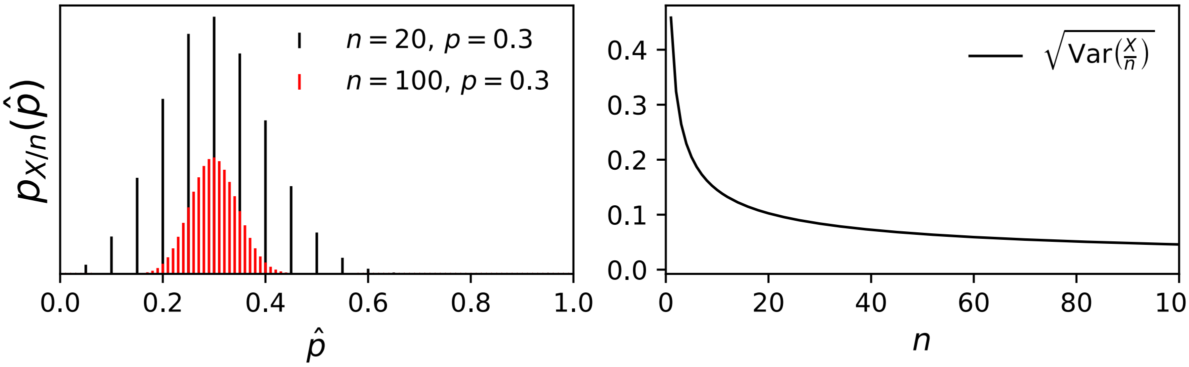
Chapter 8. Estimation

We use probability theory to describe how observations (or the results of an experiment) arise from an underlying random process. In statistics we wish to make inferences about the underlying process which generated a set of observed data.

For example, suppose we have an unfair coin that lands heads with probability p and tails with probability $1 - p$, *but we do not know the value of* p. We flip the coin some number of times and count the number of heads we observe. The statistical question is: how can we learn about the underlying (and unknown) parameter p from the observed number of heads? And equally important, how can we precisely talk about our "uncertainty" in the value of p?

Let's call x the number of heads we observe after n tosses. From our physical understanding of the experiment we postulate that x is a realization of the RV $X \sim \text{Binomial}(n, p)$. A reasonable guess for p might be $\hat{p} = x/n$, the number of heads divided by the number of tosses. This is an *estimator* of the parameter p. The key insight is that the estimator is itself a random variable, X/n, and so we can characterize it using probability theory. For instance, we can derive the probability distribution of X/n, or compute its mean and standard deviation.



To abstract the situation, let's say we measure a sample of data $\underline{x} = (x_1, \ldots, x_n)$ and we consider these observed values as realizations of corresponding random variables $\underline{X} = (X_1, \ldots, X_n)$. The x_i's might constitute a sample from an underlying population, or they might represent repeated measurements of some quantity.

If the underlying population, from which the sample has been drawn, is such that the distribution of a single random draw X has probability distribution $\text{P}_{X|\theta}(\cdot|\theta)$, where θ is a generic parameter or vector of parameters, we typically then assume that our n data point random variables $\underline{X}$ are i.i.d. $\text{P}_{X|\theta}(\cdot|\theta)$ (for instance if we are sampling with replacement from some population). This leads us to the full probability distribution for observing the data $\underline{x}$ given the parameter(s) θ.

For example, suppose some process generates the data $x_1, x_2, \ldots, x_n$ where each x_i is a realization of some random variable X_i. The RVs X_i may be distributed according to some known type of probability distribution, e.g. Poisson(λ), Exp(λ), U(a, b), N(μ, σ^2) but with *unknown parameters.*

In this case, the statistical problem is to use the observed data $\underline{x}$ to make inferences about the underlying parameters, which very often have physical significance and are what we are truly interested in.

8.1 Estimators

Consider a sequence of random variables $\underline{X} = (X_1, \ldots, X_n)$ corresponding to n i.i.d. data samples each distributed according to the distribution P_X. Let $\underline{x} = (x_1, \ldots, x_n)$ be the corresponding realized values we observe for these random variables.

Definition 8.1.1. *A* **statistic** *is a function $T = T(X_1, \ldots, X_n) = T(\underline{X})$, and is itself a random variable.*

For example, $\overline{X} = \sum_{i=1}^{n} X_i / n$ is a statistic we call the sample mean. The corresponding realised value of a statistic, e.g. $\overline{x}$, is written $t = t(\underline{x})$.

If a statistic $T(\underline{X})$ is to be used to approximate parameters of the distribution $\mathrm{P}_{X|\theta}(\cdot|\theta)$, we say T is an **estimator** for those parameters; we call the actual realised value of the estimator for a particular data sample, $t(\underline{x})$, an estimate.

8.1.1 Point Estimates

A **point estimate** is a statistic estimating a single parameter or characteristic of a distribution.

For a running example which we will return to, consider a sample of data $(x_1, \ldots, x_n)$ from an Exponential(λ) distribution with unknown λ; we might construct a point estimate for either λ itself, or perhaps for the mean of the distribution $\left(= \lambda^{-1}\right)$, or the variance $\left(= \lambda^{-2}\right)$.

Concentrating on the mean of the distribution in this example, we could propose simply the first data point we observed, X_1 as our point estimator; or we might use the sample mean, $\overline{X}$; or, if the data had been given to us already ordered we might (lazily) suggest the median, $X_{(\{n+1\}/2)}$.

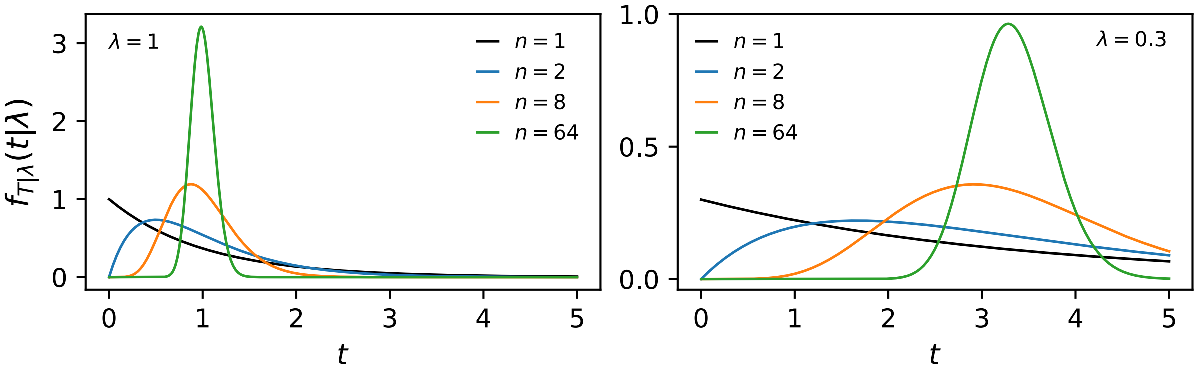
Suppose for a moment we actually knew the parameter values θ of our population distribution $\mathrm{P}_{X|\theta}(\cdot|\theta)$ (so we know λ in our exponential example).

Then since our sampled data are considered to be i.i.d. realisations from this distribution (so each $X_i \sim \mathrm{Exp}(\lambda)$), it follows that any statistic $T = T(X_1, \ldots, X_n)$ is also a random variable with some distribution which also only depends on these parameters.

If we are able to (approximately) identify this sampling distribution of our statistic, call it $\mathrm{P}_{T|\theta}$, we can then find the expectation, variance, etc of our statistic.

Sometimes $P_{T|\theta}$, will have a convenient closed-form expression which we can derive, but in other cases it will not.

In those other cases, provided that our sample size n is large, we can often use the central limit theorem (CLT) to give us an approximate distribution for $P_{T|\theta}$. If T is the sample mean, for instance, whatever the form of $P_{X|\theta}$, we know that approximately $\overline{X} \sim \mathrm{N}(\mathrm{E}[X], \mathrm{Var}[X]/n)$.



For our $X_i \sim \mathrm{Exp}(\lambda)$ example, it can be shown that our statistic $T = \overline{X}$ is a continuous random variable with pdf

$$f_{T|\lambda}(t|\lambda) = \frac{(n\lambda)^n t^{n-1} e^{-n\lambda t}}{(n-1)!}, \qquad t > 0.$$

We recognize this as the pdf of a Gamma$(n, n\lambda)$ random variable, $T \sim \mathrm{Gamma}(n, n\lambda)$.

So using the fact that Gamma(α, β) has expectation $\frac{\alpha}{\beta}$, we have

$$\mathrm{E}(\overline{X}) = \mathrm{E}_{T|\lambda}(T|\lambda) = \frac{n}{n\lambda} = \frac{1}{\lambda},$$

the same as the mean of our population distribution, $\mathrm{E}(X)$.

8.1.2 Bias, Efficiency, and Consistency

Bias

The previous result suggests that $\overline{X}$ is, at least one respect, a good statistic for estimating the unknown mean of an exponential distribution.

Formally, we define the **bias** of an estimator T for a parameter θ,

$$\mathrm{bias}(T, \theta) = \mathrm{E}(T - \theta \,|\, \theta) = \mathrm{E}(T \,|\, \theta) - \theta.$$

If, as in the exponential distribution example above (where $\theta = \lambda^{-1}$), our estimator has zero bias we say the estimator is **unbiased**. So in our example, $\bar{x}$ gives an unbiased estimate of the mean of an exponential distribution.

In fact, this is true for any distribution; the sample mean $\overline{x}$ will always be an unbiased estimate for the population mean μ:

$$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n \mathrm{E}(X_i)}{n} = \frac{n\mu}{n} = \mu.$$

Similarly, there is an estimator for the population variance σ^2 which is unbiased, irrespective of the population distribution. This estimator is not the sample variance

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

as this has one too many degrees of freedom (the n in the denominator).

Note If we knew the population mean μ, then $\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ would be unbiased for σ^2.

However, we can instead define the **bias-corrected sample variance**,

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

which is then always an unbiased estimator of the population variance σ^2.

Warning! Because of it's usefulness as an unbiased estimate of σ^2, many statistical text books and software packages refer to $s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ as the sample variance.

Efficiency

Suppose we have two unbiased estimators for a parameter θ, which we will call $\widehat{\Theta}(\underline{X})$ and $\widehat{\Psi}(\underline{X})$. And again suppose we have the corresponding sampling distributions for these estimators, $\mathrm{P}_{\widehat{\Theta}|\theta}$ and $\mathrm{P}_{\widehat{\Psi}|\theta}$, and so can calculate their means, variances, etc.

Then we say $\widehat{\Theta}$ is **more efficient** than $\widehat{\Psi}$ if:

1. $\forall\theta$, $\mathrm{Var}_{\widehat{\Theta}|\theta}\left(\widehat{\Theta}|\theta\right) \leq \mathrm{Var}_{\widehat{\Psi}|\theta}\left(\widehat{\Psi}|\theta\right)$;
2. $\exists\theta$ s.t. $\mathrm{Var}_{\widehat{\Theta}|\theta}\left(\widehat{\Theta}|\theta\right) < \mathrm{Var}_{\widehat{\Psi}|\theta}\left(\widehat{\Psi}|\theta\right)$.

That is, the variance of $\widehat{\Theta}$ is never higher than that of $\widehat{\Psi}$, no matter what the true value of θ is; and for some value of θ, $\widehat{\Theta}$ has a strictly lower variance than $\widehat{\Psi}$.

If $\widehat{\Theta}$ is more efficient than any other possible estimator, we say $\widehat{\Theta}$ is **efficient**.

Suppose we have a population with mean μ and variance σ^2, from which we are to obtain a random sample $X_1, \ldots, X_n$. Consider two estimators for μ, $\widehat{M} = \overline{X}$, the sample mean, and $\widetilde{M} = X_1$, the first observation in the sample.

We have seen $\mathrm{E}(\overline{X}) = \mu$ always, and certainly $\mathrm{E}(X_1) = \mu$, so both estimators are unbiased. We also know $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$, and of course $\mathrm{Var}(X_1) = \sigma^2$, independent of μ. So if $n \geq 2$, $\widehat{M}$ is more efficient than $\widetilde{M}$ as an estimator of μ.

In the previous example, the worst aspect of the estimate $\widetilde{M} = X_1$ is that it does not change, let alone improve, no matter how large a sample n of data is collected. In contrast, the variance of $\widehat{M} = \overline{X}$ gets smaller and smaller as n increases.

Consistency

Technically, we say an estimator $\widehat{\Theta}$ is a **consistent** estimator for the parameter θ if $\widehat{\Theta}$ **converges in probability** to θ. That is, $\forall \epsilon > 0$, $\mathrm{P}(|\widehat{\Theta} - \theta| > \epsilon) \to 0$ as $n \to \infty$.

This is hard to demonstrate, but if $\widehat{\Theta}$ is unbiased we do have:

$$\lim_{n\to\infty} \mathrm{Var}\left(\widehat{\Theta}\right) = 0 \Rightarrow \widehat{\Theta} \text{ is consistent.}$$

So returning to our example, we see $\overline{X}$ is a consistent estimator of μ for any underlying population.

8.1.3 Maximum Likelihood Estimation

Recall the case of estimating the unknown parameter p after observing a random variable $X \sim \text{Binomial}(10, p)$.

We asked how we might propose an estimator for p. Since then, we have met different criteria for measuring the relative quality of rival estimators, but no principled manner for deriving these estimates.

There are many ways of deriving estimators, but we shall concentrate on just one — **maximum likelihood** estimation.

If the underlying population is a discrete distribution with an unknown parameter θ, then each of the samples X_i are i.i.d. with probability mass function $p_{X|\theta}(x_i)$.

Since the n data samples are independent, the joint probability of all of the data, $\underline{x} = (x_1, \ldots, x_n)$, is

$$\begin{aligned} \mathrm{P}(X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_n = x_n) &= \prod_{i=1}^{n} p_{X|\theta}(x_i) \\ &\equiv L(\theta \,|\, \underline{x}) \end{aligned}$$

The function $L(\theta \,|\, \underline{x})$ is called the likelihood function and is considered as a function of the parameter θ for a fixed sample of data $\underline{x}$. $L(\theta \,|\, \underline{x})$ is the probability of observing the data we have, $\underline{x}$, if the true parameter were θ.

If, on the other hand, the underlying population is a continuous distribution with an unknown parameter θ, then each of the samples X_i are i.i.d. with probability density function $f_{X|\theta}(x_i)$, and the likelihood function is defined by

$$L(\theta \,|\, \underline{x}) \equiv \prod_{i=1}^{n} f_{X|\theta}(x_i).$$

Clearly, for a fixed set of data, varying the population parameter θ would give different probabilities of observing these data, and hence different likelihoods.

Maximum likelihood estimation seeks to find the parameter value $\widehat{\theta}_{MLE}$ which maximises the likelihood function,

$$\widehat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, L(\theta \,|\, \underline{x}).$$

This value $\widehat{\theta}_{MLE}$ is known as the **maximum likelihood estimate (MLE)**.

For maximising the likelihood function, it often proves more convenient to consider the log-likelihood, $\ell(\theta \,|\, \underline{x}) = \log\left[L(\theta \,|\, \underline{x})\right]$. Since $\log(\cdot)$ is a monotonic increasing function, the argument θ maximising ℓ maximises L.

The log-likelihood function can be written as

$$\ell(\theta \,|\, \underline{x}) = \sum_{i=1}^{n} \log\left[p_{X|\theta}(x_i)\right] \quad \text{or} \quad \ell(\theta \,|\, \underline{x}) = \sum_{i=1}^{n} \log\left[f_{X|\theta}(x_i)\right],$$

for discrete and continuous distributions respectively.

Remember, the sum is over the n observations $x_1, \ldots, x_n$, with each observation being a realization of an independent copy of the RV X.

In either the discrete or continuous case, finding the $\widehat{\theta}$ that solves $\frac{\partial}{\partial \theta}\ell(\widehat{\theta}) = 0$ yields the MLE if $\frac{\partial^2}{\partial \theta^2}\ell(\widehat{\theta}) < 0$ (i.e. $\widehat{\theta}$ is a *local maximum* of the likelihood).

Example Let's say we have $n = 100$ independent measurements of a Binomial$(10, p)$ RV X. Or in other words we consider 100 i.i.d. RVs X_i, …, X_n each distributed as $X_i \sim$ Binomial$(10, p)$.

Each of our Binomial$(10, p)$ samples X_i has pmf

$$p_X(x_i) = \binom{10}{x_i} p^{x_i}(1-p)^{10-x_i}, \qquad i = 1, 2, \ldots, 100.$$

Since the $n = 100$ data samples are assumed independent, the likelihood function for p for all of the data is

$$\begin{aligned} L(p \,|\, \underline{x}) =& L(p) = \prod_{i=1}^{n} p_X(x_i) = \prod_{i=1}^{n} \left\{ \binom{10}{x_i} p^{x_i}(1-p)^{10-x_i} \right\} \\ &= \left\{ \prod_{i=1}^{n} \binom{10}{x_i} \right\} p^{\sum_{i=1}^{n} x_i}(1-p)^{10n - \sum_{i=1}^{n} x_i}. \end{aligned}$$

So the log-likelihood is given by

$$\ell(p) = \log\left\{ \prod_{i=1}^{n} \binom{10}{x_i} \right\} + \log(p)\sum_{i=1}^{n} x_i + \log(1-p)\left(10n - \sum_{i=1}^{n} x_i\right).$$

Next, we differentiate $\ell(p)$

$$\frac{\partial}{\partial p}\ell(p) = 0 + \frac{\sum_{i=1}^{n} x_i}{p} - \frac{10n - \sum_{i=1}^{n} x_i}{1-p}.$$

Setting this derivative equal to zero, we get

$$\begin{aligned}\frac{\sum_{i=1}^n x_i}{\widehat{p}} - \frac{10n - \sum_{i=1}^n x_i}{1-\widehat{p}} = 0 \Rightarrow &(1-\widehat{p})\sum_{i=1}^n x_i = \widehat{p}\left(10n - \sum_{i=1}^n x_i\right)\\ \Rightarrow &\sum_{i=1}^n x_i = \widehat{p}\left(10n - \sum_{i=1}^n x_i + \sum_{i=1}^n x_i\right)\\ \Rightarrow &\widehat{p} = \frac{\sum_{i=1}^n x_i}{10n} = \frac{\overline{x}}{10}.\end{aligned}$$

To check this point is a maximum of ℓ, we find the second derivative

$$\frac{\partial^2}{\partial p^2}\ell(p) = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{10n - \sum_{i=1}^n x_i}{(1-p)^2} = -\frac{n\overline{x}}{p^2} - \frac{10n - n\overline{x}}{(1-p)^2} = -n\left(\frac{\overline{x}}{p^2} + \frac{10-\overline{x}}{(1-p)^2}\right)$$

(which is in fact $< 0 \; \forall p$, the likelihood is *log concave*).

Substituting $\widehat{p} = \dfrac{\overline{x}}{10}$, this gives

$$-100n\left(\frac{1}{\overline{x}} + \frac{1}{10-\overline{x}}\right) = -\frac{1000n}{(10-\overline{x})\overline{x}} \quad ,$$

which is clearly < 0. ■

Example If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

$$f_X(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

For an i.i.d. sample $\underline{x} = (x_1, \ldots, x_n)$, the likelihood function for (μ, σ^2) for all of the data is

$$L(\mu, \sigma^2) = \prod_{i=1}^n f_X(x_i) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right\}.$$

The log likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}.$$

For the MLE for μ, we can take the partial derivative wrt μ and set this equal to zero.

$$0 = \frac{\sum_{i=1}^n (x_i - \widehat{\mu})}{\sigma^2} \iff 0 = \sum_{i=1}^n (x_i - \widehat{\mu}) = \left(\sum_{i=1}^n x_i\right) - n\widehat{\mu} \iff \widehat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}.$$

To check this is a maximum, we look at the second derivative.

$$\frac{\partial^2}{\partial \mu^2}\ell(\mu, \sigma^2) = -\frac{n}{\sigma^2},$$

which is negative everywhere, so $\overline{x}$ is the MLE for μ, independently from the value of σ^2. ■

Finding the MLE

In general, we have the following procedure to find MLEs.

1. Write down the likelihood function, $L(\theta)$ where

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

   that is, the product of the n pmfs (or pdfs) viewed as a function of θ.

2. Take the natural log of the likelihood, and collect terms involving θ.

3. Find the value of θ for which log-likelihood is maximised. This is typically done by finding $\widehat{\theta}$ that solves

$$\frac{\partial}{\partial\theta}\ell(\widehat{\theta}) = \frac{\partial}{\partial\theta}\log(L(\widehat{\theta})) = 0$$

4. Check that the estimate $\widehat{\theta}$ obtained in step 3 corresponds to a maximum of the likelihood function by inspecting the second derivative of $\ell(\theta)$ wrt θ. If

$$\frac{\partial^2}{\partial\theta^2}\ell(\widehat{\theta}) < 0$$

   at $\theta = \widehat{\theta}$, then $\widehat{\theta}$ is confirmed as the MLE of θ.

Maximum likelihood estimator and CLT significance

Replace the observed values in the MLE estimate with the corresponding RVs to get an estimator: $\underbrace{\widehat{\theta}(x_1, \ldots, x_n)}_{\text{estimate}} \leftrightsquigarrow \underbrace{\widehat{\theta}(X_1, \ldots, X_n)}_{\text{estimator}}$.

We have already seen that $\overline{X}$ is always an unbiased estimator for the population mean μ. And it turns out using the CLT, that for large n, $\overline{X}$ is approximately the MLE for the population mean μ, irrespective of the distribution of X.

So how good an estimator of θ is the MLE?

- - The MLE is not necessarily unbiased.
- + For large n, the MLE is approximately normally distributed with mean θ (i.e. when the MLE is considered as a random variable, its *sampling distribution* is approximately normal with a mean equal to the *true* value of θ).
- + The MLE is consistent.
- + The MLE is always asymptotically efficient, and if an efficient estimator exists, it is the MLE.
- + Because it is derived from the likelihood of the data, it is well-principled. This is the "likelihood principle", which asserts that all the information about a parameter from a set of data is held in the likelihood.

8.2 Confidence Intervals

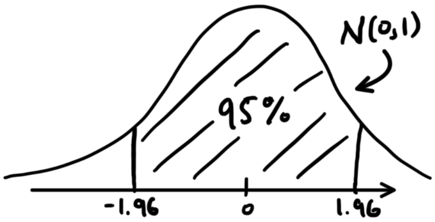
In most circumstances, it is not sufficient to report simply a point estimate $\widehat{\theta}$ for an unknown parameter θ of interest. We almost always want to also quantify our degree of uncertainty in this estimate.

For example, consider the estimator $\overline{X}$ for the population mean μ. We know by the CLT that, for any underlying distribution (mean μ, variance σ^2) for our sample, when n is large the sample mean $\overline{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. That is, $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Thus, if we knew the true population variance σ^2, we can use our standard normal tables to say that, for large n, there is a 95% probability that the random variable $\overline{X}$ will be within $1.96\frac{\sigma}{\sqrt{n}}$ of μ. That's because $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ and there is a 95% that a standard normal RV Z is in the range $-1.96 \leq Z \leq 1.96$. *And this is true no matter what the true value of μ is.* Therefore, we say that the interval

$$\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right].$$

is a 95% confidence interval for μ.

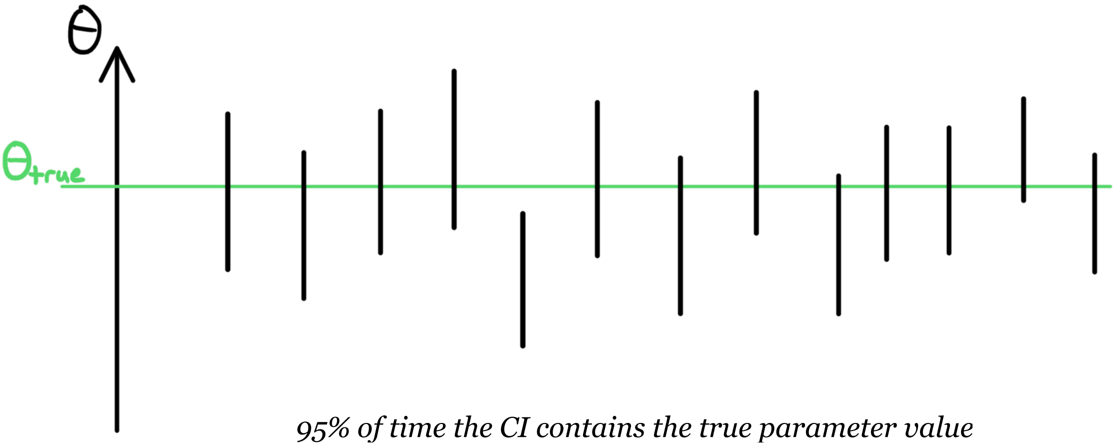


More generally, for any desired *coverage* probability level $1-\alpha$ we can define the $100(1-\alpha)\%$ **confidence interval** for μ by

$$\left[\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \quad \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

where z_α is the α-quantile of the standard normal (so before we used $\alpha = 0.05$ and hence $z_{0.975}$, which is about 1.96, to obtain our 95% C.I.).

In general, a $(1-\alpha)$ confidence interval for θ is itself a random variable, i.e. the bounds of the interval are each random variables. The key property is that there a $1-\alpha$ probability that the interval will contain the true θ, *no matter what the true value of θ is.*



95% of time the CI contains the true parameter value

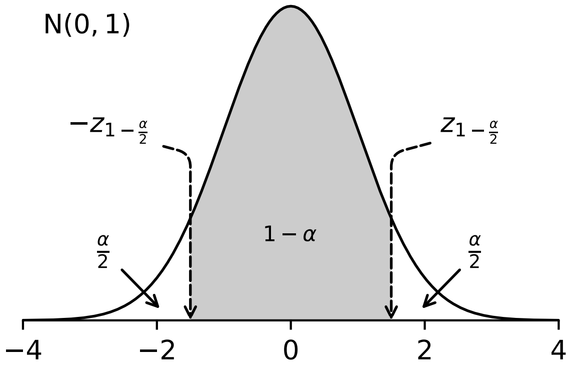


Figure 8.1: Illustration of the quantile $z_{1-\frac{\alpha}{2}}$.

Example A corporation conducts a survey to investigate the proportion of employees who thought the board was doing a good job. 1000 employees, randomly selected, were asked, and 732 said they did. Find a 99% confidence interval for the value of the proportion in the population who thought the board was doing a good job.

We can model each observation as $X_i \sim \text{Bernoulli}(p)$ for some unknown p, and we want to find a C.I. for p, which is also the mean of X_i.

We have our estimate $\widehat{p} = \overline{x} = 0.732$ for which we can use the CLT. Since the variance of $\text{Bernoulli}(p)$ is $p(1-p)$, we can use $\overline{x}(1-\overline{x}) = 0.196$ as an approximate variance.

So an approximate 99% C.I. is

$$\left[0.732 - 2.576 \times \sqrt{\frac{0.196}{1000}},\ \ 0.732 + 2.576 \times \sqrt{\frac{0.196}{1000}}\right]$$

■

Conceptually, an approximate way to go about estimating the uncertainty in an estimator is to suppose we had knowledge of the true value of our unknown parameter θ, or at least had access to the (approximate) sampling distribution of our statistic, $\mathrm{P}_{T|\theta}$. Then the variance of this distribution would give a measure of the uncertainty in our estimate. The problem is we don't know this variance without knowing θ.

A rough solution is to plug in our estimated value of θ, $\widehat{\theta}$, into $\mathrm{P}_{T|\theta}$ and hence use the (maybe further) approximated sampling distribution, $\mathrm{P}_{T|\widehat{\theta}}$. The logic, admittedly fuzzy and circular (but we will sharpen it shortly), is that the estimator T is "usually close to" the true value of θ. Therefore, we can get approximate properties of our estimator by tentatively setting the parameters of the distribution to their estimated values. For instance, we can set $\theta = \widehat{\theta}$ and calculate the variance of the distribution $\mathrm{P}_{T|\widehat{\theta}}$. This is what we did in the above example when we approximated the variance of our estimator by $\overline{x}(1-\overline{x})$.

8.2.1 Normal Distribution with Known Variance

The confidence interval given in the Bernoulli(p) example was only an approximate interval, relying on the Central Limit Theorem, and also assuming the population variance σ^2 was known.

However, if we in fact know that $X_1, \ldots, X_n$ are an i.i.d. sample from $\mathrm{N}(\mu, \sigma^2)$, then we have exactly

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{or} \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

In which case,

$$\left[\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \ \ \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

is an *exact* confidence interval for μ, assuming we know σ^2. This is because, *no matter what μ actually is,* we have that

$$1 - \alpha = \mathrm{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right) = \mathrm{P}\left(\overline{X} - \frac{\sigma}{\sqrt{n}}z_{1-\frac{\alpha}{2}} \leq \mu \ \text{ and } \ \overline{X} + \frac{\sigma}{\sqrt{n}}z_{1-\frac{\alpha}{2}} \geq \mu\right)$$

8.2.2 Normal Distribution with Unknown Variance

In any applied example where we are aiming to fit a normal distribution model to real data, it will usually be the case that both μ and σ^2 are unknown.

However, if again we have $X_1, \ldots, X_n$ as an i.i.d. sample from $\mathrm{N}(\mu, \sigma^2)$ but with σ^2 now unknown, then we have exactly

$$\frac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}} \sim \text{Student}(n-1)$$

where $S_{n-1} = \sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}}$ is the bias-corrected sample standard deviation, and Student(ν) is the Student's t-distribution with ν degrees of freedom.

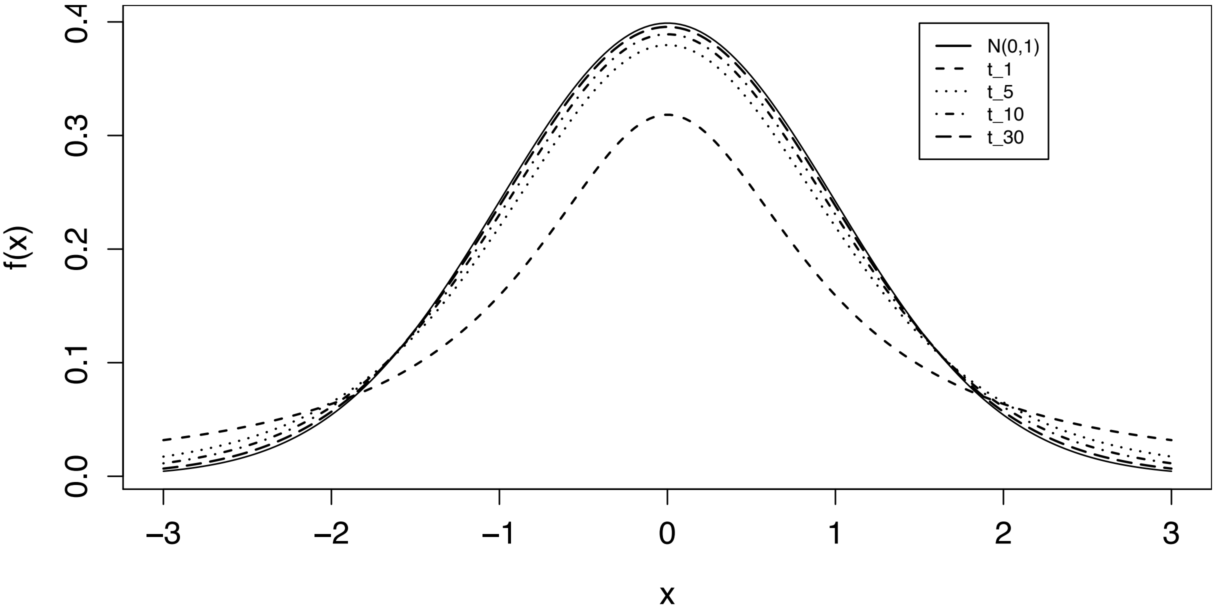
Then it follows that an exact $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\overline{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s_{n-1}}{\sqrt{n}}, \ \ \overline{x} + t_{n-1,1-\frac{\alpha}{2}}\frac{s_{n-1}}{\sqrt{n}}\right]$$

where $t_{\nu,\alpha}$ is the α-quantile of Student(ν).

ν	α				ν	α			
	0.95	0.975	0.99	0.995		0.95	0.975	0.99	0.995
1	6.31	12.71	31.82	63.66	9	1.83	2.26	2.82	3.25
2	2.92	4.30	6.96	9.92	10	1.81	2.23	2.76	3.17
3	2.35	3.18	4.54	5.84	12	1.78	2.18	2.68	3.05
4	2.13	2.78	3.75	4.60	15	1.75	2.13	2.60	2.95
5	2.02	2.57	3.36	4.03	20	1.72	2.09	2.53	2.85
6	1.94	2.45	3.14	3.71	25	1.71	2.06	2.48	2.78
7	1.89	2.36	3.00	3.50	40	1.68	2.02	2.42	2.70
8	1.86	2.31	2.90	3.36	∞	1.645	1.96	2.326	2.576

Table 8.1: Quantiles $t_{\nu,\alpha}$ of the Student t distribution with ν degrees of freedom. I.e. if the cdf is written F_ν then $F_\nu(t_{\nu,\alpha}) = \alpha$.



Notes

- Student(ν) is heavier tailed than N$(0,1)$ for any number of degrees of freedom ν.
- Hence the t-distribution CI will always be wider than the Normal distribution CI. So if we know σ^2, we should use it.
- $\lim_{\nu\to\infty}$ Student(ν) = N$(0,1)$.
- For $\nu > 40$ the difference between Student(ν) and N$(0,1)$ is so insignificant that the t distribution is not tabulated beyond this many degrees of freedom, and so there we can instead revert to N$(0,1)$ tables.

Example A random sample of 100 observations from a normally distributed population has sample mean 83.2 and bias-corrected sample standard deviation of 6.4. $\sqrt{S_{n-1}^2}$

$\bar{x}$

1. Find a 95% confidence interval for μ.
2. Give an interpretation for this interval.

Solution:

1. An exact 95% confidence interval would given by $\overline{x} \pm t_{99,0.975}\frac{s_{n-1}}{\sqrt{100}}$.

   Since $n = 100$ is large, we can approximate this by

   $$\overline{x} \pm z_{0.975}\frac{s_{n-1}}{\sqrt{100}} = 83.2 \pm 1.96 \times \frac{6.4}{10} = [81.95, 84.45].$$

2. With 95% confidence, we can say that the population mean lies between 81.95 and 84.45. Specifically, there is only a 5% chance that we got a sample of data that led to a confidence interval which did not contain the true mean.

■

8.2.3 Another way to view the confidence interval: Neyman construction

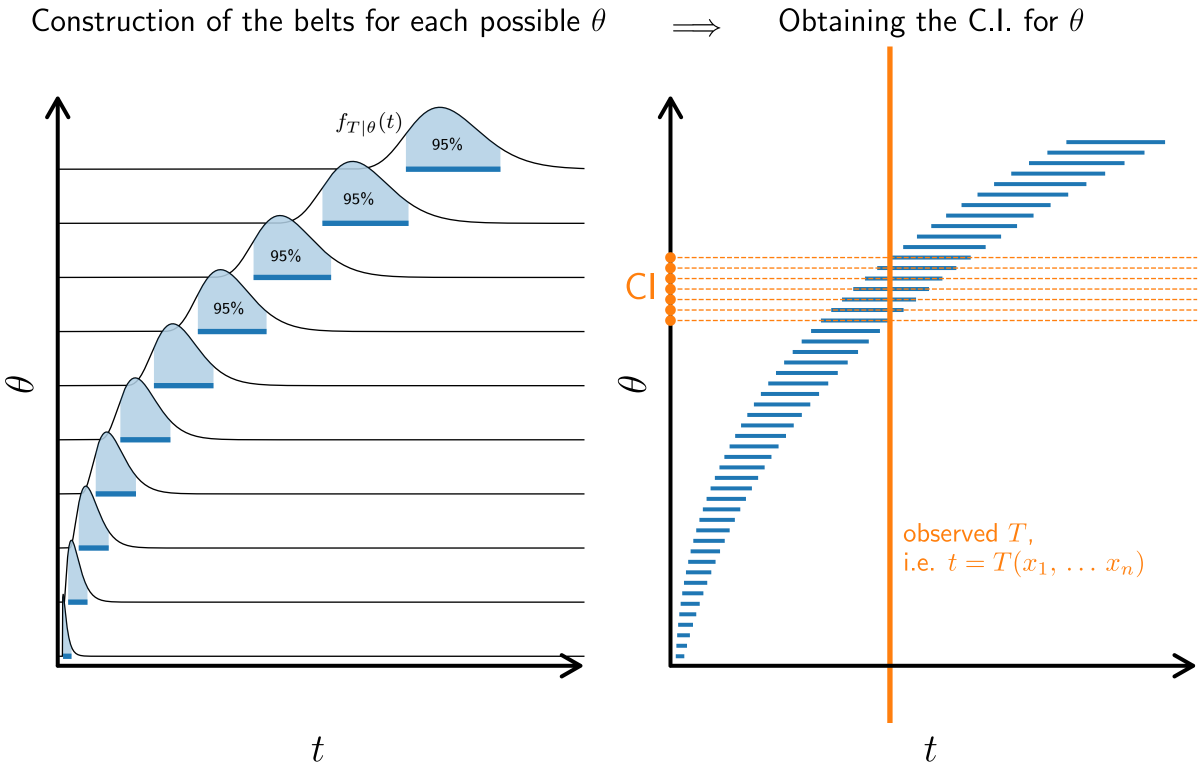
The principled definition of a **confidence interval** for the unknown parameter θ is that the confidence interval as a function of the observed data (i.e. its lower and upper boundaries are each functions of the observed data). Therefore, the confidence interval is itself a random variable — if we obtain a different realization of the X_i's we will end up constructing a different confidence interval.

The key probabilistic question we can ask is, "What is the probability that the confidence interval will contain the true value of the parameter θ?". The defining property of the confidence interval is that, *no matter what the true value of θ actually is*, the confidence interval will contain θ with a given probability called the *coverage*. For instance, there is only a 1% that a 99% confidence interval will fail to contain the true value of θ, no matter what the true value of θ actually is.

We can visualize this with the **Neyman construction**, which illustrates the construction of confidence intervals with a set of “belts”.

1. For each potential true value of θ find a range within which there is, say, a 95% of finding the estimator T. These are the belts.
2. Observe the data $x_1, \ldots x_n$ and draw a vertical line through the observed value of T, $T(x_1, \ldots, x_n)$.
3. The confidence interval is all the values of θ whose belt contains the observed T.

⇒ This procedure guarantees that no matter what the true value of θ is, there is a 95% that the resulting confidence interval contains θ (because there’s a 95% chance that T lies in the belt of the true θ).


Construction of the belts for each possible θ
$\Longrightarrow$
Obtaining the C.I. for θ
$f_{T|\theta}(t)$
95%
95%
95%
95%
θ
t
CI
θ
observed T,
i.e. $t = T(x_1, \ldots x_n)$
t