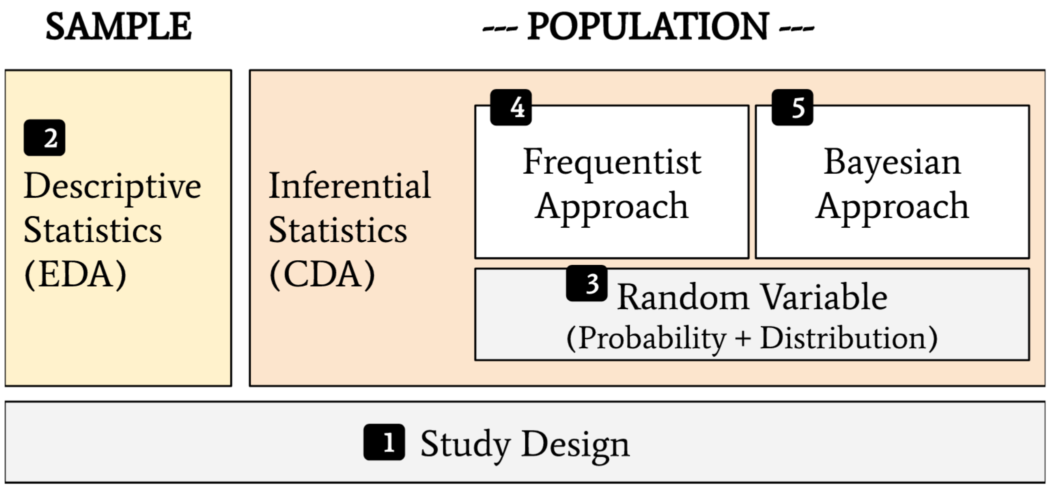
Statistics for Data Science Concept Notes

Last Edit: 2021.01.22 Written By: Sujeong Cha



* EDA: Exploratory Data Analysis, CDA: Confirmatory Data Analysis

1. Study Design

a. Study Types & Sampling

b. Variable Types

2. Descriptive Statistics

a. Types of Descriptive Stats.

b. Charts & Graphs

3. Random Variables

a. Probability

b. Random Variables

c. Expectations

d. Random Process

e. Markov Chains

4. Frequentist Inference

a. Estimation

b. Hypothesis Testing

5. Bayesian Inference

a. Estimation

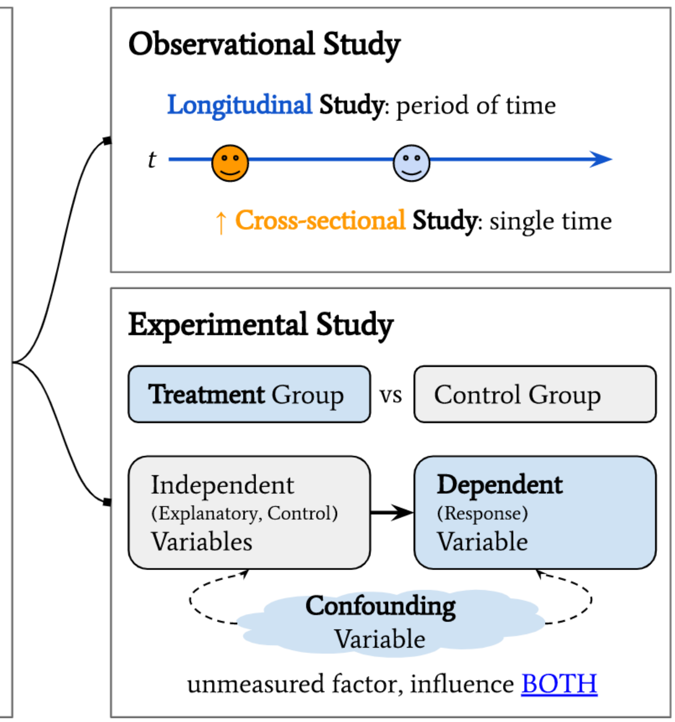
b. Bayesian Hypothesis Testing

1. Study Design

a. Study Types & Sampling Techniques

Sampling Techniques

1. **Simple Random Sampling**: equal chance for each individual
2. **Stratified Sampling**: grouping variable correlated with the measurement
3. **Cluster Sampling**: grouping variable NOT correlated with the measurement
4. **Systematic Sampling:** every nth member of the population
5. **Convenience Sampling**: unethical, biased



b. Variable Types

Types	Scale	Category	Order	Equal Intervals	True Zero	Example
Qualitative (Categorical)	Nominal	Y	N	N	N	Gender
	Ordinal	Y	Y	N	N	A, B, C
Quantitative (Numerical) Discrete/Continuous	Interval	Y	Y	Y	N	Year, IQ
	Ratio	Y	Y	Y	Y	Weight, Age

2. Descriptive Statistics

a. Types of Descriptive Statistics

Measures of Central Tendency

- **Mean / median / mode**
  * uni-/bi-/multi-modal or no mode
- **Hildebrand Rule**: is it symmetric?
  : H=(mean-median) / std
  → sufficiently symmetric if |H| < 0.2

Measures of Frequency

- **Frequency** (Counts)
- **Relative frequency** (in %)
- **Cumulative frequency**
  (only when ordered)

Measures of Position

- **Percentile**
- **Quartile**
- **IQR (Q3-Q1)**
- **z-score**

$$Z = \frac{x - \mu}{\sigma}$$

Measures of Dispersion

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

- **Range / variance / standard deviation**
  * std: same unit → easier to interpret
- To compare between variables,
  Coefficient of variation = std / mean
  ("risk/reward ratio")
- **Empirical Rule:** 68-95-99.7%
  * can apply only to bell-shaped curves
- **Chebyshev's Theorem**:
  P(within K stds) > 1-(1/K2)
  * can apply to any type of distribution

Measures of Dependence

- **Covariance**: measure of joint variability
- **Correlation coefficient**: cov / std(x)std(y)
- **Covariance matrix**:
  covariance between every pair of features

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

b. Charts & Graphs

Qualitative	Pie chart, Bar chart, Pareto chart (bars in descending order)
Quantitative	Histogram(1D), Boxplot (five-number summary) (1D) Scatter plot (2D), Line Graph (2D)

Appendix. Sample Covariance Matrix

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

- Using the sample covariance matrix, we can express the variation in every direction
  $$var(\vec{v}^T\vec{x}_1, \cdots, \vec{v}^T\vec{x}_n) = \vec{v}^T\ \Sigma(\vec{x}_1, \cdots, \vec{x}_n)\ \vec{v}$$
- Using eigendecomposition of the covariance matrix,
  $CV = VL \Rightarrow C = V\,L\,V^{-1}$: particular case of Singular Value Decomposition
  1) **PCA**: eigenvectors & eigenvalues characterize the variation of the data in every direction
  2) **Whitening**: decorrelate (eliminate the linear skew) to reveal underlying nonlinear structure
  $\vec{y}_i = \sqrt{L}^{-1}\,V^T\vec{x}_i$ (the covariance of y becomes an identity matrix)

3. Random Variables

a. Probability

Interpretation	• **Frequentists**: relative frequency in the long run : n(E) / n(S) • **Propensity**: experimental probability : freq / #OfSamples • **Bayesian** (Subjectivists): degree of belief
Rules	---------- **Probability Axioms** ---------- 1. For any event A, 0 ≤ P(A) ≤ 1 2. ΣP(A) = 1 3. Complement Rule: P(not A) = 1 – P(A) ---------- **Probability Rules** ---------- 4. Addition Rule: P(A or B) = P(A) + P(B) – P(A and B) for Disjoint: P(A or B) = P(A) + P(B) 5. Conditional Probability: P(A|B) = P(A and B) / P(B) C. Independence: P(A.B|C) = P(A|C)P(B|C) 6. Multiplication Rule: P(A and B) = P(A)P(B|A) for Independence: P(A and B) = P(A)P(B) Chain Rule: P(A and B and C) = P(A)P(B|A)P(C|A,B) (cf) Pairwise independence does not imply joint independence. Independence does not imply conditional independence or vice versa. ---------- **Applications** ---------- 7. Bayes' Rule: P(A|B) = P(A)P(B|A) / P(B) - by Rule #5 and #6 - 8. Law of Total Probability: P(S) = ΣP(S and Ai) = ΣP(Ai)P(S|Ai) - by Rule #4 and #6 -

Appendix. Basic Counting Rules

- **Combination**: order DOES NOT matter (→ divide more!)
- **Permutation**: order matters
- **Special Permutation**: number of distinguishable permutations

$${}_nC_r = \frac{n!}{r!(n-r)!} \qquad {}_nP_r = \frac{n!}{(n-r)!} \qquad \frac{n!}{n_1!n_2!\ldots n_k!}$$

b. Random Variable

Random Variables (cf) RV (uncertainty) vs Realization (revealed outcome/value)

Discrete Random Variables (Probability Mass Function)

- **Bernoulli**: $P(X=x) = p^x(1-p)^{1-x}$ $E[X] = p$ $V[X] = p(1-p)$
  only two possible outcomes

Σ

- **Binomial** B(n,p): $P(X=x) = {}_nC_x\, p^x (1-p)^{n-x}$ $E[Y] = np$ $V[Y] = np(1-p)$
  sum of n i.i.d Bernoulli (count of success)

$\fallingdotseq$

- **Poisson** Pois(λ): $P(X=x) = \frac{\lambda^k e^{-\lambda}}{k!}$ $E[X] = \lambda$ $V[X] = \lambda$
  count of successes over area/time (no upper limit)
  can be modeled as a B(n, λ/n): (success or failure within 1/n interval) * repeat n times

- **Geometric** Geo(p): $P(K=k) = (1-p)^{k-1}p$ $E[X] = 1/p$ $V[X] = (1-p)/p^2$
  success on kth Bernoulli trial
  memorylessness: the distribution of waiting time until a certain event DOES NOT depend on how much time has elapsed already ⇒ P(X>m+n | X≥m) = P(X>n)

- **Hypergeometric**: $P(X=x) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$ $E[X] = n\frac{K}{N}$ $V[X] = n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}$

  k success in n draws w/o replacement from N population (K is total possible success)

Continuous Random Variables (Probability Density Function)

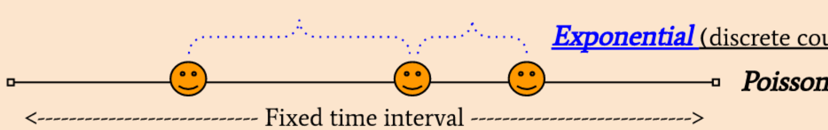
- **Uniform** unif(a,b): $P(X=x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$ $E[X] = \frac{1}{2}(a+b)$ $V[X] = \frac{1}{12}(b-a)^2$
  constant likelihood across interval
  Equivalent to a beta distribution with a=b=1

- **Exponential** exp(λ): $P(X=x) = \lambda e^{-\lambda x}$ $E[Y] = \frac{1}{\lambda}$ $V[Y] = \frac{1}{\lambda^2}$
  time between the events in a Poisson process
  memorylessness: the distribution of waiting time until a certain event DOES NOT depend on how much time has elapsed already ⇒ P(X>m+n | X≥m) = P(X>n)

- **Normal** N(μ, σ2): $P(X=x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

- **Beta** Beta(α, β): $P(X=x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}$

TIPs



Exponential (discrete counterpart: ***Geometric***)

memoryless

Appendix 1. Gaussian Mixture Model

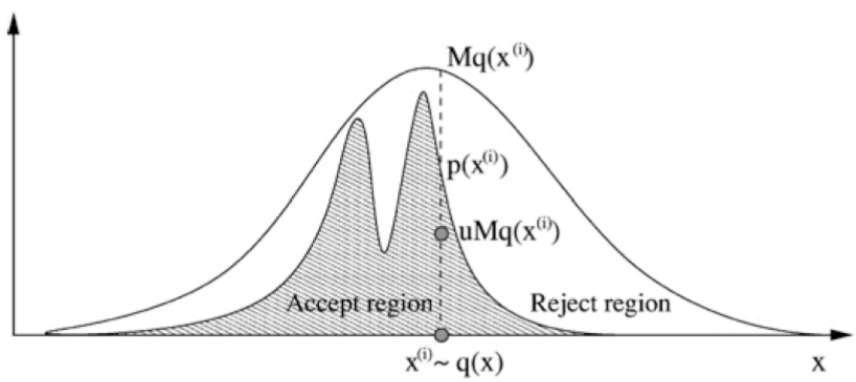
- **Gaussian random vectors**: multidimensional generalization of Gaussian random variables
  - Property 1) linear transformation of Gaussian random vectors are also Gaussian
  - Property 2) marginals of Gaussian random vectors are also Gaussian
- **Gaussian Mixture Model**: discrete marginal + continuous conditional

$$\sum_{d \in R_D} p_D(d) F_{C|D}(c|d)$$

  - data from continuous distribution whose parameters are chosen from a discrete set
  - a popular technique for clustering

Appendix 2. Sampling from Distribution

- **Inverse-transform sampling**: a sample of Uniform → set $x := F_X^{-1}(u)$
- **Rejection sampling**:
  - Why do we need it? When we cannot easily sample from f(x), there exists another density g(x) from which it is easy for us to sample (ex. Normarl, t-distribution as built-in functions)
  - How does it work? sample from a proposal function q(x) → sample from u(0,1) → y = u * M * q(x) → accept if u * M * q(x) ≤ p(x)



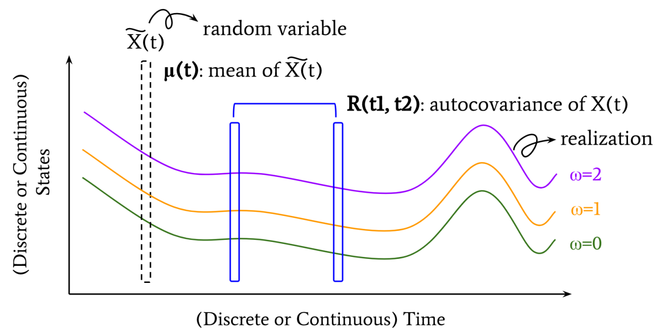
c. Expectation of Random Variables

Mean	• E[X] = $\sum$ x P(X=x) "first moment"
Median	• {x | P(X ≤ x) ≥ 0.5, P(X ≥ x) ≥ 0.5} → May not be unique for discrete R.V.
Variance	• V[X] = $\sum(x-\mu)^2$ P(X=x) "second centered moment"
Covariance	• Cov(X,Y) = E((X-E(X))(Y-E(Y)))= E(XY)-E(X)E(Y) • Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)
Correlation Coefficient	• = Cov(X,Y) / $\sigma_X\sigma_Y$ "normalized covariance" • To what extent X and Y are "linearly" related • Between -1 and 1 (can be proven with Cauchy-Schwarz inequality) • Independence implies un-correlation, but un-correlation does not imply independence (ex. U=X+Y, V=X-Y) (however, *under Gaussian*, un-correlation implies independence)
Conditional Expectation	• E(g(X,Y)|X):= h(X) where h(x):=E(g(X,Y)|X=x) • Conditional expectation is a random variable! ("function of X") • Iterated expectation: E(g(X,Y)) = E(E(g(X,Y)|X))

Appendix. Bounding Probabilities using Mean and Variance

- **Markov's inequality**: $X - a\,1_{X\geq a} \geq 0 \Rightarrow P(X \geq a) \leq \frac{E(X)}{a}$ (X is a nonnegative random variable)
- **Chebyshev's inequality**: $P(|X - E(X)| > a) \leq \frac{Var(X)}{a^2}$
  If Var(X)=Std(X)=1, equivalent to "Chebyshev's Theorem" from Part 2

d. Random Process

Random Variable Model uncertain quantities that "evolve in time"	 ● Wide/Weakly stationary: $\mu_{\tilde{X}}(t) = \mu$ **(constant)**, $R_{\tilde{X}}(t_1, t_2) = R_{\tilde{X}}(t_1 + \tau, t_2 + \tau)$ **(shift invariant)**
Examples	**IID Sequences** ● *Identical* and mutually *independent* distribution for any i (discrete-time) ● Fully characterised by the associated pdf/pmf (strictly stationary) ● $\mu_{\tilde{X}}(i) := E(\tilde{X}(i)) = \mu$ $\quad$ $R_{\tilde{X}}(i, j) = \sigma^2$ *if* $i = j$, 0 *otherwise* **Gaussian Process** ● Fully characterised by its mean function and autocovariance function ● Strict- and wide-sense stationary **Poisson Process** ● (1) $\tilde{N}(t_2) - \tilde{N}(t_1)$ is a Poisson RV with parameter $\lambda(t_2 - t_1)$ (2) $\tilde{N}(t_2) - \tilde{N}(t_1)$ & $\tilde{N}(t_4) - \tilde{N}(t_3)$ independent (if no time overlaps) ● $\mu_{\tilde{X}}(t) := E(\tilde{X}(t)) = \lambda t$ $\quad$ $R_{\tilde{X}}(t_1, t_2) := \lambda\, min(t_1, t_2)$ ● not stationary (mean not constant) **Random Walk** ● Models a sequence of steps in random directions ● $\mu_{\tilde{X}}(t) := 0$ $\quad$ $R_{\tilde{X}}(i, j) := min(i, j)$ ● not stationary (not shift invariant)
Convergence	**Convergence in Distribution** ● Cdf of $\tilde{X}(i)$ converges pointwise to the cdf of another random variable X ● Weaker than convergence with P=1 / in mean square / in probability ● ex) Binomial approximation of Poisson **Law of Large Numbers** ● Assume: (1) well-defined mean, (2) finite variance ● Weak LLN: moving average converges in mean square to μ ● Strong LLN: average converges to μ in probability

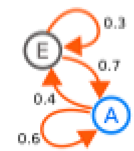
	Central Limit Theorem ● Assume: (1) finite variance, (2) independent sample w/ replacement ● The distribution of sample means tend towards Normal $\sum_{i=1}^{n} X_i \rightarrow N(n\mu,\ n\sigma^2) \Rightarrow \frac{\sum_{i=1}^{n} X_i}{n} \rightarrow N(\mu,\ \frac{\sigma^2}{n})$ ⇒ Justifies the use of Gaussian distributions to model data ● ex) Normal approximation to Binomial $X \sim Bin(n, p) \simeq N(np,\ npq) \Rightarrow\ p = X/n\ \sim\ N(p,\ pq/n)$

e. Markov Chains

Markov Chains

- **Markov property**: the future is conditionally independent from the past given the present (ex) IID sequences, Random Walk
- **Time-homogeneous Markov Chains**: the same transition probabilities for all t
  - Initial state vector and Transition matrix completely specify THMC

$$\vec{p}_{\tilde{X}(i)} = T^{i}_{\tilde{X}}\, \vec{p}_{\tilde{X}(0)}.$$



State vector for E $\begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix}$

Property

Irreducible

- for any state x, the probability of reaching every other state in a finite number of steps is non-zero (can reach every state)
- All states are recurrent
  (cf) recurrent: P(coming back)=1 *vs* transient state: P(coming back) < 1
- Have a single stationary distribution by P-F Theorem (Appendix)
  - Stationary distribution = eigenvector of T with eigenvalue = 1
  - Reversibility ($(T_{\tilde{X}})_{kj}\, \vec{p}_j = (T_{\tilde{X}})_{jk}\, \vec{p}_k.$) implies stationarity

Aperiodic

Period = 1

* **period**: the largest integer m such that it is only possible to return to x in a number of steps that is a multiple of m

Ergodic Markov Chains

its state vector converges to the stationary distribution for any initial state vector

Application: Markov Chain Monte Carlo (MCMC) Sampling

- Design an irreducible aperiodic Markov chain and sample from the stationary distribution
- Mixing(burn-in) time: takes time until convergence → discard the samples from this period
- **Metropolis-Hastings algorithm**:
  1) Generate a candidate by using the transition matrix $P(C=k|X(i-1)=j) = T_{kj}$
  2) Set X(i) = C with the acceptance probability $P_{acc}(j,k)=\min(T_{jk}P_k/T_{kj}P_j, 1)$
  → Produce a reversible, therefore stationary, Markov Chain
- Advantages:
  1) Works well when sampling from high-dimensional distributions
  2) Depends on the distribution only through the ratio (full pmf/pdf not needed)
  → Useful for sampling from posterior distributions in the Bayseian framework

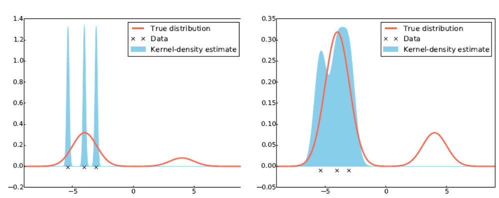
Appendix. Perron-Frobenius Theorem

If P is a stochastic matrix such that all the entries are strictly positive,

1) 1 is an eigenvalue of P and there exists an eigenvector $\mu \in \triangle n$ associated with 1
2) The eigenvector associated with 1 are unique up to scalar multiple
3) For all $x \in \triangle n$, $P^t x \rightarrow \mu$ in limit

4. Frequentist Statistics

a. Estimation

Assumption	• Parameters are unknown but fixed → cannot make probabilistic statements about the parameters
Point Estimate	**Method of Moments** • Adjust the parameters of a distribution so that the moments of the distribution coincide with the sample moments of the data • (ex) sample mean & variance: unbiased and consistent **Maximum Likelihood Estimate (MLE)** • Consistent but not always unbiased • (ex) least-square estimator, lasso estimator $\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, \widehat{L}_n(\theta\,; \mathbf{y})$
Interval Estimate	**Confidence Interval** • Quantify the estimator's accuracy ("soft estimate") ** if we had <u>infinite</u> data, we do not need a CI • Approximated using Central Limit Theorem because the sample mean follows Normal(μ, σ^2/n) • If σ is unknown, use t-score * StandardError $\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$ $\left(\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$
Non-parametric Distribution Estimate	• **[CDF] Empirical cdf**: unbiased and consistent estimator of the true cdf • **[PDF] Kernel Density Estimation** (KDE): many samples close to x → the estimate at x should be large  $\widehat{f}_{h,n}(x) := \frac{1}{n\,h} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$

Appendix. Analyzing Estimators

- **Mean Square Error** MSE[Y] = E[(Y-E(Y))2] + (E(Y)-γ)2 = variance + bias
- **Unbiased** E(Y) = γ
- **Consistent** converges to the true value as n → ∞

  * the sample median is always a consistent estimator unlike the sample mean

b. Hypothesis Testing: Reject or Fail to Reject

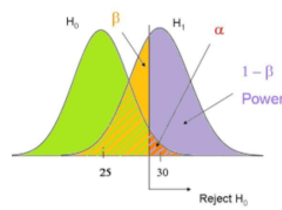
Hypothesis Testing: Did the patterns in the data came from random fluctuations?

** Frequentist perspective: either reject or fail to reject (can't compute probability)

- **Null Hypothesis** (=,≤,≥) vs. Alternative Hypothesis
- **P-value**: the probability of observing a result more extreme than the observations under H0
- **Type I Error** (α; false positive; significance level)

  Type II Error (β; false negative)

  Power (1-β): probability of rejecting H0 when it is indeed false



Parametric Testing

One sample *for mean and proportion*

z-test $$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

t-test $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$ with df = n-1

Two sample

z-test $$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Unpaired t-test $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ with df = n_1+n_2-2

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Paired t-test $$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$ with df = n-1

Non-parametric Testing

One sample *Target variance*

Chi-square for variance $$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$$

Chi-square Goodness-of-fit (right-tail)
H_0: from the hypothesized distribution

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$ with df = class-1

Two sample

Chi-square for Independence (right-tail)
H_0: two categorical variables are ind.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$ with df = (r-1)(c-1)

Permutation Test
H_0: two datasets are from the same dist.
(invariant to permutations)

Statistics of interest

$t_{\text{diff}}(\vec{x}) := t(\vec{x}_A) - t(\vec{x}_B)$ for each permutation

$$p = \mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq t_{\text{diff}}(\vec{x})\right) = \frac{1}{m}\sum_{i=1}^{m} 1_{t_{\text{diff},i} \geq t_{\text{diff}}(\vec{x})}$$

Multiple Testing

Generalization to >= 2

Analysis of Variance (ANOVA)

H_0: no difference in the means between groups (right-tail)

Source	SS	df	MS	F	Sig.
Between	SS_b	k-1	MS_b	MS_b/MS_w	p value
Within	SS_w	N-k	MS_w		
Total	$SS_b + SS_w$	N-1			

Bonferroni's Method

Guarantees that the desired significance level α holds simultaneously for all the tests.

Reject the null hypothesis if $$p_i \leq \frac{\alpha}{n}$$ *Can be proven by the union bound*

5. Bayesian Statistics

a. Estimation

Assumption	● Parameters are unknown and random → can be described probabilistically Bayes' Theorem: P(Θ|X) ∝ P(X|Θ)P(Θ) ● **Prior distribution** P(Θ): encodes uncertainty about the model ● **Likelihood** P(X|Θ): how the data(evidence) depend on the parameters ● **Posterior distribution** P(Θ|X): update uncertainty about the model ★ Conjugate priors: the prior and the posterior belong to the same family (ex) Beta distributions are conjugate priors when likelihood is binomial $\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ $\beta\left(a,b\right) := \int_u u^{a-1}\left(1-u\right)^{b-1}\,\mathrm{d}u$
Point Estimate	**Minimum mean-square-error estimation (MMSE)** ● Posterior mean E(Θ|X=x) minimizes the mean square error **Maximum-a-posteriori estimation (MAP)** ● Posterior mode (maximum of the pdf/pmf of the posterior) minimizes the probability of error $\mathrm{P}\left(\theta_{\text{other}}(\vec{X}) \neq \vec{\Theta}\right) \geq \mathrm{P}\left(\theta_{\text{MAP}}(\vec{X}) \neq \vec{\Theta}\right)$ ● Under a uniform prior, MAP equals to MLE → frequentist view can be understood as having a uniform prior
Interval Estimate	**Credible Interval** ● Interval in the posterior distribution within which an unobserved parameter value falls with a particular probability

b. Bayesian Hypothesis Testing

- Quantitative measure of much evidence there is for H_a relative to H_b given the data

$$K = \frac{P(H_b|\mathbf{y})}{P(H_a|\mathbf{y})} = \frac{P(H_b)}{P(H_a)}\boxed{\frac{P(\mathbf{y}|H_b)}{P(\mathbf{y}|H_a)}}$$

Bayes Factor
Can be calculated using MCMC method

- Model selection based on Bayes factors:
  1) calculate the expected loss for choosing H_a := $P(H_b|y)L(H_a|H_b)$ and similarly for H_b
  2) take the model which minimizes the expected loss (Bayesian decision)