Module 16 self-assessment

Question 1

Assume you are conducting a poll to determine the opinion of a large population regarding a binary decision, i.e. 'YES' or 'NO'. Such polls are usually conducted before election or a referendum. Assume that a fraction of p of the whole population will vote 'YES' while $1-p$ will vote 'NO' and that 1000 independent and representative voters are selected (in reality this is not always the case). Explain why the probability of predicting the correct outcome depends on p and use the Central Limit Theorem to approximate the probability that the poll will predict the correct outcome if $p = 0.48$.

Solution:

The probability of an accurate prediction depends heavily on p. For example, if $p = 1$ then the outcome will be correctly predicted with probability 1. If $p \approx 0.5$ this is not the case. In fact, the probability of predicting the correct outcome in this situation is approximately 50% unless we sample a very substantial amount of the population! To see this is a more formal way let's write

$$X_i = \begin{cases} 1 & \text{if individual } i \text{ votes 'YES'} \\ 0 & \text{if individual } i \text{ votes 'NO'} \end{cases}$$

making X_i independent $\mathsf{Bernoulli}(p)$ random variables. Assuming that $p < 0.5$, the outcome of the vote is accurately predicted if

$$\frac{1}{1000}\sum_{i=1}^{1000} X_i < 0.5 \iff \sum_{i=1}^{1000} X_i < 500.$$

In the case of $p = 0.48$ we have $\mathrm{Var}(X_i) = 0.48 \times 0.52 = 0.2496$ and thus by the CLT

$$\bar{X}_{1000} := \frac{1}{1000}\sum_{i=1}^{1000} X_i \approx \mathcal{N}\left(0.48, \frac{0.2496}{1000}\right)$$

which yields an estimate of 89.7%, as computed by R (R is the Matlab equivalent for statistics and data science. Although not taught in EM2B, many employers in the Data-driven Engineering sector ask for it on CVs, so you *may* want to check it out. In this course we are essentially using it an an integral calculator.)

```
> pnorm(0.5, mean=0.48, sd=sqrt(0.2496/1000))
[1] 0.8972299
```

To find this from the $\Phi(z)$ table we must standardise as

$$\mathbb{P}(\bar{X}_{1000} < 0.5) = \mathbb{P}\left(Z \leq \frac{0.5 - 0.48}{\sqrt{\frac{0.2496}{1000}}}\right) = \mathbb{P}\big(Z \leq \frac{0.02}{0.0158}\big) \approx 0.8961.$$

Note that the exact can be found from binomial quantities are

```
> pbinom(499,1000,0.48)
[1] 0.8914189
> pbinom(500,1000,0.48)
[1] 0.902746
```

The CLT solution is accurate to about 2 digits.

Question 2

Let $X_1, X_2, \dots X_n$ be iid variables drawn from the standard normal, and

$$Y_n = \sum_{i=1}^{n} X_i^2.$$

Show that $\mathbb{E}[X_j^2] = 1$, and compute $\mathbb{E}[X_j^4]$. Finally, approximate $\mathbb{P}(Y_{100} > 110)$.

Solution:

One can work out

$$\mathbb{E}[X_j^2] = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} = \dots$$

using by parts integration, or easier via the variance relation

$$\mathbb{E}[X_j^2] = \mathrm{Var}(X_j) + \mathbb{E}[X_j]^2 = 1.$$

Now

$$\mathbb{E}[X_j^4] = \int_{-\infty}^{+\infty} x^4 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \mathrm{d}x.$$

To solve this integral use integration by parts. Computationally tractable integrals involving the exponential function are of the form

$$\int x \mathrm{e}^{-\frac{1}{2}x^2} \mathrm{d}x = -\mathrm{e}^{-x^2/2} + c$$

So we set $u = x^3$ and $\mathrm{d}v = x\mathrm{e}^{-\frac{1}{2}x^2}\mathrm{d}x$. Hence, using by parts,

$$\int_{-\infty}^{+\infty} x^4 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} = \left[-x^3 \mathrm{e}^{-\frac{1}{2}x^2}\right]_{-\infty}^{+\infty} + \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{1}{2}x^2} \, 3x^2 \mathrm{d}x = 0 + \mathrm{E}[3X_j^2] = 3.$$

To find $\mathbb{P}(Y_{100} > 110)$ we appeal to the central limit theorem to assume that Y_{100} is normally distributed. Due to the independence in X, we have $\mathbb{E}[X_j^2] = 1$. The variance $\mathrm{Var}(X_j^2)$ we can compute by using the identity $\mathrm{Var}(X_j) = \mathbb{E}[X_j^2] - \mathbb{E}[X_j]^2$ with X_j replaced by X_j^2 as

$$\mathrm{Var}(X_j^2) = \mathbb{E}[X_j^4] - \mathbb{E}[X_j^2]^2 = 3 - 1 = 2.$$

This leads to a normalisation constant

$$z = \frac{110 - 100 \cdot \mathbb{E}[X_j^2]}{\sqrt{100}\sqrt{\mathrm{Var}(X_j^2)}} = \frac{1}{\sqrt{2}},$$

hence

$$\mathbb{P}(Y_{100} > 100) = \mathbb{P}(Z_{100} > \frac{1}{\sqrt{2}}) = 1 - F_Z(\frac{1}{\sqrt{2}}) = 0.24.$$