Math Statistics 2020 (Sonja Petrović)- Course notes

Romà Domènech Masana A20450272

February 2020

Contents

1 Properties of the statistics

1.1 Estimator vs estimate

Definition 1.1 (Statistic) A statistic $T(\cdot)$ is a function of a given data set $\boldsymbol{X} = X_1, ..., X_n$. More in particular, an **estimator** is the function relating the different observations, and a **estimate** is the realized value. An estimate is the same function applied to the observed values $x_1, ..., x_n$.

1.2 Sufficient statistic

For me a sufficient statistic $T(X)$ of the data X, is a function of the dataset that captures all the relevant "information" of the dataset, in terms of the likelihood. I would like to remind everyone that the bigger the likelihood for a previously fixed dataset, the more "informative" that our estimated parameter $\hat{\theta}$ is. Or the closer that this estimation $\hat{\theta}$ is to the real population (and never known!) parameter θ.

For this definition to make sense, we must agree first on the role of the likelihood on a data sample $\boldsymbol{X} = X_1, ..., X_n$ originating from a parametric distribution f with real population (and unknown) parameters θ. Well so the role that we should all agree on is the following: *In theory, if we knew the real population parameters θ, once we take a sample $\boldsymbol{X}$ from the real population, we can compute the likelihood (in the discrete case it would be just the usual probability) of obtaining such sample $\boldsymbol{X}$ from the population, according to the probability density function $f(\boldsymbol{X}|\theta)$.*

The equation

$$f(\boldsymbol{X}|\theta) = h(\boldsymbol{X})g(T(\boldsymbol{X})|\theta), \tag{1}$$

guarantees that when we maximize the likelihood f depending on the parameters θ, the function that we are actually optimizing is

$$g(T(\boldsymbol{X})|\theta),$$

because when we equate the first derivative to zero the $h(\boldsymbol{X})$ function can be dropped and is irrelevant for the maximization. It only can tell us whether we found a minimum or a maximum (because its sign matters). Well actually no, because its sign must be positive, otherwise g would be negative and this makes little sense since g is the density function of the statistic T. Equation 1 is the Theorem 6.2.6 of page 276 of the Book "Statistical Inference - Casella, Berger." Also known as Factorization Theorem.

Example 1 Let $X_1, ..., X_n$ iid $N(\mu, \sigma^2)$. Now using factorization theorem we get

$$f(\boldsymbol{x}|\mu, \sigma^2) = h(\boldsymbol{x})g(T_1(\boldsymbol{x}), T_2(\boldsymbol{x})|\mu, \sigma^2),$$

where $(T_1(\boldsymbol{x}), T_2(\boldsymbol{x})) = (\hat{X}, S_n^2)$ is the statistic, and $h(\boldsymbol{x}) = 1$.

For instance, if we are throwing a dice on the ground, and we obtain a 4, (if the dice is not loaded!) the probability of obtaining such a number is 1/6. Now imagine a dice with infinite faces, in other words, a ball. Then the probability that the ball ends with a certain point x facing upwards is of course 0 for every point on the surface of the ball. Because the ball has so many and small faces that we can never guess which one is going to end up on the top. But certainly one of the points on the surface is going to face exactly upwards. It is in order to solve this problem that the likelihood (as a way of trying to generalize the probability) was devised. the likelihood that one of the points of the ball is going to face upwards

1.3 Ancillary statistic

$T(X)$ is an ancillary statistic of the parameter θ, if its marginal distribution

$$f(T(X)|\theta) = f(T(X)), \tag{2}$$

doesn't depend on that parameter θ.

For instance, sample mean and interquantile range are ancillary with respect to the population mean μ.

1.4 Bivariate variable transformations

Assume that $X, Y : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ are two random variables with joint pdf

$$f_{X,Y}(x, y),$$

and that we have the following bivariate transformations:

$$\begin{aligned} U &= g_1(X, Y) \\ V &= g_2(X, Y), \end{aligned}$$

for some functions $h_1, h_2 : \mathbb{R}^2 \to \mathbb{R}^2$. Assume as well that we can compute the inverses

$$\begin{aligned} X &= h_1(U, V) \\ Y &= h_2(U, V), \end{aligned}$$

for some functions $g_1, g_2 : \mathbb{R}^2 \to \mathbb{R}^2$. Then the bivariate distribution of (U, V) is:

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \begin{vmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{vmatrix}$$

If we want to generalize to a $n-$dimensional vector, equation (4.6.7) of page 185 of the book "Statistical Inference - Casella, Berger" provides the result on page 185, which I don't know over what value it iterates and sums.

1.5 Distribution of the order statistics

So another interesting result from the book is the pdf of the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ so that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ and all of them are sample points. We just sorted them.

On page 246 there is a Theorem (5.4.6) that says that if the random sample $X_1, ..., X_n$ comes from a population wiht CDF $F_X(x)$ and PDF $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$ is

$$f_{X_{(i)},X_{(j)}}(u, v) = \tfrac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j}, \tag{3}$$

where $-\infty < u < v < \infty$.

1.6 Anisodimensional variable transformations

That is, variable transformations

$$\begin{array}{rcl} d:\mathbb{R}^2 & \to & \mathbb{R} \\ (u,v) & \longmapsto & d(u,v) \end{array} \tag{4}$$

between spaces of different dimensions. For the case just stated above, my trick is to use the **equidimensional variable transformation** as explained in subsection 1.4, by filling the remaining missing dimensions with the identity. In the above case we would do

$$\begin{array}{rcl} d_2:\mathbb{R}^2 & \to & \mathbb{R}^2 \\ (u,v) & \longmapsto & (u,d(u,v)) \end{array} \tag{5}$$

and then we would compute the marginal distribution of the second variable of the resulting transformation. In other words, assume we know the joint distribution of two random variables $X,Y:\Omega\to\mathbb{R}$ and we want to compute the probability density function of

$$V := d(X,Y) = X/Y, \tag{6}$$

for instance. Then I would proceed in the following way. First use the bivariate transformation rules subsection 1.4 to compute:

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v),h_2(u,v)) \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix}, \tag{7}$$

where

$$V = d(X,Y), U = X \to x = h_1(u,v) = u, \tag{8}$$

and h_2 depends on what d is. Once we get

$$f_{U,V}(u,v),$$

since we are only interested in V just integrate over U and compute the marginal:

$$f_V(v) = \textstyle\int_{\mathbb{R}} f_{U,V}(u,v)du. \tag{9}$$

The whole picture is something like:

$$\begin{array}{ccccc} \mathbb{R}^2 & \xrightarrow{(\mathrm{Id},d)} & \mathbb{R}^2 & \xrightarrow{\int_{\mathbb{R}} du} & \mathbb{R} \\ (x,y) & \longmapsto & (x,d(x,y)) & \longmapsto & \text{marginal dist over } u \end{array} \tag{10}$$

Example 1.1 Assume that

$$V = d(X,Y) = X_{(1)}/X_{(n)},$$

where $X = X_{(1)}, Y = X_{(n)}$. Then we can set $U = X$, $V = X/Y$ and so using Equation 7 with

$$h_1(u,v) = u, \quad h_2(u,v) = u/v$$

$$\begin{array}{rcl} f_{X_{(1)},X_{(1)}/X_{(n)}}(u,v) & = & f_{X_{(1)},X_{(n)}}(u,u/v) \begin{vmatrix} 1 & 0 \\ 1/v & -u/v^2 \end{vmatrix} \\ & = & f_{X_{(1)},X_{(n)}}(u,u/v)\frac{-u}{v^2} \\ & \downarrow & \\ f_{X_{(1)}/X_{(n)}}(v) & = & \int_{\mathbb{R}} f_{X_{(1)},X_{(1)}/X_{(n)}}(u,v)du \\ & = & \int_{\mathbb{R}} f_{X_{(1)},X_{(n)}}(u,u/v)\frac{-u}{v^2}du \\ & = & \int_{\mathbb{R}} n(n-1)f_X(u)f_X(u/v)[F_X(u/v)-F_X(u)]^{n-2}\frac{-u}{v^2}du. \end{array} \tag{11}$$

Where the last inequality uses the distribution of the order statistics as explained in Equation 3 for $i=1, j=n$. Then we set everything up to the distribution of X. Note: this X is another random variable than the initial abstract X from the pair X,Y.

1.7 Minimal sufficient statistic

A statistic $T(X)$ is called *minimal sufficient* if any other sufficient statistic $T'(X)$ is a function of it, or in other words, if it can be written as

$$T(X) = r(T'(X)),$$

for some injective function r.

The aim of the minimum statistic is to obtain the biggest data reduction of all the sufficient statistics.

A minimal sufficient statistic partitions the data range into the fewest possible elements so that the statistic remains sufficient. Any other sufficient statistic creates a finer partition of the data space.

The theorem is the following:

Theorem 6.2.13 on minimal sufficient statistics of θ: let $f(\boldsymbol{x}|\theta)$ be the PDF of a sample $\boldsymbol{x}$. Now, suppose there exists a function $T(\boldsymbol{x})$ (the statistic) s.t. for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, we have the following relationship

$$\begin{gathered} f(x|\theta) = c(x,y)f(y|\theta) \\ \Updownarrow \\ T(x) = T(y). \\ \downarrow \\ \text{T(X)is a minimal sufficient statistic.} \end{gathered} \tag{12}$$

Example 2 Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two iid n-dimensional samples following a $N(\mu,\sigma^2)$ distribution, then I claim that the statistic $(T_1(\boldsymbol{x}), T_2(\boldsymbol{x})) = (\hat{X}, S_n^2)$ is a minimal sufficient statistic of the parameters (μ, σ^2). Let's see this:

Let $(\hat{x}, s_y^2)$ and $(\hat{y}, s_x^2)$ be the sample means and variances of corresponding to x and y, respectively. Then

$$\begin{gathered} \frac{f(\boldsymbol{x}|\mu,\sigma^2)}{f(\boldsymbol{y}|\mu,\sigma^2)} = \frac{e^{-\frac{n(\hat{x}-\mu)^2+(n-1)s_x^2}{2\sigma^2}}}{e^{-\frac{-n(\hat{y}-\mu)^2+(n-1)s_y^2}{2\sigma^2}}} = e^{\frac{-n[\hat{x}^2-\hat{y}^2]+2\mu[\hat{x}-\hat{y}]+(1-n)[s_x^2-s_y^2]}{2\sigma^2}} = c(x,y) \\ \Updownarrow \\ -n[\hat{x}^2 - \hat{y}^2] + 2\mu[\hat{x}-\hat{y}] + (1-n)[s_x^2 - s_y^2] = 0 \\ \Updownarrow \\ \hat{x} = \hat{y} \\ s_x^2 = s_y^2 \\ \Updownarrow \\ (T_1(x), T_2(x)) = (T_1(y), T_2(y)). \end{gathered} \tag{13}$$

Therefore the statistic $(T_1(x), T_2(x))$ is a minimal sufficient statistic.

1.8 Complete distribution

A set of density functions $f(t|\theta)$ for a statistic $T(X)$ is complete if for every function g such that

$$\begin{gathered} \mathbb{E}_\theta g(T) = 0 \\ \downarrow \\ g(T) \overset{a.s.}{=} 0 \end{gathered} \tag{14}$$

1.9 Equality in L_2 iff equality a.s.

For two random variables S_0, μ_0

$$S_0 = \mu_0 \text{ a.s. } \leftrightarrow \mathbb{E}|S_0 - \mu_0|^2 = 0 \tag{15}$$

1.10 Consistent estimator

$\hat{\theta}$ is a consistent estimator of θ iff

$$\hat{\theta} \xrightarrow{P} \theta$$

1.11 Efficient estimator

An estimator is efficient if its variance attains the **Cramér-Rao** value. I.e. if

$$\operatorname{Var} \hat{\theta} = CR(\hat{\theta}).$$

The next is corollary 7.3.10 of page 337 of the book.

Cramér-Rao inequality For $X_1, ..., X_n$ are iid with pdf $f(x|\theta)$ and let $T(\boldsymbol{X})$ be an estimator of the data, then

$$\operatorname{Var}_\theta T(\boldsymbol{X}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta T(\boldsymbol{X})\right)^2}{n\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(x_1|\theta)\right)^2} = CR(\hat{\theta}). \tag{16}$$

Note that usually our estimator or statistic $T(\boldsymbol{X})$ will satisfy:

$$\mathbb{E}_\theta T(\boldsymbol{X}) = \theta,$$

(if it is unbiased!) so that the nominator in Equation 16 will usually just be 1. And then, the lower bound for the variance of an unbiased estimator, estimating θ would be something like:

$$\operatorname{Var}_\theta T(\boldsymbol{X}) \geq \frac{\left(\frac{d}{d\theta}\theta\right)^2}{n\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(x_1|\theta)\right)^2} = \frac{1}{nI(\hat{\theta})}. \tag{17}$$

This is usually the Cramér-Rao bound for θ. If we are estimating a continuous function $\tau(\theta)$ of θ, so that our estimator $T(\boldsymbol{X})$ is unbiased for it:

$$\mathbb{E}_\theta T(\boldsymbol{X}) = \tau(\theta),$$

then we will simply have that

$$CR(\hat{\theta}) = \frac{\left(\tau'(\hat{\theta})\right)^2}{nI(\hat{\theta})}. \tag{18}$$

The following is theorem 10.1.6 page 470 of book:

Theorem 1.1 (consistency of MLE estimator) *the MLE estimator $\hat{\theta}_{MLE}$ is **consistent.***

2 4 ways of estimating the parameters

2.1 Method of moments

This one is fairly simple. We just equate the **sample moments** to the **population moments**. That's to say, if we have a sample $\boldsymbol{X} = X_1, ..., X_n$ of iid random variables, then we do:

$\mathbb{E}X_1 = \bar{\boldsymbol{X}}$

$\mathbb{E}X_1^2 = \frac{1}{n}\sum_1^n X_i^2$

$\mathbb{E}X_1^3 = \frac{1}{n}\sum_1^n X_i^3$

...

$\mathbb{E}X_1^m = \frac{1}{n}\sum_1^n X_i^m$

For the m first moments, in case that we have to estimate $\theta_1, ..., \theta_m$ m parameters. And then we solve these m linear equations and we should have m estimators $\hat{\theta}_1, ..., \hat{\theta}_m$ for all the parameters.

2.2 Maximum Likelihood Estimation

Consists on taking the likelihood $f_X(x_1, ..., x_n|\theta)$ of the random variable X to be studied, given n iid observations $x_1, ..., x_n$ from it, and compute the value $\hat{\theta}_{MLE}$ that maximizes it.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} f(x_1, ..., x_n|\theta).$$

Note that the likelihood

$$f_X(x_1, ..., x_n|\theta) = \prod_1^n f_X(x_i|\theta), \tag{19}$$

where $f_X(x_i|\theta)$ is the probability density function of the variable X. This is only true if the n observations are iid.

2.3 Bayes estimators - it is not a point estimation!

The bayes way of estimating is based on the following philosophy: first we have an initial guess $p(\theta)$ of where our population parameter θ may lie, and then, we update it using the likelihood $f(\boldsymbol{X}|\theta)$ of the data given the parameters θ:

$$p(\theta|\boldsymbol{X}) \propto f(\boldsymbol{X}|\theta)p(\theta) \sim F(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), ..., T_m(\boldsymbol{X}), \boldsymbol{X}), \tag{20}$$

where I call $F(...)$ the probability distribution of this posterior distribution of the parameters θ. In other words, this posterior function tells us how the parameters we are interested in θ are distributed according to some statistics of our data $T_1, ..., T_m$ and possibly the data itself $\boldsymbol{X}$.

If we want, we can collapse all the information that the posterior pdf gives us and consider only its mean. In other words, we can now get a point estimator, the following:

$$\hat{\theta}_B = \mathbb{E}_{p(\theta|X)}(\theta|X) \tag{21}$$

As a **note** I would like to say that this Bayes mean contains less information than the posterior distribution of the parameters θ! That's why I think the posterior distribution is much reacher than just its mean.

2.4 Expectation-Maximization algorithm

It is useful for solving missing data problems, and getting good estimates for θ. Instead of maximizing the likelihood

$$\max_{\theta} L(\boldsymbol{X}, \boldsymbol{Y}|\theta),$$

where X and Y are two data sets. Now assume that some of the observations X are missing, then instead of maximixing the *completel likelihood* we will integrate over the missing observation x_1, and maximize the result. In other words, assume that all $\boldsymbol{x}$ is missing. Assume $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then instead of maximizing

$$f(\boldsymbol{x}, \boldsymbol{y}|\theta), \tag{22}$$

we wanna integrate over the missing values $\boldsymbol{x}$ and maximize

$$L^c(\boldsymbol{y}|\theta) := \int_{\mathbb{R}} f(\boldsymbol{x}, \boldsymbol{y}|\theta)k(\boldsymbol{x}|\theta^{(0)}, \boldsymbol{y})d\boldsymbol{x} = \mathbb{E}_{k(\boldsymbol{x}|\theta^{(0)}, \boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y}|\theta). \tag{23}$$

where $k(\boldsymbol{x}|\theta^{(0)}, \boldsymbol{y})$ is the pdf of the missing data $\boldsymbol{x}$ given that we know the parameter $\theta^{(0)}$ and the observed data $\boldsymbol{y}$. But now if we use the *Bayes Rule,* we observe that we can actually find out the *missing data distribution*

$$k(\boldsymbol{x}|\theta^{(0)}, \boldsymbol{y}) = \frac{f(\boldsymbol{x}, \boldsymbol{y}|\theta^{(0)})}{g(\boldsymbol{y}|\theta^{(0)})}, \tag{24}$$

where $f(\boldsymbol{x}, \boldsymbol{y}|\theta)$ is the known or assumed distribution of our data, and $g(\boldsymbol{y}|\theta)$ is the marginal distribution of $\boldsymbol{y}$ once we integrate over dx. Now note the appearance of the parameter $\theta^{(0)}$. This one is different from θ. Actually, we must assume that we know the *missing data distribution* $k(\boldsymbol{x}|\theta^{(0)}, \boldsymbol{y})$ in order to integrate over it and compute our expected likelihood, without the missing values. Otherwise, if we don't know or don't assume any specific value for θ while using the pdf of the missing values $\boldsymbol{x}$, we just can't integrate in practice, and can't cancel out the missing terms.

Once we cancelled out these missing terms, we can maximize the result and compute:

$$\hat{\theta}^{(1)} = \arg\max_\theta \mathbb{E}_{k(\boldsymbol{x}|\theta^{(0)},\boldsymbol{y})} f(\boldsymbol{x},\boldsymbol{y}|\theta), \tag{25}$$

and we will be done. Or not! Because now that we have a $\hat{\theta}^{(1)}$ we can use it to refine our missing value density function $k(\boldsymbol{x}|\boldsymbol{y},\theta)$ into $k(\boldsymbol{x}|\boldsymbol{y},\theta^{(1)})$ and repeat the step of Equation 25.

Note 1 I think that in theory at least, we could decide not to do this iterations, and just assume the distribution

$$k(\boldsymbol{x}|\boldsymbol{y},\theta)$$

for the missing values $\boldsymbol{x}$. But then I suspect that the problem is how to maximize

$$\mathbb{E}_{k(\boldsymbol{x}|\theta,\boldsymbol{y})} f(\boldsymbol{x},\boldsymbol{y}|\theta). \tag{26}$$

I think this might be a problem becase the above function is an integral of a function depending on θ multiplied by another function depending on θ. Maybe it is difficult to maximize an integral? Otherwise I don't see why we don't directly maximize Equation 26.

3 Ways of evaluating estimators

3.1 Mean Squared Error

Given an estimator

$$\hat{\theta}$$

or function from the data, that aims at estimating a parameter θ, we can actually compute its MSE by

$$MSE(\hat{\theta}) := \mathbb{E}_{p(\hat{\theta}|\theta)}(\hat{\theta}-\theta)^2 = \int_{\mathbb{R}}(\hat{\theta}-\theta)^2 p(\hat{\theta}=x|\theta)dx \tag{27}$$

It is also worth noting that

$$\begin{aligned} MSE(\hat{\theta}) \quad &= \mathbb{E}_{p(\hat{\theta}|\theta)}(\hat{\theta}-\theta)^2 = \mathbb{E}_{p(\hat{\theta}|\theta)}(\hat{\theta}-\mathbb{E}\hat{\theta})^2 + \mathbb{E}_{p(\hat{\theta}|\theta)}(\mathbb{E}\hat{\theta}-\theta)^2 \\ = \quad & \mathrm{Var}(\hat{\theta}) + Bias(\hat{\theta})^2. \end{aligned} \tag{28}$$

3.2 Unbiased estimators

Are those estimators $\hat{\theta}$ of θ such that its Bias is 0. Among these we might wonder which are the ones that minimize the MSE, but this is equivalent to minimizing

$$\mathrm{Var}_{p(\hat{\theta}|\theta)}(\hat{\theta}).$$

We call **best unbiased estimator** to those *unbiased* estimators T^* of θ that satisfy

$$\mathrm{Var}_{p(T^*|\theta)}\, T^* \leq \mathrm{Var}_{p(\hat{\theta}|\theta)}\, \hat{\theta}, \tag{29}$$

for any unbiased estimator $\hat{\theta}$ of θ. We know from Equation 16 that this value can never be smaller than the Cramér-Rao bound

$$CR(\theta) = I^{-1}(\theta).$$

And therefore if an estimator $\hat{\theta}$ is unbised and achieves this lower bound for its variance, it must automatically by a *best unbiased estimator.*

From Equation 16 we know that this is achieved by the *efficient estimators* that are *unbiased*, if there are any efficient estimators for that parameter. Because it could well be that no efficient estimator is available for a given parameter. Then, how do we know if we attained an optimal one? How do we know if we found a best unbiased estimator?

3.3 Conditioning on sufficient statistic decreases variance

That is, given an unbiased estimator T of a parameter θ, we can take a sufficient statistic T_s and then we get that:

$$\phi(T_s) := \mathbb{E}(T|T_s), \tag{30}$$

satisfies both equations:

$$\mathbb{E}\phi(T_s) = \mathbb{E}\mathbb{E}(T|T_s) = \mathbb{E}T = \theta, \text{ and}$$

$$\begin{gathered} \operatorname{Var}(\phi(T_s)) + \mathbb{E}\operatorname{Var}(T|T_s) = \operatorname{Var}(T) \\ \downarrow \\ \operatorname{Var}(\phi(T_s)) \leq \operatorname{Var}(T), \end{gathered} \tag{31}$$

using the tower property of the conditional expectation, and the total variance equation detailed at Equation 42. That is, we improve the MSE value by just conditioning our unbiased estimator T on a sufficient statistic T_s. This literally means that when estimating our parameter θ we can *just use functions of sufficient statistics* and not the data itself.
Also, this is only true for sufficient estimators T_s because only then is

$$\phi(T_s)$$

independent of he parameter θ. But this is true if we compute the conditional expectation $\mathbb{E}(T|T_s)$ and use Equation 1 with $\boldsymbol{X} = T$ and $T = T_s$.
The above is theorem 7.3.17 of the notes (**Rao-Blackwell**) page 342 of Casellas- Berger book.

3.4 Loss functions and risks

The MSE is a specific loss function. A **loss function** is any function f that relates an attained value of our estimator $\hat{\theta}$ and our target value θ and tells us an approximation of how good our estimator did in estimating θ. We would expect from a *loss function* that the closer our *estimate* is to the true parameter, the smaller the loss function's value. In the ideal case, if our estimator has an estimated value that equals our target parameter θ, the loss function

$$L(\hat{\theta}, \theta) = 0.$$

Examples of loss functions are

1. $$L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$$

2. $$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

The **risk function** is the value that tells us how good was our *estimator* in estimating the parameter θ. It is defined as the expected (or average) loss. In other words:

$$\mathbb{E}_{p(\hat{\theta}|\theta)} L((\hat{\theta}, \theta).$$

In the previous two examples it will be

1. $$\mathbb{E}_{p(\hat{\theta}|\theta)} L(\hat{\theta}, \theta) = \mathbb{E}_{p(\hat{\theta}|\theta)} |\hat{\theta} - \theta| = MAE(\hat{\theta})$$

2. $$\mathbb{E}_{p(\hat{\theta}|\theta)} L(\hat{\theta}, \theta) = \mathbb{E}_{p(\hat{\theta}|\theta)} (\hat{\theta} - \theta)^2 = MSE(\hat{\theta})$$

4 Hypotheses testing

4.1 likelihood ration

We consider

1. $H_0 : \theta = \theta_0$
2. $H_0 : \theta \neq \theta_0$

We could compare the two taking a look at the pdf of the data given that parameter, and considere the rejecting region

$$R = \{w \in \Omega : f(\boldsymbol{X}|\theta_0) < af(\boldsymbol{X}|\hat{\theta}_{MLE})\}, \tag{32}$$

or

$$R = \{w \in \Omega : \frac{\sup_{\theta\in\Theta_0} f(\boldsymbol{X}|\theta)}{\sup_{\theta\in\Theta} f(\boldsymbol{X}|\theta)} < a\}, \tag{33}$$

where $a \in [0, 1]$. The statistic would here be

$$T(\boldsymbol{X}) := \frac{\sup_{\theta\in\Theta_0} f(\boldsymbol{X}|\theta)}{\sup_{\theta\in\Theta} f(\boldsymbol{X}|\theta)}, \tag{34}$$

This would tell us, for instance let $a = 0.8$, that if the null likelihood is smalller than 80% of the Maximum Likelihood Estimation likelihood, then reject the null hypotheses. Otherwise keep it. In order to select an appropiate a, we should take a look at a few things. The first of them is...

This is the sketch idea to start with. Now, to be much more precise and practical, we will tipically be wishing $T(\boldsymbol{X})$ to follow some probability distribution $F(\theta)$ so that we can say things like:

$$\begin{gathered} F_{T((X))}(a) = \Pr(T(\boldsymbol{X}) < a) = 0.05 \\ \downarrow \\ a = 0.86. \end{gathered} \tag{35}$$

This is the type of game we will be playing with any statistic and any hypothesis testing all the time!. In the above example we would have found a boundary value $a = 0.86$ that tells us that that statistic T is only smaller than it with %5 chance. Or, in other words, it is *pretty unprobable* that under the null hypotheses, our statistic T attains a value smaller than $a = 0.86$. *Thus, if T obtains a value smaller than $a = 0.86$ it is very unprobable that the null hypotheses is true.*

5 making mathematical animations- from Jake Mcclure

https://github.com/3b1b/manim

6 Asymptotic behaviour

This is definition 10.1.11 on page 471 of the *asymptotic efficency of an estimator.*

Definition 6.1 (asymptotic efficiency) A sequence T_n of estimators of a continuous function $\tau(\cdot)$ of a parameter θ is *asymptotically efficient* for $\tau(\theta)$ if

$$\sqrt{n}\,(T_n - \tau(\theta)) \xrightarrow{d} \mathcal{N}[0,\ \ (\tau'(\theta))^2 I^{-1}(\theta)\ \ ] \tag{36}$$

Note that in particular this means T_n is unbiased! Usually we take

$$T_n = T(\boldsymbol{X}),$$

where $\boldsymbol{X}$ are n sampled observations.

The following is theorem 10.1.12 of page 472 of the book.

Theorem 6.1 (Asymptotic efficiency of MLE's) The MLE is an *efficient estimator*. In other words, let $X_1, X_2, ...$ be iid with pdf $f(x|\theta)$ and $\hat{\theta}_{MLE}$ the max. likelihood estimator of θ and $g(x)$ a continuous function. is $g : \mathbb{R}^d \to \mathbb{R}^d$ for any $d \in \mathbb{N}$? Then, if we assume good enough regularity conditions for the pdf $f(x|\theta)$, then

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \to \mathcal{N}(0, (g'(\theta))^2 I^{-1}(\theta)) \tag{37}$$

7 Max likelihood estimation

Prof. Uploaded a theorem (theorem 6.1.2 page 324) of her notes and theorem 7.2.10 of the book (page 320) that I call

Theorem 7.1 (function of MLE estimators is MLE estimator) *Let $X_1, ..., X_n$ be iid with pdf $f(x|\theta), \theta \in \Omega$. Let's now assume that we are interested in estimating a parameter η which is a function of our parameters θ, in the following way:*

$$\eta = g(\theta)$$

for some function g. (is g measurable? or continuous?) Then

$$\hat{\eta}_{MLE} = g(\hat{\theta}_{MLE}) \tag{38}$$

8 EM-Algorithm

We define $g(x|\theta)$ the joint pdf of X
$h(X, Z|\theta)$ joint pdf of observed + unobserved
$k(z, x|\theta)$ conditional pdf of unobserved. We know that

$$k(z|x, \theta) = \frac{h(X, Z|\theta)}{g(x|\theta)}$$

. Now define the observed likelihood as

$$L(\theta|x) = g(x|\theta).$$

The complete likelihood is

$$L^c(\theta|x, z) = h(x, z|\theta)$$

goal: maximize $L(\theta|x)$ using $L^c(\theta|x, z)$. Fix an arbitrary $\theta_0 \in \Omega$. Then

$$\begin{aligned}
\log L(\theta|X) &= \int \log L(\theta|X) k(z|\theta_0, x) dz \\
&= \int \log g(x|\theta) k(z|\theta_0|x) dz? \\
&= \int [\log h(x, z|\theta) - \log k(z|x, \theta)] \cdot k(z|x, \theta) dz \\
&= \int h(x, z\theta) k(z|x, \theta) dz - \int (\log k(z|x, \theta)) k(z|x, \theta_0) dz \\
&= \mathbb{E}_{\theta_0} [\log L^c(\theta|x, z)|\theta_0, x] \text{ expectation under the conditional } k(z|\theta, x) \\
&\quad - \mathbb{E}_{\theta_0} [\log k(z|\theta, x)|\theta_0, x]
\end{aligned} \tag{39}$$

Define 1st summand

$$Q(\theta|\theta_0, x) = \mathbb{E}_{\theta_0} [\log L^c(\theta|x, z)|\theta_0, x] \tag{40}$$

THis expectation defines the E step fo the EM algorithm. The$M$ refers to the maximization of $Q(\theta|\theta_0, x)$.

The next is theorem 6.1 from handout

Note 1 We cannot know the complete likelihood. However, if we take the expectation, the unobserved data *integrates out.*

Theorem 8.1

$$\max Q$$

suffices to

$$\log L(\theta|X)$$

8.1 EM Algorithm in detail

Denote

1. $\hat{\theta}^{(0)}$ = initial estimate of θ (based on observed likelihood)
2. let $\hat{\theta}^{(1)}$ the argument that maximizes $Q(\theta|\hat{\theta}^{(0)}, x)$
3. iterate: let $\hat{\theta}^{(m)}$ be the estimate for the $m-$th step. Then compute $\hat{\theta}^{(m+1)}$: E:

$$\begin{aligned} &\text{compute } Q(\theta|\hat{\theta}^{(m)}, x) \\ \text{M: } \hat{\theta}^{(m+1)} &= \arg\max Q(\theta|\hat{\theta}^{(m)}, x) \end{aligned} \tag{41}$$

Homework due 6.6.1 and 6.6.2 from handout Do this, it will be graded.

Maybe use the **goodness of fit test.**

9 Annex 1

9.1 Total variance equation

$$\text{Var}(X) = \text{Var}(\mathbb{E}(X|Y)) + \mathbb{E}(\text{Var}(X|Y)) \tag{42}$$

10 First exercise

If $X_1, ..., X_n$ are iid $N(0, \theta)$, $0 < \theta < \infty$. Show that $T(\boldsymbol{X}) = \sum_{i=1}^n X_i^2$ is s sufficent statistic for θ.

We compute the likelihood of $\boldsymbol{X} = X_1, ..., X_n$;

$$\begin{aligned} f(\boldsymbol{X} = (x_1, ..., x_n)|\theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_i^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{T(\boldsymbol{x})}{2\sigma^2}} \\ &= g(T(\boldsymbol{X})|\theta), \end{aligned} \tag{43}$$

where $h(\boldsymbol{X}) = 1$ in this case. Thus by Equation 1 or theorem 6.2.2 of page 274, $T(X)$ is a sufficient statistic.

11 Ex 6.12

N is a random variable that takes values on the natural numbers $\mathbb{N}$. The thing is, $\mathbb{P}(N = i) = p_i$, $\forall i \in \mathbb{N}$. Once we picked up a value for N, let's say $N = n$, then we take exactly n Bernoulli trials with success probability θ. Assume that the data points are $y_1, ..., y_n \in \{0, 1\}$ (1 if there is success and 0 if there is not) and that we set the statistic

$$T_1(Y) = \sum_1^n y_i.$$

as the number of successes.

(a) We must see that the pair (X, N) is a minimal sufficient statistic and N is ancillary.

But the second part is easy since N does not depend on θ, this is why it is ancilliary. In other words, if we change our value for θ, only the realizations $y_1, .., y_n$ are affected, but N was computed before these realizations took place (and does not depend on them), therefore N is an ancillary statistic.

For the first part note that

$$\begin{aligned} f((\boldsymbol{y},n)|\theta) &= f(Y_1=y_1,...,Y_n=y_n,N=n|\theta) \\ &= \Pr(Y_1=y_1,...,Y_n=y_n|N=n|\theta)\Pr(N=n|\theta) \\ &= \Pr(Y_1=y_1,...,Y_n=y_n|N=n|\theta)\Pr(N=n), \end{aligned} \tag{44}$$

the last equality is due to the fact that N is ancillary. Now

$$\begin{aligned} f((\boldsymbol{y},n)|\theta) &= p_n\theta^{y_1}(1-\theta)^{1-y_1}...\theta^{y_n}(1-\theta)^{1-y_n} \\ &= p_n\theta^{y_1+...+y_n}(1-\theta)^{n-(y_1+...+y_n)} \\ &= p_n\theta^{T_1(Y)}(1-\theta)^{n-T_1(Y)}. \end{aligned} \tag{45}$$

Now let's take another sample $(N=m,z_1,...,z_m)$ then we obtain the likelihood

$$f((\boldsymbol{z},m)|\theta) = p_m\theta^{T_1(z)}(1-\theta)^{m-T_1(z)},$$

therefore

$$\begin{gathered} \frac{f((\boldsymbol{y},n)|\theta)}{f((\boldsymbol{z},m)|\theta)} = \frac{p_n}{p_m}\theta^{T_1(y)-T_1(z)}(1-\theta)^{n-m+T_1(x)-T_1(z)} = c(\boldsymbol{y},n,\boldsymbol{z},m) \\ \Updownarrow \\ n=m \\ T_1(y)=T_1(z), \end{gathered} \tag{46}$$

therefore the duplet $(N,T_1(x))$ is a minimal sufficient statistic.

(b) Prove that the estimator $T_1(x)/N$ is unbiased for θ and has variance equal to $\theta(1-\theta)\mathbb{E}(1/N)$

Approach one 1: assuming X depends on N:

To find this use that

$$\begin{gathered} S_n = a+2a^2+3a^3+...+na^n, \\ S_n - aS_n = a+a^2+a^3+...+a^n = \frac{a(1-a^n)}{1-a}, \end{gathered} \tag{47}$$

therefore

$$S_n = a+2a^2+3a^3+...+na^n = \frac{a(1-a^n)}{(1-a)^2}. \tag{48}$$

But as far as I got was to try:

$$\begin{aligned} \mathbb{E}(T_1(x)/N) &= \sum_{n\in\mathbb{N}}\sum_{t=1}^{n}\frac{t}{n}p_n\theta^t(1-\theta)^{n-t} \\ &= \sum_{n\in\mathbb{N}}\frac{(1-\theta)^n}{n}p_n\sum_{t=1}^{n}\frac{t\theta^t}{(1-\theta)^t}. \end{aligned} \tag{49}$$

Now

$$\begin{aligned} \sum_{t=1}^{n} t\left(\frac{\theta}{(1-\theta)}\right)^t &= \frac{\frac{\theta}{1-\theta}\left(1-\left(\frac{\theta}{1-\theta}\right)^n\right)}{\left(1-\frac{\theta}{1-\theta}\right)^2} \\ &= \frac{(1-\theta)\theta}{(1-2\theta)^2}(1-\frac{\theta}{1-\theta})^n \\ &= k_1(1-\frac{\theta}{1-\theta})^n, \end{aligned} \tag{50}$$

so then

$$\mathbb{E}(T_1(x)/N) = k_1\sum_{n\in\mathbb{N}}\left[\frac{p_n}{n}[(1-\theta)^n-\theta^n]\right], \tag{51}$$

but I honestly have no idea of how to compute this last sum. And I don't see how it can yield θ.

Then

$$\mathrm{Var}(T/N) = \mathbb{E}(\frac{T^2}{N^2}) - \mathbb{E}(\frac{T}{N})^2 \tag{52}$$

I also don't know how to compute this.

12 Ex 6.15

Let $X_1, ..., X_n$ be iid $\mathbb{N}(\theta, a\theta)$ where a is a known constant and $\theta > 0$ (also $a > 0$). Then:

(a) Show that the parameter space does not contain any two dimensional open subset:

Well the parameter space is $\{(\theta, a\theta) : \theta \in (0, \infty)\}$ and a is fixed $\subset <(1, a)>$ which means, the parameter space is contained in a 1 dimensional vector space. By definition, a 1 dimensional vector space has 1 dimension, and therefore contains no subsets of dimension 2. Then it also contains no opens subsets of dimension 2.

(b)

13 HW 6 Roma Domenech Masana

13.1 Ex1

(a) Compute the Cramér-Rao bound of a Cauchy distribution.

We have the pdf:

$$f(x|\theta) = \frac{1}{\pi^n \prod_1^n (1+(x_i-\theta)^2)}, \tag{53}$$

then

$$l(x|\theta) = \log f(x|\theta) = -n \log \pi - \sum_2^n \log(1+(x_i-\theta)^2), \tag{54}$$

then

$$\frac{\partial}{\partial \theta} l(x|\theta) = 2 \sum_1^n \frac{x_i-\theta}{(1+(x_i-\theta)^2)}, \tag{55}$$

which implies

$$\frac{\partial^2}{\partial \theta^2} l(x|\theta) = 2 \sum_1^n \frac{(x_i-\theta)^2-1}{(1+(x_i-\theta)^2)^2}, \tag{56}$$

then

$$\begin{aligned} I(\theta) &= -\mathbb{E}\left(\frac{\partial}{\partial \theta} l(x|\theta)\right) = -2 \sum_1^n \mathbb{E}\left(\frac{(x_i-\theta)^2-1}{(1+(x_i-\theta)^2)^2}\right) \\ &= -2 \sum_1^n -1/4 = n/2, \end{aligned} \tag{57}$$

since

$$\begin{aligned} \mathbb{E}\left[\frac{(x_i-\theta)^2-1}{(1+(x_i-\theta)^2)^2}\right] &= 1/\pi \int_{\mathbb{R}} \frac{(x_i-\theta)^2-1}{(1+(x_i-\theta)^2)^3} dx_i \\ &= 1/\pi \int_{\mathbb{R}} \frac{(z)^2-1}{(1+(z)^2)^3} dz \end{aligned} \tag{58}$$

using the change of variables

$$z = x_i - \theta$$

now we have that

$$\begin{aligned} I_k &= \int_{\mathbb{R}} \frac{1}{(1+z^2)^k} dz \\ &= \int_{\mathbb{R}} \frac{1}{(1+z^2)^{k+1}} dz + \int_{\mathbb{R}} \frac{z^2}{(1+z^2)^{k+1}} dz \\ &= I_{k+1} + \int_{\mathbb{R}} \frac{-2kz}{(1+z^2)^{k+1}} \frac{-z}{2k} dz \end{aligned} \tag{59}$$

Now, take

$$\begin{gathered} du = \frac{-2kz}{(1+z^2)^{k+1}} dz \\ v = \frac{-z}{2k}, \\ \downarrow \\ u = \frac{1}{(1+z^2)^k} \\ dv = -2k \\ \downarrow \\ \int_{\mathbb{R}} \frac{-2kz}{(1+z^2)^{k+1}} \frac{-z}{2k} dz = 0 + \frac{1}{2k} \int_{\mathbb{R}} \frac{1}{(1+z^2)^k} dz = I_k, \end{gathered} \tag{60}$$

so

$$\begin{gathered} I_k = I_{k+1} + \frac{1}{2k} I_k, \\ \downarrow \\ I_2 = I_3 + 1/4 I_2 \\ I_1 = I_2 + 1/2 I_1 \\ \downarrow \\ I_2 = 1/2 I_1 \\ I_3 = \frac{3}{4} I_2 \end{gathered} \tag{61}$$

and

$$I_1 = \arctan |_{-\infty}^{\infty} = \pi, \tag{62}$$

Then item 58 becomes

$$\begin{aligned} 1/\pi \int_{\mathbb{R}} \frac{(z)^2-1}{(1+(z)^2)^3} dz &= \frac{1}{\pi}\left[\int_{\mathbb{R}} \frac{(z)^2+1}{(1+(z)^2)^3} dz - 2\int_{\mathbb{R}} \frac{1}{(1+(z)^2)^3} dz\right] \\ &= \frac{1}{\pi}(I_2 - 2I_3) \\ &= \frac{1}{\pi}(I_2 - 2\frac{3}{4}I_2) \\ &= \frac{1}{\pi}(I_2 - \frac{3}{2}I_2) \\ &= \frac{-1}{2\pi}I_2 \\ \\ &= \frac{-1}{4\pi}I_1 \\ &= \frac{-1}{4\pi}\pi \\ &= -1/4. \end{aligned} \tag{63}$$

Therefore, the Cramér-Rao bound is

$$CR(\theta) = I^{-1}(\theta) = n/2. \tag{64}$$

(b) Compute the asymptotic distribution of the MLE estimator $\hat{\theta}_{MLE}$.
A: By theorem 10.1.12 of page 472 of the book of Casella - Berger, the MLE estimator is *efficient* and therefore we have the following limiting distribution:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{n\to\infty} \mathcal{N}(0, I^{-1}(\theta)) = \mathcal{N}(0, 2/n). \tag{65}$$

13.2 Ex 8.1

We have $n = 1000$ tosses of a coin. We also have that

$$S := \sum_1^n x_i = 560$$

and

$$n - S = 440.$$

Can we assume that the coin is fair?

Answer: **No.** This is because if we assume the coin $X \sim Bern(\theta)$, then we can make the followign hypotheses test:

1. $H_0 : \theta = \theta_0 = 1/2$
2. $H_a : \theta \neq 1/2$

So we tried to summarize the question of the problem into making an hypotheses. Now, if $\theta = 1/2$, we would expect half of the tosses to deliver heads and half of the times to get a tail. In particular, we would expect 500 heads, and any deviation from it could be regarded as the null hypotheses $H_0 : \theta = 1/2$ not holding. We will consider the following test statistic:

$$T = |S - \theta_0 n| \geq c, \tag{66}$$

for some value c that will be our *rejection value* for our hypotheses. That's to say, if $T > c$ then we say that H_0 doesn't hold. We will try to decrease the type 1 error as much as possible, to this end, we impose the type 1 error to be smaller than 1%:

$$\alpha = 0.01.$$

This gives us a rejection value c that is the solution to the following equation, where S is under the null hypotheses $S \sim Bern(\theta_0 = 1/2)$:

$$\Pr(T = |S - \theta_0 n| \geq c) \leq \alpha \tag{67}$$

But

$$\begin{aligned} |S-\theta_0 n| \geq c & \quad \Leftrightarrow S-\theta_0 n \geq c \text{ OR } S \leq \theta_0 n - c \\ & \quad \downarrow \\ \Pr(T \geq c) = & \quad \Pr(S \leq \theta_0 n - c) + \Pr(S \geq \theta_0 n + c) \\ = & \quad F_S(\theta_0 n - c) + 1 - F_S(\theta_0 n + c) \\ \leq & \quad \alpha = 0.01, \end{aligned} \tag{68}$$

for

$$c \geq 41,$$

in particular we can choose

$$c_{0.01} = 41$$

so that we minimize the type 2 error as well. A bigger c would give us a smaller α or type 1 error, but would increase the chances of keeping the null hypotheses when there are no grounds to assume it.

This all means that if

$$T > 41$$

then we can reject the null hypotheses with a type 1 error of 1%. But according to our data

$$T = S - 500 = 560 - 500 = 60 > 41 = c_{0.01},$$

so **the null hypotheses is rejected** and it is not reasonable to claim that the coin is fair.

Code:

```
#from book Csella Berger page 402:
source("Roma/R/R useful functions/inverse1.R")

S = 560
n = 1000
teta0=0.5
tetmle = S/n # this value is
alpha=0.01

#just test whether s~500 or not, i.e.

c=0:1000
pr = pbinom(teta0*n-c, size=1000, prob=teta0) +
  1-pbinom(teta0*n+c, size=1000, prob=teta0)
alpha=0.01
c1=ceiling(inverse(c,pr,alpha))
c1#c1=41
#we check that
(pbinom(teta0*n-c1, size=1000, prob=teta0) +
  1-pbinom(teta0*n+c1, size=1000, prob=teta0) < alpha)
# So for c1=41 upwards, the chance of type 1 error is <1%
# to minimize type 2 error let's take this c=41

#now,
T = abs(560-500)
T
if(T>c1){ print(paste("the test rejects H0 with type 1 error being",
                      100*alpha,"%"))
} else print("null hypotheses cannot be rejected")
```

13.3 Ex 8.6

If $X_1, ..., X_n \sim exp(\theta)$, $Y_1, ..., Y_m \sim exp(\mu)$.

(a) Find the LRT of $H_0 : \theta = \mu$ versus $H_1 : \theta \neq \mu$

$$\lambda(x,y) = \frac{L(\theta=\mu,\mu|X,Y)}{L(\theta,\mu|X,Y)} = \frac{\hat{\mu}_1^{n+m} e^{-\hat{\mu}_1(\sum_1^n x_i + \sum_1^m y_i)}}{\hat{\theta}^n e^{-\hat{\theta}\sum_1^n x_i} \hat{\mu}_2^m e^{-\hat{\mu}_2 \sum_1^m y_i}}, \tag{69}$$

where the estimators $\hat{\theta}, \hat{\mu}_1, \hat{\mu}_2$ are the MLE estimators:

$$\begin{aligned} \hat{\mu}_1 = & \frac{n+m}{\sum_1^n x_i + \sum_1^m y_i}, \\ \hat{\theta} = & \frac{n}{\sum_1^n x_i}, \\ \hat{\mu}_2 = & \frac{m}{\sum_1^m y_i}, \end{aligned} \tag{70}$$

but then:

$$\begin{aligned} \lambda(x,y) = & \frac{\hat{\mu}_1^{n+m} e^{-n-m}}{\hat{\theta}^n e^{-n} \hat{\mu}_2^m e^{-m}} \\ = & \frac{\hat{\mu}_1^{n+m}}{\hat{\theta}^n \hat{\mu}_2^m} \\ = & \left(\frac{\hat{\mu}_1}{\hat{\theta}}\right)^n \left(\frac{\hat{\mu}_1}{\hat{\mu}_2}\right)^m \\ = & \left(\frac{\frac{n+m}{\sum_1^n x_i + \sum_1^m y_i}}{\frac{n}{\sum_1^n x_i}}\right)^n \left(\frac{\frac{n+m}{\sum_1^n x_i + \sum_1^m y_i}}{\frac{m}{\sum_1^m y_i}}\right)^m \\ = & \left(\frac{n+m}{n} \frac{\sum_1^n x_i}{\sum_1^n x_i + \sum_1^m y_i}\right)^n \left(\frac{n+m}{m} \frac{\sum_1^m y_i}{\sum_1^n x_i + \sum_1^m y_i}\right)^m \\ = & \left(\frac{n+m}{n} T(x,y)\right)^n \left(\frac{n+m}{m}[1 - T(x,y)]\right)^m, \end{aligned} \tag{71}$$

where

$$T(x,y) = \frac{\sum_1^n x_i}{\sum_1^n x_i + \sum_1^m y_i}. \tag{72}$$

(b) The last section also answers this part.

(c) Find the distribution of T given that $H_0 : X \sim exp(\mu)$ Since

$$\sum_1^n x_i \sim \Gamma(n, \mu),$$

we have that

$$\frac{2}{\mu} \sum_1^n x_i \sim \Gamma(n, 2) = \chi^2(2n). \tag{73}$$

Similarly for $\sum_1^m y_i$ we have that

$$\frac{2}{\mu} \sum_1^m y_i \sim \Gamma(m, 2) = \chi^2(2m). \tag{74}$$

This gives us an statistic

$$\begin{aligned} T = & \frac{\sum_1^n x_i}{\sum_1^n x_i + \sum_1^m y_i} \\ = & \frac{\frac{2}{\mu}\sum_1^n x_i}{\frac{2}{\mu}\sum_1^n x_i + \frac{2}{\mu}\sum_1^m y_i} \\ \sim & \frac{\chi^2(2n)}{\chi^2(2n) + \chi^2(2m)} \\ \sim & \frac{\chi^2(2n)}{\chi^2(2n+m)}. \end{aligned} \tag{75}$$

So

$$\begin{aligned} \frac{n+m}{n} T \sim & \frac{\chi^2(2n)/2n}{\chi^2(2n+m)/2(n+m)} \\ \sim & F(2n, 2(n+m)), \end{aligned} \tag{76}$$

The F of Fisher-Snedecor distribution.

13.4 Ex 8.14

We have $X \sim \text{Bern}(p)$

1. $H_0 : p_0 = 0.49$
2. $H_a : p_1 = 0.51$

Use the CLT to estimate how big your sample n should be so that the type 1 and type 2 errors are smaller than $\alpha = 0.01$. Use a test statistic like

$$T = \sum_1^n x_i \geq c.$$

Okay, so let's compute the type 1 error

$$\begin{array}{rl} Err_1 = & \Pr(\sum_1^n x_i \geq c | H_0 : p = p_0 = 0.49) \leq 0.01 \\ & \downarrow \\ = & \Pr(\frac{\sqrt{n}(\bar{x}-p)}{\sqrt{p(1-p)}} \geq \frac{\sqrt{n}(c-p)}{\sqrt{p(1-p)}} | p = p_0 = 0.49) \leq 0.01 \\ & \downarrow \\ 1 - F_Z\left(\frac{\sqrt{n}(c-p_0)}{\sqrt{p_0(1-p_0)}}\right) & \leq 0.01, \end{array} \tag{77}$$

since

$$\frac{\sqrt{n}(\bar{x}-p)}{\sqrt{p(1-p)}} \xrightarrow{n\to\infty} \mathcal{N}(0,1)$$

by the Central Limit Theorem. So under this asymptotic assumption, given a large enough n, we can find that the type 1 error is smaller than 0.01 if

$$F_Z\left(\frac{\sqrt{n}(c-p_0)}{\sqrt{p_0(1-p_0)}}\right) \geq 0.99, \tag{78}$$

and the above is true if

$$\begin{array}{c} \frac{\left(a\frac{1}{n}-p_0\right)\sqrt{n}}{\sqrt{p_0(1-p_0)}} \geq q_{0.99}^{\mathcal{N}(0,1)} \\ \Updownarrow \\ \left(a\frac{1}{n} - p_0\right)\sqrt{n} \geq \sqrt{p_0(1-p_0)}q_{0.99}^{\mathcal{N}(0,1)} \\ \Updownarrow \\ \frac{a}{n} \geq p_0 + \frac{q_{0.99}^{\mathcal{N}(0,1)}\sqrt{p_0(1-p_0)}}{\sqrt{n}} \end{array} \tag{79}$$

Similarly, for the type 2 error we get that

$$\frac{a}{n} \leq p_1 + \frac{q_{0.01}^{\mathcal{N}(0,1)}\sqrt{p_1(1-p_1)}}{\sqrt{n}} \tag{80}$$

If we divide Equation 79 by Equation 80 then we get that

$$\begin{array}{c} 1 \geq \frac{p_0+q_{0.99}\sqrt{p_0(1-p_0)}/\sqrt{n}}{p_1+q_{0.01}\sqrt{p_1(1-p_1)}/\sqrt{n}} \\ \downarrow \\ \sqrt{n}(p_1-p_0) \geq q_{0.99}\sqrt{p_0(1-p_0)} - q_{0.01}\sqrt{p_1(1-p_1)} \\ \downarrow \\ \sqrt{n} \geq \frac{q_{0.99}\sqrt{p_0(1-p_0)}-q_{0.01}\sqrt{p_1(1-p_1)}}{p_1-p_0} \simeq 8.37 \\ \downarrow \\ n \geq 71. \end{array} \tag{81}$$

14 Finite variance implies finite expectation

I think I got it. Thanks to PhD student Ziteng Cheng who actually showed it to me. I merely copy what he explained to me. It is almost like Mile's demonstration, but with the advantage that we don't have to take something like

$$\mathbb{E}\infty.$$

IN particular I just use Chebishev's inequality to reach a contradiction. See

$$\Pr(|X - \mathbb{E}X| \geq \epsilon) \leq \frac{\mathbb{E}(X - \mathbb{E}(X))^2}{\epsilon^2},$$

for any finite number

$$\epsilon.$$

Now I assume that the random variable

$$X : \Omega \to \mathbb{R}$$

has finite variance. In other words,

$$\mathbb{E}(X - \mathbb{E}(X))^2 \leq M < \infty.$$

But this allows us to say that

$$\Pr(|X - \mathbb{E}X| \geq \epsilon) \leq \frac{\mathbb{E}(X - \mathbb{E}(X))^2}{\epsilon^2} \leq \frac{M}{\epsilon^2} \xrightarrow{\epsilon \to \infty} 0.$$

Now, we want to see that if this is the case, the expectation of

$$X$$

cannot be infinite. Assume that

$$\mathbb{E}X = \infty,$$

realise now that since

$$X \in \mathbb{R}$$

then

$$|X - \mathbb{E}X| = \infty$$

But then

$$\Pr(|X - \mathbb{E}X| \geq \epsilon) = \Pr(\infty \geq \epsilon) = 1,$$

since as we said in the beggining,

$$\epsilon$$

is a finite number. But two equations earlier we just saw that this probability will eventually be smaller than 1 (it will actually get as close to 0 as we want) for a sufficiently large (and finite) number

$$\epsilon$$

. Hence the contradiction.

15 Lecture 26.03.2020

She has a poisson $X \sim Pois(\theta)$. She can test

$$\Pr(type1error) = \Pr(\bar{X} \geq 4|H_0) = 0.019$$

we also have the test

$$\Pr(type1error) = \Pr(\bar{X} \geq 3|H_0) = 0.08$$

, but imagine we wanna test just

$$\Pr(type1error) = \Pr(\bar{X} \geq c|H_0) = 0.05,$$

how do we do this? Well we can just reject H_0 as soon as $\bar{X} > 4$, or as soon as $\bar{X} = 3$ together with $w = 1$ where

$$W \sim \text{Bern}(p = \frac{\alpha - 0.19}{0.08 - 0.019})$$

Note 1 Given the above assumptions, two statisticians Stat1 Stat2, might both observe $Y = 3$, using the same test, same data and same everything, bu just because in one case $w = 1$ and in the other $w = 0$, one statistician will issue a rejection and the other an acceptance of the null hypotheses. THis is becasue the previous test is a **randomized test**.

However, if this test is done over and over again, the amount of rejections from one or other preson remains the same. If we just do it a few times the bernoulli trial might significantly influence it, but not when we do the test many times.

16 Lecture 31-03-2020

Page 399. She did and explained the proof of THeorem 8.3.27

$S(x)$ is a sufficient statistic for the model

$$\{f(x|\theta)|\theta \in \Omega_0\}$$

Under H_0,

$$X|S = s \text{ not dep. on } \theta$$

. Now let $W(x) =$ any test statistic with

$$W(X) >> 0$$

if H_1 is true.

She defines for all x:

$$p(x) := \Pr(W(X) \geq W(x)|S = s) \tag{82}$$

s is observed. She says we must show the validity

16.1 Fisher's exact test

We have two independent random variables

$$\begin{aligned} S_1 \sim Bin(n_1, p_1), \\ S_2 \sim Bin(n_2, p_2), \end{aligned} \tag{83}$$

Test

$$\begin{gathered} H_0 : p_1 = p_2 = p \\ \text{vs} \\ H_1 : p_1 > p_2 \end{gathered} \tag{84}$$

What happens under the null?? Well, under the null hypotheses

$$\begin{aligned} f(s_1, s_2|p) = \tbinom{n_1}{s_1} p^{s_1}(1-p)^{n_1-s_1} \tbinom{n_2}{s_2} p^{s_2}(1-p)^{n_2-s_2} \\ = h(X)g(S(X)|p), \end{aligned} \tag{85}$$

where $S(X) = s_1 + s_2$. This is a sufficient statistic under H_0. (factorization theorem).

Now she wants to use some test statistic that allows us the reject the null hypotheses H_0. Given value $S = s$, we can use the test statisic s_1, because if $s_1 >> 0$ compared to $s, \rightarrow s_2$ is small, since $s_2 = S - s_1$. So conditional p-value is

$$p(s_1, s_2) = \textstyle\sum_{j=s_1}^{\min(n_1,s)} f(j|s), \tag{86}$$

$$f(j|s) = \Pr(s_1 = j|S = s)$$

which is **hypergeometric**.