6 Estimation - II

Point estimators are not very useful by themselves. A single sample mean does not tell us very much about the population mean. We know that it should be close to the population mean and that larger the sample size, the more likely it is to be closer. Confidence intervals provide us a range around a point estimate that is likely to contain the population parameter that we are looking for. Let us take an example.

Consider a packaging machine that can be set to pack a certain amount of coffee per packet. The standard deviation of the machine is 1g regardless of the mean weight set. Suppose the line is set to pack 100g coffee packets. The quality assurance officer collects 64 packets during a shift. She then finds that the sample mean is 100.3 g. Can we say anything about whether on average extra coffee has been packed in the shift?

Yes we can! We know from the properties of the sample mean that it is a random variables whose expected value is the population mean, and whose variance is the population variance divided by the sample size, i.e.

$$E[\bar{X}] = \mu$$
$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

We also know using the central limit theorem that the sample mean is approximately normally distributed. As we know the population variance (square of the standard deviation), we also know the variance (and standard deviation) of the sample mean.

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{1^2}{64} = \frac{1}{64}$$
$$\sigma_{\bar{X}} = \frac{1}{\sqrt{64}} = \frac{1}{8} = 0.125$$

Now we want to use what we know about the normal distribution to construct a *confidence interval*. This means an interval which will contain the population

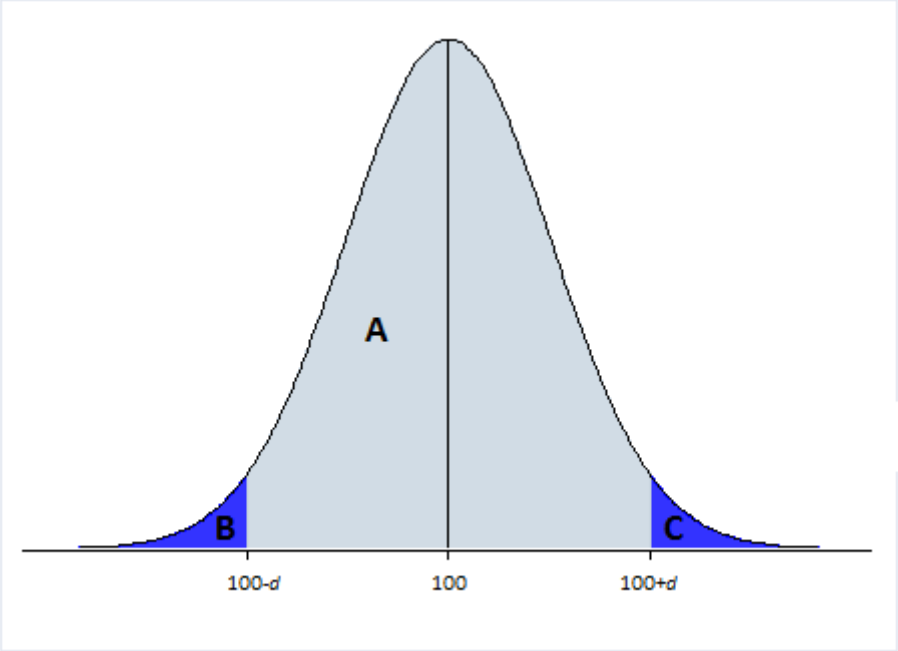


Figure 1: Distribution of the sample mean with population mean 100

mean with some probability. Let our desired probability or 'confidence' be 0.95. This means that we want to find two points a and b, $a < b$, such that there is a 95% chance this interval $[a, b]$ contains the population mean.

To do this first let us assume that the actual population mean is 100. Now we can consider the actual distribution of the sample mean, which will be a normal distribution with mean 100 and standard deviation 0.125. Using this distribution, we can construct an interval around the population mean 100, where the probability of finding the sample mean would be 0.95. This region is shown in Figure 6 as the interval between points $100 - d$ and $100 + d$, and the probability of finding the sample mean in this interval is given by the area of **A**, which would be 0.95. As is evident, we have chosen the area **A** symmetrically, i.e. with equal parts before and after 100, because we have no reason to believe that the sample mean is more or less likely to be smaller or larger than the population mean.

Now we need to find this distance d that defines this area **A**. We will use the *cdf* of the normal distribution to find this value. Note that since area of **A** is 0.95,

and **A** is symmetric about the mean 100, therefore, the areas of **B** and **C** would each be equal to 0.025. ($\because 0.95 + 0.025 + 0.025 = 1$). Hence, we can find d in two ways.

1. $P(\bar{X} \leq 100 - d) = 0.025$

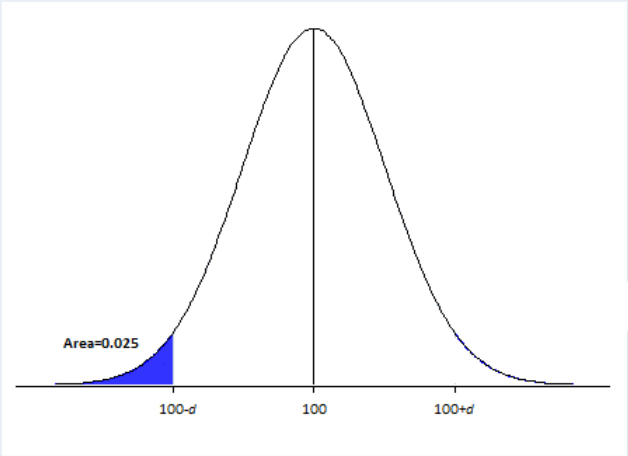


Figure 2: *cdf=0.025*

2. $P(\bar{X} \leq 100 + d) = 0.975$

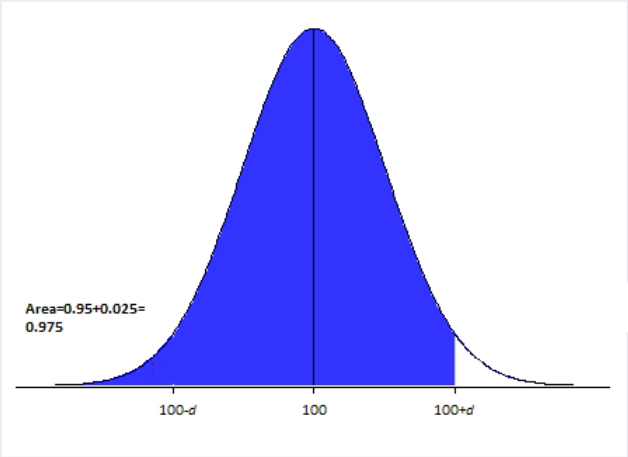


Figure 3: *cdf=0.975*

In practice we generally use the second method for convenience. We can show using simple algebra that if we write the distance d as a multiple of the standard deviation $\sigma_{\bar{X}}$ of the sample mean, i.e. $d = z\sigma_{\bar{X}}$, then we can express this *cdf* as the *cdf* of the standard normal distribution.

$$\begin{aligned} P(\bar{X} \leq 100 + d) &= P(\bar{X} - 100 \leq d) \\ &= P\left(\frac{\bar{X} - 100}{\sigma_{\bar{X}}} \leq \frac{d}{\sigma_{\bar{X}}}\right) \\ &= P\left(\frac{\bar{X} - 100}{\sigma_{\bar{X}}} \leq z\right) \end{aligned}$$

As we have seen earlier for any normal distribution X with mean μ and standard deviation σ, the random variable $Z = \dfrac{X - \mu}{\sigma}$ is a standard normal distribution, i.e. a normal distribution with mean 0 and standard deviation 1.

The *cdf* of a standard normal distribution can be read from statistical tables and is available from most statistical programs as well as on mobile apps! For our case we can see from the sample statistical table given in Figure 6

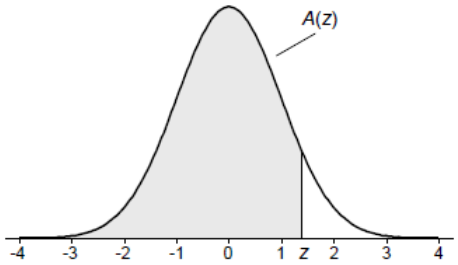
From the table we can see that $z = 1.96$. This means that if we are a distance 1.96 time the standard deviation away from the mean of a normally distributed, then the probability of the random variable taking a value beyond that is 1-0.0975=0.025. We can compare this to the Chebyshev's inequality and see that we can be much more precise in what we say because we know that the distribution is normal.

In our case then,

$$\begin{aligned} d &= z\sigma_{\bar{X}} \\ &= 1.96 * 0.125 = 0.245 \end{aligned}$$

Therefore, when the population mean is 100, there is 95% chance that the sample mean will be between 100.245 and 99.755. Clearly, the sample mean of 100.3 is outside this range. Now, for what values of the population mean would

Cumulative Standardized Normal Distribution



$A(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952

Figure 4: Snapshot of a statistical table showing the *cdf* of the standard normal distribution

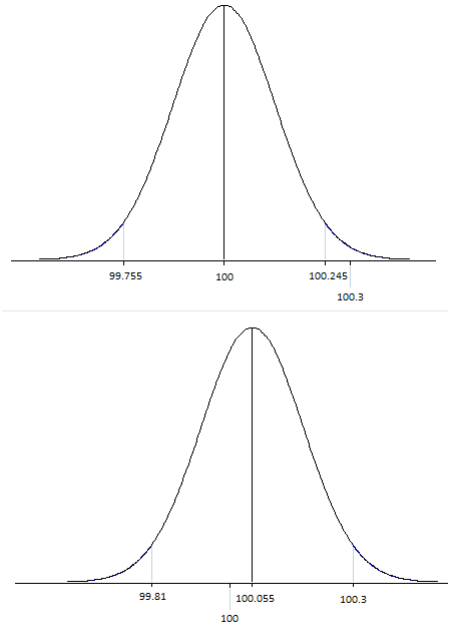


Figure 5: If the population mean was 100.055 instead of 100, then the observed sample mean 100.3 would be inside the 95% region

the sample mean of 100.3 be within this 95% range? The width range, which is $2d$ only depends on the standard deviation, with the population mean being the point it is centred around. So, if we increase the population mean from 100 to a point where it is equal to $100.3 - d = 100.055$, then the point 100.3 will be within the interval. This is shown in Figure 6. Now if we think of the highest possible value of the population mean so that 100.3 would still be in the interval, that would be $100.3 + d = 100.545$.

For the sample mean, $\bar{X}$, this interval i.e. $\bar{X} - d$ to $\bar{X} + d$, is called the 95% confidence interval. In this particular case, the observed confidence interval, based on the observed sample mean 100.3 is 100.055 to 100.545. We have used confidence intervals for population mean here, but they are a more general concept, and can be defined for any population parameter (say θ) using the corresponding estimator (say w) and the standard deviation of the estimator. The more common and useful way to think about confidence intervals is a interval around the estimator w that

will contain the population parameter of interest θ with a given probability $1-\alpha$. In our case the confidence interval $\bar{X}-d$ to $\bar{X}+d$ has a 95% chance of containing the population mean. Hence, when we observe the sample mean 100.3 and construct the confidence interval 100.055 to 100.545, we are saying that with at least 95% probability we have been packing excess coffee in the packets.

6.1 Creating confidence intervals when the population variance is unknown

In the case above, we knew the population variance and hence were able to calculate the variance of the sample mean and use it to construct the confidence interval. But when we don't know the population variance, our only measure for it is in the form of its estimator the sample variance. We can use the observed sample variance instead of the unknown population variance to calculate the confidence interval, but we will need to compensate for the fact that the population variance may be different from the observed sample variance.

We do this by using the Student's t-distribution instead of the normal distribution. Student's t-distribution is a class of distribution that is similar to the standard normal distribution, it is symmetric and is shaped like a bell-curve, but is flatter than the normal distribution. Unlike the standard normal distribution, the shape of the t-distribution depends on a parameter called the *degrees of freedom*. In our context, where we are estimating the population mean, the degrees of freedom will be one less than the sample size, i.e it will be equal to $n-1$. Intuitively, this can be understood as follows. The total degrees of freedom is the sample size n as the sample can be different in n ways. But we lose one degree of freedom when we calculate the sample mean using the sample. (The logic is similar to the reason we divide the sample variance by $n-1$ instead of n). Lower degrees of freedom imply that sample variance can be quite different from the population variance, hence the distribution is flatter to reflect this increased un-

certainty. Higher degrees of freedom implies that the sample variance is going to be fairly close to the population variance, hence the distribution is very close to a standard normal distribution.

Below we outline the procedure for finding the confidence interval with a sample of size n that has a sample mean $\bar{x}$ and sample variance s^2.

1. First decide the confidence level that we want. Let us say it is 95%.

2. Now instead of looking at the table for the standard normal distribution, we look at the table for the t-distribution with $n-1$ degrees of freedom. We find the distance t, where the *cdf* is 0.95+0.025=0.975.

3. We need to scale this distance using the standard deviation of the sample mean $\sigma_{\bar{X}}$. But we don't know the population variance σ^2, hence we don't know the variance of the sample mean. The t-distribution allows us to use the sample variance s^2 to calculate an estimate of the standard deviation of the sample mean $s_{\bar{X}}$.
$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

4. We use this estimate of the standard deviation of the sample mean to find the distance d.
$$d = ts_{\bar{X}}$$

5. Now we construct the confidence interval around the sample mean $\bar{x}$ using d. Hence, the confidence interval is $[\bar{x}-d, \bar{x}+d]$

An important point to note is that t-distribution can only be used when the population is known to be normally distributed, or at least the distribution does not significantly deviate from the normal. If n, is large, then we know that the distribution of the sample mean approximates a normal distribution regardless of the original distribution. But in those cases we mostly don't need to use the t-distribution as it also approximates the standard normal distribution when the

degrees of freedom $n-1$ is large. We mostly use t-distribution for small sample sizes and in those cases, if the original population distribution is very different from the normal then strictly speaking the t-distribution is not applicable for finding the confidence interval. In practice, we still use it assuming that the error will not be too large.