Statistics 211: Statistical Inference I

Eric K. Zhang
ekzhang@college.harvard.edu

Fall 2021

Abstract

These are notes for Harvard's *Statistics 211*, a graduate-level class taught by Lucas Janson in Fall 2021, targeted at first-year PhD students. The main focus of this class is on frequentist methods for statistical inference, i.e., how to draw mathematical conclusions from sample data based on likelihoods from a parametric model.

Course description: Foundations of frequentist and Bayesian inference, and decision theory. Likelihood, sufficiency, and ancillarity. Point estimation, unbiasedness, maximum likelihood, method of moments, minimum-variance. Parametric and non-parametric hypothesis testing, confidence intervals. Selective inference: multiple testing, familywise error rate, false discovery rate. Bayesian inference, conjugate priors, credible intervals. Admissibility, Stein's phenomenon, empirical Bayes. Time permitting: post-selection inference and the bootstrap.

Contents

1 September 1st, 2021

This is the first lecture of the course. We will discuss logistics, an overview of the class, and a bit of statistical philosophy.

1.1 Class Overview

Our professor is Lucas Janson, who is a professor in statistics at Harvard focusing on high-dimensional inference and statistical machine learning problems. Our teaching fellows are Yufan Li, Biyonka Liang, and Yash Nair. The prerequisites for this class are an undergraduate-level inference class, such as Stat 111, and some exposure to graduate-level probability (e.g., CLT, Jensen's inequality, Slutsky's theorem, continuous mapping theorem, convergence).

Most of the material in this class will come from the provided notes, but we also have three textbooks [CB21, LR06, LC06] for supplemental use. Lucas will also post modern literature that is related to each lecture topic at the end of class. Each lecture will include an "active learning" component where students think individually about a problem.

Now we discuss course topics. In most sections of this class, there will be some data y and parameter θ, and our goal is to determine some estimate of θ that is as close to the true value as possible, along with a measure of uncertainty about our estimate. The main topics are:

1. **Point estimation:** How can we find an estimate for θ?

2. **Confidence intervals:** How confident is our estimate, under Bayesian or frequentist terms?

3. **Hypothesis testing:** Is there significant evidence of a hypothesis $\theta = \theta_0$ being false?[1]

4. **Selective inference:** How can we test multiple hypotheses simultaneously?

5. **Decision making:** If we assign costs to each type of error, what estimator minimizes cost?

6. **Prediction:** If I collect new data according to $y = f_\theta(x)$, what would it look like?

Lucas emphasizes that this is a course in *good statistical thinking*, and although not all enrollees are the target audience of first-year PhD students, he hopes that the class content is broadly useful in many contexts.

1.2 Statistical Philosophy

We first discuss the difference between Bayesian and frequentist interpretations of inference. In Bayesian inference, our model is that θ is random and y is fixed. However, in frequentist thinking, we interpret our parameters as y being random and θ being fixed.

There are different techniques and notations for Bayesian and frequentist methods, like confidence intervals versus credible intervals. However, in practice, both approaches will achieve similar results, so we should not draw a hard distinction.[2] Let's pivot to one of Lucas's core beliefs.

Proposition 1.1 (First Law of Statistics)**.** *The more you know about your data's distribution, the more you can infer about that distribution from the data.*

[1] This includes *non-parametric hypothesis testing*, such as asking if a distribution is symmetric.

[2] Lucas tells us to take Stat 213 for a theoretical justification of this.

In other words, when you know some facts about the distribution, such as the fact that it is Gaussian with variance 1, you can produce better inference methods for quantities like the mean μ. If you knew nothing about the distribution, it could be the case that the mean does not exist at all, as is the case with the Cauchy distribution $p(x) = \frac{1}{\pi(1+x^2)}$.

Of course, this statement is a tautology, since a more knowledgeable individual could just pretend to know less about the data. However, it has a couple important conceptual consequences to the way we approach statistics problems:

1. When we analyze data, first ask what we know about the distribution, then ask how to use that knowledge to learn as much about the data as possible. This means that **good statistics practice is tied to knowledge of its domain of application.**

2. Domain knowledge allows us to obtain assumptions about the distribution of data being modeled, which are crucial to determining the type of inference method to use.

3. Bayesian and frequentist inference are different ways of encoding domain knowledge. Bayesian inference is better at encoding assumptions about where θ will be in the parameter space, while frequentist inference is better at problems where we do not know the prior.

In scientific literature, there are many statistical methods that are applied over and over again. We will not learn names of domain methods in this class, but they will typically be special cases of inference techniques we learn in this class. More importantly, we will understand **where these methods come from and how to extend them**.

In statistics, methodological innovation comes from doing better with the same set of assumptions, or by leveraging more assumptions within a method. Neither Bayesian nor frequentist approaches are perfect, since approximations will never exactly match the truth. Therefore, we can summarize the relevance to this class as follows:

- In Bayesian inference, there is a rigid way of specifying domain knowledge. Once we specify the domain knowledge, we're basically done, as we just examine the posterior. Our main challenge is **computational** rather than methodological, so we will not focus on Bayesian inference in this class. It's too "elegant and simple" to talk about.

- In frequentist methods, this is not true, and different methods can be better at certain parts of the parameter space. Therefore, it is often the case that complex problems do not have optimal frequentist inference methods.

That concludes the first lecture of the course. Next week, we will continue discussing frequentist and Bayesian philosophies and sufficiency.

2 September 8th, 2021

Today we discuss likelihood and relevant notation, sufficiency, and unbiased estimation (the last topic, only if time permits).

2.1 Likelihood and Notation

We start by reiterating the difference between probability and inference problems. In probability tasks, we are given a fixed distribution and are asked the probabilities of certain events in that distribution. However, in inference, the task is reversed: given certain observations from an unknown distribution, find the likelihood of the distribution having certain parameters.

Definition 2.1 (Parametric model)**.** Suppose that we have a *model* $f_\theta(y)$ representing the probability mass of some distribution at y, which is defined on some base measure.[3] We call $f_\theta(y)$ *parametric* if θ is finite-dimensional and *nonparametric* otherwise.

Now we will set up the likelihood function, which is key in statistical inference.

Definition 2.2 (Likelihood function)**.** Given a model $f_\theta(y)$ and observed data y, the *likelihood function* on parameter space is defined by

$$L(\theta) = f_\theta(y).$$

We notate the *log-likelihood* by $\ell(\theta) = \log L(\theta)$.

Oftentimes, our observations Y will be multi-dimensional, and we write them as a vector $Y = (Y_1, \dots, Y_n)$. Also, in addition to the model probability density $f_\theta(y)$, we notate the cumulative distribution function by a capital letter $F_\theta(y) = \Pr_\theta(Y \le y)$.

Note. We often parameterize probability and expectation by subscripting the relevant operator with the parameter θ. For example, we could write $\Pr_\theta(Y \in A)$ or $\mathbf{E}_\theta\left[Y\right]$.

After getting some basic notation out of the way, we can now start the first "unit" of the course, which is about sufficiency.

2.2 Sufficiency

Sufficiency is a concept related to how useful a set of observations is for predicting the parameters of a distribution. Essentially, a sufficient statistic is "good enough" for getting all the statistical information about parameters from the underlying set of variables.

Definition 2.3 (Sufficiency)**.** Given random variables $Y = (Y_1, \dots, Y_n) \sim f_\theta(y)$, we say that a statistic $T(Y)$ computed from Y is *sufficient* for θ if for all $\theta_1, \theta_2 \in \Theta$ and $A \subset \mathcal{Y}$, we have

$$\Pr_{\theta_1}\big(Y \in A \mid T = T(Y)\big) = \Pr_{\theta_2}\big(Y \in A \mid T = T(Y)\big).$$

There are a few other equivalent definitions based on notions from other fields:

- (Information Theory). The chain $\theta \to T \to Y$ is Markovian, i.e., $Y \perp\!\!\!\perp \theta \mid T$.
- (Bayesian). The conditional distribution of $\theta \mid T$ is the same as $\theta \mid Y$, for any prior on θ.

[3] Note that this definition supports both continuous and discrete random variables.

- (Measure Theory). The statistic $T(Y)$ and variable Y generate the same σ-algebra.

Example 2.4. Suppose that Y_1, Y_2 are i.i.d. $\sim \mathrm{Pois}(\lambda)$, and $T(Y) = Y_1 + Y_2$. Then, the conditional distribution is $(Y_1, Y_2) \mid T(Y) \sim \mathrm{Mult}(T, (\frac{1}{2}, \frac{1}{2}))$. We claim that $T(Y)$ is sufficient for the Poisson rate parameter λ. This is because, following the definition,

$$\begin{aligned}
\Pr_\lambda(Y_1 = k, Y_2 = t - k \mid Y_1 + Y_2 = t) &= \frac{\Pr_\lambda(Y_1 + Y_2 = t \mid Y_1 = k)\Pr_\lambda(Y_1 = k)}{\Pr_\lambda(Y_1 + Y_2 = t)} \\
&= \frac{\frac{\lambda^{t-k}e^{-\lambda}}{(t-k)!} \cdot \frac{\lambda^k e^{-\lambda}}{k!}}{\frac{(2\lambda)^t e^{-2\lambda}}{t!}} \\
&= \binom{t}{k} \Big/ 2^t.
\end{aligned}$$

This last expression does not depend on λ, so we have established sufficiency.

Example 2.5. If $Y_1, \dots, Y_n$ are i.i.d. $\sim \mathrm{Bern}(p)$, then $T(Y) = Y_1 + \cdots + Y_n$ is a sufficient statistic for p. We can verify this property in a similar manner to the above example.

Oftentimes, we may have an intuitive suspicion that a statistic is sufficient, but this requires algebraic verification to ensure that the property actually holds.

Exercise 2.1 (Pencil problem)**.** You have two minutes to think about these questions.

1. If T is sufficient, and g is injective, is $g(T)$ sufficient?
2. If $Y_i \sim \mathrm{Bern}(p_i)$ for $i = 1, 2$, is $Y_1 + 2Y_2$ sufficient for (p_1, p_2)?

Now, here is an intuitive but nontrivial statement about probability densities of random variables with a sufficient statistic.

Theorem 2.6 (Factorization theorem)**.** *A statistic $T(Y)$ is sufficient for θ if and only if the joint density $f_\theta(y)$ can be written as*

$$f_\theta(y) = g_\theta(T(y)) \cdot h(y).$$

Here, the density g_θ of the sufficient statistic is allowed to depend on θ, but the density $h(y)$ is not allowed to vary. In terms of log-likelihood, this is

$$\ell(\theta) = \log(g_\theta(T)) + \log(h(y)).$$

Proof. This theorem holds in generality, but for simplicity of proof, here assume that Y is a discrete random variable. Let's first handle the easier direction, which is the "if" statement. Following the definition of sufficiency, note that

$$\begin{aligned}
\Pr_\theta(Y = y \mid T(Y) = T(y)) &= \frac{\Pr_\theta(Y = y)}{\Pr_\theta(T(Y) = T(y))} \\
&= \frac{g_\theta(T(y))h(y)}{\sum_{y' \mid T(y') = T(y)} g(T(y'))h(y')} \\
&= \frac{h(y)}{\sum_{y'} h(y')}.
\end{aligned}$$

This does not depend on the parameter θ, so we are done. On the other hand, in the inverse direction, we can use conditional probability and the definition of sufficiency to get

$$\begin{aligned}
f_\theta(y) &= \Pr_\theta(Y = y) \\
&= \Pr_\theta(Y = y, T(Y) = T(y)) \\
&= \Pr_\theta(Y = y \mid T(Y) = T(y)) \cdot \Pr_\theta(T(Y) = T(y)).
\end{aligned}$$

The former is a function free of θ, while the latter is a function of $T(y)$, and hence we have completed the factorization. □

Note. In the above factorization, neither g_θ nor h are unique, nor are they required to be actual probability measures with physical meaning whose masses sum to 1. It helps me to intuitively think of g_θ as a measure over T and h as a probability within each quotient equivalence class of the form $\{y \in Y \mid T(y) = t\}$. In the statement of the theorem, h is more of a conditional density $Y \mid T$ shoved into the shape of a function on all of Y.

Here's another example of sufficiency, with applications to many fundamental distributions.

Example 2.7 (Exponential family)**.** Let $Y_1, \ldots, Y_n$ be i.i.d. $\sim f_\theta(y)$, belonging to an exponential family

$$f_\theta(y) = \exp\{\eta(\theta)T(y) - \psi(\eta(\theta))\}h(y).$$

Here, η is called the natural parameter of the family. This has joint density

$$\prod_{i=1}^n f_\theta(y_i) = \exp\left\{\eta(\theta)\sum_{i=1}^n T(y_i) - n\psi(\eta(\theta))\right\}\prod_{i=1}^n h(y_i).$$

We will continue this example in the next lecture.

3 September 13th, 2021

Today we will first start by redefining exponential families in a way that is hopefully more intuitive. We will then talk about unbiased estimation and minimal sufficiency.

3.1 Exponential Families and Sufficiency

Exponential families are an important topic that is related to both sufficiency and this class in general. It will show up multiple times, so we will spend some time clearly motivating and defining it from a clean slate.

Definition 3.1 (Exponential family)**.** let $Y_1, \ldots, Y_n$ be i.i.d. $\sim f_\theta(y)$, where $f_\theta(y) \propto e^{\theta T(y)} h(y)$ for some functions T and h. Then, we define the normalizing factor $\psi(\theta)$ to be

$$e^{\psi(\theta)} = \int_{-\infty}^{\infty} e^{\theta T(y)} h(y) \, \mathrm{d}y.$$

Under this definition, we can write the density as $f_\theta(y) = \exp\{\theta T(y) - \psi(\theta)\} h(y)$. We call f_θ an *exponential family* with respect to some parameter η such that $\theta = \theta(\eta)$.

In an exponential family, the joint probability density of $Y_1, \ldots, Y_n$ is

$$f_\theta(\overline{y}) = \exp\left\{ \theta \sum_{i=1}^{n} T(y_i) - n\psi(\theta) \right\} \prod_{i=1}^{n} h(y_i).$$

By Theorem 2.6, since θ only appears in the above expression once, we know that $\sum_{i=1}^{n} T(Y_i)$ is a sufficient statistic for θ.

Note. This sufficiency is a powerful result, since exponential families are very general and include many of the standard continuous distributions in statistics. It means that **we can find a sufficient statistic any exponential family by just adding together the values of $T(Y_i)$.**

To generalize further, when θ is a k-dimensional parameter vector, the probability density for an exponential family is given by

$$f_\theta(y) = \exp\left\{ \sum_{j=1}^{k} \theta_j T_j(y) - \psi(\theta) \right\} h(y).$$

In the k-dimensional scenario, the analogous k-vector is a sufficient statistic:

$$\left(\sum_{i=1}^{n} T_1(Y_i), \sum_{i=1}^{n} T_2(Y_i), \ldots, \sum_{i=1}^{n} T_k(Y_i) \right).$$

The Normal, Binomial, Gamma, and Poisson distributions are all exponential families. Note that our analysis above has elided mention of the transformation $\eta(\theta)$, which turns the parameter θ that we care about into the *natural* parameter η.

Example 3.2 (Normal distribution is an EF)**.** Recall that the normal distribution $\mathcal{N}(\mu, \sigma^2)$ has probability density

$$f_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

This is an exponential family with two parameters. The natural parameters are

$$\eta(\mu, \sigma^2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right).$$

The exponential maps are $T(y) = (y, y^2)$, and then the density can be written as

$$f_{\mu,\sigma^2}(y) = \exp\left\{\sum_{j=1}^{2} \eta_j(\mu, \sigma^2) T_j(y) - \psi(\eta)\right\}.$$

This is called *canonical form.* It shows that the sum of $T(y)$, i.e., the first and second empirical moments of the data, are a sufficient statistic for $\mathcal{N}(\mu, \sigma^2)$.

Generally, if we see a new distribution and are asked to find a sufficient statistic, we should try to write that distribution as an exponential family. This will provide a clear and methodical way to find a sufficient statistic. Furthermore, we will see later on that some exponential families have a unique sufficient statistic governed by the natural law.

Now, let's go over an example of sufficiency outside of the exponential family framework.

Example 3.3 (Uniform distribution sufficient statistic)**.** Suppose that $Y_1, \ldots, Y_n$ are i.i.d. $\sim$ Unif$[0, \theta]$. Let $Y_{(1)}, \ldots, Y_{(n)}$ be the order statistics. Then, the joint uniform density is

$$f_\theta(y) = \frac{1}{\theta^n} 1_{\{y_{(n)} \leq \theta\}} 1_{\{y_{(1)} \geq 0\}}.$$

Then, by Theorem 2.6, we immediately conclude that $Y_{(n)}$ is sufficient.

It actually turns out that order statistics are a more general technique.

Example 3.4 (Order statistics are sufficient)**.** Let f_θ be any density with respect to the Lebesgue measure, parameterized by some scalar θ. Then, if $Y_1, \ldots, Y_n$ are i.i.d. $\sim f_\theta$, then the order statistics $Y_{(1)}, \ldots, Y_{(n)}$ are a nontrivial sufficient statistic for θ. This is because the joint density is

$$\prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} f(y_{(i)}).$$

3.2 Unbiased Estimation

Estimation is a common problem in statistics, and we will place it in our inference framework.

Definition 3.5 (Unbiased estimator)**.** For an estimated quantity $g(\theta)$, we say that an estimator $T(Y)$ is *unbiased* if $\mathbf{E}_\theta\left[T(Y)\right] = g(\theta)$ for all values of θ.

Unbiased estimators tend to be "good" in terms of mean-squared error, since we can decompose the error into a bias and variance term like

$$\begin{aligned}\mathbf{E}_\theta\left[(T(Y) - g(\theta))^2\right] &= \mathbf{E}_\theta\left[(T(Y) - \mathbf{E}_\theta\left[T(y)\right])^2\right] + (\mathbf{E}_\theta\left[T(Y)\right] - g(\theta))^2 \\ &= \mathbf{Var}_\theta\left[T(Y)\right] + \text{Bias}^2.\end{aligned}$$

Now, we introduce a famous theorem that connects sufficient statistics to unbiased estimation.

Theorem 3.6 (Rao-Blackwell)**.** *Let $W(Y)$ be an unbiased estimator of $g(\theta)$ and T be a sufficient statistic for θ. Consider the estimator $\psi(T) = \mathbf{E}_\theta\left[W(Y) \mid T\right]$.*[4] *Then,*

1. $\mathbf{E}_\theta\left[\psi(T)\right] = g(\theta)$.
2. *ψ is "better" than W, meaning that $\mathbf{Var}_\theta\left[\psi(T)\right] \le \mathbf{Var}_\theta\left[W(Y)\right]$, with the equality case being when $\psi(T) = W$ for all θ.*

Proof. The first part follows directly from the law of iterated expectation, since

$$g(\theta) = \mathbf{E}_\theta\left[W(Y)\right] = \mathbf{E}_\theta\left[\mathbf{E}_\theta\left[W(Y) \mid T\right]\right] = \mathbf{E}_\theta\left[\psi(T)\right].$$

The second part follows from the law of total variance, since

$$\mathbf{Var}_\theta\left[W(Y)\right] = \mathbf{E}_\theta\left[\mathbf{Var}_\theta\left[W(Y) \mid T\right]\right] + \mathbf{Var}_\theta\left[\mathbf{E}_\theta\left[W(Y) \mid T\right]\right] \ge \mathbf{Var}_\theta\left[\psi(T)\right].$$

□

Intuitively, this theorem means that if we average the value of our unbiased estimator across fibers of a sufficient statistic, we can essentially uniformly reduce the variance of that estimator.

3.3 Minimal Sufficiency

Recall that if a statistic T is sufficient for θ, then any injective function $f(T)$ is also sufficient for θ. Some sufficient statistics convey more information than others, but we often want less information to be encoded in our statistic, rather than more.

Definition 3.7 (Minimal sufficient statistic)**.** A sufficient statistic T is called *minimal* if T is a function of any other sufficient statistic.

We present some examples below, although we do not have time to justify them:

- If $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, then
$$\left(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2\right)$$
is a minimal sufficient statistic. Equivalently, this means that the empirical mean and standard deviations (without the $n-1$ correction) are minimal sufficient.
- If $Y_1, \ldots, Y_n \sim \text{Unif}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, then $(Y_{(1)}, Y_{(n)})$ is minimal sufficient.
- If $Y_1, \ldots, Y_n \sim f$, where f is any density, then $(Y_{(1)}, \ldots, Y_{(n)})$ is minimal sufficient.

In general, minimal sufficient statistics are unique if they exist, up to bijective maps. Furthermore, they almost always exist, except for a few pathological cases.

Theorem 3.8. *Let Y be an i.i.d. vector of data with $Y_i \sim f_\theta$, and let $T(Y)$ be a sufficient statistic for θ. If for all pairs x, y such that $\frac{f_\theta(x)}{f_\theta(y)}$ is free of θ, $T(x) = T(y)$, then T is minimal sufficient.*

We'll talk more about this theorem in the next class.

[4] Notice that this is only a reasonable definition when T is a sufficient statistic, as otherwise it would depend on θ.

4 September 15th, 2021

Today we will finish talking about minimal sufficiency, introduce completeness, and discuss optimal unbiased estimation.

4.1 Minimal Sufficiency (cont.)

First, we prove the theorem from the end of last lecture. It should hopefully be fairly intuitive based on the definition of minimal sufficiency, but the formal details require some thinking.

Proof of Theorem 3.8. Let T' be another sufficient statistic for θ. Then, for any x and y such that $T'(x) = T'(y)$, we have by the factorization theorem that

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{g_\theta(T'(x))h(x)}{g_\theta(T'(y))h(y)} = \frac{h(x)}{h(y)}.$$

This is free of θ, so by the minimal sufficiency assumption, $T(x) = T(y)$. Therefore, we conclude that $T(x)$ is a function of $T'(x)$ because its fibers are at least as coarse. □

Intuitively, another way of thinking about minimal sufficiency is that the fibers of T exactly encode all dependencies of the relative distribution mass on θ. Let's provide an example of using this theorem to prove minimal sufficiency.

Example 4.1. Let $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$. Then, using the normal distribution density,

$$\frac{f_\theta(x)}{f_\theta(y)} = \exp\left[-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_i x_i - \sum_i y_i\right)\right].$$

This quotient of probability densities is independent of our parameters (μ, σ) if and only if the first and second empirical moments for x and y are equal. Therefore, $T(y) = \left(\sum_i y_i^2, \sum_i y_i\right)$ is minimal.

4.2 Complete Sufficiency

There is a different but related special case of sufficiency that we describe now.

Definition 4.2 (Complete sufficient statistic)**.** A sufficient statistic $T(\theta)$ for θ is called *complete* if the only unbiased estimator of zero that is a function of T is the zero function. More precisely, if for all θ,

$$\mathbf{E}_\theta\left[h(T(Y))\right] = 0,$$

then $h(T) = 0$ almost surely.

This means that any function whose expectation is independent of θ must essentially destroy all information about T.

Example 4.3. Consider i.i.d. Bernoulli random variables. Consider $T(y) = \sum_{i=1}^n y_i$, the sum of these random variables, which is a sufficient statistic for the Bernoulli parameter p. It is also complete sufficient, since if there is a function $h(T)$ such that $\mathbf{E}_p\left[h(T)\right] = 0$ for all $p \in (0, 1)$, then

$$\sum_{k=0}^n h(k)\binom{n}{k} p^k (1-p)^{n-k} = 0.$$

This is an n-th degree polynomial in $\frac{p}{1-p} \in (0, \infty)$. For this to vanish at all values of p, each coefficient must be zero, so $h(k) = 0$ for all k.

We will show very soon that complete sufficiency implies minimal sufficiency, except in pathological cases. The converse does not hold.

Example 4.4. Consider $Y_1, \dots, Y_n$ to be i.i.d. $\sim \mathrm{Unif}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, for some $\theta \in \mathbb{R}$. Then, as shown earlier, $T(Y) = (Y_{(1)}, Y_{(n)})$ is minimal sufficient. However, T is not complete sufficient because if $h(T) = Y_{(n)} - Y_{(1)}$, then $h(T) \mid \theta \sim \mathrm{Beta}(n-1, 2)$. This means that

$$\mathbf{E}_\theta \left[h(T) - \frac{n-1}{n+1} \right] = 0,$$

so T is not complete sufficient.

Proposition 4.5. *If a minimal sufficient statistic for some parameter θ exists, then any complete sufficient statistic for θ is also minimal.*

Proof. Let T and M be sufficient statistics such that T is complete and M is minimal. We would like to prove that T is also minimal. The key observation is to consider the function

$$h(T) = \mathbf{E}_\theta \left[T \mid M \right] - T.$$

This is well-defined, since T is a function of M because M is a minimal sufficient statistic. Note that $\mathbf{E}_\theta \left[h(T) \right] = 0$ by the law of iterated expectation. Since T is complete, we know that $h(T) = 0$ almost surely, and the equality case of Theorem 3.6 implies that T is a function of M. □

There is one more big result, which we will not prove.

Proposition 4.6. *Any minimal sufficient statistic of an exponential family is complete, except in pathological cases.*

Proof. Consult the course notes for a citation about this result. □

4.3 Optimal Unbiased Estimation

An important concept in statistical inference is the *uniformly minimum variance unbiased estimator (UMVUE)*, which is the "best" unbiased estimator for a statistic in the variance sense. Here is the big theorem that ties everything together and motivates why we care about complete sufficiency.

Theorem 4.7 (Lehmann-Scheffé)**.** *An unbiased estimator of $g(\theta)$ that is a function of a complete sufficient statistic is the unique UMVUE.*

Proof. Let W and $\tilde{W}$ be two unbiased estimators of $g(\theta)$, and let T be a complete sufficient statistic for θ. Then, $\phi(T) = \mathbf{E}_\theta \left[W \mid T \right]$ and $\tilde{\phi}(T) = \mathbf{E}_\theta \left[W \mid T \right]$. By Rao-Blackwell (Theorem 3.6), ϕ and $\tilde{\phi}$ are unbiased estimators with variance no greater than the variances of W and $\tilde{W}$, respectively.

Let $h(T) = \phi(T) - \tilde{\phi}(T)$, which means that $\mathbf{E}_\theta \left[h(T) \right] = 0$. By complete sufficiency, $\phi(T) = \tilde{\phi}(T)$ almost surely. Now, assume for the sake of contradiction that W and $\tilde{W}$ are both UMVUEs, and that W is a function of T, but $\tilde{W}$ is not a function of T. Then, $W = \phi$, but

$$\mathbf{Var}_\theta \left[W \right] = \mathbf{Var}_\theta \left[\phi \right] = \mathbf{Var}_\theta \left[\tilde{\phi} \right] \leq \mathbf{Var}_\theta \left[\tilde{W} \right] < \mathbf{Var}_\theta \left[W \right].$$

This is a contradiction, so any unbiased estimator of $g(\theta)$ is a UMVUE. Furthermore, this is unique, since we can just apply the Rao-Blackwell sledgehammer to arrive at a contradiction if multiple UMVUEs exist. □

5 September 20th, 2021

Today, we finish our discussion of the Lehmann-Scheffé theorem, then we introduce the concepts of ancillary statistics and Basu's theorem.

5.1 More on Lehmann-Scheffé

Recall that Proposition 4.5 implies that if the minimal and complete sufficient statistics both exist, then they are equivalent and unique. It is **very rare** that the MSS does not exist and **somewhat rare** that the CSS does not exist. Furthermore, no sufficient statistic in an exponential family is complete except curved exponential families.

Example 5.1. Let $Y_1, \dots, Y_n \sim \text{Pois}(\lambda)$. Then, we know that $T = \sum_{i=1}^n Y_i$ is a complete sufficient statistic for λ. Note that Y_1 is an unbiased estimator for λ, and $Y_1 \mid T \sim \text{Bin}(T, \frac{1}{n})$, so

$$\mathbf{E}_\lambda\left[Y_1 \mid T\right] = \frac{T}{n}$$

is an unbiased estimator of λ that is a function of the CSS T, so it is the unique UMVUE for λ.

Note how powerful Lehmann-Scheffé was in the above example, as it made it super easy to prove that the average of the samples is the UMVUE for the rate parameter λ, a nontrivial fact!

Example 5.2. There are some model classes for which no unbiased estimators exist, and there are also some classes for which the UMVUE is ridiculous. Let $Y \sim \text{Pois}(\lambda)$, but let $g(\lambda) = e^{-2\lambda}$. Then, $T(y) = (-1)^y$ is a UMVUE for $g(\lambda)$, even though it is clearly an absurd estimator.

Now, here's a quick pencil exercise to test our understanding of unbiased estimators.

Exercise 5.1 (Pencil problem)**.** Find a nontrivial model $\{f_\theta : \theta \in \Theta\}$ such that Θ contains an open subset of $\mathbb{R}$, and there exists a biased estimator $B(Y)$ of the mean of $\mathbf{E}_\theta\left[Y\right]$ such that $B(Y)$ always has smaller mean-squared error than the sample mean, or UMVUE?

Proof. Yes, consider the model $Y \sim \mathcal{N}(\theta, 1)$, where $\theta \in \Theta = [0, \infty)$. Then, the sample mean $\overline{y}$ has strictly worse mean-squared error than the thresholded quantity $\max(0, \overline{y})$. □

5.2 Ancillary Statistics and Basu's Theorem

So far, we have been dealing with problems in this class where the model is already known. However, sometimes we don't know the model (e.g., in goodness-of-fit tests). The following definition will be useful in cases where we want to do hypothesis testing on whether a model is satisfied.

Definition 5.3 (Ancillary statistic)**.** A statistic $A(Y)$ is called *ancillary* for parameter θ if its distribution does not depend on θ.

Here are some examples of ancillary statistics. Any constant is ancillary, but that is trivial. However, the following two examples are classic and widely used in hypothesis testing.

Example 5.4 (χ^2 test)**.** If $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, 1)$, then $A(y) = \sum_{i=1}^n (y_i - \overline{y})^2 \sim \chi^2_{n-1}$ is ancillary.

Example 5.5. If $Y_1, \dots, Y_n \sim \text{Expo}(\lambda)$, meaning that $f_\lambda(y) = \lambda e^{-\lambda y}$ for $y \geq 0$, then

$$A(y) = \frac{y_n}{y_1 + \cdots + y_n}$$

is ancillary because we can let $U_i = \lambda Y_i \sim \text{Expo}(1)$.

In general, this pattern is called a *scale family*, where we have a CDF of the form $\Pr Y \le y = F(\frac{y}{\sigma})$ for $\sigma > 0$. Then, any statistic that depends on Y only through $\frac{y_2}{y_1}, \dots, \frac{y_n}{y_1}$ is ancillary.

Example 5.6. Consider $Y_i \sim \mathrm{Unif}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, and recall that $(Y_{(1)}, Y_{(n)})$ is a minimal sufficient statistic. Note that the difference $Y_{(n)} - Y_{(1)}$ is ancillary.

Exercise 5.2 (Pencil problem)**.** Find an example of a model and two statistics where A_1 and A_2 are both ancillary, but (A_1, A_2) is not ancillary.

Proof. Let $(X_i, Y_i) \sim \mathcal{N}([\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}], [\begin{smallmatrix} 1 & r \\ r & 1 \end{smallmatrix}])$. Then, $A_1 = X \sim \mathcal{N}(0,1)$ and $A_2 = Y \sim \mathcal{N}(0,1)$ are both ancillary, but (A_1, A_2) is not, since its distribution depends on the covariance r. □

Now, we're ready to introduce a very elegant and useful result.

Theorem 5.7 (Basu's theorem)**.** *If T is a complete sufficient statistic for θ and A is ancillary for θ, then $A \perp\!\!\!\perp T$.*

Proof. For any measurable set B, consider $h_B(T) = \Pr_\theta(A \in B \mid T) - \Pr_\theta(A \in B)$. This is a function of T that does not depend on θ, since T is a sufficient statistic and A is ancillary. By the law of iterated expectation,

$$\mathbf{E}_\theta\left[h_B(T)\right] = \mathbf{E}_T\left[\Pr_\theta(A \in B \mid T)\right] - \Pr_\theta(A \in B) = 0.$$

Therefore, since T is a CSS, we conclude that $h_B(T) = 0$ almost surely, so $T \perp\!\!\!\perp A$. □

What makes Basu's theorem so useful in practice is that it allows us to find independent statistics that are not obvious at first glance. Sometimes we might even **add a variable** to our model in order to apply Basu's theorem.

Example 5.8. If $Y_i \sim \mathcal{N}(0, \sigma^2)$, and $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, while $S = \sum_{i=1}^n (Y_i - \overline{Y})^2$, then $\overline{Y} \perp\!\!\!\perp S$.

Proof. Consider the family of distributions $\mathcal{N}(\mu, \sigma^2)$ parameterized by μ, with σ^2 known. Then, $\overline{Y}$ is a complete sufficient statistic for μ, while S is ancillary. By Theorem 5.7, $\overline{Y} \perp\!\!\!\perp S$ for any μ. □

Here's the last example that we'll discuss today.

Example 5.9. Suppose that $Y_1, \dots, Y_n \sim \mathcal{N}(0, \sigma^2)$. Let M be the median of $Y_1, \dots, Y_n$. Then,

$$\begin{aligned} \mathrm{Cov}(\overline{Y}, M) &= \mathrm{Cov}(\overline{Y}, M - \overline{Y} + \overline{Y}) \\ &= \mathrm{Cov}(\overline{Y}, M - \overline{Y}) + \mathrm{Cov}(\overline{Y}, \overline{Y}). \end{aligned}$$

However, note that $M - \overline{Y}$ is ancillary for μ, so by Theorem 5.7, $\overline{Y} \perp\!\!\!\perp M - \overline{Y}$. Independence implies zero correlation, so the covariance is zero, and we conclude that

$$\mathrm{Cov}(\overline{Y}, M) = \frac{\sigma^2}{n}.$$

Exercise 5.3 (Pencil problem)**.** Based on the example above, show that $\mathrm{Cov}(Y_{(1)}, \overline{Y}) = \frac{\sigma^2}{n}$.

6 September 22nd, 2021

Today we discuss feedback, likelihood derivatives (the score function), and the Cramér-Rao lower bound for variance of unbiased estimators.

6.1 The Score Function

So far we've talked about the likelihood function $L(\theta) = f_\theta(y)$ for given data samples y, as well as the log-likelihood function $\ell(\theta)$. What happens when we take the gradient of the log-likelihood?

Definition 6.1 (Score function)**.** Given data samples $Y = (Y_1, \dots, Y_n)$ drawn from $f_\theta(y)$, the score function $S(y, \theta)$ is given by

$$S(y, \theta) = \frac{\partial \log f_\theta(y)}{\partial \theta} = \frac{\partial \ell(\theta)}{\partial \theta}.$$

Note that the score function for multiple data points equals the sum of the score functions for each individual data point.

Among other uses, the score function is useful for finding maximum likelihood estimators. In particular, if the MLE for y occurs at $\hat{\theta}$ in the interior of parameter space Θ, and the log-likelihood function ℓ is differentiable at $\hat{\theta}$, then it must satisfy $S(y, \hat{\theta}) = 0$.

Definition 6.2 (Differentiating under the integral)**.** Given a family of densities parameterized by θ for a random variable Y with respect to measure μ, we say that a function $g_\theta(y)$ satisfies the *m-th order EDI* condition if

$$\frac{\partial^m}{\partial \theta^m} \int_\mu g_\theta(y)\,\mathrm{d}y = \int_\mu \frac{\mathrm{d}^m}{\mathrm{d}\theta^m} g_\theta(y)\,\mathrm{d}y.$$

This is a useful regularity condition, and there are various real analysis results that can be used to prove this, such as the dominated convergence theorem. However, the proof of this result lies in measure theory that is out of scope for this class, so we will take the liberty of assuming necessary EDI conditions without justification.[5]

We will define a couple core regularity conditions on a model $\{f_\theta(y) : \theta \in \Theta\}$, just for convenience to prove facts about these models.

- **(A.1).** The parameter space Θ is an open set in $\mathbb{R}^n$.
- **(A.2).** The support of Y, i.e., $\{y : f_\theta(y) > 0\}$, does not depend on θ.

Proposition 6.3. *For a model $f_\theta(y)$ satisfying (A.1) and (A.2), if f_θ is differentiable on its support and the first-order EDI holds, then for all $\theta \in \Theta$,*

$$\mathbf{E}_\theta\left[S(Y, \theta)\right] = 0.$$

Proof. Pretty trivial argument, we just apply Definition 6.2 and notice that the total integral of $f_\theta(y)$ over all y is 1, so it does not change based on the value of θ. □

Note. There are some interesting structural parallels between this fact, $\mathbf{E}_\theta\left[S(Y, \theta)\right] = 0$, and the maximum likelihood estimator, which satisfies $S(y, \hat{\theta}) = 0$.

[5] Lucas says that in his research, which is statistical theory rather than probability theory, he almost always assumes the ∞-order EDI without justification. It is very rare for this condition to break.

Corollary 6.3.1. *If θ is 1-dimensional and the conditions in the previous proposition hold, then* $\mathbf{Var}_\theta\left[S(Y,\theta)\right] = \mathbf{E}_\theta\left[S^2(Y,\theta)\right]$. *(Analogous results hold for the covariance matrix of the score function when θ is higher-dimensional.)*

This corollary motivates the next topic of our lecture, which is a measure of how much *information* a random variable Y carries about its parameter θ.

6.2 Fisher Information

Intuitively, if a random variable carries a lot of information about its parameter, then its log-likelihood will change greatly between different values of θ. This means the score function will have high variance, which leads to the following.

Definition 6.4 (Fisher information)**.** The *Fisher information* for a model $Y \sim f_\theta(y)$ is a function of the parameter θ given by

$$I(\theta) = \mathbf{E}_\theta\left[S^2(Y,\theta)\right].$$

Proposition 6.5. *For a model $f_\theta(y)$ satisfying (A.1) and (A.2), such that f_θ is twice-differentiable with respect to θ on its support and satisfies the second-order EDI,*

$$I(\theta) = -\mathbf{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(Y)\right].$$

Proof. Observe that by the chain rule and product rule,

$$I(\theta) = \mathbf{E}_\theta\left[\frac{f''_\theta(Y)}{f_\theta(Y)} - \left(\frac{f'_\theta(Y)}{f_\theta(Y)}\right)^2\right].$$

By moving the derivative out of the integral sign, the first term goes to zero. Meanwhile, the second term equals the Fisher information, so we conclude. □

Corollary 6.5.1. *Consider any $\theta \in \Theta$, with two models $Y_1 \sim f_\theta(y)$ and $Y_2 \sim g_\theta(y)$, and $Y_1 \perp\!\!\!\perp Y_2$. If I_1 is the Fisher information of Y_1, and I_2 is the Fisher information of Y_2, then (Y_1, Y_2) has Fisher information $I_1 + I_2$.*

Now, let's go through some examples to get a feeling for the Fisher information metric.

Example 6.6. If $Y_1, \ldots, Y_n \sim \text{Pois}(\lambda)$, then

$$I_1(\lambda) = -\mathbf{E}_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{\lambda^{Y_1}e^{-\lambda}}{Y_1!}\right)\right] = \frac{1}{\lambda}.$$

Therefore, the Fisher information is $I_n(\lambda) = \frac{n}{\lambda}$.

Example 6.7. For a location family where $f_\theta(y) = f(y-\theta)$, the Fisher information metric is a constant function. For a scale family where $f_\theta(y) = \theta^{-1}f(\frac{y}{\theta})$, a similar calculation tells us that $I(\theta)$ is proportional to θ^{-2}.

Finally, we conclude with the big result about Fisher information, which is a lower bound on how good an unbiased estimator for a parameter can be.

Theorem 6.8 (Cramér-Rao bound). *Given a model $\{f_\theta(y) : \theta \in \Theta\}$ such that (A.1) and (A.2) hold, let $g(\theta)$ be a differentiable parametric function, and let $T(Y)$ be an unbiased estimator for $g(\theta)$. Then, if f_θ is differentiable on its support and satisfies the first-order EDI, and $I(\theta) > 0$,*

$$\mathbf{Var}_\theta\left[T\right] \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

A similar result holds for the case when θ is a vector, but it is an inequality between matrices.

Proof. First, note that

$$\begin{aligned}
\mathrm{Cov}_\theta(S(Y,\theta), T(Y)) &= \mathbf{E}_\theta\left[S(Y,\theta)T(Y)\right] - \mathbf{E}_\theta\left[S(Y,\theta)\right]\mathbf{E}_\theta\left[T(Y)\right] \\
&= \int \frac{\partial \log f_\theta(y)}{\partial \theta} T(y) f_\theta(y)\,\mathrm{d}y \\
&= \int \frac{\partial f_\theta(y)}{\partial \theta} T(y)\,\mathrm{d}y \\
&= \frac{\partial}{\partial \theta} \int T(y) f_\theta(y)\,\mathrm{d}y \\
&= \frac{\partial}{\partial \theta}\mathbf{E}_\theta\left[T(Y)\right] \\
&= g'(\theta).
\end{aligned}$$

Furthermore, as we showed previously, $\mathbf{Var}_\theta\left[S(Y,\theta)\right] = I(\theta)$. Therefore, by the Cauchy-Schwarz inequality,

$$\mathrm{Cov}_\theta(S(Y,\theta), T(Y))^2 \leq \mathbf{Var}_\theta\left[S(Y,\theta)\right]\mathbf{Var}_\theta\left[T(Y)\right].$$

Substituting in our results above, we get

$$(g'(\theta))^2 \leq I(\theta)\,\mathbf{Var}_\theta\left[T(Y)\right].$$

□

The Cramér-Rao lower bound will give us a goal to aim towards in the case when we want to find low-variance unbiased estimators. It is not always realizable, but it is still very important, and it is most of the reason behind why we care about Fisher information.

7 September 27th, 2021

Today we will finish discussing the Cramér-Rao lower bound, then introduce method of moments estimation (MOM) and maximum likelihood estimation (MLE).

7.1 More on Cramér-Rao

Recall from last lecture that the Cramér-Rao lower bound was a way to obtain a lower limit on the variance of an unbiased estimator of a model parameter, based on the Fisher information of that parameter. This lower bound is not always satisfied, but there are some cases where it is tight.

Example 7.1. If $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbf{Var}\left[\overline{Y}\right] = \frac{\sigma^2}{n} = \frac{1}{I(\mu)}$.

Example 7.2. If $Y_1, \ldots, Y_n \sim \text{Pois}(\lambda)$, then $\mathbf{Var}\left[\overline{Y}\right] = \frac{\lambda}{n} = \frac{1}{I(\lambda)}$.

Therefore, the Cramér-Rao lower bound holds for the mean of Gaussian and Poisson random variables. However, it does not hold for estimating a nonlinear function $g(\mu)$ of the mean. It turns out that maximum likelihood estimation asymptotically achieves the Cramér-Rao lower bound when the sample size is large, but not for any finite sample size.

The above examples were of unbiased estimators, but there is a variant of the Cramér-Rao lower bound that also works for biased estiamtors, as long as the bias converges quickly to zero as the sample size n increases.

Proposition 7.3 (Asymptotic Cramér-Rao). *Given an estimator $T_n(\mathbf{Y})$ of $g(\theta)$, if*

$$\sqrt{n}(T_n - g(\theta)) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

then the asymptotic variance $v(\theta) \geq \frac{(g'(\theta))^2}{I_1(\theta)}$.

Proof. We omit the proof, but see §6.2 of [LC06] for more details. □

Note that the convergence in distribution condition is much weaker than the original assumption of being an unbiased estimator in Theorem 6.8, since convergence in distribution can tolerate edge cases such as unlikely, far-out outliers.

7.2 Method of Moments Estimation

Theorizing about lower bounds is interesting, but how might we actually design estimators in practice? The *method of moments (MoM)* is a general strategy for obtaining estimators of parameters. We set sample moments equal to the population moments and solve for θ.

Definition 7.4 (Moment). Given a random variable Y parameterized by θ, the *r-th moment* of the distribution of Y is $\mathbf{E}_\theta\left[Y^r\right]$, defined for positive integers r.

Definition 7.5 (Central moment). Given a random variable Y parameterized by θ, the *r-th central moment* of the distribution of Y is $\mathbf{E}_\theta\left[(Y - \mathbf{E}_\theta\left[Y\right])^r\right]$, once again defined for positive integers r.

Sometimes, for heavy-tailed distributions like the Cauchy distribution, k-th moments may not be defined. In these cases, we may take a transformation of our samples before computing moments, which is known as the *generalized method of moments.*

Example 7.6 (MOM for $\mathcal{N}(\mu, \sigma^2)$)**.** The method of moments estimator for the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is

$$\begin{aligned}\hat{\mu}_{\text{MOM}} &= \overline{Y}, \\ \hat{\sigma}^2_{\text{MOM}} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.\end{aligned}$$

Note that the empirical variance appears, but it lacks the more subtle $\frac{1}{n-1}$ correction factor.

Example 7.7 (MOM for $\text{Binom}(k, p)$)**.** If $Y_1, \dots, Y_n \sim \text{Binom}(k, p)$, then equating the sample moments with the population moments yields

$$\begin{aligned}\overline{Y} &= kp, \\ \frac{1}{n}\sum_{i=1}^{n} Y_i^2 &= \mathbf{E}\left[Y_1^2\right] = kp(1-p) + k^2p^2.\end{aligned}$$

If we solve these equations for the parameters, we get the method of moments estimators:

$$\begin{aligned}\hat{k}_{\text{MOM}} &= \frac{\overline{Y}^2}{\overline{Y} - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}, \\ \hat{p}_{\text{MOM}} &= \overline{Y}/\hat{k}_{\text{MOM}}.\end{aligned}$$

Note that these estimators are sketchy, since the difference between the sample mean and raw sample variance in the denominator could be negative, which gives strange results in that case. But in general, finding the MLE for the binomial distribution is hard, so the MOM estimator provides a good starting point.

Example 7.8 (MOM for exponential family)**.** If $f_\theta(y)$ belongs to an exponential family

$$f_\theta(y) = \exp\{\eta(\theta)T(y) - \psi(\eta(\theta))\}h(y),$$

then differentiating under the integral sign (Definition 6.2) with respect to η yields

$$\int T(y)\exp\{\eta T(y) - \psi(y)\}h(y)\,\mathrm{d}y = \psi'(\eta)\int \exp\{\eta T(y) - \psi(y)\}h(y)\,\mathrm{d}y.$$

This means that $\mathbf{E}_\theta\left[T(Y)\right] = \psi'(\eta)$, so if we transform the samples of an exponential family by T, we get the method of moments estimator by solving:

$$\frac{1}{n}\sum_{i=1}^{n} T(Y_i) = \psi'(\eta(\theta)).$$

7.3 Maximum Likelihood Estimation

The maximum likelihood estimator (MLE) of θ is the parameter value that maximizes the likelihood function $L(\theta)$ for the observed data. Equivalently, this also maximizes the log-likelihood $\ell(\theta)$.

Example 7.9 (MLE for exponential family)**.** Once again, suppose that we have an exponential family with parameter θ, and observe that the log-likelihood function is

$$\ell_{\mathbf{Y}}(\theta) = \left(\sum_{i=1}^{n} T(Y_i)\right)\eta(\theta) - n\psi(\eta(\theta)) + \sum_{i=1}^{n}\log h(Y_i).$$

If we take the derivative with respect to η and set it to zero (to find local extrema), we get the equation for η given by

$$\frac{\partial \ell}{\partial \eta} = 0 \implies \frac{1}{n}\sum_{i=1}^{n} T(Y_i) = \psi'(\eta(\theta)).$$

Note that this is the exact same as the equation for the MOM estimator!

Here are some basic properties of the MLE.

Proposition 7.10 (Equivariance of the MLE)**.** *If $\hat{\theta}$ is an MLE for θ and $\tau = g(\theta)$, then $\hat{\tau} = g(\hat{\theta})$ is an MLE for τ.*

Proposition 7.11 (MLE and sufficiency)**.** *If there exists a unique MLE for θ, then it is a function of every sufficient statistic T of θ.*

This implies that if the MLE is a sufficient statistic, then it is minimal sufficient. Also, if there exists a complete sufficient statistic and the MLE is unbiased, then the MLE must be the unique UMVUE by Theorem 4.7.

Note. The MLE is not necessarily unbiased. For example, the conjugate prior of $Y_1, \ldots, Y_n \sim \text{Bern}(p)$ is the Beta distribution $\text{Beta}(\alpha, \beta)$, which has mode $\frac{\alpha-1}{\alpha+\beta-2}$ corresponding to the MLE, but it has mean $\frac{\alpha}{\alpha+\beta}$, which is not the same.

Finally, we state a useful theorem about maximum likelihood estimation in the context of *consistency.* Roughly speaking, this means that under some mild conditions, the MLE $\hat{\theta}$ for a sample of size n is guaranteed to converge to the true value of θ as n increases to infinity.

Definition 7.12 (Identifiability)**.** A model $\{f_\theta : \theta \in \Theta\}$ is *identifiable* if, for any two $\theta_1, \theta_2 \in \Theta$, we have $f_{\theta_1}(\mathbf{y}) = f_{\theta_2}(\mathbf{y})$ almost surely, then $\theta_1 = \theta_2$.

Essentially, this means that we can't have two parameters that produce the same distribution, since that would make it impossible to distinguish which parameter was true.

Theorem 7.13 (Consistency of MLE)**.** *Let $Y_1, \ldots, Y_n \sim f_{\theta_0}$, and let $\hat{\theta}_n$ be the MLE with respect to the model $\{f_\theta : \theta \in \Theta\}$ with $\theta_0 \in \Theta$. Then, assuming the following conditions:*

(i) *f_θ is identifiable,*

(ii) *The support of f_θ does not depend on θ,*

(iii) $\mathbf{E}_{\theta_0}\left[|\log f_\theta(Y_1)|\right] < \infty$ *for all $\theta \in \Theta$, and*

(iv) $|\Theta| < \infty$*,*

then $\hat{\theta}_n$ exists and is unique with probability tending to 1 as $n \to \infty$, and it is strongly consistent, meaning that $\hat{\theta}_n \to \theta_0$ almost surely.

8 September 29th, 2021

Today we prove the consistency of MLE, discuss the asymptotic normality of MLE, and introduce the delta method.

8.1 Consistency of MLE

First, we prove Theorem 7.13, which was our main theorem at the end of yesterday's lecture. Recall that this means if a statistical model is *identifiable* and satisfies some specific regularity conditions (in particular, we need $|\Theta| < \infty$), then the MLE is guaranteed to be strongly consistent as $n \to \infty$.

Proof of Theorem 7.13. First, let $\ell(\theta) = \log f_\theta(\mathbf{y}) = \sum_{i=1}^n \log f_\theta(y_i)$. Then,

$$\bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f_\theta(y_i).$$

Also, let $\bar{\ell}(\theta, \theta_0) = \mathbf{E}_{\theta_0}\left[\log f_\theta(Y_i)\right]$. By the weak law of large numbers, we know that $\bar{\ell}_n(\theta) \to \bar{\ell}(\theta, \theta_0)$ in probability for all parameters $\theta \in \Theta$.

Second, note that $\bar{\ell}(\theta, \theta_0)$ is uniquely maximized by θ, since KL divergence is nonnegative. This is a result known as Gibb's inequality, which can be verified by checking that

$$\mathbf{E}_{\theta_0}\left[\log f_\theta(y_i)\right] - \mathbf{E}_{\theta_0}\left[\log f_{\theta_0}(y_i)\right] = \mathbf{E}_{\theta_0}\left[\log \frac{f_\theta(y_i)}{f_{\theta_0}(y_i)}\right] \le \mathbf{E}_{\theta_0}\left[\frac{f_\theta(y_i)}{f_{\theta_0}(y_i)}\right] = 0.$$

Combining these two facts, we conclude that $\Pr_{\theta_0}(\bar{\ell}_n(\theta) \ge \bar{\ell}_n(\theta_0)) \to 0$ for all $\theta \in \Theta \setminus \{\theta_0\}$. Therefore, the probability that θ_0 uniquely maximizes the empirical likelihood function converges almost surely to 1 as $n \to \infty$, as desired. □

Note. Unfortunately, the reasoning in the theorem above does not apply to infinite parameter spaces, since we can't guarantee that the likelihood maximizer exists. We need to add a couple more conditions when $|\Theta| \not< \infty$:

(iv) θ_0 lies in the interior of Θ as a smooth manifold.

(v) $f_\theta(y)$ is differentiable with respect to θ, for all θ and almost surely in y.

These conditions imply that the likelihood has a unique maximizer when the score function is zero, with probability $p \to 1$ as $n \to \infty$.

8.2 Asymptotic Normality of MLE

The second result today will be a stricter set of conditions that provide a central limit theorem-like result for the maximum likelihood estimator as $n \to \infty$. Unlike our previous result, which only had four regularity conditions, this one will have **seven** conditions!

Theorem 8.1. *Let $Y_1, \ldots, Y_n \sim f_{\theta_0}$, and assume that $\{f_\theta : \theta \in \Theta\}$ satisfies:*

(i) θ_0 lies in the interior of Θ.

(ii) f_θ is identifiable.

(iii) The support of f_θ does not depend on θ,

(iv) For all $\theta \in \Theta$, $\log f_\theta(y)$ is almost surely three times differentiable in θ.

(v) There exists a function $M(y)$, possibly dependent on θ_0, such that for all θ in a neighborhood of θ_0,

$$\left|\frac{\partial^3 \log f_\theta(y)}{\partial\theta^3}\right| \leq M(y), \text{ and}$$
$$\mathbf{E}_{\theta_0}\left[M(Y)\right] < \infty.$$

(vi) The score function is zero at θ_0, i.e.,

$$\mathbf{E}_{\theta_0}\left[\left.\frac{\partial \log f_\theta(Y)}{\partial\theta}\right|_{\theta=\theta_0}\right] = 0.$$

(vii) The Fisher information satisfies

$$0 < I_1(\theta_0) = -\mathbf{E}_{\theta_0}\left[\left.\frac{\partial^2 \log f_\theta(Y_1)}{\partial\theta^2}\right|_{\theta=\theta_0}\right].$$

Then, there exists a consist zero $\hat{\theta}_n$ of the score function S_n, such that $S_n(\hat{\theta}_n) = 0$. As $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_1(\theta_0)}\right).$$

Proof. First, we will show the existence of the consistent root $\hat{\theta}_n$. The key in this proof is to take a Taylor expansion of the score function S_n. A second-order Taylor expansion with Lagrange form of the remainder tells us that

$$S_n(\hat{\theta}_n) = S_n(\hat{\theta}_n) + S_n'(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) + \frac{1}{2}S_n''(\hat{\theta}_n^*)(\hat{\theta}_n - \theta_0)^2,$$

where $\hat{\theta}_n^*$ is some value between θ_0 and $\hat{\theta}_n$. Roughly speaking, we can "solve" for $\hat{\theta}_n - \theta_0$ to get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{S_n(\theta_0)/\sqrt{n}}{-\frac{1}{n}S_n'(\theta_0) - \frac{1}{2n}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0)}.$$

The numerator of this fraction converges in distribution to $\mathcal{N}(0, I_1(\theta_0))$ by the central limit theorem, and the first term $-\frac{1}{n}S_n'(\theta_0)$ of the denominator converges in probability to $I_1(\theta_0)$ by the weak law of large numbers.

At this point, we would almost be done by applying Slutsky's theorem to this quotient, but there is one more issue: what about the last quadratic term $-\frac{1}{2n}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0)$? We need to bound this last term appropriately, and that is the crucial step in making this proof rigorous. We omit the rest of the proof in these notes, but you can finish the argument by a combination of fairly technical steps that concludes in an application of Slutsky's theorem. □

Lucas states that this is one of the most important results that we will improve in the class. Although only stated above for scalar θ because the notation is more convenient, it also holds in generality for cases where θ is multi-dimensional, by extending the conditions in a straightforward manner with linear algebra.

9 October 4th, 2021

Today we discuss the Delta method in asymptotics and introduce hypothesis testing.

9.1 The Delta Method

First, a remark about Theorem 8.1. Note that the normality of the MLE implies that it is unbiased, and furthermore, the variance is approximated by $(nI_1(\theta_0))^{-1}$ as $n \to \infty$, which matches with the Cramér-Rao lower bound, meaning that $\hat{\theta}_n$ is asymptotically optimal.

Definition 9.1 (Asymptotic relative efficiency)**.** If W_n and V_n be two estimators of a parametric function $g(\theta)$ such that

$$\sqrt{n}(W_n - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma_W^2),$$
$$\sqrt{n}(V_n - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma_V^2),$$

then the *asymptotic relative efficiency (ARE)* of V_n with respect to W_n is σ_W^2/σ_V^2.

The normality of the MLE for the parameter $\hat{\theta}$ is a nice result, especially since we get the asymptotic variance from the information Fisher information metric. If we apply a function to $\hat{\theta}$, the result $g(\hat{\theta})$ is also an MLE for $g(\theta)$, but what is its distribution? It turns out that it is also normally distributed, due to the following fact.

Theorem 9.2 (Delta method)**.** *If $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta))$, then if g is continuously differentiable at θ and $g'(\theta) \neq 0$,*

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

Proof. We offer a proof sketch. Take the first-order Taylor expansion of $g(T_n)$ to get

$$g(T_n) = g(\theta) + g'(\theta^*)(T_n - \theta),$$

for some mean-value term specifying θ^* between θ and T_n. Using the continuous mapping theorem and Slutsky's theorem, we can show that $\sqrt{n}(g(T_n) - g(\theta))$ converges to the same distribution as $\sqrt{n}g'(\theta)(T_n - \theta)$, so we conclude. □

Example 9.3. If $Y_1, \ldots, Y_n \sim \text{Bin}(k, p)$, for fixed k, then the MLE for p is $\hat{p} = \overline{Y}/k$. This has asymptotic variance $\frac{p(1-p)}{n}$, which is consistent with the Fisher information $I_1(p) = \frac{1}{p(1-p)}$. Now, if we want to estimate the odds ratio $g(p) = p/(1-p)$, note that $g'(p) = 1/(1-p)^2$, so

$$\sqrt{n}(g(\hat{p}) - g(p)) \xrightarrow{d} \mathcal{N}\left(0, \frac{p}{(1-p)^3}\right).$$

There are some issues with the method presented above. One is the case when we want to estimate a parameter plus error bars by one standard deviation, which is hard because the asymptotic variance is a function of the parameter, which is unknown. In this case, we can usually just plug in our MLE for the parameter and get a result that still converges in distribution.

Alternatively, we can apply a *variance-stabilizing transformation* to the parameter, which is essentially a function $h(\theta)$ such that the asymptotic variance $(h'(\theta))^2/I_1(\theta)$ is invariant as θ changes. This means that $h'(\theta) \propto \sqrt{I(\theta)}$, so $h(\theta) \sim \int \sqrt{I(\theta)}$. For the binomial parameter, this variance-stabilizing transformation is $\sin^{-1}\sqrt{p}$.

The second issue is that the rate of convergence to normal also depends heavily on p. For example, in a Binomial distribution, the sample mean is approximately Poisson for small values of the parameter p, which has a heavy skew between the lengths of the two tails.

9.2 Hypothesis Testing

In hypothesis testing, we typically test a *null hypothesis* H_0 against an *alternative hypothesis* H_1. There are several kinds of hypotheses for a scalar parameter θ; here are some examples:

- $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. (Point null versus point alternative.)
- $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_1$. (Point null versus one-sided alternative.)
- $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_1$. (Point null versus two-sided alternative.)
- $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_1$. (Composite null versus composite alternative.)
- $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \mathbb{R}$. (Point null within a composite alternative.)

These types of tests can all be useful in different scientific scenarios. Generally, a test partitions the sample space into an acceptance region A, where H_0 is accepted, and a rejection region R, where H_0 is rejected. Also, we will usually partition the space of a sufficient statistic $T(\mathbf{Y})$, rather than dealing with the details of $\mathbf{Y}$ directly.

Example 9.4 (Z-test)**.** If $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, 1)$, then one test for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ would be to use the statistic $T(\mathbf{Y}) = \overline{Y}$ and decision rule to reject H_0 if $\sqrt{n}(\overline{Y} - \mu_0) > 2$.

A test is judged by two kinds of error. Type-I error (false positive) is when H_0 is rejected when it is true, while Type-II error (false negative) is when H_0 is accepted despite being false. The *power* of a test is the probability of rejecting H_0 when it is false, and it is denoted $\beta(\theta) = \Pr_\theta(T(\mathbf{Y}) \in R)$.

10 October 6th, 2021

Today, we will continue discussing hypothesis testing and introduce the notion of significance levels, the goal of maximizing power, and offer theoretical analysis of the most powerful tests.

10.1 The Neyman-Pearson Lemma

As before, consider the same problem of hypothesis testing, with two hypotheses H_0 and H_1. Let Θ_0 be the set of parameters in the null hypothesis H_0, and let Θ_1 be the parameter space of the alternative hypothesis H_1. A common goal among hypothesis tests is to guarantee that the probability of Type-I error is bounded by a certain *significance level* α, meaning that

$$\sup_{\theta \in \Theta_0} \Pr_{\theta}(T(\mathbf{Y}) \in R) \leq \alpha.$$

(For example, a common value in some scientific fields is $\alpha = 0.05$.) The value on the left-hand side is called the *size* of the test, and we generally try to achieve equality between the size and significance level to maximize power.

Definition 10.1 (UMP)**.** A test is called *universally most powerful (UMP)* for a significance level α if it has the maximum power for all $\theta \in \Theta_1$, among all hypothesis tests of level α.

The following famous lemma is used to construct UMP tests.

Theorem 10.2 (Neyman-Pearson lemma)**.** *Suppose that we have i.i.d. samples $Y_1, \dots, Y_n \sim f_\theta(y)$, and suppose that we are testing the null hypothesis $H_0 : \theta = \theta_0$ against an alternative hypothesis $H_1 : \theta = \theta_1$, where $\theta_0 \neq \theta_1$. The rejection region*

$$R = \left\{ \mathbf{Y} : \frac{f_{\theta_1}(\mathbf{Y})}{f_{\theta_0}(\mathbf{Y})} \geq c \right\}$$

is the most powerful level α test if c satisfies the size-α condition

$$\Pr_{\theta_0}(\mathbf{Y} \in R) = \alpha.$$

Proof. Intuitively, the way to think about this statement is that the maximum-power estimator is simply the one that rejects when the likelihood ratio between H_0 and H_1 exceeds some constant ratio. This makes sense because by adding the maximum-likelihood ratio points to the rejection region, you get the most marginal power for a given test size. We leave the details of the proof as an exercise; it simply involves moving around some integral inequalities in the proper way. □

The NP lemma generalizes to randomized hypothesis tests, when there is no threshold c that exactly has size α. Instead, we can interpolate between the two nearest values to α. Assume that there are c_1, c_2 such that $\frac{f_{\theta_1}(\mathbf{Y})}{f_{\theta_2}(\mathbf{Y})}$ lies in (c_1, c_2) with probability zero assuming H_0, and

$$\Pr_{\theta_0}\left(\frac{f_{\theta_1}(\mathbf{Y})}{f_{\theta_2}(\mathbf{Y})} \geq c_1 \right) = \alpha_1 > \alpha > \alpha_2 = \Pr_{\theta_0}\left(\frac{f_{\theta_1}(\mathbf{Y})}{f_{\theta_2}(\mathbf{Y})} \geq c_2 \right).$$

Then, the maximum power level α test rejects whenever the likelihood ratio is $\geq c_2$, accepts when it is $< c_1$, and rejects with probability $\frac{\alpha - \alpha_1}{\alpha_1 - \alpha_2}$ when it is equal to c_1.

Example 10.3 (Z-tests are UMP)**.** Consider a model of $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, 1)$. If our hypotheses are $H_0 : \mu = \mu_0$, and $H_1 : \mu = \mu_1$ for some $\mu_1 > \mu_1$, then the likelihood ratio is monotone increasing in $\overline{Y}$, so the NP lemma tells us that the MP test is a simple threshold $\overline{Y} \geq c$ for some c. Similarly, this inequality would be reversed if $\mu_1 < \mu_0$, and it can be easily generalized to the case when the alternative hypothesis is a one-sided composite $H_1 : \mu \geq \mu_1$.

10.2 Testing Composite Hypotheses

We've seen how to construct MP tests for general point hypotheses using Theorem 10.2, but what about UMP tests for composite alternative hypotheses? In the one-sided alternative case of $\mathcal{N}(\mu, 1)$, a Z-test is also UMP, as we discussed above. However, for a two-sided hypothesis $H_1 : \mu \neq \mu_0$, **the UMP does not exist**, since the MP test differs based on $\mu \in \Theta_1 = \mathbb{R} \setminus \{\mu_0\}$.

Definition 10.4 (Monotone likelihood ratio)**.** The family of distributions $\{f_\theta(y) : \theta \in \Theta\}$ has *monotone likelihood ratio (MLR)* in a statistic $T(\mathbf{Y})$ if the ratio $f_{\theta_2}(\mathbf{Y})/f_{\theta_1}(\mathbf{Y})$ can be expressed as a function of $\theta_1, \theta_2, T(\mathbf{Y})$, and for each $\theta_1 < \theta_2$, the ratio is non-decreasing in $T(\mathbf{Y})$ when at least one of the numerator and denominator is positive.

Note. If a model has MLR in $T(\mathbf{Y})$, then T is a sufficient statistic, since for any fixed $\theta' \in \Theta$,

$$f_\theta(\mathbf{Y}) = \frac{f_\theta(\mathbf{Y})}{f_{\theta'}(\mathbf{Y})} \cdot f_{\theta'}(\mathbf{Y}).$$

Then, we conclude that T is sufficient by Theorem 2.6.

Example 10.5. An exponential family $f_\theta(y) = \exp\{T(y)\eta(\theta) - \psi(\eta(\theta))\}h(Y)$ has MLR in the sufficient statistic $T(\mathbf{Y})$ if the natural parameter $\eta(\theta)$ is a non-decreasing function of θ.

Here's the punchline. If we want to test a composite hypothesis with composite null, of the form $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, then the following theorem gives us a way to find a UMP.

Theorem 10.6 (Karlin-Rubin)**.** *Consider testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. If the family $\{f_\theta(y) : \theta \in \Theta\}$ has MLR in some statistic $T(\mathbf{Y})$, then for any t_0, the test with rejection region $T > t_0$ is UMP for level α, where $\alpha = \Pr_{\theta_0}(T > t_0)$.*

Proof. See Chapter 8 of [CB21] for a proof. □

11 October 13th, 2021

Today we continue discussing composite hypotheses and the likelihood-ratio test.

11.1 Karlin-Rubin Test

Similar to how the Neyman-Pearson lemma generalizes to randomized tests when there is no precise size-matches-power cutoff (happening when the test statistic T is discrete), Karlin-Rubin also generalizes to randomized tests in the same way.

Example 11.1. Let $Y_1, \dots, Y_n \sim \text{Bern}(p)$, and suppose that we want to test $H_0 : p \le p_0$ against the alternative hypothesis $H_1 : p > p_0$, using the Karlin-Rubin lemma. First, suppose that we are doing a simple point test of p_0 versus $p_1 > p_0$, where the Neyman-Pearson test statistic would be

$$\left(\frac{p_1}{p_0}\right)^T \left(\frac{1-p_0}{1-p_1}\right)^T,$$

where T is the natural sufficient statistic $T(\mathbf{Y}) = \sum_{i=1}^n Y_i$. Notice that the left-hand side is increasing with p_1 for any $p_1 > p_0$, so we conclude that T has the monotone likelihood ratio property (Definition 10.4). Therefore, the Karlin-Rubin lemma states that the most powerful test of its size for $H_0 : p \le p_0$ has rejection region $T \ge c$ for some threshold c. This test is most powerful for the specific level

$$\alpha = \sum_{j=c+1}^n \binom{n}{j} p_0^j (1-p_0)^{n-j}.$$

However, given a desired level α, there may not be a value of c that satisfies this equation exactly, since c is a discrete variable. In this case, we can do a randomized test by finding c such that

$$\sum_{j=c+1}^n \binom{n}{j} p_0^j (1-p_0)^{n-j} < \alpha < \sum_{j=c}^n \binom{n}{j} p_0^j (1-p_0)^{n-j}.$$

Then, our test rejects when $T > c$, accepts when $T < c$, and when $T = c$, it rejects with probability

$$\frac{\alpha - \sum_{j=c+1}^n \binom{n}{j} p_0^j (1-p_0)^{n-j}}{\binom{n}{c} p_0^c (1-p_0)^{n-c}}.$$

11.2 Likelihood-Ratio Test

Consider the problem of testing a null hypothesis $H_0 : \theta \in \Theta_0$ against an alternative hypothesis $H_1 : \theta \in \Theta_1$. Assume that $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. The likelihood-ratio test compares these statistical models based on the ratio of their likelihoods, i.e., with statistic

$$\Lambda = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}.$$

Here, we reject the null hypothesis when $\Lambda > c$ for some threshold c. In other words, if $\hat{\theta}_0$ is the likelihood maximizer for $L(\theta)$ over Θ_0, and $\hat{\theta}$ is the global maximizer over Θ, then the likelihood ratio is simply

$$\Lambda = \frac{f_{\hat{\theta}}(\mathbf{Y})}{f_{\hat{\theta}_0}(\mathbf{Y})}.$$

Note that $\Lambda \ge 1$, since $\hat{\theta}$ is the MLE for θ.

Example 11.2. Consider two parameters $\theta, \mu > 0$ and corresponding samples of i.i.d. random variables, $X_1, \dots, X_n \sim \text{Expo}(\theta)$ and $Y_1, \dots, Y_m \sim \text{Expo}(\mu)$. Our null hypothesis is $H_0 : \mu = \theta$, while our alternative hypothesis is $H_1 : \mu \neq \theta$. This is a complex hypothesis test, as we have multiple parameters and a difficult parameter region Θ_0, which makes it difficult to apply results like Karlin-Rubin.

Instead, we will try to directly use a likelihood-ratio test, comparing the log-likelihoods. Observe that the log-likelihood can be written as

$$\ell(\theta, \mu) = -n \log \theta - \frac{\sum_{j=1}^n X_j}{\theta} - m \log \mu - \frac{\sum_{j=1}^m Y_j}{\mu}.$$

We can immediately see that the maximum likelihood estimator over the entire parameter space is $\hat{\theta} = \overline{X}$, and $\hat{\mu} = \overline{Y}$. Meanwhile, it can be shown that the maximum likelihood estimator in the null hypothesis parameter space Θ_0 is

$$\hat{\theta}_0 = \frac{\sum_{j=1}^n X_j + \sum_{j=1}^m Y_j}{n+m}.$$

Therefore, our likelihood-ratio statistic is

$$\Lambda = \frac{L(\hat{\theta}, \hat{\mu})}{L(\hat{\theta}_0, \hat{\theta}_0)} = \frac{\hat{\theta}_0^{n+m}}{\hat{\theta}^n \hat{\mu}^m} = \frac{n^n m^m}{(n+m)^{n+m}} T^{-n} (1-T)^{-m},$$

where $T = \frac{\sum_{j=1}^n X_j}{\sum_{j=1}^n X_j + \sum_{j=1}^m Y_j}$. The log-likelihood ratio is now

$$\log \Lambda = \text{const} - n \log T - m \log(1-T).$$

If we graph this function on $T \in (0, 1)$, notice that it blows up to ∞ at either side of the interval, and it achieves its minimum when $T = \frac{n}{n+m}$. Therefore, instead of using the likelihood-ratio test directly, we might prefer a test of the form

$$R = \left\{ \left| T - \frac{n}{n+m} \right| > \alpha \right\}.$$

To determine α, we could just simulate T, whose distribution is θ-free under the null hypothesis, by ancillarity of the quotient of random variables in a scale family.

Note. In general, we often use the above formulation of the likelihood-ratio test when Θ is a higher-dimensional space, and the alternative hypothesis Θ_1 is a lower-dimensional subset of that space. For instance, in the example above, $\mu = \theta$ forms a one-dimensional subset of $\mathbb{R}^2$.

Naturally, the next question is about the asymptotic distribution of the likelihood-ratio statistic, since this informs our choice of the threshold α. Just like we can prove asymptotic normality of the MLE as $n \to \infty$, the likelihood-ratio statistic also has an asymptotic distribution.

Theorem 11.3 (Asymptotic chi-squared distribution of likelihood ratio)**.** *Consider testing $H_0 : \theta \in \Theta_0$ versus an alternative hypothesis $H_1 : \theta \in \Theta \setminus \Theta_0$. Let*

$$\Lambda_n = \frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta_0 \in \Theta_0} L_n(\theta_0)}$$

be the likelihood-ratio statistic for i.i.d. sampled data of size n from $\{f_\theta(y) : \theta \in \Theta\}$. Then, under the same conditions as Theorem 8.1 and that the global maximum of $L_n(\theta)$ is one of the consistent roots of the score equation, we have under H_0 that

$$2 \log \Lambda_n \xrightarrow{d} \chi_1^2,$$

Proof. First, observe that we can write

$$2\log\Lambda_n = -2(\ell_n(\theta_0) - \ell_n(\hat{\theta}_n)).$$

Here, we use ℓ_n for the log-likelihood and $\hat{\theta}_n$ for the maximum likelihood estimator that is a consistent root of the score equation, which is similar to the notation in Theorem 8.1. A second-order Taylor series expansion of $\ell_n(\theta_0)$ around $\hat{\theta}_n$ yields

$$\ell_n(\theta_0) = \ell_n(\hat{\theta}_n) + \frac{1}{2}(\theta_0 - \hat{\theta}_n)^2 \ell_n''(\theta_n^*),$$

for some θ_n^* between θ_0 and $\hat{\theta}_n$. Plugging this in, we get

$$2\log\Lambda_n = -(\theta_0 - \hat{\theta}_n)^2 \ell_n''(\theta_n^*) = n(\hat{\theta}_n - \theta_0)^2 \bigg(-\frac{1}{n}\ell_n''(\theta_n^*)\bigg).$$

Notice that the difference between the maximum likelihood estimator $\hat{\theta}_n$ and θ_0 is asymptotically normal by Theorem 8.1, i.e., $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \frac{1}{I_1(\theta_0)})$, so by the continuous mapping theorem,

$$n(\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \frac{1}{I_1(\theta_0)}\chi_1^2.$$

Finally, we can take care of the last term by using a combination of Taylor expansion and the weak law of large numbers to get that $-\frac{1}{n}\ell_n''(\theta_n^*) \xrightarrow{p} I_1(\theta_0)$ under H_0, so we arrive thereafter at the desired result by Slutsky's theorem. □

12 October 18th, 2021

Today, we discuss asymptotic and non-parametric hypothesis tests.

12.1 Asymptotic Hypothesis Tests

Recall that we showed the likelihood-ratio test statistic converges to an asymptotic χ^2-distribution. There is another way of thinking about hypothesis tests in terms of how quickly certain test statistics asymptotically converge to their limit distributions.

Definition 12.1 (Score test)**.** Suppose that we are testing a null hypothesis $H_0 : \theta = \theta_0$ against an alternative hypothesis $H_1 : \theta \neq \theta_0$. The likelihood-ratio test statistic in this case would be $\Lambda_n = L(\hat{\theta}_n)/L(\theta_0)$. As an alternative, the *score statistic* has asymptotic distribution

$$\frac{1}{\sqrt{n}} S_n(\mathbf{Y}, \theta_0) \xrightarrow{d} \mathcal{N}(0, I_1(\theta_0)),$$

by the definition of Fisher information as the variance of the score function. Therefore, in the score test, we reject at level α when

$$\left| \frac{S(\mathbf{Y}, \theta_0)}{\sqrt{n I_1(\theta_0)}} > z_{1-\frac{\alpha}{2}} \right|.$$

Definition 12.2 (Wald test)**.** In the *Wald test*, observe that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_1(\theta_0)}\right).$$

Therefore, in this hypothesis test, we reject at level α when

$$\left| \sqrt{n I_1(\theta_0)} (\hat{\theta}_n - \theta_0) \right| > z_{1-\frac{\alpha}{2}}.$$

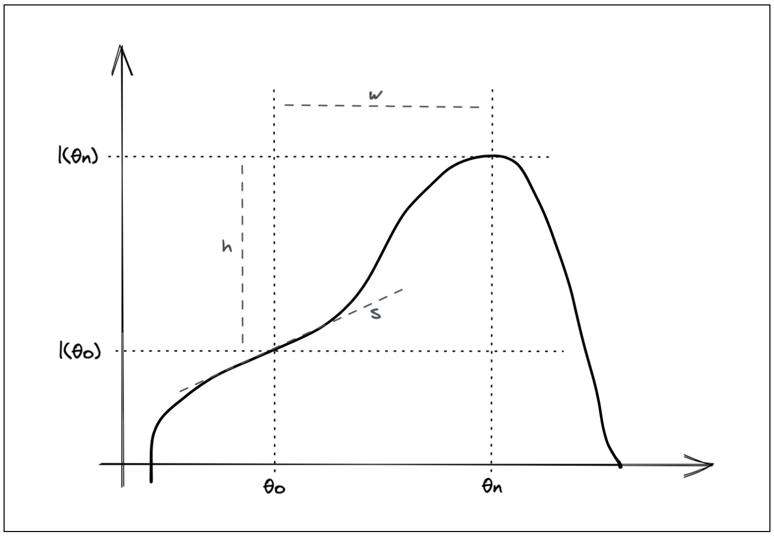

Figure 1: Schematic visualization of asymptotic hypothesis tests.

Visually, we can think of these tests as in terms of a schematic like Fig. 1. Here, the Wald test statistic is represented by w, the likelihood-ratio test statistic is visualized by $\log h$ as a quotient of likelihoods, and the score test statistic is the slope of the tangent labeled s. Each of these is nearly asymptotically equivalent as $n \to \infty$ but they have different power.

Note. We can generalize each of these asymptotic hypothesis tests to cases when the data and parameters are multi-dimensional. In these cases, the tests involve the Fisher information matrix, which is required to be positive definite (in order to be invertible). The limit distributions in these cases depend on dimensionality of the parameter sets Θ and Θ_0:

- For the LR test, we have
$$2\log\Lambda_n \xrightarrow{d} \chi^2_{\dim(\Theta)-\dim(\Theta_0)}.$$

- For the Wald test, we have
$$(\hat{\theta}-\theta_0)^\top \mathbf{I}(\theta_0)(\hat{\theta}-\theta_0) \xrightarrow{d} \chi^2_{\dim(\Theta)-\dim(\Theta_0)}.$$

- For the score test, we have
$$[\nabla\ell(\mathbf{Y},\theta_0)]^\top \mathbf{I}^{-1}(\theta_0)[\nabla\ell(\mathbf{Y},\theta_0)] \xrightarrow{d} \chi^2_{\dim(\Theta)-\dim(\Theta_0)}.$$

12.2 Nonparametric Hypothesis Tests

Before talking about non-parametric hypothesis tests, we need to introduce a rigorous notion of p-value, for the purposes of this course.

Definition 12.3 (p-value)**.** For a family of hypothesis tests of size α that defines a negation region R_α for any $\alpha \in [0,1]$, such that $R_{\alpha_1} \subseteq R_{\alpha_2}$ for any $\alpha_1 \leq \alpha_2$, the p-value for a data set $\mathbf{X}$ is the smallest significance level α such that $\mathbf{Y} \in R_\alpha$.

Most hypothesis tests are of this form, and they reject based on a test statistic threshold $T(\mathbf{Y}) \geq C(\alpha)$, for some increasing non-linear function C. Then, the p-value of a test statistic T can simply be computed by taking
$$p = \min\{\alpha : T(\mathbf{Y}) \geq C(\alpha)\} = C^{-1}(T(\mathbf{Y})).$$

In general, we know that $\sup_{\theta\in\Theta_0} \Pr_\theta(p \leq \alpha) = \sup_{\theta\in\Theta_0} \Pr_\theta(\mathbf{Y} \in R_\alpha) \leq \alpha$, for all α, since the size of a test is less than or equal to its significance level. This is the common way that hypothesis tests are formulated in applied statistics. Furthermore, if the size equals the level for all α, which is fairly common, we can also say more precisely that
$$p \xrightarrow{d} \mathrm{Unif}[0,1].$$

Unlike our previous hypothesis tests in this course, note that by writing p-values, we effectively eliminate the parameter from our mathematical expressions. Here are some considerations to keep in mind when thinking about non-parametric tests:

- **Weak assumptions:** Non-parametric tests can easily support very general null hypotheses. For example, consider data distribution $Y \sim \mathcal{N}(\theta, 1)$. While a standard parametric test could compare $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, non-parametric tests could also test hypotheses like $H_0 : Y \sim \mathcal{N}(0,1)$, or $H_0 : \mathbf{E}[Y] = 0$.

- **Power:** Under parametric assumptions, we lose a bit of power in non-parametric tests compared to their counterparts. However, non-parametric tests can gain relative power when the parametric model fails.

- **Discrete statistics:** Non-parametric tests tend to have discrete test statistics.

Let's give some concrete examples of these hypothesis tests now. Most of these examples center around some "maximally ancillary" statistic for the null hypothesis.

Example 12.4 (Sign test)**.** Suppose that we have random variables $Y_1, \ldots, Y_n \in \mathbb{R}$ sampled i.i.d. from some data distribution. Then, the *sign test* evaluates the likelihood of the null hypothesis $H_0 : \Pr(Y > \theta_0) = \Pr(Y < \theta_0)$. The test stiatistic in this case is

$$\sum_{i=1}^{n} 1_{\{Y_i > \theta_0\}} \sim \text{Bin}\left(n, \frac{1}{2}\right) \text{ under } H_0.$$

The test statistic concentrates around the mean, so we reject when the statistic is too far away from $\frac{n}{2}$, and our size is

$$\Pr\left(\left|\text{Bin}\left(n, \frac{1}{2}\right) - \frac{n}{2}\right| \geq c\right) \leq \alpha.$$

13 October 20th, 2021

Today we will finish discussing the sign test, then introduce a bunch of other famous hypothesis tests: the Wilcoxon signed-rank test, the Kolmogorv-Smirnov test, the Mann-Whitney U test, and the permutation test.

13.1 Sign Test and Signed-Rank Test

Recall that the sign test asks whether the *median* of the sampled data $\mathbf{Y}$ is equal to θ_0. This test has statistic distributed according to $\mathrm{Bin}(n, \frac{1}{2})$ under the null hypothesis, which is given by

$$T(\mathbf{Y}) = \sum_{i=1}^{n} 1_{\{Y_i > \theta_0\}}.$$

What if $\mathbf{Y}$ is a discrete variable, so there may be positive probability mass on the median point $\Pr(Y = \theta_0) > 0$, and we would like to write a sign test for whether the median equals θ_0? In this case, instead of our single additive sign statistic, we could imagine describing a trinomial statistic $\mathrm{sgn}(Y_i - \theta_0) \in \{-1, 0, 1\}$ and test for

$$\Pr(\mathrm{sgn}(Y_i - \theta_0) = -1) \le 0.5,$$
$$\Pr(\mathrm{sgn}(Y_i - \theta_0) = 1) \le 0.5.$$

Alternatively, if we wanted to instead test if the two sides of θ_0 were exactly balanced in mass, there are a couple ways that we could implement such a test in practice with just one statistic:

- Ignore all samples that are equal to θ_0, testing the fact that under H_0, since

$$\Pr(Y > \theta_0 \mid Y \neq \theta_0) = \Pr(Y < \theta_0 \mid Y \neq \theta_0).$$

- For each sample $Y_i = \theta_0$, flip a fair coin to determine its sign, in order to break ties.

Another simple variant of the sign test can be derived by applying it to paired samples.

Definition 13.1 (Paired sample sign test)**.** Given i.i.d. samples $(X_1, Y_1), \dots, (X_n, Y_n) \sim \mathcal{D}$, the *paired sample sign test* has null hypothesis

$$H_0 : \Pr(Y_i - X_i > \theta_0) = \Pr(Y_i - X_i < \theta_0).$$

The test statistic in this case would be $T(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} 1_{\{Y_i - X_i > \theta_0\}}$.

Next, suppose we wanted to test a null hypothesis H_0 stating that the data is sampled from a *symmetrical* distribution around θ_0. In other words, the variable $|Y_i - \theta_0|$ is conditionally independent from the event $Y_i - \theta_0 > 0$. This can be handled with the following test.

Definition 13.2 (Wilcoxon signed-rank test)**.** Given data $Y_1, \dots, Y_n \sim \mathcal{D}$, the Wilcoxon signed-rank test for continuous random variables has test statistic

$$W = \sum_{i=1}^{n} \mathrm{sgn}(Y_i - \theta_0) R_i,$$

where $R_i = |Y_i - \theta_0|$ for each i. Under the null hypothesis that $\mathcal{D}$ is a symmetric distribution around θ_0, each sign is a Rademacher random variable that is independent from $R_i \in \mathrm{Unif}\{1, \dots, n\}$ which are integers drawn without replacement, so

$$\mathbf{E}\left[W\right] = 0, \quad \mathbf{Var}\left[W\right] = 1^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Furthermore, by the central limit theorem, W is asymptotically normal under H_0.

Note. In the discrete case of Wilcoxon's test, there may be cases when the ranks R_i are ambiguous due to ties. To fix this, we can either use a randomized tie-breaking strategy (which preserves the distribution of W but is non-deterministic), or correct for ties by setting each rank R_i to the average of tied ranks (which is deterministic but changes the distribution of W).

13.2 Two-Sample Tests

Now, let's discuss two-sample hypothesis tests. Here, we have some data $X_1, \ldots, X_n \sim F_X$ and $Y_1, \ldots, Y_m \sim F_Y$, both samples i.i.d. and satisfying $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$. Our null hypothesis will be that the two distrbutions are in fact the same, i.e., $F_X = F_Y$.

Definition 13.3 (Kolmogorv-Smirnov test)**.** The *Kolmogorv-Smirnov (K-S) test* tests distributional null hypotheses by comparing the cumulative distribution functions. There are two major variants of K-S, for the one-sample and two-sample cases.

- **One-sample variant:** Given a distribution F_0 and samples $Y_1, \ldots, Y_n \sim F$, the one-sample K-S test has null hypothesis $H_0 : F = F_0$. The test statistic relies on the cumulative distribution functions,

$$\sup_{y\in\mathbb{R}} \left| F_0(y) - \hat{F}_Y(y) \right|,$$

  where $\hat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^n 1_{\{Y_i \le y\}}$ is the empirical cumulative distribution function. Notice that this test statistic does not depend on F_0 when Y has a continuous distribution, since by a change of variables,

$$\begin{aligned}\sup_{y\in\mathbb{R}} \left| F_0(y) - \hat{F}_Y(y) \right| &= \sup_{\hat{y}\in(0,1)} \left| \hat{y} - \frac{1}{n}\sum_{i=1}^n 1_{\{F_0(Y_i)\le \tilde{y}\}} \right| \\ &= \sup_{\hat{y}\in(0,1)} \left| \hat{y} - \frac{1}{n}\sum_{i=1}^n 1_{\{U_i\le \tilde{y}\}} \right|,\end{aligned}$$

  where $U_i = F_0(Y_i) \sim \mathrm{Unif}[0,1]$. Visually, we can imagine taking an empirical estimate of the CDF for Y as a staircase function, then finding the maximum vertical deviation between this estimate and F_0.

- **Two-sample variant:** In the two-sample variant, we simply compare the empirical cumulative distribution functions of both variables, to get test statistic

$$\sup_{y\in\mathbb{R}} \left| \hat{F}_X(y) - \hat{F}_Y(y) \right|.$$

  Here, we are comparing two staircase functions against each other. Once again, for continuous random variables, this statistic doesn't depend on F_X and F_Y under the null hypothesis H_0.

Note that both of these variants have a test statistic that is asymptotically distributed according to the maximum absolute deviation of a *Brownian bridge* process on $[0, 1]$, which makes it possible to analytically compute cutoffs for a desired confidence level α, around $\Theta(-\log \alpha/\sqrt{n})$.

Although the Kolmogorv-Smirnov test is extremely general and applies in many cases, it is not as powerful as a test that is targeted towards specific alternative distributions, if more information is known about the data. Next, we'll introduce a more specific test that is similar to the paired-sample Wilcoxon signed-rank test, but it assumes independence of X and Y.

Definition 13.4 (Mann-Whitney U test)**.** Assume that we have independent data $X_1, \ldots, X_n \sim \mathcal{D}_X$ and $Y_1, \ldots, Y_m \sim \mathcal{D}_Y$. The null hypothesis H_0 states that $\mathcal{D}_X = \mathcal{D}_Y$, and the *Mann-Whitney U test* statistic is

$$U = \sum_{i=1}^{n} R(X_i) - \frac{n(n+1)}{2}.$$

Under the null hypothesis, $\mathbf{E}\left[U\right] = \frac{nm}{2}$ and $\mathbf{Var}\left[U\right] = \frac{nm(n+m+1)}{12}$. This is asymptotically normal as $n, m \to \infty$. Here, the alternative hypothesis for this test would be that Y *stochastically dominates* X, or vice versa, in that one of their CDFs is strictly greater than the other.

We have a midterm exam next week, and review materials will be posted on the course website.

14 October 25th, 2021

Today we discuss the permutation test, a general family of statistical tests for independence of paired draws from a bivariate distribution, as well as selective inference.

14.1 Tests of Independence

Suppose that we have i.i.d. random samples $(X_1, Y_1), \dots, (X_n, Y_n)$ drawn from an unknown bivariate distribution. We want to test the null hypothesis H_0, that $X_i \perp\!\!\!\perp Y_i$ are independent, against the alternative hypothesis that that they are not independent.

In the permutation test, the essential observation is that when $X \perp\!\!\!\perp Y$, we can permute the samples according to some arbitrary permutation σ, taking

$$(X_1, \dots, X_n, Y_1, \dots, Y_n) \longmapsto (X_{\sigma(1)}, \dots, X_{\sigma(n)}, Y_1, \dots, Y_n).$$

Notice that in the null hypothesis, the likelihood of any sample is identical to that of any of its $n!$ permuted variants. Therefore, for any statistic $T : \mathbb{R}^{2n} \to \mathbb{R}$, the distribution of $T(\mathbf{X}, \mathbf{Y})$ is the same as the distribution of its permuted variants.

Definition 14.1 (Permutation test)**.** Given random samples $(X_1, Y_1), \dots, (X_n, Y_n)$ from an unknown bivariate distribution, and a test statistic $T : \mathbb{R}^{2n} \to \mathbb{R}$, we define the following for any permutation π of $\{1, \dots, n\}$:

$$T^\pi = T(X_{\pi(1)}, \dots, X_{\pi(n)}, Y_1, \dots, Y_n).$$

Then, if Π_n is the set of all permutations of length n, and $\mathrm{id} \in \Pi_n$ is the identity element, the permutation test p-value is computed as

$$\frac{1}{n!} \sum_{\pi \in \Pi_n} 1_{\{T^\pi \geq T^{\mathrm{id}}\}}.$$

Notice that the p-value lies in the range $[1/n!, 1]$, and it is conservative. In other words, under H_0, the rejection region R_α where the $p \leq \alpha$ has likelihood $\Pr_{H_0}(R_\alpha) \leq \alpha$.

Note that the above test requires computing $n!$ statistics T^π for each $\pi \in \Pi_n$, which is intractable for most values of n. Instead, we can generate estimate p-values given $N \leq n!$ permutations, denoted $\pi_1, \dots, \pi_N$, sampled uniformly from Π_n. Here, the p-value is

$$\frac{1}{N+1} \left(1 + \prod_{j=1}^{N} 1_{\{T^{\pi_j} \geq T^{\mathrm{id}}\}} \right).$$

In the above expression, the extra addition of 1 replaces the always-true $1_{\{T^{\mathrm{id}} \geq T^{\mathrm{id}}\}}$ term of the full permutation test, which makes this a conservative p-value.

Notice that the permutation test is very flexible, since T can be *any function* without restrictions. If there is some expected relationship (e.g., quadratic relationship) between X and Y in the alternative hypothesis, T can be set to the regression coefficient of a quadratic fit, which would be much higher in the original data than the shuffled data.

Note. Another way of viewing the permutation test is in terms of sufficient statistics. The minimal sufficient statistic for general univariate probability distributions is simply the order statistics of the sampled data. Therefore, we can condition on the order statistics to get an unbiased estimate of the joint distribution under the null hypothesis.

14.2 Selective Inference

Now we introduce *selective inference*, which is the problem of soundly testing many hypotheses for statistical significance on data. It is increasingly common in scientific fields to want to test many hypotheses at once, but such a procedure can often lead to fallacies if care is not taken. Consider the following scenarios:

- For a single hypothesis, we could reasonably reject a null hypothesis at significance level 5%.
- However, if we test 20 hypotheses all at this same 5% significance level, then we should expect to get a false positive at this significance level, even if the null hypothesis H_0 is true.
- Even worse, given 20000 human genes, if we did a test at the 5% level, this would generate around 1000 false positives, which could completely drown out the signal for the few significant genes that we actually care about.

Therefore, in selective inference, we choose to primarily study not a single hypothesis, but rather a *family* of hypotheses. Given m null hypotheses $H_{0,1}, \dots, H_{0,m}$, let $\mathcal{H}_0$ be the subset of those that are true null hypotheses, and let $m_0 = |\mathcal{H}_0|$. (We sometimes also slightly abuse notation and let $\mathcal{H}_0$ refer to the set of indices of true null hypotheses.)

	True Nulls	False Nulls	Total
Selected	V	S	R
Not Selected	U	T	$m - R$
Total	m_0	$m - m_0$	m

Figure 2: Notation for selective inference.

Under some hypothesis test, we select some of the null hypotheses for rejection. Here, we let R be the number of selected hypotheses, and we define quantities V, S, U, and T according to Fig. 2.

Definition 14.2 (Familywise error rate (FWER))**.** The *familywise error rate* of a selection rule is the probability that any of the hypotheses in $\mathcal{H}_0$ is selected, or $\text{FWER} = \Pr(V > 0)$.

There are two kinds of control for the familywise error rate: *strong control*, where the error is tracked for any arrangement of true and false null hypotheses, and *weak control*, where we only bound the FWER when $m = m_0$. For the rest of this section, we only discuss strong control.

Definition 14.3 (False discovery rate (FDR))**.** We define the *false discovery proportion* of a selection rule to be

$$\text{FDP} = \frac{V}{V+S} = \frac{V}{R},$$

where 0/0 is taken to be 0. The *false discovery rate* is the expectation

$$\text{FDR} = \mathbf{E}\left[\text{FDP}\right] = \mathbf{E}\left[\frac{V}{R}\right],$$

where 0/0 is once again taken to be 0.

Note. Under the global null hypothesis, the FDR is equal to the FWER, as both are either 0 or 1. In general, we can observe that FWER $\geq$ FDR, so the FDR may be easier to control as a weaker error rate that admits procedures with higher power.

Most work in selective inference focuses on taking in p-values $p_1, \dots, p_m$ and returning a selected set of indices that controls a rate like the FWER or FDR, while maximizing power. Here's a first example of such a procedure.

Definition 14.4 (Bonferroni procedure)**.** Given m p-values $p_1, \dots, p_m$, the *Bonferroni procedure* for controlling the FWER at level α rejects all hypotheses with p-values lower than α/m.

This succinctly captures the core idea that with more hypotheses, we usually need to be stricter about p-value thresholds being low enough to reject. However, this procedure is relatively primitive, as it doesn't make any dependence or independence assumptions among the p-values.

Proposition 14.5. *The Bonferroni procedure controls the FWER at level α.*

Proof. Simply use a union bound:

$$\mathrm{FWER} = \Pr(V > 0) \le \sum_{k \in \mathcal{H}_0} \Pr\left(p_k \le \frac{\alpha}{m}\right) = \frac{m_0}{m}\alpha.$$

In fact, this means that the procedure is conservative by a factor of $m_0/m < 1$. □

15 October 27th, 2021

Today, we continue discussing selective inference procedures that control the FWER and FDR.

15.1 FWER-Controlling Procedures

Recall that the Bonferroni procedure rejects all p-values below α/m, where α is the FWER level and m is the number of p-values. This does not require independent samples.

Example 15.1. Suppose that we draw m p-values independently from the uniform distribution $\mathrm{Unif}[\frac{1}{m}, 1]$. Then, we pick one of the p-values uniformly among the m and redraw it from $\mathrm{Unif}[0, \frac{1}{m}]$. These values are now marginally $\mathrm{Unif}[0, 1]$, and the FWER of the Bonferroni procedure applied to these p-values, assuming the global null hypothesis, is exactly α.

What about the case when the hypotheses are independent? We can derive a *slightly* stronger procedure under this assumption, which has a larger p-value rejection threshold.

Definition 15.2 (Šidák procedure)**.** We reject all hypotheses below $\alpha_m = 1 - (1 - \alpha)^{1/m}$.

Proposition 15.3. *The Šidák procedure controls the FWER at level α when the null hypothesis p-values are independent, and it is tight when $m_0 = m$.*

Proof. This is a simple calculation using joint probability of independent events,

$$\mathrm{FWER} = \Pr(V > 0) = 1 - \Pr(V = 0) = 1 - (1 - \alpha)^{m_0} \leq 1 - (1 - \alpha)^m.$$

□

Note that by taking a Taylor series expansion,

$$\alpha_m \approx \frac{\alpha}{m}\left(1 + \frac{\alpha}{2} + \frac{\alpha^2}{3} + \cdots\right),$$

when m is large and α is small. This means that the p-value in the Šidák procedure is about $\alpha/2$ better than the Bonferroni procedure in relative terms. For example, when $\alpha = 0.05$, the Šidák procedure has a selection threshold that is $1.025\times$ the Bonferroni threshold.

We might be a little bit disappointed, since this procedure requires a strong independence assumption, yet it doesn't improve the threshold much. Indeed, the Bonferroni procedure captures most of the selection power with a simple rule. However, the following selection rule will show that Bonferroni can be uniformly improved without making any additional assumptions.

Definition 15.4 (Holm procedure)**.** Let $p_{(1)}, \ldots, p_{(m)}$ be p-values in increasing order, and let $H_{0,(1)}, \ldots, H_{0,(m)}$ be their corresponding null hypotheses. Then, the *Holm procedure* rejects nothing if $p_{(1)} > \alpha/m$, and otherwise, it rejects $H_{0,(1)}, \ldots, H_{0,(\hat{k})}$ where

$$\hat{k} = \min\left\{k : p_{(k+1)} > \frac{\alpha}{m-k}\right\}.$$

This is called a *step-down procedure*,[6] and it can be thought of as an iterative procedure that chooses to reject each null hypothesis in turn by comparing its p-value against the current threshold, assuming all previous null hypotheses are false.

[6] The name "step-down" is confusing, since it means the opposite: going in order of increasing p-value.

Note that the Holm procedure is strictly less conservative than the Bonferroni procedure, since it rejects at least as many null hypotheses.

Proposition 15.5. *The Holm procedure controls the FWER at level α.*

Proof. Let k_0 be the sorted index of the smallest null p-value. In other words,

$$k_0 = \min\{k : H_{0,(k)} \in \mathcal{H}_0\}.$$

Then, the family-wise error rate can be bounded by

$$\text{FWER} = \Pr(V > 0) \leq \Pr\left(p_{(k_0)} \leq \frac{\alpha}{m - k_0 - 1}\right) \leq \Pr\left(\min_{k \in \mathcal{H}_0} p_k \leq \frac{\alpha}{m_0}\right).$$

Finally, by a union bound, this expression is at most

$$\Pr\left(\min_{k \in \mathcal{H}_0} p_k \leq \frac{\alpha}{m_0}\right) \leq \sum_{k \in \mathcal{H}_0} \Pr\left(p_k \leq \frac{\alpha}{m_0}\right) = \alpha.$$

□

An equivalent way to think about the Holm method is as a non-stepwise procedure, which applies threshold $\alpha/(m - \hat{k} + 1)$ to all p-values. Of course, you still need to use the same stepwise method to actually compute $\hat{k}$, but it's an interesting conceptual way of thinking about it.

15.2 FDR-Controlling Procedures

The following procedure for controlling the FDR was introduced in 1995, and it has since become one of the most cited papers in the statistical literature, ubiquitous throughout science research for selective inference. It is also interesting from a statistical point of view.

Definition 15.6 (Benjamini-Hochberg procedure)**.** Let q be a level for the FDR. The *Benjamini-Hochberg (BH) procedure* rejects nothing if $\min_k P_{(k)}/k > q/m$. Otherwise, it rejects the hypotheses $H_{0,(1)}, \ldots, H_{0,(\hat{k})}$, where

$$\hat{k} = \max\left\{k : p_{(k)} \leq \frac{kq}{m}\right\}.$$

This is called a *step-up procedure*, and it can be viewed as iterating in order of descending p-value until finding one that is lower than the significance threshold.

Note. Even with the same rejection levels, the behavior of a step-up and step-down procedure differs based on where the p-value curve intersects the threshold curve.

16 November 1st, 2021

Today, we go over the proofs of correctness for selective inference procedures that control the FDR, and we discuss confidence intervals.

16.1 More on FDR Control

Recall that the Benjamini-Hochberg procedure from last lecture is a step-up procedure, which at level q, rejects the $\hat{k}$ smallest p-values, where $\hat{k}$ is the largest index k such that $p_{(k)} \leq kq/m$.

Theorem 16.1. *The Benjamini-Hochberg procedure controls the FDR at level q, assuming that the p-values are independent.*

Proof. Let $p_1, \dots, p_m$ be p-values, and without loss of generality, let p_1 be derived from the null hypothesis. Let R be the number of p-values rejected by the BH procedure. Observe that p_1 is rejected and $R = r$ if and only if p_1 is rejected and $\tilde{R} = r$, where

$$\tilde{R} = R(\{0, p_2, \dots, p_m\})$$

is the number of rejected p-values if p_1 were replaced with 0. This means that

$$\begin{aligned}
\text{FDR} &= \sum_{r=1}^{m} \mathbf{E}\left[\frac{v}{r} 1_{\{R=r\}}\right] \\
&= \sum_{r=1}^{m} \mathbf{E}\left[\frac{1}{r} \sum_{k \in \mathcal{H}_0} 1_{\{p_k \leq \frac{qr}{m}\}} 1_{\{R=r\}}\right] \\
&= \sum_{r=1}^{m} \frac{m_0}{r} \mathbf{E}\left[1_{\{p_1 \leq \frac{qr}{m}\}} 1_{\{R=r\}}\right] \\
&= \sum_{r=1}^{m} \frac{m_0}{r} \mathbf{E}\left[1_{\{p_1 \leq \frac{qr}{m}\}} 1_{\{\tilde{R}=r\}}\right] \\
&= \sum_{r=1}^{m} \frac{m_0}{r} \Pr\Big(p_1 \leq \frac{qr}{m}\Big) \Pr\Big(\tilde{R} = r\Big) \\
&\leq q \frac{m_0}{m} \underbrace{\sum_{r=1}^{m} \Pr(\tilde{R} = r)}_{=1} \\
&= q \frac{m_0}{m}.
\end{aligned}$$

Observe that the BH procedure is conservative by a factor of m_0/m, similar to the Bonferroni procedure. We can also compute the variance of the FDP (not just its mean), which is

$$\mathbf{Var}\left[\text{FDP}\right] = \frac{q m_0}{m} \left(\mathbf{E}\left[\frac{1}{1 + \tilde{R}}\right] - \frac{1}{m}\right) \leq q \, \mathbf{E}\left[\frac{1}{\max\{R, 1\}}\right].$$

For example, if $q = 0.1$ and $R \approx 50$, then $\sqrt{\mathbf{Var}\left[\text{FDP}\right]} \approx 0.05$. □

So far, we've seen the proof assuming independence of p-values, which is a nice simplifying assumption. However, it turns out that the BH procedure also controls the FDR (up to a log-factor) when the p-values are allowed to be arbitrarily dependent, which we will show now.

Theorem 16.2. *The Benjamini-Hochberg procedure controls the FDR at level*

$$q\frac{m_0}{m}\left(\sum_{i=1}^{m}\frac{1}{i}\right) \approx q\frac{m_0}{m}(\log(m)+0.577),$$

under arbitrary p-value dependence.[7]

Proof. Once again, we can compute that

$$\begin{aligned}
\mathrm{FDR} &= \sum_{r=1}^{m}\mathbf{E}\left[\frac{v}{r}1_{\{R=r\}}\right] \\
&= \sum_{r=1}^{m}\mathbf{E}\left[\frac{1}{r}\sum_{k\in\mathcal{H}_0}1_{\{p_k\le\frac{qr}{m}\}}1_{\{R=r\}}\right].
\end{aligned}$$

Now, we rewrite this inner sum as

$$\begin{aligned}
\mathrm{FDR} &= \sum_{k\in\mathcal{H}_0}\mathbf{E}\left[\sum_{r=1}^{m}\sum_{i=1}^{r}\frac{1_{\{R=r\}}}{r}1_{\{\frac{q(i-1)}{m}<p_k\le\frac{qi}{m}\}}\right] \\
&= \sum_{k\in\mathcal{H}_0}\mathbf{E}\left[\sum_{i=1}^{m}\sum_{r=i}^{m}\frac{1_{\{R=r\}}}{r}1_{\{\frac{q(i-1)}{m}<p_k\le\frac{qi}{m}\}}\right] \\
&= \sum_{k\in\mathcal{H}_0}\mathbf{E}\left[\sum_{i=1}^{m}\frac{1_{\{R\ge i\}}}{R}1_{\{\frac{q(i-1)}{m}<p_k\le\frac{qi}{m}\}}\right] \\
&\le \sum_{k\in\mathcal{H}_0}\sum_{i=1}^{m}\frac{1}{i}\Pr\left(\frac{q(i-1)}{m}<p_k\le\frac{qi}{m}\right) \\
&= q\frac{m_0}{m}\left(\sum_{i=1}^{m}\frac{1}{i}\right).
\end{aligned}$$

In the last step, we assumed that the individual p-values were tight up to the size of the tests, but the same bound would apply to conservative p-values by the rearrangement inequality. □

16.2 Confidence Intervals

We now introduce confidence intervals, which are commonly abbreviated CIs.

Definition 16.3 (Interval estimator)**.** Let $L(\mathbf{Y})$ and $U(\mathbf{Y})$ be a pair of functions of the data $\mathbf{Y}\sim f_\theta$ such that $L(\mathbf{Y})\le U(\mathbf{Y})$ almost surely. Then, $[L(\mathbf{Y}),U(\mathbf{Y})]$ is an *interval estimator* of θ if we infer that $L(\mathbf{Y})\le\theta\le U(\mathbf{Y})$.

Definition 16.4 (Coverage probability)**.** The *coverage* of an interval estimator is the probability

$$\Pr_\theta(L(\mathbf{Y})\le\theta\le U(\mathbf{Y})).$$

Our general goal is to design confidence intervals such that the coverage is large.

[7] The approximate numerical value $\gamma\approx 0.577$ is the Euler-Mascheroni constant.

Definition 16.5 (Confidence coefficient)**.** The *confidence coefficient* of an interval estimator is

$$\inf_{\theta \in \Theta} \left[\Pr_{\theta}(L(\mathbf{Y}) \le \theta \le U(\mathbf{Y})) \right].$$

Given these definitions, we are now ready to introduce confidence intervals.

Definition 16.6 (Confidence interval)**.** A *confidence interval* of level α is an interval estimator with confidence coefficient at least $1 - \alpha$.

Example 16.7. If $Y_1, \dots, Y_n \sim \text{Unif}[0, \theta]$ for $\theta > 0$, and $Y(n) = \max_{1 \le i \le n} Y_i$, consider the following two interval estimators:

- $[aY_{(n)}, bY_{(n)}]$, where $1 \le a < b$.
- $[c + Y_{(n)}, d + Y_{(n)}]$, where $0 \le c < d$.

In the first case, the coverage probability does not depend on θ, as we could just scale the model appropriately. However, in the latter case, the coverage probability depends strongly on θ.

Example 16.8. If $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, and we are interested in a 95%-confidence interval, where $\alpha = 0.05$, for μ, there are two common methods we can use:

- (z-test). If σ^2 is known, return $\overline{Y} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$.
- (t-test). If σ^2 is unknown, return $\overline{Y} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}$, where $s^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})/(n-1)$ is the unbiased sample variance.

The first method generally yields better results, since you have more information.

After setting the level of a confidence interval, the next question to ask is about how much power we get. We'll see next time that we should generally prefer confidence intervals that have shorter *length*, when both have equal coverage.

17 November 3rd, 2021

Today, we will discuss confidence intervals in more detail, including the two most common means of constructing confidence intervals.

17.1 CIs by Inverting Hypothesis Tests

One primary method of constructing confidence intervals is by using a hypothesis test. Suppose that for every θ_0 in the parameter space, we have a level-α test for the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative $H_1 : \theta \neq \theta_0$. Furthermore, let $A(\theta_0)$ be the acceptance region for this level-α test, meaning that

$$\Pr_{\theta_0}(\mathbf{Y} \in A(\theta_0)) \geq 1 - \alpha.$$

We can then define the set $C(\mathbf{Y}) = \{\theta : \mathbf{Y} \in A(\theta)\}$. Notice that

$$\Pr_{\theta}(\theta \in C(\mathbf{Y})) = \Pr_{\theta}(\mathbf{Y} \in A(\theta)) \geq 1 - \alpha.$$

Therefore, if the set $C(\mathbf{Y})$ turns out to be an interval (which will often happen in practice), then we have shown that $C(\mathbf{Y})$ is a $(1-\alpha)$-confidence interval for θ.

We have therefore shown that given a general hypothesis test for a point null hypothesis against the natural alternative, we can naturally arrive at an equivalent confidence interval for that test. Hypothesis tests are equally as powerful as confidence intervals in these cases; neither is more informative than the other. However, there are a few limitations to confidence intervals:

- Existence of a confidence interval requires stronger assumptions than a hypothesis test. For example, calculating a confidence interval implicitly assumes that the model generating the data is correct, whereas you could reject the null hypothesis in a hypothesis test even if the model was not correct.

- To generate a confidence interval using this method, you need hypothesis tests to exist for every value of the parameter θ, which isn't always easy.

- Hypothesis tests can answer much more general questions, including non-parametric ones. Confidence intervals only work when we're analyzing a parameter in a contiguous range. Sometimes, the confidence sets $C(\mathbf{Y})$ may not be contiguous or useful at all.

A classic example of getting a CI from a hypothesis test is the t-interval from Example 16.8, which is commonly seen in high school statistics classes. You can derive this interval by applying the method above with the t-test statistic.

Example 17.1. Let $Y_1, \ldots, Y_n$ be i.i.d. $\sim \theta \cdot \text{Expo}$, where $\theta > 0$, and we are interested in obtaining some $(1-\alpha)$-confidence interval for θ. First, we need to find a hypothesis test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Since the MLE of θ is just $\overline{Y}$, the likelihood ratio test statistic is

$$\Lambda = \frac{\overline{Y}^{-n} e^{-n}}{\theta_0^{-n} e^{-n\overline{Y}/\theta_0}} \propto \left(\frac{\overline{Y}}{\theta_0} e^{-\overline{Y}/\theta_0}\right)^{-n}.$$

Note that we have $n\overline{Y}/\theta_0 \sim \text{Gamma}(n)$, so the distribution of Λ is free of the parameter θ, and we can find some cutoff c_α such that the acceptance region of the size-α LR test is

$$A(\theta_0) = \left\{\mathbf{Y} : \frac{\overline{Y}}{\theta_0} e^{-\overline{Y}/\theta_0} > c_\alpha\right\}.$$

Next, observe that the function $f(x) = xe^{-x}$ is unimodal with a maximum at $x = 1$, so the acceptance region is really just an interval, and if $c_\alpha^{(1)} \le c_\alpha^{(2)}$ are the two points where $f(c_\alpha^{(1)}) = f(c_\alpha^{(2)}) = c_\alpha$, then a confidence interval for θ would be

$$C(\mathbf{Y}) = \left[\frac{\overline{Y}}{c_\alpha^{(2)}}, \frac{\overline{Y}}{c_\alpha^{(1)}} \right].$$

This was a more complicated example of a confidence interval derived from a hypothesis test, and hopefully it illustrates how the technique is applied in general.

17.2 CIs by Using Pivotal Quantities

The second method of constructing confidence intervals is using pivots. The definition of a pivot is similar to that of an ancillary statistic, but it is also allowed to use the parameter θ.

Definition 17.2 (Pivotal quantity)**.** A *pivotal quantity*, abbreviated *pivot*, is a function $Q(\mathbf{Y}, \theta)$ whose distribution does not depend on θ.

If we can find a closed set $\mathcal{A}$ such that $\Pr_\theta(Q(\mathbf{Y}, \theta) \in \mathcal{A}) \ge 1 - \alpha$, then it follows that the set of θ where $Q(\mathbf{Y}, \theta) \in \mathcal{A}$ is a $(1 - \alpha)$-confidence set, and if Q is monotone and $\mathcal{A}$ is an interval, then this is also a confidence interval.

Example 17.3. If we have samples $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, and σ^2 is known, then the z-score of $\overline{Y}$ is a pivot:

$$Q_1(\mathbf{Y}, \mu) = \frac{(\overline{Y} - \mu)\sqrt{n}}{\sigma} \sim \mathcal{N}(0, 1).$$

Otherwise, if σ^2 is unknown, then the following is also a pivot:

$$Q_2(\mathbf{Y}, \mu) = \frac{(\overline{Y} - \mu)\sqrt{n}}{s} \sim t_{n-1},$$

where $s^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \overline{Y})^2$ is the sample variance. Both Q_1 and Q_2 recover the z-interval and t-interval from Example 16.8, respectively.

Note that we could also have used Q_1 as a pivot in the case when σ^2 is unknown, but this would produce a 2-dimensional confidence set for (μ, σ^2). This is not useful if we just want an interval for μ. The advantage of the t-statistic is that it does not depend on σ^2.

Example 17.4. In Example 17.1, we inverted the acceptance region of the likelihood ratio test to obtain a confidence interval for the parameter θ. However, we observed that the likelihood ratio Λ was actually invariant in distribution based on the parameter θ, so it is a pivotal quantity.

We also could have chosen to invert the pivot $\overline{Y}/\theta_0$ directly, which would yield a slightly different confidence interval for θ, based on quantiles of the gamma distribution.

17.3 Criteria for Selecting CIs

The power of the confidence interval we get from a pivot largely depends on the quality of the pivot. Even after picking a pivot, we still need to choose some contiguous region $\mathcal{A} \subset \mathbb{R}$ where the quantity $Q(\mathbf{Y}, \theta)$ contains at least $1 - \alpha$ of its mass. There are a few desiderata:

- **Shortest-width**: Minimize the value of $U(\mathbf{Y}) - L(\mathbf{Y})$.

- **Equal-tails**: Satisfy the requirement that $\Pr_\theta(\theta > U(\mathbf{Y})) = \Pr_\theta(\theta < L(\mathbf{Y}))$.
- **Centered on an estimator**: For some estimator $\hat{\theta}$, choose so that $\hat{\theta} = \frac{L(\mathbf{Y})+U(\mathbf{Y})}{2}$.

It's fairly straightforward to satisfy the equal-tails and centered on an estimator requirements by simply looking at appropriate regions of mass $1-\alpha$ in the probability distribution of $Q(\mathbf{Y}, \theta)$. For the shortest-width requirement, this comes down to optimization, minimizing $b-a$ subject to

$$\int_a^b f(x)\,\mathrm{d}x \geq 1-\alpha,$$

where $f(x)$ is the probability density function of the pivot. In the case when the distribution is unimodal, the solution to this optimization problem is just a level set $\mathcal{A} = \{x : f(x) \geq c_\alpha\}$ of the probability density, with c_α selected so that the mass in $\mathcal{A}$ is at least $1-\alpha$.

18 November 8th, 2021

Today we continue discussing confidence intervals and introduce Bayesian inference.

18.1 Asymptotic Confidence Intervals

Similar to how we define asymptotic level-α hypothesis tests, we can also derive methods for constructing asymptotic $(1-\alpha)$-confidence intervals. Formally, an asymptotic CI is an interval estimator $[L_n(\mathbf{Y}), U_n(\mathbf{Y})]$, where n denotes the sample size, such that

$$\liminf_{n\to\infty} \Pr_\theta(L_n(\mathbf{Y}) \le \theta \le U_n(\mathbf{Y})) \ge 1-\alpha.$$

Analogous to last lecture, we can construct asymptotic confidence intervals by inverting asymptotic hypothesis tests (see Section 12.1), or by finding an asymptotic pivot, either by using variance-stabilizing transformations or other techniques.

Example 18.1. Suppose that $Y_1, \dots, Y_n \sim \text{Pois}(\lambda)$, and we would like to find a confidence interval for λ. The mean and variance are both λ, so the asymptotic distribution of the MLE $\overline{Y}$ is

$$\sqrt{n}(\overline{Y}-\lambda) \xrightarrow{d} \mathcal{N}(0,\lambda).$$

This means that $\sqrt{n}(\overline{Y}-\lambda)/\sqrt{\lambda} \xrightarrow{d} \mathcal{N}(0,1)$, which is an asymptotic pivot. Then, an asymptotic $(1-\alpha)$-confidence interval is

$$\begin{aligned} C(\mathbf{Y}) &= \left\{\lambda : \left|\frac{\sqrt{n}(\overline{Y}-\lambda)}{\sqrt{\lambda}}\right| \le z_{1-\alpha/2}\right\} \\ &= \left\{\lambda : n\lambda^2 - (2n\overline{Y} + z_{1-\alpha/2}^2)\lambda + n\overline{Y}^2 \le 0\right\}. \end{aligned}$$

By the quadratic formula, the endpoints of this confidence interval are at

$$\frac{2n\overline{Y} + z_{1-\alpha/2}^2 \pm \sqrt{4n\overline{Y}z_{1-\alpha/2}^2 + z_{1-\alpha/2}^4}}{2n}.$$

Example 18.2. Another confidence interval of the same level can be obtained with Slutsky's theorem and the consistency of the MLE, $\overline{Y} \xrightarrow{d} \lambda$, which implies that $\sqrt{n}(\overline{Y}-\lambda)/\sqrt{\overline{Y}} \xrightarrow{d} \mathcal{N}(0,1)$. This yields the slightly different confidence interval,

$$C(\mathbf{Y}) = \left\{\lambda : \overline{Y} - \sqrt{\overline{Y}/n}\, z_{1-\alpha/2} \le \lambda \le \overline{Y} + \sqrt{\overline{Y}/n}\, z_{1-\alpha/2}\right\}.$$

Example 18.3. What if we wanted to instead use a variance-stabilizing transformation? For the Poisson distribution, this transformation is $\lambda \mapsto \sqrt{\lambda}$, from which we can use the Delta method to get

$$\sqrt{n}\left(\sqrt{\overline{Y}} - \sqrt{\lambda}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\partial\sqrt{\lambda}}{\partial\lambda}\right)^2 \lambda\right) = \mathcal{N}\left(0, \frac{1}{4}\right).$$

This asymptotic confidence interval has endpoints at

$$\left(\sqrt{\overline{Y}} \pm \frac{z_{1-\alpha/2}}{2\sqrt{n}}\right)^2.$$

Each of these three examples yields a slightly different confidence interval in the finite-sample case, but when taking the limit to asymptotic behavior as $n \to \infty$, their corresponding confidence regions are equivalent.

18.2 Bayesian Inference

In Bayesian inference, rather than carefully considering individual parameter configurations in the set Θ, we instead imagine a *distribution* of parameters over the global set. In this setting, our usual frequentist model $f_\theta(y)$ is changed to a conditional distribution $f(y \mid \theta)$. The prior probability of i.i.d. samples $Y_1, \ldots, Y_n \sim f(y \mid \theta)$ is equal to the likelihood function

$$L(\theta) = f(\mathbf{Y} \mid \theta).$$

Then, using Bayes' rule, the posterior probability distribution is given by

$$\pi(\theta \mid \mathbf{Y}) = \frac{L(\theta)\pi(\theta)}{\int_\Theta L(\tilde{\theta})\pi(\tilde{\theta})\,\mathrm{d}\tilde{\theta}} \propto L(\theta)\pi(\theta).$$

Here, the integral in the denominator is called a *normalizing constant*, and it is usually intractable to compute this exactly due to the very difficult integral over Θ.

Example 18.4. How might we compute or estimate the value of $\mathbf{E}\left[\theta \mid \mathbf{Y}\right]$ in the Bayesian inference setting? One way to do this is is to discretize the parameter space Θ and generate a grid of values, then estimate the normalizing constant as a Riemann sum over this grid. This suffers from the curse of dimensionality.

Another, often more efficient method, is to use the *Markov chain Monte Carlo (MCMC)* algorithm or other sampling techniques to generate approximately i.i.d. samples from $\pi(\theta \mid \mathbf{Y})$. Although these techniques are not exact, they usually obtain good results after a fully-polynomial number of mixing steps. Then, an estimate for $\mathbf{E}\left[\theta \mid \mathbf{Y}\right]$ is obtained by taking the sample mean.

Although the general Bayesian inference problem is intractable (application of Bayes' rule), there are some distributions for which we can analytically compute the posterior distribution.

Definition 18.5 (Conjugate prior)**.** Suppose we have a family of distributions, $\mathcal{G} = \{g_\tau(\theta) : \tau \in \mathrm{T}\}$. Then, $\mathcal{G}$ is conjugate to $f(y \mid \theta)$ if $\pi(\theta) = g_\tau(\theta)$ for some $\tau \in \mathrm{T}$ means that for every y of the data drawn from $f(\bullet \mid \theta)$,

$$\pi(\theta \mid y) = g_{\tau'}(\theta)$$

for some $\tau' \in \mathrm{T}$.

In other words, a conjugate prior is a family of distributions for θ, such that updating your prior on sampled data still results in a distribution of that family.

Example 18.6. If $Y \sim \mathrm{Bin}(n, \theta)$, where n is known, then

$$f(y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Then, a conjugate prior for this distribution is $\pi(\theta) = \mathrm{Beta}(\alpha, \beta)$ for $\alpha, \beta > 0$, with PDF

$$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)},$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta function. Under this assumption on the prior, the posterior distribution for θ has a particularly simple form:

$$\pi(\theta \mid y) \propto \pi(\theta) f(y \mid \theta) \propto \theta^{y+\alpha-1}(1-\theta)^{(n-y)+\beta-1}.$$

Therefore, the Bayesian update is simply $(\alpha, \beta) \mapsto (y + \alpha, n - y + \beta)$.

Conjugate priors are useful because they are usually already known for common distributions, and they make it computationally efficient to apply inference techniques. However, one common criticism of conjugate priors is that your actual prior knowledge of the parameters of a distribution may not match any of the existing distributions $g_\tau(\theta)$. This limitation is somewhat reduced in severity by taking *mixtures* of conjugate priors.

Proposition 18.7. *If we have k prior distributions $\pi_1(\theta), \dots, \pi_k(\theta)$, such that each corresponding posterior distribution*

$$\pi_j(\theta \mid y) = \frac{L(\theta)\pi_j(\theta)}{C_j}$$

is computable, then if $\pi(\theta) = \sum_{j=1}^k w_j \pi_j(\theta)$ for weights $w_1, \dots, w_k$ summing to 1, we have

$$\pi(\theta \mid y) = \frac{L(\theta)\sum_{j=1}^k w_j\pi_j(\theta)}{C} = \sum_{j=1}^k \frac{w_j C_j}{C}\pi_j(\theta \mid y).$$

Therefore, $\pi(\theta \mid y)$ is a mixture of $\pi_j(\theta \mid y)$ for each j, with coefficients

$$w_j^* = \frac{w_j C_j}{\sum_{\ell=1}^k w_\ell C_\ell}.$$

Example 18.8. Given a continuous probability density function f for a scalar random variable, supported on $[-c, c]$, how might you approximate f as a mixture of Gaussians? One simple method would be to divide up the interval into N bins, then add a Gaussian centered at each bin with variance c/N and weight equal to the amount of probability mass in that bin.

18.3 Conjugate Priors of NEFs

One interesting property of natural exponential families is that they have a simple conjugate prior in the Bayesian inference setting.

Proposition 18.9. *Suppose that we have an NEF with distribution*

$$f(y \mid \eta) = \exp\{\eta y - \psi(\eta)\}h(y),$$

where we draw n i.i.d. samples $Y_1, \dots, Y_n \sim f(y, \eta)$. Our natural sufficient statistic is $\overline{Y} = T$. The mean and variance of the distribution are $\mu = \mathbf{E}[Y \mid \eta] = \psi'(\eta)$ and $\mathbf{Var}[Y \mid \eta] = \psi''(\eta) = V(\mu)$, respectively, and our likelihood function is

$$L(\eta) \propto \exp\left\{n\sum_{i=1}^n Y_i - n\psi(\eta)\right\} = \exp\{n(\eta T - \psi(\eta))\}.$$

Then, the following family of distributions, parameterized by (μ_0, r), is a conjugate prior:

$$\pi(\eta) \propto \exp\{r(\eta\mu_0 - \psi(\eta))\}.$$

Alternatively, we can reparameterize this in terms of the mean $\mu = \psi'(\eta)$, which yields

$$\pi(\mu) \propto \frac{1}{V(\mu)}\exp\left\{-r\int \frac{\mu - \mu_0}{V(\mu)}\,\mathrm{d}\mu\right\}.$$

Proof. Applying Bayes rule by multiplying the likelihood and prior distribution density function, we have that the posterior distribution is proportional to

$$\pi(\eta \mid \mathbf{Y}) \propto \exp\{(nT + r\mu_0)\eta - (n+r)\psi(\eta)\}.$$

From this, we can see that the prior $\pi(\eta)$ and posterior $\pi(\eta \mid \mathbf{Y})$ are in the same algebraic form. The update can be effectively summarized by:

$$\begin{aligned} r &\longmapsto n + r, \\ \mu_0 &\longmapsto \frac{r\mu_0 + nT}{n + r}. \end{aligned}$$

Furthermore, the mean of this distribution (or any distribution in the conjugate prior) is actually $\mathbf{E}[\mu] = \mu_0$. It is not obvious to show this, but see [DY79] for the proof. Therefore, the parameter can be interpreted as the prior expectation for the mean.

Furthermore, we can interpret the posterior update rule as a weighted average of the prior mean μ_0 and the observed mean $T = \overline{Y}$. The weights are determined by the relative Fisher information of two estimates μ_0 and T, treated as if they used r and n independent samples, respectively. Then, we have $\mathbf{E}[\mu \mid \mathbf{Y}] = B\mu_0 + (1 - B)T$, where

$$B = \frac{\frac{r}{V(\mu)}}{\frac{r}{V(\mu)} + \frac{n}{V(\mu)}}.$$

□

Intuitively, B is the *shrinkage factor*, which determines how much the prior mean μ_0 contributes to the posterior distribution, compared to the new evidence $\mathbf{Y}$.

19 November 10th, 2021

Today, we finish talking about conjugacy, then introduce the Jeffreys prior, as well as Bayesian point and interval estimation.

19.1 More on Conjugacy

Let's give a couple examples of the conjugate priors corresponding to natural exponential families, as determined by application of Proposition 18.9.

Example 19.1. If $Y \sim \mathcal{N}(\mu, c)$, where the variance function is $V(\mu) = c$ for known c, then

$$\int_{-\infty}^{\infty} \frac{\mu - \mu_0}{V(\mu)} \,\mathrm{d}\mu = \frac{1}{2c}\mu^2 - \frac{\mu_0}{c}\mu.$$

Since this is the form of the negative log-likelihood function, up to a constant, we conclude that the posterior distribution of μ is normal.

Example 19.2. If $Y \sim \text{Pois}(\mu)$, then $V(\mu) = \mu$, and the corresponding integral is

$$\int_{0}^{\infty} \frac{\mu - \mu_0}{V(\mu)} \,\mathrm{d}\mu = \mu - \mu_0 \log \mu.$$

Therefore, the posterior distribution $\pi(\mu)$ is a gamma distribution, since the likelihood is proportional to $\mu^{r\mu_0} e^{-r\mu} \sim \frac{1}{r} \cdot \text{Gamma}(r\mu_0 + 1)$.

19.2 The Jeffreys Prior

The Jeffreys prior is a kind of "standard" prior, which does not depend on a specific parameterization. In other words, the prior is equivariant on bijective transformations of the parameter θ. It turns out that this is enough to uniquely specify a prior.

Definition 19.3 (Jeffreys prior)**.** Given a model parameterized by θ, with Fisher information $I(\theta)$, the *Jeffreys prior* for θ is given implicitly by

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

It turns out that the Jeffreys prior is equivariant, which we formalize and prove below.

Proposition 19.4. *Given a model with parameter θ, the Jeffreys prior for θ is the same as the prior that comes from reparameterizing the Jeffreys prior for $\beta(\theta)$ and changing variables to θ, for any smooth bijective function β.*

Proof. We can simply compute via the chain rule that

$$\pi(\theta) = \pi(\beta)\left|\frac{\mathrm{d}\beta}{\mathrm{d}\theta}\right| \propto \sqrt{I(\beta)\left(\frac{\mathrm{d}\beta}{\mathrm{d}\theta}\right)^2} = \sqrt{\mathbf{E}_\beta\left[\left(\frac{\mathrm{d}\ell}{\mathrm{d}\beta}\right)^2\right]\left(\frac{\mathrm{d}\beta}{\mathrm{d}\theta}\right)^2} = \sqrt{\mathbf{E}_\theta\left[\left(\frac{\mathrm{d}\ell}{\mathrm{d}\beta} \cdot \frac{\mathrm{d}\beta}{\mathrm{d}\theta}\right)^2\right]} = \sqrt{I(\theta)}.$$

□

Example 19.5 (Jeffreys prior for a variance-stabilizing transformation)**.** If we apply a variance-stabilizing parameter $h(\theta)$ to our parameter such that $h'(\theta) \propto \sqrt{I(\theta)}$, then the Jeffreys prior on $h(\theta)$ has constant likelihood at every point; it is a uniform distribution.

Note that in the above example, if the parameter domain is infinite, then this prior is *improper* or *uninformative*, since it cannot be normalized to form a true distribution with nonzero mass. However, we can still use improper priors in calculations, as long as we are careful to normalize the posterior distribution to be proper.

Example 19.6 (Jeffreys prior for an NEF)**.** If $Y_1, \dots, Y_n \sim \mathrm{NEF}[\mu, V(\mu)]$, then

$$\mathbf{E}_\mu[Y_i] = \mu \quad \text{and} \quad \mathbf{Var}_\mu[Y_i] = V(\mu).$$

Then, observe that $\overline{Y} \sim \mathrm{NEF}[\mu, \frac{V(\mu)}{n}]$, and the Fisher information of $\overline{Y}$ with respect to the natural parameter η is

$$\begin{aligned} I(\eta) &= -\mathbf{E}_\eta\left[\frac{\mathrm{d}^2}{\mathrm{d}\eta^2}\log f(\overline{y} \mid \eta)\right] \\ &= n\psi''(\eta) \\ &= nV(\mu). \end{aligned}$$

Therefore, we conclude that the Jeffreys prior is $\pi(\eta) \propto \sqrt{V(\mu)}$. If we wanted the prior in terms of the mean μ instead of the natural parameter η, then we could transform it by a change of variables, which yields the expression

$$\pi(\mu) \propto \pi(\eta)\left|\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right| \propto \sqrt{V(\mu)}\left|\frac{\mathrm{d}\psi'(\eta)}{\mathrm{d}\eta}\right|^{-1} = \frac{1}{\sqrt{V(\mu)}}.$$

19.3 Bayesian Point and Interval Estimation

Let's talk a bit more about how we might do estimation in Bayesian inference. Recall that there are two main types of inference, which are point and interval estimation. The former, Bayesian point estimation, is fairly simple.[8]

Example 19.7 (Minimum-variance Bayesian estimation)**.** If our goal is to find an estimator $\hat{\theta}$ that minimizes the mean-squared error

$$\begin{aligned} \mathrm{MSE} &= \mathbf{E}\left[(\hat{\theta} - \theta)^2 \mid Y\right] \\ &= \mathbf{E}\left[(\theta - \mathbf{E}[\theta \mid Y])^2\right] + (\hat{\theta} - \mathbf{E}[\theta \mid Y])^2, \end{aligned}$$

then the optimal solution, is simply $\hat{\theta} = \mathbf{E}[\theta \mid Y]$.

There are a couple other common Bayesian point estimation techniques. For example, if we wanted to optimize for the minimum *absolute error*, then we would take the median of the posterior distribution $\theta \mid Y$, rather than its expectation. Likewise, if we wanted to find the highest posterior likelihood, then we would use the mode of the posterior distribution, which is the *maximum a posteriori (MAP) estimator*.

This is essentially all we will discuss about Bayesian point estimation. Hopefully you've seen MAP estimation in previous statistics classes. Now, we'll move on to interval estimation. The Bayesian version of a confidence interval is called a *credible interval*.

[8] Notice how Bayesian methods are much simpler than point estimation in the frequentist setting, which is why we spend more time in this class on frequentist inference.

Definition 19.8 (Credible interval)**.** An estimator $C(Y)$ is called a $(1-\alpha)$-*credible interval* if

$$\Pr(\theta \in C(Y)) \geq 1 - \alpha.$$

This is just a probability interval of $\pi(\theta \mid Y)$ with mass at least $1-\alpha$.

Similar to confidence intervals created by pivoting, the smallest credible interval for a unimodal posterior distribution is just the level set. Also, we may prefer an equal-tailed $C(Y)$, or for $C(Y)$ to be centered on some estimator $\hat{\theta}$, which is analogous to the analysis in Section 17.3.

Example 19.9 (Jeffreys prior for a normal distribution)**.** Suppose that $Y \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 = 100$ and we observe some value of Y. Then, Y is the maximum likelihood estimator for μ, as well as the mean, median, and mode of the posterior distribution for μ under Jeffreys prior.

If we instead used the conjugate prior $\mathcal{N}(100, 15^2)$ in the above example, then $r = \sigma^2/\tau^2 = 100/225 = 0.44$. The posterior distribution would be $\mathcal{N}(113.8, 69.2)$. Notice how the posterior variance is equal to half the harmonic mean of $\sigma^2 = 100$ and $\tau^2 = 225$, and the posterior mean can be computed by taking a weighted average of $\frac{r\cdot 100 + 1\cdot 120}{r+1}$.

Example 19.10 (Frequentist coverage of credible intervals)**.** Let's consider the frequentist coverage of the above intervals under various priors. Essentially, the coverage probability is a function of the parameter μ, and it specifies how likely it is that μ lies in the Bayesian credible interval.

When taking the Jeffreys prior, the Bayesian 95% credible interval is simply $Y \pm 1.96\sigma$, which coincides with the frequentist 95% confidence interval. Therefore, the coverage probability is precisely 0.95, which makes sense given that the Jeffreys prior for μ in $\mathcal{N}(\mu, \sigma^2)$ is essentially uninformative.

However, with a conjugate prior of $\pi \sim \mathcal{N}(\mu_0, \tau^2)$, where $r = \sigma^2/\tau^2$, we can calculate a general formula for the frequentist coverage of Bayesian credible intervals in our model, which is

$$\begin{aligned}
&\Pr_\mu\left(\mu \in \left[\left(\frac{r}{n+r}\mu_0 + \frac{n}{n+r}\overline{Y}\right) \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n+r}}\right]\right) \\
&= \Pr_\mu\left(\mu \in \left[\left(\frac{r}{n+r}\mu_0 + \frac{n}{n+r}\left(\mu + \frac{\sigma}{\sqrt{n}}Z\right)\right) \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n+r}}\right]\right) \\
&= \Pr_\mu\left(-\frac{\sigma\sqrt{n}}{n+r}Z \in \left[\frac{r}{n+r}(\mu_0 - \mu) \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n+r}}\right]\right) \\
&= \Pr_\mu\left(Z \in \left[\frac{r}{\sigma\sqrt{n}}(\mu - \mu_0) \pm z_{1-\alpha/2}\sqrt{\frac{n+r}{n}}\right]\right) \\
&= \Phi\left(\frac{\sigma}{\tau^2\sqrt{n}}(\mu - \mu_0) + z_{1-\alpha/2}\sqrt{1 + \frac{\sigma^2}{\tau^2 n}}\right) - \Phi\left(\frac{\sigma}{\tau^2\sqrt{n}}(\mu - \mu_0) - z_{1-\alpha/2}\sqrt{1 + \frac{\sigma^2}{\tau^2 n}}\right).
\end{aligned}$$

If we graph this expression, it roughly follows one's intuition for how the coverage probability should behave based on the conjugate prior. When n is small relative to r, if our prior is roughly correct ($\mu \approx \mu_0$), the frequentist coverage is higher than $1-\alpha$. However, if $|\mu - \mu_0|$ is greater than about τ, the coverage rapidly decreases because the inaccurate prior leads our interval astray.

As n increases relative to r, the frequentist coverage of the Bayesian credible interval at any μ converges to $1-\alpha$, since the prior distribution gets less important relative to the evidence.

In general, as n gets large, the frequentist coverage of Bayesian credible intervals gets close to the true coverage. This is a consequence of the Bernstein-von Mises theorem, which states that the Bayesian credible interval for a fixed prior is asymptotically normal and converges to the Wald confidence interval (Definition 12.2), as $n \to \infty$.

20 November 15th, 2021

Today, we discuss Bayesian predictive inference and decision analysis.

20.1 Bayesian Predictive Inference

Suppose that we have observed data $\mathbf{Y} = (Y_1, \ldots, Y_n)$, as well as some new data point Y_{new}. The problem of *predictive inference* asks us to find the posterior predictive distribution,

$$f(Y_{\text{new}} \mid \mathbf{Y}) = \int_\Theta f(Y_{\text{new}}, \theta \mid \mathbf{Y}) \, \mathrm{d}\theta = \int_\Theta f(Y_{\text{new}} \mid \theta) \pi(\theta \mid \mathbf{Y}) \, \mathrm{d}\theta.$$

Assume that $Y_{\text{new}} \mid \theta$ is known, since this model is parametric. Then, the predictive inference problem is essentially equivalent to the Bayesian update framework that we have already discussed in class, using the data $\mathbf{Y}$ to find the posterior distribution $\pi(\theta)$ using Bayes' rule. Predictive intervals can also be computed using the same method.

20.2 Frequentist Decision Analysis

Decision analysis is concerned with the following choice scenario. The statistician chooses a function δ from the sample space to the decision space $\mathcal{D}$, and nature chooses a parameter θ from a *data-generating process (DGP)*. Then, depending on what the statistician cares about, our objective is to minimize a loss function

$$L(\theta, \delta(Y)) \in \mathbb{R}_{\geq 0}.$$

(This problem can be extended to multi-dimensional data, but we will consider the one-dimensional case $Y \in \mathbb{R}$ in this section because it is simpler.) Some examples of loss functions that are commonly used in statistics include:

- (Mean squared error). $L(\theta, \delta(Y)) = (\theta - \delta(Y))^2$.
- (Mean absolute error). $L(\theta, \delta(Y)) = |\theta - \delta(Y)|$.
- (Asymmetric linear error). For some $p, q > 0$,

$$L(\theta, \delta(Y)) = \begin{cases} p(\theta - \delta(Y)) & \text{if } \theta \geq \delta(Y), \\ q(\delta(Y) - \theta) & \text{if } \theta < \delta(Y). \end{cases}$$

- (Linear exponential loss). For some small value of c,

$$L(\theta, \delta(Y)) = e^{c(\theta - \delta(Y))} - c(\theta - \delta(Y)) - 1 \approx \frac{1}{2} c^2 (\theta - \delta(Y))^2.$$

Now, we discuss how to analyze and solve decision analysis problems in a frequentist context. Core to this idea is the following notion.

Definition 20.1 (Risk)**.** The *risk function* for a decision analysis problem is

$$R(\theta, \delta) = \mathbf{E}_\theta \left[L(\theta, \delta(Y)) \right].$$

Definition 20.2 (Domination)**.** We say that a decision rule δ_1 *dominates* another rule δ_2 if

$$R(\theta, \delta_1) \leq R(\theta, \delta_2) \qquad \forall \theta \in \Theta,$$

with strict inequality holding for at least one $\theta \in \Theta$. On the other hand, we say that δ_1 is *as good as* δ_2 if neither rule dominates the other.

Definition 20.3 (Admissibility)**.** A decision rule δ is called *admissible* if there is no other rule $\delta' \in \mathcal{D}$ that dominates it.

Roughly speaking, we can think of admissible rules as being better than every other decision rule for at least one value of the parameter θ. These definitions have a very frequentist flavor, as we assume no prior information about the distribution of the parameters. However, while inadmissible rules are usually poor, not all admissible rules are necessarily good either.

Example 20.4. Consider $Y_1, \ldots, Y_n \sim \mathcal{N}(\theta, 1)$, and let our loss function be the mean squared error. A simple and reasonable decision rule in this case might be $\delta(\mathbf{Y}) = \overline{Y}$, which happens to be the UMVUE. However, setting $\delta(\mathbf{Y}) = 5$ to be a constant is technically a valid decision rule, since for $\theta = 5$, this is the unique rule with the global minimum risk of $R(\theta, \delta) = 0$.

Admissibility is a simple notion that simply checks, intuitively, that δ is good somewhere. Next, we will introduce the notion of *minimaxity*, which intuitively checks that δ is terrible nowhere.

Definition 20.5 (Minimax decision rule)**.** We say that δ_m is *minimax* with respect to $\mathcal{D}$ if

$$\delta_m \in \operatorname*{argmin}_{\delta \in \mathcal{D}} \max_{\theta \in \Theta} R(\theta, \delta).$$

In other words, δ_m minimizes the global maximum risk for an unknown value of θ.

Notice that minimax decision rules are simple to find. It simply requires computing, for each $\delta \in \mathcal{D}$, the maximum of $R(\theta, \delta)$ for all $\theta \in \Theta$. Then, you take δ that results in the smallest of these maximum-risk values, and that is the minimax rule.

Proposition 20.6 (Pencil problem)**.** *If a decision rule δ^* with constant $R(\theta, \delta^*) = c(\delta^*)$ is admissible in $\mathcal{D}$, then it is minimax in $\mathcal{D}$.*

Proof. Assume for the sake of contradiction that there exists $\delta' \in \mathcal{D}$ such that δ' has smaller maximum risk than δ^*. Then,

$$R(\theta, \delta') \leq \max_{\theta \in \Theta} R(\theta, \delta') < \max_{\theta \in \Theta} R(\theta, \delta^*) = c(\delta^*) = R(\theta, \delta^*),$$

where the last step follows because the risk function is a constant for $\theta \in \Theta$. Therefore, δ' dominates δ^*, which is a contradiction, so we conclude. □

20.3 Bayesian Decision Analysis

Next, we shift gears to discuss the Bayesian formulation of decision analysis, which is fairly different because we have a prior on the parameter distribution $\theta \sim \pi$.

Definition 20.7 (Bayes risk)**.** The *Bayes risk* function is

$$B(\pi, \delta) = \mathbf{E}_\pi \left[R(\theta, \delta) \right] = \mathbf{E}_\pi \left[\mathbf{E}_\theta \left[L(\theta, \delta(Y)) \right] \right].$$

Unlike frequentist inference, where we introduced two different criteria for desirable decision rules, the Bayesian formulation has an unambiguous criterion for a "Bayes rule" — not to be confused with Bayes' rule or Bayes' theorem.

Definition 20.8 (Bayes decision rule)**.** A decision rule $\delta \in \mathcal{D}$ is called a *Bayes rule* with respect to π and $\mathcal{D}$ if

$$\delta^{\pi} \in \operatorname*{argmin}_{\delta \in \mathcal{D}} B(\pi, \delta).$$

Proposition 20.9. *Suppose that $Y \sim f_{\theta}$, for $\theta \in \Theta$, and we have a prior $\pi(\theta)$, decision space $\mathcal{D}$, and loss function L. If $\mathcal{D}$ contains a δ^* minimizing*

$$\mathbf{E}_{\pi}\left[L(\theta, \delta(Y)) \mid Y\right]$$

almost surely with respect to the marginal distribution of Y,

$$m(y) = \mathbf{E}_{\pi}\left[f_{\theta}(y)\right] = \int_{\Theta} f_{\theta}(y)\pi(\theta)\,\mathrm{d}\theta,$$

then δ^* is is a Bayes rule. In other words, Bayes rules are precisely the decision rules that minimize the expected loss with respect to the posterior distribution of Y.

Proof. If all $\delta \in \mathcal{D}$ have infinite Bayes risk, then all of them are Bayes rules. Otherwise, suppose that there exists $\delta \in \mathcal{D}$ with $B(\pi, \delta) < \infty$. Then,

$$\begin{aligned} B(\pi, \delta) &= \mathbf{E}_{\pi}\left[\mathbf{E}_{\delta}\left[L(\theta, \delta(Y)\right]\right] \\ &= \mathbf{E}_{m}\left[\mathbf{E}_{\pi}\left[L(\theta, \delta(Y) \mid Y\right]\right] \\ &\geq \mathbf{E}_{m}\left[\mathbf{E}_{\pi}\left[L(\theta, \delta^*(Y) \mid Y\right]\right] \\ &= B(\pi, \delta^*). \end{aligned}$$

Therefore, δ^* has Bayes risk that is less than or equal to the Bayes risk of δ. □

21 November 17th, 2021

Today, we illustrate a link between frequentist and Bayesian decision analysis by showing that under some reasonable regularity conditions on the model, any Bayes rule is admissible, and conversely, every admissible rule is Bayes.

21.1 Bayes Rules are Admissible

First, we show the basic result, that if a rule is unique and Bayes, then it is also admissible, meaning that it is not dominated by any other decision rule.

Theorem 21.1. *Any unique Bayes rule (up to equality, almost surely in m) is admissible.*

Proof. Suppose for the sake of argument that there is some decision rule δ' such that

$$R(\theta, \delta') \leq R(\theta, \delta^\pi),$$

for all $\theta \in \Theta$. Then, integrating this over $\theta \sim \pi$ implies that

$$\begin{aligned} B(\pi, \delta') &= \mathbf{E}_\pi \left[R(\theta, \delta')\right] \\ &\leq \mathbf{E}_\pi \left[R(\theta, \delta^\pi)\right] \\ &= B(\pi, \delta^\pi). \end{aligned}$$

However, δ^π is defined as the unique global minimizer of the Bayes risk under parameter distribution π, so this implies that $B(\pi, \delta') = B(\pi, \delta^\pi)$, and $\delta' = \delta^\pi$ almost surely. □

The above argument was fairly simple, but we can also prove variants of this fact. The following variant works for finite parameter spaces, even if for Bayes rules that are not necessarily unique. It uses the fact that a dominating rule must be different in at least one risk value.

Theorem 21.2. *Let $\Theta = \{\theta_1, \dots, \theta_k\}$ be a finite set of parameters, and let δ^π be Bayes for π with finite Bayes risk, where $\pi(\theta_j) > 0$ for all j. Then, δ^π is admissible.*

Proof. Suppose for the sake of argument that there exists a rule δ' such that

$$R(\theta_j, \delta') \leq R(\theta_j, \delta^\pi),$$

for all $j \in \{1, \dots, k\}$, with strict equality for some j_0. Then,

$$\begin{aligned} B(\pi, \delta') &= \sum_{j=1}^{k} \pi(\theta_j) R(\theta, \delta') \\ &\leq \pi(\theta_{j_0}) R(\theta_{j_0}, \delta') + \sum_{j \neq j_0} \pi(\theta_j) R(\theta, \delta^\pi) \\ &< \pi(\theta_{j_0}) R(\theta_{j_0}, \delta^\pi) + \sum_{j \neq j_0} \pi(\theta_j) R(\theta, \delta^\pi) \\ &= B(\pi, \delta^\pi). \end{aligned}$$

This is a contradiction, so we conclude that it is impossible. □

Can we generalize this same result to continuous parameter spaces? Initially, this might seem difficult, since the strict inequality is harder to use in an infinite parameter space. However, with some analysis, it turns out that the answer is yes: we can, assuming extra regularity conditions.

Theorem 21.3. *Let $\Theta \subseteq \mathbb{R}$ be an open set, such that $\pi(\theta)$ has support Θ, and $R(\theta, \delta)$ is continuous in θ, for all $\delta \in \mathcal{D}$. Then, if δ^π is a Bayes rule for π, it is admissible.*

Proof. Suppose for the sake of argument that δ' is a rule such that

$$\begin{aligned} R(\theta, \delta') &\leq R(\theta, \delta^\pi), && \forall \theta \in \Theta, \\ R(\theta_0, \delta') &< R(\theta_0, \delta^\pi), && \text{for some } \theta_0 \in \Theta. \end{aligned}$$

Then, let $\eta = R(\theta_0, \delta^\pi) - R(\theta_0, \delta') > 0$. By continuity, there exists some $\epsilon > 0$ such that for all θ in an open ball A of radius ϵ around θ_0,

$$R(\theta, \delta^\pi) - R(\theta, \delta') > \frac{\eta}{2}.$$

This region A has positive probability mass because the support of π is over all of Θ, so

$$B(\pi, \delta^\pi) - B(\pi, \delta') > \frac{\eta}{2}\pi(A) > 0.$$

Thus, we have a contradiction, as we assumed that δ^π is a Bayes rule. □

21.2 Admissible Rules are Bayes

Next, we discuss the converse result, which is that admissible rules are Bayes. Before we can prove anything, we'll look a little bit at the high-dimensional geometry of admissible decision rules.

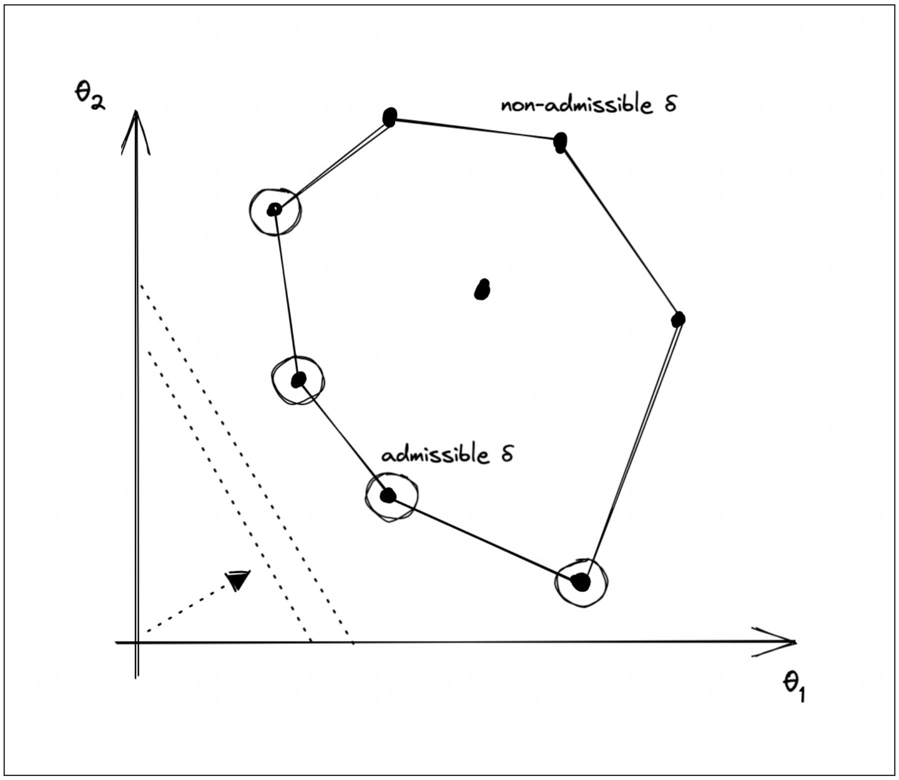

Figure 3: Geometric drawing of admissible decision rules.

As an initial example, assume that $\Theta = \{\theta_1, \theta_2\}$, and suppose that we map each decision rule δ to a geometric point $(R(\theta_1, \delta), R(\theta_2, \delta)) \in \mathbb{R}^2$. Then, each point is located in the upper-right quadrant of the Cartesian plane, and the convex hull of the points represents every risk configuration that can arise as the randomized mixture of one or more decision rules. See Fig. 3 for a diagram depicting this geometry, where each point is a rule $\delta \in \mathcal{D}$, and the lower-left boundary of the hull represents precisely the admissible rules that are not dominated by any other rule.

Proposition 21.4. *Assume that $|\Theta| = k$, and let $\mathcal{S} \subset \mathbb{R}^k$ be the convex hull of the points $\mathbf{s}_\delta = (R(\theta_1, \delta), \ldots, R(\theta_k, \delta))$ for all $\delta \in \mathcal{D}$. Let the lower orthant of a point in $\mathbb{R}^k$ be defined as*

$$Q(\mathbf{s}) = \{(x_1, x_k) \in \mathbb{R}^k \mid x_1 \leq s_q, \ldots, x_k \leq s_k\}.$$

Then, δ is admissible if and only if $Q(\mathbf{s}_\delta) \cap \mathcal{S} = \{\mathbf{s}_\delta\}$.

This proposition is fairly straightforward, and it makes sense, given that a decision rule is only valid if it occupies an optimal point on the frontier of possibilities. Now we might ask what the Bayes rules in the picture are. It turns out that they are precisely the same; the Bayes rules are solutions to linear programming problems optimizing the value of some linear form (represented by the dotted lines) on the convex set $\mathcal{S}$, so they can only occur on a lower-left boundary point of the hull. We formalize this intuition with the following theorem.

Theorem 21.5 (Complete class theorem for finite Θ)**.** *Let $|\Theta| = k$ be a set of parameters, and let $\mathcal{D}$ be a set of decision rules that is closed under randomized combinations. Furthermore, assume that $R(\theta, \delta)$ is a nonnegative risk function taking finite values for all $\delta \in \mathcal{D}$ and $\theta \in \Theta$. Then, if $\delta^* \in \mathcal{D}$ is admissible, it is a Bayes rule with respect to some proper prior.*

We will prove this theorem in the next lecture. Note that the closure under randomized combinations is an important condition, as otherwise, the set $\mathcal{S}$ would not be guaranteed to be convex, so not all admissible rules on the would be accessible as the minimum of a linear form corresponding to the supporting hyperplane.

22 November 22nd, 2021

Today, we first prove the complete class theorem for finite Θ as stated in last lecture, then we discuss the admissibility of the sample mean in univariate normals.

22.1 Proof of the Complete Class Theorem

First, we pick up from where we left off last time and prove the complete class theorem.

Proof of Theorem 21.5. Once again, we use the notation where $\mathcal{S} \subset \mathbb{R}^k$ is a convex set containing the risk values $\mathbf{s}_\delta$ of hypotheses $\delta \in \mathcal{D}$, where

$$\mathbf{s}_\delta = (R(\theta_1, \delta), \dots, R(\theta_k, \delta)).$$

Then, applying Proposition 21.4 to the admissible hypothesis δ^*, we get that $Q(\mathbf{s}^*) \cap \mathcal{S} = \{\mathbf{s}^*\}$, where we abbreviate $\mathbf{s}_{\delta^*} = \mathbf{s}^*$ for clarity. Then, if we let $\tilde{Q}(\mathbf{s}^*) = Q(\mathbf{s}^*) \setminus \{\mathbf{s}^*\}$, we have

$$\tilde{Q}(\mathbf{s}^*) \cap \mathcal{S} = \emptyset.$$

Notice that $\mathbf{s}^*$ is an extreme point of $Q(\mathbf{s}^*)$, so removing it maintains that the resultant set is still convex. By the hyperplane separation theorem,[9] there exists some nonzero normal vector $\mathbf{w} \in \mathbb{R}^k$ representing a linear form, such that

$$\sup_{\mathbf{x} \in Q(\mathbf{s}^*)} \mathbf{w}^\top \mathbf{x} \le \inf_{\mathbf{s} \in \mathcal{S}} \mathbf{w}^\top \mathbf{s}.$$

By inspecting the structure of Q, clearly all coordinates of $\mathbf{w}$ must be nonnegative. Finally, we can turn $\mathbf{w}$ into a proper prior on Θ by taking

$$\forall j : \ \pi(\theta_j) = \frac{w_j}{w_1 + \cdots + w_k}.$$

Observe that $B(\pi, \delta) \propto \mathbf{w}^\top \mathbf{s}_\delta$ for all δ, so δ^* must be a Bayes rule for the prior π, as desired. □

22.2 Admissibility of the Sample Mean

Before showing that the sample mean is admissible for any loss function, we will first do an technical thought experiment to get some intuition about its properties.

Example 22.1. Suppose that $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, and σ^2 is known. Also, suppose that our prior is $\mu \sim \mathcal{N}(\mu_0, \tau^2)$. Then, our shrinkage factor is

$$B = \frac{\sigma^2/n}{\sigma^2/n + \tau^2}.$$

As $n \to \infty$, the shrinkage factor approaches $B = 0$, so the prior makes a smaller and smaller effect on the final value of the posterior mean. Therefore, no matter what prior we choose, the limit of the Bayes risk-minimizing decision rule is just the sample mean.

Motivated by this example, here is the big result that we're excited about.

[9] This is a key property of convex sets in Euclidean space and will be our sledgehammer in this proof.

Theorem 22.2 (Blyth's method)**.** *Given i.i.d. $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, with variance σ^2 known, the sample mean $\overline{Y}$ is an admissible decision rule for μ with respect to the squared error.*

Proof. Suppose for the sake of argument that $\overline{Y}$ is inadmissible, so by definition, there exists some decision rule $t(\mathbf{Y})$ that dominates $\overline{Y}$. Therefore, for any μ,

$$R(\mu, t) = \mathbf{E}_\mu \left[(t(\mathbf{Y}) - \mu)^2\right] \leq \mathbf{E}_\mu \left[(\overline{Y} - \mu)^2\right] = R(\mu, \overline{Y}),$$

with strict inequality holding for some μ_0. Without loss of generality, assume that $\mu_0 = 0$. Then, since our risk function is continuous in μ, we can argue that there exists some $\epsilon > 0$ such that for any $|\mu| < \epsilon$, we have

$$R(\mu, \overline{Y}) - R(\mu, t) \geq \epsilon.$$

Now, assume for the sake of simplicity that $\sigma^2/n = 1$, as the particular value will not matter. Then, $R(\mu, \overline{Y}) = 1$, so we can rewrite this as

$$R(\mu, t) = \mathbf{E}_\mu \left[(t(\mathbf{Y}) - \mu)^2\right] \leq 1 - \epsilon 1_{\{\mu \in (-\epsilon, \epsilon)\}}.$$

Furthermore, given a prior $\pi_{\tau^2} \sim \mathcal{N}(0, \tau^2)$, we see that

$$\frac{\tau^2}{1+\tau^2}\overline{Y}$$

is a Bayes rule with respect to π_{τ^2} for all $\tau^2 > 0$. Hence, using the fact that Bayes rules minimize the expectation of the posterior risk given their prior, we have

$$\begin{aligned}
\mathbf{E}_{\pi_{\tau^2}} \left[1 - \epsilon 1_{\{\mu \in (-\epsilon, \epsilon)\}}\right] &\geq \mathbf{E}_{\pi_{\tau^2}} \left[\mathbf{E}_\mu \left[(t(\mathbf{Y}) - \mu)^2\right]\right] \\
&\geq \mathbf{E}_{\pi_{\tau^2}} \left[\mathbf{E}_\mu \left[\left(\frac{\tau^2}{1+\tau^2}\overline{Y} - \mu\right)^2\right]\right] \\
&= \mathbf{E}_{\pi_{\tau^2}} \left[B^2\mu^2 + (1-B)^2\right] \\
&= B^2\tau^2 + (1-B)^2 \\
&= 1 - \frac{1}{1+\tau^2}.
\end{aligned}$$

However, the original risk on the left-hand side of this inequality can also be written as

$$\begin{aligned}
\mathbf{E}_{\pi_{\tau^2}} \left[1 - \epsilon 1_{\{\mu \in (-\epsilon, \epsilon)\}}\right] &= 1 - \epsilon \mathrm{Pr}_{\pi_{\tau^2}}(-\epsilon < \mu < \epsilon) \\
&= 1 - \epsilon\Big(\Phi\Big(\frac{\epsilon}{\tau}\Big) - \Phi\Big(-\frac{\epsilon}{\tau}\Big)\Big) \\
&= 1 - \frac{2\epsilon^2}{\tau}\left(\frac{\Phi\big(\frac{\epsilon}{\tau}\big) - \Phi\big(-\frac{\epsilon}{\tau}\big)}{2\epsilon/\tau}\right).
\end{aligned}$$

Combining this with the previous inequality yields

$$1 - \frac{2\epsilon^2}{\tau}\left(\frac{\Phi\big(\frac{\epsilon}{\tau}\big) - \Phi\big(-\frac{\epsilon}{\tau}\big)}{2\epsilon/\tau}\right) \geq 1 - \frac{1}{1+\tau^2}.$$

Finally, we can rearrange to get the inequality

$$2\epsilon^2\left(\frac{\Phi\big(\frac{\epsilon}{\tau}\big) - \Phi\big(-\frac{\epsilon}{\tau}\big)}{2\epsilon/\tau}\right) \leq \frac{\tau}{1+\tau^2}.$$

Now, we examine the limit behavior of this inequality, as our prior tends towards the Jeffreys prior when $\tau^2 \to \infty$. Using the difference quotient on the left-hand side, we see that it converges to

$$2\epsilon^2\Phi'(0) = \frac{2\epsilon^2}{\sqrt{2\pi}}.$$

On the other hand, the right-hand side converges to zero, so we have a contradiction. □

Here is a nice bonus property of the sample mean, which we get for free.

Corollary 22.2.1. *If $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 known, then $\overline{Y}$ is minimax for μ with respect to the squared error.*

Proof. Recall that an admissible decision rule with constant risk must also be minimax. The sample mean $\overline{Y}$ is an admissible decision rule with constant risk σ^2/n, invariant of μ. □

22.3 Least Favorable Priors

So far, we have been looking at the relationship between Bayes rules and admissible rules. In this section, we turn our attention to minimaxity, which is connected to *least favorable priors.*

Definition 22.3 (Least favorable prior)**.** If $r_\pi = B(\pi, \delta^\pi)$, then a prior π^* is called *least favorable* if $r_{\pi^*} \geq r_\pi$ for all proper priors π. In other words,

$$\pi^* \in \operatorname*{argmax}_{\pi} \min_{\delta} B(\pi, \delta).$$

Theorem 22.4. *If δ^π is Bayes with respect to a prior π such that $r_\pi = \sup_\theta R(\theta, \delta^\pi)$, then:*

(i) δ^π is a minimax decision rule,

(ii) δ^π is the unique minimax decision rule if δ^π is the unique Bayes rule for π, and

(iii) π is a least favorable prior.

Proof. To prove (i), note that any other decision rule δ has a maximum risk value over all parameters θ that cannot be smaller than that of δ^π, since

$$\sup_\theta R(\theta, \delta) \geq \mathbf{E}_\pi\left[R(\theta, \delta)\right] \geq \mathbf{E}_\pi\left[R(\theta, \delta^\pi)\right] = r_\pi = \sup_\theta R(\theta, \delta^\pi).$$

Furthermore, if δ^π is the unique Bayes rule for π, then the second step above becomes a strict inequality $\mathbf{E}_\pi\left[R(\theta, \delta)\right] > \mathbf{E}_\pi\left[R(\theta, \delta^\pi)\right]$, so (ii) follows. Finally, to prove (iii), note that for any other prior $\tilde{\pi}$ on the parameters, we have

$$r_{\tilde{\pi}} = B(\tilde{\pi}, \delta^{\tilde{\pi}}) \leq B(\tilde{\pi}, \delta^\pi) \leq \sup_\theta R(\theta, \delta^\pi) = r_\pi.$$

□

What are the scenarios where we can use the above theorem? Well, one such scenario when r_π achieves this supremum is when our risk $R(\theta, \delta^\pi)$ is the same for all values of θ.

Corollary 22.4.1. *If a Bayes rule δ^π has constant risk, then it is minimax.*

23 November 29th, 2021

Today, we discuss the admissibility and minimaxity of the sample mean (also the maximum likelihood estimator) in any dimension and show the counterintuitive result that admissibility does not generalize to higher dimensions (Stein's paradox).

23.1 More on Minimax Decision Rules

Recall that a minimax decision rule δ^* minimizes the max risk over all parameters: $\sup_\theta R(\theta, \delta^*)$. Furthermore, a Bayes rule for a least favorable prior is also minimax, as we showed at the end of the last lecture.

Unfortunately, if we want to apply this theorem to prove that some decision rule δ is minimax, we actually need to show that our estimator is a Bayes rule with respect to some proper prior, so this doesn't work for things like the sample mean, which is Bayes with respect to the Jeffreys prior. To get around this, we will extend our definition to sequences of priors.

Definition 23.1 (Least favorable sequence)**.** A sequence of priors π_k is called *least favorable* if

$$\lim_{k\to\infty} r_{\pi_k} \geq r_\pi.$$

for all proper priors π.

Theorem 23.2 (Minimax duality)**.** *Let π_k be a sequence of priors and δ a decision rule. If*

$$\sup_\theta R(\theta, \delta) = \lim_{k\to\infty} r_{\pi_k},$$

then δ is minimax, and $\{\pi_k\}$ is a least favorable sequence of priors.

Proof. Note that the inequality version of the stated condition always holds, since if δ' is any decision rule, then for any k,

$$\sup_\theta R(\theta, \delta') \geq B(\pi_k, \delta') \geq r_{\pi_k}.$$

Therefore, we have shown that δ is minimax, since it has the lowest possible value of $\sup_\theta R(\theta, \delta)$. For the other part, note that for any prior π,

$$r_\pi = \mathbf{E}_\pi \left[R(\theta, \delta^\pi)\right] \leq \mathbf{E}_\pi \left[R(\theta, \delta)\right] \leq \sup_\theta R(\theta, \delta) = \lim_{k\to\infty} r_{\pi_k}.$$

□

Corollary 23.2.1. *If $\mathbf{Y} \sim \mathcal{N}(\mu, \sigma^2 I_k)$ with σ^2 known, then $\hat{\mu} = \mathbf{Y}$ is a minimax estimator for the mean with respect to the sum of squared errors.*

Proof. Consider a sequence of priors $\pi_{\tau^2} = \mathcal{N}(0, \tau^2 I_k)$. As $\tau^2 \to \infty$, this converges to the Jeffreys prior on the entire space, and the Bayes risk approaches $k\sigma^2$, which is the same as the constant risk of the UMVUE decision rule $\hat{\mu} = \mathbf{Y}$. Therefore, the sequence π_{τ^2} as $\tau^2 \to \infty$ is a least favorable sequence of priors, and $\hat{\mu} = \mathbf{Y}$ is a minimax estimator. □

23.2 Stein's Paradox

Although we just showed that the sample mean is minimax with respect to squared error, we did not prove admissibility, which is a different criterion. Note that the sample mean is trivially admissible in one dimension. We might also expect for it to be admissible in k dimensions, but surprisingly, this is actually not the case for $k \geq 3$!

Example 23.3 (Stein's paradox)**.** Consider a vector of independent random variables $Y_1, \ldots, Y_k \sim \mathcal{N}(\mu_i, 1)$, where our parameter vector is $\mu = (\mu_1, \ldots, \mu_k)$. The maximum likelihood estimator for μ is the sample mean

$$\hat{\mu}^{\text{MLE}} = \mathbf{Y}.$$

Suppose that our loss function is the mean-squared error, so our risk is

$$R(\mu, \hat{\mu}) = \mathbf{E}_\mu \left[\sum_{i=1}^{k} (\mu_i - \hat{\mu}_i)^2 \right].$$

For the sample mean, we have $R(\mu, \hat{\mu}^{\text{MLE}}) = k$. The paradox is that the following estimator, known as *Stein's estimator*, dominates the sample mean for $k \geq 3$, with risk $R(\mu, \hat{\mu}^{\text{JS}}) \leq k$:

$$\hat{\mu}^{\text{JS}} = \left(1 - \frac{k-2}{\|Y\|^2}\right) \mathbf{Y}.$$

Let's see why this paradox occurs. First, why does Theorem 22.2 (Blyth's method) fail for $k \geq 2$, when it worked to prove admissibility of the sample mean when $k = 1$? The issue lies in the last step, when we argued that $\epsilon \pi_{\tau^2}(A) \to 0$ at a rate of $1/\tau$, which is slower than the rate-$1/\tau^2$ convergence of the limits of Bayes risk. (Note that although Blyth's method fails when $k \geq 2$, there are other methods that can prove admissibility for $k = 2$ specifically.)

To prove Stein's theorem, which is that Stein's estimator has risk $\leq k$, we will prove two useful intermediate lemmas, known simply as *Stein's identity* and *Stein's lemma*. The former is a formula for the mean-squared error, and the latter is a step in deriving this formula.

Lemma 23.4 (Stein's identity)**.** *Given* $\mathbf{Y} \sim \mathcal{N}(\mu, \sigma^2 I_k)$*, let*

$$\hat{\mu}(\mathbf{Y}) = \mathbf{Y} + g(\mathbf{Y}),$$

for any function $g : \mathbb{R}^k \to \mathbb{R}^k$ *that is differentiable and satisfies the condition*

$$\mathbf{E}_\mu \left[\sum_{i=1}^{k} |\nabla_i g_i(\mathbf{Y})| \right] < \infty.$$

In other words, for any μ*, the expectation of each diagonal entry of the Jacobian is finite. Then, the expectation of the mean-squared error of* $\hat{\mu}$ *is*

$$\mathbf{E}_\mu \left[\|\mu - \hat{\mu}\|^2 \right] = k\sigma^2 + \mathbf{E}_\mu \left[\|g(\mathbf{Y})\|^2 + 2\sigma^2 \sum_{i=1}^{k} \nabla_i g_i(\mathbf{Y}) \right].$$

Proof. We can expand the left-hand side using linearity of expectation to get

$$\begin{aligned}
\mathbf{E}_\mu \left[\|\hat{\mu} - \mu\|^2 \right] &= \mathbf{E}_\mu \left[\|\mathbf{Y} + g(\mathbf{Y}) - \mu\|^2 \right] \\
&= \mathbf{E}_\mu \left[\|\mathbf{Y} - \mu\|^2 \right] + 2\,\mathbf{E}_\mu \left[(\mathbf{Y} - \mu)^\top g(\mathbf{Y}) \right] + \mathbf{E}_\mu \left[\|g(\mathbf{Y})\|^2 \right] \\
&= k\sigma^2 + \mathbf{E}_\mu \left[\|g(\mathbf{Y})\|^2 \right] + 2\sigma\,\mathbf{E}_\mu \left[\left(\frac{\mathbf{Y} - \mu}{\sigma} \right)^\top g(\mathbf{Y}) \right].
\end{aligned}$$

The result follows immediately from application of the following lemma. □

The following unbiased estimator for the risk is a direct corollary of Stein's identity.

Corollary 23.4.1 (Stein's unbiased risk estimate (SURE))**.** *An unbiased estimate of the squared-error risk of $\hat{\mu}(\mathbf{Y}) = \mathbf{Y} + g(\mathbf{Y})$ is*

$$\mathrm{SURE}(\hat{\mu}) = k\sigma^2 + \|g(\mathbf{Y})\|^2 + 2\sigma^2 \sum_{i=1}^{k} \nabla_i g_i(\mathbf{Y}).$$

Let's now prove Stein's lemma, which will finish the argument for both results.

Lemma 23.5 (Stein's lemma)**.** *Using the notation of the previous lemma, for any $i = 1, \ldots, k$,*

$$\mathbf{E}_\mu \left[\left(\frac{\mathbf{Y}_i - \mu_i}{\sigma} \right) g_i(\mathbf{Y}) \right] = \sigma \, \mathbf{E}_\mu \left[\nabla_i g_i(\mathbf{Y}) \right].$$

Proof. This is a result of integration by parts. Let ϕ be the density function of the multivariate standard normal distribution $\mathcal{N}(0, I_k)$, so we have

$$\begin{aligned}
\mathbf{E}_\mu \left[\left(\frac{\mathbf{Y}_i - \mu_i}{\sigma} \right) g_i(\mathbf{Y}) \right] &= \int_{\mathbb{R}^k} \left(\frac{y_i - \mu_i}{\sigma} \right) g_i(\mathbf{y}) \frac{1}{\sigma^k} \phi\left(\frac{\mathbf{y} - \mu}{\sigma} \right) \mathrm{d}\mathbf{y} \\
&= -\sigma \int_{\mathbb{R}^k} g_i(\mathbf{y}) \frac{1}{\sigma^k} \nabla_i \phi\left(\frac{\mathbf{y} - \mathbf{u}}{\sigma} \right) \mathrm{d}\mathbf{y} \\
&= \sigma \int_{\mathbb{R}^k} \nabla_i g_i(\mathbf{y}) \frac{1}{\sigma^k} \phi\left(\frac{\mathbf{y} - \mathbf{u}}{\sigma} \right) \mathrm{d}\mathbf{y} \\
&= \sigma \, \mathbf{E}_\mu \left[\nabla_i g_i(\mathbf{Y}) \right].
\end{aligned}$$

□

Finally, we use the former results to prove Stein's theorem.

Theorem 23.6 (Stein's theorem)**.** *Stein's estimator has risk $R(\mu, \hat{\mu}^{\mathrm{JS}}) \leq k$.*

Proof. This follows from applying Stein's identity to the function

$$g(\mathbf{y}) = -\frac{k-2}{\|\mathbf{y}\|^2} \mathbf{y}.$$

After plugging this into Stein's identity and doing some algebra (taking $\sigma^2 = 1$), we get

$$\mathbf{E}_\mu \left[\|\hat{\mu}^{\mathrm{JS}} - \mu\|^2 \right] = k - \mathbf{E}_\mu \left[\frac{(k-2)^2}{\|\mathbf{Y}\|^2} \right] < k.$$

□

24 December 1st, 2021

Today is the last lecture. We go over some details in proof of Stein's theorem that we glossed over last time, discuss the properties of Stein's estimator, and finally conclude the course.

24.1 Proof of Stein's Theorem

An astute reader might have noticed that in the previous proof of Stein's theorem, we seem to be missing something, since it appears as if $\hat{\mu}^{\mathrm{JS}}$ even dominates $\mathbf{Y}$ for $k = 1$. The catch here is that when we applied integration by parts in Stein's lemma, it requires the function g to be differentiable, so it only holds for $k = 2$.

Luckily, we can slightly relax the conditions to only require that for all j, $g(\mathbf{y})$ is differentiable in its j-th argument and almost surely in the remaining arguments. This criterion also holds for $k \geq 3$, but not when $k = 1$, which explains the difference in behavior for various dimensions.

Exercise 24.1 (Pencil problem)**.** The risk bound in Theorem 23.6 is not in closed form, as it involves an expectation, which make it difficult to use for analysis. Can we find a simpler upper bound for this risk function, still lower than k, but which depends on the value of μ?

Proof. The trick is to use Jensen's inequality to bound the value of the reciprocal of the squared magnitude, which is

$$\mathbf{E}_\mu\left[\frac{1}{\|\mathbf{Y}\|^2}\right] \geq \frac{1}{\mathbf{E}_\mu\left[\|\mathbf{Y}\|^2\right]} = \frac{1}{\mathbf{E}_\mu\left[\chi_k^2\right] + \|\mu\|^2} = \frac{1}{k + \|\mu\|^2}.$$

Therefore, the risk can be upper-bounded by

$$R(\mu, \hat{\mu}^{\mathrm{JS}}) \leq k - \mathbf{E}_\mu\left[\frac{(k-2)^2}{\|\mathbf{Y}\|^2}\right] \leq k - \frac{(k-2)^2}{k + \|\mu\|^2}.$$

This bound is best when $\mu = 0$ and worse when $\|\mu\|$ is large. Intuitively, the reason is that the shrinkage method always reduces the variance of the estimator by bringing points closer together, but it works best near the origin, where all values are being moved toward the true mean. □

24.2 Properties of the James Stein Estimator

Note that Stein's estimator $\hat{\mu}^{\mathrm{JS}}$ turns out to be inadmissible, since we can adjust it slightly to produce a dominating estimator

$$\hat{\mu}^{\mathrm{JS},+} = \left(1 - \frac{k-2}{\|\mathbf{Y}\|^2}\right)_+ \mathbf{Y},$$

where we use the notation $(x)_+ = \max(x, 0)$. Now, we show several examples that illustrate that Stein's phenomenon occurs in many statistical inference situations in practice, not just the basic case of minimizing mean-squared error on an isotropic standard normal.

Example 24.1. We can generalize Stein's estimator for other multivariate normal distributions $\mathbf{Y} \sim \mathcal{N}(\mu, \Sigma)$, by adjusting the ℓ_2 norm to the quadratic form

$$\hat{\mu}^{\mathrm{JS},\Sigma_i} = \left(1 - \frac{\tilde{k} - 2}{\mathbf{Y}^\top \Sigma \mathbf{Y}}\right)\mathbf{Y},$$

where we define $\tilde{k} = \frac{\mathrm{tr}(\Sigma)}{\lambda_{\max}(\Sigma)}$. This estimator also dominates the simple maximum likelihood estimator $\overline{Y}$ with respect to mean-squared error, for all $\tilde{k} \geq 3$.

Example 24.2. Even when the variance σ^2 of the standard normal distribution $\mathcal{N}(0, \sigma^2 I_k)$ is unknown, if we can approximate it by some unbiased sample variance $s^2 \sim \sigma^2 \chi^2_\nu / \nu$, then the estimator

$$\hat{\mu}^{\mathrm{JS},s^2} = \left(1 - \frac{(k-2)\frac{\nu}{\nu+2}s^2}{\|\mathbf{Y}\|^2}\right)\mathbf{Y}.$$

dominates the maximum likelihood estimator $\mathbf{Y}$ for $k \geq 3$.

Furthermore, Stein's phenomenon holds in generality even for data distributions that are not multivariate normal, as well as loss functions that are not as heavy-tailed as the mean-squared error. For example, it holds for $\mathbf{Y}$ in dimension $k \geq 3$ with respect to

$$L(\mu, \hat{\mu}) = \log\bigl(1 + \|\hat{\mu} - \mu\|^2\bigr).$$

Therefore, as a somewhat sobering point, James Stein showed that multi-dimensional decision analysis requires nontrivial adjustments to minimize risk.

Note. The philosophical message of Stein's phenomenon is that *shrinkage* is sometimes desirable, even from a purely frequentist standpoint. If you care about the loss of your estimator, you should generally be biased towards shrinking your estimator towards zero. An application of this is in machine learning, where ridge regression is a form of shrinkage that generally improves error.

24.3 Modern Statistical Research

That concludes the programmed material for this course. Lucas Janson concludes the class by summarizing the topics we covered so far and providing an overview of modern statistical research.

The goal of learning the material in a class like this one is to understand the underlying motivation behind standard methods in different scientific disciplines. For example, Lucas runs a free statistical consulting service at Harvard, where most clients come with scientific data that slightly violates some assumptions of standard methods in their discipline. By understanding the math behind course topics like sufficient statistics and frequentist inference, we can adjust methods to fit the specific needs of research.

On the research side, Lucas works in high-dimensional statistics, where current problems are typically focused on larger domains to handle constantly-growing datasets. Researchers think about many problems, including issues related to:

- **More data:** With bigger n, your methods become harder to execute in practice. For example, computing the covariance matrix is an operation that scales linearly in the number of data points and quadratically in the number of features.

- **More dimensions:** With bigger p, you end up with an exponential amount of parameters in most standard statistical models, such as polynomial regression. Even computing nearest neighbors in p dimensions is very expensive. This is known as the *curse of dimensionality.*

- **More powerful computers:** Given the increasing amount of highly available, large-scale compute resources, experiments have been getting much larger, which necessitates the development of statistical approaches that handle these datasets.

- **Machine learning:** Neural networks are not well-understood in the current statistical learning literature, but a key empirical property is that they perform shrinkage to obtain models of high-dimensional data distributions. Otherwise, their effectiveness would not make sense, given their lack of adversarial robustness. It's not understood how this shrinkage happens, but it occurs implicitly in the training methods (SGD) and architecture (Dropout).

- **Black boxes:** Many statistical learning methods of today are black-box algorithmic approaches, which perform well empirically but are not well-understood. The effectiveness of models like CNNs therefore illustrates something about the distributions that they are trained on, like image classification. In some sense, these are implicit assumptions, but not knowing the distribution formally makes it hard to apply many of the methods in this class.

Finally, we will briefly introduce an approach that handles the analysis needed for these kinds of black-box models.

Example 24.3. In *conformal inference*, we have an arbitrary function $f(\mathbf{X}_i; \mathcal{D}_\mathbf{y})$, and we develop a variant of the permutation test given this estimator. Using ideas directly from non-parametric inference and hypothesis testing, we can construct a prediction interval for $\mathbf{Y}_{\text{new}} \mid \mathbf{X}_{\text{new}}$, while leveraging the benefits of a black-box machine learning algorithm.

That concludes our statistical inference course for the semester! For the undergraduates in this class, Lucas reminds us to consider the Concurrent Masters program and suggests that we think about writing an honors thesis in Statistics.

References

[CB21] George Casella and Roger L Berger. *Statistical inference.* Cengage Learning, 2021.

[DY79] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281, 1979.

[LC06] Erich L Lehmann and George Casella. *Theory of point estimation.* Springer Science & Business Media, 2006.

[LR06] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses.* Springer Science & Business Media, 2006.