CHAPTER 3

Statistics

1. Introduction

In statistics we are faced with data, which could be measurements in an experiment, responses in a survey etc. There will be some randomness, which may be inherent in the problem or due to errors in measurement etc. The problem in statistics is to make various kinds of inferences about the underlying distribution, from realizations of the random variables. We shall consider a few basic types of problems encountered in statistics. We shall mostly deal with examples, but sufficiently many that the general ideas should become clear too. It may be remarked that we stay with the simplest "textbook type problems" but we shall also see some real data. Unfortunately we shall not touch upon the problems of current interest, which typically involve very huge data sets etc. Here are the kinds of problems we study.

General setting: We shall have data (measurements perhaps), usually of the form $X_1, \ldots, X_n$ which are realizations of independent random variables from a common distribution. The underlying distribution is not known. In the problems we consider, typically the distribution is known, except for the values of a few parameters. Thus, we may write the data as $X_1, \ldots, X_n$ i.i.d. $f_\theta(x)$ where $f_\theta(x)$ is a pdf or pmf for each value of the parameter(s) θ. For example, the density could be of $N(\mu, \sigma^2)$ (two unknown parameters μ and σ^2) or of Pois(λ) (one unknown parameter λ).

(1) Estimation: Here, the question is to guess the value of the unknown θ from the sample $X_1, \ldots, X_n$. For example, if X_i are i.i.d. from Ber(p) distribution (p is unknown), then a reasonable guess for θ would be the sample mean $\overline{X}_n$ (an *estimator*). Is this the only one? Is it the "best" one? Such questions are addressed in estimation.

(2) Confidence intervals: Here again the problem is of estimating the value of a parameter, but instead of giving one value as a guess, we instead give an interval and quantify how sure we are that the interval will contain the unknown parameter. For example, a coin with unknown probability p of turning up head, is tossed n times. Then, a confidence interval for p could be of the form $[\overline{X}_n - \frac{3}{\sqrt{n}}\sqrt{\overline{X}_n(1-\overline{X}_n)}, \overline{X}_n + -\frac{3}{\sqrt{n}}\sqrt{\overline{X}_n(1-\overline{X}_n)}]$ where $\overline{X}_n$ is the proportion of heads in n tosses. The reason for such an interval will come later. It turns out that if n is large, one can say that with probability 0.99 ("confidence level"), this interval will contain the true value of the parameter.

(3) Hypothesis testing: In this type of problem we are required to decide between two competing choices ("hypotheses"). For example, it is claimed that one batch of students is

better than a second batch of students in mathematics. One way to check this is to give the same exam to students in both exams and record the scores. Based on the scores, we have to decide whether the first batch is better than the second (one hypothesis) or whether there is not much difference between the two (the other hypothesis). One can imagine that this can be done by comparing the sample means etc., but that will come later.

A good analogy for testing problems is from law, where the judge has to decide whether an accused is guilty or not guilty. Evidence presented by lawyers take the role of data (but of course one does not really compute any probabilities quantitatively here!).

(4) Regression: Consider two measurements, such as height and weight. It is reasonable to say that weight and height are positively correlated (if the height is larger, the weight tends to be larger too), but is there a more quantitative relationship? Can we predict the weight (roughly) from the height? One could try to see if a linear function fits: wt. $= a$ ht. $+ b$ for some a, b. Or perhaps a more complicated fit such as wt. $= a$ ht. $+ b$ ht.$^2 + c$, etc. To see if this is a good fit, and to know what values of a, b, c to take, we need data. Thus, the problem is that we have some data (H_i, W_i), $i = 1, 2, \ldots, n$, and based on this data we try to find the best linear fit (or the best quadratic fit) etc.

As another example, consider the approximate law that the resistivity of a material is proportional to the temperature. What is the constant of proportionality (for a given material). Here we have a law that says $R = aT$ where a is not known. By taking many measurements at various temperatures we get data (T_i, R_i), $i = 1, 2, \ldots, n$. From this we must find the best possible a (if all the data points were to lie on a line $y = ax$, there would be no problem. In reality they never will, and that is why the choice is an issue!).

2. Estimation problems

Consider the following examples.

(1) A coin has an unknown probability p of turning up head. We wish to determine the value of p. For this, we toss the coin 100 times and observe the outcomes. How to give a guess for the value of p based on the data?
(2) A factory manufacture light bulbs whose lifetimes may be assumed to be exponential random variables with a mean life-time μ. We take a sample of 50 bulbs at random and measure their life-times $X_1, \ldots, X_{50}$. Based on this data, how can we present a reasonable guess for μ? We may want to do this so that the specifications can be printed on the product when sold.
(3) Can we guess the average height μ of all people in India by taking a random sample of 100 people and measuring their heights?

In such questions, there is an unknown parameter μ (there could be more than one unknown parameter too) whose value we are trying to guess based on the data. The data consists of i.i.d. random variables from a family of distributions. We assume that the family of distributions is known and the only unknown is (are) the value of the parameter(s). Rather than present the ideas in abstract let us see a few examples.

Example 155. Let $X_1, \ldots, X_n$ be i.i.d. random variables with Exponential density $f_\mu(x) = \frac{1}{\mu} e^{-x/\mu}$ (fro $x > 0$) where the value of $\mu > 0$ is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

This is the framework in which we would study the second example above, namely the lie-time distribution of light bulbs. Observe that we have parameterized the exponential family of distributions differently from usual. We could equivalently have considered

$g_\lambda(x) = \lambda e^{-\lambda x}$ but the interest is then in estimating $1/\lambda$ (which is the expected value) rather than λ. Here are two methods.

Method of moments: We observe that $\mu = \mathbf{E}_\mu[X_1]$, the mean of the distribution (also called *population mean*). Hence it seems reasonable to take the sample mean $\overline{X}_n$ as an estimate. On second thought, we realize that $\mathbf{E}_\mu[X_1^2] = 2\mu^2$ and hence $\mu = \sqrt{\frac{1}{2}\mathbf{E}_\mu[X_1^2]}$. Therefore it also seems reasonable to take the corresponding sample quantity, $T_n := \sqrt{\frac{1}{2n}(X_1^2 + \ldots + X_n^2)}$ as an estimate for μ. One can go further and write μ in various ways as $\mu = \sqrt{\mathrm{Var}_\mu(X_1)}$, $\mu = \sqrt[3]{\frac{1}{6}\mathbf{E}_\mu[X_1^3]}$ etc. Each such expression motivates an estimate, just by substituting sample moments for population moments.

This is called estimating by the *method of moments* because we are equating the sample moments to population moments to obtain the estimate.

We can also use other features of the distribution, such as quantiles (we may call this the "method of quantiles"). In other words, obtain estimates by equating the sample quantiles to population quantiles. For example, the median of X_1 is $\mu \log 2$, hence a reasonable estimate for μ is $M_n/\log 2$, where M_n is a sample median. Alternately, the 25% quantile of Exponential$(1/\mu)$ distribution is $\mu \log(4/3)$ and hence another estimate for μ is $Q_n/\log(4/3)$ where Q_n is a 25% sample quantile.

Maximum likelihood method: The joint density of $X_1, \ldots, X_n$ is

$$g_\mu(x_1, \ldots, x_n) = \mu^{-n} e^{-\mu(x_1 + \ldots + x_n)} \qquad \text{if all } x_i > 0$$

(since X_i are independent, the joint density is a product). We evaluate the joint density at the observed data values. This is called the likelihood function. In other words, define,

$$L_X(\mu) := \mu^{-n} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}.$$

Two points: This is the joint density of $X_1, \ldots, X_n$, evaluated at the observed data. Further, we like to think of it as a function of μ with $X := (X_1, \ldots, X_n)$ being fixed.

When μ is the actual value, then $L_X(\mu)$ is the "likelihood" of seeing the data that we have actually observed. The *maximum likelihood estimate* is that value of μ that maximizes the likelihood function. In our case, by differentiating and setting equal to zero we get,

$$0 = \frac{d}{d\mu} L_X(\mu) = -n\mu^{-n-1} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i} + \mu^{-n} \left(\frac{1}{\mu^2} \sum_{i=1}^n X_i \right) e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}$$

which is satisfied when $\mu = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}_n$. To distinguish this from the true value of μ which is unknown, it is customary to put a hat on the leter μ. We write $\hat{\mu}_{MLE} = \overline{X}_n$. We should really verify whether $L(\mu)$ is maximized or minimized (or neither) at this point, but we leave it to you to do the checking (eg., by looking at the second derivative).

Let us see the same methods at work in two more examples.

Example 156. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p) random variables where the value of p is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

Method of moments: We observe that $p = \mathbf{E}_p[X_1]$, the mean of the distribution (also called *population mean*). Hence, a method of moments estimator would be the sample mean $\overline{X}_n$.

In this case, $\mathbf{E}_p[X_1^2] = p$ again but we don't get any new estimate because $X_k^2 = X_k$ (as X_k is 0 or 1)

Maximum likelihood method: Now we have a probability mass function instead of density. The joint pmf of of $X_1, \ldots, X_n$ is $f_p(x_1, \ldots, x_n = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$ when each x_i is 0 or 1. The likelihood function is

$$L_X(p) := p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = p^{n\overline{X}_n}(1-p)^{n(1-\overline{X}_n)}.$$

We need to find the value of p that maximizes $L_X(p)$. Here is a trick that almost always simplifies calculations (try it in the previous example too!). Instead of maximizing $L_X(p)$, maximize $\ell_X(p) = \log L_X(p)$ (called the *log-likelihood function*). Since "log" is an increasing function, the maximizer will remain the same. In our case,

$$\ell_X(p) = \overline{X}_n \log p + n(1-\overline{X}_n)\log(1-p).$$

Differentiating and setting equal to 0, we get $\hat{p}_{MLE} = \overline{X}_n$. Again the sample mean is the maximum likelihood estimate.

A last example.

Example 157. Consider the two-parameter Laplace-density $f_{\theta,\alpha}(x) = \frac{1}{2\alpha} e^{-\frac{|x-\theta|}{\alpha}}$ for all $x \in \mathbb{R}$. Check that $f_{\theta,\alpha}$ is indeed a density for all $\theta \in \mathbb{R}$ and $\alpha > 0$.

Now suppose we have data $X_1, \ldots, X_n$ i.i.d. from $f_{\theta,\alpha}$ where we do not know the values of θ and α. How to estimate the parameters?

Method of moments: We compute

$$\mathbf{E}_{\theta,\alpha}[X_1] = \frac{1}{2\alpha}\int_{-\infty}^{+\infty} t e^{-\frac{|t-\theta|}{\alpha}} dt = \frac{1}{2}\int_{-\infty}^{+\infty} (\alpha s + \theta) e^{-|s|} ds = \theta.$$

$$\mathbf{E}_{\theta,\alpha}[X_1^2] = \frac{1}{2\alpha}\int_{-\infty}^{+\infty} t^2 e^{-\frac{|t-\theta|}{\alpha}} dt = \frac{1}{2}\int_{-\infty}^{+\infty} (\alpha s + \theta)^2 e^{-|s|} ds = 2\alpha^2 + \theta^2.$$

Thus the variance is $\mathrm{Var}_{\theta,\alpha}(X_1) = 2\alpha^2$. Based on this, we can take the method of moments estimate to be $\hat{\theta}_n = \overline{X}_n$ (sample mean) and $\hat{\alpha}_n = \frac{1}{\sqrt{2}} s_n$ where $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. At the moment the ideas of defining sample variance as s_n^2 may look strange and it might be more natural to take $V_n := \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ as an estimate for the population variance. As we shall see later, s_n^2 has some desirable properties that V_n lacks. Whenever we say sample variance, we mean s_n^2, unless stated otherwise.

Maximum likelihood method: The likelihood function of the data is

$$L_X(\theta,\alpha) = \prod_{k=1}^n \frac{1}{2\alpha} \exp\left\{-\frac{|X_k - \theta|}{\alpha}\right\} = 2^{-n}\alpha^{-n} \exp\left\{-\sum_{k=1}^n \frac{|X_k-\theta|}{\alpha}\right\}.$$

The log-likelihood function is

$$\ell_X(\theta,\alpha) = \log L(\theta,\alpha) = -n\log 2 - n\log\alpha - \frac{1}{\alpha}\sum_{k=1}^n |X_k - \theta|.$$

We know that[1] for fixed $X_1,\ldots,X_n$, the value of $\sum_{k=1}^n |X_k-\theta|$ is minimized when $\theta = M_n$, the median of $X_1,\ldots,X_n$ (strictly speaking the median may have several choices, all of them are equally good). Thus we fix $\hat{\theta} = M_n$ and then we maximize $\ell(\hat{\theta},\alpha)$ over α by differentiating. We get $\hat{\alpha} = \frac{1}{n}\sum_{k=1}^n |X_k-\theta|$ (the sample mean-absolute deviation about the median). Thus the MLE of (θ,α) is $(\hat{\theta},\hat{\alpha})$.

In homeworks and tutorials you will see several other estimation problems which we list in the exercise below.

Exercise 158. Find an estimate for the unknown parameters by the method of moments and the maximum likelihood method.

(1) $X_1,\ldots,X_n$ are i.i.d. $N(\mu,1)$. Estimate μ. How do your estimates change if the distribution is $N(\mu,2)$?
(2) $X_1,\ldots,X_n$ are i.i.d. $N(0,\sigma^2)$. Estimate σ^2. How do your estimates change if the distribution is $N(7,\sigma^2)$?
(3) $X_1,\ldots,X_n$ are i.i.d. $N(\mu,\sigma^2)$. Estimate μ and σ^2.

[**Note:** The first case is when σ^2 is known and μ is unknown. Then the known value of σ^2 may be used to estimate μ. In the second case it is similar, now μ is known and σ^2 is not known. In the third case, both are unknown].

Exercise 159. $X_1,\ldots,X_n$ are i.i.d. Geo(p) Estimate $\mu = 1/p$.

Exercise 160. $X_1,\ldots,X_n$ are i.i.d. Pois(λ) Estimate λ.

Exercise 161. $X_1,\ldots,X_n$ are i.i.d. Beta(a,b) Estimate a,b.

The following exercise is approachable by the same methods but requires you to think a little.

Exercise 162. $X_1,\ldots,X_n$ are i.i.d. Uniform$[a,b]$ Estimate a,b.

3. Properties of estimates

We have seen that there may be several competing estimates that can be used to estimate a parameter. How can one choose between these estimates? In this section we present some properties that may be considered desirable in an estimator. However, having these properties does not lead to an unambiguous choice of one estimate as the best for a problem.

The setting: Let $X_1,\ldots,X_n$ be i.i.d random variables with a common density $f_\theta(x)$. The parameter θ is unknown and the goal is to estimate it. Let T_n be an estimator for θ, this just means that T_n is a function of $X_1,\ldots,X_n$ (in words, if we have the data at hand, we should be able to compute the value of T_n).

Bias: Define the *bias* of the estimator as $\text{bias}_{T_n}(\theta) := \mathbf{E}_\theta[T_n] - \theta$. If $\text{Bias}_{T_n}(\theta) = 0$ for all values of the parameter θ then we say that T_n is *unbiased* for θ. Here we write θ in the

[1] If you do not know here is an argument. Let $x_1 < x_2 < \ldots < x_n$ be n distinct real numbers and let $a \in \mathbb{R}$. Rewrite $\sum_{k=1}^n |x_k - a|$ as $(|x_1-a|+|x_n-a|)+(|x_2-a|+|x_{n-1}-a|)+\ldots$. By triangle inequality, we see that

$$|x_1-a|+|x_n-a| \geq x_n - x_1, \quad |x_2-a|+|x_{n-1}-a| \geq x_{n-1}-x_2, \quad |x_3-a|+|x_{n-2}-a| \geq x_{n-2}-x_3 \ldots.$$

Further the first inequality is an equality if and only if $x_1 \leq a \leq x_n$, the second inequality is an equality if and only if $x_2 \leq a \leq x_{n-1}$ etc. In particular, if a is a median, then all these inequalities become equalities and shows that a median minimizes the given sum.

subscript of $\mathbf{E}_\theta$ to remind ourself that in computing the expectation we use the density f_θ. However we shall often omit the subscript for simplicity.

Mean-squared error: The *mean squared error* of T_n is defined as $\text{m.s.e.}_{T_n}(\theta) = \mathbf{E}_\theta[(T_n - \theta)^2]$. This is a function of θ. Smaller it is, better our estimate.

In computing mean squared error, it is useful to observe the formula

$$\text{m.s.e.}_{T_n}(\theta) = \text{Var}_{T_n}(\theta) + (\text{Bias}_{T_n}(\theta))^2.$$

To prove this, consider and random variable Y with mean μ and observe that for any real number a we have

$$\begin{aligned}\mathbf{E}[(Y-a)^2] &= \mathbf{E}[(Y-\mu+\mu-a)^2] = \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 + 2(\mu-a)\mathbf{E}[Y-\mu] \\ &= \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 = \text{Var}(Y) + (\mu-a)^2.\end{aligned}$$

Use this identity with T_n in place of Y and θ in place of a.

Example 163. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Let $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \overline{X}_n)^2$ be an estimate for σ^2. By expanding the squares we get

$$V_n = \overline{X}_n^2 + \frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{2}{n}\overline{X}_n \sum_{k=1}^n X_k = \left(\frac{1}{n}\sum_{k=1}^n X_k^2\right) - \overline{X}_n^2.$$

It is given that $\mathbf{E}[X_k] = \mu$ and $\text{Var}(X_k) = \sigma^2$. Hence $\mathbf{E}[X_k^2] = \mu^2 + \sigma^2$. We have seen before that $\text{Var}(\overline{X}_n) = \sigma^2$ and $\mathbf{E}[\overline{X}_n] = \mu$. Hence $\mathbf{E}[\overline{X}_n^2] = \mu^2 + \frac{\sigma^2}{n}$. Putting all this together, we get

$$\mathbf{E}[V_n] = \left(\frac{1}{n}\sum_{k=1}^n \mu^2 + \sigma^2\right) - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Thus, the bias of V_n is $\frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$.

Example 164. For the same setting as the previous example, suppose $W_n = \frac{1}{n}\sum_{k=1}^n (X_k - \mu)^2$. Then it is easy to see that $\mathbf{E}[W_n] = \sigma^2$. Can we say that W_n is an unbiased estimate for σ^2? There is a hitch!

If the value of μ is unknown, then W_n is *not* an estimate (cannot compute it using $X_1, \ldots, X_n$!). However if μ is known, then it is an unbiased estimate. For example, if we knew that $\mu = 0$, then $W_n = \frac{1}{n}\sum_{k=1}^n X_k^2$ is an unbiased estimate for σ^2.

When μ is unknown, we define $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{X}_n)^2$. Clearly $s_n^2 = \frac{n}{n-1}V_n$ and hence $\mathbf{E}[s_n^2] = \frac{n}{n-1}\mathbf{E}[V_n] = \sigma^2$. Thus, s_n^2 is an unbiased estimate for σ^2. Note that s_n^2 depends only on the data and hence it is an estimate, whether μ is known or unknown.

All the remarks in the above two examples apply for any distribution, i.e.,

(1) The sample mean is unbiased for the population mean.
(2) The sample variance $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{X}_n)^2$ is unbiased for the population variance. But $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \overline{X}_n)^2$ is not, in fact $\mathbf{E}[V_n] = \frac{n-1}{n}\sigma^2$.

It appears that s_n^2 is better, but the following remark says that one should be cautious in making such a statement.

Remark 165. In case of $N(\mu, \sigma^2)$ data, it turns out that although s_n^2 is unbiased and V_n is biased, the mean squared error of V_n is smaller! Further V_n is the maximum likelihood estimate of σ^2! Overall, unbiasedness is not so important as having smaller mean squared error, but for estimating variance (when the mean is not known), we always use s_n^2. The computation of the m.s.e is a bit tedious, so we skip it here.

Example 166. Let $X_1,\ldots,X_n$ be i.i.d. Ber(p). Then $\overline{X}_n$ is an estimate for p. It is unbiased since $\mathbf{E}[\overline{X}_n] = p$. Hence, the m.s.e of $\overline{X}_n$ is just the variance which is equal to $p(1-p)/n$.

A puzzle: A coin C_1 has probability p of turning up head and a coin C_2 has probability $2p$ of turning up head. All we know is that $0 < p < \frac{1}{2}$. You are given 20 tosses. You can choose all tosses from C_1 or all tosses from C_2 or some tosses from each (the total is 20). If the objective is to estimate p, what do you do?

Solution: If we choose to have all $n = 20$ tosses from C_1, then we get $X_1,\ldots,X_n$ that are i.i.d. Ber(p). An estimate for p is $\overline{X}_n$ which is unbiased and hence $\mathrm{MSE}_{\overline{X}_n}(p) = \mathrm{Var}(\overline{X}_n) = p(1-p)/n$. On the other hand if we choose to have all 20 tosses from C_2, then we get $Y_1,\ldots,Y_n$ that are i.i.d. Ber$(2p)$. The estimate for p is now $\overline{Y}_n/2$ which is also unbiased and has $\mathrm{MSE}_{\overline{Y}_n/2}(p) = \mathrm{Var}(\overline{Y}_n) = 2p(1-2p)/4 = p(1-2p)/2$. It is not hard to see that for all $p < 1/2$, $\mathrm{MSE}_{\overline{Y}_n/2}(p) < \mathrm{MSE}_{\overline{X}_n}(p)$ and hence choosing C_2 is better, at least by mean-squared criterion! It can be checked that if we choose to have k tosses from C_1 and the rest from C_2, the MSE of the corresponding estimate will be between the two MSEs found above and hence not better than $\overline{Y}_n/2$.

Another puzzle: A factory produces light bulbs having an exponential distribution with mean μ. Another factory produces light bulbs having an exponential distribution with mean 2μ. Your goal is to estimate μ. You are allowed to choose a total of 50 light bulbs (all from the first or all from the second or some from each factory). What do you do?

Solution: If we pick all $n = 50$ bulbs from the first factory, we see $X_1,\ldots,X_n$ i.i.d. Exp$(1/\mu)$. The estimate for μ is $\overline{X}_n$ which has $\mathrm{MSE}_{\overline{X}_n}(\mu) = \mathrm{Var}(\overline{X}_n) = \mu^2/n$. If we choose all bulbs from factory 2 we get $Y_1,\ldots,Y_n$ i.i.d. Exp$(1/2\mu)$. The estimate for μ is $\overline{Y}_n/2$. But $\mathrm{MSE}_{\overline{Y}_n/2}(\mu) = \mathrm{Var}(\overline{Y}_n/2) = (2\mu)^2/4n = \mu^2/n$. The two mean-squared errors are exactly the same!

Probabilistic thinking: Is there any calculation-free explanation why the answers to the two puzzles are as above? Yes, and it is illustrative of what may be called probabilistic thinking. Take the second puzzle. Why are the two estimates same by mean-squared error? Is one better by some other criterion?

Recall that if $X \sim \mathrm{Exp}(1/\mu)$ then $X/2 \sim \mathrm{Exp}(1/2\mu)$ and vice versa. Therefore, if we have data from Exp$(1/\mu)$ distribution, then we can divided all the numbers by 2 and convert it into data from Exp$(1/2\mu)$ distribution. Conversely if we have data from Exp$(1/2\mu)$ distribution, then we can convert it into data from Exp$(1/\mu)$ distribution by multiplying each number by 2. Hence there should be no advantage in choosing either factory. We leave it for you to think in analogous ways why in the first puzzle C_2 is better than C_1.

4. Confidence intervals

So far, in estimating of an unknown parameter, we give a single number as our guess for the known parameter. It would be better to give an interval and say with what confidence we expect the true parameter to lie within it. As a very simple example, suppose

we have one random variable X with $N(\mu,1)$ distribution. How do we estimate μ? Suppose the observed value of X is 2.7. Going by any method, the guess for μ would be 2.7 itself. But of course μ is not equal to X, so we would like to give an interval in which μ lies. How about $[X-1,X+1]$? Or $[X-2,X+2]$? Using normal tables, we see that $\mathbf{P}(X-1<\mu<X+1)=\mathbf{P}(-1<(X-\mu)<1)=\mathbf{P}(-1<Z<1)\approx 0.68$ and similarly $\mathbf{P}(X-2<\mu<X+2)\approx 0.95$. Thus, by making the interval longer we can be more confident that the true parameter lies within. But the accuracy of our statement goes down (if you want to know the average height of people in India, and the answer you give is "between 100cm and 200cm", it is very probably correct, but of little use!). The probability with which our CI contains the unknown parameter is called the level of confidence. Usually we fix the level of confidence, say as 0.90 and find an interval *as short as possible* but subject to the condition that it should have a confidence level of 0.90.

In this section we consider the problem of confidence intervals in Normal population. In the next we see a few other examples.

The setting: Let $X_1,\ldots,X_n$ be i.i.d. $N(\mu,\sigma^2)$ random variables. We consider four situations.

(1) Confidence interval for μ when σ^2 is known.
(2) Confidence interval for σ^2 when μ is known.
(3) Confidence interval for μ when σ^2 is unknown.
(4) Confidence interval for σ^2 when μ is unknown.

A starting point in finding a confidence interval for a parameter is to first start with an estimate for the parameter. For example, in finding a CI for μ, we may start with $\overline{X}_n$ and enlarge it to an interval $[\overline{X}_n-a,\overline{X}_n+a]$. Similarly, in finding a CI for σ^2 we use the estimate $s_n^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X}_n)^2$ if μ is unknown and $W_n=\frac{1}{n}\sum_{i=1}^n(X_i-\mu)^2$ if the value of μ is known.

4.1. Estimating μ when σ^2 is known. We look for a confidence interval of the form $I_n=[\overline{X}_n-a,\overline{X}_n+a]$. Then,

$$\mathbf{P}\left(I_n\ni\mu\right)=\mathbf{P}\left(-a\le\overline{X}_n-\mu\le a\right)=\mathbf{P}\left(-\frac{a\sqrt{n}}{\sigma}\le\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}\le\frac{a\sqrt{n}}{\sigma}\right)$$

Now we use two facts about normal distribution that we have seen before.

(1) If $Y\sim N(\mu,\sigma^2)$ then $aX+b\sim N(a\mu+b,a^2\sigma^2)$.
(2) If $Y_1\sim N(\mu,\sigma^2)$ and $Y_2\sim N(\nu,\tau^2)$ and they are independent, then $X+Y\sim N(\mu+\nu,\sigma^2+\tau^2)$.

Consequently, $\overline{X}_n\sim N(0,\sigma^2/n)$ and $\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma}\sim N(0,1)$. Therefore,

$$\mathbf{P}\left(I_n\ni\mu\right)=\mathbf{P}(-\frac{a\sqrt{n}}{\sigma}\le Z\le-\frac{a\sqrt{n}}{\sigma})$$

where $Z\sim N(0,1)$. Fix any $0<\alpha<1$ and denote by z_α the number such that $\mathbf{P}(Z>z_\alpha)=\alpha$ (in other words, z_α is the $(1-\alpha)$-quantile of the standard normal distribution). For example, from normal tables we find that $z_{0.05}\approx 1.65$ and $z_{0.005}\approx 2.58$ etc.

If we set $a=z_{\alpha/2}\sigma/\sqrt{n}$, we get

$$\mathbf{P}\left(\left[\overline{X}_n-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\overline{X}_n+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right]\ni\mu\right)=1-\alpha.$$

This is our confidence interval.

4.2. Estimating σ^2 when μ is known. Since μ is known, we use $W_n = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ to estimate σ^2. Here is an exercise.

Exercise 167. Let $Z_1,\ldots,Z_n$ be i.i.d. $N(0,1)$ random variables. Then, $Z_1^2+\ldots+Z_n^2 \sim$ Gamma$(n/2,1/2)$.

Solution: For $t>0$ we have

$$\mathbf{P}\{Z_1^2 \le t\} = \mathbf{P}\{-\sqrt{t} \le Z_1 \le \sqrt{t}\} = 2\int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}du = \frac{1}{\sqrt{2\pi}}\int_0^t e^{-s/2}s^{-1/2}ds.$$

Differentiate w.r.t t to see that the density of Z_1^2 is $h(t) = \frac{1}{\sqrt{\pi}}e^{-t/2}t^{-1/2}\sqrt{(1/2)}$, which is just the Gamma$(\frac{1}{2},\frac{1}{2})$ density.

Now, each Z_k^2 has the same Gamma$(\frac{1}{2},\frac{1}{2})$ density, and they are independent. Earlier we have seen that when we add independent Gamma random variables with the same scale parameter, the sum has a Gamma distribution with the same scale but whose shape parameter is the sum of the shape parameters of the individual summands. Therefore, $Z_1^2+\ldots+Z_n^2$ has Gamma$(n/2,1/2)$ distribution. This completes the solution to the exercise.

In statistics, the distribution Gamma$(1/2,1/2)$ is usually called the *chi-squared distribution with n degrees of freedom*. Let $\chi_n^2(\alpha)$ denote the $1-\alpha$ quantile of this distribution. Similarly, $\chi_n^2(1-\alpha)$ is the α quantile (i.e., the probability for the chi-squared random variable to fall below $\chi_n^2(1-\alpha)$ is exactly α).

When X_i are i.i.d. $N(\mu,\sigma^2)$, we know that $(X_i-\mu)/\sigma$ are i.i.d. $N(0,1)$. Hence, by the above fact, we see that

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

has chi-squared distribution with n degrees of freedom. Hence

$$\mathbf{P}\left\{\frac{nW_n}{\chi_n^2\left(\frac{\alpha}{2}\right)} \le \sigma^2 \le \frac{nW_n}{\chi_n^2\left(1-\frac{\alpha}{2}\right)}\right\} = \mathbf{P}\left\{\chi_n^2\left(1-\frac{\alpha}{2}\right) \le \frac{nW_n}{\sigma^2} \le \chi_n^2\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha.$$

Thus, $\left[\frac{ns_n^2}{\chi_{n-1}^2\left(\frac{\alpha}{2}\right)}, \frac{ns_n^2}{\chi_{n-1}^2\left(1-\frac{\alpha}{2}\right)}\right]$ is a $(1-\alpha)$-confidence interval for σ^2.

An important result: Before going to the next two confidence interval problems, let us try to understand the two examples already covered. In both cases, we came up with a random variable ($\sqrt{n}(\overline{X}_n-\mu)/\sigma$ and W_n/σ^2, respectively) which involved the data and the unknown parameter whose distributions we knew (standard normal and χ_n^2, respectively) and these distributions do not depend on any parameters. This is generally the key step in any confidence interval problem. For the next two problems, we cannot use the same two random variables as above as they depend on the other unknown parameter too (i.e., $\sqrt{n}(\overline{X}_n-\mu)/\sigma$ uses σ which will be unknown and W_n/σ^2 uses μ which will be unknown). Hence, we need a new result that we state without proof.

Theorem 168. *Let $Z_1,\ldots,Z_n$ be i.i.d. $N(\mu,\sigma^2)$ random variables. Let $\overline{Z}_n$ and s_n^2 be the sample mean and the sample variance, respectively. Then,*

$$\overline{Z}_n \sim N(\mu, \frac{\sigma^2}{n}), \quad \frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2,$$

and the two are independent.

This is not too hard to prove (a muscle-flexing exercise in change of variable formula) but we skip the proof. Note two important features. First, the surprising independence of the sample mean and the sample variance. Second, the sample variance (appropriately scaled) has χ^2 distribution, just like W_n in the previous example, but the degree of freedom is reduced by 1. Now we use this theorem in computing confidence intervals.

4.3. Estimating σ^2 when μ is unknown. The estimate s_n^2 must be used as W_n depends on μ which is unknown. Theorem thm:indepofsamplemeanandvar tells us that $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$. Hence, by the same logic as before we get

$$\mathbf{P}\left\{\frac{(n-1)s_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)} \le \sigma^2 \le \frac{(n-1)s_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right\} = \mathbf{P}\left\{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right) \le \frac{(n-1)s_n^2}{\sigma^2} \le \chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right\}$$
$$= 1-\alpha.$$

Thus, $\left[\frac{(n-1)s_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{(n-1)s_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right]$ is a $(1-\alpha)$-confidence interval for σ^2.

If μ is known, we could use the earlier confidence interval using W_n, or simply ignore the knowledge of μ and use the above confidence interval using s_n^2. What is the difference? The cost of ignoring the knowledge of μ is that the second confidence interval will be typically larger, although for large n the difference is slight. On the other hand, if our knowledge of μ was inaccurate, then the first confidence interval is invalid (we have no idea what its level of confidence is!) which is more serious. In realistic situations it is unlikely that we will know one of the parameters but not the other - hence, most often one just uses the confidence interval based on s_n^2.

4.4. Estimating μ when σ^2 is unknown. The earlier confidence interval We look for a confidence interval $[\overline{X}_n - \frac{\sigma}{\sqrt{n}}z_{\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}z_{\alpha/2}]$ cannot be used as we do not know the value of σ.

A natural idea would be to use the estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ in place of σ^2. However, recall that the earlier confidence interval (in particular, the cut-off values $z_{\alpha/2}$ in the CI) was an outcome of the fact that

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

Is it true if σ is replaced by s_n? Actually no, but we have a different distribution called *Student's t-distribution*.

Exercise 169. Let $Z \sim N(0,1)$ and $S^2 \sim \chi^2_n$ be independent. Then, the density of $\frac{Z}{S/\sqrt{n}}$ is given by

$$\frac{1}{\sqrt{n-1}\text{Beta}(\frac{1}{2}, \frac{n-1}{2})} \frac{1}{\left(1+\frac{t^2}{n-1}\right)^{\frac{n}{2}}}$$

for all $t \in \mathbb{R}$. This is known as *Student's t-distribution*.

The exact density of t-distribution is not important to remember, so the above exercise is optional. The point is that it can be computed from the change of variable formula and that by numerical integration its CDF can be tabulated.

How does this help us? From Theorem 168 we know that $\frac{\sqrt{n}(\overline{X}_n-\mu)}{\sigma} \sim N(0,1)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$, and the two are independent. Take these random variables in the above exercise to conclude that $\frac{\sqrt{n}(\overline{X}_n-\mu)}{s_n}$ has t_{n-1} distribution.

The t-distribution is symmetric about zero (the density at t and at $-t$ are the same). Further, as the number of degrees of freedom goes to infinity, the t-density converges to the standard normal density. What we need to know is that there are tables from which we can read off specific quantiles of the distribution. In particular, by $t_n(\alpha)$ we mean the $1-\alpha$ quantile of the t-distribution with n degrees of freedom. Then of course, the α quantile is $-t_n(\alpha)$.

Returning to the problem of the confidence interval, from the fact stated above, we see that (use T_n to indicate a random variable having t-distribution with n degrees of freedom).

$$\begin{aligned}
&\mathbf{P}\left(\overline{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \leq \mu \leq \overline{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbf{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbf{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \leq T_{n-1} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1-\alpha.
\end{aligned}$$

Hence, our $(1-\alpha)$-confidence interval is $\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \overline{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right]$.

Remark 170. We remarked earlier that as $n \to \infty$, the t_{n-1} density approaches the standard normal density. Hence, $t_{n-1}(\alpha)$ approaches z_α for any α (this can be seen by looking at the t-table for large degree of freedom). Therefore, when n is large, we may as well use

$$\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right].$$

Strictly speaking the level of confidence is smaller than for the one with $t_{n-1}(\alpha/2)$. However for n large the level of confidence is quite close to $1-\alpha$.

5. Confidence interval for the mean

Now suppose $X_1, \ldots, X_n$ are i.i.d. random variables from some distribution with mean μ and variance σ^2, both unknown. How can we construct a confidence interval for μ?

In case of normal distribution, recall that the $(1-\alpha)$-CI that we gave was

$$\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \overline{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right] \text{ or } \left[\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$$

Is this a valid confidence interval in general? The answer is "No" for both. If X_i are from some general distribution then the distributions of $\sqrt{n}(\overline{X}_n - \mu)/s_n$ and $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ are very complicated to find. Even if X_i come from binomial or exponential family, these distributions will depend on the parameters in a complex way (in particular, the distributions are not free from the parameters, which is important in constructing confidence intervals).

But suppose n is large. Then the sample variance is close to population variance and hence $s_n \approx \sigma$. Further, by CLT, we know that $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ has approximately $N(0,1)$ distribution. Hence, we see that

$$\mathbf{P}\left\{-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \leq z_{\alpha/2}\right\} \approx \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1-\alpha.$$

Consequently, we may say that

$$\mathbf{P}\left\{\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right\} \approx 1-\alpha.$$

Thus, $\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$ is an approximate $(1-\alpha)$-confidence interval. Further, when n is large, the difference between $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ and $\hat{s}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ is small (indeed, $s_n^2 = (n/(n-1))\hat{s}_n^2$). Hence it is also okay to use $\left[\overline{X}_n - \frac{\hat{s}_n}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{\hat{s}_n}{\sqrt{n}} z_{\alpha/2}\right]$ as an approximate $(1-\alpha)$-confidence interval.

Example 171. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Consider the problem of finding a confidence interval for p. Since each X_i is 0 or 1, observe that

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

Hence, an approximate $(1-\alpha)$-CI for p is given by

$$\left[\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

6. Actual confidence by simulation

Suppose we have a candidate confidence interval whose confidence we do not know. For example, let us take the confidence interval

$$\left[\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

for the parameter p of i.i.d. Ber(p) samples. We saw that for large n this has approximately $(1-\alpha)$ confidence. But how large is large? One way to check this is by simulation. We explain how.

Take $p = 0.3$ and $n = 10$. Simulate $n = 10$ independent Ber(p) random variables and compute the confidence interval given above. Check whether it contains the true value of p (i.e., 0.3) or not. Repeat this exercise 10000 times and see what proportion of times it contains 0.3. That proportion is the true confidence, as opposed to $1-\alpha$ (which is valid only for large n). Repeat this experiment with $n = 20$, $n = 30$ etc. See how close the actual confidence is to $1-\alpha$. Repeat this experiment with different value of p. The n you need to get close to $1-\alpha$ will depend on p (in particular, on how close p is to $1/2$).

This was about checking the validity of a confidence interval that was specified. In a real situation, it may be that we can only get $n = 20$ samples. Then what can we do? If we have an idea of the approximate value of p, we can first simulate Ber(p) random numbers on a computer. We compute the sample mean each time, and repeat 10000 times to get so many values of the sample mean. Note that the histogram of these 10000 values tells us (approximately) the actual distribution of $\overline{X}_n$. Then we can find t (numerically) such that $[\overline{X}_n - t, \overline{X}_n + t]$ contains the true value of p in $(1-\alpha)$-proportion of the 10000 trials. Then, $[\overline{X}_n - t, \overline{X}_n + t]$ is a $(1-\alpha)$-CI for p. Alternately, we may try a CI of the form

$$\left[\overline{X}_n - t\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + t\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

where we choose t numerically to get $(1-\alpha)$ confidence.

Summary: The gist of this discussion is this. In the neatly worked out examples of the previous sections, we got explicit confidence intervals. But we assumed that we knew the

data came from $N(\mu, \sigma^2)$ distribution. What if that is not quite right? What if it is not any of the nicely studied distributions? The results also become invalid in such cases. For large n, using law of large numbers and CLT we could overcome this issue. But for small n? The point is that using simulations we can calculate probabilities, distributions, etc, numerically and approximately. That is often better, since it is more robust to assumptions.