Module 2

General Sir John Kotelawala Defense University

Department of Mathematics

MA 3102 APPLIED STATISTICS

In an industrial process the diameter of a ball bearing is an important component part. The buyer sets specification on the diameter to be 3.0 ± 0.098 cm, that is (2.902, 3.098) cm. The implication is that ball bearing with diameter outside these specifications will be scrapped. It is known that in the process, the diameter of a ball bearing has a normal distribution with mean $\mu = 3.0$ cm and standard deviation $\sigma = 0.05$ cm. For each of the questions below, provide numerical answer.

a. What is the probability that a ball bearing will be scrapped?

b. If four ball bearings are selected at random what is the probability that the average diameter of these four ball bearings would be less than 3.05 cm?

How do you obtain the value for $\mu \ and \ \sigma^2$

If we don't know the $\mu \ and \ \sigma^2$

Then these values are to be estimated

Estimation

A. Point estimation
B. Interval estimation

Point estimation

- Obtaining a particular value as an estimate of a parameter.
- Usually the value of the point estimator defers from the actual value of the parameter.
- In a random sample of *n* observation, the sample mean $\bar{x}$ and sample variance s^2 are the point estimators of the population mean μ and population variance σ^2respectively.

Example

We could measure a variable (Ex: Inside diameter of the test tubes) of random sample n=20 observations. If sample mean $\bar{x}$=1.5 cm and sample variance s^2 = 0.05 cm^2 and the point estimators of μ = 1.5 cm & σ^2=0.05 cm^2

Why we go for Sampling?

- Obtained in order to derive a statistic (a mean for example) from the observations that constitute in the sample.
- Large sample – more reliable than small sample

Interval estimation

- Usually the value of the point estimator defers from the actual value of the parameter but could be high close.
- Therefore giving one value is not so satisfactory.
- To get over this difficulty we introduce interval estimators or confidence interval.
- An interval estimator of a parameter is the interval between two statistics that include the true value of the parameter with some probability.

Example: $P(\text{Lower limit} \leq \mu \leq \text{Upper limit}) = (1-\alpha) = 0.95$

Example: Average Number of Children in a family

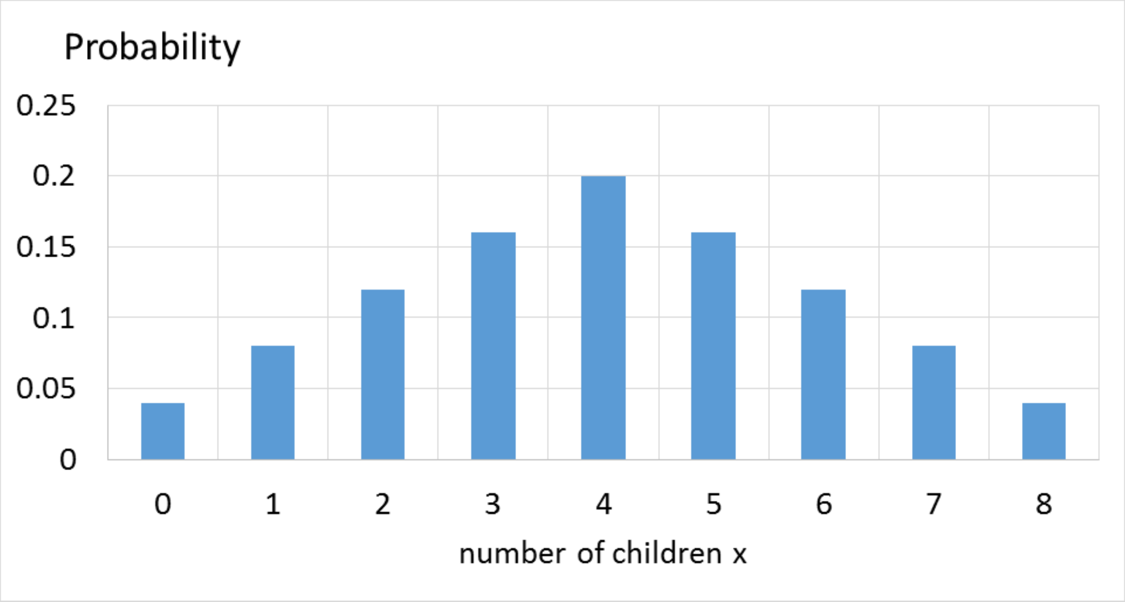
- Say our population size is 5 and we don't want to ask all families.
- The population having **0, 2, 4, 6, 8 children**
- We ask only 2 families at random coming for a super market at 8.00 am and 2.00 pm.

sample #	first	second	Avg
1	0	0	**0**
2	0	2	*1*
3	0	4	2
4	0	6	3
5	0	8	4
6	2	0	*1*
7	2	2	2
8	2	4	3
9	2	6	4
10	2	8	5
11	4	0	2
12	4	2	3
13	4	4	4
14	4	6	5
15	4	8	6
16	6	0	3
17	6	2	4
18	6	4	5
19	6	6	6
20	6	8	7
21	8	0	4
22	8	2	5
23	8	4	6
24	8	6	7
25	8	8	8

0, 2, 4, 6, 8
Calculate the mean for each possible samples for sample size 2.

X	#	Probability
0	1	0.040
1	2	0.080
2	3	0.120
3	4	0.160
4	5	0.200
5	4	0.160
6	3	0.120
7	2	0.080
8	1	0.040
	25	

All calculated means are the estimates of same population parameter. The variation between these individual estimates are called sampling error. The way in which sample statistics cluster around a population parameter is called the ***distribution of the statistic*** or the ***sampling distribution*.**



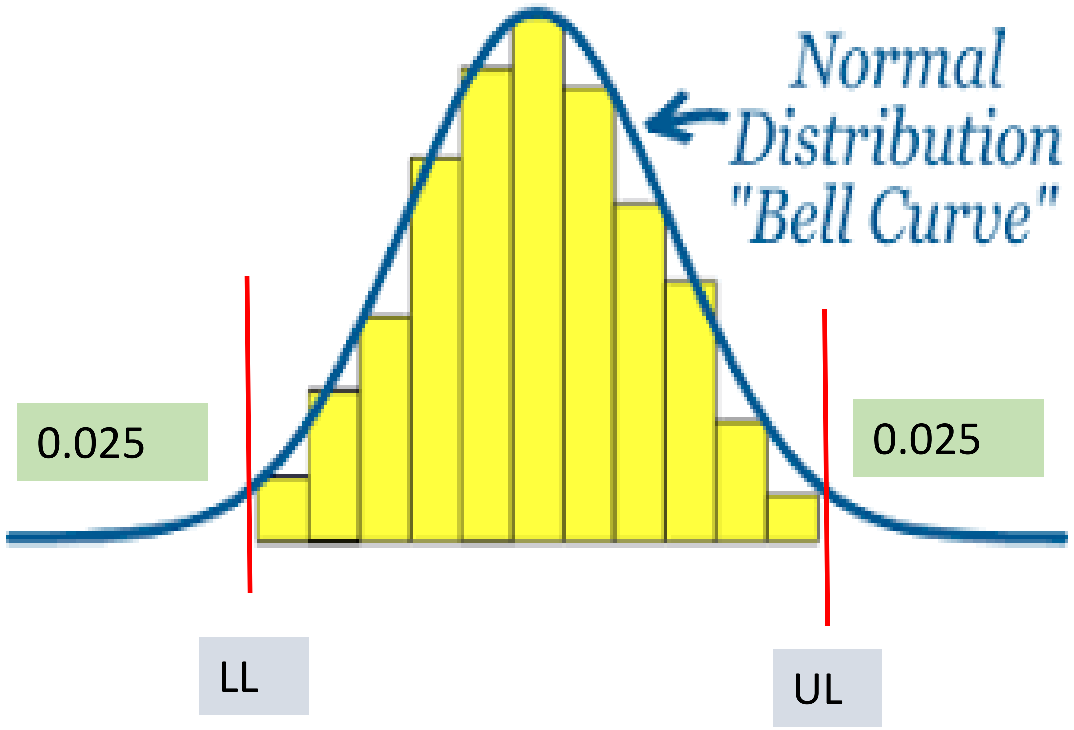
A. Interval Estimation (CI) on the Mean μ where variance σ^2 is known

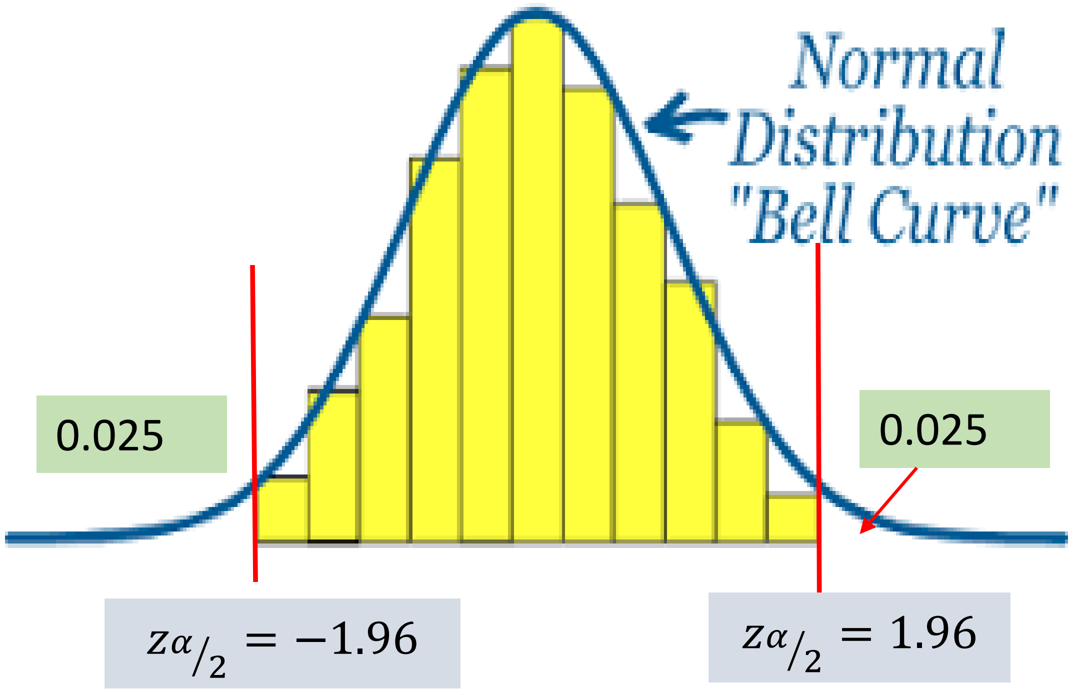
Example

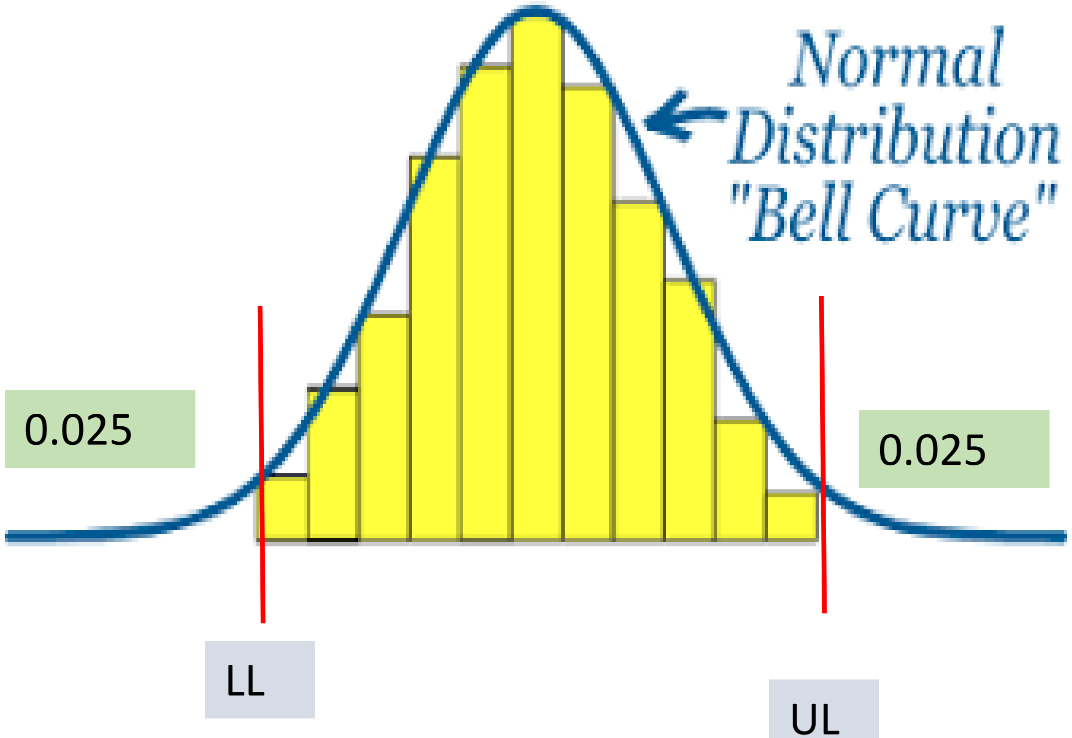
- Let's take an example to compute this.
- We take a sample of 16 stocks from a large population with a mean return of 5.2%. We know that the population standard deviation is 1.5%.
- Calculate the 95% confidence interval for the population mean.
- For 95% confidence interval, $z\alpha/2 = 1.96$
- The confidence interval will be:

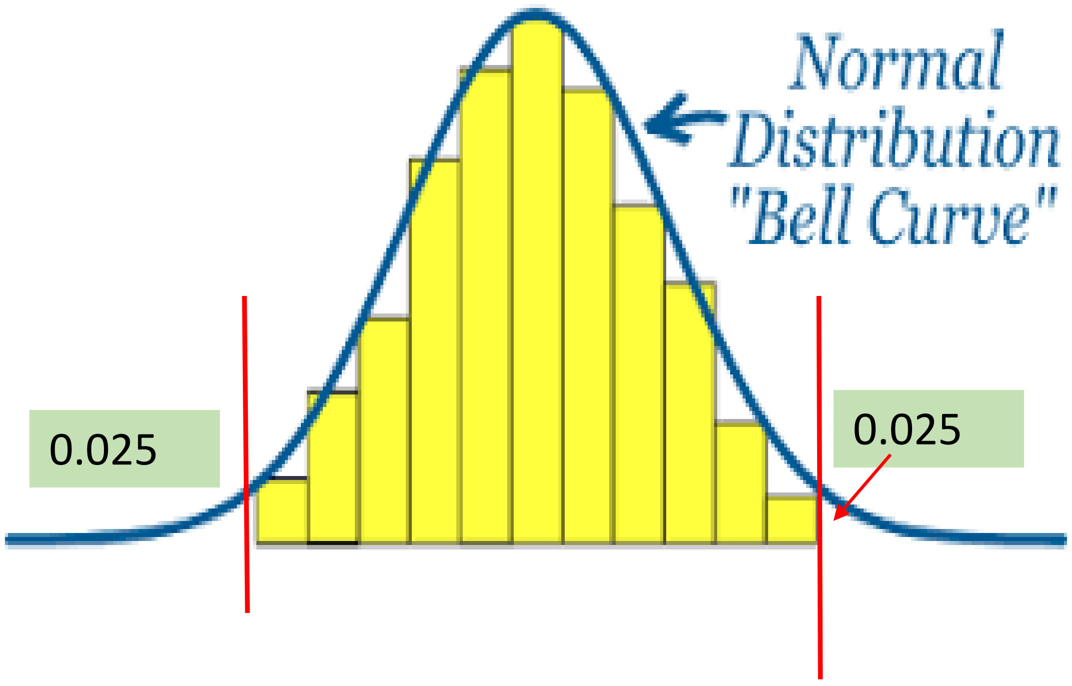
Example

- Let’s take an example to compute this.
- We take a sample of 16 stocks from a large population with a mean return of 5.2%. We know that the population standard deviation is 1.5%.
- Calculate the 95% confidence interval for the population mean.
- For 95% confidence interval, $z\alpha/2$ = 1.96
- The confidence interval will be:









$z_{\alpha/2} = -1.96$

$z_{\alpha/2} = 1.96$

$$-z_{\alpha/2} = \frac{LL - \mu}{\sigma/\sqrt{n}}$$

$$LL = \mu - z_{\alpha/2}\,\sigma/\sqrt{n}$$

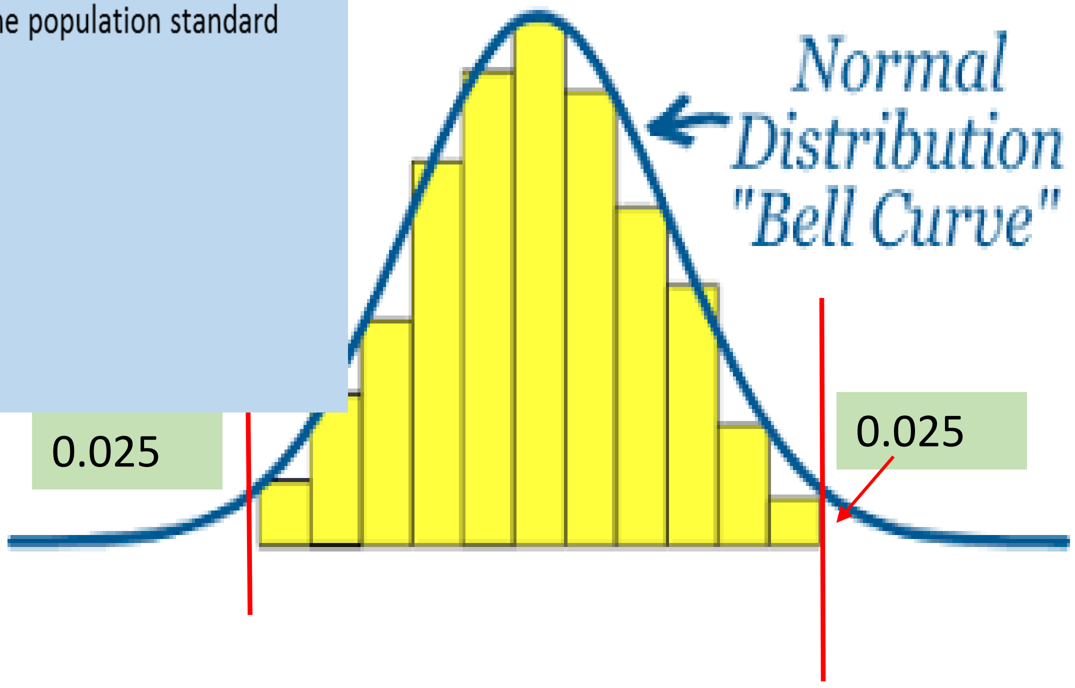
$$z_{\alpha/2} = \frac{UL - \mu}{\sigma/\sqrt{n}}$$

$$UL = \mu + z_{\alpha/2}\,\sigma/\sqrt{n}$$

$$CI = \bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$$

Example

- Let's take an example to compute this.
- We take a sample of 16 stocks from a large population with a mean return of 5.2%. We know that the population standard deviation is 1.5%.
- Calculate the 95% confidence interval for the population mean.
- For 95% confidence interval, $z\alpha/2 = 1.96$
- The confidence interval will be:



$$CI = \bar{x} \pm z\alpha/_2 \, {}^{\sigma}/_{\sqrt{n}}$$

$$Confidence\ Interval = 5.2\% \pm 1.96 \times \frac{1.5\%}{\sqrt{16}}$$

$$= 5.2\% +/- 0.735\%$$

$$\text{Confidence Interval} = 4.465\% < \mu < 5.935\%$$

$$z\alpha/_2 = -1.96 \qquad z\alpha/_2 = 1.96$$

$$-z\alpha/_2 = \frac{LL - \mu}{\sigma/\sqrt{n}} \qquad z\alpha/_2 = \frac{UL - \mu}{\sigma/\sqrt{n}}$$

$$LL = \mu - z\alpha/_2 \, {}^{\sigma}/_{\sqrt{n}} \qquad UL = \mu + z\alpha/_2 \, {}^{\sigma}/_{\sqrt{n}}$$

We are 95% confidence that the true mean is between 4.465% and 5.935%.
z is obtained from the standard normal distribution table as shown below. F(Z) value is 0.025 at z = -1.96 and F(Z) value is 0.9750 at z = 1.96.

A. Interval Estimation (CI) on the Mean μ where variance σ^2 is unknown

Example

- Let's take an example to compute this.
- A sample of size 15 is taken from a larger population
- Sample mean is calculated as 12 and the sample variance as 25.
- What is the 95% confidence interval for the population mean μ?

σ^2 is unknown and we know only s^2.

The expression for the confidence interval in this case is similar to the case where σ 2 is known:

$$CI = \bar{x} \pm t\alpha_{/_{2},n-1} {}^{s}/_{\sqrt{n}}$$

where σ^2 has been replaced with the sample variance s^2 and $t\alpha_{/_{2},n-1}$ is the appropriate value from the t–distribution tables.

Example

- Let's take an example to compute this.
- A sample of size 15 is taken from a larger population
- Sample mean is calculated as 12 and the sample variance as 25.

$$CI = \bar{x} \pm t_{\alpha/2, n-1} \, {}^{s}/_{\sqrt{n}}$$

$$t_{0.05/2, 15-1} = 2.145$$

$$12 \pm 2.145 \times \frac{5}{\sqrt{15}}$$

$$12 \pm 2.77$$

Hence, the confidence interval is (9.23, 14.77).

TABLE A-3 *t* Distribution: Critical *t* Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.1
Degrees of Freedom	**Area in Two Tails**				
	0.01	0.02	0.05	0.10	0.2
1	63.657	31.821	12.706	6.314	3.07
2	9.925	6.965	4.303	2.920	1.88
3	5.841	4.541	3.182	2.353	1.63
4	4.604	3.747	2.776	2.132	1.53
5	4.032	3.365	2.571	2.015	1.47
6	3.707	3.143	2.447	1.943	1.44
7	3.499	2.998	2.365	1.895	1.41
8	3.355	2.896	2.306	1.860	1.39
9	3.250	2.821	2.262	1.833	1.38
10	3.169	2.764	2.228	1.812	1.37
11	3.106	2.718	2.201	1.796	1.36
12	3.055	2.681	2.179	1.782	1.35
13	3.012	2.650	2.160	1.771	1.35
14	2.977	2.624	2.145	1.761	1.34
15	2.947	2.602	2.131	1.753	1.34
16	2.921	2.583	2.120	1.746	1.33
17	2.898	2.567	2.110	1.740	1.33
18	2.878	2.552	2.101	1.734	1.33
19	2.861	2.539	2.093	1.729	1.32
20	2.845	2.528	2.086	1.725	1.32
21	2.831	2.518	2.080	1.721	1.32
22	2.819	2.508	2.074	1.717	1.32
23	2.807	2.500	2.069	1.714	1.31
24	2.797	2.492	2.064	1.711	1.31
25	2.787	2.485	2.060	1.708	1.31

TABLE A-3 *t* Distribution: Critical *t* Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
	Area in Two Tails				
Degrees of Freedom	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310
31	2.744	2.453	2.040	1.696	1.309
32	2.738	2.449	2.037	1.694	1.309
34	2.728	2.441	2.032	1.691	1.307
36	2.719	2.434	2.028	1.688	1.306
38	2.712	2.429	2.024	1.686	1.304
40	2.704	2.423	2.021	1.684	1.303
45	2.690	2.412	2.014	1.679	1.301
50	2.678	2.403	2.009	1.676	1.299
55	2.668	2.396	2.004	1.673	1.297
60	2.660	2.390	2.000	1.671	1.296
65	2.654	2.385	1.997	1.669	1.295
70	2.648	2.381	1.994	1.667	1.294
75	2.643	2.377	1.992	1.665	1.293
80	2.639	2.374	1.990	1.664	1.292
90	2.632	2.368	1.987	1.662	1.291
100	2.626	2.364	1.984	1.660	1.290
200	2.601	2.345	1.972	1.653	1.286
300	2.592	2.339	1.968	1.650	1.284
400	2.588	2.336	1.966	1.649	1.284
500	2.586	2.334	1.965	1.648	1.283
750	2.582	2.331	1.963	1.647	1.283
1000	2.581	2.330	1.962	1.646	1.282
2000	2.578	2.328	1.961	1.646	1.282
Large	2.576	2.326	1.960	1.645	1.282

Example

A credit card company wants to determine the mean income of its card holders. It also wants to find out if there are any differences in mean income between males and females. A random sample of 225 male card holders and 190 female card holders was drawn, and the following results obtained:

	Mean	Standard deviation
Males	£16 450	£3675
Females	£13 220	£3050

Calculate 95% confidence intervals for the mean income for males and females. Is there any evidence to suggest that, on average, males' and females' incomes differ? If so, describe this difference

- 95% confidence interval for male income
- The true population variance, σ 2 , is unknown
- Use the t–distribution
- i.e. x¯ ± tp/2 × p s 2/n. Here, x¯ = 16450, s 2 = 36752 = 13505625 and n = 225. The value tp/2 must be found from table
- Recall that the degrees of freedom, ν = n−1, and so here we have ν = 225 − 1 = 224.
- Notice that table 1.1 only gives value of ν up to 2000;
- for higher values, we use the ∞ row.
- Since we require a 95% confidence interval, we read down the 5% column, giving a t value of 1.96
- (recall that this is the same as the value used if σ 2 were known and we used the normal distribution – that's because the t–distribution converges to the normal distribution as the sample size increases).
- Thus, the 95% confidence interval for μ is found as 16450 ± 1.96 × p 13505625/225, i.e. 16450 ± 480.2. So, the 95% confidence interval is (£15969.80, £16930.20).
- 95% confidence interval for female income Again, the true population variance, σ 2 , is unknown
- x¯ ± tp/2 × p s 2/n. Now, x¯ = 13220, s 2 = 30502 = 9302500, and n = 190. Again, since the sample size is large, we use the ∞ row of table 1.1 to obtain the value of tp/2, and so the 95% confidence interval for μ is found as 13220 ± 1.96 × p 9302500/190, i.e. 13220 ± 1.96 × 221.27, i.e. 13220 ± 433.69. So, the 95% confidence interval is (£12786.31, £13653.69).
- Since the 95% confidence intervals for males and females do not overlap, there is evidence to suggest that males' and females' incomes, on average, are different.
- Further, it appears that male card holders earn more than women.