# Hypothesis testing

January 13, 2021

## 1 Simple hypothesis testing: unknown Gaussian mean

Draw a sample $\{x_1, \ldots, x_N\}$ of $N$ independent measurements from a variable $X$ following a normal distribution with unknowm mean $\mu$ and known standard deviation $\sigma = 1$. Use this sample to choose between the two hypotheses $H_0 : \mu = \mu_0 = 1$ and $H_1 : \mu = \mu_1 = 1.5$. We will use the Neyman-Pearson test.

1. Write the test statistic.
2. Draw the test statistic distributions for $H_0$ and $H_1$.
3. Determine the critical region of the Neyman-Pearson test and compute its power when $N = 25$ and $\alpha = 0.05$.
4. Draw the ROC curve: correspondance between power and size of the test.
5. What is the minimum required sample size to have a contamination smaller than 0.1?

*Technical note*: we'll use the relation, between the cdf of the Normal distribution,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt,$$

and the error function,

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2}\, dt$$

Here are a couple of small python functions relating $\Phi$ with erf, as well as their inverse functions $\Phi^{-1}$ and $\mathrm{erf}^{-1}$:

```
import numpy as np
from scipy.special import erf, erfinv

def norm_Phi(x):
    return 0.5 * (1 + erf(x/np.sqrt(2)))

def norm_Phi_inv(x):
    return np.sqrt(2) * erfinv(2*x - 1)
```

# 2 Poisson process with unknown background: limits and p-values

We come back to the example: $\gamma\gamma \to \gamma\gamma$ from ATLAS (Nature Phys. 13 (2017) no. 9, 852-858).

- $D = 13$ observed events
- $B = 2.6 \pm 0.7$ background events (Poisson distributed, estimated from a control sample)

The unknowns are the expected background count $b$ and the expected signal count $s$. The expected event count is $d = b + s$.

1. **p-value against the background-only hypothesis.** Using the profile log-likelihood ratio,

$$q(s) = -2 \ln \frac{\mathcal{L}(x|s, \hat{\hat{b}}(s))}{\mathcal{L}(x|\hat{s}, \hat{b})}; \quad q(s, \hat{s} < 0) = 0.$$

   *Note: the actual p-value with the published numbers requires a very large number of toy experiments to be generated. In the interest of time, for this question only, we will assume* $\mathbf{D = 8}$ ***and*** $\mathbf{s_0 = 4.3}$*, instead of* $D = 13$ *and* $s_0 = 7.3$.

   (a) Using 10 000 toy experiments, compute the p-value for the observation in the data. What is the median expected significance, under the assumption that $s_0 = 4.3$?

   (b) What is the median expectation, according to the Asimov dataset? (Reminder: this means setting the observed $D$, $Q$ and $B$ to the expected values (Poisson means), even if they are not integers)

   (c) How does the p-value in data compare to the "naive" $s/\sqrt{b}$?

2. **Hypothesis test inversion.** Keeping the same test statistic $q(s)$ as before (also $q(s) = 0$ if $\hat{s} < 0$), draw the distribution of $q(s)$ using 10 000 toys, drawn from the same $s$ value, for $5 \leq s \leq 15$ (step size of 1). Then draw the observed p-value under the signal + background hypothesis[1], as a function of $s$. How can you read the Feldman-Cousins 68% C.L. interval from this plot? Compare to the 68% C.L. interval obtained last time from the profile likelihood interval ($\Delta(-2\ln\mathcal{L}_p) = 1$).

3. **Figures of merit** in a counting experiment

   (a) We mentioned earlier the "naive" $s/\sqrt{b}$. Can you see where this comes from? Consider the p-value for the background-only hypothesis, in the large $b$ limit.

   (b) You may have heard also about $s/\sqrt{s+b}$, and it is useful to realise when which one should be used. Where does $s/\sqrt{s+b}$ come from? Consider the case of a measurement of $\hat{s}$ and its "precision", again in the large sample limit.

[1] also called $\mathrm{CL}_{s+b}$