Exercise Sheet 4: Data Science Methods

Technische Universität München
Winter semester 2024-2025

Instructions

You can hand in your solutions during the next tutorial on 09/01/25 or submit them earlier at the group office 1269 in the Physics Department. For code, you can send it to `alan.zander@tum.de`.

1 Exercise: Likelihood Ratio Test (LRT)

Consider the following experiment from the lecture. In a given time interval, we want to determine the number S of photons ("on") coming from a source in the sky. This can be modeled as a simple Poisson process

$$p(n_{\text{on}}|S, B) = \text{Pois}\left(n_{\text{on}}|S + B\right), \tag{1}$$

where n_{on} is the total number of photons arriving from the portion of the sky where the source is located. S and B are the numbers of photons coming from the source and the background, respectively.

Before starting the experiment, we can count the number of photons coming from an empty ("off") patch of sky to estimate the background. Again,

$$p(n_{\text{off}}|B) = \text{Pois}\left(n_{\text{off}}|\tau B\right), \tag{2}$$

where τ is a given constant accounting for the time difference in the two measurements. Hence, since these two observations are independent, we have

$$p\left(n_{\text{on}}, n_{\text{off}}|S, B\right) = \text{Pois}\left(n_{\text{on}}|S + B\right) \text{Pois}\left(n_{\text{off}}|\tau B\right). \tag{3}$$

Now, from theoretical considerations, the number of photons coming from the source is thought to be $S0 = 60$. This is our null hypothesis H_0, regardless of B (nuisance parameter). Furthermore, after performing the experiment, we observed $n_{\text{on}} = 105$ and $n_{\text{off}} = 265$, where the background-measurement ("off") was five times longer than the source-measurement ("on"), i.e. $\tau = 5$.

1.1 Determine the global best fit for S and B, i.e. the value of the MLEs $\hat{S}_{\text{MLE}}$ and $\hat{B}_{\text{MLE}}$ with no restrictions at the observed values $n_{\text{on}} = 105$, $n_{\text{off}} = 265$.

1.2 Illustrate the distribution of the MLEs $\hat{S}_{\text{MLE}}$ and $\hat{B}_{\text{MLE}}$, assuming $S = S0$ and $B = 50$.

1.3 Determine the MLE $\hat{\hat{B}}_{\text{MLE}}$, given the null hypothesis H_0.

1.4 Construct the LRT,

$$\lambda_{S_0}(n_{\text{on}}, n_{\text{off}}) = -2 \log \left(\frac{p\left(n_{\text{on}}, n_{\text{off}}|S_0, \hat{\hat{B}}_{\text{MLE}}\right)}{p\left(n_{\text{on}}, n_{\text{off}}|\hat{S}_{\text{MLE}}, \hat{B}_{\text{MLE}}\right)} \right) \tag{4}$$

and plot it against $\hat{S}_{\text{MLE}}$ to verify the relation $\lambda_{S_0} = \frac{\left(\hat{S}_{\text{MLE}} - S_0\right)^2}{\sigma^2_{\hat{S}_{\text{MLE}}}}$.

1.5 According to Wilk's theorem, the LRT is distributed according to $\sim \chi^2_k$ under the null hypothesis H_0, where k is the difference in the dimensionality of the alternative and null hypotheses. Show that this indeed holds via MC sampling. For a fixed size $\alpha = 0.05$, can we reject H_0?

1.6 Furthermore, the LRT follows a non-central χ^2_{nc} distribution with non-central parameter $\Lambda^2_{S'} = \frac{\left(S' - S_0\right)^2}{\sigma^2_{\hat{S}_{\text{MLE}}}}$ under the alternative hypothesis. Illustrate this by assuming, say, $S' = \hat{S}_{\text{MLE}}$, $B' = \hat{B}_{\text{MLE}}$.

2 Exercise: Neyman Construction

The Neyman construction is a method for building confidence intervals (CI) for a parameter θ using the probability distribution of a statistic T, $P(T|\theta)$. The CI for θ, at a given confidence level (CL) $1-\alpha-\beta$, is then $[\theta_a, \theta_b]$, where

$$\alpha = 1 - F\left(T_{\text{obs}}; \theta_a\right), \tag{5}$$

$$\beta = F\left(T_{\text{obs}}; \theta_b\right). \tag{6}$$

Here F is the cumulative distribution function of $P(T|\theta)$ and T_{obs} is the observed value of T. Equations (5, 6) must be solved (mostly numerically) to find θ_a and θ_b.

Now imagine that a random variable X is normally distributed $X \sim \mathcal{N}(\mu_0, \sigma)$, with known variance $\sigma^2 = 1$ and unknown (to us) expectation value $\mu_0 = 5$.

2.1 Construct a CI with $\alpha = \beta = 0.025$ (central CI at 95% CL), if we observe $X_{\text{obs}} = 4.5$ (here we are using the statistic $T(X) = X$, i.e. the observation itself).

2.2 Simulate $N = 2000$ observations X_{obs} (using the real parameter μ_0) and construct for each observation a 90% CI. How often do you expect the real value μ_0 to be inside the CIs? Confirm this by counting how often your CIs included the real parameter μ_0.

2.3 Now, instead of computing the CI for each of the $N = 2000$ observations, use the MLE of μ_0 as a statistic, given these N observations, and construct its 90% central CI.

2.4 **Bonus:** Let N_0 be the **expected** number of times μ_0 should have landed in a CI and N_{obs} the number you actually counted in 2.2. Construct a 99% CI for N_0, given N_{obs}. Is N_0 inside the CI? *Hint: N_{obs} follows a binomial distribution with success probability $p = N0/N$.*

3 Exercise: Wilk's Theorem

Consider the experiment of Exercise 1, having again observed the values $n_{\text{on}} = 105$, $n_{\text{off}} = 265$. However, this time, we are also interested in the background B, i.e. it is no longer a nuisance parameter.

3.1 Construct the LRT.

3.2 Plot the 68% and 90% CI for the estimation of S and B (via MLE), using Wilk's theorem.

3.3 What is the probability that the true (and forever unknown) parameters (S_{true}, B_{true}) lie inside our CIs? (no calculation needed)

Mathematical note: Wilk's theorem holds in the infinitely large sample size. However, here we only have one observation. Why can we still use it? Well, the important point for Wilk's theorem to work is a Gaussian-shaped likelihood. This is given for our Poisson distributions as for large expectation values, they are very well approximated by normal distributions.