MAS1403

Quantitative Methods for Business Management

Semester 2, 2011—12

Lecturer: Dr. Lee Fawcett

School of Mathematics & Statistics

Chapter 1

Confidence intervals

1.1 Introduction

Statistics is the science that studies the collection and interpretation of data to enable us to draw conclusions about the population from which these data have been drawn. As the sample we draw from the population gets bigger and bigger, it becomes increasingly difficult to "picture" the data just by looking at the raw observations; the data are easier to understand if we can find some way of *summarising* them.

Recall from semester 1 that data can be summarised in two ways: graphically and numerically. Various graphical techniques available for summarising data were presented, including stem and leaf plots, bar charts, histograms, relative frequency histograms and frequency polygons. You were also introduced to two types of numerical summaries: measures of *location* and measures of *spread*.

A measure of location is a quantity which is 'typical' of the data; examples of such measures include

(i) the sample mean, usually denoted $\bar{x}$ ("add them up and divide by how many we have");

(ii) the sample median ("the one in the middle"), and

(iii) the sample mode ("the value which occurs most often").

A measure of spread is a value which quantifies the variability in the data (or how "spread out" the observations are); examples include

(i) the range ("largest minus smallest");

(ii) the variance ("average squared distance of observations from the mean" – the standard deviation, usually denoted s, is the square root of this value), and

(iii) the inter–quartile range ("upper quartile minus lower quartile" – this value represents the middle 50% of the data, as the lower quartile has one quarter of the data less than it, and the upper quartile has three–quarters of the data less than it).

1.2 Estimation

Last semester we concentrated on pure description of data, although we recognised that this might prompt us to ask pertinent questions about the population from which the sample was drawn. What exactly does the sample, often a tiny subset, tell us of the population? We can never observe the whole population, even if it is finite, except at enormous expense, and so the population mean and variance (or indeed any aspect of the population distribution) can never be known exactly. We call these unknown quantities *parameters* and use Greek letters to denote them:

- μ ("mu") is the symbol commonly used for the population mean, and
- σ ("sigma") is commonly used for the population standard deviation.

Hopefully (and if we have a representative sample), the sample mean ($\bar{x}$) will be quite close to the true population mean μ; likewise, the sample standard deviation (s) will be a good estimator for σ. **In this section, we concentrate on $\bar{x}$ as an estimator for μ.**

Before we can use our sample of n observations we must ask the question: Is $\bar{x}$ a "good" estimate of μ? How do we infer (find something out about) the unknown μ using $\bar{x}$? So long as the sample size n is fairly large, we can hope that $\bar{x}$ is close to μ. But how close is it? To answer this question, we must make some plausible assumptions about the population. But first, let's consider the following example.

Example: *The Vintage Clothing Co.*

The *Vintage Clothing Co.* are a large retailer of bespoke and retro clothing. They have 1,000 branches across the U.K., and all of their branches are open on Sundays. However, they are considering whether or not it is worthwhile staying open on Sundays. Table 1.1 overleaf shows the number of transactions at each of their shops on Sunday 29th January 2012.

Suppose the marketing department of *The Vintage Clothing Co.* are interested in the average number of transactions across all their stores on Sunday 29th January 2012. Can we work this out *exactly*?

The answer is "yes", as we have data from every single branch in Table 1.1. Actually, Table 1.1 shows we have taken a *census* – every single branch has been asked to provide us with data. So in this case, it is possible to work out the *population mean* μ:

$$\mu = \frac{282 + 258 + 399 + 271 + \ldots + 426 + 477}{1000} = 320 \text{ transactions.}$$

Now, let's suppose the company don't have the time/resources to take a census. In fact, a week before the 29th January, just five stores are selected at random, and we work out the mean number of transactions using the data from these five stores only. Let's suppose the top left–hand block in Table 1.1 above are stores 1–100, the next block along are stores 101–200, etc. We put the numbers 1–1000 into a bag and draw, without replacement, 5 numbers at random:

282	258	399	271	343	285	247	513	171	123	168	327	430	240	410	341	90	512	245	336
290	263	446	185	330	111	243	376	139	351	311	389	546	321	393	487	287	514	149	315
264	320	217	257	349	640	97	298	393	454	363	354	360	326	199	502	154	273	213	413
293	407	362	270	344	263	290	263	50	253	345	581	229	264	304	394	246	235	417	452
499	276	412	323	310	177	248	178	409	275	278	307	495	515	232	432	577	269	370	248
339	404	371	262	336	218	274	483	211	245	316	381	432	233	223	447	412	250	262	337
202	133	356	408	224	379	197	278	235	509	171	232	429	315	326	602	63	290	230	121
242	389	219	206	393	437	306	152	294	271	230	398	346	344	379	347	468	300	325	237
305	174	291	261	214	532	335	63	100	357	190	347	208	420	322	463	203	216	356	504
389	236	445	378	255	301	308	150	289	453	464	273	211	450	222	250	214	259	296	356
320	420	357	160	372	99	316	218	248	322	145	399	433	393	403	361	241	234	388	255
261	279	369	342	168	322	304	254	99	503	303	212	105	166	257	422	460	331	288	410
346	370	235	355	65	340	420	338	568	644	164	288	319	159	324	208	452	297	305	259
268	340	305	361	319	519	293	380	286	431	402	329	363	330	612	248	302	592	589	349
446	588	304	454	164	240	293	478	540	339	245	257	222	471	469	273	244	126	174	183
277	216	555	401	380	338	212	476	77	363	140	451	329	66	217	461	435	380	314	324
522	111	119	316	116	471	142	336	277	101	518	264	226	256	539	324	320	292	476	324
333	332	404	362	202	204	341	80	333	267	439	136	343	389	244	370	268	362	317	400
372	595	314	182	470	192	555	374	368	192	225	321	435	403	316	312	307	368	236	452
192	63	407	125	253	89	70	186	491	342	122	367	106	334	161	177	180	355	356	317
454	122	286	39	361	262	316	272	285	201	191	162	229	334	278	231	154	290	277	392
644	297	398	118	246	148	478	167	337	344	395	334	255	401	504	304	408	204	673	126
192	507	41	457	405	306	282	446	195	512	252	510	557	191	321	404	542	438	291	449
377	240	441	308	346	265	375	332	580	130	353	426	95	588	332	109	467	333	388	309
263	529	172	529	315	257	481	260	297	382	438	64	226	185	369	275	320	126	321	375
190	340	337	224	363	212	371	229	175	388	332	315	389	452	266	393	323	253	280	420
219	400	378	241	616	551	359	489	314	450	645	224	320	405	182	251	370	341	318	232
240	471	293	240	184	296	617	565	206	147	169	401	140	462	389	310	262	334	263	269
323	351	187	544	387	425	353	175	378	484	205	295	413	189	559	251	480	283	262	304
213	574	579	325	246	206	419	306	471	264	270	300	278	131	561	328	440	514	280	391
281	403	256	348	183	161	444	482	338	268	313	252	179	414	444	266	400	435	433	506
203	269	450	322	459	183	212	242	144	406	401	174	605	270	487	494	235	316	368	319
260	254	157	377	145	284	401	220	452	59	335	467	251	192	371	298	317	382	363	397
282	303	328	378	363	636	374	143	495	239	423	496	411	462	282	411	203	395	590	388
278	272	417	666	233	316	287	268	186	247	339	397	276	291	324	81	271	399	129	325
247	152	315	224	130	323	352	276	398	338	231	258	310	421	215	85	237	356	439	348
507	277	240	188	321	419	370	374	211	224	340	264	441	226	563	279	297	114	546	277
281	196	498	375	348	234	469	103	324	643	315	293	444	109	408	100	477	293	138	206
485	279	494	513	97	293	669	312	425	70	181	210	241	187	448	55	253	564	404	382
31	496	234	200	411	386	218	382	483	405	435	414	379	360	194	291	393	247	314	285
204	188	444	416	106	485	276	250	248	200	352	463	251	197	197	456	293	333	373	240
295	297	271	141	319	256	197	110	338	237	249	291	393	437	432	274	202	182	176	212
482	96	272	296	323	289	285	160	203	336	217	321	202	266	253	436	390	259	596	383
236	291	226	250	270	439	360	310	326	415	447	336	354	273	243	390	213	318	346	599
637	255	61	393	324	492	484	259	271	150	550	185	224	352	387	441	232	261	313	410
246	529	97	448	369	199	140	498	287	293	258	431	267	396	217	340	278	297	387	281
162	237	305	239	246	412	632	385	342	340	673	414	298	383	152	438	408	452	492	603
439	223	404	466	380	214	155	410	291	234	248	325	391	338	416	262	361	358	484	129
152	363	90	383	365	500	362	190	343	138	233	179	200	476	128	308	221	649	278	152
525	275	355	585	394	183	488	323	312	595	257	434	160	375	478	353	239	331	426	477

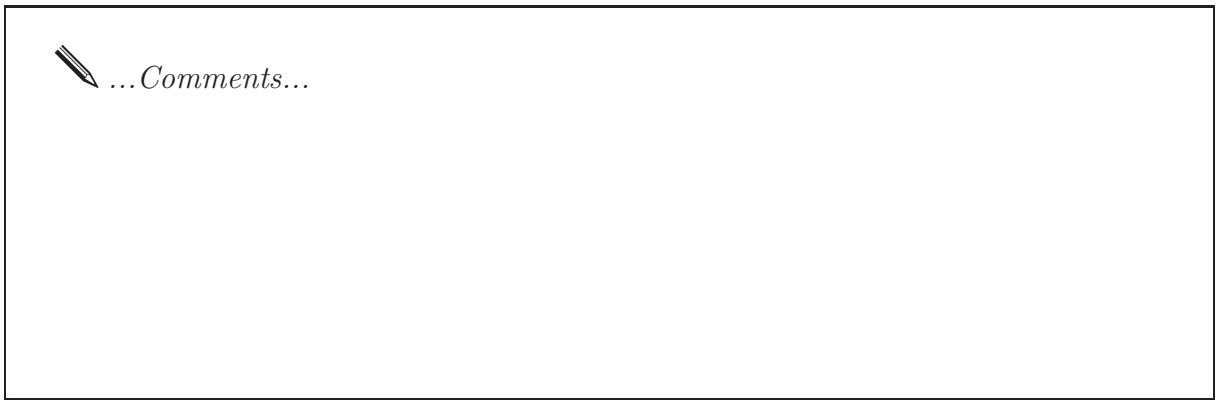
Table 1.1: The number of transaction at each branch of *The Vintage Clothing Co.*

Store	No. of transactions, X
637	$x_1 = 374$
327	$x_2 = 452$
849	$x_3 = 271$
666	$x_4 = 419$
680	$x_5 = 643$

Let's suppose this is the only information we have. It is no longer possible to work out the true population mean, as we don't have information from every single shop; we can now only work out the *sample mean* $\bar{x}$:

$$\bar{x} = \frac{374 + 452 + 271 + 419 + 643}{5} = 431.8 \approx 432 \text{ transactions.}$$

Obviously, the marketing team are not just interested in what goes on in these five shops; however, this is the only information they have, and so they use this information to draw conclusions about all 1,000 shops as a whole. This is known as the process of *statistical inference* – we are trying to *infer* things about the population, based on the limited information in our sample. Hopefully, provided we don't have a *biased sample*, $\bar{x}$ will do a good job at estimating μ. Has it done a good job here?



This begs the question: "*How accurate can our sample mean be in estimating the population mean?*". Let's take another random sample by drawing another five numbers from the bag, at random:

Store	No. of transactions, X
558	$x_1 = 253$
428	$x_2 = 446$
903	$x_3 = 251$
364	$x_4 = 256$
14	$x_5 = 185$

This gives

$$\bar{x} = \frac{253 + 446 + 251 + 256 + 185}{5} = 278.2 \approx 278 \text{ transactions.}$$

This sample mean is much closer to the population mean, but still not very close. Also, it is quite different from the mean of the previous sample. You could repeat this procedure yourself (filling in the table below) to select three more random samples of size 5, and calculate the sample means. How close are these sample means to the correct population value $\mu = 320$ transactions? In the space below, the u's are random numbers obtained using the *random number generator* button `Ran` on your calculator – I will show you how to do this in class – I wouldn't expect you to draw numbers from a bag!

Your random sample 1

$u =$	$store =$	$x_1 =$
$u =$	$store =$	$x_2 =$
$u =$	$store =$	$x_3 =$
$u =$	$store =$	$x_4 =$
$u =$	$store =$	$x_5 =$
		$\bar{x} =$

Your random sample 2

$u =$	$store =$	$x_1 =$
$u =$	$store =$	$x_2 =$
$u =$	$store =$	$x_3 =$
$u =$	$store =$	$x_4 =$
$u =$	$store =$	$x_5 =$
		$\bar{x} =$

Your random sample 3

$u =$	$store =$	$x_1 =$
$u =$	$store =$	$x_2 =$
$u =$	$store =$	$x_3 =$
$u =$	$store =$	$x_4 =$
$u =$	$store =$	$x_5 =$
		$\bar{x} =$

...Comments...

In fact, we could take many samples, and it's very likely that we'll get a different value for $\bar{x}$ each time; it's also very *unlikely* that any of our $\bar{x}$'s will be exactly the same as the true population mean μ. Figure 1.1 below shows histograms of $\bar{x}$'s taken from 100 samples from our population of 1,000 stores. In the top–left graph, we have taken samples of size $n = 5$, like we did in the last couple of pages of these notes. In the other graphs, moving from left to right, we have increased this to 100 samples of size $n = 10$, $n = 50$, $n = 100$, $n = 250$ and $n = 500$. The vertical line indicates the true population mean $\mu = 320$.

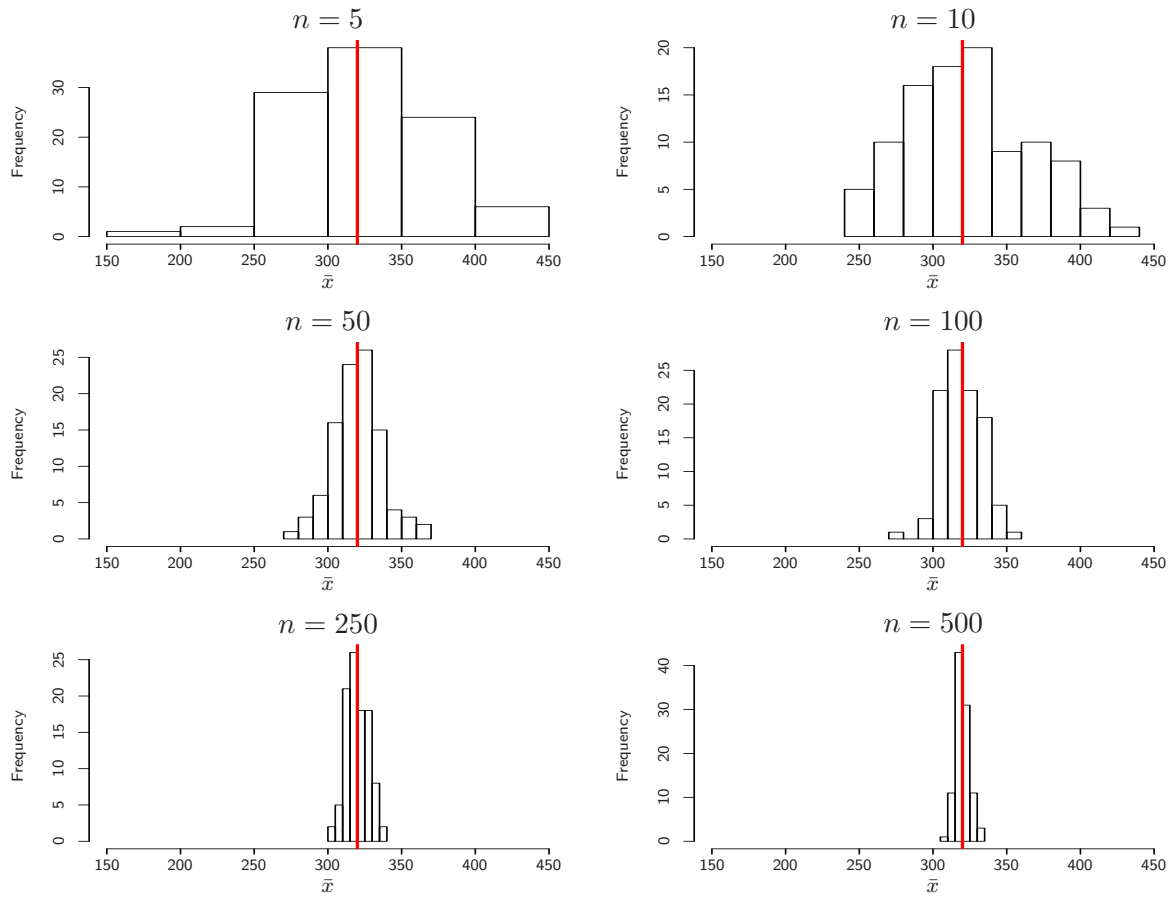


Figure 1.1: Distribution of sample means for increasing sample sizes n.

You should notice two things:

1. The distribution of $\bar{x}$'s, in all plots, looks like a Normal distribution (bell–shaped curve; see semester 1);

2. As we increase the sample size (n), the distribution for $\bar{x}$ gets more and more concentrated – around the true population value $\mu = 320$!

In fact, what we can see in action in this graph is known as the *central limit theorem*. This is a very powerful result in Statistics which tells us about the distribution of the sample mean $\bar{x}$. We now state this formally.

The Central Limit Theorem

Suppose $x_1, x_2, \ldots, x_n$ are a random sample from *any* population, with mean μ and variance σ^2. If n is large, then

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{approximately;}$$

if $x_1, x_2, \ldots, x_n$ come from a Normal distribution themselves, then this result holds for *any* n.

This means that if we were to take many samples of size n, and for each sample calculate the mean $\bar{x}$, then our histogram of $\bar{x}$'s will *always* be Normally distributed around the true population mean μ (if n is large; however, this result is true for *any* n if our random sample is Normally distributed). What's more, we also know about the *variability* of $\bar{x}$. If we know the population variance σ^2, then the variance of $\bar{x}$ is σ^2/n, giving a standard deviation for $\bar{x}$ of $\sigma/\sqrt{n}$. We call this quantity the *standard error*. We will now use this result to form *confidence intervals* for the population mean μ.

1.3 Interval estimation

The values we calculate for sample means and variances are *point estimates*; they are single values based on a limited sample of the whole population. Suppose that we wish to estimate the mean μ of a population. The natural estimate for μ is the sample mean $\bar{x}$. However, as we have seen, $\bar{x}$ is never exactly equal to μ; all we really hope is that $\bar{x}$ will be close to μ. One way of improving our inference is to construct *interval estimates*, more commonly known as *confidence intervals*. We simply place an interval over the point estimate for μ which allows us to say (with a certain level of confidence) within what range the population mean lies. The calculation of these intervals depends on the size of our sample (n), the level of confidence we choose, and whether or not the population variance (σ^2) is known.

1.3.1 Case 1: Known variance σ^2

We know from the results above that, if our random sample is drawn from a Normal distribution, or if n is large (i.e. $n \geq 30$), then

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

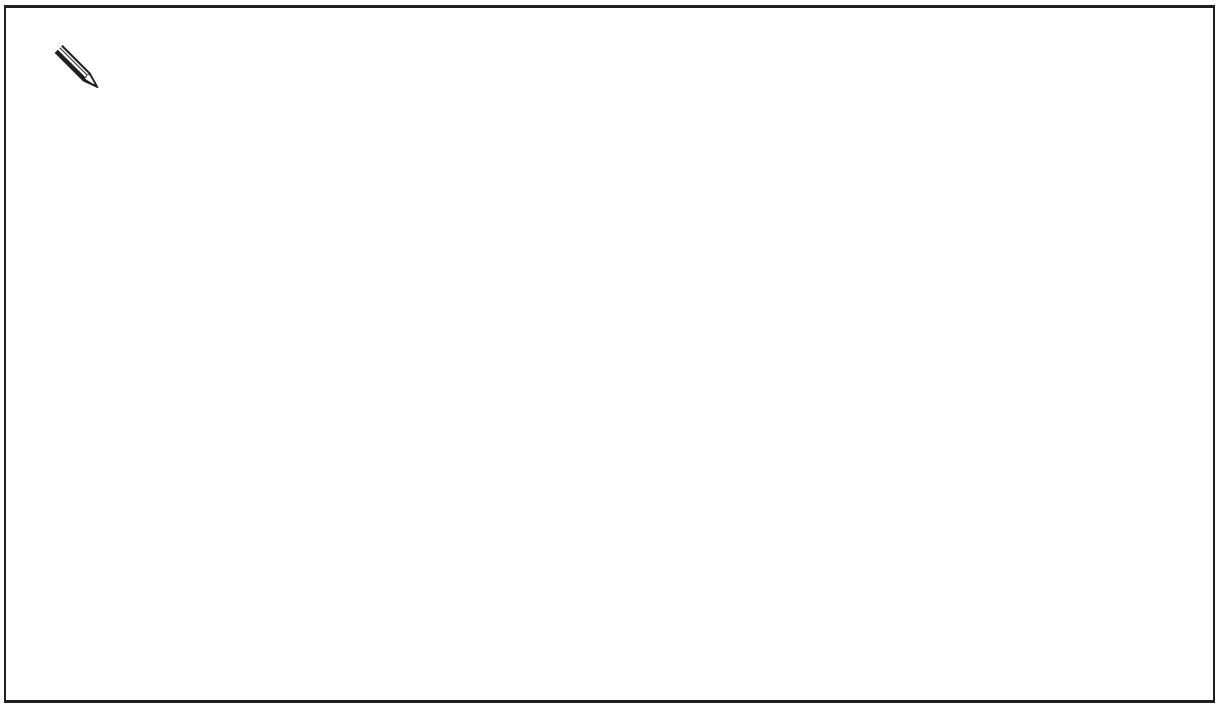
If we initially assume we know the population variance σ^2, we can standardise $\bar{x}$ as we did last semester; i.e.

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}.$$

Recall that the standard normal distribution is $Z \sim N(0, 1)$, i.e. Z has zero mean and variance (and so standard deviation) 1; also recall that approximately 95% of the standard normal distribution lies between -1.96 and 1.96, i.e.

$$\Pr(-1.96 < Z < 1.96) = 0.95.$$

This is easier to see if we draw a picture:



Since we know that $Z = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}$, we can write this as

$$\Pr\left(-1.96 < \frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} < 1.96\right) = 0.95;$$

rearranging for μ gives us an expression for the *95% confidence interval for* μ:

$$\left(\bar{x} - 1.96\sqrt{\sigma^2/n} \quad , \quad \bar{x} + 1.96\sqrt{\sigma^2/n}\right);$$

thus, we can say that the two values $\bar{x} - 1.96\sqrt{\sigma^2/n}$ and $\bar{x} + 1.96\sqrt{\sigma^2/n}$ are the lower and upper bounds (respectively) of the (95%) confidence interval.

We often write more simply as

$$\bar{x} \pm 1.96\sqrt{\sigma^2/n}.$$

Going back to *The Vintage Clothing Co.* example, this means that if we were to take 100 samples and for each one calculate a 95% confidence interval (using the formula above), then about 95 of these confidence intervals would "capture" the true population value $\mu = 320$.

Consider the following examples.

Example 1
A coffee machine fills cups with hot water; the variance of the filling process is known to be $\sigma^2 = 10$ml. A sample of 100 filled cups gives a sample mean of $\bar{x} = 40$ml. What is the 95% confidence interval of the population mean μ?

We already have a formula for the 95% confidence interval:

$$\bar{x} \pm 1.96\sqrt{\sigma^2/n}.$$

So, inputting our values, we get

$$\begin{aligned} 40 &\pm 1.96\sqrt{10/100}, \qquad \text{i.e.} \\ 40 &\pm 0.62. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is $(39.38, 40.62)$.

What would happen if the sample size increased to 200 and everything else remained the same? We'd get

$$\begin{aligned} 40 &\pm 1.96\sqrt{10/200}, \qquad \text{i.e.} \\ 40 &\pm 0.44. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is $(39.56, 40.44)$. This should be intuitive, because as the sample size increases we are becoming more sure of our estimate for the population value.

What would be the 99% confidence interval in this case? From tables for the standard Normal distribution (see semester 1), we can find that

$$\Pr(-2.58 < Z < 2.58) = 0.99;$$

hence, the 99% confidence interval is given by

$$\bar{x} \pm 2.58\sqrt{\sigma^2/n},$$

in this case giving

$$\begin{aligned} 40 &\pm 2.58\sqrt{10/200}, \qquad \text{i.e.} \\ 40 &\pm 0.58. \end{aligned}$$

Hence, the 99% confidence interval for the population mean μ is $(39.42, 40.58)$. You should note that this gives a wider range than the 95% confidence interval. This is (again) intuitive; as you increase the percentage of certainty you want, you will naturally incorporate a larger range.

Example 2
Geordie Sparkz are an electrical company based in Newcastle producing circuitboards for large plasma televisions. One of their machines punches tiny holes in these curcuitboards that should be 0.5mm in diameter. A sample of 30 circuitboards off the production line is inspected; the average diameter of the holes produced by this machine, for this sample, is 0.54mm.

Assuming the machine is set to ensure a standard deviation of $\sigma = 0.12$mm, calculate the 95% confidence interval for the population mean diameter of holes produced by this machine. Do you think there is a real problem with this machine?

...Solution to Example 2...

1.4 Exercises

1. A company packs sacks of flour. The variance of the filling process is 100g. A sample of 50 bags is taken and weighed and the resulting sample mean is 750g. Compute a 95% and 99% confidence interval for the mean weight of a bag of flour.

2. A company manufactures bolts with a process variance of 50mm. A sample of 100 bolts is taken and measured and their average length is calculated as 98mm. What is the 95% confidence interval for the mean length of bolts? If the bolts are designed to be 100mm long, is the process satisfactory?

3. Take a random sample of size $n = 5$ from the data in Table 1.1 (i.e. complete one of the tables on page 6 of these notes). Assuming we know that $\sigma = 121.9$ transactions, and that our sample comes from a Normal distribution, obtain a 95% confidence interval for the population mean number of transactions at *The Vintage Clothing Co.* Does your interval contain the true value for μ?