Course:	Mathematical Statistics
Term:	Fall 2018
Instructor:	Gordan Žitković

Lecture 11
Likelihood, MLE and sufficiency

11.1 Likelihood - definition and examples

As we have seen many times before, the statistical inference based on a random sample depends heavily on the model we assume for the data. We would estimate the unknown parameter (say, the mean of the distribution) in vastly different ways when the data are normal, compared to the case of uniformly distributed data. The information about the assumed model is captured in the joint pdf (or pmf) of our data and its dependence on the parameters. In a huge number of situations, this joint pdf/pmf has a simple explicit form and, as we will see below, many important conclusions can be reached by looking at it in a right way.

Definition 11.1.1. Given a random sample $Y_1, \dots, Y_n$ from a discrete distribution D with an unknown parameter θ, we define the **likelihood (function)** by

$$L(\theta; y_1, \dots, y_n) = p^{\theta}_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = p^{\theta}(y_1)p^{\theta}(y_2)\dots p^{\theta}(y_n),$$

where p^{θ} is the pmf of (each) Y_i.

If $Y_1, \dots, Y_n$ come from a continuous distribution, we set

$$L(\theta; y_1, \dots, y_n) = f^{\theta}_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = f^{\theta}(y_1)f^{\theta}(y_2)\dots f^{\theta}(y_n),$$

where f^{θ} is the pdf of (each) Y_i.

Simply put, the likelihood is the same as the joint pdf (pmf), but with the emphasis placed on the dependence on the parameter. When considering likelihoods, we think of $y_1, \dots, y_n$ as fixed and of $L(\theta; y_1, \dots, y_n)$ as the "likelihood" of it being the parameter θ that actually produced $y_1, \dots, y_n$. This should not be confused with a probability - as a function of θ, the likelihood $L(\theta; y_1, \dots, y_n)$ is not a pdf (or a pmf) of a probability distribution. In order to be able to interpret likelihood as a probability, we need a completely different paradigm, namely Bayesian statistics.

In these notes, $Y_1, \dots, Y_n$ are always independent, and, so the likelihood can be written as a product of individual pdfs (by the factorization criterion). In general, when $Y_1, \dots, Y_n$ are dependent random variables, the notion of a likelihood can still be used if the joint distribution (pmf or pdf) of $Y_1, \dots, Y_n$ is specified. Almost everything we cover below will apply to this case, as well.

Example 11.1.2. Here are the likelihood functions for random samples from some of our favorite distributions:

1. **Bernoulli.** Suppose that $Y_1, \dots, Y_n$ and independent and $Y_i \sim B(p)$. The pmf of Y_i can be written as

$$p(y) = \mathbb{P}[Y_i = y] = \left\{ \begin{matrix} p, & y = 1 \\ (1-p), & y = 0 \end{matrix} \right\} = p^y (1-p)^{1-y} \text{ for } y = 0, 1.$$

While it may look strange at first, the right-most expression $p^y(1-p)^{1-y}$ happens to be very useful. For example, it allows us to write the full likelihood in a very compact form

$$\begin{aligned} L(p; y_1, \dots, y_n) &= p^{y_1}(1-p)^{1-y_1} \times \dots \times p^{y_n}(1-p)^{1-y_n} \\ &= p^{\sum_i y_i} \times (1-p)^{n - \sum_i y_i}. \end{aligned}$$

2. **Normal.** For a random sample $Y_1, \dots, Y_n$ from a normal $N(\mu, \sigma)$-distribution, we have

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

and, so,

$$L(\mu, \sigma; y_1, \dots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^n e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}}.$$

We can go a step further and try to isolate the parameters μ and σ by expanding each square $(y_i - \mu)^2$:

$$L(\mu, \sigma, y_1, \dots, y_n) = \frac{1}{\sigma^2(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n y_i - n\frac{\mu^2}{2\sigma^2}}$$

3. **Uniform.** Let $Y_1, \dots, Y_n$ be a random sample from a uniform distribution $U(0, \theta)$, with an unknown $\theta > 0$. The pdf of a single Y_i is

$$f(y) = \tfrac{1}{\theta}\mathbf{1}_{\{0 \leq y \leq \theta\}},$$

and, so

$$\begin{aligned} L(\theta; y_1, \dots, y_n) &= \tfrac{1}{\theta^n} \mathbf{1}_{\{0 \leq y_1 \leq \theta\}} \times \dots \times \mathbf{1}_{\{0 \leq y_n \leq \theta\}} \\ &= \tfrac{1}{\theta^n} \mathbf{1}_{\{0 \leq y_1, \dots, y_n \leq \theta\}}. \end{aligned}$$

The condition $0 \leq y_1, \dots, y_n \leq \theta$ is equivalent to the two conditions $0 \leq \min(y_1, \dots, y_n)$ and $\max(y_1, \dots, y_n) \leq \theta$. Therefore, we can write

$$L(\theta; y_1, \dots, y_n) = \tfrac{1}{\theta^n} \mathbf{1}_{\{\min(y_1, \dots, y_n) \geq 0\}} \mathbf{1}_{\{\max(y_1, \dots, y_n) \leq \theta\}}.$$

11.2 Maximum-likelihood estimation

We mentioned that the word "likelihood" in the likelihood function refers to the parameter, but that we cannot think of it as probability without changing our entire worldview. What we can do is compare likelihoods for different values of the parameter and think of the parameters with the higher value of the likelihood as "more likely" to have produced the observations $y_1, \dots, y_n$.

Example 11.2.1. Three buses (of unknown sizes, and not necessarily the same) carrying football fans arrive at a game between the Green team and the Orange team. Suppose that 90% of the people in the first bus are fans of the Green team, and 10% fans of the Orange team. The composition of the second bus is almost exactly opposite: 15% Green team fans, and 85% Orange team fans. The third bus carries the same number of Green-team and Orange-team fans. Once the buses arrive at the game, the two populations mix (say as they enter the stadium) and a person is randomly selected from the crowd. It turns out she is a fan of the Green team. What is your best guess of the bus she came to the game in?

The situation can be modeled as follows; the (unknown) parameter θ corresponding to the bus our fan came from - can take only three values 1, 2 or 3. The observation Y can take only two values G (for the Green team fans) and O (for the Orange team fans). The likelihood function $L(\theta, y)$ is given by

$$L(\theta; G) = \begin{cases} 0.9, & \theta = 1 \\ 0.15, & \theta = 2 \\ 0.5, & \theta = 3 \end{cases} \quad \text{and} \quad L(\theta; O) = \begin{cases} 0.1, & \theta = 1 \\ 0.85, & \theta = 2 \\ 0.5, & \theta = 3 \end{cases}.$$

Since the randomly picked person was a fan of the Green team, we focus on $L(\theta; G)$. The three values it can take, namely 0.9, and 0.15

and 0.5 cannot be interpreted as probabilities, as they do not add up to 1. We can still say that, in this case, $\theta = 1$ is much more likely than $\theta = 2$, and we should probably guess that our fan came in the first bus (the one that carried mostly the Green-team fans).

The thinking we used to formulate our guess in the above example was simple: pick the value of the parameter which yields the highest likelihood, given the observed data. If we follow this procedure systematically, we arrive at one of the most important classes of estimators:

Definition 11.2.2. An estimator $\hat{\theta} = \hat{\theta}(y_1, \dots, y_n)$ is called the **maximum-likelihood estimator (MLE)** if it has the property that for any other estimator $\hat{\theta}' = \hat{\theta}'(y_1, \dots, y_n)$ we have

$$L(\hat{\theta}; y_1, \dots, y_n) \geq L(\hat{\theta}'; y_1, \dots, y_n), \text{ for all } y_1, \dots, y_n.$$

Maximum-likelihood are often easy to find whenever explicit expressions for the likelihood functions are available. Unlike in Example 11.2.1 above, the unknown parameters often vary continuously and we can use calculus to find the values that maximize the likelihood. A very useful trick is to maximize the **log-likelihood** $\log L(\theta; y_1, \dots, y_n)$ instead of the likelihood L. We get the same maximizers (as $x \mapsto \log(x)$ is an increasing function), but the expressions involved are often much simpler.

Example 11.2.3.

1. **Normal (with a known variance).** Let $Y_1, \dots, Y_n$ be a random sample from a normal model with an unknown mean μ and the known variance $\sigma = 1$. Thanks to Example 11.1.2 above, we have the following expression for the likelihood function

$$L(\mu; y_1, \dots, y_n) = \tfrac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2}.$$

To find the MLE $\hat{\mu}$ for μ, we need to find the value of μ (depending on $y_1, \dots, y_n$) which maximizes $L(\mu; y_1, \dots, y_n)$. We could use the standard technique and differentiate $L(\mu; y_1, \dots, y_n)$ in μ, set the obtained value to 0 and solve for μ. The log-likelihood

$$\log L(\mu; y_1, \dots, y_n) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\sum_{i=1}^n (y_i - \mu)^2$$

is much easier to differentiate:

$$\frac{\partial}{\partial \mu} \log L(\mu; y_1, \dots, y_n) = -\tfrac{1}{2} \sum_{i=1}^{n} \frac{\partial}{\partial \mu} (y_i - \mu)^2$$
$$= -\tfrac{1}{2} \sum_{i=1}^{n} 2(y_i - \mu) = n\mu - \sum_{i=1}^{n} y_i.$$

We set the obtained expression to 0 and solve for μ, obtaining

$$\hat{\mu} = \tfrac{1}{n} \sum_{i=1}^{n} Y_i.$$

It can be show that this $\hat{\mu}$ is indeed the maximum of $L(\mu; y_1, \dots, y_n)$ (and not a minimum or an inflection point). It should not be too surprising that we obtained the sample mean - it has already been shown to be the best estimator for μ is the mean-squared-error sense.

2. **Exponential.** The likelihood function for a random sample from the exponential distribution with parameter τ is given by

$$L(\tau; y_1, \dots, y_n) = \tfrac{1}{\tau^n} e^{-\frac{1}{\tau} \sum_{i=1}^{n} y_i} \text{ for } y_1, \dots, y_n > 0.$$

As above, the log-likelihood is going to be easier to differentiate:

$$\frac{\partial}{\partial \tau} \log L(\tau; y_1, \dots, y_n) = \frac{\partial}{\partial \tau} \Big(-n \log(\tau) - \tfrac{1}{\tau} \textstyle\sum_i y_i \Big)$$
$$= -n/\tau + \tau^{-2} \textstyle\sum_i y_i.$$

Setting this derivative to 0, we obtain the equation $0 = -n/\tau + \tau^{-2} \sum_i y_i$, with the solution

$$\hat{\tau} = \frac{\sum_i y_i}{n} = \bar{y}.$$

3. **Uniform.** We have derived in Example 11.1.2 above that the likelihood function for a random sample from a uniform $U(0, \theta)$-distribution is given by

$$L(\theta; y_1, \dots, y_n) = \tfrac{1}{\theta^n} \mathbf{1}_{\{\max(y_1, \dots, y_n) \leq \theta\}},$$

when $y_1, \dots, y_n \geq 0$ (which we can safely assume). As a function of θ it looks like this:

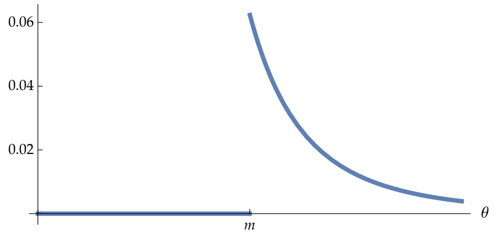

Figure 1. The likelihood function for a sample from $U(0,\theta)$, as a function of θ where $m = \max(y_1,\dots,y_n)$.

It is clear from the picture above that L attains its maximal value for $\theta = \max(y_1,\dots,y_n)$, but this fact cannot be obtained by differentiation. Indeed, as a function of θ the likelihood is not differentiable (or even continuous). Nevertheless, the MLE is well-defined and equals

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

11.3 Sufficient statistics

Consider three pollsters each of whom collects candidate preference data from a sample of $n = 1000$ voters (as always, the candidates are A and B, and each voter prefers one to the other). The first pollster reports her findings as follows:

Voter 1: **A**, Voter 2: **B**, Voter 3: **B**, ..., Voter 1000: **A**.

The second pollster decides that it is not important which particular voters prefer A and which prefer B. For him, it is enough to report the data in the following format:

Number of voters who prefer A to B: **550**.

The third pollster argues that that is an overkill, too. She simply reports

The majority prefers: **A**.

The first pollster reports every detail of his data set, while the other two summarize it to different degrees. Intuitively, however, we feel that there is no real difference between what the first and the second pollster reported, but that the third pollster left out some important information. Indeed, whether the number of votes for A is 501 or 1000, and the report of the third pollster look exactly the same. In the former case, we would not be wrong to say that the race is in "dead heat", while, in the later, we could be practically sure that A would win.

The notion of an estimator we introduced some time ago will come in handy to give a mathematical explanation of what is going on here. Remember that an estimator is any function of the data $Y_1, \dots, Y_n$. In the present context where there is no specific parameter singled out, we also use the word **statistic** as a synonym. Therefore, if $Y_1, \dots, Y_n$ denote the voters preferences encoded by $A \mapsto 1$ and $B \mapsto 0$, the quantities reported by the three pollsters are

1. $T_1 = (Y_1, \dots, Y_n)$,
2. $T_2 = Y_1 + \dots + Y_n$, and
3. $T_3 = \mathbf{1}_{\{Y_1 + \dots + Y_n > \frac{1}{2}n\}}$.

It is clear that T_2 is a function of T_1, i.e., that if we know T_1, we can easily compute T_2. Similarly, T_3 is a function of T_2. It does not work the other way around. If someone tells us the value of T_2, i.e., the number of voters in the sample who prefer A, there is no way for us to work out the exact preference of the first sampled voter, the second sampled voter, etc. Similarly, if all we know is that the majority of voters prefer A, we cannot tell what that majority actually is.

There is, however, a big difference between our inability to recover T_1 from T_2 and our inability to recover T_2 from T_3. In the first case, the missing information is irrelevant for the parameter of interest (i.e., p, the proportion of A voters in the entire population). Indeed, once the proportion T_2 is known, any arrangement of 1000 voters in a sequence (while making sure that the total number of A voters is exactly T_2) is equally likely no matter what the value of p is. In fact, for all we know, the first pollster also recorded only the number of A voters and deleted all the data about the preferences of particular voters. He might then have been told by his boss to report every detail of his sample, and, fearing for his job, cooked up a fictitious data set by randomly selecting 5800 different numbers between 1 and 10000 and claiming that the voters with those numbers prefer A, while all other voters preferred B. If he in fact did that, there would be no way of proving that he cheated, even if the true value of p were revealed.

In the second case, the information contained in T_3 is not sufficient to fake the value of T_2, let alone the value of T_1. If the third pollster found herself in the same situation as the second pollster above, she would not be able to come up with a plausible value of T_2. Indeed, the number of A voters could have been anywhere between 501 and 10000, and the knowledge of the true value of p would be very helpful to make a good guess.

If we translate the above discussion into mathematics, the two scenarios differ in one significant way:

1. the conditional distribution of $T_1 = (Y_1, \dots, Y_n)$, given $T_2 = Y_1 + \dots + Y_n$ does not depend on p, while
2. the conditional distribution of $T_1 = (Y_1, \dots, Y_n)$ given $T_3 = \mathbf{1}_{\{Y_1+\dots+Y_n > \frac{1}{2}n\}}$ does.

Let us recall the mathematical definition of the conditional probability and conditional distributions. The basic definition, known from a basic probability class is the following:

$$\mathbb{P}[A|B] = \mathbb{P}[A \cap B]/\mathbb{P}[B] \text{ when } \mathbb{P}[B] > 0.$$

We usually interpret $\mathbb{P}[A|B]$ as the probability that (the event) A will happen, if we already know that B happened. More generally, if Y and T are discrete random variables, we talk about the **conditional distribution of** a random variable Y, **given** $T = t$, as the collection of probabilities $\mathbb{P}[Y = y|T = t]$, as y "runs" through the support of Y, i.e., through the set of all possible values of Y.

In order to define the notion of sufficiency, we will need a similar concept, but with a single random variable Y replaced by an entire random vector $(Y_1, \dots, Y_n)$. Starting with the discrete case, we remember that the pmf of discrete random vector $(Y_1, \dots, Y_n, T)$ is the function of $y_1, \dots, y_n$ and t, given by

$$p_{Y_1,\dots,Y_n,T}(y_1, \dots, y_n, t) = \mathbb{P}[Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n, T = t].$$

Definition 11.3.1. Let $Y_1, \dots, Y_n, T$ be a discrete random vector with the pmf $p_{Y_1,\dots,Y_n,T}$. When $\mathbb{P}[T = t] > 0$, we define the **conditional pmf of** $Y_1, \dots, Y_2$ **given** $T = t$ by

$$p_{Y_1,\dots,Y_n|T}(y_1, \dots, y_n|t) := \mathbb{P}[Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|T = t]$$

Using the definition of the conditional probability, we see that

$$p_{Y_1,\dots,Y_n|T}(y_1, \dots, y_n|t) := \frac{\mathbb{P}[Y_1=y_1,Y_2=y_2,\dots,Y_n=y_n,T=t]}{\mathbb{P}[T=t]} = \frac{p_{Y_1,\dots,Y_n,T}(y_1,\dots,y_n,t)}{p_T(t)}$$

To give a definition for the continuous case, we replace all pmfs by pdfs:

Definition 11.3.2. Let $Y_1, \dots, Y_n, T$ be a random vector with a continuous distribution and the joint pdf $f_{Y_1,\dots,Y_n,T}(y_1, \dots, y_n, t)$. The **conditional pdf of** $Y_1, \dots, Y_n$ **given** $T = t$, with $f_T(t) > 0$ is given by

$$f_{Y_1,\dots,Y_n|T}(y_1, \dots, y_n|r) = \frac{f_{Y_1,\dots,Y_n,T}(y_1, \dots, y_n, t)}{f_T(t)},$$

where f_T is the marginal pdf of T.

We can now give a mathematical definition of the notion of a sufficient statistic:

Definition 11.3.3. A statistic (estimator) T in the random sample $Y_1, \dots, Y_n$ is said to be **sufficient** for the unknown parameter θ if the conditional distribution of $(Y_1, \dots, Y_n)$, given $T = t$, does not depend of θ (for all t for which it is defined).

In words, T is sufficient if the distribution of the "extra randomness" in the sample $(Y_1, \dots, Y_n)$, in addition to that already contained in T, does not depend on the parameter. This is why the second pollster could cook up the data in a convincing way and the third one could not.

Example 11.3.4. Let us analyze the statistics T_2 and T_3 introduced above in the light of Definition 11.3.3. Remember that $n = 1000$, $T_2 = Y_1 + \dots + Y_{1000}$ is the total number of votes of for A, and T_3 is the indicator of the event in which A wins, i.e., that $Y_1 + \dots + Y_{1000} > 500$.

- T_2 **is sufficient for** p**.** We start from the definition of the conditional pdf of $(Y_1, \dots, Y_n)$, given $T_2 = t$:

$$p_{Y_1,\dots,Y_n|T}(y_1, \dots, y_n|t) = \mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n, T_2 = t] / \mathbb{P}[T_2 = t]$$

To make progress, let us differentiate between two cases, depending on the value of t:

1. $y_1 + \dots + y_n \neq t$. In this case $\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n, T = t] = 0$ since it is impossible to have all these equalities hold at the same time.

2. $y_1 + \dots + y_n = t$. In this case we have $T = t$, as soon as $Y_1 = y_1, \dots, Y_n = y_n$, so that

$$\begin{aligned}\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n, T = t] &= \mathbb{P}[Y_1 = y_1, \dots, y_n = y_1] \\ &= p^{\sum_i y_i}(1-p)^{n-\sum_i y_i} \\ &= p^t(1-p)^{n-t}.\end{aligned}$$

The random variable T_2 is binomially distributed with parameters n and p, and, so, for $t = 0, 1, \dots, n$, we have

$$\mathbb{P}[T_2 = t] = \binom{n}{t} p^t (1-p)^{n-t}.$$

Putting everything together, we get

$$\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n | T = t] = \begin{cases} 0, & \sum_i y_i \neq t \\ \binom{n}{t}^{-1}, & \sum_i y_i = t. \end{cases}$$

Since expression does not feature p anywhere, we conclude that T_2 is a sufficient statistic.

- T_3 **is not sufficient for** p**.** The computation here parallels that from above, but we will not be able to get everything in closed form. It still helps to differentiate between two cases, depending on the relationship between $(y_1, \dots, y_n)$ and $t \in \{0, 1\}$. Since the cases $t = 0$ and $t = 1$ are very similar, we assume from now on that $t = 1$.

1. $y_1 + \dots + y_n \leq n/2$. Like above, this is an impossible case and
$$\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n | T_3 = 1] = 0.$$
2. $y_1 + \dots + y_n > n/2$. In this case, just like above, we have
$$\begin{aligned} \mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n, T_3 = 1] &= \mathbb{P}[Y_1 = y_1, \dots, y_n = y_1] \\ &= p^{\sum_i y_i} (1-p)^{n - \sum_i y_i} \end{aligned}$$

The random variable T_3 has a Bernoulli distribution whose parameter is

$$\mathbb{P}[T_3 = 1] = \mathbb{P}[Y_1 + \dots + Y_n > n/2] = \sum_{k=501} \binom{1000}{k} p^k (1-p)^{100-k}.$$

Putting everything together, we get

$$\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n | T = t] = \\ = \begin{cases} 0, & \mathbf{1}_{\{\sum_i y_i > n/2\}} \neq t \\ \frac{p^{\sum_i y_i} (1-p)^{n - \sum_i y_i}}{\sum_{k=501}^{1000} \binom{1000}{k} p^k (1-p)^{100-k}}, & \mathbf{1}_{\{\sum_i y_i > n/2\}} = t \end{cases}$$

This expression cannot be simplified any further, but we can graph it as a function of p (we pick $\sum y_i = 600$):

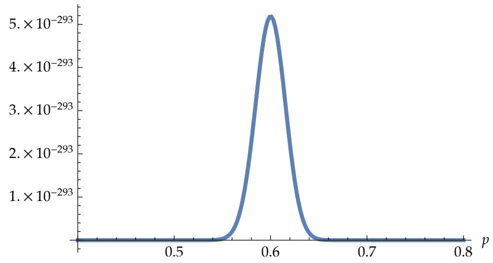

Figure 2. The conditional probability $\mathbb{P}[Y_1 = y_1, \dots, Y_n = y_n | T_3 = 1]$ when $\sum_i y_i = 600$ as a function of the parameter p.

The obtained graph shows a clear dependence on p (it is not a horizontal line), so T_3 is not a sufficient statistic for p.

The computations in the example above are tedious and, in the case of T_3, require numerical computation (graphing) to reach a conclusion. Luckily, there is a simple criterion for sufficiency - called the Fisher-Neyman criterion - which is much easier to apply.

Theorem 11.3.5 (The Fisher-Neyman factorization criterion). *Let $Y_1, \dots, Y_n$ be a random sample with the likelihood function $L(\theta; y_1, \dots, y_n)$. The statistic $T = T(Y_1, \dots, Y_n)$ is sufficient for θ if and only if L admits the following factorization*

$$L(\theta; y_1, \dots, y_n) = g(\theta, T(y_1, \dots, y_n)) \times h(y_1, \dots, y_n), \tag{11.3.1}$$

where the function h does not depend on θ.

We omit the proof, but note that it is not very complicated - one simply needs to follow the definitions and use properties of conditional probabilities. Here are some examples, though:

Example 11.3.6.

1. **Bernoulli.** Example 11.1.2 gave us the following form for the likelihood in this case:

$$L(p; y_1, \dots, y_n) = p^{\sum y_i}(1-p)^{n-\sum y_i}.$$

The Fisher-Neyman criterion is easy to apply. We simply take

$$g(p, T) = p^T(1-p)^{n-T} \text{ and } h(y_1, \dots, y_n) = 1,$$

so that

$$L(p; y_1, \dots, y_n) = g(p, \sum y_i) h(y_1, \dots, y_n).$$

In other words, the likelihood, in the form above, is already factorized as it only depends on $y_1, \dots, y_n$ through their sum $\sum y_i$.

We note that $\bar{Y}$ is also a sufficient statistic for p. Indeed, we can write

$$g(p, \bar{y}) = p^{n\bar{y}}(1-p)^{n-n\bar{y}}.$$

In fact, if k is any bijection (and not only $k(y) = y/n$) and T is a sufficient statistic, then so is $k(T)$.

2. **Normal (with a known standard deviation).** In the normal model with σ known to be equal to 1, the likelihood is given by

$$L(\mu; y_1, \dots, y_n) = \tfrac{1}{(2\pi)^{n/2}} \exp(-\tfrac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2).$$

We expand the squares and obtain the following, equivalent, expression:

$$\begin{aligned} L(\mu; y_1, \dots, y_n) &= \tfrac{1}{(2\pi)^{n/2}} \exp\Big(- \tfrac{1}{2} \big(\sum y_i^2 - 2\mu \sum y_i + n\mu^2 \big) \Big) \\ &= \underbrace{\tfrac{1}{(2\pi)^{n/2}} \exp(-\tfrac{1}{2} \sum y_i^2)}_{h(y_1,\dots,y_n)} \times \underbrace{\exp(-2\mu \sum y_i + n\mu^2)}_{g(\mu, \sum y_i)} \end{aligned}$$

It follows that $T(Y_1, \dots, Y_n) = \sum_i Y_i$ is a sufficient statistic for μ. As above, another sufficient statistic is $\bar{Y}$.

3. **Uniform.** For a random sample from $U(0, \theta)$, we derived the following form for the likelihood function in Example 11.1.2:

$$L(\theta; y_1, \dots, y_n) = \tfrac{1}{\theta^n} \mathbf{1}_{\{0 \leq \min(y_1,\dots,y_n)\}} \mathbf{1}_{\{\max(y_1,\dots,y_n) \leq \theta\}},$$

for $y_1, \dots, y_n \geq 0$. This is almost in the factorized from already. Indeed, we can take $T(y_1, \dots, y_n) = \max(y_1, \dots, y_n)$ so that

$$L(\theta, y_1, \dots, y_n) = g(\theta, T(y_1, \dots, y_n)) \times h(y_1, \dots, y_n), \qquad (11.3.2)$$

for

$$g(\theta, T) = \theta^{-n} \mathbf{1}_{\{T \leq \theta\}} \text{ and } h(y_1, \dots, y_n) = \mathbf{1}_{\{0 \leq \min(y_1,\dots,y_n)\}}.$$

Thus, $T = \max(Y_1, \dots, Y_n)$ is a sufficient statistic for θ.

11.4 (*) Some important theorems in statistics

Sufficient statistics are not only important because they point exactly where, in the data, the information about the value of the parameter lies. They can also help us build good estimators, among other things. In fact, some of the deepest theorems in theoretical statistics are about sufficient statistics. To state some of them (two, to be precise), we need three related definitions. In all of these definitions, we assume that $Y_1, \dots, Y_n$ is a random sample from a distribution with an unknown parameter θ.

Definition 11.4.1. A statistic T is said to be **minimal sufficient** if any other sufficient statistic S has the property that $S = f(T)$ for some (non-random) function f.

Intuitively, a minimal sufficient statistic captures all the information about θ more efficiently than any other statistic. It does not have "unnecessary" parts.

Our second definition is the one of completeness. We state is without worrying too much about various mathematical subtleties (a reader with appropriate background will find a few):

Definition 11.4.2. A statistic T is said to be **complete** if for any function g we have

$$\mathbb{E}^{\theta}[g(T)] = 0 \text{ for all } \theta \quad \text{if and only if} \quad g(T) = 0.$$

Finally, we need a notion somewhat opposite to that of a sufficient statistic:

Definition 11.4.3. A statistic T is said to be **ancillary** if its distribution does not depend on θ.

A statistic if ancillary if it contains no information about the parameter θ. For example, a constant statistic $T = 12$ is always ancillary. Less trivially, if $Y_1, \dots, Y_n$ is a random sample from the normal $N(\mu, 1)$, then $T = Y_2 - Y_1$ is an ancillary statistic. Indeed, its distribution is $N(0, \sqrt{2})$.

Now that we have all our terms defined, here is how to check that a given statistic in complete and minimal sufficient. We start with a large class of parametric families in which complete and sufficient statistics are easy to find:

Definition 11.4.4. A family of distributions with the unknown parameter $\theta = (\theta_1, \dots, \theta_m)$ is called an **exponential family** if its pdfs (or pmfs) have the following form:

$$f^{\theta}(y) = \exp\Big(\sum_{i=1}^{m} \eta_i(\theta) T_i(y) - A(\theta)\Big) h(y) \tag{11.4.1}$$

Proposition 11.4.5. *If $Y_1, \dots, Y_n$ is a random sample from the exponential* *family (11.4.1) with the following, additional, property:*

$$\text{if} \sum_{i=1}^{m} \eta_i(\theta)(T_i(y) - T_i(y')) = 0 \text{ for all } \theta, \text{ then } T(y) = T(y'),$$

for all y, y' with $h(y) > 0$ and $h(y') > 0$. Then the statistic $T(y_1, \dots, y_n) = (T_1(y_1) + \dots + T_1(y_n), \dots, T_m(y_1) + \dots + T_m(y_n))$ is complete and minimal sufficient.

It turns out that almost all distribution families in these notes form exponential families (try to write some of them in the form (11.4.1)). The only exceptions are the Student's t and the uniform.

We are ready now for our important theorems:

Theorem 11.4.6 (Basu)**.** *Let T be a complete and minimal sufficient statistic and S an ancillary statistic. Then T and S are independent for any value of the parameter θ.*

This theorem, e.g., tells us that $\bar{Y}$ and S^2 are independent in the normal model (how, exactly?).

Theorem 11.4.7 (Lehmann-Scheffé)**.** *Let S be an unbiased estimator for θ. If it can be written as $S = f(T)$ for some complete and sufficient statistic T, then S is the unique UMVUE for θ.*

The Lehmann-Scheffé theorem makes finding an UMVUE for θ an easy task. It is enough to pick a complete and sufficient statistic T and then apply a function f to is so that $T = f(S)$ is unbiased. It will automatically be an UMVUE.

11.5 Problems

Problem 11.5.1. Write down the likelihood functions, and then compute the MLEs (maximum likelihood estimators) for the unknown parameter in the following situations:

1. a random sample of size n from the Poisson distribution with parameter θ, i.e., the distribution with the pmf

$$p(y) = e^{-\theta} \tfrac{\theta^y}{y!},\ y \in \mathbb{N}_0.$$

2. a random sample of size n, where the pdf is given by

$$f(y) = \tfrac{1}{y\sqrt{2\pi}} e^{-\frac{1}{2}(\ln y - \theta)^2} \mathbf{1}_{\{y>0\}}.$$

(*Note:* This is known as the log-normal distribution.)

3. A random sample from the $\Gamma(k, \tau)$-distribution (where $k \in \mathbb{N}$ is considered known, so that we are looking for the MLE of τ only). (*Note:* The pdf of a $\Gamma(k, \tau)$ distribution for $k \in \mathbb{N}$ is $f(y) = \frac{y^{k-1}e^{-y/\tau}}{\tau^k (k-1)!}\mathbf{1}_{\{y>0\}}$.)

4. a random sample of size 1 (single observation) from a normal distribution where both the mean and the standard deviation are unknown, but are known to be equal to each other (with value θ), i.e. $Y_1 \sim N(\theta, \theta)$. Is the obtained estimator (the MLE) unbiased?

Problem 11.5.2. Let $Y_1, \dots, Y_n$ be a random sample from the geometric $g(p)$ distribution with the unknown parameter $p \in (0, 1)$. The MLE for p is

(a) $\frac{1}{n+\bar{Y}}$ (b) $\prod_i (\frac{Y_i}{1+Y_i})$ (c) $\frac{\bar{Y}}{n+\bar{Y}}$ (d) $\frac{\sum_i Y_i}{n+\sum_i Y_i}$ (e) none of the above

Problem 11.5.3. Let $Y_1, \dots, Y_n$ be a random sample from the normal distribution with a known mean $\mu = 0$ and an unknown standard deviation σ. The MLE (maximum likelihood estimator) for σ is

(a) $\frac{1}{n}\sum_{i=1}^n Y_i$ (b) $\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}$ (c) $\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}$ (d) $\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2}$
(e) none of the above

Problem 11.5.4. Let $Y_1, \dots, Y_n$ be a random sample from the Poisson $P(\lambda)$ distribution with an unknown parameter $\lambda > 0$. The MLE for λ is

(a) $\sum_{i=1}^n Y_i$

(b) $\sum_{i=1}^n Y_i^2$

(c) $\sum_{i=1}^n \log(Y_i)$

(d) $\frac{1}{n}\sum_{i=1}^n \log(Y_i)$

(e) none of the above

Problem 11.5.5. In all of the below, let $(Y_1, \dots, Y_n)$ is a random sample from the stated distribution. Show that T is a sufficient statistic for the unknown parameters:

1. $E(\tau)$, $T = \bar{Y}$
2. $g(p)$, $T = \sum_i Y_i$.
3. $Y_1, \dots, Y_n$ are a random sample from a distribution with pdf

$$f_Y(y) = c(\beta) y^3 (1-y)^\beta \mathbf{1}_{\{0<y<1\}},$$

with an unknown parameter $\beta > 0$, where $c(\beta)$ is the constant chosen so that $\int_0^1 f_Y(y)\, dy = 1$ (no need to compute it), and $T = \sum \log(1 - Y_i)$.

Problem 11.5.6. Let $Y_1, \dots, Y_n$ be a random sample from a distribution with pdf

$$f(y) = \tfrac{1}{2}\theta^3 y^2 e^{-\theta y}\mathbf{1}_{\{y>0\}},$$

where $\theta > 0$ is an unknown parameter. Which one of the following is **not** a sufficient statistic for θ?

(a) $\bar{Y}$ (b) $1/\bar{Y}$ (c) $\prod_{i=1}^n Y_i^2$ (d) $\prod_{i=1}^n e^{Y_i}$ (e) all of the above are sufficient

Problem 11.5.7. Let $Y_1, \dots, Y_n$ be a random sample from the uniform distribution $U(0, \theta)$ where $\theta > 0$ is an unknown parameter. Then

(a) $\bar{Y}$ is an MLE for θ

(b) $\bar{Y}$ is a sufficient statistic for θ

(c) $\min(Y_1, \dots, Y_n)$ is an MLE for θ

(d) $\max(Y_1, \dots, Y_n)$ is a sufficient statistic for θ

(e) none of the above

Problem 11.5.8. Let $Y_1, \dots, Y_n$ be a random sample from a normal distribution with the *known mean* $\mu = 0$ and an unknown standard deviation $\sigma > 0$.

1. Write down the likelihood function and find a sufficient statistic for σ.
2. What is the MLE $\hat{\sigma}$ for σ?
3. Is $\hat{\sigma}^2$ (where $\hat{\sigma}$ is as in 2.) an unbiased estimator for σ^2?

Problem 11.5.9. We consider the normal model here, as in Problem 11.5.8 above, but now both μ and σ are unknown. This makes the parameter $\theta = (\mu, \sigma)$ two-dimensional.

1. Write down the likelihood function $L(\mu, \sigma; y_1, \dots, y_n)$, the log-likelihood function and the equations for the MLE for (μ, σ). Solve them. (*Note:* Since there are two parameters, the MLE $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ itself will be two-dimensional, i.e., there are going to be two MLEs, one for μ and one for σ. There are also going to be two equations, one obtained by differentiating $\log L$ in μ and setting the result to 0 and the other, obtained by differentiating in σ.)
2. With $\hat{\mu}$ and $\hat{\sigma}$ as above, is $\hat{\mu}$ unbiased for μ? How about $\hat{\sigma}^2$? Is it unbiased for $\hat{\sigma}^2$?
3. The notion of sufficiency in this case needs to be updated in a similar way. We will need two sufficient statistics instead of the usual one to completely summarize the information in the sample. Show that the pair $(\bar{Y}, S'^2)$ is a (two-dimensional) sufficient statistic for (μ, σ). (*Note:* The factorization criterion still applies. You simply need to factorize into a

function g of $\mu, \sigma, \bar{y} = \frac{1}{n} \sum y_i$ and $s'^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$, and a function h which does not depend on parameters.) (*Hint:* First show and then use the following algebraic identity: $\frac{1}{n} \sum_i y_i^2 = s'^2 + \bar{y}^2$.)