Data Visualisation

Chapter 4

Prediction & Confidence Intervals

Dr. Tony Lyons

Autumn Semester 2019

B Sc (Hons) Multimedia Applications Development

B Sc (Hons) Creative Computing

Waterford Institute of Technology

Contents

1 Introduction

Confidence intervals are used to make predictions about various populations by collecting data on smaller samples from those populations. These predictions are not made with any specific level of confidence, rather they are made within a range of confidence levels, the is to say in a confidence interval. In many cases these confidence intervals are calculated using **the normal distribution.**

1.1 The normal distribution

The normal distribution is the graph of the function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \text{ for } -\infty < x < \infty,$$

which is shown in Figure 1 Associated with every normal distribution is a **mean** μ ("mu") and **standard deviation** σ ("sigma"). The shaded area

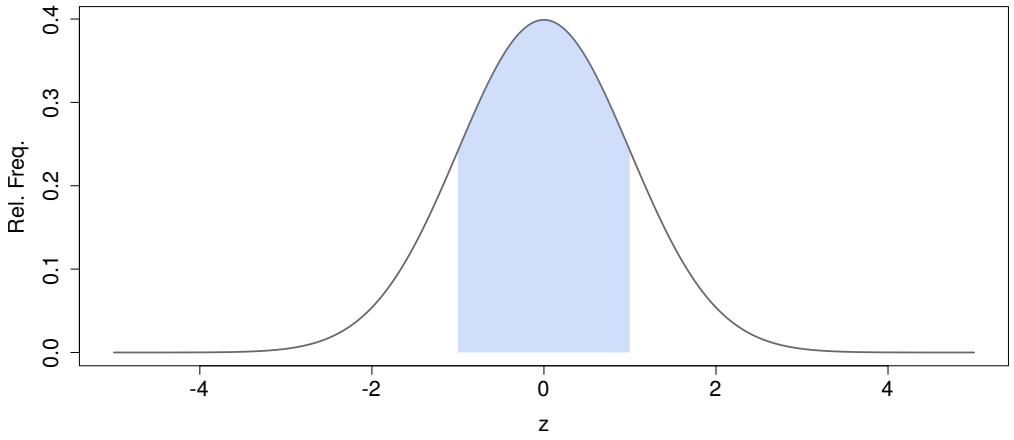

Figure 1: The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

beneath the curve lies between $z = \mu \pm \sigma$ (i.e. is one standard deviation either side of the mean) and covers 68% of the area beneath the curve. This is always the case for every normal distribution, regardless of the values of μ and σ.

1.2 Prediction intervals and z-scores

1.2.1 Confidence parameters

When we calculate a confidence interval, we do so to a certain confidence level x with $0 < x < 1$. Associated with this is the **confidence parameter**

$$\alpha = 1 - x.$$

Example 1: Confidence parameter

Evaluate the confidence parameter corresponding to a confidence level of 95%.

Solution. The confidence level 90% in decimal form is

$$x = \frac{90}{100} = 0.9.$$

This means the confidence parameter corresponding to a 95% level is

$$\alpha = 1 - 0.9 = 0.1.$$

1.2.2 z-scores

The z-score corresponding to the confidence parameter $\alpha = 0.1$ is the value $z_{\frac{\alpha}{2}}$, so that 90% of the area beneath the normal distribution lies between $\pm z_{\frac{\alpha}{2}}$. In Figure 2 the z-scores for the 90% confidence interval is shown in red. This means 90% of the area is shaded blue, while the remaining 10% is split evenly between the left and right tails.

In practice, when we seek a z-score, we calculate it from the normal distribution tables which gives an area corresponding to the area in the left tail, that is to say, the z-score which gives an area 0.95. This z-score is approximately

$$z_{0.05} = 1.645,$$

with $-z_{0.05}$ corresponding to the red line in Figure 4.

We can now use this z-score to calculate 90% confidence intervals for various populations, even those with mean and standard deviation other than 0 and 1 respectively.

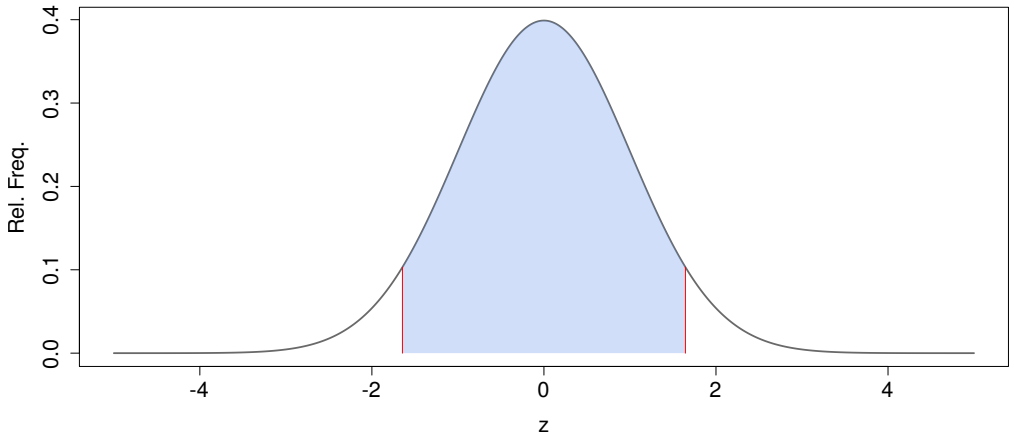

Figure 2: The $z_{\frac{\alpha}{2}}$ boundaries for the 95% confidence interval.

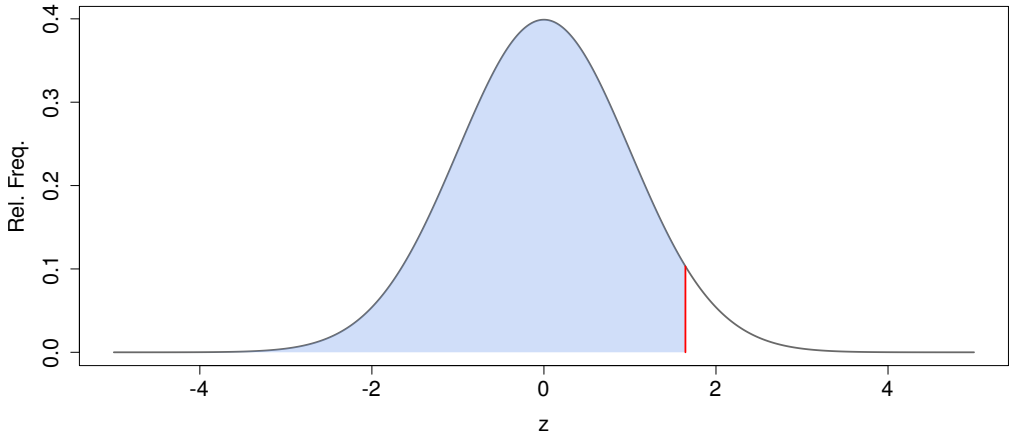

Figure 3: The shaded region of area 0.95 with corresponding z-score $z_{0.05} = 1.645$.

1.3 Prediction intervals in general

In practice it is very rare the data we work with belong to a normal distribution with $\mu = 0$ and $\sigma = 1$. If x is a data value from a normal distribution

with mean μ and standard deviation σ, the z-scores for the confidence intervals of x, can be calculated from the z-score in the normal distribution with $\mu = 0$ and $\sigma = 1$. To do so we define the **z-score** associated with x, given by

$$z = \frac{x - \mu}{\sigma}. \tag{Z Score}$$

The variable z is normally distribution with **mean** $\mu = 0$ and **standard deviation** $\sigma = 1$.

Example 2: Confidence interval

A sample is taken with sample mean $\mu = 152$ and standard deviation $\sigma = 4.5$. Find the 95% confidence interval for this data set.

Solution.

(1) The confidence parameter associated with the 95% confidence level is

$$\alpha = 1 - \frac{95}{100} = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025.$$

(2) The confidence interval for the normal distribution with $\mu = 0$ and $\sigma = 1$ is the z-value which gives an area of 0.975 beneath the normal curve, and its numerical values is

$$z_{0.025} = 1.96,$$

approximately.

(3) Transposing the formula (Z Score) we have

$$x = \mu + \sigma z \tag{1.1}$$

and so we find

$$x_{0.025} = 152 + 4.5(1.96) = 160.82. \tag{1.2}$$

This means the 95% confidence interval is

$$160.82 - 152 = 8.82$$

either side of the mean. Hence the 95% confidence interval for the data set is

$$\text{CI} = [143.81, 160.82]. \tag{1.3}$$

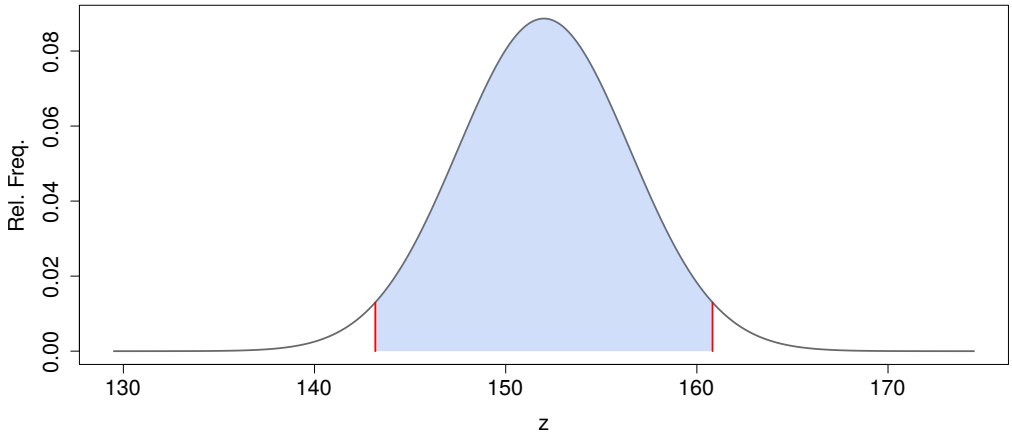

Figure 4: The shaded area is the probability that x is in the confidence interval.

2 Confidence intervals for samples

In many cases, we do not measure the mean value of some parameter for an entire population, instead, we collect data on a subset of the population called a **sample**. Then by evaluating properties of the sample we may infer the corresponding values for the entire population, to within a certain confidence interval. To do so, we introduce the **standard error** of a sample of size n as

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} \qquad \text{(Standard Error)}$$

where σ is the standard deviation of the sample. The standard error is then used to form the **confidence interval** for the mean of the entire population (μ) based on a sample mean $\bar{x}$ according to

$$\mathrm{CI} = \bar{x} \pm z_{\frac{\alpha}{2}} \mathrm{SE} = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Example 3: Sample mean and population mean

The monthly salaries of **ten** people in a company were measures, and the following data collected

€23790	€48210	€75860	27650	€110350
€72510	€33200	€41000	€29110	€22500

The data above has a **mean** €48418 and **standard deviation** €1520. Using this, estimate the mean yearly salary at the company, to a confidence level of 95% and to a confidence level of 99%.

Solution. The **95% confidence parameter is given by**

$$\alpha = 1 - \frac{95}{100} = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025.$$

This means we first need to find the z-score corresponding to a probability (or area) of

$$P = 0.95 + 0.025 = 0.975.$$

This is found from the tables (or using R) to be

$$z_{0.975} = 1.96.$$

So the z-score corresponding to the 95% confidence for the mean company salary lies between ± 1.96. We infer that the 95% confidence interval for the average salary in the company is

$$\bar{x} \pm z_{0.975}\frac{\sigma}{\sqrt{n}} = 48418 \pm 1.96\frac{1520}{\sqrt{10}}$$

and so we are 95% confident the average yearly salary at the company is within the range

$$\mu \in [€47475.89, €49360.11]\,.$$

The **99% confidence parameter is given by**

$$\alpha = 1 - \frac{99}{100} = 0.01 \Rightarrow \frac{\alpha}{2} = 0.005.$$

This means we first need to find the z-score corresponding to a probability (or area) of

$$P = 0.99 + 0.005 = 0.995.$$

This is found from the tables (or using R) to be

$$z_{0.995} = 2.58.$$

So the z-score corresponding to the 99% confidence for the mean company salary lies between ± 2.58. We infer that the 99% confidence interval for the average salary in the company is

$$\bar{x} \pm z_{0.995}\frac{\sigma}{\sqrt{n}} = 48418 \pm 258\frac{1520}{\sqrt{10}}$$

and so we are 99% confident the average yearly salary at the company is within the range

$$\mu \in [€47177.88, €49658.12]\,.$$

Notice that the 99% confidence interval is wider than the 95% confidence interval. This means if we want to be more confident that the mean salary lies within a certain interval, then the wider we must make that interval. This is a general feature of normal distributions.

Interpretation of Confidence Intervals

- We generate 20 random samples with 5 sample values each, from population data following a normal distribution with

$$\mu = 2.45 \qquad \sigma = 4.3.$$

```
N=20
n=5
set.seed(203)
Samples=replicate(N,rnorm(n, mean=2.45,sd=4.3))
Samples
```

```
##           [,1]      [,2]        [,3]     [,4]      [,5]       [,6]
## [1,] -1.644400  2.121679  1.19194033 1.585888 2.5484803 -1.3599034
## [2,]  2.056893  6.948298  9.68120673 4.697561 6.3796938  6.6468498
## [3,]  1.589452 -1.372787 -1.68465210 4.658463 1.0433567 -5.5577432
## [4,]  4.460478  5.616777 -0.08352911 7.414669 2.3629508 -0.8533518
## [5,] -2.860914 -1.019045  2.02500289 7.110482 0.8638923  5.1960145
##             [,7]      [,8]       [,9]     [,10]     [,11]     [,12]
## [1,] -3.88682016 12.503633  8.3752639 -3.972467  6.900888  8.827341
## [2,] -0.03719397 -5.218122  7.5407216 -1.852655 -2.707283 11.251193
## [3,]  0.52488820  3.427453 -2.6668947  4.349549  3.806416  3.150171
## [4,]  2.03300531  6.342675  0.6888546 -3.447608  6.310154  0.564886
## [5,]  8.02475970  6.772056  0.6699252  1.004934  7.261719  7.409925
##            [,13]     [,14]     [,15]     [,16]     [,17]      [,18]
## [1,] -2.04389281 -1.737172  3.508731 -5.754591  5.845637  4.7497142
## [2,]  1.58707827  3.583186  1.024592  1.989778 -6.233986 10.0079044
## [3,]  6.67788884  6.158757 -1.880273  2.827771  6.043789  7.6134616
## [4,]  2.74362161  8.893343  8.277448  3.806495  7.677186  2.6913766
## [5,]  0.06111038  8.160548  1.076858 -2.617160  4.485949  0.4161922
##          [,19]     [,20]
## [1,] 11.973866 -1.550063
## [2,]  1.642661  4.110023
## [3,]  5.245932 10.074673
## [4,] -4.152731  6.965926
## [5,]  1.776122  4.811878
```

- The mean of each sample is given by

```
M <- numeric(length = length(1:N))
for (i in 1:N) {
  M[i] <- mean(Samples[,i])
}
M
```

```
##  [1]  0.72030179  2.45898455  2.22599375  5.09341254  2.63967479
##  [6]  0.81437314  1.33172782  4.76553902  2.92157411 -0.78364947
## [11]  4.31437876  6.24070317  1.80516126  5.01173259  2.40147102
## [16]  0.05045878  3.56371490  5.09572980  3.29717002  4.88248740
```

- We now calculate the 90% confidence intervals for the true mean ($\mu = 2.45$) from these sample means:
  1. The z-score for this confidence interval is

```
z<-qnorm(0.95,mean=0,sd=1)
z
```

```
## [1] 1.644854
```

2. The standard error is

```
sigma=4.3
SE<-sigma/sqrt(n)
SE
```

```
## [1] 1.923018
```

3. The lower limits of the confidence intervals are

```
Lower<-M-z*SE
Lower
```

```
##  [1] -2.4427821 -0.7040993 -0.9370901  1.9303287 -0.5234091 -2.3487108
##  [7] -1.8313561  1.6024551 -0.2415098 -3.9467334  1.1512949  3.0776193
## [13] -1.3579226  1.8486487 -0.7616129 -3.1126251  0.4006310  1.9326459
## [19]  0.1340861  1.7194035
```

4. The upper limits of the confidence intervals are

```
Upper<-M+z*SE
Upper
```

```
##  [1] 3.883386 5.622068 5.389078 8.256496 5.802759 3.977457 4.494812
##  [8] 7.928623 6.084658 2.379434 7.477463 9.403787 4.968245 8.174816
## [15] 5.564555 3.213543 6.726799 8.258814 6.460254 8.045571
```

5. The 90% confidence intervals for the true mean, obtained from these samples are

```
paste(Lower,Upper,sep=' to ')
```

```
##  [1] "-2.44278210003587 to 3.88338567935312"
##  [2] "-0.704099343028834 to 5.62206843636015"
##  [3] "-0.937090142237961 to 5.38907763715102"
##  [4] "1.93032865475635 to 8.25649643414533"
##  [5] "-0.523409100036313 to 5.80275867935267"
##  [6] "-2.34871075034187 to 3.97745702904711"
##  [7] "-1.83135607356881 to 4.49481170582017"
##  [8] "1.60245512616353 to 7.92862290555251"
##  [9] "-0.241509781844536 to 6.08465799754445"
## [10] "-3.9467333617393 to 2.37943441764968"
## [11] "1.15129486766207 to 7.47746264705105"
## [12] "3.0776192778049 to 9.40378705719388"
## [13] "-1.35792263196282 to 4.96824514742617"
## [14] "1.84864870130381 to 8.17481648069279"
## [15] "-0.761612873535727 to 5.56455490585325"
## [16] "-3.11262511316115 to 3.21354266622783"
## [17] "0.400631011919433 to 6.72679879130841"
## [18] "1.93264591481243 to 8.25881369420141"
## [19] "0.134086126801717 to 6.4602539061907"
## [20] "1.71940351158355 to 8.04557129097253"
```

- These are the 90% Confidence Intervals for each sample of size 5.
- Notice that **most** but **not all** of these confidence intervals contain the true mean of the full population.

- In fact, approximately 90% of these confidence intervals contain the true mean.
- This gives us the correct interpretation for the 90% confidence interval.