**1**

**Assume a sample of real numbers is a realization of independent repetitions of a random variable with Lebesgue density:**

$$f(x; a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \cdot \mathbf{1}_{(0,1)}(x)$$

**for parameter** $\vartheta = (a, b) \in (0, \infty) \times (0, \infty)$.

**a. Find a complete sufficient statistic.**
**b. Consider the maximum likelihood estimator problem.**
**c. Compute the method of moments estimator** $\hat{\vartheta}MM$ ***for*** $\vartheta$. ***Comment on it.***
***d. Use the delta method to consider the (two-dimensional) normal approximation of*** $\hat{\vartheta}MM$**: formulate a suitable limit theorem.**
**e. Using the normal approximation of** $\hat{\vartheta}_{MM}$**, construct an approximate 95% confidence region. Transform the problem into a multivariate normal distribution with a known variance-covariance matrix. Describe the resulting region as precisely as possible.**

1.a) $B(a,b) = \dfrac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$

Let $X = (X_1, \dots, X_n) \overset{iid}{\sim} B(a,b)$, $a, b > 0$.

The joint density is:

$$f(x_1, \dots, x_n; a, b) = \prod_{i=1}^{n} \left[ \frac{1}{B(a,b)} x_i^{a-1} (1-x_i)^{b-1} \mathbb{1}_{(0,1)}(x_i) \right]$$

$$= B(a,b)^{-n} \left( \prod_{i=1}^{n} x_i^{a-1} \right) \left( \prod_{i=1}^{n} (1-x_i)^{b-1} \right) \cdot \prod_{i=1}^{n} \mathbb{1}_{(0,1)}(x_i)$$

$$= B(a,b)^{-n} \exp\left[ (a-1) \sum_{i=1}^{n} \ln(x_i) + (b-1) \sum_{i=1}^{n} \ln(1-x_i) \right] \mathbb{1}_{(0,1)^n}(x)$$

Since we want to apply the Fisher-Neyman factorization theorem, we write

$$f(x; a, b) = h(x) \cdot g(T(x); a, b)$$

Where:

$$h(x) = \mathbb{1}_{(0,1)^n}(x)$$

$$T(x) = (T_1, T_2) = \left( \sum_{i=1}^{n} \ln x_i ; \sum_{i=1}^{n} \ln(1-x_i) \right)$$

$$g(T(x); a, b) = B(a,b)^{-n} \cdot \exp\left( (a-1) T_1 + (b-1) T_2 \right)$$

So by the Fisher - Neyman factorization theorem, the statistic

$$T(X) = \left( \sum_{i=1}^{n} \ln X_i \, ; \, \sum_{i=1}^{n} \ln(1-X_i) \right)$$

is sufficient for $\theta = (a,b)$

To show that $T = (T_1, T_2)$ is complete we need to show for any measurable function $g(T_1, T_2)$ that

$$\mathbb{E}_{a,b}\left[ g(T_1, T_2) \right] = 0 \quad \forall a,b \implies g(T_1, T_2) = 0 \quad \text{almost surely.}$$

But we can as well show that the Beta distribution is part of the exponential family of distribution:

$$f(x, a, b) = \exp\left[ (a-1) \ln x + (b-1) \ln(1-x) - \ln B(a,b) \right] \cdot \mathbb{1}_{(0,1)}(x)$$

$$= h(x) \exp\left( \langle Q(a,b), T(x) \rangle - A(a,b) \right)$$

with:

$$h(x) = \mathbb{1}_{(0,1)}(x)$$

$$Q(a,b) = (a-1, \, b-1)$$

$$T(x) = (\ln(x) \, , \, \ln(1-x))$$

$$A(a,b) = \ln B(a,b)$$

Moreover, the natural parameters $(a-1)$ and $(b-1)$ can vary over an open space in $\mathbb{R}^2$, the Beta distribution is a full-rank exponential family. Hence, the statistic is complete.

The complete sufficient statistic for $\theta = (a,b)$ is

$$\boxed{T(X) = \left( \sum_{i=1}^{n} \ln X_i \, ; \, \sum_{i=1}^{n} \ln(1-X_i) \right)}$$

1.b) Let us compute the Maximum Likelihood Estimator (ML E) of $\theta = (a,b) \in (0,\infty) \times (0,\infty)$ for a sample :

$$X = (X_1, \dots, X_n) \overset{iid}{\sim} \mathcal{B}(a,b), \quad a, b > 0.$$

with density :

$$f(x; a,b) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}, \quad x \in (0,1)$$

The likelihood is

$$L(\theta) = \prod_{i=1}^{n} f(x_1, \dots, x_n ; \theta)$$

The log-likelihood is

$$\ell(\theta) = \ln\left(L(\theta)\right)$$
$$= -n \ln\left(B(a,b)\right) + (a-1) \underbrace{\sum_{i=1}^{n} \ln(x_i)}_{S_1} + (b-1) \underbrace{\sum_{i=1}^{n} \ln(1-x_i)}_{S_2}$$

Now we look for :

$$\hat{\theta}_{ML} = \underset{\theta \in (0,\infty)^2}{\operatorname{argmax}} \ \ell\left(f(x,\theta)\right)$$

We know that :

$$\frac{\partial}{\partial a} B(a,b) = \Gamma(b) \cdot \frac{\Gamma'(a)\Gamma(a+b) - \Gamma'(a+b)\Gamma(a)}{\Gamma(a+b)^2}$$

$$= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \frac{\Gamma'(a+b)}{\Gamma(a+b)}$$

$$= B(a,b)\left(\frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)}\right)$$

$$= B(a,b)\left(\Psi(a) - \Psi(a+b)\right)$$

with $\Psi : z \mapsto \frac{d}{dz} \ln \Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$

We can write :

$$\ell(x_1,\dots,x_n;a,b) = n\left(\ln\Gamma(a+b) - \ln\Gamma(a) - \ln\Gamma(b)\right) + (a-1)S_1 + (b-1)S_2$$

So computing the partial derivatives

$$\frac{\partial \ell}{\partial a} = n\left(\psi(a+b) - \psi(a)\right) + S_1$$

$$\frac{\partial \ell}{\partial b} = n\left(\psi(a+b) - \psi(b)\right) + S_2$$

To find $\hat{\theta}_{ML}$, we need to solve :

$$\begin{cases} \dfrac{\partial \ell}{\partial a} = n\left(\psi(a+b) - \psi(a)\right) + S_1 = 0 \\ \dfrac{\partial \ell}{\partial b} = n\left(\psi(a+b) - \psi(b)\right) + S_2 = 0 \end{cases}$$

$$\Longleftrightarrow \begin{cases} \psi(a) - \psi(a+b) = \dfrac{1}{n}\sum_{i=1}^{n} \ln X_i \\ \psi(b) - \psi(a+b) = \dfrac{1}{n}\sum_{i=1}^{n} \ln(1-X_i) \end{cases}$$

This system doesn't have algebraic solution for $a,b$.

Numerically (check exercise 1.ipynb), on the data provided:

$\hat{a}_{ML} = 2.432755$

$\hat{b}_{ML} = 3.81518$

1.c) Method of moments.

To estimate $\theta = (a,b)$, we can use the first and second moment to estimate $a$ and $b$.

For a beta distribution $B(a,b)$, the population first and second moment are given by:

$$\mu_1 = \mathbb{E}[X] = \frac{a}{a+b}$$

$$\mu_2 = \mathbb{E}[X^2] = \frac{a(a+1)}{(a+b)(a+b+1)}$$

Suppose a sample size of $n$ is drawn, resulting in the values $X_1, \dots, X_n$. Let

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

be the first and second sample moment.

Now, we solve:

$$\frac{a}{a+b} = \hat{\mu}_1 \quad \Rightarrow \quad a = \hat{\mu}_1 (a+b) = \hat{\mu}_1 S$$

$$\Rightarrow \quad b = \frac{a(1-\hat{\mu}_1)}{\mu_1}$$

with $S = a+b$

$$= (a+b)(1-\hat{\mu}_1) = (1-\hat{\mu}_1) S$$

Inject into the second moment equation:

$$\hat{\mu}_2 = \frac{\hat{\mu}_1 S(\hat{\mu}_1 S + 1)}{S(S+1)} = \frac{\hat{\mu}_1 (\hat{\mu}_1 S + 1)}{S+1}$$

$$\Rightarrow \quad \hat{\mu}_2 S + \hat{\mu}_2 = \hat{\mu}_1^2 S + \hat{\mu}_1$$

$$\Rightarrow \quad S(\hat{\mu}_2 - \hat{\mu}_1^2) = \hat{\mu}_1 - \hat{\mu}_2$$

$$\Rightarrow \quad S = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\mu}_2 - \hat{\mu}_1^2}$$

Substitute back in expressions for $a$ and $b$:

$$\hat{a} = \frac{\hat{\mu}_1 (\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\mu}_2 - \hat{\mu}_1^2} \quad \text{and} \quad \hat{b} = \frac{(\hat{\mu}_1 - \hat{\mu}_2)(1 - \hat{\mu}_1)}{\hat{\mu}_2 - \hat{\mu}_1^2}$$

Therefore:

$$\boxed{\hat{\theta}_{MM} = \left( \frac{\hat{\mu}_1 (\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\mu}_2 - \hat{\mu}_1^2} , \frac{(\hat{\mu}_1 - \hat{\mu}_2)(1 - \hat{\mu}_1)}{\hat{\mu}_2 - \hat{\mu}_1^2} \right)}$$

Comments:

- Here we have an explicit estimator, straightforward to compute from distribution sample.

- This estimator is consistent as long as the sample size is sufficiently large (relies on the strong law of numbers), but may be less efficient than the MLE estimator.

To conclude, this estimator, simple to compute, provides a intuitive way to compute the parameters a and b. But it may not be as efficient and robust as the MLE since this method is sensitive to outliers and to small sample sizes.

Numerically we find:

$\hat{a}_{MM} = 2.39$

$\hat{b}_{MM} = 3.77$

1.e) We will use delta method to find the asymptotic distribution of $\hat{\theta}_{MM}$.

By the Central Limit Theorem, for we have

$$\sqrt{n}\left(\begin{pmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right)$$

Where $\mu_1 = \mathbb{E}[X] = \dfrac{a}{a+b}$ and $\mu_2 = \mathbb{E}[X^2] = \dfrac{a(a+1)}{(a+b)(a+b+1)}$

and $\Sigma$ is the asymptotic covariance matrix of the sample moments. $\Sigma$ can be derived from the Beta distribution:

$$\Sigma = \begin{pmatrix} Var(X) & Cov(X, X^2) \\ Cov(X, X^2) & Var(X^2) \end{pmatrix}$$

with $Var(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$

$$Cov(X, X^2) = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2]$$

$$Var(X^2) = \mathbb{E}[X^4] - (\mathbb{E}[X^2])^2$$

$Cov(X,Y)$

$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

$\Sigma$ depends on a and b.

Let us apply the Delta Method to the function

$$g: \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \mapsto \begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix} = \left(\frac{\mu_1(\mu_1 - \mu_2)}{\mu_2 - \mu_1^2}, \frac{(\mu_1 - \mu_2)(1-\mu_1)}{\mu_2 - \mu_1^2}\right)$$

g is differentiable on $\mathbb{R}^2$, so the jacobian of g is:

$$\nabla g(\mu_1, \mu_2) = \begin{pmatrix} \dfrac{\partial \hat{a}}{\partial \mu_1} & \dfrac{\partial \hat{a}}{\partial \mu_2} \\ \dfrac{\partial \hat{b}}{\partial \mu_1} & \dfrac{\partial \hat{b}}{\partial \mu_2} \end{pmatrix}$$

$$\frac{\partial \hat{a}}{\partial \mu_1} = \frac{\mu_1 - \mu_2}{\mu_2 - \mu_1^2} + \frac{\mu_1\left(\mu_2 - \mu_1^2\right) - \mu_1\left(\mu_1 - \mu_2\right)\left(-2\mu_1\right)}{\left(\mu_2 - \mu_1^2\right)^2}$$

$$= \frac{\mu_1\mu_2 - \cancel{\mu_1^3} - \mu_2^2 + \mu_2\mu_1^2 + \mu_1\mu_2 - \cancel{\mu_1^3} + 2\cancel{\mu_1^3} - 2\mu_1^2\mu_2}{\left(\mu_2 - \mu_1^2\right)^2} = \frac{2\mu_1\mu_2 - \mu_1^2\mu_2 - \mu_2^2}{\left(\mu_2 - \mu_1^2\right)^2}$$

$$\frac{\partial \hat{a}}{\partial \mu_2} = \mu_1 \frac{-\left(\mu_2 - \mu_1^2\right) - \left(\mu_1 - \mu_2\right)}{\left(\mu_2 - \mu_1^2\right)^2}$$

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2} \Rightarrow \left| \begin{array}{l} u = f \cdot g \\ u' = f' g + f g' \end{array} \right.$$

$$\frac{\partial \hat{b}}{\partial \mu_1} = \frac{\left(\left(\mu_2 - \mu_1\right) + \left(1 - \mu_1\right)\right)\left(\mu_2 - \mu_1^2\right) - \left(1 - \mu_1\right)\left(\mu_1 - \mu_2\right)\left(-2\mu_1\right)}{\left(\mu_2 - \mu_1^2\right)^2}$$

$$= \frac{\mu_2^2 - \mu_1\mu_2 + \mu_2 - \mu_1\mu_2 - \mu_1^2\mu_2 + \cancel{\mu_1^3} - \mu_1^2 + \cancel{\mu_1^3} + 2\mu_1^2 - 2\mu_1\mu_2 - 2\cancel{\mu_1^3} + 2\mu_1^2\mu_2}{\left(\mu_2 - \mu_1^2\right)^2}$$

$$= \frac{\mu_1^2 + \mu_1^2\mu_2 + \mu_2^2 - 4\mu_1\mu_2 + \mu_2}{\left(\mu_2 - \mu_1^2\right)^2}$$

$$\frac{\partial \hat{b}}{\partial \mu_2} = \frac{-\left(1 - \mu_1\right)\left(\mu_2 - \mu_1^2\right) - \left(1 - \mu_1\right)\left(\mu_1 - \mu_2\right) \times 1}{\left(\mu_2 - \mu_1^2\right)^2} = \frac{-\mu_2 + \mu_1^2 + \mu_1\mu_2 - \mu_1^3 - \mu_1 + \mu_1\mu_2 + \mu_1^2 - \mu_1\mu_2}{\left(\mu_2 - \mu_1^2\right)^2}$$

$$= \frac{-\mu_1^3 + 2\mu_1^2 + \mu_1\mu_2 - \mu_1 - \mu_2}{\left(\mu_2 - \mu_1^2\right)^2}$$

Anticipating but we can see that $\left(\frac{\partial \hat{a}}{\partial \mu_1}, \frac{\partial \hat{a}}{\partial \mu_2}\right)$ and $\left(\frac{\partial \hat{b}}{\partial \mu_1}, \frac{\partial \hat{b}}{\partial \mu_2}\right)$ aren't colinear, i.e.

$$\left\langle \left(\frac{\partial \hat{a}}{\partial \mu_1}, \frac{\partial \hat{a}}{\partial \mu_2}\right), c \right\rangle \neq \left(\frac{\partial \hat{b}}{\partial \mu_1}, \frac{\partial \hat{b}}{\partial \mu_2}\right) \qquad \forall c \in \mathbb{R}^2$$

Thus, the rank of $\nabla g$ is 2 and $\nabla g$ is full rank.

We have that $g(\hat{\mu}_1, \hat{\mu}_2) = \hat{\theta}_{MM}$. Since the Central Limit Theorem gives $\sqrt{n}\left(\hat{\theta}_{MM} - \theta\right) \xrightarrow{d} \mathcal{N}(0_{\mathbb{R}^2}, \Sigma)$, then the delta method states:

$$\sqrt{n}\left( g\begin{pmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{pmatrix} - g\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(0_{\mathbb{R}^2}, \nabla g(\mu_1,\mu_2)^T \Sigma \nabla g(\mu_1,\mu_2)\right)$$

Conclusion:

$$\boxed{\sqrt{n}\left(\begin{pmatrix}\hat{a}_{MM}\\ \hat{b}_{MM}\end{pmatrix} - \begin{pmatrix}a\\ b\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \nabla g(\mu_1,\mu_2)^T \Sigma \nabla g(\mu_1,\mu_2)\right)}$$

1.d) We want to compute an approximate 95% confidence region for $\hat{\theta}_{MM} = (\hat{a}_{MM}, \hat{b}_{MM})$ using the normal approximation above.

Denoting $V = \nabla g(\mu_1,\mu_2)^T \Sigma \nabla g(\mu_1,\mu_2)$, and remembering the quadratic form of a multi-variate normal: if $Z \sim \mathcal{N}(0, V)$ then

$$Z^T V^{-1} Z \sim \chi^2_2$$

Substituting $Z = \sqrt{n}\left(\hat{\theta}_{MM} - \theta\right)$ gives:

$$n\left(\hat{\theta}_{MM} - \theta\right)^T V^{-1} \left(\hat{\theta}_{MM} - \theta\right) \sim \chi^2_2$$

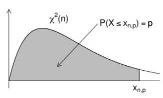
But we have to check if $V$ is invertible as $\Sigma$ is. $*$

The 95% confidence region for $\theta = (a,b)$ is the ellipsoid defined by:

**Loi du khi-deux**

Table des quantiles de $X \sim \chi^2(n)$ :

$$\mathbb{P}\left(X \leq x_{n,p}\right) = p.$$




| $n$ \ $p$ | 0.005 | 0.01 | 0.025 | 0.05 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.95 | 0.975 | 0.99 | 0.995 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.10 | 0.45 | 1.32 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 |
| 2 | 0.01 | 0.02 | 0.05 | 0.10 | 0.21 | 0.58 | 1.39 | 2.77 | 4.61 | 5.99 | 7.38 | 9.21 | 10.60 |
| 3 | 0.07 | 0.11 | 0.22 | 0.35 | 0.58 | 1.21 | 2.37 | 4.11 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 |
| 4 | 0.21 | 0.30 | 0.48 | 0.71 | 1.06 | 1.92 | 3.36 | 5.39 | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 |
| 5 | 0.41 | 0.55 | 0.83 | 1.15 | 1.61 | 2.67 | 4.35 | 6.63 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 |
| 6 | 0.68 | 0.87 | 1.24 | 1.64 | 2.20 | 3.45 | 5.35 | 7.84 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 |
| 7 | 0.99 | 1.24 | 1.69 | 2.17 | 2.83 | 4.25 | 6.35 | 9.04 | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 3.49 | 5.07 | 7.34 | 10.22 | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 |
| 9 | 1.73 | 2.09 | 2.70 | 3.33 | 4.17 | 5.90 | 8.34 | 11.39 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 |
| 10 | 2.16 | 2.56 | 3.25 | 3.94 | 4.87 | 6.74 | 9.34 | 12.55 | 15.99 | 18.31 | 20.48 | 23.21 | 25.19 |
| 11 | 2.60 | 3.05 | 3.82 | 4.57 | 5.58 | 7.58 | 10.34 | 13.70 | 17.28 | 19.68 | 21.92 | 24.72 | 26.76 |
| 12 | 3.07 | 3.57 | 4.40 | 5.23 | 6.30 | 8.44 | 11.34 | 14.85 | 18.55 | 21.03 | 23.34 | 26.22 | 28.30 |
| 13 | 3.57 | 4.11 | 5.01 | 5.89 | 7.04 | 9.30 | 12.34 | 15.98 | 19.81 | 22.36 | 24.74 | 27.69 | 29.82 |
| 14 | 4.07 | 4.66 | 5.63 | 6.57 | 7.79 | 10.17 | 13.34 | 17.12 | 21.06 | 23.68 | 26.12 | 29.14 | 31.32 |
| 15 | 4.60 | 5.23 | 6.26 | 7.26 | 8.55 | 11.04 | 14.34 | 18.25 | 22.31 | 25.00 | 27.49 | 30.58 | 32.80 |
| 16 | 5.14 | 5.81 | 6.91 | 7.96 | 9.31 | 11.91 | 15.34 | 19.37 | 23.54 | 26.30 | 28.85 | 32.00 | 34.27 |
| 17 | 5.70 | 6.41 | 7.56 | 8.67 | 10.09 | 12.79 | 16.34 | 20.49 | 24.77 | 27.59 | 30.19 | 33.41 | 35.72 |
| 18 | 6.26 | 7.01 | 8.23 | 9.39 | 10.86 | 13.68 | 17.34 | 21.60 | 25.99 | 28.87 | 31.53 | 34.81 | 37.16 |
| 19 | 6.84 | 7.63 | 8.91 | 10.12 | 11.65 | 14.56 | 18.34 | 22.72 | 27.20 | 30.14 | 32.85 | 36.19 | 38.58 |
| 20 | 7.43 | 8.26 | 9.59 | 10.85 | 12.44 | 15.45 | 19.34 | 23.83 | 28.41 | 31.41 | 34.17 | 37.57 | 40.00 |
| 21 | 8.03 | 8.90 | 10.28 | 11.59 | 13.24 | 16.34 | 20.34 | 24.93 | 29.62 | 32.67 | 35.48 | 38.93 | 41.40 |
| 22 | 8.64 | 9.54 | 10.98 | 12.34 | 14.04 | 17.24 | 21.34 | 26.04 | 30.81 | 33.92 | 36.78 | 40.29 | 42.80 |
| 23 | 9.26 | 10.20 | 11.69 | 13.09 | 14.85 | 18.14 | 22.34 | 27.14 | 32.01 | 35.17 | 38.08 | 41.64 | 44.18 |
| 24 | 9.89 | 10.86 | 12.40 | 13.85 | 15.66 | 19.04 | 23.34 | 28.24 | 33.20 | 36.42 | 39.36 | 42.98 | 45.56 |
| 25 | 10.52 | 11.52 | 13.12 | 14.61 | 16.47 | 19.94 | 24.34 | 29.34 | 34.38 | 37.65 | 40.65 | 44.31 | 46.93 |
| 26 | 11.16 | 12.20 | 13.84 | 15.38 | 17.29 | 20.84 | 25.34 | 30.43 | 35.56 | 38.89 | 41.92 | 45.64 | 48.29 |
| 27 | 11.81 | 12.88 | 14.57 | 16.15 | 18.11 | 21.75 | 26.34 | 31.53 | 36.74 | 40.11 | 43.19 | 46.96 | 49.64 |
| 28 | 12.46 | 13.56 | 15.31 | 16.93 | 18.94 | 22.66 | 27.34 | 32.62 | 37.92 | 41.34 | 44.46 | 48.28 | 50.99 |
| 29 | 13.12 | 14.26 | 16.05 | 17.71 | 19.77 | 23.57 | 28.34 | 33.71 | 39.09 | 42.56 | 45.72 | 49.59 | 52.34 |
| 30 | 13.79 | 14.95 | 16.79 | 18.49 | 20.60 | 24.48 | 29.34 | 34.80 | 40.26 | 43.77 | 46.98 | 50.89 | 53.67 |
| 40 | 20.71 | 22.16 | 24.43 | 26.51 | 29.05 | 33.66 | 39.34 | 45.62 | 51.81 | 55.76 | 59.34 | 63.69 | 66.77 |
| 50 | 27.99 | 29.71 | 32.36 | 34.76 | 37.69 | 42.94 | 49.33 | 56.33 | 63.17 | 67.50 | 71.42 | 76.15 | 79.49 |
| 60 | 35.53 | 37.48 | 40.48 | 43.19 | 46.46 | 52.29 | 59.33 | 66.98 | 74.40 | 79.08 | 83.30 | 88.38 | 91.95 |
| 70 | 43.28 | 45.44 | 48.76 | 51.74 | 55.33 | 61.70 | 69.33 | 77.58 | 85.53 | 90.53 | 95.02 | 100.4 | 104.2 |
| 80 | 51.17 | 53.54 | 57.15 | 60.39 | 64.28 | 71.14 | 79.33 | 88.13 | 96.58 | 101.9 | 106.6 | 112.3 | 116.3 |
| 90 | 59.20 | 61.75 | 65.65 | 69.13 | 73.29 | 80.62 | 89.33 | 98.65 | 107.6 | 113.1 | 118.1 | 124.1 | 128.3 |
| 100 | 67.33 | 70.06 | 74.22 | 77.93 | 82.36 | 90.13 | 99.33 | 109.1 | 118.5 | 124.3 | 129.6 | 135.8 | 140.1 |

$$\boxed{n\left(\hat{\theta}_{MM} - \theta\right)^T V^{-1} \left(\hat{\theta}_{MM} - \theta\right) \leqslant 5.99}$$

(*) To verify this, we have shown that $\nabla g$ is of full rank, then

$$V = \nabla g(\mu_1, \mu_2)^T \Sigma \nabla g(\mu_1, \mu_2)$$

is symetric positive-definite, then invertible.

An ellipsoid, defined as

$$(x - v)^T A (x - v) = 1$$

where:

- $x \in \mathbb{R}^2$ are the coordinates $(x_1, x_2)$
- $v \in \mathbb{R}^2$ is the center of the ellipsoid
- $A \in \mathbb{R}^{2\times 2}$ is a symetric positive-definite matrix. The eigenvectors of $A$ give the direction of the major and minor axes, and the eigenvalues give the length of these axes.

In our case, $(\hat{a}_{MM}, \hat{b}_{MM})$ define the center of the ellipse, the shape and orientation is defined by matrix $V$. Finally, the value of the quantile of the $\chi^2_2$ defines the size of the ellipsoid, here 5.99. If we had chosen a 99 % confidence interval, the quantile of the Chi Squared (9.21) is bigger, hence resulting is a bigger zone since we want to have more confidence in the fact that the real $\theta = (a, b)$ lies in the ellipsoid.

↳ Ellipsoid plotted on the notebook

## 2

For two independent samples $X_1, , X_m$ (i.i.d. Bernoulli with parameter $p$) and $Y_1, , Y_n$ (i.i.d. Bernoulli with parameter $q$), we test the hypothesis $p = q$ using a chi-squared test of homogeneity at significance level $\alpha$.

a. For $m = 95$, $n = 100$, and $\alpha = 0.1$, compute or estimate the size of the test as precisely as possible.
b. Suppose we want power at least 0.85 for $|p - q| > 0.05$. Determine the required sample sizes (assume $m = n$).
c. For $m = 95$, $n = 100$, use the same test statistic but determine a critical value such that the test has an exact significance level $\alpha = 0.1$.
d. (*) The asymptotic distribution of the test statistic (under $p = q$) is derived using Wilks' theorem. Determine the asymptotic distribution of the test statistic for general $p$ and $q$.

a) Formally, we want to test the null hypothesis $H_0 : p = q$ against $H_1 : p \neq q$.

The outcome of two Bernoulli samples can be summarized in this table:

| | Success | Failure | Total |
|---|---|---|---|
| Sample X | $S_X$ | $m - S_X$ | $m$ |
| Sample Y | $S_Y$ | $n - S_Y$ | $n$ |
| Total | $S = S_X + S_Y$ | $(m+n) - S$ | $m + n$ |

$r$ population : X and Y (2)
$c$ levels : S and F (2)

with $S_X = \sum_{i=1}^{m} X_i$ the number of successes in the first sample.
$S_Y = \sum_{i=1}^{n} Y_i$ the number of success in the second sample.

We now compute the expected counts under the null hypothesis $H_0$:

$$E_{S_X} = \frac{m \cdot S}{m+n} \qquad E_{S_Y} = \frac{n \cdot S}{m+n}$$

$$E_{F_X} = m - E_{S_X} \qquad E_{F_Y} = n - E_{S_Y}$$

The Degrees of freedom of our $\chi^2$ test is given by:

$$df = (r-1)(c-1) = (2-1)(2-1) = 1$$

The chi-squared test statistic is:

$$\chi^2 = \frac{(S_x - E_{S_x})^2}{E_{S_x}} + \frac{(S_y - E_{S_y})^2}{E_{S_y}} + \frac{(F_x - E_{F_y})^2}{E_{F_y}} + \frac{(F_y - E_{F_y})^2}{E_{F_y}}$$

The quantile or critical value for $\alpha = 0.1$ for a $\chi^2_1$ is

$$\chi^2_{1,1-\alpha} = \chi^2_{1,0.9} \approx 2.71$$

The size of test is the supremum probability of rejecting $H_0$ when it is true,

$$\sup_{p,q \in [0,1]} \mathbb{P}\left(\text{reject } H_0 \mid H_0 \text{ is true}\right)$$

In order to find it we must run simulations. (cf. exercice2-R.ipynb)

After simulating for $p = q = \pi$ with $\pi$ ranging from $0.1$ to $0.9$,
we have that the empirical size approximately $0.10$ but can be higher for $\pi$ in the extremes ($0.1$ or $0.9$). I founded that the supremum here is $0.1067$ for $\pi = 0.8$

2.b) The type II error is the failure to reject a false $H_0$, ie:

$$\beta = \mathbb{P}\left(\text{Fail to reject } H_0 \mid H_1 \text{ is true}\right)$$

Power $= 1 - \beta$.

We want power $\geqslant 0.85$, which means we want at most $15\%$ chance of missing a real difference of $5\%$ ($|p-q| = 0.05$)

The question is, how large each sample must be so that the test rejects $H_0 : p = q$ at least $85\%$ of the time when the true difference is $5\%$ $|p-q| = 0.05$.

With the chi squared test of homogenity, I don't think we could isolate $m$ $(=n)$ to have a explicit value, as we can do for other tests. We will approach this $m$ numerically:

For differents $m$:

1. Generate $X \sim \text{Binomial}(m, p)$ $\quad Y \sim \text{Binomial}(m, q)$
2. Compute the test statistic $\chi^2$.
3. Compare it to the critical value $\chi^2_1 \approx 2.71$ $(\alpha = 0.10)$ or the p-value to alpha
4. The rate of rejections (for 10 000 simulation) is the power.

To fix $p$ and $q$, since we are under $\mathcal{H}_1$, we want the case where it is the harder to tell the two proportion apart. This is when the variance of the Binomial is maximum $(p(1-p))$.

→ The Binomial variance is maximized near $p = 0.5$ $\quad (\mathrm{Var}(X) = p(1-p))$

A choice would be:

$$p = 0.5 + \frac{0.05}{2} = 0.525$$

$$q = 0.5 - \frac{0.05}{2} = 0.475$$

In exercice 2-R.ipynb, I find $\hat{m}$ such that:

$$\hat{m} = \underset{m \in [100;\, 2000]}{\operatorname{argmin}} \left| \mathbb{P}\left(\text{reject } \mathcal{H}_0 \mid p = 0.525,\ q = 0.475,\ n = m\right) - 0.85 \right|$$

With 50 000 simulations.

The minim $m = n$ to ensure that the power of our test is 85 % is $\hat{m} \approx 1448$.

2.c) Now, with the same test statistic $\chi^2$, we want to find $c$ such that

$$\sup_{p \in [0,1]} \mathbb{P}(\chi^2 \geq c) = 0.1$$

Here, we compute $\mathbb{P}(\chi^2 \geq c)$ over $p = q \in \{0.1, 0.2, \dots, 0.9\}$ and find the $c$ that minimizes $\left| \sup_{p=q \in [0,1]} \mathbb{P}(\chi^2 \geq c) - 0.1 \right|$ with 50 000 simulations.

This gives $c = 2.745$ for a size of $0.1 \pm 0.003$

1, d) We have two Bernoulli samples, with parameters $(p, q)$.

- Under $H_1$ : $\ell(p, q)$ is the loglikelihood of the model under $H_1$
- Under $H_0$ : $\ell(\pi)$ is the loglikelihood of the model under $H_0$

The MLE under $H_1$ is $\hat{p}, \hat{q}$ and under $H_0$ is $\hat{\pi}$.

The likelihood ratio test statistic is:

$$T = 2\left(\ell(\hat{p}, \hat{q}) - \ell(\hat{\pi})\right)$$

Under regularity conditions, assuming $H_0 : p = q$ is true, Wilk's theorem says

$$T \xrightarrow{d} \chi^2_r \qquad \text{with } r = \underset{\hat{p},\hat{q}}{2} - \underset{\hat{\pi}}{1} = 1$$

the number of restrictions (difference in dimension between the alternative and the null hypothesis)

Hence, asymptotically, we can compare $T$ to $\chi^2_{1,1-\alpha}$ critical value to perform the test at level $\alpha = 0.1$.

We can make a link between the $\chi^2$ test of homogeneity for a $2 \times 2$ table and the LRT statistic $T$ as they are asymptotically equivalent. Both tests, under $H_0$ goes to $\chi^2(1)$ for large samples.

3) We can't assume normality so we use rank tests.

We want to test if the distribution of two independent samples is equal.

→ Kolmogorov-Smirnov test is not base on the rank

→ Wilcoxon - Mann - Whitney Test is based on ranks and meet the requirements.

In the course there is the Mann-Whitney U test and the Wilcoxon rank sum test.

So we have:

- Sample 1: $X_1, \dots, X_{n_1}$ with $n_1 = 10$. These are iid from a non-normal distribution $F$ with mean $\mu_1$ and variance $\sigma^2 = 1$
- Sample 2: $Y_1, \dots, Y_{n_2}$ with $n_1 = 15$. These are iid from a non-normal distribution $G$ with mean $\mu_2$ and variance $\sigma^2 = 1$

The hypothesis to be tested is

$$H_0 : F = G \qquad \text{against} \qquad H_1 : F \neq G.$$

We use the two sided Wilcoxon - Mann - Whitney test in the exact form.

$$U_{(n_1, n_2)} = \min \left( U^{X<Y}_{(n_1, n_2)} , U^{X>Y}_{(n_1, n_2)} \right)$$

where

$$U^{X<Y}_{(n_1, n_2)} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}_{X_i < Y_j}$$

$$U^{X>Y}_{(n_1, n_2)} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}_{X_i > Y_j} = n_1 n_2 - U^{X<Y}_{(n_1, n_2)}$$

The law is of $U_{(n_1, n_2)}$ under $H_0$ is then computed by reccurence and $U_{(n_1, n_2)}$ is then compared to the critical value of this distribution $c$ that verifies, at level $\alpha$:

$$\mathbb{P}_{H_0}(U \leq c) \leq \alpha$$

Then we reject $H_0$ if the observed $U \leq c$ because $U$ is the smaller of the two possible counts.

Now, with the power function, we want to know to know $\mathbb{P}(\text{reject } \mathcal{H}_0)$ given a true difference between the two distributions.

Since $\sigma^2 = 1$ for both distributions, then their only difference is the shift $\Delta$.

So we can define the power function as:

$$\text{Power}(\Delta) = \mathbb{P}\left(\text{reject } \mathcal{H}_0 \;\middle|\; \mu_1 - \mu_2 = \Delta\right) \quad , \quad \Delta \neq 0$$
$$= \mathbb{P}\left(\min\left(U^{X<Y}_{n_1,n_2}, U^{X>Y}_{n_1,n_2}\right) \geq u_{1-\alpha} \;\middle|\; \mu_1 - \mu_2 = \Delta\right)$$

This function depends on the level of the test $\alpha$ and the shift $\Delta$.

Then, to analyze the power function analytically, I would need to choose a distribution (lognormal, gamma, etc) such that $Y \overset{d}{=} X + \Delta$.

Then compute the distribution of the test statistic $U$ by reccurrence.

Then integrate that distribution in the rejection region to have the power as a function of $\alpha$ and $\Delta$.

In the last part, I believe it would be nearly impossible to find a close form so let us use a numeric approach (exercise 3. ipynb).

## 4

a. Assume a model with $n$ independent repetitions of a Bernoulli random variable with parameter $p$. Consider a family of tests $\phi_{p_0} : \mathbb{R}^n \to [0,1]$ for the point null hypothesis $H_0 : p = p_0$ against alternative $A : p \neq p_0$, at significance level 0.05, of the form:

$$\phi_{p_0}(x_1, , x_n) = \begin{cases} 1, & \sum x_i < C_1(p_0) \text{ or } \sum x_i > C_2(p_0) \\ 0, & \text{otherwise} \end{cases}$$

Your diagram shows the functions $C_1(p_0)$ and $C_2(p_0)$ for $n = 20$. Use inversion to construct a confidence interval for the Bernoulli parameter for your $n$. →

b. Let $C$ be a confidence region for parameter $\vartheta \in \Theta$. The coverage at $\vartheta_0 \in \Theta$ is $P_{\vartheta_0}(\{\mathbf{X} | \vartheta_0 \in C(\mathbf{X})\})$. The confidence level is the exact lower bound of this coverage. For the confidence interval from part a, draw the coverage graph (do your best!) and estimate the confidence coefficient. You may also calculate it exactly.

b) We have that:

$$\text{Coverage}(p_0) = \mathbb{P}_{p_0}\left(\{X \mid p_0 \in C(X)\}\right)$$

From part one, we built confidence intervals based on:

$$S = \sum x_i \sim \text{Bin}(n=20, p).$$

We inverted this test and for each observed value $s$, we found all $p_0$ that verify

$$C_1(p_0) \leq s \leq C_2(p_0)$$

Here, on the other hand, we fix $p_0$ and check for which $s \in \{0, .., 20\}$ does $p_0$ belongs to the confidence intervals (in red on the graph).

So the process is for each $p_0$, capture all $s$ (on the diagram ← in green) such that $p_0$ falls into their CI.

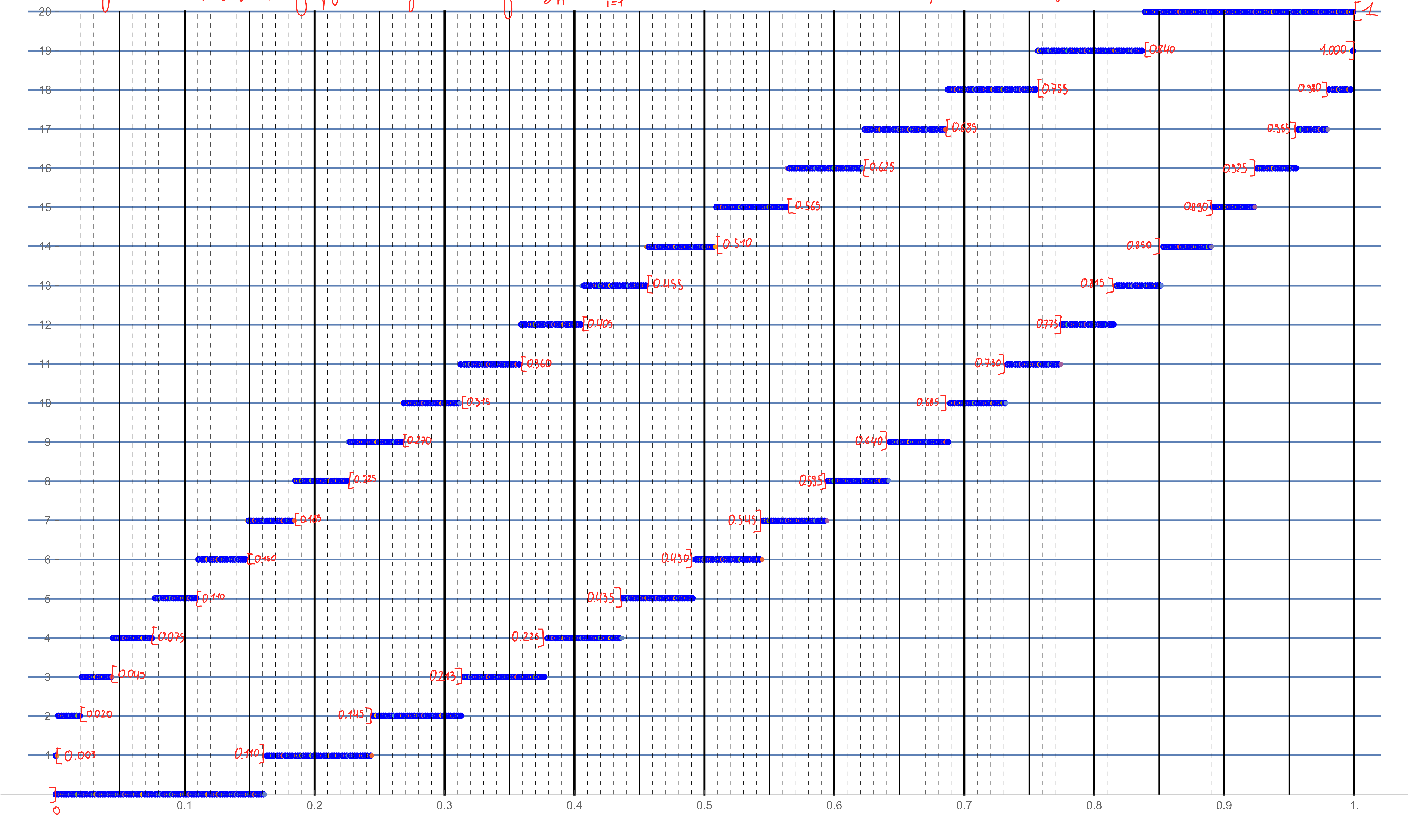
Then compute:

$$\text{Coverage}(p_0) = \sum_s \mathbb{P}(S = s) = \sum_s \binom{20}{s} p^s (1-p)^{20-s}$$

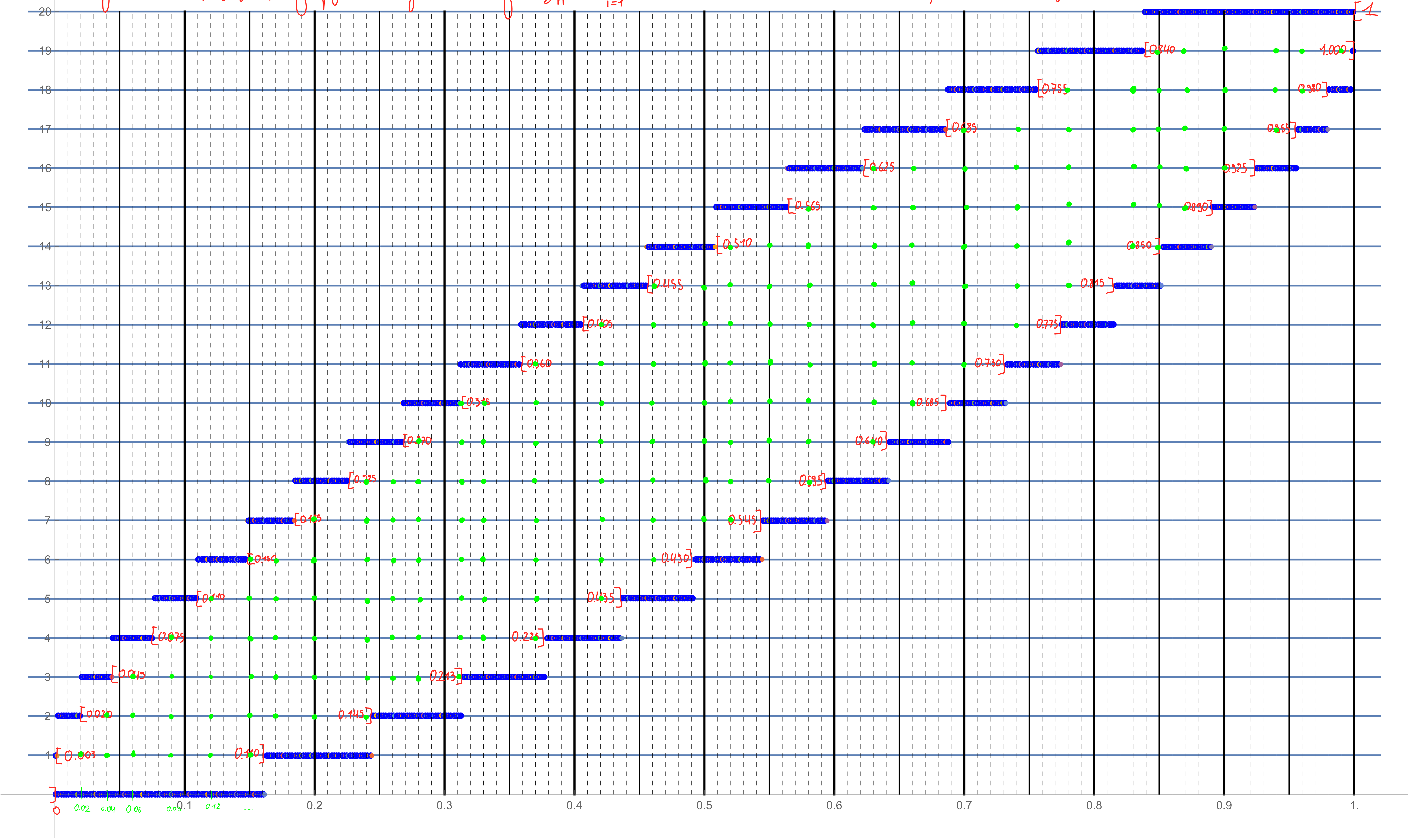
Finally, plot the coverage against $p_0$ and extract the minimum on the curve to have the confidence coefficient:

I will record the green dots in exercise4.csv and plot the coverage function in exercise4.ipynb

Confidence intervals of $p_0$ as a function of $S_n = \sum_{i=1}^{n} x_i$ with $x = \{x_1, \dots, x_n\}$ the result of $n$ independents repetitions of Bernoulli R.V.



Confidence intervals of $p_0$ as a function of $S_n = \sum_{i=1}^{n} x_i$ with $x = \{x_1, \dots, x_n\}$ the result of $n$ independants repetitions of Bernoulli R.V.



## 5

Consider a discrete distribution on fixed $m+1$ points $\xi_0, , \xi_m$ with probabilities $(p_0, , p_m)$.

a. Test the hypothesis of a specific distribution $H_0 : (p_0, , p_4) = (1/6, 1/6, 1/6, 3/10, 2/10)$ using asymptotic tests at significance level 0.05. For sample sizes $n = 40, 50, 70, 80$, calculate:

(i) The exact size of the likelihood ratio test for $H_0$.
(ii) The exact size of the Pearson test:

$$\sum_j \frac{(T_j - n\pi_j)^2}{n\pi_j}$$

b. Test the hypothesis $H : p_1 = 4p_2$ (with your $m$).
(i) Construct a likelihood ratio test for $H$. Does the asymptotic distribution theorem apply?
(ii) Choose a parameter set $(p_0, , p_4)$ that satisfies $H$. Simulate $N = 13000$ samples of size $n = 45$ and compute the proportion where your test rejects $H$.

5. a) The size of the test is:

$$\alpha^* = \sup_{\pi \in \mathcal{H}_0} P_\pi(\text{reject } \mathcal{H}_0)$$

Here $\mathcal{H}_0$ is the hypothesis:

Test a dice with 5 faces.

$$p = (p_0, p_1, p_2, p_3, p_4) = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{3}{10}, \frac{2}{10}\right)$$

So the size is the probability of rejecting that vector.

We draw $X_1, \dots, X_n$ iid samples that follows the unknowed distribution $p$.
We set $j$ the count of observation equal to $\xi_j$:

$$T_j = \sum_{i=1}^{n} \mathbb{1}\{X_i = \xi_i\}$$

We have $\sum_{j=0}^{4} T_j = n$

Then the vector $T = (T_0, \dots, T_4) \sim \text{Multinomial}(n; p_0, \dots, p_4)$

(i) LRT:

- Under $H_0$:

$$\ell(\hat{p}_0) = \ln\left(\frac{n!}{T_0! \cdots T_4!} p_0^{T_0} p_1^{T_1} \cdots p_4^{T_4}\right)$$

$$= \ln\left(\frac{n!}{T_0! \cdots T_4!}\right) + \sum_{j=0}^{4} T_j \ln(p_j)$$

- Under the alternative:

$$\ell(\hat{p}) = \ln\left(\frac{n!}{T_0! \cdots T_4!}\left(\frac{T_0}{n}\right)^{T_0} \cdots \left(\frac{T_4}{n}\right)^{T_4}\right)$$

$$= \ln\left(\frac{n!}{T_0! \cdots T_4!}\right) + \sum_{j=0}^{4} T_j \ln\left(\frac{T_j}{n}\right)$$

Thus:

$$\Lambda_n = -2\left(\ell(\hat{p}_0) - \ell(\hat{p})\right)$$

$$= 2\sum_{j=0}^{4} T_j\left[\ln\left(\frac{T_j}{n}\right) - \ln(p_j)\right]$$

$$= 2\sum_{j=0}^{4} T_j \ln\left(\frac{T_j}{n p_j}\right)$$

In the continious case, Wilk's theorem tells us that under $H_0$, $\Lambda_n$ converges in distribution to a $\chi^2_k$ variable, where $k$ is the number of degrees of freedom. Therefore, we'd approximate the exact size by

$$P_{H_0}(\Lambda_n > c) \simeq 1 - F_{\chi^2_k}(c)$$

where $F_{\chi^2_k}$ is the cumulative distribution of a $\chi^2_k$ distribution.

However, in our case $\Lambda_n$ is a discrete random variable (a function of the multinomial counts $T_j$). So instead of approximate the distribution asymptotically we compute the exact size by summing over all realizations $(t_0, \ldots, t_4)$ such that

$$\sum_j t_j = n \quad \text{and} \quad \Lambda_n(t) > c$$

Each term has probability:

$$\mathbb{P}_{H_0}(T = t) = \frac{n}{t_0!\, t_1!\, t_2!\, t_3!\, t_4!} \, p_0^{t_0} \, p_1^{t_1} \, p_2^{t_2} \, p_3^{t_3} \, p_4^{t_4}$$

So the exact size is:

$$\alpha^* = \sum_{\substack{t=(t_0,\ldots,t_4)\in\mathbb{N}^5 \\ \sum t_j = n \\ \Lambda_n(t) > c}} \frac{n}{t_0! \cdots t_4!} \prod_{j=-}^{4} p_j^{t_j}$$

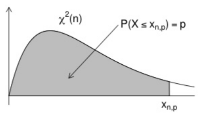
Where $c \simeq 9.49$ the critical value from the $\chi^2_4$ distribution at 0.05 significance.

number of restrictions: 4 $\left( p_4 = 1 - \sum_{j=0}^{3} p_j \text{, so it's not a restriction} \right)$

**Loi du khi-deux**

Table des quantiles de $X \sim \chi^2(n)$ :

$$\mathbb{P}(X \leq x_{n,p}) = p.$$




| $n$ \ $p$ | 0.005 | 0.01 | 0.025 | 0.05 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.95 | 0.975 | 0.99 | 0.995 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.10 | 0.45 | 1.32 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 |
| 2 | 0.01 | 0.02 | 0.05 | 0.10 | 0.21 | 0.58 | 1.39 | 2.77 | 4.61 | 5.99 | 7.38 | 9.21 | 10.60 |
| 3 | 0.07 | 0.11 | 0.22 | 0.35 | 0.58 | 1.21 | 2.37 | 4.11 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 |
| 4 | 0.21 | 0.30 | 0.48 | 0.71 | 1.06 | 1.92 | 3.36 | 5.39 | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 |
| 5 | 0.41 | 0.55 | 0.83 | 1.15 | 1.61 | 2.67 | 4.35 | 6.63 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 |
| 6 | 0.68 | 0.87 | 1.24 | 1.64 | 2.20 | 3.45 | 5.35 | 7.84 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 |
| 7 | 0.99 | 1.24 | 1.69 | 2.17 | 2.83 | 4.25 | 6.35 | 9.04 | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 3.49 | 5.07 | 7.34 | 10.22 | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 |
| 9 | 1.73 | 2.09 | 2.70 | 3.33 | 4.17 | 5.90 | 8.34 | 11.39 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 |
| 10 | 2.16 | 2.56 | 3.25 | 3.94 | 4.87 | 6.74 | 9.34 | 12.55 | 15.99 | 18.31 | 20.48 | 23.21 | 25.19 |
| 11 | 2.60 | 3.05 | 3.82 | 4.57 | 5.58 | 7.58 | 10.34 | 13.70 | 17.28 | 19.68 | 21.92 | 24.72 | 26.76 |
| 12 | 3.07 | 3.57 | 4.40 | 5.23 | 6.30 | 8.44 | 11.34 | 14.85 | 18.55 | 21.03 | 23.34 | 26.22 | 28.30 |
| 13 | 3.57 | 4.11 | 5.01 | 5.89 | 7.04 | 9.30 | 12.34 | 15.98 | 19.81 | 22.36 | 24.74 | 27.69 | 29.82 |
| 14 | 4.07 | 4.66 | 5.63 | 6.57 | 7.79 | 10.17 | 13.34 | 17.12 | 21.06 | 23.68 | 26.12 | 29.14 | 31.32 |
| 15 | 4.60 | 5.23 | 6.26 | 7.26 | 8.55 | 11.04 | 14.34 | 18.25 | 22.31 | 25.00 | 27.49 | 30.58 | 32.80 |
| 16 | 5.14 | 5.81 | 6.91 | 7.96 | 9.31 | 11.91 | 15.34 | 19.37 | 23.54 | 26.30 | 28.85 | 32.00 | 34.27 |
| 17 | 5.70 | 6.41 | 7.56 | 8.67 | 10.09 | 12.79 | 16.34 | 20.49 | 24.77 | 27.59 | 30.19 | 33.41 | 35.72 |
| 18 | 6.26 | 7.01 | 8.23 | 9.39 | 10.86 | 13.68 | 17.34 | 21.60 | 25.99 | 28.87 | 31.53 | 34.81 | 37.16 |
| 19 | 6.84 | 7.63 | 8.91 | 10.12 | 11.65 | 14.56 | 18.34 | 22.72 | 27.20 | 30.14 | 32.85 | 36.19 | 38.58 |
| 20 | 7.43 | 8.26 | 9.59 | 10.85 | 12.44 | 15.45 | 19.34 | 23.83 | 28.41 | 31.41 | 34.17 | 37.57 | 40.00 |
| 21 | 8.03 | 8.90 | 10.28 | 11.59 | 13.24 | 16.34 | 20.34 | 24.93 | 29.62 | 32.67 | 35.48 | 38.93 | 41.40 |
| 22 | 8.64 | 9.54 | 10.98 | 12.34 | 14.04 | 17.24 | 21.34 | 26.04 | 30.81 | 33.92 | 36.78 | 40.29 | 42.80 |
| 23 | 9.26 | 10.20 | 11.69 | 13.09 | 14.85 | 18.14 | 22.34 | 27.14 | 32.01 | 35.17 | 38.08 | 41.64 | 44.18 |
| 24 | 9.89 | 10.86 | 12.40 | 13.85 | 15.66 | 19.04 | 23.34 | 28.24 | 33.20 | 36.42 | 39.36 | 42.98 | 45.56 |
| 25 | 10.52 | 11.52 | 13.12 | 14.61 | 16.47 | 19.94 | 24.34 | 29.34 | 34.38 | 37.65 | 40.65 | 44.31 | 46.93 |
| 26 | 11.16 | 12.20 | 13.84 | 15.38 | 17.29 | 20.84 | 25.34 | 30.43 | 35.56 | 38.89 | 41.92 | 45.64 | 48.29 |
| 27 | 11.81 | 12.88 | 14.57 | 16.15 | 18.11 | 21.75 | 26.34 | 31.53 | 36.74 | 40.11 | 43.19 | 46.96 | 49.64 |
| 28 | 12.46 | 13.56 | 15.31 | 16.93 | 18.94 | 22.66 | 27.34 | 32.62 | 37.92 | 41.34 | 44.46 | 48.28 | 50.99 |
| 29 | 13.12 | 14.26 | 16.05 | 17.71 | 19.77 | 23.57 | 28.34 | 33.71 | 39.09 | 42.56 | 45.72 | 49.59 | 52.34 |
| 30 | 13.79 | 14.95 | 16.79 | 18.49 | 20.60 | 24.48 | 29.34 | 34.80 | 40.26 | 43.77 | 46.98 | 50.89 | 53.67 |
| 40 | 20.71 | 22.16 | 24.43 | 26.51 | 29.05 | 33.66 | 39.34 | 45.62 | 51.81 | 55.76 | 59.34 | 63.69 | 66.77 |
| 50 | 27.99 | 29.71 | 32.36 | 34.76 | 37.69 | 42.94 | 49.33 | 56.33 | 63.17 | 67.50 | 71.42 | 76.15 | 79.49 |
| 60 | 35.53 | 37.48 | 40.48 | 43.19 | 46.46 | 52.29 | 59.33 | 66.98 | 74.40 | 79.08 | 83.30 | 88.38 | 91.95 |
| 70 | 43.28 | 45.44 | 48.76 | 51.74 | 55.33 | 61.70 | 69.33 | 77.58 | 85.53 | 90.53 | 95.02 | 100.4 | 104.2 |
| 80 | 51.17 | 53.54 | 57.15 | 60.39 | 64.28 | 71.14 | 79.33 | 88.13 | 96.58 | 101.9 | 106.6 | 112.3 | 116.3 |
| 90 | 59.20 | 61.75 | 65.65 | 69.13 | 73.29 | 80.62 | 89.33 | 98.65 | 107.6 | 113.1 | 118.1 | 124.1 | 128.3 |
| 100 | 67.33 | 70.06 | 74.22 | 77.93 | 82.36 | 90.13 | 99.33 | 109.1 | 118.5 | 124.3 | 129.6 | 135.8 | 140.1 |

I will proceed the calculation in R under exercise 5.ipynb.

**Results**

| $n$ | $\alpha^*$ |
|---|---|
| 40 | 0.0573 |
| 50 | 0.0545 |
| 60 | 0.0535 |
| 70 | 0.0535 |
| 80 | 0.0527 |

(ii) Pearson Test.

Here we have directly that the test is defined as:

$$Q_n = \sum_{j=0}^{4} \frac{(T_j - n p_j)^2}{n p_j}$$

Which under $H_0$ is also asymptotically a $\chi^2_4$. We have the same critical value and the exact size is

$$\alpha^* = \mathbb{P}_{H_0}(Q_n > 9.49)$$

$$= \sum_{\substack{t=(t_0,\dots,t_4)\in\mathbb{N}^5 \\ \sum t_j = n \\ Q_n(t) > 9.49}} \frac{n}{t_0! \cdots t_4!} \prod_{j=-}^{4} p_j^{t_j}$$

Results

| $n$ | $\alpha^*$ |
|---|---|
| 40 | 0.0479 |
| 50 | 0.0483 |
| 60 | 0.0489 |
| 70 | 0.0491 |
| 80 | 0.0490 |

5.b. (i) We have now $H_0 : p_1 = 4p_2$

LRT :

- Under the alternative :

$$\ell(\hat{p}) = \sup_{p} \ell(p) = \ln\left(\frac{n!}{T_0! \, ... \, T_4!}\right) + \sum_{j=0}^{4} T_j \ln\left(\frac{T_j}{n}\right)$$

- Under $H_0$:

$$\ell(\hat{p}_0) = \sup_{\substack{p \\ p_1 = 4p_2}} \ell(p)$$

To find it, we need to use constrained optimization, ie Lagrange Multipliers:

$$\begin{cases} \sum_{j=0}^{4} p_j = 1 & (1) \\ p_1 - 4p_2 = 0 & (2) \end{cases}$$

Given the counts $T = (T_0, .., T_4)$, the log likelihood is:

$$\ell(p) = \sum_{j=1}^{4} T_j \ln(p_j)$$

Let $\lambda$ and $\mu$ be the two Lagrange multipliers such that:

- $\lambda$ for the contraint (1)
- $\mu$ for the contraint (2)

Thus:

$$\mathcal{L}(p, \lambda, \mu) = \sum_{j=0}^{4} T_j \ln(p_j) + \lambda\left(1 - \sum_{j=0}^{4} p_j\right) + \mu(p_1 - 4p_2)$$

We differentiate $\mathcal{L}$ with respect to $p_j$, and set to zero to find $p_j$ :

- $j = 0$ : $\frac{\partial \mathcal{L}}{\partial p_0} = \frac{T_0}{p_0} - \lambda = 0 \Rightarrow p_0 = \frac{T_0}{\lambda}$
- $j = 1$ : $\frac{\partial \mathcal{L}}{\partial p_1} = \frac{T_1}{p_1} - \lambda + \mu = 0 \Rightarrow p_1 = \frac{T_1}{\lambda - \mu}$
- $j = 2$ : $\frac{\partial \mathcal{L}}{\partial p_2} = \frac{T_2}{p_2} - \lambda + 4\mu = 0 \Rightarrow p_2 = \frac{T_2}{\lambda - 4\mu}$
- $j = 3$ : $p_3 = T_3/\lambda$
- $j = 4$ : $p_4 = T_4/\lambda$

Then we find $\lambda$ and $\mu$:

$$\begin{cases} \dfrac{T_4}{\lambda - \mu} = \dfrac{T_2}{\lambda - 4\mu} \cdot 4 \implies \dfrac{\lambda - 4\mu}{4\lambda - 4\mu} = \dfrac{T_2}{T_1} & (1) \\ \dfrac{T_0 + T_3 + T_4}{\lambda} + \dfrac{T_1}{\lambda - \mu} + \dfrac{T_2}{\lambda - 4\mu} = 1 & (2) \end{cases}$$

$\vdots$

$\lambda = \dots$

$\mu = \dots$

That gives us $\hat{p}_{MLE,j}$ the Maximum Likelihood Estimators under the constraint $p_1 = 4p_2$.

Then, the LRT statistic is:

$$\Lambda_n = 2 \sum_{j=0}^{4} T_j \left( \ln\left(\frac{T_j}{n}\right) - \ln\left(\hat{p}_{MLE,j}\right)\right)$$

The Wilks theorem applies here because with this constrained optimization, we made sure that all estimated parameters $\hat{p}_{MLE,j}$ are in the interior of the parameter space.

Moreover, with one restriction, we have under $H_0$:

$$\Lambda_n \xrightarrow{d} \chi^2_1$$

8. Using the definition directly, prove that the statistic $T(X_1, \dots, X_n) = X_1 + \dots + X_n$ is sufficient in the non-normal model with known variance $\sigma^2$.

The definition is:

> Let $X = (X_1, \dots, X_n)$ be a random sample from a probability distribution that depends on a parameter $\theta$.
>
> A statistic $t = T(X)$ is sufficient for underlying parameter $\theta$ precisely if the conditional probability distribution of the data $X$, given the statistic $t = T(X)$, does not depend on the parameter $\theta$.

In other words

$$f_{X|T(X)=t}(x) \text{ does not depend on } \theta.$$

$$\iff f_{X|T(X)=t}(x) = h(x,t) \quad \text{where } h \text{ is a function that doesn't involve } \theta.$$

Using the joint probability density function of $X$ parameterized by $\theta$, we have

$$f_X(x;\theta) = f_{T(X)}(t,\theta) \cdot f_{X|T(X)=t}(x)$$

My intuition building this proof is that many non linear models belongs to the exponential family. Then we have a density in canonical form to work on.

So, the density of each iid observation $X_i$ in the exponential family can be written in canonical form as:

$$f_{X_i}(X_i;\theta) = h(X_i)\exp\left(\eta(\theta)X_i - A(\theta)\right)$$

Since the observations are iid for the density of $X$, the joint density can be written as:

$$f_X(X;\theta) = \prod_{i=1}^{n} f_{X_i}(X_i;\theta) = \prod_{i=1}^{n} h(X_i)\exp\left(\eta(\theta)X_i - A(\theta)\right)$$

$$= \left(\prod_{i=1}^{n} h(X_i)\right)\exp\left(\eta(\theta)\sum_{i=1}^{n} X_i - nA(\theta)\right)$$

Defining:

- $H(X) = \prod_{i=1}^{n} h(X_i)$
- $T(X) = \sum_{i=1}^{n} X_i$

We have

$$f_X(X;\theta) = H(X)\exp\left(\eta(\theta)T(X) - nA(\theta)\right)$$

Now, we fix $T(X) = t$. The conditional density of $X$ given $T(X) = t$ is:

$$f_{X|T(X)=t}(x;\theta) = \frac{f_X(x;\theta)}{f_{T(X)}(t;\theta)} \qquad (1)$$

We must show that this expression doesn't depend on $\theta$.

So:

$$f_X(X;\theta) = H(X)\exp\left(\eta(\theta)t - A(\theta)\right)$$

And

$$f_{T(X)}(t;\theta) = \int_{T(X)=t} f_X(u;\theta)\,du$$

$$= \int_{T(X)=t} H(u)\exp\left(\eta(\theta)t - A(\theta)\right) du$$

$$= \exp\left(\eta(\theta)t - A(\theta)\right)\int_{T(X)=t} H(u)\,du$$

Then denoting $C(t) := \int_{T(X)=t} H(u)\,du$, which does not depend on $\theta$, and substituting in (1):

$$f_{X|T(X)=t}(X;\theta) = \frac{f_X(X;\theta)}{f_{T(X)}(t;\theta)}$$

$$= \frac{H(X)\exp(\eta(\theta)t - A(\theta))}{\exp(\eta(\theta)t - A(\theta))\,C(t)}$$

$$= \frac{H(X)}{C(t)} \quad \text{which is independent of } \theta.$$

Then we proved that the conditional density given $T(X) = t$ depends only on $X$ through $H(X) = \prod_{i=1}^{n} h(X_i)$. Therefore using the definition, the statistic

$$T(X) = \sum_{i=1}^{n} X_i$$

is sufficient for $\theta$.

<u>Remark</u>:

- This proof applies to all iid model with density that can be written as:

$$f_X(x;\theta) = h(x)\exp\big(\eta(\theta)\, x - A(\theta)\big)$$

Which includes many non-normal distributions: Poisson, Binomial, Geometric, Exponential, Gamma.

- But for any model with known variance that isn't part of this family, such as the Laplace, this proof doesn't apply and the $T(X) = X_1 + \dots + X_n$ statistic might not be sufficient.