Representations.

- $X \equiv -\log U \sim \text{Expo}$
- $U \sim \text{Beta}(1,1)$
- $B \equiv U^{1/a} \sim \text{Beta}(a,1)$
- $X_1 + \cdots + X_n \sim \text{Gam}(n)$
- $\frac{G_a}{G_a+G_b} \sim \text{Beta}(a,b)$
- $\text{Beta}(a,b) \cdot \text{Beta}(a+b,c) \sim \text{Beta}(a,b+c)$
- $\chi^2_n \sim 2\text{Gam}(n/2)$
- $Z_1^2 + \cdots + Z_n^2 \sim \chi^2_n$
- $T \equiv \frac{Z}{\sqrt{\chi^2_n/n}}$
- For $U_1, \ldots, U_n \overset{iid}{\sim} \text{Unif}$, $U_{(j)} \sim \text{Beta}(j, n-j+1)$

Descriptive Statistics.

- Sample SD: $s^2 = \dfrac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$
- Sample Covariance: $s_{xy} = \dfrac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$
- Empirical CDF: $\hat{F}_n(x) = \dfrac{1}{n}\sum_{j=1}^{n} I(x_j \le x)$

Models. A **model** is a family of probability distributions indexed by a parameter θ. The **parameter space** is the set of all θ, and we call a model **parametric** if θ is finite dimensional, and **non-parametric** if θ is infinite dimensional.

- $\{\text{Pois}(\theta) : \theta > 0\}$ is a 1-dimensional parametric model.
- If we want to write down a model, we might do something like $Y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.
- Our models are almost always flawed, but it's useful theoretically to assume that there is some real **data generating process** out there, which we are trying to approximate with our model. We'll write this as $G_Y(y)$. For right now, we make the assumption that we have correctly specified our model, so that $G_Y(y) = F_Y(y|\theta)$.

Intro Definitions.

- An object that we wish to make inference about from data is an **estimand** denoted θ.
- A **statistic** is a function of only the random variables and possibly known parameters, which we write $T(\mathbf{Y})$.
- An **estimator** $\hat{\theta} = T(\mathbf{Y})$ is a statistic used to proxy the estimand θ.
- An **estimate** is a statistic evaluated at a realization of our data $\mathbf{Y} = \mathbf{y}$ so that $T(\mathbf{y})$ is a proxy of the estimand θ based on the data $\mathbf{y}$ we have on hand.

Likelihood. Let $\mathbf{y}$ be the observed value of $\mathbf{Y}$. The likelihood function is a function of θ defined

$$L(\theta; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}|\theta).$$

This is reminiscent of Bayes' Rule

$$g(\theta|y) = \frac{f(y|\theta)g(\theta)}{f(y)} \propto f(y|\theta)g(\theta) = \text{likelihood} \cdot \text{prior}.$$

Also, we have the **log-likelihood**

$$\ell(\theta; \mathbf{y}) = \log L(\theta; \mathbf{y}).$$

Invariance property of likelihood. We have the likelihood function $L(\theta; \mathbf{y})$ with $\psi = g(\theta)$ an injective function, then $L(\psi; \mathbf{y}) = L(\theta; \mathbf{y})$.

Method of Moments. We set the sample moments equal to the theoretical moments, so

$$\mathbb{E}[Y_i] = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$\mathbb{E}[Y_i^2] = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

Then we solve for the parameters.

Maximum Likelihood Estimator. The MLE of θ is $\hat{\theta}$ that maximizes the likelihood function.

- $Y \sim \text{Bin}(n,p)$ with n known, p unknown.

$$L(p; y) = p^y(1-p)^{n-y}$$
$$\implies \ell(p; y) = y\log p + (n-y)\log(1-p)$$
$$\implies \frac{y}{p} - \frac{n-y}{1-p} = 0$$
$$\implies \hat{p} = \frac{y}{n}.$$

- For n unknown parameters, we let

$$\frac{\partial \ell}{\partial \theta_1} = \cdots = \frac{\partial \ell}{\partial \theta_n} = 0.$$

Some properties of the MLE are as follows

- **Invariant.** If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
- **Consistent.** $\hat{\theta}$ converges in probability to θ.
- Asymptotically normal as $n \to \infty$.
- Asymptotically unbiased as $n \to \infty$.
- Asymptotically efficient (no other asymptotically unbiased estimator will have a lower standard error asymptotically).

Evaluating an Estimator.

- **Bias.** $\mathbb{E}[\hat{\theta}] - \theta$
- **SE.** $\sqrt{\text{Var}\,\hat{\theta}}$
- **MSE.** $\mathbb{E}[\hat{\theta} - \theta]^2$

The **bias-variance tradeoff** is illustrated by the fact that the MSE is equal to bias^2 + variance.

An estimator $\hat{\theta}$ is **consistent** for the estimand θ if $\hat{\theta} \overset{P}{\to} \theta$ as $n \to \infty$. That is, for any $\epsilon > 0$, $P(|\hat{\theta} - \theta| > \epsilon) \to 0$.

The **Kullback-Leibler divergence** is a metric between distributions,

$$KL(f_0, f_1) = \int f_0 \log \frac{f_0(x)}{f_1(x)}\, dx \ge 0.$$

CMT. $X_n \overset{D}{\to} X_0$ implies that $h(X_n) \overset{D}{\to} h(X_0)$ for all continuous functions.

Slutsky. If $X_n \overset{D}{\to} X$ and $Y_n \overset{D}{\to} c$, then $X_nY_n \overset{D}{\to} cX$, and $X_n + Y_n \overset{D}{\to} X + c$.

Change of Variables. $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$.

The **score function** is defined as follows

$$s(\theta; \mathbf{y}) = \frac{\partial \log L(\theta; \mathbf{y})}{\partial \theta} = \frac{1}{L(\theta; \mathbf{y})} \frac{\partial L(\theta; \mathbf{y})}{\partial \theta}.$$

Some properties of the score function are

$$\mathbb{E}[s(\theta^*; \mathbf{Y})] = 0$$
$$\mathrm{Var}(s(\theta^*; \mathbf{Y})) = -\mathbb{E}[s'(\theta^*; \mathbf{Y})]$$

where s' denotes the partial derivative wrt θ. The **Fisher information** for a parameter θ is

$$\mathcal{I}(\theta) = \mathrm{Var}_\theta(s(\theta; \mathbf{Y})).$$

When the sample size is n, we write $\mathcal{I}_n(\theta)$. When Y_i are iid, then $\mathcal{I}_n(\theta) = n \cdot \mathcal{I}_1(\theta)$.

We can also reparameterize. Let $\tau = g(\theta)$, where g differentiable with $g'(\theta) \neq 0$. Then

$$\mathcal{I}(\tau) = \frac{\mathcal{I}(\theta)}{(g'(\theta))^2}.$$

The **Cramer-Rao lower bound** provides a fundamental limit to how well we can estimate a parameter unbiasedly, based on Fisher information. Let $\hat{\theta}$ be an unbiased estimator of θ. Under regularity conditions,

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta^*)}.$$

If $\hat{\theta}$ is such that $\mathbb{E}[\hat{\theta}] = g(\theta^*)$, then under regularity conditions,

$$\mathrm{Var}(\hat{\theta}) \geq \frac{(g'(\theta^*))^2}{\mathcal{I}(\theta^*)}.$$

Maximum Likelihood CLT. Let our data be iid $Y_1, \ldots, Y_n$, and let $\hat{\theta}_{MLE}$ be the MLE for θ. If we have the following **regularity conditions**

- The support of Y_i does not depend on θ
- θ^* does not lie on the boundary of the parameter space. For example, if $\Theta = [0, \infty)$ and $\theta^* = 0$, then weird things happen.
- $f_\theta(y)$ should be smooth.
- DUtHIS
- θ is of fixed dimension.

Then we have the asymptotic distribution of $\hat{\theta}$ is given by

$$\sqrt{n}(\hat{\theta} - \theta^*) \to \mathcal{N}(0, \mathcal{I}_1^{-1}(\theta^*)).$$

As an approximation,

$$\hat{\theta} \sim \mathcal{N}\left(\theta^*, \frac{1}{n\mathcal{I}_1(\theta^*)}\right).$$

Natural Exponential Families. The Normal (σ^2 known), Poisson, Gamma (a known), Binomial (n fixed), and Negative Binomial models can be unified into one framework. A model with density $f_\theta(y)$ is a **natural exponential family** if we can write

$$f_\theta(y) = e^{\theta y - \psi(\theta)} h(y)$$

where h does not depend on θ. The parameter θ is called the natural parameter, which may be a reparameterization of how the model was originally specified.

NEF Fun Facts.

- Let $Y_1, \ldots, Y_n$ be iid rvs from the NEF, then the joint density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is
$$f_\theta(\mathbf{y}) = e^{n(\theta \bar{y} - \psi(\theta))} h_n(\mathbf{y})$$
where $h_n(\mathbf{y}) = h(y_1) \cdots h(y_n)$.
- Let Y follow the NEF $f_\theta(y) = e^{\theta y - \psi(\theta)} h(y)$. Then $\mathbb{E}[Y] = \psi'(\theta)$ and $\mathrm{Var}(Y) = \psi''(\theta)$.
- $\bar{Y}$ is a sufficient statistic for θ.
- Let $\mu = \mathbb{E}[Y] = \psi'(\theta)$, then the MLE of μ is $\hat{\mu} = \bar{Y}$.
- $\mathcal{I}_1(\theta) = \psi''(\theta)$

A model with density $f_\theta(x)$ is an **exponential family** if we can write

$$f_\theta(x) = e^{\theta T(x) - \psi(\theta)} g(x)$$

where g does not depend on θ. Every NEF is an EF (when $T(x) = x$), but not every EF is an NEF.

Sufficient statistics. A statistic T is **sufficient** for θ if the conditional distribution of $(Y_1, \ldots, Y_n) \mid T$ does not depend on θ. We can find sufficient statistics using the **factorization criterion**

$$f_\theta(\mathbf{y}) = g_\theta(\mathbf{t}) h(\mathbf{y})$$

where $\mathbf{t}$ is the observed value of $\mathbf{T}$ and h does not depend on θ.

Rao-Blackwell. Let $\hat{\theta}$ be an estimator for θ and T a sufficient statistic for θ. Then the **Rao-Blackwellized estimator** is $\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta}|T]$, which is better or equal to $\hat{\theta}$ in MSE, and strictly better unless $\hat{\theta}$ is a function of T.

Confidence Interval Miscellany. Note that for $Z \sim \mathcal{N}(0, 1)$, we have $P(-1.96 \leq Z \leq 1.96) \approx 0.95$. To find a confidence interval, we can use:

- Asymptotics
  - CLT
  - CLT for MLE
  - Delta Method
- Distributional Calculation
  - Pivot: A quantity whose distribution does not depend on an unknown parameter. Usually, pivotal quantities involve theta.

CLT. Let $X_i \overset{iid}{\sim} [\mu, \sigma^2]$, then as $n \to \infty$,

$$\sqrt{n}(\bar{X} - \mu) \overset{D}{\to} \mathcal{N}(0, \sigma^2).$$

LLN. $\overline{X}_n \to \mu$. Weak is $\overset{P}{\to}$ and Strong is $\overset{as}{\to}$. An assumption is X_i independent, $\mathbb{E}[X_n] = 0$ with $\mathbb{E}[X_n^2] \leq M < \infty$.

Delta Method. If $\sqrt{n}(T_n - \theta_0) \overset{D}{\to} Z$, where Z is a rv (typically Normal) then

$$\sqrt{n}(g(T_n) - g(\theta_0)) \overset{D}{\to} g'(\theta_0) Z.$$

Inequalities.

- Markov $P(X > a) < E(\varphi(X))/\varphi(a)$.
- Chebyshev $P(|X - \mathbb{E}X| \geq c) \leq \mathrm{Var}(X)/c^2$.
- Chernoff $P(X \geq a) \leq \mathbb{E}[e^{tX}]/e^{ta}$.
- Cauchy-Schwarz $\mathbb{E}|X_1 X_2| \leq \sqrt{EX_1^2}\sqrt{EX_2^2}$.
- Jensen $\mathbb{E}g(X) \geq g(\mathbb{E}X)$ if g convex.

Interval estimation.

- **Coverage probability.** The proportion of the time that the interval contains the true value of interest.
- **Confidence interval.** Predicting the response of some statistic (usually the mean). We interpret the confidence level by "When repeating the process of generating the confidence interval, we expect $1-\alpha$ of the confidence intervals to contain the true parameter value." Parameter is fixed, and interval is random.

  Confidence intervals capture the uncertainty about the interval we have obtained (i.e., whether it contains the true value or not). Thus, they cannot be interpreted as a probabilistic statement about the true parameter values.
- **Credible interval.** With $f(\theta)$ prior and $f(\theta|y)$ posterior, we have $P(\theta \in (c_L, c_U)|y) \geq 1-\alpha$. We interpret as: "There is a 90% chance that the parameter θ lies in our credible interval based on our prior beliefs and data." Interval is fixed and parameter is random.

  Credible intervals capture our current uncertainty in the location of the parameter values and thus can be interpreted as probabilistic statement about the parameter.
- **Prediction interval.** Predicting the value of some new response. Wider than a confidence interval because there is more uncertainty in predicting a single response than the mean.
- **Bootstrap confidence interval.** Can be created with asymptotics or with percentile.

Hypothesis Testing.

- **Null and Alternative hypotheses.** We have a parameter $\theta \in \Theta$ parameter space. We partition the parameter space into Θ_0, Θ_1. The null is $H_0 : \theta \in \Theta_0$ and alternative is $H_1 : \theta \in \Theta_1$.
- **Type I error.** Reject null when it's true.
- **Type II error.** Do not reject null when it's false.
- **Power function.** We select a rejection region R before collecting data, and the power of our test is

$$\beta(\theta) = P(\vec{Y} \in R|\theta).$$

  This is the ability to reject H_0. We can compute the Type I error rate (the size/level of a test) as

$$\alpha = \max_{\theta \in \Theta_0} \beta(\theta).$$

- **Z-test.** For a one or two-sided test for hypothesis about θ, we can use a z-statistic

$$t(Y) = \frac{\hat{\theta} - \theta}{\hat{\sigma}/\sqrt{n}}$$

  where $\hat{\theta}$ and $\hat{\sigma}$ are consistent estimators for θ and the std dev of $\sqrt{n} \cdot \hat{\theta}$ respectively. Apply CLT.
- **t-test.** The t-statistic is any parameter of the form

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})}.$$

- **Wald test.** Assess the constraints based on the squared difference between the estimated and hypothesized parameter values, weighted by the precision of the estimate. Intuitively, the larger this weighted distance, the less likely it is that the constraint is true. The test statistic is

$$W = \frac{(\theta - \theta_0)^2}{\text{Var}(\hat{\theta})}.$$

  When H_0 is true, $W \sim \chi^2_1$. Note that $\sqrt{W}$ can be compared to $\mathcal{N}(0,1)$.
- **Score test.** Near the maximum of the likelihood function, the gradient of the likelihood function (the score function) evaluated at the restricted estimator should be close to zero. Our test statistic to test $H_0 : \theta = \theta_0$ is

$$S(\theta_0) = \frac{s(\theta_0; \mathbf{y})^2}{I(\theta_0)}.$$

  This has an asymptotic distribution of χ^2_1 when H_0 is true.
- **Likelihood ratio test.** A high quality test (of all tests with same Type I error, the lowest Type II error rate) with estimator

$$\Lambda(Y) = 2\log\left(\frac{\max_{\theta \in \Theta} L(\theta; Y)}{\max_{\theta \in \Theta_0} L(\theta; Y)}\right)$$

$$= 2\log\left(\frac{L(\hat{\theta}; Y)}{L(\hat{\theta}_0; Y)}\right)$$

  This is close to 1 when null is true, and substantially greater than 1 when it is false. Based on the MLE for the parameter. Efficient for large samples.

  When dimension of Θ is p and dimension of Θ_0 is p_0, then if H_0 is true, under mild conditions,

$$\Gamma(Y) \xrightarrow{D} \chi^2_{p-p_0}.$$

Linear Regression.

- We have a model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with $\mathbb{E}[\epsilon_i | X_i = x] = 0$ and $\text{Var}(\epsilon_i | X_i = x) = \sigma^2$. We can find $\hat{\beta}_0$ and $\hat{\beta}_1$ for which we can get $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. We get the **residuals** $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, and the RSS measures how well the line fits the data $\sum \hat{\epsilon}_i^2$.
- We can compute the **least squares estimator** for $\hat{\beta}_0$ and $\hat{\beta}_1$ in which we minimize

$$\sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2.$$

  We get

$$\hat{\beta}_1 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\hat{\sigma}^2 = \frac{1}{n-2} \sum \hat{\epsilon}_i^2$$

  where $\hat{\sigma}^2$ is unbiased for σ^2. $\hat{\beta}_1$ is also unbiased, and

$$\text{Var}(\hat{\beta}_1 | X = x) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

- When we suppose that $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, then $Y_i | X_i = x_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. The MLE for σ^2 is $\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2$. The one with $n-2$ is unbiased, but has a higher MSE. The MLE for β_1, β_0 is the same. Additionally, we have the following MLE distributions

$$\hat{\beta}_1 | \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$
$$\hat{\beta}_0 | \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)\right).$$

- Here's the student-t trick. We have

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma / s_{xx}} \sim \mathcal{N}(0,1) \text{ with } s_{xx}^2 = \sum (x_i - \bar{x})^2$$
$$V = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

and $V \perp\!\!\!\perp Z$, so we have

$$\sqrt{n-2} \cdot \frac{Z}{\sqrt{V}} \sim t_{n-2}.$$

Sampling.

- In the **design based approach** to sampling we treat every element of the population as a fixed value rather than a random variable.
- **Simple random sample.** All $\binom{N}{n}$ samples are equally likely. We define the following

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$$
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \mu)^2.$$

Then $\mathbb{E}[\bar{Y}] = \mu$, $\text{Cov}(Y_1, Y_2) = -\sigma^2/(N-1)$, and

$$\text{Var}(\bar{Y}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.$$

- **Stratified sampling.** Consider L strata of sizes N_i for $i = 1, \ldots, L$. Each subgroup i has a mean of μ_i and variance of σ_i^2 defined as in SRS. We do a simple random sample of size n_i for $i = 1, \ldots, L$. We then can define

$$\bar{Y}_{\text{stratified}} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{Y}_i.$$

Then we have

$$\mathbb{E}[\bar{Y}_{\text{stratified}}] = \mu$$
$$\text{Var}(\bar{Y}_{\text{stratified}}) = \sum_{i=1}^{L} \frac{N_i}{N} \text{Var}(\bar{Y}_i).$$

- **Horvitz-Thompson estimator.** Suppose our estimand is $\tau = \sum_{i=1}^{N} y_i$. Then the HT estimator is

$$\hat{\tau} = \sum_{i \in S} \frac{Y_i}{\pi_i}$$

where S is the sample, and π_i is the probability that Y_i is included in the sample. This estimator is unbiased. If N is known, then $\hat{\tau}/N$ is an unbiased estimator for the population mean μ.

Resampling.

- **Permutation test.** Let $X_1, \ldots, X_m \overset{iid}{\sim} F_x$ and $Y_1, \ldots, Y_n \overset{iid}{\sim} F_y$ be two independent samples. We test $H_0 : F_x = F_y$ vs $H_1 : F_x \neq F_y$. We have a test statistic T. We compute t_0 from the data, then for each permutation of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ we compute $t_1, \ldots, t_{(m+n)!}$. The p-value is

$$P(T \geq t_0) = \frac{1}{(m+n)!} \sum_{j=1}^{(m+n)!} I(t_j \geq t_0).$$

- **Bootstrap Procedure.** You have some iid $Y_1, \ldots, Y_n$ and calculate an estimator $\hat{\theta}$. We can create D bootstrapped estimates by doing for $i = 1, \ldots, D$:
  - Resample n elements with replacement from our sample $Y_1, \ldots, Y_n$ and call them $Y^*_{i,1}, \ldots, Y^*_{i,n}$.
  - Compute θ_i^* from $Y^*_{i,1}, \ldots, Y^*_{i,n}$.

  We can then use these D bootstrapped estimates to learn about the distribution of $\hat{\theta}$.
- **Non-parametric bootstrap.** We have a sample coming from true CDF F for which we compute the estimate $\hat{\theta}$. Then we approximate the CDF using the ECDF $\hat{F}$ with

$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x).$$

We generate bootstrapped samples from this $\hat{F}$ to understand the distribution of $\hat{\theta}$.
- **Parametric bootstrap.** Same as above, but we approximate the CDF using $\hat{F} = F_{\hat{\theta}}$.
- **Bootstrap SE.** The sample standard deviation $\hat{S}$ of $\{\theta_i^*\}$ is an estimator for $\text{SE}(\hat{\theta})$.
- **Bootstrap CI.** We can use asymptotics to get $\hat{\theta} + 1.96\hat{S}$, or we can use the quantiles of the bootstrapped replications, though this might not be great when $\hat{\theta}$ is biased.

Also, we know $\dfrac{\hat{\theta} - \theta}{\text{SE}(\hat{\theta})}$ has approximately the same distribution as $\dfrac{\hat{\theta}^* - \hat{\theta}}{\text{SE}(\hat{\theta}^*)}$, so we use bootstrap within bootstrap to get $\text{SE}(\hat{\theta}^*)$ then use the quantiles of the latter.

Causal Inference.

- In general, $\alpha \neq \theta$, but when we randomize, that is $0 < P(X = 1) < 1$, then $\alpha = \theta$. Then we have $\hat{\theta} = \bar{Y}_1 - \bar{Y}_0$ is an unbiased estimator for $\hat{\alpha}$.
- In the **design-based approach**, the statistical model is known, and the focus is on the experiment. In the **population model-based approach**, the model is unknown, which is criticized because your sample may not be representative of the entire population, and your procedure may have bias.
- We use the **potential outcomes framework** for which we label our potential outcomes $Y_i(1)$ and $Y_i(0)$ – there are two potential observable Y values for each individual, one for treatment world and one for no treatment world. We can only observe <u>one</u>! Then we have $Y_i = Y_i(0)1_{X_i=0} + Y_i(1)1_{X_i=1}$ with

$$\theta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1) - Y(0)]$$
$$\alpha = \mathbb{E}[Y(1)|X_i = 1] - \mathbb{E}[Y(0)|X_i = 0].$$

- A solution to the problems that come with population based approaches is the **sharp null hypothesis**. We assign n units to treatments $A(X = 1)$ and $B(X = 0)$ randomly and we consider $(Y_i(1), Y_i(0))$ with the sharp null hypothesis $H_0 : Y_i(1) = Y_i(0)$. We then perform a **randomization test**, which if $P(X_i = 1)$ is the same for all i, is a permutation test. The procedure is as follows:
  1. Compute all possible ways individuals could have been randomized as per our procedure.
  2. For each way, compute a metric (test statistic) of the treatment difference.
  3. p-value is the proportion of those test statistics that equal or exceed the observed.

Bayesian Inference.

- In the **Bayesian framework** we treat θ as a rv. We have a model/likelihood along with a prior distribution. We want to find the posterior distribution of θ given the data. We just apply Bayes' Rule!
- We have some estimators.
  - **Posterior mean.**
$$\hat{\theta}_{PM} = \mathbb{E}[\theta|y] = \int \theta f(\theta|y)\, d\theta.$$
This minimizes the average posterior squared loss $\mathbb{E}[(\theta - \hat{\theta})^2|y]$.
  - **Posterior mode (MAP).**
$$\hat{\theta}_{MAP} = \max_{\theta} f(\theta|y).$$
To compute this we can maximize the log prior $\log f(\theta|y) = \log L(\theta; y) + \log f(\theta)$.
  - **Posterior median.**
$$\hat{\theta}_{M} = F^{-1}_{\theta|y}(1/2).$$
This minimizes the average posterior absolute loss $\mathbb{E}[|\theta - \hat{\theta}||Y]$.
- Let $\mathcal{G}$ be some family of PDFs, and suppose your prior $g(\theta) \in \mathcal{G}$. You observe data from some distribution $f(x|\theta)$. We say g is a **conjugate prior** for the likelihood f iff
$$g(\theta|x) \propto g(\theta) \cdot f(x|\theta) \in \mathcal{G}.$$
- We have the **Beta-Binomial Conjugacy** which has prior $\mathrm{Beta}(a, b)$ and observed k of n trials to have posterior $\mathrm{Beta}(a + k, b + n - k)$.
- We have the **Gamma-Poisson Conjugacy** which has prior $\mathrm{Gam}(r, n)$ with $f(y|\theta) \sim \mathrm{Pois}(\theta)$, then we have posterior $\mathrm{Gam}(y + r, n + 1)$.
- Generalizing to the **Exponential Family Conjugate Priors**. Let $Y_1, \ldots, Y_n$ follow the NEF
$$f(y|\theta) = \exp(\theta y - \psi(\theta))h(y).$$
Assume $Y_1, \ldots, Y_n$ independent conditioned on θ, so the likelihood function is $L(\theta|y) = \exp(n(\theta\bar{y} - \psi(\theta)))$. Conjugate prior on θ is
$$\pi \propto \exp(r_0\theta\mu_0 - \psi(\theta)),$$
and the posterior mean of the mean parameter $\mu = \mathbb{E}[Y_1|\theta] = \psi'(\theta)$ is the weighted average
$$\mathbb{E}[\mu|y] = (1 - B)\bar{y} + B\mu_0$$
where $B = r_0/(r_0 + n)$.

Decision Theory.

- A **loss function** $L(\theta, \hat{\theta})$ is the loss or cost associated with using the estimate $\hat{\theta}$ when the true parameter is θ. It must hold that $L(\theta, \hat{\theta}) \geq 0$ and $L(\theta, \theta) = 0$.
  - Squared Error Loss – $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$
  - Absolute Error Loss – $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$
  - Zero-One Loss – $L(\theta, \hat{\theta}) = I(\hat{\theta} \neq \theta)$
- A **risk function** $r(\theta)$ is defined for loss function L
$$r(\theta) = \mathbb{E}[L(\theta, \hat{\theta})|\theta].$$
- An estimator $\hat{\theta}$ is **inadmissible** if there exists some other estimator $\tilde{\theta}$ dominates $\hat{\theta}$ in risk. For all θ,
$$r_{\tilde{\theta}}(\theta) \leq r_{\hat{\theta}}(\theta).$$
If $\hat{\theta}$ is not inadmissible, then it is **admissible**.
- The **James-Stein estimator** for $\mu = (\mu_1, \ldots, \mu_K)$ unknowns where we have $Y_i \sim \mathcal{N}(\mu_i, V)$ independent is defined
$$\hat{\mu}_{JS} = \left(1 - \frac{(K-2)V}{S}\right) Y_i$$
where $S = \sum_{i=1}^{n} Y_i^2$. This guy shows that the MLE is inadmissible as the risk of the JS estimator is always less than or equal to the risk of the MLE with respect to the loss function
$$L(\mu, \hat{\mu}) = \sum_{i=1}^{K} (\mu_i - \hat{\mu}_i)^2.$$