S&DS 242 Study Guide

Varun Varanasi

May 7, 2022

1 Review of Probability Theory

- A **statistic** is a value computed from data
- The distribution of a statistic is its **probability distribution**
- A statistic is said to be unbiased if $\mathbb{E}(\hat{p}) = p$
- A **random variable** is a function from Ω to $\mathbb{R}$
- Discrete random variables take a finite number of values
- **Probability mass functions** outputs the probabilities of given inputs
- Variables are said to be **independent** if $P(X = x_i and Y = y_i) = P(X = x_i)P(Y = y_i)$
- A **cumulative distribution function** is defined as $F(X) = P(X \leq x)$
- $F_X(x) = \int_{-\infty}^{x} f_x(y)dy$
- The inverse of $F_x(x)$ is the **quantile function**
- A **joint distribution** is specified by a joint pmf/pdf = $\mathbb{P}[X_1 = x_1, ...X_k = x_k]$
- If $X_1, ..., X_n$ are independent then $f_{X_1,...,X_n}(x_1, ..., x_n) = f_{X_1}(x_1) \cdot ... \cdot f_{X_n}(x_n)$

Expected Value

- The **expected value** of a random variable X is given by $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_x(x)dx$
- **The Law of the Unconcious Statistician (LOTUS)**: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_x(x)dx$
- Expectation is linear: $\mathbb{E}(aX_1 + bX_2) = a\mathbb{E}(X_1) + b\mathbb{E}(X_2)$

Variance and Covariance

- The **variance** of a random variable is given by $Var[X] = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}(X^2) - (\mathbb{E}X)^2$
- Variance is invariant to translations (i.e $Var[X + c] = Var[X]$)
- Variance with a constant multiple: $Var[cX] = c^2Var[X]$
- $Var[X_1 + X_2] = Var[X_1] + Var[X_2]$ only if X_1 and X_2 are independent
- Otherwise, the variance of a sum is given by: $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$
- **Covariance** is defined as $Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
  - Translational Invariance: $Cov[X + a, Y + b] = Cov[X, Y]$
  - $Cov[X, X] = Var[X]$
  - If, X and Y are independent then $Cov[X, Y] = 0$; however, the converse is not true

  - Bilinearity: $Cov[X_1 + ... + X_n, Y_1 + ... + Y_m] = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov[X_i, Y_j]$
  - Constant multiple: $Cov[aX, bY] = ab\ Cov[X, Y]$
- Generalized Variance: $Var[X_1 + ... + X_n] = \sum_{i=1}^{n} Var[X_i] + 2\sum_{i<j} Cov[X_i, X_j]$
- **Standard deviation** is defined as $\sigma = \sqrt{Var[X]}$
- **correlation** between variables is their covariance normalized by the product of their standard deviations: $corr(X, Y) = \frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$
- **Cauchy-Schwarz Inequality**: $Cov[X, Y]^2 \leq Var[X]Var[Y]$

Moment Generating Functions

- A **moment generating function** of random variable X is defined as $M_X(t) = \mathbb{E}[e^{tX}]$
- Moment generating functions provide a more convinent way of studying distributions
- Two variables with the same moment generating functions have the same distribution
- The MGF of a sum of independent variables is the product of their individual MGFs (i.e $M_{X_1+...+X_n}(t) = M_{X_1}(t) \cdot ... \cdot M_{X_n}(t)$)

Multivariate Normal Distribution

- A multivariate normal distribution is characterized by a mean vector, μ and a symmetric covariance matrix, Σ
- A set of random variables is said to be multivariate normal if any linear combination of the variables has a normal distribution
- Furthermore, $\mathbb{E}[X_i] = \mu_i$, $Var[X_i] = \Sigma_{ii}$, and $Cov[X_i, X_j] = \Sigma_{ij}$

Large-sample Approximations

(Weak) Law of Large Numbers

Theorem: Suppose $X_1, ...X_n$ are iid with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] < \infty$. Let $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$. Then $\hat{X} \to \mu$ in probability, as $n \to \infty$.

Central Limit Theorem

Theorem: Suppose $X_1, ...X_n$ are iid with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Let $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$. Then, $\sqrt{n}(\frac{\bar{X}-\mu}{\sigma}) \to N(0, 1)$ in distribution, as $n \to \infty$.

Continuous Mapping Theorem

Theorem: Let g(x) be a continuous function of x. As $n \to \infty$,

- If $S_n \to Z$ in distribution, then $g(S_n) \to g(Z)$ in distribution
- If $S_n \to \mu$ in probability, then $g(S_n) \to g(\mu)$ in probability

2 Hypothesis Testing

- A **hypothesis test** is a binary question about the distribution of the data
- The goal of a hypothesis test is to either accept a **null hypothesis**, H_0 or reject in favor of an **alternative hypothesis**, H_1.
- Under the Neyman-Pearson paradigm, the default assumption is that H_0 is true. Therefore, the burden of the study is to disprove it.
- A **test statistic** T is any statistic computed from data of which extreme values provide evidence against H_0
- We can compute T from data and compare it against the distribution of T if H_0 were true. We refer to this proposed distribution as the **null distribution** of T
- For a given test statistic T, we divide its possible values into an acceptance and rejection region. If our calculated value of T belongs to the rejection region, then we reject H_0 in favor of H_1
- **Type I Error** is the probability that we wrongly reject H_0
- Type 1 Error = $\mathbb{P}_{H_0}$[T belongs to rejection region]
- In the Neyman-Pearson paradigm, we choose our rejection region such that Type 1 Error $\leq \alpha$ for a specificed α
- We refer to this value α as the **significance level** of our test
- Alternatively, we can define a **p-value** which is the significant level at which we would reject H_0
- In other words, the p-value is the probability that null distribution selects a value more extreme than the computed statistic
- A hypothesis is said to be **simple** if it completely specifies the distribution of the data
- **Type II Error** is the probability of accepting the null H_0 when H_1 is true.
- We define the **power** of a test to be the probability of correctly rejecting H_0
- Power: $1 - \beta = \mathbb{P}_{H_1}[reject\ H_0]$
- The goal of a proposed test statistic is to maximize the power of the test under the condition that type 1 error is $\leq \alpha$
- For a generic Normal Distributions, $T \sim N(\mu, \sigma^2)$, the rejection region of T is $T < \mu - \sigma \cdot z(\alpha)$ for a one sided test and $|T - \mu| > \sigma \cdot z(\frac{\alpha}{2})$

The Neyman-Pearson Lemma

Let H_0 and H_1 be simple hypotheses, and fix a significant level α. Suppose there exists a value c > 0, such that the likelihood ratio test which rejects H_0 when L(x) $<$ c and accepts H_0 when L(x) $\geq$ c has Type 1 error = α. Then for any other test with Type 1 error $\leq \alpha$, its power against H_1 is at most the power of the likelihood ratio test.

- The Liklihood ratio statistic: $L(x) = \frac{f_0(X)}{f_1(X)}$
- A hypothesis is referred to as **composite** if it is not simple
- For a hypothesis test with signficance level α on composite hypotheses, every possible distribution described by H_0 must have type 1 error less than the significance level

Pivotal Test Statistics

- A work around for this restriction is the use of a **pivotal** statistic which is a test statistic T whose sampling distribution is identical for every distribution in H_0
- The **one-sample t-statistic**, $T = \frac{\sqrt{n}\bar{X}}{S}$ is an example of a pivotal statistic for the $H_0 : X \sim N(0, \sigma^2)$
- S is the sample variance where $S = \frac{1}{n-1}((X_1 - \bar{X})^2 + ... + (X_n - \bar{X})^2)$
- Sample variance is independent of sample mean for iid draws from a normal distribution
- **t-distribution with n degrees of freedom** is given by $\frac{Z}{\sqrt{\frac{U}{n}}}$ where $Z \sim N(0, 1)$ and $U \sim \chi^2_n$

Non-parametric Test Statistics

- Another alternative to working with composite hypotheses is to rephrase them as nonparametric hypotheses
- A nonparametric hypothesis is one that does not specify the distribution with a particular form (i.e. H_0: f has median 0)
- **Sign Statistic** is an example of a nonparametric test statistic
- $S = \sum_{i=1}^{n} 1X_i > 0$
- We can define our rejection region based on Binomial(n, 1/2)
- For large n, we can use the CLT to estimate the distribution of S via $\sqrt{4n}(\frac{S}{2} - \frac{1}{2}) \to N(0, 1)$
- For large n, the type 1 error of this statistic will approach α (asymptotic level-α test)

Two-sample t-test

- Test the means of two normal distributions via the **pooled two-sample t-statistic**: $T = \frac{\bar{X} - \bar{Y}}{S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}}$
- Pooled sample variance: $S^2_{pooled} = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_i - \bar{Y})^2\right)$
- The pooled two-sample t-test is pivital under $H_0 : \mu_x = \mu_y$
- This test statistic $T \sim t_{m+n-2}$
- Welch's t-Test corrects for the assumption of the same variance between thw two distribution (also known as unequal variances t-test)
- Welch's t-test: $T_{welch} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n}S^2_X + \frac{1}{m}S^2_Y}}$

Mann-Whitney-Wilcoxon rank-sum Test

- Mann-Whitney-Wilcoxon rank-sum test is a nonparametric method for a two-sample test
  1. Sort a pooled sample of all observations $X_1, ...X_n, Y_1, ..., Y_m$ in increasing order
  2. Rank the sorted list with the smallest observation with a rank of 1
  3. Define T as the sum of ranks for the Y values
- Under the null-hypothesis the $\binom{m+n}{m}$ possibilities are equally distributed
  - $\mathbb{E}(T) = \frac{m(m+n+1)}{2}$
  - $Var[T] = \frac{mn(m+n+1)}{12}$
  - $\frac{T - \mathbb{E}[T]}{\sqrt{Var[T]}} \to N(0, 1)$ as $n, m \to \infty$

Permutation Tests

- Consider a generic test statistic $T(X_1, ..., X_n, Y_1, ..., Y_m)$
- The **permutation null distribution** of T is the distribution of $T(X_1^*, ..., X_n^*, Y_1^*, ..., Y_m^*)$ where each input is a random permutation of the datapoints
    1. Randomly permute pooled data many times and compute the value of T for each permutation
    2. Compute p-value as the fraction of the simulations whete T $\geq t_{obs}$ where t_{obs} is the value of T for the original data
    3. Reject H_0 if the p-value is $\leq \alpha$
- Permutation test is a type of conditional test
- Although the permutation test is not pivotal, it is pivotal over its conditional inputs
- Permutation test over all (m+n)! permutations would have type 1 error $\leq \alpha$
- Measures of distances between pairwise points can be used to estimate whether the distributions are approximately equivalent

Fisher's Exact Test

- Fisher's exact test is another example of a conditional test
- Suppose $X_1, ...X_{N_A} \sim Bern(p)$ and $Y_1, ...Y_{N_B} \sim Bern(q)$
- Our hypotheses are as follows: $H_0 : p = q$ and $H_1 : p > q$
- Let's consider our test statistic N_{A_1} the number of outcomes 1 in the group A
- Under random permutation, $P[N_{A_1} = k] = \frac{\binom{n_1}{k}\binom{n-n_1}{n_A-k}}{\binom{n}{n_A}}$ which takes the distribution $N_{A_1} \sim hypergeometric(n, n_1, n_A)$
- Reject H_0 when $N_a > U_{n,n_1,n_A}(\alpha)$

Power and Effect Size

- For a test of $\bar{X} \sim N(0, \frac{\sigma^2}{n})$ vs. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Analytically, this test rejects when $\frac{\sqrt{n}}{\sigma}\bar{X} > z(\alpha)$
- Under this paradigm, the power is given by $\Phi(\sqrt{n}\frac{\mu}{\sigma} - z(\alpha))$
- **Effect size** $\frac{\mu}{\sigma}$ is the shift in the mean between the tested hypothesis divided by the noise standard deviation
- There exist many power formulas dependent on the distributions and statistics being tested
- Paired Design is a testing method used to improve power of an experiment
- For $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x > \mu_y$ we can consider the paired differences $D_i = X_i - Y_i$
- Power for this paired distance simplifies to $\Phi(\frac{1}{\sqrt{1-\rho}} \cdot \sqrt{\frac{n}{2}} \cdot \frac{\mu_x - \mu_y}{\sigma} - z(\alpha))$
- $1 - \rho$ is known as the relative efficiency of the unpaired design

The Multiple Testing Problem

- For n different hypothesis tests, you will (on average) falsely reject αn of them
- p-values of the n tests should be uniformally distributed between [0,1]

The Bonferroni Correction

- Instead of using a significance level of α use a significance level of $\frac{\alpha}{n}$
- **Family-wise error rate** is the probability that we reject at least one of the true nulls
- Bonferroni method controls FWER at level α (guarantees FWER α for all null hypotheses)
- Controlling FWER is over-restrictive

The Benjamini-Hochberg Procedure

- **False Discovery Rate** is the number of true null hypotheses rejected divided by the number of total hypotheses rejected
- Controlling at FDR at a level α means that FDR $\leq \alpha$
- Estimate FDR by $\hat{FDP} = \frac{tn}{R(t)} \leq \alpha$ for the largest cutoff t where R(t) is the number of rejected hypotheses
  1. Sort the n total p-values from smallest to largest
  2. Find the largest r s.t. $P_{(r)} \leq \frac{\alpha r}{n}$
  3. Reject the first r null hypotheses

3 Parametric Models

- **Parametric model** is a family of distributions that are described by a few parameters
- A general parametric model with k parameters is denoted $f(x|\theta)$ where $\theta \in \mathbb{R}^k$
- We define the set of allowable values for our model as the parameter space (subset of $\mathbb{R}^k$)
- The study of parametric models is focused on estimating θ given $X_1, X_2, ..., X_n \sim f(x|\theta)$

Method of Moments

- Estimate k coefficients by considering the first k moments of the distribution
- Recall: $\mu_1 = \mathbb{E}[X], \mu_2 = \mathbb{E}[X^2], ..., \mu_k = \mathbb{E}[X^k]$
- We then estimate these moments using our data via $\hat{\mu_1} = \frac{1}{n}(X_1 + ... + X_n), \hat{\mu_2} = \frac{1}{n}(X_1^2 + ... + X_n^2) ... \hat{\mu_k} = \frac{1}{n}(X_1^k + ... + X_n^k)$

Bias, Variance, and Mean-Squared-Error

- Since each estimate of θ is dependent on the data, we can consider them statistics with inherent randomness based on the variability of our data
- We can subsequently evaluate the accuracy of our estimate

Bias

  - Bias is given by $\mathbb{E}_\theta[\hat{\theta}] - \theta$
  - Bias is a measure of how close the average value of our estimate is to the true parameter

Standard Error

  - Standard Error is given by $\sqrt{Var_\theta[\hat{\theta}]}$
  - Standard error is a measure of how variable our estimate is around the true value

Mean-Squared Error

- MSE is given by $\mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$
- $MSE = Variance + Bias^2$

- An estimate is said to be unbiased if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all possible $\theta \in \mathbb{R}^k$
- Jensen's inequality: $\mathbb{E}[g(Y)] > g(\mathbb{E}[\mathbb{Y}])$ for strictly convex functions

Maximum Liklihood Estimator

- For $X_1, X_2, ..., X_n \sim f(x|\theta)$ we define the joint PDF/PMF as the likelihood function: $lik(\theta) = f(X_1|\theta) \times f(X_2|\theta) \times ... \times f(X_n|\theta)$
- We define the Maximum Likelihood Estimator as the θ that maximizes this function
- Computing the MLE is an optimization problem that often uses the log-likelihood function: $l(\theta) = \sum_{i=1}^{n} log\ f(X_i|\theta)$
- Always verify that the MLE estimate exists within the parameter space and that the extremum point is indeed a maximum

Newton Raphson Method

1. Begin with an initial guess: $\alpha^{(0)}$ (Often times MOM estimator)
2. Iterate $\alpha^{(t+1)}$ via $\alpha^{(t+1)} = \alpha^{(t)} - \frac{f(\alpha^{(t)})}{f'(\alpha^{(t)})}$
3. If $\alpha^{(t+1)}$ extends outside of the parameter space, reset it to a small value within it

- For multiple parameters we generalize this form via: $\theta^{(t+1)} = \theta^{(t)} - (\nabla^2 l(\theta^{(t)}))^{-1} \nabla l(\theta^{(t)})$

Lagrange Multipliers

- Lagrange multipliers are used to optimize under a given constraint
- We define the Lagrangian as: $L(\theta, \lambda) = \sum_{i=1}^{n} log\ f(X_i|\theta) + \lambda(constraint\ equation)$
- For each θ_i we solve $0 = \frac{\partial L}{\partial \theta_i}$ finding θ_i as a function of λ
- We then substitute our expressions for each θ_i into $0 = \frac{\partial L}{\partial \lambda}$ to solve for λ via our initial constraint
- Finally, we use our expression for λ to find θ_i from our original equations

Confidence Intervals

- If we have the bias and variance of our estimate we can make an asympotiotic prediction on the statistical behavior of our estimate
- Specifically, by the Law of Large Numbers we have that $\lambda \to \mathbb{E}_{\lambda_0}[\hat{\lambda}]$ as $n \to \infty$
- We can then invoke the Central Limit Theorem: $\sqrt{n}(\hat{\lambda} - \mathbb{E}_{\lambda_0}) \to N(0, Var_{\lambda_0}[\hat{\lambda}] * n)$
- Alternatively, this reveals that for large n, the distribution of $\hat{\lambda}$ is apprximately $N(\mathbb{E}_{\lambda_0}, Var_{\lambda_0}[\hat{\lambda}])$
- We can use this distribution to construct a **confidence interval** for our estimate
- We define $z(\alpha/2)$ to be the upper alpha point of the standard normal distribution.
- By defining our coverage interval as : $-SE \cdot z(\alpha/2) \leq \hat{\lambda} - \mathbb{E}_{\lambda_0} \leq SE \cdot z(\alpha/2)$ we ensure that our estimate will be in the interval with probability $1 - \alpha$ for large n

Fisher Information

- Under smoothness conditions a Theorem tells us that as $n \to \infty$ we have two properties of the MLE of the model
  (a) The MLE is consistent, $\hat{\theta} \to \theta_0$ in probability
  (b) The MLE is asmptotically normal and $\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \frac{1}{I(\theta_0)})$
- We define $I(\theta)$ as the Fischer Information where $I(\theta) = Var_\theta\left[\frac{\partial}{\partial\theta} logf(X|\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} logf(X|\theta)\right]$
- We also call the quantity $\frac{\partial}{\partial\theta} logf(X|\theta)$ as the score
- We have that $\hat{\theta}$ is asympotiotically unbiased where the bias of $\hat{\theta}$ is less than the order $\frac{1}{\sqrt{n}}$ such that $\sqrt{n}(\hat{\theta} - \theta)$ converges to a distribution with mean 0
- Our standard error is approximately $\sqrt{\frac{1}{nI(\theta_0)}}$
- Notice that our standard error is on the order $\frac{1}{\sqrt{n}}$ which means it is the key contributer to the mean-squared error
- Finally, we have that under the true parameter θ_0 the distribution of $\hat{\theta}$ is approximately $N(\theta_0, \frac{1}{nI(\theta_0)})$
- We can then use this normal approximation to construct confidence intervals for our MLE estimates
- For example, our coverage interval for θ with coverage $1 - \alpha$ is $\hat{\theta} \pm \sqrt{\frac{1}{nI(\hat{\theta})}} \cdot z(\alpha/2)$

Geometrical Interpretation

- We can also consider the Fischer Information to be the curvature around the true parameter
- Therefore, large Fischer Information values indicate that small perturbations of θ away from θ_0 lead to large decreases in the log-likelihood
- Alternatively, a large Fischer Information can be interpreted as our data containing more "information" about the parameter

Fisher Information for Multiple Parameters

- For a model with k parameters we extend the Fischer Information into a k x k Fischer Information matrix where $I(\theta)_{ij} = Cov_\theta\left[\frac{\partial}{\partial\theta_i} logf(X|\theta), \frac{\partial}{\partial\theta_j} logf(X|\theta)\right] = \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} logf(X|\theta)\right]$
- Now, under the assumption that $I(\theta)$ is invertible we have $\sqrt{n}(\hat{\theta} - \theta) \to N(0, I(\theta)^{-1})$

Fisher Information for Non-Identically Distributed Observations

- In the case that we are dealing with non-identially distributed data, we can define the Fischer Information as $I_Y(\theta) = Var_\theta[l'(\theta)] = -\mathbb{E}[l''(\theta)]$
- We can then approximate our MLE via $N(\theta_0, I_Y(\theta_0)^{-1})$
- For multiple parameters we use $I_Y(\theta)_{ij} = Cov_\theta\left[\frac{\partial}{\partial\theta_i} l(\theta), \frac{\partial}{\partial\theta_j} l(\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta)\right]$

Delta Method

- The Delta Method is used to quantify the uncertainity of a plug-in estimates
- If a function $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at $\theta \in \mathbb{R}$ and if $\sqrt{n}(\hat{\theta} - \theta) \to N(0, v(\theta))$ in distribution, then as $n \to \infty$ we have $\sqrt{n}(g(\hat{\theta}) - g(\theta)) \to N(0, g'(\theta)^2 v(\theta))$
- We can quantify the uncertainity in our MLE via the Fisher Information
- The delta method can be applied to find the standard errors for method-of-moments estimates
- We find that $\sqrt{n}(\hat{\theta}-\theta) \to N(0, g'(h(\theta))^2 \cdot v(\theta))$ which yields a $1-\alpha$ confidence interval of $\hat{\theta} \pm \sqrt{\frac{g'(h(\hat{\theta})^2 \cdot v(\hat{\theta}))}{n}} \cdot z(\alpha/2)$ where h is the function relating the theoretical mean to θ and g is its inverse

Cramer-Rao Bound and Asymptotic Efficiency

- **Cramer-Rao Lower Bound** states that any unbiased estimator of θ must have a variance of at least $\frac{1}{nI(\theta)}$
- For two estimators $\hat{\theta}$ and $\tilde{\theta}$ from the same data that satisfy $\sqrt{n}(\hat{\theta}-\theta) \to N(0, u(\theta))$ and $\sqrt{n}(\tilde{\theta}-\theta) \to N(0, v(\theta))$ we refer to the ratio of their variances as **asymptotic relative efficiency**
- This value can be interepreted as the ratio of the sample sizes needed for the two estimators to have the same variance
- An estimator is said to be asympotiotically efficient if $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \frac{1}{I(\theta)})$
- The MLE is asympotiotically efficient but the MOM estimators are not necessarily so
- Cramer-Rao bound holds for plug-in estimators where we have $\frac{g'(\theta)^2}{nI(\theta)}$ as the new lower bound

Bayesian Inference

- Unlike the frequentist paradigm of statistical inference, in the Bayesian paradigm we treat θ as a random variable
- Key Tenant of Bayesian Inference is the expression: $f_{\Theta|X}(\theta|x) \propto f_{(X|\Theta)}(x|\theta) f_{\Theta}(\theta)$ which can be interpreted as the Posterior distribution is proportional to the product of Likelihood and Prior
- To extract a singular estimate for our parameter we consider the posterior mean and posterior mode of our posterior Distributions
- Often times confidence intervals are is just the upper-$\alpha/2$ and lower-$\alpha/2$ points of our posterior distributions
- Bayes mean can be interpretted as a weighted average between the sample mean and the prior mean
- A prior distribution is said to be a conjugate prior if the resulting posterior distribution for a given model is in the same family as the Porior
- Improper priors (prior distributions that do not describe valid PDFs) can be used in bayesian analysis to produce valid posterior distributions
- In contrast to the frequentist approach, the Bayesian model assigns randomness to the parameter rather than the data
- Bayesian is conditional on data!
- Bayesian credible interval does not guarantee frequentist coverage for a fixed true parameter, but it does for the average case
- The influence of prior distributions diminishes as n increases
- For large n the frequentist and Bayesian approaches converge. Specifically, the posterior distribution from the Bayesian approach resembles a normal distribution centered at the MLE with variance $\frac{1}{nI(\hat{\theta})}$

Kullback-Leibler Divergence

- The Kullback-Leibler Divergence is a notion of how "close" two probability distributions are
- The KL Divergence from f to g is given by $D_{KL}(g||f) = \int g(x) log \frac{g(x)}{f(x)} dx = \mathbb{E}[log \frac{g(x)}{f(x)}] = \mathbb{E}[log\ g(X)] - \mathbb{E}[log\ f(X)]$
- KL Divergence is asymmetric where $D_{KL}(g||f) \neq D_{KL}(f||g)$
- The KL Divergence for a parameter θ and its true parameter $\hat{\theta}$ is approximately $\frac{I(\theta_0)}{2}(\theta - \theta_0)^2$
- In a mis-specified model the MLE converges to the value of θ that minimizes $D_{KL}(g(x)||f(x|\theta))$
- Variance is no longer given by our Fischer Information metric when we estimate our parameter by minimizing the KL Divergence
- Instead, we can use the delta method to find the variance

Bootstrap

- Bootstrap methods are a series of computational methods used to estimate statistics or develop confidence intervals from a given dataset
- The main idea behind the Bootstap is to simulate new data based on the existing data

Parametric Bootstrap

- The parametric bootstrap method estimates the model $f(x|\theta)$ via $f(x|\hat{\theta})$
- It is akin to using the plug-in principle to simulate new draws and then calculating your statistic from the new draws

Nonparametric Bootstrap

- Instead of assuming a parametric model, in this method we sample our existing data values with replacement
- Continue to make n draws, but now some of the values may be repeated
- The estimated distribution from the nonparametric bootstrap is an empirical distribution
- While empirical distributions are always discrete and do not provide useful information about the mode, max value, or min value, but they can be useful tools to estimate CDF, mean, and variance
- Guards against model misspecification

Bootstrap Confidence Intervals

- For an estimate $\hat{\theta}$ we can estimate $\hat{se}$ via a bootstrap method and use the confidence interval: $\hat{\theta} \pm z(\alpha/2) \cdot \hat{se}$
- For B bootstrap simulations we have $\hat{\theta_1^*}, ...\hat{\theta_B^*}$. We can define our confidence interval by looking at the $\alpha/2$ and $1 - \alpha/2$ percentiles
- Finally, let $q^{\alpha/2}$ and $q^{1-\alpha/2}$ be the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\hat{\theta_1^*} - \hat{\theta}, ...\hat{\theta_B^*} - \hat{\theta}$ from which we define our interval $[\hat{\theta} - q^{1-\alpha/2}, \hat{\theta} - q^{\alpha/2}] = [2\hat{\theta} - \hat{\theta}^{*(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^{*(\alpha/2)}]$
- Basic bootstrap accounts for bias better than the percentile Bootstrap

Generalized Likelihood Ratio Test

- Suppose we have the parametric model $f(x|\theta)$ and a null parameter θ_0 for n IID observations. We want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$
- We define the **Generalized Likelihood Ratio Test (GLRT)** to be the test that rejects H_0 for small values of $\Lambda = \frac{lik(\theta_0)}{max_\theta lik(\theta_0)}$
- This statistic is testing whether the likelihood of the MLE differs from the likelihood of our null parameter by an amount greater than random chance
- GLRT can equivalently be rephrased to reject H_0 for large values of $-2log\Lambda = 2l(\hat{\theta}) - 2l(\theta_0)$
- GLRT rejects at level α when $2log\Lambda \geq \chi^2_k(\alpha)$ where k is the dimension of the parameter space
- When we are dealing with multidimensional parameters we can further generalize the likelihood ratio test to evaluate sub-models
- For $H_0 : \theta \in \Omega_0$ vs. $H_1 : \theta \notin \Omega_0$ we define $\Lambda = \frac{max_{\theta \in \Omega_0} lik(\theta)}{max_{\theta \in \Omega} lik(\theta)}$ where we reject using the χ^2_k distribution. However, unlike before, k is now the difference in dimensions between the full model Ω and the sub-model Ω_0

Test of Independence

	Dem	Rep	Independent
female	422	381	273
male	299	365	232

- We want to model each of the counts p_{ij} as multinomial with 1972 total observations where i = 1,2 and j = 1,2,3
- We also have the constraints that the sum across i and the sum across j are both equal to 1
- Our null hypothesis is that there is no association between gender and party affiliations. Alternatively, $H_0 : p_{ij} = p_{i.}p_{.j}$ where $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$
- These constraints satisfiy a sub-model of a full multinomial model with 3 dimensions (5 variables and 2 equations of constraint)
- Proceed with GLRT as normal

Pearson chi-squared test

- The Pearson chi-squared test is an alternative to the GLRT that is often used in multinomial testing problems
- Uses the test statistic $X^2 = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i}$
- Has the same asympotiotic distribution as the GLRT

The Bradley-Terry Model

- Let β_i represent the strength of team i
- We model the outcome of a game between teams i and j as a Bernoulli random variable with distribution Bernoulli(p_{ij}) where we define $\beta_i - \beta_j$ as the log odds of p_{ij}
- This yields the following equations: $log\frac{p_{ij}}{1-p_{ij}} = \beta_i - \beta_j$ and $p_{ij} = \frac{e^{\beta_i - \beta_j}}{1+e^{\beta_i - \beta_j}}$
- The strengths of each team is defined relatively so we can set one $\beta_i = 0$ as a standard of comparison
- We can also encode a home team advantage by ensuring i is at home for every game (i,j) and including an intercept α

Generalized Linear Models

- Response values $Y_i, ..., Y_n$ are modeled as independent observations drawn from $Y_i \sim f(y|\theta_i)$
- For the ith response we have covariates $x_{i1}, ..., x_{ip}$ which are used to model θ_i by a link function, g, via $g(\theta_i) = \alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip}$
- The cannonical link aims to have PDF/PMF of the form $f(y|\eta) = e^{\eta y - A(\eta)} h(y)$

Common Test Statistics

Test	Hypotheses	Test Statistic	Distribution
One-sample t-test	$H_0: \ N(0, \sigma^2)$ $H_1: \ N(\mu, \sigma^2), \mu > 0$	$T = \frac{\sqrt{n}\bar{X}}{S}$	$T \sim t_n$
Sign Test	$H_0: \ f \ has \ median = 0$ $H_1: \ f \ has \ median > 0$	$S = \sum_{i=1}^{n} \mathbf{1}\{X_i > 0\}$	$S \sim Binom(n, \frac{1}{2})$ $\sqrt{4n}(\frac{S}{n} - \frac{1}{2}) \to N(0,1)(large \ n)$
Two-sample t-test	$N \quad (\mu_x, \sigma^2) \ vs. \ N(\mu_y, \sigma^2)$ $H_0 \ : \mu_x = \mu_y$ $H_1 \ : \mu_x > \mu_y$	$T = \frac{\bar{X} - \bar{Y}}{S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}}$	$T \sim t_{m+n-2}$
Welch's unequal variance Test	$N \quad (\mu_x, \sigma_x^2) \ vs. \ N(\mu_y, \sigma_y^2)$ $H_0 \ : \mu_x = \mu_y$ $H_1 \ : \mu_x > \mu_y$	$T_{welch} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n}S_x^2 + \frac{1}{m}S_y^2}}$	Similar to t distribution
Mann-Whitney-Wilcoxon Test	$H_0: \ f = g$ $H_1: \ f \ stochastically \ dominates \ g$	1. Pool and rank X and Y observations 2. Take note of the ranks of Y and sum their values	Central Limit Theorem
Fischer's Exact Test	$Binom \quad (n, p) \ vs. \ N(m, q)$ $H_0: \quad p = q$ $H_1: \quad p > q$	N_{A1}, number of successes in first Binom	Hypergeometric for N_{A1}

Common Distributions

Bernoulli Random Variables

- Takes 0 or 1 with probability (1-p) and p respectively
- $p(x) = p^x(1-p)^{1-x}$ for x = 0 and 1
- Can serve as an indicator variable
- Mean: p
- Variance: $p(1-p)$

Binomial Distribution

- n trials with a fixed probability of p for each trials
- Models the number of successes, X, in n trials
- $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Generalized as a multinomial distribution for multiple outcomes

Geometric and Negative Binomial Distributions

- Geoemtric distribution is an infinite sequence of Bernoullii trials (Run repeated trials until the first success)
- $p(k) = (1-p)^{k-1}p$
- Negative Binomial distribution is a generalization of the geometric distribution for r successes
- $p(k) = \binom{k-1}{r-1}p^r(1-p)^{k-r}$

Hypergeometric Distribution

- Consider a total of n balls with r black balls and n-r white ones. A hypergeometric distribution is the number of black balls drawn when selecting m balls without replacement
- $P(k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$

Poisson Distribution

- Limit of a binomial distribution as $n \to \infty$ and $p \to 0$ s.t. $np = \lambda$
- Poisson distribution with parameter λ: $P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $p(x) = p^x(1-p)^{1-x}$ for x = 0 and 1
- Can serve as an indicator variable

Uniform Random Variable

- Uniform distribution across an interval
- $f(x) = frac1b - a$ for $a \leq x \leq b$

Exponential Density

- With parameter λ: $f(x) = \lambda e^{-\lambda}$ for $x \geq 0$
- $F(x) = 1 - e^{-\lambda x}$
- Memoryless Property: probability that it will last t more time is independent of time elapsed

Gamma Density

- With parameters α and λ: $g(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}t^{\alpha-1}e^{-\lambda t}$
- $\Gamma(\alpha) = \int_0^\infty u^{x-1}e^{-u}du$ for $x > 0$
- Gamma density with $\alpha = 1$ is simply the exponential distribution
- α and λ are sometimes referred to as the shape and scale paramaters
- Mean: $\frac{\alpha}{\lambda}$
- Variance: $\frac{\alpha}{\lambda^2}$

Normal Distribution

- For a mean, μ and standard deviation, σ :$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Beta Density

- $f(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$ for $0 \leq u \leq 1$
- Useful for variables restricted to [0,1]
- a = b = 1 is the uniform distribution

Chi-squared Distribution

- Suppose $X_1, ...X_n \overset{iid}{\sim} N(0,1)$, then the distribution of $X_1^2 + ... + X_n^2$ is the chi-sqared distribution with n degrees of freedom, χ_n^2