Math 357: Statistics

Shereen Elaidi

Winter 2020 Term

Contents

1 Introduction

Data consists of observations $x_1, ..., x_n$. These are regarded as realisations of random phenomena modelled by the random variables $X_1, ..., X_n$. In this course, the X_i's will be random variables in $\mathbb{R}^d$, usually with $d = 1$.

Definition 1 (Random Sample)**.** The random variables $X_1, ..., X_n$ are called a **random sample** from a distribution F for $i = 1, ..., n$ (or, if we want, we write $X \sim F$).

In this course, the data will be a realisation of the random sample F. The basic issue is that F is unknown, so our task is to learn $\mathcal{F}$ from the realisations $x_1, ..., x_n$. A **model** for F is $\mathcal{F}$, which is a collection of (certain) probability distributions such that $F \in \mathcal{F}$. It is always an (artificial) **approximation to reality**.

Example 1 (U.S. 2016 Election Poll)**.** . $n = 2,000$, and let $x_1, ..., x_n$ be the realisations of $X_1, ..., X_n$, which are iid. Then, assuming there are only two candidates:

$$F \in \mathcal{F} = \{ \text{ Bernoulli}(p) \mid p \in]0, 1[\ \}$$

This is an example of a **parametric model**, since p, which is the probability of success, is an *unknown* parameter. We know that for each x_i, we have $x_i \in \{0, 1\}$. To estimate p, we have:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

i.e., the sample mean. By the Weak Law of Large Numbers, which we can use thanks to the iid assumption, we can see how good the estimator is by observing that it will $\hat{p}$ will converge in probability to p as $n \to \infty$:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} p \tag{1}$$

2 Part 1: Properties of Random Samples

We first remark that the assumption that the $X_1, ..., X_N$ are iid is also called **sampling from an infinite population.** To see why, consider that our population were finite, say $\{x_1, ..., x_N\}$, and we sample the $X_1, ..., X_N$ with replacement. Then, the probability of choosing a specific x_k would be:

$$\mathbb{P}[X_1 = x_k] = \frac{1}{N} \ \forall \ k \ \in \{1, ..., N\}$$
$$\mathbb{P}[X_2 = x_k \mid X_1 = x_j] = \begin{cases} 0 & \text{if } k = j \\ \frac{1}{N-1} & \text{if } k \neq j \end{cases}$$

Using the law of total probability, we can obtain $\mathbb{P}[X_2 = x_k]$:

$$\begin{aligned}\mathbb{P}[X_2 = x_k] &= \sum_{j=1}^{N} \mathbb{P}[x_2 = x_k \mid X_1 = x_j]\mathbb{P}[X_1 = x_j] \\ &= \sum_{j=1, j\neq k}^{N} \frac{1}{N-1}\frac{1}{N} = \frac{1}{N}\end{aligned}$$

These samples are identically distributed but not independent. When $N >> n$, the dependence between $X_1, ..., X_n$ plays *essentially no role.* The following example illustrates this:

Example 2. Let $N = 2,000$ and $n = 10$. If we assume that they are not independent:

$$\mathbb{P}[X_1 > 2,000, ..., X_n > 2,000] = \frac{\binom{800}{10}\binom{200}{0}}{\binom{1000}{0}} \approx 0.106164$$

If do assume they are independent:

$$\mathbb{P}[X_1 > 2,000, ..., X_n > 2,000] = [\mathbb{P}[X_1 > 200]]^{10} = \left(\frac{800}{1000}\right)^{10} \approx 0.107374$$

Definition 2 (Statistic)**.** Let $X_1, ..., X_n$ be a random sample from F, a distribution on $\mathbb{R}^d$. For a measurable function:

$$T : \underbrace{\mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{n \text{ times}} \to R^k \tag{2}$$

the random variable $T(X_1, ..., X_n)$ is called a **statistic**. The distribution of the statistic, $T(X_1, ..., X_n)$ is called a **sampling distribution** of the statistic $T(X_1, ..., X_n)$.

CAUTION: $T(X_1, ..., X_n)$ is a function of $X_1, ..., X_n$ ONLY. That is, $T(X_1, ..., X_n)$ must be a *vector of numbers.*

Example 3. Let's go back to the opinion poll example. The estimator

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is a statistic. A realisation is $0.47 = T(X_1, ..., X_n)$. Note that the following is *not* a statistic:

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i - p\right)^2 \tag{3}$$

since there is a dependence on a parameter p.

Definition 3 (Sample Mean, Sample Variance, Sample Standard Deviation)**.** The average:

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4}$$

is called the **sample mean**. The quantity:

$$s^2 := \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{5}$$

is called the **sample variation**. The statistic $s := \sqrt{s^2}$ is called the **sample standard deviation.** When these values are realised, they are denoted $\overline{x}, s^2$, and s. $\overline{x}$ and s^2 are measures of central tendency and variability, respectively.

Theorem 1. Suppose that $x_1, ..., x_n \in \mathbb{R}$ and let $\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$. Then:

1.

$$\min_{\alpha}\sum_{i=1}^{n}(x_i - \alpha)^2 = \sum_{i=1}^{\infty}(x_i - \overline{x})^2 \tag{6}$$

2.

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x} \tag{7}$$

Proof. 1.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \alpha)^2 &= \sum_{i=1}^{n}(x_i \pm \overline{x} - \alpha)^2 \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^2 + (\overline{x} - a)n + 2(\overline{x} - \alpha)\underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{:=n\overline{x} - n\overline{x} = 0} \\ &\geq \sum_{i=1}^{n}(x_i - \overline{x})^2\end{aligned}$$

2. Set $\alpha = 0$ in the preceding calculation. Then:

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

□

Lemma 2. Let $X_1, ..., X_n$ be iid from F on $\mathbb{R}$, and let $g : \mathbb{R} \to \mathbb{R}$ be measurable, and $X \sim F$. Suppose that $\text{Var}[g(X)] < \infty$. Then:

1.
$$\mathbb{E}\left[\sum_{i=1}^{n} g(X_i)\right] = n\mathbb{E}[g(X)] \tag{8}$$

2.
$$\text{Var}\left[\sum_{i=1}^{n} g(X_i)\right] = n\text{Var}[g(X)] \tag{9}$$

Theorem 3. Let $X_1, ..., X_n$ be a random sample from F, and assume that $X \sim F$. Suppose that $\text{Var}[X] = \sigma^2 < \infty$ and let $\mu := \mathbb{E}[X]$. Then:

1. $\mathbb{E}\left[\overline{X}\right] = \mu$
2. $\text{Var}[\overline{X}] = \sigma^2/n$.
3. $\mathbb{E}\left[s^2\right] = \sigma^2$.

Remark. The reason why we divided by $(n-1)$ and not n in s^2 was because we wanted property (3) to be true.

Proof. 1. Obvious.

2.
$$\text{Var}[\overline{X}] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$$

3.
$$\begin{aligned}\mathbb{E}\left[\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x}^2)\right] &\underbrace{=}_{\text{Thm 1.2}} \mathbb{E}\left[\frac{1}{(n-1)}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}\right]\right] \\ &= \frac{1}{(n-1)}\left[n\mathbb{E}\left[X^2\right] - n\mathbb{E}\left[\overline{X}\right]^2\right] \\ &= \frac{1}{(n-1)}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} - \mu^2\right)\right] \\ &= \sigma^2\end{aligned}$$

□

Example 4 (Simon Newcomb trying to measure the speed of light, 1835-1909)**.** In 1882, he attempted to carry out measurements to determine the speed of light. He collected 66 data points. *Question: given the data, what would be the model for the distribution of the speed of light? How would you describe the stochastic mechanism?*

We have n random variables, $X_1, ..., X_n$, where $n = 66$. The paramteric statistical model is:

$$\begin{aligned}\mathcal{F} &= \{\ N(\mu, \sigma^2),\ \mu \in \mathbb{R}^2, \sigma^2 > 0\} \\ T_i &= (24800 + X_i) \cdot 10^{-9}\end{aligned}$$

We think that a normal distribution is reasonable. That is, $X_i \sim N(\mu, \sigma^2)$. We will write each X_i as the sum of the mean and an error term, ε:

$$X_i = \mu + \varepsilon_i$$

where we assume $\varepsilon_i \sim N(0, \sigma^2)$. The ε_i has some distribution function with CDF F_0 with $\mathbb{E}[\varepsilon_i] = 0$. We can express the distribution $\mathcal{F}$ in an alternative way in terms of shifts:

$$\mathcal{F} = \{F_0(\cdot - \mu),\ \mu \in \mathbb{R},\ F_0 \text{ is a CDF with expectation zero } \}$$

This gives us something that we call a **semi-parametric model**. We then obtain the $\overline{x}$, sample mean. We can use this to estimate μ.

Question: How confident are we in the obtained sample mean? In order to answer this question, we need to know something about the sampling distribution of the statistics. We can estimate the variance of the sample mean using the previous theorem and the assumption about the errors being normally distributed as follows. Recall that the sample variance is given by:

$$\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

So, the expected value of s^2 is:

$$\mathbb{E}\left[s^2\right] = \text{Var}[X] = \sigma^2 \text{ (if the errors are normally distributed)}$$

and,

$$\text{Var}[\overline{X}] = \sigma^2/n$$

This means that the uncertainty of $\overline{X}$ depends on the underlying distribution since the σ^2's are the variances of the X_i's.

Theorem 4. Suppose that $X_1, ..., X_n$ is a random sample from an underlying distribution F. Let $X \sim F$. Suppose also that X has a moment generating function $M_X(t)$ for $t \in I$. Then, the moment generating function of $\overline{X}$ is:

$$M_{\overline{X}}(t) = \{M_x(t/n)\}^n,\ t/n \in I \tag{10}$$

Proof. The proof follows from the IID property of the random variables $X_1, ..., X_n$.

$$\begin{aligned}
\text{MGF} = \mathbb{E}\left[e^{t\overline{X}}\right] &= \mathbb{E}\left[e^{t/n(X_1+...+X_n)}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{n} e^{t/nX_i}\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[e^{t/nX_i}\right] \\
&= \{M_X(t/n)\}^n
\end{aligned}$$

□

The next example will give us some concrete examples of applying the previous theorem.

Example 5. 1. Let $F = N(\mu, \sigma^2)$. So, we know that $X \sim N(\mu, \sigma^2)$. From Math 356 we know the moment generating function is:

$$M_X(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}\ t \in \mathbb{R}$$

Invoking the theorem and simplifying:

$$\begin{aligned}
M_{\overline{X}}(t) &= \left(e^{\frac{t}{n}\mu + \frac{1}{2}\sigma^2 \frac{t^2}{n^2}}\right)^n\ t \in \mathbb{R} \\
&= e^{t\mu + \frac{1}{2}\sigma^2 \frac{t^2}{n}}
\end{aligned}$$

since the MGF uniquely determines the underlying distribution, this gives us that:

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

2. If $X \sim \text{Binomial}(m, p)$:

$$M_X(t) = (1 - p + pe^t)^m$$
$$M_{\overline{X}}(t) = (1 - p + pe^{t/n})^{m \cdot n}$$

a) A modification of the $\overline{X}$ will be distributed binomially. Namely, $n \times \overline{X}$ will get rid of the n in the denominator:

$$M_{n \times \overline{X}}(t) = \mathbb{E}\left[e^{nt\overline{X}}\right] = (1 - p + pe^t)^{m \cdot n}$$

$\Rightarrow n \times \overline{X} \sim \text{binomial}(m \cdot n, p)$.

3. If $X \sim \text{Gamma}(\alpha, \beta)$:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$
$$M_X(t) = (1 - t\beta)^{-\alpha}, \ t < \frac{1}{\beta}$$
$$M_{\overline{X}}(t) = \left(1 - \frac{t}{n}\beta\right)^{-\alpha n}, \ t < \frac{n}{\beta}$$
$$\Rightarrow \overline{X} \sim \text{Gamma}\left(\alpha \cdot n, \frac{\beta}{n}\right)$$

2.1 Sampling from the Normal Population

The setup for this section will be as follows: let $X_1, ..., X_n$ be iid from $N(\mu, \sigma^2)$.

Theorem 5. Let $X_1, ..., X_n$ be iid from $\text{N}(\mu, \sigma^2)$. Then:

1. $\overline{X} \sim N(\mu, \sigma^2)$. (Shown using the MGF).
2. $\overline{X}$ and s^2 are independent.

Proof. WLOG for (a), we can assume that $\mu = 0$ and $\sigma^2 = 1$, since we standardise random variables. Why? The standardisation of a random variable X_i, denoted by X_i^*, is an affine transformation given by:

$$X_i^* = \frac{X_i - \mu}{\sigma}$$

X_i^* is still normally distributed, which can be shown using an MGF argument. Thus:

$$\overline{X}_i^* = \frac{\overline{X} - \mu}{\sigma}$$

$(\overline{X} = \sigma\overline{X}^* + \mu)$. Moreover, for $(s^*)^2$:

$$(s^*)^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i^* - \overline{X}^*)^2 = \frac{1}{(n-1)} \sum_{i=1}^n \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{s^2}{\sigma^2}$$

This justifies why we can say WLOG. To show that (b) holds when $\mu = 0$ and $\sigma^2 = 1$, we will play with s^2:

$$s^2 = \frac{1}{(n-1)} \left[\sum_{i=2}^n (X_i - \overline{X})^2 + (X_1 - \overline{X})^2\right]$$

because we also know that

$$\sum_{i=1}^n (X_i - \overline{X}) = 0 \Rightarrow (X_1 - \overline{X}) = -\sum_{i=2}^n (X_i - \overline{X})$$

This implies that we can re-write s^2 as:

$$s^2 = \frac{1}{(n-1)}\left[\sum_{i=2}^{n}(X_i - \overline{X})^2 + \left(\sum_{i=2}^{n}(X_i - \overline{X})\right)^2\right] := h(X_2 - \overline{X}, ..., X_n - \overline{X})$$

h is clearly measurable. We now have a function of only $n-1$ random variables, which is why we normalise s^2 by $(n-1)$.

Lemma 6 (Core of the argument)**.** If $X_1, ..., X_n$ are iid $N(0,1)$, then

$$\overline{X} \perp (X_2 - \overline{X}, ..., X_n - \overline{X}) \tag{11}$$

From this lemma 1.14, we can then immediately conclude that $\overline{X} \perp s^2$.

Proof. This is where we use the normality assumption. Define a one-to-one function $g : \mathbb{R}^n \to \mathbb{R}^n$:

$$g(x_1, ..., x_n) \mapsto (\overline{x}, x_2 - \overline{x}, ..., x_n - \overline{x})$$

Since g is one-to-one, we can invert it:

$$g^{-1}(y_1, ..., y_n) \mapsto \left(g_n - \sum_{i=2}^{n} y_i, y_n + y_1, ..., y_n + y_1\right)$$

Need to calculate the Jacobian of the transformation:

$$\text{Jac}(g) = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \cdots & 1 \end{bmatrix}$$

$\det(\text{Jac}(g)) = n$. So, by the transformation laws:

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = f_{(X_1,...,X_n}(g^{-1}(y_1, ..., y_n))|\text{Jac}(g)|$$

Here, $Y_1 = \overline{X}, Y_2 = X_2 - \overline{X}, ..., Y_n = X_n - \overline{X}$, and so by the IID of the X_i:

$$\begin{aligned} f_{(X_1,...,X_n)}(x_1, ..., x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \end{aligned}$$

where the second inequality follows by the standard normal assumption.

$$\begin{aligned} f_{(Y_1,...,Y_n)}(y_1, ..., y_n) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\left(y_1 - \sum_{i=2}^{n} y_i\right)^2 + (y_2 + y_1)^2 + ... + (y_n + y_1)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{\frac{-ny_1^2}{2}\right\} \prod_{i=2}^{n} \exp\left\{-\frac{1}{2}\left(\sum_{i=2}^{n} y_i^2 - \left(\sum_{i=2}^{n} y_i\right)^2\right)\right\} \end{aligned}$$

(the cross terms will drop out). Thus, we have factored the densities of the Y's into the product of two functions with difference dependencies (y_1 vs. $y_2, ..., y_n$). Thus, $Y_1 \perp (Y_2, ..., Y_n)$ □

By the following theorem from Chapter 4 of the textbook, we obtain the desired result:

Theorem 7 (Generalisation of Theorem 4.3.2)**.** Let $X_1, ..., X_N$ be independent random vectors. Let $g_i(x_i)$ be only a factor of x_i, $i = 1, ..., n$. Then, the random variables $U_i := g_i(X_i)$, $i = 1, ..., n$ are mutually independent.

□

Definition 4 (Chi Squared Distribution with v degrees of freedom)**.**

$$f_v(x) := \frac{1}{2^{v/2}\gamma(v/2)} x^{v/2-1} e^{-x/2} \tag{12}$$

$x > 0$.

Remark: χ^2_v is a Gamma$(v/2, 2)$ distribution (special case of the gamma distribution). Recall from earlier that the MGF of χ^2_v is $(1-2t)^{-v/2}$ for $t < 1/2$.

Lemma 8 (Facts about χ^2)**.**
1. If $X \sim \chi^2_v$, then $\mathbb{E}[X] = v$ and $\text{Var}[X] = 2v$
2. If $X_1 \sim \chi^2_{v_1}$ and $X_2 \sim \chi^2_{v_2}$, $X_1 \perp X_2$, then $X_1 + X_2 \sim \chi^2_{(v_1+v_2)}$
3. If $X \sim N(0,1)$ then $X^2 \sim \chi^2_1$. (Proof of this one is on assignment 1).

The following theorem is very important since it leads to the chi squared test.

Theorem 9. Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Then:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)} \tag{13}$$

Reality check:

$$\mathbb{E}\left[\frac{(n-1)s^2}{\sigma^2}\right] = (n-1)$$

which implies that:

$$\frac{(n-1)}{\sigma^2}\mathbb{E}\left[s^2\right] = (n-1) \Rightarrow \mathbb{E}\left[s^2\right] = \sigma^2$$

We can elegantly prove this using moment generating functions:

Proof. From the preceding Lemma and the first theorem of the section, we have that we can standardise:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Squaring this, we obtain:

$$n\left(\frac{(\overline{X} - \mu)^2}{\sigma^2}\right) \sim \chi^2_1$$

So, summing the random variables gives:

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n$$

Therefore, by adding and subtracting the sample mean and simplifying:

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2} \\ &= \frac{(n-1)s^2}{\sigma^2} + \underbrace{\frac{n(\overline{X} - \mu)^2}{\sigma^2}}_{\sim\chi^2_1}\end{aligned}$$

From the first theorem of the section, these are independent. We cannot subtract to solve for the quantity that we are interested in, so we will work with moment generating functions. The MGF of the left hand side is:

$$(1-2t)^{n/2},\ t<1/2$$

and the MGF of the right hand side is:

$$M_{\frac{(n-1)s^2}{\sigma^2}}(t)\cdot(1-2t)^{-1/2},\ t<1/2$$

equating these, we obtain:

$$M_{\frac{(n-1)s^2}{\sigma^2}}(t)=(1-2t)^{-(n-1)/2},\ t<1/2$$

However, this is the MGF of $\chi^2_{(n-1)}$, and since the MGF uniquely determines the distribution, we obtain that $\frac{(n-1)s^2}{\sigma^2}\sim\chi^2_{(n-1)}$, which is what we wanted to show. □

Observe: this proof heavily relies on the normality of the distribution, independence, and the chi squared.

2.2 Standardisation

Motivation: quality control problem. Let $\overline{X}$ be the random variable estimating the underlying mean. Say that we set a quality cutoff, μ_0, and we want to answer the question:

$$\text{does } \mu=\mu_0?$$

What if we try to standardise:

$$\frac{\overline{X}-\mu_0}{\sigma/n}\sim N(0,1)$$

and then compare the quantities? The problem here is that this is not a statistic. We do not know what σ is. However, if you don't know something, estimate it:

$$\sqrt{n}\frac{\overline{X}-\mu_0}{\sqrt{s^2}}$$

The problem with this approach is that it is not distributed $N(0,1)$. Especially when the sample is not too large. This motivates the following definition:

Definition 5 (Student-t distribution with v degrees of freedom)**.** has the density:

$$f_v(x):=\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{(v+1)/2}$$

where $x\in\mathbb{R}$ and $v>0$.

Remark: if you set $v=1$, you obtain the Cauchy Lorentz distribution. $\mathbb{E}[X]=\infty$. Also observe that the student-t is a heavy-tailed distribution.

Lemma 10. Let $X\sim t_v$. Then:

1. $\mathbb{E}[X]=0$, provided that $v>1$ (otherwise it does not exist).
2. $\mathrm{Var}[X]=v/(v-2)$ when $v>2$ (otherwise, var$[X]$ does not exist).
3. (Assignment): if $Z\sim N(0,1)$, $V\sim\chi^2_v$, $Z\perp V$, then:

$$\frac{Z}{\sqrt{v/v}}\sim T_v \tag{14}$$

This is the **most important part of the distribution**. You can prove it using the transformation theorem for densities.

Theorem 11. Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Then:

$$\frac{\overline{X} - \mu}{\sqrt{s^2/n}} \sim t_{(n-1)}$$

The proof is pretty obvious using the lemma:

Proof. We will express the ratio as a standard normal Z. We already have that:

$$\sqrt{n}\frac{\overline{X} - \mu}{\sigma} \sim N(0,1) \perp \frac{s^2(n-1)}{\sigma^2} \sim \chi^2_{(n-1)}$$

So, taking the ratio in the form of what is given to us in the previous lemma gives:

$$\frac{\sqrt{n}\frac{\overline{X}-\mu}{\sigma}}{\sqrt{\frac{s^2(n-1)}{\sigma^2(n-1)}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sqrt{s^2}} \sum t_{(n-1)}$$

where we obtain the distribution from the previous lemma. □

The student-t model forms a nice statistical model for certain types of data.

2.3 F-Test Basis

If we want to compare quality, we need to standardise by variance. The following section helps answer the question, **is the variance between two samples the same?**

Definition 6 (The Fischer-Snedecor's F Distribution with Two Parameters v_1 and v_2 degrees of freedom)**.** This distribution is denoted by F_{v_1,v_2}. This is the distribution of the following:

$$\frac{V_1/v_1}{V_2/v_2}$$

where $V_1 \sim \chi^2_{v_1}$, $V_2 \sim \chi^2_{v_2}$, and $V_1 \perp V_2$.

This will lead to the F-test.

Theorem 12. If $X_1, ..., X_n$ is a random sample from $N(\mu_x, \sigma_x^2)$, and $Y_1, ..., Y_m$ is a random sample from $N(\mu_y, \sigma_y^2)$, and the two samples are *independent*, s_x^2 and s_y^2 are the sample variances, then:

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} \sim F_{n-1,m-1}$$

Proof. From the previous result and independence, we have:

$$\frac{(n-1)s_x^2}{\sigma_x^2} \sim \chi^2_{(n-1)}$$
$$\frac{(m-1)s_y^2}{\sigma_y^2} \sim \chi^2_{(m-1)}$$

are independent. If we divide by the degrees of freedom and invoke the definition of the F distribution, we obtain the desired result.

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} \sim F_{(n-1),(m-1)} \tag{15}$$

□

This forms the basis of the F-test. If we want them to have the same variances, then s_y^2 and s_x^2 had better be close. The way to assess this is to look at ratios, and then the σ^2's will drop out and we can then test hypotheses. Moreover, if $\sigma_x^2 = \sigma_y^2$, then:

$$\frac{s_x^2}{s_y^2} \sim F_{(n-1),(m-1)}$$

we will later use this to construct the so-called **F-test**.

2.4 Limiting Sample Distributions (Asymptotics)

What happens as $n \to \infty$? These questions are answered by **Weak Law of Large Numbers** and **Convergence in Distribution.**

More precisely, let $X_1, ..., X_n$ be iid from F. Then, define $T_n := T(X_1, ..., X_n)$ be a real-valued statistic. Then:

Q1: Does T_n converge in probability to an estimator $\theta \in \mathbb{R}$?

$$\forall \varepsilon > 0 \ \mathbb{P}[|T_n - \theta| > \varepsilon] \to 0 \ \text{ as } n \to \infty \tag{16}$$

Q2: What happens to the distribution as $n \to \infty$? In other words, if (r_n) is a sequence of real numbers, typically such that $r_n \to \infty$ ad $n \to \infty$, does

$$r_n(T_n - \theta) \xrightarrow{d} T \tag{17}$$

What the (r_n) is doing is that it is "zooming" into $T_n \to \theta$. **Idea:** when n is "large enough":

$$r_n(T_n \to \theta) \xrightarrow{d} T_n \frac{T}{r_n} + \theta$$

where the final term is called the **location and scale model**. This may have a nice distribution. The distribution of $\frac{T}{r_n} - \theta$ is often called the **large-sample** or **(asymptotic) distribution of** T_n or the **limiting distribution** of T_n. This is fundamental for quantifying uncertainty and for hypothesis testing.

2.4.1 Question 1

First we will have a refresher from Math 356. The prime tool for convergence in probability is the Weak Law of Large Numbers.

Theorem 13 (Weak Law of Large Numbers)**.** Let $X_1, X_2, ...$ be iid random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2 < \infty$. Define

$$\overline{X_n} := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{18}$$

then, $\overline{X_n}$ converges in probability to the expected value of X.

Theorem 14 (Continuous Mapping Theorem)**.** If $T_n \xrightarrow{P} T$ and g is continuous on the set C such that $\mathbb{P}[T \in X] = 1$, then $g(T_n) \xrightarrow{P} g(T)$.

In particular, if $T_n \xrightarrow{P} \theta$ and if g is continuous at θ, then $g(T_n) \xrightarrow{P} g(\theta)$.

Example 6 (Another justification for the sample variance)**.** Let $X_1, ..., X_n$ be a random sample from X (i.e., from F with $X \sim F$), and assume that $\mathbb{E}\left[X^2\right] < \infty$. Then, by the WLLN, $\overline{X} \xrightarrow{P} \mathbb{E}[X]$. For the sample variance:

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

Applying the weak law of large numbers to the first term and the continuous mapping theorem to the second term gives us that the difference will converge in probability to:

$$\mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \text{Var}[X]$$

since the square root is continuous:

$$s = \sqrt{s^2} \xrightarrow{P} \sqrt{\text{Var}[X]}$$

Example 7. Let $X_1, ..., X_n$ be a random sample from X. Suppose that we are after $\mathbb{P}[X \in A]$. This can be estimated by the **empirical probability**, which counts how many x's fall into A. Set $Z_i := \chi_{X_i \in A}$. Then, the Z_i are iid Bernoulli random variables. Using the Weak Law of Large Numbers:

$$\mathbb{P}_n \xrightarrow{P} \mathbb{P}[X \in A]$$

which is the expected value of a Bernoulli random variable. By the Strong Law of Large numbers:

$$\mathbb{P}_n \to \mathbb{P}[X \in A] \text{ a.s.}$$

Since F is the CDF of X, we can try to learn it from the data. For $x \in \mathbb{R}$, $F(x) = \mathbb{P}[X \leq x]$. The **empirical distribution function** F_n is given by:

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \chi_{X_i \leq x} \tag{19}$$

This is a central object in mathematical statistics.

Given an observed sample $x_1, ..., x_n$, a sample empirical CDF is given by:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi_{x_i \leq x} \tag{20}$$

This is a CDF itself. It is a CDF with a discrete distribution with support $\{x_1, ..., x_n\}$ and $\mathbb{P}[x_i] = 1/n$. By the Weak Law of Large Numbers:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi_{X_i \leq x} \xrightarrow{P} F(x)$$

for $\forall x \in \mathbb{R}$. Actually, by the strong law of large numbers:

$$F_n(x) \to F(x) \ a.s.$$

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

Theorem 15 (Glivenlco-Cantelli Theorem)**.** We have uniform convergence:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \text{ as } n \to \infty \tag{21}$$

Notice that:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \int x dF_n(x) \tag{22}$$

2.4.2 Question 2

We will need the following tools from probability: the Central Limit Theorem and Slutsky's theorem.

Theorem 16. Let $Z_1, ..., Z_n$ be iid with $\mathbb{E}[Z_i] < \infty$, $\mathbb{E}[X_1] = \mu$, $\text{var}[Z_1] < \infty$, and $\text{var}[Z_i] = \sigma^2$. Then, the **Central Limit Theorem** states:

$$\sqrt{n}\frac{\overline{Z} - \mu}{\sigma} \xrightarrow[CLT]{} N(0,1) \tag{23}$$

or

$$\sqrt{n}(\overline{Z} - \mu) \xrightarrow[CLT]{} N(0,\sigma^2) \tag{24}$$

Theorem 17 (Slutsky's Theorem)**.** Assume that $T_n \xrightarrow{d} T$, $Y_n \xrightarrow{p} c$ where $c \in \mathbb{R}$. Then:

1. $T_n + Y_n \xrightarrow{d} T + c$
2. $T_n \cdot Y_n \xrightarrow{d} T \cdot c$
3. $T_n / Y_n \xrightarrow{d} T/c$ if $c \neq 0$.

Remarks: If $T_n \xrightarrow{d} c$ where $c \in \mathbb{R}$, then $T_n \xrightarrow{p} c$. Moreover, we can say something with the CMT. If $T_n \xrightarrow{d} T$, g is continuous on C with $\mathbb{P}[T \in C] = 1$. Then, $g(T_n) \xrightarrow{d} g(T)$.

Important remark: if (r_n) is a sequence of numbers, $r_n \to \infty$ as $n \to \infty$. Then, if $r_n(T_n - \theta) \xrightarrow{d} T$, then we have:

$$T_n - \theta = \underbrace{r_n(T_n - \theta)}_{\xrightarrow{d} T} \underbrace{\frac{1}{r_n}}_{\to 0} \xrightarrow[\text{slutskys's}]{d} 0$$

So, $T_n - \theta \xrightarrow{p} 0$, or $T_n \xrightarrow{d} \theta$. This means that the second statement implies the first statement.

Example 8 (Alternative proof of the CLT)**.** Let $X_1, ..., X_N$ be iid from X, $\mathbb{E}[X] = \mu$, $\text{Var}[X] = \sigma^2 < \infty$ (this assumption is very important!)

$$\sqrt{n}\frac{\overline{X} - \mu}{\sqrt{s^2}} = \underbrace{\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}_{\text{By vanilla CLT, this goes to} N(0,1)} \cdot \underbrace{\frac{\sigma}{\sqrt{s^2}}}_{\xrightarrow{p} 1} \tag{25}$$

and so by Slutsky's theorem

$$\xrightarrow{d} N(0,1)$$

Corollary 1. Assume that each $T_n \sim t_n$ (student-t with n degrees of freedom). Then $T_n \xrightarrow{d} N(0,1)$.

Example 9. Suppose $X_1, ..., X_n$ are iid from Bernoulli(p). Suppose that we want to estimate $\text{var}[X_1] = p(1-p)$. **Q:** could we estimate it by:

$$\overline{X}(1 - \overline{X}) \xrightarrow[P]{CMT} p(1-p)$$

More precisely, can we find a sequence (r_n) such that

$$(r_n)(\overline{X}(1 - \overline{X}) - p(1-p)) \xrightarrow{d} ??$$

This is not exactly the CMT; this example provides the motivation for developing the **delta method**. Suppose that we want $r_n(T_n - \theta) \xrightarrow{d} T$ but we are interested in $g(T_n) - g(\theta)$. What could we do? We could use the Taylor Expansion:

$$g(T_n) - g(\theta) \approx g'(\theta)(T_n - \theta)$$
$$(r_n)(g(T_n) - g(\theta)) \approx g'(\theta)\{r_n(T_n - \theta)\}$$

By Slutsky's theorem, $r_n(T_n - \theta)$ converges in distribution to $T \cdot g'(\theta)$.

Theorem 18 (Delta Method)**.** Suppose $(r_n)(T_n - \theta) \xrightarrow{d} T$, where (r_n) is a real sequence with $r_n \to \infty$. Let g be a function and T_n takes values in the domain of g, and assume that g is *differentiable* at θ. Then:

$$r_n(g(T_n) - g(\theta)) \xrightarrow{d} T \cdot g'(\theta) \tag{26}$$

Proof will be postponed. Back to the example. Here:

$$\begin{aligned} g(x) &= x(1-x) \\ g'(x) &= 1 - 2x \end{aligned}$$

By the central limit theorem

$$(\overline{X} - p)\sqrt{n} \xrightarrow{d} N(0, p(1-p))$$

Let $\sqrt{n} := (r_n)$. Then, by the delta method

$$\begin{aligned} \sqrt{n}(\overline{X}(1-\overline{X}) - p(1-p)) &\xrightarrow{d} N(0, p(1-p)) \cdot (1-2p) \\ &= N(0, (1-2p)^2 p(1-p)) \end{aligned}$$

When $p = 1/2$, the statement is uninteresting since the derivative is zero. We will have both convergence in probability and distribution to zero, which doesn't give us any information really.

We are no ready to prove the delta method.

Proof. This proof uses several common arguments that you should learn for stats :-). Observe that by the continuous mapping theorem, $T_n \xrightarrow{p} \theta$, also $g(T_n) \xrightarrow{p} g(\theta)$. Define the following function:

$$\begin{aligned} &f : \mathbb{R} \to \mathbb{R} \\ &h \mapsto \begin{cases} \frac{g(\theta+h)-g(\theta)}{h} - g'(\theta) & \text{if } h \neq 0 \\ 0 & \text{if } h = 0 \end{cases} \end{aligned}$$

Here, the interesting domain is small and is about zero. Observe that f is continuous at zero. Now we can use the continuous mapping theorem:

$$f(T_n - \theta) \xrightarrow{p} f(0) = 0$$

since the first term is equal to

$$\frac{g(T_n) - g(\theta)}{T_n - \theta} - g'(\theta)$$

By Slutsky's theorem:

$$\begin{aligned} &r_n(g(T_n) - g(\theta)) - g'(\theta) r_n[T_n - \theta] \\ &= \underbrace{r_n[T_n - \theta]}_{\xrightarrow{d} T} \underbrace{f(T_n - \theta)}_{\xrightarrow{p} \theta} \xrightarrow{d} T \cdot 0 = 0 \end{aligned}$$

where the last convergence follows from Slutsky's theorem. Therefore:

$$r_n(g(T_n) - g(\theta)) - g'(\theta) r_n[T_n - \theta] \xrightarrow{p} 0$$

Now:

$$r_n[g(T_n) - g(\theta)] = \underbrace{r_n[g(T_n) - g(\theta)] - g'(\theta) r_n(T_n - \theta)}_{\xrightarrow{p} 0} + \underbrace{g'(\theta)(T_n - \theta) \cdot r_n}_{\xrightarrow{d} g'(\theta) T \ (CMT)}$$

and so by Slutsky's theorem, we obtain convergence in distribution to $g'(\theta)T$, which completes the proof. □

2.5 Order Statistics

Definition 7 (Order Statistics)**.** The **order statistics** of a random sample $X_1, ..., X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, ..., X_{(n)}$. In other words, they are random variables that satisfy $X_{(1)} \leq \cdots \leq X_{(n)}$. In particular:

$$\begin{aligned} X_{(1)} &= \min_{1 \leq i \leq n} X_i \\ X_{(2)} &= \text{ second smallest } X_i \\ &\vdots \\ X_{(n)} &= \max_{1 \leq i \leq n} X_i \end{aligned}$$

Definition 8 (Sample Range)**.** The **sample range** is defined as $R := X_{(n)} - X_{(1)}$.

Definition 9 (Sample Median)**.** The **sample median**, denoted by M, is the number such that approx one half of the observations are less than M and one half sare greater. It can be defined in terms of the order statistics as:

$$M := \begin{cases} X_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{(X_{(n/2)} + X_{(n/2+1)})}{2} & \text{if } n \text{ is even} \end{cases}$$

3 Point Estimation

Let $X_1, ..., X_n$ be a random sample from F, and we will *assume*:

$$F \in \mathcal{F} = \{F_\theta,\ \theta \in \Theta\}$$

This means that our distribution is known up to a parameter θ. θ is an unknown vector of parameters. Θ is called the **parameter space**, $\Theta \subseteq \mathbb{R}^k$. **Question:** where does Θ come from? This is the "art" of statistical modelling. The objective in this section is to learn θ from the data.

Notation: F_θ is a CDF. P_θ is the corresponding probability measure. F_θ has a density / probability mass function, which we will denote by $f(x, \theta)$. This notation emphasises the dependence on θ.

Example 10. Assume that we have the normal model. Then:

$$\mathcal{F} = \{N(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$$

Here, $\theta = (\mu, \sigma^2)$ and the parameter space $\Theta = \mathbb{R} \times]0, \infty[$.

Definition 10 (Point Estimator)**.** A **point estimator** is any statistic $T(X_1, ..., X_n)$ (function of only the data) taking values in $\mathbb{R}^k$ that has been constructed with the aim of estimating θ. The observed value $T(x_1, ..., x_n)$ is called an **estimate** of θ, which is a concrete number.

This definition is vague since we do not want to eliminate important estimates a priori. Often, estimators are denoted with $\hat{\theta}, \tilde{\theta}, \theta_n$, or $\hat{\theta_n}$.

3.1 Ways We Can Construct Estimators

We will explore two of these methods in depth in this class. The third we will briefly mention. The first two are frequentist methods and the final one is a bayesian method.

1. Method of moments
2. Method of maximum likelihood
3. Bayesian estimation method

3.1.1 Method of moments

This is the oldest method. It was developed in the 19th century by Karl Pearson. It has been what we heuristically have been doing all along. Suppose that $X \sim F_\theta$. Then, as an idea we could first calculate the jth moment μ_j of X for $j = 1, ..., k$, that is:

$$\mu_j = \mathbb{E}_\theta \left[X^j\right] = \int x^j dF_\theta(x) = \int x^j dP_\theta = \int x^j f(x;\theta)dx$$

How would we estimate μ_j from the data? By the weak law of large numbers, μ_j can be estimated by the jth sample moment:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

observe that this μ_j is a function of θ. Thus, the general idea of the method of moments is:

Compute the jth moment of the parametric model. Then, solve for θ:

1. Calculate the population moments μ_j, $j = 1, ..., k$ (the first k moments), and observe that μ_j is a function of θ.
2. Set $\mu_j = m_j$ for $j = 1, ..., k$ and solve for θ. Start with $j = 1$ and see where it takes us.

Note that sometimes you may need more moments than the first k¡ since some moments may not actually depend on θ (they could be zero, for example).

Example 11 (Method of moments for the normal distribution)**.** Let $\mathcal{F} = \{N(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 > 0\}$. Here, since we have two parameters to estimate, we need at least two moments. We have:

$$\mathbb{E}[X] = \mu \; \mathbb{E}\left[X^2\right] = \text{var}[X] + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2$$

Now equate the population moments and solve for θ. That is, we need to solve:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$
$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2$$

and this gives us

$$\hat{\mu} = m_1$$
$$\hat{\sigma}^2 = m_2 - \hat{\mu}^2$$

Substituting in what all of this means gives us:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)}{n}s^2$$

Hence, the estimator (**method of moments estimator**) for (μ, σ^2) is:

$$(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{X}, \frac{(n-1)}{n}s^2\right)$$

Q: What are some advantages and disadvantages of the method of moments?
Some advantages include:

1. It is intuitive.
2. Population moments are easy to estimate and we know how they behave asymptotically.

a) Weak law of large numbers and the CLT.

3. Asymptotic properties of $\hat{\theta}$ can often be derived using the CMT and the Delta Method.

Some disadvantages include ("the bitter reality"):

1. There is a systematic bias for n small, which is encoded by the $(n-1)/n$ term.
2. These estimators are often not optimal.

Example 12 (Method of moments for binomial)**.** Let our model be $\mathcal{F} = \{B(N,p),\ p \in [0,1],\ N = \{1,...\}\}$. There are two cases: N can either be known or known. If N is known, then it is easy. In this case, $\theta = p$ and $\Theta = [0,1]$. In this case:

$$\begin{aligned}\mu_1 &= Np \\ m_1 = Np \Rightarrow \hat{p} &= \frac{m_1}{N} = \frac{\overline{x}}{N}\end{aligned}$$

The not so easy case is when both p and N are unknown. We have the following system of equations:

$$\begin{aligned}\mu_1 &= Np \\ \mu_2 &= Np(1-p) + (Np)^2\end{aligned}$$

We thus have the following system of equations that we need to solve:

$$\begin{aligned}m_1 &= Np \\ m_2 &= Np - Np^2 + N^2p^2\end{aligned}$$

Solving for $\hat{p}$ gives:

$$\hat{p} = \frac{m_1}{N}$$

Substituting this estimation into the second equation and simplifying gives:

$$\begin{aligned}m_2 &= m_1 - \frac{m_1^2}{N^2}N(n-1) \\ \Rightarrow m_2 &= m_1 + \frac{m_1^2}{N}(N-1) \\ \Rightarrow Nm_2 &= Nm_1 + m_1^2(N-1) \\ \Rightarrow N(m_2 - m_1 - m_1^2) &= -m_1^2\end{aligned}$$

What if $X_1 = ... = X_n = 0$? This event has non-zero probability, and so we must account for it. In that case, there is nothing more we can do. Otherwise, we obtain:

$$\begin{aligned}\hat{N} &= \begin{cases}\frac{-m_1^2}{m_2 - m_1 - m_1^2} \\ \text{undefined if } X_1 = ... = X_n = 0\end{cases} \\ \hat{p} &= \frac{-m_2 - m_2 - m_1^2}{m_1} \text{ or undefined in the case } X_1 = ... = X_n = 0\end{aligned}$$

So, our method of moments (MoM) estimator of (p, N) is:

$$\left(\frac{-m_2 + m_1 + m_1^2}{m_1}, \frac{m_1^2}{-m_2 + m_1 + m_1^2}\right) = \left(\frac{\overline{X} - \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2}{\overline{X}}, \frac{(\overline{X})^2}{\overline{X} - \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2}(*)\right)$$

Observations from this example:

- Estimating N was not very intuitive.

- There is no guarantee that the second component of (*) is an integer or non-negative. This is why we chose such a vague definition of an estimator; because we don't want to exclude this method.
- Estimating N for the binomial is quite difficult.

Here is an example where θ determines the support of the random variable.

Example 13. Let $\mathcal{F} = \{U(]-\theta, \theta[),\ \theta \in]0, \infty[\}$. Since the expected value is zero, we will need more moments. Recall the following from probability theory: if $X \sim U(]a, b[)$, then:

$$\mathbb{E}[X] = \frac{a+b}{2} \text{ and } \text{var}[X] = \frac{1}{2}(b-a)^2$$

then the moments are:

$$\mu_1 = 0$$
$$\mu_2 = \frac{1}{2}(2\sigma)^2 = 2\sigma$$

this is good since it is a genuine function of θ. Set:

$$\mu_2 = m_2$$

and solve the following to obtain the method of moment estimator:

$$2\sigma^2 = m_2 \Rightarrow \hat{\theta} = \sqrt{\frac{m_2}{2}} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{2}}$$

Observe that θ could theoretically be zero. But, since this is a continuous distribution, the probability of θ actually being zero is zero.

3.1.2 Method of Maximum Likelihood

This is more optimal, but it can be harder to obtain. Assume that F_θ has a PDF or PMF for any $\theta \in \Theta$. Recall that we denoted this PMF by $f(x;\theta)$ to emphasise the dependence on θ. Before proceeding, a word of caution:

The likelihood function is NOT RANDOM!!!

Definition 11 (Likelihood Function)**.** Given that $(X_1, ..., X_n) = (x_1, ..., x_n)$ have been observed, the function of θ defined by $L : \Theta \to [0, \infty[$,

$$L(\theta) := \prod_{i=1}^{n} f(x_i; \theta) \tag{27}$$

is called the **likelihood function** for a fixed $x = (x_1, ..., x_n)$ as a function of θ.

Example 14. Suppose that $x_1 = 1$, $x_2 = 2$, $x_3 = 2$, and $x_4 = 5$ and that we assume the Poisson model:

$$\mathcal{F} = \{P(\lambda),\ \lambda \in]0, \infty[\}$$

then, the likelihood function is:

$$L(\lambda; x) = e^{-\lambda}\frac{\lambda^1}{1!} + e^{-\lambda}\frac{\lambda^2}{2!} + e^{-\lambda}\frac{\lambda^3}{2!} + e^{-\lambda}\frac{\lambda^4}{5!}$$

We can think of the Likelihood function as a sort of "function summary" whose domain is the parameter space.

Interpretation of the Likelihood Function:

1. If F_θ is discrete, then:

$$L(\theta; x) = \mathbb{P}[X_1 = x_1, ..., X_n = x_n]$$

   It is simply the product of the PMFs. It thus represents the probability of observing the sample that we actually observed. Moreover, if for some $\theta_1, \theta_2 \in \Theta$, if $L(\theta_1; x) > L(\theta_2; x)$, then it means that we were more likely to observe that data if the parameter is θ_1 instead of θ_2. Thus, *information is encoded in the function.*
2. If the distribution F_θ is continuous, then this interpretation will obviously not work. However, we can do the usual "trick":

$$\mathbb{P}_\theta[X \in]x - \varepsilon, x + \varepsilon[] = \int_{x-\varepsilon}^{x+\varepsilon} f(t; \theta)dt \approx f(x; \sigma) \cdot 2\varepsilon$$

   It is thus proportional up to a constant depending on ε. Mathematically:

$$L(\theta; x) \propto \mathbb{P}[X_1 \in]x_1 - \varepsilon, x_1 + \varepsilon[, \ ..., X_n \in]x_n - \varepsilon, x_n + \varepsilon[]$$

   Similar to the discrete case, comparing $L(\theta_1; x)$ with $L(\theta_2; x)$ will give us a comparison of the probability of the actual observed sample.

So the aim is to find a way maximise the likelihood function in θ.

Definition 12 (Maximum Likelihood Estimate)**.** For an observed sample $x = (x_1, ..., x_n)$, let $\hat{\theta}(x)$ be the value for which L is maximised. In other words:

$$L(\hat{\theta}(x); x) := \sup_{\theta \in \Theta} L(\theta; x) \tag{28}$$

Then, $\hat{\theta}$ is called the **maximum likelihood estimate**. (Grounding: it is a number, not an object). If $\hat{\theta}(x)$ exists for almost all samples, and as a function:

$$\hat{\theta} : \mathbb{R}^n \to R^k, \ x \mapsto \hat{\theta}(x)$$

is a measurable function, then the *function* $\hat{\theta}(X)$ is called the **Maximum Likelihood Estimator (MLE)** of θ.

Example 15 (Continued from the Poisson Case)**.** Let $x_1, ..., x_n$ be a sample from the Poisson distribution. Then, the aim is to maximise:

$$L(\theta; x) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \cdots x_n!}$$

You *could* differentiate this monstrosity, but differentiating products is in general not fun. However, sums are not too bad to differentiate. So, observe that this function is:

1. Positive
2. Wherever $\log(L)$ is maximised, L is maximised.

So we can equivalently maximise $\log(L(\lambda; x))$. Thus:

$$\log L(\lambda; x) = \ell(\lambda; x) = -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \log(x_1! \cdots x_n!)$$

$$\Rightarrow \frac{\partial \ell(\lambda; x)}{\partial \lambda} = -n + \left(\sum_{i=1}^{n} x_i \right) \frac{1}{\lambda} = 0 \iff n\lambda = \sum_{i=1}^{n} x_i$$

$$\Rightarrow \hat{\lambda}(x) = \overline{x}$$

Take the second partial derivative to verify that this is indeed a maximum.

$$\frac{\delta \ell^2}{\delta \lambda} = -\frac{1}{\lambda^2} \sum_{i=1}^{n} x_i < 0$$

This, the MLE is $\overline{x}$.

Some observations for calculating maximum likelihood estimators:

1. It is often simpler to maximise the **log likelihood function**, which is defined as:

$$\ell(\theta; x) := \log L(\theta; x) \tag{29}$$

2. If ℓ is differentiable, we can look for the maximum by solving the so-called **likelihood equations**:

$$\frac{\partial \ell}{\partial \theta_j} = 0 \ j = 1, ..., k \tag{30}$$

Example 16. Let $x_1, ..., x_n$ be a random sample from $B(N, p)$, where N is *known*. Then, $\Theta = [0, 1]$ and

$$L(p; x) = \prod_{i=1}^{n} \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}$$

thus, the log likelihood function ℓ is:

$$\ell(p; x) = \sum_{i=1}^{n} \left[x_i \log(p) + (N - x_i) \log(1-p) + \underbrace{\log \binom{N}{x_i}}_{\text{if } N \text{ were unknown, this would be a big mess}} \right]$$

We are searching for critical points. Let's first assume that $\overline{x} \notin \{0, N\}$. We obtain:

$$\frac{\partial \ell}{\partial p} = \left(\sum_{i=1}^{n} x_i \right) \frac{1}{p} + \left(nN - \sum_{i=1}^{n} x_i \right) \left(\frac{-1}{1-p} \right)$$

setting the partial derivative equal to zero and solving, we obtain:

$$\begin{aligned}
&\frac{n\overline{x}}{p} - \frac{n(N - \overline{x})}{1-p} = 0 \\
\Rightarrow &\frac{\overline{x}(1-p) - (N - \overline{x})}{p(1-p)} = 0 \\
\Rightarrow &\overline{x}(1-p) - (N\overline{x})p = 0 \\
\Rightarrow &\overline{x} - \overline{x}p - Np + \overline{x}p = 0 \\
\Rightarrow &p = \frac{\overline{x}}{N}
\end{aligned}$$

Thus, $p = \overline{x}/N$ is the method of moments estimator. We now need to verify that this is indeed a maximum with the second derivative test:

$$\frac{\partial^2 \ell}{\partial^2 p} = -n\overline{x}\frac{1}{p^2} + \underbrace{(nN - n\overline{x})}_{\geq 0} \frac{-1}{(1-p^2)} < 0 \text{ (concave in } p)$$

We now need to check the boundary cases. If $\overline{x} = 0$, then

$$\ell(p; x) = nN \log(1-p)$$

This function is decreasing in p, which implies that it is maximised at $p = 0 = \frac{\overline{x}}{N}$. If $\overline{x} = N$, then $x_i = N \ \forall \ i \in \{1, ..., N\}$. This is an increasing function in p, which implies that it is maximised at $p = 1$, which implies that $\hat{p} = 1 = \overline{x}/N$

For the past two examples, we've seen that the MLE estimator agrees with the method of moments examples. The following example will illustrate that this is not always the case.

Example 17. Let $x_1, ..., x_n$ be an observation of a random sample taken from the model:

$$\mathcal{F} = \{\mathcal{U}]0, \theta[\mid \theta \geq 0\}$$

If $X \sim \mathcal{U}]0, \theta[$, then using the method of moments we have that:

$$\mathbb{E}[X] = \theta/2$$
$$\frac{\theta}{2} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \hat{\theta} = 2\overline{x}$$

In contrast, using the method of maximum likelihood gives:

$$L(\theta; x) = \prod_{i=1}^{n} \frac{1}{\theta}\chi_{x_i \in [0,\theta]}$$

since χ depends on θ, taking the log will not be a good idea. Instead, this becomes:

$$= \left(\frac{1}{\theta}\right)^n (\chi_{\min_{1\leq i\leq n}(x_i)\geq 0})(\chi_{\max_{1\leq i\leq n}(x_i)\leq \theta})$$

For almost all samples, we have:

$$\min_{1\leq i\leq n} x_i \geq 0$$

but for $\theta < \max(x_i)$ we have that $L = 0$ but for $\theta \geq \max x_i$, L is decreasing. Thus, the MLE is maximised at $\max_{1\leq i\leq n} x_i$. Thus,

$$\hat{\theta(x)} = \max_{1\leq i\leq n} X_i = X_{(n)}$$

is the MLE (to do: insert graph).

Example 18. Let $x_1, ..., x_n$ be a random sample from $N(\mu, \sigma^2)$ from the model

$$\mathcal{F} = \{N(\mu, \sigma^2) \mid \mu \in \mathbb{R},\ \sigma^2 \in]0, \infty[\}$$

Then:

$$L(\mu, \sigma^2; x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma}}\frac{1}{\sqrt{\sigma^2}}\exp\left\{\frac{-(x_i-\mu)^2}{2\sigma}\right\}$$

and so the log likelihood is:

$$\ell(\mu, \sigma^2; x) = \sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\log\sigma^2 - \frac{(x_i-\mu^2)}{2\sigma^2}\right]$$

taking the partials, we obtain:

$$\frac{\partial \ell}{\partial \mu} = -\sum_{i=1}^{n}\frac{2(x_i-\mu)^2(-1)}{2\sigma^2} = \frac{n\overline{x} - n\mu}{\sigma^2}$$
$$\frac{\partial \ell}{\partial \sigma^2} = \sum_{i=1}^{n}\left[-\frac{1}{2}\frac{1}{\sigma^2} + \frac{(x_i-\mu)^2}{2}\frac{1}{\sigma^4}\right]$$

setting $\frac{\partial \ell}{\partial \mu} = 0$, we obtain that $\hat{\mu} = \overline{x}$. For the other one:

$$\frac{-n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0$$
$$n\sigma^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$$
$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

How do we prove that our values $\hat{\mu}$ and $\hat{\sigma}^2$ are indeed maximums? The trick is to use somethig called the **profile likelihood**. Define:

$$\ell^*(\sigma^2) := \sup_{\mu \in \mathbb{R}} \ell(\mu, \sigma^2) \tag{31}$$

we observe that if we maximise ℓ^*, then this amounts to maximising the whole thing, since $\overline{x}$ minimises the sum $\sum_{i=1}^n (x_i - \mu)^2$. So all we need to do is maximise this:

$$\ell^*(\sigma^2) = \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(x_i - \overline{x})^2$$

the above is called the **profile likelihood univariate problem**. Solving the above gives you the $\hat{\sigma}^2$ that is the solution.

3.1.3 Invariance of the MLE Theorem

Motivation: suppose that $X_1, ..., X_n$ are iid taken from Bernoulli(p). In this case, we have that the method of moments and MLE method agree:

$$\hat{p} = \overline{x} = \hat{p}_{MLE}$$

The standard deviation of $X_1 = \sqrt{p(1-p)}$. We can estimate this by $\sqrt{\hat{p}(1-\hat{p})}$. This is a non-bijective function of p. In this sense, this is the MLE of the standard deviation, which motivates the idea of an invariance principle.

Theorem 19 (Invariance of the MLE)**.** Consider a statistical model $\{F_\theta \mid \theta \in \Theta\}$ and suppose that $g : \Theta \to \mathbb{R}^m$ is an *arbitrary* function. Set $\Gamma := g(\Theta)$ and $\gamma := g(\theta)$. Then, if $\hat{\theta}(x_1, ..., x_n)$ is the maximum likelihood estimate of θ, then $g(\hat{\theta}(x_1, ..., x_n)$ is the ML estimate of γ in the following sense: if

$$L^*(\gamma; x_1, ..., x_n) = \sup_{\theta \in \Theta \mid g(\theta) = \gamma} \{L(\theta; x_1, ..., x_n)\} \tag{32}$$

then:

$$L^*(g(\hat{\theta}(x_1, ..., x_n)); x_1, ..., x_n) = \sup_{\gamma \in \Gamma} L^*(\gamma; x_1, ..., x_n) \tag{33}$$

Proof.

$$\begin{aligned} L^*(g(\hat{\theta}(x_1, ..., x_n)); x_1, ..., x_n) &= \sup_{\theta \mid g(\theta) = g(\hat{\theta})} \{L(\theta; x_1, ..., x_n)\} \\ &= L(\hat{\theta}; x_1, ..., x_n) \text{ (since } \hat{\theta} \text{ is the maximiser)} \\ &= \sup_{\theta \in \Theta} L(\theta; x_1, ..., x_n) \end{aligned}$$

Sup over the range of g, and then through the pre-image (just running through the θ's in a systematic way):

$$= \sup_{\gamma \in \Gamma} \underbrace{\sup_{\theta \in \Theta \mid g(\theta) = \gamma} L(\theta; x_1, ..., x_n)}_{= L^*(\gamma; x_1, ..., x_n)}$$

□

Example 19. Let $X_1, ..., X_n$ be iid from Bernoulli(p). If $\hat{p} = \overline{X} = p_{\text{MLE}}$, then $\sqrt{\hat{p}(1-\hat{p})}$ is the MLE of the standard deviation $\sqrt{p(1-p)}$.

The next example will be a nice philosophical example.

Example 20. Among 20 tosses, assume that there were 7 heads. Suppose that we want to estimate p in two ways:

- Experiment # 1: the coin was tossed 20 times and it had 7 heads. However, we don't know *when* those 7 heads happened. In this case, think of likelihood as the probability of observing what you have observed:

$$\binom{20}{7} p^7 (1-p)^{13}$$

implies that the log likelihood is

$$\log \binom{20}{7} + \underbrace{7\log(p) + 13\log(1-p)}_{\text{this is called the } \underline{\textbf{kernel}} \text{ of the log-likelihood}}$$

- Experiment # 2: suppose you waited until 7 heads were tossed, and you are told it took 20 tosses to get there. Therefore, the difference here is that *you know that the 20th toss was a head.* This probability is modelled by the **negative binomial**:

$$\binom{19}{6} p^7 (1-p)^{13}$$

and so the log likelihood is:

$$\log \binom{19}{6} + \underbrace{7\log(p) + 13\log(1-p)}_{\text{kernel}}$$

Even though the likelihoods are not the same, the MLE between the two will be the same! We have:

$$L_1 \propto p^7 (1-p)^{13} \tag{34}$$

that is, $L_1(p) = c_1 \cdot p^7(1-p)^{13}$. Similarly, $L_2(p) = c_2 \cdot p^7(1-p)^{13} \propto p^7(1-p)^{13}$. Therefore, the MLE estimates are the same in the two experiments, since $L_1(p) = cL_2(p)$. This is called the **likelihood principle**: when two likelihood functions of two experiments are equal up to a multiplicative constant, they contain the same information about the unknown parameter.

3.1.4 Bayesian Method

In the **Bayesian philosophy**, we quantify the lack of knowledge of a parameter θ with a probability distribution or density $\pi(\theta)$. This is called a **prior**. A prior distribution is *your* belief of what the probability is. Once the data has been collected– say, $x_1, x_2, ..., x_n$, we can update the prior to incorporate this information, and this leads to the **posterior density**.

The **posterior density function** is given by:

$$\pi(\theta \mid x_1, ..., x_n) := \frac{f(\theta, x_1, ..., x_n)}{f(x_1, ..., x_n)} = \frac{f(x_1, ..., x_n; \theta)\pi(\theta)}{\int_\Theta f(x_1, ..., x_n; \theta)\pi(\theta) d\theta} \tag{35}$$

Example 21. Consider $X_1, ..., X_n$ taken from a Bernoulli(p) distribution with a prior from the Beta distribution. Recall that the **beta function** is given by:

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

and so the **beta distribution** B(α, β) has the density:

$$f(x; \alpha, \beta) = x^{\alpha-1}(1-x)^{(\beta-1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \chi_{x \in]0,1[} \tag{36}$$

The set of beta distributions has a lot of different shapes as you vary the parameters, which makes it a nice distribution to choose priors from. For example, the uniform is a special case of a beta distribution. Given the data $x_1, ..., x_n$, we obtain:

$$\begin{aligned} f(x_1, ..., x_n; p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{n\overline{x}}(1-p)^{n-n\overline{x}} \end{aligned}$$

We thus obtain the posterior:

$$\begin{aligned} \pi(p|x) &= \frac{p^{n\overline{x}}(1-p)^{n-n\overline{x}}p^{\alpha-1}(1-p)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}{\int_0^1 p^{n\overline{x}}(1-p)^{n-n\overline{x}}p^{\alpha-1}(1-p)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}dp} \\ &\propto c(x_1, ..., x_n, \alpha, \beta)p^{n\overline{x}+\alpha-1}(1-p)^{n-n\overline{x}+\beta-1} \end{aligned}$$

The posterior is a beta density:

$$\pi(p|x_1, ..., x_n) \sim \text{Beta}(n\overline{x} + \alpha, n - n\overline{x} + \beta)$$

here, the Beta is the so-called **conjugate prior**; that means, the posterior belongs to the same class of densities as the prior. Observe that we no longer have a point estimate of p; we have a distribution. To obtain a point estimate, we can take the expected value of the density:

$$\begin{aligned} \hat{p}_B = &\text{ expected value of the posterior.} \\ \Rightarrow \hat{p}_B &:= \mathbb{E}[Y], \ Y \sim \text{Beta}(n\overline{x} + \alpha, n - n\overline{x} + \beta) \end{aligned}$$

Note that the expected value of $B(\alpha, \beta)$ is $\frac{\alpha}{\alpha+\beta}$ and:

$$\hat{p}_B = \frac{n\overline{x} + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta}\underbrace{\overline{x}}_{\text{MLE}} + \frac{\alpha + \beta}{n + \alpha + \beta}\underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{expectation of prior}}$$

Notice the weights. The more data that you have, the more the weights favour the data. Thus, a small sample means that the prior has a larger effect.

Example 22 (Poisson Likelihood Example - Continued)**.** Consider the Poisson(λ) model with parameter $\lambda > 0$ and data $x_1, ..., x_n$. In order to go Bayesian, we need to multiply the prior estimate of λ, and we want it to be a conjugate prior. *What do we multiply it by?* $\pi(\lambda) \sim \text{Gamma}(\alpha, \beta)$ works:. We obtain:

$$f(x_1, ..., x_n; \lambda)\pi(\lambda) = \frac{1}{\Gamma(\alpha)}\frac{\lambda^{n\overline{x}}e^{-n\lambda}}{x_1! \cdots x_n!}\underbrace{\lambda^{\alpha-1}e^{-\lambda\beta}\beta^{\alpha}}_{\pi(\lambda)\sim\text{Gamma}(\alpha,\beta)} \propto \alpha\lambda^{\alpha+n\overline{x}-1}e^{-\lambda(n+\beta)} \tag{37}$$

The posterior is again a Gamma, so the Gamma prior is a conjugate prior. Thus:

$$\pi(\lambda|x) \sim \text{Gamma}(\alpha + n\overline{x}, n + \beta)$$

which gives:

$$\hat{\lambda}_B = \frac{\alpha + n\overline{x}}{n + \beta} = \frac{n}{n + \beta}\overline{x} + \frac{\beta}{n + \beta}\frac{\alpha}{\beta}$$

where the final term comes from the fact that the expected value of Gamma(α, β) is α/β.

3.2 Method of Evaluating Estimators

Definition 13 (Unbiased, Consistent, and MSE)**.** Let $\mathcal{F} = \{F_\theta \mid \theta \in \Theta\}$ be our statistical model, let $\gamma : \Theta \to \mathbb{R}^n$, and also let $T(x_1, ..., x_n)$ be an estimator of $\gamma(\theta)$. Then:

1. An estimator $T(X_1, ..., X_n)$ is called **unbiased** if:
$$\mathbb{E}_\theta [T(X_1, ..., X_n)] = \gamma(\theta) \tag{38}$$
On average, does it get the right thing?
2. T is called **consistent** if
$$T(X_1, ..., X_n) \xrightarrow{P} \gamma(\theta) \text{ as } n \to \infty \tag{39}$$
Some kind of limiting statement – generally uses WLLN techniques
3. The **mean squared error (MSE)** of $T(X_1, ..., X_n)$ is:
$$\text{MSE}(T) := \mathbb{E}_\theta [\{T(X_1, ..., X_n) - \gamma(\theta)\}]^2 \tag{40}$$

Before going on, need to make some important remarks:

1. If X_1 has an expectation μ and variance σ^2, then we already had that the sample mean is an unbiased estimator of μ and s^2 is an unbiased estimator of σ^2 (Theorem 1.9).
2. Sometimes the consistency as defined here is called **weak consistency** to differentiate it from **strong consistency**, which is the case when $T(X_1, ..., X_n) \to \gamma(\theta)$. In general, it is hard to prove strong consistency. It generally requires the use of the strong law of large numbers.
   a) *Caution!*: beware of asymptotics! You can have quite stupid examples. For example, consider $\text{N}(\mu, \sigma^2)$. Suppose that we want to estimate μ. We can construct this silly unbiased estimator:
   $$T(X_1, ..., X_n) = X_1$$
   And we can construct this silly consistent estimator:
   $$T(X_1, ..., X_n) = \begin{cases} X_1 & \text{if } n \leq 10^6 \\ \overline{X} & \text{if } n > 10^k \end{cases}$$
   b) (Deriving bias variance decomposition of MSE). Write $T = T(X_1, ..., X_n)$. Then, for any θ:
   $$\begin{aligned} \text{MSE}(T) &= \mathbb{E}_\theta [T(X_1, ..., X_n) + \mathbb{E}_\theta [T] - \mathbb{E}_\theta [T] + \gamma(\theta)]^2 \\ &= \underbrace{\mathbb{E}_\theta [T - \mathbb{E}_\theta [T]]^2}_{\text{variance of T}} + \mathbb{E}_\theta [\mathbb{E}_\theta [T] - \gamma(\theta)]^2 \end{aligned}$$
   T is the only random part.
   $$\begin{aligned} &= \text{var}_\theta[T] + (\mathbb{E}_\theta [T] - \gamma(\theta))^2 + 2(\mathbb{E}_\theta [T] - \gamma(\theta)) \underbrace{\mathbb{E} [T - \mathbb{E}_\theta [T]]}_{=0} \\ &= \text{var}_\theta[T] + [\mathbb{E}_\theta [T] - \gamma(\theta)]^2 \end{aligned}$$
   We call $\mathbb{E}_\theta [[] T] - \gamma(\theta)$ the **bias** of T, and so:
   $$\text{MSE}_\theta = \text{var}_\theta[T] + (\text{Bias}_\theta(T))^2 \tag{41}$$

Example 23. Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Then:
$$\text{MSE}(\overline{X}) = \text{var}_\theta[\overline{X}] = \sigma^2/n$$

this converges to zero as $n \to \infty$. Moreover:

$$\text{MSE}[s^2] = \text{var}[s^2] = \text{var}\left[\underbrace{\frac{n-1}{\sigma^2}s^2}_{\sim\chi^2_{n-1}} \cdot \frac{\sigma^2}{(n-1)}\right] = \frac{\sigma^4}{(n-1)^2} \underbrace{2(n-1)}_{\text{variance of } \chi^2_{(n-1)}} = \frac{2\sigma^4}{(n-1)}$$

This converges at the same rate as the sample mean. Recall the MLE of σ^2:

$$\hat{\sigma}^2 = \frac{(n-1)}{n}s^2$$

and so the bias is:

$$\begin{aligned}
\text{Bias}(\hat{\sigma}^2) &= \mathbb{E}\left[\frac{n-1}{n}s^2\right] - \sigma^2 \\
&= \frac{n-1}{n}\sigma^2 - \sigma^2 \\
&= \frac{-\sigma}{n}
\end{aligned}$$

Thus, the bias is always negative and so it will systematically underestimate σ^2. Thus, the MSE($\hat{\sigma}^2$) is:

$$\begin{aligned}
\text{MSE}(\hat{\sigma}^2) &= \mathbb{E}\left[(\hat{\sigma}^2 - \hat{\sigma})^2\right] \\
&= \text{var}[\hat{\sigma}^2] + [\text{Bias}(\hat{\sigma}^2)]^2 \\
&= \left(\frac{n-1}{n}\right)^2 \text{var}[s^2] + \frac{\sigma^4}{n^2} \\
&= \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{(n-1)} + \frac{\sigma^4}{n^2} \\
&= \frac{\sigma^4(2(n-1)+1}{n^4} \\
&= \frac{\sigma^4}{n^2}(2n-1) \\
&= \frac{2\sigma^4}{n-1} \cdot \underbrace{\frac{2n^2-3n+1}{2n^2}}_{<1} \\
&< \frac{2\sigma^4}{n-1}
\end{aligned}$$

So, $\hat{\sigma}^2$ is closer, on average, to the true value of θ than s^2 is. Thus, the MLE is preferable. This illustrates the idea of a tradeoff between bias and variability.

To do: insert figure

Example 24. Let $X_1, ..., X_n$ be taken from Bernoulli(p). Recall that $\hat{p}_{\text{MLE}} = \overline{x}$ and $\text{MSE}(\hat{p}_{\text{MLE}}) = \frac{p(1-p)}{n}$. In contrast, a Bayesian estimator with a Beta(α, β) prior, we have:

$$\hat{p}_B = \frac{n}{n+\alpha+\beta}\overline{x} + \frac{\alpha}{n+\alpha+\beta} = \frac{n\overline{x}-\alpha}{n+\alpha+\beta}$$

Calculating the MSE gives:

$$\begin{aligned}
\text{MSE}(\hat{p}_B) &= \text{var}\left[\frac{n\overline{x}-\alpha}{n+\alpha+\beta}\right] + \left(\text{Bias}\left[\frac{n\overline{x}-\alpha}{n+\alpha+\beta}\right]\right)^2 \\
&= \underbrace{\frac{np(1-p)}{(n+\alpha+\beta)^2}}_{\text{variance}} + \left(\frac{np+\alpha}{n+\alpha+\beta} - p\right)^2 \\
&= \frac{np(1-p)}{(n+\alpha+\beta)^2} + \frac{(\alpha - p\alpha - p\beta)^2}{(n+\alpha+\beta)^2}
\end{aligned}$$

To compare, we need to choose α and β to be independent of p.

$$\text{MSE}(\hat{p}_B) = ... = \frac{\alpha^2 + p(n - 2\alpha^2 - 2\alpha\beta) + p^2(-n + \alpha^2 + \beta^2 + 2\alpha\beta}{(n + \alpha + \beta)^2}$$

Thus, if we want the above equation to be independent of p, set the coefficients of p and p^2 to equal zero. We thus must solve:

$$\begin{aligned} n - 2\alpha^2 - 2\alpha\beta &= 0 \\ -n + \alpha^2 + \beta^2 + 2\alpha\beta &= 0 \end{aligned}$$

which gives:

$$\alpha = \sqrt{\frac{n}{4}} \qquad\qquad \beta = \sqrt{\frac{n}{4}}$$

and therefore:

$$\text{MSE}(\hat{p}_B) = \frac{n}{4(n + \sqrt{n})^2} \qquad\qquad \hat{p}_B = \frac{n\overline{x} + \sqrt{n/4}}{n + \sqrt{n}}$$

To do: include figures. Thus, for small sample sizes, $\text{MSE}(\hat{p}_B) << \text{MLE}(\hat{p}_{\text{MLE}})$. The advantage shrinks as sample size increases.

Theorem 20. Suppose $\mathbb{E}_\theta\left[T(X_1, ..., X_n)\right] \to \gamma(\theta)$ as $n \to \infty$ (that is, $T(X_1, ..., X_n)$ is asymptotically unbiased) and $\text{var}_\theta[T] \to 0$ as $n \to \infty$. Then, T is a consistent estimator or $\gamma(\theta)$.

Proof. Application of the Markov Inequality. Fix $\varepsilon > 0$. Then:

$$\mathbb{P}[|T(X_1, ..., X_n) - \gamma(\theta)| > \varepsilon] \leq \frac{\mathbb{E}\left[(T(X_1, ..., X_n) - \gamma(\theta))^2\right]}{\varepsilon^2} = \frac{\text{var}_\theta[T] - (\text{Bias}_\theta[T])^2}{\varepsilon^2} \to 0 \text{ as } n \to 0$$

□

Q: *Does there exist a strategy of choosing the best estimator?* **A:** No! Consider the other "stupid" estimator: Suppose we want to estimate θ and $17 \in \Theta$. Then the statistic:

$$T(X_1, ..., X_n) = 17$$

will always have

$$\text{MSE}_{17}(T) = 0$$

You can never beat that silly statistic. More generally, you can never *systematically* minimise the MSE uniformly over all values of θ. Thus, we need to restrict the class of estimators that we consider, which leads us to the next section.

3.3 Best Unbiased Estimators

Cutoff for Midterm Material

4 Sufficiency and Completeness