Stat 111 Notes

Elvin Lo

Spring 2024

2024.5.19

Preface

These notes follow the Stat 111 Book by Blitzstein and Shephard, the text accompanying STAT 111 at Harvard College.

Contents

1 Introduction

1.1 Overview

1.2 The big picture

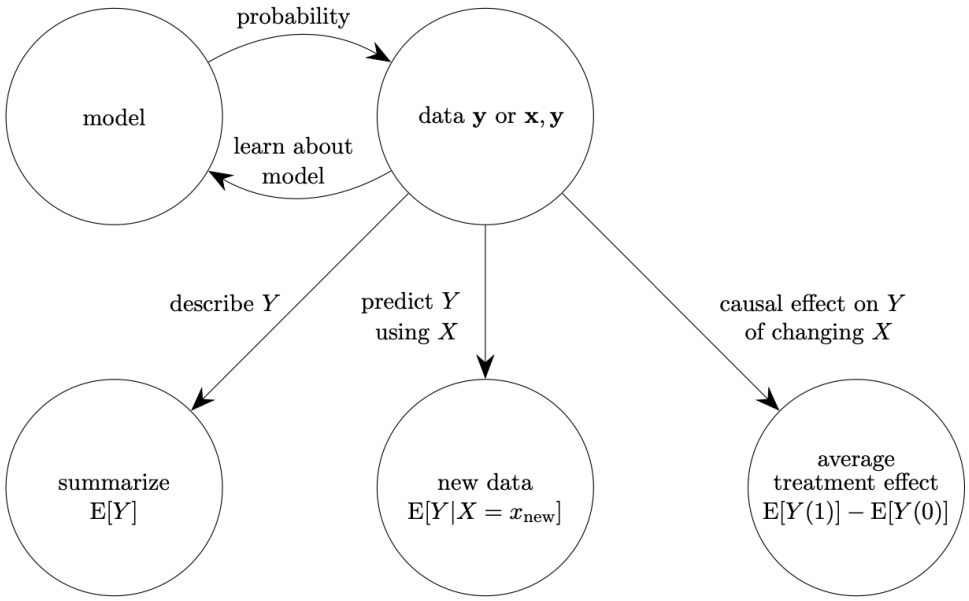

Figure 1.1: Roadmap of the relationships between some of the most fundamental concepts in statistics.

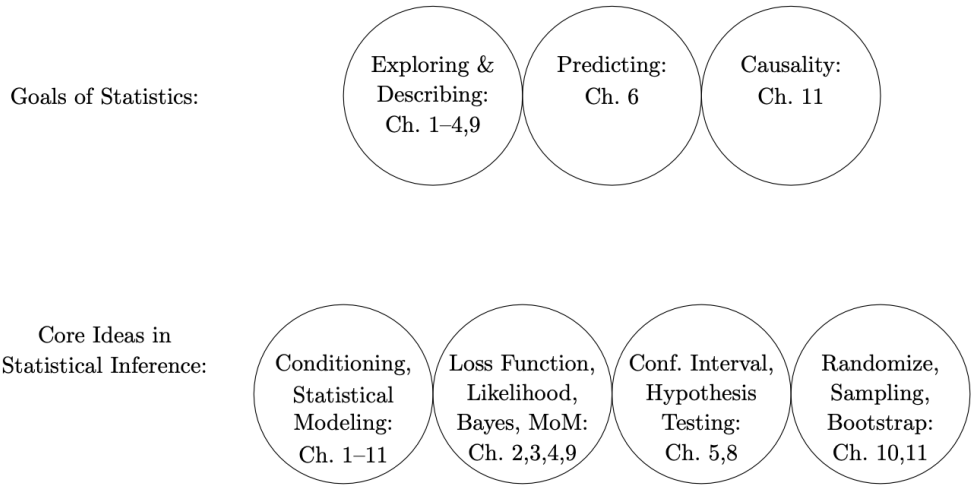

Figure 1.2: The core goals of statistics together with the main tools developed in statistical inference supporting them.

Notation 1.1.
We use capital letters for r.v.s and lowercase letters for the observed data corresponding to those r.v.s. Because we are discussing data in statistics, we generally use the letter Y when we have only one set of r.v.s.

Definition 1.2. Fundamental statistics tasks
Let $Y_1, \ldots, Y_n$ be the r.v.s that will "crystallize" into the observed values $y_1, \ldots, y_n$.

- A *statistical model* is a collection of possible joint distributions for $Y_1, \ldots, Y_n$, possibly indexed by some parameter θ.
- A statistic is a function of the data.
- An *estimand* is a particular quantity that we wish to learn, e.g., the model parameters or the mean of a random variable.
- Based on the model, *probability* lets us determine how likely various events are and what the typical values of our random variables are.
- *Exploring* provides interpretable summaries of the data, either visually or numerically.
- *Describing* goes in the reverse direction to probability, addressing the fact that θ is typically unknown. We can consider questions such as (1) how should we estimate θ given some data, (2) how should we estimate some probability like $P(Y_1 = 0)$, (3) how confident should we be about our estimates, (4) how good are our model's assumptions, and so on. Note that various strategies for estimating an estimand are possible.
- *Prediction* considers how to use observed data to predict not yet observed data.
- *Causality* asks what the effect on a variable will be if we intervene to change another variable.

1.3 Learning and deciding: frequentist and Bayesian inference

Definition 1.3. Learning and deciding
There are two major kinds of tasks for inference of unknown quantities: learning and deciding. Let θ be an estimand, that is an unknown quantity of interest we wish to learn.

1. Learning about θ.
   (a) *Point estimation*: we choose an estimator $\hat{\theta}$, as a function of $Y_1, \ldots, Y_n$. Ideally, we will be able to prove that $\hat{\theta}$ will be close to θ with high probability.
   (b) *Interval estimation*: provide an interval that contains θ with high probability.
2. Deciding about θ. In some applications, the goal is to make a decision,e.g., is it plausible or not that $\theta = 0$, or which of two rival hypotheses $\theta = \theta_0$ and $\theta = \theta_1$ should we choose.

Definition 1.4.
Learning and deciding can be approached from either frequentist or Bayesian perspectives.

- The *frequentist approach* focuses on coming up with procedures that work well in the long run; this requires considering drawing new datasets over and over again.
- The *Bayesian approach* focuses on the data at hand. We model θ as a random variable with some prior distribution, and then use Bayes' rule to update our probabilities for θ based on the data. Once we have the posterior distribution, we might estimate θ using the posterior mean, median, or even create an interval that has, say, 95% chance of containing θ, given the data.

1.4 Exploring and describing Y

We discuss some useful summaries of a dataset $y_1, \ldots, y_n$ (aside from visualizations like histograms and scatter plots).

Definition 1.5. Sample mean, sample standard deviation
The sample mean of $y_1, \ldots, y_n$ is

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j.$$

The sample standard deviation s is the square root of the sample variance

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

Definition 1.6. Sample covariance, sample correlation, linear regression
The sample covariance is

$$s_{x,y} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x}) (y_j - \bar{y})$$

and the sample correlation is

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}.$$

The linear regression of y on x is

$$b_{y \sim x} = \frac{s_{x,y}}{s_x^2},$$

where s_x and s_y are the sample standard deviations of $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, respectively.

Definition 1.7. Order statistics
The order statistics of $y_1, y_2, \ldots, y_n$ are the same data points, sorted in increasing order:

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}.$$

Some examples of quantities based on order statistics are the sample minimum $y_{(1)}$, the sample maximum $y_{(n)}$, the range $y_{(n)} - y_{(1)}$, and the sample median $y_{((n+1)/2)}$ (if n is odd; there are different conventions about what to do if n is even).

Definition 1.8. Quantile function
Let F be a CDF. The quantile function of F is defined by

$$F^{-1}(p) = Q(p) = \min\{y : F(y) \geq p\}.$$

The value $Q(p)$ is called the p-quantile of the distribution. For a random variable Y, we will use Q_Y to denote the quantile function of Y. Note that for a general CDF F (that may not be continuous or invertible), we have

$$F\left(F^{-1}(p)\right) \geq p.$$

Definition 1.9. Sample quantile
The corresponding sample quantity is the p-sample quantile, defined to be a value such that approximately proportion p of the sample is less than or equal to that value. The p-sample quantile of the dataset $y_1, \ldots, y_n$ is the order statistic $y_{(\lceil np \rceil)}$, denoted by $\hat{Q}(p)$:

$$\hat{Q}(p) = y_{(\lceil np \rceil)}.$$

There are different conventions for the sample quantile, though if n is large then it is unlikely to matter which convention is used. We will use a simple convention.

Definition 1.10. Empirical CDF
The empirical CDF (ECDF) of the data set is the CDF of a r.v. obtained by choosing one of the n data points $y_1, \ldots, y_n$ uniformly at random,

$$\hat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} I\left(y_j \leq y\right).$$

Note that the ECDF is always a step function, jumping every time it reaches one of the data points.

Note that the ECDF converges to the true CDF as the sample size grows. If we have a model under which the data are realizations of i.i.d. r.v.s $Y_1, \ldots, Y_n$ with CDF F, then the strong LLN implies that for each $y \in \mathbb{R}$, we have with probability 1 that

$$\lim_{n \to \infty} \hat{F}(y) = E\left[I\left(Y_1 \leq y\right)\right] = P\left(Y_1 \leq y\right) = F(y).$$

1.5 Predicting Y from X

Remark 1.11. What is prediction?
Think of it is as: having seen some data X, what are the likely values of Y? *Forecasting* is a special case of prediction where X is an aspect of the past and Y is an aspect of the future.

Definition 1.12. Regression models
In a regression model, we have predictor variables $X_1, \ldots, X_k$ and an outcome variable Y, and we try to use the predictor variables to predict the outcome variable. We can think of

$$E\left[Y \mid X_1, \ldots, X_k\right]$$

as the best prediction of Y as a function of $X_1, \ldots, X_k$.

1.6 Causal impact on Y of manipulating X

To help differentiate prediction and causality, let us establish some notation for thinking about causal effects.

In causal studies, X is called the assignment and Y is the outcome. As a simple example, consider a binary treatment with binary outcomes. Then we have $X \in \{0, 1\}$ where the event $X = 1$ is receiving the treatment and the event $X = 0$ is the control, and we might write the outcome Y as 1 for a success and 0 otherwise.

Definition 1.13. Potential outcomes, treatment effect
The pair $\{Y(0), Y(1)\}$ are called the potential outcomes. The random variable

$$\tau = Y(1) - Y(0)$$

is the treatment effect (or causal effect) on the outcome of moving a person from control to treatment. Across a population, moving everyone from control to treatment, the

$$\mathrm{E}[\tau] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$$

is the population's average treatment effect.

Definition 1.14. Counterfactual
A major challenge is that even after the study is over we never see both $Y(0)$ and $Y(1)$: we only see one of them as the individual is either under treatment or control, not both. So we cannot compute τ directly, which makes it tricky to estimate $\mathrm{E}[\tau]$.

To infer $\mathrm{E}[\tau]$, we have the assignment X and an outcome, mathematically written as

$$\begin{aligned} Y &= \begin{cases} Y(1), & \text{if } X = 1 \\ Y(0), & \text{if } X = 0 \end{cases} \\ &= Y(X) \\ &= XY(1) + (1 - X)Y(0). \end{aligned}$$

The potential outcome we do not see, $Y(1 - X)$, is called a counterfactual. It is impossible to observe a counterfactual.

2 Models, Likelihood, Estimation, and Method of Moments

2.1 Statistical models

Definition 2.1. Statistical model
A statistical model views $\mathbf{y}$ as a realization of the random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from their joint cumulative distribution function (CDF) $F_{\mathbf{Y}}$. The model specifies a collection of possibilities for $F_{\mathbf{Y}}$. Before making the observation, we have a random vector $\mathbf{Y}$. After making the observation, $\mathbf{Y}$ crystallizes into the data $\mathbf{y}$. We say that the model generated the data, and often we want to use the data to learn about the model.

Definition 2.2. Estimand
An estimand is an aspect of $F_{\mathbf{Y}}$ that we wish to learn about from the data that we will observe.

Estimands are often denoted by Greek letters.

Definition 2.3. Parametric model
A parametric statistical model is a family of probability distributions for $\mathbf{Y}$, indexed by a finite-dimensional parameter θ. The distributions in a model are usually specified by their joint CDFs or by their joint densities (joint PMFs in the discrete case, joint PDFs in the continuous case). The parameter space, denoted by Θ, is the set of all allowable values of θ. Thus each $\theta \in \Theta$ picks out a single probability distribution for $\mathbf{Y}$.

If in the above definition we instead allow θ to be infinite-dimensional, then we have a non-parametric model.

Notation 2.4. CDF notation
To make explicit the fact that the CDF in a parametric model depends on θ, we sometimes write the CDF as $F_{\mathbf{Y};\theta}$. Then the CDF evaluated at $\mathbf{y}$ is denoted by $F_{\mathbf{Y};\theta}(\mathbf{y})$ or $F_{\mathbf{Y}}(\mathbf{y};\theta)$. If θ is being modeled as a random variable, as in the Bayesian approach, then ; is replaced by the conditioning bar | in the notation, yielding $F_{\mathbf{Y}|\theta}$ for the CDF (which technically is, from a Bayesian point of view, the conditional CDF of $\mathbf{Y}$ given θ). Then the CDF evaluated at $\mathbf{y}$ is written as $F_{\mathbf{Y}|\theta}(\mathbf{y})$ or $F_{\mathbf{Y}}(\mathbf{y} \mid \theta)$.

Example 2.5. Data $Y_1, \ldots, Y_n$ are i.i.d.
Often it is scientifically plausible to assume that $Y_1, \ldots, Y_n$ are independent and identically distributed. Under the i.i.d. assumption,

$$Y_j \overset{\text{i.i.d.}}{\sim} F_{Y_1;\theta}, \quad j = 1, \ldots, n,$$

where $F_{Y_1;\theta}$ is the CDF of each Y_j. By independence, the joint CDF is the product of the marginal CDFs:

$$F_{\mathbf{Y}}(\mathbf{y};\theta) = \prod_{j=1}^{n} F_{Y_1}(y_j;\theta).$$

2.2 Likelihood

Definition 2.6. Likelihood function
Let $\mathbf{y}$ be the observed value of $\mathbf{Y}$. The function given by

$$L(\theta; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \theta),$$

regarded as a function of the parameter, with the data held constant, is called the likelihood function. That is, the likelihood function is the probability or probability density of the data given the parameters, as a function of the parameters. So the likelihood function is a function of θ, with $\mathbf{y}$ treated as fixed.

Notationally, it is conventional to separate θ and y by a semicolon: $L(\theta; \mathbf{y})$. Often we even write $L(\theta)$ for the likelihood function, leaving the $\mathbf{y}$ implicit, to simplify the notation further and to emphasize that the likelihood function is regarded as a function of θ. Also note that two likelihood functions are viewed as equivalent if one is a positive constant times the other. In fact, the "constant" can even be a function of the data (it just cannot depend on the parameter)!

Remark 2.7. Bayesian vs frequentist perspectives
Bayesian perspective: In a Bayesian approach, we have a prior density $\pi(\theta)$ for θ, and use Bayes' rule to obtain the posterior density:

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

where the proportionality stems from the fact that we are treating $\mathbf{y}$ as fixed (so the denominator $f(\mathbf{y})$, which is the marginal density of $\mathbf{y}$, is viewed as a constant). That is, Bayes' rule says, in words: The posterior is proportional to likelihood times prior.
So the two key ingredients for a Bayesian analysis are the prior distribution $\pi(\theta)$ and the likelihood function $L(\theta; \mathbf{y})$. Combining the likelihood and the prior, we obtain the posterior distribution, which we then base our inferences on.

Frequentist perspective: In a frequentist approach, θ does not have a posterior distribution, but we can use the likelihood function as a surrogate for assessing how plausible various possible values of θ are. One of the most widely used estimation techniques in statistics is maximum likelihood estimation, which says to estimate θ using

$$\hat{\theta} = \arg\max_{\theta} L(\theta; \mathbf{y}),$$

the parameter value that maximizes the likelihood function. This value is called the maximum likelihood estimate (MLE).

Definition 2.8. Log-likelihood
It is very common when working with likelihood to work with the log-likelihood.

$$L(\theta; \mathbf{y}) = \log L(\theta; \mathbf{y}).$$

One reason is numerical stability, since extremely small probabilities often come up in likelihood calculations. But also, the log lets us consider sums rather than products, and since log is a continuous, strictly increasing function, maximizing the likelihood is equivalent to maximizing the log-likelihood (for when we study MLE later). If the Y_j are independent, then the likelihood function is

$$L(\theta; \mathbf{y}) = \prod_{j=1}^{n} f_{Y_j}(y_j; \theta)$$

and the log-likelihood function is

$$l(\theta; \mathbf{y}) = \sum_{j=1}^{n} \log f_{Y_j}(y_j; \theta).$$

Theorem 2.9. Invariance of likelihood under transformation of the parameter
The likelihood function is unchanged under reparameterization, in the following sense. Consider a likelihood function $L(\theta; \mathbf{y})$ and let $\psi = g(\theta)$ be a reparameterization, where g is a known one-to-one function. Then

$$L(\psi; \mathbf{y}) = L(\theta; \mathbf{y}).$$

Theorem 2.10. Invariance of likelihood under transformation of the data
Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the observed data, coming from a model with parameter θ. Let h be a known one-to-one function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Use h to transform the data, letting $\mathbf{x} = h(\mathbf{y})$. Then taking the dataset to be $\mathbf{x}$ rather than y has no effect on the likelihood function:

$$L(\theta; \mathbf{x}) = L(\theta; \mathbf{y})$$

Proof. For simplicity, we will only write the proof in the case of a single observation y from a continuous distribution, with h a differentiable, strictly increasing function. Let Y be the r.v. that "crystallizes" to $y, x = h(y)$, and $X = h(Y)$. By the change of variables formula,

$$L(\theta; x) = f_X(x; \theta) = f_Y(y; \theta)\frac{1}{h'(y)}.$$

But $\frac{1}{h'(y)}$ is a multiplicative "constant" (not depending on θ), so it can be dropped. Then we can take the likelihood function for θ, based on the data x, to be

$$L(\theta; x) = f_Y(y; \theta) = L(\theta; y)$$

□

2.3 Statistics, estimators, and estimates

Definition 2.11. Statistic
A statistic is a function of $Y_1, \ldots, Y_n$ (and possibly other known quantities). We can write a statistic as $T(\mathbf{Y})$, where computing the function T must not require knowledge of any unknown parameters.

Definition 2.12. Estimator
Suppose that we use the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ to construct a statistic

$$\widehat{\theta} = T(\mathbf{Y})$$

with the intention that this statistic should estimate an estimand θ. The statistic $\widehat{\theta}$ is called an estimator.

Definition 2.13. Bias
The bias of an estimator $\widehat{\theta}$ for θ is

$$\text{bias}(\widehat{\theta}) = \text{E}(\widehat{\theta}) - \theta.$$

We say that $\widehat{\theta}$ is unbiased for θ if its bias is 0, i.e., its expected value is θ.
To compute the bias, recall that by LOTUS,

$$\text{E}[\widehat{\theta}] = \int T(\mathbf{y}) f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

if $\mathbf{Y}$ is continuous, and

$$\text{E}[\widehat{\theta}] = \sum_{\mathbf{y}} T(\mathbf{y}) P(\mathbf{Y} = \mathbf{y}; \theta),$$

if $\mathbf{Y}$ is discrete, where the integral and sum are over the support of $\mathbf{Y}$. Typically, the bias depends on θ (so we may be able to compute the bias theoretically but may not know the actual value of the bias due to θ being unknown).

Definition 2.14. Standard error
The standard error of an estimator $\hat{\theta}$ for θ is its standard deviation:

$$\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

This is a measure of how variable the estimator is.

Lemma 2.15. Sum of squares identity
For any random variables $Y_1, \ldots, Y_n$ and any constant c,

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

Definition 2.16. Estimate
An estimate is a realization of an estimator. So if our data $\mathbf{y}$ is a realization of $\mathbf{Y}$ and $T(\mathbf{Y})$ is an estimator of some estimand θ, then $T(\mathbf{y})$ is an estimate of θ.

2.4 Sample moments and method of moments

Definition 2.17. Method of moments
Consider the following strategy, the method of moments principle, for obtaining an estimator. Express the estimand in terms of moments, and then replace the estimand with the estimator and the theoretical moments with the sample moments. An estimator obtained in this way is called a method of moments estimator (MoM).

3 Loss Functions, Bias-Variance Tradeoff, and Asymptotics

3.1 Bias and variance of sample p-quantiles

Theorem 3.1. Asymptotic distribution of sample quantile
If $Y_1, \ldots, Y_n$ be i.i.d. Uniform random variables on $[0,1]$ and $p \in (0,1)$. Then as $n \to \infty$,

$$\sqrt{n}\left\{Y_{(\lceil np \rceil)} - Q_{Y_1}(p)\right\} \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Proof. The proof of this theorem uses the change of variables formula and Taylor approximations. The details of the proof get pretty technical, so the proof is in the starred Section 3.8. □

3.2 Bias and variance are sometimes in conflict

Definition 3.2. Kernel density estimator
Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. with a CDF $F_{Y_1}(y)$ and PDF $f_{Y_1}(y)$. Let the estimand be $\theta = f_{Y_1}(y)$ for some value of y, and let $h > 0$. Let

$$\hat{\theta} = \frac{1}{hn} \sum_{j=1}^{n} I\left(Y_j \in (y - h/2, y + h/2]\right).$$

The estimator $\hat{\theta}$ is called the kernel density estimator (KDE) of $f_{Y_1}(y)$, with rectangular kernel. The number h is called a bandwidth. More generally, let K be a nonnegative function, called the kernel function. Then the kernel density estimator for this choice of K is

$$\hat{\theta} = \frac{1}{hn} \sum_{j=1}^{n} K\left(\frac{Y_j - y}{h}\right).$$

A widely used kernel aside from the rectangular kernel is to let K be the standard Normal PDF. In this book, by default when discussing KDE we will assume we are using the rectangular kernel. In terms of the ECDF, we can write the KDE as

$$\hat{\theta} = \frac{\hat{F}(y + h/2) - \hat{F}(y - h/2)}{h}.$$

Taking the limit as $h \to 0$ on the right-hand side would yield the definition of the derivative of $\hat{F}$ at y. As noted earlier, the derivative of the ECDF is not useful, but here we fix h rather than letting $h \to 0$. Thus, the KDE can be thought of as a way to get an approximate notion of "slope" for the ECDF without actually taking the derivative.

3.3 Loss functions, risk, and mean square error

Definition 3.3. Loss function
A loss function is a function

$$\text{Loss}(\theta, \hat{\theta})$$

interpreted as the loss or cost associated with using the estimate $\hat{\theta}$ when the estimand is θ. We require $\text{Loss}(\theta, \hat{\theta}) \geq 0$ and $\text{Loss}(\theta, \theta) = 0$.

We take a frequentist approach for now, treating θ as an unknown constant (not having a distribution). The estimator $\hat{\theta} = T(\mathbf{Y})$ is, of course, a random variable. So $\text{Loss}(\theta, \hat{\theta})$ is also a random variable, as it is a function of $\hat{\theta}$ (for each fixed θ).

Definition 3.4. Risk function
For the estimator $\hat{\theta} = T(\mathbf{Y})$, the risk function is the expected loss

$$\text{Risk}(\theta) = \text{E}_\theta[\text{Loss}(\theta, \hat{\theta})] = \int \text{Loss}(\theta, T(\mathbf{y})) f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

where the integral is over the support of $\mathbf{Y}$.

Definition 3.5. MSE
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

is called the squared error loss. Its expected value is the mean square error (MSE) of $\hat{\theta}$ is:

$$\text{MSE}(\hat{\theta}) = \text{E}_\theta \left[(\hat{\theta} - \theta)^2 \right].$$

Sometimes the square root of the MSE is used instead, in order to have the units be the same as that of θ; this is called the root mean square error (RMSE).

Definition 3.6. MAE
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$$

is called the absolute error loss. Its expected value is the mean absolute error (MAE) of $\hat{\theta}$ is

$$\text{MAE}(\hat{\theta}) = \text{E}_\theta[|\hat{\theta} - \theta|].$$

Absolute error loss punishes small errors more severely and large errors more gently than the RMSE. Which one of the two makes more sense in practice depends, of course, on the application.

Definition 3.7. $0-1$ **loss**
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = I(\hat{\theta} \neq \theta)$$

is called $0-1$ loss. Note that its expected value is $\text{E}_\theta[\text{Loss}(\theta, \hat{\theta})] = P_\theta(\hat{\theta} \neq \theta)$.

Definition 3.8. Check loss
The loss function

$$\text{Loss}(\theta, \hat{\theta}) = |\hat{\theta} - \theta| \left\{ c_+ I(\hat{\theta} > \theta) + c_- I(\hat{\theta} < \theta) \right\}, \quad c_+, c_- \geq 0,$$

is called the check loss. Unlike the previous few, this last loss function is not symmetric.

3.4 Bias-Variance tradeoff

Theorem 3.9. Bias-variance tradeoff
The mean square error of an estimator $\hat{\theta}$ for the estimand θ is the variance plus the square of the bias:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2.$$

Proof. Let $V = \hat{\theta} - \theta$. Then

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \text{E}_\theta \left[V^2 \right] \\ &= \text{Var}_\theta(V) + (\text{E}_\theta[V])^2 \\ &= \text{Var}(\hat{\theta}) + \{\text{E}(\hat{\theta}) - \theta\}^2 \\ &= \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2. \end{aligned}$$

□

3.5 Consistency of estimators

Definition 3.10. Consistency
An estimator $\hat{\theta}$ is consistent for the estimand θ if $\hat{\theta}$ converges in probability to θ as the sample size $n \to \infty$, i.e., for every $\epsilon > 0$ we have

$$P(|\hat{\theta} - \theta| \geq \epsilon) \to 0$$

as $n \to \infty$. This is written in shorthand as

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

Note that implicitly $\hat{\theta}$ depends on n. Sometimes it is clearer to write the dependence explicitly, with notation such as $\hat{\theta}_n$ that explicitly indicates the sample size n.

Theorem 3.11. Sufficient condition for consistency
If $\hat{\theta}$ is an estimator for the estimand θ and $\text{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is consistent. In particular, since MSE is variance plus squared bias, to show that $\hat{\theta}$ is consistent it suffices to show that both the bias and the variance go to 0 as $n \to \infty$.

Proof. Suppose that the MSE of $\hat{\theta}$ goes to 0 . Then by Markov's inequality, for any $\epsilon > 0$ we have

$$P(|\hat{\theta} - \theta| \geq \epsilon) = P\left((\hat{\theta} - \theta)^2 \geq \epsilon^2\right) \leq \frac{\mathrm{E}\left[(\hat{\theta} - \theta)^2\right]}{\epsilon^2} = \frac{\text{MSE}(\hat{\theta})}{\epsilon^2} \to 0$$

□

Remark 3.12.
It may be tempting to believe that $\hat{\theta} \xrightarrow{p} \theta$ implies that $\mathrm{E}[\hat{\theta}] \to \theta$, since if two r.v.s are very likely to be very close to each other, then it may seem intuitively that their means should also be close. But this implication is false. For a counterexample, let the data be realizations of independent r.v.s $U, Y_1, \ldots, Y_n$ where $U \sim \text{Unif}(0, 1)$ and $Y_j \sim \text{Bern}(\theta)$. Let

$$\hat{\theta} = \bar{Y} + nI(U \leq 1/n).$$

Then $\hat{\theta} \xrightarrow{p} \theta$, since $\bar{Y} \xrightarrow{p} \theta$ and for n large, the second term in $\hat{\theta}$ is very likely to be 0 . But

$$\mathrm{E}[\hat{\theta}] = \mathrm{E}[\bar{Y}] + nP(U \leq 1/n) = \theta + 1$$

Example 3.13.
As we mentioned earlier, many estimators have $\text{bias}(\hat{\theta}) \approx b/n$ and $\text{Var}(\hat{\theta}) \approx a/n$, which would mean that $\text{MSE}(\hat{\theta}) \approx a/n$. Other estimators have slower rates at which $\text{MSE}(\hat{\theta})$ converges to zero as n gets large. Table 3.1 gives some core examples of this for some descriptive statistics.

Statistics	Approximate MSE	Consistent?
sample mean	$\text{Var}(Y_1)/n$	Yes
sample variance	$\text{Var}\left[\{Y_1 - \mathrm{E}[Y_1]\}^2\right]/n$	Yes
sample covariance	$\text{Var}[\{X_1 - \mathrm{E}[X_1]\}\{Y_1 - \mathrm{E}[Y_1]\}]/n$	Yes
sample p-quantile for uniform	$p(1-p)/n$	Yes
kernel density estimator	$d/n^{4/5}$	Yes
Y_1	$\text{Var}(Y_1)$	No
$\frac{1}{2}(Y_1 + Y_2)$	$\text{Var}(Y_1)/2$	No

Theorem 3.14. Continuous mapping theorem
Let $X, X_1, X_2, \ldots$ be a sequence of r.v.s and let g be a continuous function. If

$$X_n \xrightarrow{p} X$$

then

$$g(X_n) \xrightarrow{p} g(X).$$

Also, if

$$X_n \xrightarrow{d} X,$$

then

$$g(X_n) \xrightarrow{d} g(X).$$

In particular, it follows that if $\hat{\theta}$ is a consistent estimator for θ and g is a continuous function, then $g(\hat{\theta})$ is a consistent estimator for $g(\theta)$.

Theorem 3.15. Properties of convergence in probability
Let $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then

$$X_n + Y_n \xrightarrow{p} X + Y,$$
$$X_n - Y_n \xrightarrow{p} X - Y,$$
$$X_n Y_n \xrightarrow{p} XY,$$

and, if $P(Y_n = 0) = P(Y = 0) = 0$,

$$X_n / Y_n \xrightarrow{p} X/Y.$$

3.6 Large sample (asymptotic) approximations

Definition 3.16. Convergence in distribution
Let $X_1, X_2, \ldots$ be a sequence of random variables and F_{X_n} be the CDF of X_n. Let X be a random variable with CDF F_X. Then X_n converges in distribution to the random variable if

$$F_{X_n}(x) \to F_X(x),$$

for all $x \in \mathbb{R}$ such that F_X is continuous at x. This is written in shorthand as $X_n \xrightarrow{d} X$.

Theorem 3.17.
Let $X_n \xrightarrow{p} X$. Then $X_n \xrightarrow{d} X$. The converse is false in general. But if X is a constant c (i.e., X is a degenerate r.v. that always equals c), then $X_n \xrightarrow{p} X$ is equivalent to $X_n \xrightarrow{d} X$.

Remark 3.18.
Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. It does not follow that $X_n + Y_n \xrightarrow{d} X + Y$. As a simple counterexample, let $X_n = Y_n \sim \mathcal{N}(0,1)$ and X, Y be i.i.d. $\mathcal{N}(0,1)$. Then $X_n + Y_n = 2X_n \sim \mathcal{N}(0,4)$, whereas $X + Y \sim \mathcal{N}(0,2)$. Clearly, the $\mathcal{N}(0,4)$ distribution does not converge to $\mathcal{N}(0,2)$. The problem is that the statement $X_n \xrightarrow{d} X$ is about the marginal distributions of X_n and X, and similarly for the statement $Y_n \xrightarrow{d} Y$, whereas the distribution of $X_n + Y_n$ depends heavily on the joint distribution of X_n and Y_n.

Theorem 3.19. Slutsky's Theorem
If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then

- $X_n + Y_n \xrightarrow{d} X + c$;
- $X_n - Y_n \xrightarrow{d} X - c$;
- $X_n Y_n \xrightarrow{d} cX$;
- $X_n / Y_n \xrightarrow{d} X/c$, if $c \neq 0$.

Theorem 3.20. Delta method
Suppose that g is a differentiable function and

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

Then

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

As an approximation, this says that

$$g(\hat{\theta}) \dot{\sim} \left(g(\theta), \left(g'(\theta)\right)^2 \frac{\omega^2}{n}\right),$$

for n large.

Proof. If n is large, then $\hat{\theta}$ is close to θ (with high probability). Taylor expand $g(\hat{\theta})$ about θ, yielding the approximation

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

as the higher order terms should be smaller as they involve squares, cubes, etc. of $\{\hat{\theta} - \theta\}$ which is going to zero at rate $n^{-1/2}$. Rearranging,

$$\begin{aligned}\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} &\approx g'(\theta)\sqrt{n}\{\hat{\theta} - \theta\} \\ &\xrightarrow{d} g'(\theta)\omega Z,\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. □

3.7 Multivariate asymptotic approximations*

3.8 A couple of technical proofs*

3.9 Concentration inequalities*

4 Maximum Likelihood Estimation

4.1 Defining and finding the maximum likelihood estimate (MLE)

Definition 4.1. Maximum likelihood estimator
The maximum likelihood estimate (MLE) of θ is the value $\hat{\theta}$ that maximizes the likelihood function $L(\theta; \mathbf{y})$. Mathematically, this is written as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{y})$$

4.2 Properties of the MLE

Some of the main properties that the MLE enjoys are as follows. All of these require some technical assumptions known as regularity conditions. Under these assumptions, which we will discuss more later, we have the following.

- The MLE is invariant, which means that if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
- The MLE $\hat{\theta}$ is consistent, which means that it converges in probability to the true θ.
- The MLE is asymptotically Normal (so its distribution is approximately Normal if the sample size is large).
- The MLE is asymptotically unbiased (the bias approaches 0 as the sample size grows).
- The MLE is asymptotically efficient (no other asymptotically unbiased estimator will have a lower standard error asymptotically).

Theorem 4.2. Invariance of MLE
Let $\hat{\theta}$ be the MLE of θ, and let g be a one-to-one function. Then the MLE of $g(\theta)$ is $g(\hat{\theta})$.

Proof. This result follows from the invariance property of likelihood. Let $\tau = g(\theta)$. Each point on the reparameterized likelihood curve $L(\tau)$ has a corresponding point on the original likelihood function $L(\theta)$, such that the likelihood value for τ is the same as the likelihood value for the corresponding θ. In particular, the value $\hat{\tau}$ that maximizes $L(\tau)$ is $\hat{\tau} = g(\hat{\theta})$. □

Definition 4.3. MLE under a parameter transformation that is not one-to-one
Invariance of the MLE is so convenient, in fact, that we define it to be true even when g is not one-to-one. If $\hat{\theta}$ is the MLE of θ and g is not a one-to-one function, then we define the MLE of $g(\theta)$ to be $g(\hat{\theta})$.

4.3 Kullback-Leibler divergence

Notation 4.4.
Let θ^* be the estimand, such that the random variables $Y_1, \ldots, Y_n$ are generated by the joint CDF $F_{\mathbf{Y};\theta^*}$. The hope, of course, is that our estimators will be close to θ^*. The distribution $F_{\mathbf{Y};\theta^*}$ is called the data generating process. To summarize our notation:

- $\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta; \mathbf{Y})$ is an estimator,
- θ is the argument in the likelihood function $L(\theta; \mathbf{Y})$, and
- θ^* is the true value or estimand, generating $\mathbf{Y}$ through $F_{\mathbf{Y};\theta^*}$.

Definition 4.5. KL divergence
The Kullback-Leibler divergence (KL divergence) from the CDF F to the CDF G is

$$D_{\mathrm{KL}}(F\|G) = \mathrm{E}\left(\log \frac{f(\mathbf{Y})}{g(\mathbf{Y})}\right) = \mathrm{E}[\log f(\mathbf{Y}) - \log g(\mathbf{Y})] = \int \{\log f(\mathbf{y}) - \log g(\mathbf{y})\} f(\mathbf{y}) d\mathbf{y},$$

where f and g are the PDFs corresponding to F and G (in the case of discrete distributions, we replace the PDFs with PMFs). The expectations are computed with $\mathbf{Y}$ generated according to F, not G.

An important case of KL divergence is when $F = F_{\mathbf{Y};\theta^*}$ and $F = F_{\mathbf{Y};\theta}$. Then

$$D_{\mathrm{KL}}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right) = \mathrm{E}\left[\log L\left(\theta^*; \mathbf{Y}\right) - \log L(\theta; \mathbf{Y})\right].$$

Intuitively, $D_{\mathrm{KL}}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right)$ measures how much the expected log-likelihood is higher at θ^* than at θ, computing the expectation under the true distribution $F_{\mathbf{Y};\theta^*}$.

Theorem 4.6. Additivity of KL divergence
If we observe independent $Y_1, \ldots, Y_n$, then

$$D_{KL}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right) = \sum_{j=1}^{n} D_{KL}\left(F_{Y_j;\theta^*} || F_{Y_j;\theta}\right),$$

where $D_{KL}\left(F_{Y_j;\theta^*} \| F_{Y_j;\theta}\right)$ is the Kullback-Leibler divergence for the j th observation:

$$D_{KL}\left(F_{Y_j;\theta^*} \| F_{Y_j;\theta}\right) = \mathrm{E}\left(\log \frac{L\left(\theta^*; Y_j\right)}{L\left(\theta; Y_j\right)}\right)$$

In the i.i.d. case,

$$D_{KL}\left(F_{\mathbf{Y};\theta^*} \| F_{\mathbf{Y};\theta}\right) = n D_{KL}\left(F_{Y_1;\theta^*} \| F_{Y_1;\theta}\right).$$

Theorem 4.7. Nonnegativity of KL divergence
For any CDFs F and G,

$$D_{KL}(F\|G) \geq 0.$$

This inequality is strict unless $F = G$, that is they are the same distribution functions.

Theorem 4.8. Consistency of MLE
Suppose that the parameter space is finite and that the observations $Y_1, \ldots, Y_n$ are i.i.d. Also assume that for $\theta_1 \neq \theta_2$, the distribution function $F_{Y;\theta_1}$ is different from the distribution of $F_{Y;\theta_2}$ (this is known as identifiability of the model). Then the MLE $\hat{\theta}$ is consistent:

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

as the sample size $n \to \infty$.

Score function

Definition 4.9. Score function
The score function is

$$s(\theta; \mathbf{y}) = \frac{\partial \log L(\theta; \mathbf{y})}{\partial \theta} = \frac{1}{L(\theta; \mathbf{y})} \frac{\partial L(\theta; \mathbf{y})}{\partial \theta}$$

Theorem 4.10. Information equality
Under some regularity conditions (mainly that the $f_{\mathbf{Y}}(\mathbf{y}; \theta)$ is a smooth function in θ, the support of $\mathbf{Y}$ does not depend on θ, that the expected values needed below exist, and that we can differentiate under the integral sign when needed below),

$$\begin{aligned} \mathrm{E}\left[s\left(\theta^*; \mathbf{Y}\right)\right] &= 0, \\ \mathrm{Var}\left\{s\left(\theta^*; \mathbf{Y}\right)\right\} &= -\mathrm{E}\left[s'\left(\theta^*; \mathbf{Y}\right)\right]. \end{aligned}$$

The prime in s' denotes taking the partial derivative with respect to θ, that is $s'\left(\theta^*; \mathbf{Y}\right) = \partial s\left(\theta^*; \mathbf{Y}\right) / \partial \theta$.

Fisher information

Definition 4.11. Fisher information
The Fisher information in the sample for a parameter θ in a parametric statistical model $F_{\mathbf{Y};\theta}$ is

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \mathrm{Var}\left\{s(\theta^*;\mathbf{Y})\right\} = \mathrm{E}\left[s(\theta^*;\mathbf{Y})^2\right],$$

where we compute the variance under the assumption that the true parameter value is θ. Let

$$\mathcal{J}_{\mathbf{Y}}(\theta^*) = -\mathrm{E}\left[s'(\theta^*;\mathbf{Y})\right].$$

Then

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \mathcal{J}_{\mathbf{Y}}(\theta^*),$$

the information equality. In statistics it is traditional to suppress θ^* and write $\mathcal{I}_{\mathbf{Y}}(\theta)$ and $\mathcal{J}_{\mathbf{Y}}(\theta)$, without the stars, implicitly understanding the role of θ^*.

It is not obvious from looking at the definition why Fisher information is a measure of information. Some intuition for this can be gleaned from thinking about the curvature of $\mathrm{E}\log L(\theta;\mathbf{Y})$. If the expected log-likelihood function has a sharp peak at θ^*, the data can be very informative about θ. If the $\mathrm{E}\log L(\theta;\mathbf{Y})$ is quite flat at θ^*, the data do not seem to be giving us much information that we can use for pinpointing the true parameter value.

Definition 4.12. Fisher information when transforming the parameter
Let $\tau = g(\theta)$, where g is a differentiable function with $g'(\theta) \neq 0$. Then

$$\mathcal{I}_{\mathbf{Y}}(\tau) = \frac{\mathcal{I}_{\mathbf{Y}}(\theta)}{\{g'(\theta)\}^2}.$$

Cramér-Rao lower bound

Theorem 4.13. CRLB
Let $\hat{\theta}$ be an unbiased estimator of θ in a parametric statistical model $F_{\mathbf{Y};\theta}$. Under regularity conditions,

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_{\mathbf{Y}}(\theta)}.$$

REVIEW PROOFS IN THIS SECTION

Asymptotic distribution of the MLE

Theorem 4.14. Asymptotic distribution of the MLE
In addition to the good properties we have already seen, such as invariance and consistency, the MLE has excellent asymptotic properties: for large sample size, it is approximately the case that the MLE is Normal, unbiased, and achieves the CRLB. Again we use the θ^* notation at the start, to make the exposition clear.

Let $\hat{\theta}$ be the MLE of a scalar parameter θ, based on i.i.d. observations $Y_1, \ldots, Y_n$ from $F_{\mathbf{Y};\theta^*}$. Under regularity conditions, the asymptotic distribution of $\hat{\theta}$ is given by the following:

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{Y_1}^{-1}(\theta^*)\right),$$

(that is converges in distribution) as the sample size $n \to \infty$. As an approximation, this result says that for large n,

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta^*, \frac{1}{n\mathcal{I}_{Y_1}(\theta^*)}\right)$$

4.4 Likelihoods based on conditional distributions

Suppose the data segments into pieces called $\mathbf{x}$ and $\mathbf{y}$ (e.g., $\mathbf{y}$ are outcomes and $\mathbf{x}$ are predictors), then the statistical model is the joint distribution $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})$. Of course, the joint density equals the marginal density times the conditional density:

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}).$$

Statisticians often decide to study solely the conditional distribution, either out of convenience (we will see an example of this at the end of this section) or because their scientific focus is on the conditional distribution (our next example). In parametric models the statistical model becomes:

$$f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}; \theta).$$

Then we can define the likelihood for this conditional density

$$L(\theta; \mathbf{y} \mid \mathbf{x}) = f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}; \theta).$$

4.5 Numerical optimization of the likelihood*

4.6 Multiple parameter version*

4.7 Estimation when model approximates the truth*

5 Confidence Intervals

5.1 Introduction

Definition 5.1. Interval estimation
An interval estimate $C(\mathbf{y})$ of a scalar estimand θ based on data $\mathbf{y}$ is an interval $[L(\mathbf{y}), U(\mathbf{y})]$, where the lower bound $L(\mathbf{y})$ and upper bound $U(\mathbf{y})$ are functions of the data, such that $L(\mathbf{y}) \leq U(\mathbf{y})$ for all $\mathbf{y}$. The corresponding random interval $[L(\mathbf{Y}), U(\mathbf{Y})]$, where $\mathbf{Y}$ are the random vectors that give rise to the data, is called an interval estimator, written $C(\mathbf{Y})$. Intuitively, the goal is that the probability should be high that $C(\mathbf{Y})$ contains the estimand.

Definition 5.2. Coverage
Let $C(\mathbf{Y})$ be an interval estimator for θ and $C(\mathbf{y})$ be the corresponding interval estimate. If $\theta \in C(\mathbf{y})$, we say that the interval covers θ. The probability of the interval estimator covering θ if θ is the true estimand, $P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y}))$, is called the coverage probability of the interval estimator. Note that the coverage probability is a function of θ.

Definition 5.3. Confidence Interval
Fix a number α with $0 < \alpha < 1$. (In practice, the most common choice is $\alpha = 0.05$.) The interval estimator $C(\mathbf{Y}) = [L(\mathbf{Y}), U(\mathbf{Y})]$ is a $(1-\alpha)$ confidence interval (CI) if it has coverage probability $1-\alpha$ for all possible values of θ:

$$P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1 - \alpha.$$

The constant $1-\alpha$ is called the level of the confidence interval. The half-width $0.5\{U(\mathbf{Y}) - L(\mathbf{Y})\}$ is called the margin of error.

Remark 5.4.
Confidence intervals are widely misinterpreted. For example, if the 95% CI $[0.1, 0.4]$ for θ is calculated from the data, a common mistake is to say that we can be 95% confident that θ is between 0.1 and 0.4 , or that the probability is 0.95 that θ is between 0.1 and 0.4. This is a category error since the statement "θ is between 0.1 and 0.4" is deterministic, either true or false: we are currently working in a frequentist setting, so θ is fixed. It is the interval estimator that is random here, not the estimand.

5.2 Constructing confidence intervals

Example 5.5. Ideal case scenario: a Normal estimator
We wish to create a $1-\alpha$ confidence interval for θ. Suppose that we are in the happy situation where $\hat{\theta}$ is Normal. Specifically, let

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \sigma^2\right),$$

with σ^2 known. A simple but powerful approach is then to standardize $\hat{\theta}$ to get a standard Normal random variable:

$$\frac{\hat{\theta}-\theta}{\sigma} \sim \mathcal{N}(0,1).$$

Then

$$P\left(a \leq \frac{\hat{\theta}-\theta}{\sigma} \leq b\right) = F_{\mathcal{N}(0,1)}(b) - F_{\mathcal{N}(0,1)}(a),$$

and so letting

$$c_p = Q_{\mathcal{N}(0,1)}(1-p),$$

we derive that

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\sigma, \hat{\theta} + c_{\alpha/2}\sigma\right] = \hat{\theta} \pm c_{\alpha/2}\sigma$$

is a $1-\alpha$ CI for θ with margin of error $c_{\alpha/2}\sigma$, centered at the MLE with $c_{\alpha/2}$ standard errors of slack in each direction.

Definition 5.6. Pivot
A pivotal quantity or pivot is a random variable whose distribution is known. In contrast to a statistic,

- A pivot typically depends on unknown parameters but its distribution cannot depend on unknown parameters.
- A statistic cannot depend on unknown parameters but its distribution typically depends on unknown parameters.

For example, suppose Y_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Then $\bar{Y}$ statistic, but

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) \sim \mathcal{N}(0,1)$$

is a pivot.

5.3 Asymptotic approximations

Definition 5.7. Approximate Pivot
Suppose that n is large and, based on asymptotics, we know that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

Then

$$\frac{\hat{\theta} - \theta}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

and the expression on the left-hand side is called an approximate pivot
When we have an approximate pivot, we can obtain an approximate CI by proceeding as in Example 5.2.1 (with approximate equalities in place of equalities). It follows that if σ is known then the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\sigma/\sqrt{n}, \hat{\theta} + c_{\alpha/2}\sigma/\sqrt{n}\right]$$

is an approximate $1 - \alpha$ CI. Usually in practice σ is unknown, but can be estimated with some consistent estimator $\hat{\sigma}$. Then it is natural to use the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\hat{\sigma}/\sqrt{n}, \hat{\theta} + c_{\alpha/2}\hat{\sigma}/\sqrt{n}\right]$$

though it is not obvious what effect plugging in $\hat{\sigma}$ for σ has on the coverage probability for fixed n; this may need to be studied via simulation. Asymptotically this substitution is fine though, since if

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0, 1)$$

and $\hat{\sigma} \xrightarrow{p} \sigma$, then by the continuous mapping theorem and Slutsky's theorem,

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}}\right) = \sqrt{n}\left(\frac{\hat{\theta} - \theta}{\sigma}\right)_{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Remark 5.8.
The asymptotic approach only gives approximate confidence intervals. The mathematical statement is that, if $C(\mathbf{Y})$ is the interval estimator, then

$$\lim_{n\to\infty} P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1 - \alpha$$

as the sample size $n \to \infty$. This does not say, for a fixed n, how close the coverage probabilities are to $1 - \alpha$.

Confusingly, people often say "confidence interval" when they mean "approximate confidence interval". Some so-called $1 - \alpha$ confidence intervals are just aspirational, and in reality the coverage probability is far from $1 - \alpha$ for at least some possible values of θ. In such situations it is clearer to call $1 - \alpha$ the nominal level of the interval estimator. The hope is that the coverage probabilities will be close to $1 - \alpha$ for all θ but this may not be true, or it may be true but not yet demonstrated.

5.4 Pivots with non-Gaussian distributions

6 Regression

6.1 Regression

6.2 Predictive regression

Definition 6.1. Predictive regression
The task of estimating the conditional expectation

$$\mu(\mathbf{x}) = \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

is called predictive regression. The variable Y is called the outcome variable, while the $\mathbf{X}$ variables are called predictors, covariates, regressors, or features. Running a statistical method in this setting is sometimes called regressing Y on $\mathbf{X}$.

Definition 6.2. Homoskedasticity and heteroskedasticity
Assume that

$$\sigma^2(\mathbf{x}) = \mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x})$$

exists. If $\sigma^2(\mathbf{x})$ does not vary with $\mathbf{x}$, then the predictive regression is called homoskedastic. Otherwise it is heteroskedastic.

Definition 6.3. Regression error
For predictive regression, the regression error is the random variable

$$U(\mathbf{x}) = Y - \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

Theorem 6.4. Signal-noise decomposition
From our definitions, we may decompose Y into signal (the predicted part $\mu(\mathbf{x})$) and noise (the random error $U(\mathbf{x})$):

$$Y = \mu(\mathbf{x}) + U(\mathbf{x}).$$

Broadly speaking, a lot of statistical work is about trying to separate signal from noise.

Theorem 6.5. Regression error: mean 0, uncorrelated with predictors
For a random pair $(\mathbf{X}, Y)$, write the regression error (for $\mathbf{X}$ random) as

$$U(\mathbf{X}) = Y - \mathrm{E}[Y \mid \mathbf{X}].$$

Then

$$\mathrm{E}[U(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}] = 0,$$
$$\mathrm{E}[U(\mathbf{X})] = 0,$$

and for each predictor variable X_j,

$$\mathrm{Cov}\left(U(\mathbf{X}), X_j\right) = 0.$$

Proof. By construction,

$$\mathrm{E}[U(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}] = \mathrm{E}[U(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}] = \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}] - \mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}] = 0,$$

for all $\mathbf{x}$. By Adam's law, we also have $\mathrm{E}[U(\mathbf{X})] = 0$ unconditionally. Again by Adam's law, as long as the covariance exists, for each predictor variable X_j we have

$$\mathrm{Cov}\left(U(\mathbf{X}), X_j\right) = \mathrm{E}\left[X_j U(\mathbf{X})\right] = \mathrm{E}\left[\mathrm{E}\left[X_j U(\mathbf{X}) \mid \mathbf{X}\right]\right] = \mathrm{E}\left[X_j \mathrm{E}[U(\mathbf{X}) \mid \mathbf{X}]\right] = \mathrm{E}\left[0 X_j\right] = 0.$$

□

Proposition 6.6. Variance of Y
It is also important to consider the variance of Y, both conditionally and unconditionally. Recall $\sigma^2(\mathbf{x}) = \mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x})$. Then $\sigma^2(\mathbf{x}) = \mathrm{Var}(U \mid \mathbf{X} = \mathbf{x})$, so by Eve's law,

$$\mathrm{Var}(U) = \mathrm{E}[\mathrm{Var}(U \mid \mathbf{X} = \mathbf{x})] + \mathrm{Var}(\mathrm{E}[U \mid \mathbf{X} = \mathbf{x}]) = \mathrm{E}\left[\sigma^2(\mathbf{X})\right],$$

if this variance exists. Likewise, Eve's law says

$$\mathrm{Var}(Y) = \mathrm{E}\left[\sigma^2(\mathbf{X})\right] + \mathrm{Var}(\mu(\mathbf{X})),$$

so a summary measure of the unconditional effectiveness of the prediction is

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

the share of the variation of Y contributed by the variation in the prediction.

Definition 6.7. Linear regression model

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

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_K)^{\mathrm{T}}$. This is called a linear regression model and the elements of $\boldsymbol{\theta}$ are called the regression coefficients. Typically, θ_0 is called the intercept and $\theta_1, \ldots, \theta_K$ are called slopes. Note that linear regression means linear in the parameters θ; the function can be nonlinear in the predictors.

Definition 6.8. Logit function
The logit function is defined by

$$\mathrm{logit}(p) = \log(p/(1-p)),$$

for $0 < p < 1$. The inverse logit function, which is also called the logistic function, sigmoid function, or expit function, is the inverse of the logit function:

$$\mathrm{logit}^{-1}(x) = \frac{e^x}{1 + e^x},$$

for all $x \in \mathbb{R}$.

Definition 6.9. Logistic regression
The logistic regression model assumes that the probability of success, given the predictor variables, is

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) = \mu(\mathbf{x} \mid \boldsymbol{\theta}) = \text{logit}^{-1}\left(\theta_0 + \theta_1 x_1 + \ldots + \theta_K x_K\right)$$

6.3 Statistical models of predictive regression

Gaussian linear regression without intercept

The Gaussian linear regression model assumes the scatter around $\mathrm{E}(Y \mid \mathbf{X} = \mathbf{x}, \boldsymbol{\theta})$ is Gaussian. For now consider a single predictor and no intercept,

$$Y_j \mid (X_1 = x_1, \ldots, X_n = x_n), \theta \overset{\text{ind.}}{\sim} \mathcal{N}\left(\theta x_j, \sigma^2\right).$$

Our parameter has MLE

$$\hat{\theta} = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \left(\sum_{j=1}^{n} x_j y_j\right).$$

Definition 6.10. Residuals
The value $\hat{\theta} x_j$ is called the fitted value or predicted value of Y_j. The difference between the actual value of y_j and the predicted value is called the residual, and denoted by

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

Theorem 6.11. Residuals are orthogonal to predictors
With notation as above, we have

$$\sum_{j=1}^{n} x_j \hat{U}_j = 0.$$

Proof. Note that

$$\sum_{j=1}^{n} x_j^2 \hat{\theta} = \hat{\theta} \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j$$

so

$$\sum_{j=1}^{n} x_j \hat{U}_j = \sum_{j=1}^{n} \left(x_j y_j - x_j^2 \hat{\theta}\right) = \sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} x_j^2 \hat{\theta} = 0.$$

□

Theorem 6.12. Properties of the least squares estimator
Assume that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ have conditionally independent outcomes. Inference will condition the random variables $\mathbf{X}$ at the observed values $\mathbf{x} = (x_1, \ldots, x_n)$. Write $\mu_j = \mathrm{E}[Y_j \mid X_j = x_j]$ and $\sigma_j^2 = \mathrm{Var}(Y_j \mid X_j = x_j)$. Then

$$\mathrm{E}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \left(\sum_{j=1}^{n} x_j \mu_j\right), \quad \mathrm{Var}(\hat{\theta} \mid \mathbf{X} = \mathbf{x}) = \left(\sum_{j=1}^{n} x_j^2\right)^{-2} \left(\sum_{j=1}^{n} x_j^2 \sigma_j^2\right).$$

Proof. Conditioning on the predictors, $\hat{\theta}$ is linear in the outcomes, so (as expectations of sums are sums of expectations)

$$\mathrm{E}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \sum_{j=1}^{n} x_j \mathrm{E}[Y_j \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \left(\sum_{j=1}^{n} x_j \mu_j\right).$$

Conditioning on the predictors $\hat{\theta}$ is linear in the conditionally independent outcomes, so

$$\mathrm{Var}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-2} \sum_{j=1}^{n} x_j^2 \mathrm{Var}[Y_j \mid \mathbf{X} = \mathbf{x}] = \left(\sum_{j=1}^{n} x_j^2\right)^{-2} \left(\sum_{j=1}^{n} x_j^2 \sigma_j^2\right).$$

□

Lemma 6.13. More properties
Properties of Gaussian linear regression model under successively stronger conditions.

(a) (a) If $\mu_j = \theta x_j$, then

$$\mathrm{E}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \theta$$

so then the estimator is conditionally unbiased for θ.

(b) Under homoskedasticity, i.e., $\sigma_j^2 = \sigma^2$,

$$\mathrm{Var}[\hat{\theta} \mid \mathbf{X} = \mathbf{x}] = \sigma^2 \left(\sum_{j=1}^{n} x_j^2\right)^{-1}$$

(c) If $Y_j \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left(\theta X_j, \sigma^2\right)$, then $\hat{\theta}$ is conditionally unbiased for θ and conditionally achieves the Cramér-Rao lower bound

$$\mathrm{Var}(\hat{\theta} \mid \mathbf{X} = \mathbf{x}) = \mathcal{I}(\theta)^{-1}$$

Furthermore,

$$\hat{\theta} \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left(\theta, \sigma^2 \left(\sum_{j=1}^{n} x_j^2\right)^{-1}\right).$$

Proof. Special case of Theorem 6.3.5. □

Gaussian linear regression with intercept

Now consider that we have an intercept, yielding the model

$$Y \mid (X = x, \theta_0, \theta_1) \sim \mathcal{N}\left(\theta_0 + \theta_1 x, \sigma^2\right).$$

We may derive the MLE of the parameters

$$\hat{\theta}_1 = \frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}.$$

6.4 Linear regression, method of moments, and least squares

In the previous section, we studied linear predictive regression using MLE. Of course, other estimation strategies exist. This section discusses the method of moments and least squares strategies; they both yield the same estimator as the MLE.

6.5 Linear projection and descriptive regression

Definition 6.14. Descriptive regression
In comparison to predictive regression which models $\mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$, descriptive regression looks at $(\mathbf{X}, \mathbf{Y})$. The descriptive regression statistic is

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

note that this is not a parametrized statistical model. As discussed below, the descriptive regression appears in a regression approach called linear projection.

Definition 6.15. Linear projection
Assume that the random variables X, Y each have a finite variance. Then the linear projection of Y on X at $X = x$ is defined as

$$\mu_L(x) = \mathrm{E}[Y] + \beta_{Y \sim X}(x - \mathrm{E}[X]).$$

The linear projection $\mu_L(x)$ is not the conditional expectation $\mathrm{E}[Y \mid X = x]$ in general. The conditional expectation $\mathrm{E}[Y \mid X = x]$ is the function of x that best approximates Y, in the sense of minimizing the expected square error; the linear projection $\mu_L(X)$ is the best linear function of x for approximating Y. Writing the linear error as

$$U_L = Y - \mu_L(X),$$

then by construction $\mathrm{E}[U_L] = 0$ and $\mathrm{E}[XU_L] = 0$, due to the derivatives.

6.6 Multiparameter regression*

6.7 Additional regressions*

7 Exponential Families and Sufficiency

7.1 Natural Exponential Families

Definition 7.1. Natural exponential families (NEFs)
A density $f(y;\theta)$ follows a natural exponential family (NEF) if we can write

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

where the nonnegative function h does not depend on θ. The parameter θ is called the natural parameter and may be a reparameterization of how the model was originally specified.
Another way to think of a density of this form is that it factors as a function of y (not involving θ) times a function of θ (not involving y) times a function of both y and θ, where the function of both y and θ takes the simple form $e^{\theta y}$.

Theorem 7.2. Mean and variance in an NEF
Let Y follow the NEF $f(y;\theta) = e^{\theta y - \psi(\theta)} h(y)$. Then

$$\mathrm{E}[Y] = \psi'(\theta), \quad \mathrm{Var}(Y) = \psi''(\theta).$$

Proof. Consider that Y is continuous; the discrete case is analogous. Densities must integrate to 1, so

$$\int_{-\infty}^{\infty} e^{\theta y} h(y) dy = e^{\psi(\theta)}.$$

By DUThIS, differentiating both sides with θ, we have

$$\int_{-\infty}^{\infty} y e^{\theta y} h(y) dy = \psi'(\theta) e^{\psi(\theta)}.$$

Therefore,

$$\psi'(\theta) = \int_{-\infty}^{\infty} y e^{\theta y - \psi(\theta)} h(y) dy = \mathrm{E}[Y],$$

by the definition of expectation. For the variance, we can DUThIS again:

$$\psi''(\theta) = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \left(y e^{\theta y - \psi(\theta)} h(y) \right) dy = \int_{-\infty}^{\infty} y \left(y - \psi'(\theta) \right) e^{\theta y - \psi(\theta)} h(y) dy.$$

Let $\mu = \mathrm{E}[Y] = \psi'(\theta)$. By LOTUS, we then have

$$\psi''(\theta) = \mathrm{E}[Y(Y-\mu)] = \mathrm{E}\left[Y^2\right] - \mu \mathrm{E}[Y] = \mathrm{E}\left[Y^2\right] - \mu^2 = \mathrm{Var}(Y),$$

as desired. □

Theorem 7.3. MLE of an NEF
Suppose we have data $Y_1, \ldots, Y_n$ i.i.d. from the NEF $f(y;\theta) = e^{\theta y - \psi(\theta)} h(y)$. Let $\mu = \mathrm{E}[Y_1] = \psi'(\theta)$ be the mean parameter (it is a reparameterization of θ). Then

- The MLE of μ is its MoM estimator:

$$\hat{\mu} = \bar{Y}$$
$$\hat{\theta} = \left(\psi'\right)^{-1}(\bar{Y})$$

- The Fisher information per observation is

$$\mathcal{I}_{Y_1}(\theta) = \psi''(\theta) = \mathrm{Var}(Y_1)$$
$$\mathcal{I}_{Y_1}(\mu) = \left(\frac{\partial\theta}{\partial\mu}\right)^2 \mathcal{I}_{Y_1}(\theta) = \frac{1}{\psi''(\theta)} = \frac{1}{\mathrm{Var}(Y_1)}$$

- The asymptotic distribution of the MLE is

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \psi''(\theta)^{-1}\right)$$
$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \psi''(\theta)\right).$$

- The MLE of μ achieves the CRLB (with equality, not just asymptotically).

Proof. The log-likelihood function is

$$l(\theta; \mathbf{y}) = n\{\theta\bar{y} - \psi(\theta)\},$$

and the score function is

$$s(\theta; \mathbf{y}) = n\left\{\bar{y} - \psi'(\theta)\right\}.$$

Setting the score equal to 0 , we have that the MLE of θ is as claimed. To check that we have found the maximum, use the second derivative test:

$$\frac{\partial}{\partial\theta} s(\theta; y) = -n\psi''(\theta) < 0,$$

since $\psi''(\theta) = \mathrm{Var}(Y) > 0$. So the log-likelihood function is concave and the Fisher information is as stated in the theorem by the information equality. The MLE is unique as the function ψ' has an inverse since it is continuous (because it is differentiable) and strictly increasing (because $\psi''(\theta) = \mathrm{Var}(Y) > 0$). By invariance, the MLE of $\mu = \psi'(\theta)$ is

$$\hat{\mu} = \psi'(\hat{\theta}) = \psi'\left(\left(\psi'\right)^{-1}(\bar{Y})\right) = \bar{Y}.$$

The asymptotic properties are implied by the standard MLE properties using the Fisher information. To show that $\hat{\mu} = \bar{Y}$ achieves the CRLB, note that $\hat{\mu}$ is unbiased, with

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}(Y_1) = \frac{1}{n\mathcal{I}_{Y_1}(\mu)}.$$

□

Definition 7.4. Exponential family (EF)
A density $f(y;\theta)$ follows an exponential family (EF) if we can write

$$f(y;\theta) = e^{\theta T(y) - \psi(\theta)} g(y),$$

where g does not depend on θ. A generalization of NEFs, an EF is obtained by transforming the variable in a natural exponential family. The difference between an NEF and an EF is that the observation appears as itself in the exponent in an NEF, whereas it appears in some transformed form in the exponent in an EF. If $T(y) = y$ then we recover the definition of an NEF.

7.2 Sufficient statistics

Definition 7.5. Sufficient statistic
For $Y_1, \ldots, Y_n$ from the parametric statistical model $F_{\mathbf{Y};\boldsymbol{\theta}}$, a statistic $T(\mathbf{Y})$ is sufficient for θ if the conditional distribution of

$$(Y_1, \ldots, Y_n) \mid T$$

does not depend on θ.

Here the conditional distribution of $\mathbf{Y}$ given T does not involve θ, so once we know T there is no further statistical information to be gained about θ from looking at the entire vector $(Y_1, \ldots, Y_n)$.

- Sufficient statistics are not unique, for any one-to-one transformation preserves sufficiency.
- A sufficient statistic of the smallest dimension possible is called a minimal sufficient statistic.

Theorem 7.6. Factorization criterion
When computing the conditional distribution $\mathbf{Y}$ given T is difficult, we have a simpler criterion. For θ in $F_{\mathbf{Y};\boldsymbol{\theta}}$, the statistic T is sufficient iff we can factor

$$f(\mathbf{y};\theta) = g_\theta(t)h(\mathbf{y}),$$

where t is the observed value of T and the function h does not depend on θ.

Proof. We will prove the factorization criterion in the discrete case. The continuous case is analogous but more technical to prove. Suppose that T is sufficient. Then

$$f(\mathbf{y};\theta) = P(\mathbf{Y} = \mathbf{y};\theta) = P(\mathbf{Y} = \mathbf{y}, T = t;\theta),$$

since T is a deterministic function of $\mathbf{Y}$. So

$$f(\mathbf{y};\theta) = P(T = t;\theta)P(\mathbf{Y} = \mathbf{y} \mid T = t) = g_\theta(t)h(\mathbf{y}),$$

where $g_\theta(t) = P(T = t;\theta)$ and $h(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y} \mid T = t)$. This is a valid choice for h since it does not depend on θ (since the conditional distribution of $\mathbf{Y}$ given T does not involve θ) and since t is a deterministic function of $\mathbf{y}$. Conversely, suppose that

$$f(\mathbf{y};\theta) = g_\theta(t)h(\mathbf{y}).$$

Let $T = s(\mathbf{Y})$. The conditional PMF of $\mathbf{Y} \mid T$ is

$$P(\mathbf{Y} = \mathbf{y} \mid T = t; \theta) = \frac{P(\mathbf{Y} = \mathbf{y}, T = t; \theta)}{P(T = t; \theta)} = \frac{P(\mathbf{Y} = \mathbf{y}; \theta)}{P(T = t; \theta)}$$

for $t = s(\mathbf{y})$, and 0 otherwise. We need to show that this conditional distribution does not depend on θ. To do so, we can expand the denominator based on all possible values of $\mathbf{Y}$ that are compatible with the observed t :

$$\frac{P(\mathbf{Y} = \mathbf{y}; \theta)}{P(T = t; \theta)} = \frac{P(\mathbf{Y} = \mathbf{y}; \theta)}{\sum_{\tilde{\mathbf{y}}: s(\tilde{\mathbf{y}}) = t} P(\mathbf{Y} = \mathbf{y}; \theta)} = \frac{g_\theta(t) h(\mathbf{y})}{\sum_{\tilde{\mathbf{y}}: s(\tilde{\mathbf{y}}) = t} g_\theta(t) h(\tilde{\mathbf{y}})} = \frac{h(\mathbf{y})}{\sum_{\tilde{\mathbf{y}}: s(\tilde{\mathbf{y}}) = t} h(\tilde{\mathbf{y}})},$$

which does not depend on θ. Thus, T is a sufficient statistic for θ. □

Theorem 7.7. Likelihood function based on a sufficient statistic
Let t be a sufficient statistic for the model $f(\mathbf{y}; \theta)$. Then the likelihood function can be expressed as a function of t (up to a multiplicative constant that does not depend on θ). In particular, knowing the sufficient statistic suffices for knowing the likelihood function, and we can take the $g(\theta; t)$ appearing in the factorization criterion as our likelihood function.

Proof. Note that if t is sufficient then, with notation as in the factorization criterion,

$$\begin{aligned} L(\theta; \mathbf{y}) &= f(\mathbf{y}; \theta) \\ &= g(\theta; t) h(\mathbf{y}), \end{aligned}$$

where $h(\mathbf{y})$ does not depend on θ. Since for likelihood purposes we can drop multiplicative constants (including functions of the data), we can take $g(\theta; t)$ as our likelihood function. In this function, the data only appears through t. □

Example 7.8. Factorization criterion in an NEF
We show that the sample mean is a sufficient statistic for NEFs. Let the observations $Y_1, \ldots, Y_n$ be i.i.d. random variables from the NEF

$$f(y; \theta) = e^{\theta y - \psi(\theta)} h(y).$$

The joint density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is

$$f(\mathbf{y}; \theta) = e^{n(\theta \bar{y} - \psi(\theta))} h_n(\mathbf{y}).$$

This is exactly in the form needed for the factorization criterion: the $h_n(\mathbf{y})$ factor does not depend on θ, and the exponential factor depends on $\mathbf{y}$ only through the sample mean $\bar{y}$. Hence the sample mean $\bar{Y}$ is a sufficient statistic.

Corollary 7.9. MLE depends on data only through sufficient statistic
Suppose that T is a sufficient statistic for θ and that we have factored the likelihood function as

$$L(\theta; \mathbf{y}) = cg(\theta; t).$$

Then the MLE depends upon the data only through T, and the Fisher information in the sample is

$$\mathcal{I}_{\mathbf{Y}}(\theta) = \mathrm{Var}\left[\frac{\partial \log g(\theta; T)}{\partial \theta}\right].$$

Corollary 7.10. Posterior depends on data only through sufficient statistic
Suppose that T is a sufficient statistic for θ. Then the posterior for $\theta, \pi(\theta \mid \mathbf{y})$, depends upon the data only through T :

$$\pi(\theta \mid \mathbf{y}) = \pi(\theta \mid t)$$

which depends on the data only through t.

Proof. Let $\pi(\theta)$ be the prior density. Then the posterior density is

$$\pi(\theta \mid \mathbf{y}) \propto g(\theta; t)\pi(\theta).$$

The data enter the right-hand side only through t, and normalizing $g(\theta; t)\pi(\theta)$ so that it integrates to 1 (with respect to θ) does not introduce any additional involvement of the data. □

Theorem 7.11. Rao-Blackwell
Let $\hat{\theta}$ be an estimator for θ and T be a sufficient statistic for θ for the parametric model $F_{\mathbf{Y};\theta}$. Then the Rao-Blackwellized estimator

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

is better than or equal to $\hat{\theta}$ in MSE. The estimator $\hat{\theta}_{\mathrm{RB}}$ is strictly better than $\hat{\theta}$ in MSE unless $\hat{\theta}$ is already a deterministic function of T.

Proof. The result follows from Adam's law and Eve's law. By Adam's law,

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{RB}}\right] = \mathrm{E}[\mathrm{E}[\hat{\theta} \mid T]] = \mathrm{E}[\hat{\theta}]$$

so the bias of $\hat{\theta}_{\mathrm{RB}}$ is the same as that of $\hat{\theta}$. Therefore, if $\hat{\theta}_{\mathrm{RB}}$ has lower variance than $\hat{\theta}$, then it also has lower MSE. To compare the variances, we can use Eve's law:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{E}[\mathrm{Var}(\hat{\theta} \mid T)] + \mathrm{Var}(\mathrm{E}[\hat{\theta} \mid T]) = \mathrm{E}[\mathrm{Var}(\hat{\theta} \mid T)] + \mathrm{Var}\left(\hat{\theta}_{\mathrm{RB}}\right) \geq \mathrm{Var}\left(\hat{\theta}_{\mathrm{RB}}\right),$$

with strict inequality unless $\mathrm{Var}(\hat{\theta} \mid T) = 0$ with probability 1 . If $\mathrm{Var}(\hat{\theta} \mid T) = 0$ then, given T, the estimator $\hat{\theta}$ is constant, which means that $\hat{\theta}$ is a deterministic function of T. □

8 Hypothesis Testing

8.1 Introduction

Suppose we want to make decisions from some data. In *decision theory*, we specify the possible actions and calculate the expected utility (or expected loss) of each action, averaging out randomness and then taking the action delivering the highest expected utility. This approach is especially useful in situations based on the decision makers' personal utility (e.g., deciding how to build an investment portfolio or spam filter).

Alternatively, *frequentist hypothesis testing* or null hypothesis significance testing (NHST) loses the flexibility of using a utility function, but of course also avoids having to specify a utility function. In NHST, we use the data to either retain or reject the null. We say "retain" rather than "accept" since if we retain the null hypothesis we are not definitively concluding that it is true; rather, we are concluding that there is not sufficient evidence to reject it.

8.2 Hypotheses, tests, critical values, and power

Definition 8.1. Statistical hypothesis
Partition the parameter space Θ into two disjoint pieces: $\Theta = \Theta_0 \cup \Theta_1$. Then test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

$H_0 : \theta \in \Theta_0$ is called the null hypothesis and $H_1 : \theta \in \Theta_1$ is called the alternative hypothesis. The null hypothesis is simple if it consists of a single point, say $\Theta_0 = \{\theta_0\}$; we then write the null hypothesis as $H_0 : \theta = \theta_0$. The null hypothesis is composite if it is not simple. Similarly, the alternative hypothesis can be simple or composite.

Definition 8.2. Statistical hypothesis test
A hypothesis testing procedure, or test for short, specifies which values of $\mathbf{y}$ lead to H_0 being rejected and which lead to H_0 being retained.

Definition 8.3. Critical region
The retention region A is the set of possible values of the data $\mathbf{y}$ such that we retain the null hypothesis if $\mathbf{y} \in A$. The rejection region or critical region is then A^C, and the hypothesis is rejected if $\mathbf{y} \notin A$.

Our tests are generally based on a test statistic $T(\mathbf{y})$. Often $T(\mathbf{y})$ is univariate, and then our test is of the form $A^C = \{\mathbf{y} : T(\mathbf{y}) > c\}$ or $A^C = \{\mathbf{y} : T(\mathbf{y}) < c_L \text{ or } T(\mathbf{y}) > c_U\}$.

Definition 8.4. Critical values
Let $T(\mathbf{y})$ be a test statistic. If the rejection region A^C is of the form $\{\mathbf{y} : T(\mathbf{y}) > c\}$ or of the form $\{\mathbf{y} : T(\mathbf{y}) < c_L \text{ or } T(\mathbf{y}) > c_U\}$, then the fixed numbers c, c_L, c_U are called critical values.

Definition 8.5. Power function
Assume the data are generated by the model $F_{\mathbf{Y};\theta}$. Suppose we have formulated our null and alternative hypotheses and selected a retention region A. The power function of our test is

$$\beta(\theta) = P_{\mathbf{Y};\theta}(\mathbf{Y} \notin A) = \int_{A^C} f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

for $\theta \in \Theta$, recalling A^c, the rejection region, is the complement of A.

Definition 8.6. One-sided tests and two-sided tests
A test of the hypotheses

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

is called a two-sided test. Tests of the hypotheses

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

(or the reverse inequalities) are called one-sided tests. Note the two-sided tests are a simple null hyopthesis against a composite alternative hypothesis, while the one-sided test is a composite against a composite.

8.3 Hypothesis testing errors and size

Definition 8.7. False positive, false negative
For statistical hypothesis testing there are two types of errors:

- Type I error (false positive): $\theta \in \Theta_0$ but $\mathbf{y} \notin A$,
- Type II error (false negative): $\theta \in \Theta_1$ but $\mathbf{y} \in A$.

The probabilities of these errors, for each element of θ in Θ_0 and Θ_1, can be read off from the power function:

$$P_{\mathbf{Y};\theta}(\text{Type I error}) = P_{\mathbf{Y};\theta}\left(\mathbf{Y} \in A^C\right), \quad \text{for } \theta \in \Theta_0,$$

and

$$P_{\mathbf{Y};\theta}(\text{Type II error}) = P_{\mathbf{Y};\theta}(\mathbf{Y} \in A), \quad \text{for } \theta \in \Theta_1.$$

Definition 8.8. Size
The size or level of the test is the maximum possible Type I error probability:

$$\alpha = \max_{\theta \in \Theta_0} \beta(\theta).$$

An α-sized test is said to be valid if the size is indeed α and invalid if not.

8.4 Calibrating the size of testing procedures

Hypothesis testing procedures can be compared across commonly sized procedures, favoring those which are most powerful under the alternative. Oftentimes, though, we want to calibrate our tests

to hit a prespecified size α?

In the case of a simple null hypothesis $H_0 : \theta = \theta_0$, i.e., Θ_0 contains just the value θ_0, the size is

$$\beta(\theta_0) = P_{\mathbf{Y};\theta_0}(\text{Type I error}) = P_{\mathbf{Y};\theta_0}\left(\mathbf{Y} \in A^C\right),$$

and a correctly sized test finds a rejection region A^C such that

$$\beta(\theta_0) = \alpha.$$

In this case, a common approach is to set a rejection region of the form $T(\mathbf{Y}) > c$, and let the critical value c be whatever it needs to be in order to make $\beta(\theta_0) = \alpha$.

Definition 8.9. Nominal size of a test
Of course, to exactly control the size of a test, we need to know the distribution of the test statistic under the null. Oftentimes we know the distribution of the p-quantile and MLE asymptotically but not exactly, so we instead try to achieve the correct size only asymptotically as the sample size $n \to \infty$,

$$\lim_{n\to\infty} P_{\mathbf{Y};\theta_0}\left(\mathbf{Y} \in A^C\right) = \alpha.$$

In this situation, we say the test has nominal size α, while the actual size is $P_{\mathbf{Y};\theta_0}\mathbf{Y}(\mathbf{Y} \in A^C)$.

In practical applications, if you develop a new test with nominal size α, it would be expected that the actual size of new tests will be calculated using simulation for some realistic cases. Sometimes the asymptotics provide a great approximation, but other times it will take a very large n for the nominal size to be close to the actual size.

Theorem 8.10. (t**-statistic has a** t**-distribution)**
Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$. Then

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\bar{Y} - \mu)}{\hat{\sigma}} \sim t_{n-1},$$

where t_{n-1} is the Student-t distribution with $n-1$ degrees of freedom (see Definition 10.4.4 in the Stat 110 book) and $\hat{\sigma}^2 = (n-1)^{-1}\sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2$.

8.5 Duality between hypothesis tests and confidence intervals

Theorem 8.11. Inverting a confidence interval
Suppose $C(\mathbf{Y})$ is a $1-\alpha$ confidence interval for θ. Retaining the null if $\theta_0 \in C(\mathbf{Y})$ is a α sized test of the null $H_0 : \theta = \theta_0$.

Proof. The definition of a $1-\alpha$ CI is that $P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1-\alpha$. Specializing this result to $\theta = \theta_0$ gives the desired size for the hypothesis test.

The other way around is slightly more complicated. It says find a $1-\alpha$ confidence interval by finding all the null hypothesis values of θ which are not rejected by a size α test of θ. □

Theorem 8.12. Inverting a test
For any θ_0, define the set $A(\theta_0)$ as the retention region (i.e., H_0 is retained if $\mathbf{y} \in A(\theta_0)$) for a α sized test of the null $H_0 : \theta = \theta_0$. Then for a fixed $\mathbf{y}$, let

$$C(\mathbf{y}) = \{\theta : \mathbf{y} \in A(\theta)\},$$

the set of all parameter values θ which this data $\mathbf{y}$ would lead to the null being retained. Then $C(\mathbf{Y})$ is a $1 - \alpha$ confidence interval for θ.

Proof. By construction $\mathbf{y} \in A(\theta)$ if and only if $\theta \in C(\mathbf{y})$. This implies

$$P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = P_{\mathbf{Y};\theta}(\mathbf{Y} \in A(\theta)).$$

But the test has size α, so this establishes the theorem. □

Remark 8.13. Hypothesis testing versus confidence intervals
Both Joe and Neil prefer the use of confidence intervals to hypothesis testing. They are equivalent, but hypothesis testing yields a binary outcome (rejection or retention), while confidence intervals yield much more, in fact providing the information to carry out an infinite number of hypothesis tests–telling you a whole range of θ where the null would be retained.

8.6 Testing using likelihood-based quantities

Definition 8.14. Wald test
We turn the MLE $\hat{\theta}$ into a test statistic using the asymptotic pivot under the null distribution,

$$T(\mathbf{Y}) = \sqrt{\mathcal{I}_{\mathbf{Y}}(\theta_0)}\left(\hat{\theta} - \theta_0\right) \dot{\sim} \mathcal{N}(0, 1),$$

noting that under the null the true value of θ is $\theta^* = \theta_0$ and so the information is evaluated at θ_0. The test has a nominal α size and rejects if

$$T(\mathbf{Y}) < Q_{\mathcal{N}(0,1)}(\alpha/2) \text{ or } T(\mathbf{Y}) > Q_{\mathcal{N}(0,1)}(1 - \alpha/2),$$

or equivalently, reporting using the square of $T(\mathbf{Y})$, if

$$W(\mathbf{Y}) = \mathcal{I}_{\mathbf{Y}}(\theta_0)\left(\hat{\theta} - \theta_0\right)^2 > Q_{\chi^2_1}(1 - \alpha).$$

Definition 8.15. Score test
Using a Taylor expansion, it may be shown that under the null hypothesis and some regularity conditions, then

$$s\left(\theta_0; \mathbf{Y}\right) \dot{\sim} \mathcal{N}\left(0, \mathcal{I}_{\mathbf{Y}}\left(\theta_0\right)\right).$$

The score test evaluates the score at the null value θ_0,

$$T(\mathbf{Y}) = \frac{s\left(\theta_0; \mathbf{Y}\right)}{\sqrt{\mathcal{I}_{\mathbf{Y}}\left(\theta_0\right)}} \dot{\sim} \mathcal{N}(0,1),$$

rejecting the null if

$$T(\mathbf{Y}) < Q_{\mathcal{N}(0,1)}(\alpha/2) \text{ or } T(\mathbf{Y}) > Q_{\mathcal{N}(0,1)}(1-\alpha/2),$$

or equivalently if

$$\frac{s\left(\theta_0; \mathbf{Y}\right)^2}{\mathcal{I}_{\mathbf{Y}}\left(\theta_0\right)} > Q_{\chi^2_1}(1-\alpha),$$

giving a test with nominal α size. Intuitively, the expected value of the score is 0 at the true parameter value, which our null takes to be $\theta = \theta_0$, so a very large value of $|s\left(\theta_0, \mathbf{y}\right)|$ makes it seem implausible that $\theta = \theta_0$.

Definition 8.16. Likelihood ratio (LR) test
Consider a parameter space $\Theta = \{\theta_0, \theta_1\}$, yielding a simple null $H_0 : \theta = \theta_0$ and simple alternative $H_1 : \theta = \theta_1$. The likelihood ratio is

$$LR(\mathbf{y}) = \frac{L\left(\theta_1; \mathbf{y}\right)}{L\left(\theta_0; \mathbf{y}\right)},$$

and we reject the null if $LR(\mathbf{y}) > c$, for c determined so that $P_{\mathbf{Y};\theta_0}(LR(\mathbf{Y}) > c) = \alpha$. To determine c, instead of calculating analytically the distribution of $LR(\mathbf{Y})$, we can use simulation since under the null we know $F_{\mathbf{Y};\theta_0}$. That is, we can simulate under the null R distinct i.i.d. copies $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(R)}$, and then choose the smallest c so that

$$\frac{1}{R}\sum_{j=1}^{R} I\left(LR\left(\mathbf{Y}^{(j)}\right) > c\right) = \alpha.$$

Theorem 8.17. Neyman-Pearson lemma
For a simple null $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ the most powerful test, at size α, is the likelihood ratio $LR(\mathbf{y})$.

The proof is outside the textbook scope, but Neyman-Pearson lets us know that the LR test is the best test for parametric testing with a simple null and simple alternative; the theorem extends to some one-sided tests, but not generally two-sided tests.

Definition 8.18. Likelihood ratio test statistic
Consider now the two-sided test problem: $H_0 : \theta = \theta_0$, versus $H_1 : \theta \neq \theta_0$. Then the likelihood ratio test statistic is

$$\mathrm{LR}(\mathbf{y}) = \frac{L(\hat{\theta}; \mathbf{y})}{L(\theta_0; \mathbf{y})}.$$

Note that $\mathrm{LR}(\mathbf{y})$ is always at least 1 . The likelihood ratio test rejects H_0 if $\mathrm{LR}(\mathbf{y})$ is "large", which makes sense intuitively since a large value of $\mathrm{LR}(\mathbf{y})$ means that some parameter value other than θ_0 has much higher likelihood than θ_0. Like the Wald test, the LR test is based on the MLE $\hat{\theta}$.

To instead work with the log-likelihood ratio

$$\log\{\mathrm{LR}(\mathbf{y})\} = \log L(\hat{\theta}; \mathbf{y}) - \log L(\theta_0; \mathbf{y}),$$

we have the following theorem.

Theorem 8.19. Asymptotic distribution of the log likelihood ratio
Let

$$\Lambda(\mathbf{y}) = 2\left[\log L(\hat{\theta}; \mathbf{y}) - \log L(\theta_0; \mathbf{y})\right]$$

be twice the log of the likelihood ratio. Under the null $H_0 : \theta = \theta_0$ and some mild regularity conditions, rejecting the null if

$$\Lambda(\mathbf{y}) > Q_{\chi^2_1}(1 - \alpha)$$

has a nominal α size.

Proof. Under the null, $\hat{\theta} \xrightarrow{p} \theta_0$, so using a Taylor expansion of θ_0 about $\hat{\theta}$ (we do it this way to knock out a term in the expansion)

$$\log L(\theta_0; \mathbf{y}) \approx \log L(\hat{\theta}; \mathbf{y}) + \left(\theta_0 - \hat{\theta}\right) s(\hat{\theta}; \mathbf{y}) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)^2 s'(\hat{\theta}; \mathbf{y})$$

Now for regular problems, by the definition of the MLE, $s(\hat{\theta}; \mathbf{y}) = 0$, so rearranging produces

$$\Lambda(\mathbf{y}) \approx \left(\hat{\theta} - \theta_0\right)^2 \left[-s'(\hat{\theta}; \mathbf{y})\right] = \left\{\sqrt{n}\left(\hat{\theta} - \theta_0\right)\right\}^2 \left\{-n^{-1} s'(\hat{\theta}; \mathbf{y})\right\}.$$

To be clear about our logic, think about the i.i.d. case. Then

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, 1/\mathcal{I}_{Y_1}(\theta_0)\right)$$

and, by the strong law of large numbers,

$$-n^{-1} s'(\theta_0; \mathbf{Y}) = -n^{-1} \sum_{j=1}^{n} s'(\theta_0; Y_j) \xrightarrow{p} -\mathrm{E}\left[s'(\theta_0; Y_1)\right] = \mathcal{I}_{Y_1}(\theta_0),$$

where the last equality holds by the information equality. So by the continuous mapping theorem, under the null,

$$-n^{-1} s'(\hat{\theta}; \mathbf{Y}) \xrightarrow{p} \mathcal{I}_{Y_1}(\theta_0).$$

Hence the result holds by Slutsky's Theorem. □

Comparing the three tests

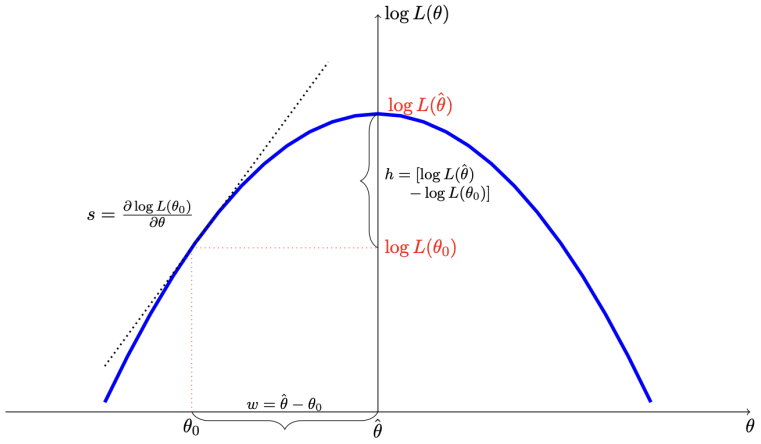

Figure 8.1: The log-likelihood function is depicted as a parabola for concreteness. The Wald test uses width w, the LR test uses height h, and the score test uses slope s.

Remark 8.20. The LR test is generally the best
In practice the Fisher information is often not available in closed form, and having to estimate it might cause some additional distortion. Thus, the score, despite being a relatively simple statistic, may not be great. The Wald test is usually viewed as the worst behaving of the three tests; not only is the estimation of Fisher information an issue, but the MLE is not the sum of items and so the asymptotic distribution can take quite a while to kick in.

Fascinatingly, the likelihood ratio test is usually expected to have the size closest to the nominal value. It looks quite complicated but it does not require finding or estimating the Fisher information, and it enjoys a nice invariance property due to the invariance property of likelihood functions.

8.7 p-values

One of the most widely used and controversial communication devices in statistics is the p-value statistic.

Definition 8.21. *p*-value for simple null
For a simple null hypothesis H_0, suppose that there is a test statistic $T(\mathbf{Y})$, such that the test rejects H_0 when $T(\mathbf{Y})$ is large, i.e., there is a critical value c such that the rejection region is $\{\mathbf{y} : T(\mathbf{y}) > c\}$. Let $\mathbf{y}$ be the observed data and t be the observed value of T. Then the observed p-value is

$$p(\mathbf{y}) = P\left(T \geq t \mid H_0\right).$$

Similarly, if the test rejects H_0 when $T(\mathbf{Y})$ is small, then the observed p-value is

$$p(\mathbf{y}) = P\left(T \leq t \mid H_0\right).$$

That is, the p-value is the probability of a result at least as extreme as what was actually observed, assuming the null hypothesis.

Theorem 8.22. *p*-values are Uniform under the null
Let $T(\mathbf{Y})$ be a continuous test statistic, and suppose we are using a hypothesis test that rejects $H_0 : \theta = \theta_0$ when $T(\mathbf{y})$ is large. Then the p-value is Uniform under the null:

$$p(\mathbf{Y}) \sim \text{Unif}(0, 1),$$

under H_0.

Proof. Let t_0 be the observed value of T. The p-value that gets computed from the data is then $P\left(T \geq t_0\right) = 1 - F_T\left(t_0\right)$, where F_T is the CDF of T under H_0. So as a random variable, the p-value is $1 - F_T(T)$. By universality of the Uniform,

$$F_T(T) \sim \text{Unif}(0, 1)$$

if H_0 is true. Since $1 - U \sim \text{Unif}(0, 1)$ for $U \sim \text{Unif}(0, 1)$, it follows that the p-value is $\text{Unif}(0, 1)$ under the null hypothesis. □

Definition 8.23. *p*-value for general null
For a null hypothesis H_0, let A_α be a retention region for each α, such that the test has size $\alpha : P_{\mathbf{Y};\theta_0}\left(\mathbf{Y} \in A_\alpha^C\right) = \alpha$. For data $\mathbf{y}$, the p-value is the smallest α (Type I error rate) at which we could have rejected H_0. That is,

$$p(\mathbf{y}) = \min\left\{\alpha : \mathbf{y} \in A_\alpha^C\right\}.$$

(It is easy to check that if the null hypothesis is simple and the test statistic T is a continuous r.v., then the two definitions of p-value are equivalent.)

8.8 Multiparameter testing*

8.9 Testing when model approximates the truth*

9 Bayesian Inference

9.1 Introduction

9.2 Prior to posterior

Definition 9.1. Prior and posterior; marginal likelihood
Consider a parametric model $F_{Y|\theta}$ for data $\mathbf{y}$ with parameter θ. In taking a Bayesian approach, we posit a joint distribution for the pair

$$\mathbf{Y}, \theta.$$

Then $f(\mathbf{y}; \theta) = f(\mathbf{y} \mid \theta)$ is the conditional distribution of $\mathbf{y}$ given θ. The prior for θ is the marginal distribution of θ. The posterior for θ is the conditional distribution of θ given $\mathbf{y}$. The prior density for θ is often denoted by $\pi(\theta)$, in which case the posterior density is $\pi(\theta \mid \mathbf{y})$. The marginal likelihood $f(\mathbf{y})$ is the marginal distribution of the data.

Theorem 9.2. Bayes' rule
Consider a parametric model $F_{Y|\theta}$ for data $\mathbf{y}$, and let $\pi(\theta)$ be the prior density on the parameter θ. Let $L(\theta; \mathbf{y}) = f(\mathbf{y} \mid \theta)$ be the likelihood function. Then the posterior density for θ is proportional to the likelihood times the prior:

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

and we may calculate the normalizing constant $f(\mathbf{y})$ by LOTP:

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} L(\tilde{\theta}; \mathbf{y})\pi(\tilde{\theta})d\tilde{\theta}.$$

9.3 Point estimation

Definition 9.3. Prior and posterior mean, median, and mode
Let the estimand θ have a continuous prior density $\pi(\theta)$. Then define

$$\begin{aligned}
\text{prior mean } &= \mathrm{E}[\theta] = \int_{-\infty}^{\infty} \theta\pi(\theta)d\theta, \quad \text{if this integral exists,} \\
\text{prior median } &= Q_\theta(0.5), \quad \text{which always exists,} \\
\text{prior mode } &= \underset{\theta}{\mathrm{argmax}}\pi(\theta), \quad \text{if this value exists and is unique.}
\end{aligned}$$

Likewise, let θ have a continuous posterior density $\pi(\theta \mid \mathbf{y})$. Then define

$$\begin{aligned}
\text{posterior mean } &= \mathrm{E}[\theta \mid \mathbf{y}] = \int_{-\infty}^{\infty} \theta\pi(\theta \mid \mathbf{y})d\theta, \quad \text{if this integral exists,} \\
\text{posterior median } &= Q_{\theta|\mathbf{y}}(0.5), \quad \text{which always exists,} \\
\text{posterior mode } &= \underset{\theta}{\mathrm{argmax}}\pi(\theta \mid \mathbf{y}), \quad \text{if this value exists and is unique.}
\end{aligned}$$

Note the MAP (the posterior mode) has the very convenient property that it can be computed without knowing the normalizing constant in Bayes theorem, since multiplying a function by a positive constant has no effect on where the function is maximized.

Theorem 9.4. Posterior mean minimizes squared error loss; posterior median minimizes absolute error loss
For squared error loss $\mathrm{Loss}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, the posterior mean is

$$\mathrm{E}[\theta \mid \mathbf{y}] = \underset{c}{\mathrm{argmin}}\mathrm{E}\left[(\theta - c)^2 \mid \mathbf{y}\right].$$

For absolute error loss $\mathrm{Loss}(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, the posterior median is

$$Q_{\theta|\mathbf{y}}(0.5) = \underset{c}{\mathrm{argmin}}\mathrm{E}[|\theta - c| \mid \mathbf{y}].$$

Proof. The result follows from Theorem 6.1.4 in the Stat 110 book, which says that for any r.v. X,

$$\mathrm{E}[X] = \underset{c}{\mathrm{argmin}}\mathrm{E}\left[(X - c)^2\right]$$

and

$$Q_X(0.5) = \underset{c}{\mathrm{argmin}}\mathrm{E}[|X - c|].$$

These results and a more general result for $Q_X(p)$ are also proven in Chapter 6. □

9.4 Computing Bayesian estimators

Remark 9.5. The power of simulation
Sometimes we want to compute quantities like the posterior mean and posterior median. Doing so analytically may be very difficult for several reasons: the integrals may be difficult even if we know the posterior density, or we may not know the normalizing constant of the posterior density. If we are interested in a specific component of the parameter, then marginalizing to that component is also often difficult.

Simulation is powerful to estimate quantities in this situation. For eaxmple, a common approach is Markov chain Monte Carlo (MCMC), such as the Metropolis-Hastings algorithm, where the input is the unnormalized posterior

$$L(\theta; \mathbf{y})\pi(\theta).$$

Then MCMC creates a simulated path from a Markov chain whose stationary distribution is $\pi(\theta \mid \mathbf{y})$. If this chain is run for a long time, then we can use the simulated θ values

$$\theta^{[1]}, \ldots, \theta^{[B]}$$

from the chain to approximate $\mathrm{E}[\theta \mid \mathbf{y}]$ or $Q_{\theta|\mathbf{y}}(0.5)$ (or other summaries of the posterior distribution) using sample versions, e.g., use the sample mean of $\theta^{[1]}, \ldots, \theta^{[B]}$ to approximate the posterior mean, and use the sample median of $\theta^{[1]}, \ldots, \theta^{[B]}$ to approximate the posterior median.

9.5 Credible intervals

Definition 9.6. Credible interval
Let $0 < \alpha < 1$. A $1 - \alpha$ credible interval or posterior probability interval for an estimand θ is an interval estimate $[a(\mathbf{y}), b(\mathbf{y})]$ such that

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

While the randomness of confidence comes from the interval itself, the randomness here comes from our uncertainty about θ. In general, for a $1-\alpha$ credible interval we can choose the interval

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

though this is not a unique choice nor necessarily the shortest credible interval. If the computation is carried out using by sampling B times from the posterior, then this credible interval can be reported as

$$\left[\theta_{(\lceil 0.025B \rceil)}, \theta_{(\lceil 0.975B \rceil)}\right].$$

Remark 9.7. Coverage probability of credible interval
There is no guarantee that a 95% credible interval will be a 95% confidence interval, nor vice versa. On average though, the coverage probability of a 95% credible interval will be 95%, where the averaging is over both θ and $\mathbf{Y}$. To see this, we can just apply Adam's law: letting I be the indicator of the credible interval covering θ,

$$P(I = 1) = \mathrm{E}[I] = \mathrm{E}[\mathrm{E}[I \mid \mathbf{Y}]] = \mathrm{E}[P(I = 1 \mid \mathbf{Y})] = \mathrm{E}[0.95] = 0.95.$$

Once we have developed a credible interval estimator, we can assess its coverage probability (often via simulation).

9.6 Conjugate priors

Definition 9.8. Conjugate prior
A family of priors is conjugate for a particular statistical model if choosing a prior in the family always results in a posterior that is in the same family.

Theorem 9.9. Normal-Normal conjugacy with general sample size
Let $Y_1, \ldots, Y_n$ be random variables with

$$Y_j \mid \mu \overset{i.i.d.}{\sim} \mathcal{N}\left(\mu, \sigma^2\right).$$

Let the prior be $\mu \sim \mathcal{N}\left(\mu_0, \tau_0^2\right)$, with $\sigma^2, \mu_0, \tau_0^2$ known. Then the posterior distribution for μ is

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

Write the posterior variance divided by the prior variance as

$$b_n = \frac{\tau_n^2}{\tau_0^2} = \frac{\tau_0^{-2}}{n\sigma^{-2} + \tau_0^{-2}} = \frac{\sigma^2}{\sigma^2 + n\tau_0^2}.$$

Then we can also write the posterior distribution in the following nice form:

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

Proof. To show this, we can calculate the posterior directly as we did in the $n = 1$ case, but a more elegant way is to note that $\bar{Y}$ is a sufficient statistic and

$$\bar{Y} \,\Big|\, \mu \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

so the conditional distribution of $\mu \mid (\mathbf{Y} = \mathbf{y})$ is the same as the conditional distribution of $\mu \mid (\bar{Y} = \bar{y})$, which we already know from the $n = 1$ case. □

Theorem 9.10. Normal model with heteroskedasticity
Suppose $Y_j \mid \mu \overset{\text{ind.}}{\sim} \mathcal{N}\left(\mu, \sigma_j^2\right)$ and $\mu \sim \mathcal{N}\left(m_0, \tau_0^2\right)$. Then $\mu \mid y_1, \ldots, y_n \sim \mathcal{N}\left(m_n, \tau_n^2\right)$,where

$$\tau_n^{-2} = \tau_0^{-2} + \sum_{j=1}^{n} \sigma_j^{-2}, \quad m_n = \tau_n^2 \left(\tau_0^{-2} m_0 + \sum_{j=1}^{n} \sigma_j^{-2} y_j \right),$$

where we assume that $\sigma_1^2, \ldots, \sigma_n^2, m_0, \tau_0$ are known.

Theorem 9.11. Bayesian Gaussian linear regression
Assume that

$$Y_j \mid (\mathbf{X} = \mathbf{x}, \theta) \overset{\text{ind.}}{\sim} \mathcal{N}\left(\theta x_j, \sigma^2\right), \quad j = 1, \ldots, n,$$

and $\theta \mid (\mathbf{X} = \mathbf{x}) \sim \mathcal{N}\left(m_0, \tau_0^2\right)$. Then $\theta \mid (\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}) \sim \mathcal{N}\left(m_n, \tau_n^2\right)$, where

$$\tau_n^{-2} = \tau_0^{-2} + \sigma^{-2} \sum_{j=1}^{n} x_j^2, \quad m_n = \tau_n^2 \left(\tau_0^{-2} m_0 + \sigma^{-2} \sum_{j=1}^{n} x_j y_j \right).$$

where we assume that m_0, τ_0 are known.

9.7 Bayesian model choice

Suppose that Bill advocates for a statistical model and Jose prefers a very different model. How can we learn from the data which researcher (if either) to believe? Is Bill a better statistical modeler, or is Jose? In statistics, this type of problem is called model choice.

In Bayesian model choice, we follow Lindley's advice and quantify the uncertainty about model choice using probability. To formalize this, write Bill's likelihood and prior as $f(\mathbf{y} \mid \theta, \text{Bill})$ and $f(\theta \mid \text{Bill})$, and Jose's likelihood and prior as $g(\mathbf{y} \mid \lambda, \text{Jose})$ and $g(\lambda \mid \text{Jose})$. There is no need for there to be any connection between Bill's model and Jose's model, which we have emphasized by using different letters and parameters for their models. Importantly, there is no need for the dimensions of θ and λ to be the same.

Definition 9.12. Bayes factor
The Bayes factor is defined as the ratio of marginal likelihoods for the two models:

$$\frac{f(\mathbf{y} \mid \text{ Bill })}{g(\mathbf{y} \mid \text{ Jose })}.$$

The Bayes factor is the factor that converts the ratio of prior probabilities of the two models to the ratio of posterior probabilities:

$$\frac{P(\text{ Bill } \mid \mathbf{y})}{P(\text{ Jose } \mid \mathbf{y})} = \frac{P(\text{ Bill })}{P(\text{ Jose })} \frac{f(\mathbf{y} \mid \text{ Bill })}{g(\mathbf{y} \mid \text{ Jose })},$$

The above equation follows from applying Bayes' rule separately to the numerator and denominator; note that the $f(\mathbf{y})$ cancels. To find the marginal likelihood of Bill's model, we can use the law of total probability (and likewise for Jose's model):

$$f(\mathbf{y} \mid \text{ Bill }) = \int f(\mathbf{y} \mid \theta, \text{ Bill }) f(\theta \mid \text{Bill}) d\theta$$

interesting examples to read

9.8 Bayesian prediction

Definition 9.13. Posterior predictive distribution
Let Y be a variable that we wish to predict, given some observed data. In the Bayesian framework, the posterior predictive distribution of Y is the conditional distribution of Y, given the observed data.

9.9 Hierarchical models

We introduce hierarchical models by example.

Definition 9.14. Two-level Gaussian hierarchical model
Assume that

$$Y_j \mid \mu_1, \ldots, \mu_K \overset{\text{ind.}}{\sim} \mathcal{N}\left(\mu_j, \sigma^2\right), \quad j = 1, \ldots, K$$

where each conditional mean is

$$\mu_j \overset{\text{ind.}}{\sim} \mathcal{N}\left(\gamma, \lambda_0^2\right).$$

Then the pair $(\mathbf{Y}, \boldsymbol{\mu})$ follow a two-level Gaussian hierarchical model, also known as a two-level multilevel model. The model is indexed by $\psi = (\sigma, \gamma, \lambda_0)$, which are called hyperparameters. Here we will think of them as known.

In this hierarchical model the Y_j are conditionally independent given $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, and the μ_j are i.i.d. By the result from Theorem 9.6.5,

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

Marginally,

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

Definition 9.15. Three-level Gaussian hierarchical model
Assume that

$$Y_j \mid \mu_1, \ldots, \mu_K, \gamma \overset{\text{ind.}}{\sim} \mathcal{N}\left(\mu_j, \sigma^2\right), \quad j = 1, \ldots, K,$$

where each conditional mean is

$$\mu_j \mid \gamma \overset{\text{ind.}}{\sim} \mathcal{N}\left(\gamma, \lambda_0^2\right),$$

and

$$\gamma \sim \mathcal{N}\left(g_0, \tau_0^2\right).$$

Then the triple $(\mathbf{Y}, \boldsymbol{\mu}, \gamma)$ follow a three-level Gaussian hierarchical model. It is indexed by hyperparameters $\psi = (\sigma, \lambda_0, g_0, \tau_0)$.

Intuitively, we may visualize this model as a tree with three levels. The hyperparameter γ is at the top (level 3), then the μ's (level 2), and finally the data Y's (level 1).

Theorem 9.16. Parameters of three-level Gaussian hierarchical model
Assume a three-level Gaussian hierarchical model. Then

$$\gamma \mid (\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}\left(g_K, \tau_K^2\right), \quad g_K = \frac{\tau_0^{-2} g_0 + \left(\sigma^2 + \lambda_0^2\right)^{-2} K\bar{y}}{\tau_0^{-2} + K\left(\sigma^2 + \lambda_0^2\right)^{-2}},$$

and

$$\mu_j \,\Big|\, (\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}\left(\frac{\lambda_0^{-2} g_K + \sigma^{-2} y_j}{\lambda_0^{-2} + \sigma^{-2}}, \frac{\lambda_0^{-4}\tau_K^2}{\left(\lambda_0^{-2} + \sigma^{-2}\right)^2} + \frac{1}{\lambda_0^{-2} + \sigma^{-2}}\right),$$

where $\tau_K^{-2} = \tau_0^{-2} + K\left(\sigma^2 + \lambda_0^2\right)^{-2}$.

Proof. Conditional on $\gamma, (\mathbf{Y}, \boldsymbol{\mu}) \mid \gamma$ is a two-level Gaussian hierarchical model, so

$$Y_j \mid \gamma \overset{\text{ind.}}{\sim} \mathcal{N}\left(\gamma, \sigma^2 + \lambda_0^2\right).$$

The first relation holds by the result for the Normal-Normal posterior. Now recall from the result for the two-level model that

$$\mu_j \,\Big|\, (\mathbf{Y} = \mathbf{y}, \gamma) \sim \mathcal{N}\left(\frac{\lambda_0^{-2}\gamma + \sigma^{-2} y_j}{\lambda_0^{-2} + \sigma^{-2}}, \frac{1}{\lambda_0^{-2} + \sigma^{-2}}\right)$$

The result for the parameters of $\mu_j \mid (\mathbf{Y} = \mathbf{y})$ then follow from Adam's law and Eve's law:

$$\begin{aligned}
\mathrm{E}\left[\mu_j \mid (\mathbf{Y} = \mathbf{y})\right] &= \mathrm{E}\left[\mathrm{E}\left[\mu_j \mid (\mathbf{Y} = \mathbf{y}, \gamma)\right]\right] \\
\mathrm{Var}\left(\mu_j \mid (\mathbf{Y} = \mathbf{y})\right) &= \mathrm{E}\left[\mathrm{Var}\left(\mu_j \mid (\mathbf{Y} = \mathbf{y}, \gamma)\right)\right] + \mathrm{Var}\left(\mathrm{E}\left[\mu_j \mid (\mathbf{Y} = \mathbf{y}, \gamma)\right]\right)
\end{aligned}$$

□

9.10 Stein's Paradox

Definition 9.17. Risk function for Bayesian estimators
Suppose we are working with respect to a loss function $L(\theta, \hat{\theta})$. Recall that in the frequentist perspective, the risk function of the estimator $\hat{\theta}$ is its expected loss,

$$R(\theta) = \mathrm{E}_{\mathbf{Y};\theta}[\mathrm{Loss}(\theta, \hat{\theta})].$$

This is the average loss incurred by using $\hat{\theta} = T(\mathbf{Y})$, averaged over the random $\mathbf{Y}; \theta$. In the Bayesian approach we can look at the expected loss given the data,

$$\mathrm{E}_{\theta|\mathbf{Y}=\mathbf{y}}[\mathrm{Loss}(\theta, \hat{\theta})] = \mathrm{E}[\mathrm{Loss}(\theta, \hat{\theta}) \mid y],$$

averaging over the random θ given $\mathbf{Y} = \mathbf{y}$.

Definition 9.18. Admissibility
An estimator $\hat{\theta}$ is inadmissible if there exists another estimator whose risk function is less than or equal to 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.

Theorem 9.19. Stein's theorem
Let

$$Y_j \sim \mathcal{N}\left(\mu_j, \sigma^2\right)$$

for $j = 1, \ldots, K$ be independent, where $K \geq 3$, the μ_j are unknown, and σ^2 is known. Let the estimand be the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and the loss function be the total squared error loss,

$$\mathrm{Loss}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{j=1}^{K} \left(\mu_j - \hat{\mu}_j\right)^2.$$

Then the MLE, which is $\mathbf{Y} = (Y_1, \ldots, Y_K)$ itself, is inadmissible.

Theorem 9.20. James-Stein estimator
In the setup of the previous theorem, let

$$S = \sum_{j=1}^{K} Y_j^2.$$

Let $\hat{\boldsymbol{\mu}}_{\mathrm{JS}} = (\hat{\mu}_{\mathrm{JS},1}, \ldots, \hat{\mu}_{\mathrm{JS},K})$ be the James-Stein estimator, defined by

$$\hat{\mu}_{\mathrm{JS},j} = \left[1 - \frac{(K-2)\sigma^2}{S}\right] Y_j.$$

Then $\hat{\boldsymbol{\mu}}_{\mathrm{JS}}$ has strictly lower risk than $\mathbf{Y}$ for all $\boldsymbol{\mu} \in \mathbb{R}^K$. Specifically, the risk function of $\mathbf{Y}$ is the constant $K\sigma^2$, whereas the risk function of $\hat{\boldsymbol{\mu}}_{\mathrm{JS}}$ is

$$\left[K - (K-2)^2\sigma^2 \mathrm{E}\left(\frac{1}{S}\right)\right] \sigma^2.$$

10 Sampling and Resampling

10.1 Introduction

10.2 Design-based inference

Definition 10.1. Model-based versus design-based inference
Throughout statistics and its applications, we encounter problems where a random sample is drawn from a population of interest, and we wish to use the sample to learn about an aspect of the population.

- Sometimes the population is a hypothetical, infinite population, such as when we consider i.i.d. draws from a parametric statistical model $F_{Y;\theta}$. So far in this book we have focused on model-based inference, where we introduce a model for the data $Y_1, Y_2, \ldots, Y_n$, with unknown CDF $F_{Y;\theta}$.
- In contrast, design-based inference involves sampling from a specific, finite population. The values in the population of the variable of interest are fixed numbers $y_1, y_2, \ldots, y_N$, where N is the population size. The randomness comes entirely from drawing a random sample from the population, not from modeling the data.

A model-based approach with an approximately correct model may use the data more efficiently, while also relying less on assumptions about how the sampling was done. But sometimes we do not think we have enough information to come up with a reasonable model, and then a design-based approach can reduce concerns about whether we have a decent model and about subjectivity in the choice of the model.

Problem setup: Throughout this section and the next, consider the following setup. There is a fixed, finite population of interest, consisting of N individuals. The individuals have been given ID numbers $1, 2, \ldots, N$, so that each individual is uniquely labeled. Let y_i be the quantity of interest for individual i (e.g., the i th person's age or income), so the entire finite population is

$$y_1, \ldots, y_N.$$

Definition 10.2. Finite sample estimands
A finite sample estimand is an estimand that is defined as a function of $y_1, \ldots, y_N$. Some notable examples of finite sample estimands are

$$\mu = \frac{1}{N}\sum_{j=1}^{N} y_j, \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N} (y_j - \mu)^2, \quad F(y) = \frac{1}{N}\sum_{j=1}^{N} I(y_j \leq y),$$

which are called the population mean, population variance, and population CDF of the y variable, respectively.

Definition 10.3. Census
In a census, we get to observe all of $y_1, y_2, \ldots, y_N$.

10.3 Sampling design

Definition 10.4. Sampling design
Suppose that we will collect a random sample of size n. Let I_j be the ID number of the j th individual selected. The sampling design is the joint probability mass function of $I_1, \ldots, I_n$. That is, writing $\mathbf{i}_{1:n} = (i_1, \ldots, i_n)$, the sampling design specifies

$$P(I_1 = i_1, \ldots, I_n = i_n), \quad \text{for all } i_{1:n} \in \{1, \ldots, N\}^n.$$

This joint probability can be written more compactly as $P(\mathbf{I}_{1:n} = \mathbf{i}_{1:n})$. The sampling design determines the statistical properties of the outcomes

$$\mathbf{Y}_{1:n} = (Y_1, \ldots, Y_n) = (y_{I_1}, \ldots, y_{I_n}),$$

as we regard $y_1, \ldots, y_N$ as fixed, but the sampled IDs $\mathbf{I}_{1:n}$ as random.

Definition 10.5. Equal probability sample
A sampling design such that the marginal PMFs satisfy

$$P(I_j = k) = 1/N, \quad \text{for all } j = 1, \ldots, n, \quad k = 1, \ldots, N$$

is called an equal probability sample.

Definition 10.6. Uniformly random draw
An equal probability sample of size 1 is called a uniformly random draw. Explicitly, let I be a random variable whose possible values are $1, 2, \ldots, N$, with equal probabilities. That is, I is Discrete Uniform on $\{1, 2, \ldots, N\}$. Let

$$Y = y_I.$$

Then Y is the y value for an individual chosen at random from the population. We call I a uniformly random draw from the population of ID numbers and Y a uniformly random draw from the population of y values. The values $y_1, \ldots, y_N$ are regarded as fixed, so all of the randomness in Y comes from the randomness in I. This is the essential perspective of design-based inference.

Theorem 10.7. Some properties of equal probability samples
All equal probability samples have the property that, for $j = 1, \ldots, n$, the

$$\mathrm{E}_I[Y_j] = \mu, \quad \mathrm{Var}_I(Y_j) = \sigma^2, \quad \mathrm{E}_I[\bar{Y}] = \mu, \quad \mathrm{E}_I[\hat{F}(y)] = F(y),$$

where the expectation is over the sampling design, regarding $y_1, \ldots, y_N$ as fixed. (However, the equal probability sample assumption is not enough to tell us what the variance of $\bar{Y}$ or $\hat{F}(y)$ are, nor the expectation of S^2.)

Proof. By linearity,

$$\mathrm{E}_I[\bar{Y}] = \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}_I\left[y_{I_j}\right].$$

But

$$\mathrm{E}_I\left[Y_j\right] = \mathrm{E}_I\left[y_{I_j}\right] = \sum_{k=1}^{N} P\left(I_j = k\right) y_k = \frac{1}{N}\sum_{k=1}^{N} y_k = \mu,$$

by the equal probability sample assumption. The same argument yields $\mathrm{E}_I[\hat{F}(y)] = F(y)$. □

Now we discuss SRSs with and without replacement, which both yield equal probability samples.

Definition 10.8. SRS with replacement
Simple random sampling (SRS) with replacement draws i.i.d. $I_1, \ldots, I_n$ from a Discrete Uniform distribution $\{1, 2, \ldots, N\}$, and sets

$$Y_j = y_{I_j}, \quad j = 1, \ldots, n.$$

The sampling is with replacement since the same ID number can be chosen multiple times. Conveniently, $Y_1, \ldots, Y_n$ are i.i.d., with each being a uniformly random draw from the population. The sampling design is

$$P\left(I_1 = i_1, \ldots, I_n = i_n\right) = \prod_{j=1}^{n} P\left(I_j = i_j\right) = 1/N^n, \quad i_{1:n} \in \{1, \ldots, N\}^n.$$

When calculating an expectation under SRS with replacement, we sometimes write E_{with} as a reminder that the sampling is with replacement. So SRS with replacement is an equal probability sample.

Definition 10.9. SRS without replacement
A simple random sample without replacement of size $n \leq N$ from the population is a sample of size n, chosen without replacement, such that all $N!/(N-n)!$ possible permutations are equally likely. Then set

$$Y_j = y_{I_j}, \quad j = 1, \ldots, n.$$

When calculating an expectation under SRS without replacement, we sometimes write $\mathrm{E}_{\text{w/o}}$. In terms of the sampling design, the SRS without replacement scheme has

$$P\left(I_1 = i_1, \ldots, I_n = i_n\right) = \frac{1}{N!/(N-n)!}$$

By symmetry, SRS without replacement is an equal probability sampling scheme, but

$$\mathrm{Cov}_{\text{w/o}}\left(Y_j, Y_k\right) \neq 0, \quad j, k \in \{1, 2, \ldots, N\},$$

since the sampled IDs are dependent.

Theorem 10.10. Properties of sample mean for SRS without replacement
For SRS without replacement, for $j, k \in \{1, \ldots, n\}$ and $j \neq k$,

$$\mathrm{E}_{\mathrm{w/o}}[Y_j] = \mu, \quad \mathrm{Var}_{\mathrm{w/o}}(Y_j) = \sigma^2, \quad \mathrm{Cov}_{\mathrm{w/o}}(Y_j, Y_k) = -\frac{\sigma^2}{N-1}.$$

Consequently,

$$\mathrm{E}_{\mathrm{w/o}}[\bar{Y}] = \mu, \quad \mathrm{Var}_{\mathrm{w/o}}(\bar{Y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

Definition 10.11. Finite population correction
The factor $(N-n)/(N-1)$ that appears in the variance for sampling without replacement is called the finite population correction.

This factor is less than 1 for $n > 1$, so the sample average has lower variance for sampling without replacement than for sampling with replacement. This fact makes sense intuitively due to the negative correlation between the Y_j, and since it seems redundant to sample the same individual more than once.

Definition 10.12. Strata
Partition the IDs into L subsets, called strata. Index the strata by ℓ, and refer to the ID within the ℓ th stratum by the pair

$$(j, l), \quad j = 1, \ldots, N_l, \quad \ell = 1, \ldots, L,$$

where N_l is the size of the ℓ th stratum. Assume that each $N_\ell \geq 1$ and $\sum_{\ell=1}^{L} N_\ell = N$. Within the ℓ th stratum, the fixed population is

$$y_{1,\ell}, \ldots, y_{N_\ell,\ell}.$$

Define the population quantities within the ℓ th stratum as:

$$\mu_\ell = N_\ell^{-1} \sum_{j=1}^{N_\ell} y_{j,\ell}, \quad \sigma_\ell^2 = N_\ell^{-1} \sum_{j=1}^{N_\ell} (y_{j,\ell} - \mu_\ell)^2, \quad F_\ell(y) = N_\ell^{-1} \sum_{j=1}^{N_\ell} I(y_{j,\ell} \leq y).$$

Theorem 10.13. Reconstructing population-specific quantities from strata
The stratum-specific quantities can be used to produce the corresponding population specific quantities via

$$\mu = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \mu_\ell, \quad \sigma^2 = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \sigma_\ell^2 + \sum_{\ell=1}^{L} \frac{N_\ell}{N} (\mu_\ell - \mu)^2, \quad F(y) = \sum_{\ell=1}^{L} \frac{N_\ell}{N} F_\ell(y).$$

Definition 10.14. Stratified sampling design
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.

Collect all the sampled IDs in the ℓ th stratum as

$$\mathbf{I}_{1:n_\ell,\ell} = (I_{1,\ell}, \ldots, I_{n_\ell,\ell}), \quad \ell = 1, \ldots, L.$$

A sampling design is a stratified sampling design if the sampling is done independently across strata. If $P(I_{j,\ell} = k) = 1/N_\ell$ for all $k = 1, \ldots, N_\ell, j = 1, \ldots, n_\ell$ and $\ell = 1, \ldots, L$, then the design is an equal probability stratified sample.

The outcomes can be used as the input into stratum-specific estimators (of μ_l and $F_l(y)$), and they can be pooled as the following population estimators of μ and $F(y)$:

$$\widehat{\mu}_{\text{strat}} = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \bar{Y}_\ell, \quad \widehat{F}_{\text{strat}}(y) = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \widehat{F}_\ell(y).$$

10.4 Horvitz–Thompson estimator

Definition 10.15. Horvitz-Thompson estimator
The Horvitz-Thompson estimator is a very general way to construct an unbiased estimator of μ, for any sampling design such that for each individual there is a known, positive probability that the individual will be included in the sample.

Let the sampling design be

$$P(I_1 = i_1, \ldots, I_n = i_n).$$

Let

$$C_j = I(I_1 = j) + \cdots + I(I_n = j), \quad j = 1, \ldots, N,$$

be the number of times that ID j is selected for the sample. (So $C_j \leq 1$ if the sampling is without replacement.) Let the inclusion probability of ID j be

$$\pi_j = P(C_j \geq 1) = 1 - P(C_j = 0).$$

Assume that N and π_j are known, with $\pi_j > 0$ for all j. The Horvitz-Thompson estimator of μ is

$$\hat{\mu}_{\text{HT}} = \frac{1}{N} \sum_{j=1}^{N} \frac{I(C_j \geq 1)}{\pi_j} y_j,$$

basically saying that we sum up y_j/π_j over all individuals j that appear at least once in the sample, and then divide by N.

Theorem 10.16. Unbiasedness of Horvitz-Thompson estimator
With notation as above,

$$\mathrm{E}\left[\hat{\mu}_{HT}\right] = \mu.$$

Proof. Simply linearity of expectation. □

10.5 The bootstrap

Definition 10.17. Bootstrap
Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the observed dataset, assumed to be i.i.d. from an unknown CDF F. We make no parametric assumptions about F (parametric versions of the bootstrap also exist and can be useful, but here we will focus on the nonparametric bootstrap). Create a synthetic dataset

$$\mathbf{Y}^* = \left(Y_1^*, \ldots, Y_n^*\right),$$

by performing a simple random sample with replacement from $(y_1, \ldots, y_n)$. Equivalently, let

$$Y_j^* \overset{\text{i.i.d.}}{\sim} \hat{F}, \quad j = 1, \ldots, n$$

where $\hat{F}$ is the ECDF of $\mathbf{y}$:

$$\hat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} I\left(y_j \leq y\right).$$

Then $(Y_1^*, \ldots, Y_n^*)$ is called a bootstrap sample. The bootstrap entails generating some large number B of independent bootstrap samples (each of size n), and then using the bootstrap samples for inferential tasks such as approximating the standard error of an estimator.

Intuitively, population is to sample as sample is to bootstrap sample (that is F is to Y as $\hat{F}$ is to Y^*).

Remark 10.18. Bootstrap confidence intervals
The bootstrap can also be used to construct approximate confidence intervals. Some methods are:

- Normal interval with bootstrap standard error
- Percentile method
- Bootstrap t interval, also known as the studentized bootstrap interval

The first two are intuitive, quick, and dirty. The third is more computationally intensive but generally yields better performance.

Definition 10.19. Permutation tests for hypothesis testing
Suppose that we are comparing two groups, group 0 and group 1, and wish to test whether the distribution that generated the data in group 0 is the same as the distribution that generated the data in group 1. To conduct a permutation test, let the data for group 0 be the observed values of

$$X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F_X,$$

and the data for group 1 be the observed values of

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_Y.$$

Also assume that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent. The CDFs F_X and F_Y are unknown, and no parametric assumptions are made about them. A permutation test for

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

can be conducted as follows.

1. Let T be a test statistic (chosen before looking at the data). For simplicity, assume that T is nonnegative. The test statistic T should be chosen so that large values of T are evidence against H_0. For example, a natural choice is $T = |\bar{Y} - \bar{X}|$. If we are more interested in medians than means, we could instead let T be the absolute difference between the sample medians.

2. Compute the observed value t_0 of T from the data.

3. Generate B random permutations of the data, for some large number B. Each random permutation is a completely random shuffle of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. In particular, this scrambles which data points belong to which group. The random permutations should be a simple random sample without replacement, though for computational convenience a simple random sample with replacement can be used instead.

4. For each of the B random permutations, compute the test statistic T. Call these $t_1, \ldots, t_B$.

5. The p-value is

$$P_0\left(T \geq t_0\right) \approx \frac{1}{B} \sum_{j=1}^{B} I\left(t_j \geq t_0\right),$$

   where P_0 denotes probability under the permutation distribution of T, i.e., the distribution under random shuffles of the data rather than under repeated sampling.

Intuitively, the idea behind the test is that under the null, the X_i's and Y_j's are completely interchangeable, so it would be surprising if the observed value of the test statistic were extreme compared with simulated values of the test statistic obtained by randomly shuffling the data.

11 Experiments and Causality

11.1 Causality

11.2 Causal framework

We assume binary treatment and neglect the possibillity of non-compliance (i.e., when patients are assigned a treatment but forget to take it).

Definition 11.1. Assignment
Consider n patients in a study, labeled from 1 through n. Define the j^{th} assignment

$$W_j \in \{0, 1\}$$

to be 1 if patient j receives the treatment (e.g., new drug) and 0 if patient j receives the control (e.g., standard of care). Collect all the assignments in the study as the treatment assignment vector

$$\mathbf{W} = (W_1, \ldots, W_n) \in \{0, 1\}^n.$$

Definition 11.2. Potential outcome and treatment effect
Let $Y_j(w_1, \ldots, w_n)$ be a potential outcome for patient j, defined to be the (possibly random) outcome for patient j if the assignments for all the patients were $w_1, \ldots, w_n$. There are 2^n potential outcomes for patient j, corresponding to the 2^n possible different assignments.

Definition 11.3. Non-interference assumption
The non-interference assumption says that for each j, the treatment of others has no impact on the outcome for the j^{th} individual:

$$Y_j(w_1, \ldots, w_{j-1}, w_j, w_{j+1}, \ldots, w_n) = Y_j\left(w'_1, \ldots, w'_{j-1}, w_j, w'_{j+1}, \ldots, w'_n\right),$$

for all $(w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_n) \in \{0, 1\}^{n-1}$ and $\left(w'_1, \ldots, w'_{j-1}, w'_{j+1}, \ldots, w'_n\right) \in \{0, 1\}^{n-1}$. Under the non-interference assumption, we will write the potential outcomes in the vastly simpler notation

$$Y_j(w_j),$$

and then write the collections of potential outcomes under the control versus treatment as

$$\mathbf{Y}(0) = \{Y_1(0), \ldots, Y_n(0)\}, \mathbf{Y}(1) \quad = \{Y_1(1), \ldots, Y_n(1)\}.$$

Remark 11.4.
Non-interference is a very powerful assumption. For cancer drugs it may be reasonable to assume that treating one individual does not impact the medical outcome of other individuals. However, if the drug impacts a virus, e.g., giving the patient immunity, then this treatment may not only improve the health of that patient, but also improve the health of other patients to whom that patient may have spread the virus. In such a situation, the non-interference assumption would be violated. In applying the potential outcomes framework in practice, thought is always needs about whether non-interference is a plausible assumption.

Definition 11.5. Treatment effect
The treatment effect for the j^{th} patient of moving from assignment $(w_1, \ldots, w_n)$ to assignment $(w_1', \ldots, w_n')$ is the difference

$$\tau_j = Y_j\left(w_1', \ldots, w_n'\right) - Y_j\left(w_1, \ldots, w_n\right),$$

or under the non-interference assumption,

$$\tau_j = Y_j(1) - Y_j(0).$$

The average treatment (causal) effect for the sample is

$$\bar{\tau} = n^{-1} \sum_{j=1}^{n} \tau_j.$$

In the rather extraordinary situation where $\tau_j = \tau$ for all j and some τ, then the treatment effects are said to be homogeneous; otherwise, they are said to be heterogeneous.
The average causal effect $\bar{\tau}$ in the sample is called a finite sample estimand; it applies only to the specific group of patients in the study. Another causal quantity is the population quantity $\mathrm{E}[\tau_1]$, where the expectation averages over some population.

Theorem 11.6. Switching equation

$$Y_j = Y_j(W_j) = W_j Y_j(1) + (1 - W_j) Y_j(0), \quad j = 1, \ldots, n.$$

This identity just says that if $W_j = 1$ then the actual outcome is $Y_j(1)$, while if $W_j = 0$, then the actual outcome is $Y_j(0)$.

Definition 11.7. SUTVA
The Stable Unit Treatment Value Assumption (SUTVA) is shorthand for the assumptions of non-interference and that the form of the treatment is the same across all units (and likewise for the form of the control). This second assumption is important because we need to know that outcomes do not vary with how the treatment was administered.

Definition 11.8. Assignment mechanism
The assignment mechanism is the joint probability mass function of the assignments given the potential outcomes:

$$P(\mathbf{W} = \mathbf{w} \mid \{\mathbf{Y}(0), \mathbf{Y}(1)\}).$$

This PMF is interesting because it reflects the probabilistic link between the potential outcomes and the assignments. For example, is each individual randomized to treatment or control by flipping a fair coin (independent of all the potential outcomes)? Or does a doctor with decades of experience examine individuals before assignment and then assign an individual to treatment only if the doctor believes the individual would not respond well to the control?

11.3 Ethics of experimentation

11.4 Randomized control trials

Definition 11.9. Randomization
If the assignment mechanism satisfies

$$P(\mathbf{W} = \mathbf{w} \mid \{\mathbf{Y}(0), \mathbf{Y}(1)\}) = P(\mathbf{W} = \mathbf{w}),$$

i.e., the assignments are independent of the potential outcomes, then the assignments have been randomized.

Definition 11.10. RCT
Experiments where the assignments are randomized are called randomized experiments or randomized control trials (RCTs).

Remark 11.11. Causal estimands: finite sample and population-based
It is important to distinguish carefully between the estimands $\bar{\tau}$ and $\mathrm{E}[\tau_1]$. This distinction is similar to the distinction in Chapter 10 between the design-based perspective and the population model-based perspective in sampling.

- 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. The finite sample quantity says, in principle, nothing about the possible causal effect on any units which are not in the study. The analysis is entirely self-contained, making conclusions only about the n units. This style of study is limited but powerful.

  On the one hand, it can be carried out without making strong assumptions such as how the units in the study were sampled from the wider population. On the other hand, it is severely limited by the fact that it provides no general knowledge about how the treatment might work beyond these n units. If we are focusing on a finite sample estimand, we typically take an approach similar to the design-based perspective from Chapter 10, treating the potential outcomes as fixed and letting the randomness be only due to randomness in the assignments.

- The population quantity $\mathrm{E}[\tau_1]$ 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. This kind of inference is of course highly desirable to help inform decisions about, e.g., whether a new drug should be approved and widely prescribed. But it requires stronger assumptions than does inference for the finite sample quantity. If we are focusing on a population estimand, we typically take a model-based approach, building a statistical model for $\{W_1, Y_1(0), Y_1(1)\}, \ldots, \{W_n, Y_n(0), Y_n(1)\}$, such as assuming that they are i.i.d. draws from some parametric or nonparametric model.

11.5 A population-based statistical model for experiments

Definition 11.12. Statistical model for population modeling
In population-based modeling, we assume a statistical model for the triples

$$\{W_1, Y_1(0), Y_1(1)\}, \ldots, \{W_n, Y_n(0), Y_n(1)\},$$

and we assume that $\{W_j, Y_j(0), Y_j(1)\}$ are independent across the subscript j. If this is combined with the randomization assumption, then

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

With randomization, we then have

$$\mathrm{E}[\tau_1] = \mathrm{E}[Y_1(1)] - \mathrm{E}[Y_1(0)] = \mathrm{E}[Y_1 \mid W_1 = 1] - \mathrm{E}[Y_1 \mid W_1 = 0],$$

which is easily estimated.

Proposition 11.13. MLE of $E(\tau_1)$
Suppose our outcomes are binary. Denoting

$$\theta_1 = \mathrm{E}\left[Y_1(1)\right] = \mathrm{E}\left[Y_1 \mid W_1 = 1\right] = P\left(Y_1 = 1 \mid W_1 = 1\right),$$
$$\theta_0 = \mathrm{E}\left[Y_1(0)\right] = \mathrm{E}\left[Y_1 \mid W_1 = 0\right] = P\left(Y_1 = 1 \mid W_1 = 0\right),$$

then we may write

$$\mathrm{E}\left[\tau_1\right] = \theta_1 - \theta_0,$$

with

$$\widehat{\mathrm{E}\left[\tau_1\right]} = \hat{\theta}_1 - \hat{\theta}_0$$

for

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

Proof. Thus let us calculate the MLEs $\hat{\theta}_0$ and $\hat{\theta}_1$. We have the PMF

$$P\left(Y_1 = y_1, W_1 = w_1; \theta_0, \theta_1\right) = P\left(Y_1 = y_1 \mid W_1 = w_1; \theta_0, \theta_1\right) P\left(W_1 = w_1\right),$$

and to calculate the likelihood we may drop $P(W_1 = w_1)$. Thus our likelihood is

$$\begin{aligned}
\log L\left(\theta_0, \theta_1\right) &= \log P\left(\mathbf{Y} = \mathbf{y} \mid \mathbf{W} = \mathbf{w}; \theta_0, \theta_1\right) \\
&= \sum_{j=1}^n \log P\left(Y_j = y_j \mid W_j = w_j; \theta_0, \theta_1\right) \\
&= \sum_{j:w_j=0} \log \theta_0^{y_j}\left(1 - \theta_0\right)^{1-y_j} + \sum_{j:w_j=1} \log \theta_1^{y_j}\left(1 - \theta_1\right)^{1-y_j} \\
&= \sum_{j=1}^n \left(1 - w_j\right) \log \theta_0^{y_j}\left(1 - \theta_0\right)^{1-y_j} + \sum_{j=1}^n w_j \log \theta_1^{y_j}\left(1 - \theta_1\right)^{1-y_j} \\
&= \log L_0\left(\theta_0\right) + \log L_1\left(\theta_1\right)
\end{aligned}$$

where

$$\log L_0\left(\theta_0\right) = \sum_{j=1}^n \left(1 - w_j\right)\left(y_j \log\left(\theta_0\right) + \left(1 - y_j\right) \log\left(1 - \theta_0\right)\right),$$
$$\log L_1\left(\theta_1\right) = \sum_{j=1}^n w_j\left(y_j \log\left(\theta_1\right) + \left(1 - y_j\right) \log\left(1 - \theta_1\right)\right).$$

The result is two log-likelihoods for Bernoulli experiments, with probabilities of success θ_0 and θ_1, and sample sizes $\sum_{j=1}^n \left(1 - w_j\right)$ (the number of control units) and $\sum_{j=1}^n w_j$ (the number of treated units), respectively. Then by the MLE of Bernoullis, our conclusion follows. □

Theorem 11.14. Properties of the MLEs $\hat{\theta}_0, \hat{\theta}_1$
With the above assumptions and notation, we have

$$\mathrm{E}\left[\hat{\theta}_0 \mid \mathbf{W}=\mathbf{w}\right]=\theta_0, \quad \mathrm{E}\left[\hat{\theta}_1 \mid \mathbf{W}=\mathbf{w}\right]=\theta_1,$$

with

$$\mathrm{Var}\left(\hat{\theta}_0 \mid \mathbf{W}=\mathbf{w}\right)=\frac{\theta_0\left(1-\theta_0\right)}{\sum_{j=1}^{n}\left(1-w_j\right)}, \quad \mathrm{Var}\left(\hat{\theta}_1 \mid \mathbf{W}=\mathbf{w}\right)=\frac{\theta_1\left(1-\theta_1\right)}{\sum_{j=1}^{n} w_j},$$

and the Fisher information in the sample is

$$\mathcal{I}_{\mathbf{Y}|\mathbf{W}=\mathbf{w}}\left(\theta_0, \theta_1\right)=\begin{pmatrix} \mathcal{I}_{\mathbf{Y}|\mathbf{W}=\mathbf{w}}\left(\theta_0\right) & 0 \\ 0 & \mathcal{I}_{\mathbf{Y}|\mathbf{W}=\mathbf{w}}\left(\theta_1\right) \end{pmatrix}=\begin{pmatrix} \frac{\sum_{j=1}^{n}\left(1-w_j\right)}{\theta_0\left(1-\theta_0\right)} & 0 \\ 0 & \frac{\sum_{j=1}^{n} w_j}{\theta_1\left(1-\theta_1\right)} \end{pmatrix}.$$

For the MLE of $\mathrm{E}\left[\tau_1\right]$, note that $\hat{\theta}_1, \hat{\theta}_0$ are conditionally independent given $\mathbf{W}$, so

$$\mathrm{E}\left[\hat{\theta}_1-\hat{\theta}_0 \mid \mathbf{W}=\mathbf{w}\right]=\mathrm{E}\left[\tau_1\right], \quad \mathrm{Var}\left(\hat{\theta}_1-\hat{\theta}_0 \mid \mathbf{W}=\mathbf{w}\right)=\frac{\theta_1\left(1-\theta_1\right)}{\sum_{j=1}^{n} w_j}+\frac{\theta_0\left(1-\theta_0\right)}{\sum_{j=1}^{n}\left(1-w_j\right)}.$$

Remark 11.15. Population inference for $E(\tau)$
Approximate inference on $\widehat{\mathrm{E}\left[\tau_1\right]}$ can be based on the usual approximate pivot

$$\left.\frac{\widehat{\mathrm{E}\left[\tau_1\right]}-\mathrm{E}\left[\tau_1\right]}{\sqrt{\frac{\hat{\theta}_1\left(1-\hat{\theta}_1\right)}{\sum_{j=1}^{n} w_j}+\frac{\hat{\theta}_0\left(1-\hat{\theta}_0\right)}{\sum_{j=1}^{n}\left(1-w_j\right)}}}\right| \left(\mathbf{W}=\mathbf{w}\right) \dot{\sim} \mathcal{N}(0,1),$$

where for the asymptotics we suppose that both $\sum_{j=1}^{n} w_j$ and $\sum_{j=1}^{n}\left(1-w_j\right)$ get large. The approximate pivot can be used to obtain population confidence intervals for $\mathrm{E}\left[\tau_1\right]$ or to perform population testing about the value of $\mathrm{E}\left[\tau_1\right]$ (i.e., to test the population null hypothesis of no causality, $\mathrm{E}\left[\tau_1\right]=0$).

11.6 A finite sample approach for experiments

Definition 11.16. Finite sample approach
Again, the focus is on the outcomes, the assignments, and the potential outcomes for the n units, which we collect together as $\mathbf{Y}, \mathbf{W}, \mathbf{Y}(1), \mathbf{Y}(0)$. We will now make a big switch, carrying out inference conditional on the potential outcomes

$$\mathbf{Y}(1) = \mathbf{y}(1), \quad \mathbf{Y}(0) = \mathbf{y}(0)$$

where $\mathbf{y}(1) = (y_1(1), \ldots, y_n(1))$ and $\mathbf{y}(0) = (y_1(0), \ldots, y_n(0))$. That is, we are viewing the potential outcomes as fixed, even though half of them will never be observed. Once we have conditioned, we move to the using the finite sample approach and make no assumptions about how the potential outcomes are generated. As in the design-based framework from Chapter 10, we make no assumptions about the finite population from which are sampling; instead, inference is driven by the sampling mechanism.

Typically the inferential focus in this approach is the finite sample average treatment effect

$$\bar{\tau} = n^{-1} \sum_{j=1}^{n} \{y_j(1) - y_j(0)\}.$$

Note that once we have conditioned on the potential outcomes $\mathbf{Y}(1), \mathbf{Y}(0)$, the only things left to write down probabilities for are $\mathbf{Y}$ and $\mathbf{W}$. But once we also condition on $\mathbf{W}$, the $\mathbf{Y}$ are determined by the switching equation

$$Y_j = W_j Y_j(1) + (1 - W_j) Y_j(0).$$

Theorem 11.17. Finite sample approach
We focus on the case that our assignments are randomized independently over the units:

$$P(\mathbf{W} = \mathbf{w}) = \prod_{j=1}^{n} P\left(W_j = w_j\right).$$

Defining

$$G_1 = \frac{W_1 Y_1}{\mathrm{E}\left[W_1\right]} - \frac{\left(1 - W_1\right) Y_1}{\mathrm{E}\left[1 - W_1\right]},$$

Then

$$\mathrm{E}\left[G_1 \mid \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}\right] = y_1(1) - y_1(0),$$

and

$$\mathrm{Var}\left(G_1 \mid \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}\right) = \frac{y_1^2(1)}{\mathrm{E}\left[W_1\right]} + \frac{y_1^2(0)}{1 - \mathrm{E}\left[W_1\right]} - \left\{y_1(1) - y_1(0)\right\}^2.$$

These results suggest the estimator

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

conditionally unbiased given the potential outcomes and having conditional variance

$$\mathrm{Var}\left(\hat{\tau}_{\mathrm{MoM}}(\mathbf{W}) \mid \mathbf{Y}(1) = \mathbf{y}(1), \mathbf{Y}(0) = \mathbf{y}(0)\right) = \frac{1}{n^2} \sum_{j=1}^{n} \left(\frac{y_j^2(1)}{\mathrm{E}\left[W_j\right]} + \frac{y_j^2(0)}{1 - \mathrm{E}\left[W_j\right]} - \left\{y_j(1) - y_j(0)\right\}^2 \right),$$

which will typically become small as n gets large.

In the finite sample setting, there are two widely-used choices of null hypothesis.

Definition 11.18. Fisher null and randomization test
The first is the Fisher null, which states that the treatment effect is zero for every individual in the sample:

$$H_0 : \tau_j = 0, j = 1, \ldots, n,$$

where $\tau_j = y_j(1) - y_j(0)$. This is tested against the composite alternative $H_1 : \sum_{j=1}^{n} |\tau_j| > 0$, i.e., there is a causal effect on at least one patient.

Definition 11.19. Neyman null
The second finite sample null is the Neyman null, which states that the average causal effect across the sample is zero:

$$H_0 : \bar{\tau} = 0$$

which is tested against the alternative $H_1 : \bar{\tau} \neq 0$. Of course, the Fisher null implies the Neyman null, but not conversely.

11.7 Observational studies

Definition 11.20. Observational studies
In observational studies, we have only data for which the assignments of the individuals were outside of the researcher's control. Observational data are often used to provide descriptions or summaries, as well as generate predictions through predictive regressions. When we seek causal conclusions, the task is much harder. In RCTs, we used assignments, potential outcomes, and non-interference to define causal terms and randomization to drive causal inference. In observational studies we can still phrase causal statements using potential outcomes and non-interference, but it is fairly rare that we see useful forms of randomization appearing in nature or through societal actions.

This approach to observational studies is to condition on some covariates, and then assume that given those covariates the assignments are independent of the potential outcomes. We will see that causal inference is possible under these assumptions, and we can use statistical techniques analogous to those we deployed to analyze the RCTs. However, the conditional independence assumption is often hard to justify or verify. Write the covariates as

$$X = (X_1, ..., X_n),$$

where the covariates must be pretreatment variables. As with RCTs, we will take a potential outcomes approach and assume non-interference. But now we will generalize the assignment mechanism to

$$P(\mathbf{W} \mid \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{X}),$$

so the assignment probabilities can depend on the covariates. The assignments are called unconfounded or ignorable if they are conditionally independent of the potential outcomes, given the covariates:

$$[\mathbf{W} \perp\!\!\!\perp \{\mathbf{Y}(0), \mathbf{Y}(1)\}] \mid \mathbf{X}.$$