Masco: follow-along with R

2025-10-10

1. Random variables

R includes standard functions to give the p.m.f (probability mass function), the p.d.f (probability density function), the c.d.f (cumulative distribution function) and to generate realizations of random variables from the most common distributions.

Questions:

1. Generate random samples of size $N = 10, 20, \cdots, 2000$ from a normal distribution with mean 1 and standard deviation 1.
2. Plot the sample mean against the sample size.
3. What are your conclusions.
4. Repeat the same steps, but using samples from a Binomial distribution.

2. Mean

In a game, a player rolls a fair 6-faced dice. The player:

- Wins 10 points if the result is 6.
- Loses 3 points if the result is 5 or 4.
- Loses 1 point if the result is 3, 2, or 1.

Questions:

1. Build an experiment that shows the score as a function of the number of dice tosses.
2. At the end of the game, the player wins if the total score is positive and loses otherwise. How many times would you recommend playing to increase the chance of winning? Plot the score against the number of tosses.
3. If the dice is unfair, with probability $\frac{8}{59}$ for face 6 and the other faces being equally likely, would you recommend playing? Run an experiment to justify your choice.

3. Example: Spike Trains of a Cockroach Antennal Lobe Neurons

The following spike train data was generated using the `STAR` (Spike Train Analysis with R) library. It represents the spike trains of four neurons, each recorded during 30 seconds of spontaneous activity (see `CAL1S`).

You can load this data using the command

```
load("spikes_data.RData")
```

Make sure the the data is in your 'working directory'. If you look at 'Environment' section on the top right of your screen, you will see a variable named `sp_time`.

1. Plot the *raster plot* of the neuron firing times in the same figure.

2. Compute the mean and standard deviation of the waiting times between consecutive spikes for each neuron, then use these results to comment on the previous figure.

4. Variance

The Fano Factor (FF) is defined as the ratio of the variance of the spike count ($\mathrm{Var}(N)$) to the mean spike count ($\mathrm{E}[N]$) over a fixed time window and repeated trials:

$$\mathrm{FF} = \frac{\mathrm{Var}(N)}{\mathrm{E}[N]}.$$

Suppose you have recorded the number of spikes from a single neuron in the primary visual cortex (V1) during a 500ms presentation of a preferred stimulus over 15 separate trials.

```
spike_counts <- c(8,10,9,14,11,11,8,10,9,15,9,12,7,10,9)
```

1. Calculate the Fano Factor (FF) for this neuron's response.
2. Based on the FF value, describe the variability of the neuron's spike train relative to a random (Poisson) process.

5. Linear regression

In this part, we consider the real dataset `cars` from the package `datasets`, it can be loaded using the command

```
data(cars)
```

To understand this dataset, use the `help` command

```
help(cars)
```

Consider the simple linear regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \qquad 1 \leq i \leq N.$$

The y_i's are the observations of the response variable, and the x_i's of the explanatory variable.
The regression coefficients α and β are respectively the intercept and the slope. The ε_i terms can be interpreted as random errors or deviations from the regression line, representing small measurement errors or unobserved influences.

Questions:

1. Create a dataset in R consisting of $N = 100$ observations with $(\alpha, \beta, \sigma) = (0.5, 1.5, 0.25)$.
2. Plot the observed data and superimpose the true regression curve.
3. Compute the correlation between car speed and stopping distance.
4. Fit a linear regression model to the `cars` dataset, with stopping distance as the response variable.

6. EEG Data

We consider EEG (electroencephalography) data from the package `eegkitdata`.
To load this dataset, first install and load the package:

```
library(eegkitdata)
data(eegdata)
help(eegdata)
```

Questions:

1. Give a brief description of this dataset.
2. Compute the correlation between sensors for a subject from group **a** and a subject from group **c**.
3. Plot the correlation matrices for the subjects you computed in the previous question, and compare the two plots.
4. Do you think that the EEG recordings from group **a** differ from those of group **c** ?

Bonus

Read and complete the exercises in the following resources:

- Intro to R – Part 21: Descriptive Statistics
- Intro to R – Part 22: Probability Distributions