STAT 111 Final Cheatsheet

Compiled by Zad Chin. Material based on Professor Joe Blitzstein's lectures, Jamie Liu's midterm cheatsheet, and Joseph K. Blitzstein and Neil Shephard Lecture Notes for STAT 111.

Models

Key terms

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

$$Y = (Y_1, \dots, Y_n),$$

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

$$y = (y_1, \dots, y_n).$$

A *statistic* is a function T of the random vector 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: $\bar{y}$.

Sufficient Statistics

Let $Y = (Y_1, \dots, Y_n)$ be a sample from some model. A statistic $T(Y)$ is *sufficient* for θ if the conditional distribution of $Y \mid T$ does not depend on θ.

Factorization Criterion

The statistic $T(Y)$ is sufficient if and only if we can factor the joint density of Y as $p(y \mid \theta) = g(T(y), \theta)h(y)$.

Sufficiency of Order Statistics

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

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

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 (the logarithm of the MGF of Y).

Properties of NEFs

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

- $\mathbb{E}Y = \Psi'(\theta)$, $\mathbb{V}Y = \Psi''(\theta)$, and the MGF $M_Y(t) = e^{\Psi(\theta+t)-\Psi(\theta)}$.
- If $Y_1, \dots, Y_n \sim Y$ are i.i.d., $\bar{Y}$ is a sufficient statistic for θ.
- The MLE of $\mu = \mathbb{E}Y$ is $\hat{\mu} = \bar{Y}$.
- The Fisher information is $\mathcal{I}_Y(\theta) = \Psi''(\theta)$.
- We call the process where we fix $h(y)$ as a baseline distribution from which we construct the entire NEF *exponential tilting*.
- Some examples of NEFs include the Normal (with σ^2 known), the Poisson, the Binomial (with n fixed), the Negative Binomial (with r fixed), and last, but not least, the Gamma (with a known).

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).

Likelihood

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 y:

$$L(\theta) = p(y \mid \theta).$$

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

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

Log Likelihoods

Loglikelihood ℓ is the logarithm of the likelihood:

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

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

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

Finding the MLE

The steps you can carry out to find the MLE of θ given the data y:

1. Find $L(\theta)$ and take logs to find $\ell(\theta)$.
2. Find $\ell'(\theta)$, set it to zero, and solve for θ (call this $\hat{\theta}$).
3. Find $\ell''(\theta)$ and check that $\ell''(\hat{\theta}) < 0$.
4. With this, $\hat{\theta}$ is your maximum likelihood estimate!
5. To find the maximum likelihood estimator, convert the y_j's into Y_j's.

Reparameterization

Likelihood Invariance Allow θ to be an estimand of interest, and let $\psi = g(\theta)$, where g is injective. Then, $L(\psi) = L(\theta)$.

MLE Invariance If g is injective and $\psi = g(\theta)$, the MLE of ψ is equal to the MLE of θ evaluated at g. After all, maximizing $L(\psi)$ is equivalent to maximizing $L(\theta)$ because applying g is an inequality preserving operation.

Score Function

The *score* function is defined to be the first derivative of the loglikelihood:

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

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

Information Inequality

Let $Y = (Y_1, \dots, Y_n)$ be a random vector of i.i.d. random variables, and suppose that the following regularity conditions hold:

- The support of Y does not depend on θ.
- All expectations and derivatives exist.

Then, both equalities below hold; the latter is known as the information equality.

$$\mathbb{E}s(\theta) = 0, \quad \mathbb{V}s(\theta) = -\mathbb{E}s'(\theta).$$

Fisher Information

The *Fisher information* for a parameter θ is defined as $\mathcal{I}_Y(\theta) = \mathbb{V}s(\theta)$

Remarks:

- Letting $\mathcal{J}_Y(\theta)$ denote $-\mathbb{E}s'(\theta)$, we have $\mathcal{I}_Y(\theta) = \mathcal{J}_Y(\theta)$ (the information equality!) if the regularity conditions mentioned earlier hold.
- You might sometimes see θ^* used instead of θ to really emphasize the fact that the entry to $\mathcal{I}_Y$ is the "true value" of θ.
- Be wary not to confuse $\mathcal{I}_Y(\theta)$ and $\mathcal{I}_{Y_1}(\theta)$. The former is the Fisher information with respect to the entire data vector $Y = (Y_1, \dots, Y_n)$, while the latter is the Fisher information with respect to the single observation Y_1.
- In fact, if our random variables $Y_1, \dots, Y_n$ are i.i.d., we have $\mathcal{I}_Y(\theta) = n\mathcal{I}_{Y_1}(\theta)$.

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

Asymptotics

Convergence Equivalence

Fix some constant c, and allow $Y_1, \dots, Y_n$ to be random variables. Then, convergence in probability implies convergence in distribution.

Central Limit Theorem

Let $Y_1, \dots, Y_n$ be i.i.d random variables with mean $\mathbb{E}Y_1 = \mu$ and finite variance $\mathbb{V}Y_1 = \sigma^2 < \infty$. Then, the convergence

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

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

Law of Large Numbers

Let $Y_1, \dots, Y_n$ be i.i.d. random variables with finite first moments, i.e. $\mathbb{E}|Y_1| < \infty$. Then, $\bar{Y} \to_{\mathcal{A}} \mathbb{E}Y_1$ and $\bar{Y} \to_{\mathcal{P}} \mathbb{E}Y_1$.

Continuous Mapping Theorem

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

1. $Y_n \to_{\mathcal{A}} Y$ implies $g(Y_n) \to_{\mathcal{A}} g(Y)$.
2. $Y_n \to_{\mathcal{P}} Y$ implies $g(Y_n) \to_{\mathcal{P}} g(Y)$.
3. $Y_n \to_{\mathcal{D}} Y$ implies $g(Y_n) \to_{\mathcal{D}} g(Y)$.

Slutsky's Theorem

Suppose $X_n \to_{\mathcal{D}} X$ and $Y_n \to_{\mathcal{D}} c$, where c is a constant (recall that the latter condition is equivalent to $Y_n \to_{\mathcal{P}} c$). Then,

1. $X_n + Y_n \to_{\mathcal{D}} X + c$.
2. $X_n Y_n \to_{\mathcal{D}} cX$.
3. For $c \neq 0$, $X_n / Y_n \to_{\mathcal{D}} X/c$.

Delta Methods

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

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

Point Estimation

Methods of Moments

Finding MoM

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables. Recall that we can freely use the notation

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

In a similar manner, for any kth moment, we can, without rederivation, notate

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

Then, the *method of moments* (MoM) estimator for some parameter θ is found by:

1. Replacing each component of θ with a corresponding component of $\hat{\theta}$.
2. Replacing the first $\dim\theta$ moments $\mathbb{E}Y_1^k$ from the model with $\bar{Y}^k$.

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

Properties of MoM

If $\mathbb{V}Y_1^k < \infty$, $\mathbb{E}\bar{Y}^k = \mathbb{E}Y_1^k$ and $\mathbb{V}\bar{Y}^k = \mathbb{V}(Y_1^k)/n$. Moreover, by the CLT, we obtain

$$\sqrt{n}(\bar{Y}^k - \mathbb{E}Y_1^k) \to_{\mathcal{D}} \mathcal{N}(0, \mathbb{V}Y_1^k).$$

Note that, in general, $\mathbb{E}\hat{\theta} \neq \theta$ if $\hat{\theta}$ is an MoM estimator, even though $\mathbb{E}\bar{Y}^k = \mathbb{E}Y_1^k$

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 $Y_{(1)}, \ldots, Y_{(n)}$, when the rth *sample quantile* is $Y_{\lceil nr \rceil}$. This is a nonparametric estimator for $Q(r)$.

Consistency of Estimators

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

$$\hat{\theta} \to_{\mathcal{P}} \theta$$

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. Recognize that $\theta = \mathbb{E}Y_1$ and $\hat{\theta} = \bar{Y}$ for some i.i.d. $Y_1, \ldots, Y_n$, when $\hat{\theta} \to_{\mathcal{P}} \theta$ follows immediately from the WLLN.
3. Recognize that $\theta = g(\mathbb{E}Y_1)$ and $\hat{\theta} = g(\bar{Y})$ for some continuous function g and i.i.d. $Y_1, \ldots, Y_n$, when $\hat{\theta} \to_{\mathcal{P}} \theta$ follows from the WLLN and CMT.
4. Fix some $\epsilon > 0$, and show that $\mathbb{P}(|\hat{\theta} - \theta| > \epsilon) \to 0$ directly.

Bias, Standard Error and Loss

- The **bias** of an estimator $\hat{\theta}$ for an estimand θ is $\text{Bias}(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta$. .
- The **standard error** of an estimator $\hat{\theta}$ is $\text{SE}(\hat{\theta}) = (\mathbb{V}\hat{\theta})^{1/2}$. 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: $\text{Loss}(\theta, x) = \mathbb{I}(\theta \neq x)$.
  - 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):
  $$\text{MSE}(\theta, \hat{\theta}) = \mathbb{E}(\theta - \hat{\theta})^2.$$
- **Bias-Variance Decomposition:**The MSE can be decomposed as
  $$\text{MSE}(\theta, \hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \mathbb{V}\hat{\theta}.$$

Cramer Rao Lowe Bound

CRLB for Unbiased Estimators Under the regularity conditions mentioned at the start, if $\hat{\theta}$ is unbiased for θ,

$$\mathbb{V}\hat{\theta} \geq \frac{1}{\mathcal{I}_Y(\theta)}.$$

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

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

where $g(\theta) = \mathbb{E}\hat{\theta}$. In the zero bias case, it was the case that $g(\theta) = \theta$, so $g'(\theta) = 1$.

Rao-Blackwellization

Let T be a sufficient statistic and $\hat{\theta}$ be any estimator (in both cases with respect to the estimand θ). Then, setting

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

it is the case that the inequality of MSEs given by $\text{MSE}(\hat{\theta}_{RB}, \theta) \leq \text{MSE}(\hat{\theta}, \theta)$ holds.

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 as we've seen above.
- To find the Rao-Blackwell estimator, you usually need to determine conditional distributions of the form $Y \mid T$. In Statistics 111, this is usually done by citing a Statistics 110 story, so do make a relevant list of those!

Interval Estimation

Confidence Interval An interval estimator with a coverage probability at least $1 - \alpha$ for all possible values of θ is called a $100(1 - \alpha)\%$ *confidence interval* (CI). We say that $1 - \alpha$ is the *level* of our CI, and that $(U(Y) - L(Y))/2$ is the *margin of error* of our CI.

Exact CIs

A *pivot* is a random variable that is free of any parameters (e.g. $\mathcal{N}(0, 1)$, $\text{Unif}(6, 9)$). Suppose we want to build a 95% CI for θ, where we observe $Y \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, say equal to 4, the strategy is as follow:

1. Do some algebraic manipulation to get a pivot. In our example, we get
   $$\frac{Y - \theta}{\sigma} \sim \mathcal{N}(0, 1).$$
2. Find the values of the quantile function of the pivot evaluated at 0.025 and 0.975. In our example, $Q_{\mathcal{N}(0,1)}(0.025) \approx -1.96$ and $Q_{\mathcal{N}(0,1)}(0.975) \approx 1.96$.
3. With these values, rest assured that a 95% CI will surface from the inequality you get when you set the 0.025 quantile value as the lower bound and the 0.975 quantile as the upper bound. In our example, a 95% CI starts from
   $$Q_{\mathcal{N}(0,1)}(0.025) \leq \frac{y - \theta}{\sigma} \leq Q_{\mathcal{N}(0,1)}(0.975),$$
   where y is the observed value (in an experiment) of the random variable Y.
4. Rearrange this inequality to get just the parameter of interest in the middle. In our example, some rearrangement yields the following inequality:
   $$y - Q_{\mathcal{N}(0,1)}(0.975)\sigma \leq \theta \leq y - Q_{\mathcal{N}(0,1)}(0.025)\sigma.$$
5. Plug in all the values you know and assert that a 95% CI has been found! In our example, suppose we observe $y = 1$. Plugging in known values,
   $$1 - (1.96)(2) \leq \theta \leq 1 + (1.96)(2),$$
   so we can conclude that a 95% CI for θ is given by $[-2.92, 4.92]$.

Asymptotic CIs

Sometimes, finding a suitable pivot is hard. In these cases, we appeal to the CLT, which *always* allows us 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. With this, we can construct a CI as we did before, with the caveat that the resulting interval is not exact for finite n.

Bootstrap Confidence Interval

- *Normal approximation*: construct an $(1 - \alpha) \cdot 100\%$ interval with endpoints
  $$\theta \pm \text{ qnorm } (1 - \alpha/2) \cdot \hat{SE}(\hat{\theta})$$
  more accurate if we have Normal asymptotics of $\frac{\hat{\theta} - \theta}{SE(\theta)}$
- *Percentile interval:* Construct an interval with the empirical quantiles of the values $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$
- *Bootstrap t:* Simulate the following pivot to ascertain its distribution, calculate the quantiles, and then use the usual method for constructing a CI from a pivot:
  $$\frac{\hat{\theta}^* - \hat{\theta}}{\text{SE}(\hat{\theta}^*)} \text{ is an estimator of } \frac{\hat{\theta} - \theta}{\text{SE}(\hat{\theta})}$$
  Note that in the left-hand expression, the randomness comes from the resampling, which gives random values of $\hat{\theta}^*$

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. $Y_1, \dots, Y_n \sim [\theta, \sigma^2]$, with σ^2 known, we have

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

by the CLT. Then, an asymptotic 95% CI for θ (which you can just quote) 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 $\bar{Y}$.

Regression

Linear Regression

When the regression function is a linear function of the parameters, i.e. when

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

we enter the realm of *linear regression*
We often assume $\mathbb{V}(U_j \mid X = x) = \sigma^2$ for all j. This simplifying assumption is commonly known as *homoskedasticity*, and in its absence (the variance is different for each j), we say that the data exhibits *heteroskedasticity.*
The difference between the true value and its predicted value yields the *residual*

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

The *residual sum of squares* (RSS), measures the quality of the regression line's fit to the data. It is given by the following sum: $\mathrm{RSS}(\hat{\theta}) = \sum_{j=1}^n \hat{U}_j^2$.

Predictive Regression Models

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

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

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

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

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.

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:

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

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

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

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

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

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 $\theta = (\theta_0, \dots, \theta_K)$. In other words, we have the model

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

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

Hypothesis Testing

Here, we assume you are familiar with null/alternate hypothesis, one sided vs two sided test, as well as rejection region, and p-value;

Type I/II error

There are two types of errors that could happen in hypothesis testing.

- A Type I error is typically known as a *false positive.* Formally, such an error occurs when $\theta \in \Theta_0$ but $y \in R$, i.e. we reject when it is true.
- A Type II error is typically known as a false negative. Formally, this error occurs when $\theta \in \Theta_1$ but $y \notin R$, i.e. we fail to reject the null when it is false.

The probability of each error can be calculated as follows.

- Type I error is calculated by $\max_{\theta \in \Theta_0} \mathbb{P}(Y \in R \mid \theta)$.
- Type II error is calculated by $\mathbb{P}(Y \notin R \mid \theta = \theta_1)$ for some $\theta_1 \in \Theta_1$.

Test Level and Power

The *power function* of a test is defined as $\beta(\theta) = \mathbb{P}(Y \in R \mid \theta)$. The *Type I error*, also called the *level* or *size* of a test, is $\alpha = \max_{\theta \in \Theta_0} \beta(\theta)$. This should usually be determined prior to looking at the data. For a given $\theta \in \Theta_1$, the *power* of a test is $\beta(\theta) = 1 - \mathbb{P}(Y \notin R \mid \theta)$, while for $\theta \in \Theta_0$, the power is the Type I error.
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 $T(Y)$ and its distribution under the null $T(Y) \mid \theta = \theta_0$.
3. Determine the rejection region by either finding critical values or p-values. Support or reject the null hypothesis based on what you find.

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 $T(Y)$ 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

$$T(Y) = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{s} \dot{\sim} \mathcal{N}(0, 1).$$

;
But when the sample size is small, we can no longer use the z-test. Instead, there is the t-statistic

$$T(Y) = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{s} \sim t_{n-1},$$

and with this statistic, we can instead perform what is known as the t-test.
Asymptotic Hypothesis Testing 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
In this test, we use the asymptotic pivot of the MLE $\hat{\theta}$ under the null to obtain

$$T(Y) = \sqrt{\mathcal{I}_Y(\theta_0)}(\hat{\theta} - \theta_0) \dot{\sim} \mathcal{N}(0, 1).$$

Score Test
In this test, we use the asymptotic pivot of the score $s(\theta_0)$ under the null to get

$$T(Y) = \frac{s(\theta_0)}{\sqrt{\mathcal{I}_Y(\theta_0)}} \dot{\sim} \mathcal{N}(0, 1).$$

Likelihood Ratio Test
In this test, we use the asymptotic pivot of the likelihood ratio under the null:

$$T(Y) = 2\log\frac{L(\hat{\theta})}{L(\theta_0)} = 2(\ell(\hat{\theta}) - \ell(\theta_0)) \dot{\sim} \chi_1^2.$$

Bayesian Inference

In the Bayesian framework we treat θ as a rv. We have a model/likelihood along with a prior distribution. We want to find the posterior distribution of θ given the data. We just apply Bayes' Rule! We have some estimators.
Posterior mean

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

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

$$\hat{\theta}_{MAP} = \max_\theta f(\theta \mid y).$$

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

$$\hat{\theta}_M = F_{\theta|y}^{-1}(1/2)$$

This minimizes the average posterior absolute loss $\mathbb{E}[| \ \theta - \hat{\theta} \| Y]$
Conjugacy
Let $\mathcal{G}$ be some family of PDFs, and suppose your prior $g(\theta) \in \mathcal{G}$. You observe data from some distribution $f(x \mid \theta)$. We say g is a conjugate prior for the likelihood f iff

$$g(\theta \mid x) \propto g(\theta) \cdot f(x \mid \theta) \in \mathcal{G}$$

Some common conjugacies are listed in the table below. Note that in STAT 111, we focus on Beta-Binomial, Gamma-Poisson, and Normal-Normal

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

If there is no conjugacy, follow the following 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. Suppose we have a loss function $\operatorname{Loss}(\theta, \hat{\theta})$. The *risk function* of $\hat{\theta}$ is defined as $r_{\hat{\theta}}(\theta) = \mathbb{E}(\operatorname{Loss}(\theta, \hat{\theta}) \mid \theta)$.

An estimator $\hat{\theta}$ is *inadmissible* if there exists another estimator whose risk function is at most that of $\hat{\theta}$ for all possible θ, with strict inequality for at least one possible value of θ. An estimator is *admissible* if it is not inadmissible. "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

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

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

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

Sampling

Different Sampling Strategy

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).

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

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

Stratified Sampling
In sampling, it is often desirable to partition the population $y_1, \ldots, y_n$ into subpopulations. This is called *stratification*, and each subpopulation is called a *stratum*. In stratified sampling, a sample is drawn from each *stratum*, with these samples independent across *strata*. Sampling can be carried out with replacement in each stratum, or without replacement in each stratum.

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

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

is an unbiased estimator for the population total

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

If N is known, then

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

is an unbiased estimator for the population mean μ.

Resampling

Bootstrap Procedures

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

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

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

Bootstrap with Hypothesis Testing: Permutation Test

Suppose we have $X_1, \ldots, X_m \overset{i.i.d}{\sim} F_X$ 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:

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

Performing Permutation Test The complete permutation test can be conducted as:

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. Compute p-value:

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

Causal Inference

Randomized Control Trials

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

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

which is equivalent to saying that the assignment mechanisms satisfies:

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

Population Based Modelling

The *population quantity*, $\mathbb{E}(\tau_1)$ 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 $\mathbb{E}(\tau_j)$ 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:

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

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

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

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

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:

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

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

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

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

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

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

We have that $Y_1=W_1Y_1(1)+(1-W_1)Y_1(0)$, thus,

$$W_1Y_1=W_1Y_1(1)$$

$$(1-W_1)Y_1=(1-W_1)Y_1(0)$$

Neyman Null and Fisher Null

- We use the *Neyman Null* $H_0:\bar{\tau}=0$ against $H_1:\bar{\tau}\neq 0$. Note that Neyman null allows individual causal effects to be non-zero, but they must balance out over the finite sample.
- We use the Fisher null $H_0:\tau_j=0,\forall j=1,\cdots,n$, which says there's no treatment effect at all for any individual (i.e. $Y_j(1)=Y_j(0)=Y_j$), against $H_1:\sum_{j=1}^n|\tau_j|>0$.

Randomized testing with Fisher Null

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.$W^{(1)},\cdots,W^{(M)}$ 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

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

Mathematical Tools

Taylor Approximation

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

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

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.

Differentiation under the internal sign

For any function f, by DuThIS, we have that

$$\frac{d}{dx}\int_a^b f(x,t)dt=\int_a^b\frac{d}{dx}f(x,t)dt$$

Sum of Squares Identity

Now, let's talk about some additional notation that might pop up here and there. 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

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

You are encouraged to use the expressions $\bar{Y}$ and S^2, along with their crystallized analogues $\bar{y}$ and s^2, freely without having to rederive them! 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.

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

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

Nomral 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.
$L(t)=\frac{1}{\left(\begin{array}{l}t\\n\end{array}\right)}$ if $Y_1,Y_2,\ldots Y_n\in\{1,2,\ldots t\}$ and 0 otherwise

$$=\frac{\operatorname{Ind}\left(Y_{(n)}\leq t\right)}{\left(\begin{array}{l}t\\n\end{array}\right)}$$

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.

$$P\left(Y_{(n)}=m\right)=\frac{\left(\begin{array}{c} m-1 \\ n-1 \end{array}\right)}{\left(\begin{array}{c} t \\ n \end{array}\right)}$$

$$\mathrm{E}\left(Y_{(n)}\right)=\sum_{m=n}^{t} m\left(\begin{array}{c} m-1 \\ n-1 \end{array}\right)=\frac{n}{n+1}(t+1)$$

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: $M_n \dot{\sim} \mathcal{N}\left(\theta, \frac{\pi}{2} \frac{\sigma^2}{n}\right)$ 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 $Y_1, Y_2, \ldots Y_n \sim \operatorname{Pois}(\theta)$. via Mean: 1. $\theta=\mathrm{E}(Y)$ 2. $\hat{\theta}_{MoM}=\bar{Y}$ via Variance: 1. $\theta=\operatorname{Var}(Y)=\mathrm{E}\left(Y^2\right)-(\mathrm{E}(Y))^2$ 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 $\sqrt{T}$?

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

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

Pivot based on Student-t distribution

Let the data be i.i.d $Y_1, \cdots, Y_n \sim N\left(\mu, \sigma^2\right)$ 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*

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

Gamma–Poisson story for buses in Blotchville (Story 8.4.5 Stat 110)

In Blotchville, buses arrive at a certain bus stop according to a Poisson process with rate λ buses per hour, where λ is unknown. Based on his very memorable adventures in Blotchville, Fred quantifies his uncertainty about λ using the prior $\lambda \sim Gamma(r_0, b_0)$, where r_0 and b_0 are known, positive constants with r_0 an integer. To better understand the bus system, Fred wants to learn more about λ. He is a very patient person, and decides that he will sit at the bus stop for t hours and count how many buses arrive in this time interval. Let Y be the number of buses in this time interval, and suppose Fred observes that $Y=y$
Results:

- Then the *posterior distribution* for λ (which is also equivalent to the conditional distribution of λ given the data y is Gamma (r_0+y, b_0+t)
- Since conditionals PDF are PDFs, the posterior mean

$$E(\lambda | Y=y)=\frac{r_0+y}{b_0+t} \text{ and } Var(\lambda | Y=y)=\frac{r_0+y}{(b_0+t)^2}$$

- Knowing the above, and simplifying, we also have that the marginal distribution for Y is

$$Y \sim NBin(r_0, b_0/(b_0+t))$$

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 $\hat{\mu}_{\mathrm{JS}}$ is then obtained from $\hat{\mu}_{\text{Bayes}}$ by replacing μ_0 by $\bar{Y}$ and b by $\hat{b}$:

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

Table of Distributions

Note: A table of distributions will be provided in the final exam. An example is at pg 3 of the final review handout.

Distribution	PMF/PDF and Support	Expected Value	Variance	MGF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q=1-p$	p	pq	$q+pe^t$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k q^{n-k}$ $k \in \{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^k p$ $k \in \{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t}$, $qe^t<1$
Negative Binomial NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^r q^n$ $n \in \{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r$, $qe^t<1$
Hypergeometric Hypergeometric(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k \in \{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\left(\frac{w+b-n}{w+b-1}\right)n\frac{\mu}{n}(1-\frac{\mu}{n})$	messy
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k \in \{0,\ 1,\ 2,\ \dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal $N(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in (-\infty,\infty)$	μ	σ^2	$e^{t\mu+\frac{\sigma^2 t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x \in (0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$, $t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x \in (0,\infty)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^a$, $t<\lambda$
Beta Beta(a,b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x \in (0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	messy
Log-Normal $\mathcal{LN}(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}$ $x \in (0,\infty)$	$\theta=e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$	doesn't exist
Chi-Square χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x \in (0,\infty)$	n	$2n$	$(1-2t)^{-n/2}$, $t<1/2$
Student-t t_n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$ $x \in (-\infty,\infty)$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$	doesn't exist

Conjugate Priors and Posteriors

Conjugate Prior for some discrete distribution

Sample Space	Likelihood	Conjugate Prior	Posterior
$X = \{0, 1\}$	Bernoulli(θ)	Beta(α, β)	Beta$(\alpha + n\bar{X}, \beta + n(1 - \bar{X}))$
$X = \{0, 1\}$	NBinom(k, θ)	$\theta \sim$Beta(α, β)	Beta$(\alpha + rn, \beta + \sum_{i=1}^{n} x_i))$
$X = \mathbb{Z}_+$	Poisson(λ)	Gamma(α, β)	Gamma$(\alpha + n\bar{X}, \beta + n)$
$X = \mathbb{Z}_{++}$	Geometric(θ)	Gamma(α, β)	Gamma$(\alpha + n, \beta + n\bar{X})$
$X = \mathbb{H}_k$	Multinomial(θ)	Dirichlet(α)	Dirichlet$(\alpha + n\bar{X})$

Conjugate Priors for some continuous distribution

Likelihood	Conjugate Prior	Posterior
Uniform(θ)	Pareto(ν_0, k)	Pareto $\left(\max\{\nu_0, X_{(n)}\}, n + k\right)$
Exponential(θ)	Gamma (α, β)	Gamma$(\alpha + n, \beta + n\bar{X})$
N(μ, σ^2), known σ^2	N(μ_0, σ_0^2)	$\mathrm{N}\left(\left(\frac{1}{\sigma_0^2 + \frac{n}{\sigma^2}}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{X}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$
N(μ, σ^2), known μ	InvGamma(α, β)	$\mathrm{InvGamma}\left(\alpha + \frac{n}{2}, \frac{n}{2}\overline{(X - \mu)^2}\right)$
N(μ, σ^2), known μ	ScaledInv-$\chi^2(\nu_0, \sigma_0^2)$	ScaledInv-$\chi^2 \left(\nu_0 + n, \frac{\nu_0 \sigma_0^2}{\nu_0 + n} + \frac{n\overline{(X - \mu)^2}}{\nu_0 + n}\right)$
N(μ, Σ), known Σ	N(μ_0, Σ_0)	$\mathrm{N}\left(K(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\overline{X}), K\right), K = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$
N(μ, Σ), known μ	InvWishart(ν_0, S_0)	InvWishart$(\nu_0 + n, S_0 + n\overline{S})$, $\overline{S}$ sample covariance