Chapter 14 : Introduction to Statistical Inference

Note : Here the 4-th and 5-th editions of the text have different chapters, but the material is the same.

Data $x_1, x_2, \ldots, x_n$ is a random sample from some population with mean μ.

We saw in Chapter 11 that the random variable $\bar{x}$ has expected value μ no matter what the value of μ is. $\bar{x}$ is an unbiased estimator of μ. It is also called a point estimator of μ.

$\bar{x}$ is a random variable and it is typically not equal to the population parameter μ. What is the real value of μ? Can we say anything about the value of μ based on a random sample of data?

Statistical Inference studies and gives methods to draw conclusions (make inferences) about a population from a sample of data.

Remark : An inference is similar to a deduction, but the conclusion is not guaranteed with certainty, as opposed to a deduction where the conclusion is guaranteed with certainty as in a mathematical deduction. Some of this distinction will be clear from our examples. Specifically inferences are based on random samples and so if an experiment is repeated the data will change and the inferences will change, but typically not in a very major or important way.

Some Very Simple Conditions for Inferences about a Population Mean

1. A simple random sample is obtained from a population. There is no non-response or other practical difficulties with the data
2. the variable we measure has exactly a normal distribution $N(\mu,\ \text{sd}\ = \sigma)$
3. we do not know μ, but we know σ.

Items 2 and 3 are typically not true. After seeing how we can make inferences about μ in this special case we will then see how to make more realistic conditions for making

inferences. In particular as in Chapter 11, we will have in place of the *exact normal distribution* for $\bar{x}$ a property of an approximate normal distribution for $\bar{x}$. For point 3, we will have another way of approximating the population mean, in particular by using the sample variance to estimate the population variance σ^2.

Confidence Interval

Here is an overview of the reasoning and the method used in inference

1. For a given value of μ, the random variable $\bar{x}$ has distribution Normal with mean μ, standard deviation $\frac{\sigma}{\sqrt{n}}$

2. For a given value of μ and using the 95% rule we have with probability 0.95 that $\bar{x}$ will fall into the interval

$$\mu - 1.96 * \frac{\sigma}{\sqrt{n}}, \mu + 1.96 * \frac{\sigma}{\sqrt{n}}$$

   Written in another fashion we have

$$\mu - 1.96 * \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96 * \frac{\sigma}{\sqrt{n}} \tag{1}$$

   with probability 0.95

   The 95% rule will replace the *critical value* 1.96 by the value 2 and so we will have

$$\mu - 2 * \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 2 * \frac{\sigma}{\sqrt{n}}$$

   with probability 0.95

3. For a given sample we observe the value of $\bar{x}$. We then use equation (1) to "solve" for μ in terms of this observed value of $\bar{x}$. This gives

$$\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}} \tag{2}$$

   or in terms of the 95% rule with 1.96 replaced by 2

$$\bar{x} - 2 * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2 * \frac{\sigma}{\sqrt{n}}$$

Equation (2) is called a 95% confidence interval, since it is based on a 0.95 probability or proportion central interval for the normal distribution of $\bar{x}$. Notice a probability or proportion of 0.95 is also the same as 95% probability.

Item or property 2 gives the values in terms of an interval that $\bar{x}$ falls into with large probability, that is the ones that are reasonably or "consistent" with the given value of μ. This is equation (1). Item or property 3 then *inverts* this statement to ask which values of μ are "consistent" with the observed value of $\bar{x}$. This is Equation (2).

Aside: For those who are interested in a little of the algebra we can see how we get equation (2) from equation (1)? For those not interested in the algebra just skip these next few lines.

Equation (1) is actually two inequalities

$$\mu - 1.96 * \frac{\sigma}{\sqrt{n}} \leq \bar{x}$$

and

$$\bar{x} \leq \mu + 1.96 * \frac{\sigma}{\sqrt{n}}$$

From the first we get, by adding $1.96 * \frac{\sigma}{\sqrt{n}}$ to both sides, we obtain

$$\mu \leq \bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}}$$

From the second we get, by subtracting $1.96 * \frac{\sigma}{\sqrt{n}}$ to both sides we obtain

$$\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}} \leq \mu$$

Putting these inequalities together two together we get equation (2).

End of Aside

For now *pretend* that the grades for test 1 followed a normal distribution with mean μ and standard deviation $\sigma = 4.04$, the actual standard deviation for the grades. The 95% confidence interval for the true population grade mean, based on a sample of size $n = 10$ is then given by (each line will follow from the previous line)

$$\begin{aligned} \bar{x} - 1.96 * \frac{4.04}{\sqrt{10}} &\leq \mu \leq \bar{x} + 1.96 * \frac{4.04}{\sqrt{10}} \\ \bar{x} - 1.96 * 1.278 &\leq \mu \leq \bar{x} + 1.96 * 1.278 \\ \bar{x} - 2.50 &\leq \mu \leq \bar{x} + 2.50 \end{aligned}$$

Notice that we can write this formula down even before taking a random sample of size $n = 10$.

Take a random sample from the test 1 population. When I did this I obtained data

$$27, 25, 16, 25, 24, 29, 27, 27, 28, 24$$

For this data the sample mean is $\bar{x} = 25.2$. Based on the normal assumptions above we have a 95% confidence interval

$$\bar{x} \pm \frac{4.04}{\sqrt{10}} = 25.2 \pm 2.50 = [22.70, 27.70]$$

Thus based on this random sample of $n = 10$ data points we have learned (actually inferred) that the actual population mean in reasonably thought to be between 22.7 and 27.7.

This calculation means that with 95% confidence the *true* population mean (which we typically do not know) is a number between 22.70 and 27.70. In particular a claim that the *true* population mean parameter value μ falls into this interval and not outside this interval; thus for example it is not reasonable (at 95% confidence level) that the value of μ is 20.

Confidence Interval

Our *estimate of an interval of reasonable or consistent values* of μ is of the form

$$\bar{x} \pm \text{ margin or error}$$

The margin of error depends on

- the sample size through $\sqrt{n}$
- the population standard deviation σ.

  Aside : when we generalize this method the dependence on the *population variance* (again typically unknown) will be through the sample standard deviation $\sqrt{s^2} = s$. Recall from our earlier discussion in Chapter 11 that the sample variance s^2 is an unbiased estimator of the *true* population variance.
- the confidence level as determined by the corresponding critical value for the sampling distribution of the estimator $\bar{x}$.

This interval is called a *confidence interval*. We can choose the probability interval (central 0.90, central 0.95 or central 0.99) and this yields the corresponding confidence interval through the relationships (1) and (2). The central 0.90 probability interval corresponds to a 90% confidence interval, a central 0.95 probability interval corresponds to a 95% confidence interval, and central 0.99 probability interval corresponds to a 99% confidence interval. A confidence interval is a *random interval*. The true value μ falls into this interval with the corresponding probability level.

$$\bar{x} - z^* \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z^* \times \frac{\sigma}{\sqrt{n}} \tag{3}$$

or in another notation

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \ .$$

The critical value z^* is chosen so the probability interval for $\bar{x}$ corresponds to the given confidence level. For the simple exact normal distribution assumptions we have the corresponding critical values

Confidence interval	Probability interval	probability in upper tail	Critical value z^*
99%	0.99	$\frac{1-.99}{2} = .005$	2.58
95%	0.95	$\frac{1-.95}{2} = .025$	1.96
90%	0.90	$\frac{1-.90}{2} = .05$	1.65

For a 95% confidence interval the true value of the parameter falls into a confidence interval with probability 0.95. Thus on average 1 out 20 confidence intervals will not contain the true value of μ. A confidence interval is a *random interval*. Similarly for a 90% (or 99%) confidence interval, on average 9 out of 10 (99 out of 100) of these intervals contains the *true* population mean value μ.

To illustrate this idea we have done a simulation experiment with $M = 20$ replicates. For each replicate a simple random sample of size $n = 10$ is taken from a normal distribution with mean $= \mu = 23.6$ and standard deviation $= \sigma = 4.04$. For each random sample the corresponding 95% confidence interval is calculated. According to the probability rules about 1 in 20 (that is 5%) of such intervals on average will NOT contain the true value of $\mu = 23.6$. This is shown in Figure 1. Each line in this plot is one of the confidence intervals. A centre dashed vertical line corresponding to $\mu = 23.6$. In this plot all the confidence intervals overlap the value $\mu = 23.6$, except for 1 interval.

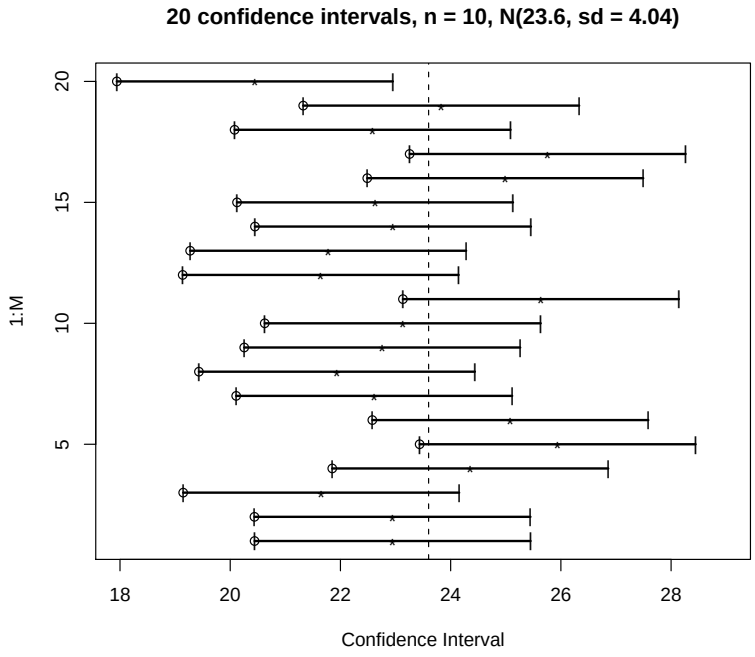

Figure 1: $M = 20$ random confidence intervals

How can we obtain a more precise estimate of the *true* value of the population mean? In terms of our confidence interval our *estimate* of the values of μ that are consistent with the observed data is of the form

$$\bar{x} \pm \text{ margin or error}$$

or more precisely

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \ .$$

Thus we can make the estimate more precise by

- use a smaller value of z^*, which means a lower confidence level, and hence less likely to contain the true value of μ
- use a larger value for n, that is increase the sample size.

  This will be more expensive, so it may not be possible.

  The increase in precision is proportional to 1 over $\sqrt{n}$. Thus to make the confidence interval one half as long requires that $\sqrt{n}$ gets changed to $2\sqrt{n} = \sqrt{4n}$, so that 4 imtes as many data points are required.
- make the population variance smaller.

  This typically cannot be done, as we are working with the population that is given, and the random sample that we obtain from it. However sometimes an experimental design such as matched pairs or a paired design will allow us to obtain data with the given population mean but smaller variability. Recall for example the shock absorbers example where this is possible.

How can we guarantee that our confidence interval contains the true value? If we use a 100% confidence interval then $z^* = \infty$ and our confidence interval is $\bar{x} \pm \infty$, or every possible value of μ. This interval is of course useful in helping us to learn or understand what value of μ are reasonable and which values of μ are not reasonable.

How many observations should we take? Ideally we want to take as many as possible. However in many cases it costs resources (typical scientific or engineering experiment, experimental units), time (all types of studies) and money (typically all types of experiments : lab assistants, poll questioners). Thus for practical considerations one cannot take arbitrarily large samples.

On the other hand we might need to obtain a certain degree of precision. For example an opinion poll might want to measure the proportion of voters who favour the ruling party, but it is sufficient to know this to a margin or precision of plus or minus 3 percentage points. In using a drug to control blood pressure we might want to know the blood pressure to a precision of 5 units.

We can now translate this question into the following : for a given precision m at say 95% confidence level, how big should the sample size n so that

$$m = z^* \times \frac{\sigma}{\sqrt{n}}$$

Where does this come from? The confidence interval form is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm m$$

and so we match up m and $z^* \frac{\sigma}{\sqrt{n}}$. Since n is the only unknown we the solve for n, yielding

$$n = \frac{(z^*)^2 \sigma^2}{m^2} = \left(\frac{z^* \sigma}{m} \right)^2 .$$

Since we can only take whole numbers (integers) of observations (how can one take .6 of an observation?) we will then take n to be the value of the right hand side, but rounded up to the next integer.

Consider the test 1 grades example again. For the purpose of this calculation we pretend the distribution of grades is normal and that the population size is very large.

Here we have the population standard deviation $\sigma = 4.04$. How big should the sample size be to have precision $m = 3$, that is our 95% confidence interval will be $\bar{x} \pm 3$.

We will need to take

$$\begin{aligned} n &= \left(\frac{z^*\sigma}{m}\right)^2 \\ &= \left(\frac{4.04 \times 1.96}{m}\right)^2 \\ &= \left(\frac{4.04 * 1.96}{3}\right)^2 \\ &= 2.64^2 = 6.97 \end{aligned}$$

or more specifically since n is an integer we take $n = 7$ by rounding up. We would have rounded up even if the calculated of the expression were 6.01. That is because we need an integer or whole number of observations and it has to be at least 6.01 (bigger than 6.01).

For different value of precision, again at confidence level 95%, we have

m	$\left(\frac{z^*\sigma}{m}\right)^2$	n
3.0	7.0	7
2.0	15.7	16
1.0	62.7	63
0.5	250.8	251

Aside : Comment on Opinion Polls

It is for this reason that opinion polls take a random sample of approximately 1600 individuals. This will result in a 95% confidence interval of a population proportion (which is the same as a sample mean of "success" and "failure" counts) which is of the form

$$\hat{p} \pm .03$$

This is often reported as a margin of error of 3% 19 times out of 20.

Recall the beginning of our discussion of confidence intervals. We had some Very Simple Conditions for Inferences about a Population Mean, which are given here again.

1. A simple random sample is obtained from a population. There is no non-response or other practical difficulties with the data
2. the variable we measure has exactly a normal distribution $N(\mu,\ \text{sd}\ = \sigma)$
3. we do not know μ, but we know σ.

Suppose that instead of property 3 we do not known σ. This is much more realistic. What can we do now?

Recall also that be based our confidence interval on the following idea. For a given value of μ and using the 95% rule we have with probability 0.95 that $\bar{x}$ will fall into the interval

$$\mu - 1.96 * \frac{\sigma}{\sqrt{n}}, \mu + 1.96 * \frac{\sigma}{\sqrt{n}}$$

This used the property that $\bar{x}$ has exactly a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, or equivalently

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

Here $\sim$ is a shorthand for saying "distributed as".

When σ is not known we can use in place of σ the sample variance $s = \sqrt{s^2}$, where s^2 is the sample variance. However the random variable

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

no longer has a standard normal distribution, but instead a distribution called the Student's t distribution with degrees of freedom $n-1$. The $n-1$ is related to the divisor $n-1$ in the formula for the sample variance. The critical values for the Student's t distribution are given in Table C near the end of the text. We discuss later how the corresponding confidence interval gets changed.