Bayesian tests

Exercise 1 (Bernoulli model– Simple hypotheses)**:**

1. Posterior distribution: $\pi(\theta_0|x) \propto \frac{2}{3}\theta_0^s(1-\theta_0)^{n-s}$, $\pi(\theta_1|x) \propto \frac{1}{3}\theta_1^s(1-\theta_1)^{n-s}$, with $s = \sum_{i=1}^n x_i$.

$$\pi(\theta_1|x) > \pi(\theta_0|x) \quad \iff \quad \frac{\theta_1^s(1-\theta_1)^{n-s}}{\theta_0^s(1-\theta_0)^{n-s}} > 2 \quad \iff \quad \rho^s > 2\left(\frac{1-\theta_0}{1-\theta_1}\right)^n,$$

with

$$\rho = \frac{\theta_1}{1-\theta_1} / \frac{\theta_0}{1-\theta_0}.$$

Let $\bar{x} = \frac{s}{n}$ be the empirical mean. Bayesian test:

$$\delta(x) = 1_{\{\bar{x} > c\}}$$

with

$$c = \frac{n\log(\frac{1-\theta_0}{1-\theta_1}) + \log 2}{n\log(\frac{1-\theta_0}{1-\theta_1}) - n\log(\frac{\theta_0}{\theta_1})}.$$

2. For $n = 100$, $c \approx 0.307$ so you accept the null hypothesis.

3. The risk is:

$$r = \frac{2}{3}\alpha + \frac{1}{3}\beta.$$

Now:

$$\alpha = P(S > s|\ \theta = \theta_0).$$

Using a Gaussian approximation, $S \approx \mathcal{N}(n\theta, n\theta(1-\theta))$, that is $S \approx n\theta + \sqrt{n\theta(1-\theta)}Z$ with $Z \sim \mathcal{N}(0,1)$. We get:

$$\alpha \approx P(Z > \frac{s - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}) \approx 0.09.$$

Similarly:

$$\beta \approx P(Z > \frac{n\theta_1 - s}{\sqrt{n\theta_1(1-\theta_1)}}) \approx 0.41.$$

Thus, $r \approx 0.2$.

4. When $n \to +\infty$, we get:

$$c \to \frac{\log(\frac{1-\theta_0}{1-\theta_1})}{\log(\frac{1-\theta_0}{1-\theta_1}) - \log(\frac{\theta_0}{\theta_1})}.$$

The type-I error rate is:

$$\alpha = P_0(\bar{X} > c).$$

Since $\bar{X}$ tends a.s. to θ_0 under H_0, we need to compare θ_0 to the limiting value of c:

$$\theta_0 > \frac{\log(\frac{1-\theta_0}{1-\theta_1})}{\log(\frac{1-\theta_0}{1-\theta_1}) - \log(\frac{\theta_0}{\theta_1})} \quad \iff \quad \theta_0 \log(\frac{\theta_0}{\theta_1}) + (1-\theta_0)\log(\frac{1-\theta_0}{1-\theta_1}) < 0.$$

This last expression is the Kullback-Leibler divergence between Bernoulli distributions with respective parameters θ_0, θ_1, which is non-negative. We conclude that $P_0(\bar{X} > c)$ tends to 0 when $n \to +\infty$: the type-I error rate vanishes.

Similarly, the type-II error rate is:

$$\alpha = P_1(\bar{X} \leq c).$$

Since $\bar{X}$ tends a.s. to θ_1 under H_1, we need to compare θ_1 to the limiting value of c:

$$\theta_1 \leq \frac{\log(\frac{1-\theta_0}{1-\theta_1})}{\log(\frac{1-\theta_0}{1-\theta_1}) - \log(\frac{\theta_0}{\theta_1})} \quad \iff \quad \theta_1 \log(\frac{\theta_1}{\theta_0}) + (1-\theta_1)\log(\frac{1-\theta_1}{1-\theta_0}) \leq 0.$$

This last expression is positive whenever $\theta_0 \neq \theta_1$ so that $P_1(\bar{X} \leq c)$ tends to 0 when $n \to +\infty$: the type-II error rate vanishes.

Overall, the Bayes risk tends to 0.

Exercise 2 (Exponential model – One-tailed hypothesis)**:**

1. Posterior distribution: $\pi(\theta|x) \propto \theta^n e^{-\theta S} e^{-\lambda\theta}$ with $S = \sum_{i=1}^n x_i$.
   Gamma distribution with parameters $(n+1, S+\lambda)$.
   Equivalently, distributed as $Z/(S+\lambda)$ with $Z \sim \Gamma(n+1, 1)$.
   Thus $\pi(\Theta_0|x) = P(Z/(S+\lambda) > \frac{1}{2}) = P(Z > \frac{1}{2}(S+\lambda))$ and $\pi(\Theta_1|x) = P(Z \leq \frac{1}{2}(S+\lambda))$.

   $$\pi(\Theta_1|x) > \pi(\Theta_0|x) \quad \iff \quad P(Z > \frac{1}{2}(S+\lambda)) > P(Z \leq \frac{1}{2}(S+\lambda)) \quad \iff \quad m < \frac{1}{2}(S+\lambda),$$

   with m the median of Z.
   Bayesian test is:
   $$\delta(x) = 1_{\{S > 2m-\lambda\}}.$$

2. For $n = 6$, we get $m \approx 6.7$ and $2m - \lambda \approx 12.3$. Thus $S > 2m - \lambda$ and you reject the null hypothesis. For any value of λ, you get the same conclusion.

Exercise 3 (Bernoulli model – Two-tailed hypothesis)**:**

1. $\int \pi(\theta)\mathrm{d}\mu(\theta) = \frac{1}{2} + \frac{1}{2}\int_0^1 \mathrm{d}x = 1$

2. Let $\Theta_0 = \{\frac{1}{2}\}$, $\Theta_1 = [0,1] \setminus \{\frac{1}{2}\}$. Posterior distribution:

   $$\pi(\Theta_0|x) = \int_{\Theta_0} \pi(\theta)p(x|\theta)\mathrm{d}\mu(\theta) = \frac{1}{2}\binom{n}{s}\frac{1}{2^n}$$

   $$\pi(\Theta_1|x) = \int_{\Theta_1} \pi(\theta)p(x|\theta)\mathrm{d}\mu(\theta) = \frac{1}{2}\int_{\Theta_1}\binom{n}{s}\theta^s(1-\theta)^{n-s}\mathrm{d}\theta,$$

with $s = \sum_{i=1}^n x_i$. We get:

$$\frac{\pi(\Theta_1|x)}{\pi(\Theta_0|x)} = \frac{2^n}{B(s+1, n-s+1)},$$

with

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a+b-1)!}{(a-1)!(b-1)!}$$

for positive integers a, b, that is:

$$\frac{\pi(\Theta_1|x)}{\pi(\Theta_0|x)} = \frac{2^n s!(n-s)!}{(n+1)!}.$$

Bayesian test:

$$\delta(x) = 1_{\{\frac{2^n s!(n-s)!}{(n+1)!} > 1\}}.$$

3. You accept the null hypothesis for $s = 1$ and reject the null hypothesis for $s = 0$ or 2.

χ^2 tests

Exercise 4 (A dice – Test for fit)**:**

1. χ^2-test:
$$\delta(x) = 1_{\{T(x)>c\}},$$
where
$$T(x) = \sum_{k=1}^{6} \frac{(N_k - n\theta_k)^2}{n\theta_k}$$
and $N_k = \sum_{i=1}^{n} 1_{\{x_i=k\}}$ is the number of samples equal to k.
Level α:
$$c = Q(1-\alpha)$$
with Q the quantile function of the $\chi^2(5)$ distribution.

2. For $\alpha = 5\%$, $c \approx 11.1$ while $T(x) \approx 7.4$ so you accept the null hypothesis (the dice is fair).

Exercise 5 (Uniform model – Test for fit)**:**

1. χ^2-test:
$$\delta(x) = 1_{\{T(x)>c\}},$$
where
$$T(x) = \sum_{k=1}^{5} \frac{(N_k - np_k)^2}{np_k}$$
with $p = (2,3,3,2,2) \times \frac{1}{12}$.
Level α:
$$c = Q(1-\alpha)$$
with Q the quantile function of the $\chi^2(4)$ distribution.

2. A level $\alpha = 1\%$, $c \approx 13.3$ while $T(x) \approx 8.22$ so you accept the null hypothesis (the distribution is uniform).

Exercise 6 (Poisson model – Test for fit)**:** 1. Parameter estimation by ML:
$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
χ^2-test:
$$\delta(x) = 1_{\{T(x)>c\}},$$

where

$$T(x) = \sum_{k=1}^{4} \frac{(N_k - np_k)^2}{np_k}$$

with p_k the probability to be in the k-th range for a Poisson distribution with parameter $\widehat{\theta}$.
Level α:

$$c = Q(1-\alpha)$$

with Q the quantile function of the $\chi^2(2)$ distribution (since the parameter θ is of dimension 1).

2. A level $\alpha = 1\%$, $c \approx 9.2$ while $T(x) \approx 9.1$ (since $(p_1, p_2, p_3, p_4) \approx (0.16, 0.40, 0.33, 0.11)$) so you accept the null hypothesis (the distribution is Poisson).

Exercise 7 (Gaussian model – Test for fit)**:**

1. ML gives mean 100 and standard deviation 20.
χ^2-test:

$$\delta(x) = 1_{\{T(x)>c\}},$$

where

$$T(x) = \sum_{k=1}^{4} \frac{(N_k - np_k)^2}{np_k}$$

with $p \approx (0.16, 0.34, 0.34, 0.16)$.
Level α:

$$c = Q(1-\alpha)$$

with Q the quantile function of the $\chi^2(1)$ distribution (because of the parameter has dimension 2).

2. A level $\alpha = 1\%$, $c \approx 6.6$ while $T(x) \approx 0.57$ so you accept the null hypothesis.

Exercise 8 (Test for independence)**:**
χ^2-test for independence:

$$\delta(x) = 1_{\{T(x)>c\}},$$

where

$$T(x) = \sum_{i,j=1,2} \frac{(N_{ij} - \frac{N_i N_j}{n})^2}{\frac{N_i N_j}{n}}$$

Level α:

$$c = Q(1-\alpha)$$

with Q the quantile function of the $\chi^2(1)$ distribution for question 1 and $\chi^2(4)$ distribution for question 2.

1. Genre:

	Non-intensive	Intensive
Men	1 119	281
Women	1 128	102

At level $\alpha = 1\%$, $c \approx 6.6$ while $T(x) \approx 72$ so you reject independence.

Observe that you also reject the independence between genre and the fact to go to hospital for Covid-19 (assuming equal numbers of men and women in the population for simplicity):

$$T(x) = \sum_{i=1,2} \frac{(N_i - \frac{n}{2})^2}{\frac{n}{2}} \approx 2.7.$$

2. Age: at level $\alpha = 1\%$, $c \approx 13.3$ while $T(x) \approx 10\,880$ so you reject independence.