Section 1: Statistics

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.[1]

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).

1.3 Notation

Unless noted otherwise, we use the following conventions:

- Capital letters for random variables and events. E.g., $X \sim \text{Bern}(p)$, $P(A) = 0.5$.
- Lower-case letters for crystallizations (i.e., realized values). E.g., $P(X = x) = p^x(1-x)^{1-x}$.
- $\mathbb{R}$ for the set of real numbers, $\forall$ for "for all," and $\exists$ for "there exists." E.g., $\forall\ x \in \mathbb{R}$, $\exists\ y \in \mathbb{R}$ such that $y > x$.
- Double right arrows for implications. E.g., $x + 2 = 5 \implies x = 3$.
- $\log(x)$ for the natural logarithm of x. E.g., $\log(e) = 1$.
- $F(x)$ for CDF and $f(x)$ for PDF or PMF for X continuous or discrete, respectively. E.g., for $X \sim \text{Bern}(p)$, $f(x) = P(X = x) = p^x(1-p)^{1-x}$ while for $X \sim \text{Expo}(\lambda)$, $f(x) = \lambda e^{-\lambda x}$.
- $\mathbb{1}\{A\}$ for indicators. E.g., $\mathbb{E}[\mathbb{1}\{A\}] = P(A)$.
- {expression: rule} for sets. E.g., $\{x \in \mathbb{N} : x > 1\} = \{2, 3, 4, \ldots\}$.
- $\lceil x \rceil$ as the ceiling of x (rounded up to the nearest integer). E.g., $\lceil 3.14 \rceil = 4$ (analogously, $\lfloor x \rfloor$ is the floor of x).

1.4 Join GUSH!

The Group for Undergraduates in Statistics at Harvard (GUSH) hosts events throughout the school year, open to all students! Join the mailing list at http://www.gushclub.org.

[1] Roughly, we will follow the structure of *Introduction to Statistics: Inference, Description, Prediction, and Causality* by Joseph K. Blitzstein and Neil Shephard.

2 Big Picture

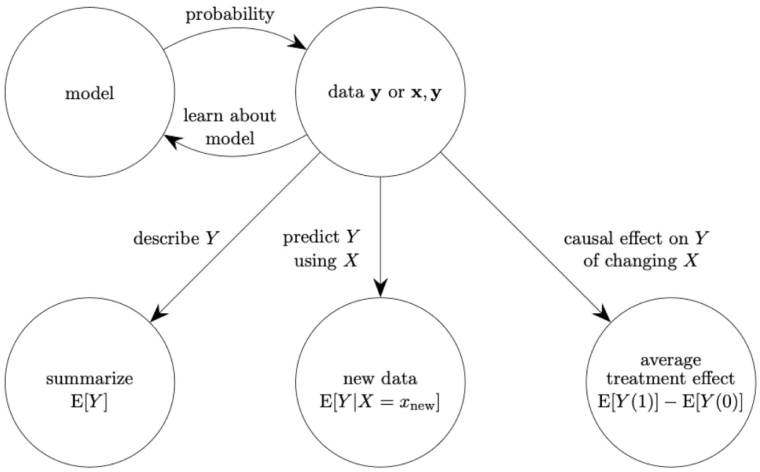

FIGURE 1: A roadmap of statistics.[2]

Statistics is the "other side" of probability, aimed at quantifying uncertainty in the world. Whereas in *probability* (i.e., STAT 110), we'd be given a fully-specified distribution and asked to calculate certain probabilities—e.g., if $X \sim \text{Expo}(\lambda = 4)$, what is $P(X \leq \frac{1}{4})$?—in *statistics* (i.e., STAT 111), we'll be given a model along with realizations as data and asked to estimate unknown parameters—e.g., if $X \sim \text{Expo}(\lambda)$ and we observe $X_1 = \frac{1}{3}$, $X_2 = \frac{1}{4}$, $X_3 = \frac{1}{5}$, how can we best estimate λ? We estimate these unknown *estimands* using *estimators*, which are often denoted with hats—e.g., we can estimate $\lambda = \frac{1}{\mathbb{E}[X_1]}$ with $\hat{\lambda} = \frac{1}{\bar{X}}$.

Broadly, the three goals of statistics are *describing* data—e.g., what is the expected salary in all U.S. adults?—*predicting* a variable from others—e.g., what is the expected salary for an adult who is 40 years old?—and drawing *causal* conclusions—e.g., what is the expected causal effect of attending college on salary? These goals involve unknown quantities, which we must *infer* about! We can use either *model-based inference* or *design-based inference* (or a mix of both). Similarly, we can approach inference through the *Frequentist* paradigm or the *Bayesian* paradigm (or a mix of both).

3 Mathematics

Idea. Mathematics is the *language* of probability and statistics. So that we're all on the same page, let's review some important concepts.

Definition 1 (Sum of squares identity)**.** $\sum_{i=1}^{n}(Y_i - c)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + n(\bar{Y} - c)^2$ for constant c.

Definition 2 (Square of a sum)**.** $\left(\sum_{i=1}^{n} Y_i\right)^2 = \sum_{i=1}^{n} Y_i^2 + \sum_{i \neq j} Y_i Y_j$, where $\sum_{i \neq j} Y_i Y_j = 2\sum_{i<j} Y_i Y_j$.

[2]This figure is from "Section #1: Introduction" by Danielle Paulson.

Definition 3 (Inverse of a function)**.** For a function $f(x) = y$, the inverse is $f^{-1}(y) = x$ such that $f^{-1}(f(x)) = f(f^{-1}(x)) = x$.

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

Concept Checker 1. Use the sum of squares identity to rewrite $\sum_{i=1}^{n} Y_i^2$.

Solution

Concept Checker 2. Let $f(x) = \log(x) + 5$. What is $f^{-1}(x)$?

Solution

4 Quantities

Idea. In statistics, there are many quantities we're interested in, often used to describe a random phenomenon. We distinguish between *theoretical quantities* (which are unknown) and *sample quantities* (which can be calculated with our limited sample data).

4.1 Important Quantities

Definition 4 (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 5 (Cumulative distribution function)**.** For a random variable Y, its cumulative distribution function (CDF) is $F(y) = P(Y \leq y)$.[3]

[3]Importantly, every random variable has a CDF, but only continuous random variables have a PDF. Thus, we often use CDF.

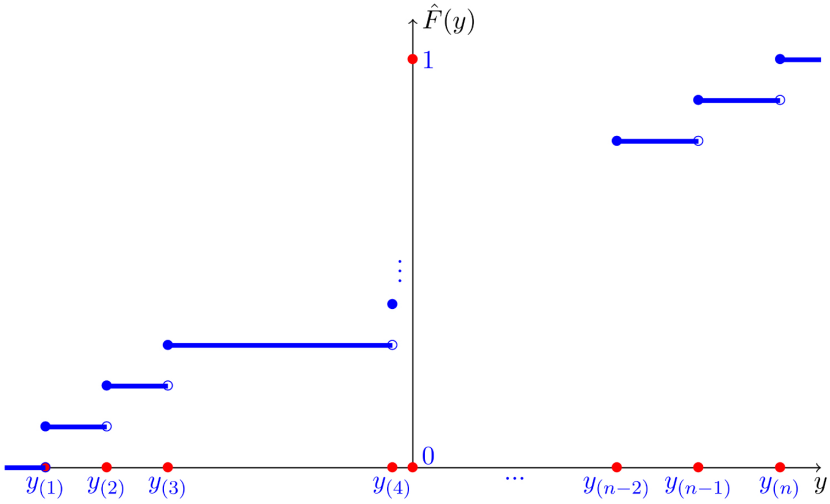

FIGURE 3: An empirical CDF.[4]

Definition 6 (Empirical CDF)**.** For data $\vec{Y} = (Y_1, \ldots, Y_n)$, the empirical CDF (ECDF) is $\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y\}$.

• As illustrated in Figure 3, the ECDF is always a step function, jumping every time it reaches one of the data points.

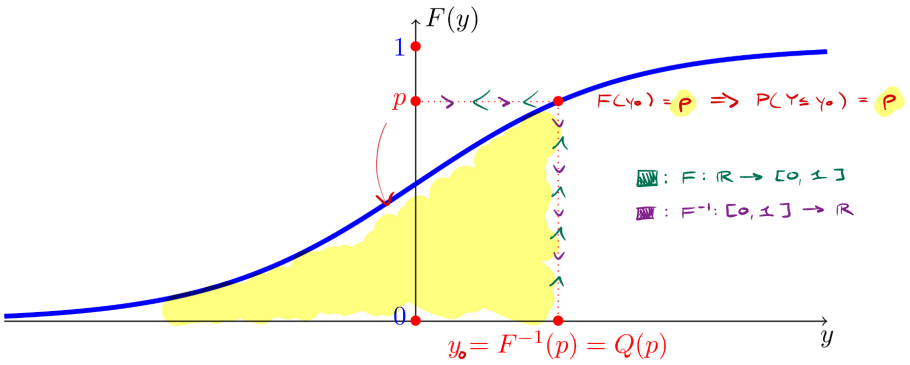

FIGURE 4: The relationship between CDF and quantile function.[5]

Definition 7 (Quantile function)**.** 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 $F^{-1} : [0, 1] \to \mathbb{R}$.

[4]This figure is from *Introduction to Statistics: Inference, Description, Prediction, and Causality* by Joseph K. Blitzstein and Neil Shephard.

[5]This figure is from *Introduction to Statistics: Inference, Description, Prediction, and Causality* by Joseph K. Blitzstein and Neil Shephard.

• 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 $y = 1000$.

• Notice in Figure 4 the relationship between CDF and quantile function.

Definition 8 (Sample 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 largest data point). Notice 50% of the data lies to the left of $y = Y_{(6)}$.

Concept Checker 3. 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	$F(y) = P(Y \leq y)$	$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i \leq y\}$
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	$P(Y = y)$	$\hat{P}(Y = y) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i = y\}$

5 Inference

Idea. In statistics, we want to *infer* about unknown quantities (i.e., *estimands*). The course begins with *model-based inference*, followed by *design-based inference*. Also, the course begins with the *Frequentist* paradigm, followed by the *Bayesian* paradigm. Regardless of the approach we use, the basic vocabulary stays the same.

5.1 Estimands, Estimators, and Estimates

Definition 9 (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 10 (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. $\Theta = [0, 24]$.

Definition 11 (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 12 (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.[6] They can be calculated with our limited sample data.

• Using the same "hat" notation, we may sometimes see the sample mean $\bar{Y}$ denotes as $\hat{\mu}$ to emphasize it's an estimator for the theoretical mean μ.[7]

Concept Checker 4. Let the data be $Y_1, \ldots, Y_n \sim \mathcal{N}(\theta, 1)$. Which of the following quantities are valid estimators 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. $\hat{\theta}_4 = 3$

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

[7] 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 (we will learn about the pros and cons of different estimators later in the course).

Solution

5.2 Model-Based vs. Design-Based Inference

Definition 13 (Model-based inference)**.** We use a probability distribution as a model for a phenomenon of interest. This distribution has a parameter (or multiple), which we aim to learn more about.

• E.g., we can model the amount of emails received in an hour as $Y \sim \text{Pois}(\lambda)$. Thus, we need to find what λ is!

Definition 14 (Design-based inference)**.** Rather than from a model, we sample from a specific, finite population of size N. We need to infer about the estimand because we don't have all the information needed.

• E.g., we want to know the average hours of sleep Harvard students get per night. With $N = 8000$, our estimand is $\frac{1}{N}\sum_{i=1}^{N} y_i$, but we only observe a random sample of size n.

5.3 Models

Definition 15 (Statistical model)**.** A statistical model views the data $\vec{y}$ as realizations of the random variables $\vec{Y} = (Y_1, \ldots, Y_n)$. The model constitutes a collection of joint distributions for $\vec{Y}$. Usually denoted as $\{F_{\vec{Y}}(\vec{y}) : F_{\vec{Y}}(\vec{y}) \text{ is a joint distribution on } \mathbb{R}^n\}$.

Definition 16 (Parametric model)**.** A statistical model where the collection of distributions is indexed by a finite-dimensional parameter θ. Usually denoted as $\{F_{\vec{Y};\theta}(\vec{y}) : \theta \in \Theta\}$.[8]

• I.e., we assume the data come from a specific family, so we only need to estimate a finite amount of numbers.

• E.g., we can model our data as $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, with p unknown. This is a one-dimensional model with $\theta = p$ and $\Theta = [0, 1]$.

Definition 17 (Nonparametric model)**.** A model where the collection of distributions is indexed by an infinite-dimensional parameter θ.

• I.e., we don't assume a shape at all—the entire distribution is unknown, so the "parameter" is the function itself.

• E.g., we can model our data as $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F$, with F as an arbitrary distribution function on $\mathbb{R}$. Thus, we make no parametric assumptions about the form of F. The parameter is $\theta = F$, which is an infinite-dimensional function (one degree of freedom for each real number).

Concept Checker 5. Assume we use a $\mathcal{N}(\mu, \sigma^2)$ model, where both parameters are unknown. What is θ and Θ?

[8] $F_{\vec{Y};\theta}(\vec{y})$ is sometimes written as $F_{\vec{Y}}(\vec{y};\theta)$ or $F_{\vec{Y}}(\vec{y} \mid \theta)$.

Solution

5.4 Frequentist vs. Bayesian

Definition 18 (Frequentist paradigm)**.** We regard θ as a fixed but unknown value.

• E.g., let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, with p unknown. Now, p could be $0, 0.1, 0.01, \ldots$, but at the end of the day, it's just a number.

Definition 19 (Bayesian paradigm)**.** We model θ as a random variable with a prior distribution. After observing the data, we use that information to update the posterior distribution.

• E.g., let $Y_1, \ldots, Y_n \mid p \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. With no information, we can let the prior be $p \sim \text{Unif}(0, 1)$. If we observe many successes (i.e., $Y_i = 1$), the posterior may look like $p \mid \vec{Y} \sim \text{Beta}(8, 4)$ (i.e., left-skewed).

6 Practice Problems

Problem 1. An important result is the expectation and variance of the sample mean. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma^2 < \infty$. Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ be the sample mean.

(a) Find $\mathbb{E}[\bar{Y}]$.

(b) Find $\text{Var}[\bar{Y}]$.

(c) Suppose the data are i.i.d. Normal. I.e., $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Do we know the distribution of $\bar{Y}$ in this case?

(d) Now suppose the data are i.i.d. Bernoulli. I.e., $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. How can we rewrite $\mathbb{E}[\bar{Y}]$ and $\text{Var}[\bar{Y}]$ in terms of p? Do we know the distribution of $\bar{Y}$ in this case?

(e) Now suppose the data are identically distributed Bernoulli but NOT independent. I.e., $Y_1, \ldots, Y_n \sim \text{Bern}(p)$. What changes about $\mathbb{E}[\bar{Y}]$ and $\text{Var}[\bar{Y}]$ in this case?

(f) **Challenge**: Consider a clinical trial, where a unit is either treated or not ($Y_i = 1$ if unit i is treated, $Y_i = 0$ else). As opposed to Bernoulli randomization, in a complete randomization, out of the n units, exactly n_1 are treated while n_0 are not, with n_1 chosen prior. Thus, $\forall\ i$, $P(Y_i = 1) = \frac{n_1}{n}$. Here, the data are identically distributed Bernoulli but NOT independent. I.e., $Y_1, \ldots, Y_n \sim \text{Bern}(\frac{n_1}{n})$. Find $\text{Var}[\bar{Y}]$ in this case.

Hint: Build on the variance derivation without $\text{Cov}[Y_i, Y_j] = 0$. How can we rewrite covariance, and how can we relate our given probabilities to expectations?

Solution

Problem 2. 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 3. The Weibull distribution is a generalization of the Exponential. I.e., for $T \sim \text{Expo}(\lambda)$, $Y = T^{\frac{1}{\gamma}} \sim \text{Weibull}(\lambda, \gamma)$, with γ known and λ unknown. Let $\gamma > 0$. (Different parameterizations of the Weibull exist, so be careful if looking at other sources.)[9]

(a) Find the CDF of Y: $F_Y(y)$.

(b) Find the quantile function of Y: $Q(p)$.

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

Solution