7 Estimation

7.1 Population and Sample (P.91-92)

Suppose that we wish to study a particular health problem in Australia, for example, the average serum cholesterol level for all 40 to 60 year old males in Australia.

In this case the entire collection of all males ages 40-60 in Australia is called the *target population* or simply the *population*. In other words a population is the group we wish to study.

Definition: A *parameter* is a numerical feature of a population. Two important parameters of a population are the *mean* μ and *variance*, σ^2 or sd, σ.

Q: *How can we find the population mean serum cholesterol level of all males ages 40-60 in Australia?*

A difficult (but possible) approach is to take the serum cholesterol levels of all males ages 40 to 60 in Australia. This is called a *census*. This gives the *true* or *exact* value of this mean (parameter) cholesterol level of all males ages 40 to 60 in Australia.

As this is not always possible, statisticians can get a reasonable estimate (NOT the true or exact value) by taking a sample of males ages 40 to 60. However, this selection must be random in order to reduce extra errors such as bias.

Definition: A *random sample* from a population is a set of measurements selected such that each member has the *same chance* of being selected.

In this week, we study the problem of *parameter estimation*. The estimated parameters are then used in *statistical inference* which refers to *the process of drawing conclusions from data that are subject to random variation.*

Note: The population mean and variance are considered to be two important parameters in many statistical analysis. The following notation is used in practice:

- The population mean is denoted by μ.
- The population variance is denoted by σ^2. The population standard deviation (SD) is, therefore, σ.

7.2 Estimation the population mean, μ of a distribution (P.92-94)

Suppose that we have a random sample of size n from the population of interest. That is, we have n numerical values $x_1, x_2, \ldots, x_n$ from n random variables $X_1, X_2, \ldots, X_n$.

A natural estimator for the mean μ is

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

This $\bar{X}$ is known as an estimator for μ and we write $\hat{\mu} = \bar{X}$.

Note:

1. A sample is only a part of the population. Therefore, the sample mean $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ calculated from n observed data cannot be expected to give the exact value for the parameter μ.

2. The observed value of $\bar{x}$ depends on the particular sample to be selected and it varies from sample to sample.

3. As there is variability in the value of sample mean over different samples, the behaviour of all possible sample means is called the *sampling distribution* of the mean.

The following example illustrates the idea of sampling distribution and the variability of sample mean across samples.

Consider the outcomes of throwing a dice: $\{1, 2, 3, 4, 5, 6\}$.

Example: Write down the probability distribution of the average of two independent drawings with replacement from the above set.

There are $6^2 = 36$ possible samples drawing with replacement from the population as shown below:

x_i	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6.1	6,2	6,3	6,4	6,5	6,6

The means of the above samples are given in the following table:

$\overline{x}$	1	2	3	4	5	6
1	1.0	1.5	2.0	2.5	3.0	3.5
2	1.5	2.0	2.5	3.0	3.5	4.0
3	2.0	2.5	3.0	3.5	4.0	4.5
4	2.5	3.0	3.5	4.0	4.5	5.0
5	3.0	3.5	4.0	4.5	5.0	5.5
6	3.5	4.0	4.5	5.0	5.5	6.0

This gives the sampling distribution of $\bar{X}$ as below:

$\bar{X}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
$p(\bar{X} = \bar{x})$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This table shows that the probability distribution of $\bar{X}$ is symmetric. The next section considers the sampling distribution of $\bar{X}$.

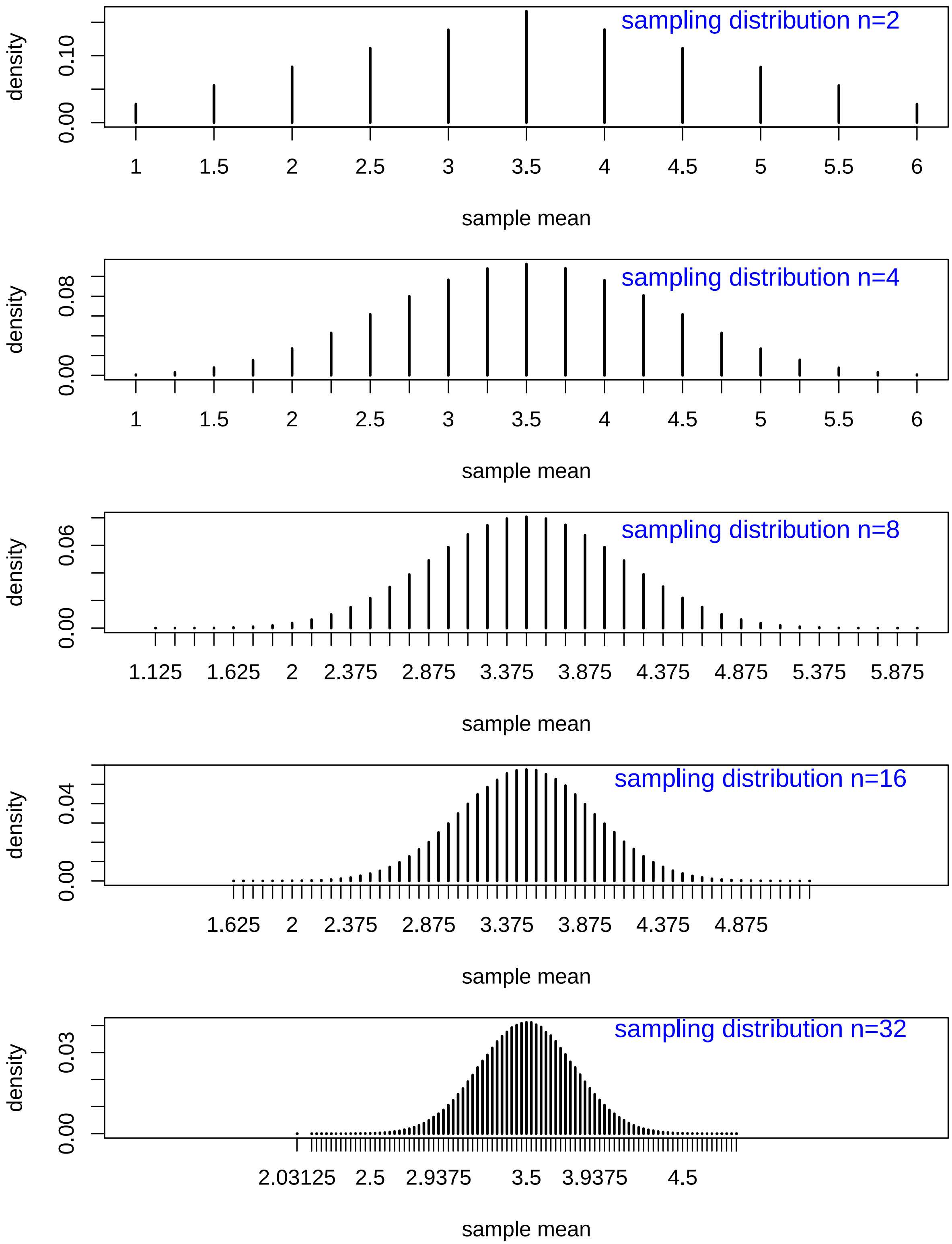
7.3 Sampling distribution of $\bar{X}$ (P.94-96)

The probability distribution of all possible values of $\bar{x}$ is called the *sampling distribution* of $\bar{X}$. It can be seen that for large n (or samples), the sampling distribution of $\bar{X}$ is approximately *symmetric* and approaches the *normal* distribution regardless of the *shape* of data distribution.

Taking our previous example of throwing a dice, the data distribution is uniform with

$$P(x = i) = \frac{1}{6}, \ i = 1, \dots, 6.$$

The sampling distribution of $\bar{X}$ for $n = 2$ independent draws is given in the table above. As the sample size n increases, the distribution of $\bar{X}$ gradually approaches normal as illustrated below:



This remarkable result as illustrated from the dice example is known as *The Central Limit Theorem* (CLT) and plays a main role in statistics.

The Central Limit Theorem (CLT): Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from *any* population with mean μ and variance σ^2. Then for large n, the distribution of $\bar{X}$ is approximately normal such that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Note:

1. As a rough guide in practice, any $n \geq 30$ is considered as large.

2. From the CLT we have,

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

   that is, Z is a *standard normal* random variable.

3. $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ is known as the *standard error* (se) of $\bar{X}$. With increasing sample size n or sample information, the se decreases as the sample distribution is more concentrated within the center. Hence the sample mean is a more precise estimate of the true mean μ.

4. The idea of CLT as illustrated in the dice example shows that even though the data distribution is uniform, the distribution of $\bar{X}$ gradually approximates normal as the sample size n increases.

Read P.96 to 98.

Example: Suppose that a random sample of size 64 was taken from a population with mean $\mu = 82$ and variance, $\sigma^2 = 144$.

(a) What is the sampling distribution of $\bar{X}$, the sample mean?
(b) Find the probability that the sample mean will lie between 80.8 and 83.2?

Solution: We have, $\mu = 82$, $\sigma^2 = 144$ and $n = 64$. Since $n = 64$ is large, the CLT can be applied. We have

(a) $\bar{X} \sim$ ____________

or $Z =$

____

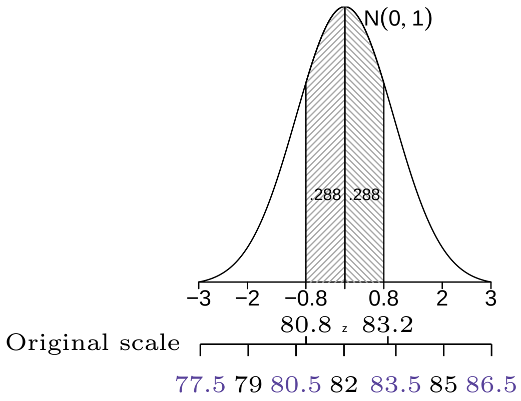
(b) $P(80.8 < \bar{X} < 83.2)$

$=$ ____________________________

$=$ ____________________

$=$ _________________

$=$ ___________________________



Exercise: Find $P(80.8 < \bar{X} < 83.2)$ when the sample size is 100.
Ans: 0.6826

Exercise: Book P.109 Q1 and Q3

7.4 An application of the CLT

Example: The weights of pears in an orchard are normally distributed with mean, $\mu = 120$g and sd, $\sigma = 32$g.

(a) If one pear is selected at random, what is the probability that its weight will be between 88g and 144g?
(b) If $\bar{X}$ denotes the average weight of a random sample of 4 pears,

(i) write down the distribution of $\bar{X}$.
(ii) give the se of $\bar{X}$.
(iii) find the probability that $\bar{X}$ will be between 88g and 144g.

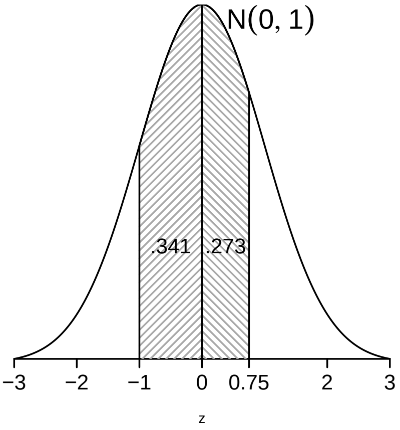
Solution: (a) Let X be the weight of a randomly selected pear. Therefore, $X \sim N(120, 32^2)$.

$P(88 < X < 144)$

$=$

$=$

$=$



(b)(i) $\bar{X} \sim N(\mu, \sigma^2/n)$ or ____________ or __________.

(ii) $\text{se}(\bar{X}) =$ ____________________.

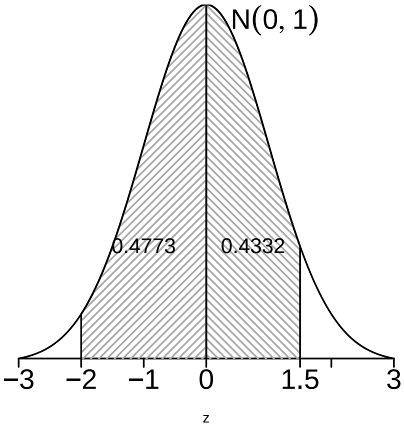
(iii) $P(88 < \bar{X} < 144)$

$=$ ______________________

$=$ ________________

$=$ ___________________________

$=$ ______



7.5 Small samples from normal population (P.99-101)

Let $X_1, X_2, ..., X_n$ be a random sample from any population (not necessarily normal) with mean μ and variance σ^2. By the CLT we know that if n is large, then

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

When n is large, the sample sd s approximates well the population sd σ. Therefore

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \to \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

However, in many practical situations, we can only draw small samples of sizes n less than 30 ($n < 30$).

What if n is small and we don't know the value of the population sd σ?

Then the corresponding sample sd, s is not a good approximation to the population sd, σ. Therefore the above approximation does not provide satisfactory results.

With the extra assumption that the sample is drawn from a normal population, one can modify the above approach using a slightly different distribution called the t distribution or *Student's* t distribution. This kind of inference based on small sample is known as the *inference for small samples.*

7.6 The t distribution

Suppose that we draw a small sample (ie $n < 30$) from a normal population with unknown sd. It can be shown that the distribution of

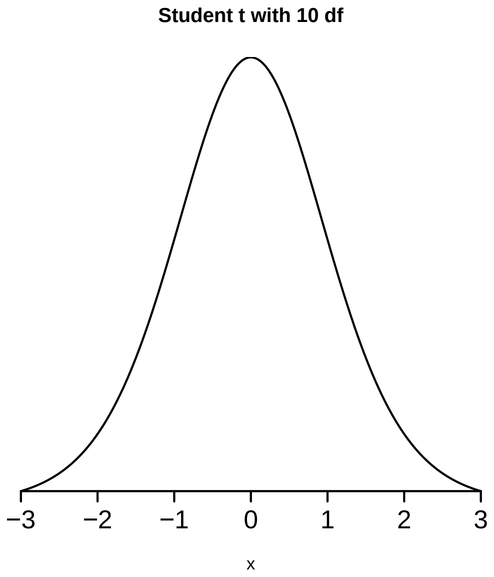
$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is t with $n - 1$ degrees of freedom (df). This is denoted by $T_n \sim t_{n-1}$.

Remember: The df is always $n - 1$ or 1 less than the sample size. That is, df $= n - 1$.

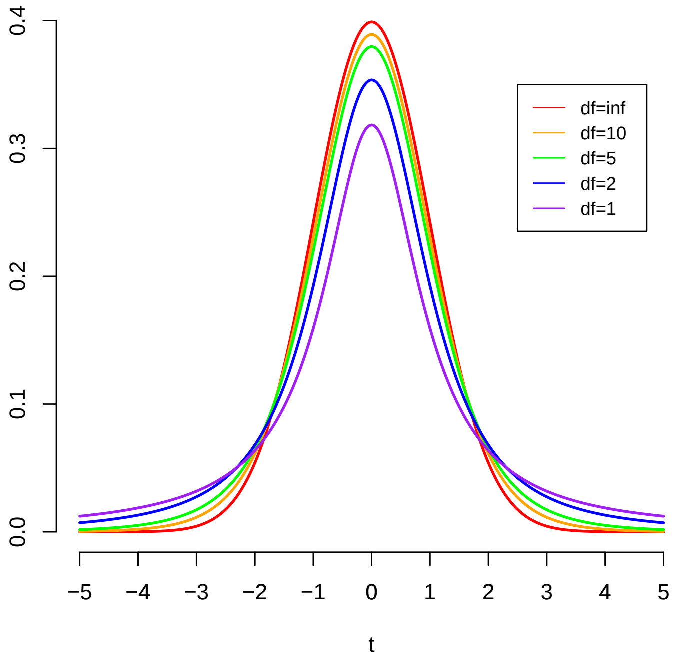
Shape of a t distribution

1. The curve is symmetric about zero with fat tails.
2. The larger the df (sample size), the thinner is the tail.
3. The total area under this curve is also one.



Note: The t distribution approaches normal if the df approaches infinity.

The following graph shows the shapes of different t distributions.



The t-table

For our convenience, similar to that of standard normal tables, the probabilities under a t curve for various df are tabulated. However, the shaded area is given under the t curve is the right tail.

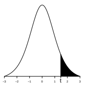
Notes:

1. The t distribution is very useful for many problems with small samples, especially when $n < 30$.

2. To obtain $p = P(t_d < k)$, use the R command `pt(k,d)`

3. To find a percentile value k for a given probability p use the command `qt(p,d)`.

Table 3: Student's t Distribution Table



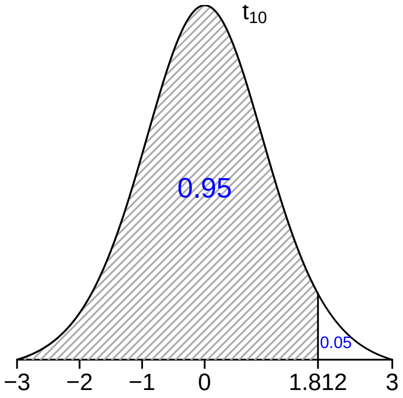
Percentile $P(t_\nu > t) = p$ for Student's t-distributions with ν d.f.

ν,p	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	1.000	3.078	6.314	12.706	31.821	63.656	318.289
2	0.816	1.886	2.920	4.303	6.965	9.925	22.328
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.894
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
35	0.682	1.306	1.690	2.030	2.438	2.724	3.340
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
45	0.680	1.301	1.679	2.014	2.412	2.690	3.281
50	0.679	1.299	1.676	2.009	2.403	2.678	3.261
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Example: Find the 95th percentile (or the upper 5th percentile) of a t distribution with 10 df.

Solution: From the t-table, the value with 10 df and 95th percentile or upper 5th percentile is 1.812. Hence

$$P(t_{10} < \underline{\hspace{2cm}}) \quad = \quad 0.95$$



Exercises:

1. Find the 1st percentile of t_{22} using table 3.
Answer: 2.508.

2. Find k such that
(i) $P(t_{15} < k) = 0.90$ and
(ii) $P(t_{25} > k) = 0.01$ using table 3.
Answers: (i) 1.341; (ii) 2.485

7.7 Confidence interval for μ on small samples (P.101-105)

A *point* estimate of μ can be obtained by the sample mean $\bar{x}$. As different samples provide different $\bar{x}$, we do not know whether our sample estimate is good enough to represent the population mean μ. Further, it does not show the variability of $\bar{x}$ across all possible samples. Moreover, the chance that it equals exactly to the true mean μ is (almost) zero.

To avoid this uncertainty, statisticians calculate an interval containing the true parameter with a given high probability. This is called an interval estimate or confidence interval (CI).

That is, an *interval* estimate, shows with a high level of probability (confidence) that it includes the true parameter.

We first look at the problem of constructing CI for the population mean μ based on a small sample from a normal population.

The key result we use in small sample inference is

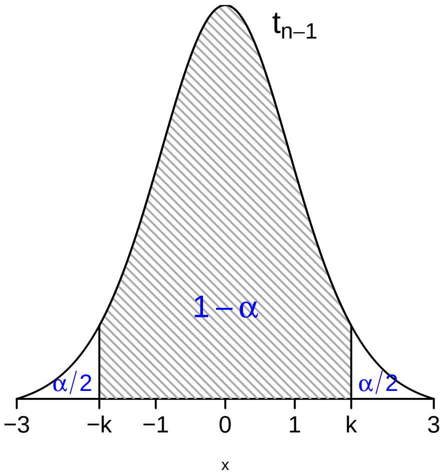
$$T_{n-1} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where $X_1, X_2, \dots, X_n$ is a random sample of size n (n is small or less than 30) from $N(\mu, \sigma^2)$ and σ^2 is unknown.

From the t table, we can find the value k such that

$$P(-k < t_{n-1} < k) = 1 - \alpha$$

equals a given (or known) probability. Note that α is the sum of area on two sides and $\alpha/2$ is the area on each side.



Note: In general a confidence interval (CI) is defined as follows:

Definition: The random interval $[\hat{\theta}_L, \hat{\theta}_R]$ is called a $100(1-\alpha)\%$ *confidence interval* (CI) for θ if

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R) = 1 - \alpha,$$

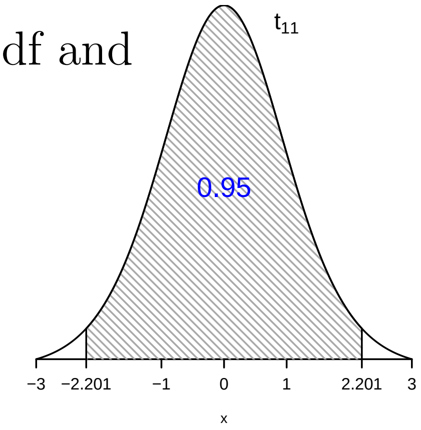
where $\hat{\theta}_L$ and $\hat{\theta}_R$ are two statistics.

Then α is called the *significant level* of the CI whereas $1 - \alpha$ is the *confident* level. For a 95% CI, say, the significant level $\alpha = 0.05$.

Example: Find the value of k such that $P(-k < t_{11} < k) = 0.95$.

Solution: From the t table, the value with 11 df and 95th percentile or upper $(1 - 0.95)/2$, i.e. 0.025 percentile is $k = 2.201$.

Therefore, $P(\_\_\_\_\_ < t_{11} < \_\_\_\_\_) = 0.95$.



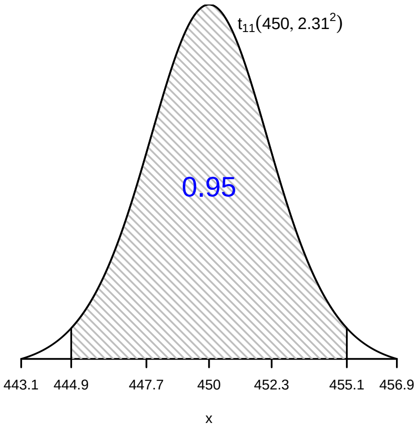
Note: In this example, we allocate 95% of the probability in the middle (symmetric) of the t_{11} curve. We use this approach to find a symmetric confidence interval (CI) for μ.

Example: Heights were measured for 12 plants grown under the treatment of a particular nutrient. The sample mean, $\bar{x} = 450$ cm and the standard deviation, $s = 8$ cm. Construct a 95% CI for μ.

Solution: Note that:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{12}} \sim t_{11}.$$

From the t table, the value of k such that $P(-k < t_{11} < k) = 0.95$ is $k = 2.201$. Now we have,



____

or ____

or ____

This gives a 95% CI for μ as

____

or ______________________________

Since

$$-t_{n-1,\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1,\alpha/2} \Leftrightarrow \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},$$

where $t_{n-1,\alpha/2}$ is the value of the t distribution with $n-1$ df and the middle area of $(1-\alpha)$ or equivalently the upper area of $\alpha/2$. Hence the $(1-\alpha)$% CI for μ is

$$\left(\bar{X} - t_{n-1,\alpha/2}\,\frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1,\alpha/2}\,\frac{S}{\sqrt{n}}\right).$$

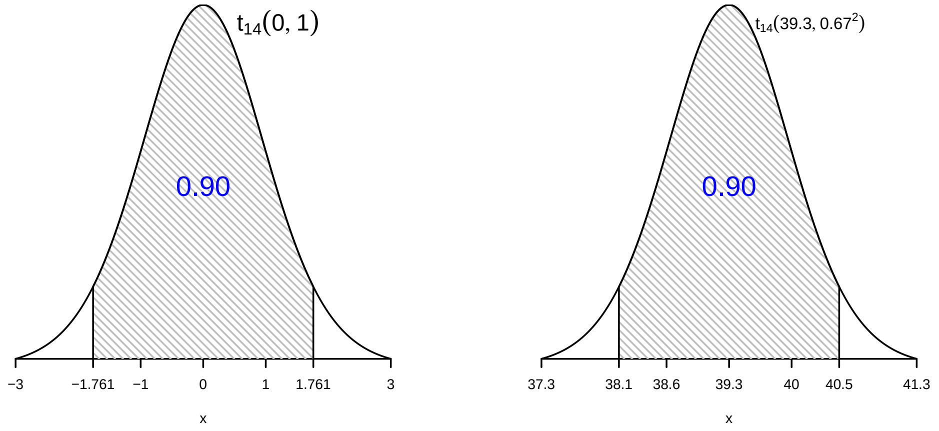
Example: A random sample of $n = 15$ from a normal population gave $\bar{x} = 39.3$ and $s = 2.6$. Find a 90% CI for μ.

Solution: We have df=$15 - 1 = 14$ and upper percentile is $(1 - 0.90)/2 = 0.05$. The t table gives, $t_{14} = 1.761$. Therefore, a 90% CI for μ is:

$$\left(\bar{X} - t_{n-1} \times S/\sqrt{n},\ \bar{X} + t_{n-1} \times S/\sqrt{n}\right)$$

= ____

= ______________

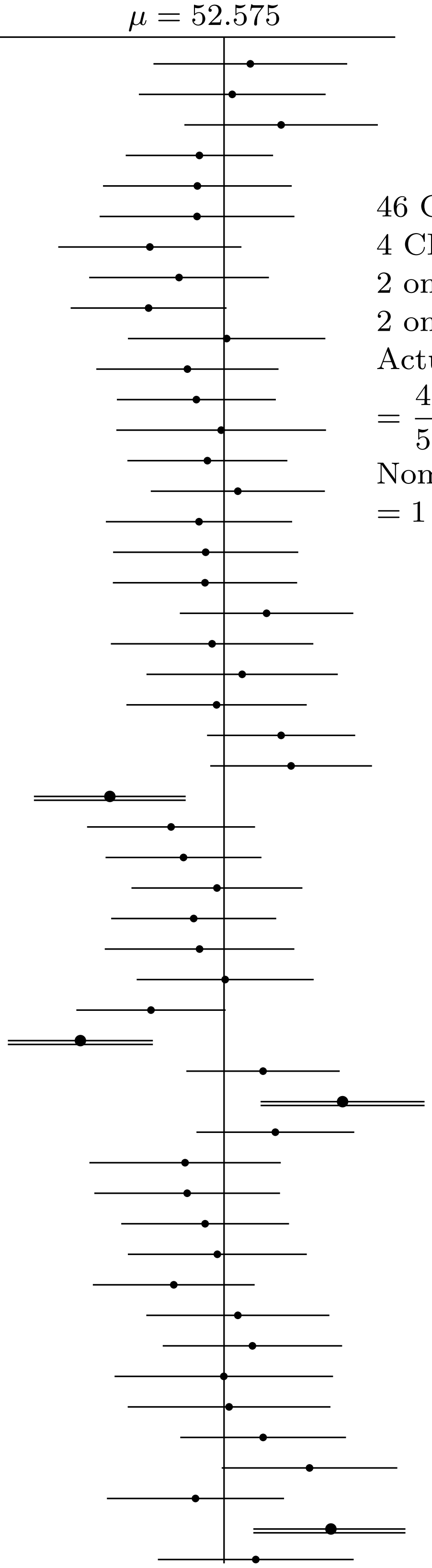


7.8 Interpretation of a CI

Over the collection of all CIs (at a given sig. level, say 95%) that could be constructed from repeated random samples of size n, 95% will contain the true parameter μ. It does not mean that $P(a < \mu < b) = 0.95$ since the population mean μ is fixed and hence the probability about a fixed value does not make sense. The following example demonstrates the meaning of CI.

Example: In a population with $\mu = 52.575$ and $\sigma^2 = 886.847$, 50 random samples of size $n = 20$ are drawn. The 50 CIs and their coverage of the true mean μ are shown below.

$\overline{x}$	s^2	C.I.
56.020	1047.629	(43.332, 68.708)
53.650	973.679	(41.418, 65.882)
60.052	1044.769	(47.381, 72.722)
49.350	606.324	(39.697, 59.002)
49.082	994.433	(36.721, 61.444)
49.038	1058.878	(36.282, 61.794)
42.857	937.009	(30.858, 54.856)
46.682	901.619	(34.911, 58.453)
42.694	677.978	(32.487, 52.901)
52.922	1086.781	(39.999, 65.844)
47.778	926.727	(35.845, 59.712)
48.950	705.443	(38.539, 59.362)
52.200	1227.258	(38.467, 65.933)
50.395	714.205	(39.919, 60.871)
54.384	845.914	(42.982, 65.785)
49.296	968.221	(37.099, 61.494)
50.167	957.080	(38.040, 62.295)
50.082	948.243	(38.010, 62.153)
58.146	840.061	(46.785, 69.508)
51.010	1144.449	(37.749, 64.271)
54.947	1021.469	(42.418, 67.476)
51.596	907.564	(39.787, 63.405)
60.053	612.693	(50.350, 69.756)
61.360	730.304	(50.767, 71.954)
37.612	642.730	(27.674, 47.550)
45.641	788.646	(34.632, 56.640)
47.266	678.076	(37.059, 57.474)
51.645	815.394	(40.452, 62.839)
48.601	760.584	(37.790, 59.412)
49.368	1003.110	(36.953, 61.784)
52.723	874.174	(41.133, 64.313)
43.005	622.081	(33.228, 52.782)
33.760	586.996	(24.262, 43.257)
57.683	656.446	(47.639, 67.726)
68.100	750.229	(57.363, 78.837)
59.298	695.199	(48.962, 69.634)
47.474	1021.986	(34.942, 60.006)
47.749	962.295	(35.588, 59.909)
50.098	785.590	(39.111, 61.085)
51.697	893.741	(39.978, 63.416)
45.989	731.062	(35.390, 56.588)
54.382	735.614	(42.392, 66.373)
56.294	898.002	(44.547, 68.041)
52.548	1333.015	(38.236, 66.860)
53.236	1147.398	(39.958, 66.514)
57.694	766.730	(46.840, 68.548)
63.771	860.750	(52.270, 75.271)
48.835	875.848	(37.234, 60.437)
66.575	645.377	(56.416, 76.333)
56.731	1070.385	(43.906, 69.556)



46 CIs cover μ
4 CIs not cover μ
2 on left &
2 on right
Actual coverage
$= \frac{46}{50} = 0.92$
Nominal coverage
$= 1 - \alpha = 0.95$

Note that the CIs change in both location and length as we move from sample to sample. Hence CI is also *random* and in repeated sampling, roughly 95% of the CIs contain the true mean μ.

7.9 CI for proportion (P.106-108)

The ideas of CLT and CI can be applied to binary variables. When the outcomes are binary, the sample mean becomes the sample proportion of "success" for certain event of interest, for example, the germination of seeds.

Let $X_1, \ldots .X_n$ be the sample of size n (independent trials) and $X_i = 1$ if the i-th observation is a "success" and 0 otherwise. Then $X_i \sim Bern(p)$ where p is the true but unknown probability or population proportion of "success" which is estimated by the sample proportion $\frac{X}{n}$ where X is the sample total. From Chapter 5, we know that

$$X = X_1 + \cdots + X_n \sim B(n, p).$$

By CLT, the sample mean of $X_1, \ldots, X_2$ is

$$\hat{p} = \frac{X}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ when } n \text{ is large.}$$

What should the mean μ and variance σ^2 be?

Answer: They refer to the random variable $Y_i \sim Bern(p)$ such that

$$E(Y_i) = \mu = p \quad \text{and} \quad Var(Y_i) = \sigma^2 = p(1-p).$$

Hence

$$\hat{p} = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right) \text{ when } n \text{ is large.}$$

Then the $(1-\alpha)\%$ CI for the population proportion p is given by:

$$\left(\hat{p} - Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

Example: Find the 95% confidence interval for the germination rate if 65 out of 100 seeds are germinated in an experiment.

Solution: We have $X = 65$, $n = 100$, $\alpha = 0.05$ and $z_{0.975} = 1.96$. The sample proportion of germination is

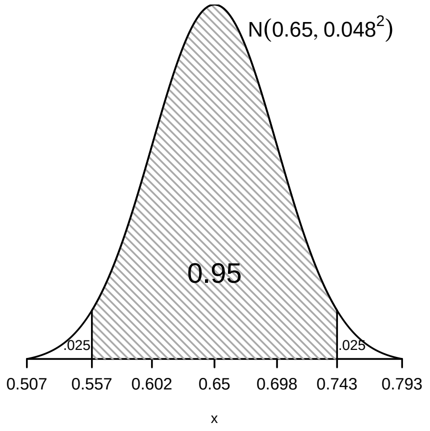
$$\hat{p} = \frac{X}{n} = \underline{\qquad\qquad}$$

Hence the 95% CI for the population proportion p is

$$\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$= \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

$$= \underline{\qquad\qquad\qquad}$$



Exercise: The experiment is repeated with another type of seeds. Find the 90% CI for the germination rate if 105 out of 120 seeds are germinated.

Answer: $(0.8253,\ \ 0.9247)$

Exercises: Book P.109 Q2, Q4-8.

In next week, we will look at the problem of making decisions about the population parameters, for example, the population mean μ and population proportion p, based on the sample information. This problem of statistical inference is known as *hypothesis testing*.