TD1: Introduction to statistical models

Exercise 1

We aim to determine whether ENSAI students have any preference for cats or dogs. We assume that, a priori, they have no preference on average. We ask n students what their preferences are, and we let X be the number of "cat" answers.

1. Define H_0 and H_1. Is it a one-sided (unilatéral) or two-sided (bilatéral) test ?
2. We observe 10 students and $X = 8$ "cat" answers. Compute the pvalue in this specific case.
3. Write the expression of the pvalue in terms of n and X and F, the cdf of $\mathrm{Bin}(n, 0.5)$.
4. Write a line of code to compute the pvalue in Julia, Python or R.
5. What is the pvalue if H_1 is instead

   a. "Students prefer cats" or
   b. "Students prefer dogs"

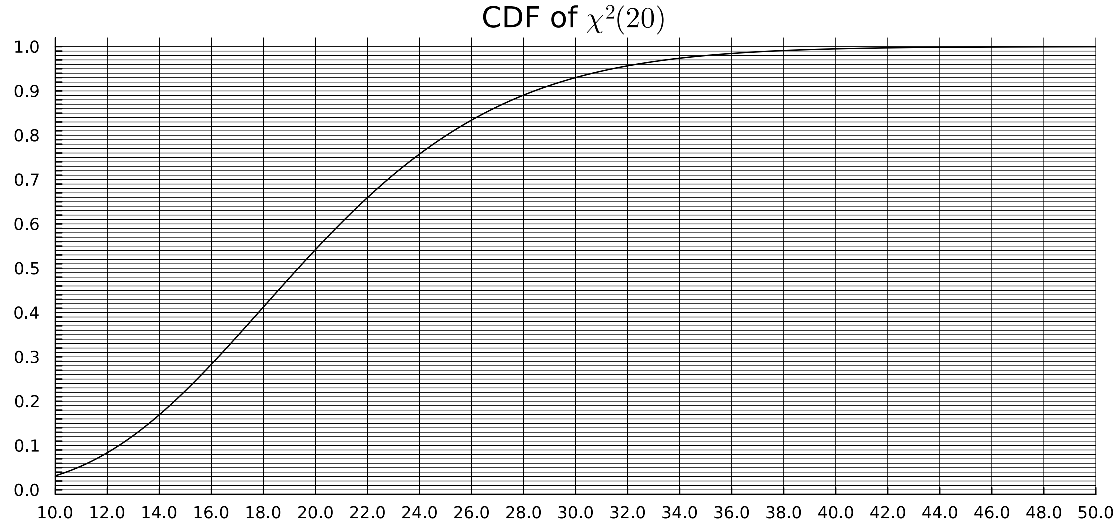
Exercise 2

Let $(X_1, X_2, \dots, X_n)$ be a sample from an exponential distribution $\mathcal{E}(\lambda)$. We want to test:

$$H_0 : \lambda = \frac{1}{2} \quad \text{vs.} \quad H_1 : \lambda = 1.$$

1. Show that if $X \sim E(\lambda)$ and $Y \sim \Gamma(k, \lambda)$ with $k \in \mathbb{N}^*$, then $X + Y \sim \Gamma(k + 1, \lambda)$. *We recall that the density of* $\Gamma(\lambda, k)$ *is given by* $p(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$
2. Deduce that $S_n = \sum_{i=1}^{n} X_i$ follows a Gamma distribution $\Gamma(n, \lambda)$.
3. For a sample of size $n = 10$, What is the rejection region of S_n at the 0.05 significance level for the simple likelihood ratio test?
   We admit that a Gamma distribution $\Gamma(n, \frac{1}{2})$ *is a chi-squared distribution with* $2n$ *degrees of freedom,* $\chi^2(2n)$. **Bonus**: Show this fact for $n = 1$, using a polar change of variable.
4. The empirical mean is $\bar{x}_{10} = 2.5$. What can we conclude?

5. Recall what a cdf is, and read the p-value on the cdf of the $\chi^2(20)$ distribution



6. Compare the p-value if we use a Gaussian approximation of $\sum X_i$ with the TCL.

Exercise 3

Let $X_1, X_2, \ldots, X_n$ be random variables drawn from a normal distribution $N(\theta, 1)$. To test $H_0 : \theta = 5$ against $H_1 : \theta > 5$, we propose the following test:

$$T = \mathbf{1}\{\bar{x} > 5 + u\},$$

where $\bar{x}$ is the empirical mean and u is to be fixed.

1. a. Derive the function $t \to \mathbb{P}(Z \geq t) - e^{-t^2/2}$, where $Z \sim \mathcal{N}(0, 1)$
   b. Deduce that $\mathbb{P}(Z \geq t) \leq e^{-t^2/2}$ for all t
2. Deduce a value of u such that the type I error of this test is smaller than a given α. Rewrite the test T in function of α.
3. Fix $\alpha = 1/e$ (and $u = \sqrt{2/n}$). Compute the power function.

Exercise 4

Let the family of Pareto distributions with **known** parameter a and **unknown** parameter θ:

$$f(x) = \begin{cases} \frac{\theta}{a}\left(\frac{a}{x}\right)^{\theta+1}, & \text{if } x \geq a, \\ 0, & \text{if } x < a. \end{cases}$$

1. Compute the mean and variance of X, if X follows a Pareto distribution of parameter a and θ.
2. Rewrite the density in the form $f(x) = a(x)b(\theta)e^{c(\theta)d(x)}$ and identify a,b,c and d.
3. Deduce the general form of the uniformly most powerful test UMP_α for $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$.
4. For $a = 1$, construct the test for the null hypothesis: the mean of the distribution is smaller than or equal to 2.
5. What is the density of $d(X_1)$? d *is defined in Q.2*
6. Write a line of code in Julia, Python or R to compute the rejection region at level $\alpha = 0.05$.