Understanding Data and Statistical Design (60117)

Lab 1: Introduction to statistics

This lab is not assessed.

In this week's lab we look at some of the fundamentals of data and statistics.

Question 1

Binomial RV

Consider the toss of a dice where A represents obtaining a 1 or a 3 and B represents obtaining a 2, 4, 5, or 6. Let this dice be "fair", by which that each face has an equal chance of being thrown. As the dice is fair we have

$$\text{Prob}(A) = 1/6 + 1/6 = 1/3$$

and

$$\text{Prob}(B) = 1 - 1/3 = 2/3,$$

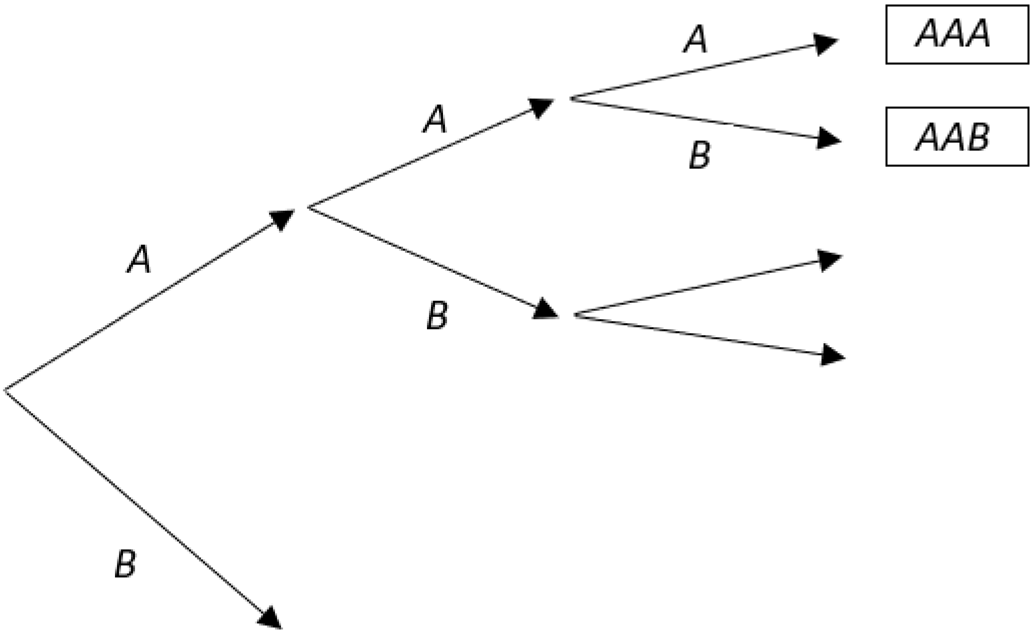
the last because probabilities sum to 1.

Note that each outcome of a dice toss is "independent" of others – the outcome of one dice toss has no impact on the outcome of another (the dice toss "events" are "independent").

Now suppose an experiment is conducted where a sequence of three dice tosses is performed.

(a) Write down the eight possible outcomes (in terms of events A and B) of this sequence of three dice tosses (this is the "sample space" of the experiment).

Hint. Two of the eight possible outcomes are AAA (1 or 3 on all three dice tosses) and AAB (1 or 3 on first two dice tosses and 2, 4, 5 or 6 on last). A tree diagram can often be helpful, shown incomplete below.



To calculate the probability of the 8 possible outcomes of the 3 dice tosses we use the "multiplication rule". For example

$$\text{Prob}(AAA) = \text{Prob}(A) \times \text{Prob}(A) \times \text{Prob}(A) = \frac{1}{3} * \frac{1}{3} * \frac{1}{3} = \frac{1}{27},$$
$$\text{Prob}(AAB) = \text{Prob}(A) \times \text{Prob}(A) \times \text{Prob}(B) = \frac{1}{3} * \frac{1}{3} * \frac{2}{3} = \frac{2}{27}$$

etc. (The multiplication operation above is valid only in the case of "independent events", as we have for the outcome of each dice toss.)

Now let the random variable (RV) X be the number of 1s or 3s (number of events A) obtained in the sequence of three dice tosses. The RV X follows what is called the "binomial distribution" and to signify this we could write $X \sim \text{Bin}(3,1/3)$ where 3 and 1/3 are the "parameters of the distribution": 3 trials, probability of success (event A) 1/3.

(b) Write down the probability mass function of X

$$p(x) = \text{Prob}(X = x) = \begin{cases} & x = 0 \\ 12/27 & x = 1 \\ & x = 2 \\ & x = 3 \end{cases}.$$

Hint. Referring to your answer in (a), identify the outcomes corresponding to $X = x$ and add their probabilities together. For example

$$p(1) = \text{Prob}(X = 1) = P(ABB) + P(BAB) + P(BBA)$$
$$= \left(\frac{1}{3} * \frac{2}{3} * \frac{2}{3}\right) + \left(\frac{2}{3} * \frac{1}{3} * \frac{2}{3}\right) + \left(\frac{2}{3} * \frac{2}{3} * \frac{1}{3}\right) = \frac{12}{27}.$$

(The addition operation above is valid only in the case of "independent events", as we have for the outcome of each dice toss.)

Check your answer by ensuring that the four probabilities $p(x)$ sum to 1.

(c) Using your answer from (b), calculate the expected value of X.

Hint. Calculate the weighted sum

$$\mu = E[X] = \sum_{x=0}^{3} x * \text{Prob}(X = x) = \sum_{x=0}^{3} xp(x).$$

(d) Using your answers from (b) and (c), calculate the variance of X.

Hint. Calculate the weighted sum

$$\sigma^2 = \text{var}(X) = \sum_{x=0}^{3} (x - \mu)^2 * \text{Prob}(X = x) = \sum_{x=0}^{3} (x - \mu)^2 p(x).$$

Question 2

Uniform RV

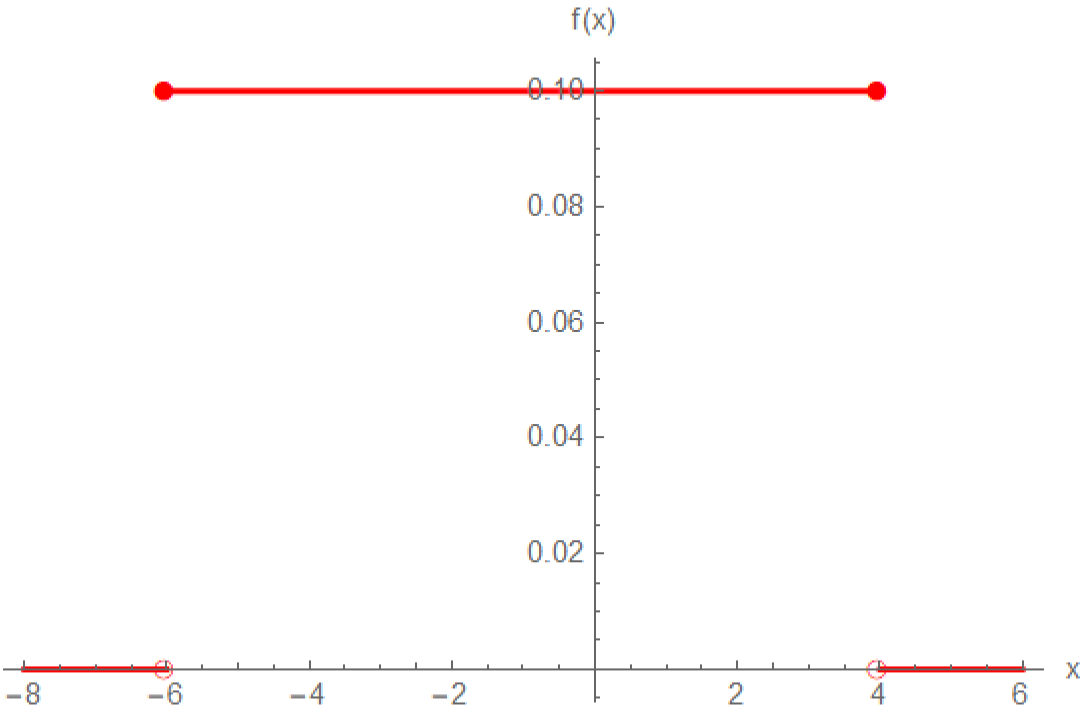
Consider a RV X that follows the "uniform distribution" on the interval $[-6,4]$, which we can signify by writing $X \sim U(-6,4)$. The RV X has probability density function (PDF) given by

$$f(x) = \begin{cases} 0 & x < -6 \\ 1/10 & -6 \le x \le 4. \\ 0 & x > 4 \end{cases}$$

(a) Show that

$$\text{Prob}(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = \int_{-6}^{4} \frac{1}{10} dx = 1.$$

Hint. For those who don't know calculus, the "integral" above is the area between the curve $f(x)$ and the x-axis for $-\infty < x < \infty$. We can find this using geometric arguments (see diagram below).



(b) Calculate

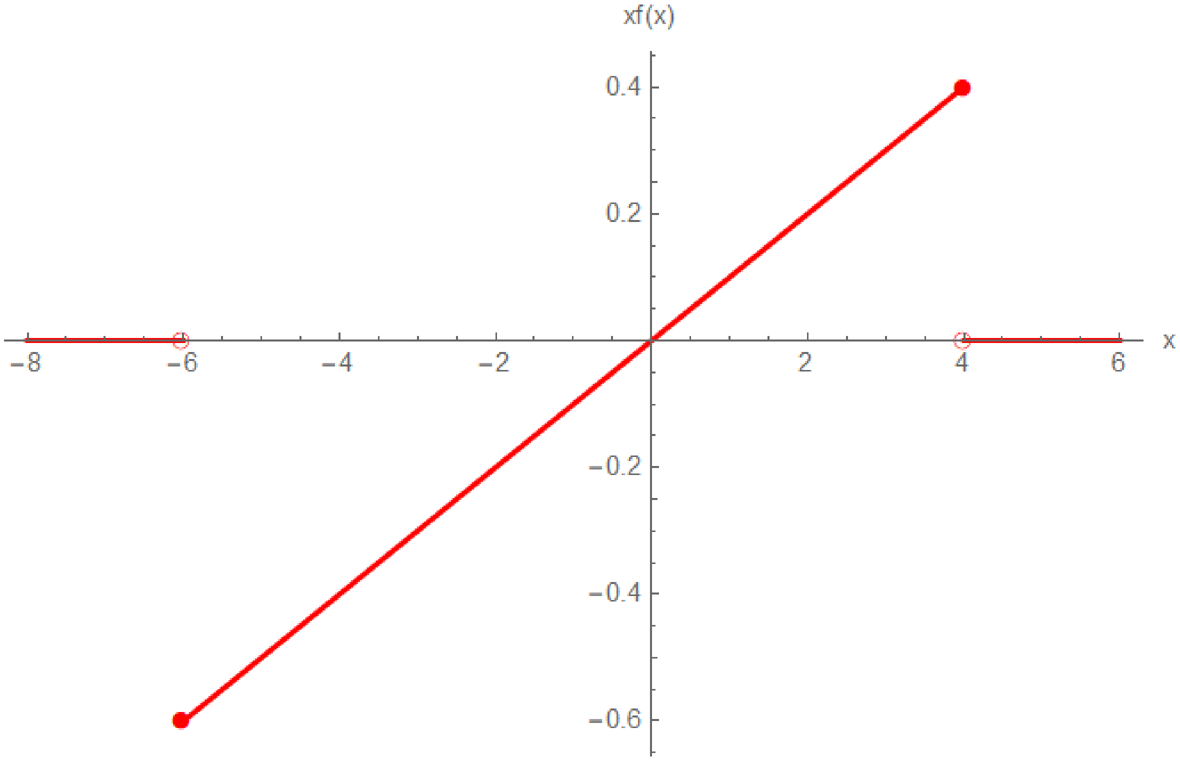
$$\text{Prob}(-3.5 \le X \le 2.5) = \int_{-3.5}^{2.5} f(x)dx = \int_{-3.5}^{2.5} \frac{1}{10} dx.$$

Hint. This is the area between the curve $f(x)$ and the x-axis for $-3.5 \le x \le 2.5$ which can be found using geometric arguments and the diagram provided above.

(c) Calculate the expected value of X, i.e.

$$\mu = E[X] = \int_{-\infty}^{\infty} xf\,(x)dx = \int_{-6}^{4} x\frac{1}{10}dx.$$

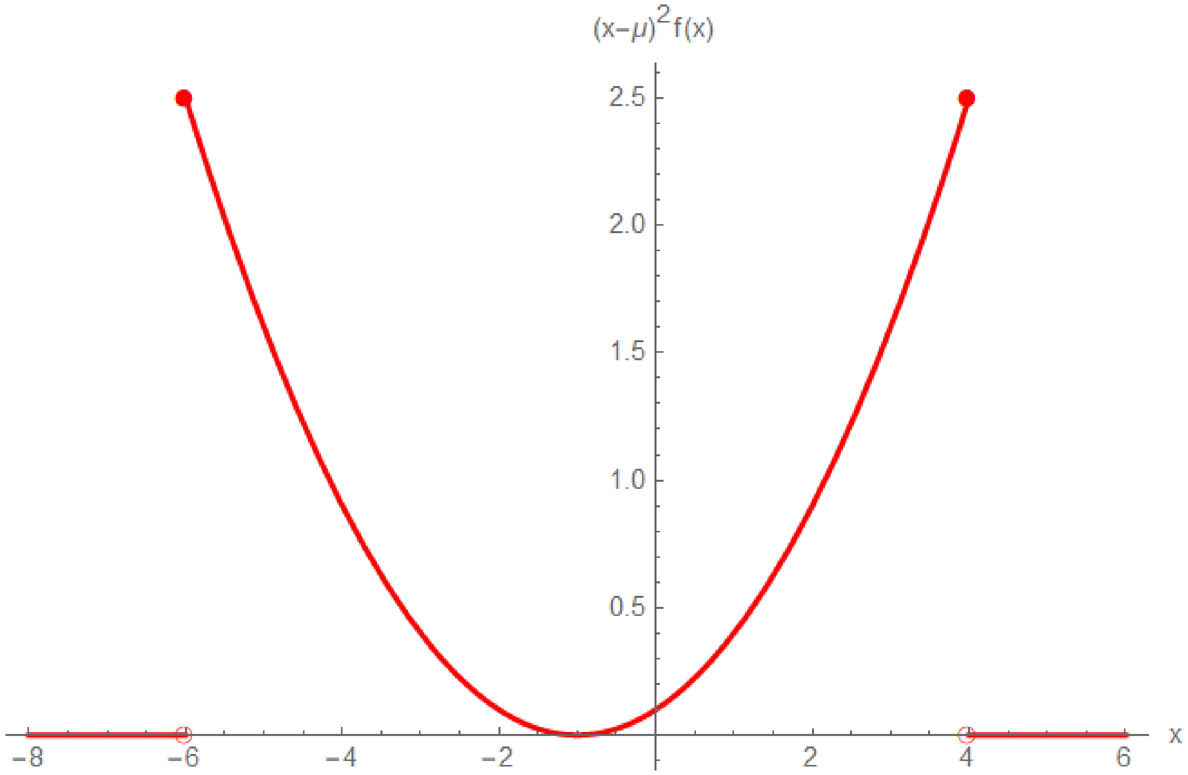
Hint. For those who don't know calculus, the integral above is the area between the curve $xf(x)$ and the x-axis for $-\infty < x < \infty$. We can find this using geometric arguments (see diagram below). Be careful, as there are both "positive" and "negative" areas.



(d) Calculate the variance of X, i.e.

$$\sigma^2 = \mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f\,(x)dx = \int_{-6}^{4} (x-\mu)^2 \frac{1}{10} dx.$$

Hint. For those who don't know calculus, the integral above is the area between the curve $(x-\mu)^2 f(x)$ and the x-axis for $-\infty < x < \infty$.



This is harder to calculate using geometric arguments. However, with calculus it is easy.

Question 3

Normal RVs

For this question we use the data in “lab1.csv”.

The data was generated by a random number generator according to the bivariate joint-normal distribution

$$\binom{X}{Y} \sim N\left(\binom{5}{-1}, \begin{pmatrix} 4 & 9/2 \\ 9/2 & 9 \end{pmatrix}\right)$$

which means that:

- $X \sim N(5,2)$, i.e. normally distributed with mean $\mu_X = 5$ and standard deviation $\sigma_X = 2$ (variance $\sigma_X^2 = 4$)
- $Y \sim N(-1,3)$, i.e. normally distributed with mean $\mu_Y = -1$ and standard deviation $\sigma_Y = 3$ (variance $\sigma_Y^2 = 9$)
- covariance of X, Y $\text{covar}(X,Y) = 9/2$
- sample size $n = 100$.

(a) Compute the sample mean $\bar{x}$ and sample standard deviation s_x of the X data sample.

Hint. Use the R functions `mean` and `sd`.

(b) Calculate $\text{corr}(X,Y)$, i.e. the population correlation of X and Y. Compare this to the sample correlation of X and Y.

Hint. For the first part see lecture notes and for the second part use the R function `cor`.

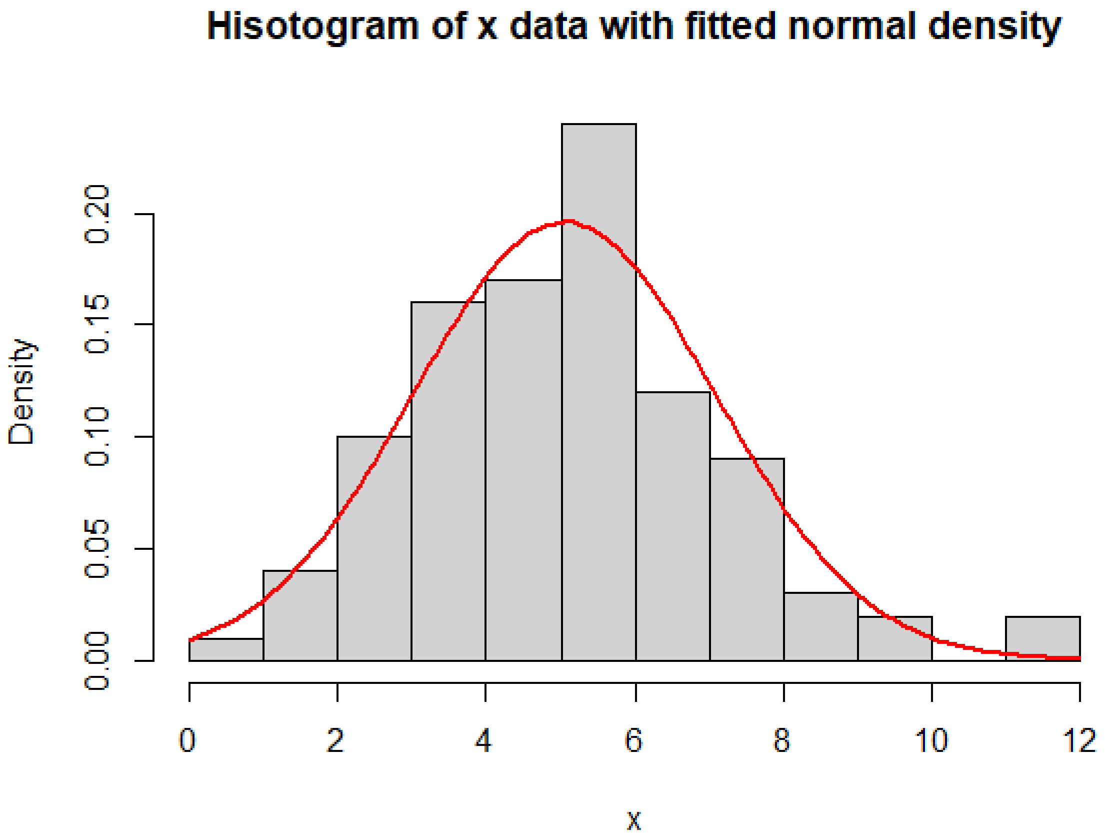
(c) Using R, create a box plot of the X data sample and copy into your answer. Describe the main features of the box plot.

Hint. Use R function `boxplot` and refer to lecture R code file.

(d) Using R, create a scatter plot of the X and Y data sample and copy into your answer. Describe the type and sign of the relationship between the variables.

Hint. Use R function `plot` and refer to lecture R code file.

(e) The following is a histogram of the X data sample with normal density fitted over the top (see accompanying R code file).



Determine if the plot suggests the X data sample is from a normal population.

Hint. The closer the sample points to the red line the stronger the evidence of normality.