Final Review

Ricky Truong (rickytruong@college.harvard.edu),
Emily Xing (exing@college.harvard.edu)

1 Introduction

1.1 Logistics

Welcome! The goal of this session is to review the important concepts from STAT 111 ("Introduction to Statistical Inference"), especially in preparation for the final exam.

We aim for this session to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

2 Sufficient Statistics, Rao-Blackwell, and NEF

2.1 Sufficient Statistics

Idea. A statistic $T(\vec{Y})$ *"suffices"* for θ if it captures everything the data can tell us about θ. Once we know T, the remaining randomness in $\vec{Y}$ is irrelevant to finding θ.

Definition 1 (Sufficient statistic)**.** For $Y_1, \ldots, Y_n$ from a parametric model $F_{Y;\theta}$, a statistic $T(\vec{Y})$ is **sufficient** for θ if the conditional distribution of $(Y_1, \ldots, Y_n) \mid T$ does not depend on θ.

Example 1 (Bernoulli trials)**.** Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$ and $T = \sum_{i=1}^n Y_i$. Is T sufficient for p?

For any $\vec{y}$ with $\sum y_i = t$,

$$P(\vec{Y} = \vec{y} \mid T = t) = \frac{P(\vec{Y} = \vec{y},\, T = t)}{P(T = t)} = \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}},$$

which does not depend on p. Thus $T = \sum_{i=1}^n Y_i$ is sufficient. Intuitively, all that matters is the *total* number of successes, not which specific trials succeeded.

- Knowing T renders the ordering and identity of individual observations irrelevant for learning about θ.
- The full data $\vec{Y}$ is always sufficient, but this is trivial. We seek the most efficient sufficient statistic possible.
- Sufficient statistics are not unique: if T is sufficient, then so is any one-to-one function $g(T)$.
- The *minimal* sufficient statistic achieves the greatest simplification.

Theorem 1 (Factorization Criterion)**.** *$T(\vec{Y})$ is sufficient for θ if and only if the joint density (or PMF) factors as*

$$f_{\vec{Y}}(\vec{y};\, \theta) = g_\theta(T(\vec{y})) \cdot h(\vec{y}),$$

where g_θ may depend on θ (but only through $T(\vec{y})$), and h does not depend on θ at all.

• ***Strategy**: Write down the joint density/PMF. Factor it into a piece that only depends on $(\theta, T(\vec{y}))$ and a leftover piece free of θ.*

• ☣*: The factorization only needs to hold up to multiplicative constants that are free of θ. Constant factors can always be absorbed into h.*

Example 2 (Poisson)**.** Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$. The joint PMF is

$$f_{\vec{Y}}(\vec{y};\, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i} \cdot \prod_{i=1}^{n} \frac{1}{y_i!}.$$

Set $g_\lambda(t) = e^{-n\lambda}\lambda^t$ (with $t = \sum y_i$) and $h(\vec{y}) = \prod_{i=1}^n \frac{1}{y_i!}$. Since h is free of λ, $T = \sum_{i=1}^n Y_i$ (equivalently, $\bar{Y}$) is sufficient for λ.

Concept Checker 1. For each model below, use the Factorization Criterion to identify a sufficient statistic for the named parameter.

1. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$, sufficient statistic for λ.
2. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 **known**, sufficient statistic for μ.
3. $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$, sufficient statistic for θ.

Solution

2.2 Rao-Blackwell

Idea. Got an unbiased estimator, but it doesn't use all the data we have? Rao-Blackwell says: *condition on the sufficient statistic*. At worst, you'll end up with the same estimator. At best, you'll almost always make things strictly better (lower MSE)!

Theorem 2 (Rao-Blackwell)**.** *Let $\hat{\theta}$ be any estimator of θ, and let T be a sufficient statistic for θ. Define the Rao-Blackwellized estimator*

$$\hat{\theta}_{RB} = \mathbb{E}[\hat{\theta} \mid T].$$

Then:

*(i) $\hat{\theta}_{RB}$ has the **same bias** as $\hat{\theta}$ (in particular, unbiasedness is preserved).*

*(ii) $\hat{\theta}_{RB}$ has **lower or equal variance** than $\hat{\theta}$.*

(iii) Therefore, $MSE(\hat{\theta}_{RB}) \leq MSE(\hat{\theta})$, with equality iff $\hat{\theta}$ is already a function of T.

Proof. **(i) Bias.** By Adam's Law: $\mathbb{E}[\hat{\theta}_{\text{RB}}] = \mathbb{E}[\mathbb{E}[\hat{\theta} \mid T]] = \mathbb{E}[\hat{\theta}]$. So $\text{Bias}(\hat{\theta}_{\text{RB}}) = \text{Bias}(\hat{\theta})$.

(ii) Variance. By Eve's Law:

$$\mathrm{Var}(\hat{\theta}) = \underbrace{\mathbb{E}[\mathrm{Var}(\hat{\theta} \mid T)]}_{\geq 0} + \mathrm{Var}(\mathbb{E}[\hat{\theta} \mid T]) = \mathbb{E}[\mathrm{Var}(\hat{\theta} \mid T)] + \mathrm{Var}(\hat{\theta}_{\mathrm{RB}}) \geq \mathrm{Var}(\hat{\theta}_{\mathrm{RB}}).$$

Equality holds iff $\mathrm{Var}(\hat{\theta} \mid T) = 0$ a.s., meaning $\hat{\theta}$ is already a deterministic function of T. □

• **Why does this work?** The sufficient statistic captures all the information about θ in the data. If your estimator doesn't fully exploit T, it contains residual randomness that is uninformative about θ—and that noise increases variance. Conditioning on T averages out this uninformative variation.

• **Intuition**: If $\hat{\theta}$ is a function of T, then we can take out what's known. E.g., let $\hat{\theta} = \bar{Y}$ and $T = \sum_{i=1}^n Y_i$. Thus $\hat{\theta}_{\mathrm{RB}} = \mathbb{E}[\bar{Y} \mid \sum_{i=1}^n Y_i] = \bar{Y}$, the original estimator!

• ☣: $\hat{\theta}_{\mathrm{RB}} = \mathbb{E}[\hat{\theta} \mid T]$ is a function of T only (since T is sufficient, the conditional expectation cannot depend on θ). This is what makes it a valid statistic.

Example 3 (Rao-Blackwell for Poisson)**.** Each page of a book has a $\mathrm{Pois}(\lambda)$ number of typos, independently. We observe $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(\lambda)$ and wish to estimate $\theta = P(Y_1 = 0) = e^{-\lambda}$. A naive (but unbiased) estimator uses only the first observation:

$$\hat{\theta} = \mathbb{1}\{Y_1 = 0\}.$$

This is unbiased ($\mathbb{E}[\hat{\theta}] = e^{-\lambda}$), but it throws away $Y_2, \ldots, Y_n$. We know $T = \sum_{i=1}^n Y_i$ is sufficient. By the Chicken-Egg story, $Y_1 \mid T = t \sim \mathrm{Bin}(t, \frac{1}{n})$, so

$$\hat{\theta}_{\mathrm{RB}} = \mathbb{E}[\hat{\theta} \mid T] = P(Y_1 = 0 \mid T) = \left(1 - \frac{1}{n}\right)^T = \left(1 - \frac{1}{n}\right)^{n\bar{Y}}.$$

For large n, $\left(1 - \frac{1}{n}\right)^{n\bar{Y}} \approx e^{-\bar{Y}} = \hat{\theta}_{\mathrm{MLE}}$, so the Rao-Blackwellized estimator is essentially the MLE.

Concept Checker 2. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p)$. Consider the estimator $\hat{p} = Y_1$ (using only the first observation).

1. Show $\hat{p} = Y_1$ is unbiased for p.
2. What is the sufficient statistic T?
3. Compute $\hat{p}_{\mathrm{RB}} = \mathbb{E}[Y_1 \mid T]$.
4. Does this result make intuitive sense?

Solution

Concept Checker 3. Can the MLE be improved by Rao-Blackwellization?

Solution

2.3 Natural Exponential Family (NEF)

Idea. Many of the most important distributions in statistics share a common algebraic structure in PMF/PDF. This structure—the *Natural Exponential Family*—makes it straightforward to identify sufficient statistics, derive MLEs, and compute means and variances all at once.

Definition 2 (Natural Exponential Family)**.** A distribution belongs to the **Natural Exponential Family (NEF)** if its density (or PMF) can be written as

$$f_Y(y;\, \theta) = e^{\theta y - \psi(\theta)} h(y),$$

where θ is the **natural parameter**, $\psi(\theta)$ is the **log-partition function**, and $h(y) \geq 0$ is the **base measure** (does not depend on θ).

• **Strategy**: Rewrite the density as $\exp(\log(f_Y(y)))$.

Theorem 3 (Mean and variance via ψ)**.** *If $Y \sim NEF(\theta)$, then $\mathbb{E}[Y] = \psi'(\theta)$ and $Var(Y) = \psi''(\theta)$.*

• **Intuition**: $\psi(\theta)$ encodes all distributional information. Its first derivative gives the mean, its second gives the variance.

Theorem 4 (Sufficient statistic and MLE in NEF)**.** *Let $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} NEF(\theta)$. Then: (i) $\bar{Y}$ is sufficient for θ; (ii) $\hat{\mu}_{MLE} = \bar{Y}$.*

Proof. **(i)** The joint density is

$$f_{\vec{Y}}(\vec{y};\,\theta) = e^{n(\theta\bar{y}-\psi(\theta))} \cdot \prod_{i=1}^{n} h(y_i).$$

The first factor depends on the data only through $\bar{y}$; the second does not involve θ. By the Factorization Criterion, $\bar{Y}$ is sufficient.

(ii) Setting $\ell'(\theta) = n(\bar{y} - \psi'(\theta)) = 0$ gives $\psi'(\hat{\theta}_{\text{MLE}}) = \bar{y}$, so $\hat{\mu}_{\text{MLE}} = \bar{y}$. □

Example 4 (Bernoulli as NEF)**.** Let $Y \sim \text{Bern}(p)$. We have

$$f_Y(y;\,p) = p^y(1-p)^{1-y} = \exp\left(y \log\frac{p}{1-p} - \log\frac{1}{1-p}\right) \cdot 1.$$

Setting $\theta = \log\frac{p}{1-p}$ and $\psi(\theta) = \log(1+e^\theta)$: $\psi'(\theta) = p = \mathbb{E}[Y]$ and $\psi''(\theta) = p(1-p) = \text{Var}(Y)$. ✓

Example 5 (Poisson as NEF)**.** Let $Y \sim \text{Pois}(\lambda)$. We have

$$f_Y(y;\,\lambda) = \exp(y\log\lambda - \lambda) \cdot \frac{1}{y!}.$$

Setting $\theta = \log\lambda$ and $\psi(\theta) = e^\theta$: $\psi'(\theta) = \lambda = \mathbb{E}[Y]$ and $\psi''(\theta) = \lambda = \text{Var}(Y)$. ✓

Concept Checker 4. Let $Y \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known.

1. Show that the Normal distribution is in the NEF by identifying θ, $\psi(\theta)$, and $h(y)$.
2. Use the NEF properties to verify $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$.

Solution

3 Hypothesis Tests

3.1 Fundamentals

Definition 3 (Statistical hypothesis)**.** Partition the parameter space Θ into two disjoint sets: $\Theta = \Theta_0 \cup \Theta_1$ (i.e., *null set* and *alternative set*, respectively). We test $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, where H_0 is the *null hypothesis* and H_1 is the *alternative hypothesis*.[1]

• The null is *simple* if $\Theta_0 = \{\theta_0\}$ (i.e., if we test only one null value) and *composite* otherwise.

• Tests of the form $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ (i.e., with a *composite null*) are *one-sided* while tests of the form $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ (i.e., with a *simple null*) are *two-sided*.

[1] Equivalently, the alternative hypothesis is sometimes denoted as H_A.

Distribution	θ	$\psi(\theta)$	$\mathbb{E}[Y] = \psi'(\theta)$	$\text{Var}(Y) = \psi''(\theta)$	**Suff. stat.**
Bern(p)	$\log \frac{p}{1-p}$	$\log(1+e^\theta)$	p	$p(1-p)$	$\sum Y_i$
Bin(m, p), m known	$\log \frac{p}{1-p}$	$m\log(1+e^\theta)$	mp	$mp(1-p)$	$\sum Y_i$
Pois(λ)	$\log \lambda$	e^θ	λ	λ	$\sum Y_i$
$\mathcal{N}(\mu, \sigma^2)$, σ^2 known	$\frac{\mu}{\sigma^2}$	$\frac{\sigma^2\theta^2}{2}$	μ	σ^2	$\bar{Y}$
Expo(λ)	$-\lambda$	$-\log(-\theta)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\sum Y_i$
Gamma(α, λ), α known	$-\lambda$	$-\alpha\log(-\theta)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\sum Y_i$

Table 1: Common NEF distributions, where $f_Y(y;\,\theta) = e^{\theta y - \psi(\theta)}h(y)$. The $\mathbb{E}[Y]$ and $\text{Var}(Y)$ columns are expressed in terms of the original parameter for readability. Sufficient statistics assume $Y_1, \ldots, Y_n$ i.i.d.

Definition 4 (Retention and rejection regions)**.** The *retention region* A is the set of data values for which we retain H_0. The *rejection/critical region* A^c is the complement. We reject H_0 if and only if $y \in A^c$. If a test is based on a scalar test statistic $T(\vec{y})$, typical rejection regions are:

- For a one-sided test, $\{y : T(\vec{y}) > c\}$ with *critical value* c.
- For a two-sided test, $\{y : |T(\vec{y})| > c\}$ with critical values $-c, c$.

Definition 5 (Power function and size)**.** The *power function* of a test is $\beta(\theta) = P_{Y;\theta}(Y \in A^c) = \int_{A^c} f_{Y;\theta}(y)\,dy$ (i.e., the probability of rejecting H_0 as a function of possible values of θ).

- E.g., $\beta(111)$ is the probability of rejecting H_0 if, hypothetically, θ were truly 111.
- The *size/level* of a test is $\alpha = \sup_{\theta\in\Theta_0} \beta(\theta)$ (i.e., the max Type I error probability over all null values).
- **Strategy**: Our test statistic uses the null value θ_0, but when evaluating $\beta(\theta)$, think of θ_0 as a constant and θ as the *truth*. Only in $\beta(\theta_0)$ is θ_0 the *truth*.

Concept Checker 5. Suppose we test $H_0 : \theta = 111$ vs. $H_1 : \theta \neq 111$. Do we want $\beta(111)$ to be small or large? Do we want $\beta(110)$ to be small or large?

Solution

Definition 6 (Type I and Type II errors)**.** Type I error is a false positive (i.e., rejecting the null when it's actually true) while Type II error is a false negative (i.e., retaining the null when it's actually false).

• $P(\text{Type I error} \mid \theta \in \Theta_0) = \beta(\theta)$ and $P(\text{Type II error} \mid \theta \in \Theta_1) = 1 - \beta(\theta)$.

3.2 Confusion Matrix

	H_0 **is true**	H_0 **is false**
Reject H_0	Type I error (false positive)	Correct decision
Retain H_0	Correct decision	Type II error (false negative)

3.3 Calibrating Size

Example 6 (Two-sided test with Normal data)**.** Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta^*, \sigma^2)$, with σ^2 known and θ^* unknown. We want to test $H_0 : \theta^* = \theta_0$ vs. $H_1 : \theta^* \neq \theta_0$. We'll use the test statistic $T(\vec{Y}) = \frac{\bar{Y} - \theta_0}{\sigma/\sqrt{n}}$. We always test assuming the null value of θ^* (i.e., θ_0). In a two-sided test, θ_0 is the only point in Θ_0. To achieve a test with size α, use critical value $c = \Phi^{-1}(1 - \frac{\alpha}{2})$.

Proof.

$$
\begin{aligned}
Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta^*, \sigma^2) &\implies \bar{Y} \sim \mathcal{N}\left(\theta^*, \frac{\sigma^2}{n}\right) && \text{by sample mean} \\
&\implies \frac{\bar{Y} - \theta_0}{\sigma/\sqrt{n}} \sim \mathcal{N}\left(\frac{\theta^* - \theta_0}{\sigma/\sqrt{n}}, 1\right) && \text{by shifting/scaling}
\end{aligned}
$$

$$
\begin{aligned}
\beta(\theta) &= P(\vec{Y} \in A^c) && \text{by rejection region} \\
&= P(|T(\vec{Y})| > c) && \text{by critical value} \\
&= P(T(\vec{Y}) > c) + P(T(\vec{Y}) < -c) && \text{by separating} \\
&= P\left(\frac{\bar{Y} - \theta_0}{\sigma/\sqrt{n}} > c\right) + P\left(\frac{\bar{Y} - \theta_0}{\sigma/\sqrt{n}} < -c\right) && \text{by substituting} \\
&= P\left(\frac{\bar{Y} - \theta_0}{\sigma/\sqrt{n}} + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&\quad + P\left(\frac{\bar{Y} - \theta_0}{\sigma/\sqrt{n}} + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} < -c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) && \text{by algebra} \\
&= P\left(\frac{\bar{Y} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + P\left(\frac{\bar{Y} - \theta}{\sigma/\sqrt{n}} < -c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) && \text{by simplifying} \\
&= 1 - P\left(\frac{\bar{Y} - \theta}{\sigma/\sqrt{n}} \leq c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + P\left(\frac{\bar{Y} - \theta}{\sigma/\sqrt{n}} < -c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) && \text{by complement} \\
&= 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) && \text{by Standard Normal CDF}
\end{aligned}
$$

In $\beta(\theta)$, we plug in different values of θ as the "truth" (i.e., $\theta^* = \theta$) such that $\frac{\bar{Y} - \theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$. Recall $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$ (i.e., the max Type I error probability over all null values), so $\alpha = \beta(\theta)|_{\theta = \theta_0}$ since θ_0 is the only value in Θ_0. Since $\alpha = \beta(\theta)|_{\theta=\theta_0} = 1 - \Phi(c) + \Phi(-c) = 2 - 2\Phi(c)$, we have $c = \Phi^{-1}(1 - \frac{\alpha}{2})$. □

Example 7 (One-sided test with Normal data)**.** With the same setup, we want to test $H_0 : \theta^* \leq \theta_0$ vs. $H_1 : \theta^* > \theta_0$. We'll use the test statistic $T(\vec{Y}) = \frac{\bar{Y}-\theta_0}{\sigma/\sqrt{n}}$. We always test assuming the null value of θ^* (i.e., θ_0). In a one-sided test, θ_0 is the boundary point. To achieve a test with size α, use critical value $c = \Phi^{-1}(1-\alpha)$.

Proof.

$$\begin{aligned}
\beta(\theta) &= P(\vec{Y} \in A^c) && \text{by rejection region} \\
&= P(T(\vec{Y}) > c) && \text{by critical value} \\
&= 1 - P(T(\vec{Y}) \leq c) && \text{by complement} \\
&= 1 - P\left(\frac{\bar{Y}-\theta_0}{\sigma/\sqrt{n}} \leq c\right) && \text{by substituting} \\
&= 1 - P\left(\frac{\bar{Y}-\theta_0}{\sigma/\sqrt{n}} + \frac{\theta_0-\theta}{\sigma/\sqrt{n}} \leq c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right) && \text{by algebra} \\
&= 1 - P\left(\frac{\bar{Y}-\theta}{\sigma/\sqrt{n}} \leq c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right) && \text{by simplifying} \\
&= 1 - \Phi\left(c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right) && \text{by Standard Normal CDF}
\end{aligned}$$

In $\beta(\theta)$, we plug in different values of θ as the "truth" (i.e., $\theta^* = \theta$) such that $\frac{\bar{Y}-\theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Notice $\theta \uparrow \implies c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}} \downarrow \implies \Phi\left(c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right) \downarrow \implies 1 - \Phi\left(c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right) \uparrow$, so $\beta(\theta)$ strictly increases as θ does. Recall $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$ (i.e., the max Type I error probability over all null values), so $\alpha = \beta(\theta)|_{\theta=\theta_0}$ since θ_0 is the largest value in Θ_0. Since $\alpha = \beta(\theta)|_{\theta=\theta_0} = 1 - \Phi(c)$, we have $c = \Phi^{-1}(1-\alpha)$. □

3.4 z-test vs. t-test

Definition 7 (z-test)**.** Let $\hat{\theta}$ be a *consistent* estimator with *asymptotic Normality* for θ and $\widehat{\text{SE}}[\hat{\theta}]$ be a *consistent* estimator for $\text{SE}[\hat{\theta}]$. To test a null value θ_0, the *z-test statistic* is $z = \frac{\hat{\theta}-\theta_0}{\widehat{\text{SE}}[\hat{\theta}]} \dot{\sim} \mathcal{N}(0,1)$ under H_0 for large n.

• Notice we don't assume Normal data but large n as well as a consistent and asymptotic Normal estimator (e.g., $\bar{Y}$ for i.i.d. data).

Definition 8 (t-test)**.** Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$, with θ and σ^2 unknown. To test a null value θ_0, the *t-test statistic* is $t = \frac{\sqrt{n}(\bar{Y}-\theta_0)}{\hat{\sigma}} \sim t_{n-1}$ under H_0, where $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}$.[2]

• Notice we don't assume large n but Normal data.

Concept Checker 6. Let $Y_1, \ldots, Y_n$ be i.i.d., with $\mathbb{E}[Y_i] = \theta$ and $\text{Var}[Y_i] = \sigma^2 < \infty$ unknown and n large. Suppose $\hat{\theta} = \bar{Y}$ and $\widehat{\text{SE}}[\hat{\theta}] = \frac{\hat{\sigma}}{\sqrt{n}}$ are consistent estimators (specifically, for $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$). Show $z = \frac{\hat{\theta}-\theta_0}{\widehat{\text{SE}}[\hat{\theta}]} \dot{\sim} \mathcal{N}(0,1)$ under $H_0 : \theta = \theta_0$.

[2]This holds by the story of the t distribution.

Solution

3.5 Setups

Let $\mu = \mathbb{E}[Y_i]$ and $\sigma^2 = \text{Var}[Y_i]$. Use the following estimators: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 \xrightarrow{p} \sigma^2$ and $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2} \xrightarrow{p} \sigma$. Suppose we're interested in testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Data	**Test Stat and Distribution under H_0**	$100(1-\alpha)\%$ **CI**
$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, μ unknown, σ^2 known	$\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$	For μ: $\bar{Y} \pm (\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2})$
$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, μ and σ^2 unknown	$\frac{\bar{Y}-\mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$	For μ: $\bar{Y} \pm (\frac{\hat{\sigma}}{\sqrt{n}})Q_{t_{n-1}}(1-\frac{\alpha}{2})$
$Y_1, \ldots, Y_n$ i.i.d., μ unknown, σ^2 known	$\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \dot{\sim} \mathcal{N}(0,1)$ for large n	For μ: $\bar{Y} \pm (\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2})$
$Y_1, \ldots, Y_n$ i.i.d., μ and σ^2 unknown	$\frac{\bar{Y}-\mu_0}{\hat{\sigma}/\sqrt{n}} \dot{\sim} \mathcal{N}(0,1)$ for large n	For μ: $\bar{Y} \pm (\frac{\hat{\sigma}}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2})$

3.6 Three Likelihood-Based Tests

Definition 9 (Wald test)**.** To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ with $\hat{\theta}_{\text{MLE}}$, use $W(\vec{Y}) = \sqrt{\mathcal{I}_{\vec{Y}}(\theta_0)}(\hat{\theta}_{\text{MLE}} - \theta_0) \dot{\sim} \mathcal{N}(0,1)$ under H_0.

• Reject if $|W(\vec{Y})| > Q_{\mathcal{N}(0,1)}(1-\frac{\alpha}{2})$ or equivalently if $W(\vec{Y})^2 = \mathcal{I}_{\vec{Y}}(\theta_0)(\hat{\theta}-\theta_0)^2 > Q_{\chi^2_1}(1-\alpha)$.

Definition 10 (Score test)**.** To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ with $\hat{\theta}_{\text{MLE}}$, use $S(\vec{Y}) = \frac{s(\theta_0;Y)}{\sqrt{\mathcal{I}_{\vec{Y}}(\theta_0)}} \dot{\sim} \mathcal{N}(0,1)$ under H_0.

• Reject if $|S(\vec{Y})| > Q_{\mathcal{N}(0,1)}(1 - \frac{\alpha}{2})$.

Definition 11 (Likelihood ratio test)**.** To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ with $\hat{\theta}_{\text{MLE}}$, use $\Lambda(\vec{Y}) = 2\left(\ell(\hat{\theta}_{\text{MLE}}; \vec{Y}) - \ell(\theta_0; \vec{Y})\right) \dot{\sim} \chi^2_1$ under H_0.

• Reject if $\Lambda(\vec{Y}) > Q_{\chi^2_1}(1 - \alpha)$.

• $\Lambda(\vec{Y}) \geq 0$, since $\hat{\theta}_{\text{MLE}}$ maximizes the likelihood.

Concept Checker 7. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$, with n large and θ unknown. Suppose we want to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Find the test statistic for whichever test is more appropriate: z-test or t-test. Also, find the test statistics for the Wald test, score test, and likelihood ratio test.

Solution

3.7 p-value

Definition 12 (p-value for simple null)**.** For a *simple null* H_0 with test statistic $T(Y)$ that rejects for large T, the *p-value* is $p(\vec{y}) = P(T \geq t(\vec{y}) \mid H_0)$ (i.e., the probability of observing a result at least as extreme as the data, assuming H_0 is true).

Definition 13 (p-value for general null)**.** For *any null* H_0 with *retention region* A_α for each *size* α, the *p-value* is $p(\vec{y}) = \min\{\alpha : y \in A^c_\alpha\}$ (i.e., the smallest size at which the test would've rejected H_0).

• ☣: The p-value is *not* the probability that H_0 is true or the probability that the result is "just due to chance". Also, the threshold $\alpha = 0.05$ is only a convention and not a mathematical truth!

Concept Checker 8. Let $T(\vec{Y})$ be a continuous test statistic. Show $p(\vec{Y}) \sim \text{Unif}(0,1)$ under H_0.

Solution

4 Bayesian Inference

4.1 Big Picture

In the *frequentist* framework, θ is an unknown but fixed constant. We build estimators and confidence intervals whose performance we assess over hypothetical repeated experiments. In the *Bayesian* framework, θ is treated as a random variable, and we use probability to quantify our uncertainty about it.

The workflow is:

- *Prior* $\pi(\theta)$: our beliefs about θ before seeing data, encoded as a probability distribution.
- *Likelihood* $L(\theta; y) = f(y \mid \theta)$: how likely the observed data are under each value of θ.
- *Posterior* $\pi(\theta \mid y)$: our updated beliefs about θ after seeing data.

4.2 Prior to Posterior

Definition 14 (Prior, posterior, marginal likelihood)**.** Consider a parametric model $f(y \mid \theta)$ with unknown parameter θ. The **prior** $\pi(\theta)$ is the marginal distribution of θ, encoding our beliefs before seeing data. The **posterior** $\pi(\theta \mid y)$ is the conditional distribution of θ given the observed data y. The **marginal likelihood** is $f(y) = \int f(y \mid \theta)\pi(\theta)\, d\theta$.

Theorem 5 (Bayes' rule)**.**

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

We almost always work up to proportionality, since $f(y)$ does not depend on θ.

• **Strategy**: Write $\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$, simplify, and pattern-match to a known distribution.

• ☣: **Cromwell's rule.** If $\pi(\theta_0) = 0$, then $\pi(\theta_0 \mid y) = 0$ no matter what the data say. Never assign prior probability of exactly 0 or 1 to something unless it is logically impossible or certain.

• **As** $n \to \infty$: the likelihood dominates the prior, and the posterior concentrates near the MLE.

Concept Checker 9. Let $Y \mid \theta \sim \text{Bern}(\theta)$ and $\theta \sim \text{Beta}(2, 2)$.

1. Write out $\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$ and identify the posterior distribution.
2. What does the $\text{Beta}(2, 2)$ prior encode about our beliefs for θ?
3. What happens to the posterior as we observe more and more data?

Solution

4.3 Point Estimation

Definition 15 (Posterior mean, median, and mode)**.** Let θ have a continuous posterior density $\pi(\theta \mid y)$. Then:

$$\begin{aligned}\text{Posterior mean} &= \mathbb{E}[\theta \mid y] = \int \theta\, \pi(\theta \mid y)\, d\theta \\ \text{Posterior median} &= Q_{\theta|y}(0.5) \\ \text{Posterior mode (MAP)} &= \arg\max_{\theta}\, \pi(\theta \mid y) = \arg\max_{\theta} \{\log L(\theta; y) + \log \pi(\theta)\}\end{aligned}$$

Theorem 6 (Optimal loss functions)**.**

- *Squared error loss $(\theta - \hat{\theta})^2$: minimized by the **posterior mean** $\mathbb{E}[\theta \mid y]$.*
- *Absolute error loss $|\theta - \hat{\theta}|$: minimized by the **posterior median** $Q_{\theta|y}(0.5)$.*
- *0–1 loss (in a limit): minimized by the **posterior mode (MAP)**.*

• **MAP vs. MLE.** With a flat (Uniform) prior on a bounded interval, MAP = MLE. With an informative prior, MAP regularizes the MLE toward the prior mean.

• **MAP and LASSO.** With a Laplace prior $\pi(\theta) \propto e^{-d|\theta|}$, the MAP is the LASSO estimator. With a Normal prior, the MAP/posterior mean is ridge regression.

• ☣: MAP is **not** invariant to reparameterization.

Concept Checker 10. For the Beta-Binomial model $Y \mid p \sim \text{Bin}(n, p)$, $p \sim \text{Beta}(a, b)$, write down the posterior mean, median, and MAP explicitly. For large n, what do all three converge to?

Solution

4.4 Credible Intervals

Definition 16 (Credible interval)**.** Let $0 < \alpha < 1$. A $100(1-\alpha)\%$ **credible interval** for θ is an interval $[a(y), b(y)]$ such that

$$P(a(y) \leq \theta \leq b(y) \mid y) = 1 - \alpha.$$

The standard choice is the equal-tailed interval $[Q_{\theta|y}(\alpha/2),\ Q_{\theta|y}(1-\alpha/2)]$.

• **Direct probability interpretation.** A 95% credible interval means: given the data, there is a 95% probability that θ lies in the interval.

• **Average frequentist coverage.** A 95% credible interval has *on average* 95% frequentist coverage. By Adam's law: $P(\theta \in C(Y)) = \mathbb{E}[P(\theta \in C(Y) \mid Y)] = \mathbb{E}[0.95] = 0.95$.

• ☣: A 95% credible interval is *not* guaranteed to have 95% frequentist coverage at a specific fixed θ.

Concept Checker 11. Suppose $\theta \sim \mathcal{N}(0, 2^2)$ and $Y \mid \theta \sim \mathcal{N}(\theta, 1)$. We observe $Y = y$.

1. Find the posterior distribution $\theta \mid y$.
2. Write down a 95% credible interval for θ.
3. If the true $\theta = 1$, is the coverage of this credible interval exactly 95%?

Solution

4.5 Conjugate Priors

Idea. A conjugate prior is one where the posterior stays in the same distributional family as the prior. We only need to update the parameters, not the family itself.

Definition 17 (Conjugate prior)**.** A family of priors is **conjugate** for a particular likelihood if choosing a prior in the family always results in a posterior in the same family.

Theorem 7 (Beta-Binomial conjugacy)**.** *If $p \sim Beta(a, b)$ and $Y \mid p \sim Bin(n, p)$, then*

$$p \mid (Y = y) \sim Beta(a + y,\ b + n - y).$$

Posterior mean: $\dfrac{a+y}{a+b+n}$. *Interpret $a-1$ as prior successes, $b-1$ as prior failures.*

Theorem 8 (Gamma-Poisson conjugacy)**.** *If $\lambda \sim Gamma(r_0, b_0)$ (rate b_0) and $Y_1, \ldots, Y_n \mid \lambda \overset{i.i.d.}{\sim} Pois(\lambda)$, then with $S = \sum y_i$:*

$$\lambda \mid y \sim Gamma(r_0 + S,\ b_0 + n).$$

Posterior mean: $\dfrac{r_0+S}{b_0+n}$. *Predictive:* $\tilde{Y} \mid y \sim NBin\left(r_0 + S,\ \dfrac{b_0+n}{b_0+n+1}\right)$.

Theorem 9 (Normal-Normal conjugacy)**.** *Let $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known, and prior $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$. Then*

$$\mu \mid y \sim \mathcal{N}(\mu_n, \tau_n^2), \qquad \tau_n^{-2} = n\sigma^{-2} + \tau_0^{-2}, \qquad \mu_n = \tau_n^2\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right).$$

Writing $b_n = \tau_n^2/\tau_0^2 = \sigma^2/(\sigma^2 + n\tau_0^2)$ (shrinkage factor): $\mu_n = (1 - b_n)\bar{y} + b_n\mu_0$.
As $n \to \infty$, $b_n \to 0$ and $\mu_n \to \bar{y}$.

• **Intuition.** The posterior mean compromises between $\bar{y}$ and μ_0, weighted by their relative precisions.

• **Normal-Normal predictive.** $\tilde{Y} \mid y \sim \mathcal{N}(\mu_n,\ \sigma^2 + \tau_n^2)$. The extra τ_n^2 reflects parameter uncertainty.

Concept Checker 12. A manufacturer claims product weights are $\mathcal{N}(\theta, 100)$ grams. Your prior for θ is $\mathcal{N}(200, 400)$. You weigh $n = 25$ items and find $\bar{y} = 190$.

1. Find the posterior distribution of θ.
2. Interpret the shrinkage factor b_n.
3. Find the posterior predictive distribution for a new item's weight.

Solution

4.6 Bayesian Model Choice

Definition 18 (Bayes factor)**.** Suppose Bill uses likelihood $f(y \mid \theta)$ and prior $f(\theta \mid \text{Bill})$, and Jose uses $g(y \mid \lambda)$ and $g(\lambda \mid \text{Jose})$. The **Bayes factor** is

$$\text{BF} = \frac{f(y \mid \text{Bill})}{g(y \mid \text{Jose})} = \frac{\int f(y \mid \theta)\, f(\theta \mid \text{Bill})\, d\theta}{\int g(y \mid \lambda)\, g(\lambda \mid \text{Jose})\, d\lambda}.$$

It converts prior odds to posterior odds:

$$\frac{P(\text{Bill} \mid y)}{P(\text{Jose} \mid y)} = \frac{P(\text{Bill})}{P(\text{Jose})} \times \text{BF}.$$

• **Candidate's formula.** For any fixed value t: $f(y) = \frac{f(y|\theta=t)\,\pi(\theta=t)}{\pi(\theta=t|y)}$.

• **Lindley's paradox.** For large n, the Bayes factor penalizes complex models more heavily than a frequentist test. The log Bayes factor is the BIC.

4.7 Posterior Predictive Distribution

Definition 19 (Posterior predictive)**.** Given observed data y, the **posterior predictive distribution** of a future observation $\tilde{Y}$ is

$$f(\tilde{y} \mid y) = \int_{\theta \in \Theta} f(\tilde{y} \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

• By Adam's law: $\mathbb{E}[\tilde{Y} \mid y] = \mathbb{E}[\theta \mid y]$ (in the Normal case).

• By Eve's law: $\text{Var}(\tilde{Y} \mid y) = \underbrace{\mathbb{E}[\text{Var}(\tilde{Y} \mid \theta)]}_{\text{aleatoric}} + \underbrace{\text{Var}(\mathbb{E}[\tilde{Y} \mid \theta])}_{\text{epistemic}} = \sigma^2 + \tau_n^2$ (in Normal-Normal).

• **Simulation strategy.** Draw $\theta^{[b]} \sim \pi(\theta \mid y)$, then $\tilde{y}^{[b]} \sim f(\tilde{y} \mid \theta^{[b]})$.

Concept Checker 13. Let $Y_1, \ldots, Y_n \mid \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$, with σ^2, μ_0, τ_0^2 known. We want $Y_{n+1} \mid Y_1, \ldots, Y_n$. Is $Y_{n+1} \perp\!\!\!\perp (Y_1, \ldots, Y_n)$? What is the distribution?

Solution

4.8 Decision Theory

Definition 20 (Risk function)**.** For an estimator $\hat{\theta} = T(Y)$, the **risk function** is the expected loss:

$$\text{Risk}(\theta) = \mathbb{E}_\theta\Big[\text{Loss}(\theta, \hat{\theta})\Big].$$

Definition 21 (Admissibility)**.** An estimator $\hat{\theta}$ is **inadmissible** if there exists another estimator with risk $\leq$ that of $\hat{\theta}$ for all θ, with strict inequality for at least one θ. It is **admissible** if no such dominating estimator exists.

4.9 Hierarchical Models & Stein's Paradox

Idea. When data come from multiple groups, a hierarchical model places a prior on the group-level parameters. This induces *partial pooling*: groups with few observations are pulled toward the overall mean, while groups with many observations stay close to their own sample mean.

Definition 22 (Two-level Gaussian hierarchical model)**.** For $j = 1, \ldots, K$:

$$Y_j \mid \mu_j \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_j, \sigma^2), \qquad \mu_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\gamma, \lambda_0^2).$$

Hyperparameters $(\sigma, \gamma, \lambda_0)$ known. The posterior is

$$\mu_j \mid y \sim \mathcal{N}(m_j, \lambda_K^2), \quad m_j = \lambda_K^2(\lambda_0^{-2}\gamma + \sigma^{-2}y_j), \quad \lambda_K^{-2} = \lambda_0^{-2} + \sigma^{-2}.$$

Marginally, $Y_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\gamma, \sigma^2 + \lambda_0^2)$.

• Unconditionally, Y_1 and Y_2 are *not* independent—they share information about μ. Given μ_1, μ_2, they *are* conditionally independent.

Concept Checker 14. You are given a mysterious die that may or may not be loaded. You observe n rolls that are all 6. Your friend will roll next, and denote this value as Y_{n+1}. Is $Y_{n+1} \perp\!\!\!\perp \vec{Y}_n$? What if we condition on $p = \frac{1}{6}$?

Solution

Theorem 10 (Stein, 1956)**.** *Let $Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$ independently for $j = 1, \ldots, K$ with $K \geq 3$ and σ^2 known. Under total squared error loss $\sum_j (\mu_j - \hat{\mu}_j)^2$, the MLE $\hat{\mu} = Y$ is* ***inadmissible****.*

Theorem 11 (James-Stein estimator)**.** *Let $S = \sum_j Y_j^2$. The estimator*

$$\hat{\mu}_j^{JS} = \left(1 - \frac{(K-2)\sigma^2}{S}\right) Y_j$$

has strictly lower risk than Y for all $\mu \in \mathbb{R}^K$.

- ☣: The MLE is inadmissible for $K \geq 3$, but is admissible for $K \leq 2$.

Concept Checker 15 (Hierarchical Normal model)**.** For $j = 1, \ldots, J$, suppose $\bar{y}_j \mid \theta_j \overset{\text{ind.}}{\sim} \mathcal{N}(\theta_j, \sigma_j^2)$ with known σ_j^2, and

$$\theta_j \mid \mu, \tau \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \tau^2), \qquad \pi(\mu, \tau) \propto 1.$$

1. Derive $\theta_j \mid \mu, \tau, \bar{y}_j$. Write the posterior mean as $\hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j)\mu$ and identify λ_j. What happens as $\tau^2 \to \infty$? As $\tau^2 \to 0$?
2. Why are groups with larger σ_j^2 shrunk more toward μ?
3. Write down $p(\tilde{y}_j \mid y)$ for an existing group j, and $p(\tilde{y}_{\text{new}} \mid y)$ for a new exchangeable group.

Solution

5 Sampling

5.1 Design-Based Inference

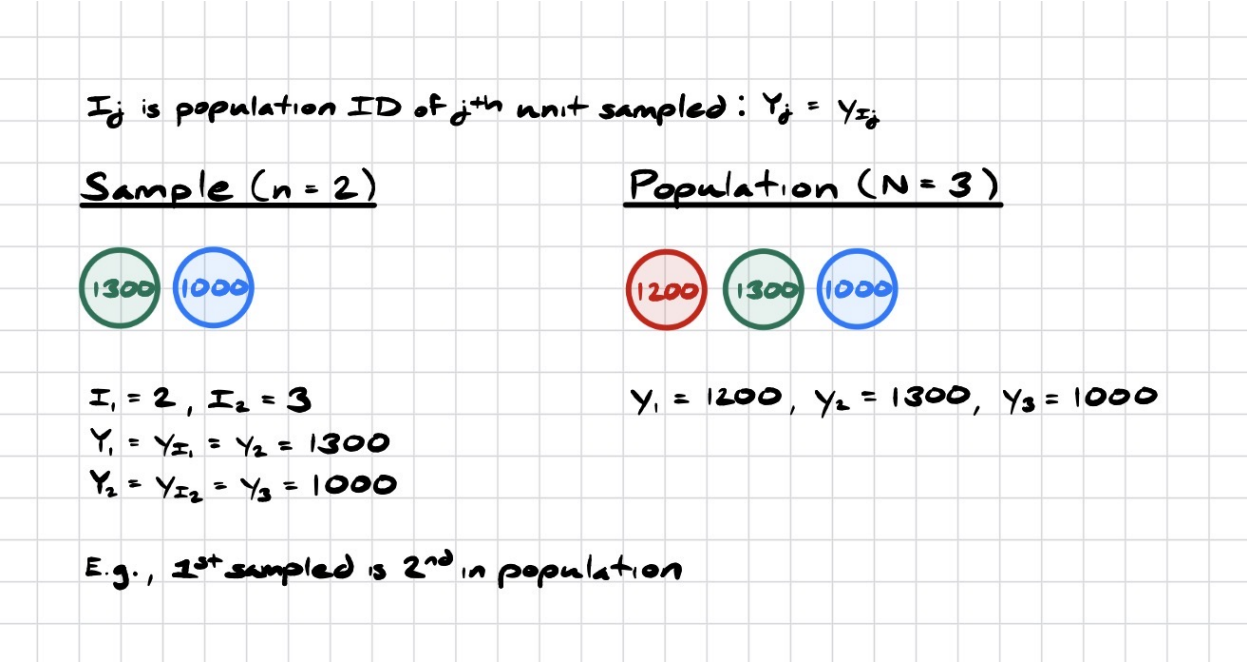

FIGURE 1: The relationship between sample and population in design-based inference.

Definition 23 (Finite sample estimand)**.** A function of $y_1, \ldots, y_N$.

- E.g., population mean is $\mu = \frac{1}{N}\sum_{j=1}^{N} y_j$.
- E.g., population variance is $\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - \mu)^2$.
- E.g., population CDF is $F(y) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\{y_j \leq y\}$.

Definition 24 (Sampling design)**.** The joint PMF of $I_1, \ldots, I_n$, where I_j is the population ID of the jth unit sampled: $P(I_1 = i_1, \ldots, I_n = i_n)$.

- We regard $y_1, \ldots, y_N$ as fixed, but notice $Y_j = y_{I_j}$, so we can describe the randomness from the sampling with the random variables $I_1, \ldots, I_n$.
- In Figure 1, we have a population of $N = 3$ students with the following SAT scores: $\{1200, 1300, 1000\}$. Notice $y_1 = 1200$, $y_2 = 1300$, $y_3 = 1000$. These scores are fixed. We randomly sample $n = 2$ scores: $\{1300, 1000\}$. Since the first person we sampled was the second person in the population, $I_1 = 2$, so $Y_1 = y_{I_1} = y_2 = 1300$.

Concept Checker 16. What values can I_j take on? What values can j take on?

Solution

Definition 25 (Equal probability sample)**.** A *sampling design* such that the marginal PMF satisfies $P(I_j = k) = \frac{1}{N}\ \forall\ j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, N\}$.

• I.e., the unconditional probability the jth unit sampled is the kth unit in the population is equally likely for any j and k—there's no reason why, e.g., the first unit sampled is more likely to be the first unit in the population.

• All *equal probability samples* have $\mathbb{E}_I[Y_j] = \mu$, $\text{Var}_I[Y_j] = \sigma^2$, $\mathbb{E}_I[\bar{Y}] = \mu$, and $\mathbb{E}_I[\hat{F}(y)] = F(y)\ \forall\ j \in \{1, \ldots, n\}$, where expectation is with respect to the sampling design (i.e., with respect to the randomness in I).

Concept Checker 17. Recall $\mathbb{E}\left[g(X)\right] = \sum_{\text{supp}(X)} g(x)P(X = x)$ for X discrete by LOTUS. Use this to show all equal probability samples have $\mathbb{E}_I[Y_j] = \mu$.

Solution

5.2 Simple Random Sampling

Definition 26 (SRS with replacement)**.** An *equal probability sample* where we draw the n population IDs as $I_j \overset{\text{i.i.d.}}{\sim} \text{DUnif}(1, N)$ and set $Y_j = y_{I_j}\ \forall\ j \in \{1, \ldots, n\}$.

• The *sampling design* is given by $P(I_1 = i_1, \ldots, I_n = i_n) = \frac{1}{N^n}$.

• All *SRS with replacement* have $\mathbb{E}_{\text{with}}[S^2] = \sigma^2$, $\text{Var}_{\text{with}}[\bar{Y}] = \frac{\sigma^2}{n}$, and $\text{Var}_{\text{with}}[\hat{F}(y)] = \frac{F(y)(1-F(y))}{n}$. Like before, expectation is with respect to the randomness in I, but we sometimes write $\mathbb{E}_{\text{with}}$ as a reminder the sampling is with replacement.

Definition 27 (SRS without replacement)**.** An *equal probability sample* where we draw the n population IDs such that all $\frac{N!}{(N-n)!}$ permutations are equally likely and set $Y_j = y_{I_j}\ \forall\ j \in \{1, \ldots, n\}$.

• The sampling design is given by $P(I_1 = i_1, \ldots, I_n = i_n) = \frac{1}{N!/(N-n)!}$.[3]

• All *SRS without replacement* have $\text{Cov}_{\text{w/o}}[Y_j, Y_k] = \frac{-\sigma^2}{N-1}$ and $\text{Var}_{\text{w/o}}[\bar{Y}] = \frac{\sigma^2}{n}(\frac{N-n}{N-1})\ \forall\ j, k \in \{1, \ldots, n\}$ with $j \neq k$, where $\frac{N-n}{N-1}$ is the *finite population correction.* Like before,

[3]The number of samples of size n from a population of size N—without replacement and where order matters—is given by $\frac{N!}{(N-n)!}$.

expectation is with respect to the randomness in I, but we sometimes write $\mathbb{E}_{\text{w/o}}$ as a reminder the sampling is without replacement.

5.3 Stratified Sampling

Definition 28 (Strata)**.** Partition the population IDs—$I_1, \ldots, I_N$—into *strata*, each of size $N_\ell \; \forall \; \ell \in \{1, \ldots, L\}$, where L is the number of strata. Assume $N_\ell \geq 1$—i.e., no stratum has fewer than 1 unit—and $\sum_{\ell=1}^{L} N_\ell = N$—i.e., all stratum sizes add up to the population size. Within the ℓth stratum, we denote the fixed population values as $\{y_{1,\ell}, \ldots, y_{N_\ell,\ell}\}$. Within the ℓth stratum, we define the stratum quantities as the following:

- *Stratum mean* is $\mu_\ell = \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} y_{j,\ell}$, which we can relate to *population mean*: $\mu = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \mu_\ell$.
- *Stratum variance* is $\sigma_\ell^2 = \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} (y_{j,\ell} - \mu_\ell)^2$, which we can relate to *population variance*: $\sigma^2 = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \sigma_\ell^2 + \sum_{\ell=1}^{L} \frac{N_\ell}{N} (\mu_\ell - \mu)^2$.
- *Stratum CDF* is $F_\ell(y) = \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} \mathbb{1}\{y_{j,\ell} \leq y\}$, which we can relate to *population CDF*: $F(y) = \sum_{\ell=1}^{L} \frac{N_\ell}{N} F_\ell(y)$.

Concept Checker 18. Show $\sum_{\ell=1}^{L} \frac{N_\ell}{N} \mu_\ell = \mu$. Additionally, what role does the $\frac{N_\ell}{N}$ term play? E.g., if $N = 100$, $N_1 = 2$, and $N_2 = 98$ with $L = 2$, what should we take into consideration?

Solution

Definition 29 (Stratified sampling design)**.** A sampling design is a *stratified sampling design* if the sampling is done independently across strata. We define stratum-specific estimators as the following:

- *Stratum sample mean* is $\bar{Y}_\ell = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} Y_{j,\ell}$, which we can pool to obtain an estimator for *population mean*: $\hat{\mu}_{\text{strat}} = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \bar{Y}_\ell$.
- *Stratum empirical CDF* is $\hat{F}_\ell(y) = \frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \mathbb{1}\{Y_{j,\ell} \leq y\}$, which we can pool to obtain an estimator for *population CDF*: $\hat{F}_{\text{strat}}(y) = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \hat{F}_\ell(y)$.

• All *stratified sampling designs* have $\mathbb{E}_I[\hat{\mu}_{\text{strat}}] = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \mathbb{E}_I[\bar{Y}_\ell]$ and $\text{Var}_I[\hat{\mu}_{\text{strat}}] = \sum_{\ell=1}^{L} \left(\frac{N_\ell}{N}\right)^2 \text{Var}_I[\bar{Y}_\ell]$, where expectation is with respect to the randomness in I.[4]

Definition 30 (Equal probability stratified sampling design)**.** A sampling design is an *equal probability stratified sampling design* if $P(I_{j,\ell} = k) = \frac{1}{N_\ell}\ \forall\ j \in \{1, \ldots, n_\ell\}$, $\ell \in \{1, \ldots, L\}$, and $k \in \{1, \ldots, N_\ell\}$, where $I_{j,\ell}$ is the population ID of the jth unit in the ℓth stratum.

• I.e., the unconditional probability the jth unit sampled is the kth unit in the population within the ℓth stratum is equally likely for any j, ℓ, and k—there's no reason why, e.g., the first unit sampled is more likely to be the first unit in the population within the first stratum.

• All *equal probability stratified sampling designs* have $\mathbb{E}_I[\bar{Y}_\ell] = \mu_\ell$, so $\mathbb{E}_I[\hat{\mu}_{\text{strat}}] = \mu$. Additionally, $\text{Var}_{\text{with}}[\bar{Y}_\ell] = \left(\frac{1}{n_\ell}\right)\sigma_\ell^2$ for *SRS with replacement* and $\text{Var}_{\text{w/o}}[\bar{Y}_\ell] = \left(\frac{1}{n_\ell}\right)\left(\frac{N_\ell - n_\ell}{N_\ell - 1}\right)\sigma_\ell^2$ for *SRS without replacement*.

5.4 Horvitz-Thompson Estimator

Definition 31 (Horvitz-Thompson estimator)**.** Let the sampling design be given by $P(I_1 = i_1, \ldots, I_n = i_n)$—i.e., any sampling design, with or without replacement. Let $C_j = \mathbb{1}\{I_1 = j\} + \ldots + \mathbb{1}\{I_n = j\}$—i.e., the number of times we sample the jth unit in the population. Let $\pi_j = P(C_j \geq 1)$—i.e., the probability we sample the jth unit in the population. Assume N and $\pi_j > 0$ are known $\forall\ j \in \{1, \ldots, N\}$. The *Horvitz-Thompson estimator* for μ is given by $\hat{\mu}_{\text{HT}} = \frac{1}{N}\sum_{j=1}^{N} \frac{\mathbb{1}\{C_j \geq 1\}}{\pi_j} y_j$.

• The Horvitz-Thompson estimator is unbiased for μ (i.e., $\mathbb{E}_I[\hat{\mu}_{\text{HT}}] = \mu$).

• ☣: $\hat{\mu}_{\text{HT}}$ is for $\mu = \frac{1}{N}\sum_{j=1}^{N} y_j$, but we can estimate $\sum_{j=1}^{N} y_j$ by removing the $\frac{1}{N}$ term.

• **Strategy**: If you know π_j (i.e., the probability of sampling the jth unit). If you want an unbiased estimator in the context of sampling.

Concept Checker 19. A treasure trove consists of N gems, labeled $1, \ldots, N$. Each gem got a quick, rough appraisal. For gem j, let v_j be its true value and b_j be its appraised value. Currently, N and b_j are known while v_j is not. Let $v_{\text{total}} = v_1 + \ldots + v_N$ and $b_{\text{total}} = b_1 + \ldots + b_N$. We wish to estimate v_{total} by sampling n gems based on their appraised value. We include gem j in our sample with probability $\frac{b_j}{b_{\text{total}}}$. Propose an estimator for v_{total} and show it is unbiased.[5]

[4] For some intuition, the sampling is done independently across strata, so the covariance terms from the bilinearity of variance disappear.

[5] Inspired by Problem 2 in "Stat 111 Homework 10, Spring 2026" by Joseph K. Blitzstein and Neil Shephard.

Solution

6 Resampling

6.1 Bootstrapping

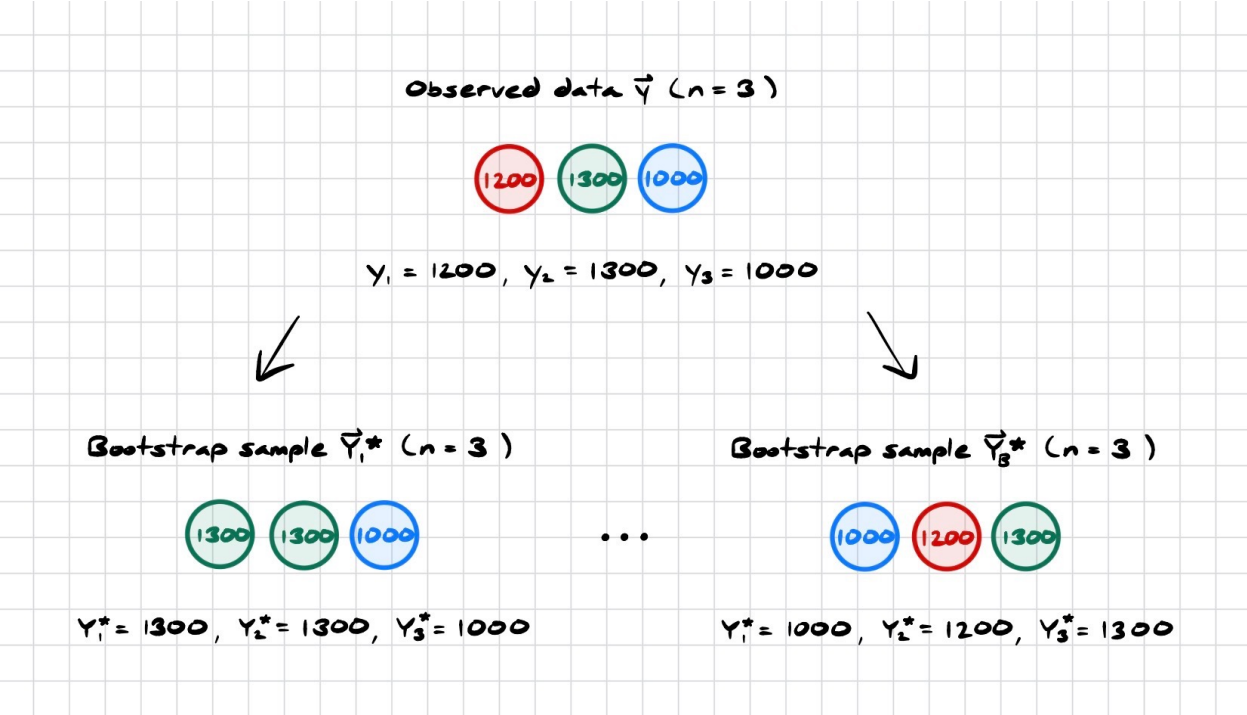

FIGURE 2: The idea behind bootstrapping with $n = 3$.

Definition 32 (Bootstrap)**.** Let $\vec{y} = (y_1, \ldots, y_n)$ be the *observed* dataset, assumed to be i.i.d. from some unknown CDF F. We create a synthetic dataset $\vec{Y}^* = (Y_1^*, \ldots, Y_n^*)$ by performing an *SRS with replacement* from $\vec{y}$.[6] We call $\vec{Y}^*$ a *bootstrap sample*.

• *Bootstrapping* involves generating some large number B of independent bootstrap samples, each of size n, and using the bootstrap samples for inferential tasks.

Definition 33 (Real world)**.** F generates $\vec{Y}$, which generates $\hat{\theta}$. We regard $\hat{\theta}$ as an estimate for θ.

Definition 34 (Bootstrap world)**.** $\hat{F}$ generates $\vec{Y}^*$, which generates $\hat{\theta}^*$. We regard $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ as estimates for $\hat{\theta}$, where $\hat{\theta}_j^*$ is a statistic calculated from the jth bootstrap sample $\vec{Y}_j^*$. We define the following bootstrap-world quantities, where we use the *bootstrap expectation* (i.e., *sampling with replacement* from $\hat{F}$, conditional on the observed data $\vec{y}$):

• $\text{Bias}_{\text{boot}}[\hat{\theta}^*] = \mathbb{E}_{\text{boot}}[\hat{\theta}^*] - \hat{\theta}$.

• $\text{SE}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\mathbb{E}_{\text{boot}}[(\hat{\theta}^* - \mathbb{E}_{\text{boot}}[\hat{\theta}^*])^2]}$.

• $F_{\text{boot}}(\theta) = P_{\text{boot}}(\hat{\theta}^* \leq \theta)$.

[6] Equivalently, $Y_j^* \overset{\text{i.i.d.}}{\sim} \hat{F}$, where $\hat{F} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i \leq y\}$ is the ECDF of $\vec{y}$. The ECDF places mass $\frac{1}{n}$ on each of the n observations.

Concept Checker 20. Suppose we generate one bootstrap sample $\vec{Y}^* = (Y_1^*, \ldots, Y_n^*)$. Show the bootstrap expectation of an arbitrary point in the bootstrap sample is the observed sample mean (i.e., $\mathbb{E}_{\text{boot}}[Y_j^*] = \bar{y}$).

Solution

Definition 35 (Computational cost of bootstrap)**.** Suppose we want the bootstrap expectation of some statistic $\hat{\theta}^* = T(\vec{Y}^*)$, but this can be very computationally expensive for even moderately large n.[7] Thus, in practice, we use *Monte Carlo estimation* (i.e., we use only a subset of all possible bootstrap samples).

Definition 36 (Monte Carlo estimation)**.** *Monte Carlo estimation* works since the bootstrap replications $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ are i.i.d. draws from the bootstrap world, so each estimator converges in probability to its respective bootstrap-world quantity as $B \to \infty$ by LLN. We use the bootstrap replications to calculate the following estimators:

- $\hat{\mathbb{E}}_{\text{boot}}[\hat{\theta}^*] = \bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^* \xrightarrow{p} \mathbb{E}_{\text{boot}}[\hat{\theta}^*]$.
- $\widehat{\text{Bias}}_{\text{boot}}[\hat{\theta}^*] = \bar{\theta}^* - \hat{\theta} \xrightarrow{p} \text{Bias}_{\text{boot}}[\hat{\theta}^*]$.
- $\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2} \xrightarrow{p} \text{SE}_{\text{boot}}[\hat{\theta}^*]$.
- $\hat{F}_{\text{boot}}(\theta) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\hat{\theta}_b^* \leq \theta\} \xrightarrow{p} F_{\text{boot}}(\theta)$.

Example 8 (Standard error)**.** We want $\text{SE}[\hat{\theta}] = \sqrt{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}$. We can approximate this with $\text{SE}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\mathbb{E}_{\text{boot}}[(\hat{\theta}^* - \mathbb{E}_{\text{boot}}[\hat{\theta}^*])^2]}$. But this is often computationally expensive, so in practice, we estimate it with $\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*] = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2}$. Crucially, there are two different sources of error.

- The difference between $\text{SE}_{\text{boot}}[\hat{\theta}^*]$ and $\text{SE}[\hat{\theta}]$ is caused by the use of $\hat{F}$ over F. This error falls as n increases.

[7] E.g., for $n = 30$, $30^{30} \approx 2 \times 10^{44}$.

• The difference between $\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*]$ and $\text{SE}_{\text{boot}}[\hat{\theta}^*]$ is caused by the use of Monte Carlo simulation. This error falls as B increases, which is under our control.

Definition 37 (Bootstrap confidence intervals)**.** Let θ be the estimand and $\hat{\theta}$ be an estimator for θ. Suppose we want to use the bootstrap to get an approximate 95% confidence interval for θ.[8] We have three different procedures:

• Normal interval with bootstrap standard error: $\hat{\theta} \pm 1.96 \times \widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*]$.

• Percentile method: $\left[\hat{\theta}^*_{(\lceil 0.025B \rceil)}, \hat{\theta}^*_{(\lceil 0.975B \rceil)}\right]$.

• Bootstrap t interval: $\left[\hat{\theta} - \hat{Q}^*(0.975)\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*], \hat{\theta} - \hat{Q}^*(0.025)\widehat{\text{SE}}_{\text{boot}}[\hat{\theta}^*]\right]$, where $\hat{Q}^*(p) = T^*_{(\lceil Bp \rceil)}$ is the sample p-quantile of the bootstrap replications $T^*_1, \ldots, T^*_B$ of $T = \frac{\hat{\theta}-\theta}{\widehat{\text{SE}}[\hat{\theta}]}$.[9]

6.2 Permutation Tests

Definition 38 (Permutation test)**.** Suppose there are two groups. We observe $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F_X$ for group 0 and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_Y$ for group 1 and assume $\vec{X} \perp\!\!\!\perp \vec{Y}$. We can use a *permutation test* for the hypotheses $H_0 : F_X = F_Y$ vs. $H_A : F_X \neq F_Y$.

• **Strategy**: Let T be a test statistic, chosen such that large values of T are evidence against H_0.[10] Compute the observed value t_0 of T. Generate a large number B of random permutations of the data, each an SRS without replacement of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.[11] For each permutation, compute the test statistic and call these $t_1, \ldots, t_B$. The p-value is $P_0(T \geq t_0) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{t_b \geq t_0\}$.[12]

• ~~Be~~: When determining which values are "more extreme" in the p-value, consider the direction specified in the hypotheses.

7 Causal Inference

7.1 Fundamentals

Definition 39 (Outcome)**.** For n units, Y_i is the *outcome* for unit i.

• E.g., suppose we are interested in the causal effect of tutoring on SAT score. Then Y_1 is the SAT score for unit 1.

Definition 40 (Treatment)**.** For n units, W_i is the *treatment* for unit i. The *treatment vector* is $\vec{W} = (W_1, \ldots, W_n)$.

• E.g., if $W_i \in \{0, 1\}$ is binary, then $W_1 = 1$ indicates unit 1 received tutoring while $W_1 = 0$ indicates unit 1 did not.

[8] We choose $\alpha = 0.05$ to simplify notation, but this can be generalized to any α.

[9] $T^*_j = \frac{\hat{\theta}^*_j - \hat{\theta}}{\text{SE}_{\text{boot}}[\hat{\theta}^*_j]}$, but like before, $\text{SE}_{\text{boot}}[\hat{\theta}^*_j]$ may be computationally expensive, so we replace it with the Monte Carlo estimate, which may involve a second "level" of bootstrapping.

[10] For example, one common choice is $T = |\bar{X} - \bar{Y}|$.

[11] This scrambles which data points belong to which groups.

[12] P_0 denotes probability under the permutation distribution of T (i.e., the distribution under random shuffles of the data rather than under repeated samples).

Definition 41 (Potential outcome)**.** For n units, $Y_i(w_1, \ldots, w_n)$ is the *potential outcome* for unit i. It is a function of all possible *treatments* (i.e., the potential outcome if $W_1 = w_1, \ldots, W_n = w_n$).

• E.g., for $n = 3$, $Y_1(1, 0, 1)$ is the potential SAT score for unit 1 if, hypothetically, units 1 and 3 received tutoring while unit 2 did not.

Definition 42 (Individual treatment effect)**.** For unit i, $\tau_i = Y_i(1) - Y_i(0)$.

Concept Checker 21. Which of the following are causal quantities?

1. $Y_i(1) - Y_i(0)$
2. $\mathbb{E}[Y_i(0)]$
3. $\mathbb{E}[Y_i \mid W_i = 0]$
4. $\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$

Solution

Definition 43 (Consistency)**.** Under *consistency*, $Y_i = Y_i(W_i)$ (i.e., unit i's outcome is their potential outcome under observed treatment W_i).

• E.g., units 1 and 2 both receive tutoring, but unit 1 gets multiple hours of hands-on practice while unit 2 only gets a worksheet, so consistency is violated.

Definition 44 (Non-interference)**.** Under *non-interference*, $Y_i(W_1, \ldots, W_n) = Y_i(W_i)$ (i.e., unit i's potential outcome is only a function of their own treatment).

• E.g., units 1 and 2 are best friends who study together and share exam advice. If unit 2 gets tutoring while unit 1 does not, $Y_1(0, 1)$ is still different from $Y_1(0, 0)$, so non-interference is violated.

Definition 45 (Stable unit treatment value assumption)**.** Under *SUTVA*, both *consistency* and *non-interference* hold.

Definition 46 (Unconfoundedness)**.** Under *unconfoundedness*, $Y_i(0), Y_i(1) \perp\!\!\!\perp W_i$ (i.e., unit i's treatment is independent of their potential outcomes).

• E.g., unit 1 doesn't like tutoring while unit 2 desperately wants to improve their score. Magically, we know $Y_1(0) = 1000$ and $Y_1(1) = 1050$ while $Y_2(0) = 1000$ and $Y_2(1) = 1500$. If students choose whether to get tutoring, there is self-selection since $Y_2(1)$ being high affects W_2, so unconfoundedness is violated.

Definition 47 (Binary treatment)**.** Under *binary treatment*, $W_i \in \{0, 1\}$.

• Unless otherwise noted, we will often assume *binary treatment* and *SUTVA* since this dramatically simplifies notation.

Definition 48 (Switching equation)**.** Assume *binary treatment* and *SUTVA*. $Y_i = W_i Y_i(1) + (1 - W_i)Y_i(0)$.

Concept Checker 22. Assume binary treatment. Suppose there are n units. How many potential outcomes are there for each unit i? What if we assume SUTVA?

Solution

Definition 49 (Set of potential outcomes)**.** Assume *binary treatment* and *SUTVA*. $\{\vec{Y}(0), \vec{Y}(1)\}$, where $\vec{Y}(0) = (Y_1(0), \ldots, Y_n(0))$ and $\vec{Y}(1) = (Y_1(1), \ldots, Y_n(1))$.

Definition 50 (Randomization)**.** An *assignment mechanism* where $P(\vec{W} = \vec{w} \mid \{\vec{Y}(0), \vec{Y}(1)\}) = P(\vec{W} = \vec{w})$.

- An experiment where treatment is randomized is called a *randomized control trial* (RCT).
- Notice the assumption of *unconfoundedness* holds under *randomization*.

7.2 The Two Approaches

	Super-Population	**Finite-Sample**
Population	Units are drawn i.i.d. from a larger super-population	The n units in the study are fixed and finite
Inference	ATE for super-population	ATE for sample
Source(s) of randomness	Treatment assignment, sampling	Treatment assignment
Potential outcomes	$\{Y_i(0), Y_i(1)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}^*$	A fixed set: $\vec{Y}(0) = \vec{y}(0), \vec{Y}(1) = \vec{y}(1)$
Random variables	$W_i, Y_i(0), Y_i(1)$	W_i
Data	$Y_i = W_i Y_i(1) + (1 - W_i)Y_i(0)$	$Y_i = W_i y_i(1) + (1 - W_i)y_i(0)$
Estimand	$\tau_{\text{PATE}} = \mathbb{E}[Y_i(1) - Y_i(0)]$	$\tau_{\text{SATE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i(1) - y_i(0))$
Estimator	DIM/MOM/MLE: $\hat{\tau}_{\text{PATE,DIM}} = \bar{Y}_1 - \bar{Y}_0$	MOM/DIM: $\hat{\tau}_{\text{SATE,MOM}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i W_i}{\mathbb{E}[W_i]} - \frac{Y_i(1-W_i)}{\mathbb{E}[1-W_i]}\right)$

7.3 Super-Population Approach

Definition 51 (Setup for super-population approach)**.** Assume *binary treatment*, *SUTVA*, and *randomization*. Let $\{Y_i(0), Y_i(1)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}^*$, where $\mathbb{P}^*$ is the super-population (i.e., we assume the potential outcomes are an i.i.d. sample from a statistical model).

Definition 52 (Population average treatment effect)**.** $\tau_{\text{PATE}} = \mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$.

- I.e., the expected treatment effect across all units in the super-population.

Definition 53 (DIM estimator for PATE)**.** Assume *binary treatment*, *SUTVA*, and *randomization*. Let $n_0 = \sum_{i=1}^{n}(1-W_i)$ and $n_1 = \sum_{i=1}^{n} W_i$ such that $n = n_0+n_1$. $\hat{\tau}_{\text{PATE,DIM}} = \bar{Y}_1 - \bar{Y}_0$, where $\bar{Y}_w = \frac{1}{n_w}\sum_{i:W_i=w} Y_i = \frac{1}{\sum_{i=1}^{n}\mathbb{1}\{W_i=w\}}\sum_{i=1}^{n}\mathbb{1}\{W_i = w\}Y_i$.

• We can show $\hat{\tau}_{\text{PATE,DIM}} = \hat{\tau}_{\text{PATE,MOM}}$ and, under *binary outcome*, $\hat{\tau}_{\text{PATE,DIM}} = \hat{\tau}_{\text{PATE,MLE}}$.

Definition 54 (Properties of DIM estimator for PATE)**.** Assume *binary treatment*, *SUTVA*, and *randomization*. The DIM estimator is unbiased and achieves CRLB.

• $\mathbb{E}[\hat{\tau}_{\text{PATE,DIM}}] = \tau_{\text{PATE}}$.

• $\text{Var}[\hat{\tau}_{\text{PATE,DIM}}] = \frac{\text{Var}[Y_i|W_i=1]}{n_1} + \frac{\text{Var}[Y_i|W_i=0]}{n_0}$.

Concept Checker 23. Y_i is deterministic from W_i in the finite-sample approach, but this is not true in the super-population approach. Explain this mathematically and intuitively.

Solution

7.4 Finite-Sample Approach

Definition 55 (Setup for finite-sample approach)**.** Assume *binary treatment*, *SUTVA*, and *randomization*. We condition on $\vec{Y}(0) = \vec{y}(0), \vec{Y}(1) = \vec{y}(1)$. Importantly, by the *switching equation*, Y_i is deterministic from W_i given the potential outcomes.[13]

Definition 56 (Sample average treatment effect)**.** $\tau_{\text{SATE}} = \bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}(y_i(1) - y_i(0))$.

Definition 57 (MOM estimator for SATE)**.** Assume *binary treatment*, *SUTVA*, and *randomization*. $\hat{\tau}_{\text{SATE,MOM}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i W_i}{\mathbb{E}[W_i]} - \frac{Y_i(1-W_i)}{\mathbb{E}[1-W_i]}\right)$.

• It can be shown $\hat{\tau}_{\text{SATE,MOM}}$ is the difference in means estimator under *complete randomization*, where $P(W_i = 1) = \frac{n_1}{n}\ \forall\ i \in \{1,\ldots,n\}$.

Definition 58 (Properties of MOM estimator for SATE)**.** Assume *binary treatment*, *SUTVA*, and *randomization*. The MOM estimator is conditionally unbiased given $\vec{Y}(w) = \vec{y}(w)$.

• $\mathbb{E}[\hat{\tau}_{\text{SATE,MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \tau_{\text{SATE}}$.

• $\text{Var}[\hat{\tau}_{\text{SATE,MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{(y_i(1))^2}{\mathbb{E}[W_i]} + \frac{(y_i(0))^2}{\mathbb{E}[1-W_i]} - (y_i(1) - y_i(0))^2\right)$.

Concept Checker 24. Show $\mathbb{E}[\hat{\tau}_{\text{SATE,MOM}} \mid \vec{Y}(w) = \vec{y}(w)] = \tau_{\text{SATE}}$.

[13] E.g., I have a weight y that is not random. It's just a number at the end of the day. We regard what my weight would be with and without medication—$y(1)$ and $y(0)$, respectively—as also not random. What is random is Y, which of the two will be observed. The randomness comes from treatment assignment: $Y = y(0)$ if $W = 0$ and $Y = y(1)$ if $W = 1$.

Solution

7.5 Confidence Intervals and Hypothesis Tests

Definition 59 (Fisher's null hypothesis)**.** In the *finite-sample approach*, there is no treatment effect for each unit. I.e., $H_0 : \tau_i = 0 \ \forall \ i \in \{1, \ldots, n\}$.

• Alternatively, there is a treatment effect for at least one unit. I.e., $H_A : \sum_{i=1}^{n} |\tau_i| > 0$.

• Under H_0, $Y_i(0) = Y_i(1) = Y_i$ because $\tau_i = Y_i(1) - Y_i(0) = 0$.

• **Key insight:** Fisher's null is called the *sharp* null because it fills in the missing potential outcomes. Under H_0, we know both $Y_i(0)$ and $Y_i(1)$ for every unit (they're both just Y_i). This means we can compute $\hat{\tau}_{\text{SATE,MOM}}$ for *any* hypothetical assignment vector $\vec{W}$, not just the one we observed.

Definition 60 (Randomization test)**.** To test *Fisher's null hypothesis*, generate B i.i.d. draws $\vec{W}^{(1)}, \ldots, \vec{W}^{(B)}$ from the assignment mechanism. For each draw b, compute $\hat{\tau}^{(b)}$ using the observed Y_i's and the hypothetical assignment $\vec{W}^{(b)}$. The randomization p-value is $p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{\left|\hat{\tau}^{(b)}\right| \geq |\hat{\tau}_{\text{SATE,MOM}}|\right\}$.

• Intuitively, if the null is true, the observed $\hat{\tau}_{\text{SATE,MOM}}$ should look like a typical draw from this randomization distribution. A small p-value means our observed test statistic is unusually large, which is hard to explain by chance alone.

Definition 61 (Neyman's null hypothesis)**.** In the *finite-sample approach*, the SATE is 0. I.e., $H_0 : \tau_{\text{SATE}} = 0$.

• Alternatively, $H_A : \tau_{\text{SATE}} \neq 0$.

• Fisher's null implies Neyman's null, but the converse is not true since individual effects can cancel. E.g., $\tau_1 = 1, \tau_2 = -1, \ldots$ gives $\tau_{\text{SATE}} = 0$, but Fisher's null fails.

• **Key insight:** Neyman's null is weaker since it only cares about the average, not every individual. We test it via the asymptotic Normal pivot, using the conservative variance estimator from before:

$$T = \frac{\hat{\tau}_{\text{SATE,MOM}}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}} \dot{\sim} \mathcal{N}(0,1).$$

Reject H_0 when $|T| > z_{\alpha/2}$. A nominal $100(1-\alpha)\%$ CI for τ_{SATE} is $\hat{\tau}_{\text{SATE,MOM}} \pm z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}$.

Definition 62 (Super-population MLE estimator)**.** In the *super-population approach*, recall the conditional expectation and variance of $\hat{\tau}_{\text{PATE,MLE}}$. We use the asymptotic Normality of the MLE and the "plug-in" principle to substitute in estimators to construct a pivot: $\frac{\hat{\tau}_{\text{PATE,MLE}} - \tau_{\text{PATE}}}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{\sum_{i=1}^n w_i} + \frac{\hat{\theta}_0(1-\hat{\theta}_0)}{\sum_{i=1}^n (1-w_i)}}} \mid \vec{W} = \vec{w} \dot{\sim} \mathcal{N}(0,1)$.

• To test $H_0 : \tau_{\text{PATE}} = 0$, we use $T = \frac{\hat{\tau}_{\text{PATE,MLE}}}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{\sum_{i=1}^n w_i} + \frac{\hat{\theta}_0(1-\hat{\theta}_0)}{\sum_{i=1}^n (1-w_i)}}} \mid \vec{W} = \vec{w} \dot{\sim} \mathcal{N}(0,1)$.

Concept Checker 25. Suppose our observed data are $\vec{Y} = (3,6,7,2)$ and $\vec{W} = (0,1,1,0)$. What is $\hat{\tau}_{\text{DIM}}$? If we assume Fisher's null, what would $\hat{\tau}^{(1)}_{\text{DIM}}$ be from the randomly-generated treatment $\vec{W}^{(1)} = (1,0,1,0)$?

Solution

8 Practice Problems

Problem 1. Suppose we are interested in the causal effect of hands-on tutoring on SAT score for all high school students. There are 3 treatment levels: $W = 2$ for hands-on tutoring, $W = 1$ for hands-off tutoring, and $W = 0$ for control. We randomly sample n students and randomize treatment assignment. Let Y_i be the SAT score for student i. Assume SUTVA and $P(W = 2) > 0$.

(a) Would the super-population approach or finite-sample approach be more appropriate for this problem?

(b) Suppose we are interested in $\mathbb{E}[Y(2)]$. What type of quantity is $\mathbb{E}[Y(2)]$? State what it is clearly in words.

(c) Identify $\mathbb{E}[Y(2)]$ as a statistical quantity.

(d) Construct a consistent estimator: $\hat{\mathbb{E}}[Y(2)]$.

Solution

Problem 2. A company has been accused of gender discrimination, allegedly tending to pay higher salaries to men than to women. Suppose $X_1, \ldots, X_n$ are i.i.d. draws from a theoretical distribution of salaries for men at the company and $Y_1, \ldots, Y_m$ are i.i.d. draws from a theoretical distribution of salaries for women at the same company. The salaries $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent but not necessarily i.i.d. Let $T = \bar{X}_n - \bar{Y}_m$. Intuitively, we would like to know whether T has such a large positive value it would be implausible to observe that value (or an even larger value) if the two theoretical distributions were equal. However, n and m are small. The CLT is not applicable with such small sample sizes, so it is not reasonable to assume $\bar{X}_n$ and $\bar{Y}_m$ are approximately Normal.[14]

(a) Explain how we can use a permutation test to answer our scientific question of interest.

(b) Suppose $n = m = 2$ and the data are $X_1 = 8, X_2 = 4, Y_1 = 6, Y_2 = 2$, measured in tens of thousands of dollars. Find the p-value for the permutation test.

Solution

Problem 3 (Normal posterior predictive)**.** A random sample of n students is drawn from a population whose weights are $\mathcal{N}(\theta, 400)$ ($\sigma = 20$). The sample mean is $\bar{y} = 150$. Use prior $\theta \sim \mathcal{N}(180, 1600)$ ($\tau_0 = 40$).

[14]Inspired by Problem 6 in "Stat 111 Final Exam, Spring 2025" by Joseph K. Blitzstein and Neil Shephard.

(a) Find the posterior distribution $\theta \mid y$ as a function of n.

(b) Find the posterior predictive distribution $\tilde{y} \mid y$ for a new student's weight, justifying parameters using Adam's and Eve's laws.

(c) For $n = 10$: give a 95% posterior interval for θ and a 95% posterior predictive interval for $\tilde{y}$.

Solution

Problem 4 (Posterior as a compromise)**.** Let Y be the number of heads in n coin flips with unknown probability θ.

(a) With a Unif$(0, 1)$ prior, derive the prior predictive distribution $P(Y = k)$.

(b) With $\theta \sim \text{Beta}(\alpha, \beta)$, show algebraically that the posterior mean lies strictly between the prior mean and observed frequency y/n.

(c) Show that if the prior is Unif$(0, 1)$, the posterior variance is always less than the prior variance.

(d) Give an example of a Beta(α, β) prior and data (n, y) where the posterior variance *exceeds* the prior variance.

Solution

Problem 5 (Censored and uncensored data)**.** Suppose $Y \mid \theta \sim \text{Expo}(\theta)$ with conjugate prior $\theta \sim \text{Gamma}(\alpha, \beta)$.

(a) We observe $Y \geq 100$. Find $\pi(\theta \mid Y \geq 100)$ and the posterior mean and variance.

(b) Now suppose we observe $Y = 100$ exactly. Find $\pi(\theta \mid Y = 100)$ and the posterior mean and variance.

(c) Why is the posterior variance *higher* in (b), even though more information was given?

Solution

Problem 6 (Coverage of posterior intervals)**.** Consider any parametric model with scalar parameter θ.

(a) Prove that a Bayesian 50% credible interval contains the true θ with probability exactly 50% when θ is drawn from the prior.

(b) Suppose $\theta \sim \mathcal{N}(0, 4)$ and $Y \mid \theta \sim \mathcal{N}(\theta, 1)$. The true value is $\theta_0 = 1$. What is the exact frequentist coverage of the posterior 50% interval?

Solution

Problem 7 (Hierarchical model for SAT scores)**.** Randomized control trials on a tutoring program are run in $J = 7$ schools. Let y_j be the estimated average treatment effect with known standard error σ_j:

School j	y_j	σ_j
1	3	8
2	16	6
3	−6	12
4	−7	6
5	4	8
6	8	25
7	−1	5

Assume $Y_j \mid \theta_j \overset{\text{ind.}}{\sim} \mathcal{N}(\theta_j, \sigma_j^2)$, $\theta_j \mid \mu, \lambda \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \lambda^2)$, $\mu \mid \lambda \sim \mathcal{N}(0, 6^2)$, with λ known.

(a) Find $\theta_j \mid \mu, Y, \lambda$. Explain why $\mathbb{E}[\theta_j \mid \mu, Y, \lambda]$ makes sense.

(b) Find $\mu \mid Y, \lambda$. Explain why $\mathbb{E}[\mu \mid Y, \lambda]$ makes sense.

(c) Find $\theta_j \mid Y, \lambda$. Explain why $\mathbb{E}[\theta_j \mid Y, \lambda]$ makes sense.

(d) Plot $\mathbb{E}[\theta_j \mid Y, \lambda]$ vs. $\lambda \in [0, 40]$ for all 7 schools. Hint: `matplot` in R.

(e) Describe a principled way to infer λ if it is unknown.

Solution

Problem 8 (James-Stein and batting averages)**.** A sabermetrician estimates batting averages $\mu_1, \dots, \mu_k$ of $k > 3$ players. Let $Y_j \mid \mu_j \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_j, V)$ with $\mu_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, \sigma_\mu^2)$ a priori. Use total squared error loss.

(a) Find $\hat{\mu}_{\text{MLE}}$ and risk $r_{\text{MLE}}(\mu)$.

(b) The Bayes estimator has jth component $\mathbb{E}[\mu_j \mid Y] = b\mu_0 + (1-b)Y_j$ where $b = V/(V+\sigma_\mu^2)$. Find $r_{\text{Bayes}}(\mu)$. Is it always smaller than $r_{\text{MLE}}(\mu)$?

(c) Explain intuitively how $\hat{\mu}_{\text{Bayes}}$ relates to regression toward the mean.

(d) Show $\hat{b} = (k-3)V/S$ is unbiased for b, where $S = \sum_j (Y_j - \bar{Y})^2$.

Solution