STAT4106 Course Notes

John W Smith Jr

Lecture 1: Order Statistics

Motivation

As career statisticians, we should always be interested in the best case and worst case scenarios.

Imagine that you working at an insurance company and you are in charge of determining whether or not someone should be given a policy. It is important to be able to examine the worst case scenario of this client, and use that to determine if they will receive an insurance policy.

This is one of many situations where it is important to be able to characterize particular behaviour, and this is not possible without order statistics. Order statistics allow us to compute distributions of "k^{th}" case scenarios.

Order Statistics

Let $Y_1, Y_2, ..., Y_n$ be independent continuous random variables with CDF (cumulative distribution function) denoted $F(y)$ and corresponding PDF (probability density function) $f(y)$, i.e. $Y_1, Y_2, ..., Y_n \sim$ i.i.d. $f(y)$. We denote the **ordered** variables by $Y_{(1)}, Y_{(2)}, ..., Y_{(N)},$, where $Y_{(1)} < Y_{(2)} ... < Y_{(N)}$. These ordered variables are referred to as the **order statistics**.

Some common statistics are based on the order statistics. These include the **sample range**, $Y_{(N)} - Y_{(1)}$, the **sample midrange**, $\frac{Y_{(N)}+Y_{(1)}}{2}$, and the **sample median**. The sample median is defined as $Y_{(\frac{N+1}{2})}$ if n is odd and $\frac{1}{2}(Y_{(\frac{N+1}{2})} + Y_{(\frac{N}{2})})$ if n is even. Much of our focus in this course will be on the first order statistic, or the minmum, $Y_{(1)}$ and the max order statistic, or maximum, $Y_{(N)}$.

Let us first consider $Y_{(N)}$.

$$F_{(Y_{(N)})}(y) = P(Y_{(N)} \leq y) = P(Y_1 \leq y, Y_2 \leq y, ...Y_N \leq y) = \text{(by i.i.d.)}$$

$$P(Y_1 \leq y)P(Y_2 \leq y)...P(Y_n \leq y) = F(y)^N$$

Then, to find the pdf of $Y_{(N)}$, we take the derivative of the cdf using the chain rule:

$$g_{(N)}(y) = \frac{d}{dy}F(y)^N = N \cdot (F(y)^{N-1}) \cdot \frac{d}{dy}F(y) = N \cdot (F(y)^{N-1}) \cdot f(y),$$

with the last step coming from the fact that the pdf can be defined as the first derivative of the cdf.

Now let us consider $Y_{(1)}$.

$$F_{Y_{(1)}}(y) = P(Y_{(1)} \leq y) = 1 - P(Y_{(1)} > y) =$$

$$1 - P(Y_1 > y, Y_2 > y, ..., Y_N > y) = \text{(by i.i.d.) } 1 - P(Y_1 > y)...P(Y_N > y)$$

$$= 1 - (1 - F(y))^N$$

Then, to find the pdf of $Y_{(1)}$, we once again take the derivative of the cdf using the chain rule:

$$g_{(N)}(y) = -N \cdot (1 - F(y))^{N-1} \cdot \frac{d}{dy}(-F(y)) = N \cdot (1 - F(y))^{N-1} \cdot f(y),$$

with the last step once again coming from the fact that the pdf can be defined as the first derivative of the cdf.

Now that we have shown the form of the first and last order statistics, we will show how these formulas can be used in order to find the distribution of these statistics for some simple cases.

First and Last Order Statistics of the Uniform Distribution

Consider $Y_1, ..., Y_n \sim Unif(0, 1)$ and are i.i.d. Find $g_{(1)}(y)$.

Solution

We have that $Y_1, Y_2, ...Y_N$ are uniform, thus $f(y) = 1, 0 \leq y \leq 1$. Then, we need $F(y)$. By definition, $F(y) = \int_{-\infty}^{y} f(y)dy$. Then,

$$F(y) = \int_{-\infty}^{y} 1 \; dy = y\Big|_0^y = y$$

Then, using the formula we derived above, we see that $g_{(1)}(y) = N \cdot (1 - y)^{N-1} \cdot 1 = N(1 - y)^{N-1}, 0 \leq y \leq 1$.

Now let us consider the same problem, but instead we will find $g_{(N)}(y)$.

Solution

We have done most of the heavy lifting in answering the question above - namely finding $F(y)$ and $f(y)$. Then, we can use the result derived earlier to obtain the pdf of the last order statistic.

$$g_{(N)}(y) = N \cdot y^{N-1} \cdot 1 = Ny^{N-1}, 0 \leq y \leq 1$$

It is worth noting that there is something special about the distribution of these order statistics - they both follow a **beta** distribution. The beta distribution can be parameterized as

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 \leq x \leq 1$$

Thus we have that $g_{(1)} \sim Beta(1, N)$, and $g_{(N)} \sim Beta(N, 1)$. The uniform distribution itself is also a beta distribution, with $\alpha = \beta = 1$. This relationship will be explored further in a homework problem.

Now that we have shown the distribution of the first and last order statistics, we will cover the general case.

In general, the joint density of $Y_{(1)}, Y_{(2)}, ..., Y_{(N)}$ is

$$g_{(1)(2)...(N)}(y_1, ..., y_N) = \begin{cases} n! f(y_1) f(y_2) ... f(y_N), & \text{if } y_1 < y_2 ... < y_n \\ 0, & \text{otherwise} \end{cases}$$

Theorem 6.5 Let $Y_{(k)}$ represent the k^{th} order statistic. Then, the pdf of $Y_{(k)}$ is

$$g_{(k)}(y_k) = \frac{n!}{(k-1)!(n-k)!} \Big(F(y_k)\Big)^{k-1} \Big(1 - F(y_k)\Big)^{n-k} f(y_k), -\infty < y_k < \infty$$

The proof is quite involved, and is given in Casella and Berger, Page 229, Theorem 5.4.4. For now, we will attempt to use this formula to compute some order statistics from another familiar distribution.

Exponential Order Statistics

Consider $Y_1, ..., Y_N \sim \exp(\lambda)$ and are i.i.d. Find $g_{(k)}(y)$.

Solution

We have that $Y_1, Y_2, ..., Y_N$ are exponentially distributed, thus $f(y) = \lambda \exp(-\lambda y)$, $0 \leq y < \infty$. Then, we must find $F(y)$, and we know $F(y) = \int_{-\infty}^{y} f(y) dy$

$$F(y) = \int_0^y \lambda \exp(-\lambda y) dy = -\exp(-\lambda y)\Big|_0^y = -\exp(-\lambda y) - (-1) = 1 - \exp(-\lambda y)$$

Now, equipped with both the pdf and cdf, we use the result of Theorem 6.5 to obtain the exponential order stat pdf.

$$g_{(k)}(y_k) = \frac{n!}{(k-1)!(n-k)!}\Big(1-\exp(-\lambda y)\Big)^{k-1}\Big(\exp(-\lambda y)\Big)^{n-k}\lambda\exp(-\lambda y) \quad \square$$

Practice Problems

These problems will be covered in class if there is time, and otherwise are expected to be completed as practice. Written solutions to these problems will not be provided on Canvas, but I am happy to discuss them during office hours and review sessions.

Parabolic Order Statistics

Suppose we have a random variable x with probability density function $f(x) = \frac{x^2}{9}$, $x \in [0, 3]$, and we observe $X_1, X_2, ..., X_N$

Compute the n^{th} order statistic and its mean and variance.

Compute the general formula for the k^{th} order statistic.

Order statistics for a certain class of scale beta distributions

Suppose that we have a random variable x with probability density function given by:

$$f_x(x) = \frac{a}{\theta^a}x^{a-1}, 0 < x < \theta,$$

with realizations $X_1, X_2, ..., X_N$. This is a **scale beta distribution** with parameters θ, $\alpha = a$ and $\beta = 1$.

Verify that x is a random variable.

Compute the general formula for the k^{th} order statistic and use this to find the distribution of the first and last order statistic.

Lecture 2: The Moment Generating Function

Lets start by recalling the definition of a moment generating function (or M.G.F.). Let X be a random variable, and let $\mathbb{E}[\cdot]$ denote the expectation operator. The **moment generating function**, $M_X(t)$ is defined as

$$M_X(t) = \mathbb{E}_X[\exp(tX)], t \in \mathbb{R}$$

The M.G.F. has some useful properties, including $\mathbb{E}[X^n] = \frac{d^n}{dt^n} M_x(t)\Big|_{t=0}$. This means that in the case that we are having trouble getting tractable integrals for the moments, if we can find the M.G.F. we are still able to obtain the moments. Another important result regarding moment generating functions is given and proved below.

Lemma 1: Let X be a random variable with M.G.F $M_x(t)$, and $a, b \in \mathbb{R}$. Then, $M_{aX+b}(t) = \exp(bt) \cdot M_X(at)$.

Proof: By definition, $M_{aX+b}(t) = \mathbb{E}_X[\exp((aX+b)t)]$. Then:

$$\mathbb{E}_X[\exp((aX+b)t)] = \mathbb{E}_X[\exp(aXt+bt)] = \mathbb{E}_X[\exp(aXt)\exp(bt)]$$

$$= \exp(bt)\mathbb{E}_X[\exp((at)X)] = \exp(bt)M_X(at) \quad \square$$

This lemma will come in handy when we are looking at distributions of functions of random variables, and more importantly it has given us a little bit of insight on how to work with proofs involving the M.G.F. We will now cover a theorem that will prove to be very important for determining sampling distributions.

Theorem 6.2: Let $Y_1, Y_2, ..., Y_N$ be independent random variables with moment generating functions $M_{Y_1}(t), M_{Y_2}(t), ..., M_{Y_N}(t)$. Let $U = Y_1 + Y_2 + ... + Y_N$. Then, $M_U(t) = \prod_{i=1}^N M_{Y_i}(t)$.

Proof:

$$M_U(t) = \mathbb{E}_U[\exp(tU)] = \mathbb{E}_U[\exp(t(Y_1 + ... + Y_N))] = \mathbb{E}[\exp(tY_1) \cdot ... \cdot \exp(tY_N)]$$

$$= \mathbb{E}[\exp(tY_1)] \cdot ... \cdot \mathbb{E}[\exp(tY_N)] = M_{Y_1}(t) \cdot ... \cdot M_{Y_N}(t) = \prod_{i=1}^N M_{Y_i}(t) \quad \square$$

A Useful Application: The Distribution of Sum of R.V.s

Suppose $Y_1, ..., Y_N \sim$ i.i.d. $Poisson(\lambda)$ Let $U = \sum_{i=1}^N Y_i$. What is the distribution of U?

Solution: We know that the Poisson has M.G.F. $M_{Y_i}(t) = \exp(\lambda(e^t - 1))$. Then, applying Theorem 6.2 we have that:

$$M_U(t) = \prod_{i=1}^N M_{Y_i}(t) = \prod_{i=1}^N \exp(\lambda(e^t - 1)) = \exp(n\lambda(e^t - 1))$$

This is the M.G.F. of a $Poisson(n\lambda)$. Thus we have that $U \sim Poisson(n\lambda)$.

Suppose now that $Y_1, ..., Y_N$ are independent, but $Y_i \sim Poisson(\lambda_i)$. Let $U = \sum_{i=1}^N Y_i$.

What is the distribution of U?

Solution: We know that each Y_i has corresponding M.G.F. $M_{Y_i}(t) = \exp(\lambda_i(e^t - 1))$. Then, applying Theorem 6.2 again we have that:

$$M_U(t) = \prod_{i=1}^{N} M_{Y_i}(t) = \prod_{i=1}^{N} \exp(\lambda_i(e^t - 1)) = \exp(\sum_{i=1}^{N} \lambda_i(e^t - 1)) = \exp((e^t - 1)\sum_{i=1}^{N} \lambda_i)$$

This is the M.G.F. for a $Poisson(\sum_{i=1}^{N} \lambda_i)$, and thus we have that $U \sim Poisson(\sum_{i=1}^{N} \lambda_i)$.

Suppose that $Y_1, Y_2, ..., Y_N \sim$ i.i.d. $exponential(\beta)$. Let $U = \sum_{i=1}^{N} Y_i$. What is the distribution of U?

Solution: The M.G.F. of the exponential distribution is $M_Y(t) = (1 - \beta t)^{-1}$. Then, by Theorem 6.2:

$$M_U(t) = \prod_{i=1}^{N} M_{Y_i}(t) = \prod_{i=1}^{N} (1 - \beta t)^{-1} = (1 - \beta t)^{-N}$$

This is the M.G.F. for a $Gamma(N, \beta)$, and thus $U \sim Gamma(N, \beta)$.

Deriving Properties of the Normal Distribution with the MGF

Recall some the elementary properties of the Normal distribution that were taught to you in your introductory statistics courses. For example:

1. If $X \sim N(\mu, \sigma^2)$, $aX \sim N(a\mu, a^2\sigma^2)$
2. If $X \sim N(\mu, \sigma^2)$, $aX + b \sim N(a\mu + b, a^2\sigma^2)$
3. If $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim N(\mu_2, \sigma_2^2)$, and X and Y are independent, then for any $a, b \in \mathbb{R}$, $aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$

We can prove these elementary results using Lemma 1 and Theorem 6.2.

Proof of 1: The MGF of a normal random variable is $M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$. By Lemma 1 we have

$$M_{aX}(t) = M_X(at) = \exp\left(\mu at + \frac{\sigma^2(at)^2}{2}\right) = \exp\left((\mu a)t + \frac{(a^2\sigma^2)t^2}{2}\right)$$

Thus the MGF of aX is a Normal MGF with $\mu^* = a\mu, \sigma^{2*} = a^2\sigma^2$. Thus $aX \sim N(a\mu, a^2\sigma^2)$.

Proof of 2: We can directly apply Lemma 1 to the normal MGF to obtain:

$$M_{aX+b}(t) = \exp(bt) \cdot M_X(at) = \exp(bt) \cdot \exp\left(a\mu t + \frac{a^2\sigma^2 t^2}{2}\right) = \exp\left((a\mu + b)t + \frac{a^2\sigma^2 t^2}{2}\right)$$

This is a Normal MGF with $\mu^* = a\mu + b, \sigma^{2*} = a^2\sigma^2$. Thus $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

Proof of 3: Using a combination of Lemma 1 and Theorem 6.2 we have (by independence of X and Y).

$$M_{aX+bY}(t) = M_X(at) \cdot M_Y(bt) = \exp\left(a\mu_1 t + \frac{a^2\sigma_1^2 t^2}{2}\right) \cdot \exp\left(b\mu_2 t + \frac{b^2\sigma_2^2 t^2}{2}\right)$$

$$= \exp\left((a\mu_1 + b\mu_2)t + \frac{(a^2\sigma_1^2 + b^2\sigma_2^2)t^2}{2}\right)$$

Thus $aX + bY \sim N(a\mu_1 + a\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

Practice Problems

1. Suppose $Y_1, ..., Y_N$ are independent, with $Y_i \sim Exponential(\beta_i)$. Let $U = \sum_{i=1}^N Y_i$. What is the distribution of U?
2. Suppose $Y_1, ..., Y_N$ are independent, with $Y_i \sim Normal(\mu_i, \sigma_i^2)$. Let $U = \sum_{i=1}^N Y_i$. What is the distribution of U?
3. Let $Y \sim Bern(p)$. Find $M_Y(t)$.
4. Let $Y_1, ..., Y_N \sim$ i.i.d. $Bern(p)$. Let $U = \sum_{i=1}^N Y_i$. What is the distribution of U?
5. Suppose Y_1, Y_2, Y_3 are independent random variables that represent the number of sharks that are tagged by 3 different ecological conservation groups in a particular week. Let $Y_1 \sim Poisson(.5)$, $Y_2 \sim Poisson(1)$, $Y_3 \sim Poisson(2)$. What is the probability that there are no sharks tagged in a given week?
6. Let $X_1, X_2, ..., X_N$ be i.i.d. random variables for the failure time for a particular brand of lightbulbs, with $X_i \sim Exp(1)$. What are the mean and variance for a single observation of the failure time? Let $\bar{X} = \sum_{i=1}^N X_i/n$ denote the average failure time for a particular batch of lightbulbs. What are the mean and variance for the average failure time? How does this differ from the single observation of a failure time?

One Last Note on MGFs

The moment generating function is a special case of something called a **Laplace transform**. Those who are interested in reading more on properties of the m.g.f. by virtue of being a special case of the Laplace transform can find a thorough treatment in William Feller's text "An introduction to probability theory and its applications".

Lecture 3: Sampling Distributions and the CLT

We assume that we have a random sample from a population. We write this mathematically as $Y_1, Y_2, ..., Y_N \sim$ i.i.d. $f(y)$.

We use numerical consequences of our sample, i.e. **statistics**, to obtain information about the numerical characteristics of the population ($f(y)$), i.e. its **parameters**. Statistics are random variables and (at least in this course) we will treat parameters as unknown constants. This process is referred to as **statistical inference**.

To estimate and make decisions about parameters, it is important to know the probability distributions of the statistics. These are referred to as **sampling distributions**.

One of most important statistics, in practice and in theory, is the **sample mean**, $\bar{Y} = \sum_{i=1}^{N} Y_i/N$. We will focus on the sample mean first.

Let $Y_1, Y_2, ..., Y_N \sim$ i.i.d. $f(y)$, with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$. Then, let us find $\mathbb{E}[\bar{Y}]$ and $\mathbb{V}[\bar{Y}]$.

$$\mathbb{E}[\bar{Y}] = \frac{1}{N}(\mathbb{E}[Y_1] + \mathbb{E}[Y_2] + ... + \mathbb{E}[Y_N]) = \frac{N\mu}{N} = \mu$$

$$\mathbb{V}[\bar{Y}] = \frac{1}{N^2}(\mathbb{V}[Y_1] + \mathbb{V}[Y_2] + ... + \mathbb{V}[Y_N]) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

Thus we have shown that for any distribution that has a mean and a variance, that the sample mean will have the same mean and a variance scaled by the number of components being averaged.

7.2: Sampling Distributions Related to the Normal

Now let us assume that $Y_1, Y_2, ..., Y_N \sim$ i.i.d. $N(\mu, \sigma^2)$. Then, $\bar{Y} \sim N(\mu, \sigma^2)$. The proof will make use of the moment generating function, and this result is a special case of one of your homework problems, with $\mu_i = \mu, \sigma_i^2 = \sigma^2, a_i = \frac{1}{N}, \forall i \leq N$.

We will use the notation $\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{N}$, and $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{N}}$. The standard deviation of a statistic is often referred to as the **standard error**. Thus the standard error of $\bar{Y}$ is $\frac{\sigma}{\sqrt{N}}$.

Applying Lemma 1 from last lecture, we can obtain the distribution of an old friend,

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1)$$

Now, we will look at some distributions of functions of normal random variables.

Theorem 7.2 Let $Z_i = \frac{Y_i - \mu}{\sigma}$. Then, $Z_1, Z_2, ..., Z_N \sim$ i.i.d. $N(0,1)$. Furthermore, $Z_i^2 \sim \chi^2(1)$, and $\sum_{i=1}^N Z_i^2 \sim \chi^2(N)$.

Proof: The proof is omitted here, but is covered in Example 2.1.9 in Casella and Berger's Statistical Inference (pdf available online). It is a good exercise in transformations of random variables, which was covered in STAT4105.

The χ^2 distribution also comes into play when we want to estimate variances. Recall the **sample variance**,

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Theorem 7.3: If $Y_1, Y_2, ..., Y_N \sim$ i.i.d. $N(\mu, \sigma^2)$, then

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

In addition, $\bar{Y}$ and S^2 are independent random variables.

Proof: The proof is once again omitted here, but is covered in Casella and Berger Theorem 5.3.1.

If we are sampling from a normal distribution and σ^2 is known, then we can make inferences about μ through the fact that $Z = \sqrt{N}\frac{\bar{Y}-\mu}{\sigma} \sim N(0,1)$. When σ is unknown, we estimate it using the sample variance, i.e. $\hat{\sigma} = \sqrt{S^2}$. Substituting this estimate in for σ surely ruins the normality, but what is the resulting distribution? It turns out to be a familiar friend once again:

$$\sqrt{N}\frac{\bar{Y}-\mu}{S} \sim t(n-1)$$

Thus when the population variance is unknown, we will use the t distribution to make inferences about the true population mean μ.

Definition 7.2 Suppose $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$. Then, if Z and W are independent, we say that $T = \frac{Z}{\sqrt{W/\nu}}$ has a t distribution with ν degrees of freedom.

It turns out that the t distribution with ν degrees of freedom will have only its first $\nu - 1$ moments. In the case that $\nu > 2$, the t distribution has a finite mean and variance, which are $\mathbb{E}[T] = 0$ and $\mathbb{V}[T] = \frac{\nu}{\nu-2}$

Definition 7.3 Let W_1 and W_2 be two independent χ^2 random variables with ν_1 and ν_2 degrees of freedom respectively. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F_{\nu_1,\nu_2}$$

We say that F follows an F distribution with $\{\nu_1, \nu_2\}$ degrees of freedom. The F distribution arises quite naturally in ANOVA, design of experiments, and regression.

Corollary 7.3: Suppose $Y_1, Y_2, ..., Y_{N_1} \sim$ i.i.d. $N(\mu_1, \sigma_1^2)$ and $X_1, X_2, ..., X_{N_2} \sim$ i.i.d. $N(\mu_2, \sigma_2^2)$, and let the samples be independent. Then

$$\frac{S_1^2/\sigma_1^2}{S_1^2/\sigma_1^2} \sim F_{N_1-1, N_2-1}$$

Proof: From Theorem 7.3 we have that $S_1^2 = \frac{(N_1-1)S^2}{\sigma_1^2} \sim \chi^2(N_1 - 1)$ and similarly $S_2^2 = \frac{(N_2-1)S^2}{\sigma_2^2} \sim \chi^2(N_2-1)$. Then, dividing through by the degrees of freedom, N_1-1 and $N_2 - 1$ respectively, we get $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$. By Definition 7.3, since we have the ratio of two independent χ^2 distributions divided by their degrees of freedom, we have an F_{N_1-1,N_2-1}. $\square$.

We have built up quite a lot of machinery here, so let's take some time to make sure that we understand with some examples.

Example: Construction of Various Distributions

Suppose we have $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6 \sim N(0, 1)$, and they are independent. Let us construct the following:

- A χ^2 distribution with 3 degrees of freedom
- A t distribution with 5 degrees of freedom
- An F distribution with $\{2, 2\}$ degrees of freedom

Solution: We can construct a χ_3^2 distribution using Theorem 7.2. Since we have that Y_i are independent $N(0, 1)$ random variables, we know that $Y_1^2 + Y_2^2 + Y_3^2 \sim \chi_3^2$.

To build a t_5 distribution, we invoke the definition given above. We know that $\sqrt{N}\frac{\bar{Y}-\mu}{S} \sim t(n-1)$. In our case, $\mu = 0$ and $N = 6$. Thus $\sqrt{6}\frac{\bar{Y}}{S} \sim t_5$.

To build an $F_{2,2}$, we invoke Definition 7.3. We can build two χ_2^2 random variables through summing 2 independent squared standard normals. Thus $Y_1^2 + Y_2^2 \sim \chi_2^2$, $Y_3^2 + Y_4^2 \sim \chi_2^2$. Thus $\frac{(Y_1^2+Y_2^2)/2}{(Y_3^2+Y_4^2)/2} \sim F_{2,2}$.

Lecture 4: Central Limit Theorem (CLT)

The Central Limit Theorem is a result that we have been using since introductory statistics classes. It is a powerful theorem that allows us to make inferences about large population sizes using a common distribution that we are experienced working with - the Normal distribution. While we have taken it for granted thus far, it really is one of the most shocking and important results to grace statistics. We will not build the proper machinery to prove the general case of the CLT, but we are able to prove a special case, making use of the mgf.

Theorem 4.1: The (Weak) Central Limit Theorem Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with mgf $M_X(t)$, $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2 < \infty$. Let $U_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. Then, $U_n \sim N(0,1)$.

Proof: We start by making a transformation. Let $Y_i = \frac{X_i-\mu}{\sigma}$. Note that Y_i must also have a moment generating function, and by construction we have that $\mathbb{E}[Y_i] = 0$ and $\mathbb{V}[Y_i] = 1$. Note that we can now write $\bar{X} = \sum_{i=1}^n X_i/n = (\sum_{i=1}^N \sigma Y_i + \mu)/n$. Then,

$$U_n = \frac{\sqrt{n}}{\sigma}\Big(\bar{X} - \mu) = \frac{\sqrt{n}}{\sigma}\Big(\frac{1}{n}(\sum_{i=1}^N \sigma Y_i + \mu) - \mu) = \frac{\sqrt{n}}{\sigma}\Big(\sigma\frac{\sum_{i=1}^n Y_i}{n} + \frac{n\mu}{n} - \mu\Big)$$

$$= \frac{\sqrt{n}}{\sigma} \cdot \frac{\sigma \sum_{i=1}^n Y_i}{n} = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$$

Thus we have shown that $U_n = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$. Then, we can find the mgf of U_n using this relation. From properties of the mgf covered in Notes 2, we have

$$M_{U_n}(t) = M_{\sum_{i=1}^N Y_i/\sqrt{n}} = (M_Y\Big(\frac{t}{\sqrt{n}}\Big))^n$$

Though we don't have any information on $M_Y(t/\sqrt{n})$ (except existence), we can compute it through the Taylor expansion.

$$M_Y\Big(\frac{t}{\sqrt{n}}\Big) = \sum_{k=0}^{\infty} M_Y(0)\frac{(t/\sqrt{n})^k}{k!} = 1 + \mathbb{E}[Y]\frac{t}{\sqrt{n}} + \mathbb{V}[Y]\frac{t^2}{2n} + R_Y(t)$$

Since we constructed Y_i to have a mean of 0 and a variance of 1, we have

$$M_Y\Big(\frac{t}{\sqrt{n}}\Big) = 1 + \frac{t^2}{2n} + R_Y(t)$$

Then,

$$M_{U_n}(t) = \Big(1 + \frac{t^2}{2n} + R_Y(t)\Big)^n$$

A consequence of Taylor's theorem is that this remainder will go to 0 when exponentiated, and thus

$$M_U(t) = \left(1 + \frac{t^2}{2n}\right)^n = \exp\left(\frac{t^2}{2}\right)$$

This is the mgf of a standard normal distribution. Thus we have shown that U_n converges to a standard normal distribution. $\square$.

A Stronger Central Limit Theorem Let $X_1, X_2, ..., X_n \sim$ i.i.d. $f(x)$ be a random sample with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2 < \infty$. Let $U_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. Then, $U_n \sim N(0, 1)$.

We will not prove this stronger form of the CLT, but we will note some important differences between this one and the one that we just proved. For the weak version, we required a moment generating function. This can be a tall order, as many functions do not even always have all of their moments. With this weak central limit theorem, all that we require is a finite mean or variance to ensure convergence. We will demonstrate this with an example below.

The CLT in action

Suppose $Y_1, Y_2, ..., Y_n \sim$ i.i.d. *exponential*(β), with pdf $f(y) = \lambda \exp(-\lambda y)$. Find the distribution of $U_n = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$.

Solution: In order to apply the CLT, we must confirm that the first two moments of this distribution exist.

$$\mu = \mathbb{E}[Y_i] = \int_0^\infty y\lambda \exp(-\lambda y)dy =$$

Another CLT example

Suppose $X_1, X_2, ..., X_n \sim$ i.i.d. $f(x) = x^2 + x + \frac{1}{6}, 0 \leq x \leq 1$. What does the central limit theorem tell us about the distribution of $U_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$?

Solution: In order to apply the CLT, we must confirm that the first two moments of this distribution exist.

$$\mu = \mathbb{E}[X_i] = \int_0^1 x(x^2 + x + \frac{1}{6})dx = \frac{x^4}{4} + \frac{x^3}{3} + \frac{x^2}{12}\Big|_0^1 = \frac{2}{3}$$

Thus we have shown the first moment exists. Now we will find the variance, using the fact that $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$\mathbb{V}[X_i] = \int_0^1 x^2(x^2 + x + \frac{1}{6})dx = \frac{x^5}{5} + \frac{x^4}{4} + \frac{x^3}{18}\Big|_0^1 = \frac{91}{180}$$

Then, $V[X_i] = \frac{91}{180} - \left(\frac{2}{3}\right)^2$. Then, the variance and mean are both finite and thus we know that the limiting distribution of U_n is a standard normal by the central limit theorem.

A canonical counterexample

Let us consider the **Cauchy distribution**. The Cauchy is a heavy-tailed distribution that belongs to the family of t distributions - a t distribution with one degree of freedom. It has pdf $f(x) = \frac{1}{\pi(1+x^2)}, x \in \mathbb{R}$. It is the most common example of a distribution that "breaks" the clt. Let's see why.

Suppose $X_1, X_2, ..., X_n \sim$ i.i.d. Cauchy. Let us compute the first moment.

$$\mathbb{E}[X_i] = \int_{\mathbb{R}} \frac{x}{\pi(1+x^2)} dx$$

We can compute this with a u substitution, $u = 1 + x^2$.

$$\int_{\mathbb{R}} \frac{x}{\pi(1+x^2)} dx = \int_1^\infty \frac{1}{2\pi(1+u)} du = \frac{\log(u)}{2\pi}\Big|_1^\infty = \infty$$

Thus we have shown that the Cauchy does not even have a first moment! Though we will not cover it in this course, it is a consequence of a result called **Jensens' inequality** that if a distribution does not have a k^{th} moment, then it will not have any moments $m \geq k+1$. Since the Cauchy does not have even a first moment, we are able to conclude that it has no moments!

Lecture 5: Estimation

In general, it is very important to distinguish between parameters (μ, σ^2, α etc) and statistics ($\bar{y}, S^2$). A key aspect of **statistical inference** is to use statistics to estimate parameters.

If we use the value of $\bar{y}$, say for example $\bar{y} = 100.1$, to estimate μ, then this is an example of a **point estimate**. If instead we calculated an interval, say $(98.1, 102.1)$, which is intended to contain the true μ value, then we have an **interval estimate**. Our point estimate has a probability of 0 to match the true parameter value (why?), and thus interval estimates are preferred.

The function of our sample values that provides the estimate is referred to as the **estimator**. For example, we may want to use $\bar{y} = (Y_1 + ... + Y_n)/n$ as the estimator for μ. The numerical value of the estimator is called the **estimate**.

Bias and MSE of Point Estimators

Without loss of generality, let θ represent the parameter of interest, and let $\hat{\theta}$ represent an estimator of θ.

Definition: The **bias** $B(\hat{\theta})$ of an estimator $\hat{\theta}$ is $B(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.

Definition: If $B[\hat{\theta}] = 0$, we call $\hat{\theta}$ an unbiased estimator of θ.

If we had to choose between 2 unbiased estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, we would want to pick the one with a smaller variance. This means that with repeated sampling, the estimator with smaller variance will gives estimates closer to the unknown value. But what about if one or both of the estimators are biased? How do we pick which one is "better"?

Definition: The **mean squared error (MSE)** of a point estimator is the expected value of $(\hat{\theta} - \theta)^2$,

$$MSE(\hat{\theta} = \mathbb{E}[(\hat{\theta} - \theta)^2])$$

A little bit of manipulation shows that we may write the MSE in the alternate form

$$MSE(\hat{\theta}) = \mathbb{V}[\hat{\theta}] + B(\hat{\theta})^2$$

Being "unbiased" is a good property of a point estimator, but there can a numer of unbiased estimators for a given parameter. The minimum MSE estimator in some cases will be a biased estimator.

Comparison of two σ^2 estimators

Let us consider first the estimator S^2, where our population is normally distributed with mean μ and variance σ^2. We want to find the MSE of S^2. Recall the earlier result that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$$\mathbb{E}\Big[\frac{(n-1)S^2}{\sigma^2}\Big] = n - 1$$

$$\frac{n-1}{\sigma^2}\mathbb{E}[S^2] = n - 1$$

$$\mathbb{E}[S^2] = \frac{\sigma^2}{n-1}n - 1 = \sigma^2$$

Thus we have shown that S^2 is unbiased. Now, for the variance:

$$\mathbb{V}\Big[\frac{(n-1)S^2}{\sigma^2}\Big] = 2(n-1)$$

$$\mathbb{V}[S^2] = 2(n-1)\frac{\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}$$

Then, combining these we get:

$$MSE(S^2) = \frac{2\sigma^4}{n-1}$$

Now, let us consider a second estimator, $S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. It is worth noting that $S_*^2 = \frac{n-1}{n}S^2$, as this will allow us to re-use some calculations from above.

$$\mathbb{E}[S_*^2] = \frac{n-1}{n}\mathbb{E}[S^2] = \frac{n-1}{n}\sigma^2$$

$$B(S_*^2)^2 = \Big(\sigma^2 - \frac{\sigma^2(n-1)}{n}\Big)^2 = \frac{\sigma^4}{n^2}$$

Then, computing the variance of S_*^2:

$$\mathbb{V}[S_*^2] = \frac{n-1}{n}\mathbb{V}[S^2] = \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$$

$$MSE(S_*^2) = \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}$$

We see that the MSE for the biased estimator of σ^2 is actually lower for all values of n. This is an example of a case where a biased estimator has MSE lower than a common unbiased estimator everywhere. This is however, **not** the minimum MSE estimator of the form $\hat{\theta} = cS^2, c \in \mathbb{R}$. What is?

A uniform estimator

Suppose $X_1, X_2, ..., X_n \sim$ i.i.d. $Unif(0, \theta)$. Find the MSE of $\hat{\theta} = 2 \cdot \bar{X}$. Is $\hat{\theta}$ unbiased?

We start by finding $\mathbb{E}[\hat{\theta}]$. Recall that the $Unif(\theta_1, \theta_2)$ has expected value $\frac{1}{2}(\theta_1 + \theta_2)$

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[2 \cdot \bar{X}] = 2\mathbb{E}[\bar{X}] = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{2}{n}\Big(\frac{n\theta}{2}\Big) = \theta$$

Thus $\hat{\theta}$ is unbiased, and $B(\hat{\theta}) = 0$. Then, to compute the MSE we must find the variance of $\hat{\theta}$. Recall that the variance of a uniform random variable is $\frac{(\theta_2-\theta_1)^2}{12}$. Then,

$$\mathbb{V}[2 \cdot \bar{X}] = 4 \cdot \mathbb{V}[\bar{X}] = 4 \cdot \frac{\theta^2}{12n} = \frac{\theta^2}{3n}$$

Combining the squared bias and the variance of the estimator we get $MSE(\hat{\theta}) = \frac{\theta^2}{3n}$.

Practice Problem: Minimizing variance in estimators

(Problem taken from Wackerly et al 7th edition).

Suppose that $\mathbb{E}[\hat{\theta}_1] = \mathbb{E}[\hat{\theta}_2] = \theta$, with $\mathbb{V}[\hat{\theta}_1] = \sigma_1^2, \mathbb{V}[\hat{\theta}_2] = \sigma_2^2$.
Consider $\hat{\theta}_3 = a\hat{\theta}_1 + (1-a)\hat{\theta}_2$.

Is $\hat{\theta}_3$ unbiased?
If $\hat{\theta}_1$ is independent of $\hat{\theta}_2$, how do we choose a to minimize the variance of $\hat{\theta}_3$?

Lecture 6: Interval Estimation

Evaluating the Goodness of a Point Estimator

Definition: The **error of estimation**, ϵ, is the distance between an estimator and the parameter being estimated. It is defined as

$$\epsilon = |\hat{\theta} - \theta|$$

This is a random quantity, but we can evaluate

$$P(\epsilon < b) = P(|\hat{\theta} - \theta| < b) = P(-b < \hat{\theta} - \theta < b) = P(\theta - b < \hat{\theta} < \theta + b)$$

by using the sampling distribution of $\hat{\theta}$. Thus if we want to find a value of b such that $P(\epsilon < b) = .90$, then we must find b such that

$$\int_{\theta-b}^{\theta+b} f(\hat{\theta}) d\hat{\theta} = .90$$

It turns out that whether or not we know the sampling distribution of $\hat{\theta}$, **if $\hat{\theta}$ is unbiased**, then we can use, as an approximation,

$$\hat{\theta} \pm k\sigma_{\hat{\theta}}$$

to find interval estimates. Markov's inequality tells us that the probability that this interval contains θ is **at least** $1 - k^{-2}$. For $k = 2$, the probability is close to 95% for the normal distribution.

Confidence Intervals

We want intervals such that if a sample is taken, our probability of the interval containing the paramater is a specified value.

Two-sided case:

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha,$$

where $\hat{\theta}_L$ is called the lower confidence limit, $\hat{\theta}_U$ is called the upper confidence limit. This is a $(1-\alpha)$ confidence level. The CI is of course $(\hat{\theta}_L, \hat{\theta}_U)$.

One-sided cases:

$$P(\hat{\theta}_L \leq \theta) = 1 - \alpha,$$

the CI here is $(\hat{\theta}_L, \infty)$.

$$P(\theta \leq \hat{\theta}_U) = 1 - \alpha,$$

the CI here is $(-\infty, \hat{\theta}_U)$
One useful way to find confidence intervals is to use something called the **pivotal method**.

Definition: A **pivotal quantity** of an unknown parameter θ is: a function of the sample observations and θ, with θ being the only unknown quantity, and has a probability distribution that does not depend on θ.

We use the percentiles of a pivotal quantity to determine an interval and then simplify to get the interval in our desired form.

Example

Show that if $X_1, X_2, ..., X_n \sim$ i.i.d. $N(\mu, \sigma^2)$, with σ^2 known, then $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is a pivotal quantity for μ.

We see that Z is a both a function of μ and the sample observations. Furthermore, since σ is known, μ is the only unknown quantity. We have shown in an earlier lecture that for this case, $Z \sim N(0,1)$, therefore the probability distribution of Z is not a function of μ. Therefore Z is a pivotal quantity for μ.

Example: Finding σ^2 Intervals

Suppose that $X_1, ..., X_n \sim$ i.i.d. $N(\mu, \sigma^2)$. Recall that under this scenario, we have that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

We want to compute a $1-\alpha$ percent confidence interval for σ^2. In order to do this, we need to first show that $\frac{(n-1)S^2}{\sigma^2}$ is a pivotal quantity for σ^2.

This is relatively straightforward. We see that this quantity is a function of the sample observations (though it is masked by the shorthand expression S^2) and that the only unknown is σ^2, since n is given. Thus $\frac{(n-1)S^2}{\sigma^2}$ is a pivotal quantity for σ^2. We will now use this pivotal quantity to compute a confidence interval for σ^2.

We want:

$$P\big[\chi^2_L \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_U\big] = 1-\alpha,$$

where U, L stand for upper and lower, respectively. χ^2 functions are not symmetric, and thus we have a lot of freedom in picking these - there are an uncountably infinite number of points χ^2_L and χ^2_U that satisfy this equality. We will add the restriction that $P[\frac{(n-1)S^2}{\sigma^2} \geq \chi^2_U] = P[\frac{(n-1)S^2}{\sigma^2} \leq \chi^2_L] = \frac{\alpha}{2}$. An image representing this is shown below.

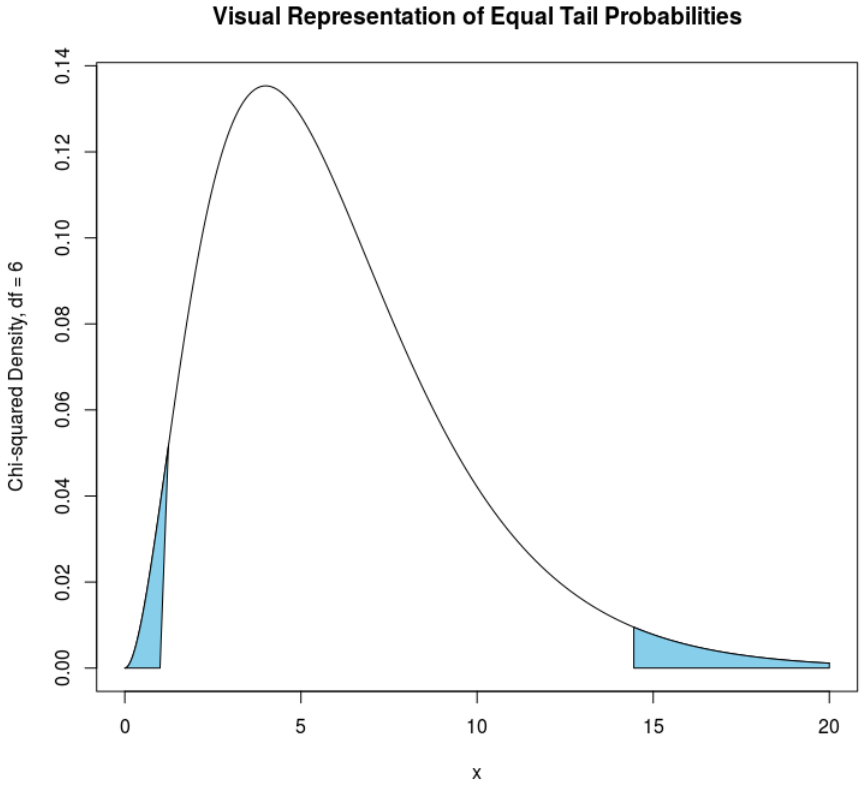

Figure 1: Visual representation of what it looks like to restrict the interval to have equal probability in the tails of the distribution

Then, we can now write out probability as $P\big[\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}\big] = 1-\alpha$. We can take the reciprocal of everything and flip the inequality to get:

$$P\left[\frac{1}{\chi^2_{1-\frac{\alpha}{2}}} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi^2_{\frac{\alpha}{2}}}\right] = 1-\alpha$$

Then, moving some terms around we can get an interval for σ^2.

$$P\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\right] = 1-\alpha$$

Practice: Confidence Intervals for β in a Gamma distribution

Suppose that $Y \sim Gamma(\alpha, \beta)$, with $\alpha = 2$ given. Show that $\frac{2Y}{\beta}$ is a pivotal quantity for β, and compute its distribution. Using this distribution, find a 90 percent confidence interval for β.

Lecture 7: Large Sample Confidence Intervals

Motivation

So far, we have learned about order statistics, sampling distributions, limiting distributions (the CLT), estimators (point and interval), how to assess estimators, and how to compute confidence intervals. This section will, simplistically, tie together a great deal of concepts that have been covered so far.

It is not always easy to find pivotal quantities and get exact confidence intervals on parameters. Pivotal quantities can be tricky to find and in some cases it may be too difficult to check the assumptions - for example when your data are assumed to come from a pdf that is very difficult to work with. With enough samples we still may be able to get approximate confidence intervals, using the Central Limit Theorem and treating our collection of samples as approximately normal. From here, we already have a known pivotal quantity, Z, to help us get approximate confidence intervals. We also have the practical tools to assess the performance of these estimators, and the theoretical tools to diagnose poor performance.

Large Sample Confidence Intervals

In our introduction to stat classes, we learned that we can can do hypothesis tests on parameters of non-normal distributions by using the central limit theorem and treating them as approximately normal. The table below lists a few of these that we may remember:

Target Parameter θ	Sample Size	$\hat{\theta}$	$\mathbb{E}[\hat{\theta}]$	$\sigma_{\hat{\theta}}$
μ	n	$\bar{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{p(1-p)}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\bar{Y}_1 - \bar{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^*$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Table 1: * requires the populations to be independent

In general, if we have the central limit theorem, we have

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \dot{\sim} N(0,1)$$

and Z is (approximately) a pivotal quantity for θ. Using the distribution of our pivotal quantity we can make two sided confidence intervals of the form

$$(\theta_L, \theta_U) = (\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}})$$

In the cases where $\sigma_{\hat{\theta}}$ is unknown, we will esimate it and have confidence intervals of the form

Interpretation of Confidence Intervals

The interpretation of a confidence interval is very important. In the business world, for example, poor confidence interval interpretations may mislead shareholders or lose clients money. As career statisticians, we should have an ironclad interpretation of a frequentist confidence interval (or, alternatively, be Bayesian!).

The correct interpretation is as follows: if one takes many samples of size n from a population and forms a $(1-\alpha)100\%$ confidence interval on θ for each sample, we expect that **on average** $(1-\alpha)100\%$ of these confidence intervals will contain the true value of θ.

We will now build a bit of intuition on confidence intervals.

Fact 1: As the sample size n increases, the confidence intervals will get more narrow.

Fact 2: As α gets larger, the width of the $1-\alpha$ confidence interval will get smaller.

Fact 3: As the variation in the population increases, the width of the confidence interval will get larger.

It is important to understand that we are trying to capture the value of the unknown parameter θ in our confidence interval. For a given calculated interval, we do not know if we have been successful or not. We only have an idea of what happens with **repeated** sampling.

Selecting Sample Sizes

At some point in their lives, every statistician is asked "I am planning an experiment, how large should my sample be?" This question is always very difficult to answer and is often problem specific, with financial or time constraints on producing samples, and we cannot answer appropriately unless we are told how much accuracy is needed in estimating the parameter of interest. In the case where the experimentor can tell us what bound they want on ϵ, the **error of estimation**, we can give them a hand.

We assume that the error of estimation is one-half of the width on the $(1-\alpha)\%$ confidence interval. When our samples come from a normal distribution and we are computing a confidence interval for the mean, we have

$$\epsilon = \frac{\sigma \cdot z_{\alpha/2}}{\sqrt{n}}$$

We can rearrange this equation with a bit of algebra to get the required sample size n,

$$n = \left\lceil \frac{(z_{\alpha/2})^2 \sigma^2}{\epsilon^2} \right\rceil$$

The **ceiling** is taken because as statisticians we always want to be conversative. Often times in practice we don't know σ^2 either, in which case it is common to use one of two estimators of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{N} (X_i - \bar{X})^2 \tag{1}$$

$$\hat{\sigma}^2 = \frac{Range^2}{16} \tag{2}$$

(1) is more commonly used currently, (2) is an old quality control rule-of-thumb which has mostly fallen out of favor. Nonetheless it is important to be exposed to it.

If we are using a Normal approximation to the Binomial distribution to get confidence intervals for p, then we have the following expression for the error of estimation

$$\epsilon = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$$

This gives us the following expression for n:

$$n = \left\lceil (z_{\alpha/2})^2 \frac{p(1-p)}{\epsilon^2} \right\rceil$$

An interesting concern that arises here is that we are estimating p, and different values of p can give drastically different required sample sizes. Once again we want to be conservative. We can do this by always choosing the maximum value of $p(1-p)$. This is simple, and is done by taking the derivative.

$$\frac{d}{dp}\{p(1-p)\} = \frac{d}{dp}\{p - p^2\} = 1 - 2p = 0, p = \frac{1}{2}$$

This gives us that the most conservative estimate for n is

$$n = \frac{(z_{\alpha/2})^2}{4\epsilon^2}$$

We can make similar arguments for some of the other cases in the Table. For example, we can get a solution for $\mu_1 - \mu_2$ in the special case where $n_1 = n_2 = n; \sigma_1^2 = \sigma_2^2 = \sigma^2$.

$$\epsilon = z_{\alpha/2}\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = z_{\alpha/2}\sqrt{\frac{2\sigma^2}{n}}$$

This gives us

$$n = \left\lceil (z_{\alpha/2})^2 \frac{2\sigma^2}{\epsilon^2} \right\rceil$$

We can do the same for $p_1 - p_2$ when we have $n_1 = n_2 = n$

$$\epsilon = z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$$

Solving for this gives us

$$n = \left\lceil (z_{\alpha/2})^2 \frac{p_1(1-p_1) + p_2(1-p_2)}{\epsilon^2} \right\rceil$$

In the interest of being conservative, we let $p_1 = p_2 = .5$ to obtain

$$n = \left\lceil (z_{\alpha/2})^2 \frac{1}{2\epsilon^2} \right\rceil$$

Lecture 8: Efficiency, Consistency, Sufficiency

Motivation

So far we have learned how to compare any two estimators, and how to make confidence intervals with any unbiased estimator, so long as we have a sufficiently large sample and have a sampling distribution for our estimator that obeys the central limit theorem. Something that we have not yet covered are the immutable properties that we want for a good statistical estimator. What makes an estimator good? We will discuss the idea of **efficiency**, as well as give another way to compare two estimators using something called the relative efficiency. We will also talk about consistency of estimators, which is exactly what you might think it is. It tells us when we can expect an estimator to perform well when we use it over and over. Lastly we will talk about sufficiency. We want our estimators to use ALL of the available information from the sample, otherwise we are throwing away information that may be pertinent to estimation.

Efficiency

If we have two unbiased estimators of θ, say $\hat{\theta}_1$ and $\hat{\theta}_2$, then we would certainly prefer the estimator with the smaller variance. An estimator $\hat{\theta}_1$ is said to be relatively more **efficient** than $\hat{\theta}_2$ if $MSE[\hat{\theta}_1] < MSE[\hat{\theta}_2]$. We give a more formal definition below.

Definition: Given two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, with mean squared errors of $MSE(\hat{\theta}_1)$ and $MSE(\hat{\theta}_2)$ respectively, then the **efficiency** of $\hat{\theta}_1$ compared to $\hat{\theta}_2$ is defined as

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}$$

We say that $\hat{\theta}_1$ is a better estimator of θ if $eff(\hat{\theta}_1, \hat{\theta}_2) > 1$.

In many ways, this is very similar to what we have been doing already by comparing the MSE's of two estimators. We will show this with an example.

Efficiency Example

Suppose that we have $Y_1, ..., Y_n \sim Unif(0, \theta)$ are i.i.d. Let $\hat{\theta}_1 = 2\bar{Y}$, and $\hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$. Find $eff(\hat{\theta}_1, \hat{\theta}_2)$.

Solution
We will start by finding the MSE of $\hat{\theta}_1 = 2\bar{Y}$.

$$\mathbb{E}[2\bar{Y}] = 2\mathbb{E}[\bar{Y}] = 2 \cdot \frac{\theta}{2} = \theta$$

$$\mathbb{V}[2\bar{Y}] = 2^2\mathbb{V}[\bar{Y}] = 4\mathbb{V}[\bar{Y}] = 4 \cdot \frac{\theta^2}{12n} = \frac{\theta^2}{3n}$$

$$B(\hat{\theta}_1) = \mathbb{E}[\hat{\theta}_1] - \theta = \theta - \theta = 0$$

$$MSE(\hat{\theta}_1) = \frac{\theta^2}{3n}$$

Now we will find the MSE of $\hat{\theta}_2$. We have from a previous homework assignment that $\mathbb{E}[X_{(n)}] = \frac{n}{n+1}\theta$ and $\mathbb{V}[X_{(n)}] = \frac{n\theta^2}{(n+1)^2(n+2)}$. Then,

$$\mathbb{E}\left[\frac{n+1}{n}X_{(n)}\right] = \frac{n+1}{n} \cdot \mathbb{E}[X_{(n)}] = \frac{n+1}{n}\frac{n}{n+1}\theta = \theta$$

$$\mathbb{V}\left[\frac{n+1}{n}X_{(n)}\right] = \frac{(n+1)^2}{n^2}\mathbb{V}[X_{(n)}] = \frac{(n+1)^2}{n^2}\frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$$

$$MSE(\hat{\theta}_2) = \frac{\theta^2}{n(n+2)}$$

Then,

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)} = \frac{\theta^2/(n(n+2))}{\theta^2/3n} = \frac{3}{n+2}$$

So, for $n = 1$, the estimators have the same MSE. If we have more than one observation, $\hat{\theta}_2$ performs better.

Definition: An estimator $\hat{\theta}_1$ is said to **dominate** $\hat{\theta}_2$ if:

- $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$ for at least one value of θ.
- $MSE(\hat{\theta}_1) \leq MSE(\hat{\theta}_2)$ for every value of θ.

In other words, for an estimator to dominate another, it needs to do strictly better for at least one value and it can't do worse for any other values. Looking at our example above, we see that neither estimator dominates the other for $n = 1$, but for $n > 1$, $\hat{\theta}_1$ is dominated by $\hat{\theta}_2$.

Consistency of an estimator

As we take more and more samples, we want our estimator to get closer and closer to the true value. If this isn't happening then there is no point in taking extra samples, right? This property is referred to as **consistency**. Consistency is an asymptotic property that is defined as the sample size increases without bound ($n \to \infty$).

Mathematically, we say that an estimator $\hat{\theta}_n$ is consistent for θ if $\forall \epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \leq \epsilon) = 1,$$

or equivalently,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0$$

For intuition's sake, we can think about a consistent estimator as being able to concentrate the distribution of $\hat{\theta}_n$ arbitrarily tightly around θ by increasing the sample size.

Theorem 9.1 An unbiased estimator $\hat{\theta}_n$ for θ is consistent if

$$\lim_{n\to\infty} \mathbb{V}[\hat{\theta}_n] = 0$$

Proof: Recall Markov's theorem. For a random variable Y with $\mathbb{E}[Y] = \mu, \mathbb{V}[Y] = \sigma^2 < \infty$, $\forall k > 0$ we have that

$$P(|Y - \mu| > k\sigma) \leq \frac{1}{k^2}$$

Let's replace Y with $\hat{\theta}_n$. Since $\hat{\theta}_n$ is unbiased, we know that it's expectation is θ. We can replace σ with $\sqrt{\mathbb{V}[\hat{\theta}]}$ to obtain

$$P\Big(|\hat{\theta}_n - \theta| > k\sqrt{\mathbb{V}[\hat{\theta}]}\Big) \leq \frac{1}{k^2}$$

If we pick $k = \frac{\epsilon}{\sqrt{\mathbb{V}[\hat{\theta}]}}$, we have

$$P(|\hat{\theta}_n - \theta| > \epsilon) \leq \frac{\mathbb{V}[\hat{\theta}_n]}{\epsilon^2}$$

Taking the limit on both sides:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) \leq \lim_{n\to\infty} \frac{\mathbb{V}[\hat{\theta}_n]}{\epsilon^2}$$

By assumption, $\mathbb{V}[\hat{\theta}_n] \to 0$, thus

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) \leq 0$$

Since probabilities are bounded below by 0, we have

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

Thus we have the definition of a consistent estimator. $\square$

Definition: An estimator $\hat{\theta}_n$ is an asymptotically unbiased estimator of θ if

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \theta$$

Theorem 9.2: An asymptotically unbiased estimator $\hat{\theta}_n$ for θ is consistent if

$$\lim_{n\to\infty} \mathbb{V}[\hat{\theta}_n] = 0$$

Definition: If $\hat{\theta}_n$ is a consistent estimator of θ, we say that $\hat{\theta}_n$ **converges in probability** to θ.

Theorem 9.3 (Continuous Mapping Theorem): Suppose $\hat{\theta}$ converges in probability to θ. Then, for any $g(\cdot) \in \mathcal{F}(\mathbb{R})$ (real valued functions) that is continuous at θ, $g(\hat{\theta}_n)$ converges in probability to $g(\theta)$.

This is a powerful theorem for our purposes. We now know that asymptotically we can estimate functions of our parameters.

Lemma 9.1 Suppose $\hat{\theta}_n$ converges in probability to θ and $\hat{\tau}_n$ converges in probability to τ. Then,

- $\hat{\theta}_n + \hat{\tau}_n$ converges in probability to $\theta + \tau$
- $\hat{\theta}_n \cdot \hat{\tau}_n$ converges in probability to $\theta \cdot \tau$
- $\hat{\theta}_n / \hat{\tau}_n$ converges in probability to θ / τ

Sufficiency

So far, we have picked estimators based on properties such as their MSE, whether or not they are unbiased, and in general whether or not they seem reasonable (mostly the first 2, though). What we have not talked about is whether or not these estimators are using all of the available information. We would hate to be using estimators that don't make use of all of the data, as that means that we are missing out.

A statistic that uses all of the available information in a sample are called **sufficient statistics**. An estimator based off of a sufficient statistic is said to have the property of **sufficiency**.

Definition: Let $Y_1, Y_2, ..., Y_n$ denote a random sample from a probability distribution with some unknown parameter θ. The statistic $U(Y_1, .., Y_n)$ is said to be **sufficient** for θ if the conditional distribution of $Y_1, .., Y_n$ given U does not depend on θ.

This definition can be a bit hard to verify as it currently stands. Calculations to look at conditional distributions may be tedious. It turns out that we are able to verify sufficiency through other means, but first we need to introduce the concept of a likelihood function.

Let $y_1, ..., y_n$ be observations taken from corresponding random variables $Y_1, ..., Y_n$ with unknown parameters $\Theta = \{\theta_1, ..., \theta_k\}$. The **likelihood** of $\mathbf{Y} = \{y_1, ..., y_n\}$ is defined as

$$\mathscr{L}(\Theta|\mathbf{Y}) = \prod_{i=1}^{n} f(y_i|\Theta)$$

Example: Exponential Likelihoods

Suppose that $Y_1, ..., Y_n \sim Exp(\lambda)$ are independent, with $f(y_i) = \lambda^{-1} \exp(\lambda^{-1} y_i)$. Find the likelihood function, $\mathscr{L}(\lambda|\mathbf{Y})$.

Solution

We have by the definition of the likelihood that

$$\mathscr{L}(\lambda|\mathbf{Y}) = \prod_{i=1}^{n} f(y_i|\lambda) = \prod_{i=1}^{n} \lambda^{-1} \exp(\lambda^{-1} y_i)$$

Using properties of the exponential, we can combine this to get

$$\mathscr{L}(\lambda|\mathbf{Y}) = \lambda^{-n} \exp(\lambda^{-1} \sum_{i=1}^{n} y_i) \quad \square$$

Now that we have a bit of a feel for the likelihood, we will look at a new way to determine whether a statistic U is sufficient or not.

Theorem 9.4: The Factorization Theorem

Let U be a statistic based on the random sample $Y_1, ..., Y_n$. U is a sufficient statistic for estimating θ if and only if

$$\mathscr{L}(y_1, ..., y_n|\theta) = g(U, \theta) \cdot h(y_1, ..., y_n),$$

where $g(U, \theta)$ is a function **only** of U, θ, and $h(y_1, ..., y_n)$ is not a function of θ.

Revisiting the exponential example

Using the factorization theorem, show that $U = \bar{Y}$ is sufficient for λ.

Lecture 9: Exponential Families

Motivation

Throughout this two semester course, many distributions have been mentioned. Binomial, Normal, Gamma, Beta, Poisson, Geometric are just a few that we can name. It is natural to take the opportunity to wonder: do any of these distributions share certain properties? Is there an overarching statement that we can make about a lot of these distributions? Of course we know that there are necessary properties for their probability mass/density functions, i.e. they must sum/integrate to one, they must be right continuous, and they must be non-negative - but is that all? It turns out that all of the distributions mentioned in the second sentence come are from the same **family** of distributions: the exponential family. Distributions from the exponential family have nice properties that we can leverage, especially when we are talking about sufficient statistics.

The Exponential Family

A family of pdfs or pmfs is called an **exponential family** if it can be expressed in the form

$$f(x|\Theta) = h(x)c(\Theta)\exp\Big(\sum_{i=1}^{k} w_i(\Theta)t_i(x)\Big)$$

Here $h(x) \geq 0$ and $t_1(x), ..., t_k(x)$ are real-valued functions of the observation x (that cannot depend on Θ), $c(\Theta) \geq 0$, and $w_1(\Theta), ..., w_k(\Theta)$ are real-valued functions of Θ (that cannot depend on x). In order to show that a distribution is in the exponential family, we have to identify these functions and show that it can be written in the form above. Below we will cover 2 examples.

Normal Exponential Family

The normal distribution is one of the first things that should come to mind when we look at how an exponential distribution is defined. It already largely has the form that we need, and will give us an easy example to gain comfort with how to find the $w_i(\cdot), t_i(\cdot), h(\cdot)$ and $c(\cdot)$ functions.

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2 \Big)$$

We can expand the square in the numerator to get

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) \Big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \Big) =$$

First we will pick $c(\mu, \sigma^2)$ and $h(x)$, as they should be the easiest. We see that the function outside of the exponential is

$$\frac{1}{\sqrt{2\pi}\sigma} = 1 \cdot \frac{1}{2\pi\sigma}$$

Thus we can pick $h(x) = 1$ and $c(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}$. Then, looking inside of the exponential we have $-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}$. We can pick

$$t_1(x) = x^2, w_1(\mu, \sigma^2) = -\frac{1}{2\sigma^2}$$

$$t_2(x) = x, w_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}$$

$$t_3(x) = 1, w_3(x) = -\frac{\mu}{2\sigma^2}$$

Thus we've shown that the normal distribution is in the exponential family. Let's try a bit of a more difficult example now.

Beta Exponential Family

Unlike the normal distribution, we look at the beta distribution pdf and think to ourselves: this doesn't look like it has the proper form at all. And that's true, it is not immediately clear that the beta is a member of the exponential family. We will use a couple of tricks with exponentials and logarithms to show that it has the form that we want.

$$f(x|\alpha, \beta) = \frac{1}{\beta(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, x \in (0, 1)$$

We can start by remembering a couple of easy properties: $\exp(\log(x)) = x$, and $\log(x^a) = a\log(x)$. We can look at the beta pdf and write it using this first fact as:

$$f(x|\alpha, \beta) = \frac{1}{\beta(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} = \frac{1}{\beta(\alpha, \beta)} \exp(\log(x^{\alpha-1})) \exp(\log((1-x)^{\beta-1}))$$

We can further simplify this by combining across the exponential and using the second fact:

$$\frac{1}{\beta(\alpha,\beta)}\exp(\log(x^{\alpha-1}))\exp(\log((1-x)^{\beta-1})) = \frac{1}{\beta(\alpha,\beta)}\exp((\alpha-1)\log(x)+(\beta-1)\log(1-x))$$

We will once again start by looking to see what $c(\alpha,\beta)$ and $h(x)$ are. Outside of the exponent we have:

$$\frac{1}{\beta(\alpha,\beta)} = 1 \cdot \frac{1}{\beta(\alpha,\beta)}$$

Then we can pick $h(x) = 1$ and $c(\alpha,\beta) = \frac{1}{\beta(\alpha,\beta)}$. Looking into the exponent we have:

$$(\alpha-1)\log(x) + (\beta-1)\log(1-x)$$

We can pick

$$t_1(x) = \log(x), w_1(\alpha,\beta) = \alpha - 1$$
$$t_2(x) = \log(1-x), w_2(\alpha,\beta) = \beta - 1$$

Thus we have shown that the beta distribution is a member of the exponential family. It is important to remember these little tricks when we are trying to show that distributions are members of the exponential family - it is a required exercise for many other exponential family distributions such as the binomial, gamma, and negative binomial.

So far we have looked at how to show that a distribution is a member of the exponential family but we have not yet explored any of the properties that all of these distributions share. The first of these is given below in the form of a theorem.

Theorem 9.5: Suppose that $X_1, X_2, ..., X_n$ are i.i.d. observations from a random variable that belongs to the exponential family, with the canonical parameterization

$$f(x_j|\Theta) = h(x_j)c(\Theta)\exp\Big(\sum_{i=1}^{k} w_i(\Theta)t_i(x_j)\Big)$$

Then,

$$T(\mathbf{X}) = \Big(\sum_{j=1}^{n} t_1(X_j), \sum_{j=1}^{n} t_2(X_j), ..., \sum_{j=1}^{n} t_k(X_j)\Big)$$

is a sufficient statistic for Θ.

Proof We can start by writing the likelihood of $\mathbf{X}$,

$$\mathscr{L}(\Theta|\mathbf{X}) = \prod_{j=1}^{n} h(X_j)c(\Theta)\exp\Big(\sum_{i=1}^{k} w_i(\Theta)t_i(X_j)\Big)$$

$$= \Big(\prod_{j=1}^{n} h(X_j)\Big) c(\Theta)^n \prod_{j=1}^{n} \exp\Big(\sum_{i=1}^{k} w_i(\Theta) t_i(X_j)\Big)$$

$$= \Big(\prod_{j=1}^{n} h(X_j)\Big) c(\Theta)^n \exp\Big(\sum_{j=1}^{n} (\sum_{i=1}^{k} w_i(\Theta) t_i(X_j))\Big)$$

$$= \Big(\prod_{j=1}^{n} h(X_j)\Big) c(\Theta)^n \exp\Big(w_1(\Theta)\sum_{j=1}^{n} t_1(X_j) + ... + w_k(\Theta)\sum_{j=1}^{n} t_k(X_j)\Big)$$

We can invoke the factorization theorem, choosing $h(X_1, .., X_n) = \Big(\prod_{j=1}^{n} h(X_j)\Big)$, and the rest to be $g(U, \mathbf{X})$, with $U = T(\mathbf{X})$. $\square$

Example: Beta Sufficient Statistics using Theorem 9.5

Suppose $X_1, .., X_n$ are i.i.d. observations from a $Beta(\alpha, \beta)$ distribution. Find a sufficient statistic for $\{\alpha, \beta\}$.

Solution: Last class we showed that the beta distribution is a member of the exponential family, and derived the t_i functions to be $t_1(X_j) = \log(X_j)$, and $t_2(X_j) = \log(1 - X_j)$. Using the theorem above, we have that

$$T(\mathbf{X}) = \Big(\sum_{j=1}^{n} \log(X_j), \sum_{j=1}^{n} \log(1 - X_j)\Big)$$

is a sufficient statistic for α, β.

Gamma Sufficient Statistics

We will now use Theorem 9.5 to look at sufficient statistics for the Gamma distribution. First we will have to show that the Gamma distribution is a member of the exponential family.

$$f(x|\alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\Big(-\frac{x}{\beta}\Big), x \in (0, \infty)$$

We will use some tricks from last lecture to write $x^{\alpha-1} = \exp(\log(x^{\alpha-1})) = \exp((\alpha - 1)\log(x))$. Then, we can rewrite the pdf as

$$f(x|\alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \exp\Big((\alpha - 1)\log(x) - \frac{x}{\beta}\Big), x \in (0, \infty)$$

Looking at the terms outside of the exponential, we have $(\beta^\alpha \Gamma(\alpha))^{-1}$. We can pick $h(x) = 1, c(\alpha, \beta) = (\beta^\alpha \Gamma(\alpha))^{-1}$. Inside of the exponential we have:

$$(\alpha - 1)\log(x) - \frac{x}{\beta}$$

We can pick $t_1(x) = \log(x), w_1(\alpha, \beta) = \alpha - 1, t_2(x) = x, w_2(\alpha, \beta) = -\beta^{-1}$. Thus we have shown that the gamma distribution is a member of the exponential family. We can apply Theorem 9.5 to show that

$$T(\mathbf{X}) = \Big(\sum_{j=1}^{n} \log(X_j), \sum_{j=1}^{n} X_j \Big)$$

is a sufficient statistic for $\{\alpha, \beta\}$.

An Interesting Case: The Weibull

The Weibull distribution is a distribution that commonly arises from applications in physics and quality control. Though the distribution is named after the Swedish mathematician Waloddi Weibull, it was first discovered by Frechet in 1927, and used by Rosin and Rammler in 1933 to describe a particle size distribution. The Weibull is the resulting transformation from taking the k^{th} root of an exponential random variable. The pdf is given by

$$f(x|\lambda, k) = \frac{k}{\lambda^k} x^{k-1} \exp\Big(- \frac{x^k}{\lambda^k} \Big), x \geq 0, \lambda > 0, k > 0$$

The Weibull is **not** a member of the exponential family in general, as it is not possible to break apart apart the k from the x inside of the exponent in the pdf. In the case where k is fixed, the Weibull distribution is a member of the exponential family. We can rewrite $x^{k-1} = \exp((k-1)\log(x))$ to obtain

$$f(x|\lambda, k) = \frac{k}{\lambda^k} \exp\Big((k-1)\log(x) - \frac{x^k}{\lambda^k} \Big),$$

Examining the values outside of the distribution we have $k\lambda^{-k}$, thus we can pick $h(x) = 1, c(\lambda, k) = k\lambda^{-k}$. Inside of the exponential we have

$$(k-1)\log(x) - \frac{x^k}{\lambda^k}$$

We can pick $t_1(x) = \log(x)$, $w_1(\lambda, k) = k - 1$, $t_2(x) = x^k$ and $w_2(\lambda, k) = -\lambda^{-k}$. Then, we know that

$$T(\mathbf{X}) = \Big(\sum_{j=1}^{n} \log(x), \sum_{j=1}^{n} x^k \Big)$$

is a sufficient statistic for λ.

Lecture 10: Rao-Blackwell, MVUEs, and Cramer Rao

Motivation

So far, we have talked a lot about desirable properties of estimators. Unbiasedness, low variance, sufficiency, and consistency have all been discussed in great detail. One thing that has not been discussed so far is how to improve our existing estimators. We have talked a great deal about how to verify certain properties of estimators, but haven't had much to say on how to make them better. A natural thing to think about when talking about making estimators "better" is lowering the variance, particularly when we already have an unbiased estimator. It turns out that lowering the variance is intimately related to the concept of sufficiency, and we will look at a way to lower the variance of an unbiased estimator by conditioning the estimator on a sufficient statistic.

Quick Refresher on Conditional Means and Variances

In order to prove the Rao-Blackwell Theorem, we need results regarding Conditional Means and Variances. These should have been covered in the pre-requisite course, but in the event that they were not, the results are given below.

Let X and Y be any two random variables. Then,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

$$\mathbb{V}[X] = \mathbb{V}[\mathbb{E}[X|Y]] + \mathbb{E}[\mathbb{V}[X|Y]]$$

These are often referred to as the law of total expectation and the law of total variance.

The Rao-Blackwell Theorem

When we discussed sufficiency, one of the key ideas was that estimators based on sufficient statistics used all of the available information from the observed data. Using all of the available information, we expect to have less variation than a statistic than an estimator that is not based on a sufficient statistic. This is the idea behind the Rao-Blackwell Theorem.

Rao-Blackwell Theorem Let $\hat{\theta}$ be an unbiased estimator for θ with $\mathbb{V}[\hat{\theta}] < \infty$. If U is a sufficient statistic for θ, define $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|U]$. Then, for all θ,

$$\mathbb{E}[\hat{\theta}^*] = \theta, \mathbb{V}[\hat{\theta}^*] \leq \mathbb{V}[\hat{\theta}]$$

Proof The proof follows quickly from the definition of sufficiency and conditional expectation and variance laws. Because U is sufficient for θ, the conditional distribution of any statistic (including $\hat{\theta}^*$), given U, does not depend on θ. Therefore $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|U]$ is not a function of θ, and is therefore a statistic. Invoking the laws of total expectation and variance,

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\mathbb{E}[\hat{\theta}|U]] = \mathbb{E}[\hat{\theta}] = \theta$$

Thus $\hat{\theta}^*$ is unbiased.

$$\mathbb{V}[\hat{\theta}] = \mathbb{V}[\mathbb{E}[\hat{\theta}|U]] + \mathbb{E}[\mathbb{V}[\hat{\theta}|U]] = \mathbb{V}[\hat{\theta}^*] + \mathbb{E}[\mathbb{V}[\hat{\theta}|U]]$$

Since variances are always non-negative, and expectations of non-negative random variables are always non-negative, it follows from rearrangement of the variance terms that $\mathbb{V}[\hat{\theta}] \geq \mathbb{V}[\hat{\theta}^*]$. $\square$

After seeing this theorem, a tempting thought for most is to take the resulting estimator $\hat{\theta}^*$ and once again apply the Rao-Blackwell Theorem. This will unfortunately not work, as the second application will result once again in $\hat{\theta}^*$.

A natural follow up question after covering the Rao-Blackwell Theorem is: what sufficient statistic should we be using to condition on? The answer comes from another result that we just covered: we should be conditioning on the sufficient statistic that we obtain from applying the factorization theorem. We do not have the complete set of machinery built up to give a full theoretical justification, but here is the main idea. The factorization theorem identifies a statistic U that does the best job about summarizing the information available in the data. This is called the **minimal sufficient statistic**, and can be thought of as the sufficient statistic with the smallest dimensionality. Furthermore, in the cases we will be covering in this class, when we condition on this minimal sufficient statistic U, we will not only receive an estimator with a smaller variance, but it will be the estimator for θ that has the **minimum variance** for all unbiased estimators of θ. This is referred to intuitively as the **minimum variance unbiased estimator**. This is a powerful line of attack for finding MVUEs, as conditional expectation calculations can be difficult and tedious.

The above paragraph is very dense with information, so the idea is summarized below.

- Start with unbiased estimator $\hat{\theta}$
- Use factorization theorem to find a sufficient statistic U
- Apply Rao-Blackwell Theorem by conditioning on U
- Obtain minimum variance unbiased estimator $\hat{\theta}^*$

Example: Bernoulli MVUE

Let $Y_1, .., Y_n$ denote an independent random sample from a $Bern(p)$. Use the factorization theorem to find a sufficient statistic that best summarizes the data, and use this to find an MVUE for p.

We will start by writing the likelihood:

$$\mathscr{L}(p|\mathbf{Y}) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}$$

We can pick $U = \sum_{i=1}^n y_i$ to write the likelihood as

$$\mathscr{L}(p|\mathbf{Y}) = p^U(1-p)^{n-U}$$

Applying the factorization theorem, we let $g(U,p) = p^U(1-p)^{n-U}$ and $h(y_1, .., y_n) = 1$. From the factorization theorem, we have that $U = \sum_{i=1}^n Y_i$ is a minimal sufficient statistic for p. Thus the Rao-Blackwell Theorem tells us that if we can make an unbiased estimator that is a function of U, we have an MVUE for p.

$$\mathbb{E}[U] = \mathbb{E}[\sum_{i=1}^{n} Y_i] = np \rightarrow \mathbb{E}\Big[\frac{U}{n}\Big] = p$$

Thus $U/n = \bar{Y}$ is the minimum variance unbiased estimator for p in a binomial distribution.

Example: Normal MVUE

Suppose that we want to find the MVUE for μ and σ^2 for a Normal distribution from $X_1, .., X_n$ i.i.d. observations. Previously, we showed that the normal distribution was an exponential family member with $t_1(x) = x$ and $t_2(x) = x^2$. From sufficiency properties of exponential families, we have that

$$T(\mathbf{X}) = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$$

is a sufficient statistic for (μ, σ^2). Then, we want to build unbiased estimators from these sufficient statistics so that we can obtain the MVUE. We will start first with μ.

$$\mathbb{E}[\sum_{i=1}^{n} X_i] = n\mu \rightarrow \mathbb{E}\Big[\frac{\sum_{i=1}^{n} X_i}{n}\Big] = \mu$$

Thus for μ we have that the MVUE is $\bar{X}$. Computing an unbiased estimator for σ^2 is a bit more involved, but is nothing that we can't handle. We know from previous work that S^2 is unbiased for σ^2, but is it a function of the sufficient statistics?

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$

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

Thus S^2 can be written as a function of our sufficient statistics. This tells us that $\bar{X}$ is the MVUE for μ and S^2 is the MVUE for σ^2.

Cramer-Rao Lower Bounds

We can imagine a situation where it is very difficult to even obtain results for U from the factorization theorem. In cases like these, we want a theoretical lower bound on the variance of all unbiased estimators. This tells us that if the estimator we built attains that lower bound, then it must be the MVUE. Luckily for us, this result is available by way of a major theorem proved by Cramer, Rao, Frechet, and Darmois called the Cramer-Rao Inequality (aptly named?).

Theorem: Cramer-Rao Inequality: Let $X_1, .., X_n$ be a sample from a random variable with pdf $f(x|\theta)$, and let $W(\mathbf{X}) = W(X_1, .., X_n)$ be any estimator that satisfies

$$\frac{d}{d\theta}\mathbb{E}_\theta[W(\mathbf{X})] = \int_{\mathscr{X}} \frac{d}{d\theta}[W(x)f(x|\theta)]dx, \ \mathbb{V}_\theta[W(\mathbf{X})] < \infty$$

Then,

$$\mathbb{V}_\theta[W(\mathbf{X})] \geq \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[W(\mathbf{X})])^2}{\mathbb{E}_\theta[(\frac{d}{d\theta}\log(f(\mathbf{X}|\theta)))^2]}$$

The proof comes from a clever application of the Cauchy-Schwarz Inequality. We will not prove it here, but those who are interested may find the proof on Page 335 of Casella and Berger (Section 7.3).

Corollary: Cramer-Rao Inequality for iid Exponential Families Suppose that

the assumptions of Cramer-Rao are met, and that in addition $f(x|\theta)$ is a member of the exponential family with independent observations. Then,

$$\mathbb{V}_\theta[W(\mathbf{X})] \geq \frac{(\frac{d}{d\theta}\mathbb{E}_\theta[W(\mathbf{X})])^2}{-n\mathbb{E}_\theta[\frac{d}{d\theta^2}\log(f(X|\theta))]}$$

Furthermore, if we restrict $W(\mathbf{X})$ to be unbiased:

$$\mathbb{V}_\theta[W(\mathbf{X})] \geq \frac{1}{-n\mathbb{E}_\theta[\frac{d}{d\theta^2}\log(f(X|\theta))]}$$

Example: Poisson Cramer-Rao Bounds

Consider $X_1, .., X_n \sim Pois(\lambda)$ are i.i.d. Let $\hat{\lambda} = \bar{X}$. Is $\hat{\lambda}$ an MVUE for λ?

Solution: We will first compute the variance of $\hat{\lambda}$. We know from properties of the sample mean that

$$\mathbb{V}[\bar{X}] = \frac{\sigma^2}{n} = \frac{\lambda}{n}$$

Then, we want to compute the Cramer-Rao Lower Bound for the Poisson. Since we have the i.i.d. assumption and the Poisson is an exponential family member, for any statistic $W(\mathbf{X})$ we have

$$\mathbb{V}_\theta[W(\mathbf{X})] \geq \frac{1}{-n\mathbb{E}_\theta[\frac{d}{d\theta^2}\log(f(X|\theta))]} = \frac{1}{-n\mathbb{E}_\lambda[\frac{d}{d\lambda^2}\log(\exp(-\lambda)\lambda^x/x!)]}$$

Differentiating and rearranging gives that

$$\mathbb{V}_\lambda[W(\mathbf{X})] \geq \frac{\lambda}{n}$$

Then, since our estimator attains the minimum possible variance for an unbiased estimator, we have that $\bar{X}$ is an MVUE for λ.

Lecture 11: Method of Moment Estimators

Motivation

We have talked about some sophisticated ways of finding estimators, including through the factorization theorem and Rao-Blackwellization. These methods are very powerful and can give us fantastic properties that we want from estimators, but they won't always work so well. Sometimes the distributions that we work with are just painstakingly complicated, which is amplified by the additional calculations that we have to perform to invoke the factorization theorem or Rao-Blackwell. Powerful results like these do not come out of nowhere, they are usually built out of necessity.

Today we will learn about the **Method of Moments** estimators, the oldest and most primitive method of finding estimators, dating back to Karl Pearson in the late 1800s. The Method of Moments has the advantage of being simple and intuitive, but as we will see does not guarantee (most) of the desirable properties for estimators that we seek.

Method of Moments (MOM)

The idea of the method is moments is very simple, and it should not be surprising that it was the first line of attack. Recall that the k^{th} moment of a random variable is given by

$$\mu_k' = \mathbb{E}[Y^k]$$

The k^{th} sample moment of an estimator is given by

$$m_k' = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

The basic idea of method of moments is that our sample moments should provide us with fairly good estimates of the population moments. Since $\mu_1', .., \mu_k'$ are functions of the population parameters, we can equate population and sample moments in an attempt to solve for estimators of the desired parameters in terms of the sample moments. The idea is summarized below.

Method of moments
Choose as estimates the values of the parameters that are solutions of the system of equations $\mu_k' = m_k'$ for $k = 1, .., t$, where t is the number of parameters that you are estimating.

Immediately we see the simplicity of the idea - rather than having to deal with likelihoods and computation of conditional and joint distributions, we are tasked with solving a system of equations. So long as our parameter space is not collosal, we should able to solve these by hand.

Example: Uniform Estimator

Throughout this class we have used the example where $Y_1, .., Y_n \sim Unif(0, \theta)$ are independent and we are interested in estimating θ. The baseline estimator that we have always compared our new estimators to is $\hat{\theta} = 2\bar{Y}$. This result actually comes from a method of moments approach.

Our first moment is given by

$$\mu_1' = \mu = \frac{\theta}{2}$$

Note that since our parameter space of interest is only 1 dimension, this is actually all that we need. The corresponding first sample moment is given by

$$m_1' = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Now, setting $m_1' = \mu_1'$ and solving for θ, we get

$$\mu_1' = m_1' \rightarrow \frac{\theta}{2} = \bar{Y}, \hat{\theta} = 2\bar{Y} \quad \square$$

Example: Gamma MOM Estimators

One distribution that arises frequently but is difficult to work with is the Gamma distribution. Maximum likelihood methods have no closed form solution for the joint estimation of α and β, and optimization of the likelihood function in general is difficult because it involves working with polygamma functions. The method of moments is actually able to give us a pretty simple pair of estimators for α and β.

We start by looking at the moments of the Gamma distribution. Recall that $\mathbb{E}[X^2] = \mathbb{V}[X] + \mathbb{E}[X]^2$. Then,

$$\mu_1' = \mathbb{E}[X] = \alpha\beta$$
$$\mu_2' = \mathbb{E}[X^2] = \alpha\beta^2 + (\alpha\beta)^2$$

We set these equal to the population moments to get the system of equations

$$\alpha\beta = \bar{Y}$$

$$\alpha\beta^2 + \alpha^2\beta^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

We can solve the first equation to obtain $\hat{\beta} = \frac{\bar{Y}}{\hat{\alpha}}$. Then, plugging in this estimate for the second equation we get

$$\frac{\bar{Y}^2}{\hat{\alpha}^2} + \bar{Y}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

$$\frac{\bar{Y}^2}{\hat{\alpha}^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2$$

$$\hat{\alpha} = \frac{\bar{Y}^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2}$$

We can now plug this back into the estimate for β to obtain

$$\hat{\beta} = \frac{\bar{Y}}{\frac{\bar{Y}^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2}{\bar{Y}}$$

While we were not able to explicity use the factorization theorem to create an estimator, we can still use it in order to check whether or estimator is a function of sufficient statistics, a property that we certainly desire. To do this, we look at the likelihood:

$$\mathscr{L}(\alpha,\beta|\mathbf{Y}) = \prod_{i=1}^{n} \frac{1}{\beta^\alpha \Gamma(\alpha)} Y_i^{\alpha-1} \exp\left(-\frac{Y_i}{\beta}\right) = \left(\frac{1}{\beta^\alpha \Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} Y_i\right)^{\alpha-1} \exp\left(-\frac{\sum_{i=1}^{n} Y_i}{\beta}\right)$$

We can invoke the factorization theorem with $U = (\prod_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i)$ by picking $h(Y_1, ..., Y_n) = 1$, and the remainder to be $g(U, \alpha, \beta)$. We see that our estimator from the method of moments is not an MVUE, as $\prod_{i=1}^{n} Y_i$ does not appear anywhere.

From our due diligence in checking if the estimator was based on sufficient statistics, we have uncovered one of a couple of problems that come with using the MOM. We are not able to make assertions regarding the properties of an estimator. We cannot ever say with certainty that an estimator that we get from applying with method will be unbiased, consistent, or sufficient. On the other hand, making estimators from applying the Rao-Blackwell Theorem or the Factorization Theorem can at the very least guarantee us that our estimator is sufficient, and can often guarantee that it has mimimum variance in the case that it is unbiased.

Another property that the method of moments does not have is that the estimator does not preserve the range of the parameter that is being estimated. Consider the estimator that we obtained above for θ for the uniform distribution:

$$\hat{\theta} = 2\bar{Y}, \bar{Y} \in (0, \theta)$$

Then, the resulting range for our estimator $\hat{\theta}$ is $(0, 2\theta)$. We are capable (though unlikely) of estimating θ with a 100% error. Preserving the range of the parameter you are estimating is quite important. Not only can it have absolutely no physical intepretation, i.e. a case where we get a negative estimate for a variance, but it will also prevent us from making practically useful confidence intervals and forecasts. For these (and other) reasons, the MOM is usually left behind for more sophisticated methods like maximum likelihood estimation, which preserve the range of the parameters that are being estimated.

Lecture 12: Maximum Likelihood Estimation

Motivation

The most powerful methods that we have talked about for finding estimators utilize the Likelihood function. This should give indication that the likelihood is a very powerful tool in its own right. What if, rather than using the factorization theorem or the Rao-Blackwell theorem in order to find good unbiased estimators, we just looked at

the likelihood and maximized that in order to find our estimates? This is the premise behind maximum likelihood estimation, which is by far the most popular technique for finding estimators. The likelihood also forms the basis for Bayes estimators, which regularize the likelihood through the use of priors in order to find estimators with desirable properties.

Maximum Likelihood Estimation

Recall the definition of the likelihood:

Let $y_1, ..., y_n$ be observations taken from corresponding random variables $Y_1, ..., Y_n$ with unknown parameters $\Theta = \{\theta_1, ..., \theta_k\}$. The **likelihood** of $\mathbf{Y} = \{y_1, ..., y_n\}$ is defined as

$$\mathscr{L}(\Theta|\mathbf{Y}) = \prod_{i=1}^{n} f(y_i|\Theta)$$

The idea behind maximum likelihood estimation is that we want to treat this likelihood as a function of Θ and maximize it. The resulting point where the maximum is achieved is the maximum likelihood estimator. The idea seems like a simple application of calculus, so let us give it a try.

Example: Exponential MLE

Suppose $X_1, .., X_n \sim Exp(\lambda)$ are iid. Find the maximum likelihood estimator, $\hat{\lambda}_{MLE}$.

Naturally, the first step is to compute the likelihood:

$$\mathscr{L}(\lambda|\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\lambda} \exp\Big(- \frac{X_i}{\lambda}\Big) = \frac{1}{\lambda^n} \exp\Big(- \frac{\sum_{i=1}^{n} X_i}{\lambda}\Big)$$

Now that we have the likelihood, we want to take its derivative with respect to λ and set it equal to 0.

$$\frac{d}{d\lambda}\mathscr{L}(\lambda|\mathbf{X}) = \frac{d}{d\lambda}\Big\{ \frac{1}{\lambda^n} \exp\Big(- \frac{\sum_{i=1}^{n} X_i}{\lambda}\Big)\Big\} = 0$$

Computing the derivative using the product rule:

$$\frac{d}{d\lambda}\Big\{ \frac{1}{\lambda^n} \exp\Big(-\frac{\sum_{i=1}^{n} X_i}{\lambda}\Big)\Big\} = -\frac{n}{\lambda^{n+1}} \exp\Big(-\frac{\sum_{i=1}^{n} X_i}{\lambda}\Big)+\frac{\sum_{i=1}^{n} X_i}{\lambda^{n+2}} \exp\Big(-\frac{\sum_{i=1}^{n} X_i}{\lambda}\Big) = 0$$

Then,

$$\frac{n}{\lambda^{n+1}} \exp\Big(- \frac{\sum_{i=1}^{n} X_i}{\lambda}\Big) = \frac{\sum_{i=1}^{n} X_i}{\lambda^{n+2}} \exp\Big(- \frac{\sum_{i=1}^{n} X_i}{\lambda}\Big)$$

Rearranging this gives us

$$\frac{\sum_{i=1}^n X_i}{\lambda^{n+2}} = \frac{n}{\lambda^{n+1}},\ \hat{\lambda}_{MLE} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

We see that the MLE estimator for λ is $\bar{X}$, which is the same as the MVUE (computed in Lecture 9).

A Motivating Example for Improvement: Normal MLEs

Suppose that $X_1, .., X_n \sim N(\mu, \sigma^2)$ are iid, and we want an MLE for σ^2. The likelihood is given by

$$\mathscr{L}(\sigma^2|\mathbf{X}, \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\right)$$

We can already see that the derivative calculation here will be a bit messy. We have a lot of constants to keep track of while performing the product rule, leaving us with a lot of places to make a small error and get a result that doesn't make any sense. This leads us to the question: are there other functions that we can maximize in order to get the maximum likelihood estimator?

Monotone Increasing Functions

Definition: A function $f : \mathbb{R} \to \mathbb{R}$ is **monotone increasing** if $f(x) < f(y)$ whenever $x < y$.

In layman's terms, a monotone increasing function is always increasing. Using a little bit of our calculus knowledge we can rewrite this condition as: A function $f : \mathbb{R} \to \mathbb{R}$ is **monotone increasing** if $f'(x) > 0, \forall x \in \mathbb{R}$.

Monotone increasing functions have some great properties that we can leverage to simplify our MLE calculations. The most important one is given below as a theorem.

Theorem: Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function with a maximum of $f(a)$ that is attained at $a \in \mathbb{R}$. If $g : \mathbb{R} \to \mathbb{R}$ is a continuous monotone increasing function, then $g(f(a))$ is also a local maximum.

Proof Since f has a maximum at a, we have that

$$f'(a) = 0, f''(a) < 0$$

Looking at the derivative of the composition function $g(f(x))$ using the chain rule, we have

$$\frac{d}{dx} g(f(x)) = g'(f(x))f'(x)$$

Evaluating this at a, we have

$$\frac{d}{dx}g(f(x))\Big|_{x=a} = g'(f(a))f'(a) = g'(0)\cdot 0$$

Thus we see that $g(f(x))$ has a critical point at $x = a$. To show that this is a maximum, we must show that the second derivative is less than 0.

$$\frac{d^2}{dx^2}g(f(x)) = \frac{d}{dx}g'(f(x))f'(x) = g''(f(x))(f'(x))^2 + f''(x)g'(f(x))$$

Evaluating this at $x = a$, we have

$$g''(f(a))(f'(a))^2 + f''(a)g'(f(a)) = f''(a)g'(f(a))$$

By assumption, we have that $f''(a) < 0$, since $f(a)$ is a maximum. Since g is monotonic, it's derivative is always positive, thus $f''(a)g'(f(a)) < 0$. $\square$

The above theorem lets us transform our likelihood with a monotone increasing function in order to find our maximum, which allows us to utilize the log likelihood, as the log as a monotone increasing function. Let us revisit the 2 examples above with our new knowledge.

The Exponential MLE (Again)

Suppose $X_1, .., X_n \sim Exp(\lambda)$ are iid. Find the maximum likelihood estimator, $\hat{\lambda}_{MLE}$ using the log likelihood.

We have that the likelihood is given by

$$\mathscr{L}(\lambda|\mathbf{X}) = \frac{1}{\lambda^n}\exp\Big(-\frac{\sum_{i=1}^n X_i}{\lambda}\Big)$$

Then,

$$\log(\mathscr{L}(\lambda|\mathbf{X})) = -n\log(\lambda) - \frac{\sum_{i=1}^n X_i}{\lambda}$$

Taking the derivative and setting it equal to zero yields

$$\frac{d}{d\lambda}\log(\mathscr{L}(\lambda|\mathbf{X})) = -\frac{n}{\lambda} + \frac{\sum_{i=1}^n X_i}{\lambda^2} = 0$$

$$\frac{\sum_{i=1}^n X_i}{\lambda^2} = \frac{n}{\lambda},\ \hat{\lambda}_{MLE} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

As we can see, the calculation is much easier when dealing with the log likelihood rather than the traditional likelihood.

Normal Variance MLE

Suppose that $X_1, .., X_n \sim N(\mu, \sigma^2)$ are iid, and we want an MLE for σ^2 and for μ. The likelihood is given by

$$\mathscr{L}(\sigma^2|\mathbf{X}, \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(X_i-\mu)^2\Big) = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\Big)$$

We will start with the MLE for σ^2. In order to make things a little less complicated, let's write the likelihood in terms of $v = \sigma^2$,

$$\mathscr{L}(v|\mu, \mathbf{X}) = \Big(\frac{1}{\sqrt{2\pi v}}\Big)^n \exp\Big(-\frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu)^2\Big)$$

The log likelihood is given by

$$\log(\mathscr{L}(v|\mu, \mathbf{X})) = -\frac{n}{2}\log(v) - \frac{n}{2}\log(2\pi) - \frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu)^2$$

Taking the derivative and setting it equal to zero gives

$$\frac{d}{dv}\Big(-\frac{n}{2}\log(v) - \frac{n}{2}\log(2\pi) - \frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu)^2\Big) = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$

Rearranging gives

$$\frac{n}{2v} = \frac{1}{2v^2}\sum_{i=1}^{n}(X_i-\mu)^2, \ \hat{v} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

Since this is in terms of μ, we must find $\hat{\mu}$ and plug it into the estimate for σ^2.

Taking the derivative of the log likelihood with respect to μ gives

$$\frac{d}{d\mu}\Big(-\frac{n}{2}\log(v) - \frac{n}{2}\log(2\pi) - \frac{1}{2v}\sum_{i=1}^{n}(X_i-\mu)^2\Big) = \frac{2}{2v}\sum_{i=1}^{n}(X_i-\mu) = 0$$

Rearranging and simplifying gives

$$\sum_{i=1}^{n} X_i = n\mu, \ \hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

Thus our MLE for μ and σ^2 for a normal are given by

$$\hat{\mu}_{MLE} = \bar{X}, \ \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

The Uniform MLE

One of the most unintuitive examples when first talking about maximum likelihood estimation is the MLE for the uniform distribution. Consider $X_1, ..., X_n \sim Unif(0, \theta)$ are iid. Let's find the maximum likelihood estimator for θ.

$$\mathscr{L}(\theta|\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}_{X_i \in (0,\theta)}$$

Here, the $mathbf1$ is the indicator function, telling us that each X_i must be between 0 and θ. Then, we can write the likelihood as

$$\mathscr{L}(\theta|\mathbf{X}) = \frac{1}{\theta^n},\ 0 \leq X_i \leq \theta, i = 1, ..., n$$

Let's look at the condition that $0 \leq X_i \leq \theta$. This is telling us that each individual X_i is greater than 0 (which we already know), and that each individual X_i is less than θ. Note that this second statement is equivalent to saying that the largest X_i is smaller than θ, i.e. $X_{(n)} \leq \theta$. This allows us to write the likelihood as

$$\mathscr{L}(\theta|\mathbf{X}) = \frac{1}{\theta^n} \mathbf{1}_{X_{(n)} \leq \theta}$$

Then, we want to maximize this function. θ^{-n} is a decreasing function, so we want to pick the smallest value in our restricted range, which is $X_{(n)}$. Thus $\hat{\theta} = X_{(n)}$.

Properties of MLE Estimators

An estimator $\hat{\theta}$ that is obtained for maximum likelihood estimation has the following properties:

- $\hat{\theta}$ is consistent
- $\hat{\theta}$ is support preserving. This means that if $\theta \in (\theta_L, \theta_U)$ then $\hat{\theta} \in (\theta_L, \theta_U)$
- $\hat{\theta}$ is functionally invariant. If $\hat{\theta}$ is an MLE for θ, then $g(\hat{\theta})$ is the MLE for $g(\theta)$.

Lecture 14: Bayes Estimators

Motivation

Bayesian statistics is a very hot area of research, and has found its way into a great deal of applications due to its flexibility. Fields ranging from weather forecasting all the way to the spam filter on your smart phone will utilize Bayesian techniques. It is (in my opinion) a very natural way of statistical interpretation. Rather than trying to treat your models as being correct but having these unknown (constant) parameters

that are driving them, we get to treat our data as fixed observations with uncertainty belonging to our parameters and model. Bayesian statistics gives us another method of finding estimators in conjunction with a tool that we've already been using: the likelihood.

Prior Distributions

The first ingredient for any Bayesian recipe is a prior distribution. Generally speaking, a prior distribution, $\pi(\theta)$, is a distribution that we are putting on our unknown parameters that tell us some information about them *a priori*. In general it is bad to try to incorporate too much prior information, so we want to try to make our priors as uninformative as possible.

Posterior Distribution

We will define the posterior distribution, $\pi(\theta|\mathbf{X})$, as follows

$$\pi(\theta|\mathbf{X}) \propto \pi(\theta)\mathscr{L}(\theta|\mathbf{X})$$

The posterior distribution is the result of penalizing our likelihood function with our prior distribution, which is reflecting our prior information about the process. Depending on what type of prior we use, we can sometimes get a closed form distribution for the posterior that we can use to make point estimates and confidence intervals. This (usually) arises from using a certain type of prior distribution.

Definition: A **conjugate prior** $\pi(\theta)$ is a prior distribution such that $\pi(\theta)$ and $\pi(\theta|\mathbf{X})$ live in the same family of distributions.

To supplement some of these abstract definitions, let's do an example. This should help build some intuition.

Example: Poisson Gamma System

Suppose $X_1, ..., X_n \sim$ iid $Pois(\lambda)$. Let's pick our prior to be a gamma distribution, $\pi(\lambda) \sim Gamma(a, b)$. Now, let's take a look at the form of our posterior.

First, we need the likelihood,

$$\mathscr{L}(\lambda|\mathbf{X}) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{X_i}}{X_i!} = \frac{\exp(-n\lambda)\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!}$$

Now, writing out the distribution of the prior we have

$$\pi(\lambda) = \frac{1}{b^a\Gamma(a)}\lambda^{a-1}\exp\Big(-\frac{\lambda}{b}\Big)$$

Multiplying our likelihood and prior, we have

$$\pi(\lambda|\mathbf{X}) \propto \mathscr{L}(\lambda|\mathbf{X}) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{X_i}}{X_i!} = \frac{\exp(-n\lambda)\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!} \frac{1}{b^a \Gamma(a)} \lambda^{a-1} \exp\Big(-\frac{\lambda}{b}\Big)$$

First off, we can ditch any terms that do not have λ in them, as we don't actually need them to figure out the posterior distribution. We can combine across some common terms, giving us

$$\pi(\lambda|\mathbf{X}) \propto \exp(-\lambda(n + b^{-1}))\lambda^{\sum_{i=1}^{n} X_i + a - 1}$$

Looking at the form of this distribution, we see that it is a Gamma distribution, with parameters

$$\alpha^* = \sum_{i=1}^{n} X_i + a, \beta^* = \frac{1}{n + b^{-1}}$$

Since the prior was a Gamma distribution and the posterior was also a Gamma distribution, we conclude that the Gamma distribution is a conjugate prior for the Poisson distribution.

Bayes Estimators

We have had a brief look at prior distributions and posterior distributions, the fundamental backbone of Bayesian analysis. Now we will look at how to get estimators from these distributions.

Definition: The **Bayes estimator**, $\hat{\theta}_b$ from a posterior distribution $\pi(\theta|\mathbf{X})$ is given by

$$\hat{\theta}_b = \mathbb{E}_\theta[\pi(\theta|\mathbf{X})]$$

We see that the Bayes estimator is given by the expected value of the posterior distribution. For example, our Bayes estimator from the Poisson Gamma system example is given by

$$\hat{\lambda}_b = \alpha^* \beta^* = \frac{\sum_{i=1}^{n} X_i + a}{n + b^{-1}}$$

Now that we know how to compute Bayes estimators, let's try another example.

Example: Beta Bernoulli System

In Homework 4, I alluded to a bit of Bayesian analysis by asking you to compute the distribution of the likelihood when we treated p as a random variable and $\mathbf{X}$ as fixed. Here we will look at that example again, armed with new knowledge.

Suppose that $X_1, .., X_n \sim$ iid $Bern(p)$. We will perform a Bayesian analysis with $\pi(p) \sim Beta(a, b)$.

$$\mathscr{L}(p|\mathbf{X}) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$$

The distribution of our prior is:

$$\pi(p) \propto p^{a-1}(1-p)^{b-1}$$

Then,

$$\pi(p|\mathbf{X}) \propto p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} \cdot p^{a-1}(1-p)^{b-1}$$

Combining, we have

$$\pi(p|\mathbf{X}) \propto p^{\sum_{i=1}^n X_i+a-1}(1-p)^{n-\sum_{i=1}^n X_i+b-1}$$

Looking at the kernel, we see that this is also a beta distribution, with parameters

$$\alpha^* = \sum_{i=1}^n X_i + a, \beta^* = n - \sum_{i=1}^n X_i + b$$

Thus we see that the beta distribution is the conjugate prior for the bernoulli / binomial distributions. The Bayes estimator is given by

$$\hat{p}_b = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\sum_{i=1}^n X_i + a}{n + a + b}$$

The homework problem was the case where $a = b = 1$. It is also worth noting that when comparing bernoulli estimators in homework 3, the second estimator was obtained from the case where $a = b = \frac{1}{2}$.

Practice: Exponential Gamma System

The last example that we will go over for our crash course on Bayes estimators is the exponential gamma system.

Suppose that $X_1, .., X_n \sim$ iid $Exp(\lambda)$. We will perform a Bayesian analysis with $\pi(\lambda) \sim Gamma(a, b)$. We start with the likelihood,

$$\mathscr{L}(\lambda|\mathbf{X}) = \prod_{i=1}^{n} \lambda \exp(-\lambda X_i) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} X_i)$$

Our prior distribution has the form

$$\pi(\lambda) \propto \lambda^{a-1} \exp\Big(-\frac{\lambda}{b}\Big)$$

Then,

$$\pi(\lambda|\mathbf{X}) \propto \lambda^n \exp(-\lambda \sum_{i=1}^{n} X_i) \cdot \lambda^{a-1} \exp\Big(-\frac{\lambda}{b}\Big) = \lambda^{n+a-1} \exp\Big(-\lambda\big(\sum_{i=1}^{n} X_i + \frac{1}{b}\big)\Big)$$

We see that this is also a gamma distribution, with

$$\alpha^* = n + a, \beta^* = \sum_{i=1}^{n} X_i + \frac{1}{b}$$

What is the resulting Bayes estimator? Is the Gamma the conjugate prior to the exponential?

Some Bonus Information

Something that we have not discussed much is how to pick prior distributions, and/or how to pick the parameters for our prior distributions (generally called **hyperparameters**). We will not extensively discuss these methods in this course, we will only need to know how to compute a Bayes estimator when we are given a specific prior distribution. Nonetheless, it is an interesting topic. Prior distributions want to generally be chosen so that we are affecting the inference as little as possible. This can be done by attempting to maximize the distance between the prior distribution and the posterior distribution according to some metric. The priors that do this are called **reference priors**. Another common way of choosing priors is to make them transformation invariant. These priors are called Jeffreys priors, and it turns out that when you are choosing 1-dimensional priors, Jeffreys and reference priors are identical. If this has sparked your interest, additional information about choosing priors can be found in *A First Course in Bayesian Analysis* by Peter Hoff, and *Bayesian Theory* by Adrian Smith and Jose Bernardo.

Lecture 14: Introduction to Hypothesis Testing

Motivation

p-values and hypothesis tests are a hot subject currently - and not necessarily in a positive light. There is a natural need to satisfy our scientific curiosity by conducting well-designed experiments to examine our hypotheses. Academic science is largely interested in statistically significant results, and as such scientists are abusing p-values and hypothesis tests, as well as often lacking the statistical knowledge to properly interpret results. As statisticians, it is important that we have a fundamental understanding of hypothesis tests, both for our own analyses and for collaboration with others.

Hypothesis Testing

So far, our inference has largely focused on point estimation. We have looked at how to build estimators and assess the expected performance of these estimators, but we haven't talked about how to test our scientific hypotheses about parameters of interest. This is where hypothesis testing comes into play.

Definition 8.1.1 A **hypothesis** is a statement about a population parameter.

This definition is rather general, but the important point is that a hypothesis makes a statement about a population. Our goal when performing hypothesis testing will be to decide, based on our sample from the population, which of two disjoint hypotheses are true.

Definition 8.1.2 The two complementary hypotheses in a hypothesis testing problem are called the **null hypothesis** and the **alternative hypothesis**. We will denote these by H_0 and H_1 respectively.

If we let θ be the population parameter of interest, the null and alternative hypotheses are usually formatted $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_0^c$, where θ is some subset of the parameter space and θ_0^c is its complement. A couple examples of how we might format these based on a specific problem statement is given below.

Examples

Suppose that a scientist is interested in the average change in a patient's blood pressure after taking some medication. The scientist wants to be able to say that his medication makes a positive impact, so for this example we might have

$$H_0 : \theta = 0, H_1 : \theta > 0$$

As another example, suppose that an (angry) consumer is interested in the proportion of items that are defective from a particular supplier. Given the imperfection of factory machinery, there is usually an acceptable baseline threshold of defective items, given by θ_0. For this problem we would have something like

$$H_0 : \theta \geq \theta_0, H_1 : \theta < \theta_0$$

In hypothesis testing problems, we set our null and alternative hypotheses and then perform an experiment and collect sample data. After doing this, we are tasked with whether we are going to accept H_0 as true or decide that we have sufficient evidence that H_0 is false.

Definition 8.1.3 A **hypothesis testing procedure** or **hypothesis test** is a rule that specifies:

- For which sample values the decision is made to accept H_0 as true
- For which sample values H_0 is rejected and H_1 is accepted as true

The part of the sample space where H_0 is rejected is called the rejection region, or critical region. The complement of the rejection region is called the acceptance region.

We are taking an inherently probabilistic approach to assessing these hypotheses, and therefore it is not surprising that we can make mistakes when performing these tests. There are two types of mistakes that we can make: a **Type I** and **Type II** error. A Type I error is made when we reject H_0 when H_0 is actually true. We refer to this error rate as α. A Type II error rate is made if we accepted H_0 when H_1 is actually true. This quantity can be a bit tricky to find, and we will refer to it as β. We can write these α and β relations as a probabilistic statement,

$$P(\text{Reject } H_0 | H_0 \text{ True}) = \alpha$$

$$P(\text{Accept } H_0 | H_1 \text{ True}) = \beta$$

Now that we have formally defined all of the elements of a statistical test, we can look at an example.

Example

Suppose that a political candidate, Mr. Jones, claims that he will receive more than 50% of the vote and will become the elected official. Someone is interested in testing the claim made by Mr. Jones, and goes out to sample $n = 15$ voters to see if they

are voting for Mr. Jones. We note that this is a sample from a Binomial distribution, with parameters $n = 15$ and p. This gives us that our hypotheses are

$$H_0 : p = .5, H_1 : p < .5$$

The Test Statistic here will be Y, the number of people who answer the poll saying that they will be voting for Mr. Jones. The conductor of the survey decides that they will refute Jones claim if $Y \leq 2$. This tells us that $RR = \{0, 1, 2\}$.

What is α, the Type I error rate?

By definition, $\alpha = P(Y \in RR | H_0 \text{ True})$. Since we are conditioning on H_0 being true, we may set $p = .5$. This gives us that:

$$\alpha = \sum_{y=0}^{2} \binom{15}{y} (.5)^y (.5)^{15-y} \approx .004$$

We see that we have subjected ourselves to a small chance of a Type I error, meaning that it is very unlikely that we will conclude that Jones will lose when he actually wins.

Just because we have a small chance of making a Type I error does not mean that we have a fantastic test. We need to check that we are aptly protected from concluding that Jones is a winner when he will actually lose (a Type II error). Suppose that the true value for p is actually .3.

$$\beta = P(\text{Type II error}) = P(\text{Accept } H_0 | H_1 \text{ True})$$

We can find β for the case where we have that the true value of $p = .3$,

$$\beta = P(Y > .2 | p = .3) = \sum_{y=3}^{1} 5 \binom{15}{y} .3^y (.7)^{15-y} \approx .873$$

Thus if we use $RR = \{0, 1, 2\}$, our test will lead us to conclude with high probability that Jones is a winner, even if p is quite far from .5.

In general, if the true value is $p = p^*$, we can compute a curve that shows us our Type II error rate for various values of the true probability p^*

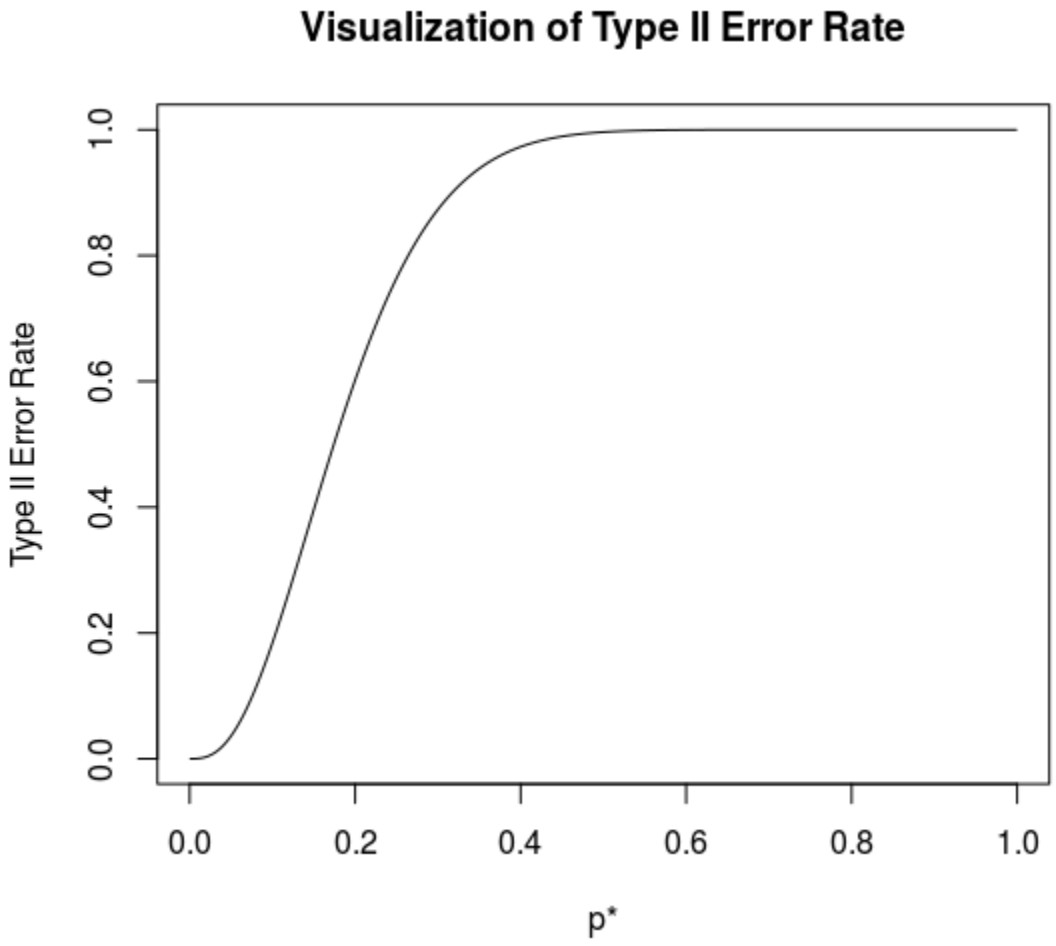

Figure 2: Visual representation Type II error for different p^* values

Lecture 15: Common Large Sample Tests

Motivation

A lot of the theory that we have built throughout this course will be relevant to hypothesis testing. Hypothesis testing inherently requires us to have estimators for the parameters that we are testing, and we have discussed that extensively. Some distributions that we use will not have closed form tests, but with a reasonable sample size we can almost always find an "approximately" normal test using the central limit theorem, and build confidence intervals.

Common Large Sample Tests

Suppose that we are interested in testing a set of hypotheses concerning a parameter θ, based on sample data $\mathbf{Y} = \{Y_1, ..., Y_n\}$. This section will focus on developing hypothesis testing procedures that are based on unbiased estimators that have an (approximately) normal distribution, by invoking the Central Limit Theorem.

Recall some of the large sample estimators discussed in lecture 7,

Target Parameter θ	Sample Size	$\hat{\theta}$	$\mathbb{E}[\hat{\theta}]$	$\sigma_{\hat{\theta}}$
μ	n	$\bar{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{p(1-p)}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\bar{Y}_1 - \bar{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^*$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Table 2: * requires the populations to be independent

We may want to test hypotheses of the form $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$. If our estimate $\hat{\theta}$ is close to θ_0, it seems rather reasonable that we will accept H_0. If the reality is that $\theta > \theta_0$, it is more likely that we will get a large(r) value of $\hat{\theta}$. Larger values of $\hat{\theta}$ will favor the rejection of $H_0 : \theta = \theta_0$. We can formalize these elements of our hypothesis test as follows:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta > \theta_0$$

$$\text{Test Statistic: } \hat{\theta}$$

$$\text{Rejection Region: RR} = \{\hat{\theta} > k\}, \text{ for some choice of } k$$

The value of k for the rejection region is determined by fixing our Type I error rate (α), and choosing k accordingly. If H_0 is true and our estimator $\hat{\theta}$ is approximately normally distributed, then $\hat{\theta} \sim N(\theta_0, \sigma_{\hat{\theta}})$. This allows us to pick our k value using the quantiles of a standard normal distribution,

$$k = \theta_0 + z_\alpha \sigma_{\hat{\theta}},$$

where z_α is defined as

$$Z \sim N(0,1), P(Z > z_\alpha) = \alpha$$

This gives us a rejection region of $RR = \{\hat{\theta} | \hat{\theta} > \theta_0 + z_\alpha \sigma_{\hat{\theta}}\}$. If we instead choose to use the test statistic $Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\theta_0}}$, then our rejection region is simply $\{z | z > z_\alpha\}$.

Example

Suppose that a vice president in charge of sales for a large corporation claims that salespeople are averaging no more than 15 sales contacts per week, and he would like to increase this figure. As a check on his claim, $n = 36$ salespeople are selected at random, and the number of contacts made by each is recorded for a single randomly selected week. The mean and variance of the 36 measurements were $\bar{X} = 17$ and $S^2 = 9$. Does the evidence observed contradict the vice president's claim? Use $\alpha = .05$.

Solution: By the statement of the problem, we can write the null and alternative hypotheses as

$$H_0 : \mu = 15 \text{ versus } H_a : \mu > 15$$

For sufficiently large n, the sample mean $\bar{X}$ is approximately normally distributed, with a mean of μ and a variance of σ^2/n, giving us a test statistic of

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

We can substitute in our estimate for σ by using S, which gives us $Z = \frac{17-15}{3/\sqrt{36}} = 4$. The z_α value for a one sided test with confidence level $\alpha = .05$ is given by $z_\alpha \approx 1.645$. Thus we reject H_0.

Example

A machine in a factory must be repaired if it produces more than 10% defective items among the large lot that it produces in a day. A random sample of 100 items from the day's production contains 15 defectives, and the supervisor claims that they need to call someone to repair the machine. Does the sample evidence support his decision? The machine repair is expensive, and thus the company wants to make sure that they don't call a repairman to fix it when there are no actual problems, thus we use $\alpha = .01$.

Solution: If we let Y denote the number of observed defective items out of 100, then Y follows a binomial distribution, with p being the probability that a randomly selected item is defective. This means that our null and alternative hypotheses are:

$$H_0 : p = .1 \text{ versus } H_a : p > .1$$

Using the table from above, we see that an unbiased estimator for p is $\frac{Y}{n}$. In order to compute our test statistic, we need to find $\sigma^2_{\hat{p}}$.

$$\mathbb{V}[\hat{p}] = \frac{1}{n^2}\mathbb{V}[Y] = \frac{p(1-p)}{n}$$

Then, we want to build our test statistic Z. Note that Z is built under the assumption that the **null hypothesis is true**. Thus anywhere that we see a p, we will replace it with p_0. This gives us that

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = 1.667$$

We can compute the rejection region from the 99th quantile of a normal distribution to be 2.32. This tells us that the supervisor should fail to reject the null hypothesis, and not call a repairman to fix the machine.

So far, we have really only discussed the case where we have $H_0 : \theta_0$ against $H_a : \theta > \theta_0$, but fear not. The case where we are testing against $H_a : \theta < \theta_0$ and $H_a : \theta \neq \theta_0$ are performed in an analogous manner, summarized below.

Large Sample α-Level Hypothesis Tests

$$H_0 : \theta = \theta_0$$
$$H_a\text{: } \{\theta > \theta_0, \theta < \theta_0, \theta \neq \theta_0\}$$
$$\text{Test statistic: } Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$
$$\text{Rejection region: } (\{z > z_\alpha\}, \{z < -z_\alpha\}, \{|z| > z_{\frac{\alpha}{2}}\})$$

Lecture 16: Power

Motivation

We want to be able to make statements about the error rate of our tests. We are less likely to trust the results from tests with high Type I and Type II error rates. We have discussed the idea of sample size calculations for the situations where we are interested in pinning down our confidence intervals between a certain error threshold with some probability. However, now we will discuss sample size calculations in the context of attempting to achieve certain levels of Type I and Type II error rates. We will see that these problems are only slightly more complicated than their previous counterparts.

Power

Suppose we are testing $H_0 : \theta = \theta_0$ vs $H_a : \theta > \theta_0$, and we want to calculate the Type II error, β, when $\theta = \theta_a$. Then, our rejection region is

$$RR = \{\hat{\theta} | \hat{\theta} > k\},$$

for some choice of k. We recall that

$$\beta = P(\theta \notin RR | H_a \text{ True}) = P(\hat{\theta} \leq k | \theta = \theta_a)$$

$$= P\Big(\frac{\hat{\theta} - \theta_a}{\sigma_\theta} \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}} | \theta = \theta_a\Big)$$

We note that with adequate sample size, the LHS of this probability statement is distributed roughly standard normal.

We define **power** of a test to be

$$Power = 1 - \beta$$

Example: Power Calculation

Let us consider testing $H_0 : \mu = 50$ vs $H_a : \mu > 50$, when $n = 100$ and $\sigma = 10$. Assume that $\alpha = .05$ and $\mu_a = 52$.

If we use $\hat{\mu} = \bar{X}$, we can write the rejection region,

$$RR = \{\hat{\mu} | \hat{\mu} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\} = \{\hat{\mu} | \hat{\mu} > 50 + 1.645 \cdot \frac{10}{\sqrt{100}}\}$$

$$= \{\hat{\mu} | \hat{\mu} > 51.645\}$$

We can use the formula from above to find β,

$$\beta = P\Big(\frac{\hat{\theta} - \theta_a}{\sigma_\theta} \leq \frac{k - \theta_a}{\sigma_{\hat{\theta}}} | \theta = \theta_a\Big) = P(Z \leq \frac{51.645 - 52}{10/\sqrt{100}})$$

$$= P(Z \leq -.355) = .3632$$

This tells us that the power is given by $1 - \beta = .6368$.

For this example, the larger that μ_a is, the higher the power of our test will be. Increasing power corresponds to lowering β. In order to lower the value of β for a given value of μ_a, we need to either increase α (a tradeoff between Type I and Type II error rates), or increase the sample size n.

Picking Sample Sizes for Specified α, β

A question that is continuously asked to us as statisticians is "I am designing a study and want to know how many samples I should take". (Good) scientists want to make sure that their results have not just appeared by chance, and want to be able to make statements about the probability that their results can be trusted.

Suppose that we are testing $H_0 : \mu = \mu_0$ vs $H_a : \mu > \mu_0$. We specify a value of α and a desired value of β for when $\mu = \mu_a$. We will use $\hat{\mu} = \bar{X}$. Then,

$$\alpha = P(\bar{X} > k | \mu = \mu_0)$$

$$= P\Big(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{k - \mu_0}{\sigma/\sqrt{n}} | \mu = \mu_0\Big) = P(Z > z_\alpha)$$

Similarly,

$$\beta = P(\bar{X} \leq k | \mu = \mu_a)$$

$$= P\Big(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{k - \mu_a}{\sigma/\sqrt{n}} | \mu = \mu_a\Big) = P(z \leq -z_\beta)$$

Thus we have

$$z_\alpha = \frac{k - \mu_0}{\sigma/\sqrt{n}}, -z_\beta = \frac{k - \mu_a}{\sigma/\sqrt{n}}$$

This gives us 2 equations that we can use to solve for n,

$$k = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_a - z_\beta \frac{\sigma}{\sqrt{n}}$$

$$(z_\alpha + z_\beta)\frac{\sigma}{\sqrt{n}} = \mu_a - \mu_0$$

$$\sqrt{n} = \frac{(z_\alpha + z_\beta)\sigma}{(\mu_a - \mu_0)} \rightarrow n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}$$

The same formula will apply for lower-tail tests as well. For two-sided tests, this solution is approximately correct if we replace z_α with $z_{\alpha/2}$

Sample Size Example

Suppose that we want to test $H_0 : \mu = 15$ vs $H_a : \mu > 15$. Let $\mu_a = 16$, $\sigma^2 = 9$ What sample size is required if we use $\alpha = \beta = .05$?

Solution: We know that

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}$$

Then, for this example, we have $z_\alpha = z_\beta = 1.645$, yielding

$$n = \frac{(1.645 + 1.645)^2 \cdot 9}{(16 - 15)^2} = 97.4$$

Since we do not want to undershoot, we pick $n = 98$ as the appropriate sample size to achieve these levels.

The p-value

The elephant in the room thus far has been the infamous p-value. In our discussion of hypothesis testing thus far we have not yet actually defined the p-value.

Definition: The **p-value** is the probability of obtaining a sample result at least as contradictory to the null hypothesis as the obtained result, assuming that the null hypothesis is true.

Note: Definitions referring to the probability of obtaining sample results at least as extreme as those obtained are not correct unless they refer to "extreme in the direction of the alternative hypothesis".

We reject H_0 if the p-value is less than or equal to α.

Definition: The p-value is the smallest value of α for which we would reject H_0.

The p-value contains more information than whether or not one rejects H_0 for a given value of α. The use of p-values allows the reader of a research report to make their own decision about the result, based on their own personal α value.

As I mentioned in the introduction for the hypothesis testing section, p-value methods have been heavily criticized recently. Part of these criticisms is that very small deviations from $H_0 : \theta = \theta_0$ can lead to very small $p-$values for large sample sizes. It is recommended that we use confidence intervals to assess the size of an effect, because low (including significant) p-values do not necessarily imply that there is practical significance. Understanding the difference between statistical significance and practical significance is very important to use as career statisticians.

Lecture 17: Likelihood Ratio Tests

Motivation

Of all the topics that we have discussed this semester, one of the most powerful tools has been the likelihood function. It should be no surprise that the likelihood has applications in hypothesis testing. We can look at ratios of likelihoods in the null and alternative hypothesis space in order to decide the form and rejection rules for our statistics. We will also see that statistics built off of these likelihood ratios will have nice properties in terms of the power of our tests.

LRTs

Likelihood ratio methods of hypothesis testing are closely related to maximum likelihood estimation, which we have talked about in this course. Recall that if we have a random sample $X_1, .., X_n$ with pdf (or pmf) $f(x|\theta)$, then we define the likelihood function to be

$$\mathscr{L}(\theta|\mathbf{X}) = \prod_{i=1}^{n} f(x_i|\theta)$$

If we let Θ denote the entire parameter space, we can define the likelihood ratio test as follows:

Definition: The **likelihood ratio test statistic** for testing $H_0 : \theta \in \Theta_0$ vs $H_a : \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{X}) = \frac{\sup_{\Theta_0} \mathscr{L}(\theta|\mathbf{X})}{\sup_{\Theta} \mathscr{L}(\theta|\mathbf{X})}$$

A **likelihood ratio test** is any test that has a rejection region of the form

$$RR = \{\mathbf{X}|\lambda(\mathbf{X}) \leq c\}, c \in [0, 1]$$

Though these definitions may seem complicated, the idea is beautifull intuitive. If Θ_0 is a simple hypothesis, i.e. $\mu = 5$, then the numerator of λ is just the joint density of our data evaluated at what we think our parameters are based on the null hypothesis. The denominator is simply the maximum value of the likelihood across the entire parameter space. If our likelihood evaluated at the Θ_0 value is close to the maximum likelihood value over the entire parameter space, then our value for λ will be close to 1, and we will be less likely to reject. However, if our likelihood under the null hypothesis deviates significantly from the maximum of the likelihood function over the entire parameter space, we will have a small value for λ, potentially giving us evidence that our null hypothesis is not true.

Example: Normal LRT

Some of the tests that we have built are *already* created from likelihood ratio tests and we just didn't know it. Here we will show the equivalence of a normal two-tailed test and the normal LRT.

Let $X_1, ..., X_n$ be a random sample from a Normal distribution with mean μ and variance σ^2, with σ^2 known. We will test $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$. Compute a likelihood ratio test statistic.

Solution: Recall that the LRT test statistic is given by:

$$\lambda(\mathbf{X}) = \frac{\sup_{\Theta_0} \mathscr{L}(\theta|\mathbf{X})}{\sup_{\Theta} \mathscr{L}(\theta|\mathbf{X})}$$

We start by writing a general equation for the normal likelihood:

$$\mathscr{L}(\mu|\mathbf{X}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x_i-\mu)^2\Big) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big)$$

Since the null region Θ_0 contains only the point θ_0, the numerator of the LRT is simply the likelihood evaluated at θ_0,

$$\sup_{\Theta_0} \mathscr{L}(\theta|\mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2\Big)$$

The denominator of the LRT will be given by the maximum value of the likelihood over every point outside of the null region. We know from our experience with maximum likelihood estimation that the normal likelihood is maximized at $\bar{X}$, giving us

$$\sup_{\Theta} \mathscr{L}(\theta|\mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\Big)$$

Then, we have that

$$\lambda(\mathbf{X}) = \frac{\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2\Big)}{\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\Big)} = \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\Big)$$

Thus we reject the null hypothesis for

$$\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\Big) \leq c$$

We can simplify this further by using the fact that

$$\sum_{i=1}^{n}(x_i-\theta_0)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2,$$

giving us that we reject H_0 for

$$\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}+\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\Big)\leq c$$

$$\exp\Big(-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\Big)\leq c, -\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\leq \log(c), \frac{(\bar{x}-\theta_0)^2}{\sigma^2/n}\geq -2\log(c)$$

$$|\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}|\geq\sqrt{-2\log(c)}$$

Thus we see that this is equivalent to the normal α level confidence test when we use $c=\exp(-\frac{1}{2}(z_{\alpha/2})^2)$.

It is a bit of a surprise that LRT methods lead to the same test statistic and rule that classical methods lead to. In fact, LRT methods for the hypotheses $H_0:\mu=\mu_0$ vs $H_a:\mu<\mu_0$ or $\mu>\mu_0$ will also lead to the same results that we have already derived.

Lecture 18: Generating Random Samples and Monte Carlo Methods

Motivation

We have performed several simulations in this class, to showcase results like the central limit theorem. Simulations are an integral part of statistics - they help to give us visualizations of results crafted from careful theory and can give us a much needed sanity check. We have learned quite a bit about random variables throughout these two semesters - how to compute their means, variances, moment generating functions, how to estimate their parameters, and many more. Something that we haven't talked about is how we generate random samples from these distributions. If you look at a histogram of a random sample from a normal distribution in R using *rnorm*, you will see (approximately) a bell curve. Today we will look at elementary methods of generating samples.

Inverse CDF Sampling

We will start off by assuming that we know how to generate a random sample from a uniform distribution. The most common algorithm for generating from a uniform distribution is called the **Mersenne Twister** algorithm, and it is the psuedo random number generator of choice for (almost) all relevant programming languages and softwares. More information on the Mersenne Twister can be found here: `https://`

blogs.mathworks.com/cleve/2015/04/17/random-number-generator-mersenne-twister/.

Being able to sample from a uniform distribution may not seem important at first, but there is a lot of significance in being able to generate a random sample over 0 to 1 for many reasons. Recall the cumulative distribution function for a random variable X, $F_X(x) = P(X \leq x)$. The CDF lives between 0 and 1, and represents the total area that is enclosed by the probability density function from $-\infty$ to x. The idea behind inverse CDF sampling is to generate a random sample of uniform numbers between 0 and 1 and then solving backwards to get the quantiles of the distribution, which will generate a random sample from $f(x)$. We can phrase this below as an algorithm, show an example, and verify that the example works.

The Inverse CDF Method:
Suppose X is a random variable with cumulative distribution function $F_x(x)$. Then, we can generate random samples from X by performing the following steps

- Generate $u \sim Unif(0, 1)$
- Compute $F_X^{-1}(x)$
- Compute $X = F_X^{-1}(u)$. This resulting $X \sim F_X(x)$

Example: Exponential Inverse CDF Method

Suppose that $X \sim Exp(\lambda)$, with pdf

$$f(x|\lambda) = \lambda \exp(-\lambda x), x \in (0, \infty)$$

Give an algorithm to produce a random sample from this distribution.

Solution Our first step is to solve for the cdf.

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy = \int_0^x \lambda \exp(-\lambda y)dy = -\exp(-\lambda y)\Big|_0^x = 1 - \exp(-\lambda x)$$

Now we need to find the inverse cdf, $F_X^{-1}(x)$. By definition, we have

$$F_X(F_X^{-1}(x)) = x$$

$$1 - \exp(-\lambda F_X^{-1}(x)) = x \rightarrow 1 - x = \exp(-\lambda F_X^{-1}(x))$$

$$F_X^{-1}(x) = -\frac{1}{\lambda}\log(1 - x)$$

Then, our algorithm to generate a sample from an exponential distribution is as follows:

- Generate $u \sim Unif(0, 1)$
- Set $X_i = -\frac{1}{\lambda}\log(1 - x)$
- Repeat until we have an adequate number of samples

Below is some R code showing how we would implement this in practice:

```
len <- 10000
x <- rep(NA, len)
for (i in 1:len){
  u <- runif(1)
  x[i] <- -.5*log(1-u)
}
hist(x, breaks = 80, prob = TRUE, main = 'Exponential Samples, lam = 2')
xp <- seq(from = .001, to = 5, length.out = 1000)
points(xp, dexp(xp, 2), type = 'l', col = 'red')
```

Example: Sampling from a Parabola

Suppose that we are interested in generating samples from the parabola,

$$f(x) = x^2, x \in (-(3/2)^{1/3}, (3/2)^{1/3})$$

We can do this using the inverse CDF method. We start by computing the cdf:

$$F_X(x) = \int_{-(3/2)^{1/3}}^{x} y^2 dy = \frac{y^3}{3}\Big|_{(-3/2)^{1/3}}^{x} = \frac{x^3}{3} + \frac{1}{2}$$

Then, we can solve for the inverse cdf using the definition:

$$F_X(F_X^{-1}(x)) = x$$

$$\frac{F_X^{-1}(x)^3}{3} + \frac{1}{2} = x$$

$$F_X^{-1}(x) = (3\Big(x - \frac{1}{2}\Big))^{\frac{1}{3}}$$

Then, we can generate random samples from this using the following algorithm:

- Generate $u \sim Unif(0, 1)$
- Set $X_i = (3\Big(u - \frac{1}{2}\Big))^{\frac{1}{3}}$
- Repeat until we have an adequate number of samples

Lastly, it is worth noting that the inverse CDF method does generate **exact** samples from the distribution when it can be used.

Monte Carlo Methods

The inverse CDF method has quite a few downfalls. Many distributions do not have an analytic expression for their cumulative density function, such as the normal, gamma, and beta. The CDF is also quite hard to generalize to higher dimensions, so without an analagous expression to the CDF in more than one dimension, we cannot use the inverse CDF method in more than one dimension.

For distributions where we cannot generate exact samples, we have to settle for approximations, using usually a technique called Monte Carlo. Broadly defined, Monte Carlo Methods are methods based on generating random samples for a set number (n) of iterations. Methods where samples are only generated until a condition is met (for example, generating samples from a normal distribution until we have a sample where exactly 50 samples have a value less than the mean) are called Las Vegas methods. Monte Carlo methods can be used to solve a wide variety of problems, including helping us to generate random samples from a distribution, which we will go over next class. Below is a simple example of a Monte Carlo method.

Example: The Dartboard Problem

Suppose that we have a dartboard, with a radius of 1/2 foot, enclosed in a 1x1 foot square. We are not very good at throwing darts, so our throws are essentially random in the x and y coordinates. What is the probability that a given throw lands on the board?

We know that if our throws are uniformally distributed, then we can just analytically calculate the probability as $\pi r^2 \approx .78$. This can be confirmed by the simulation below:

```
> x <- rep(NA, len)
> y <- rep(NA, len)
> in.board <- rep(0, len)
> for (i in 1:len){
+   x[i] <- runif(1)
+   y[i] <- runif(1)
+   if (((x[i] - .5)^2 + (y[i] - .5)^2 ) <= .25) in.board[i] <- 1
+ }
> mean(in.board)
[1] 0.7848
```

Now, what if we were fairly good at throwing darts? Maybe our dart throwing capacity actually follows a $Beta$ distribution, with $\alpha = \beta = 2$. Now, if we wanted to evaluate the exact probability that one of our darts hits the board, we would need to integrate a bivariate distribution that is the product of two distributions with no tractable integrals over a nontrivial constraint. This is just about impossible to do without numerical approximation in the first place, and Monte Carlo Methods are certainly the easiest way to attack this problem. We can modify our psuedo-code from above slightly to obtain an approximation,

```
> x <- rep(NA, len)
> y <- rep(NA, len)
> in.board <- rep(0, len)
> for (i in 1:len){
+   x[i] <- rbeta(1, 2, 2)
+   y[i] <- rbeta(1, 2, 2)
+   if (((x[i] - .5)^2 + (y[i] - .5)^2 ) <= .25) in.board[i] <- 1
+ }
> mean(in.board)
[1] 0.9532
```

At their root, these Monte Carlo methods are really just approximating integrals. Probabilities correspond to the total area/volume occupied over a space, which is exactly what we think of an integral as.

Lecture 19: The Metropolis Algorithm

Motivation

The 20th and 21st centuries have been filled with fantastic innovation in science, technology, engineering, and mathematics. We have evolved into a society that regularly uses computers, and behind the scenes all of these computers utilize a slew of source code and algorithms. Algorithms are massively important, and many have changed the computing game forever. We think about algorithms like the Fast Fourier Transform, quicksort, various matrix algorithms, and the simplex method. But what algorithm has been the most influential? The most influential algorithm would have to be one that changed the landscape completely, changing the field fundamentally. This algorithm is the Metropolis / Metropolis-Hastings algorithm, an algorithm that allows us to (approximately) produce random samples from **any** distribution. This has been particularly relevant in statistics, making Bayesian computing extremely relevant after work by Gelfand and Smith and Geman and Geman.

The Metropolis Algorithm

Imagine that we have a distribution, $f(x)$. We know the form of the function, but have no idea how to generate samples from this distribution. Now, suppose that we have a function $g(x)$ that lives in the same space as $f(x)$ (common support), and we DO know how to generate samples from $g(x)$. What if I told you that we could use $g(x)$ to help us generate random samples from $f(x)$? The idea seems preposterous at first, but that is **exactly** what the Metropolis Algorithm says. The formal algorithm is given below:

The Metropolis Algorithm
Suppose that we have $Y \sim f_Y(y)$ and $V \sim g_V(v)$, and f_Y and g_V have common support. Then, to generate $Y \sim f_Y$:

- Generate $V \sim f_V$. Set $Z_0 = V$.
- For some number of iterations, generate $U_i \sim Unif(0,1)$, $V_i \sim f_V$, and calculate

$$\rho_i = \min\left\{\frac{f_Y(V_i)g_V(Z_{i-1})}{g_V(V_i)f_Y(Z_{i-1})}, 1\right\}$$

- If $U_i \leq \rho_i$, set $Z_i = V_i$, otherwise set $Z_i = Z_{i-1}$

This algorithm won't produce exact samples from f_Y, but it is guaranteed to produce a convergent sequence of samples.

Hiding under these confusing equations and acceptance rules is a beautifully intuitive idea. We don't know how to generate samples from f, but we do know the value

of its density at any given point. We know that samples from f will be most dense in areas where it has high density. What we are doing is generating points to investigate using g. If we look at the form of ρ_i, we have

$$\rho_i = \min\left\{\frac{f_Y(V_i)}{f_Y(Z_{i-1})}\frac{g_V(Z_{i-1})}{g_V(V_i)}, 1\right\}$$

The first fraction is looking at the ratio of the density values of the point that we are proposing and the value that we are currently sitting at. If we have proposed moving to a point with higher density, this value will be greater than 1. The second fraction is helping us quantify how difficult it will be to move backwards from Z_{i-1} to V_i if we were to accept that point. We use a combination of these ratios to make a decision on whether or not we will move to the point that we have proposed. The distribution f is often referred to as the **target distribution**, while g is called the **proposal distribution**.The proof of the algorithm requires some Markov Chain theory, and thus we will omit it.

This is a very powerful and very general algorithm, but given that this is the last lecture of the semester, we will not be able to explore in great detail. We will look at a few common cases where we can apply the Metropolis algorithm.

Common Frameworks

Suppose that we have some distribution $f(x), x \in \mathbb{R}$. One of the most common proposal schemes that we use utilizes a normal distribution. The algorithm we will outline here is generally referred to as **Random Walk Metropolis**, or RWM. It has a lot of caveats: it helps to simplify the algorithm steps, it is easy to tune, and since we have worked with the normal distribution so much we have more intuition about our proposal.

Random Walk Metropolis

- Generate $Z_0 \sim N(0, \sigma^2)$
- Generate $U_i \sim Unif(0,1)$, $V_i \sim N(Z_{i-1}, \sigma^2)$, calculate

$$\rho_i = \min\left\{\frac{f_Y(V_i)}{f_Y(Z_{i-1})}, 1\right\}$$

- If $U_i \leq \rho_i$, set $Z_i = V_i$. Otherwise set $Z_i = Z_{i-1}$

We see that we are now making our acceptance rule solely based on the ratio of the density at our proposed point to our current point. If we propose a point with higher density, we are guaranteed to move there. If we propose a point with lower density,

we will still move there with some probability, despite it being a "worse" point than our current position.

These algorithms, while powerful, will require careful tuning.
Examples here