## 7.2. The methods of moments and maximum likelihood

In this section, we introduce two methods for obtaining point estimators: the method of moments and the method of maximum likelihood. Although maximum likelihood estimators are generally preferable to moment estimators because of efficiency, they often require more computation than moment estimators. Sometimes, two methods produce the same estimator.

### The method of moments

**Definition.** Let $X_1, \dots, X_n$ be a random sample from some distribution. For $k = 1, 2, \dots,$ the kth population moment or kth moment of the distribution, is $E(X^k)$. The kth sample moment is $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.

- $E(X) = \mu \rightarrow$ first population moment
- $\frac{1}{n}\sum X_i = \bar{X} \rightarrow$ first sample moment

**Definition.** Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution depending on parameters $\theta_1, \dots, \theta_m$ whose values are unknown. Then the method of moments estimators (mmes) $\hat{\theta}_1, \dots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \dots, \theta_m$.

**Example.** Let $X_1, \dots, X_n$ be a random sample of service times from an exponential distribution with parameter $\lambda$. Find the mme of $\lambda$.

**Solution.** Since there is only one parameter to be estimated, the estimator is obtained by equating $E(X)$ to $\bar{X}$. For exponential dist., $E(X) = 1/\lambda$. Then, we obtain

$$\frac{1}{\lambda} = \bar{X} \Rightarrow \lambda = \frac{1}{\bar{X}}.$$

Thus, the mme of $\lambda$ is $\hat{\lambda} = 1/\bar{X}$.

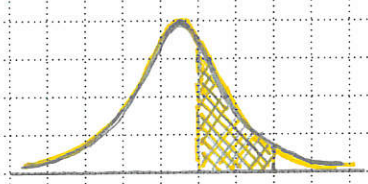
## Maximum likelihood estimation

This method is recommended, at least when the sample size is large, since the resulting estimators have certain desirable efficiency properties.

NOTE. Probability and likelihood are different concepts.

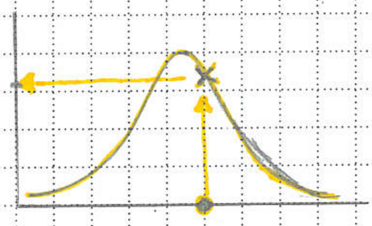
Probabilities are the areas under a fixed distribution

P( data | distribution)



Likelihoods are the y-axis values for fixed data points with distributions that can be moved.

L( distribution | data)



Definition. Let $X_1, \dots, X_n$ have a joint distribution (pmf or pdf) that depends on a parameter $\theta$ whose value is unknown. The joint dist., regarded as a function of $\theta$, is called the likelihood function and is denoted by $L(\theta)$. The maximum likelihood estimate (mle) $\hat{\theta}$ is the value of $\theta$ that maximizes the likelihood function.

The likelihood function tells us how likely the observed sample is, as a function of the possible parameter value. Maximizing the likelihood gives the parameter value for which the observed sample is most likely to have been generated, that is, the parameter value that "agrees most closely" with the observed data. Maximizing likelihood equals maximizing the logarithm of the likelihood, $\ell(\theta)$.

NOTATION. $\ell(\theta) = \ln[L(\theta)]$ : log-likelihood function.

Example. Suppose $X_1, \dots, X_n$ is a random sample from an exponential distribution with parameter $\lambda$.

- Find mle of $\lambda$, and compare it with mme of $\lambda$.

Solution. Because of independence, the likelihood function is a product of the individual pdfs:

$$L(\lambda) = f(x_1, \dots, x_n)$$
$$= (\lambda e^{-\lambda x_1}) \cdot \dots \cdot (\lambda e^{-\lambda x_n})$$
$$= \lambda^n e^{-\lambda \sum x_i}$$

Now, we determine the value of $\lambda$ that maximizes $\ln[L(\lambda)]$:

$$l(\lambda) = \ln[L(\lambda)]$$
$$= \ln[\lambda^n e^{-\lambda \sum x_i}]$$
$$= n \ln(\lambda) - \lambda \sum x_i$$

When we equate the derivative of $l'(\lambda)$ to zero, we obtain

$$l'(\lambda) = \frac{n}{\lambda} - \sum x_i = 0$$

$$\Rightarrow \quad \lambda = \frac{n}{\sum x_i} = \frac{n/n}{1/n \sum x_i} = \frac{1}{\bar{x}}.$$

Then the mle is $\hat{\lambda} = \frac{1}{\bar{X}}$. In previous example (page 1), we found that the mme is $\hat{\lambda} = \frac{1}{\bar{X}}$. The two methods often gives the same estimator.

....../....../........

The definition of the mles can be extended to distributional families that include two or more parameters.

Some properties of MLEs

The mle of $\sigma^2$, as well as many other mles, can be easily derived using the following proposition.

MLE Invariance Principle. Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ be the mles of the parameters $\theta_1, \theta_2, \ldots, \theta_m$. Then the mle of any function $h(\theta_1, \ldots, \theta_m)$ of these parameters is the function $h(\hat{\theta}_1, \ldots, \hat{\theta}_m)$ of the mles.

Example. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution, which includes the two parameters $\mu$ and $\sigma$.

a) Find the mles of $\mu$ and $\sigma$.

b) Find the mles of $\sigma^2$ and the population coefficient of variation $h(\mu, \sigma) = 100\mu/\sigma$.

Solution. a) The likelihood function is

$$L(\mu,\sigma) = f(x_1, \ldots, x_n; \mu, \sigma)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum (x_i-\mu)^2}{2\sigma^2}}$$

and

$$\ell(\mu,\sigma) = \ln[L(\mu,\sigma)] = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum (x_i-\mu)^2}{2\sigma^2}$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum (x_i-\mu)^2$$

$$= -\frac{n}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2\sigma^2}\sum (x_i-\mu)^2$$

To find mise of $\mu$ and $\sigma$, we take the partial derivatives and equate them zero:

NEXT PAGE

$$\frac{\partial \ell}{\partial \mu} = + \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \mu$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu$$

$$\Rightarrow \mu = \bar{x}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (x_i - \mu)^2 = 0 \Rightarrow \frac{1}{\sigma^3} \sum (x_i - \mu)^2 = \frac{n}{\sigma}$$

$$\Rightarrow \sum (x_i - \mu)^2 = \sigma^2 n$$

$$\Rightarrow \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

$$\Rightarrow \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Thus, the mles of $\mu$ and $\sigma$ are

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}.$$

b) The mle of the function $h(\mu, \sigma) = \sigma^2$ is

$$\widehat{\sigma^2} = \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

The mle of the population coefficient of variation $h(\mu, \sigma) = 100\mu/\sigma$ is $100\hat{\mu}/\hat{\sigma}$.

Theorem. Under very general conditions on the joint distribution of the sample, when the sample size is large, the maximum likelihood estimator of any parameter $\theta$

i) is close to $\theta$ (consistency),

ii) is approximately unbiased, and

iii) has variance that is nearly as small as can be achieved by any unbiased estimator. In other words, the mle $\hat{\theta}$ is at least approximately the MVUE of $\theta$.

## Some complications

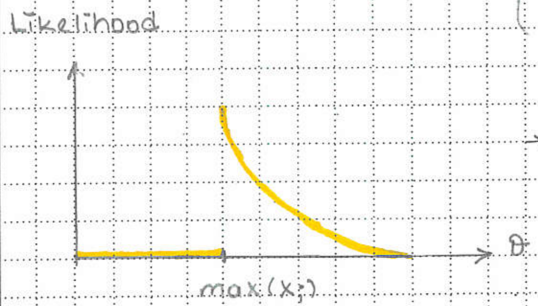
Sometimes calculus cannot be used to obtain mles.

**Example.** Suppose the waiting time for a bus is uniformly distributed on $[0, \theta]$. Let $X_1, \ldots, X_n$ denote the observed waiting times from a random sample drawn from this distribution. Find the mle of $\theta$.

**Solution.** The likelihood function is

$$L(\theta) = f(x_1, \ldots, x_n; \theta)$$

$$= \begin{cases} \dfrac{1}{\theta^n} & , 0 \leq x_1 \leq \theta, \ldots, 0 \leq x_n \leq \theta \\ 0 & \text{otherwise} \end{cases}$$




→ If $\theta \geq \max(x_i)$, then $L(\theta) = \dfrac{1}{\theta^n} > 0$

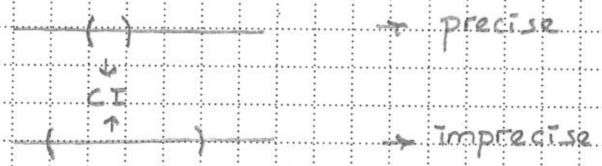
If $\theta < \max(x_i)$, then $L(\theta) = 0$.

In this case, calculus does not work because the maximum of the likelihood occurs at the discontinuity point. The figure shows that the mle is $\hat{\theta} = \max(x_i)$. Thus, if the waiting times are 2.3, 3.7, 1.5, 0.4, and 3.2, then the mle is $\hat{\theta} = 3.7$.

## 8. Statistical Intervals Based on a Single Sample

A point estimate by itself provides no information about the precision and reliability. An alternative way is to calculate and report an entire interval of plausible values — an interval estimate or confidence interval (CI).

A confidence interval is calculated by first selecting a confidence level, which is a measure of the degree of reliability of the interval. The higher the confidence level, the more strongly we believe that the value of the parameter being estimated lies within the interval.




### 8.1. Basic properties of confidence intervals

In this part, we consider a simple case to introduce the foundational concepts. Suppose a random sample from a $N(\mu, \sigma)$ distribution. We, then, know that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \rightarrow Z \text{ has a standard normal dist.}$$

Because the area under the standard normal curve between $-1.96$ and $1.96$ is $.95$,

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95 \qquad \ldots (1)$$

If we manipulate this equation, we obtain

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95 \qquad \ldots (2)$$

or, in interval notation,

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \; \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \qquad \ldots (3)$$

This is a random interval as the endpoints involve a r.v. However, the interval's width $(2 \cdot 1.96\sigma/\sqrt{n})$ is not random.

Definition. If after observing $X_1 = x_1, \dots, X_n = x_n$, we compute the observed sample mean $\bar{x}$ and then substitute $\bar{x}$ into (3) in place of $\bar{X}$, the resulting fixed interval is called a 95% confidence interval for $\mu$. This CI can be expressed either as

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \text{ is a 95\% CI for } \mu$$

or as

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \quad \text{with 95\% confidence.}$$

A concise expression for the interval is $\bar{x} \pm 1.96\,\sigma/\sqrt{n}$.

Interpreting a confidence level

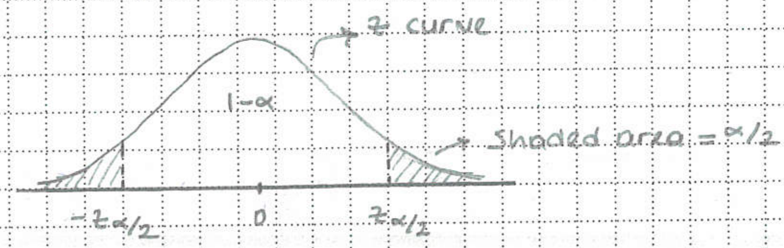
A confidence interval gives a range for the parameter, and a confidence level that the range covers the true value. Confidence levels involve thinking not only about the actual sample but about other samples that could have been drawn. Imagine constructing a 95% CI from 100 different random samples. The interval will vary from sample to sample. For about 95% of these samples, the resulting intervals will contain the true population parameter. Here, 95% is called the confidence level.

Other levels of confidence

Definition. A $100(1-\alpha)\%$ confidence interval for the mean $\mu$ of a normal population when the value of $\sigma$ is known is given by

$$\left(\bar{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \ \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

or, equivalently, by $\bar{x} \pm z_{\alpha/2}\cdot\sigma/\sqrt{n}$.




$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$$

Confidence level, precision, and choice of sample size

• Higher confidence levels require wider confidence intervals. A 99% CI is wider than 95% CI because it captures the true parameter more often.

• Confidence (reliability) and precision (narrowness) are inversely related. Namely, higher confidence ⇒ lower precision.

• A 100% CI is uninformative $(-\infty, \infty)$.

• A good strategy is to specify both the desired confidence level and interval width, and then determine the necessary sample size.

Example. Consider a normal population distribution with the value of $\sigma$ known.

a) What is the confidence level for the interval $\bar{x} \pm 2.81\,\sigma/\sqrt{n}$ ?

b) What value of $z_{\alpha/2}$ in the CI formula results in a confidence level of 99.7% ?

Solution.

a) $z_{\alpha/2} = 2.81$ implies that $\alpha/2 = 1 - \overbrace{\Phi(2.81)}^{.9975} = .0025$.

by z table

Now, we have $\alpha/2 = .0025$, then $\alpha = .005$.

Then, the confidence level is $100(1-\alpha)\% = 100(1-.005)\% = 99.5\%$

b) 99.7% implies that $\alpha = .003$, $\alpha/2 = .0015$, and $z_{.0015} = 2.96$.

Example. Extensive monitoring of a certain operating system has suggested that response time to a particular editing command is normally distributed with stand. dev. 25 ms. A new operating system has been installed, and an estimate of the true average response time $\mu$ for the new environment is desired. Assuming that response times are still normally distributed with $\sigma = 25$, what sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10 ?

Solution. Recall that $CI = 2(1.96)\frac{\sigma}{\sqrt{n}}$ is the width of interval associated with 95% C.I. Then, the sample size n must satisfy

$$10 = 2(1.96)\frac{25}{\sqrt{n}}$$

$$\Rightarrow \sqrt{n} = \frac{2(1.96)\,25}{10} = 9.80$$

$$\Rightarrow n = (9.80)^2 = 96.04.$$

Since n must be an integer, a sample size of 97 is required.

The general formula for the sample size n necessary to ensure an interval width w is obtained from $w = 2 z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ as

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{w/2}\right)^2.$$

→ The half-width $1.96\frac{\sigma}{\sqrt{n}}$ of the 95% CI is sometimes called the margin of error associated with a 95% confidence level.

Deriving a general confidence interval

The general strategy for deriving a C.I. is pivotal quantity.

Definition. Suppose a rv satisfying the following two properties can be found:

1. The variable is a function of both $X_1, \dots, X_n$ and $\theta$.
2. The probability dist. of the variable does not depend on $\theta$ or on any other unknown parameters.

Such a rv is called a pivotal quantity.

Example. If the population distribution is normal with $\sigma$ known and $\theta = \mu$ unknown, the variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a pivotal quantity: 1. The function depends on $X_i$'s and $\theta = \mu$.

2. $Z \sim N(0,1)$; that is, Z does not depend on $\mu$.

In general, the form a pivotal quantity is usually suggested by examining the distribution of an appropriate estimator $\hat{\theta}$.

Let $h(X_1, \dots, X_n, \theta)$ be a general pivotal quantity. For $\alpha$, $0 \le \alpha \le 1$, and constants $a$ and $b$, not involving $\theta$, the lower and upper confidence limits can be written as

$$P(a < h(X_1, \dots, X_n, \theta) < b) = 1-\alpha$$

$$\Rightarrow \quad P(l(X_1, \dots, X_n) < \theta < u(X_1, \dots, X_n)) = 1-\alpha$$

Obtaining the second equation follows the same steps as the derivation of a CI for $Z = (\bar{X}-\mu)/(\sigma/\sqrt{n})$, and represents a general formulation of CIs based on pivotal quantities. For instance, in the normal example, we saw that $l(X_1, \dots, X_n) = \bar{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}$ and $u(X_1, \dots, X_n) = \bar{X} + z_{\alpha/2}\, \sigma/\sqrt{n}$, where $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$.

Example. A theoretical model suggests that the time-to-breakdown of an insulating fluid between electrodes at a particular voltage has an exponential distribution with unknown parameter $\lambda$. A random sample of $n=10$ breakdown times yields the following sample data (in min):

$x_1 = 41.53$, $x_2 = 18.73$, $x_3 = 2.99$, $x_4 = 30.34$, $x_5 = 12.33$

$x_6 = 117.52$, $x_7 = 73.02$, $x_8 = 223.63$, $x_9 = 4.00$, $x_{10} = 26.78$

Derive a 95% CI for both $\lambda$ and for the true average breakdown time.

Solution. Let $h(X_1, \dots, X_n, \lambda) = 2\lambda \sum X_i$. Here $X_i \sim Exp(\lambda)$, $i = 1, \dots, n$.

Then, $M_{X_i}(t) = \dfrac{\lambda}{\lambda - t}$, $t < \lambda$, and $M_{\sum X_i}(t) = \left(\dfrac{\lambda}{\lambda - t}\right)^n = \left(1 - \dfrac{t}{\lambda}\right)^{-n}$.

This is the gamma mgf with $\alpha = n$ and $\beta = 1/\lambda$. Recall that $\chi^2_\nu \sim Gamma(\alpha = \nu/2, \beta = 2)$. Scaling $\sum X_i$ by $2\lambda$ yields $Gamma(\alpha = n, \beta = 2)$, which means $2\lambda \sum X_i$ has a $\chi^2_{2n} \sim Gamma(n, 2)$.

Since $h$ is a function of both $X_i$'s and $\lambda$, yet its distribution $\chi^2_{2n}$ does not depend on $\lambda$, it is a pivotal quantity.

95% means that $1-\alpha = .9500$, then $\alpha = .05$. Then $\alpha/2 = .025$.
So we need the quantiles $\chi^2_{.025,20}$ and $\chi^2_{.975,20}$. Also, we have $\nu = 2n = 2(10) = 20$. Then, by Table A.5, we obtain

$$P(9.591 < 2\lambda \sum X_i < 34.170) = .95$$

$$\Rightarrow P\left(\frac{9.591}{2\sum X_i} < \lambda < \frac{34.170}{2\sum X_i}\right) = .95,$$

which represents the CI for $\lambda$. Here, the lower limit is $l = \frac{9.591}{2\sum X_i}$ and the upper limit is $u = \frac{34.170}{2\sum X_i}$.

For the given data, $\sum x_i = 550.87$ and the relevant interval is $(.00871, .03101)$. Based on the data, we are 95% confident that the true value of the parameter $\lambda$ is within $(.00871, .03101)$.

The mean of an exponential rv is $\mu = 1/\lambda$. Then, we write

$$P\left(\frac{2\sum X_i}{34.170} < \frac{1}{\lambda} < \frac{2\sum X_i}{9.591}\right) = .95$$

and

$$\left(\frac{2\sum X_i}{34.170}, \frac{2\sum X_i}{9.591}\right).$$

By setting the data values in this interval, we get the 95% CI for true average breakdown time as $\left(\frac{2\sum x_i}{34.170}, \frac{2\sum x_i}{9.591}\right) = (32.24, 114.87)$,

which is a quite wide interval. This means that variability is high. Also, the endpoints are not equidistant from the point estimate unlike the normal case.

A general large-sample confidence interval

Let $X_1, \dots, X_n$ be a random sample from any population having a mean $\mu$ and and standard deviation $\sigma$. For large $n$, CLT implies $\bar{X}$ is approximately normal, regardless of the population. $Z = (\bar{X}-\mu)/(\sigma/\sqrt{n})$ is approximately standard normal. Then, we have

$$P\left(-z_{\alpha/2} < (\bar{X}-\mu)/(\sigma/\sqrt{n}) < z_{\alpha/2}\right) \approx 1-\alpha,$$

This yields an approximate $100(1-\alpha)\%$. That is, when $n$ is large, the CI, $\bar{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$ for $\mu$ remains valid whatever the population distribution.

Now, let us generalize the discussion above. Suppose that $\hat{\theta}$ is an estimator of a parameter $\theta$ with the properties

1. $\hat{\theta}$ has approximately a normal dist.
2. $\hat{\theta}$ is (at least approximately) unbiased for $\theta$; and
3. $\sigma_{\hat{\theta}}$, the standard deviation of $\hat{\theta}$, is available.

We know that, under very general conditions, mle $\hat{\theta}$ satisfies the first two properties when $n$ is large.

Then,

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

has approximately a standard normal dist., which makes $Z$ an approximate pivotal quantity. Then,

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1-\alpha,$$

so that a $100(1-\alpha)\%$ CI for $\theta$ may potentially be obtained. But, this depends on the formula for $\sigma_{\hat{\theta}}$:

Case I. $\sigma_{\hat{\theta}}$ does not involve any unknown parameters.
Solving directly for $\theta$ yields $\hat{\theta} \pm z_{\alpha/2}\, \sigma_{\hat{\theta}}$.

Case II. $\sigma_{\hat{\theta}}$ involves at least one unknown parameter but not $\theta$.
Replacing unknown parameters with their estimates gives $s_{\hat{\theta}}$.
Under general conditions, we obtain $\hat{\theta} \pm z_{\alpha/2}\, s_{\hat{\theta}}$. For instance, the interval $\bar{x} \pm z_{\alpha/2}\, s/\sqrt{n}$ is an example of this case.

Case III. $\sigma_{\hat{\theta}}$ involves the unknown $\theta$ itself.
Solving $\dfrac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = z_{\alpha/2}$ can be difficult. Replacing $\theta$ in $\sigma_{\hat{\theta}}$ by its estimate $\hat{\theta}$ and getting an approximate solution is often used.
This results in $s_{\hat{\theta}}$; accordingly the interval $\hat{\theta} \pm z_{\alpha/2}\, s_{\hat{\theta}}$.

One-sided confidence intervals (confidence bounds)

Sometimes we may want to have only one of the bounds. For instance, a surgeon may want only a lower confidence bound for true average remission time after colon cancer surgery.

Suppose a random sample $X_1, \dots, X_n$ and a parameter $\theta$. In general, an upper and a lower confidence bounds for $\theta$ with confidence level $100(1-\alpha)\%$ are, respectively, $u(X_1, \dots, X_n)$ and $l(X_1, \dots, X_n)$ such that

$$P(\theta < u(X_1, \dots, X_n)) = 1-\alpha,$$

$$P(l(X_1, \dots, X_n) < \theta) = 1-\alpha.$$

One-sided confidence bounds are often obtained by identifying a pivotal quantity and manipulating an appropriate inequality statement to isolate $\theta$.

Example. Suppose that $X_1, \dots, X_n$ is a random sample from a normal distribution for which $\sigma$ is known. Since the cumulative area under the standard normal curve to the left of 1.645 is .95, we have

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.645\right) = .95$$

$$\Rightarrow P\left(\underbrace{\bar{X} - 1.645\frac{\sigma}{\sqrt{n}}}_{\text{lower confidence bound for } \mu} < \mu\right) = .95$$

— • —

You made it. Well done!