

Chapter 6

Introduction to Inference

This chapter discusses procedures for finding confidence intervals and carrying out significance tests for the mean of a population. The methods require use of the normal distribution and hence are applicable only when the underlying population may be assumed to be approximately normal or when the sample size is so large that the normal approximation given by the central limit theorem may be invoked.

6.1 Estimating with Confidence

The data $\{x_1, \dots, x_n\}$ are assumed to come from a $N(\mu, \sigma)$ population with mean μ and a known standard deviation σ . A level C confidence interval for the mean μ is given by

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

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean, n is the sample size, and z^* is the value such that the area between $-z^*$ and z^* under a standard normal curve equals C .

The Excel function required is `=CONFIDENCE(α, σ, n)` which calculates the margin of error (half width of interval) associated with a $C = 1 - \alpha$ level confidence interval for the mean of a normal distribution with standard deviation σ based on a sample of size n , that is $z^* \frac{\sigma}{\sqrt{n}}$

It is also possible to calculate each component in the confidence interval directly. This more complicated approach is useful if intermediate results such as margin of error are needed. The formula for the normal critical value z^* is `NORMSINV(p)`, which returns the inverse of the standard normal cumulative distribution function $\Phi^{-1}(p)$, where $0 < p < 1$. In the first example below a confidence interval is constructed both ways.

Example 6.1. (Example 6.2, page 422 in the text.) Tim Kelley has been weighing himself once a week for several years. Last month his four measurements (in pounds) were

190.5 189.0 195.5 187.0

	A	B	C	D
1	Confidence Interval for a Normal Mean Using Z			
2				
3		values	formulas	
4	User Input			
5	sigma	3		Data
6	conf	0.90		190.5
7	Summary Statistics			189.0
8	n	4	=COUNT(Data)	195.5
9	xbar	190.5	=AVERAGE(Data)	187.0
10	Calculations			
11	SE	1.500	=sigma/SQRT(n)	
12	z	1.64	=NORMSINV(0.5+conf/2)	
13	ME	2.47	=z*SE	
14	Excel ME	2.47	=CONFIDENCE(1-conf,sigma,n)	
15	Confidence Limits			
16	lower	188.03	=xbar-ME	
17	upper	192.97	=xbar+ME	

Figure 6.1: Confidence Interval for a Normal Mean

Give a 95% confidence interval for his mean weight last month.

Solution. In Figure 6.1 the Excel formulas required are given in column C. These are entered into the adjacent cells in column B to create the workbook template to solve this problem. The Excel output is in column B. The user inputs required are the standard deviation and the confidence level. If the data have already been summarized, then you can enter the values for the sample size n and the average in cells B8:B9, respectively. Otherwise Excel will read the data and calculate n and \bar{x} . The data can be located in a convenient place on the same sheet or it can be located on another sheet. The latter is particularly useful for large data sets. In this example, with only four data points, we have recorded them on the same sheet as the calculations.

The following steps describe how to construct the workbook.

1. Enter the labels as shown in column A.
2. **Name** the cell ranges to be used. Select cells A5:B6, A8:B9, A11:B13 and A16:B17. To select **noncontiguous blocks** of cells, make the first selection A5:B6, then hold down the **Control (Windows)** or **Command (Macintosh)** key while selecting the other ranges. From the Menu Bar choose **Name – Create**, select **Left Column**, and then click OK. Next, select D5:D9, and from the Menu Bar choose **Name – Create**, select **Top Row**, and then click OK to name the data range.
3. Enter the formulas shown Figure 6.1 into columns B8:B9, B11:B14, and B16:B17, and enter the data in D5:D9. Since you have named the data range you can refer to the cells D5:D9 as “Data,” for instance as in the formula = COUNT(Data). Otherwise, you would type = COUNT(D5:D9) giving the actual

locations. These formulas are sufficiently simple that you can enter them by hand rather than use the Formula Palette or the Function Wizard.

The only input needed once the workbook has been constructed are the population standard deviation (σ) and the confidence level. Type “3” and “0.90” into cells B5 and B6, respectively. The results are immediately recorded in cells B16:B17, showing a lower confidence limit 188.03 and an upper confidence limit 192.97 for the population mean μ .

Explanation

The formula = COUNT(Data) gives the sample size by counting the number of cells named by the variable Data. You could also type the integer “4” instead. Likewise, = AVERAGE(Data) is the Excel formula for the sample mean \bar{x} . For comparison we have also provided the corresponding formula = CONFIDENCE(α, σ, n) in cell C14.

How Confidence Intervals Behave

A confidence interval is a random interval that has a specified probability of containing an unknown parameter. Thus, a 90% confidence interval for a population mean has probability 0.90 of containing the mean. So, in repeated confidence intervals, in the long run approximately 90% of these confidence intervals would contain the population mean.

Example 6.2. Take 100 SRS of size 3 from an $N(3.0, 0.2)$ population and construct a 90% confidence interval for the mean. Count how many times the confidence interval contains the mean 3.0.

Solution

- Following the instructions given in Example 4.6 for simulating samples from a specified distribution, choose **Tools – Data Analysis – Random Number Generation** from the Menu Bar, complete a box like the one shown in Figure 4.8, but for normal not discrete random numbers, with “3” for the **Number of Variables**, “100” for the **Number of Random Numbers**, “3.0” for the **Mean**, “0.2” for the **Standard Deviation**, and choose a convenient range for the output. We have selected the range A8:C107.
- In cell E8 enter

$$= \text{AVERAGE}(A8:C8) - \text{NORMSINV}(0.5+0.9/2)*0.2/\text{SQRT}(3)$$
In cell F8 enter

$$= \text{AVERAGE}(A8:C8) + \text{NORMSINV}(0.5+0.9/2)*0.2/\text{SQRT}(3)$$
- Select cells E8:F8, click the fill handle and drag the contents to F107. The cells in column F will contain the value 1 if the confidence interval for the data in the corresponding row contains the true value 3.0, while the cells will contain 0 otherwise.

4. Count the number of times 1 appears by entering = SUM(G8:G107) in an empty cell (H8, for example).

Figure 6.2 shows a portion of a workbook with the simulation for which 92 times out of 100 the true mean was within the 90% confidence limits.

	A	B	C	D	E	F	G	H
1	Behavior of Repeated Confidence Intervals							
2								
3	lower	= AVERAGE(A8:C8) - NORMSINV(0.5+0.90/2)*0.2/SQRT(3)						
4	upper	= AVERAGE(A8:C8) + NORMSINV(0.5+0.9/2)*0.2/SQRT(3)						
5	G8	=IF(AND(E8<3, 3<F8), 1,0)						
6								
7					lower	upper		
8	3.1772	2.7218	3.3097		2.880	3.259	1	92
9	3.0863	3.0417	2.8220		2.793	3.173	1	
10	2.8207	2.8480	2.9353		2.678	3.058	1	
11	2.9131	3.2380	3.0292		2.870	3.250	1	
12	3.0904	3.1497	2.9295		2.867	3.246	1	
13	2.8767	3.0868	3.2555		2.883	3.263	1	
14	2.8937	2.7254	2.9995		2.683	3.063	1	
15	3.1976	3.0303	2.8750		2.844	3.224	1	
16	2.8378	2.9206	2.7565		2.648	3.028	1	
17	2.7972	2.9133	3.1956		2.779	3.159	1	
18	2.5773	3.2215	3.0810		2.770	3.150	1	
19	3.1855	2.6901	3.0221		2.776	3.156	1	
20	3.1750	3.1092	3.0020		2.905	3.285	1	

Figure 6.2: Repeated Confidence Intervals

6.2 Tests of Significance

Significance tests are used to judge whether a specified (null) hypothesis is consistent with a data set.

We create a workbook for testing the null hypothesis $H_0 : \mu = \mu_0$ for a specified null value μ_0 against one-sided or two-sided alternatives. The data $\{x_1, x_2, \dots, x_n\}$ are assumed to come from an $N(\mu, \sigma)$ population where σ is known. The same procedure can also be used to carry out a large sample test. In the workbook in Figure 6.3, the user can either test at a specified level of significance or determine a P -value.

The user inputs are the sample size, sample mean, standard deviation (which may be input as values, as formulas, or as named references depending on the context), null hypothesis, and level of significance.

Example 6.3. (Example 6.16, page 450 in the text.) Bottles of a popular cola drink are supposed to contain 300 ml of cola. There is some variation from bottle to bottle because the filling machinery is not precise. The distribution of the contents is normal with standard deviation $\sigma = 3$ ml. A student who suspects that the bottle is underfilling measures the contents of six bottles. The results are 299.4, 297.7, 310.0, 298.9, 300.2, 297.0. Is this convincing evidence that the mean content of cola bottles is less than the advertised 300 ml?

	A	B	C	D
1	Z Test for a Normal Mean - Values and Formulas			
2				
3		values	formulas	
4	User Input			
5	sigma	3.0		Data
6	null	300.0		299.4
7	alpha	0.05		297.7
8				301.0
9	Summary Statistics			298.9
10	n	6	=COUNT(Data)	300.2
11	xbar	299.03	=AVERAGE(Data)	297.0
12	Calculations			
13	SE	1.225	=sigma/SQRT(n)	
14	z	-0.789	=(xbar-Null)/SE	
15	Lower Test			
16	lower_z	-1.645	=NORMSINV(alpha)	
17	Decision	Do Not Reject H0	=IF(z<lower_z,"Reject H0","Do Not Reject H0")	
18	Pvalue	0.215	=NORMSDIST(z)	
19	Upper Test			
20	upper_z		=-NORMSINV(alpha)	
21	Decision		=IF(z>upper_z,"Reject H0","Do Not Reject H0")	
22	Pvalue		=1-NORMSDIST(z)	
23	Two-Sided Test			
24	two_z		=ABS(NORMSINV(alpha/2))	
25	Decision		=IF(ABS(z)>two_z,"Reject H0","Do Not Reject H0")	
26	Pvalue		=2*(1-NORMSDIST(ABS(z)))	

Figure 6.3: Significance Test for a Normal Mean

Solution. The workbook template in Figure 6.3 shows all formulas required in column C for any type of alternative: lower, upper, and two-sided tests, respectively. These go into the adjacent cells of column B. Then enter the values $\sigma = 3$, $\alpha = 0.05$, and null hypothesis = 300.0. Here the alternative is $H_a : \mu < \mu_0$ and only the values for the lower test are therefore shown in column B. The critical value at the 5% level of significance is $-z^* = -1.645$ in cell B16 and since the computed z in B14 is -0.789 we do not reject H_0 . Note that the P -value 0.215 is also given in B18.

Explanation

We encountered the function `NORMSINV` previously. The formula `= NORMSDIST` returns the cumulative normal distribution function $\Phi(z)$. For a one-sided lower test, the P -value is the area to the left of the computed z score $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ and is thus given by `= NORMSDIST(z)`. For an upper test, the P -value is the area to the right of $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ and is given by `= 1 - NORMSDIST(z)`. For a two-sided test, the formula for the P -value is `= 2*(1 - NORMSDIST(|z|))`. The formula `= ABS(z)` returns the absolute value of z . The decision rule uses the logical `= IF(statement, true, false)`, which returns the string designated as true if the statement is true or else it returns false.