7.1 Overlaying densities on histograms

1. Recall in Lab Week 4, that we considered the Weibull distribution as a probability model for time between earthquakes at a location in Turkey. Specifically, we considered $T \sim \text{WEI}(281.1, 0.657)$ so that the **scale** parameter, α, was 281.1 and the **shape** parameter, β, was 0.657. We can randomly generate, for example, 1000 observations from this Weibull distribution in R by using the command

```
times <- rweibull(1000, shape = 0.657, scale = 281.1)
```

which will store the 1000 generated values in the vector `times`.

(a) Randomly generate 10,000 observations from the $\text{WEI}(281.1, 0.657)$ distribution and store these values in the vector called `times`.

(b) Copy and paste the below command into your R script and run it.

```
hist(times, freq = FALSE, breaks = 100, xlim = c(0, 4000))
```

Note that:

- We are setting `breaks = 100` to divide the entire range of the generated values into approximately 100 bins.
- We are also setting `xlim = c(0, 4000)` so that we just plot the bins and data less than 4000 to get a better look at where most of the data lies.
- We are setting `freq = FALSE` so that the histogram is scaled so that the total area in the columns is equal to one. This is with respect to all bins, and not just those that we are looking at by restricting `xlim = c(0, 4000)`.

(c) Now overlay the corresponding Weibull density using the below command.

```
curve(dweibull(x, shape = 0.657, scale = 281.1), from = 0,
  add = TRUE, lwd = 2, col = "red")
```

What do you see? Do the histogram and density look similar? Try running the code again with different sample sizes to see what happens.

2. Repeat Question 1, but this time take a look at the normal distribution. The R function `rnorm` can be used to simulate data from normal distributions. Try varying sample sizes and choices for μ and σ.

7.2 Understanding standard deviation and mean with histograms

Before we look at some more histograms, we will consider a probability result for the normal distribution.

3. Assume $X \sim N(\mu, \sigma^2)$.

(a) For each of the choices of μ and σ below, use the `pnorm` function to calculate $P(\mu - 2\sigma < X < \mu + 2\sigma)$.

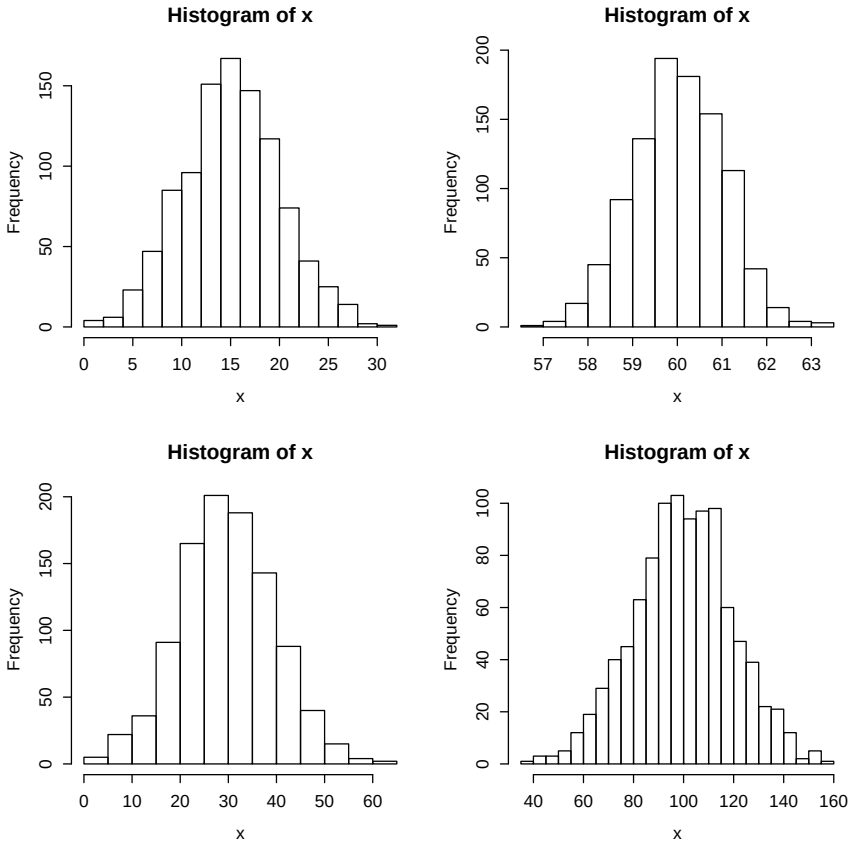
i. $\mu = 10$ and $\sigma = 2$.

ii. $\mu = 100$ and $\sigma = 15$.

(b) What did you find above? In fact, the result will be the same for any μ and σ.

(c) What is the probability that X is within 2 standard deviations of the mean?

(d) If you were to randomly sample many observations from the $N(\mu, \sigma^2)$ distribution, what proportion do you think will fall within two standard deviations of the mean?



4. In the histograms above, $n = 1000$ observations have been randomly generated from a $N(\mu, \sigma^2)$ distribution for different choices of μ and σ. In no particular order, the μ parameters used for each are one of $30, 60, 15, 100$ and the σ parameter for each are one of $20, 1, 5, 10$. Keeping in mind that your answer to 3d should have revealed that approximately 95% of observations will be generated within two standard deviations of the mean, match the μ and σ parameters that must have been used to generate the sample data used for each histogram.

   Top-left histogram: $\mu = \ldots\ldots$ and $\sigma = \ldots\ldots$

   Top-right histogram: $\mu = \ldots\ldots$ and $\sigma = \ldots\ldots$

   Bottom-left histogram: $\mu = \ldots\ldots$ and $\sigma = \ldots\ldots$

   Bottom-right histogram: $\mu = \ldots\ldots$ and $\sigma = \ldots\ldots$

7.3 Estimating the mean

5. Consider a random sample denoted $X_1, X_2, \ldots, X_n$ to be sampled from a population with unknown mean μ.
   (a) Show that $E(\overline{X}) = \mu$.
   (b) Is $\overline{X}$ an unbiased estimator of μ? Explain.

7.4 Estimating the variance

Before we look at estimators of the variance, we will first provide the following result.

Result 7.4.1. *Consider two random variables, X and Y. If X and Y are independent, then $E(XY) = E(X)E(Y)$.*

We can show that the above is true via the definition of covariance and noting that, due to independence, $\text{Cov}(X, Y) = 0$.

6. Consider a random sample denoted X_1, X_2, X_3 to be sampled from a population with unknown mean μ and unknown variance σ^2. Recall an intuitive estimator for σ^2 in (1) of Readings 7.3 is, in the case of $n = 3$,

$$\widehat{\sigma}^2 = \frac{1}{3}\sum_{i=1}^{3}(X_i - \overline{X})^2 = \frac{1}{3}\sum_{i=1}^{3}X_i^2 - \overline{X}^2$$

   where the sample mean estimator is $\overline{X} = (X_1 + X_2 + X_3)/3$. In this question you will be stepped through a proof that $E(\widehat{\sigma}^2) \neq \sigma^2$. This proof can be easily generalised for any n, not just $n = 3$.

(a) Using the definition of variance, show that $E(X_i^2) = \sigma^2 + \mu^2$ for $i = 1, 2, 3$.

(b) Using Result 7.4.1 above, show that $E(X_i X_j) = \mu^2$ for $i, j = 1, 2, 3$ and $i \neq j$.

(c) Consider $E\left(\overline{X}^2\right)$. Expand $\overline{X}^2 = (X_1 + X_2 + X_3)^2/9$ so that it is in terms of $X_1^2, X_1 X_2$ etc.

(d) Now, using some of your results above, show that $E(\overline{X}^2) = \sigma^2/3 + \mu^2$.

(e) Now show that $E[(X_1^2 + X_2^2 + X_3^2)/3] = \sigma^2 + \mu^2$.

(f) Finally, put this altogether to show that

$$E(\widehat{\sigma}^2) = \sigma^2 - \frac{\sigma^2}{3} = \frac{2}{3}\sigma^2.$$

Using this result, is $\widehat{\sigma}^2$ an unbiased estimator of σ^2? Explain.

(g) Note that, for a general n, $S^2 = n\widehat{\sigma}^2/(n-1)$. Use this with $n = 3$ to show that S^2 is an unbiased estimator of σ^2.

7.5 Estimating the mean, variance and standard deviation in R

We will now look at the pressure data set that was considered in Applications 7.1 from Pearson (2011). The data set, and others, can be found on the LMS in the Lab folder for Week 7. These were originally downloaded from the Oxford University Press website (**click here**). To load the data, follow these steps:

- Download the file "pressure1.csv" from the LMS to a users folder or USB etc. that you have access to.
- In RStudio, go to the `File` menu, select `Import Dataset` and then choose `From Text File...`
- Locate the `.csv` file you saved, select the file and press the `Open` button.
- There are some options here that you can change, but you should be able to leave them untouched. In the bottom-right you should see a Data Frame with two columns (X: observation number; Pressure: the pressure values). Select `Import` and the data should now be stored in a data frame called "pressure1" (unless you changed the name).
- A new window should now be open displaying the data. Go back to your working script file for this lab.

If you are not using RStudio, you can also use an R command to read in the file. For example,

```
pressure1 <- read.csv("C:/pressure1.csv", header = TRUE)
```

will read in the data if it is located in the top level of the C drive. Note that in file paths 'forward slashes' are used, not 'back-slashes' like in Windows. The option `header = TRUE` tells R that the first row is the column headings.

7. (a) The values we wish to look at are in the column labeled `Pressure` in the data frame `pressure1`. We can access these values via `pressure1$Pressure`. The below commands will plot the histogram and calculate the sample mean, sample variance and sample standard deviation, which is the square root of the sample variance.

```
hist(pressure1$Pressure, freq = FALSE, xlab = "Pressure", breaks = 20)
mean(pressure1$Pressure) # Sample mean
var(pressure1$Pressure)  # Sample variance
sd(pressure1$Pressure)   # Sample standard deviation
```

Run these commands in R.

(b) Overlay a normal probability density curve with the parameters set to your estimates from above. Does it appear to fit the data well?

(c) Try changing the argument `breaks = 20` to choose some different numbers of bins.

(d) It can be difficult to guess a suitable mean and standard deviation from skewed histograms. On the LMS you will find three more data sets called `pressure2.csv`, `pressure3.csv` and `pressure4.csv`. Create histograms of these datasets can see if you can guess close to the sample means and sample standard deviations for each. Compare your guesses to the true estimates found using R. How did you go?

8. To access the solutions to this lab on Overleaf, **click here**.

Once you have finished the lab, have another go at any Assignment 2 questions you need to still do.

References

PEARSON, R. 2011. *Exploring Data in Engineering, the Sciences, and Medicine*. New York: Oxford University Press.