Chapter 9. Estimation

In statistics we typically analyse a set of data by considering it as a random sample from a larger, underlying population about which we wish to make inference.

1. Chapter 3 on numerical summaries considered various summary sample statistics for describing a particular sample of data. We defined quantities such as the sample mean $\overline{x}$, and sample variance s^2.

2. Chapters 6 and 7 on random variables, on the other hand, were concerned with characterising the underlying population. We defined corresponding population parameters such as the population mean $\mathrm{E}(X)$, and population variance $\mathrm{Var}(X)$.

We noticed a duality between the two sets of definitions of statistics and parameters.

In particular, we saw that they were equivalent in the extreme circumstance that our sample exactly represented the entire population. Away from this extreme circumstance, the sample statistics can be seen to give approximate values for the corresponding population parameters. We can use them as **estimates**.

For convenient modelling of populations, we met several simple parameterised probability distributions e.g. Poisson(λ), Exp(λ), U(a, b), N(μ, σ^2). There, population parameters such as mean and variance are functions of the distribution parameters. So more generally, we may wish to use the data, or just their sample statistics, to estimate distribution parameters.

For a sample of data $\underline{x} = (x_1, \ldots, x_n)$, we can consider these observed values as realisations of corresponding random variables $\underline{X} = (X_1, \ldots, X_n)$.

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 $\mathrm{P}_{X|\theta}(\cdot|\theta)$, 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. $\mathrm{P}_{X|\theta}(\cdot|\theta)$.

Suppose I was interested in the ages of people on this course. Then I could ask every person in this population their age, and thus calculate the population mean μ, variance σ^2, etc.

More realistically I might just, randomly select people, say 20, in the class and ask them their age. That is, collect a sample of 20 observations from the population, and thus calculate a sample mean $\overline{x}$, variance s^2, etc.

If this sampling had to be done *with replacement*, then we know by the CLT that the sample mean $\overline{x}$ is a random variable whose distribution is well approximated by a normal distribution with mean equal to the population mean μ and variance one twentieth of the population variance σ^2.

Note If the sampling were done *without replacement* (which is preferable), the samples would be (marginally) identically distributed but not independent.

Suppose we have a random variable X (representing a random draw from our underlying population), s.t. $X \sim \text{Binomial}(10, p)$, where the probability parameter p is unknown.

Then suppose we are able to draw a sample of size 100 from the population, that is, observe 100 independent Binomial$(10, p)$ random variables, and that we observe the data in the table below.

x	0	1	2	3	4	5	6	7	8	9	10
Freq.	2	16	35	22	21	3	1	0	0	0	0

Using these data, how might we estimate p?

9.1 Estimators

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

Definition 9.1.1. *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. 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.

9.1.1 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 $X_i \sim \text{Exp}(\lambda)$), 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 at least use the CLT to give us an approximate distribution for $P_{T|\theta}$ if T is the sample mean. Whatever the form of $P_{X|\theta}$, we know that approximately $\overline{X} \sim N(E[X], Var[X]/n)$.

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

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

This is the pdf of a Gamma$(n, n\lambda)$ random variable, a well known continuous random variable distribution, $T \sim \text{Gamma}(n, n\lambda)$.

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

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

the same as the mean of our population distribution, $E(X)$.

9.1.2 Bias, Efficiency and Consistency

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,

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

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, $\overline{x}$ 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:

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

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.

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**,

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

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 (and indeed your formula sheet for the exam) refer to $s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ as the sample variance.

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

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

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

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

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 $E(\overline{X}) = \mu$ always, and certainly $E(X_1) = \mu$, so both estimators are unbiased. We also know $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$, and of course $\text{Var}(X_1) = \sigma^2$, 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.

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, $\forall\epsilon > 0, P(|\widehat{\Theta} - \theta| > \epsilon) \to 0$ as $n \to \infty$.

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

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

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

9.1.3 Maximum Likelihood Estimation

Recall the Binomial$(10, p)$ example, where p was unknown.

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

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

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

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

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,

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

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, $\ell(\theta|\underline{x}) = \log\{L(\theta|\underline{x})\}$. Since $\log(\cdot)$ is a monotonic increasing function, the argument θ maximising ℓ maximises L.

The log-likelihood function can be written as

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

for discrete and continuous distributions respectively.

In either case, finding $\widehat{\theta}$ that solves $\dfrac{\partial}{\partial\theta}\ell(\widehat{\theta}) = 0$ yields the MLE if $\dfrac{\partial^2}{\partial\theta^2}\ell(\widehat{\theta}) < 0$.

Example Continuing the Binomial question... each of our Binomial$(10, p)$ samples X_i have 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

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

which is clearly < 0. So the MLE for p is $\widehat{p} = \dfrac{\overline{x}}{10} = 0.257$.

■

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

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

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.

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

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 mass/density functions 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 in the (log) likelihood function by inspecting the second derivative of $\ell(\widehat{\theta})$ 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.

CLT significance

We have already seen that $\overline{X}$ is always an unbiased estimator for the population mean μ. And now, using the CLT, we also have that for large n that $\overline{X}$ is always 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 approximately normally distributed with mean 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.

9.2 Confidence Intervals

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

If we were again to suppose we had knowledge of the true value of our unknown parameter θ, or at least had access to the (approximate) true sampling distribution of our statistic, $\mathrm{P}_{T|\theta}$, then the variance of this distribution would give such a measure.

The solution we consider is to plug in our known estimated value of θ, $\widehat{\theta}$, into $\mathrm{P}_{T|\theta}$ and hence use the (maybe further) approximated sampling distribution, $\mathrm{P}_{T|\widehat{\theta}}$.

In particular, we know by the CLT that for any underlying distribution (mean μ, variance σ^2) for our sample, the sample mean $\overline{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. We now further approximate this by imagining $\overline{X} \sim \mathrm{N}(\overline{x}, \frac{\sigma^2}{n})$.

Then, if we knew the true population variance σ^2, we would be able to say that *had the true mean parameter μ been $\overline{x}$*, then for large n, from our standard normal tables, with 95% probability we would have observed our statistic $\overline{X}$ within the interval

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

This is known as the 95% confidence interval for μ.

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

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

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

Loose interpretation: Amongst all the possible intervals $\overline{x} \pm z_{1-\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$ *we might have observed* (that is, from samples we *might* have drawn), 95% would have contained the unknown true parameter value μ.

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.

Clearly each observation $X_i \sim \text{Bernoulli}(p)$ 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 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

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

■

9.2.1 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 $\text{N}(\mu, \sigma^2)$, then we have exactly

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

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.

9.2.2 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

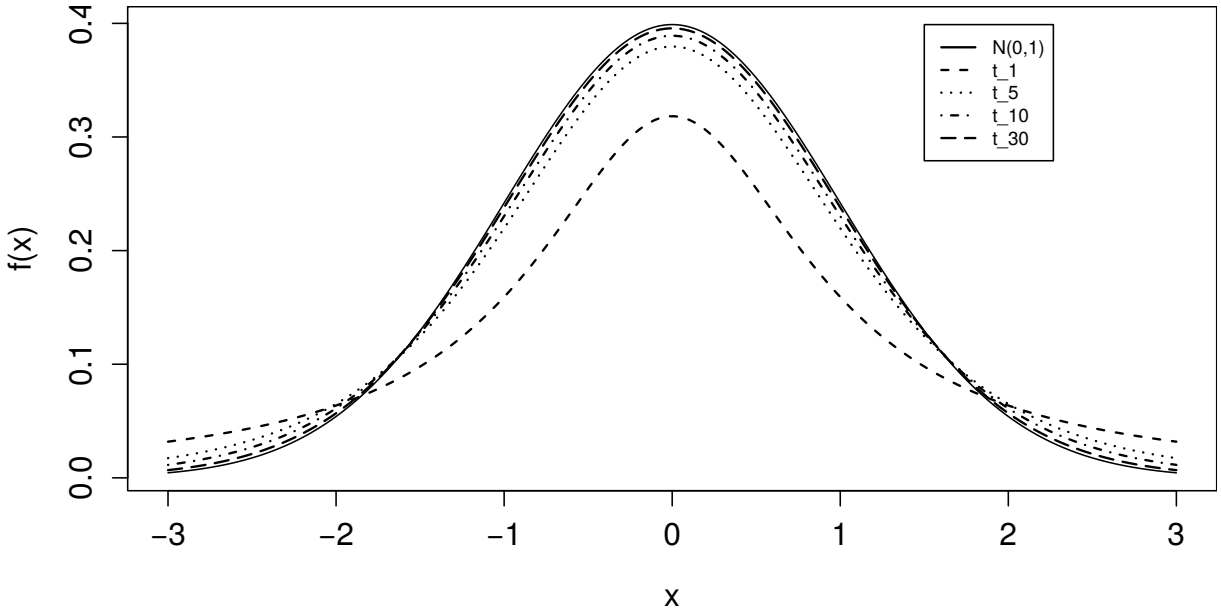
$$\frac{\overline{X} - \mu}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$$

where $s_{n-1} = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$ is the bias-corrected sample standard deviation, and t_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 t_ν.



Notes

- t_ν is heavier tailed than $\mathrm{N}(0,1)$ 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.
- $\lim_{\nu\to\infty} t_\nu \equiv \mathrm{N}(0,1)$.
- For $\nu > 40$ the difference between t_ν and $\mathrm{N}(0,1)$ is so insignificant that the t distribution is not tabulated beyond this many degrees of freedom, and so there we can instead revert to $\mathrm{N}(0,1)$ tables.

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

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.

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

■