_petal_length, and virginica_petal_length.

We performed the ANOVA test using the **f_oneway()** function from the **scipy.stats** library. The test calculated the F-statistic and p-value for the analysis, which provided information about the statistical significance of the differences between the means of the petal length for the three species.

Here is the code.

```
from scipy.stats import f_oneway

# load the 'iris' dataset
iris = sns.load_dataset('iris')

# perform the ANOVA test
setosa_petal_length = iris[iris['species'] == 'setosa']['petal_length']
versicolor_petal_length = iris[iris['species'] == 'versicolor']['petal_length']
virginica_petal_length = iris[iris['species'] == 'virginica']['petal_length']

f_statistic, p_value = f_oneway(setosa_petal_length, versicolor_petal_length, virginica_petal_length)

# print the results
print('F-statistic:', f_statistic)
print('p-value:', p_value)
```

Here is the output.

```
F-statistic: 1180.161182252981
p-value: 2.8567766109615584e-91
```

The resulting F-statistic was 1180.16, which is a large value indicating that there is a significant difference between the means of the petal length for the three species.

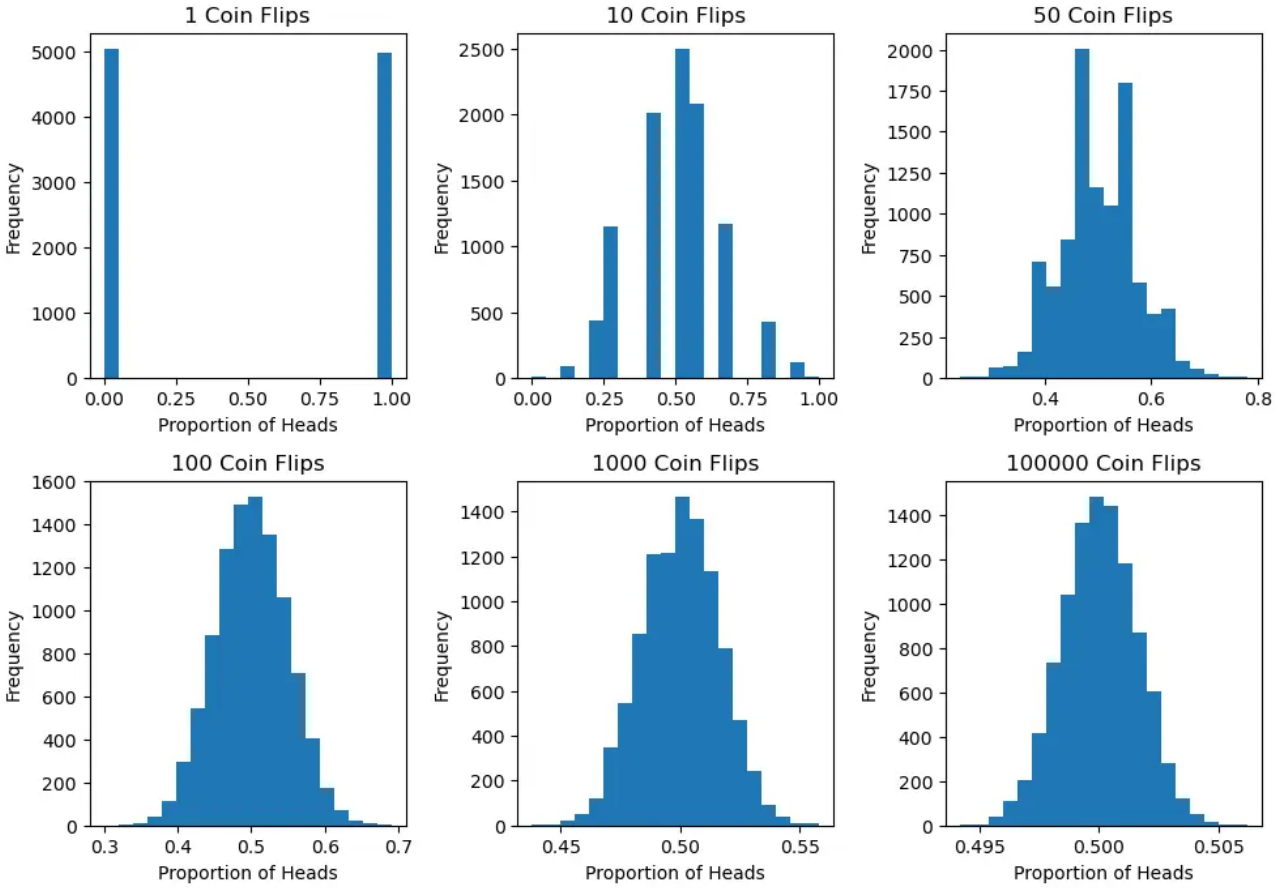
The p-value was 2.86e-91, which is a very small value close to zero, indicating that the probability of observing such a large F-statistic by chance alone is very low.

Therefore, we can reject the null hypothesis that there is no significant difference between the three species' means of petal length. We conclude that there is a significant difference between the means of the petal length for at least one pair of species.

Central Limit Theorem

The central limit theorem states that if a random sample is drawn from any population, regardless of its distribution, the distribution of the sample means will be approximately normally distributed as the sample size increases.

Example: Coin Flip



Suppose we flip a fair coin multiple times and record the number of times it lands heads up.

We repeat this process many times and record the average number of heads for each set of coin flips. In that case, the distribution of these averages will approximate a normal distribution as the number of coin flips increases.

For example, if we flip a coin once, the probability of getting heads might be one. If we flip the coin 10 times, the probability of getting heads will vary, which is not normally distributed.

However, if we repeat this process many times and record the average number of heads for each set of coin flips, the distribution of these averages will be approximately normal.

As we continue to increase the number of coin flips, the distribution of the sample means will continue to approach a normal distribution, as predicted by the central limit theorem.

This is a useful concept in statistics because it allows us to make inferences about a population based on a sample, even if we don't know the distribution of the population.

By using the central limit theorem, we can assume that the sample means will be normally distributed and use this information to perform hypothesis tests or construct confidence intervals.

Probability Rules

There are several fundamental properties and four probability rules that you should know. These probability rules serve as the foundation for more complex (but still fundamental) equations, like Bayes Theorem, which will be covered after.

Note: this does not review joint probability, union of events, or intersection of events. Review them beforehand if you do not know what these are.

Basic Properties

1. Every probability is between 0 and 1.
2. The sum of the probabilities of all possible outcomes equals 1.
3. If an event is impossible, it has a probability of 0.
4. Conversely, certain events have a probability of 1.

The Four Probability Rules

1. Addition Rule

The addition rule in probability says that if there are two events, A and B, which cannot happen at the same time, then the probability of either event happening is equal to the total of their probabilities.

This means that if the events are independent, we can add their probabilities to calculate the overall probability of either event happening.

Here is the formula.

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example: Consider a single toss of a fair coin. The probability of getting heads or tails is:

$$P(heads\ or\ tails) = \frac{1}{2} + \frac{1}{2} = 1$$

2. Complementary Rule

The complementary rule of probability says that when the probability of event A is represented by p, then the probability of event A not happening is represented by 1-p.

$$P(not\ A) = 1 - P(A)$$
$$P(\neg A) = 1 - P(A)$$

Example: The probability of getting a 2 on a standard six-sided die is 1/6. The probability of getting other than 2 is:

$$P(not\ 2) = 1 - \frac{1}{6} = \frac{5}{6}$$

3. Conditional Rule

Bayes theorem helps us calculate the conditional probability of the given events. Now let's look at the formula and the example to make it clearer.

$$P(B\ given\ A) = \frac{P(A\ and\ B)}{P(A)}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Example: A medical test for a certain disease has a false positive rate of 5% and a false negative rate of 10%. If 1% of the population has the disease, what is the probability that a person who tests positive actually has the disease?

Let's explain this by giving numbers to these rates. We have a population of 10,000 people, and 1% or 100 people have the disease, while the other 9,900 do not.

Out of the 100 people who have the disease, 90 will test positive, and 10 will test negative.

Out of the 9,900 people who do not have the disease, 495 will test positive even though they do not have the disease, and 9,405 will test negative.

Using the formula

$$P(B\ given\ A) = \frac{P(A\ and\ B)}{P(A)}$$

we can calculate the probability that a person who tests positive actually has the disease as follows:

- P(B given A) = P(positive test and has the disease) / P(positive test)

$$P(B\ given\ A) = \frac{P(positive\ test\ and\ has\ the\ disease)}{P(positive\ test)}$$

- P(positive test and has the disease) = 90

$$P(positive\ test\ and\ has\ the\ disease) = 90$$

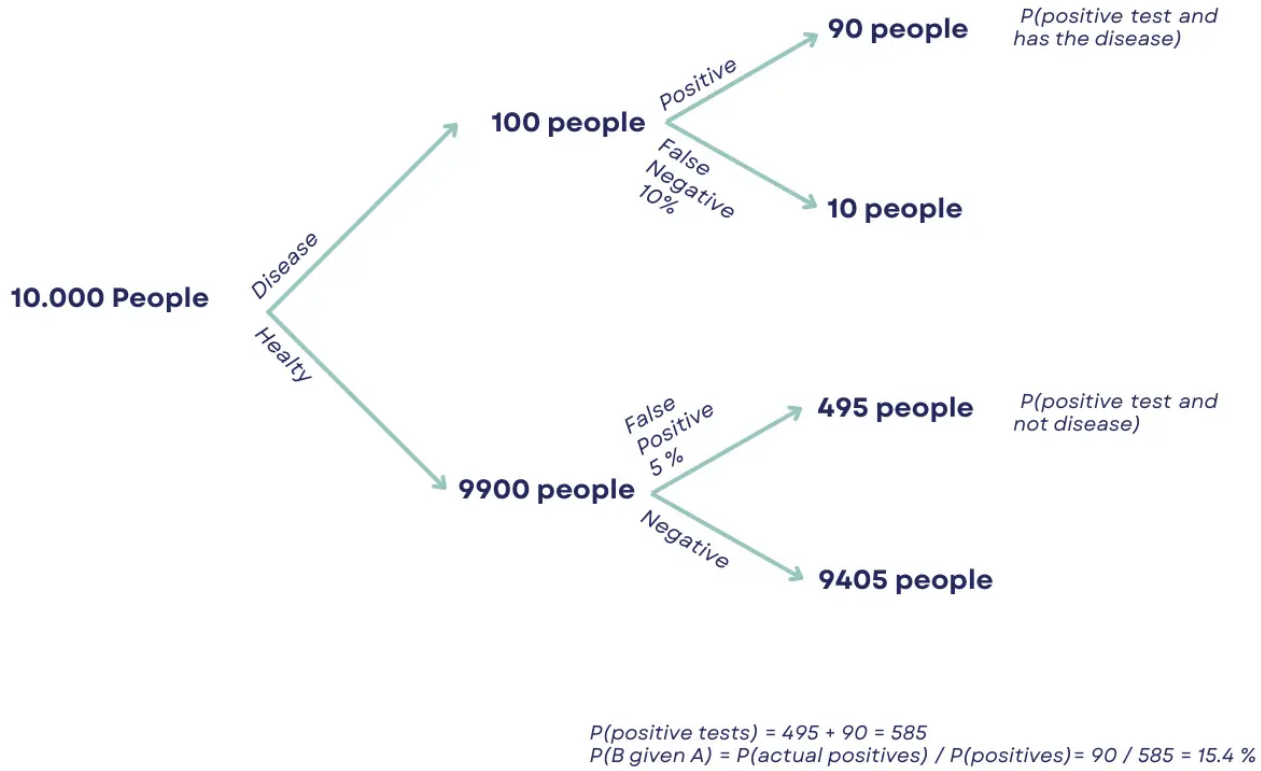
- P(positive test) = (90 + 495) = 585

$$P(positive\ test) = (90 + 495) = 585$$

- P(B given A) = 90 / 585 = approximately 0.154 or 15.4%

$$P(B\ given\ A) = \frac{90}{585} = 0.154 = 15.4\%$$

Therefore, the probability that a person who tests positive actually has the disease is approximately 15.4%.



4. Multiplication Rule

If events A and B are not related to each other, they are called independent events. In such cases, the probability of both events happening together is equal to the product of their individual probabilities.

$$P(A\ and\ B) = P(A) \cdot P(B\ given\ A)$$

$$P(A \cap B) = P(A) \cdot P(B|A)$$

Example: Let's assume that you are drawing two cards from a standard deck of 52 cards. The probability of drawing an ace on the first draw is 4/52. The probability of drawing another ace on the second draw (assuming you did not replace the first card) is 3/51. The probability of drawing two aces is the product of these probabilities is:

$$P(A\ and\ B) = \frac{4}{52} \cdot \frac{3}{51} = \frac{1}{221}$$

To practice using these equations, you can check out this resource.

Bayes Theorem

Bayes theorem is a conditional probability statement, essentially it looks at the probability of one event (B) happening given that another event (A) has already happened. The formula is as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- P(A) is the **prior**, which is the probability of A being true.
- P(B|A) is the **likelihood**, the probability of B being true given A.
- P(B) is the **marginalization** or the **normalizing constant**
- P(A|B) is the **posterior**.

What you'll find in a lot of practice problems is that the normalizing constant, P(B), is not given. In these cases, you can use the alternative version of Bayes Theorem, which is below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

To get a better understanding of Bayes Theorem and follow along with some practice questions, check out here.

Combinations and Permutations

Combinations and permutations are two slightly different ways that you can select objects from a set to form a subset. Permutations take into consideration the order of the subset whereas combinations do not.

Combinations and permutations are extremely important if you're working on network security, pattern analysis, operations research, and more. Let's review what each of the two are in further detail:

Permutations

Definition: A permutation of n elements is any arrangement of those n elements in a **definite order**. There are n factorial (n!) ways to arrange n elements. *Note the bold: order matters!*

The number of permutations of n things taken r-at-a-time is defined as the number of r-tuples that can be taken from n different elements and is equal to the following equation:

$$P_{n,r} = \frac{n!}{(n-r)!}$$

Example Question: How many permutations does a license plate have with 6 digits?

$$P_{9,6} = \frac{9!}{(9-6)!} = 60480$$

Combinations

Definition: The number of ways to choose r out of n objects where **order doesn't matter**.

The number of combinations of n things taken r-at-a-time is defined as the number of subsets with r elements of a set with n elements and is equal to the following equation:

$$C_r^n = \frac{n!}{(n-r)!r!}$$

Example Question: How many ways can you draw 6 cards from a deck of 52 cards?

$$C_6^{52} = \frac{52!}{(52-6)!6!} = 20358520$$

Note that these are very very simple questions and that it can get much more complicated than this, but you should have a good idea of how it works with the examples above!

Summary

In this article, we covered all the important statistics concepts you'll most likely get at any data science interview.

Statistics is one of the pillars of data science. There's no serious data science project that doesn't require the application of at least several statistics methods we discussed.

In real life, you'll do statistical analysis, and build, train, and evaluate models in real life. Techniques available for sampling and hypothesis testing are crucial for these data science tasks, so no wonder they come up at the interviews very often.

To make the understanding easier, we backed up most of the theory with practical examples taken from experience in data science.

Statistics Cheat Sheet

Population

The entire group one desires information about

Sample

A subset of the population taken because the entire population is usually too large to analyze
Its characteristics are taken to be representative of the population

Mean

Also called the arithmetic mean or average
The sum of all the values in the sample divided by the number of values in the sample/population
μ is the mean of the population; $\bar{x}$ is the mean of the sample

Median

The value separating the higher half of a sample/population from the lower half
Found by arranging all the values from lowest to highest and taking the middle one (or the mean of the middle two if there are an even number of values)

Variance

Measures dispersion around the mean
Determined by averaging the squared differences of all the values from the mean
Variance of a population is σ^2

$$\sigma^2 = \frac{\sum (x-\mu)^2}{n}$$

Can be calculated by subtracting the square of the mean from the average of the squared scores:

$$\sigma^2 = \frac{\sum x^2}{n} - \mu^2$$

Variance of a sample is s^2; note the *n-1*

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

Can be calculated by:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Standard Deviation

Square root of the variance
Also measures dispersion around the mean but in the same units as the values (instead of square units with variance)
σ is the standard deviation of the population and *s* is the standard deviation of the sample

Standard Error

An estimate of the standard deviation of the sampling distribution—the set of all samples of size *n* that can be taken from a population
Reflects the extent to which a statistic changes from sample to sample

For a mean, $\frac{s}{\sqrt{n}}$

For the difference between two means,

Assuming equal variances $\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$; unequal variances $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

T-test

One-Sample

Tests whether the mean of a normally distributed population is different from a specified value

Null Hypothesis (H_0): states that the population mean is equal to some value (μ_0)
Alternative Hypothesis (H_a): states that the mean does not equal/is greater than/is less than μ_0
t-statistic: standardizes the difference between $\bar{x}$ and μ_0

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$ Degrees of freedom (df) = *n-1*

Read the table of t-distribution critical values for the p-value (probability that the sample mean was obtained by chance given μ_0 is the population mean) using the calculated t-statistic and degrees of freedom.

- H_a: $\mu > \mu_0$ → the t-statistic is likely positive; read table as given
- H_a: $\mu < \mu_0$ → the t-statistic is likely negative; the t-distribution is symmetrical so read the probability as if the t-statistic were positive
- Note: if the t-statistic is of the 'wrong' sign, the p-value is 1 minus the *p* given in the chart
- H_a: $\mu \neq \mu_0$ → read the p-value as if the t-statistic were positive and double it (to consider both less than and greater than)

If the p-value is less than the predetermined value for significance (called α and is usually 0.05), reject the null hypothesis and accept the alternative hypothesis.

Example:
You are experiencing hair loss and skin discoloration and think it might be because of selenium toxicity. You decide to measure the selenium levels in your tap water once a day for one week. Your results are given below. The EPA maximum contaminant level for safe drinking water is 0.05 mg/L. Does the selenium level in your tap water exceed the legal limit (assume α=0.05)?

Day	Selenium mg/L
1	0.051
2	0.0505
3	0.049
4	0.0516
5	0.052
6	0.0508
7	0.0506

H_0: $\mu = 0.05$; H_a: $\mu > 0.05$
Calculate the mean and standard deviation of your sample:

$$\bar{x} = 0.0508$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{(0.051 - 0.0508)^2 + (0.0505 - 0.0508)^2 + etc...}{6} = 9.15 \times 10^{-7}$$

$$s = \sqrt{s^2} = 9.56 \times 10^{-4}$$

The t-statistic is: $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.0508 - 0.05}{\frac{9.56 \times 10^{-4}}{\sqrt{7}}} = 2.17$ and the degrees of freedom are *n-1* = 7-1 = 6

Looking at the t-distribution of critical values table, 2.17 with 6 degrees of freedom is between *p*=0.05 and *p*=0.025. This means that the p-value is less than 0.05, so you can reject H_0 and conclude that the selenium level in your tap water exceeds the legal limit.

T-test

Two-Sample

Tests whether the means of two populations are significantly different from one another

Paired

Each value of one group corresponds directly to a value in the other group; ie: before and after values after drug treatment for each individual patient

Subtract the two values for each individual to get one set of values (the differences) and use $\mu_0 = 0$ to perform a one-sample t-test

Unpaired

The two populations are independent

H_0: states that the means of the two populations are equal ($\mu_1 = \mu_2$)

H_a: states that the means of the two populations are unequal or one is greater than the other ($\mu_1 \neq \mu_2$, $\mu_1 > \mu_2$, $\mu_1 < \mu_2$)

t-statistic:

assuming equal variances: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ assuming unequal variances: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}}$

degrees of freedom = $(n_1\text{-}1)+(n_2\text{-}1)$

Read the table of t-distribution critical values for the p-value using the calculated t-statistic and degrees of freedom. Remember to keep the sign of the t-statistic clear (order of subtracting the sample means) and to double the p-value for an H_a of $\mu_1 \neq \mu_2$.

Example:

Consider the lifespan of 18 rats. 12 were fed a restricted calorie diet and lived an average of 700 days (standard deviation=21 days). The other 6 had unrestricted access to food and lived an average of 668 days (standard deviation=30 days). Does a restricted calorie diet increase the lifespan of rats (assume α=0.05)?

μ_1=700, s_1=21, n_1=12; μ_2=668, s_2=30, n_2=6

H_0: $\mu_1=\mu_2$

H_a: $\mu_1>\mu_2$ (because we are only asking if a restricted calorie diet increases lifespan)

We cannot assume that the variances of the two populations are equal because the different diets could also affect the variability in lifespan.

The t-statistic is: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}} = \dfrac{700 - 668}{\sqrt{\dfrac{21^2}{12} + \dfrac{30^2}{6}}} = 2.342$

Degrees of freedom = $(n_1\text{-}1)+(n_2\text{-}1)$ = (12-1)+(6-1)=16

From the t-distribution table, the p-value falls between 0.01 and 0.02, so we do reject H_0. The restricted calorie diet does increase the lifespan of rats.

Chi-Square Test

For Goodness of Fit

Checks whether or not an observed pattern of data fits some given distribution

H_0: the observed pattern fits the given distribution

H_a: the observed pattern does not fit the given distribution

The chi-square statistic is: $\chi^2 = \sum \dfrac{(O-E)^2}{E}$ (*O* is the observed value and *E* is the expected value)

Degrees of freedom = number of categories in the distribution – 1

Get the p-value from the table of χ^2 critical values using the calculated χ^2 and df values. If the p-value is less than α, the observed data does not fit the expected distribution. If $p>\alpha$, the data likely fits the expected distribution

***Example 1*:**

You breed puffskeins and would like to determine the pattern of inheritance for coat color and purring ability. Puffskeins come in either pink or purple and can either purr or hiss. You breed a purebred, pink purring male with a purebred, purple hissing female. All individuals of the F_1 generation are pink and purring. The F_2 offspring are shown below. Do the alleles for coat color and purring ability assort independently (assume α=0.05)?

Pink and Purring	Pink and Hissing	Purple and Purring	Purple and Hissing
143	60	55	18

Independent assortment means a phenotypic ratio of 9:3:3:1, so:

H_0: the observed distribution of F_2 offspring fits a 9:3:3:1 distribution

H_a: the observed distribution of F_2 offspring does not fit a 9:3:3:1 distribution

The expected values are:

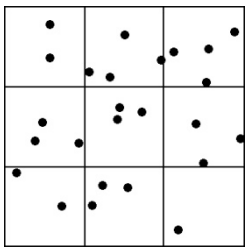
Pink and Purring	Pink and Hissing	Purple and Purring	Purple and Hissing
155.25	51.75	51.75	17.25

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(143-155.25)^2}{155.25} + \frac{(60-51.75)^2}{51.75} + \frac{(55-51.75)^2}{51.75} + \frac{(18-17.25)^2}{17.25} = 2.519$$

df=4-1=3

From the table of χ^2 critical values, the p-value is greater than 0.25, so the alleles for coat color and purring ability do assort independently in puffskeins.

Example 2:



You are studying the pattern of dispersion of king penguins and the diagram on the right represents an area you sampled. Each dot is a penguin. Do the penguins display a uniform distribution (assume α=0.05)?

H_0: there is a uniform distribution of penguins

H_a: there is not a uniform distribution of penguins

There are a total of 25 penguins, so if there is a uniform distribution, there should be 2.778 penguins per square. There actual observed values are 2, 4, 4, 3, 3, 3, 2, 3, 1, so the χ^2 statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(1-2.778)^2}{2.778} + 2\left(\frac{(2-2.778)^2}{2.778}\right) + 4\left(\frac{(3-2.778)^2}{2.778}\right) + 2\left(\frac{(4-2.778)^2}{2.778}\right) = 2.72$$

df=9-1=8

From the table of χ^2 critical values, the p-value is greater than 0.25, so we do not reject H_0. The penguins do display a uniform distribution.

Chi-Square Test

For Independence

Checks whether two categorical variables are related or not (independence)

H_0: the two variables are independent

H_a: the two variables are not independent

Does not make any assumptions about an expected distribution

The observed values ($\#_1$, $\#_2$, $\#_3$, and $\#_4$) are usually presented as a table. Each row is a category of variable 1 and each column is a category of variable 2.

		Variable 1		Totals
		Category *x*	Category *y*	
Variable 2	Category *a*	$\#_1$	$\#_2$	$\#_1+\#_2$
	Category *b*	$\#_3$	$\#_4$	$\#_3+\#_4$
Totals		$\#_1+\#_3$	$\#_2+\#_4$	$\#_1+\#_2+\#_3+\#_4$

The proportion of category *x* of variable 1 is the number of individuals in category *x* divided by the total number of individuals $\left(\frac{\#_1+\#_3}{\#_1+\#_2+\#_3+\#_4}\right)$. Assuming independence, the expected number of individuals that fall within category *a* of variable 2 is the proportion of category *x* multiplied by the number of individuals in category *a* $\left(\frac{\#_1+\#_3}{\#_1+\#_2+\#_3+\#_4}\right)(\#_1+\#_2)$. Thus, the expected value is:

$$E = \frac{(\#_1+\#_3)(\#_1+\#_2)}{\#_1+\#_2+\#_3+\#_4} = \frac{(row\ total)(column\ total)}{grand\ total}$$

Degrees of freedom = (*r*-1)(*c*-1) where *r* is the number of rows and *c* is the number of columns

The chi-square statistic is still $\chi^2 = \sum \frac{(O-E)^2}{E}$

Read the p-values from the table of χ^2 critical values.

Example:

Given the data below, is there a relationship between fitness level and smoking habits (assume α=0.05)?

	Fitness Level				
	Low	Medium-Low	Medium-High	High	
Never smoked	113	113	110	159	495
Former smokers	119	135	172	190	616
1 to 9 cigarettes daily	77	91	86	65	319
≥ 10 cigarettes daily	181	152	124	73	530
	490	491	492	487	1960

H_0: fitness level and smoking habits are independent
H_a: fitness level and smoking habits are not independent
First, we calculate the expected counts. For the first cell, the expected count is:

$$E = \frac{(row\ total)(column\ total)}{grand\ total} = \frac{(495)(490)}{1960} = 123.75$$

	Fitness Level			
	Low	Medium-Low	Medium-High	High
Never smoked	123.75	124	124.26	122.99
Former smokers	154	154.31	154.63	153.06
1 to 9 cigarettes daily	79.75	79.91	80.08	79.26
≥ 10 cigarettes daily	132.5	132.77	133.04	131.69

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(113-123.75)^2}{123.75} + \frac{(113-124)^2}{124} + \frac{(110-124.26)^2}{124.26} + etc... = 91.73$$

df=(*r*-1)(*c*-1)=(4-1)(4-1)=9

From the table of χ^2 critical values, the p-value is less than 0.001, so we reject H_0 and conclude that there is a relationship between fitness level and smoking habits.

Type I error

The probability of rejecting a true null hypothesis
Equals α

Type II error

The probability of failing to reject a false null hypothesis

Probability

Joint Probability

The probability of events A and B occurring
$P(A \text{ and } B) = P(A) \times P(B)$ when events A and B are independent

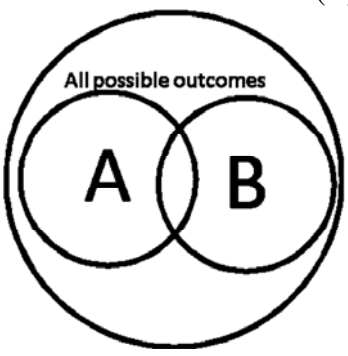
Union of Events

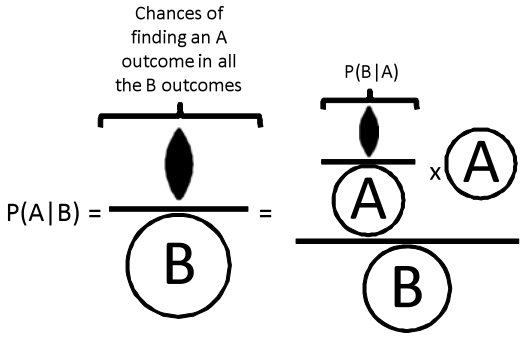
The probability of either event A or event B occurring
$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Conditional Probability

The probability of event A occurring given that event B has occurred

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{or} \quad P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$





Example 1:
Assume that eye color is an autosomally inherited trait controlled by one gene with two alleles. Brown is dominant to blue. A brown-eyed man with genotype Bb and a blue-eyed woman have three children. The first has blue eyes. What is the probability that all three children have blue eyes?

Without considering the first child, the probability that the couple has three children with blue eyes is
$0.5 \times 0.5 \times 0.5 = 0.125 = P(A \text{ and } B) = P(2 \text{ children} = \text{bb and 1st child bb})$
With his parents, the probability that the 1st child is bb is: $P(B) = P(1\text{st child} = \text{bb}) = 0.5$
Therefore, $P(2 \text{ children} = \text{bb} \mid \text{1st child bb}) = P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)} = \dfrac{0.125}{0.5} 0.25$

Example 2:
Based on an analysis of her pedigree, it is determined that a woman has a 70% chance of being Zz and a 30% chance of being ZZ for a sex-linked trait, where Z is dominant to z. If she now has a son with the Z phenotype, what is the probability of her being Zz?

We're looking for: P(W=Zz|S=Z)
But it's hard to find P(W=Zz and S=Z) because the two events are not independent. Instead, let us use:
$P(A \mid B) = \dfrac{P(B \mid A) \times P(A)}{P(B)}$
$P(S = Z \mid W = Zz) = 0.5$ (50% chance of passing on the Z allele)
$P(W = Zz) = 0.7$ (given)
$P(S = Z) = (0.7 \times 0.5) + (0.3 \times 1) = 0.65$ (son can be Z from the woman being either Zz or ZZ)
$P(W = Zz \mid S = Z) = \dfrac{0.5 \times 0.7}{0.65} = 0.538$

Multiple Experiments

Binomial distribution

For when you are not concerned about the order of the events, only that they occur

$$P(X = m) = \frac{n! \times p^m \times (1-p)^{(n-m)}}{m! \times (n-m)!}$$

for *m* outcomes of event X in *n* total trials with *p*=probability of X occurring once

Example:
What is the probability that a couple has one boy out of five children?

$$P(1 \text{ boy of 5 children}) = \frac{5! \times 0.5^1 \times 0.5^4}{1! \times (4)!} = 0.15625$$

Poisson distribution

The binomial distribution works for a small number of trials but as *n* gets too large, the factorials become unwieldy.
The Poisson distribution is an estimate of the binomial distribution for large *n*.

$$P(X = m) = \frac{e^{-np} \times (n \times p)^m}{m!}$$

Note: *np* is also known as the number of expected outcomes for event X

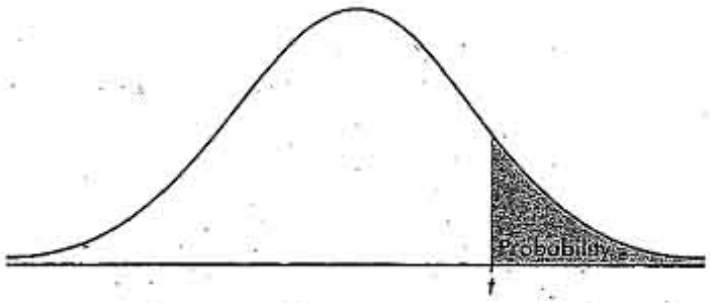


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

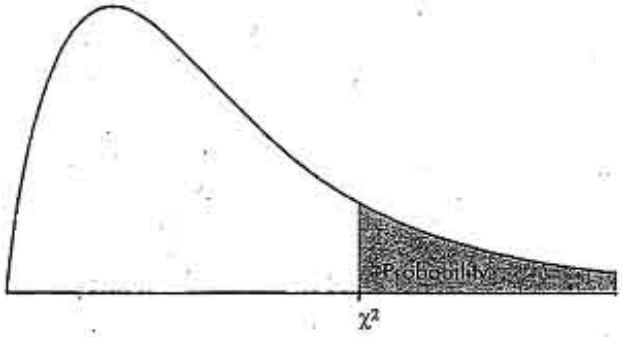


TABLE C: χ^2 CRITICAL VALUES

	Tail probability p										
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4

The following table shows general guidelines for choosing a statistical analysis. We emphasize that these are general guidelines and should not be construed as hard and fast rules. Usually your data could be analyzed in multiple ways, each of which could yield legitimate answers. The table below covers a number of common analyses and helps you choose among them based on the number of dependent variables (sometimes referred to as outcome variables), the nature of your independent variables (sometimes referred to as predictors). You also want to consider the nature of your dependent variable, namely whether it is an interval variable, ordinal or categorical variable, and whether it is normally distributed (see What is the difference between categorical, ordinal and interval variables? (https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/) for more information on this). The table then shows one or more statistical tests commonly used given these types of variables (but not necessarily the only type of test that could be used) and links showing how to do such tests using SAS, Stata and SPSS.

Number of Dependent Variables	Nature of Independent Variables	Nature of Dependent Variable(s)*	Test(s)	How to SAS	How to Stata	How to SPSS	How to R

1	0 IVs (1 population)			SAS (/sas/whatstat/what-statistical-analysis-	Stata (/stata/whatstat/what-	SPSS (/spss/whatstat/what-	R (/r/whatstat/what-statistical-analysis-

interval & normal	one-sample t-test	should-i-usestatistical-analyses-using-sas/#1sampt)	statistical-analysis-should-i-usestatistical-analyses-using-stata/#1sampt)	statistical-analysis-should-i-usestatistical-analyses-using-spss/#1sampt)	should-i-usestatistical-analyses-using-r/#1sampt)
ordinal or interval	one-sample median	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#1sampm)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#1sampm)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#1sampm)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#1sampm)
categorical (2 categories)	binomial test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#bitest)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#bitest)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#bitest)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#bitest)
categorical	Chi-square goodness-of-fit	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#chifit)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#chifit)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#chifit)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#chifit)

1 IV with 2 levels (independent groups)		2	SAS (/sas/whatstat/what-statistical-analysis-	Stata (/stata/whatstat/what-	SPSS (/spss/whatstat/what-	R (/r/whatstat/what-statistical-analysis-

interval & normal	independent sample t-test	should-i-usestatistical-analyses-using-sas/#2ittest)	statistical-analysis-should-i-usestatistical-analyses-using-stata/#2ittest)	statistical-analysis-should-i-usestatistical-analyses-using-spss/#2ittest)	should-i-usestatistical-analyses-using-r/#2ittest)
ordinal or interval	Wilcoxon-Mann Whitney test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#wilc)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#wilc)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#wilc)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#wilc)
categorical	Chi-square test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#chisq)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#chisq)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#chisq)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#chisq)
	Fisher's exact test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#exact)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#exact)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#exact)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#exact)
1 IV with 2 or more levels (independent groups)		SAS (/sas/whatstat/what-statistical-analysis-	Stata (/stata/whatstat/what-	SPSS (/spss/whatstat/what-	R (/r/whatstat/what-statistical-analysis-

	interval & normal	one-way ANOVA	should-i-usestatistical-analyses-using-sas/#1anova)	statistical-analysis-should-i-usestatistical-analyses-using-stata/#1anova)	statistical-analysis-should-i-usestatistical-analyses-using-spss/#1anova)	should-i-usestatistical-analyses-using-r/#1anova)
	ordinal or interval	Kruskal Wallis	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#kw)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#kw)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#kw)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#kw)
	categorical	Chi-square test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#chisq)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#chisq)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#chisq)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#chisq)
1 IV with 2 levels (dependent/matched groups)	interval & normal	paired t-test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#pairt)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#pairt)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#pairt)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#pairt)
	ordinal or interval	Wilcoxon signed ranks test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#wilcsign)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#wilcsign)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#wilcsign)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#wilcsign)
	categorical	McNemar	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#Mcnemar)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#Mcnemar)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#Mcnemar)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#Mcnemar)
1 IV with 2 or more levels (dependent/matched		one-way	SAS (/sas/whatstat/what-statistical-analysis-	Stata (/stata/whatstat/what-	SPSS (/spss/whatstat/what-	R (/r/whatstat/what-statistical-analysis-

groups)	interval & normal	repeated measures ANOVA	should-i-usestatistical-analyses-using-sas/#1repanova)	statistical-analysis-should-i-usestatistical-analyses-using-stata/#1repanova)	statistical-analysis-should-i-usestatistical-analyses-using-spss/#1repanova)	should-i-usestatistical-analyses-using-r/#1repanovA)
	ordinal or interval	Friedman test	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#fried)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#fried)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#fried)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#fried)
	categorical (2 categories)	repeated measures logistic regression	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#1replog)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#1replog)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#1replog)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#1replog)
2 or more IVs (independent groups)	interval & normal	factorial ANOVA	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#factanov)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#factanov)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#factanov)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#factanov)
	ordinal or interval	ordered logistic regression	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#orderedlogistic)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#orderedlogistic)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#orderedlogistic)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#orderedlOgistic)
	categorical (2 categories)	factorial logistic regression	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#faclogistic)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#faclogistic)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#faclogistic)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#faclogisTic)
1 interval IV			SAS (/sas/whatstat/what-statistical-analysis-	Stata (/stata/whatstat/what-	SPSS (/spss/whatstat/what-	R (/r/whatstat/what-statistical-analysis-

interval & normal	correlation	should-i-usestatistical-analyses-using-sas/#corr)	statistical-analysis-should-i-usestatistical-analyses-using-stata/#corr)	statistical-analysis-should-i-usestatistical-analyses-using-spss/#corr)	should-i-usestatistical-analyses-using-r/#corr)
interval & normal	simple linear regression	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#simpreg)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#simpreg)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#simpreg)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#simpreg)
ordinal or interval	non-parametric correlation	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#nonparr)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#nonparr)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#nonparr)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#nonparr)
categorical	simple logistic regression	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#simplog)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#simplog)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#simplog)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#simplog)
1 or more interval IVs and/or 1 or more categorical IVs	**interval & normal**		SAS (/sas/whatstat/what-statistical-analysis-	Stata (/stata/whatstat/what-	SPSS (/spss/whatstat/what-	R (/r/whatstat/what-statistical-analysis-

		multiple regression	should-i-usestatistical-analyses-using-sas/#multreg)	statistical-analysis-should-i-usestatistical-analyses-using-stata/#multreg)	statistical-analysis-should-i-usestatistical-analyses-using-spss/#multreg)	should-i-usestatistical-analyses-using-r/#multreg)
		analysis of covariance	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#ancova)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#ancova)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#ancova)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#ancova)
	categorical	multiple logistic regression	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#logistic)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#logistic)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#logistic)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#logistic)
		discriminant analysis	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#discrim)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#discrim)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#discrim)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#discrim)
2+	**1 IV with 2 or more levels (independent groups)**			SAS (/sas/whatstat/what-statistical-analysis-	Stata (/stata/whatstat/what-	SPSS (/spss/whatstat/what-	R (/r/whatstat/what-statistical-analysis-

		interval & normal	one-way MANOVA	should-i-usestatistical-analyses-using-sas/#manova)	statistical-analysis-should-i-usestatistical-analyses-using-stata/#manova)	statistical-analysis-should-i-usestatistical-analyses-using-spss/#manova)	should-i-usestatistical-analyses-using-r/#manova)
	2+	interval & normal	multivariate multiple linear regression	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#mmreg)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#mmreg)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#mmreg)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#mmreg)
	0	interval & normal	factor analysis	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#factor)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#factor)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#factor)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#factor)
2 sets of 2+	**0**	interval & normal	canonical correlation	SAS (/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/#cancor)	Stata (/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/#cancor)	SPSS (/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#cancor)	R (/r/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-r/#cancor)

*Technically, assumptions of normality concern the errors rather than the dependent variable itself. Statistical errors are the deviations of the observed values of the dependent variable from their true or expected values. These errors are unobservable, since we usually do not know the true values, but we can estimate them with residuals, the deviation of the observed values from the model-predicted values. Additionally, many of these models produce estimates that are robust to violation of the assumption of normality, particularly in large samples.

This page was adapted from *Choosing the Correct Statistic* developed by James D. Leeper, Ph.D. We thank Professor Leeper for permission to adapt and distribute this page from our site.

ST1232 cheat sheet

Statistics for Life Sciences (National University of Singapore)

Hypothesis Testing

1. Assumptions:
   - Data comes from randomization
   - Shape of population distribution is normal
   - Sample size large enough
2. Hypothesis:
   - H_0: null hypothesis
   - H_1: parameter falls in alternative range of values
3. Test statistic:
   - How far point estimate falls from the parameter value given in null hypothesis
   - Measured by number of standard errors between point estimate and parameter value in H_0
4. Interpreting test statistic:
   - Presume H_0 is true
   - But if sample test statistic falls out in the tail of sampling distribution, value is unlikely
   - P-value is the probability of "what we saw or something more extreme" given H_0 is true

Significance level:

- a number such that we reject H_0 if p-value is less than or equal to that number
- if we reject H_0, the results are statistically different

Hypothesis Testing for proportions

1. Assumptions:
   - Variable collected is categorical
   - Data comes from randomization
   - Sample size n is sufficiently large that the sampling distribution of sample proportion $\hat{p}$ is approximately normal if $np_0(1-p_0) \geq 5$
2. Hypothesis
   - H_0: $p = p_0$
   - H_1: $p \neq p_0$
3. Test statistic:
   - $\hat{p}$ $N(p_0, \frac{p0(1-p_0)}{n})$ when H_0 is true
   - $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, z $(0,1)$

Hypothesis testing for mean (One sample t-test)

1. Assumptions:
   - Variable collected is quantitative
   - Data comes from randomization
   - Population distribution is approximately normal
   - n is small → one sided t-test
2. Hypothesis:
   - H_0: $\mu = \mu_0$
3. Test statistic:
   - Distance between sample mean $\acute{X}$ and H_0 value of population mean
   - Measured in terms of standard errors
   - $T = \frac{\acute{X}-\mu_0}{s/\sqrt{n}}$ with n-1 degrees of freedom
4. Conclusion:
   - If H_0 is true
   - Mean is not significantly different from μ_0

Normality assumption

1. Histogram with a Normal pdf overlaid
   - Symmetric, unimodal
2. Q-Q plot
   - Standardized sample quantiles against theoretical quantiles of N(0,1) distribution
   - If fall on a straight line, the data is normal
   - Right tail below straight line: fatter
   - Left tail below straight line: thinner

Errors in conclusions

1. Type 1 error: occurs if we reject H_0 when in fact it is true
   - when H_0 is assumed to be true
2. Type 2 error: occurs if we do not reject H_0 when it is false
   - When H_1 is assumed to be true
3. ↑sample size → type 2↓
4. ↓significance level, type 1↓ but type 2↑

2 independent sample t-test, Equal variance

1. Assumptions:
   - Independent samples with randomized experiment
   - Sample standard deviations not twice…
   - Population distribution is approximately normal and n is small
2. Hypothesis:
   - H_0: $\mu_1 = \mu_2$
3. Test statistic:
   - Point estimate of difference between the population means $\acute{X} - \acute{Y}$
   - Estimate of common variance: $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-}{n_1+n_2-2}$
   - $T = \frac{(\acute{X}-\acute{Y})-0}{se}$ with n_1+n_2-2 degrees of freedom
   - se= $s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

2 independent sample t-test, Unequal variance

1. Assumptions:
   - Quantitative variable for 2 groups, independent, random
   - Sample standard deviation twice…
   - Approximately normal and n is small
2. Hypothesis:
   - H_0: $\mu_1 = \mu_2$
3. Test statistic:
   - Point estimate of difference between the population means $\acute{X} - \acute{Y}$
   - $T = \frac{(\acute{X}-\acute{Y})-0}{se}$ with complicated number of degrees of freedom (df)
   - se= $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
4. Conclusion:
   - If do not reject H_0, no statistically difference between means

Dependent sample test

-Testing if $\mu_1 = \mu_2$

1.Construct new variable

-$D_i = X_i - Y_i$

2.Hypothesis:

-H_0: $\mu_D = 0$

3.Test statistic:

- $T = \frac{\acute{X}-\mu_0}{s/\sqrt{n}}$ with n-1 degrees of freedom

4.Pairing up reduced standard error in $\acute{D}$

ANOVA: Compare more than 2 Group means

1. Assumptions:
   - Population distribution for the response for k groups are normal
   - Largest group sd less twice of smallest sd
   - Data randomized
   - N observations in total: n observations in k group, N=nk
2. Hypothesis:
   - H_0: $\mu_1 = \mu_2 = \mu_k$
   - H_1: At least 2 means are not equal
3. Test statistic:
   - Involves within grp &

Multiple comparison test

-To control Experiment error rate **a**: the probability of making at least 1 Type 1 error when all m hypothesis are true

1. Bonferroni correction:

- Not doing ANOVA
- perform each test at a/m significance level
- for each CI at (1-a/m)100% level
- p-value ¿ $\frac{a}{m}$

2.Tukey:

- replacing use of t-distribution for each pairwise distribution with a multiplier from another

Wilcoxon signed rank For matched pairs

1. Assumptions:
   - 2 groups are independent
   - Data random
   - Difference of observations can be ranked
   - Population distribution of difference is symmetric
2. Hypothesis:
   - H_0: $q_{0.5} = 0$
3. Test statistic:
   - Sum of the ranks of the difference that were

2-sample Wilcoxon rank sum (Mann-Whitney U)

1. Assumptions:
   - Independent observations $X_1, X_2 \ldots X_{n1}$ and $Y_1, Y_2 \ldots Y_{n2}$ from 2 groups
   - Random data

Hypothesis:

- H_0: 2 samples are from same distribution

2. Test statistic:
   - Minimum of R_X and R_Y, R_X e sum of

Kruskal-Wallis

-If more than 2 groups of independent samples

-Not normal and unequal variance

1. Assume:

-Independent observations from k groups where k ¿ 2

-Random data

2.Hypothesis:

-H_0: The distributions of all groups

Simple linear Regression

-F-test equals t-test

1. Assumptions:
   - Random data
   - R/S between X and Y is linear
   - $\varepsilon \ N(0,\sigma^2)$, normally distributed subpopulation for each explanatory variable and same variance
   - $Y \ N(\beta_0+\beta_1 X$
2. t-test Hypothesis:

Multiple linear Regression

-more than one explanatory variable

-t-test for individual coefficient(quantitative)

-F-test for overall significance (categorical)

-Use adjusted R^2 to compare models

Indicator variables

$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2$

when reference=female (0) &male (1)

1. Hypothesis:
   - H_0: all coefficients=0
   - H_1: at least 1 not zero

Point estimates

- $\hat{\beta}_0$ (constant) and $\hat{\beta}_1$ are point estimates
- $\hat{Y}$ is a point estimate of the mean of response at a particular X value

-when X increases by 1 unit, mean response $\hat{Y}$ increases by $\hat{\beta}_1$ unit

Compare within 2 groups

-H_0: $\mu_i = \mu_j$ if F-test is significant

- $T = \frac{\acute{Y}_i - \acute{Y}_j}{se}$

where se= $s_p\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$ with N-k degrees of freedom

- between grp variability
- F= $\frac{Btwn\ grp\ vari}{within\ grp\ vari}$ with numerator k-1 and denominator N-k degrees of freedom

4. p-value: always the areas to the right of it; larger F → against H_0

distribution (a studentized range distribution)
- input: no. of k grps; output: all significant pairwise comparisons

3. Dunnett:
-k groups with one control, conduct m=k-1 comparisons
-same as Tukey but will provide shorter CI

- positive (W+)
- If W+ ≈ W-, do not reject H_0

For one sample

1. Hypothesis:
   - H_0: $q_{0.5} = m_0$
2. Test statistic:
   -sum of ranks ¿ m_0 known as (W+)

sample while R_Y is the sum of ranks of Y-sample
- H_0 is correct if $R_X \approx R_Y$

3.Test stats:
-Uses the variability of sample-mean ranks
-Under H_0, follows a X^2_{k-1} distribution (chi-square)

- H_0: $\beta_1 = 0$ and t-distribution follows n-2 degrees of freedom

3. F-test Hypothesis:
   - H_0: $\beta_1 = 0$ and F distribution with 1 and n-2 degrees of freedom

2. **Adjusted** R^2 =0.068 means only 6.8% of variation in the marks change is explained by regression model

With interaction term

$\hat{Y} = \beta_1 + \beta_1 X_1 + \beta_2$

Estimating σ^2

-how variable the response measurements are around fitted line

- $\hat{\sigma^2}$ is computed using residuals

$e_i = Y_i - \hat{Y}$

-Variance is minimized at $X_0 = \acute{X}$, smaller the better

Scatterplot

1. Linearity violated:
   - Add higher order terms in X eg. X^2
2. Variance not constant:
   - Transform the response by taking ln(Y), $\sqrt{Y}$ or (1/Y)
   - 1 unit increase in X, mean response $\hat{Y}$ increase by $e^{\hat{\beta}_1}$ times

Residuals plot

- Detect non-normality
- Check for non-constant variance and the need to transform Y
- Check for the need to add higher order terms in X

Plots to make

1. Plot r_i's against $\hat{Y}_i$ on
2. Plot r_i's against X_i
3. Create histogram of r_i
4. Create QQ plot using r_i

- Transform Y when there is funnel shape
- Add higher order X term when there is a curve band
- Non-normal when points are outside 3 and -3

*Note that r_i are not independent

Cook's distance plot

- To check for influential point

R^2: The proportion of total variation of the response that is explained by the fitted regression.

4.Boxplot
-Minimum, lower quartile, median, upper quartile and maximum
-For comparing between quantitative vs categorical variable
-Skewness of distributions (mention if obvious)

Chi-square test for 2 × 2 with continuity correction

1. Assumptions:
- 2 categorical variables
- all expected cell count >5
- Data random
2. Hypothesis:
- H_0: 2 Variables are independent
- H_1: 2 Variables are dependent
3. Test statistic:
- $\chi^2 = \sum \frac{(|observed\ cou}{ex}$
4. P-value:
- Right tail probability of χ^2 distribution with 1 degree of freedom

Pearson chi-square

- r rows and c columns that define 2 categorical random variables
- use $\chi^2_{(r-1)(c-1)}$ distribution

Fisher's exact

1. Assume:
- ≥ 20% expected cell count <5
- 2 binary categorical variables
- Data random
2. Hypothesis:
- H_0: 2 variables are independent
3. Test stats:
- The first cell count

Linear by linear

-There is at least 1 ordinal variable

Strength of association

-How different is P(disease|exposure) is from P(disease|no exposure)

- $\hat{p}_1 = \frac{a}{a+b}$ is point estimate of P(disease|exposure)
- $\hat{p}_2 = \frac{c}{c+d}$ is a point estimate of P(disease|no exposure)

Prospective study:
-randomly assign the exposure variable to subjects or record their exposure variable status

Retrospective study:
-case control study
-cannot obtain $\hat{p}_1$ and $\hat{p}_2$ from 2 × 2 table

Difference of proportion:

- $\hat{p}_1 - \hat{p}_2$

Relative risk:

- $\frac{\hat{p}_1}{\hat{p}_2}$

Odds of success:

-odds= $\frac{p}{p-1}$ where p is the probability of success and p-1 is failure

-odds ratio= $\frac{a/b}{c/d} = \frac{ad}{bc}$

-The odds of having disease in expose is n times the odds of having disease in unexposed

Computing CI for OR

-estimated se for ln(OR) is se(ln(OR))= $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

-Formula for a (1- α)100% CI:

$e^{\ln(\frac{ad}{bc}) \pm q_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+}}$

Logistic regression

-Response variable Y is a binary random variable (categorical)

1.Assumptions:
-Data is randomized
-Observations are independent
-Denote p_i=P(Y_i=1) (success probability) given a value of X_i

2.Odds of success:

- $\frac{P(Y_i=1)}{P(Y_i=0)} = \frac{p_i}{1-p_i}$
- ln ($\frac{p_i}{1-p_i}$)= $\beta_0 + \beta_1 X_i$
- ln ($\frac{p_i}{1-p_i}$)= $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + .$

3.Distributional assumption:

- Y_i $Bin(1, p_i)$
- $p_i = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$

Maximum likelihood estimation:
-Test fitness of model using R^2

Model interpretation

-For every increase in X_i, the ln odds increase by β_1 times

-For every increase in X_i, the odds of success will increase by e^{β_1} times

Wald Hypothesis Test

-Test if an individual term is significantly different from zero

1. Hypothesis:
-H_0: $\beta_1 = 0$
2. Test statistic:
- $\frac{1}{\hat{\beta}_¿ \over s.e(¿¿)}^2$ with $\frac{\hat{\beta}_1}{¿¿}$ sampling distribution χ^2_1 distribution

Confidence Interval of Wald Test

- $\hat{\beta}_1 \pm q_{1-\frac{\alpha}{2}} \times s.e(\hat{\beta}_1)$

-odds ratio corresponding to β is (e^L, e^U ¿

-95% CI for odds ratio corresponding to 1 unit increase in X_i is (e^L, e^U ¿

Explanatory data analysis

1.Frequency table:
-For categorical or quantitative variable
-Modal category: highest frequency

1.Mean | 6.pth-quantile | **Probability** | **Random variable:** | **Binomial Distribution** | **Confidence interval**

-Look at proportion of modal category, compare it via the difference

2.Bar Plot:

-To display single categorical variable

-Mention group of categories with high/low proportions

-Mention any apparent trend in proportions if there is ordering

3.Histogram

-use bars to portray the frequency or relative frequency of the possible outcomes for quantitative variable

-Overall pattern: Gap, outlier?

-Skewed or symmetric?

-Unimodal? Bimodal?

-Spread of data

-Median and IQR

-No. of outliers

-Outlier: 1.5 $\times IQR$ below and above upper/lower quartile

5.Contingency table

-To display 2 categorical variables

6.Scatterplot

-For association between quantitative vs quantitative variable

-correlation:

$$\frac{Y_i - \acute{Y}}{s_y}$$

$$\left(\frac{X_i - \acute{X}}{s_x}\right)(¿)$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n} ¿$$

-r is between 1 and -1

-Sum of observations divided by number of observations

- $\acute{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

2.Median

-Middle value of the observations when observations ordered from smallest to largest

- $\left(\frac{n+1}{2}\right)$ th largest if odd,

average of $\left(\frac{n}{2}\right)$ th and

$\left(\frac{n}{2}+1\right)$ th if even →

$X_{0.5}$

*If date is highly skewed, report median and if symmetric, report mean

3.Range

-Difference between largest and smallest observations

-Measure of spread but sensitive to extreme observations

4.Variance

-Average of the squared deviation from the mean

-

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \acute{X}\right)^2$$

5.Standard deviation

-Represents average distance from mean

-68% observations fall between 1 sd of mean, between $\acute{X} - s$ and $\acute{X} + s$

-95% fall between 2 sd of mean

-All fall between 3 sd of mean

-value that p percent of the observations fall below or at that value ($\widehat{q_{0.5}} = X_{0.5}¿$

7.IQR

-Distance between upper and lower quartile $\left(\widehat{q_{0.75}} - \widehat{q_{0.25}}\right)$

-50% of sample points fall within IQR

-how spread the 'middle' data is

Lurking variable

-Variable that is usually unobserved that influence the association between variables of primary interest

Confounding

-2 explanatory variables are associated with a response variable but also associated with each other

Poor sources of bias

-Sampling bias

-Non-response bias

-Response bias

1.Mutually exclusive (disjoint)

-A and B has no common outcome

$A \cap B = \varnothing$

-If P(A) ≥ 0 , P(S)=1,

$P(A \cup B) = P(A)$

2.Not mutually exclusive

-

$P(A \cup B) = P(A)$

-

$P(A) = P(A \cap B)$

3.Independent event

-2 events do not influence one another

-

$P(A \cap B) = P(A)$

-

$P(A|B) = P(A)$

4.Conditional probability

-Probability of A given B

-

$$P(A|B) = \frac{P(A \cap$$

5.Bayes Theorem

6.Sensitivity

-Probability that the test is positive given that the person has disease P(A|B)

7.Specificity

-Probability that the test is negative given no disease $P(A^c|B^c)$

8.Prevalance

-Having Disease P(B)

-Numerical measurement of outcome of experiment via random sampling

- $p_z = \frac{e^{-20}20^z}{z!}$

for barplot

1.Mean:

-sum of probabilities multiplied by possibilities

$$\mu = \sum_x x\, p_x$$

2.Expected value of X, E(X): The mean of the probability distribution of a random variable X

$$E(\acute{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i =$$

*mean of variables $\acute{X}$ =mean of each random variable

3.Standard deviation

-measures variability of a random variable from the mean

$$\sigma^2 = \sum_x (X-\mu)^2 p$$

4.Variance

-large standard deviation has larger variability

$$Var(\acute{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$

Probability density function pdf

-curve that determines probabilities of intervals

Mean of continuous variable X

$$\mu = \int x f(x)\, dx$$

Variance of continuous variable

$$\sigma^2 = \int (x-\mu)^2 f(x$$

Pth quantile of cont variable

$$P(x \leq q_p) = p$$

Combination:

$$C_k^n = \frac{n!}{k!(n-k)!} =$$

-n trials have 2 possible outcomes, independent

-same probability of success

$X \quad Bin(n, p)$

$P(X = x) = C_x^n p^x (1-$

$1-p$

$E(X) = np, Var(X) =$

)

Normal distribution

-Symmetric, bell-shaped

$X \quad N(\mu, \sigma^2)$

$$\acute{X} = \frac{X_1 + .. X_n}{n} \quad N(\mu$$

$$Z = \frac{X-\mu}{\sigma} \quad N(0,1)$$

Probability distribution

P(X=1)=p, P(X=0)=1-p

Sampling distribution:

-Probability distribution that specifies the probabilities for possible values that statistic can take

Sampling proportion $\hat{p}$ for random sample size n:

Mean=p, sd= $\sqrt{\frac{p(1-p)}{n}}$,

$p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2$ if $\hat{p} \quad N¿$

n is sufficiently large that $n\hat{p}(1-\hat{p}) \geq 5$, the sampling distribution is approximately normal

CLT: $\acute{X} \quad N\left(\mu, \frac{\sigma^2}{n}\right)$

-independent observations

-n $\geq$ 30

-X is approximately normal

*a long run interpretation

Margin of error:

-measures how accurate the point estimate is likely to be estimating a parameter→sd

Standard error:

-an estimated deviation of a sampling distribution

Population proportion: p

Sample proportion: $\hat{p}$

→for cat

When $n\hat{p}(1-\hat{p}) \geq 5$

,

$$\hat{p} \pm q_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

99%=2.58,80%=1.28,95%=1.96

Width: higher the CI, smaller α and wider the width

$$2 \times q_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n = \left(\frac{2 \times q_{1-\frac{\alpha}{2}}}{D}\right) p(1-$$

*use p=0.5 (highest variability to obtain highest n)

t-distribution:

-probability on the degree of freedom(df)

-thicker tail, more variability than N(0,1)

-larger the df, closer to N(0,1)

$$\acute{X} \pm t_{n-1,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

$$2 \times t_{n-1,1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \leq D$$

$$n \geq \left(\frac{t_{n-1,1-\frac{\alpha}{2}} \times s}{D}\right)^2$$