STAT 111 Final Cheatsheet

Adapted from https://github.com/zadchin/STAT-111-Cheatsheet, Will Nickols' review and Ethan Tan's sheet. Edited by Je Qin Chooi (2023).

1 Models and Likelihood

Key terms

In the classic inference problem, we start by considering the i.i.d. observations

$$\vec{Y} = (Y_1, \ldots, Y_n),$$

which are random variables representing the data, which then crystallize to

$$\vec{y} = (y_1, \ldots, y_n).$$

A *statistic* is a function T of the random vector $\vec{Y}$, and common statistics include the sample mean, sample median, sample mode, sample variance, and various quantiles of the data. We assume that the data we collect behave according to a *model*. This model is *parametric* if θ is finite-dimensional and *nonparametric* if θ is infinite-dimensional. Then,

- An *estimand* is a quantity of interest. Example: θ.
- An *estimator* is a random variable that encapsulates the method we use to estimate the estimand. Example: $\bar{Y}$.
- An *estimate* is a number that represents the crystallized version of some constructed estimator. Example: $\bar{y}$.

Likelihoods

The *likelihood* function describes the probability of observing the data. In other words, it is a function L of the estimand θ given fixed data $\vec{y}$:

$$L(\theta) = p(\vec{y} \mid \theta).$$

In the special case where $\vec{y} = (y_1, \ldots, y_n)$, with the y_j's coming from i.i.d. random variables, we can factor the joint density $p(\vec{y} \mid \theta) = p(y_1, \ldots, y_n \mid \theta)$ and get

$$L(\theta) = \prod_{j=1}^{n} p(y_j \mid \theta).$$

Log-Likelihoods

$$\ell(\theta) = \log L(\theta)$$

In the usual case of $y_1, \ldots, y_n$ coming from i.i.d. random variables, we find that the log-likelihood is a sum of n terms, and taking derivatives is now easy:

$$\ell(\theta) = \log \prod_{j=1}^{n} p(y_j \mid \theta) = \sum_{j=1}^{n} \log p(y_j \mid \theta).$$

Reparameterization

Given two estimands θ and ψ, and an injective function g such that $\psi = g(\theta)$, we have $L(\psi; \vec{y}) = L(\theta; \vec{y})$. To find $L(\psi; \vec{y})$, plug in $g^{-1}(\psi)$ for θ everywhere in $L(\theta; \vec{y})$. Example: $L(\sigma^2; \vec{y}) = L(\sigma; \vec{y})$

Data Transformation

If the original $\vec{y}$ can be reconstructed from $h(\vec{y})$, then $L(\theta; \vec{y}) = L(\theta; h(\vec{y}))$. Example: $L(\sigma^2; \vec{y}) = L(\sigma^2, \exp(\vec{y}))$

Censored Data

Some of the data is observed and some is in an unobserved region.

- Strategy: Find the likelihood function by multiplying the PDF/PMF of known values by the CDF (or difference in CDFs) for the unobserved region.
- Example: Let $Y_1, \ldots, Y_n \sim \text{Expo}(\lambda)$ be observed particle emission times, but the clock broke from time t_1 to t_2 and we know m decays occurred during that time. Then

$$L(\lambda; \vec{y}) = \left(\prod_{i=1}^{n} f_y(y_i) \right) (F(t_2) - F(t_1))^m$$

Bias, Standard Error and Loss

- The **bias** of an estimator $\hat{\theta}$ for an estimand θ is $\text{Bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta$. .
- The **standard error** of an estimator $\hat{\theta}$ is $\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$. You might notice that this is the same as the standard deviation of $\hat{\theta}$.
- A **loss function** is a function $\text{Loss}(\theta, x) \geq 0$ that is assumed to be convex in x with the property that $\text{Loss}(x, x) = 0$ for all x. Examples of loss functions include:
  - 0-1 loss: $\text{Loss}(\theta, x) = I(\theta \neq x)$.
  - Absolute error loss: $\text{Loss}(\theta, x) = |\theta - x|$.
  - Squared error loss: $\text{Loss}(\theta, x) = (\theta - x)^2$.
- The expectation of the squared error loss is called the **mean squared error** (MSE):

$$\text{MSE}(\theta, \hat{\theta}) = \text{E}[(\theta - \hat{\theta})^2]$$

- **Bias-Variance Decomposition**

$$\text{MSE}(\theta, \hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}).$$

Consistency of Estimators

An estimator $\hat{\theta}$ is said to be *consistent* for the estimand θ if the convergence

$$\hat{\theta} \xrightarrow{p} \theta$$

holds; that is, if $\hat{\theta}$ converges in probability to the true value of the estimand.

Proving Consistency To show that some $\hat{\theta}$ is consistent for a corresponding θ:

1. Show that $\text{MSE}(\hat{\theta}, \theta) \to 0$.
2. Use WLLN if $\theta = \text{E}(Y_1)$ and $\hat{\theta} = \bar{Y}$ for some i.i.d. $Y_1, \ldots, Y_n$.
3. Use WLLN and CMT if $\theta = g(\text{E}(Y_1))$ and $\hat{\theta} = g(\bar{Y})$ for some continuous function g and i.i.d. $Y_1, \ldots, Y_n$.
4. Fix some $\epsilon > 0$, and show that $\mathbb{P}(|\hat{\theta} - \theta| > \epsilon) \to 0$ directly.

Nonparametric methods

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

Sample Quantile If $Y \sim F$, the rth *quantile* is $Q(r) = \min\{x \mid F(x) \geq r\}$. Then, for i.i.d. random variables $Y_1, \ldots, Y_n$, one can consider the order statistics $Y_{(1)}, \ldots, Y_{(n)}$, and the rth *sample quantile* is $\hat{Q}(r) = Y_{(\lceil nr \rceil)}$. The sample median is thus $\hat{Q}(0.5) = Y_{(\lceil n/2 \rceil)}$.

2 MLE and MOM

MLE Invariance

If g is injective and $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$. After all, maximizing $L(\psi)$ is equivalent to maximizing $L(\theta)$ because applying g is an inequality preserving operation.

Methods of Moments

Finding MoM

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables. The *method of moments* (MoM) estimator for some parameter θ is found by:

1. Put hats on: replace components of θ with components of $\hat{\theta}$.
2. Replace theoretical moments $\text{E}(Y_1^k)$ with sample moments $\overline{Y^k}$.

Finally, one can solve for each component of $\hat{\theta}$ in terms of sample moments.

Properties of MoM

If $\text{Var}(Y_1^k) < \infty$, then $\text{E}(\overline{Y^k}) = \text{E}(Y_1^k)$ and $\text{Var}(\overline{Y^k}) = \text{Var}(Y_1^k)/n$. Moreover, by the CLT, we obtain

$$\sqrt{n}(\overline{Y^k} - \text{E}(Y_1^k)) \xrightarrow{d} \mathcal{N}(0, \text{Var}(Y_1^k)).$$

Note that, in general, $\text{E}(\hat{\theta}) \neq \theta$ if $\hat{\theta}$ is an MoM estimator (i.e. biased), even though $\text{E}(\overline{Y^k}) = \text{E}(Y_1^k)$.

Score Function

The *score* function is defined to be the first derivative of the log-likelihood:

$$s(\theta) = \ell'(\theta).$$

To find the MLE, we set $s(\theta) = 0$ and solve for θ, which we call $\hat{\theta}$.

Remark: $\text{E}[s(\theta^*; \vec{Y})] = 0$

Regularity Conditions

1. The data is i.i.d. $f_\theta(y)$.
2. The support does not depend on θ. (unlike $\text{Unif}(0, \theta)$).
3. $\frac{\partial^3}{\partial \theta^3} f_\theta(y)$ exists.
4. θ^* is not on the boundary of the parameter space (unlike $\sigma = 0$).
5. We can differentiate under the integral sign.

Consistency of MLE: With regularity conditions and a correctly specified model, the MLE is consistent. $\hat{\theta} \xrightarrow{p} \theta$.

Information Equality

Under regularity conditions,

$$\text{E}(s(\hat{\theta}; \vec{y}) = 0$$

$$\text{Var}(s(\theta^*; \vec{Y})) = -\text{E}(s'(\theta^*; \vec{Y})).$$

The latter is known as information equality.

Fisher Information

The *Fisher information* for a parameter θ is defined as

$$\mathcal{I}_Y(\theta) = \text{Var}(s(\theta; \vec{Y}))$$

Remarks:

- Here we write $\mathcal{I}_Y(\theta)$ to mean the Fisher Information from the entire data set, and $\mathcal{I}_1(\theta)$ from one data point.
- Under regularity conditions, invoke the information equality and get $\mathcal{I}_Y(\theta) = -\text{E}[s'(\theta^*; \vec{Y})]$
- For i.i.d. $Y_1, \ldots, Y_n$, we have $\mathcal{I}_Y(\theta) = n\mathcal{I}_1(\theta)$.

Fisher under Reparameterization: Suppose that $\tau = g(\theta)$, where g is injective and differentiable. Then, we have the relation $\mathcal{I}_Y(\tau) = \mathcal{I}_Y(\theta)/g'(\theta)^2$.

Cramér-Rao Lower Bound

CRLB for Unbiased Estimators Under regularity conditions, if $\hat{\theta}$ is unbiased for θ, then

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_Y(\theta)}.$$

CRLB in the General Case If we make no assumptions about the bias of $\hat{\theta}$ for θ, we instead obtain the bound

$$\text{Var}(\hat{\theta}) \geq \frac{g'(\theta)^2}{\mathcal{I}_Y(\theta)},$$

where $g(\theta) = \text{E}[\hat{\theta}]$. In the zero bias case, it was the case that $g(\theta) = \theta$, so $g'(\theta) = 1$. We call estimators efficient if they achieve the CRLB.

3 Asymptotics

Convergence Equivalence

$Y_n \xrightarrow{p} Y$ (in probability) implies $Y_n \xrightarrow{d} Y$ (in distribution). Given a constant c, and r.v. Y_n, we have $Y_n \xrightarrow{d} c$ implies $Y_n \xrightarrow{p} c$.

Central Limit Theorem

Let $Y_1, \ldots, Y_n$ be i.i.d random variables with mean $\text{E}(Y_1) = \mu$ and finite variance $\text{Var}(Y_1) = \sigma^2 < \infty$. Then, the convergence

$$\frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

holds. Equivalently, we have $\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ or $\bar{Y} \dot{\sim} \mathcal{N}(\mu, \sigma^2/n)$.

- Useful with the factor $\sigma/\hat{\sigma}$ and Slutsky's Theorem to get consistent estimator for σ.
- Since raising an r.v. to a power is still an r.v., we also have for the general case of moments

$$\sqrt{n}(\overline{Y^k} - \text{E}(Y_1^k)) \xrightarrow{d} \mathcal{N}(0, \text{Var}(Y_1^k)$$

Law of Large Numbers

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with finite first moments, i.e. $\text{E}(|Y_1|) < \infty$. Then,

- SLLN $\bar{Y} \xrightarrow{a} \text{E}(Y_1)$
- WLLN $\bar{Y} \xrightarrow{p} \text{E}(Y_1)$

Similar to CLT we have for general moments

$$\overline{Y^k} \xrightarrow{p} \text{E}[Y_1^k]$$

Note that LLN gives convergence to a constant, so useful with Slutsky's Theorem.

Continuous Mapping Theorem

Let $g : \mathbb{R} \to \mathbb{R}$ be continuous on a set A, where $\mathbb{P}(Y \in A) = 1$. Then:

1. $Y_n \xrightarrow{p} Y$ implies $g(Y_n) \xrightarrow{p} g(Y)$.
2. $Y_n \xrightarrow{d} Y$ implies $g(Y_n) \xrightarrow{d} g(Y)$.

Slutsky's Theorem

Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is a constant (recall that the latter condition is equivalent to $Y_n \xrightarrow{p} c$). Then,

1. $X_n + Y_n \xrightarrow{d} X + c$.
2. $X_n Y_n \xrightarrow{d} cX$.
3. For $c \neq 0$, $X_n / Y_n \xrightarrow{d} X/c$.

This also holds for $Y_n \xrightarrow{d} Y$ if X_n and Y_n are independent.

Delta Methods (Indestructibility of the Normal)

Suppose that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$, and let g be a function that is continuously differentiable in a neighborhood of θ. Then,

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

Useful for asymptotic confidence intervals

Other Asymptotics

Asymptotics of the MLE With regularity conditions and a correctly specified model, the MLE has the asymptotic distribution

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

The MLE is asymptotically Normal, asymptotically unbiased, and asymptotically efficient (achieves the CRLB; MLE has the lowest SE asymptotically among all asymptotically unbiased estimators).

$$\text{Var}(\hat{\theta}) \approx \frac{1}{n\mathcal{I}_1(\theta^*)}$$

Asymptotics of Sample Quantiles Suppose X has the continuous CDF $F(x)$. By CMT, the pth sample quantile $\hat{Q}(p)$ has the asymptotic distribution

$$\sqrt{n}(\hat{Q}(p) - Q(p)) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{(f(Q(p)))^2}\right)$$

Asymptotics of the Empirical CDF

- $\hat{F}(y) \xrightarrow{p} F(y)$ by LLN
- $\sqrt{n}(\hat{F}(y) - F(y)) \xrightarrow{d} \mathcal{N}(0, F(y)(1 - F(y)))$ by CLT

4 Interval Estimation

Interval Estimator: An *interval estimator* is an interval $C(\vec{Y}) = [L(\vec{Y}), U(\vec{Y})]$ with $L(\vec{Y}) \leq U(\vec{Y})$.

Coverage Probability: $P(\theta \in C(\vec{Y})|\theta)$. It is a function of θ.

Confidence Interval: An interval estimator with a coverage probability at least $1 - \alpha$ for all possible values of θ is a $100(1-\alpha)\%$ *confidence interval* (CI). We say that $1 - \alpha$ is the *level* of our CI.

The CI is random, θ is fixed. The probability that the CI generated by repeated draws of data contains the fixed θ is $1 - \alpha$. CI is not unique.

Exact CIs

A *pivot* is a random variable that is a function of the data and the parameter(s), but the pivot itself has a known distribution that does not depend on the parameters (e.g. $\mathcal{N}(0, 1)$, Unif$(6, 9)$).

1. The Normal, t, Gamma, and χ^2 distrinutions are commonly used as pivots.
2. For $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ with both μ, σ^2 unknown, we have

$$\frac{\bar{Y} - \mu}{\sqrt{\hat{\sigma}^2/n}} \sim t_{n-1}$$

This gives a 95% CI for μ

$$[\bar{Y} - Q(0.975)\sqrt{\hat{\sigma}^2/n}, \bar{Y} - Q(0.025)\sqrt{\hat{\sigma}^2/n}]$$

where Q is the quantile function for t_{n-1}.

Asymptotic CIs

Use the CLT to find an asymptotic $\mathcal{N}(0, 1)$ pivot. When this pivot is not nice enough, we can further improve it by using the delta method, the CMT, or Slutsky's theorem. Construct a CI as we did before, with the caveat that the resulting interval is not exact for finite n.

Example: Let $Y_1, \ldots, Y_n \sim \text{Expo}(\lambda)$. By the CLT,

$$\frac{\sqrt{n}(\bar{Y} - 1/\lambda)}{\sqrt{1/\lambda^2}} \xrightarrow{d} \mathcal{N}(0, 1) \implies \sqrt{n}(\bar{Y}\lambda - 1) \xrightarrow{d} \mathcal{N}(0, 1)$$

With n large we can move n around and have the approximate interval

$$\left[\frac{Q(0.025)/\sqrt{n} + 1}{\bar{Y}}, \frac{Q(0.975)/\sqrt{n} + 1}{\bar{Y}}\right]$$

A useful shortcut Finally, we look at the asymptotic 95% CI that you will end up using 95% of the time. For i.i.d. $Y_1, \ldots, Y_n \sim [\theta, \sigma^2]$, with σ^2 known, we have

$$\sqrt{n}(\bar{Y} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

by the CLT. Then, an asymptotic 95% CI for θ is

$$\left[\bar{y} - \frac{1.96\sigma}{\sqrt{n}}, \bar{y} + \frac{1.96\sigma}{\sqrt{n}}\right],$$

where $\bar{y}$ is just the crystallized version of the sample mean random variable $\bar{Y}$ (CIs can be estimators or estimates depending on context).

5 Sufficient statistics and EFs

Sufficient Statistics

Let $\vec{Y} = (Y_1, \ldots, Y_n)$ be a sample from some model. A statistic $T(\vec{Y})$ is *sufficient* for θ if the conditional distribution of $\vec{Y} \mid T(\vec{Y})$ does not depend on θ. Sufficient statistics are not unique.

Factorization Criterion
The statistic $T(\vec{Y})$ is sufficient if and only if we can factor the joint density of $\vec{Y}$ as $p(\vec{y} \mid \theta) = g(T(\vec{y}), \theta)h(\vec{y})$. The likelihood can then be written purely in terms of $T(\vec{y})$ and θ.

Sufficiency of Order Statistics
Let p be *any* density parameterized by some scalar θ. Then, if $Y_1, \ldots, Y_n$ are i.i.d. with density p, it is the case that $(Y_{(1)}, \ldots, Y_{(n)})$ is sufficient for θ. After all,

$$L(\theta) \propto \prod_{j=1}^{n} p(y_j) = \prod_{j=1}^{n} p(y_{(j)}).$$

Rao-Blackwellization

Let T be a sufficient statistic and $\hat{\theta}$ be any estimator (in both cases with respect to the estimand θ). The Rao-Blackwellized estimator is

$$\hat{\theta}_{RB} = \text{E}(\hat{\theta} \mid T),$$

and the following holds: $\text{MSE}(\hat{\theta}_{RB}, \theta) \leq \text{MSE}(\hat{\theta}, \theta)$.

Remarks

- To Rao-Blackwellize an estimator, one must condition on sufficient statistics for the theorem to hold. The theorem fails for arbitrary statistics.
- A Rao-Blackwellized estimator will have the same bias but may have an improved (smaller variance). This follows from Adam's law and Eve's law.
- Rao-Blackwellization will not change an estimator if it was already a function of the sufficient statistic T in the first place. This follows directly from the "taking out what's known" property of conditional expectation.
- In particular, Rao-Blackwellization will not improve the MLE because the MLE is always a function of the sufficient statistics.
- To find the Rao-Blackwell estimator, you usually need to determine conditional distributions of the form $Y \mid T$, usually done by citing a Statistics 110 story.

NEFs

A random variable Y follows a *natural exponential family* (NEF) if one can write

$$p(y \mid \theta) = e^{\theta y - \Psi(\theta)} h(y).$$

We call θ the natural (canonical) parameter and note that $\Psi(\theta)$ is the cumulant generating function of Y_1 (the logarithm of the MGF of Y_1).

Properties of NEFs

Let Y_1 be NEF with the canonical form defined above. Then, we can deduce that

- $\mathrm{E}[Y_1] = \Psi'(\theta)$, $\mathrm{Var}(Y_1) = \Psi''(\theta)$, and the MGF $M_Y(t) = \mathrm{E}[e^{tY_1}] = e^{\Psi(\theta+t)-\Psi(\theta)}$.
- If $Y_1, \ldots, Y_n$ are i.i.d., $\bar{Y}$ is a sufficient statistic for θ.
- The MLE of $\mu = \mathrm{E}[Y_1]$ is $\hat{\mu} = \bar{Y}$.
- The Fisher information is $\mathcal{I}_1(\theta) = \Psi''(\theta) = \mathrm{Var}(Y_1)$.
- We call the process where we fix $h(y)$ as a baseline distribution from which we construct the entire NEF *exponential tilting*.
- Examples of NEFs: the Normal (with σ^2 known), the Poisson, the Binomial (with n fixed), the Negative Binomial (with r fixed), the Gamma (with a known), the First Success. Then χ^2 and Exponential follow from Gamma, Bernoulli follows from Binomial, and Geometric follows from Negative Binomial,

Exponential Families

A random variable Y follows an *exponential family* (EF) if one can write

$$p(y \mid \theta) = e^{\theta T(y) - \Psi(\theta)} h(y).$$

Some examples of EFs include the Weibull, the Normal (but now with μ known and σ^2 unknown), and the Normal (with both μ and σ^2 unknown). And to prove that a distribution follows a NEF or an EF, you need to manipulate a given density and pattern match to a general functional form (as when you find sufficient statistics).

6 Regression

For this section assume m data points $(Y_1, \vec{X}_1), \ldots, (Y_m, \vec{X}_m)$, with K covariates for each data point $\vec{X}_j = (\vec{X}_{j,1}, \ldots, \vec{X}_{j,K})$. For brevity we write $\vec{X}_j = \vec{X} = (X_1, \ldots, X_K)$. Outcome variable Y is always one dimensional.

Predictive regression is the task of estimating the conditional expectation

$$\mu(\vec{x}) = \mathrm{E}[Y | \vec{X} = \vec{x}]$$

where Y is called the *outcome variable* and X the *predictors* or *covariates*.

Regression error is the difference between random outcome and predicted outcome given predictors

$$U(\vec{x}) = Y - \mu(\vec{x})$$

We also write $Y = \mu(\vec{x}) + U(\vec{x})$ and interpret the random outcome as signal (known) plus noise (random).

- $U(\vec{x})$ is still r.v. with randomness from Y.
- $U(\vec{x})$ is unobservable since require knowing true $\mu(\vec{x})$ that we can only get an estimate of.
- $\mathrm{E}[U(\vec{X}) \mid \vec{X} = \vec{x}] = 0$, $\mathrm{E}[U(\vec{X})] = 0$.
- $\mathrm{Cov}(U(\vec{X}), \vec{X}) = 0$. Predictors and noise are uncorrelated.

From Eve's Law, variation in outcome is the sum of the variation in prediction and the variation in random noise.

$$\mathrm{Var}(Y) = \mathrm{Var}(\mu(\vec{X})) + \mathrm{Var}(U(\vec{X}))$$

The R^2 statistic is defined as

$$R^2 = \frac{\mathrm{Var}(\mu(\vec{X}))}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(U(\vec{x})}{\mathrm{Var}(Y)}$$

Linear Regression

When the regression function is a linear function of the parameters, we have *linear regression*

$$\mu(\vec{x}) = \mathrm{E}(Y \mid \vec{X} = \vec{x}, \vec{\theta}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_K x_K$$

where $\vec{\theta} = (\theta_0, \ldots, \theta_k)^T$. The task is now to estimate $\vec{\theta}$. Linear as in parameters, not predictors (e.g. $\theta_1 x_1 x_2$ or $\theta_1 x_1^2$ are ok).

Homoskedasticity is when we assume $\mathrm{Var}(U_j \mid \vec{X} = x) = \sigma^2$ for all j. If the variance is different for each j, we say that the data exhibits *heteroskedasticity*.

Residual is the difference between the true value and its predicted value (here we write for the jth data point):

$$\hat{U}_j = Y_j - \hat{\vec{\theta}} \vec{x}_j.$$

which is computable with data (unlike the regression error). If the predictors are one-dimensional we have $\hat{U}_j = Y_j - \hat{\theta} x_j$.

Residual sum of squares (RSS) measures the quality of the regression line's fit to the data:

$$\mathrm{RSS}(\hat{\theta}) = \sum_{j=1}^{n} \hat{U}_j^2.$$

Predictive Regression Models

For continuous data, the joint density for the outcomes conditioned on the predictors is

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

Therefore, the MLE for θ in this setup would be given by the expression below:

$$\hat{\theta} = \mathrm{argmax}_\theta \sum_{j=1}^{n} \log p(y_j \mid X_j = x_j, \theta).$$

A Gaussian Example

Suppose, for this example, that we have one predictor and no intercept. For this Gaussian linear regression, suppose also that the noise is distributed as *independent* Normals; that is, we have $Y_j \mid X_j = x_j, \theta \sim \mathcal{N}(\theta x_j, \sigma^2)$. Then, we obtain the MLE

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

Check that $\hat{\theta}$ is conditionally unbiased and conditionally achieves the CRLB. Note also that under different assumptions (homo/hetero-skedasticity), even if we use the same MLE, the standard error of the estimator would be different.

Least Squares Regression

For the model $Y_j = \theta X_j + U_j$, the *least squares estimator* for θ is given by

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

For Gaussian linear regression, the MLE coincides with the least squares estimator, and in fact, the MoM estimator also coincides with the least squares estimator. In general, $\hat{\theta}_{LS} = \hat{\theta}_{MLE}$ for homscedastic Normal errors.

Logistic Regression

Logistic regression says that the probability of success is a logistic function of $(\theta_0 + \theta_1 x_1 + \cdots + \theta_K x_K)$, with $\theta = (\theta_0, \ldots, \theta_K)$. In other words, we have the model

$$P(Y = 1 \mid X = x, \theta) = \frac{\exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_K x_K)}{1 + \exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_K x_K)}.$$

Recall that the logit function is $\mathrm{logit}(r) = \log \frac{r}{1-r}$. Then, the logistic function is $\mathrm{logit}^{-1}(r) = \exp(r)/(1 + \exp(r))$. This is also known as the sigmoid curve.

Descriptive Regression

Descriptive regression is interested in the joint distribution of (X, Y), utilizing summaries such as $\mathrm{Cov}(X, Y)$. We define the following regression model:

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

We can interpret this as follows. Suppose that we are using $a + bX$ to mimic the behavior of Y. Then, if we set $\alpha = \mathrm{E}[Y] - \beta_{Y \sim X} \mathrm{E}[X]$, we actually discover that

$$(\alpha, \beta_{Y \sim X}) = \mathrm{argmin}_{(a,b)} \mathrm{E}[(Y - (a + bX))^2].$$

We report the goodness of fit of this regression with the R^2 statistic given by

$$R^2 = \mathrm{Cor}(X, Y)^2.$$

7 Hypothesis Testing

Consider a partition of the parameter space Θ into two disjoint sets Θ_0 and Θ_1 such that $\Theta = \Theta_0 \cup \Theta_1$. The null hypothesis is $H_0 : \theta \in \Theta_0$, the alternative hypothesis is $H_1 : \theta \in \Theta$.

One-sided test: $H_0 : \theta \leq \theta_0$ v.s. $\theta > \theta_0$ (or $\geq$ v.s. $<$)
Two-sided test: $H_0 : \theta = \theta_0$ v.s. $H_1 : \theta \neq \theta_0$.
Simple hypothesis: $\Theta_0 = \{\theta_0\}$
Composite hypothesis: Θ_0 is an interval or intervals.
Rejection region: A subset R of the data $\vec{y}$ such that we reject H_0 if $\vec{y} \in R$ and retain H_0 if $\vec{y} \notin R$. Note that rejection region is shaped entirely by the null.

Hypothesis Testing

- Find a **test statistic** $T(\vec{Y})$ and use the rejection region with the **critical values** below
- One-sided: $R = \{\vec{y} : T(\vec{y}) > c\}$
- Two-sided: $R = \{\vec{y} : T(\vec{y}) < c_L \text{ or } T(\vec{y}) > c_U\}$
- Test statistic should have a known distribution

Type I/II error

- **Type I error** or *false positive* means rejecting the null when the null is true. Formally, $\theta \in \Theta_0$ but $y \in R$. This is controlled by the α level of the test.
- **Type II error** or *false negative* means not rejecting the null when the null is false. Formally, $\theta \in \Theta_1$ but $y \notin R$.

Test Level and Power

Power function of a test (given R) is the probability of rejecting the null under a given value of θ, $\beta(\theta) = P(\vec{Y} \in R \mid \theta)$.

- If $\theta \in \Theta_0$, then $\beta(\theta) = P(\text{Type I error})$.
- If $\theta \in \Theta_1$, then $\beta(\theta) = P(\vec{Y} \in R \mid \theta \in \Theta_1) = 1 - P(\vec{Y} \notin R \mid \theta \in \Theta_1) = 1 - P(\text{Type II error})$

The **size** or **level** of a test is the maximum probability of Type I error occurring, $\alpha = \max\{\beta(\theta) : \theta \in \Theta_0\}$. This should usually be determined prior to looking at the data.

P-values

Given data $\vec{y}$, the **p-value** is the smallest α at which we can reject the null.

$$p(\vec{y}) = \inf\{\alpha : \vec{y} \in R_\alpha\}$$

where R_α is the rejection region for the test at the α level.

- The probability of obtaining data at least as extreme as the current data under the null.
- Given α, we say reject H_0 at the $\alpha-$ level if $p < \alpha$.
- (Universality of p-values) Consider two-sided test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ with a continuously distributed $T(\vec{Y})$. Then the p-value is uniform under the null, $p(\vec{y}) \sim \text{Unif}(0, 1)$.
- Large p-values could be (1) H_0 is true, or (2) the test has low power.
- E.g. for one-sided test $p(\vec{y}) = \inf\{\alpha : T(\vec{y}) > c\}$ where c is the critical value corresponding to R_α.

p-hacking: Testing many hypothesis increases the probability of observing a low p-value by chance.

Constructing Hypothesis Tests

To construct a hypothesis test, you would usually follow the following steps.

1. Figure out and clearly state your null and alternate hypotheses.
2. Find the test statistic $T(\vec{Y})$ and its distribution under the null $T(\vec{Y}) \mid (\theta = \theta_0)$.
3. Determine the rejection region by either finding critical values or p-values. Support or reject the null hypothesis based on what you find. e.g. for one-sided test find R (e.q. c) such that

$$P(\vec{Y} \in R \mid \theta = \theta_0) = P(T(\vec{y}) > c \mid \theta = \theta_0) \leq \alpha$$

Step (a) was covered before. For step (b), constructing $T(Y)$ can be tricky, but you can usually do this by finding some estimator for the parameter θ and building a pivot out of that. Also, you'd want $T(\vec{Y})$ to "differ" under H_0 vs. under H_1. Finally, if the sample size n is large, you can also use the asymptotic distribution of $T(Y)$ under the null. We now cover the most common types of hypothesis tests.

Z-test vs t-test

When σ is not known, we then need to consider the sample size. If the sample size is reasonably large, i.e. when $n \geq 30$, we can still appeal to the CLT and asymptotic tools to estimate σ and get the estimated z-statistic under H_0

$$T(\vec{Y}) = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\hat{\sigma}} \dot{\sim} \mathcal{N}(0, 1)$$

where $\hat{\theta}$ (e.g. $\bar{Y}$) is a consistent estimator to θ_0 (e.g. μ_0) and $\hat{\sigma}$ (e.g. sample standard deviation) is a consistent estimator to the standard deviation of $\sqrt{n}\hat{\theta}$. But when the sample size is small, we can no longer use the z-test. Instead, there is the t-statistic

$$T(\vec{Y}) = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\hat{\sigma}} \sim t_{n-1},$$

and with this statistic, we can instead perform what is known as the t-test. This requires

- $\hat{\theta} \sim \mathcal{N}(\theta_0, \sigma^2)$ under H_0
- $\hat{\sigma}^2 \sim \sigma^2 \chi^2(m)$
- $\hat{\theta} \perp\!\!\!\perp \hat{\sigma}^2$ under H_0

Likelihood-based/Asymptotic Hypothesis Tests

For large n, we can use test statistics whose distributions are only asymptotically valid. For the tests below, suppose that we are testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

Wald Test

Use the asymptotic pivot of the MLE $\hat{\theta}$ under the null to obtain

$$T(\vec{Y}) = \sqrt{n\mathcal{I}_1(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1).$$

We reject the null if

$$\left|\sqrt{n\mathcal{I}_1(\theta_0)}(\hat{\theta} - \theta_0)\right| > Q_{\mathcal{N}(0,1)}(1 - \alpha/2)$$

Score Test

Use the asymptotic pivot of the score $s(\vec{Y}, \theta_0)$ under the null and regularity conditions to get

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

We reject the null if

$$\left|\frac{s(\vec{Y}, \theta_0)}{\sqrt{n\mathcal{I}_1(\theta_0)}}\right| > Q_{\mathcal{N}(0,1)}(1 - \alpha/2)$$

Likelihood Ratio Test

Here allow the general $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$. We use the asymptotic pivot of the likelihood ratio under the null:

$$\Lambda(\vec{Y}) = 2 \log \frac{L(\hat{\theta}; \vec{Y})}{L(\theta_0; \vec{Y})} = 2(\ell(\hat{\theta}) - \ell(\theta_0))$$

where $\hat{\theta}$ is the MLE. Under regularity conditions

$$\Lambda(\vec{Y}) \xrightarrow{d} \chi^2_1.$$

We reject the null if

$$2 \log \frac{L(\hat{\theta}; \vec{Y})}{L(\theta_0; \vec{Y})} > Q_{\chi^2_1}(1 - \alpha)$$

This means we reject the null if the likelihood is higher under the MLE than under the null.

Hypothesis Tests and Confidence Intervals

Consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. The set

$$C_{\vec{Y},\alpha} = \{\theta : \vec{Y} \notin R_{\theta,\alpha}\}$$

is a $100(1 - \alpha)\%$ confidence interval for the unknown θ.

- Hypothesis tests correspond 1-1 with confidence intervals.
- If a CI contains θ_0 then we would retain the null in the corresponding hypothesis test.
- If we retain the null, the corresponding CI contains θ_0.

8 Bayesian Inference

In the Bayesian framework we treat θ as a r.v.. We have a model/likelihood along with a prior distribution $f(\theta)$. We want to find the posterior distribution of θ given the data $f(\theta \mid \vec{y})$

Bayes' Rule

$$f(\theta \mid \vec{y}) = \frac{f(\vec{y} \mid \theta) f(\theta)}{f(\vec{y})} \propto L(\theta; \vec{y}) f(\theta)$$

$f(\vec{y})$ is just a normalizing constant.

Posterior distribution The distribution of $\theta \mid \vec{y}$.

Posterior probability $P(\theta \in [a, b] \mid \vec{y}) = \int_a^b f(\theta \mid \vec{y})\, d\theta$

Posterior predictive distribution The distribution of $Y_{n+1} \mid (Y_1, \ldots, Y_n)$.

Point Estimators

Posterior mean

$$\hat{\theta}_{PM} = \mathbb{E}[\theta \mid \vec{y}] = \int \theta f(\theta \mid \vec{y}) d\theta.$$

This minimizes the average posterior squared loss $\mathbb{E}\left[(\theta - \hat{\theta})^2 \mid \vec{y}\right]$.

Biased assuming proper prior and finite variance.

Posterior Median

$$\hat{\theta}_M = Q_{\theta|y}(1/2)$$

This minimizes the average posterior absolute loss $\mathbb{E}[|\theta - \hat{\theta}| \mid \vec{y}]$, so another formulation is

$$\hat{\theta}_M = \text{argmin}_\theta\, \mathbb{E}[|\theta - \hat{\theta}| \mid \vec{y}]$$

Posterior mode (MAP)

$$\hat{\theta}_{MAP} = \text{argmax}_\theta\, f(\theta \mid \vec{y}).$$

To compute this we can maximize the log prior $\log f(\theta \mid \vec{y}) = \log L(\theta; \vec{y}) + \log f(\theta)$.

Interval Estimators

Credible intervals: Let $\alpha \in (0, 1)$. A $1 - \alpha$ credible interval or posterior probability interval for parameter is an interval estimate $[L(\vec{y}), U(\vec{y})]$ such that $P(L(\vec{y}) \leq \theta \leq U(\vec{y}) \mid \vec{y}) = 1 - \alpha$. Generally, find the credible interval using the quantile function $Q_{\theta|\vec{y}}$ of the posterior distribution:

$$[Q_{\theta|\vec{y}}(\alpha/2), Q_{\theta|\vec{y}}(1 - \alpha/2)]$$

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.

Conjugate Priors

Suppose we have $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \pi(\theta)$. If the posterior $\pi(\theta \mid \vec{Y} = \vec{y})$ is in the same family as $\pi(\theta)$, we call π a **conjugate prior** for the sampling distribution F.

Beta-Binomial

Suppose

$$p \sim \text{Beta}(a, b)$$
$$Y_i \mid p \sim \text{Bin}(n_i, p)$$

We have the posterior

$$p \mid (\vec{Y} = \vec{y}) \sim \text{Beta}(a + \sum_{i=1}^n y_i, b + \sum_{i=1}^n n_i - \sum_{i=1}^n y_i)$$

If we have only one data point this simplifies to

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

Poisson-Gamma

Suppose

$$\lambda \sim \text{Gamma}(a, b)$$
$$Y_i \mid \lambda \sim \text{Pois}(\lambda t_i)$$

We have the posterior

$$\lambda \mid (\vec{Y} = \vec{y}) \sim \text{Gamma}(a + \sum_{i=1}^n y_i, b + \sum_{i=1}^n t_i)$$

and marginally

$$Y_{n+1} \mid \vec{Y} \sim \text{NBin}\left(a + \sum_{i=1}^n y_i, \frac{b + \sum_{i=1}^n t_i}{b + t_{n+1} + \sum_{i=1}^n t_i}\right)$$

One data point simplifies the posterior to

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

and marginally

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

Normal-Normal
Suppose

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$
$$Y_i \mid \mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2) \qquad \forall i \in \{1, \ldots, n\}$$

where σ^2, μ_0 and τ_0^2 are known constants. We have the posterior

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

where

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

Letting $b_m = \tau_n^2/\tau_0^2$, we can also rewrite the posterior as

$$\mu \mid (\vec{Y} = \vec{y}) \sim \mathcal{N}((1-b_n)\bar{y} + b_n\mu_0, b_n\tau_0^2)$$

Marginally

$$Y_{n+1} \mid \vec{Y} \sim \mathcal{N}(\mu_n, \tau_n^2 + \sigma^2)$$

Generalizing to the Exponential Family Conjugate Priors.
Let $Y_1, \ldots, Y_n$ follow the NEF

$$f(y \mid \theta) = \exp(\theta y - \psi(\theta))h(y).$$

Assume $Y_1, \ldots, Y_n$ independent conditioned on θ, so the likelihood function is $L(\theta \mid y) = \exp(n(\theta\bar{y} - \psi(\theta)))$. Conjugate prior on θ is

$$\pi \propto \exp\left(r_0\theta\mu_0 - \psi(\theta)\right)$$

and the posterior mean of the mean parameter $\mu = \mathbb{E}\left[Y_1 \mid \theta\right] = \psi'(\theta)$ is the weighted average

$$\mathbb{E}[\mu \mid y] = (1-B)\bar{y} + B\mu_0$$

where $B = r_0/\left(r_0 + n\right)$.

Inference with Hierarchical Models

Usually use conjugacy. If there is no conjugacy, follow these steps.

1. Write down the joint density of all the unknown parameters and data:

$$p(y_1, \ldots, y_j, \theta_1, \ldots, \theta_j, \mu) = p(\mu)\prod_{i=1}^{j} p(y_i \mid \theta_i)p(\theta_i \mid \mu).$$

   This factorization follows the structure of conditional independence.

2. Use this joint distribution to get an expression for the conditional density you're interested in. So if we are interested in $\theta_1, \ldots, \theta_j, \mu \mid Y$,

$$p(\theta_1, \ldots, \theta_j, \mu \mid y) = \frac{p(y, \theta_1, \ldots, \theta_j, \mu)}{p(y)}.$$

   The denominator can be obtained by integrating out all of the θ_i's and μ.

Risk, Admissibility

Let $\hat{\theta}$ be an estimator for θ.
Loss function: A convex function $\text{Loss}(\theta, \hat{\theta}) \geq 0$ with the property $\text{Loss}(x, x) = 0$ for all x.
Risk function of $\hat{\theta}$ is $r_{\hat{\theta}}(\theta) = \text{E}(\text{Loss}(\theta, \hat{\theta}) \mid \theta)$.
An estimator $\hat{\theta}$ is **inadmissible** if there exists another estimator $\tilde{\theta}$ with $r_{\tilde{\theta}}(\theta) \leq r_{\hat{\theta}}(\theta)$ for all possible θ, with $r_{\tilde{\theta}}(\theta) < r_{\hat{\theta}}(\theta)$ for at least one possible value of θ. Otherwise it is admissable. "Admissible" intuitively means "not dominated in risk by any other estimator."

Stein Paradox with Normals

Suppose that we have

$$Y_i \mid \mu_i, \sigma^2 \sim \mathcal{N}(\mu_i, \sigma^2)$$

independently for $i = 1, \ldots, k$, $k \geq 3$, with σ^2 known and μ_i unknown. Let

$$\mu = (\mu_1, \ldots, \mu_k), \quad \hat{\mu} = (Y_1, \ldots, Y_k),$$

where $\hat{\mu}$ is meant to be an estimator for μ. Consider the squared error loss

$$\text{Loss}(\mu, \hat{\mu}) = \sum_{i=1}^{k}(\mu_i - \hat{\mu}_i)^2.$$

Then, $\hat{\mu}$ is inadmissible and is dominated by the James-Stein estimator given by

$$(\hat{\mu}_{JS})_i = \left(1 - \frac{(k-2)\sigma^2}{\sum_{i=1}^{k} Y_i^2}\right) Y_i.$$

The factor shows that the James-Stein estimator exhibits shrinkage towards zero.

9 Sampling

So far we used **model-based** inference, where randomness comes from the modeled distribution of data with infinite/super-population. Now we enter **design-based** inference by artificially introducing randomness from sampling/randomization, and with a finite population.

Survey Sampling

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

- Estimators all unbiased and obey CLT
- Sample average $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
  - Unbiased

$$\text{E}[\bar{Y}] = \frac{1}{n}\sum_{i=1}^{n}\text{E}[Y_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

  - Variance of sample mean $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$
- Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- ECDF

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$$

Sampling without Replacement
A SRS without replacement of size n from a total population of size N is a random sample chosen without replacement such that all $\binom{N}{n}$ possible samples are equally likely. We can think of the sample as being chosen all at once, or one at a time without replacement. This yields $P(Y_i = y_j) = 1/n$ for all i, j. Furthermore, observe that by symmetry, we obtain

$$\text{Cov}(Y_i, Y_j) = \frac{-\sigma^2}{N-1}.$$

- Sample average $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
  - Unbiased

$$\text{E}[\bar{Y}] = \frac{1}{n}\sum_{i=1}^{n}\text{E}[Y_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

  - Variance of sample mean $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$ where the factor is called the finite population correction, found using covariance.
- Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Stratified Sampling

Divide the population into L strata such that $\sum_{\ell=1}^{L} N_\ell = N$ where the N_ℓ are known, and let μ_ℓ and σ_ℓ^2 be the mean and variance of stratum ℓ. Take a sample of n_ℓ from stratum ℓ for each ℓ. Sampling can be carried out with replacement in each stratum, or without replacement in each stratum. For SRS within each stratum without replacement, we get the estimator for population mean

$$\hat{\mu} = \sum_{i=\ell}^{L}\frac{N_\ell}{N}\bar{Y}_\ell$$

This estimator is unbiased ($\text{E}[\hat{\mu}] = \mu$) and has the variance

$$\text{Var}(\hat{\mu}) = \sum_{\ell=1}^{L}\left[\left(\frac{N_\ell}{N}\right)^2 \cdot \frac{\sigma_\ell^2}{n_\ell} \cdot \frac{N_\ell - n_\ell}{N_\ell - 1}\right]$$

This variance is minimized when $n_\ell/n \propto N_\ell\sigma_\ell$, i.e. sample more from strata with larger size and variance.

Horvitz-Thompson estimator

The Horvitz-Thompson estimator is a very general way to construct an unbiased estimator for the population total $y_1 + y_2 + \cdots + y_N$, provided that for each individual we know the probability that the individual will be included in the sample. Theorem 10.2.15. Let

$$\pi_j = P(j \in S)$$

be the probability that individual j is included in a sample drawn from a set of distinct ID numbers, S. Assume that the π_j are known in advance, and that $\pi_j > 0$ for all j. Then the Horvitz-Thompson estimator

$$\hat{\tau}_{\text{HT}} = \sum_{j \in S}\frac{y_j}{\pi_j}$$

is an unbiased estimator for the population total

$$\tau = y_1 + y_2 + \cdots + y_N$$

If N is known, then

$$\hat{\mu}_{\text{HT}} = \frac{\hat{\tau}_{\text{HT}}}{N}$$

is an unbiased estimator for the population mean μ. However, it can have very bad variance, see Basu's elephants.

10 Resampling

Bootstrap Procedures

Here are the general steps you would need to conduct a non-parametric bootstrap.

1. Create a "bootstrapped" sample by randomly selecting observations from the original sample with replacement.
2. Calculate the statistic of interest, $\hat{\theta}$,(e.g. mean, median, standard deviation) for the bootstrapped sample. Repeat these 2 steps a large number of times, typically at least $B = 10^4$ times, to create a distribution of the statistic.
3. The mean of this distribution is an estimate of the population statistic, and the standard deviation can be used to create confidence intervals.
4. The standard error of $\hat{\theta}$ can be calculated from the sample standard deviation of the bootstrap estimates.

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

For a parametric bootstrap, instead of SRS with replacement, generate the bootstrap samples using the estimate of the parameter in a distribution. e.g. if the data is from Pois(λ) and we have an estimate $\hat{\lambda}$, generate the bootstrap samples from Pois($\hat{\lambda}$).

Bootstrap Confidence Interval

Assume n large for bootstrap to be accurate (involves LLN).

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

  $$\hat{\theta} \pm Q_{\mathcal{N}(0,1)}(1-\alpha/2)\cdot\widehat{\text{SE}(\hat{\theta})}$$

  more accurate if we have Normal asymptotics of $\frac{\hat{\theta}-\theta}{\widehat{\text{SE}(\hat{\theta})}}$
- *Percentile interval:* Construct an interval with the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the values $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$
- *Bootstrap t interval:* Simulate the following pivot to ascertain its distribution, calculate the quantiles, and then use the usual method for constructing a CI from a pivot:

  $$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\widehat{\text{SE}(\hat{\theta}^*)}} \text{ is an estimator of } T = \frac{\hat{\theta}-\theta}{\widehat{\text{SE}(\hat{\theta})}}$$

  Note that in the left-hand expression, the randomness comes from the resampling, which gives random values of $\hat{\theta}^*$. Since $\widehat{\text{SE}(\hat{\theta}^*)}$ is usually unknown, we can run an additional layer of bootstrapping to estimate it. The bootstrapped interval is then

  $$\left[\hat{\theta} - \hat{Q}^*(1-\alpha/2)\widehat{\text{SE}(\hat{\theta})}, \hat{\theta} - \hat{Q}^*(\alpha/2)\widehat{\text{SE}(\hat{\theta})}\right]$$

  where $\hat{Q}^*$ is the bootstrapped quantile of T^*. This has the best performance.

Bootstrap with Hypothesis Testing: Permutation Test

Suppose we have $X_1, \ldots, X_m \overset{\text{i.i.d}}{\sim} F_X$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} F_Y$, two independent samples. Also assume that $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ are independent. The CDFs F_X and F_Y are unknown, and no parametric assumptions are made about them. Consider the hypotheses:

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

Complete Permutation Test

1. Find a test statistic $T(\mathbf{X}, \mathbf{Y})$ such that large values of T are evidence against H_0 (e.g., $T(\mathbf{X}, \mathbf{Y}) = |\bar{Y} - \bar{X}|$)
2. Compute observed $t_0 = T(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})$ from data.
3. Compute T from each permutation of $(x_1, \ldots, x_m, y_1, \ldots, y_n)$ to get values $t_1, \ldots, t_{(m+n)!}$
4. P-value is the proportion of times that t was at least as extreme as t_0.

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

Remarks

- Since the total number of permutations is large, we can do *sampled permutation test* instead, where we randomly choose K permutations.
- Permutation test has the flexibility of choosing any test statistic, non-parametric, no asymptotics.
- Limitations: strong null, i.e. H_0 says that F_X and F_Y are the same distribution.

11 Causal Inference

Assignment: The assignment W_j is 1 if subject j is in the treatment group and 0 otherwise.

Potential outcomes: $Y_j(w_1, \ldots, w_n)$ is the outcome for patient Y_j if the assignments were $w_1, \ldots, w_n$.

Treatment effect: $\tau_j = Y_j(w_1, \ldots, w_n) - Y_j(w_1', \ldots, w_n')$ is the effect of moving from assignment to another.

Non-interference: The assignment of others has no effect on the potential outcomes of a particular subject:
$Y_j(w_1, \ldots, w_j, \ldots, w_n) = Y_j(w_1', \ldots, w_j, \ldots, w_n') = Y_j(w_j)$

Assignment mechanism: $P(\vec{W} = \vec{w} \mid Y\vec{(0)}, Y\vec{(1)})$ is the joint PMF of outcomes given the potential outcomes.

Switching equation: $Y = Y(W) = Y(1)W + Y(0)(1-W)$

This gives

$$WY = WY(1)$$

$$(1-W)Y = (1-W)Y(0)$$

Unconfoundedness: $\vec{W} \perp\!\!\!\perp (Y\vec{(0)}, Y\vec{(1)})$ given $\vec{X}$

Randomized Control Trials

We say that the assignments have been *randomized* if the assignments are independent of the potential outcomes, i.e.

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

which is equivalent to saying that the assignment mechanisms satisfies:

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

In observational studies this is not necessarily the case (so have to find all confounders to assume unconfoundedness). Note that $\vec{W} \perp\!\!\!\perp \{Y\vec{(0)}, Y\vec{(1)}\}$ does not mean $\vec{W} \perp\!\!\!\perp \vec{Y}$.

Population Based Modelling

The *population quantity*, $\text{E}(\tau_1)$ is the causal quantity for all units in a wider population beyond the sample. This is extrapolative: inference will take data from the n units and extrapolate to the entire population. The population quantity $\text{E}(\tau_j)$ is a causal quantity for all patients in a wider population beyond the sample.

Average treatment effect 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

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

Then, we can express the average treatment effect of the population as:

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

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

MLE estimator for $\text{E}(\tau_1)$

Note that under randomization assumption:

$$\theta_0 = P\left(Y_1 = 1 \mid W_1 = 0\right) = P\left(Y_1(0) = 1\right) = p_{10} + p_{11}$$

$$\theta_1 = P\left(Y_1 = 1 \mid W_1 = 1\right) = P\left(Y_1(1) = 1\right) = p_{01} + p_{11}$$

Hence we can estimate the population causal quantity via

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

The MLE estimator for θ_0 and θ_1 are shown to be

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

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:

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

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

$$\text{Var}(\widehat{\text{E}(\tau_1)}) = \frac{\theta_1(1-\theta_1)}{\sum_{i=1}^n w_i} - \frac{\theta_0(1-\theta_0)}{\sum_{i=1}^n(1-w_i)}$$

Finite Sample Modelling

The *finite sample*, or design-based, quantity $\bar{\tau}$ is specific to the units in the study, i.e., the average outcome if all the n units in the study are given the treatment minus the average outcome if all the n units in the study are given the control. We treat the $y_j(0)$ and $y_j(1)$ as fixed, and the randomness comes from W_j. Assume that the assignments are independent of the potential outcomes.

Average Treatment Effect The average treatment effect of a finite sample of size n is:

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

This is our usual estimand.

MoM estimator

Based on the above setup, the MoM estimator for $\bar{\tau}$ is:

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

Neyman Null and Fisher Null

- **Fisher null** $H_0 : \tau_j = 0$ for all j vs. $H_1 : \sum_{j=1}^n |\tau_j| > 0$. No treatment effect at all for any individual (i.e. $Y_j(1) = Y_j(0) = Y_j$) vs. at least one individual has treatment effect. We can use a permutation test for $\hat{\tau}_{MOM}$, which we now call a *randomization test*.
- **Neyman null** $H_0 : \bar{\tau} = 0$ vs. $H_1 : \bar{\tau} \neq 0$. Note that Neyman null allows individual causal effects to be non-zero, but they must balance out over the finite sample.

Randomized testing with Fisher Null

Use this to test finite sample treatment effect.

We use $T = |\hat{\tau}_{MoM}(w)|$ and reject the null if $T > Q(1-\alpha)$ where α is the pre-specified size of the test and Q is the quantile function of $\hat{\tau}_{MoM}$.

Then, we carry out the randomization test mechanistically similar to a permutation test we discussed early. We draw i.i.d. $\vec{W}^{(1)}, \cdots, \vec{W}^{(M)}$ M times and compute T for each of the draw. If p-value is needed, we define it to be the proportion similarly in the permutation test, which is

$$\text{p-value} = \frac{1}{M}\sum_{j=1}^{M} I(|\hat{\tau}_{MoM}(\vec{W}^{(j)})| \geq |\hat{\tau}_{MoM}(\vec{w})|)$$

12 Mathematical Tools

Taylor Approximation

First order Taylor expansion gives a linear approximation of a function g near some point x_0 as

$$g(x) \approx g(x_0) + \frac{\partial g(x_0)}{\partial x}(x - x_0).$$

For a fixed x_0, the Taylor expansion is linear in x. This approximation should be reasonably accurate when x is close to x_0.

Sum of Squares Identity

Let $Y_1, \cdots, Y_n$ be random variables. The *sample mean*, $\bar{Y}$, is the random variable

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

On the other hand, the *sample variance*, S^2, is the random variable given by

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2.$$

When $Y_1, \ldots, Y_n$ crystallize into the numbers $y_1, \ldots, y_n$, we can analogously define

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \quad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2.$$

Now, we obtain

$$\sum_{j=1}^{n}(Y_j - c)^2 = (n-1)S^2 + n(\bar{Y} - c)^2$$

for all $c \in \mathbb{R}$. This turns out to be a really important identity that appears all the time in statistics e.g. when deriving the posterior when the prior is Uniform on $(\mu, \log\sigma)$ and the data is Normal. Also,

$$\sum_{j=1}^{n}(Y_j - c)^2 = \sum_{j=1}^{n}(Y_j - \bar{Y})^2 + n(\bar{Y} - c)^2$$

Setting $c = 0$

$$\sum_{j=1}^{n}(Y_j - \bar{Y})^2 = \sum_{j=1}^{n}(Y_j^2) - n(\bar{Y})^2 = \sum_{j=1}^{n}(Y_j^2 - (\bar{Y})^2)$$

13 Stat 110 Concepts

Conditional Probability

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

Law of Total Probability (LOTP)

$$P(A) = \sum_{k=1}^{n} P(A \mid B_k)P(B_k)$$

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Marginal Distribution of R.V.s

$$f_X = \int f_{X,Y}\, dY$$

Expectation and Variance

Expectation

$$\mathrm{E}[X] = \int x f_X(x)\, dx$$

For i.i.d. X_i we have $\mathrm{E}(\bar{X}) = \mathrm{E}(X_i)$

Linearity

$$\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c$$

Conditional Expectation

$$\mathrm{E}[X \mid A] = \int x f_X(x \mid A)\, dx$$

LOTUS

$$\mathrm{E}[g(X)] = \int g(x) f_X(x)\, dx$$

Variance

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}(X^2) - (E[X])^2$$

Covariance

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$

Correlation

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

Properties

- $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$
- For $X \perp\!\!\!\perp Y$, $\mathrm{Var}(X+Y) = \mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(X+a, Y+b) = \mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(W+X, Y+Z) =$ $\mathrm{Cov}(W,Y) + \mathrm{Cov}(W,Z) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$
- $\mathrm{Corr}(aX+b, cY+d) = \mathrm{Cov}(X,Y)$
- $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i) + 2\sum_{i<j}\mathrm{Cov}(X_i, X_j)$
- If X_i identically distributed, then $\mathrm{Var}(\sum_i X_i) = n\,\mathrm{Var}(X_1) + 2\binom{n}{2}\mathrm{Cov}(X_1, X_2)$
- If X_i uncorrelated (or more strongly, independent), then $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i)$
- If X_i uncorrelated (or more strongly, independent) and have the same variance, then $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X_i)/n$

Adam's Law (LOTE)

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}(Y \mid A)] = \sum_{k=1}^{n}\mathrm{E}(Y \mid A_k)P(A_k)$$

Eve's Law (LOTV)

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathrm{E}[Y \mid X])$$

Jensen's Inequality

$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$ for convex g. $\leq$ for concave g.

Poisson Processes

For a Poisson process of rate λ arrivals per unit time:

- The number of arrivals in a time interval of length t is $\mathrm{Pois}(\lambda t)$.
- Number of arrivals in disjoint time intervals are independent.
- Inter-arrival times are i.i.d. $\mathrm{Expo}(\lambda)$.
- CDF is $P(X \leq x) = 1 - e^{-\lambda x}$.

Convolutions of Random Variables

A convolution of n r.v.s is simply their sum. Let X and Y be independent.

- $X \sim \mathrm{Pois}(\lambda_1), Y \sim \mathrm{Pois}(\lambda_2) \implies X + Y \sim \mathrm{Pois}(\lambda_1 + \lambda_2)$
- $\mathrm{Bin}(n,p)$ can be thought of as a sum of n i.i.d. $\mathrm{Bern}(p)$ r.v.s.
- $X \sim \mathrm{Bin}(n_1,p), Y \sim \mathrm{Bin}(n_2,p) \implies X + Y \sim \mathrm{Bin}(n_1+n_2,p)$
- $\mathrm{Gamma}(n,\lambda)$ with integer n can be thought of as a sum of i.i.d. $\mathrm{Expo}(\lambda)$ r.v.s.
- $X \sim \mathrm{Gamma}(a_1,\lambda), Y \sim \mathrm{Gamma}(a_2,\lambda) \implies X + Y \sim \mathrm{Gamma}(a_1+a_2,\lambda)$
- $\mathrm{NBin}(r,p)$ can be thought of as a sum of r i.i.d. $\mathrm{Geom}(p)$ r.v.s.
- $X \sim \mathrm{NBin}(r_1,p), Y \sim \mathrm{NBin}(r_2,p) \implies X + Y \sim \mathrm{NBin}(r_1+r+2,p)$
- $X \sim \mathcal{N}(\mu_1,\sigma_1^2), Y \sim \mathcal{N}(\mu_2,\sigma_2^2) \implies X + Y \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

Transformations of Random Variables

- $Y \sim \mathrm{Expo}(\lambda) \implies \lambda Y \sim \mathrm{Expo}(1) \implies kY \sim \mathrm{Expo}(\lambda/k)$
- $X \sim \mathcal{N}(\mu,\sigma^2) \implies \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

Special Cases of Distributions

- $\mathrm{Bin}(1,p)$ is the same distribution as $\mathrm{Bern}(p)$
- $\mathrm{Beta}(1,1)$ is the same distribution as $\mathrm{Unif}(0,1)$
- $\mathrm{Gamma}(1,\lambda)$ is the same distribution as $\mathrm{Expo}(\lambda)$
- $\mathrm{NBin}(1,p)$ is the same distribution as $\mathrm{Geom}(p)$
- χ_n^2 is the sum of squares of n i.i.d. $\mathcal{N}(0,1)$
- χ_n^2 is the same distribution as $\mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$
- $X_i \overset{\text{ind}}{\sim} \mathrm{Expo}(\lambda_i) \implies \min(X_1,\ldots,X_k) \sim \mathrm{Expo}(\lambda_1 + \ldots + \lambda_k)$
- $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Expo}(\lambda) \implies \max(X_1,\ldots,X_k) \sim Y_1 + \cdots + Y_k$ where $Y_j \sim \mathrm{Expo}(j\lambda)$ and the Y_j are independent.
- For $X \sim \mathrm{Expo}(\lambda)$, $X^{1/\gamma} \sim \mathrm{Weibull}(\lambda,\gamma)$
- For $X \sim \mathcal{N}(\mu,\sigma^2)$, $e^X \sim \text{Log-Normal}(\mu,\sigma^2)$
- Let $X \sim \mathrm{Bin}(n,p), Y \sim \mathrm{Bin}(m,p)$ with $X \perp\!\!\!\perp Y$. Then $X \mid (X+Y=r) \sim \mathrm{HGeom}(n,m,r)$
- Let $X \sim \mathrm{Pois}(\lambda_1), Y \sim \mathrm{Pois}(\lambda_2)$ with $X \perp\!\!\!\perp Y$. Then $X \mid (X+Y=n) \sim \mathrm{Bin}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$
- **Chicken-egg:** If there are $Z \sim \mathrm{Pois}(\lambda)$ items and we randomly and independently accept each item with probability p, then the number of accepted items $Z_1 \sim \mathrm{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \mathrm{Pois}(\lambda(1-p))$, and $Z_1 \perp\!\!\!\perp Z_2$.
- **Bank-Post Office:** If $X \sim \mathrm{Gamma}(a,\lambda)$, $Y \sim \mathrm{Gamma}(b,\lambda)$ and $X \perp\!\!\!\perp Y$, then $\frac{X}{X+Y} \sim \mathrm{Beta}(a,b)$ and $X+Y \perp\!\!\!\perp \frac{X}{X+Y}$
- **Beta-Binomial Conjugacy:** For $X \mid p \sim \mathrm{Bin}(n,p)$ and $p \sim \mathrm{Beta}(a,b)$, the posterior $p \mid (X=x) \sim \mathrm{Beta}(a+x, b+n-x)$.
- **Binomial-Poisson:** $\mathrm{Bin}(n,p)$ is approximately $\mathrm{Pois}(np)$ if p is small.
- **Binomial-Normal:** $\mathrm{Bin}(n,p)$ is approximately $\mathcal{N}(np, np(1-p))$ if n is large and p is not near 0 or 1.

14 Important Examples

MLE and MoM for Normal Model

Normal with known variance Let $Y_1, \cdots, Y_n$ be iid $N(\mu, \sigma^2)$ with $\theta = \mu$ unknown but σ^2 is known. The likelihood function, dropping normalizing constant is

$$L(\mu; y) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\}$$

and the log-likelihood is

$$\ell(\mu; \mathbf{y}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_j - \mu)^2 = -\frac{1}{2\sigma^2}\left\{\sum_{j=1}^{n}(y_j - \bar{y})^2 + n(\bar{y} - \mu)^2\right\}$$

It is easy to maximize $\ell(\mu; \mathbf{y})$, just set $\mu = \bar{y}$, and we observe that

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and so $\hat{\mu}$ is unbias with standard error

$$\text{SE}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

Normal with both parameters unknown Let $Y_1, \cdots, Y_n$ be iid $N(\mu, \sigma^2)$ with both parameters unknown. We will parameterize the model in terms of the mean and standard deviation, $\theta = (\mu, \sigma)$ instead of (μ, σ^2). Then, we observe that

$$L(\mu, \sigma; \mathbf{y}) = \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\}$$

and that the log likeligood is

$$\ell(\mu, \sigma; \mathbf{y}) = -\frac{1}{2\sigma^2}\left\{\sum_{j=1}^{n}(y_j - \bar{y})^2 + n(\bar{y} - \mu)^2\right\} - n\log\sigma$$

By multivariate calculus derivation (which we will skip here), we have the MLE as

$$\hat{\mu} = \bar{Y}, \hat{\sigma} = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

Sufficient Statistic and MLE in an NEF

The PMF/PDF of an NEF can be written as $f_\theta(y) = e^{\theta y - \psi(\theta)}h(y)$, so the joint log-likelihood is:

$$L(\theta) = e^{\theta\sum Y_j - n\psi(\theta)}$$

$$\ell(\theta) = \theta\sum_{j=1}^{n} Y_j - n\psi(\theta)$$

$$s(\theta) = \sum_{j=1}^{n} Y_j - n\psi'(\theta) = 0$$

$$\frac{1}{n}\sum_{j=1}^{n} = \psi'(\theta) = E(Y)$$

$$\hat{\mu}_{MLE} = \bar{Y}$$

So, $\bar{Y}$ is a sufficient statistic.

Censored data

Suppose there are $n = 30$ devices. They are observed for 7 months, at which point 21 have failed while 9 still work. Assume each device's lifetime $Y_j \sim_{iid} Expo(\lambda)$ and the estimand is $\mu = 1/\lambda$.
For each observation:

$$L_j(\lambda) = \begin{cases} f(y) \text{ if observed} \\ 1 - F(7) \text{ if not observed} \end{cases}$$

$$L(\lambda) = \left(\prod_{j=1}^{21}\lambda e^{-\lambda t_j}\right)\left(e^{-7\lambda}\right)^9 = \lambda^{21}e^{-21\lambda\bar{t}}e^{-63\lambda}$$

$$\ell(\lambda) = 21\log(\lambda) - 21\lambda\bar{t} - 63\lambda$$

$$s(\lambda) = \frac{21}{\lambda} - 21\bar{t} - 63 = 0$$

$$\hat{\lambda}_{MLE} = \frac{1}{\bar{t} + 3}$$

$$\hat{\mu}_{MLE} = \bar{t} + 3, \text{ by invariance}$$

German Tank Problem

Suppose n tanks are captured, with serial numbers $Y_1, Y_2, \ldots Y_n$. Assume the population serial numbers are $1, 2, \ldots t$ and that the data is a simple random sample. Estimate the total number of tanks t.
$L(t) = \frac{1}{\binom{t}{n}}$ if $Y_1, Y_2, \ldots Y_n \in \{1, 2, \ldots t\}$ and 0 otherwise

$$= \frac{\text{Ind}\left(Y_{(n)} \le t\right)}{\binom{t}{n}}$$

The likelihood of t is 0 for $t < Y_{(n)}$ because we would have already observed a tank with a higher serial number. However, the likelihood function is decreasing, so the maximum likelihood estimate must be $\hat{t}_{MLE} = Y_{(n)}$. However, this estimator is biased. The PMF for $Y_{(n)}$ is the number of ways to choose $n-1$ tanks with serial numbers less than $Y_{(n)}$ divided by the total number of ways to choose n tanks from t.

$$P\left(Y_{(n)} = m\right) = \frac{\binom{m-1}{n-1}}{\binom{t}{n}}$$

$$\text{E}\left(Y_{(n)}\right) = \sum_{m=n}^{t} m\binom{m-1}{n-1} = \frac{n}{n+1}(t+1)$$

So, we can correct our estimator to $\frac{n+1}{n}Y_{(n)} - 1$, which is unbiased.

Sample Mean vs. Sample Median

Let $Y_1, Y_2, \ldots Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta, \sigma^2\right)$; estimand θ Sample mean: $\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ Sample median: $M_n \sim \mathcal{N}\left(\theta, \frac{\pi}{2}\frac{\sigma^2}{n}\right)$ by asymptotic distribution of sample quantiles
Sample mean is a more efficient estimator as it has a lower variance, but in cases when the assumption of Normal is wrong (e.g. Cauchy), sample median may be more robust.

Poisson Method of Moments - 2 Ways

Let $Y_1, Y_2, \ldots Y_n \sim \text{Pois}(\theta)$. via Mean: 1. $\theta = \text{E}(Y)$ 2. $\hat{\theta}_{MoM} = \bar{Y}$ via Variance: 1. $\theta = \text{Var}(Y) = \text{E}\left(Y^2\right) - (\text{E}(Y))^2$ 2.
$\hat{\theta}_{MoM} = \frac{1}{2}\sum Y_j^2 - \bar{Y}^2 = \frac{1}{n}\sum\left(Y_j - \bar{Y}\right)^2$

Variance-Stabilizing of Poisson

Let $T \sim Pois(\lambda) \approx N(\lambda, \lambda)$ for large λ. What is the approximate distribution of $\sqrt{T}$?

$$T \to_d N\left(\lambda, \frac{\lambda}{n}\right), \text{ by CLT}$$

$$\sqrt{T} \to_d N\left(\sqrt{\lambda}, \frac{1}{4}\right), \text{ by Delta Method}$$

Pivot based on Student-t distribution

Let the data be i.i.d $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$ with both parameters unknown. Suppose that we want a $1 - \alpha$ CI for μ. Since σ is unknown, we can replace σ by the standard deviation $\hat{\sigma}$, but then we can only have an approximate CI. Instead, let us construct a pivot, the *t-statistics*

$$T = \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{\hat{\sigma}}$$

Basu's elephant and Horowitz-Thompson

In Lecture 20, we consider that we have $y_i, y_2, \cdots, y_N$, where each y_i represents the volume of a tree in the forest, and we wish to estimate $\mu = \frac{1}{N}\sum_{j=1}^{N} y_j$. Then

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{j=1}^{N}\frac{I\{j \in S\}y_j}{\pi_j}$$

Recall that by construction, $\hat{\mu}_{HT}$ will always be unbiased. However, for small π_i, some of the weights, $\frac{1}{\pi_i}$ could be crazy. This is one of the downsides of HT estimator and one of the most famous problem thinking and explaining this concept is Basu's Elephant by D. Basu (1971).

James-Stein estimator for batting averages

From Homework 9 Problem 1 A sabermetrician wants to estimate the batting averages of $k > 3$ baseball players, based on data from early in the season. Let μ_j be the theoretical batting average of player j (i.e., the number of hits divided by number of times at bat that would result from a hypothetical very large number of times at bat). Let Y_j be the proportion of hits that player j gets out of n times at bat (i.e., the number of hits divided by n). It would be natural to model the number of hits as Binomial, but for simplicity and to connect with material discussed in class, we will use a Normal approximation to the Binomial. Assume the following model:

$$Y_j|\mu_j \sim N(\mu_j, \sigma^2), \text{ for } j = 1, 2, \ldots, k,$$

with $Y_1, \ldots, Y_k$ conditionally independent given $\mu_1, \ldots, \mu_k$. A priori, let the μ_j be i.i.d. with

$$\mu_j \sim N(\mu_0, \tau_0^2).$$

Assume that the hyperparameters μ_0 and τ_0^2 are unknown, though σ^2 is still known. In class we discussed the James-Stein estimator that shrinks the MLE toward 0. If we know μ_0, it would make more sense to shrink toward μ_0 rather than toward 0. Since the marginal distribution of Y_i is

$$Y_j \sim N(\mu_0, \sigma^2 + \tau_0^2),$$

we will estimate μ_0 with $\bar{Y}$ and shrink the MLE toward $\bar{Y}$. Let

$$S = \sum_{i=1}^{k}(Y_i - \bar{Y})^2.$$

At homework 9 Q1(e), we have shown that

$$\hat{b} = \frac{(k-3)\sigma^2}{S}$$

is an unbiased estimator for b. The James-Stein estimator $\hat{\mu}_{\text{JS}}$ is then obtained from $\hat{\mu}_{\text{Bayes}}$ by replacing μ_0 by $\bar{Y}$ and b by $\hat{b}$:

$$\hat{\mu}_{j,\text{JS}} = \hat{b}\bar{Y} + (1 - \hat{b})Y_j.$$