Stat 111 **Spring 2022**

Final Review

1 Estimators

1.1 MLE

1.1.1 Finding the MLE.

(a) Write down the likelihood $L(\theta; \mathbf{y})$. Drop constants and take the log to get the log-likelihood $\ell(\theta; \mathbf{y})$. *Note:* Be careful about dropping *multiplicative* constants in the likelihood vs dropping *additive* constants in the log-likelihood.

(b) Take the derivative with respect to the parameter and find the critical point by setting the derivative equal to zero.

(c) Show that the second derivative is less than 0 to verify that the critical point is a maximum.

Invariance of Likelihood: Let $\psi = g(\theta)$ be a reparameterization where g is a one-to-one function, then

$$L(\theta; \mathbf{y}) = L(\psi; \mathbf{y})$$

Invariance of the MLE: if we are trying to find the MLE of $g(\theta)$ and we have already found $\hat{\theta}_{MLE}$, we can just plug $\hat{\theta}_{MLE}$ into the function g to get the answer.
Consistency of MLE: The MLE $\hat{\theta}$ is consistent for true θ^*, i.e.,

$$\hat{\theta} \xrightarrow{p} \theta^*, \; n \to \infty.$$

Asymptotic Distribution of MLE: As $n \to \infty$, the MLE $\hat{\theta}$

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N\left(0, \frac{1}{\mathcal{I}_{Y_1}(\theta^*)}\right)$$

where $\mathcal{I}_{Y_1}(\theta^*)$ is the Fisher information for *one* observation. The MLE is asymptotically unbiased and achieves the CRLB.

1.1.2 NEFs to find the MLE:

- MLE for mean parameter $\mu = E(Y)$ is $\hat{\mu} = \bar{Y}$.
- Fisher Information $\mathcal{I}_1(\theta) = \Psi''(\theta) = \text{Var}(Y)$.

1.2 MoM

Replace expectations (estimands) with sample moments from the data (estimators); replace θ with $\hat{\theta}$.

- Mean: $g(\theta) = E(Y) \longleftrightarrow g(\hat{\theta}) = \bar{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- General k^{th} moment: $h(\theta) = E(Y^k) \longleftrightarrow h(\hat{\theta}) = \bar{\boldsymbol{Y}^k} = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$.

2 Properties of Estimators

2.1 Score Function, Fisher Info. & CRLB

2.1.1 Score function

$$s(\theta; \mathbf{Y}) = \frac{\partial \ell(\theta; \mathbf{Y})}{\partial \theta} = \ell'(\theta; \mathbf{Y}).$$
$$\mathbb{E}\left[s(\theta^*; \boldsymbol{Y})\right] = 0$$

2.1.2 Fisher Information

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \text{Var}\left(s(\theta^*; \mathbf{Y})\right) = -\text{E}\left[s'(\theta^*; \boldsymbol{Y})\right]$$

If the data are i.i.d, then $\mathcal{I}_{\mathbf{Y}}(\theta^*) = n\mathcal{I}_{Y_1}(\theta^*)$

Fisher Information under Parameter Transformation If we transform parameter θ by $\tau = g(\theta)$, where g is one-one and differentiable, we have:

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

2.1.3 Cramer Rao Lower Bound

For a biased estimator $\hat{\theta}$ with $g(\theta) = \mathbb{E}(\hat{\theta})$

$$\text{Var}(\hat{\theta}) \geq \frac{[g'(\theta^*)]^2}{\mathcal{I}_{\mathbf{Y}}(\theta^*)},$$

Note: for unbiased estimator $\hat{\theta}$, let $g(\theta) = \theta$ so $g'(\theta) = 1$.

3 Sufficiency, Rao-Blackwell, NEF

3.1 Sufficient Statistics

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be a sample from model $F_{\boldsymbol{Y}}(\boldsymbol{y}|\theta)$. A statistic $T(\boldsymbol{Y})$ is a sufficient statistic for θ if conditional distribution of $\boldsymbol{Y}|T$ does NOT depend on θ.

Theorem 3.1 (Factorization criterion)**.** $T(\mathbf{Y})$ is sufficient statistic if and only if we can factor

$$f_{\mathbf{Y}}(\mathbf{y}|\theta) = g(T(\mathbf{Y}), \theta)h(\mathbf{y}),$$

where $f_{\mathbf{Y}}(\mathbf{y}|\theta)$ is the joint PMF/PDF of $\mathbf{Y}$.

3.2 Rao-Blackwell

Let T be a sufficient statistic and $\hat{\theta}$ be any estimator for θ. Then the MSE of Rao-Blackwellized estimator

$$\hat{\theta}_{RB} = E(\hat{\theta}|T)$$

does not exceed that of the original estimator $\hat{\theta}$, i.e.,

$$\text{MSE}(\hat{\theta}_{RB}, \theta) \leq \text{MSE}(\hat{\theta}, \theta).$$

3.3 Natural Exponential Families (NEF)

A r.v. Y follows NEF if its PDF is in the form

$$f_Y(y|\theta) = e^{\theta y - \Psi(\theta)} h(y).$$

where the nonnegative function h does not depend on θ. The natural parameter θ may be a reparameterization of how the model was originally specified. $\Psi(\theta)$ is the cumulant generating function, which is the log of MGF.

Named distributions that are NEFs: Normal, Poisson, Gamma, Binomial, and Negative Binomial

Theorem 3.2. If Y is NEF in canonical form defined above:

(a) $E(Y) = \Psi'(\theta)$, $\text{Var}(Y) = \Psi''(\theta)$, corresponding to first and second cumulant. In fact, MGF $M_Y(t) = E(e^{Yt}) = e^{\Psi(\theta+t)-\Psi(\theta)}$.

(b) $\bar{Y}$ is a sufficient statistic for θ.

(c) MLE for mean parameter $\mu = E(Y)$ is $\hat{\mu} = \bar{Y}$.

(d) Fisher Information $\mathcal{I}_1(\theta) = \Psi''(\theta)$.

4 Asymptotics

4.1 Tools for asymptotics

CLT for simple estimators For simple estimators, such as the sample average, we typically use CLT to obtain a result of the form

$$X_n = \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$$

Often we want to obtain the asymptotic distribution of a more complicated estimator, and can use the following asymptotic tools:

Theorem 4.1 (Continuous Mapping Theorem)**.** If $X_1, X_2, \ldots$ are sequences of random variables and g is a continuous function, then

(i) if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$,

(ii) if $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.

Theorem 4.2 (Slutsky's Theorem)**.** If $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are sequences of random variables, such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ (c a constant), then

(i) $X_n + Y_n \xrightarrow{d} X + c$,

(ii) $X_n Y_n \xrightarrow{d} cX$,

(iii) if $c \neq 0$, then $X_n / Y_n \xrightarrow{d} X/c$.

Biohazard: In general, $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$ does NOT imply that $X_n + Y_n \xrightarrow{d} X + Y$

Theorem 4.3 (Delta Method)**.** Suppose that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$$

as $n \to \infty$ and g is a continuously differentiable function. Then as $n \to \infty$,

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\partial g(\theta)}{\partial \theta}\right)^2 \omega^2\right).$$

5 Hypothesis Testing

Null and Alternative Hypotheses: null is the status quo that we want to disprove

- One-sided: H_0: $\theta \leq \theta_0$ vs. H_1: $\theta > \theta_0$ or H_0: $\theta \geq \theta_0$ vs. H_1: $\theta < \theta_0$
- Two-sided: H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$

Power: the power function of a test is defined as

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

Type I and Type II Error:

- Type I error: $\theta \in \Theta_0$, but $\mathbf{y} \in R$, i.e. reject the null when null is true, (false positive).
  - also called the level or size of a test and is calculated as

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

- Type II error: $\theta \in \Theta_1$, but $\mathbf{y} \notin R$, i.e., fail to reject the null when null is false, (false negative).

Rejection Region: A subset of the range of data $\mathbf{y}$, where we reject H_0 if $\mathbf{y} \in R$ and fail to reject H_0 if $\mathbf{y} \notin R$.

- Often expressed in terms of a statistic, called *test statistic*, $T(\mathbf{Y})$: e.g. $R = \{\mathbf{y} : T(\mathbf{y}) > c\}$ or $R = \{\mathbf{y} : T(\mathbf{y}) < c_L \text{ or } T(\mathbf{y}) > c_U\}$, where c, c_U, c_L are *critical values*.

P-value: Let R_α be the rejection region for a test with Type I error of α. The p-value for data $\mathbf{y}$ is the smallest α at which we can reject H_0:

$$p = \min\{\alpha : T(\mathbf{y}) \in R_\alpha\}.$$

We can also think of the p-value as the probability of observing more extreme data than current data $\mathbf{y}$ if H_0 is true.

5.1 Constructing Hypothesis Tests

To construct a hypothesis test, we need to specify both the null hypothesis (and alternative) and the rejection region:

(a) Determine H_0 and H_1 (e.g. one-sided or two-sided, based on the scientific question, before seeing the data);

(b) Find a test statistic $T(\mathbf{Y})$ and find its distribution under the null, i.e. $T(\mathbf{Y})|(\theta = \theta_0)$;

(c) Determine the rejection region $R = \{\mathbf{y} : T(\mathbf{y}) > c\}$ (can be other forms), more specifically the critical value c, usually by controlling Type I error such that $P(T(\mathbf{y}) > c|\theta = \theta_0) \leq \alpha$.

- Rejection region for one-sided test: $R = \{\mathbf{y} : T(\mathbf{y}) > c\}$ or $R = \{\mathbf{y} : T(\mathbf{y}) < c\}$; for two-sided: $R = \{\mathbf{y} : T(\mathbf{y}) < c_L \text{ or } T(\mathbf{y}) > c_U\}$.

5.2 Test Statistics

z-test: Test statistic based on CLT and normal approximation/asymptotics. Under H_0

$$z(\mathbf{T}) = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n}} \dot{\sim} N(0,1),$$

where $\hat{\theta}$ is consistent estimator to θ and $\hat{\sigma}$ is consistent estimator to the standard deviation of $\sqrt{n}\hat{\theta}$.

t-test: Suppose we have (1) estimator $\hat{\theta}$ for θ such that $\hat{\theta} \sim N(\theta_0, \sigma^2)$ under H_0 (true for finite sample size n), (2) estimator $\hat{\sigma}^2$ for σ^2 such as $\hat{\sigma}^2 \sim \sigma^2\chi^2(m)$ and (3) $\hat{\theta} \perp\!\!\!\perp \hat{\sigma}^2$ under H_0. We can construct a t-test statistic:

$$t(\mathbf{T}) = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{m}} \sim t_m.$$

5.2.1 Asymptotic hypothesis tests:

Assuming $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, and we have sufficiently large n, we can use the following asymptotic tests.

Wald test: The asymptotic distribution of the MLE under the null is

$$\hat{\theta} \xrightarrow{d} \mathcal{N}(\theta_0, \mathcal{I}^{-1}(\theta_0)).$$

$$\sqrt{\mathcal{I}(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1).$$

To control the Type I error rate at α, we reject H_0 if

$$\left|\sqrt{n\mathcal{I}_1(\theta_0)}(\hat{\theta} - \theta_0)\right| > \Phi^{-1}(1 - \alpha/2),$$

Score test: By the asymptotic normality of the score function (under the null and regularity conditions)

$$\frac{s(\boldsymbol{Y}, \theta_0)}{\sqrt{n\mathcal{I}_1(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1),$$

and we can reject analogously to before.

Likelihood ratio test: Now allow the more general $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$. The likelihood ratio test statistic is

$$\Lambda(\boldsymbol{Y}) = 2\log\left(\frac{\max_{\theta\in\Theta} L(\theta; \boldsymbol{Y})}{\max_{\theta\in\Theta_0} L(\theta; \boldsymbol{Y})}\right) = 2\log\left(\frac{L(\hat{\theta}; \boldsymbol{Y})}{L(\hat{\theta}_0; \boldsymbol{Y})}\right),$$

where $\hat{\theta}$ is the MLE for θ and $\hat{\theta}_0$ is the MLE in which we only consider the null parameter space. Under regularity conditions, we have

$$\Lambda(\boldsymbol{Y}) \xrightarrow{d} \chi^2_{p-p_0},$$

where p is the dimension of Θ and p_0 is the dimension of p_0. Observe that $\Lambda(\boldsymbol{Y})$ is large when the likelihood under the alternative is much higher than the likelihood under the null, so we reject (with Type I error rate controlled at α) when

$$\Lambda(\boldsymbol{Y}) > \text{qchisq}(1-\alpha, \text{df} = p - p_0).$$

6 Confidence Intervals

6.1 Constructing Confidence Intervals

Say we want to find a $100(1-\alpha)\%$ confidence interval for a parameter θ.

1. Find a pivot $Q(\mathbf{Y}, \theta)$ (remember, the distribution of this pivot must be known, and in particular *cannot* contain any θ's!)
   - For an exact CI, we can use techniques like standardization to get a pivot.
   - For an asymptotic CI, we usually use CLT, delta method, asymptotic normality of the MLE, or some other result to get an asymptotic pivot. *Note:* this interval is not exact for finite n
2. Find an interval $[A, B]$ such that

   $$\mathbb{P}(A \le Q(\boldsymbol{Y}, \theta) \le B) = 1 - \alpha.$$

   (Typically this is done by letting A and B be the quantile function of the pivot's known distribution, evaluated at $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$.)
3. Rearrange the inequality $A \le Q(\boldsymbol{Y}, \theta) \le B$ to be of the form

   $$L(\mathbf{Y}) \le \theta \le U(\mathbf{Y}).$$

   (Note in particular that θ should not appear on the left or right end of this inequality! Only in the middle.)

4. Your confidence interval is $[L(\mathbf{Y}), U(\mathbf{Y})]$.

When doing step 3, be careful when manipulating the inequality. If you multiply each side of the inequality by a negative number or take the reciprocal of each side of the inequality, you need to flip all the inequality signs.

7 Linear Regression

7.1 Predictive Regression

Given some predictor variables $\boldsymbol{X}$ that we observe, how can we predict the outcome variable Y?

Remark Let $\mu(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$ be the predicted outcome and $U(\mathbf{x}) = Y - E[Y|\mathbf{X} = \mathbf{x}]$ be the random noise/regression error.

- $\mathbb{E}[U(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}] = 0$, which means that $\mathbb{E}[U(\boldsymbol{X})] = 0$ unconditionally (by Adam's Law).
- $\text{Cov}(U(\boldsymbol{X}), \boldsymbol{X}) = 0$. This means that predictors and noise are uncorrelated.

7.1.1 Linear Regression

When the regression function is a linear function of the parameters (not the predictors), we have linear regression:

$$\mu(\mathbf{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K$$

where $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_K)^T$ are the regression coefficients, with θ_0 as the intercept and the other entries as the slopes. We would like to estimate $\boldsymbol{\theta}$.

Statistical Model Assume independent data pairs $(\boldsymbol{X_1}, Y_1), \ldots, (\boldsymbol{X_n}, Y_n)$. For continuous data, the joint density for the outcome conditioned on predictors is:

$$f(y_1, \ldots, y_n|\boldsymbol{X_1} = \boldsymbol{x_1}, \ldots, \boldsymbol{X_n} = \boldsymbol{x_n}, \boldsymbol{\theta}) = \prod_{j=1}^{n} f(y_j|\boldsymbol{X_j} = \boldsymbol{x_j}, \boldsymbol{\theta})$$

with the following MLE:

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} \sum_{j=1}^{n} \log f(y_j|\boldsymbol{X_j} = \boldsymbol{x_j}, \boldsymbol{\theta})$$

Gaussian Linear Regression Assume that the noise is distributed as *independent* Normals (and that we are working with a single predictor):

$$Y_j|X_j = x_j, \theta \sim \mathcal{N}(\theta x_j, \sigma^2)$$

By substituting in the Normal PDF, the MLE for θ is

$$\hat{\theta}_{MLE} = \frac{\sum_{j=1}^{n} x_j Y_j}{\sum_{j=1}^{n} x_j^2}$$

Note that this is the same as the least squares estimator

$$\hat{\theta}_{LS} = \underset{\theta}{\operatorname{argmin}} \Big(\sum_{j=1}^{n} (Y_j - \theta x_j)^2 \Big) = \frac{\sum_{j=1}^{n} x_j Y_j}{\sum_{j=1}^{n} x_j^2}$$

Definition 7.1 (Residual)**.** The difference between the true value and its predicted value yields the residual (observable value)

$$\hat{U}_j = Y_j - \hat{\theta} x_j$$

The Residual Sum of Squares (RSS) measures the quality of the regression line's fit to the data:

$$\text{RSS} = \sum_{j=1}^{n} \hat{U}_j^2$$

7.2 Descriptive Regression

In this setting, we are less concerned with prediction and instead, we want to describe the joint distribution of $\mathbf{X}$ and Y, where the $(\mathbf{X}_i, Y_i)$ data pairs are i.i.d. For simplicity, assume that X is scalar.
One summary of the joint distribution is the *descriptive regression*:

$$\beta_{Y \sim X} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Remark We can interpret this as follows. Suppose we are using $a + bX$ to mimic the behavior of Y. The values of (a, b) that minimize the expected squared error $S(a, b) = \mathbb{E}[(Y - a - bX)^2]$ are $(\alpha, \beta_{Y \sim X})$, where $\alpha = \mathbb{E}[Y] - \beta_{Y \sim X} \mathbb{E}[X]$.

Assuming $E[X] = 0$ and $E[Y] = 0$, we can also write a MoM estimator as follows

$$\hat{\theta}_{MoM} = \frac{\sum_{j=1}^{n} X_j Y_j}{\sum_{j=1}^{n} X_j^2}$$

8 Bayesian Statistics

In Bayesian inference, we think of the parameters as random variables in order to use probability and Bayes' Rule to quantify our uncertainty about them. In an inference setting, we can write Bayes' Rule as

$$f(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)\pi(\theta)}{f(\boldsymbol{y})} \propto f(\boldsymbol{y}|\theta)\pi(\theta) = L(\theta;\boldsymbol{y})\pi(\theta).$$

where $\pi(\theta)$ is the prior distribution of θ, $f(\theta|y)$ is the posterior distribution of θ afteer observing data y and $f(y|\theta)$ is equivalent to the likelihood $L(\theta;y)$. Since we are conditioning on observed data, note that we drop the $f(y)$ since it is just a normalizing constant (free from θ) which we call the marginal likelihood of y:

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} L(\tilde{\theta};\mathbf{y})\pi(\tilde{\theta})d\tilde{\theta}$$

The posterior probability of θ being in a range of values $[a,b]$ is

$$P(\theta \in [a,b]|\mathbf{y}) = \int_a^b f(\theta|\mathbf{y})d\theta$$

8.1 Point Estimators

(a) Posterior mean: $\hat{\theta} = E(\theta|\mathbf{y}) = \int \theta f(\theta|\mathbf{y})d\theta$

- Minimizes expected square loss, i.e. $\hat{\theta} = \text{argmin}_{\tilde{\theta}} E((\theta - \tilde{\theta})^2|\mathbf{y})$
- Biased, assuming proper prior and finite variance.

(b) Posterior median: $\hat{\theta} = \text{median}(\theta|\mathbf{y}) = Q_{\theta|\mathbf{y}}(0.5)$

- Minimizes expected absolute loss, i.e., $\hat{\theta} = \text{argmin}_{\tilde{\theta}} E(|\theta - \tilde{\theta}||\mathbf{y})$

(c) Posterior mode: (aka *maximum a posterior*, or MAP)

- $\hat{\theta}_{MAP} = \text{mode}(\theta|\mathbf{y}) = \text{argmax}_{\theta} f(\theta|\mathbf{y})$
- MAP is related to MLE and the prior through the following:

  $$\hat{\theta}_{MAP} = \text{argmax}_{\theta}\{\log L(\theta;\boldsymbol{y}) + \log \pi(\theta)\}$$

  Note that MAP is equivalent to MLE if we use a flat prior. ($p(\theta) \propto 1$).
- Easy to compute since we don't need to know the normalizing constant

8.2 Credible Intervals

A $1-\alpha$ credible interval or posterior probability interval for parameter θ is an interval estimate $[a(\boldsymbol{y}), b(\boldsymbol{y})]$ such that $P[\theta \in [a(\boldsymbol{y}), b(\boldsymbol{y})]|\boldsymbol{y}] = 1-\alpha$

Intervals are typically constructed as $[Q_{\theta|\boldsymbol{y}}(\alpha/2), Q_{\theta|\boldsymbol{y}}(1-\alpha/2)]$ where $Q_{\theta|\boldsymbol{y}}$ is the quantile function of posterior. Alternatively, the credible interval can be found via simulation by sampling from the posterior distribution, in which case you can get the 95% credible interval in R via $[\theta_{(\lceil 0.025B \rceil)}, \theta_{(\lceil 0.975B \rceil)}]$

Credible Vs. Confidence Intervals: It is important to keep the distinction between credible and confidence intervals clear. A 95% confidence interval says that for repeated trials we can construct intervals and 95% of the time, the true parameter value will be in the interval we constructed. A 95% credible interval says that after updating the prior with the data, we think that the parameter will fall within that particular interval with 95% probability.

8.3 Conjugacies

8.3.1 Beta-Binomial

If $Y|p \sim \text{Bin}(n, p)$ and $p \sim \text{Beta}(a, b)$, the posterior is

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

8.3.2 Poisson-Gamma

If $Y|\lambda \sim \text{Pois}(\lambda t)$ and $\lambda \sim \text{Gamma}(a, b)$, the posterior is

$$\lambda|(Y = y) \sim \text{Gamma}(a + y, b + t)$$

and marginally

$$Y \sim \text{NBin}(a, \frac{b}{b+t})$$

8.3.3 Normal-Normal Conjugacy

Suppose

$$Y_j|\mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2) \quad \text{for } i = 1, \ldots, n$$
$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$

where σ^2, μ_0 and τ_0^2 are known constants. The posterior is

$$\mu|(\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}(\mu_n, \tau_n^2)$$

where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}, \quad \text{and} \quad \mu_n = \tau_n^2 \left(\frac{n\bar{y}}{\sigma^2} + \frac{1}{\tau_0^2}\mu_0 \right)$$

8.4 Hierarchical models

Bayesian hierarchical models describes a data generating process with sub-models of multiple levels, which are integrated together by Bayes theorem to get the posterior distribution. The following is a common example of a hierarchical model:

$$Y_j|\theta_1, \ldots, \theta_J, \mu, \lambda \overset{indep}{\sim} \mathcal{N}(\theta_j, \sigma^2)$$

$$\theta_j|\mu, \lambda \overset{iid}{\sim} \mathcal{N}(\mu, \lambda^2)$$
$$\mu|\lambda \sim \text{Pois}(\lambda)$$

8.4.1 Conditional independence

To think about the model's dependence structure, we should think backwards and draw out a diagram of *what generates what*. In the above example, the following data generating process can be drawn:

$$\begin{array}{ccccc} (\mu) & \rightarrow & (\theta_1) & \rightarrow & (Y_1) \\ & \searrow & (\theta_2) & \rightarrow & (Y_2) \\ & & \vdots & & \end{array}$$

- Unconditionally Y_1 and Y_2 are not independent because they share information of μ.
- Given θ_1 and θ_2, Y_1 and Y_2 are conditionally independent, because all of μ's information is carried forward through θ_1 and θ_2, but values of them are now given and fixed.

8.4.2 Joint, marginal, conditional distributions

- The marginal distribution: Integrating out the other random variable

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

- The conditional distribution: Bayes' Rule

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

- The joint distribution:

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

With these tools at hand, we are able to find any posterior we want for hierarchical models!

8.5 Inference with hierarchical models

Before doing anything: *check for conjugacy!* If there's no conjugacy, then follow these steps:

1. *Write down the joint density of all the unknown parameters and data.*
   Here we want

   $$P(Y_1,\ldots,Y_J,\theta_1,\ldots,\theta_J,\mu) = P(Y_1|\theta_1)\ldots P(Y_J|\theta_J)P(\theta_1|\mu)\ldots P(\theta_J|\mu)P(\mu)$$

   Note that this factorization is according to the structure of conditional independence.
2. *Use the joint distribution from Step 1 to get an expression for the conditional density you're interested in.*
   Here suppose that we are interested in $\theta_1,\ldots,\theta_J,\mu|\mathbf{Y}$, then

   $$P(\theta_1,\ldots,\theta_J,\mu|\mathbf{Y}) = \frac{P(\mathbf{Y},\theta_1,\ldots,\theta_J,\mu)}{P(\mathbf{Y})}$$

Then we get the denominator by integrating out all of θ's and μ.

3. *Evaluate the numerator and denominator from Step 2 . Usually this involves marginalizing out some variables so there are some integrals, which can be done analytically, numerically, or via simulation.*

9 Sampling & Bootstrap

9.1 Sampling

Suppose the entire finite population is of size N with variables of interest: $y_1, \ldots, y_N$, fixed and constant. We have population-level estimands:

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

9.1.1 Simple Random Sample (SRS) With Replacement

Let the entire population size be N and the sample size be n. Let $Y_1, \ldots, Y_n$ be the variables in the sample. Pick an ID number from $\{1, ..., N\}$ and observe y_i. Repeat this n times to get a SRS *with* replacement (the same ID number can get picked multiple times).

- **Implementation:** Draw i.i.d. $U_1, ..., U_n \sim \text{Unif}(0, 1)$. Let the jth sampled index be $\lceil NU_j \rceil$.
- **Estimators:** all unbiased and all obey CLT
  - Sample average: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. $E(\bar{Y}) = \mu$ and $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$.
  - Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - ECDF: $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$

9.1.2 SRS Without replacement

Random sample of size n chosen from a total population of size N without replacement such that all $\binom{N}{n}$ possible samples are equally likely.

- $P(Y_i = y_j) = \frac{1}{N}, \forall i, j$, since equal probability across different samples.
- Covariance: by symmetry $\text{Cov}(Y_i, Y_j) = \frac{-\sigma^2}{N-1}$ since $\text{Cov}(Y_1, Y_2) = \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2]$.
- **Estimators:**
  - Sample average: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. $E(\bar{Y}) = \mu$ and $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$, where $\frac{N-n}{N-1}$ is called *finite population correction*.
  - Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

9.2 Stratified Sampling

Divide the population into L strata with N_ℓ as the size of stratum ℓ and total population size $N = \sum_{\ell=1}^{L} N_\ell$. Let μ_ℓ and σ_ℓ^2 be the mean and variance of stratum ℓ. We can conduct SRS (with or without replacement) **independently** in each stratum, with sample size n_ℓ for stratum ℓ. Assume that all n_ℓ, N_ℓ are known.

SRS within each stratum without replacement.

- Estimator for population mean μ: $\bar{Y}_{\text{stratified}} = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \bar{Y}_\ell$
  - Unbiased: $E(\bar{Y}_{\text{stratified}}) = \sum_{l=1}^{L} \frac{N_l}{N} \mu_\ell = \mu$.
  - Variance: $\text{Var}(\bar{Y}_{\text{stratified}}) = \sum_{\ell=1}^{L} (\frac{N_l}{N})^2 \text{Var}(\bar{Y}_l) = \sum_{l=1}^{L} [(\frac{N_l}{N})^2 \cdot \frac{\sigma_j^2}{n_j} \cdot \frac{N_l - n_j}{N_l - 1}]$.

9.3 Non-parametric Bootstrap

In inference, we often construct an estimator $\hat{\theta}$ of an estimand θ and want to quantify our uncertainty about it. If we can derive the distribution of $\hat{\theta}$, then the estimator's standard error is easy to calculate. Often, the distribution of $\hat{\theta}$ is difficult to find analytically.

Bootstrapping is a re-sampling technique for estimating the distribution of an estimator. While we can only run an experiment once in practice, and thereby only observe one value of $\hat{\theta}$, we can bootstrap by recalculating the estimate many times from re-sampled versions of the data. We choose some large B and calculate estimates $\hat{\theta}_b^*$ for $b = 1, \ldots, B$.

Re-sampling the data happens **with replacement** (otherwise we would just permute our values!). Suppose we had

$$\mathbf{Y} = \begin{pmatrix} 4 & 9 & 8 & 7 & 2 \end{pmatrix} \implies \hat{\theta} = \bar{\mathbf{Y}} = 6$$

Then by re-sampling we might get

$$\mathbf{Y}_1^* = \begin{pmatrix} 8 & 9 & 8 & 2 & 2 \end{pmatrix} \implies \hat{\theta}_1^* = \bar{\mathbf{Y}}_1^* = 29/5$$
$$\vdots$$
$$\mathbf{Y}_B^* = \begin{pmatrix} 4 & 4 & 2 & 9 & 7 \end{pmatrix} \implies \hat{\theta}_B^* = \bar{\mathbf{Y}}_B^* = 26/5$$

Then to estimate the standard error of $\hat{\theta}$ we could calculate the sample standard deviation of the bootstrapped estimates:

$$\widehat{\text{SE}(\hat{\theta})} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \overline{\hat{\theta}^*})^2}, \text{where } \overline{\hat{\theta}^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

9.3.1 Confidence Intervals From Bootstrap

We can also use bootstrapping to construct confidence intervals.

- Normal approximation: construct an $(1-\alpha)\cdot 100\%$ interval with endpoints

$$\hat{\theta} \pm \Phi^{-1}(1-\alpha/2)\cdot \widehat{\mathrm{SE}(\hat{\theta})}.$$

  More accurate if we have Normal asymptotics of $\frac{\hat{\theta}-\theta}{\mathrm{SE}(\hat{\theta})}$.

- Percentile interval: Construct an interval with the empirical quantiles of the values $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

- Bootstrap t: Simulate the following pivot to ascertain its distribution, calculate the quantiles, and then use the usual method for constructing a CI from a pivot:

$$\frac{\hat{\theta}^* - \hat{\theta}}{\mathrm{SE}(\hat{\theta}^*)} \text{ is an estimator of } \frac{\hat{\theta}-\theta}{\mathrm{SE}(\hat{\theta})}.$$

  Note that in the left-hand expression, the randomness comes from the *resampling*, which gives random values of $\hat{\theta}^*$.

Note that we assume n to be large for the bootstrap to be accurate, otherwise the Law of Large Numbers does not apply and the empirical CDF of the observed data is not a good estimator of the true CDF.

9.4 Two-Sample Permutation Test

The permutation test works by computing all possible values of the test statistic under possible reorderings of the labels on the data.

For two-sample permutation test, suppose we have $X_1, \ldots, X_m \overset{i.i.d}{\sim} F_X$ and $Y_1, \ldots, Y_n \overset{i.i.d}{\sim} F_Y$, two independent samples. Consider the hypotheses:

$$H_0 : F_X = F_Y \text{ vs. } H_1 : F_X \neq F_Y$$

9.4.1 Complete Permutations

The complete permutation test can be conducted as:

(a) Find a test statistic $T(\mathbf{X}, \mathbf{Y})$ such that large values of T are evidence *against* H_0 (e.g., $T(\mathbf{X}, \mathbf{Y}) = |\bar{Y} - \bar{X}|$).

(b) Compute observed $t_0 = T(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})$ from data.

(c) Compute T from each permutation of $(x_1, \ldots, x_m, y_1, \ldots, y_n)$ to get values $t_1, \ldots, t_{(m+n)!}$.

(d) Compute p-value:

$$p = \Pr(T \geq t_0) = \frac{1}{(m+n)!} \sum_{i=1}^{(m+n)!} I_{t_i \geq t_0}$$

9.4.2 Sampled Permutations

The complete permutation test is computationally inefficient or even inapplicable with large sample size (thinking about $(m+n)!$). The Sampled Permutation Test is very similar, but only considers a *subset* of permutations rather than considering *all possible* permutations.

(a) First steps same as complete permutation. Then, compute T from K random permutations of $(x_1, \ldots, x_m, y_1, \ldots, y_n)$ to get values $t_1, \ldots, t_K$.

(b) Compute p-value:

$$p = \Pr(T \geq t_0) \approx \frac{1}{K} \sum_{i=1}^{K} I_{t_i \geq t_0}$$

Remarks.

- **Advantages**: flexibility of choosing any test statistic, no parametric assumptions (non-parametric), no asymptotics, no complex math.
- **Limitation**: Strong null, i.e. H_0 says that F_X and F_Y are the **same distribution**.

10 Potential Outcomes

10.1 Causal Framework

Consider a drug study with n patients labeled from 1 to n.

Definition 10.1 (Assignment)**.** The jth assignment $W_j = \{0, 1\}$ is 0 if patient j is in the control group and 1 if patient j receives the treatment. We can denote all assignments in the study as $\boldsymbol{W} = (W_1, \ldots, W_n)$.

Definition 10.2 (Potential outcome)**.** Define $Y_j(W_1, \ldots, W_n) \in \{0, 1\}$ as a potential outcome for patient j.

Remark Note that we treat the potential outcome of patient j also as a function of the assignment of all other patients. Subsequently, there are 2^n potential outcomes for patient j. We do not mean that the potential outcome can take 2^n values, but each of the 2^n possible assignments to n patients gives us one potential outcome.

Definition 10.3 (Treatment effect)**.** The treatment effect of moving from assignment $\boldsymbol{W}$ to $\boldsymbol{W}'$ is

$$\tau_j = Y_j(\boldsymbol{W}) - Y_j(\boldsymbol{W}') \in \{-1, 0, 1\}$$

Definition 10.4 (Non-interference)**.** In some experiments, it is reasonable to assume that for each j, the treatment of others has no impact on jth individual, i.e.

$$Y_j(w_1, \ldots, w_{j-1}, w_j, w_{j+1}, \ldots, w_n) = Y_j(w'_1, \ldots, w'_{j-1}, w_j, w'_{j+1}, \ldots, w'_n)$$

Under this assumption of non-interference, we can simplify the notation for patient j's potential outcome to $Y_j(w_j)$, which only has two potential outcomes $Y_j(0)$ and $Y_j(1)$. We can then write the actual outcome of patient j as

$$Y_j = W_j Y_j(1) + (1 - W_j) Y_j(0)$$

and the treatment effect as

$$\tau_j = Y_j(1) - Y_j(0)$$

Definition 10.5 (Assignment mechanism)**.** The assignment mechanism is the joint PMF of the assignments given the potential outcomes:

$$P(\boldsymbol{W} = \boldsymbol{w} | \{\boldsymbol{Y}(0), \boldsymbol{Y}(1)\}$$

10.2 Randomized control trials (RCT)

Definition 10.6 (Randomization)**.** We say that the assignments have been randomized if the assignments are independent of the potential outcomes, i.e.

$$\boldsymbol{W} \perp\!\!\!\perp \{\boldsymbol{Y}(0), \boldsymbol{Y}(1)\}$$

This is equivalent to saying that the assignment mechanism satisfies:

$$P(\boldsymbol{W} = \boldsymbol{w} | \{\boldsymbol{Y}(0), \boldsymbol{Y}(1)\} = P(\boldsymbol{W} = \boldsymbol{w}) = \prod_{j=1}^{n} P(W_j = w_j)$$

10.3 Population based and finite sample modeling

In this section, we assume that the data is observed in pairs $(w_1, y_1), \ldots, (w_n, y_n)$.

10.3.1 Population based modeling

The population quantity $E(\tau_j)$ is a causal quantity for all patients in a wider population beyond the sample.

Definition 10.7. We assume a statistical model where (W_j, Y_j) are i.i.d. across j, and we assume that the study is randomized, then we can define

$$p_{ik} := P(Y_1(0) = i, Y_1(1) = k), i, k \in \{0, 1\}$$

Hence we can express the average treatment effect of the population as:

$$E(\tau_j) = E(Y_j(1) - Y_j(0)) = (p_{01} + p_{11}) - (p_{10} + p_{11}) = p_{01} - p_{10}$$

$$\text{Var}(\tau_j) = E[\tau_j^2] - (E[\tau_j])^2 = (p_{01} + p_{10}) - (p_{01} + p_{10})^2$$

10.3.2 MLE estimator for $E(\tau_1)$

Note that under randomization assumption:

$$\theta_0 := P(Y_1 = 1|W_1 = 0) = P(Y_1(0) = 1) = p_{10} + p_{11}$$
$$\theta_1 := P(Y_1 = 1|W_1 = 1) = P(Y_1(1) = 1) = p_{01} + p_{11}$$

Hence we can estimate the population causal quantity via

$$E(\tau_1) = p_{01} - p_{10} = \theta_1 - \theta_0$$

The MLE estimator for θ_0 and θ_1 are shown to be

$$\hat{\theta}_0 = \frac{\sum_{j=1}^n Y_j(1-w_j)}{\sum_{j=1}^n (1-w_j)}, \qquad \hat{\theta}_1 = \frac{\sum_{j=1}^n Y_j w_j}{\sum_{j=1}^n w_j}$$

Which are ratio of counts: e.g. $\hat{\theta}_1$ is the fraction of actual outcomes which are 1 among people who received the treatment, since the conditional likelihood is Bernoulli. Subsequently:

$$\widehat{E(\tau_1)} = \frac{\sum_{j=1}^n Y_j w_j}{\sum_{j=1}^n w_j} - \frac{\sum_{j=1}^n Y_j(1-w_j)}{\sum_{j=1}^n (1-w_j)}$$

and we can derive the variance, FI, devise pivot for confidence intervals, and carry out hypothesis testing for population level average causal effect as discussed previously in Stat 111.

10.3.3 Finite sample modeling

We treat finite sample quantity as specific only to the patients in the study. In principle, a finite sample quantity says nothing about causal effect on patients that are not in the study.

Definition 10.8. The average treatment effect of a finite sample of size n is:

$$\bar{\tau}_j = \frac{1}{n}\sum_{j=1}^n \tau_j = \frac{1}{n}\sum_{j=1}^n \{y_j(1) - y_j(0)\}$$

In the finite sample approach, we condition on potential outcomes, but note that we are conditioning on something that we do not always observe. We also assume RCT as before.

The MoM estimator for $\bar{\tau}$ is:

$$\hat{\tau}_{\text{MoM}}(\boldsymbol{W}) = \frac{1}{n}\sum_{j=1}^n \left[\frac{W_j Y_j}{E(W_j)} - \frac{(1-W_j)Y_j}{E(1-W_j)}\right]$$

Remark Since $Y_1 = W_1 Y_1(1) + (1-W_1)Y_1(0)$, we have :

$$W_1 Y_1 = W_1 Y_1(1), \quad (1-W_1)Y_1 = (1-W_1)Y_1(0)$$