Mathematical Statistics

MAS 713

Chapter 4

Previous lectures

1. Probability theory
2. Random variables

Questions?

This lecture

4. Interval estimation

- 4.1 Interval estimation
- 4.2 Estimation and sampling distribution
- 4.2 Confidence interval on the **mean** of a normal distribution, variance known
- 4.3 Confidence interval on the mean of a normal distribution, variance unknown
- 4.4. The Student's *t*-distribution
- 4.5. General method to derive a confidence interval
- 4.6. Central Limit Theorem
- 4.7 Confidence interval on the mean of an arbitrary population
- 4.8 Prediction intervals

Additional reading : Chapter 9 in the textbook

Interval estimation : introduction

Introduction

The purpose of most statistical inference procedures is to generalise from information contained in an observed random sample about the population from which the samples were obtained

This can be divided into two major areas :

- **estimation**, including point estimation and interval estimation
- **tests of hypotheses**

In this chapter we will present interval estimation.

Interval estimation : introduction

There are two types of estimators:

1. Point estimator: defined by a single value of a statistic.
2. Interval estimator: defined by two numbers, between which the parameter is said to lie

Statistical Inference : Introduction

Populations are often described by the distribution of their values

$\leadsto$ for instance, it is quite common practice to refer to a 'normal population', when the variable of interest is thought to be normally distributed

In statistical inference, we focus on drawing conclusions about one parameter describing the population

Statistical Inference : Introduction

Often, the parameters we are mainly interested in are

- the **mean** μ of the population
- the **variance** σ^2 (or standard deviation σ) of the population
- the **proportion** π of individuals in the population that belong to a class of interest
- the **difference in means of two sub-populations,** $\mu_1 - \mu_2$
- the **difference in two sub-populations proportions,** $\pi_1 - \pi_2$

Statistical Inference : Introduction

Obviously, those parameters are unknown (otherwise, no need to make inferences about them) $\rightsquigarrow$ the first part of the process is thus to estimate the unknown parameters

Random sampling

Before a sample of size n is selected at random from the population, the observations are modeled as random variables $X_1, X_2, \ldots, X_n$

Definition

The set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample if

1. the X_i's are independent random variables, and
2. every X_i has the same probability distribution

Random sampling

This is often abbreviated to i.i.d., for 'independent and identically distributed' $\rightsquigarrow$ it is common to talk about an i.i.d. sample

We also apply the terms 'random sample' to the set of observed values

$$x_1, x_2, \ldots, x_n$$

of the random variables, but this should not cause any confusion

Note: as usual,
the lower case distinguishes the realization of a random sample from the upper case, which represents the random variables before they are observed

Statistic, estimator and sampling distribution

Any **numerical measure** calculated **from the sample** is called a **statistic**

Denote the unknown parameter of interest θ (so this can be μ, σ^2, or any other parameter of interest to us)

The only information we have to estimate that parameter θ is the information contained in the sample

An **estimator** of θ is thus a statistic, i.e. a function of the sample

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

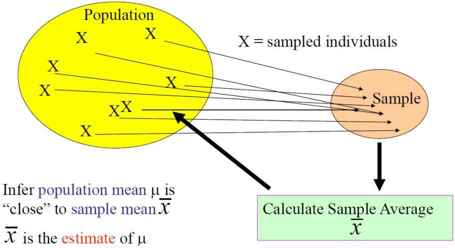
Statistic, estimator and sampling distribution

Note that an **estimator is a random variable**, as it is a function of random variables $\rightsquigarrow$ it must have a **probability distribution**

That probability distribution is called a **sampling distribution**, and it generally depends on the population distribution and the sample size

After the sample has been selected, $\hat{\Theta}$ takes on a particular value $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$, called the **estimate** of θ

An example : estimating μ in a normal population



Calculate Sample Average $\bar{x}$

Infer population mean μ is "close" to sample mean $\bar{x}$

$\bar{x}$ is the estimate of μ

Interval estimation : introduction

Non-formal example:

1. Point estimate: the temperature today is 32^o.
2. Interval estimate: the temperature today is between 28^o and 34^o with probability 0.95.

Interval estimation gives up certainty and precision regarding the value of the parameter, but gains confidence regarding the parameter being inside a pre-specified interval.

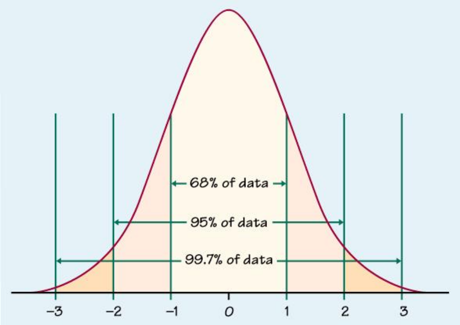
Interval estimation : introduction

When the estimator is normally distributed, we can be ‘reasonably confident’ that the true value of the parameter lies within two standard errors of the estimate

$\Longrightarrow$ 68-95-99-rule - see Chapter 3.3

The **68-95-99 rule** for normal distributions

$$\begin{aligned}\mathbb{P}(\mu-\sigma<X<\mu+\sigma)&\simeq 0.6827\\ \mathbb{P}(\mu-2\sigma<X<\mu+2\sigma)&\simeq 0.9545\\ \mathbb{P}(\mu-3\sigma<X<\mu+3\sigma)&\simeq 0.9973\end{aligned}$$



Interval estimation : introduction

Even in cases in which the estimator is not normally distributed, **Chebyshev's inequality** guarantees that the estimate of the parameter will deviate from the true value by as much as 4 standard errors at most 6 percent of the time (choose $k := 4\sigma$).

Chebyshev's inequality

Let X be a random variable with mean μ and variance σ^2. Then, for any value $k > 0$,

$$\mathbb{P}(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Interval estimation : introduction

As we see, it is often easy to determine an interval of plausible values for a parameter

$\rightsquigarrow$ such observations are the basis of **interval estimation**

$\rightsquigarrow$ instead of giving a point estimate $\hat{\theta}$, that is a single value that we know not to be equal to θ anyway, we give

an interval in which we are very confident to find the true value

Measure of centre and of variability : the sample mean and sample variance

Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Non-formal example

Basic interval estimation : example

Example

Suppose we are given a sample of size n from a population with a Normal distribution :

$$X \sim \mathcal{N}(\mu, 1)$$

The sample mean is given by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

which means that

$$\bar{X} \sim \mathcal{N}(\mu, \frac{1}{\sqrt{n}})$$

Basic interval estimation : example

Assume that the sample is:

`41.60, 41.48, 42.34, 41.95, 41.86, 42.18, 41.72, 42.26, 41.81, 42.04`

The sample mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(41.60 + 41.48 + \ldots + 42.04) = 41.924$$

and therefore

$$\bar{X} \sim \mathcal{N}(\mu, \frac{1}{\sqrt{10}})$$

Basic interval estimation : example

- 2 times the standard deviation is

$$2\sigma_{\bar{X}} = 2 \times 0.3162 = 0.6324,$$

  and we are thus 'highly confident' ($\rightsquigarrow$ 68-95-99 rule) that the true mean is within the interval

$$[41.924 \pm 0.6324] = [41.291, 42.556]$$

- if we **cannot assume** that population is **normally distributed**, then we use **4 times the standard error** $4 \times 0.3162 = 1.2648$, and we remain 'highly confident' ($\rightsquigarrow$ Chebyshev) that the true mean is within the interval

$$[41.924 \pm 1.2648] = [40.6592, 43.188]$$

$\rightsquigarrow$ the term 'highly confident' obviously needs to be quantified

Confidence intervals

Confidence intervals

The above intervals are called confidence intervals[*]

Definition

A **confidence interval** is an interval for which we can assert with a reasonable degree of certainty (or confidence) that it will contain the true value of the population parameter under consideration

[*] there exist other types of interval estimates, such as prediction intervals, see later

Confidence intervals

A confidence interval is always calculated by first selecting a confidence level, which measures the degree of reliability of the interval

$\rightsquigarrow$ a confidence interval of level $100 \times (1-\alpha)\%$ means that we are $100 \times (1-\alpha)\%$ confident that the true value of the parameter is included into the interval (obviously, α is a real number in $[0,1]$)

The most frequently used confidence levels are 90%, 95% and 99%

$\rightsquigarrow$ the higher the confidence level, the more strongly we believe that the value of the parameter being estimated lies within the interval

Confidence intervals : remarks

Information about the precision of estimation is conveyed by the length of the interval.

a short interval implies precise estimation,
a wide interval however means that there is a great deal of uncertainty concerning the parameter that we are estimating

Note that the higher the level of the interval, the wider it must be!

Confidence intervals : remarks

Fact

The $100 \times (1 - \alpha)$% refers to the percentage of all samples of the same size possibly drawn from the population which would produce an interval containing the true θ

Confidence intervals : remarks

Remark 2 : (ctd.)

$\rightsquigarrow$ if we consider taking sample after sample from the population and use each sample separately to compute $100 \times (1 - \alpha)\%$ confidence intervals, then in the long-run roughly $100 \times (1 - \alpha)\%$ of these intervals will capture θ

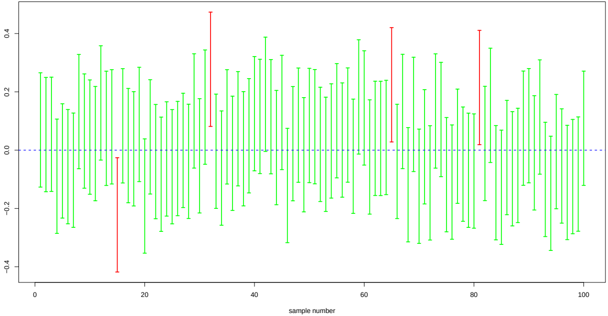
A correct probabilistic interpretation lies in the realization that a confidence interval is a random interval, because its end-points are calculated from a random sample and are therefore random variables

However, once the confidence interval has been computed, the true value either belongs to it or does not belong to it, and any probability statement is pointless

$\rightsquigarrow$ that is why we use the term "confidence level" instead of "probability"

Confidence intervals : remarks

As an illustration, we successively computed 95%-confidence intervals for μ for 100 random samples of size 100 independently drawn from a $\mathcal{N}(0, 1)$ population



$\rightsquigarrow$ 96 intervals out of 100 contain the true value $\mu = 0$

Obviously, in practice we do not know the true value of μ, and we cannot tell whether the interval we have computed is one of the 'good' 95% intervals or one of the 'bad' 5% intervals

Confidence interval on the mean of a normal distribution, variance known

Confidence interval on the mean of a normal distribution, variance known

The basic ideas for building confidence intervals are most easily understood by first considering a simple situation :

Suppose we have a normal population with

- **unknown** mean μ and
- **known** variance σ^2

Note that this is somewhat unrealistic, as typically both the mean and the variance are unknown

$\leadsto$ we will address more general situations later

Confidence interval on the mean of a normal distribution, variance known

We have thus a random sample $X_1, X_2, \ldots, X_n$, such that, for all i,

$$X_i \sim \mathcal{N}(\mu, \sigma),$$

with μ unknown and σ a known constant

Confidence interval on the mean of a normal distribution, variance known

Suppose we desire a confidence interval for μ of level $100 \times (1 - \alpha)\%$

From our random sample, this can be regarded as a ‘random interval’, say $[L, U]$, where L and U are statistics, and such that

$$\mathbb{P}(L \leq \mu \leq U) = 1 - \alpha$$

Confidence interval on the mean of a normal distribution, variance known

In that situation, we know that, for any integer $n \geq 1$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

We may standardise this normally distributed random variable :

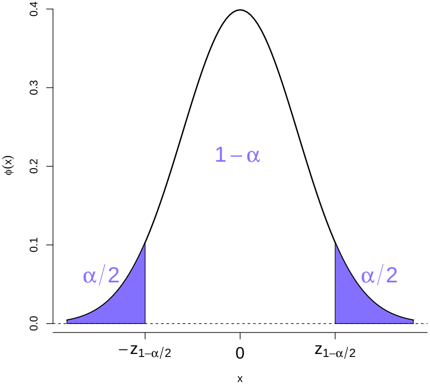
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Confidence interval on the mean of a normal distribution, variance known

In our situation, because $Z \sim \mathcal{N}(0,1)$, we may write

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

where $z_{1-\alpha/2}$ is the quantile of level $1-\alpha/2$ of the standard normal distribution,



that is,

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

Confidence interval on the mean of a normal distribution, variance known

Hence,

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$\rightsquigarrow$ we have found L and U, two statistics (random variables depending on the sample) such that

$$\mathbb{P}(L \leq \mu \leq U) = 1 - \alpha$$

$\rightsquigarrow$ L and U will yield the bounds of the confidence interval !

Confidence interval on the mean of a normal distribution, variance known

$\leadsto$ if $\bar{x}$ is the sample mean of an observed random sample of size n from a normal distribution with known variance σ^2, a confidence interval of level $100 \times (1-\alpha)\%$ for μ is given by

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

Confidence interval on the mean of a normal distribution, variance known

From Chapter 3.3: $z_{0.95} = 1.645$, $z_{0.975} = 1.96$ and $z_{0.995} = 2.575$

$\rightsquigarrow$ a confidence interval for μ of level 90% is

$$\left[\bar{x} - 1.645 \times \tfrac{\sigma}{\sqrt{n}}, \bar{x} + 1.645 \times \tfrac{\sigma}{\sqrt{n}}\right]$$

$\rightsquigarrow$ a confidence interval for μ of level 95% is

$$\left[\bar{x} - 1.96 \times \tfrac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \tfrac{\sigma}{\sqrt{n}}\right]$$

$\rightsquigarrow$ a confidence interval for μ of level 99% is

$$\left[\bar{x} - 2.575 \times \tfrac{\sigma}{\sqrt{n}}, \bar{x} + 2.575 \times \tfrac{\sigma}{\sqrt{n}}\right]$$

We see that the respective lengths of these intervals are

$$3.29\frac{\sigma}{\sqrt{n}},\ 3.92\frac{\sigma}{\sqrt{n}} \text{ and } 5.15\frac{\sigma}{\sqrt{n}}$$

Confidence interval on the mean of a normal distribution, variance known : choice of sample size

The length of a CI is a measure of the precision of the estimation
$\rightsquigarrow$ the precision is inversely related to the confidence level

However, it is desirable to obtain a confidence interval that is both

- short enough for decision-making purposes
- of adequate confidence level

$\rightsquigarrow$ one way to reduce the length of a confidence interval with prescribed confidence level is by choosing *n* large enough

Confidence interval on the mean of a normal distribution, variance known : choice of sample size

From the above, we know that in using $\bar{x}$ to estimate μ,
the error $\mathfrak{e} = |\bar{x} - \mu|$ is less than $z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ with confidence $1 - \alpha$

$\rightsquigarrow$ in other words, we can be $100 \times (1 - \alpha)\%$ confident that the error will not exceed a given amount e when the sample size is

$$\boxed{n = \left(\frac{z_{1-\alpha/2}\sigma}{\mathfrak{e}}\right)^2}$$

Confidence interval on the mean of a normal distribution, variance known : example

Example

Ten measurements of temperature are as follows :

$$\texttt{64.1}, \texttt{64.7}, \texttt{64.5}, \texttt{64.6}, \texttt{64.5}, \texttt{64.3}, \texttt{64.6}, \texttt{64.8}, \texttt{64.2}, \texttt{64.3}$$

Assume that temperature is normally distributed with $\sigma = 1$c.
a) Find a 95% CI for μ, the mean temperature

An elementary computation yields $\bar{x} = 64.46$.
With $n = 10$, $\sigma = 1$ and $\alpha = 0.05$, direct application of the previous results gives a 95% CI as follows :

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[64.46 - 1.96 \times \frac{1}{\sqrt{10}}, 64.46 + 1.96 \times \frac{1}{\sqrt{10}}\right]$$
$$= [63.84, 65.08]$$

Confidence interval on the mean of a normal distribution, variance known : example

Example (ctd.)

b) Determine how many measurements we should take to ensure that the 95% CI on the mean temperature μ has a length of at most 1,

The length of the CI in part a) is 1.24.
If we desire a higher precision, namely an confidence interval length of 1, then we need more than 10 observations

The bound on error estimation $\mathfrak{e}$ is one-half of the length of the CI, thus use expression on Slide 34 with $\mathfrak{e} = 0.5$, $\sigma = 1$ and $\alpha = 0.05$:

$$n = \left(\frac{z_{\alpha/2}\sigma}{\mathfrak{e}}\right)^2 = \left(\frac{1.96 \times 1}{0.5}\right)^2 = 15.37$$

$\rightsquigarrow$ as n must be an integer, the required sample size is 16

Confidence interval on the mean of a normal distribution, variance known : example

Example (ctd.)

b) Determine how many measurements we should take to ensure that the 95% CI on the mean temperature μ has a length of at most 1,

The length of the CI in part a) is 1.24.
If we desire a higher precision, namely an confidence interval length of 1, then we need more than 10 observations

The bound on error estimation $\mathfrak{e}$ is one-half of the length of the CI, thus use expression on Slide 34 with $\mathfrak{e} = 0.5$, $\sigma = 1$ and $\alpha = 0.05$:

$$n = \left(\frac{z_{\alpha/2}\sigma}{\mathfrak{e}} \right)^2 = \left(\frac{1.96 \times 1}{0.5} \right)^2 = 15.37$$

$\rightsquigarrow$ as n must be an integer, the required sample size is 16

Confidence interval on the mean of a normal distribution, variance known : remarks

Remark 1 : if the population is normal, the confidence interval

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad (\star)$$

is valid for all sample sizes $n \geq 1$

Confidence interval on the mean of a normal distribution, variance known : remarks

Remark 2 : this interval is not the only $100 \times (1-\alpha)\%$ confidence interval for μ. For instance, starting from $\mathbb{P}(z_{\alpha/4} \leq Z \leq z_{1-3\alpha/4}) = 1-\alpha$, another $100 \times (1-\alpha)\%$ CI could be

$$\left[\bar{x} - z_{1-3\alpha/4}\frac{\sigma}{\sqrt{n}}, \bar{x} - z_{\alpha/4}\frac{\sigma}{\sqrt{n}}\right]$$

However, interval $(\star)$ is often preferable, as it is symmetric around $\bar{x}$

Confidence interval on the mean of a normal distribution, variance known : remarks

Remark 3 : in the same spirit, we have
$\mathbb{P}(Z \leq z_{1-\alpha}) = \mathbb{P}(-z_{1-\alpha} \leq Z) = 1 - \alpha$

Hence, $]-\infty, \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}]$ and $[\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty[$ are also $100 \times (1-\alpha)\%$ CI for μ

$\rightsquigarrow$ these are called one-sided confidence intervals,
as opposed to $(\star)$ (two-sided CI) $\rightsquigarrow$ Slide 45.

Confidence interval on the mean of a normal distribution, variance unknown

Confidence interval on the mean of a normal distribution, variance unknown

Suppose now that the population variance σ^2 is not known (meaning that now bothbμ and σ^2 are unknown).

$\rightsquigarrow$ we can no longer make practical use of the core result

$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Confidence interval on the mean of a normal distribution, variance unknown

However, from the random sample $X_1, X_2, \ldots, X_n$ we have a natural estimator of the unknown σ^2 : the **sample variance**

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

which will provide an estimated sample variance $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ upon observation of a sample $x_1, x_2, \ldots, x_n$

$\leadsto$ a logical procedure is thus to replace σ with the sample standard deviation S, and to work with the random variable

$$T = \sqrt{n}\, \frac{\bar{X} - \mu}{S}$$

Confidence interval on the mean of a normal distribution, variance unknown

The fact is that, if Z was a standardised version of a normal r.v. $\bar{X}$ and was therefore normally distributed, T is now a ratio of two random variables ($\bar{X} - \mu$ and S)

$\leadsto$ T is not $\mathcal{N}(0, 1)$-distributed !

Confidence interval on the mean of a normal distribution, variance unknown

Indeed, T cannot have the same distribution as Z, as the estimation of the constant σ by a random variable S introduces some extra variability

$\rightsquigarrow$ the random variable T varies much more in value from sample to sample than does Z

It turns out that, for $n \geq 2$,

T **follows** the so-called ***t*-distribution with** $n-1$ **degrees of freedom**

The Student's *t*-distribution

The Student's *t*-distribution

The first who realized that replacing σ with an estimation effectively affected the distribution of Z was William Gosset (1876-1937), a British chemist and mathematician who, worked at the Guinness Brewery in Dublin

Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery, so that Guinness prohibited its employees from publishing any papers regardless of the contained information

$\rightsquigarrow$ Gosset used the pseudonym *Student* for his publications to avoid their detection by his employer

Thus his most famous achievement is now referred to as Student's *t*-distribution, which might otherwise have been Gosset's *t*-distribution

The Student's *t*-distribution

A random variable, say T, is said to follow the Student's *t*-distribution with ν degrees of freedom, i.e.

$$\boxed{T \sim t_\nu}$$

if its probability density function is given by

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \rightsquigarrow S_T = \mathbb{R}$$

for some integer ν

Note : the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1} e^{-x}\, dx, \qquad \text{for } y > 0$$

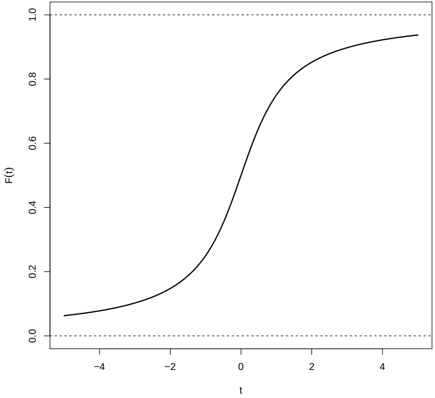
It can be shown that $\Gamma(y) = (y-1) \times \Gamma(y-1)$, so that, if y is a positive integer n,

$$\Gamma(n) = (n-1)!$$

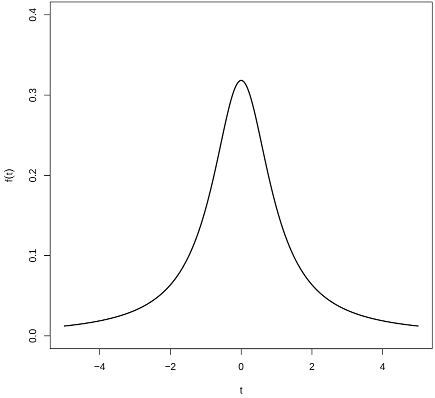
- There is no simple expression for the Student's *t*-cdf

The Student's t-distribution

Student's t distribution with 1 degree of freedom

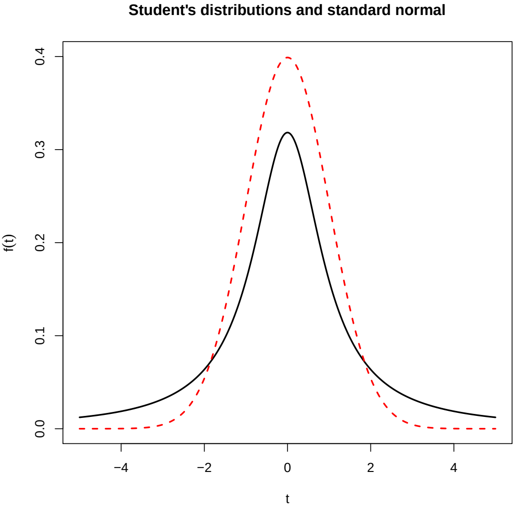


cdf $F(t)$

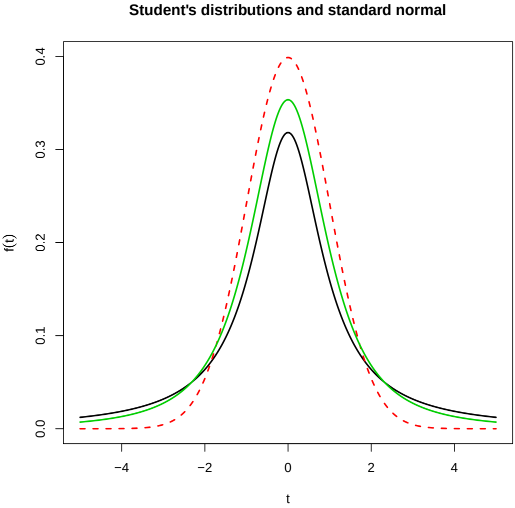


pdf $f(t) = F'(t)$

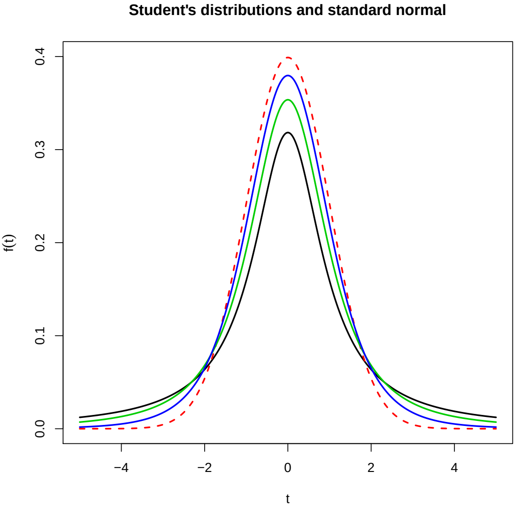
The Student's t-distribution



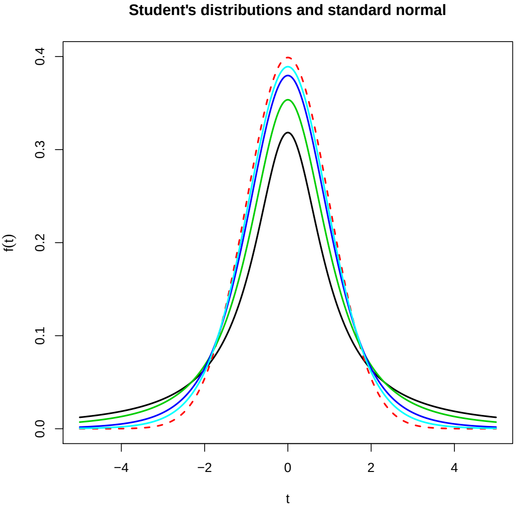
The Student's *t*-distribution



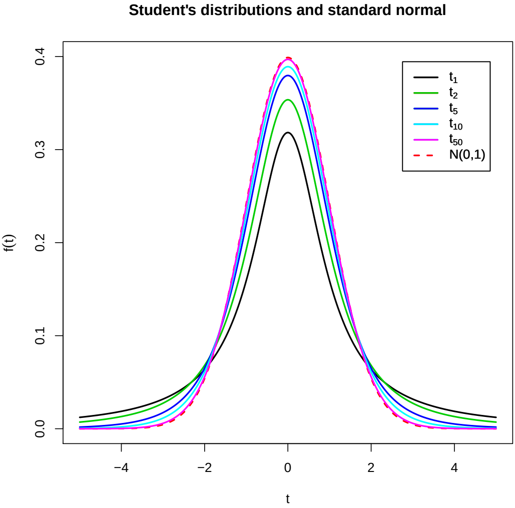
The Student's *t*-distribution



The Student's *t*-distribution



The Student's *t*-distribution



The Student's *t*-distribution

Mean and variance of the t_ν-distribution

If $T \sim t_\nu$, $$\mathbb{E}(T) = 0 \qquad \text{and} \qquad \mathbb{V}\mathrm{ar}(T) = \frac{\nu}{\nu - 2}$$

The Student's *t*-distribution

The Student's *t* distribution is similar to the standard normal distribution in that both densities are symmetric and unimodal, and the maximum value is reached at $t = 0$

However, the Student's *t* distribution has heavier tails than the normal

$\leadsto$ there is more probability to find the random variable T 'far away' from 0 than there is for Z

This is more marked for small values of ν

As the number ν of degrees of freedom increases, t_ν-distributions look more and more like the standard normal distribution

In fact, it can be shown that the Student's *t* distribution with ν degrees of freedom approaches the standard normal distribution as $\nu \to \infty$

The Student's *t*-distribution : quantiles

Similarly to what we did for the Normal distribution, we can define the quantiles of any Student's *t*-distribution

The Student's *t*-distribution : quantiles

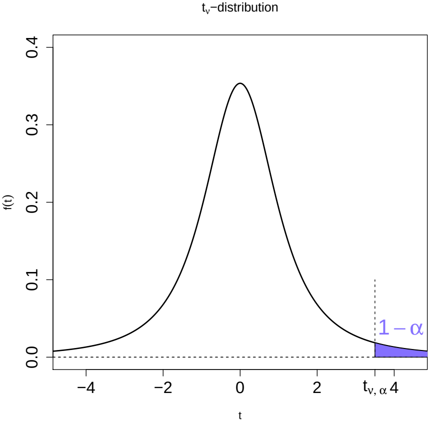
Let $t_{\nu;\alpha}$ be the value such that

$$\mathbb{P}(T > t_{\nu;\alpha}) = 1 - \alpha$$

for $T \sim t_\nu$

Like for the standard normal distribution, the symmetry of any t_ν-distribution implies that

$$\boxed{t_{\nu;1-\alpha} = -t_{\nu;\alpha}}$$



For any ν, the main quantiles of interest may be found in the *t*-distribution critical values tables

t-distribution critical values tables

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

Confidence interval on the mean of a normal distribution, variance unknown

So we have, for any $n \geq 2$,

$$T = \sqrt{n}\,\frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

Note : the number of degrees of freedom for the *t*-distribution is the number of degrees of freedom associated with the estimated variance S^2

Confidence interval on the mean of a normal distribution, variance unknown

It is now easy to find a $100 \times (1-\alpha)\%$ confidence interval for μ by proceeding essentially as we did when σ^2 was known

We may write

$$\mathbb{P}\left(-t_{n-1;1-\alpha/2} \leq \sqrt{n}\,\frac{\bar{X}-\mu}{S} \leq t_{n-1;1-\alpha/2}\right) = 1-\alpha$$

or

$$\mathbb{P}\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Confidence interval on the mean of a normal distribution, variance unknown

$\leadsto$ if $\bar{x}$ and s are the sample mean and sample standard deviation of an observed random sample of size n from a normal distribution, a confidence interval of level $100 \times (1-\alpha)\%$ for μ is given by

$$\boxed{\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

This confidence interval is sometimes called *t*-confidence interval, as opposed to ($\star$) (*z*-confidence interval)

Confidence interval on the mean of a normal distribution, variance unknown

Because t_{n-1} has heavier tails than $\mathcal{N}(0,1)$, $t_{n-1;1-\alpha/2} > z_{1-\alpha/2}$

$\rightsquigarrow$ this renders the extra variability introduced by the estimation of σ (less accuracy)

Note : One can also define one-sided $100 \times (1-\alpha)\%$ *t*-confidence intervals

$$]-\infty, \bar{x} + t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}] \text{ and } [\bar{x} - t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}, +\infty[$$

Confidence interval on the mean of a normal distribution, variance unknown : example

Example

A sample with 22 measurements of the temperature is as follows :

```
7.6, 8.1, 11.7, 14.3, 14.3, 14.1, 8.3, 12.3, 15.9, 16.4,
 11.3, 12.0, 12.9, 15.0, 13.2, 14.6, 13.5, 10.4, 13.8,
                    15.6, 12.2, 11.2
```

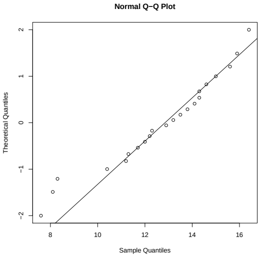
Construct a 99% confidence interval for the mean temperature

Elementary computations give

$$\bar{x} = 12.67 \qquad \text{and } s = 2.47$$

Confidence interval on the mean of a normal distribution, variance unknown : example

The quantile plot below provides good support for the assumption that the population is normally distributed



Since $n = 22$, we have $n - 1 = 21$ degrees of freedom for t. In the table, we find $t_{21;0.995} = 2.831$. The resulting CI is

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right] = \left[12.67 \pm 2.831 \times \frac{2.47}{\sqrt{22}}\right]$$
$$= [11.18, 14.16]$$

$\leadsto$ we are 99% confident that the mean temperature lies between 11.18 and 14.16 degress.

General method to derive a confidence interval

General method to derive a confidence interval

By looking at the previous two situations (CI for μ in a normal population, σ^2 known or unknown), it is now easy to give a general method for finding a confidence interval for any unknown parameter θ

General method to derive a confidence interval

Let $X_1, X_2, \ldots, X_n$ be a random sample, and suppose we can find a statistic $g(X_1, X_2, \ldots, X_n; \theta)$ with the following properties :

1. $g(X_1, X_2, \ldots, X_n; \theta)$ depends on both the sample and θ ;
2. the probability distribution of $g(X_1, X_2, \ldots, X_n; \theta)$ does not depend on θ or any other parameter

- Now, one must find constants c and u such that

$$\mathbb{P}(c \leq g(X_1, X_2, \ldots, X_n; \theta) \leq u) = 1 - \alpha$$

Because Property 2, c and u do not depend on θ

General method to derive a confidence interval

- Finally, we must manipulate the inequalities in the probability statement so that

$$\mathbb{P}(L(X_1, X_2, \dots, X_n; c, u) \leq \theta \leq U(X_1, X_2, \dots, X_n; c, u)) = 1 - \alpha$$

This gives $L(X_1, X_2, \dots, X_n; c, u)$ and $U(X_1, X_2, \dots, X_n; c, u)$ as the lower and upper limits defining a $100 \times (1 - \alpha)\%$ confidence interval for θ

The quantity $g(X_1, X_2, \dots, X_n; \theta)$ is often called a "pivotal quantity" because we pivot on it to produce the confidence interval

General method to derive a confidence interval

Example:

- $\theta = \mu$ in a normal population with known variance, then

$\Longrightarrow g(X_1, X_2, \ldots, X_n; \theta) = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$

- We hace $c = -z_{1-\alpha/2}$ and $u = z_{1-\alpha/2}$

The statistics L and U are

- $L(X_1, X_2, \ldots, X_n; c, u) = \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$,
- $U(X_1, X_2, \ldots, X_n; c, u) = \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

Central Limit Theorem

Introduction

So far we have assumed that the population distribution is normal. In that situation, we have

$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1) \qquad \text{and} \qquad T = \sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim t_{n-1}$$

where $\bar{X}$ is the sample mean and S^2 is the sample variance of a random sample of size n

These sampling distributions are the cornerstone when deriving confidence intervals for μ, and directly follow from $X_i \sim \mathcal{N}(\mu, \sigma)$

Introduction

A natural question is now :

What if the population is not normal?

$\leadsto$ surprisingly enough, the above results still hold most of the time, *at least approximately*, due to the so-called

Central Limit Theorem

The Central Limit Theorem

The Central Limit Theorem (CLT) is certainly one of the most remarkable results in probability ("*the unofficial sovereign of probability theory*"). Loosely speaking, it asserts that

the sum of a large number of independent random variables has a distribution that is approximately normal

It was first postulated by Abraham de Moivre who used the bell-shaped curve to approximate the distribution of the number of heads resulting from many tosses of a fair coin

The Central Limit Theorem

However, this received little attention until the French mathematician Pierre-Simon Laplace (1749-1827) rescued it from obscurity in his monumental work "*Théorie Analytique des Probabilités*", which was published in 1812

But it was not before 1901 that it was defined in general terms and formally proved by the Russian mathematician Aleksandr Lyapunov (1857-1918)

The Central Limit Theorem

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample (meaning **i.i.d.**) taken from a population with **finite** mean μ and **finite** variance σ^2,
and if $\bar{X}$ denotes the sample mean, then the limiting distribution of

$$\frac{1}{\sqrt{n}} \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma} = \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma}$$

as $n \to \infty$, is the **standard normal distribution**

The Central Limit Theorem

When $X_i \sim \mathcal{N}(\mu, \sigma)$, $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$ for all n

What the CLT states is that, when X_i's are not normal,

$$\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1) \text{ when } n \text{ is infinitely large}$$

$\rightsquigarrow$ the standard normal distribution provides a reasonable approximation to the distribution of $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$ when "n is large"

The Central Limit Theorem

The power of the CLT is that it holds true **for any population distribution, discrete or continuous** ! For instance,

$$X_i \sim \text{Exp}(\lambda) \quad (\mu = \frac{1}{\lambda}, \sigma = \frac{1}{\lambda}) \implies \sqrt{n}\,\frac{\bar{X} - 1/\lambda}{1/\lambda} \overset{a}{\sim} \mathcal{N}(0,1)$$

$$X_i \sim U_{[a,b]} \quad (\mu = \frac{a+b}{2}, \sigma = \frac{b-a}{\sqrt{12}}) \implies \sqrt{n}\,\frac{\bar{X} - \frac{a+b}{2}}{\frac{b-a}{\sqrt{12}}} \overset{a}{\sim} \mathcal{N}(0,1)$$

$$X_i \sim \text{Bern}(\pi) \quad (\mu = \pi, \sigma = \sqrt{\pi(1-\pi)}) \implies \sqrt{n}\,\frac{\bar{X} - \pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

Note : $\overset{a}{\sim}$ is for ‘approximately follows’

(or ‘asymptotically ($n \to \infty$) follows’)

The Central Limit Theorem

Facts :

- the larger n, the better the normal approximation
- the closer the population is to being normal, the more rapidly the distribution of $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$ approaches normality as n gets large

The Central Limit Theorem : illustration

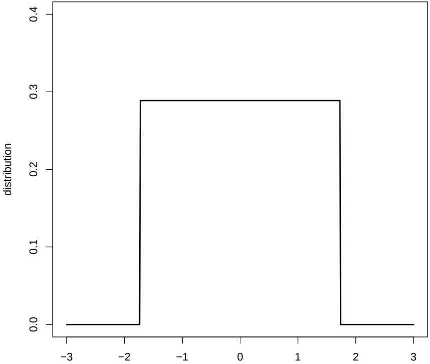
Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \mathrm{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

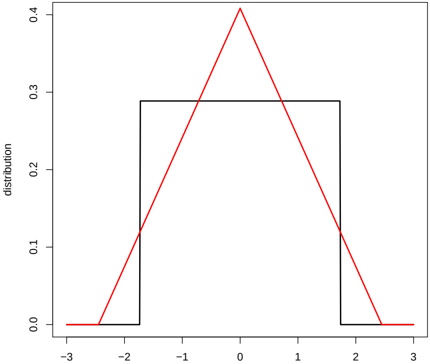
$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$



The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

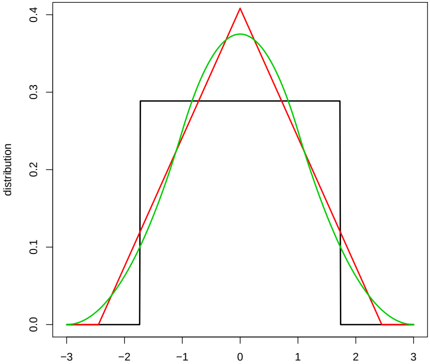
$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $\quad(\mu = 0, \sigma = 1)$ $\qquad$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $\quad(\mu = 0, \sigma = 1)$



The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

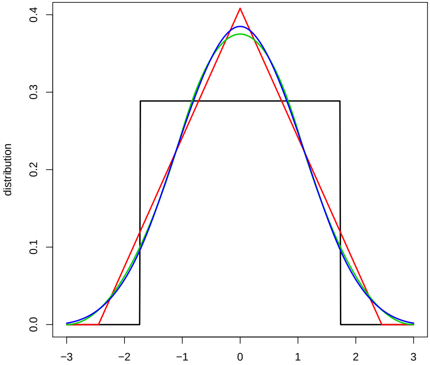
$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$



The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

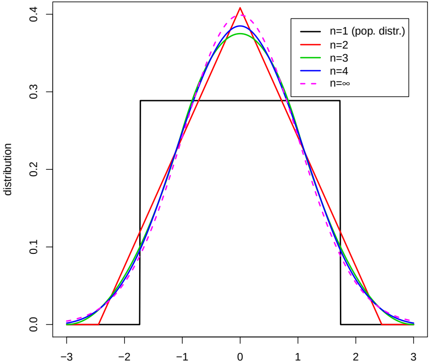
$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$



The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

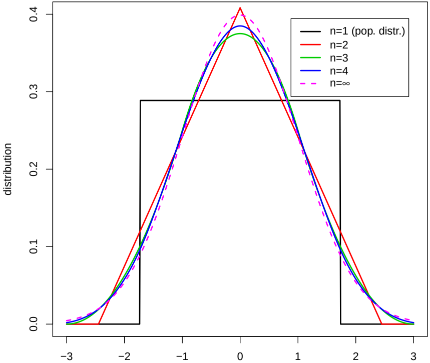
$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

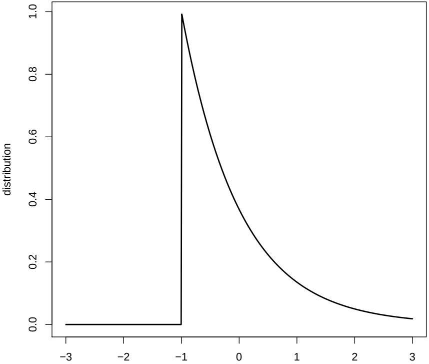


The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

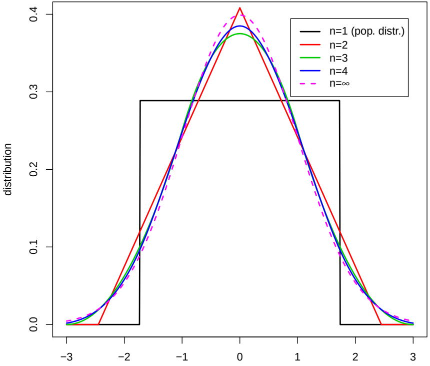


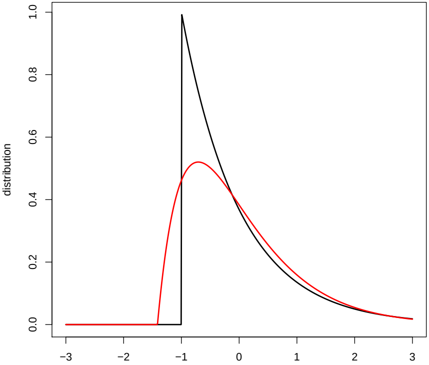


The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

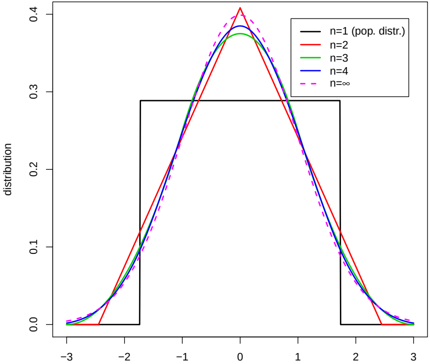


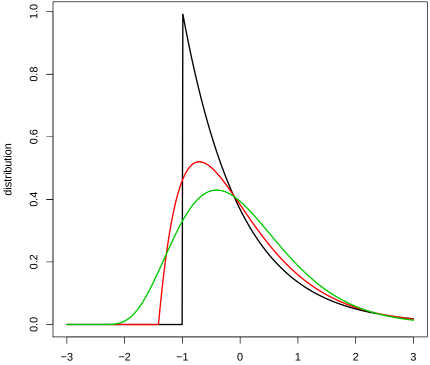


The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \mathrm{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

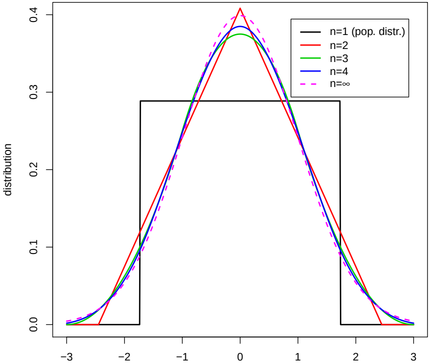


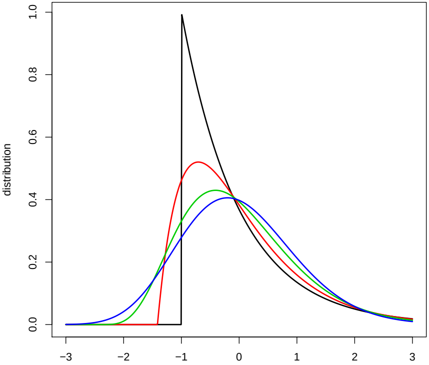


The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

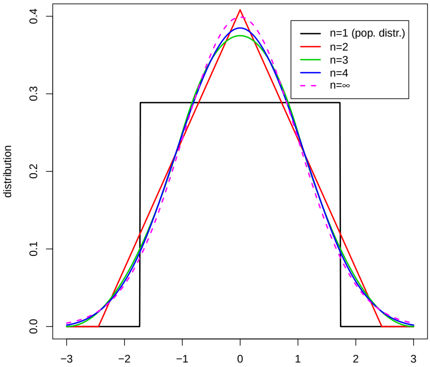


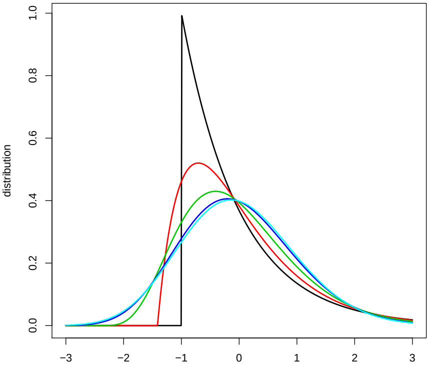


The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \mathrm{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

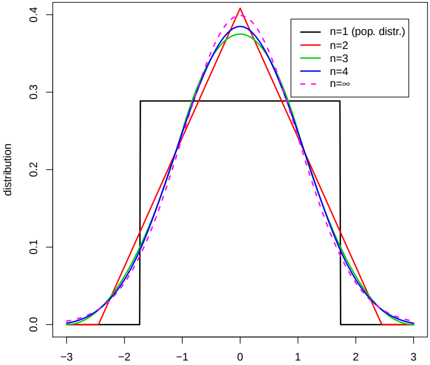


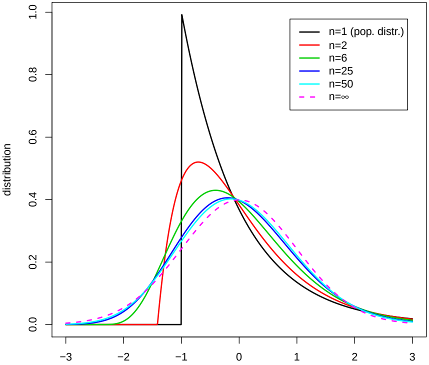


The Central Limit Theorem : illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

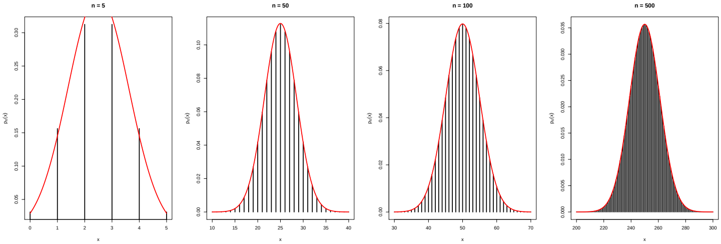




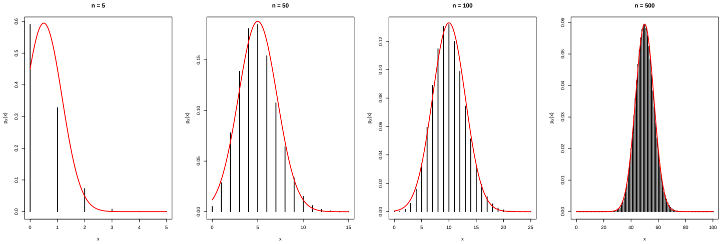
The Central Limit Theorem : illustration

Probability mass functions for $\sum_{i=1}^{n} X_i$, $X_i \sim \text{Bern}(\pi)$

$\pi = 0.5$:



$\pi = 0.1$:



The Central Limit Theorem : further illustration

Matlab example

The Central Limit Theorem : remarks

Remark 1 :

The Central Limit Theorem not only provides a simple method for computing approximate probabilities for sums or averages of independent random variables

It also helps explain why so many natural populations exhibit a bell-shaped (i.e., normal) distribution curve :

indeed, as long as the behaviour of the variable of interest is dictated by a large number of independent contributions, it should be (at least approximately) normally distributed

The Central Limit Theorem : remarks

For instance, a person's height is the result of many independent factors, both genetic and environmental. Each of these factors can increase or decrease a person's height, just as each ball in Galton's board can bounce to the right or the left. The Central Limit Theorem guarantees that the sum of these contributions has approximately a normal distribution

The Central Limit Theorem : remarks

Remark 2 :

a natural question is '**how large n needs to be**' for the normal approximation to be valid

$\rightsquigarrow$ that depends on the population distribution !

A general rule-of-thumb is that one can be confident of the normal approximation whenever the sample size n is at least 30

$$\boxed{n \geq 30}$$

Note that, in favourable cases (population distribution not severely non-normal), the normal approximation will be satisfactory for much smaller sample sizes (like $n = 5$ in the uniform case, for instance)

The rule "$n \geq 30$" just guarantees that the normal distribution provides a good approximation to the sampling distribution of $\bar{X}$ regardless of the shape of the population

Confidence interval on the mean of an arbitrary population

Confidence interval on the mean of an arbitrary distribution

The Central Limit Theorem also allows to use the procedures described in the previous slides to derive confidence intervals for μ in an arbitrary population, bearing in mind that these will be **approximate confidence intervals** (whereas they were exact in a normal population)

Confidence interval on the mean of an arbitrary distribution

Indeed, we have, if n is large enough,

$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1)$$

Hence,

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \leq z_{1-\alpha/2}\right) \simeq 1-\alpha,$$

where $z_{1-\alpha/2}$ is the quantile of level $1-\alpha/2$ of the standard normal distribution

Confidence interval on the mean of an arbitrary distribution

It follows

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \simeq 1 - \alpha,$$

so that if $\bar{x}$ is the sample mean of an observed random sample of size n from any distribution with known variance σ^2, an approximate confidence interval of level $100 \times (1-\alpha)\%$ for μ is given by

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

Confidence interval on the mean of an arbitrary distribution

Note : because this result requires "*n* large enough" to be reliable, this type of interval, based on the CLT, is often called large-sample confidence interval

One could also define large-sample one-sided confidence intervals of level $100 \times (1-\alpha)\%$: $(-\infty, \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}]$ and $[\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty)$

Confidence interval on the mean of an arbitrary distribution

What if the population standard deviation is unknown ?

Confidence interval on the mean of an arbitrary distribution

$\leadsto$ as previously, it is natural to replace σ by the sample standard deviation S and to work with

$$T = \sqrt{n}\,\frac{\bar{X} - \mu}{S}$$

One might then expect to base the derivation of the CI on $T \sim t_{n-1}$

Confidence interval on the mean of an arbitrary distribution

However, remind that, when ν is large, t_ν is very much like $\mathcal{N}(0,1)$

$\rightsquigarrow$ in **large samples**, estimating σ with S has very little effect on the distribution of T, which in turn is well approximated by the standard normal distribution :

$$\boxed{T \overset{a}{\sim} \mathcal{N}(0,1)}$$

Confidence interval on the mean of an arbitrary distribution

Consequently, if $\bar{x}$ and s are the sample mean and the sample standard deviation of an observed random sample of size n from any distribution, an approximate confidence interval of level $100 \times (1 - \alpha)\%$ for μ is given by

$$\boxed{\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

This expression holds regardless of the population distribution, as long as n is large enough

- As usual, corresponding one-sided confidence intervals could be defined : $(-\infty, \bar{x} + z_{1-\alpha}\frac{s}{\sqrt{n}}]$ and $[\bar{x} - z_{1-\alpha}\frac{s}{\sqrt{n}}, +\infty)$

Confidence interval on the mean of an arbitrary distribution : example

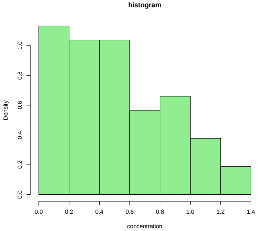
Example

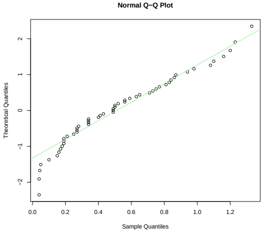
A lab reports the results of a study to investigate mercury contamination levels in fish. A sample of 53 fish was selected from some Florida lakes, and mercury concentration in the muscle tissue was measured (in ppm) :

`1.23, 0.49, 1.08, ..., 0.16, 0.27`

Find a confidence interval of level 95% on μ, the mean mercury concentration in the muscle tissue of fish

- An histogram and a quantile plot for the data





Confidence interval on the mean of an arbitrary distribution : example

$\leadsto$ both plots indicate that the distribution of mercury concentration is not normal (positively skewed)

- However, the sample is large enough ($n = 53$) to use the Central Limit Theorem and derive approximate confidence interval for μ

- Elementary computations give $\bar{x} = 0.525$ ppm and $s = 0.3486$ ppm. A large sample confidence interval is given by $\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

- With $z_{0.975} = 1.96$ and the above values, we have

$$\left[0.525 - 1.96\frac{0.3486}{\sqrt{53}}, 0.525 + 1.96\frac{0.3486}{\sqrt{53}}\right] = [0.4311, 0.6189]$$

$\leadsto$ we are $\pm$ 95% confident that the true average mercury concentration is between 0.4311 and 0.6189 ppm

Confidence intervals for the mean : summary

Confidence intervals for the mean : summary

The several situations leading to different confidence intervals for the mean can be summarised as follows :

Confidence intervals for the mean : summary

The first thing is : **is the population normal ?** (check from a histogram and/or a quantile plot, for instance)

- if yes, it is normal, is σ known ?

- if yes, use a *z*-confidence interval like $\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
- if no, use a *t*-confidence interval like $\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

- if no, it is not normal, use an approximate *z*-confidence interval like

$$\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$

provided the sample size is **large** (*large sample CI*)

Confidence intervals for the mean : summary

What if the sample size is small **and** the population is not normal ?
$\rightsquigarrow$ check on a case by case basis (beyond the scope of this course)

Prediction intervals

Prediction interval for a future observation

In some situations, we may be interested in predicting a future observation of a variable

$\leadsto$ different than estimating the mean of the variable !

$\leadsto$ instead of confidence intervals, we are after $100 \times (1-\alpha)\%$ prediction interval on a future observation

Prediction interval for a future observation

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and standard deviation σ

$\rightsquigarrow$ we wish to predict the value X_{n+1}, a single future observation

As X_{n+1} comes from the same population as $X_1, X_2, \ldots, X_n$, information contained in the sample should be used to predict X_{n+1}

$\rightsquigarrow$ the predictor of X_{n+1}, say X^*, should be a statistic

Prediction interval for a future observation

Let's define an estimator for μ as the sample mean, so we take it as predictor :

$$X^* = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Now, let's look at the error term

$$\mathfrak{e} = X_{n+1} - \bar{X}$$

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

$$X_{n+1} \sim \mathcal{N}(\mu, \sigma)$$

Prediction interval for a future observation

So we have that

$$\mathfrak{e} \sim \mathcal{N}(\mu_{\mathfrak{e}}, \sigma_{\mathfrak{e}})$$

We have that $\mu_{\mathfrak{e}} = 0$ and
the variance of the prediction error is

$$\begin{aligned}\mathbb{V}\text{ar}(\mathfrak{e}) = \mathbb{V}\text{ar}(X_{n+1} - X^*) = \mathbb{V}\text{ar}(X_{n+1} - \bar{X}) &= \mathbb{V}\text{ar}(X_{n+1}) + \mathbb{V}\text{ar}(\bar{X}) \\ &= \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{1}{n}\right)\end{aligned}$$

(because X_{n+1} is independent of $X_1, X_2, \ldots, X_n$ and so of $\bar{X}$)

$$\mathfrak{e} \sim \mathcal{N}\left(0, \sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}\right)$$

Prediction interval for a future observation

Hence,

$$Z = \frac{X_{n+1} - \bar{X}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim \mathcal{N}(0, 1)$$

Replacing the possibly unknown σ with the sample standard deviation S yields

$$T = \frac{X_{n+1} - \bar{X}}{S\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$$

Prediction interval for a future observation

Manipulating Z and T as we did previously for CI leads to the $100 \times (1-\alpha)\%$ z- and t-prediction intervals on the future observation :

$$\left[\bar{x} - z_{1-\alpha/2}\,\sigma\,\sqrt{1+\tfrac{1}{n}}, \bar{x} + z_{1-\alpha/2}\,\sigma\,\sqrt{1+\tfrac{1}{n}}\right]$$

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\,s\,\sqrt{1+\tfrac{1}{n}}, \bar{x} + t_{n-1;1-\alpha/2}\,s\,\sqrt{1+\tfrac{1}{n}}\right]$$

Prediction interval for a future observation : remarks

Remark 1 :

Prediction intervals for a single observation will always be longer than confidence intervals for μ, because there is more variability associated with one observation than with an average

Prediction interval for a future observation : remarks

Remark 2 :

As n gets larger ($n \to \infty$):

- the width of the CI for μ decreases to 0

(we are more and more accurate when estimating μ),

but

- this is not the case for a prediction interval :

the inherent variability of X_{n+1} never vanishes, even when we have observed many other observations before!

Objectives

Now you should be able to :

- Understand the basics of interval estimation □
- Explain what a confidence interval of level $100 \times (1 - \alpha)\%$ for a given parameter is □
- Construct confidence intervals on the mean of a normal distribution, advisedly using either the normal distribution or the Student's *t* distribution □
- Understand the Central Limit Theorem □
- Explain the important role of the normal distribution as a sampling distribution □
- Construct large sample confidence intervals on a mean of an arbitrary distribution □
- Explain the difference between a confidence interval and a prediction interval □
- Construct prediction intervals for a future observation in a normal population □

Put yourself to the test ! $\rightsquigarrow$ Q34 p.457, Q35 p.457, Q57 p.462, Q57 p.462,