Estimation

INTRODUCTORY EXAMPLE

Delays dataset

- Download the file ”delays.csv” and open it in PowerBI. Use the ”Table” pane and find the maximum and minimum values of the third column.
- Negative delays are not of interest for us, nor delays higher than 3h, that are outliers. Create a new table with the command:

```
pos_delays =
FILTER(delays,delays[Column3]>=0 && delays[Column3]<=10000)
```

- Using an ”R” visual, plot the histogram of the delays in the new table with the option `freq=FALSE` to display probabilities. Since PowerBI removes duplicate values, be sure to include all columns in the plot.

Introductory example

Modeling the delays

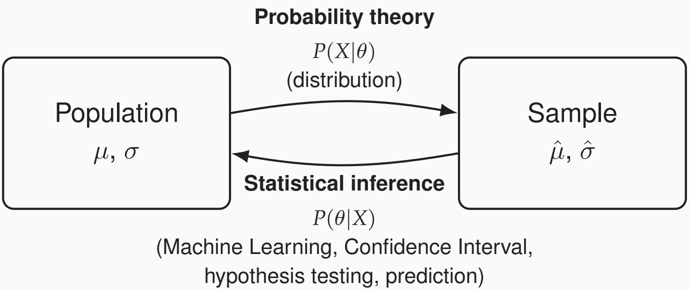
- Compute the inverse of the average of the delays.
- Use an R script to generate an exponentially distributed sample of size 100000 (use the `rexp` command) with rate the above value.
- Compare the histogram of the values with the one coming from the delays dataset.
- Do you think that assuming an exponential distribution for the delays is reasonable ?

ESTIMATION

Population vs sample

- In the previous example, the dataset of delays is a sample, i.e. a set of **observations**.
- The associated population is an hypothetical measure space describing all possible delays.

Illustration: Population vs sample



- **Goal:** Model and quantify uncertainty using data.
- Descriptive statistics: summarize the sample
- Probability theory: from population assumptions $\rightarrow$ data behavior.
- Statistical inference: from observed data $\rightarrow$ population insights.

Estimation

Statistical model

- An statistical model is a triple $(\Omega, \mathcal{T}, \mathcal{P})$ where Ω is a sample space, $\mathcal{T}$ a σ-algebra on Ω and $\mathcal{P}$ a set of probabilities on $\mathcal{T}$.
- The population associated with the studied sample hopefully belongs to $\mathcal{P}$...

Kinds of models

- If $\mathcal{P}$ can be fully described by a finite number of parameters, the model is said to be **parametric**. As an example, the set of exponential distributions is parameterized by the rate.
- If $\mathcal{P}$ can be partly described by a finite number of parameters, the model is said to be **semi-parametric**.
- In all the other cases, the model is **non-parametric**.
- In the course, only parametric models are considered.

Estimation

Estimation

- Let the population has probability measure $P \in \mathcal{P}$.
- Assume that independent random variables $X_1, \ldots, X_n$ have the identical distribution P.
- Estimation is the process of finding P using the sample $X_1, \ldots, X_n$.
- When the model is parametric, an estimator of a parameter is a random variable $Y = f(X_1, \ldots, X_n)$ where f does not depend on the parameter.

Bias and consistency

- Assuming θ is the true value of the parameter to be estimated, and $f\left(X_1, \ldots, X_n\right)$ is the estimator, the bias is:

$$E\left[f\left(X_1, \ldots X_n\right)\right] - \theta. \tag{1}$$

- If the bias vanishes, the estimator is said to be unbiased.
- The mean square error of the estimator is:

$$MSE_f(X_1, \ldots, X_n) = E\left[\left(f(X_1, \ldots X_n) - \theta\right)^2\right] \tag{2}$$

- An estimator is said to be consistent if, for any $\epsilon > 0$:

$$\lim_{n \to +\infty} P\left(\left|f(X_1, \ldots, X_n) - \theta\right| > \epsilon\right) = 0. \tag{3}$$

Consistency

Bienaymé–Chebyshev inequality
Let X be a random variable. Then, for any $\epsilon > 0$:

$$P\left(|X - E\left[X\right]| \geq \epsilon\right) \leq \frac{V(X)}{\epsilon^2}. \tag{4}$$

MSE and consistency

- The next two properties are consequences of 4.
- If an estimator $f\left(X_1, \ldots, X_n\right)$ is unbiased, then it is consistent if its variance goes to 0 as n goes to $+\infty$.
- For any estimator, if $\lim_{n \to +\infty} MSE_f\left(X_1, \ldots, X_n\right) = 0$, then the estimator is consistent.
- From now, an estimator of θ will be denoted $\hat{\theta}$, letting the sample be implicit.

Method of moments

- This procedure is one of the most common in practice.
- If the parameter θ to be estimated is a moment $E\left[X^k\right]$ of the population probability measure, then the sample mean:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{5}$$

  is an unbiased and consistent estimator.
- If g is a continuous function, $g\left(\hat{\theta}\right)$ is generally not an unbiased estimator of $g(\theta)$, it is, however, consistent.

Exercise with R: Understand estimators
Objectives:

- Understand the concept of estimation in statistics.
- Differentiate between sample and population parameters.
- Use R to estimate the population mean and variance.
- Explore the effect of sample size on estimation accuracy.
- Visualize and interpret the consistency of an estimator.

Part 1: Generating Data

Task (5 min): Simulate a population using

```
population <-
data.frame(n=1:100000,
           x=rnorm(n=100000, mean = 50, sd = 10))
```

to create a dataset in PowerBI.

Answer questions:

- Plot the histogram of the population.
- What probability distribution does the population follow?
- Can you relate the population parameters (μ, σ^2) to the moments?
- Do we usually know these values in practice?

Part 2: Sample Estimation

Task (10 min): Draw samples of different sizes (10, 50, and 200), and compute the sample mean and variance using DAX functions `SAMPLE`, `AVERAGEX`, `VARX.S`.

Answer questions:

- Do larger samples give more stable estimates of the mean?
- How close are the sample means to the true mean (50)?
- Does sample variance approximate the population variance well?

Part 3: Sampling Distribution

Task (10 min): Examine variability in estimates using simulation.

1. Set the sample size $n = 30$
2. Set the number of simulations nsim $= 1000$
3. Draw 1000 random samples (of size 30) from the population and record each sample mean.
4. Plot a histogram of the sample means.
5. Compute and report the mean and standard deviation of the sample means.

Answer questions:

- What is the average of the sample means?
- Does it approximate the true mean (unbiasedness)?
- What happens to the spread (variability) of the means if n increases?
- Can you relate this to the Central Limit Theorem?

Part 4: Understanding Consistency

Task (5 min): Investigate the consistency of the sample mean estimator. Recall: An estimator $\hat{\theta}_n$ is **consistent** if, for any $\epsilon > 0$,

$$\lim_{n \to +\infty} P\left(|\hat{\theta}_n - \theta| > \epsilon\right) = 0.$$

Run R codes:

```
sample_sizes <- c(10, 50, 200, 1000, 5000)
true_mean <- mean(population$x)
errors <- sapply(sample_sizes, function(n) {
 means <- replicate(1000, mean(sample(population$x, n)));
 mean(abs(means - true_mean) > 1) })
deviation_proba <- data.frame(n = sample_sizes, P_error = errors)
```

Answer questions:

- What happens to $P(|\bar{X}_n - \mu| > 1)$ as n increases?
- How does this result illustrate the definition of consistency?
- Why is the sample mean a consistent estimator of the population mean?

Summary and Reflection
Key Takeaways:

- Sample estimates approximate unknown population parameters.
- Larger samples yield more precise estimates.
- Sampling distributions provide a foundation for inference.
- Consistency ensures that estimators converge to the true value as sample size grows.

Reflection:

- What assumptions underlie these estimations?
- How would results differ for non-normal populations?
- How might bias arise in practical data collection?