Notes to Mathematical Statistics

Matthew Cocci

Contents

Notation Throughout the notes, I will adhere to a few conventions for the sake of clarity. In particular, capital letters like Y will typically denote *random variables*, while lower case letters like y will denote realizations of random variables—observed data points. Almost always, θ will be used to denote a parameter or vector of parameters that we hope to estimate.

1. Estimation of Parameters

Most of statistics is about optimally estimating parameters using observed data. Specifically let's define a few terms/concepts:

Definition An **Estimator** is a function of random variables that maps to a random variable for the parameter. Specifically, we could describe an estimator as a function

$$\hat{\theta} : \{Y_1, \dots Y_n\} \to RV.$$

For example, if take want to guess the mean height of adult women and we observe n women, we might define the estimator: $\bar{Y} = (Y_1 + \cdots + Y_n)/n$. More generally, an *estimator* is denoted

$$\hat{\theta}(Y_1, \dots, Y_n).$$

- **Note** It's important to recognize that in this context, $\bar{Y}$ and $\hat{\theta}(Y_1, \dots, Y_n)$ are *random variables* with their own specific distributions, not particular, unique values.

Definition Next, we define an **estimate** to be a computed value, an actual number, that we get by plugging the observed data into the estimator function. For example, we measured womens height and plug that into the above function to get $\bar{y} = (y_1 + \cdots + y_n)/n$. More generally, an *estimate* is denoted

$$\hat{\theta}(y_1, \dots, y_n).$$

This is a specific unique value derived from observed data.

2. Parameter Bias

2.1. Definition

It appears that we would desire our estimators to be **unbiased**, which means that

$$\underbrace{\int \cdots \int}_{n \text{ times}} \hat{\theta}(y_1, \dots, y_n)\ f_{Y_1,\dots,Y_n}(y_1, \dots, y_n)\ dy_1 \dots dy_n = \theta$$

Note that $\int \cdots \int$ is called the *support*, and it ranges over the values for which the joint distribution, $f_{Y_1,\dots,Y_n}(\cdot)$ is positive.

2.2. Problems

There are reasons why the criterion of unbiasedness might not be a good idea:

1. The estimator might be crazy. For example, suppose that Y follows a Poisson distribution:
$$P(Y = y) = e^{-\theta}\frac{\theta^y}{\theta!}.$$
Now suppose that we want an unbiased estimator of $e^{-2\theta}$. Well, it turns out that the only unbiased estimator of $e^{-2\theta}$ is the function
$$g(Y) = \begin{cases} +1 & y \text{ even} \\ -1 & y \text{ odd} \end{cases}.$$
But this is clearly ridiculous, as $e^{-2\theta}$ will be between 0 and 1. So those choices are nuts.

2. There might not be *any* unbiased estimate of what you're trying to estimate.

3. Sometimes, the mean square error of a biased estimator may be less than the mean square error of *any* unbiased estimator, where we define the mean square error to be
$$\underbrace{\int \cdots \int}_{\text{the support}} \Big(\hat{\theta}(y_1, \ldots, y_n) - \theta\Big)^2 f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)\, dy_1 \ldots dy_n$$
$$= Var\left[\hat{\theta}(Y_1, \ldots, Y_n)\right] + [\text{Bias of } \hat{\theta}(Y_1, \ldots, Y_n)]^2$$
where the bias is defined to be
$$\text{Bias of } \hat{\theta}(Y_1, \ldots, Y_n) = [\text{Mean of } \hat{\theta}(Y_1, \ldots, Y_n)] - \theta.$$
So it's totally conceivable that the variance of the biased estimator would be very small, despite the bias.

2.3. Conclusion

Despite the critiques just given, asking for an unbiased is often not so problematic. In fact, we'll often try to find the *minimum-variance unbiased estimator*, or the MVU estimator of a given parameter, θ. The MVU estimator is defined to be that estimator of θ that has the smallest possible variance of the entire class of unbiased estimators.

But before we go about finding them, we will have to take a detour into the realm of order statistics, which have several nice properties when it comes to finding the MVU estimator.

3. Order Statistics

Order Statistics are part of a larger field called *non-parametric* statistics, or sometimes *distribution-free* statistics.

3.1. Definitions

Suppose that $Y_1, \ldots, Y_n$ are i.i.d RV's. From there, define

$$Y_{(1)} = \min\{Y_1, \ldots, Y_n\} = Y_{(\min)}$$

$$Y_{(n)} = \max\{Y_1, \ldots, Y_n\} = Y_{(\max)}$$

Similarly, if we rank-order all of them in a row, the ith order statistic is the ith in line and is denoted $Y_{(i)}$.

Important to note: even though the Y_i were all independent, the order statistics $Y_{(i)}$ will *not* be independent. In fact, the order statistics will *all* have different distributions than that of the original Y_i, and they will all differ from *each other* as well.

3.2. Density Functions of Order Statistics

Let's consider the ith order statistic and try to find its distribution, $f_{Y_{(i)}}\left(y_{(i)}\right)$. To do so, we will want to formulate two equivalent expressions, then solve them out for the pdf. So first,

$$P\left(y \leq Y_{(i)} \leq y + \delta y\right) = f_{Y_{(i)}}(y)\,\delta y + O(\delta y)^2 \tag{1}$$

Equation 1 is relatively clear, once we define $O(\delta y)^2$ to be an error term of order $(\delta y)^2$. Now for the second formulation, we'll recall the cdf of Y_i

$$F_Y(y) = \int_{-\infty}^{t} f_Y(y)\,dy = P(Y \leq y) \tag{2}$$

In order for the ith order statistic to be in the bucket $[y, y+\delta y]$, you need $i-1$ observations before or at y, and $n-i$ after or at $y+\delta y$. This corresponds to the probability

$$\begin{aligned} P\left(y \leq Y_{(i)} \leq y + \delta y\right) =& \frac{n!}{(i-1)!1!(n-i)!}\left[(F_Y(y))^{i-1}\right] \\ & \times [f_Y(y)\delta y]\left[(1 - F_Y(y+\delta y))^{n-i}\right] + O(\delta y)^2 \end{aligned} \tag{3}$$

Now that we have two equivalent expressions for the relevant probability, we can equate Expressions 1 and 3, cancel out the errors, divide both sides by δy, then take the limit as $\delta y \to 0$ to get

$$f_{Y_{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!}\left[(F_Y(y))^{i-1}\right][f_Y(y)]\left[(1 - F_Y(y))^{n-i}\right] \tag{4}$$

3.3. Range

Next, a very useful Random Variable to know would be the *range* defined

$$R = Y_{(n)} - Y_{(1)}$$

The expectation is rather straightforward:

$$ER = E\left[Y_{(n)}\right] - E\left[Y_{(1)}\right].$$

The variance, however, is a little more complicated. In order to compute the variance, we'll need the joint density function of $Y_{(1)}$ and $Y_{(n)}$. Using an direct generalization and extension of the method in which we found the pdf of the ith order statistic, it can be show that

$$\begin{aligned} f_{Y_{(i)},Y_{(j)}}(u,v) = &\frac{n!}{(i-1)!(j-i-1)!(n-j)!}\left[\left(F_Y(u)\right)^{i-1}\right] f_Y(u) \\ &\times \left[\left(F_Y(v) - F_Y(u)\right)^{j-i-1}\right] f_Y(v) \left[\left(1 - F_Y(v)\right)^{n-j}\right] \end{aligned} \tag{5}$$

where $u = y_{(i)}$ and $v = y_{(j)}$, $i < j$. This simplifies considerably if we restrict ourselves to the case where $i = 1$ and $j = n$:

$$f_{Y_{(1)},Y_{(n)}}(u,v) = n(n-1)\left[f_y(u,v)\left(F_Y(v) - F_Y(u)\right)^{n-2} f_Y(v)\right]$$

where, in this case, $u = y_{(1)}$ and $v = y_{(n)}$.

From there, we can try to get at things like $Var(R)$.

4. The Cramer-Rao Inequality

In general, we hope to find minimum-variance unbiased (MVU) estimators. One approach is the Cramer-Rao inequality approach. But before we get there, let's go through some notation.

4.1. Notation

We'll be consider n random variables $(Y_1, \ldots, Y_n)$ which don't necessarily have to be iid. Together, they have joint density function

$$f = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta)$$

which depends on parameter θ. Note that f is just shorthand for the expression above.

Second, we let S represent the *support* of f, where the support is the set of all values such that $f > 0$.

Next, we let $\tau(\theta)$ be *any* arbitrary function of θ that we wish to estimate. The function $\tau(\theta)$ could be θ, $2\theta^2$, etc.

Finally, we specify a function that takes as its arguments the data, then returns an unbiased estimate for $\tau(\theta)$. This is written

$$\hat{\tau} = \hat{\tau}(Y_1, \ldots, Y_n)$$

where $\hat{\tau}$ represents any arbitrary *unbiased* estimator of $\tau(\theta)$. [1]

4.2. Statement of the Cramer-Rao Inequality

We now come to an important result in estimation, the Cramer-Rao Inequality. Note that this is a *very* general result, as it does not assume independence or identical distribution of the different Y_i that we use to estimate a certain parameter.

The *Cramer-Rao Inequality* states that, provided the support S does *not* depend upon the parameter θ,

$$Var(\hat{\tau}) \geq \frac{-\left[\frac{d\tau(\theta)}{d\theta}\right]^2}{E\left[\frac{d^2(\ln f)}{d\theta^2}\right]} \tag{6}$$

for any choice of $\hat{\tau}$.[2] Now why is this useful? Well, if we happen to know an unbiased estimator of $\tau(\theta)$ whose variance equals the RHS of Equation 6, then it is in fact *the* MVU estimator of $\tau(\theta)$.

[1] Note, there are possibly infinitely many of them.

[2] The proof is given in the appendix.

4.3. Special Cases

Case 1: Suppose the Y_i are, in fact, iid. Then we have that

$$f = f_{Y_1}(y_1;\theta) \times \cdots f_{Y_n}(y_n;\theta).$$

From there, it can be show that the expectation in the denominator simplifies to give us a special case of the Cramer-Rao inequality:

$$Var(\hat{\tau}) \geq \frac{-\left[\frac{d\tau(\theta)}{d\theta}\right]^2}{n \cdot E\left[\frac{d^2(\ln f_{Y_i}(y_i))}{d\theta^2}\right]} \tag{7}$$

Case 2: Next, suppose we're trying to estimate simply the parameter θ so that $\tau(\theta) = \theta$. Then the Cramer-Rao inequality simplifies to

$$Var(\hat{\tau}) \geq \frac{-1}{E\left[\frac{d^2(\ln f)}{d\theta^2}\right]} \tag{8}$$

4.4. Limitations

While useful, the Cramer-Rao Inequality may be too restrictive in some cases. For instance,

1. We might be able to think of an unbiased estimator whose variance achieves the CR bound.
2. There might not even exist and unbiased estimator that can achieve the CR bound.
3. The CR bound only applies if the support does *not* depend upon θ.

4.5. When is the Cramer-Rao Bound Achievable?

A natural question to ask is "When exactly can we hit the Cramer-Rao bound?" This requires a little thought. So first, we go back, to where we started in the proof to get the inequality in the first place:

$$0 \leq (Corr(W,V))^2 \leq 1.$$

When is this inequality an equality on the upper end? Well, whenever there is perfect correlation—when one is a linear function of the other:

$$V = a + bW.$$

Now that we know the conditions, recall what we took to be W and V when we plugged into the above inequality to the CR inequality:

$$W = \hat{\tau}, \qquad V = \frac{1}{f}\frac{df}{d\theta}$$

Since one has to be a linear function of the other, this implies

$$\frac{1}{f}\frac{df}{d\theta} = A(\theta)+B(\theta)\hat{\tau}(Y_1,\ldots,Y_n) \quad \Rightarrow \quad \ln f(\theta) = C(\theta)+D(\theta)\hat{\tau}(Y_1,\ldots,Y_n)+g(y_1,\ldots,y_2)$$

$$\Rightarrow f(\theta) = h(y_1,\ldots,y_n)e^{C(\theta)+D(\theta)\hat{\tau}} \tag{9}$$

where $C(\theta) = \int A(\theta)\, d\theta$ and $D(\theta) = \int B(\theta)\, d\theta$.

Now, we just need to note one more thing. Assume that the Cramer-Rao bound is, in fact, achievable. Again, this means

$$\frac{1}{f}\frac{df}{d\theta} = A(\theta) + B(\theta)\hat{\tau}.$$

Now let's take expectations (with respect to the random variables, not θ) throughout:

$$E\left[\frac{1}{f}\frac{df}{d\theta}\right] = E\left[A(\theta) + B(\theta)\hat{\tau}\right] \tag{10}$$

On the LHS, everything simplifies to[3]

$$\begin{aligned} E\left[\frac{1}{f}\frac{df}{d\theta}\right] &= \int\cdots\int_S \left[\frac{1}{f}\frac{df}{d\theta}\right] f\, d\mathbf{y} = \int\cdots\int_S \frac{df}{d\theta}\, d\mathbf{y} \\ &= \frac{d}{d\theta}\left[\int\cdots\int_S f\, d\mathbf{y}\right] = \frac{d}{d\theta}\left[1\right] \\ &= 0 \end{aligned}$$

Turning to the RHS of Equation 10, we use the fact that $\hat{\tau}$ is an *unbiased* estimator of $\tau(\theta)$ to write

$$\begin{aligned} E\left[A(\theta) + B(\theta)\hat{\tau}\right] &= E[A(\theta)] + E\left[B(\theta)\hat{\tau}(\theta)\right] \\ &= A(\theta) + B(\theta)\tau(\theta) \end{aligned}$$

Putting the LHS and RHS together and solving for $\tau(\theta)$, we get

$$\tau(\theta) = -\frac{A(\theta)}{B(\theta)} \tag{11}$$

[3] Note, we could interchange the differential and the integral signs because S does not depend upon θ. Also, k just stood for some unknown constant that doesn't depend upon θ.

Conclusion If we consider everything in this section, we now know that *if and only if* we can write the jdf f in the form of Equation 9,[4]

$$\Rightarrow f(\theta) = h(y_1, \ldots, y_n)e^{C(\theta)+D(\theta)\hat{\tau}},$$

then the Cramer Rao bound will be achievable by *some* $\hat{\tau}$.[5] That we can just read off of the equation just written. But that begs the question: "Just what $\tau(\theta)$ admits Cramer-Rao bound estimation?" Well that we just also saw in equation 11:

$$\tau(\theta) = -\frac{A(\theta)}{B(\theta)}$$

$$A(\theta) = \frac{d}{d\theta}\left[C(\theta)\right] \qquad B(\theta) = \frac{d}{d\theta}\left[D(\theta)\right]$$

Finally, it can easily be show that if we assume that we can achieve the Cramer-Rao bound with some estimator $\hat{\tau}$, then in the notation we used in this section

$$Var(\hat{\tau}) = -\frac{1}{B(\theta)}\frac{d}{d\theta}\left[\frac{A(\theta)}{B(\theta)}\right]$$

Limitations Of course, we should consider when this approach of identification and computation is of limited use.

1. First, there might not be a way to write the jdf in the desired exponential form, so a function of θ might not exist, as we saw with the Cauchy distribution.

2. The function of θ that we get, $\tau(\theta)$, might be really weird and not all that interesting in practice. So you find a function $\tau(\theta)$ that you *can* find an unbiased estimator for, but that doesn't mean you want to or should.

[4] And that's a big "if" which is by no means automatic.

[5] This also means that if $\tau(\theta)$ admits Cramer-Rao bound estimation, then so does any linear function of it.

5. Sufficient Statistics and the Rao-Blackwell Theorem

This is the next approach to MVU Estimation, which uses sufficient statistics and their optimality properties. This approach is much more powerful than Cramer-Rao, especially because it doesn't require that the support be independent of θ.

5.1. Definition of Sufficient Statistics

Suppose that we have n random variables: $Y_1, \ldots, Y_n$. Together, they have a joint density function $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta)$. Then W is a *sufficient statistic* for θ if and only if the conditional distribution of $Y_1, \ldots, Y_n$ given $W(Y_1, \ldots, Y_n)$ is independent of θ. Mathematically, this is expressed

$$\begin{aligned} & f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n | w; \theta) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n | w) \\ \Leftrightarrow \quad & \frac{f_{Y_1,\ldots,Y_n,W}(y_1, \ldots, y_n, w; \theta)}{f_W(w; \theta)} = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n | w) \end{aligned} \tag{12}$$

Note that often when we compute the numerator in Equation 12, the W term will give you no additional information over and above the $Y_1, \ldots, Y_n$ individually. In that case, the numerator reduces the problem to checking that

$$\frac{f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta)}{f_W(w; \theta)} = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n | w) \tag{13}$$

If it turns out that the conditional distribution of

5.2. Properties and Intuition

Vector Sufficient Statistics: Note that we can generalize so that a sufficient statistic is a vector, such as $(W_1, W_2, \ldots, W_k)$.

Uniqueness: Sufficient statistics aren't necessarily unique. For example, suppose W_1 is a sufficient statistic. Then the vector (W_1, W_2) will be too for any choice of W_2 since W_1 is sufficient on its own. Also, it's immediately clear that the vectors $(Y_1, \ldots, Y_n)$ and $(Y_{(1)}, \ldots, Y_{(n)})$ will also be sufficient statistics since they encapsulate all of the data trivially.

Intuition: Suppose there are two people, some data $(y_1, \ldots, y_n)$, and some function of the data $W = W(Y_1, \ldots, Y_n)$. Now suppose that I give Person 1 all of the n data points. Next suppose I give Person 2 just $W(y_1, \ldots, y_n)$. If the information I gave Person 1 *equals* the information I gave Person 2 from the standpoint of information about θ, then W is a sufficient statistic for θ.

5.3. Rao-Blackwell Theorem, Part 1

Suppose that a sufficient statistic, W, for θ exists. Let $\tau(\theta)$ be any function of θ admitting unbiased estimation, i.e. for which an unbiased estimator, $\hat{\tau}$, of $\tau(\theta)$ exists. Then a *unique*, *minimum variance* unbiased estimator of $\tau(\theta)$ is some function of W.

So if we find a sufficient statistic, W, and can fiddle around to get some function of W that does in fact have an expectation of θ, thank the gods of statistics, because you just found *the* unique MVU estimator of θ.

Improvement over Cramer-Rao: The Rao-Blackwell Theorem allows us to estimate parameters $\tau'(\theta)$ that are *not* linear functions of whatever $\tau(\theta)$ that the Cramer-Rao approach indicates admits CR bound estimation. It's much more general. Moreover, it works *whether or not* the support depends upon θ.

5.4. Minimal, Non-Trivial Sufficient Statistics (MNTSS)

We have a few requirements for finding really good, powerful sufficient statistics.

1. We want them to be *minimal*, in that we compress information as much as possible. So if both W_1 and (W_1, W_2) are sufficient statistics, don't use the latter. It doesn't compress information enough and is non-minimal.

2. We want our sufficient statistics to be non-trivial. So the vector $(W_1, \ldots, W_k)$ should be such that $k < n$ so we achieve *some* compression.

Note that all of this implies that any *scalar* SS must be non-minimal and non-trivial.

We want to know whether, in any given situation, a scalar MNTSS exists, and, if there is one, how can we find it? To do so, there are two situations we must consider separately:

- The support S does *not* depend upon θ.

- The support S *does* depend upon θ.

Both situations unfortunately require very different approaches, so we'll handle each in turn.

5.5. Finding a MNTSS, S Independent of θ

Here, we document three cases to find an MNTSS whenever S is independent of θ: the Factorization approach, the Smith-Jones approach, and the Exponential Family approach.

5.5.1. Factorization Approach

W is a scalar SS for θ if and only if

$$\begin{aligned} \frac{f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta)}{f_W(w;\theta)} &= k(y_1,\ldots,y_n) \\ \Rightarrow f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta) &= k(y_1,\ldots,y_n) f_W(w;\theta) \end{aligned} \tag{14}$$

where $k(y_1,\ldots,y_n)$ is some constant function of the data that is independent of θ and where $f_W(w;\theta)$ is some function of θ and the sufficient statistic, W, only.

5.5.2. Smith-Jones Approach

Again, we hope to find a scalar sufficient statistic, which would automatically be a MNTSS. Now suppose we have two statisticians, Smith and Jones.

Person	Has observed values for $Y_1,\ldots,Y_n$ of
Smith	$y_{11}, y_{12}, \ldots, y_{1n}$
Jones	$y_{21}, y_{22}, \ldots, y_{2n}$

Next we define the ratio R as the ratio of the probability of Smith's data to that of Jones' data:

$$R = \frac{f(y_{11}, y_{12}, \ldots, y_{1n};\theta)}{f(y_{21}, y_{22}, \ldots, y_{2n};\theta)}$$

Then the necessary and sufficient condition that the function $g(Y_1,\ldots,Y_n)$ be a scalar sufficient statistic for θ is that that the ratio R is independent of θ *if and only if*

$$g(Y_{11}, Y_{12}, \ldots, Y_{1n}) = g(Y_{21}, Y_{22}, \ldots, Y_{2n})$$

In practice, this means that we form the ratio, R. Then we try to find the function $g(Y_1,\ldots,Y_n)$ such that R has no θ term in it whenever the sufficient statistics match for both Smith and Jones.[6]

[6] This follows from the factorization approach. Because see by Equation 14 that if we form the ratio of the two likelihoods, the $f_W(w;\theta)$ will cancel out, removing all occurrences of θ with it, provided

$$g(Y_{11}, Y_{12}, \ldots, Y_{1n}) = g(Y_{21}, Y_{22}, \ldots, Y_{2n}).$$

All that will be left is a ratio of two constants that come from the data.

Generalization to θ as a Vector: Suppose $\theta = (\theta_1, \ldots, \theta_k)$. Form R as above. Now suppose that $g_i(Y_1, \ldots, Y_n)$ for $i = 1, \ldots, s$ are s linearly independent functions of $Y_1, \ldots, Y_n$. Then the necessary and sufficient condition that the vector

$$\begin{pmatrix} g_1(Y_1, \ldots, Y_n) & g_2(Y_1, \ldots, Y_n) & \ldots & g_s(Y_1, \ldots, Y_n) \end{pmatrix} \tag{15}$$

is a non-trivial sufficient statistic for θ is that $s < n$ and also that R is independent of θ *if an only if* the following s conditions hold:

$$g_i(Y_{11}, \ldots, Y_{1n}) = g_i(Y_{21}, \ldots, Y_{2n}), \qquad i = 1, \ldots, s \tag{16}$$

Now that we have a non-trivial sufficient statistic, we can ask "Is it minimal?" To that question, the vector in expression 15 is minimal if there does not exist another vector of length $t < s$ such that R is independent of θ if and only if the analogous conditions in Equations 16 hold for the new vector.

5.5.3. Exponential Family Approach

Suppose that the jdf can be written as[7]

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = h(y_1, \ldots, y_n) e^{C(\theta) + D(\theta)\hat{\theta}}.$$

Then $\hat{\theta}$ is a scalar, sufficient statistic for θ.

Generalization to θ as a Vector: Suppose we can write the jdf (in abbreviated vector notation) as

$$f_{\mathbf{Y}}(\mathbf{y}; \Theta) = h(\mathbf{y}) \cdot \exp\left\{ C(\Theta) + \sum_{j=1}^{s} D_j(\Theta) g_j(\mathbf{y}) \right\}, \qquad s < n$$

$$\mathbf{Y} = (Y_1, \ldots, Y_n), \qquad \mathbf{y} = (y_1, \ldots, y_n), \qquad \Theta = (\theta_1, \ldots, \theta_k)$$

where the functions $D_1, \ldots, D_s$ are linearly independent functions of Θ. Then

$$\begin{pmatrix} g_1(Y_1, \ldots, Y_n) & g_2(Y_1, \ldots, Y_n) & \ldots & g_s(Y_1, \ldots, Y_n) \end{pmatrix}$$

is a MNTSS for $\Theta = (\theta_1, \ldots, \theta_k)$.

[7] This should look familiar to the Cramer-Rao approach.

5.6. Finding a MNTSS, S Dependent Upon θ

In this case, we will only consider cases where $Y_1, \ldots, Y_n$ are iid due to the complex nature of the relationship between S and θ.

Support of the Form $[a, b(\theta)]$ Suppose that $f_Y(y;\theta)$ factorizes into

$$f_Y(y;\theta) = g(y)h(\theta).$$

Then $Y_{(n)}$ is a MNTSS for θ. Intuitively, if you know that $Y_{(n)}$ is the largest value of your observations, nothing else tells you *anything* extra about the upper bound, $b(\theta)$. Rao-Blackwell then ensures that your resulting estimator from the MNTSS is unique, provided it is unbiased.

Support of the form $[a(\theta), b(\theta)]$ For this case, we deal with two concrete examples:

1. Suppose $f_Y(y) = 1$ and $\theta < y < \theta + 1$, then your MNTSS is going to be the vector

$$\begin{pmatrix} Y_{(1)} & Y_{(n)} \end{pmatrix}.$$

   Note, this is interesting because you're trying to estimate *one* parameter, but your MNTSS is a *two*-dimensional vector.

2. Next, suppose that $f_Y(y) = 1/2\theta$ and that $-\theta < y < \theta$. Then the MNTSS will be

$$\max\{-Y_{(1)}, Y_{(n)}\}$$

5.7. Full Rao-Blackwell Theorem

The full theorem has two parts. We already saw part 1, but we'll it repeat here for completeness. So suppose we want to estimate $\tau(\theta)$ and some sufficient statistic, which we denote W, exists for θ. Then

1. The unique MVU estimator of $\tau(\theta)$ is some function of W.

2. Let $\hat{\tau} = \hat{\tau}(Y_1, \ldots, Y_n)$ be *any* unbiased estimator of $\tau(\theta)$. Then the *unique* MVU estimator of $\tau(\theta)$ is guaranteed (!) to be

$$E\left[\hat{\tau}(Y_1, \ldots, Y_n)|W\right]$$

   That's really powerful and pretty awesome.

Applying the Rao-Blackwell Theorem: The two different components give us two different approaches to find a unique MVU estimate. Here they are numbered by the part of the RB Theorem to which they apply:

1. Find the SS. Manipulate and fiddle until you get some $\hat{\tau}$ such that $E[\hat{\tau}] = \tau(\theta)$.

2. Find *any* old unbiased estimator $\hat{\tau}$. And it could be crazy![8] Like one you'd *never* think of using to estimate $\tau(\theta)$ in a million years. But just so long as its expectation is $\tau(\theta)$, RB Part 2 tells you it will work (and what to do!).

[8] See the appendix for one example.

5.8. Potential Problems

As always, no method is perfect, and there are a few issues with the sufficient statistic and Rao-Blackwell approach to finding MVU estimators:

1. There might not *be* an MNTSS to search for.
2. We have to find an unbiased estimator of $\tau(\theta)$ if we want to use Rao-Blackwell, Part 2. However, there might not *be* such an unbiased estimator.
3. Even if there is an unbiased estimator of $\tau(\theta)$, we might not be smart enough to find it.

6. Maximum Likelihood Estimation

Broadly speaking, maximum likelihood approach is a practical approach that overcomes the difficulties associated with sufficient statistics—the very same difficulties we outlined above. It is also an improvement over the method of moments, which was popular in statistics prior to MLE methods.

6.1. Finding Maximum Likelihood Estimators

Our main goal will be to find maximum likelihood estimators: functions of random variables that can be use to estimate certain parameters. So to start, suppose we have n random variables (not necessarily iid) with some jdf:

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta) \tag{17}$$

which is dependent upon some fixed, but unknown parameter, θ.

Now comes the big insight: turn that logic around. Rather than think of the jdf as a function of $y_1,\ldots,y_n$ with θ fixed, assume the $y_1,\ldots,y_n$ are fixed and θ is a variable in what we rename the *likelihood function*. This will have the same functional form as the jdf in Equation 17, but we change our idea of what's fixed. This gives us

$$L(\theta;y_1,\ldots,y_n) \quad \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta;y_1,\ldots,y_n) \tag{18}$$

Note, that while this works, we'll often apply the same method to the *log* of the likelihood function, which turns the product in the jdf into a sum. This is often more algebraically and computationally tractable, so we often solve

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell(\theta;y_1,\ldots,y_n), \qquad \ell = \ln L(\theta;y_1,\ldots,y_n)$$

Often, we'll find the maximum likelihood estimator by taking the derivative, setting the result equal to zero, and solving for the parameter. However, sometimes we'll have a *boundary maximum*, most often when the support depends upon the parameter.[9]

Multiparameter Case: Finally, we can generalize to allow for multiple parameters:

$$f_{Y_1,\ldots Y_n}(y_1,\ldots,y_n;\theta_1,\ldots,\theta_k) \Rightarrow L(\theta_1,\ldots,\theta_k;y_1,\ldots,y_n)$$

nn which case, we take partial derivatives, set them equal to zero, and solve for $\theta_1,\ldots,\theta_k$.

Constraints: In the multiparameter case particularly, we will sometimes want to allow for constraints, which can be easily implemented by a Lagrange multiplier technique when we do the maximization.

[9] Note that MLEs aren't guaranteed to biased at all. But often they will be, and when they aren't they're usually pretty close and do well enough.

6.2. Properties of Maximum Likelihood Estimators

Maximum likelihood estimators have some very nice properties, which is why we like to use them in the first place.[10] Let's review those properties now:

1. *Invariance Property*: Suppose that $\tau = \tau(\theta)$ is a monotonically increasing or decreasing function of θ. Then

$$\hat{\tau}_{\text{MLE}} = \tau\left(\theta_{\text{MLE}}\right)$$

   where $\hat{\tau}_{\text{MLE}}$ is the MLE for $\tau(\theta)$ and θ_{MLE} is the MLE for θ.

2. *Relation to MNTSS's*: If a scalar MNTSS exists, then it is a function of the maximum likelihood estimator.

   Proof. Here's an informal sketch. We know by the factorization criterion that the jdf, and hence the likelihood, will look like

$$f = L = g(y_1, \ldots, y_n)h(\hat{\theta}_{\text{MNTSS}}, \theta)$$

   where g is independent of θ. Well, to max L with respect to θ, we will be maxing h with respect to θ, which implies that θ is some function of the MNTSS. □

3. *Optimality Properties of MLEs*: If we let $\tau(\theta)$ be any function of θ and let $\hat{\tau}_{\text{MLE}}$ be the maximum likelihood estimator of $\tau(\theta)$, then under the regularity conditions assumed in the Cramer-Rao bound section,

$$\lim_{n\to\infty} E[\hat{\tau}_{\text{MLE}}] = \tau(\theta)$$

$$\lim_{n\to\infty} \frac{Var(\hat{\tau}_{\text{MLE}})}{\text{Cramer-Rao bound for } \tau(\theta)} \to 1$$

   while the distribution of $\hat{\tau}_{\text{MLE}}$ approaches the normal distribution as $n \to \infty$. This is really powerful stuff. So in words, what this means is

   a) Asymptotically unbiased.

   b) Asymptotically normal.

   c) Asymptotically has variance approaching the CR bound.

[10] When I use the shorthand "MLE," I'm explicitly talking about maximum likelihood estimat**ors**, which are functions of random variables, and more general that estimat**es**, which are simply numbers computed from the data.

6.3. Application: Simple Linear Regression

Simple linear regression provides a great testing ground to apply some of the techniques described above. So let's assume we have n random variables Y_i that are *independently*, but *not* identically distributed as follows

$$Y_i \sim N(\alpha + \beta x_i,\ \sigma^2)$$

If we form the likelihood and solve for our three unknowns, we get MLEs of

$$\hat{\beta} = \frac{\sum Y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \qquad \hat{\sigma}^2 = \frac{\sum (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n}$$

Our estimators of α and β are both unbiased; however, the estimator of σ^2 is biased a bit.

Strategy in Selecting Covariates: We know that our estimator of β is unbiased, but what about the variance of $\hat{\beta}$? It's not hard to see from the formula for $\hat{\beta}$ that, since the x_i terms are known constants,

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

This suggests that to *minimize* the variance of our estimator, we want to *maximize* the variance of our covariates.[11]

[11] However, we shouldn't do this blindly by clustering our observations at two extreme x_i values. This wouldn't allow us to see whether the relationship was linear, quadratic, or some other higher order relationship. Therefore, we should have values along the spectrum of possible x_i values, but we should, at the same time, try to make that spectrum of possible values as wide as we can.

7. Transformation Theory

Throughout this section, we'll only consider continuous random variables. So let's start with n random variables, $X_1, \ldots, X_n$, (not necessarily iid) which have jdf

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

Now let's suppose that we have Y, which is some function

$$Y = g(X_1, \ldots, X_n)$$

Let's review some reasons why we might want to find such a function, Y:

1. Often, an MLE—call it Y—will be a function of the X_i. To know the properties of Y, we have to know it's distribution.
2. Many test statistics are, in fact, *derived* statistics with parameters dependent on the variables X_i.

7.1. Univariate Case, Monotonic

Throughout, we'll assume that X has a known density function, $f_X(x)$. If we define $Y = g(X)$, we want to find $f_Y(y)$.

Monotonically Increasing/Decreasing: We can group these two cases together to get two equivalent formulations,

$$\begin{aligned} f_Y(y) &= \left[f_X(x) \left| \frac{dx}{dy} \right| \right]_{x=g^{-1}(y)} \\ &= \left[f_X(x) / \left| \frac{dy}{dx} \right| \right]_{x=g^{-1}(y)} \end{aligned}$$

7.2. Univariate Case, Non-Monotonic

Suppose that we have some RV, X, with distribution $f_X(x)$. We now want the distribution for $Y = g(X)$, where g is some non-monotonic function of X. Then

$$f_Y(y) = \sum_i \left[f_X(x_i) \left| \frac{dx_i}{dy} \right| \right]_{x_i = g_i^{-1}(y)} \tag{19}$$

where each $g_i^{-1}(y)$ is a different functional inverse of y. Note that the range of i in the sum can vary depending upon how non-monotonic the function is and how many functional inverses exist.[12]

[12] Non-monotonic transformations allow for discontinuous probability density functions.

7.3. Multivariate Case

Suppose that we have $X_1, \ldots, X_n$ with jdf $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$. Now let

$$\begin{aligned} Y_1 &= Y_1(X_1, \ldots, X_n) \\ Y_2 &= Y_2(X_1, \ldots, X_n) \\ \vdots &= \quad \vdots \\ Y_n &= Y_n(X_1, \ldots, X_n) \end{aligned}$$

We assume that the functions are 1-to-1, akin to the monotonicity assumption in the univariate case. Now we want to know the jdf, $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$. This we compute

$$\begin{aligned} f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) &= [f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)|J|]_{X_i} \qquad X_i = X_i(Y_1, \ldots, Y_n) \\ &= \left[f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \frac{1}{|J^*|} \right]_{X_i} \end{aligned}$$

where the J and J^* terms are the determinants of the Jacobian matrices

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix} \qquad J^* = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix} \tag{20}$$

Often, we will even use this to do transformations that are not 1-to-1. That will involve doing a transformation that *is* 1-to-1 using this method with dummy variables, then integrating out those dummy variables.

7.4. Applications of Univariate Case

7.4.1. Simulations via Uniform Distribution

Suppose we want to draw from a distribution, $f_X(x)$ for the variable X. Now let's consider the random variable, where we suppose the integral has an analytical solution:

$$Y = \int_a^x f_X(t)\, dt = h(X) \tag{21}$$

We can apply our transformation rule to get the distribution of Y:

$$\begin{aligned} f_Y(y) &= \left[f_X(x) / \left| \frac{dy}{dx} \right| \right] = \left[f_X(x) / \frac{d}{dx} \left(\int_a^x f_X(t)\, dt \right) \right] \\ &= [f_X(x)/f_X(x)] = 1 \end{aligned}$$

In other words, the density function of Y is the *uniform distribution*. This means that we can draw from a uniform distribution, compute the integral in Equation 21, invert, and set

$$Y = h^{-1}(Y)$$

in order to obtain a draw from X. This allows us to draw from any arbitrary distribution, so long as we can compute the integral in Equation 21 and invert the resulting function.

7.4.2. Square of a Normal Random Variable

Suppose the $X \sim \text{NID}(0,1)$, and $Y = X^2$. Clearly, we're in the non-monotonic case, so let's find the distribution of Y with that in mind:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} \left[e^{-\frac{1}{2}x_1^2} / \left| \frac{dy}{dx_1} \right| \right]_{x_1=\sqrt{y}} + \frac{1}{\sqrt{2\pi}} \left[e^{-\frac{1}{2}x_1^2} / \left| \frac{dy}{dx_2} \right| \right]_{x_2=-\sqrt{y}} \\ &= \frac{2}{\sqrt{2\pi}} \left[e^{-\frac{1}{2}y} / 2\sqrt{y} \right] \\ &= \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{1}{2}y} \end{aligned}$$

which is a χ_1^2 distribution.[13]

[13] Of course, you have to be able to simulate a normal random variable to simulate a χ_1^2 random variable, and that's a bit tricky. But we'll do that below.

7.5. Applications of Multivariate Case

In this case, we'll have to use Jacobian matrices and also dummy variables to generate random variables and compute the distributions for derived statistics.

7.5.1. Generating a N(0,1) Random Variable

We can't actually generate a *single* N(0,1) RV, but we can generate two and throw one out, so let's do that. Assume $X_1, X_2 \sim \text{NID}(0,1)$:

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)}$$

Now to do our change of variables, let's define

$$\begin{aligned} X_1 &= R\cos\theta, \qquad R > 0 \\ X_2 &= R\sin\theta, \qquad \theta \in [0, 2\pi) \end{aligned}$$

If we take $R = Y_1$ and $\theta = Y_2$, then we form the Jacobian:

$$\begin{aligned} J &= \det \begin{bmatrix} \frac{\partial X_1}{\partial R} & \frac{\partial X_1}{\partial \theta} \\ \frac{\partial X_2}{\partial R} & \frac{\partial X_2}{\partial \theta} \end{bmatrix} = \det \begin{bmatrix} \cos\theta & -R\sin\theta \\ \sin\theta & R\cos\theta \end{bmatrix} \\ &= R(\cos^2\theta + \sin^2\theta) = R \end{aligned} \tag{22}$$

Now, applying the multivariate transformation above, we get our joint density and distribution functions under the transformed variables:

$$f_{R,\theta}(r,\theta) = \frac{1}{2\pi} r e^{-\frac{1}{2}r^2}$$

$$F_R(r) = \int_0^R \int_0^{2\pi} \frac{1}{2\pi} r e^{-\frac{1}{2}r^2} \, d\theta \, dr = 1 - e^{-\frac{1}{2}R^2}$$

$$F_\theta(\theta) = \int_0^\theta \int_0^\infty \frac{1}{2\pi} r e^{-\frac{1}{2}r^2} \, dr \, d\theta = \frac{\theta}{2\pi}$$

To get normal random variables, we draw U_1 and U_2 from the Uniform(0,1) distribution, set

$$U_1 = 1 - e^{-\frac{1}{2}R^2} \quad \Rightarrow \quad R = \sqrt{-2\ln(1-U_1)}$$

$$U_2 = \frac{\theta}{2\pi} \quad \Rightarrow \quad \theta = 2\pi U_2$$

$$X_1 = \sqrt{-2\ln(1-U_1)}\cos(2\pi U_2), \qquad X_2 = \sqrt{-2\ln(1-U_1)}\sin(2\pi U_2)$$

Which is how you generate two standard normal random variables.

7.5.2. Student's t-Distribution

In order to generate a variable that has t distribution with $n-1$ degrees of freedom, we start by supposing that $Y_i \sim \text{NID}(\mu, \sigma^2)$. From there, we define

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$= \frac{1}{n-1}\left(Y_1^2 + \cdots + Y_n^2 - n\bar{Y}^2\right)$$

The statistics $\bar{Y}$ and S^2 are useful because they are both unbiased estimators of μ and σ^2, respectively. And before we get to the derivation, we make one final note: *Cochran's Theorem* says that $\bar{Y}$ and S^2 are independent for the normal distribution.[14]

Now let's define the statistic, T, whose distribution we hope to find:

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \qquad Y_i \sim \text{NID}(\mu, \sigma^2)$$

But after a bit of rearranging and variable definitions, we get

$$T = \frac{(\bar{Y} - \mu)\sqrt{n}}{S} = \frac{\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}\cdot\frac{n-1}{n-1}}}$$
$$= \frac{W_1}{\sqrt{W_2/(n-1)}}$$

Now it's clear by how we defined W_1 and W_2 that W_1 has a N(0,1) distribution, while W_2 has a χ^2 distribution with $n-1$ degrees of freedom. Since W_1 and W_2 are *independent* by Cochran's theorem, the jdf of W_1, W_2 is

$$f_{W_1,W_2}(w_1, w_2) = \left(\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}w_1^2}\right)\left(\frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}\, w_2^{\frac{n-1}{2}-1} e^{-\frac{1}{2}w_2}\right)$$

Now we just hope to find the jdf of T. Since transformations must be 1-to-1 in the number of variables, we'll choose a dummy variable, V, and define to simplify the mess of fractions in the jdf:

$$V = W_2, \qquad k = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}$$

This leads us to from the Jacobian matrix to go from W_1 and W_2 to T and V:

$$J^* = \det\begin{bmatrix} \frac{\partial T}{\partial W_1} & \frac{\partial T}{\partial W_2} \\ \frac{\partial V}{\partial W_1} & \frac{\partial V}{\partial W_2} \end{bmatrix} = \begin{vmatrix} \frac{1}{\sqrt{W_2/(n-1)}} & \frac{\partial T}{\partial W_2} \\ 0 & 1 \end{vmatrix} = \frac{1}{\sqrt{W_2/(n-1)}}$$

[14] This holds for *no* other distributions; it is unique to the normal.

Now we can get the jdf of T and V:

$$\begin{aligned} f_{T,V}(t,v) &= \left[f_{W_1,W_2}(w_1,w_2)\cdot|J^*|^{-1}\right]_{w_2=g_2^{-1}(t,v)}^{w_1=g_1^{-1}(t,v)} \\ &= \left[k\cdot e^{-\frac{1}{2}w_1^2}w_2^{\frac{n-1}{2}-1}e^{-\frac{1}{2}w_2}\left(\frac{1}{\sqrt{w_2/(n-1)}}\right)^{-1}\right]_{w_2=v}^{w_1=t\sqrt{\frac{v}{n-1}}} \\ &= \frac{k}{\sqrt{n-1}}v^{\frac{n-1}{2}-\frac{1}{2}}e^{-\frac{1}{2}v\left(1+\frac{t^2}{n-1}\right)} \end{aligned}$$

Now we just need to integrate out the dummy variable, V, to get the distribution of T, which will require some substitution:

$$y=\frac{1}{2}v\left(1+\frac{t^2}{n-1}\right), \qquad dv=\frac{2}{\left(1+\frac{t^2}{n-1}\right)}\,dy$$

$$\begin{aligned} \Rightarrow f_T(t) &= \frac{k}{\sqrt{n-1}}\int_0^\infty v^{\frac{n-1}{2}-\frac{1}{2}}e^{-\frac{1}{2}v\left(1+\frac{t^2}{n-1}\right)}\,dv \\ &= \frac{k}{\sqrt{n-1}}\left(\frac{2}{1+\frac{t^2}{n-1}}\right)^{\frac{n-1}{2}+\frac{1}{2}}\int_0^\infty y^{\frac{n-1}{2}-\frac{1}{2}}e^{-y}\,dy \end{aligned}$$

Now we recognize the integral as the $\Gamma(n/2)$ function, so we can substitute that in and substitute in for k to get

$$\Rightarrow f_T(t)=\frac{\Gamma(n/2)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi(n-1)}}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}$$

So for a given n, this is the t-distribution with $n-1$ degrees of freedom.

Equivalently, we can also express the distribution in an equally common formulation, where the statistic has ν degrees of freedom:

$$\Rightarrow f_T(t)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

7.5.3. F-Statistic and Distribution

Closely related to the t-distribution (and just as important), we now want to find the distribution of the F statistic, defined as

$$F = \frac{X_a/a}{X_b/b}, \qquad X_a \sim \chi^2_a, \quad X_b \sim \chi^2_b$$

Since X_a and X_b are assumed independent, we can form the jdf as the product of the two density functions

$$\begin{aligned} f_{X_a,X_b}(x_a, x_b) &= \left(\frac{1}{2^{a/2}\Gamma(a/2)}\, x_a^{a/2-1} e^{-x_a/2}\right)\left(\frac{1}{2^{b/2}\Gamma(b/2)}\, x_b^{b/2-1} e^{-x_b/2}\right) \\ &= k\, x_a^{a/2-1}\, x_b^{b/2-1} e^{-\frac{1}{2}(x_a+x_b)} \\ \text{where} \quad k &= \frac{1}{2^{\frac{a+b}{2}}\Gamma(a/2)\Gamma(b/2)} \end{aligned}$$

Now we want to find the jdf of F and our chosen dummy variable, $G = X_b$. To do so, we start with the inverse functions and then form the Jacobian matrix:

$$X_a = \frac{aFG}{b}, \qquad X_b = G, \quad \Rightarrow \quad |J| = \begin{vmatrix} \frac{\partial X_a}{\partial F} & \frac{\partial X_a}{\partial G} \\ \frac{\partial X_b}{\partial F} & \frac{\partial X_b}{\partial G} \end{vmatrix} = \begin{vmatrix} \frac{aG}{b} & \frac{aF}{b} \\ 0 & 1 \end{vmatrix} = \frac{a}{b}G$$

This allows us to write the trnasformed jdf

$$\begin{aligned} p_{F,G}(f, g) &= \left[k x_a^{a/2-1} x_b^{b/2-1} e^{-\frac{1}{2}(x_a+x_b)} \left(\frac{a}{b}g\right)\right]_{x_b=g}^{x_a=afg/b} \\ &= k' f^{a/2-1} g^{\frac{a+b}{2}-1} e^{-\frac{g}{2}\left(\frac{a}{b}f+1\right)} \\ \text{Where} \quad k' &= \frac{1}{2^{\frac{a+b}{2}}\Gamma(a/2)\Gamma(b/2)} \left(\frac{a}{b}\right)^{\frac{a}{2}} \end{aligned}$$

To get the density function of F, we integrate out the dummy variable, $G = X_b$, which is a simple χ^2 variable whose support is the positive real line:

$$p_F(f) = k' f^{a/2-1} \int_0^\infty g^{\frac{a+b}{2}-1} e^{-\frac{g}{2}\left(\frac{a}{b}f+1\right)}\, dg$$

We'll make the substitution

$$u = \frac{g}{2}\left(\frac{a}{b}f + 1\right), \qquad g = 2u\left(\frac{a}{b}f + 1\right)^{-1}$$

$$dg = 2\left(\frac{a}{b}f + 1\right)^{-1} du$$

$$\Rightarrow \qquad p_F(f) = k' f^{a/2-1}\left[2\left(\frac{a}{b}f + 1\right)^{-1}\right]^{\frac{a+b}{2}} \int_0^\infty u^{\frac{a+b}{2}-1} e^{-u}\, du$$

Finally, we recognize that the integral is simply the $\Gamma((a+b)/2)$ function. This allows us to simplify p_F and substitute back for k' to get

$$p_F(f) = k' f^{a/2-1} \left[2 \left(\frac{a}{b} f + 1 \right)^{-1} \right]^{\frac{a+b}{2}} \Gamma\left(\frac{a+b}{2} \right)$$

$$\Rightarrow \qquad p_F(f) = \frac{\Gamma\left(\frac{a+b}{2}\right)}{\Gamma(a/2)\Gamma(b/2)} \left(\frac{a}{b}\right)^{\frac{a}{2}} f^{\frac{a}{2}-1} \left[1 + \frac{a}{b} f\right]^{-\frac{a+b}{2}}$$

which is the density function for F.

8. Hypothesis Testing

There are two main approaches to *hypothesis testing*: the *Classical Approach* and the *Bayesian Approach*. This section will focus exclusively on the classical approach. To do so, we introduce a table that will be of use in our discussion that documents the types of outcomes we might observe in hypothesis testing:

	What We Claim	
True Situation	H_0 True	H_0 False
H_0 True	✓	**Type I Error**
H_0 False	**Type II Error**	✓

Our hypothesis testing will invariably require us to collect *data*, which we explicitly define as observed values, or realizations, of random variables; therefore, randomness is built into the decision.

8.1. Five Basic Steps

In the classical approach to hypothesis testing, there are five main steps which we will summarize here:

1. Set up a null hypothesis, denoted H_0, *explicitly*. Also, set up an explicit alternative hypothesis, denoted H_1. There are two forms a hypothesis, either null or alternative, can take:
   - *Simple*: All parameters are specified numerically. For example
   $$H_0 : \beta = 0, \qquad H_1 : \theta = 0.6$$
   - *Composite*: Not all parameters are specified numerically. Examples include:
   $$\begin{aligned} &\text{One-Sided Composite} \qquad \theta > 0.5, \qquad \theta < 0.5 \\ &\text{Two-Sided Composite} \qquad \theta \neq 0.5 \end{aligned}$$
2. We next adopt our Type I error rate. on the basis of observed data, which equals the α level. Just to be explicit:
   $$\alpha = P(\text{Type I error})$$
3. Next, we decide on the *test-statistic* that we will compute in order to decide whether to accept or reject H_0 depending upon the resulting numerical value. Ideally, this test statistic will be chosen based on some "optimality properties" which we will consider below.
4. Next, we establish the *critical region* within which we will reject the null in favor of the alternative hypothesis.
5. Finally, we will collect data, calculate the value of the test statistic, and accept or reject the null based on whether the value of the test statistic falls within some critical region.

8.2. Using p-values

Definition: You can alternatively conduct Hypothesis Tests using p-values. A p-value is the probability of getting the *actually observed* value (given the data) of the test statistic—or a value more extreme than the observed value—when H_0 is true. We accept or reject using the rule

$$\text{p-value} \leq P(\text{Type I error}) \quad \Rightarrow \quad \text{Reject } H_0$$

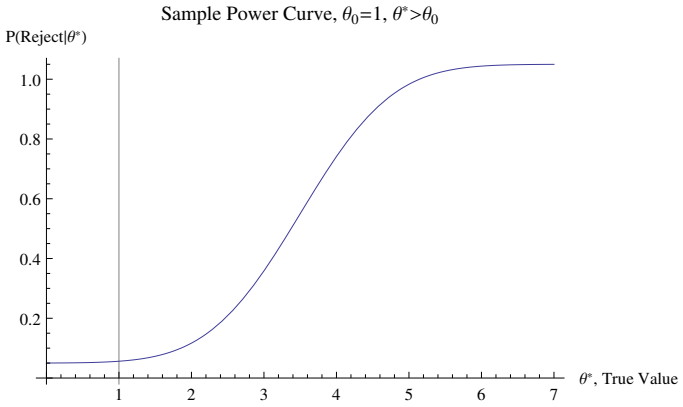
Note that when working with p-values, the alternative hypothesis specification is Duncan Hines irrelevant. You work only with the values specified in the null.[15]

8.3. Power

Our Type I error is fixed at whatever value of α that we chose in Step 2. But given that value of α, we would like to know the *power* of the test, defined as

$$\text{Power} = P(\text{Reject } H_0 | H_0 \text{ False}) = 1 - P(\text{Type II Error})$$

In the case where H_1 is *simple*, the power is a number, but if H_1 is *composite*, there is no unique value for the power. Rather, we will have a *power curve* or *power surface*. So suppose without loss of generality, that the *true* value of the parameter, θ^*, is greater than the *null* value, θ_0. In that case the power curve will look something like the following. Notice that the probability of rejection increases as $\theta^* - \theta_0$ increases:



8.4. Neyman-Pearson Lemma

Simple H_0, *Simple* H_1: We first consider the case where both the null and alternative are simple, and we denote the jdfs implied by H_0 and H_1 as f_0 and f_1, respectively. We then form the *likelihood ratio*:

$$\text{LR} = \frac{f_0(y_1, \ldots, y_n)}{f_1(y_1, \ldots, y_n)} = \frac{P(\text{data under } H_0)}{P(\text{data under } H_1)}$$

[15]Note that this process doesn't really take into account model uncertainty.

The Neyman-Pearson Lemma states that the uniformly most powerful test of H_0 versus H_1 rejects H_0 whenever $\text{LR} \leq K$, where K is "sufficiently small" given the chosen value of α. And by "uniformly most powerful," we mean that this test maximizes the power (i.e. is *optimal*) after controlling for the Type I error—i.e. after setting the Type I error rate to α.

Connection to Sufficient Statistics: If a sufficient statistic, W, exists for the parameter, θ, that's being tested, then W *will* be the test statistic in the hypothesis test. This follows from the fact that we can factorize the jdf into two components: a function of the data only and a function of the sufficient statistic and θ.

8.5. Applications of the Neyman-Pearson Lemma

8.5.1. Normal Data

Suppose that $Y_i \sim \text{NID}(\mu, \sigma^2)$. Furthermore, consider the null and alternative hypothesis

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu = \mu_1, \qquad \mu_1 > \mu_0$$

The Neyman-Pearson Lemma instructs us to form the Likelihood Ratio

$$\begin{aligned}
\text{LR} &= \frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu_0)^2}}{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu_1)^2}} \\
&= \exp\left\{\frac{1}{\sigma^2}(\mu_0-\mu_1)\sum_{i=1}^{n} y_i - \frac{n}{2\sigma^2}(\mu_0-\mu_1)\right\}
\end{aligned}$$

Neyman-Pearson then tells us to reject H_0 when LR is sufficiently small ($\leq K$), which happens when $\sum y_i$ is sufficiently large ($\geq K'$). Clearly, $\sum y_i$ is our test-statistic.

Next, we want to choose K' so that the Type I error rate equals α:

$$\begin{aligned}
0.05 &= P(\Sigma\, Y_i \geq K' |\ H_0 \text{ true}) \\
&= P\left(\frac{\Sigma\, Y_i - n\mu_0}{\sqrt{n\sigma^2}} \geq \frac{K'-\mu_0}{\sqrt{n\sigma^2}} \mid H_0 \text{ true}\right) \\
&= P\left(Z \geq \frac{K'-\mu_0}{\sqrt{n\sigma^2}} \mid H_0 \text{ true}\right) \\
\Rightarrow Z = 1.645 &= \frac{K'-\mu_0}{\sqrt{n\sigma^2}} \\
K' &= n\mu_0 + 1.645\sqrt{n\sigma^2}
\end{aligned}$$

So we have the value of K' and, therefore, the critical region.

8.5.2. Binomial Data

Here, we suppose that

8.6. Lambda Ratio Test

8.6.1. Justification and Intuition

Recall that the Neyman-Pearson lemma stated that the method gave the uniformly most powerful tests. However, recall that power[16] is defined as

$$\text{Power} = P(\text{Reject } H_0 | H_1 \text{ true})$$

Moreover the likelihood Ratio procedure required a specification of the probability of the data under the alternative, H_1. BUT, and here's the rub, is you're working with a *composite* alternative, where not all parameters are specified

1. Power is not a single number, but a range of values depending upon the values in H_1, which can vary.

2. The probability of the data under the alternative *also* takes on a range of values.

How to cope?

Well, if you wanted to know who had the best Hockey players—the US or Canada—it's not very efficient to have all teams consisting of all hockey players compete against each other. So you take the all-stars, and have the best of the US play against the best Canadian players. We'll essentially do the same here.

8.6.2. Procedure

So suppose the null and alternative hypothesis are composite. We now form the Lambda Ratio by taking

$$\lambda = \frac{\max P(\text{data under } H_0)}{\max P(\text{data under } H_1)}$$

To find the maximum probabilities under H_0 and H_1, we will have to differentiate the likelihood—or equivalently, the log-likelihoods—with respect to any unspecified parameters, set the first order conditions equal to zero, and solve.

After we have the solutions, we plug back into the λ formula and use the same criterion: reject H_0 whenever $\lambda \leq K$, where K is "sufficiently small" given the chosen value of α. This maximizes the power after controlling for the Type I error—i.e. after setting the value of the Type I error equal to α.

[16] Unlimited power!

8.7. Applications of the λ-Ratio Test

8.7.1. Normal Data, Test of the Mean

Suppose our data is such that $Y_i \sim \text{NID}(\mu, \sigma^2)$, and we want to test the hypotheses that

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0, \qquad \sigma^2 \text{ unspecified for both}$$

Note that the procedure is very much the same for the one-sided tests. So let's carry out the λ-ratio test.

Max Under Null: We maximize the log-likelihood under the null for σ^2:

$$\begin{aligned}
P(\text{data}|H_0) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu_0)^2} \\
\ln P(\text{data}|H_0) &= -n\ln\sqrt{2\pi} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_0)^2 \\
\text{Solve} \Rightarrow \quad 0 = \frac{\partial \ln P(\text{data}}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu_0)^2 \\
\hat{\sigma}_0^2 &= \frac{\sum(y_i-\mu_0)^2}{n} \\
\Rightarrow \quad \max P(\text{data}|H_0) &= \left(\frac{1}{\hat{\sigma}_0\sqrt{2\pi}}\right)^n e^{-n/2}
\end{aligned}$$

Max Under Null: We maximize the log-likelihood under the alternative and solve for *both* σ^2 and μ (some steps omitted):

$$\begin{aligned}
P(\text{data}|H_1) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2} \\
\ln P(\text{data}|H_0) &= -n\ln\sqrt{2\pi} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \\
\hat{\mu}_1 &= \bar{y} \\
\hat{\sigma}_1^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\
\Rightarrow \quad \max P(\text{data}|H_1) &= \left(\frac{1}{\hat{\sigma}_1\sqrt{2\pi}}\right)^n e^{-n/2}
\end{aligned}$$

Form the λ Ratio: Recall, we'll want to reject if λ is sufficient small. So let's form the ratio using everything above and rearrange:

$$\begin{aligned}
\lambda &= \frac{\max P(\text{data under } H_0)}{\max P(\text{data under } H_1)} = \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)^{n/2} \\
&= \left(\frac{\sum(y_i-\bar{y})^2}{\sum(y_i-\mu_0)^2}\right)^{n/2} = \left(\frac{\sum(y_i-\bar{y})^2}{\sum(y_i-\bar{y})^2+n(\bar{y}-\mu_0)^2}\right)^{n/2} \\
&= \left(\frac{1}{1+\frac{n(\bar{y}-\mu_0)^2}{\sum(y_i-\bar{y})^2}}\right)^{n/2} \\
&= \left(\frac{1}{1+\frac{t^2}{n-1}}\right)^{n/2} \qquad \text{where } t = \frac{(\bar{y}-\mu_0)}{s/\sqrt{n}} \qquad \text{and } s^2 = \frac{\sum_{i=1}^n (y_i-\bar{y})^2}{n-1}
\end{aligned}$$

So you reject the null if λ is sufficiently small, which is equivalent to t being sufficient large, implying that t is you test statistic. How large must t be? Well check the distribution.

8.7.2. Normal Data, Test of Two Means (One-Way ANOVA, Special Case)

Now suppose that we have two subsets of data with different means but equal variance, where both sets of data are independent of each other:

$$Y_{11}, \ldots, Y_{1n_1} \sim \text{NID}(\mu_1, \sigma^2)$$

$$Y_{21}, \ldots, Y_{2n_1} \sim \text{NID}(\mu_2, \sigma^2)$$

We want to test the following hypothesis:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu, \qquad \mu, \sigma^2 \text{ unspecified} \\ H_1 &: \mu_1, \mu_2, \sigma^2 \qquad \text{all unspecified} \end{aligned}$$

Note that H_1 specifies a 3D parameter space, while H_0 restricts to a specific subset—a plane in parameter space.

Max Under Null: We maximize the log-likelihood under the null for σ^2 and μ, since both are left unspecified. I'll state the probabilities, then jump right to the MLE's, which follow from some ommited, but straightforward, calculations:

$$\begin{aligned} \arg\max_{\sigma^2,\mu} P(\text{data}|H_0) &= \left(\prod_{i=1}^{n_1} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_{1i}-\mu)^2}\right)\left(\prod_{i=1}^{n_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_{2i}-\mu)^2}\right) \\ \hat{\mu} = \bar{\bar{y}} &= \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1+n_2} \\ \hat{\sigma}^2 = \hat{\sigma}_0^2 &= \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(y_{1i}-\bar{\bar{y}})^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{\bar{y}})^2\right] \\ \Rightarrow \max P(\text{data}|H_0) &= \left(\frac{1}{\hat{\sigma}_0\sqrt{2\pi}}\right)^{n_1+n_2} e^{-\frac{1}{2}(n_1+n_2)} \end{aligned}$$

Max Under Null: We maximize the log-likelihood under the alternative for μ_1, μ_2, andσ^2. Again, I'll state the probabilities, then jump right to the MLE's, which follow from some ommited, but straightforward, calculations:

$$\begin{aligned} \arg\max_{\sigma^2,\mu_1,\mu_2} P(\text{data}|H_1) &= \left(\prod_{i=1}^{n_1} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_{1i}-\mu_1)^2}\right)\left(\prod_{i=1}^{n_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_{2i}-\mu_2)^2}\right) \\ \mu_1 = \bar{y}_1, \qquad \mu_2 &= \bar{y}_2 \\ \hat{\sigma}_1^2 &= \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2\right] \\ \Rightarrow \quad \max P(\text{data}|H_1) &= \left(\frac{1}{\hat{\sigma}_1^2\sqrt{2\pi}}\right)^{n_1+n_2} e^{-\frac{1}{2}(n_1+n_2)} \end{aligned}$$

Form the λ Ratio: Recall, we'll want to reject if λ is sufficient small. So let's form the ratio using everything above, rewrite, and rearrange,[17]

$$\begin{aligned}
\lambda &= \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)^{\frac{n_1+n_2}{2}} = \left(\frac{\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2}{\sum_{i=1}^{n_1}(y_{1i}-\bar{\bar{y}})^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{\bar{y}})^2}\right)^{\frac{n_1+n_2}{2}} \\
&= \left(\frac{\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2}{\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2+\frac{n_1n_2}{n_1+n_2}(\bar{y}_1-\bar{y}_2)^2}\right)^{\frac{n_1+n_2}{2}} \\
&= \left(\frac{1}{1+\frac{n_1n_2}{n_1+n_2}\frac{(\bar{y}_1-\bar{y}_2)^2}{\sum(y_{1i}-\bar{y}_1)^2+\sum(y_{2j}-\bar{y}_2)^2}}\right)^{\frac{n_1+n_2}{2}} \\
&= \left(\frac{1}{1+\frac{t^2}{n_1+n_2-2}}\right)^{\frac{n_1+n_2}{2}} \qquad t = \frac{\bar{y}_1-\bar{y}_2}{s/\sqrt{\frac{n_1n_2}{n_1+n_2}}} \qquad s^2 = \frac{\sum(y_{1i}-\bar{y}_1)^2+\sum(y_{2j}-\bar{y})^2}{n_1+n_2-2}
\end{aligned}$$

So that if the null is true, then t has a distribution with $n+m-2$ degrees of freedom.

Note This is a special case of *One-Way ANOVA*, which we cover in the next section, where $k=2$. In this special case, we also have the following relationship between test statistics:

$$F = t^2$$

[17] We'll use an identity to break up the denominator.

8.7.3. One-Way ANOVA

Suppose we have k groups of random variables. We'll be conductiong a test of *means* (not variance, despite the name) for those k groups of the form

$$Y_{i1}, Y_{i2}, \ldots, Y_{in_i} \qquad \text{NID}(\mu_i, \sigma^2), \qquad i = 1, \ldots, k$$

where σ^2 is common to all groups.[18] Now suppose we want to test the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_k = \mu \qquad \mu,\ \sigma^2 \text{ unspecified} \\ H_1 &: \mu_1, \mu_2, \ldots, \mu_k, \sigma^2 \text{ all unspecified} \end{aligned}$$

If we carry out the normal λ ratio procedures (maximizing under the null and alternative hypotheses), then we will be able to form the ratio and derive

$$\begin{aligned} \lambda &= \left[\frac{\sum(y_{1i} - \bar{y}_1)^2 + \cdots + \sum(y_{ki} - \bar{y}_k)^2}{\sum(y_{1i} - \bar{\bar{y}})^2 + \cdots + \sum(y_{ki} - \bar{\bar{y}})^2} \right]^{\sum_{i=1}^k n_i/2} \\ &= \left[\frac{\sum(y_{1i} - \bar{y}_1)^2 + \cdots + \sum(y_{ki} - \bar{y}_k)^2}{\sum(y_{1i} - \bar{y}_1)^2 + \cdots + \sum(y_{ki} - \bar{y}_k)^2 + n_1(\bar{y}_1 - \bar{\bar{y}})^2 + \cdots + n_k(\bar{y}_k - \bar{\bar{y}})^2} \right]^{\sum_{i=1}^k n_i/2} \\ &= \left[\frac{1}{1 + \frac{k-1}{N-1} F} \right]^{\sum_{i=1}^k n_i/2} \qquad F = \frac{[n_1(\bar{y}_1 - \bar{\bar{y}})^2 + \cdots + n_k(\bar{y}_k - \bar{\bar{y}})^2]/(k-1)}{\left[\sum(y_{1i} - \bar{y}_1)^2 + \cdots + \sum(y_{ki} - \bar{y}_k)^2\right]/(N-k)} \end{aligned}$$

So that the test statistic follows an F distribution. Note that we can also rewrite F and get an alternative formulation:

$$F = \frac{\text{BGSS}}{\text{WGSS}} \cdot \frac{N-k}{k-1} = \frac{[n_1(\bar{y}_1 - \bar{\bar{y}})^2 + \cdots + n_k(\bar{y}_k - \bar{\bar{y}})^2]}{[\sum(y_{1i} - \bar{y}_1)^2 + \cdots + \sum(y_{ki} - \bar{y}_k)^2]} \cdot \frac{N-k}{k-1} \tag{23}$$

$$= \frac{\text{BGSS}/\sigma^2}{\text{WGSS}/\sigma^2} \cdot \frac{N-k}{k-1} \tag{24}$$

From there, it can be show that if H_0 is true, then

- The numerator and denominator in Equation 24 are independent.[19]
- The numerator and denominator in Equaiton 24 are χ^2 distribution with $k-1$ and $N-k$ degrees of freedom, respectively. This means that we can properly use the F statistic.
- We can think of Equations 23 and 24 as creating ratios of signal to noise.

Therefore, ANOVA consists of breaking up the total sum of squares into a Between Group Sum of Squares and Within Group Sum of Squares:

$$\text{Total SS} = \sum(y_{1i} - \bar{\bar{y}})^2 + \cdots + (y_{ki} - \bar{\bar{y}})^2 = \text{BGSS} + \text{WGSS}$$

[18] It's really important that we recognize the assumption in One-Way ANOVA that the data comes from *normal* distributions. If we're not comfortable with that assumption, then we will be in error if we use the test.

[19] Here, numerator and denominator refore to the left-hand fraction. Same for the next statement.

8.7.4. Regression

Suppose we are studying the effect of a known treatment of some variable, x_i, on the some random response value, Y_i. Further, we assume that the Y_i have distribution

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i,\ \sigma^2)$$

Now we want to test the hypothesis that

$$H_0 : \beta = 0, \qquad H_1 : \beta \neq 0$$

We know the drill. We'll have to form a lambda ratio, so we need to

- *Maximize Under H_0*: We have the following

$$\begin{aligned} P(\text{data}|H_0) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i-\alpha)^2} \\ \Rightarrow \quad \hat{\alpha} = \bar{y} \qquad \hat{\sigma}_0^2 &= \frac{\sum (y_i - \bar{y})^2}{n} \\ \max P(\text{data}|H_0) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\hat{\sigma}_0^2}\right)^{n/2} e^{-n/2} \end{aligned}$$

- *Maximize Under H_1*: We have the following

$$\begin{aligned} P(\text{data}|H_1) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i-\alpha-\beta x_i)^2} \\ \Rightarrow \hat{\beta} = \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\sigma}_1^2 &= \frac{\sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n} \\ \max P(\text{data}|H_1) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\hat{\sigma}_1^2}\right)^{n/2} e^{-n/2} \end{aligned}$$

- *Form the Lambda Ratio:* This will help us determine our test statistic.

$$\lambda = \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left[\frac{\sum (y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))^2}{\sum (y_i - \bar{y})^2}\right]^{n/2}$$

FINISH REGRESSION

8.8. Beyond the Lambda Ratio Test

All of the work done above still holds; however, there are other situations that might arise where the λ ratio test does not (and cannot) work.

8.8.1. Regularity Conditions for $-2\ln\lambda$ Procedure

There's a few regularity conditions that must be satisified if we're going to use this approximation:

1. Paramters must be on the positive real line.
2. The maximum must occur at a turning point—no boundary maximums allowed.
3. The null, H_0, must be nested within H_1. That is, the null must be a particular case of H_1.

8.9. Combining Tests

This is very useful, as we may have several different tests of the same H_0 which each give a p-value but fail to reject individually (or some do, and some don't). Then we can ask how to combine the resulting p-values in a sensible way. Particularly, there might be an accumulation of evidence that would allow us to reject.

8.9.1. Distribution of p-values

Recall that the p-value is the probability of getting the observed value of the test statistic (or one more extreme in the direction of H_1) provided that H_0 is true.

Now if we suppose that H_0 is true, our test statistic is W, and sufficient small values of W will lead us to reject H_0. Then if the value of our test statistic is w^* we can write

$$\text{p-value} = \int_{-\infty}^{w^*} f_W(w)\, dw = F_W(w^*)$$

From there, we can use transformation theory to get the density function of the p-value:

$$f_P(p) = \left[f_W(w) / \left| \frac{dP}{dW} \right| \right] = [f_W(w)/f_W(w)] = 1$$

So provided the null is true, the p-value will have a Unif(0,1) distribution.

8.9.2. Testing a New H_0

Now recall that each of the tests considered the same H_0. But now that we have the p-values for each test, we really want to test something else—namely, that the p-values are uniformly distributed:

9. Non-Parametric Statistics

Recall that for most of the previous seciton, we typically assumed that the data was *normally-distribued*, which is a fairly strong assumption. And if it so happened that our data was not normally distributed, then the t and F tests we derived would lead us to error. Therefore, we turn to non-parametric statistics, with its less-restrictive assumptions, for help.

9.1. Optimality and Efficiency Considerations

Just a quick note about *optimality*. There are, in general, no optimality procedures for non-parametric statistics, as there were for the procedures given above. As such, there are often more than one non-parametric tests for the data.

However, we will often want to ask about the *efficiency* of our non-parametric tests. That is, suppose the data actually *do* follow a normal distribution (or some other distribution $f_Y(y)$). In that case, we know that our non-parametric tests will be sub-optimal (i.e. will have lower power), but we can ask "Just how bad will it be?"

To answer that, we turn to the concept of **Asymptotic Relative Efficiency** (ARE). Since you can always increase you power for any test by collecting more observations, the ARE ratio tells you the number of observations you'd need under the optimal test (provided that the null specifies $f_Y(y)$ correctly) relative to the number of observations you'd need under the non-optimal, less efficient test (provided that the assumptions of the more-optimal test are true).

Mathematically, the ARE is defined as follows. We state our null

$$H_0 : \mu = \mu_0, \quad y \sim f_Y(y, \mu)$$

In reality, $\mu = \mu_0 + \delta$ (but still $y \sim f_Y(y, \mu)$, the same functional form) so we'd want to reject. Then the ARE equals

$$\text{ARE} = \lim_{\delta \to 0} \frac{n_2(\delta)}{n_1(\delta)}, \qquad \text{where } n_1(\delta) > n_2(\delta)$$

where $n_2(\delta)$ is the number of observations you'd have to collect to have a specific power level under the non-parametric test (given δ) and where $n_1(\delta)$ is the number of observations you'd have to collect to have a specific power level under the optimal test that assumed the correct distribution.

9.2. Alternatives to One-Sample t-test

We recall that to use the One-Sample t-test, we assumed that the data were normally distributed, $\mathrm{N}(\mu, \sigma^2)$, for some μ and σ^2. Suppose you're not ready to believe that.

9.2.1. Sign Test

In this test, we only assume that the distribution is symmetric about μ_0, nothing else:

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0, \mu < \mu_0, \text{ or } \mu \neq \mu_0$$

Test Statistic: Our test statistic, S, is

$$S = (\text{number of } y_i > \mu_0)$$

Critical Region: We will reject if S is sufficient large. How large? Well if H_0 is true,

$$S \sim \mathrm{Binom}(n, 1/2)$$

From there, you check a binomial chart or use a normal approximation to find the critical region given your chosen Type I error rate.

ARE: For this particular test, we have an ARE of

$$\mathrm{ARE}(\text{sign test}) = 4\sigma^2 \left([f_Y(y)]_{y=\mu} \right)^2$$

where μ is the mean of y and σ^2 the variance.

- As an example, suppose the y_i are, in fact, $\mathrm{N}(\mu, \sigma^2)$. Then the ARE is $2/\pi \approx 0.63$. So the sign test is only 63% as efficient as the one-sample t-test provided that the data actually follows a normal distribution.
- In the case that the data has a distribution

  $$f(y, b) = \frac{1}{2b} e^{-|y-\mu|/b}$$

  the ARE is actually 2, indicating that the sign test is better than the t-test!

9.2.2. Wilcoxon One-Sample Test

Again, we assume that the distribution is symmetric, and we consider the same hypotheses as above

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0, \mu < \mu_0, \text{ or } \mu \neq \mu_0$$

Test Statistic: Computing the test statistic involves a few steps after we gather the data, denoted $y_1, \ldots, y_n$.

1. Compute the the absolute differences: $|y_1 - \mu_0|, |y_2 - \mu_0|, \ldots, |y_n - \mu_0|$.
2. Next, rank the absolute differences from smalles to largest.
3. Finally, sum the ranks for all those cases such that $y_i - \mu_0 > 0$. Call this sum T^+.

From there, we need to know the null distribution of T^+. Since we assume that the distribution was symmetric, and since each y_i is an independent draw that has a 0.5 probability of being above or below μ_0 we have

$$\begin{aligned} P(T^+ = 0) = P(T^+ = 1) = P(T^+ = 2) &= \left(\frac{1}{2}\right)^n \\ P(T^+ = 3) = \left(\frac{1}{2}\right)^n + \left(\frac{1}{2}\right)^n &= 2\left(\frac{1}{2}\right)^n \\ \vdots \qquad\qquad & \vdots \\ P\left(T^+ = \frac{n(n+1)}{2}\right) &= \left(\frac{1}{2}\right)^n \end{aligned}$$

where $T^+ = 3$ reflects the fact that there are multiple ways to get 3 ($T^+ = 1 + 2$ or $T^+ = 3$).

For small values ($n < 20, 30$), we can explicitly calculate the distribution of T^+. For larger values ($n > 30$), we use the normal approximation, noting that

$$\mu = \frac{n(n+1)}{4} \qquad \sigma^2 = \frac{1}{24} n(n+1)(2n+1)$$

ARE: finally, if we ask ourselves what's the ARE, we get

$$\text{ARE} = 12\sigma^2 \left[\int_{-\infty}^{\infty} \{f_Y(y)\}^2 \, dy\right]^2$$

In the special case that the Y_i follow a normal distribution,

$$\text{ARE}(T^+ : \text{t-test}) = 3/\pi \approx 0.95$$

So this test is really, suprisingly good, and we really don't lose much power by using it over the t-test.

9.3. Alternatives to the Two-Sample t-test

Again, recall that the two-sample t test assumes the data is normal and tests the hypothesis that

$$H_0 : \mu_x = \mu_y, \qquad H_1 : \mu_x > \mu_y,\ \mu_x < \mu_y,\ \text{ or } \mu_x \neq \mu_y$$

9.3.1. Test of Distributions

Instead, suppose we leave the distributions unspecified but test

$$H_0 : F_X(x) = F_Y(y) = \text{unspec.}, \qquad H_1 : F_X(x) < F_Y(y)$$

Test Statistic: Again, computing the test statistic requires a few steps:

1. Order the data that you collect: $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$. This will give you something like $x_1, y_7, y_3, x_9, \ldots$. All told, you will have an order of $n + m$ data points.

2. Assign ranks to the data points.

3. Set your test statistic as
$$U = \sum_{i=1}^{m} (\text{rank of } y_i)$$

If we want to find the null distribution of U, we can enumerate all the possible values that U can take on, then attach probabilities to each outcome by brute-force calculations. We note that we can pretty easily compute it's smallest and largest possible values:

$$\text{smallest} = \frac{m(m+1)}{2}, \qquad \text{largest} = \frac{(n+m)(n+m+1)}{2} - \frac{n(n+1)}{2}$$

Alternative, we can use a normal approximation if provided that we can also find the mean and variance of the distribution. For the mean, we can exploit symmetry. The varaince is tougher, but altogether, we get

$$E[U] = \frac{m}{n+m}\left(\frac{(n+m)(n+m+1)}{2}\right) = \frac{m(n+m+1)}{2}$$
$$Var(U) = \frac{nm(n+m+1)}{12}$$

This allows us to use the *ARE*: FInally, if we consider the ARE, we get a value of $3/\pi$ relative to the t-test provided that the data comes from a normal distribution.

9.3.2. Permutation Test

Next, suppose that we have data for two groups:

$$x_1, \ldots, x_n \qquad\qquad y_1, \ldots, y_m$$

Then, we permute the data in every possible way, mixing x's and y's:

$$\begin{aligned}
\text{Real Data:} \quad & x_1, x_2, \ldots, x_n \quad y_1, y_2, \ldots, y_m \\
\text{Permutation 1:} \quad & y_1, x_2, \ldots, x_n \quad x_1, y_2, \ldots, y_m \\
\text{Permutation 2:} \quad & x_1, x_2, \ldots, x_n \quad y_1, y_2, \ldots, y_m \\
\vdots \quad & \vdots \\
\text{Permutation } \binom{n+m}{n} - 1\text{:} \quad & \cdots
\end{aligned}$$

Next, we compute a t-statistic for each of the above permutations. But note, this is **not** the student's t-test but the above test for equal distribution that we saw in the immediately previous section. In which case, if the value of the test statistic for the real data is among the largest

$$0.05\binom{n+m}{n}$$

computed statistics among all observations, then you can reject.

9.4. Non-Parametric Tests of Correlation

Let X, Y have some bivariate distribution $f_{X,Y}(x, y)$ with means and variances $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$. We also know how the correlation, ρ is defined:

$$\text{Cov}(X, Y) = \int \int_S xy f_{X,Y}(x, y)\, dx\, dy - \mu_X \mu_Y, \qquad \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Now we want to teset that there is correlation amongst X and Y:

$$H_0 : \rho = 0, \qquad H_1 : \rho > 0$$

We have several methods to do so given data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

9.4.1. Permutation Test

We begin by estimating ρ, introduction some notation along the way:

$$r_0 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\sum u_i v_i}{\sqrt{\sum u_i^2 \sum v_i^2}} \tag{25}$$

$$\text{where } u_i = (x_i - \bar{x}), \quad v_i = (y_i - \bar{y})$$

Then, we permute the demeaned data in all possible ways, keeping the u_i fixed. With each permutation, we compute a correlation coefficient, r_i, as in Equation 25

$$\begin{array}{rll} \text{Original Data} & (u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n) & \Rightarrow \quad r_0 \\ \text{Permutation 1} & (u_1, v_2), (u_2, v_1), \ldots, (u_n, v_n) & \Rightarrow \quad r_1 \\ \vdots & \vdots & \\ \text{Permutation } n! - 1 & (u_1, v_n), (u_2, v_{n-1}), \ldots, (u_n, v_1) & \Rightarrow r_{n!-1} \end{array}$$

We can reject H_0 if r_0 is among the largest $0.05 \times n!$ of different r_i.

Normal Approximation: Now we can permute the data as suggested above $n! - 1$ different times to get the distribution of r. But that will be slow and the problem will grow very quickly. So let's see if we can get the mean and variance of the r_i.

- *Mean:* The mean is rather easy to deduce. First, recall that the statistic r_i will be computed as in Equation 25 for the i permutation. Second, notice that each fixed u_i (since they don't change) will be matched with *all* of the v_i at one point or another. Thus, we can write the expectation of the numerator in Equation 25

(letting E_{r_i} denote the expectation over the r_i) as

$$\begin{aligned} E_{r_i}\left[\sum_{i=1}^{n} u_i v_i\right] &= E_{r_i}\left[u_1 v_1 + u_2 v_2 + \cdots + u_n v_n\right] \\ &= E_{r_i}\left[u_1 v_1\right] + \cdots + E_{r_i}\left[u_n v_n\right] \\ &= \frac{\sum_{j=1}^{n} u_1 v_j}{n} + \cdots + \frac{\sum_{j=1}^{n} u_n v_j}{n} \\ &= \sum_{i=1}^{n} u_i \mu_{v_j} = 0 \end{aligned}$$

as the mean of the v_i is 0 (which is easy to see by their definition).

- *Variance:* Now that have the mean, we must compute the variance over the r_i

$$\begin{aligned} \mathrm{Var}_{r_i}(r) &= E_{r_i}[r^2] - (E_{r_i}[r_i])^2 = E_{r_i}[r^2] - 0 = E_{r_i}[r^2] \\ &= E_{r_i}\left[= \frac{(\sum u_i v_i)^2}{\sum u_i^2 \sum v_i^2}\right] = E_{r_i}\left[= \frac{(u_1 v_1 + \cdots + u_n v_n)^2}{\sum u_i^2 \sum v_i^2}\right] \\ &= E_{r_i}\left[\frac{\sum_{i=1}^{n} u_i^2 v_i^2 + \sum \sum_{i \neq j} u_i u_j v_i v_j}{\sum u_i^2 \sum v_i^2}\right] \\ &= E_{r_i}\left[\frac{\sum_{i=1}^{n} u_i^2 v_i^2}{\sum u_i^2 \sum v_i^2}\right] + E_{r_i}\left[\frac{\sum \sum_{i \neq j} u_i u_j v_i v_j}{\sum u_i^2 \sum v_i^2}\right] \end{aligned}$$

As before, the u_i and u_j terms will remain fixed. However, the v_i and v_j will be permuted. So we can rewrite the last line as

$$\mathrm{Var}_{r_i}(r) = \frac{\sum_{i=1}^{n} u_i^2 \cdot E_{r_i}\left[v_i^2\right]}{\sum u_i^2 \sum v_i^2} + \frac{\sum \sum_{i \neq j} u_i u_j \cdot E_{r_i}\left[v_i v_j\right]}{\sum u_i^2 \sum v_i^2} \tag{26}$$

So we're left to compute the expectations in the above expression, recalling that the mean of the v_i is 0:

$$\begin{aligned} E_{r_i}\left[v_i^2\right] &= \sum_{j=1}^{n} v_j^2 / n \\ E_{r_i}\left[v_i v_j\right] &= \sum \sum_{k \neq \ell} \frac{v_k v_\ell}{n(n-1)} \end{aligned} \tag{27}$$

Finally, before we substitue back into Equation 26, one more identity that will prove useful:

$$\begin{aligned} \text{We know} \quad & 0 = u_1 + \cdots + u_n \\ \Rightarrow \quad & 0^2 = (u_1 + \cdots + u_n)^2 = \sum_{i=1}^{n} u_i^2 + \sum \sum_{i \neq j} u_i u_j \\ \Rightarrow & \sum \sum_{i \neq j} u_i u_j = -\sum_{i=1}^{n} u_i^2 \end{aligned}$$

Alright, so all that's left is to plug in the results from the last line and the expectations in Equations 27 into the variance expression, Equation 26:

$$\begin{aligned}
\mathrm{Var}_{r_i}(r) &= \frac{\sum_{i=1}^{n}\left[u_i^2 \cdot \left(\sum_{j=1}^{n} v_j^2/n\right)\right]}{\sum u_i^2 \sum v_i^2} + \frac{\sum\sum_{i\neq j}\left[u_i u_j \cdot \left(\sum\sum_{k\neq \ell} \frac{v_k v_\ell}{n(n-1)}\right)\right]}{\sum u_i^2 \sum v_i^2} \\
&= \frac{(\sum_{i=1}^{n} u_i^2)\left(\sum_{j=1}^{n} v_j^2/n\right)}{\sum u_i^2 \sum v_i^2} + \frac{\left(\sum\sum_{i\neq j} u_i u_j\right)\left(\sum\sum_{k\neq \ell} \frac{v_k v_\ell}{n(n-1)}\right)}{\sum u_i^2 \sum v_i^2} \\
&= \frac{1}{n} + \frac{(-\sum u_i^2)\left(-\sum \frac{v_i^2}{n(n-1)}\right)}{\sum u_i^2 \sum v_i^2} \\
&= \frac{1}{n} + \frac{1}{n(n-1)} = \frac{1}{n-1}
\end{aligned}$$

A. Proof of the Cramer-Rao Inequality

Using the notation given in the Cramer-Rao section above, we have some unbiased estimator $\hat{\tau}$ of $\tau(\theta)$ and we want to prove it's variance is bounded from below by the RHS of Equation 6. But first, we'll need two results.

Lemma A.1. *Suppose that we have any two random variables W and V with no restrictions on their joint density. Then we know*

$$-1 \leq Corr(W,V) \leq 1, \quad \Leftrightarrow \quad 0 \leq [Corr(W,V)]^2 \leq 1$$

$$\Rightarrow 0 \leq \frac{[Cov(W,V)]^2}{Var(W)Var(V)} \leq 1$$

$$Var(W) \geq \frac{[Cov(W,V)]^2}{Var(V)} \tag{28}$$

Lemma A.2. *Next, suppose that $EV = 0$. Then we know by the expansion of the covariance that*

$$Cov(W,V) = E[WV] - EW \cdot EV = E[WV] - EW \cdot 0$$

$$\Rightarrow Cov(W,V) = E[WV]$$

Using this fact, we can plug that result into Equation 28 to get

$$Var(W) \geq \frac{(E[WV])^2}{Var(V)} \tag{29}$$

again, provided that $EV = 0$.

Now, one more piece of shorthand. We'll being doing integration in n dimensions, so to simplify things, I'll write

$$d\mathbf{y} = dy_1 \; dy_2 \ldots dy_n$$

Now let's get to the proof.

Proof. To begin, we know that $\hat{\tau}$ is an unbiased estimator of $\tau(\theta)$, which means

$$\int \cdots \int_S \hat{\tau} \cdot f \; d\mathbf{y} = \tau(\theta) \tag{30}$$

Next, since f is a proper joint density function

$$\int \cdots \int_S f \; d\mathbf{y} = 1 \tag{31}$$

Now here's the only time we'll make a restrictive assumption. Namely, we assume the regularity condition that support of S **does not** depend upon θ. This allows us to

differentiate the LHS of 30 and 31 under the integral.[20]

So with that assumption, let's differentiate under the integrals with respect to θ. First do so for Equation 30:[21]

$$\int \cdots \int_S \hat{\tau} \cdot \frac{df}{d\theta} \, d\mathbf{y} = \frac{d\tau(\theta)}{d\theta} \tag{32}$$

Now we do the same for Equation 31

$$\int \cdots \int_S \frac{df}{d\theta} \, d\mathbf{y} = 0 \tag{33}$$

Next, we alter equations both Equations 32 and 33 with a simple identity, and then restate them

$$\int \cdots \int_S \hat{\tau} \left(\frac{1}{f} \frac{df}{d\theta} \right) f \, d\mathbf{y} = \frac{d\tau(\theta)}{d\theta}, \quad \Rightarrow \int \cdots \int_S \hat{\tau} \left(\frac{d(\ln f)}{d\theta} \right) f \, d\mathbf{y} = \frac{d\tau(\theta)}{d\theta} \tag{34}$$

$$\int \cdots \int_S \left(\frac{1}{f} \frac{df}{d\theta} \right) f \, d\mathbf{y} = 0, \quad \Rightarrow \quad \int \cdots \int_S \left(\frac{d(\ln f)}{d\theta} \right) f \, d\mathbf{y} = 0 \tag{35}$$

Finally, let's take Expression 35, differentiate with respect to θ using the chain rule, then shift things around:

$$\begin{aligned}
0 &= \frac{d}{d\theta} \left[\int \cdots \int_S \left(\frac{d(\ln f)}{d\theta} \right) f \, d\mathbf{y} \right] \\
&= \int \cdots \int_S \left(\frac{d^2(\ln f)}{d\theta^2} \right) f \, d\mathbf{y} + \int \cdots \int_S \left(\frac{d(\ln f)}{d\theta} \right) \frac{df}{d\theta} \, d\mathbf{y} \\
&= \int \cdots \int_S \left(\frac{d^2(\ln f)}{d\theta^2} \right) f \, d\mathbf{y} + \int \cdots \int_S \left(\frac{d(\ln f)}{d\theta} \right) \left(\frac{1}{f} \frac{df}{d\theta} \right) f \, d\mathbf{y} \\
&= \int \cdots \int_S \left(\frac{d^2(\ln f)}{d\theta^2} \right) f \, d\mathbf{y} + \int \cdots \int_S \left(\frac{d(\ln f)}{d\theta} \right)^2 f \, d\mathbf{y}
\end{aligned}$$

$$\Rightarrow \int \cdots \int_S \left(\frac{d(\ln f)}{d\theta} \right)^2 f \, d\mathbf{y} = - \int \cdots \int_S \left(\frac{d^2(\ln f)}{d\theta^2} \right) f \, d\mathbf{y} \tag{36}$$

Now, we're finally at the home stretch. We just start plugging in from here on out.

- So to start, we can interpret the LHS of 36 as the second moment for $d(\ln f)/d\theta$. We also know from 35 that the first moment, the expectation, is 0. Putting this together, we get the variance

$$Var \left(\frac{d(\ln f)}{d\theta} \right) = E \left[\frac{d^2(\ln f)}{d\theta^2} \right] = - \int \cdots \int_S \left(\frac{d^2(\ln f)}{d\theta^2} \right) f \, d\mathbf{y} \tag{37}$$

[20] Note, if the support, S, *does* depend upon θ, then **none** of this holds.
[21] Note that $\hat{\tau}$ is a function of the data, so it is constant with respect to θ.

- Next, we recycle the fact that the expectation of $d(\ln f)/d\theta$ is 0 from Equation 35. So we can use that along with Equation 34 to show that

$$Cov\left(\hat{\tau}, \frac{d(\ln f)}{d\theta}\right) = \frac{d\tau(\theta)}{d\theta} \tag{38}$$

- Finally, we use our lemmas and Equation 29, taking $W = \hat{\tau}$ and $V = d(\ln f)/d\theta$ to write

$$Var(\hat{\tau}) \geq \frac{-\left[\frac{d\tau(\theta)}{d\theta}\right]^2}{E\left[\frac{d^2(\ln f)}{d\theta^2}\right]}$$

which concludes the proof of the Cramer-Rao Inequality. QED, bitches. □

B. Example: Rao-Blackwell Theorem Part 2

As we mentioned before, the Rao-Blackwell Theorem (Part 2) is pretty awesome because it let's you get crazy with your estimators. Here's one example.

Observations We suppose that the data Y_i are iid Poission(θ), $i = 1, \ldots, n$.

Objective Function We want to find the MVU estimator of $e^{-\theta}$.

Sufficient Statistic We know that $W = \sum_{i=1}^{n}$ is a sufficient statistic for θ. Because of Rao-Blackwell (Part 2), we also know that if we find *some* unbiased estimator, X, for $e^{-\theta}$, then $E[X|W]$ is *the* MVU estimator.

Choice of X Here's where the craziness kicks in. Let's choose our unbiased estimator to be

$$X = \begin{cases} 1 & \text{if } Y_1 = 0 \\ 0 & \text{otherwise} \end{cases}$$

Now it's clear that this is a really, really stupid estimator of $e^{-\theta}$. It ignores almost all of the observations $(Y_2, \ldots, Y_n)$ and gives only two possible values. But whatever, Rao-Blackwell tells us to use it provided that it's unbiased, so let's just check that simple fact:

$$\begin{aligned} EX &= 1 \cdot e^{-\theta} + 0 \cdot [\text{all other outcomes}] \\ &= e^{-\theta} \end{aligned}$$

The expectation above follows because Y_1 is Poisson (as are all the other Y_i) and because we don't have to account for any other terms.

Apply Rao-Blackwell Now let's use the RB Theorem to get the MVU estimator:

$$E[X|W] = P(X = 1) = P(Y_1 = 0|W) \tag{39}$$

Now it's not hard to prove that if the Y_i are all iid Poisson and if we know the sum, then each of the Y_i are *equally likely* to have contributed any given count in W. Mathematically, this means that

$$Y_1|W \sim \text{Binom}(W, 1/n)$$

So we will use that for Equation 39:

$$\begin{aligned} E[X|W] &= P(Y_1 = 0|W) \\ &= \left(\frac{n-1}{n}\right)^W \end{aligned}$$

This is the unique MVU estimator of $e^{-\theta}$, which is pretty nuts.

C. Useful Tricks and Identities

The following identities are often used in deriving distributions or test statistics under the Neyman-Pearson and Lambda-Ratio tests. Here's the first.

$$\sum_{i=1}^{n}(y_i-\mu)^2=\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{y}-\mu)^2$$

Next, suppose we have k groups of observations, where the jth group has n_j observations. Then

$$\begin{aligned}\sum_{i=1}^{n_1}(y_{1i}-\bar{\bar{y}})^2+\cdots+\sum_{i=1}^{n_k}(y_{ki}-\bar{\bar{y}})^2=&\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2+\cdots+\sum_{i=1}^{n_k}(y_{ki}-\bar{y}_k)^2\\&+n_1(\bar{y}_1-\bar{\bar{y}})^2+\cdots+n_k(\bar{y}_k-\bar{\bar{y}})^2\end{aligned}$$

In the case where $k=2$, we can also write the above expression as

$$\sum_{i=1}^{n_1}(y_{1i}-\bar{\bar{y}})^2+\sum_{i=1}^{n_2}(y_{2i}-\bar{\bar{y}})^2=\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2+\sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2+\frac{nm}{n+m}(\bar{y}_1-\bar{y}_2)^2$$

Handy dandy thing for regression

D. Useful Statistics

t-Distribution: Let's detail some frequnetly used statistics. First, provided that the data comes from a $\mathrm{N}(\mu,\sigma^2)$ distribution. This statistic is useful if you know the mean (or specify it in a Hypothesis test), but don't know (or leave unspecified) the variance.

$$T=\frac{\bar{Y}-\mu}{S/\sqrt{n}}\qquad n-1 \text{ degrees of freedom}$$

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