Midterm Review

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

1 Introduction

1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

1.2 Office Hours

- Mondays, 7:30 - 9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM - 12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30 – 11:30 AM in Cabot D-Hall (Emily).

2 Mathematics

2.1 Important Formulas

- $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \ldots$ by Taylor series.
- $e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$ by compound interest.
- $\sum_{n=0}^{\infty} x^n = 1 + x + x^2 + \ldots = \frac{1}{1-x}$ for $|x| < 1$ by infinite geometric series.
- $\sum_{k=0}^{n-1} ar^k = a + ar + ar^2 + \cdots + ar^{n-1} = a\frac{1-r^n}{1-r}$ for $r \neq 1$ by finite geometric series.
- $\sum_{j=1}^{n} j = 1 + 2 + \ldots + n = \frac{n(n+1)}{2}$ by arithmetic series.
- $\sum_{j=1}^{n} \frac{1}{j} = 1 + \frac{1}{2} + \ldots + \frac{1}{n} \approx \log(n) + 0.577$ by harmonic series.
- $\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ by beta function.
- $\int_0^1 \binom{n}{k} x^k(1-x)^{n-k}dx = \frac{1}{n+1}$ by Bayes' Billiards.
- For $a > 0$, $\Gamma(a) = \int_0^{\infty} x^{a-1}e^{-x}dx$, $\Gamma(a+1) = a\Gamma(a)$ by gamma function.
- For $n \in \mathbb{Z}^+$, $\Gamma(n) = (n-1)!$ by gamma function.
- For $n = \frac{1}{2}$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ by gamma function.
- $\sum_{i=1}^{n}(Y_i - c)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - c)^2$ for constant c by sum of squares identity.
- $\left(\sum_{i=1}^{n} Y_i\right)^2 = \sum_{i=1}^{n} Y_i^2 + \sum_{i\neq j} Y_iY_j$, where $\sum_{i\neq j} Y_iY_j = 2\sum_{i<j} Y_iY_j$ by square of sum.
- $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$.

2.2 Functions

Definition 1 (Inverse of a function)**.** For a function $f(x) = y$, the inverse is $f^{-1}(y) = x$ such that $f^{-1}(f(x)) = x$ and $f(f^{-1}(y)) = y$.

• **Strategy**: Replace $f(x)$ with y, swap x and y, solve for y, and replace y with $f^{-1}(x)$.

Definition 2 (Maximum of a function)**.** The largest value a function attains within its domain.

• E.g., for $f(x) = -x^2 + 3$, $\max f(x) = 3$.

Definition 3 (Arg max of a function)**.** The input value that achieves the maximum of a function. It is good notation to specify the set of possible input values.

• E.g., for $f(x) = -x^2 + 3$, $\arg\max_{x \in \mathbb{R}} f(x) = 0$.

Definition 4 (Inequalities)**.** Flip the inequality sign when multiplying by a negative number and, if both sides are the same sign, when taking the reciprocal.

• E.g., $2 < 3 \implies -2 > -3$. E.g., $2 < 3 \implies \frac{1}{2} > \frac{1}{3}$.

2.3 Operations

Definition 5 (Logarithm)**.** $\log_b(a) = c \iff b^c = a$.

• $\log_b(1) = 0, \log_b(b) = 1, \log_b(0)$ is undefined.

• $\log(xy) = \log(x) + \log(y)$.

• $\log(\frac{x}{y}) = \log(x) - \log(y)$.

• $\log(x^k) = k \log(x)$.

Definition 6 (Derivative)**.** $\frac{d}{dx} f(x) = f'(x)$.

• $\frac{d}{dx}(ax + b) = a$ for constants a, b by linearity of differentiation.

• $\frac{d}{dx}(x^n) = nx^{n-1}$ by power rule.

• $\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x)$ by product rule.

• $\frac{d}{dx} f(g(x)) = f'(g(x))g'(x)$ by chain rule.

• $\frac{\partial}{\partial x}(axy + bx + cy + d) = ay + b$ for constants a, b, c, d by partial differentiation.

Definition 7 (Integral)**.** $\int_a^c f(x)dx = F(c) - F(a)$, where $F'(x) = f(x)$.

• $\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx$ for constants a, b, c where $a \leq b \leq c$.

• $\int_a^c f(g(x))g'(x)dx = \int_{g(a)}^{g(c)} f(u)du$, where $u = g(x)$ and $du = g'(x)dx$ by u-substitution.

3 Probability

Definition 8 (Expectation)**.** Informally, the weighted average of a random variable.

• $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ for constants a, b by linearity.

• $\mathbb{E}[X] = \sum_{\text{supp}(X)} xP(X = x)$ for X discrete by definition.

- $\mathbb{E}[X] = \int_{\text{supp}(X)} x f_X(x) dx$ for X continuous by definition.
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{1}_1 + \cdots + \mathbb{1}_n]$ where $X = \mathbb{1}_1 + \cdots + \mathbb{1}_n$ by substitution.
- $\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X \mid Y]\right]$ by Adam's Law.
- $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] P(A_i)$ by Law of Total Expectation.
- $\mathbb{E}[X^2] = \text{Var}[X] + (\mathbb{E}[X])^2$ by definition of variance.
- $\mathbb{E}[XY] = \mathbb{E}\left[\mathbb{E}[XY \mid X]\right] = \mathbb{E}\left[X \cdot \mathbb{E}[Y \mid X]\right]$ by taking out what's known.
- $\mathbb{E}[XY] = \text{Cov}[X, Y] + \mathbb{E}[X]\mathbb{E}[Y]$ by definition of covariance.
- $\mathbb{E}[X \mid Y] = \mathbb{E}\left[\mathbb{E}[X \mid Z, Y] \mid Y\right]$ by Adam's Law with extra conditioning on Y.
- $\mathbb{E}[X \mid Y = y] = \frac{\mathbb{E}[X \mathbb{1}\{Y=y\}]}{P(Y=y)}$ for Y discrete and $P(Y = y) > 0$ by rearranging LOTE.
- $\mathbb{E}\left[g(X)\right] = \sum_{\text{supp}(X)} g(x) P(X = x)$ for X discrete by Law of the Unconscious Statistician.
- $\mathbb{E}\left[g(X)\right] = \int_{\text{supp}(X)} g(x) f_X(x) dx$ for X continuous by Law of the Unconscious Statistician.
- $\mathbb{E}\left[g(X)\right] \geq g\left(\mathbb{E}[X]\right)$ for g convex by Jensen's Inequality.
- $\mathbb{E}\left[g(X)\right] \leq g\left(\mathbb{E}[X]\right)$ for g concave by Jensen's Inequality.

Definition 9 (Conditional expectation)**.** Informally, expectation takes a random variable as input and returns a number as output.

- This is true for unconditional expectation and expectation conditional on an event.
- However, expectation conditional on another random variable is a random variable (specifically, a function of possible crystallizations of that other random variable).
- E.g., let X be someone's yearly salary and Y be someone's daily salary. Then $\mathbb{E}[X \mid Y] = 365Y$, a random variable. Notice $\mathbb{E}[X \mid Y = 1] = 365$, $\mathbb{E}[X \mid Y = 2] = 730$, etc.

Definition 10 (Variance)**.** Informally, a measure of the spread of a random variable, which must be non-negative.

- $\text{Var}[aX + b] = a^2 \text{Var}[X]$ for constants a, b by bilinearity.
- $\text{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ by definition.
- $\text{Var}[X] = (\text{SD}[X])^2$ by definition of standard deviation.
- $\text{Var}[X] = \mathbb{E}\left[\text{Var}[X \mid Y]\right] + \text{Var}\left[\mathbb{E}[X \mid Y]\right]$ by Eve's Law.
- $\text{Var}[X \mid Y] = \mathbb{E}\left[\text{Var}[X \mid Z, Y] \mid Y\right] + \text{Var}\left[\mathbb{E}[X \mid Z, Y] \mid Y\right]$ by Eve's Law with extra conditioning on Y.
- $\text{Var}[X_1 + \ldots + X_n] = \sum_{i=1}^{n} \text{Var}[X_i] + 2 \sum_{i<j} \text{Cov}[X_i, X_j]$ by bilinearity.

Definition 11 (Covariance)**.** Informally, a measure of the co-movement of two random variables.

- $\text{Cov}[X, Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ by definition.
- $\text{Cov}[X, Y] = \text{Cov}[Y, X]$.
- $\text{Cov}[X, c] = 0$ for constant c.
- $\text{Cov}[cX + b, Y] = c\text{Cov}[X, Y]$ for constants c, b.
- $\text{Cov}[X, X] = \text{Var}[X]$.

- $\text{Cov}[W+X, Y+Z] = \text{Cov}[W,Y] + \text{Cov}[W,Z] + \text{Cov}[X,Y] + \text{Cov}[X,Z]$.
- $X \perp\!\!\!\perp Y \implies \text{Cov}[X,Y] = 0$, but generally, $\text{Cov}[X,Y] = 0 \not\implies X \perp\!\!\!\perp Y$.

Definition 12 (Correlation)**.** Informally, a standardized measure of covariance, bounded from -1 to 1.

- $\text{Corr}[X,Y] = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$.

4 Quantities

4.1 Important Quantities

Definition 13 (Order statistic)**.** The order statistics of $\vec{Y} = (Y_1, \ldots, Y_n)$ are the same data points, sorted in increasing order: $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$, where $Y_{(j)}$ is the jth order statistic.

- E.g., for $\vec{Y} = (5, 3, 10)$, $Y_{(1)} = Y_2 = 3$.

Definition 14 (CDF)**.** For a random variable Y, its cumulative distribution function (CDF) is $F(y) = P(Y \leq y)$.

Definition 15 (Empirical CDF)**.** For data $\vec{Y} = (Y_1, \ldots, Y_n)$, the empirical CDF (ECDF) is $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y\}$.

Definition 16 (p-quantile)**.** For a random variable Y, let F be its CDF. The quantile function of Y is $Q(p) = \min\{y : F(y) \geq p\}$, where $Q(p)$ is the p-quantile of the distribution.

- I.e., the p-quantile is the smallest y such that the CDF at y attains at least a value of p.
- If F is continuous and strictly increasing, then F^{-1} exists (as a "true" inverse), so we use $Q(p) = F^{-1}(p)$ such that $P(Y \leq y) = p \iff F(y) = p \iff y = F^{-1}(p) \iff Q(p) = y$. Notice $F : \mathbb{R} \to [0,1]$ while $F^{-1} : [0,1] \to \mathbb{R}$.
- If F^{-1} doesn't "truly" exist (e.g., when Y is discrete), we use a "generalized" inverse for the quantile function: $Q(p) = \min\{y : F(y) \geq p\}$ such that $P(Y \leq y) \geq p \iff Q(p) \leq y$.
- E.g., suppose SAT score is distributed $\mathcal{N}(1000, 200^2)$. If we're interested in the 0.5-quantile (i.e., median), then $Q(0.5) = 1000$ since symmetric distributions have mean = median. Notice 50% of the distribution lies to the left of $y = 1000$.

Definition 17 (Sample p-quantile)**.** For data $\vec{Y} = (Y_1, \ldots, Y_n)$, the sample p-quantile is $\hat{Q}(p) = Y_{(\lceil np \rceil)}$.

- E.g., suppose SAT score is distributed $\mathcal{N}(\mu, \sigma^2)$ for some unknown μ, σ^2. We observe $Y_1, \ldots, Y_{11}$. If we're interested in the sample 0.5-quantile (i.e., sample median), then $\hat{Q}(0.5) = Y_{(\lceil 11/2 \rceil)} = Y_{(6)}$ (i.e., the 6th smallest data point). Notice 50% of the data lies to the left of $y = Y_{(6)}$.

Concept Checker 1. Rewrite the following probabilities as quantiles.

1. Assume F is continuous and strictly increasing. For Z continuous, $P(Z \leq 1.96) = 0.975 \iff$ _____
2. Assume $F(3) < F(4)$. For X discrete, $P(X > 4) = 0.05 \iff$ _____

Solution

4.2 Theoretical vs. Sample Quantities

Quantity	Theoretical	Sample
Mean	$\mu = \mathbb{E}[Y]$	$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
kth moment	$\mu'_k = \mathbb{E}[Y^k]$	$M_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$
Variance	$\sigma^2 = \text{Var}[Y]$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
Standard deviation	$\sigma = \text{SD}[Y]$	$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
Covariance	$\sigma_{XY} = \text{Cov}[X, Y]$	$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
Correlation	$\rho = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$	$r_{XY} = \frac{S_{XY}}{S_X S_Y}$
CDF	$F(y) = P(Y \leq y)$	$\hat{F}(y) = d$
Median	$Q(\frac{1}{2}) = F^{-1}(\frac{1}{2})$	$\hat{Q}(\frac{1}{2}) = Y_{(\lceil n/2 \rceil)}$
p-quantile	$Q(p) = F^{-1}(p)$	$\hat{Q}(p) = Y_{(\lceil np \rceil)}$
Probability	$P(Y = y)$	$\hat{P}(Y = y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i = y\}$

Concept Checker 2. It can feel weird to think of probability as a theoretical quantity, so consider $Y_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$, with λ unknown.[1] Suppose we observe $y_1 = 5$, $y_2 = 5$, and $y_3 = 4$. What is $P(Y_i = 5)$? What is $\hat{P}(Y_i = 5)$?

Solution

5 Inference

5.1 Estimands, Estimators, and Estimates

Definition 18 (Statistic)**.** A random variable that is a function of the data $\vec{Y}$. Usually denoted as $T(\vec{Y})$, where the function T must not involve any unknown parameters.

Definition 19 (Estimand)**.** The unknown quantity (often, a parameter in a model). Usually denoted as θ, with Θ as the set of possible values for θ (i.e., the parameter space).

• E.g., let θ be the average hours of sleep Harvard students get per night. $\Theta = [0, 24]$.

[1] $Y_i \overset{\text{i.i.d.}}{\sim} F$ is shorthand notation for $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F$ for some model F.

Definition 20 (Estimator)**.** A statistic used to estimate the estimand (i.e., we can use the data to calculate this). Usually denoted as $\hat{\theta}$.

• E.g., we will observe 5 values, $\vec{Y} = (Y_1, Y_2, \ldots, Y_5)$, so I propose we use $\hat{\theta} = \frac{1}{5}\sum_{i=1}^{5} Y_i$.

Definition 21 (Estimate)**.** A crystallization of the estimator from the observed data.

• E.g., after the experiment, the data will crystallize, and we can calculate our estimate as $\hat{\theta} = \frac{1}{5}\sum_{i=1}^{5} y_i$.

Example 1 (Theoretical and sample quantities)**.** Each theoretical quantity fits the definition of an estimand. Thus, the sample quantities are estimators.[2] They can be calculated with our limited sample data.

• Using the same "hat" notation, we may sometimes see the sample mean $\bar{Y}$ denoted as $\hat{\mu}$ to emphasize it's an estimator for the theoretical mean μ.[3]

Concept Checker 3. Let the data be $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$, with θ unknown. Which of the following quantities is not a valid estimator for θ?

1. $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i$
2. $\hat{\theta}_2 = \mathbb{E}[Y_1]$
3. $\hat{\theta}_3 = \Phi(\frac{1}{Y_1}) - \arcsin(e^{Y_1})$
4. $\hat{\theta}_4 = 3$

Solution

5.2 Evaluation of Estimators

Definition 22 (Standard error)**.** $\text{SE}[\hat{\theta}] = \text{SD}[\hat{\theta}] = \sqrt{\text{Var}[\hat{\theta}]}$.

Definition 23 (Bias)**.** $\text{Bias}[\hat{\theta}] = \mathbb{E}[\hat{\theta} - \theta]$.[4]

• We say $\hat{\theta}$ is unbiased for θ if $\text{Bias}[\hat{\theta}] = 0$ (i.e., $\mathbb{E}[\hat{\theta}] = \theta$).

Definition 24 (Mean absolute error)**.** $\text{MAE}[\hat{\theta}] = \mathbb{E}[|\hat{\theta} - \theta|]$ is the risk function for the absolute error loss $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$.

[2] Or estimates, depending on whether we consider the data as crystallized ($\vec{y}$) instead of random ($\vec{Y}$).

[3] There are many possible estimators for an estimand. For example, to estimate the theoretical variance $\sigma^2 = \text{Var}[Y_1]$, we can use the method of moments estimator $\hat{\sigma}^2_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ or the ordinary least squares estimator $\hat{\sigma}^2_{\text{OLS}} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. In contrast, we generally agree upon the sample variance as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (i.e., the OLS estimator). This is arguably the "best" estimator.

[4] You may see bias denoted as $\text{Bias}[\hat{\theta}, \theta]$ or $\text{Bias}_\theta[\hat{\theta}]$ to make it explicit this is a function of the estimand as well.

Definition 25 (Mean square error)**.** $\text{MSE}[\hat{\theta}] = \mathbb{E}[(\hat{\theta}-\theta)^2]$ is the risk function for the squared error loss $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$.

• Equivalently, $\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}])^2$, which demonstrates bias-variance tradeoff.

Definition 26 (Consistency)**.** An estimator $\hat{\theta}$ is consistent for the estimand θ if $\hat{\theta} \xrightarrow{p} \theta$.

• Also, $\hat{\theta}$ is consistent for the estimand θ if $\lim_{n\to\infty} \text{MSE}[\hat{\theta}] = 0$.

Concept Checker 4. For any unbiased estimator $\hat{\theta}_{\text{UB}}$, what is $\text{MSE}[\hat{\theta}_{\text{UB}}]$ always equal to?

Solution

6 Likelihood

Definition 27 (Likelihood)**.** $\mathcal{L}(\theta; \vec{y}) = f_{\vec{Y}}(\vec{y}; \theta)$, where $f_{\vec{Y}}(\vec{y}; \theta)$ is the joint density of the data. Likelihood is very analogous to probability, but they're not the same.

• We drop multiplicative constants that are not functions of the parameter since they don't affect the argmax. E.g., for $Y \sim \text{Bin}(3, \theta)$, $\mathcal{L}(\theta; y) = \binom{3}{y}\theta^y(1-\theta)^{3-y}$ is equivalent to $\mathcal{L}(\theta; y) = \theta^y(1-\theta)^{3-y}$.[5]

• Likelihood is invariant under one-to-one transformation of the parameter. I.e., $\mathcal{L}(\psi; \vec{y}) = f_{\vec{Y}}(\vec{y}; \psi) = f_{\vec{Y}}(\vec{y}; \theta) = \mathcal{L}(\theta; \vec{y})$, where $\psi = g(\theta)$ and g is a known one-to-one function.

• Likelihood is invariant under one-to-one transformation of the data. I.e., $\mathcal{L}(\theta; \vec{x}) = \mathcal{L}(\theta; \vec{y})$, where $\vec{y}$ is the data from a model parameterized by θ, $\vec{x} = h(\vec{y})$, and h is a known one-to-one function from $\mathbb{R}^n$ to $\mathbb{R}^n$.

Definition 28 (Log-likelihood)**.** $\ell(\theta; \vec{y}) = \log(\mathcal{L}(\theta; \vec{y}))$.

• $f(x) = \log(x)$ is monotonically increasing (i.e., as the input increases, the output never decreases).

• We drop additive constants that are not functions of the parameter since they don't affect the argmax. E.g., for $Y \sim \text{Bin}(3, \theta)$, $\ell(\theta; y) = \log\left(\binom{3}{y}\right) + \log(\theta^y) + \log((1-\theta)^{3-y})$ is equivalent to $\ell(\theta; y) = \log(\theta^y) + \log((1-\theta)^{3-y})$.

7 Estimators

7.1 Maximum Likelihood Estimator (MLE)

Definition 29 (Maximum likelihood estimator)**.** The estimator that maximizes the likelihood (i.e., the "most likely" value for θ, given the data): $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta; \vec{Y})$.

• Equivalently, $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \ell(\theta; \vec{Y})$ since $f(x) = \log(x)$ is monotonically increasing. In practice, working with log-likelihood is often easier.

[5] We say two likelihood functions are "equivalent (=)" even if, more formally, they are "proportional ($\propto$)" up to a multiplicative constant that is not a function of the parameter.

• If $\hat{\theta}_{\text{MLE}}$ is the MLE for θ and g is a known function, then $g(\hat{\theta}_{\text{MLE}})$ is the MLE for $g(\theta)$ by invariance. When g is one-to-one, this follows from invariance of likelihood. When g is not one-to-one, we simply "define" $g(\hat{\theta}_{\text{MLE}})$ to be the MLE for $g(\theta)$.

• **Strategy**: Try invariance, German Tank Problem, and calculus (in that order).

7.2 Known MLEs

• $Y_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, with p unknown $\implies \hat{p}_{\text{MLE}} = \bar{Y}$

• $Y_i \overset{\text{i.i.d.}}{\sim} \text{Geom}(p)$, with p unknown $\implies \hat{p}_{\text{MLE}} = \frac{1}{\bar{Y}}$

• $Y_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$, with λ unknown $\implies \hat{\lambda}_{\text{MLE}} = \bar{Y}$

• $N \sim \text{Pois}(\lambda t)$, with t known and λ unknown $\implies \hat{\lambda}_{\text{MLE}} = \frac{N}{t}$

• $Y_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a, \lambda)$, with a known and λ unknown $\implies \hat{\lambda}_{\text{MLE}} = \frac{a}{\bar{Y}}$

• $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with μ and σ^2 unknown $\implies \hat{\mu}_{\text{MLE}} = \bar{Y}, \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

7.3 Strategies to Find MLE

Invariance: If $\theta = f(\lambda)$, then $\hat{\theta}_{\text{MLE}} = f(\hat{\lambda}_{\text{MLE}})$ by invariance.

• Hints: If we know $\hat{\lambda}_{\text{MLE}}$.

German Tank Problem: If $\mathcal{L}(\theta; \vec{Y})$ strictly decreases as θ increases, then its first possible input (that doesn't result in $\mathcal{L}(\theta; \vec{Y}) = 0$) is $\hat{\theta}_{\text{MLE}}$.

• Hints: First, write out $\mathcal{L}(\theta; \vec{Y})$. If there is an indicator in the likelihood (so that $\mathcal{L}(\theta; \vec{y})$ is not differentiable at the jump discontinuity). If θ must be discrete (so that $\mathcal{L}(\theta; \vec{y})$ is not differentiable). If θ is involved in the support.

Calculus: $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \ell(\theta; \vec{Y})$ (i.e., set $\ell'(\theta; \vec{Y}) = 0$ and check $\ell''(\theta; \vec{Y}) < 0$).

• Hints: If the strategies above don't apply.

Concept Checker 5. Which strategy would be most efficient for finding $\hat{\theta}_{\text{MLE}}$ for the following scenarios:

1. For i.i.d. data $Y_1, \ldots, Y_n$, $\mathcal{L}(\theta; \vec{Y}) = \left(\frac{5}{\theta^5}\right)^n \left(\prod_{i=1}^{n} e^{Y_i}\right)^5 \mathbb{1}\{Y_{(n)} \leq \theta\}$, where $\theta \geq 0$.
2. For $Y_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, $\theta = P(Y_i = 0)$.
3. $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$.
4. For i.i.d. data $Y_1, \ldots, Y_n$, $f_{Y_i}(y) = \theta y^{\theta - 1}$, where $0 < y < 1$ and $\theta > 0$.

Solution

7.4 Method of moments estimator

Definition 30 (Method of moments estimator)**.** The estimator where the theoretical moments are replaced with sample moments.

• For i.i.d. data $Y_1, \ldots, Y_n$, $\mu'_k = \mathbb{E}[Y_i^k]$ is the theoretical kth moment whereas $M_k = \frac{1}{n}\sum_{i=1}^n Y_i^k$ is the sample kth moment.[6]

• **Strategy**: Write the estimand in terms of theoretical moments and replace them with their sample analogues.

Concept Checker 6. What is the sample moment analogue for $\mathbb{E}[Y_i^2]$?

Solution

Concept Checker 7. Let $Y_i \overset{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$. Provide the justification at each step for the derivation of the MOM estimator.

$$\mathbb{E}[Y_i] = \frac{1}{\lambda} \quad \text{by } \_\_\_\_\_$$

$$\lambda = \frac{1}{\mathbb{E}[Y_i]} \quad \text{by } \_\_\_\_\_$$

$$\hat{\lambda}_{\text{MOM}} = \frac{1}{\bar{Y}} \quad \text{by } \_\_\_\_\_$$

[6]Sometimes, the kth sample moment can be denoted as $\overline{Y^k}$. E.g., $\overline{Y^2} = \frac{1}{n}\sum_{i=1}^n Y_i^2$. Notice this is different from $(\bar{Y})^k$. E.g., $(\bar{Y})^2 = (\frac{1}{n}\sum_{i=1}^n Y_i)^2$.

Solution

8 Likelihood Theory

Definition 31 (Regularity conditions)**.** For data $Y_1, \ldots, Y_n \sim F_{\vec{Y};\theta}$, the following must be true:

- $\mathcal{L}(\theta; \vec{y})$ is a smooth function on Θ.
- $\mathbb{E}[s(\theta^*; \vec{Y})]$ and $\text{Var}[s(\theta^*; \vec{Y})]$ exist.
- The support of $\vec{Y}$ doesn't depend on θ.
- We can differentiate under the integral sign (i.e., DUThIS).

Concept Checker 8. Do regularity conditions hold for $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$?

Solution

Concept Checker 9. Let $f_X(x)$ be the PDF for a random variable X. Assume we can DUThIS. How can we rewrite $\frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx$.

Solution

Definition 32 (Score function)**.** $s(\theta; \vec{y}) = \ell'(\theta; \vec{y}) = \frac{1}{\mathcal{L}(\theta;\vec{y})}(\mathcal{L}'(\theta; \vec{y}))$, where the derivative is with respect to θ.

- Previously, we've regarded $s(\theta; \vec{y})$ as a function of θ and set it to 0. However, we can also consider θ as fixed at its true value θ^* and regard $s(\theta^*; \vec{Y})$ as a function of the random data $\vec{Y}$.
- For $\theta = \theta^*$ under regularity conditions, $\mathbb{E}[s(\theta; \vec{Y})] = 0$.
- For $\theta = \theta^*$ under regularity conditions, $\text{Var}[s(\theta; \vec{Y})] = \mathbb{E}[(s(\theta; \vec{Y}))^2]$.

Definition 33 (Information equality)**.** For $\theta = \theta^*$ under regularity conditions, $\text{Var}[s(\theta; \vec{Y})] = -\mathbb{E}[s'(\theta; \vec{Y})]$.

Definition 34 (Fisher information)**.** For $\theta = \theta^*$, $\mathcal{I}_{\vec{Y}}(\theta) = \text{Var}[s(\theta; \vec{Y})]$, where $\mathcal{I}_{\vec{Y}}(\theta)$ is the Fisher information for θ from $\vec{Y}$.

• Under regularity conditions, $\mathcal{I}_{\vec{Y}}(\theta) = \mathbb{E}[(s(\theta; \vec{Y}))^2]$.

• Under regularity conditions, $\mathcal{I}_{\vec{Y}}(\theta) = -\mathbb{E}[s'(\theta; \vec{Y})]$.

• Fisher information is additive, so for i.i.d. data, $\mathcal{I}_{\vec{Y}}(\theta) = n\mathcal{I}_{Y_1}(\theta)$, where $\mathcal{I}_{Y_1}(\theta)$ is the Fisher information for θ from Y_1.

• Fisher information is not invariant under reparameterization, but if $\tau = g(\theta)$ where g is a differentiable function with $g'(\theta) \neq 0$, then $\mathcal{I}_{\vec{Y}}(\tau) = \frac{\mathcal{I}_{\vec{Y}}(\theta)}{(g'(\theta))^2}$.

• ☣: If you are finding the fisher information for τ, don't leave your final answer for $\mathcal{I}_{\vec{Y}}(\tau)$ in terms of θ!

Definition 35 (Cramér–Rao lower bound)**.** For any unbiased estimator $\hat{\theta}_{\text{UB}}$ under regularity conditions, $\text{Var}[\hat{\theta}_{\text{UB}}] \geq \frac{1}{\mathcal{I}_{\vec{Y}}(\theta)}$.

• For some intuition, variance decreases as we gain more information about θ, which is illustrated in the inverse relationship.

Definition 36 (Maximum likelihood estimator)**.** $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; \vec{Y})$. Under regularity conditions, the MLE has the following properties:

• MLE is invariant, meaning if $\hat{\theta}_{\text{MLE}}$ is the MLE of θ, then $g(\hat{\theta}_{\text{MLE}})$ is the MLE of $g(\theta)$.

• MLE is consistent, meaning $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$.

• MLE is asymptotically unbiased, meaning $\lim_{n\to\infty} \text{Bias}[\hat{\theta}_{\text{MLE}}] = 0$.

• MLE is asymptotically efficient, meaning no other asymptotically unbiased estimator has a lower asymptotic variance.

• MLE is asymptotically Normal, meaning for i.i.d. data $Y_1, \ldots, Y_n$, $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_{Y_1}(\theta)}\right)$.

• ☣: Notice the Fisher information in the Normal distribution is from Y_1, not $\vec{Y}$. For i.i.d. data, $\mathcal{I}_{\vec{Y}}(\theta) = n\mathcal{I}_{Y_1}(\theta)$. Keep the notation clear and consistent!

Concept Checker 10. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_{Y;\mu}$ for some model $F_{Y;\mu}$ parameterized by μ, with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma^2 < \infty$. Assume regularity conditions hold. Suppose $\hat{\mu}_{\text{MLE}} = \bar{Y}$. How can we find $\mathcal{I}_{\vec{Y}}(\mu)$, the Fisher information for μ from $\vec{Y}$?

Solution

Concept Checker 11. Let $Y_i \overset{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$, with λ unknown. We can show $\bar{Y}$ is an unbiased estimator for λ. First, what is the lowest possible variance for $\bar{Y}$? Does $\bar{Y}$ achieve this?

Next, suppose we're interested in estimating $\theta = P(Y_i = 0) = e^{-\lambda}$, for which we use $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i = 0\}$. We can show $\hat{\theta}$ is an unbiased estimator for θ. What is the lowest possible variance for $\hat{\theta}$? Does $\hat{\theta}$ achieve this?[7]

Solution

9 Asymptotics

9.1 Types of Distributions

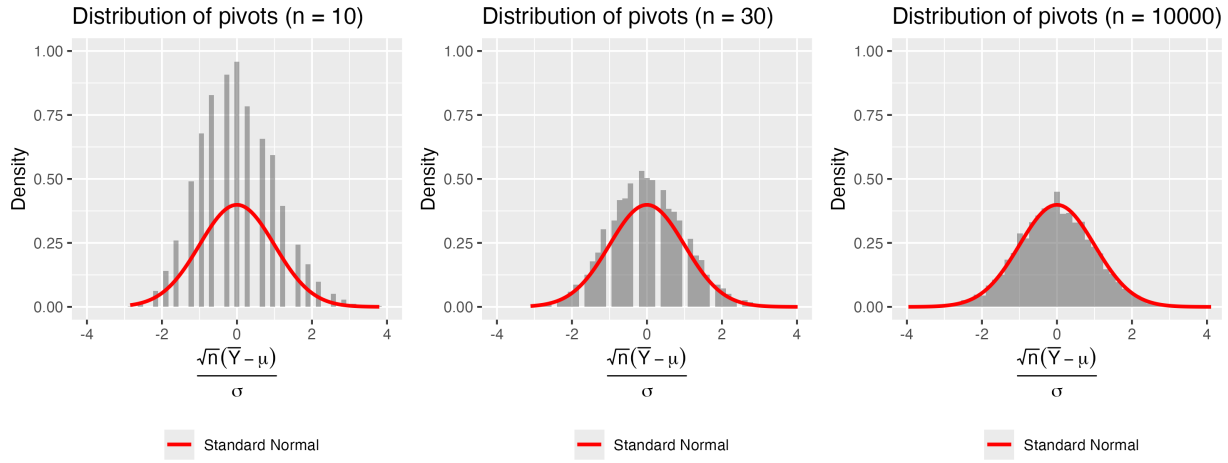

FIGURE 1: $\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma}$ (in gray) converges in distribution to $\mathcal{N}(0,1)$ (in red). If we "freeze" at $n = 30$, the approximate distribution is close to the asymptotic distribution.

Definition 37 (Exact distribution)**.** We can perfectly describe the behavior of the distribution (no approximation, no limit, no dependence on n).

• E.g., $Z \sim \mathcal{N}(0,1)$ is an exact distribution, so $Q(0.975) = 1.96$ (i.e., exactly 97.5% of the distribution falls to the left of 1.96).

Definition 38 (Asymptotic distribution)**.** The finite-sample distribution changes for each n, and the asymptotic distribution is the limit of these distributions as $n \to \infty$.

• E.g., $\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$ is an asymptotic distribution, so $Q(0.975) \to 1.96$ as $n \to \infty$.

[7]Inspired by Problem 2 in "Stat 111 Homework 4, Spring 2026" by Joseph K. Blitzstein and Neil Shephard.

Definition 39 (Approximate distribution)**.** We "freeze" the finite-sample distribution at a sufficiently large n such that the distribution is close to (but not perfectly) the asymptotic distribution.

- $\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma} \dot{\sim} \mathcal{N}(0,1)$ is an approximate distribution for large n, so $Q(0.975) \approx 1.96$.

Example 2 (Approximations of pi)**.** As an analogy, let $\hat{\pi}_n$ be the nth digit of π. E.g., $\hat{\pi}_1 = 3$, $\hat{\pi}_2 = 3.1$, $\hat{\pi}_3 = 3.14$, and so on.

- Think of π as the asymptotic goal; $\hat{\pi}_n$ "converges to" π as $n \to \infty$.
- We can "freeze" at a sufficiently large n such that $\hat{\pi}_n$ is close to (but not perfectly) π.

9.2 Types of Convergences

Definition 40 (Convergence in probability)**.** Let $X_n = X_1, X_2, \ldots$ be a sequence of random variables. X_n converges in probability to a limit X if $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$, denoted $X_n \xrightarrow{p} X$.

Definition 41 (Convergence in distribution)**.** Let $X_n = X_1, X_2, \ldots$ be a sequence of random variables with CDF $F_{X_n}(x)$. X_n converges in distribution to a limit X if $F_{X_n}(x)$ converges to a limiting CDF $F_X(x)$ (i.e., $\forall\ x$ where $F_X(x)$ is continuous, $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$), denoted $X_n \xrightarrow{d} X$.

- Importantly, this is the "weakest" type of convergence as $\xrightarrow{p} \implies \xrightarrow{d}$, but generally, $\xrightarrow{d} \not\implies \xrightarrow{p}$. However, $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$ for constant c.
- For some intuition, if X is the number of heads and Y is the number of tails in n coin flips, X and Y share the same distribution—i.e., $\text{Bin}(n, 0.5)$—but are different quantities!

Example 3 (Sample mean)**.** For n random variables $X_1, \ldots, X_n$, let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be their sample mean. Notice $\bar{X}_n$ is a sequence of random variables.

- For $n = 1$, $\bar{X}_1 = X_1$, which is a random variable.
- For $n = 2$, $\bar{X}_2 = \frac{1}{2}(X_1 + X_2)$, which is a random variable.
- For $n = 3$, $\bar{X}_3 = \frac{1}{3}(X_1 + X_2 + X_3)$, which is a random variable.
- As a sequence, we have $\bar{X}_n = X_1, \frac{1}{2}(X_1 + X_2), \frac{1}{3}(X_1 + X_2 + X_3), \ldots$

9.3 Asymptotic Tools

Definition 42 (Law of Large Numbers)**.** For n i.i.d. random variables $X_1, \ldots, X_n$ with finite expectation $\mathbb{E}[X_i] = \mu$ and sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{X}_n \xrightarrow{p} \mu$.

Definition 43 (Central Limit Theorem)**.** For n i.i.d. random variables $X_1, \ldots, X_n$ with finite expectation $\mathbb{E}[X_i] = \mu$, finite variance $\text{Var}[X_i] = \sigma^2$, and sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

- This implies $\bar{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$ for large n.
- **Tip**: "$\sqrt{n}(\ldots)$" usually means to pattern-match to CLT while "$\frac{1}{n}\sum_{i=1}^n(\ldots)$" usually means to pattern-match to LLN.

• **Tip**: Whenever you see a complicated expression that's a function of a known random variable, define it as a new random variable to pattern-match to one of the asymptotic tools! E.g., if $Y_1, \ldots, Y_n$ are i.i.d., then $Y_1^4, \ldots, Y_n^4$ are also i.i.d., so let $A_i = Y_i^4$. From there, $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n A_i - \mathbb{E}[A_i]\right)$ matches the form of CLT (as long as the assumptions hold).

• ☣: It is incorrect to say $\bar{X}_n \xrightarrow{d} \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Why? Recall convergence in distribution is in the limit as n approaches infinity, but here, n is on the right side of the equation! Thus, for each n, the "thing being converged to" keeps changing.

Definition 44 (Continuous Mapping Theorem)**.** Let $X_n = X_1, X_2, \ldots$ be a sequence of random variables and $g(x)$ be a continuous function. If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$. Additionally, if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

Definition 45 (Theorem 3.5.7.)**.** Let $X_n = X_1, X_2, \ldots$ and $Y_n = Y_1, Y_2, \ldots$ be sequences of random variables. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$, $X_n - Y_n \xrightarrow{p} X - Y$, and $X_n Y_n \xrightarrow{p} XY$. Additionally, if $P(Y_n = 0) = P(Y = 0) = 0$, then $\frac{X_n}{Y_n} \xrightarrow{p} \frac{X}{Y}$.

Definition 46 (Slutsky's Theorem)**.** Let $X_n = X_1, X_2, \ldots$ and $Y_n = Y_1, Y_2, \ldots$ be sequences of random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for any constant c, then $X_n + Y_n \xrightarrow{d} X + c$, $X_n - Y_n \xrightarrow{d} X - c$, and $X_n Y_n \xrightarrow{d} cX$. Additionally, if $c \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.

Definition 47 (Delta Method)**.** If $g(x)$ is a differentiable function and $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$, then $\sqrt{n}\left(g(\hat{\theta}) - g(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0, (g'(\theta))^2\omega^2\right)$.

• This implies $g(\hat{\theta}) \dot{\sim} \mathcal{N}\left(g(\theta), (g'(\theta))^2\frac{\omega^2}{n}\right)$ for large n.

• **Tip**: "Find the asymptotic distribution of $\hat{\theta}$" usually means to write $\sqrt{n}\left(\hat{\theta} - \theta\right)$ and pattern-match to CLT (and possibly Delta Method if $\hat{\theta}$ is a function of something whose asymptotic distribution we know).

• ☣: Don't confuse CMT with Delta Method. CMT is for applying a function to the entire sequence of random variables while Delta Method is for applying a function to only the estimator and its estimand. E.g., if $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, then $e^{\left(\sqrt{n}(\bar{X}_n - \mu)\right)} \xrightarrow{d} e^{\left(\mathcal{N}(0,\sigma^2)\right)}$ by CMT with $g(x) = e^x$, but $\sqrt{n}(e^{\bar{X}_n} - e^{\mu}) \xrightarrow{d} \mathcal{N}(0, (e^{\mu})^2\sigma^2)$ by Delta Method with $g(x) = e^x$.

Concept Checker 12. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with finite expectation $\mathbb{E}[Y_1] = \mu \neq 0$ and finite variance $\text{Var}[Y_1] = \sigma^2 > 0$. Assume $\text{Var}[Y_i^4] < \infty$. What do the following sequences converge to?[8]

1. $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i^4 - \mathbb{E}[Y_i^4]\right) \xrightarrow{d}$ _____ by _____
2. $\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^3 \xrightarrow{p}$ _____ by _____
3. $\frac{\sqrt{n}(\bar{Y}_n - \mu)}{(\bar{Y}_n)^5 + \mu^5} \xrightarrow{d}$ _____ by _____

[8]Inspired by Problem 2 in "Stat 111 Homework 3, Spring 2025" by Joseph K. Blitzstein and Neil Shephard.

Solution

10 Confidence Intervals

10.1 Constructing Confidence Intervals

Definition 48 (Confidence interval)**.** $C(\vec{Y})$ is a $100(1-\alpha)\%$ confidence interval (CI) for θ if it has coverage probability of at least $(1-\alpha)$ for all possible values of θ (i.e., $P(\theta \in C(\vec{Y})) \geq 1-\alpha \ \forall \ \theta \in \Theta$).

Definition 49 (Statistic)**.** A random variable that is a function of the data $\vec{Y}$. Usually denoted as $T(\vec{Y})$, where the function T must not involve any unknown parameters.

• I.e., the statistic/random variable itself cannot involve unknown parameters, but its distribution may.

• E.g., $\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Definition 50 (Pivot)**.** A random variable whose exact distribution is known.

• I.e., the pivot/random variable itself may involve unknown parameters, but its distribution cannot.

• E.g., $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

• An asymptotic pivot is a random variable whose asymptotic distribution is known. We can "freeze" at a sufficiently large n to construct an approximate/nominal confidence interval.

• **Strategy**: Find a pivot (exact or asymptotic) that involves the estimand of interest, write a probabilistic statement using quantiles, and isolate the estimand. Finding the pivot usually involves properties of distributions (e.g., shifting and scaling).

Concept Checker 13. Which of the following random variables is not a pivot? Assume n is known.

1. $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$
2. $U \sim \text{Unif}(0, \theta)$
3. $\frac{1}{\theta}U \sim \text{Unif}(0,1)$

Solution

Concept Checker 14. For any random variable X whose CDF F is continuous and strictly increasing (i.e., F^{-1} exists), what does $P(Q(\frac{\alpha}{2}) \leq X \leq Q(1-\frac{\alpha}{2}))$ equal?

Solution

Example 4 (Exact pivot, Normal data, one parameter unknown)**.** For $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with μ unknown and σ^2 known, $\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ by property of the Normal. This is an exact distribution. By standardizing, we have an exact pivot: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} = Z \sim \mathcal{N}(0,1)$. Suppose we want a $100(1-\alpha)\%$ confidence interval for μ.

$$
\begin{aligned}
1-\alpha &= P\left(Q_Z(\frac{\alpha}{2}) \leq \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq Q_Z(1-\frac{\alpha}{2})\right) && \text{by quantiles} \\
&= P\left((\frac{\sigma}{\sqrt{n}})Q_Z(\frac{\alpha}{2}) \leq \bar{Y}-\mu \leq (\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2})\right) && \text{by algebra} \\
&= P\left((\frac{\sigma}{\sqrt{n}})Q_Z(\frac{\alpha}{2}) - \bar{Y} \leq -\mu \leq (\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2}) - \bar{Y}\right) && \text{by algebra} \\
&= P\left(-((\frac{\sigma}{\sqrt{n}})Q_Z(\frac{\alpha}{2}) - \bar{Y}) \geq \mu \geq -((\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2}) - \bar{Y})\right) && \text{by algebra} \\
&= P\left(\bar{Y} - (\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2}) \leq \mu \leq \bar{Y} - (\frac{\sigma}{\sqrt{n}})Q_Z(\frac{\alpha}{2})\right) && \text{by simplifying}
\end{aligned}
$$

Thus, our $100(1-\alpha)\%$ confidence interval for μ is $\left[\bar{Y} - (\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2}), \bar{Y} - (\frac{\sigma}{\sqrt{n}})Q_Z(\frac{\alpha}{2})\right]$. By the symmetry of the Normal (i.e., $Q_Z(\frac{\alpha}{2}) = -Q_Z(1-\frac{\alpha}{2})$), we can rewrite this as $\bar{Y} \pm (\frac{\sigma}{\sqrt{n}})Q_Z(1-\frac{\alpha}{2})$.

10.2 Other Setups

Let $\mu = \mathbb{E}[Y_i]$ and $\sigma^2 = \text{Var}[Y_i]$. Use the following estimators: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \xrightarrow{p} \sigma^2$ and $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \xrightarrow{p} \sigma$.

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

Concept Checker 15. Let $Y_i \overset{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$, with λ unknown. Unfortunately, we do not live in Asymptopia, so n is not large enough for the CLT to apply. First, what is a pivot we can use? Next, what is a 95% confidence interval for λ?

Solution

11 Predictive Regression

11.1 Fundamentals

Definition 51 (Predictive regression)**.** The task of estimating the conditional expectation of the outcome given the predictors: $\mu(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}]$, where $\mu(\vec{x})$ is the (theoretical) prediction function.

Definition 52 (Regression error)**.** The difference between the random outcome and the theoretical prediction as a function of the predictors: $U(\vec{x}) = Y - \mathbb{E}[Y \mid \vec{X} = \vec{x}]$.

• This implies $\sigma^2(\vec{x}) = \text{Var}[U(\vec{x}) \mid \vec{X} = \vec{x}]$.

• ☣: The notation is a bit misleading since the regression error $U(\vec{X})$ isn't perfectly deterministic given $\vec{X} = \vec{x}$, so we can't "take out what's known" when conditioning on $\vec{X}$.

Definition 53 (Homoskedastic)**.** Let $\sigma^2(\vec{x}) = \text{Var}[Y \mid \vec{X} = \vec{x}]$. If $\sigma^2(\vec{x})$ does not vary with $\vec{x}$ (i.e., if the conditional variance $\sigma^2(\vec{x})$ is some constant σ^2), then the regression error is homoskedastic. Otherwise, it is heteroskedastic.

• This implies $\sigma^2(\vec{x}) = \text{Var}[U(\vec{x}) \mid \vec{X} = \vec{x}]$.

Definition 54 (Signal-noise decomposition)**.** By rearranging, we can decompose Y into the signal (i.e., the theoretical prediction $\mu(\vec{x})$) and the noise (i.e., the regression error $U(\vec{x})$): $Y = \mu(\vec{x}) + U(\vec{x})$.

Definition 55 (Properties of regression error)**.** As a random variable, regression error has an expectation of 0 (conditionally and unconditionally) and a covariance of 0 with each predictor: $\mathbb{E}[U(\vec{X}) \mid \vec{X} = \vec{x}] = 0$, $\mathbb{E}[U(\vec{X})] = 0$, and $\text{Cov}[U(\vec{X}), X_j] = 0 \ \forall \ X_j$.

Definition 56 (Variance of outcome)**.** Conditionally, we already defined $\text{Var}[Y \mid \vec{X} = \vec{x}] = \sigma^2(\vec{x})$. Unconditionally, $\text{Var}[Y] = \text{Var}[\mu(\vec{X})] + \text{Var}[U(\vec{X})]$.

Definition 57 (R^2 statistic)**.** The share of the variation in Y accounted for by the variation in the theoretical prediction: $R^2 = \frac{\text{Var}[\mu(\vec{X})]}{\text{Var}[Y]} = 1 - \frac{\text{Var}[U(\vec{X})]}{\text{Var}[Y]}$.

• If R^2 is close to 1, then very little variation in the outcome is due to the regression error (i.e., random noise), so the model explains the data well.

Definition 58 (Linear regression model)**.** A model where the conditional expectation of the outcome is a linear function of the parameters (i.e., "linear in the parameters," not necessarily in the predictors): $\mu(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}] = \theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K$, where $\vec{\theta} = (\theta_0, \ldots, \theta_K)^\top$ is the vector of parameters/regression coefficients.

• In order to predict Y given $\vec{X} = \vec{x}$, we must estimate $\vec{\theta}$ with $\hat{\vec{\theta}}$. Once we do, our estimator (before $\vec{X}$ crystallizes) is $\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K$.

Definition 59 (Residual)**.** The difference between the true outcome and the predicted outcome: $\hat{U}(\vec{X}) = Y - \hat{Y} = Y - (\hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K)$.

• Residuals are orthogonal to the predictors, so in the no-intercept simple linear regression model, $\sum_{i=1}^n X_i \hat{U}_i(X_i) = 0$.

• ☣: Regression error is unobservable (as a theoretical quantity) while residual is observable (as a statistic). Don't mix up the two!

Concept Checker 16. Which of the following models is not linear in their parameters?

1. $Y_i = \theta X_i + U_i$
2. $Y_i = \theta_0 + \theta_1 X_i^{\theta_2} + U_i$

3. $Y_i = \theta_0 + \theta_1 X_{i,1} + \theta_2 X_{i,2} + \theta_3 X_{i,1} X_{i,2} + U_i$
4. $Y_i = \theta \sin(X_i) + U_i$

Solution

11.2 No-Intercept Simple Linear Regression Model

Definition 60 (Model)**.** We model observation i's response with no intercept (i.e., with one parameter), one predictor, and homoskedastic error: $Y_i = \theta X_i + U_i$.

• We estimate with $\hat{\theta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$, which is the OLS and MOM estimator. We have $\mathbb{E}[\hat{\theta} \mid \vec{X} = \vec{x}] = \theta$ and $\text{Var}[\hat{\theta} \mid \vec{X} = \vec{x}] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$.

• If $Y_i \mid X_i = x_i \overset{\text{ind.}}{\sim} \mathcal{N}(\theta x_i, \sigma^2)$, with θ and σ^2 unknown, then $\hat{\theta}$ is also the MLE and $\hat{\theta} \mid \vec{X} = \vec{x} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$.

11.3 Extensions to Other Setups

Setup	Model	Estimator
No-intercept simple linear regression	$Y_i = \theta X_i + U_i$	$\hat{\theta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$
Simple linear regression	$Y_i = \theta_0 + \theta_1 X_i + U_i$	$\hat{\theta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$, $\hat{\theta}_0 = \bar{Y} - \hat{\theta}_1 \bar{X}$
Matrix form with K predictors	$\vec{Y} = \mathbf{X}\vec{\theta} + \vec{U}$	$\hat{\vec{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \vec{Y}$

Concept Checker 17. First, show $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Next, use this to show $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) y_i$. Finally, use this to rewrite $\hat{\theta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ in the simple linear regression setup.[9]

[9]Inspired by Problem 4 in "Stat 111 Homework 5, Spring 2026" by Joseph K. Blitzstein and Neil Shephard.

Solution

11.4 Logistic Regression

Definition 61 (Logit function)**.** $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ for $0 < p < 1$, where $\frac{p}{1-p}$ is the odds.

- $\text{logit} : (0,1) \to \mathbb{R}$.

Definition 62 (Logistic function)**.** $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$ for $x \in \mathbb{R}$.

- $\text{logit}^{-1} : \mathbb{R} \to (0,1)$.

Definition 63 (Logistic regression model)**.** Let Y be binary and $\mu(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}] = P(Y = 1 \mid \vec{X} = \vec{x})$. Logistic regression models the logit of the conditional expectation of the outcome as a linear function of the parameters: $\text{logit}(\mu(\vec{x})) = \log\left(\frac{\mu(\vec{x})}{1-\mu(\vec{x})}\right) = \theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K$.

- By applying logit^{-1} to both sides, we have $\mu(\vec{x}) = \text{logit}^{-1}(\theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K) = \frac{\exp(\theta_0+\theta_1 x_1+\ldots+\theta_K x_K)}{1+\exp(\theta_0+\theta_1 x_1+\ldots+\theta_K x_K)}$, so our estimator (before $\vec{X}$ crystallizes) is $\hat{\mu}(\vec{X}) = \text{logit}^{-1}(\hat{\theta}_0 + \hat{\theta}_1 X_1 + \ldots + \hat{\theta}_K X_K) = \frac{\exp\left(\hat{\theta}_0+\hat{\theta}_1 X_1+\ldots+\hat{\theta}_K X_K\right)}{1+\exp\left(\hat{\theta}_0+\hat{\theta}_1 X_1+\ldots+\hat{\theta}_K X_K\right)}$.

12 General Problem-Solving Strategies

Strategy 1. Ask what you have and what you want to find. See if you can relate these with formulas, pattern-matching, etc.

Concept Checker 18. Let $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$. We can show $\hat{\theta}_{\text{MLE}} = Y_{(n)}$. What is the distribution of $\frac{1}{\theta}\hat{\theta}_{\text{MLE}}$?[10]

Solution

Strategy 2. Ask if you can cite existing info.

Concept Checker 19. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables, with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma^2 < \infty$ unknown. Suppose we use $\hat{\sigma}_{\text{MOM}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$. What is the asymptotic distribution of $\hat{\sigma}_{\text{MOM}}$?[11]

Solution

Strategy 3. Consider defining new quantities based on complicated expressions.

Concept Checker 20. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables, with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma^2 < \infty$ unknown. Assume $\text{Var}[Y_i^{111}] < \infty$. What does $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^{111} - \mathbb{E}[Y_i^{111}]\right)$ converge to?

Solution

[10] Inspired by Problem 3 in "Stat 111 Homework 4, Spring 2026" by Joseph K. Blitzstein and Neil Shephard.
[11] Inspired by Problem 1 in "Stat 111 Homework 3, Spring 2026" by Joseph K. Blitzstein and Neil Shephard.

13 Practice Problems

Problem 1. Suppose SAT score is distributed $\mathcal{N}(1000, 200^2)$.

(a) Find the score that would put someone in the top 1% of the distribution. Derive the answer in terms of Φ^{-1}, the inverse CDF of a Standard Normal.

(b) Use $\Phi^{-1}(0.99) \approx 2.326$ to derive an approximation.

Solution

Problem 2. Let $Y_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a, \lambda)$, with a known and λ unknown.

(a) Find the method of moments estimator: $\hat{\lambda}_{\text{MOM}}$.

(b) **Challenge**: Now suppose both a and λ are unknown. Find the method of moments estimators: $\hat{a}_{\text{MOM}}, \hat{\lambda}_{\text{MOM}}$.

Hint: There are two unknown parameters, so write out the first two moments.

Solution

Problem 3. Let $Y_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, with p unknown. We already showed $\hat{p}_{\text{MLE}} = \bar{Y}$.

(a) Find the MSE of $\hat{p}_{\text{MLE}}$.

(b) Is $\hat{p}_{\text{MLE}}$ a consistent estimator for p?

Solution