Exponential distribution

- This exercise is an illustration of the method of moments.
- Using PowerBI and R, generate a population from an exponential distribution with rate 2:

```
sim_exp <- data.frame(n=1:100000,x=rexp(100000,rate=2))
```

- Compute the mean m and the standard deviation σ (square root of the variance) on 1000 samples of size 1000.
- Estimate the number of times $|m - 0.5|$ (resp. $|\sigma - 0.5|$) is greater than 0.026.
- In your opinion, which estimator of the inverse of the rate is the best one ?

Confidence intervals

- A consistent estimator $\hat{\theta}$ will fall within a given interval around the true value θ with a probability close to 1 if the sample size is large enough.
- In practice, the sample size is fixed, so that one wants to find c such that:
$$P\left[|\hat{\theta} - \theta| > c\right] = \alpha, \tag{1}$$
where $\alpha \in [0, 1]$ is a fixed value.
- The interval:
$$\left[\hat{\theta} - c, \hat{\theta} + c\right]$$
contains θ with probability $1 - \alpha$.
- $1 - \alpha$ is called the confidence level.

Confidence intervals

- Generally speaking, a confidence interval for θ at level $1-\alpha$ is a pair of estimators $\hat{\theta}_1, \hat{\theta}_2$ such that:

$$P\left[\theta \in \left[\hat{\theta}_1, \hat{\theta}_2\right]\right] = 1 - \alpha.$$

- A classical situation encountered is when $\hat{\theta}$ is the sample mean:

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Using the TCL

- When the sample mean is an unbiased estimator of θ, then, in the limit of large samples:

$$Z = \frac{\sqrt{n}\left(\bar{X} - \theta\right)}{\sigma} \tag{2}$$

  is $\mathcal{N}(0,1)$-distributed.

- For a given confidence level $1-\alpha$, one can find c such that:

$$P\left[|Z| \leq c\right] = 1 - \alpha. \tag{3}$$

- A confidence interval for θ is then:

$$\left[\bar{X} - \frac{\sigma c}{\sqrt{n}}, \bar{X} + \frac{\sigma c}{\sqrt{n}}\right] \tag{4}$$

- The approximation is correct provided $n \geq 30$.

Estimating the expected value

- Going back to the example of the exponential distribution, and assuming one wants to estimate the inverse of the rate λ^{-1}, the sample mean is an unbiased estimator.
- Using the TCL, can you find a confidence interval for λ^{-1} at confidence level 0.90 ?
- Is the threshold 0.026 used before coherent with what you found ?

The delta method

- Sometimes, the estimator is a function g of the sample mean and the TCL cannot be applied directly.
- However, if g is continuously differentiable and the TCL gives:

$$\sqrt{n}\left(\bar{X} - \theta\right) \to \mathcal{N}(0, \sigma^2), \tag{5}$$

  then:

$$\sqrt{n}\left[g\left(\bar{X}\right) - g(\theta)\right] \to \mathcal{N}\left(0, \left(g'(\theta)\right)^2 \sigma^2\right) \tag{6}$$

- θ can be replaced by $\bar{X}$ in the right hand side.

- Use the delta method to find a confidence interval at confidence level 0.9 for the rate in the exponential distribution example.
- Compare it with the one obtained by applying the inverse on the bounds of the confidence interval for λ^{-1}.

Slutsky's theorem

- Let Y_n be a family of random variables converging to a constant c in the sense that, for any $\epsilon > 0$:

$$P\left[|Y_n - c| > \epsilon\right] \to 0,$$

- Let $X_1, \ldots, X_n$ be an iid sample of a population distribution with expected value μ and variance σ^2.
- Then:

$$\sqrt{n}\left(Y_n \bar{X} - c\mu\right) \to \mathcal{N}\left(0, c^2\sigma^2\right). \quad (7)$$

- This theorem justifies the use of the sample variance instead of the true one in the TCL.

Exercice
In an operational setting, an approach is said to be abnormal if the aircraft crosses the boundary of a tube centered on the nominal trajectory. 1000 flights were recorded at Charles de Gaulle airport and 8 were found to be abnormal. All the flights are assumed to be independent and the probability of one flight to be abnormal is p. Let X_i be the random variable that takes the value 1 if the i-th flight is abnormal, 0 otherwise.

- Let $\bar{X}$ be the sample mean of $X_1, \ldots, X_{1000}$. Find the expected value and the variance of $\bar{X}$.
- Is $\bar{X}$ unbiased ? Consistent ?
- Assuming p is known, use the TCL to find an approximate distribution of $\bar{X}$.

ESTIMATING A PROPORTION (CTD.)

- Invoke Slutsky theorem to show that the true variance in the TCL can be replaced by $\bar{X}\left(1-\bar{X}\right)$.
- Deduce a confidence interval at level 0.95 for p.
- What happens to the size of the confidence interval if the size of the sample is reduced ? Increased ?

Estimating the mean

- In the case of normally distributed samples, it is possible to derive confidence intervals even for small samples.
- For a sample of size n, if S is the sample standard deviation, then the confidence interval for the mean at level $1-\alpha$ is:

$$\left[\bar{X} - c\frac{S}{\sqrt{n}}, \bar{X} + c\frac{S}{\sqrt{n}}\right] \tag{8}$$

where c is the value such that:

$$P\left[|T| > c\right] = \alpha, \tag{9}$$

T being distributed according to a **Student's distribution** with $n-1$ degrees of freedom.

Exercise

Sample mean of a normal sample
A milling machine produces mechanical parts of average length 40 mm with a standard deviation of 10^{-2} mm. The distribution of the length is assumed to be normal. Each day, 9 parts are sampled from the production batch and their average mean $\bar{\mu}$ computed.

- Give a confidence inteval for μ at confidence level 0.9.
- For the same confidence level, how many parts must the manufacturer sample each day to have a confidence interval of length 10^{-3} mm ?
- Same questions when the standard deviation is only an estimation based on the unbiased estimator of variance.

Estimating the variance

- Using the same notations, a confidence interval at level $1-\alpha$ for the variance is:

$$\left[\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right] \tag{10}$$

where c_1, c_2 are such that:

$$P\left[|Z| > c_2\right] = P\left[|Z| < c_1\right] = \frac{\alpha}{2} \tag{11}$$

X being distributed according to a **Chi square distribution** with $n-1$ degrees of freedom.

Exercise

Sample variance of a normal sample
For the same milling machine as in the previous exercise, the unbiased sample variance on a batch of 10 parts is found to be 2×10^{-5}.

- Give a confidence inteval for the true variance σ at confidence level 0.9
- If the sample variance becomes 5×10^{-5}, can we conclude that the machine is malfunctionning ?