

STATISTISCHE METHODEN DER DATENANALYSE

Excercise Sheets

Physik

UNI WIEN

24. November 2023

1 Sheet

1.1 ML Estimator

1.1.1 Calculating the Estimator

The Joint probability density of an exponentially distributed sample is

$$g(x_1, \dots, x_m | \tau) = \prod_{i=1}^n \frac{1}{\tau} e^{-\frac{x_i}{\tau}} \quad (1.1)$$

The logarithm of this density interpreted as a likelihood function is

$$l(\tau) = \ln g(x_1, \dots, x_n | \tau) = \sum_{i=1}^n \ln \left(\frac{1}{\tau} e^{-\frac{x_i}{\tau}} \right) \quad (1.2)$$

$$= \sum_{i=1}^n -\ln(\tau) - \frac{1}{\tau} x_i \quad (1.3)$$

Maximizing the chance to draw this particular sample:

$$\frac{\partial l}{\partial \tau} = \sum_{i=1}^n -\frac{1}{\tau} + \frac{1}{\tau^2} x_i \stackrel{!}{=} 0 \quad (1.4)$$

$$\Leftrightarrow -n + \sum_{i=1}^n \frac{1}{\tau} x_i = 0 \quad (1.5)$$

$$\Leftrightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.6)$$

Inserting $s = \sum x_i = 395.25$ and $n = 250$ yields

$$\hat{\tau} = 1.581 \quad (1.7)$$

1.1.2 Showing that the Estimator is efficient

Showing $\hat{\tau}$ is unbiased

$$E_{\tau} \left[\frac{1}{n} \sum_{i=1}^n x_i \right] = \frac{1}{n} \sum_{i=1}^n E[x_i] \quad (1.8)$$

$$= \frac{1}{n} \sum_{i=1}^n \tau \quad (1.9)$$

$$= \tau \quad (1.10)$$

Showing $\hat{\tau}$ is efficient Calculate the Fisher Information

$$I_{\tau} = E \left[-\frac{\partial^2 \ln g(x_1, \dots, x_n | \tau)}{\partial \tau^2} \right] \quad (1.11)$$

$$= E \left[-\sum_{i=1}^n \frac{1}{\tau^2} - \frac{2}{\tau^3} x_i \right] \quad (1.12)$$

$$= -\sum_{i=1}^n \frac{1}{\tau^2} - \frac{2}{\tau^3} E[x_i] \quad (1.13)$$

$$= -\sum_{i=1}^n \frac{1}{\tau^2} - \frac{2}{\tau^2} \quad (1.14)$$

$$= -\sum_{i=1}^n \frac{1}{\tau^2} \quad (1.15)$$

$$= -\frac{n}{\tau^2} \quad (1.16)$$

Comparing with estimator variance

$$V[\hat{\tau}] = V \left[\frac{1}{n} \sum_{i=1}^n x_i \right] \quad (1.17)$$

$$= \frac{\tau^2}{n} \quad (1.18)$$

Conclusion, estimator is as efficient as it gets.

1.2 Laplace distribution

1.2.1 Expectation Value and Variance

The Laplace distribution density is

$$f(x; m, s) = \frac{1}{2s} \exp \left[-\frac{|x - m|}{s} \right] \quad (1.19)$$

Expectation Value The expectation value of laplace distributed variable X is

$$E[X] = \int_{-\infty}^{\infty} \frac{x}{2s} \exp \left[-\frac{|x - m|}{s} \right] dx \quad (1.20)$$

$$= \frac{1}{2s} \int_{-\infty}^{\infty} (u + m) \exp \left[-\frac{|u|}{s} \right] du \quad (1.21)$$

$$= \frac{1}{2s} \int_{-\infty}^{\infty} u \exp \left[-\frac{|u|}{s} \right] du + m \quad (1.22)$$

$$= m \quad (1.23)$$

Variance

$$V[X] = E[(X - m)^2] \quad (1.24)$$

$$= \frac{1}{2s} \int_{-\infty}^{\infty} (x - m)^2 \exp\left\{-\frac{|x - m|}{s}\right\} dx \quad (1.25)$$

$$= \frac{1}{2s} \int_{-\infty}^{\infty} u^2 \exp\left\{-\frac{|u|}{s}\right\} du \quad (1.26)$$

$$= \frac{1}{s} \int_0^{\infty} u^2 \exp\left\{-\frac{u}{s}\right\} du \quad (1.27)$$

$$= s^3 \int_0^{\infty} v^2 e^{-v} dv \quad (1.28)$$

$$= 2s^3 \quad (1.29)$$

1.2.2 Estimators for m and s

For a given sample the joint density is

$$g(x_1, \dots, x_n | s, m) = \prod_{i=1}^n \frac{1}{2s} e^{-\frac{|x_i - m|}{s}} \quad (1.30)$$

The log likelihood function is

$$\ln g = \sum_{i=1}^n \ln \left(\frac{1}{2s} \exp\left\{-\frac{|x_i - m|}{s}\right\} \right) \quad (1.31)$$

$$= \sum_{i=1}^n \ln \left(\frac{1}{2s} \right) - \frac{|x_i - m|}{s} \quad (1.32)$$

$$= n \ln \left(\frac{1}{2s} \right) - \frac{1}{s} \sum_{i=1}^n |x_i - m| \quad (1.33)$$

The maximum likelihood estimator for m can now be found

$$\frac{\partial \ln g}{\partial s} = \frac{\partial}{\partial s} \left[n \ln \left(\frac{1}{2s} \right) - \frac{1}{s} \sum_{i=1}^n |x_i - m| \right] \quad (1.34)$$

$$= -\frac{n}{s} + \frac{1}{s^2} \sum_{i=1}^n |x_i - m| \stackrel{!}{=} 0 \quad (1.35)$$

$$\Rightarrow \hat{s} = \frac{1}{n} \sum_{i=1}^n |x_i - m| \quad (1.36)$$

The maximum likelihood estimator for s can also be found

$$\frac{\partial \ln g}{\partial m} = \frac{\partial}{\partial m} \left[n \ln \left(\frac{1}{2s} \right) - \frac{1}{s} \sum_{i=1}^n |x_i - m| \right] \quad (1.37)$$

$$= -\frac{1}{s} \sum_{i=1}^n \frac{\partial}{\partial m} |x_i - m| \quad (1.38)$$

$$= -\frac{1}{s} \sum_{i=1}^n \text{sgn}(x - m) \quad (1.39)$$

1.3 Survey

The multimodal distribution is defined as

$$f(n_A, n_B, n_C, n_D | p_A, p_B, p_C, p_D) = n! \prod_{i=A,B,C,D} \frac{1}{n_i!} p_i^{n_i} \quad (1.40)$$

with the log density

$$\ln f = \ln \left(n! \prod_{i=A,B,C,D} \frac{1}{n_i!} p_i^{n_i} \right) \quad (1.41)$$

$$= \ln(n!) + \sum_{i=A,B,C,D} \ln \left(\frac{1}{n_i!} \right) + \ln(p_i^{n_i}) \quad (1.42)$$

$$= \ln(n!) + \sum_{i=A,B,C,D} \ln \left(\frac{1}{n_i!} \right) + n_i \ln(p_i) \quad (1.43)$$

Now the estimators for the voter shares can be found with

$$\frac{\partial \ln f}{\partial p_i} = \frac{n_i}{p_i} \quad (1.44)$$

2 Sheet

2.1 Bernoulli

2.1.1 Clopper and Pearson confidence interval

A Bernoulli experiment is repeated $n = 200$ times with $k = 121$ successes. Calculate the symmetric 95% interval for the parameter p . The Interval boundaries can be calculated with the inverse beta distribution.

$$G_1(k) = \beta \left(\frac{\alpha}{2}; k, n - k + 1 \right) = 0.534 \quad (2.1)$$

$$G_2(k) = \beta \left(\frac{1 - \alpha}{2}; k + 1, n - k \right) = 0.673 \quad (2.2)$$

2.1.2 Approximation by normal distribution (bootstrap and robust)

Estimate p :

$$\hat{p} = \frac{k}{n} = \frac{121}{200} \quad (2.3)$$

With that estimate σ

$$\sigma[\hat{p}] = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (2.4)$$

$$= \sqrt{\frac{9559}{8 \cdot 10^6}} \approx 0.035 \quad (2.5)$$

$$z_{1-\frac{\alpha}{2}} = f_{\text{norm}}\left(1 - \frac{\alpha}{2}\right) = 0.248 \quad (2.6)$$

Now the interval boundaries are for the bootstrap method:

$$G_1(k) = \hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.591 \quad (2.7)$$

$$G_1(k) = \hat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.619 \quad (2.8)$$

and for the robust method

$$G_1(k) = \hat{p} - z_{1-\frac{\alpha}{2}} \frac{1}{2\sqrt{n}} \approx 0.585 \quad (2.9)$$

$$G_1(k) = \hat{p} + z_{1-\frac{\alpha}{2}} \frac{1}{2\sqrt{n}} \approx 0.625 \quad (2.10)$$

2.1.3 Agresti-Coull

$$G_1(k) \approx 0.585 \quad (2.11)$$

$$G_2(k) \approx 0.624 \quad (2.12)$$

2.2 Biased and unbiased Estimators for uniform distribution interval borders

The joint probability of the sample is

$$g(X_1, \dots, X_N | a, b) = \prod_n \frac{1}{b-a} \cdot I_{[a,b]}(X_n) \quad (2.13)$$

The ML estimator is

$$\hat{a} = \arg \max_a \prod_n \frac{1}{b-a} \cdot I_{[a,b]}(X_n) \quad (2.14)$$

This would not take a minimum if not for the constraint

$$\hat{a} \leq \min_n \{X_n\} \quad (2.15)$$

Therefore

$$\hat{a} = \min_n \{X_n\} \quad (2.16)$$

Similarly

$$\hat{b} \geq \max_n \{X_n\} \quad (2.17)$$

and

$$\hat{b} = \max_n \{X_n\} \quad (2.18)$$

Showing the estimators are biased To show the estimators are biased, calculate their distribution functions:

$$F_{\hat{a}}(x) = P(\hat{a} \leq x) = 1 - \prod_n P(X_n > x) \quad (2.19)$$

$$= 1 - P^N(X > x) \quad (2.20)$$

$$= 1 - \left(\frac{b-x}{b-a}\right)^N \quad (2.21)$$

The density is

$$\frac{\partial F_{\hat{a}}}{\partial x} = \frac{N}{b-a} \left(\frac{b-x}{b-a}\right)^{N-1} \quad (2.22)$$

Now the expectation value is

$$E(\hat{a}) = \int_a^b x \frac{n}{b-a} \left(\frac{b-x}{b-a}\right)^{N-1} dx \quad (2.23)$$

$$= \left[-\left(\frac{b-x}{b-a}\right)^N x \right]_a^b + \int_a^b \left(\frac{b-x}{b-a}\right)^N dx \quad (2.24)$$

Here is

$$\left[-\left(\frac{b-x}{b-a}\right)^N x \right]_a^b = \left[\underbrace{-\left(\frac{b-b}{b-a}\right)^N b}_{=0} + \underbrace{\left(\frac{b-a}{b-a}\right)^N a}_{=1} \right] = a \quad (2.25)$$

and

$$\int_a^b \left(\frac{b-x}{b-a} \right)^N dx = \left[-\frac{b-a}{N+1} \left(\frac{b-x}{b-a} \right)^{N+1} \right]_a^b \quad (2.26)$$

$$= -\frac{b-a}{N+1} \left[\underbrace{\left(\frac{b-b}{b-a} \right)^{N+1}}_{=0} - \underbrace{\left(\frac{b-a}{b-a} \right)^{N+1}}_{=1} \right] \quad (2.27)$$

$$= \frac{b-a}{N+1} \quad (2.28)$$

Together

$$E(\hat{a}) = a + \frac{b-a}{N+1} \quad (2.29)$$

Similarly for \hat{b} :

$$F_{\hat{b}}(x) = P(\hat{b} \leq x) \quad (2.30)$$

$$= \prod_n P(X_n \leq x) \quad (2.31)$$

$$= P^N(X \leq x) \quad (2.32)$$

$$= \left(\frac{x-a}{b-a} \right)^N \quad (2.33)$$

The density is

$$\frac{\partial F_{\hat{b}}}{\partial x} = \frac{N}{b-a} \left(\frac{x-a}{b-a} \right)^{N-1} \quad (2.34)$$

The expectation value of \hat{b} is

$$E(\hat{b}) = \int_a^b x \frac{N}{b-a} \left(\frac{x-a}{b-a} \right)^{N-1} dx \quad (2.35)$$

$$= \left[\left(\frac{x-a}{b-a} \right)^N x \right]_a^b - \int_a^b \left(\frac{x-a}{b-a} \right)^N dx \quad (2.36)$$

where

$$\left[\left(\frac{x-a}{b-a} \right)^N x \right]_a^b = \left[\underbrace{\left(\frac{b-a}{b-a} \right)^N}_=1 b - \underbrace{\left(\frac{a-a}{b-a} \right)^N}_=0 a \right] = b \quad (2.37)$$

and

$$\int_a^b \left(\frac{x-a}{b-a} \right)^N dx = \left[\frac{b-a}{N+1} \left(\frac{x-a}{b-a} \right)^{N+1} \right]_a^b \quad (2.38)$$

$$= \frac{b-a}{N+1} \left[\underbrace{\left(\frac{b-a}{b-a} \right)^{N+1}}_{=1} - \underbrace{\left(\frac{a-a}{b-a} \right)^{N+1}}_{=0} \right] \quad (2.39)$$

$$= \frac{b-a}{N+1} \quad (2.40)$$

Together

$$E(\hat{b}) = b - \frac{b-a}{N+1} \quad (2.41)$$

Conclusion Both estimators are biased, but asymptotically unbiased. This makes intuitivly sense. We would like to correct the estimators \hat{a} and \hat{b} .

$$\hat{a}_c = \hat{a} - \frac{b-a}{N+1} \quad (2.42)$$

$$\hat{b}_c = \hat{b} + \frac{b-a}{N+1} \quad (2.43)$$

but a and b are not a priori known. Lets just guess an estimator.

$$E(\hat{b} - \hat{a}) = E(\hat{b}) - E(\hat{a}) \quad (2.44)$$

$$= b - a - \frac{2(b-a)}{N+1} \quad (2.45)$$

$$= \frac{(b-a)(N+1)}{N+1} - \frac{2(b-a)}{N+1} \quad (2.46)$$

$$= (b-a) \frac{N-1}{N+1} \quad (2.47)$$

$$\Leftrightarrow E \left(\frac{N+1}{N-1} (\hat{b} - \hat{a}) \right) = b - a \quad (2.48)$$

Not quite, but this means, that

$$\frac{N+1}{N-1} (\hat{b} - \hat{a}) \quad (2.49)$$

is unbiased estimator for $b - a$. Inserting the newly found estimator yields

$$\hat{a}_c = \hat{a} - \frac{N+1}{N-1} \frac{\hat{b} - \hat{a}}{N+1} \quad (2.50)$$

$$= \frac{\hat{a}(N-1)}{N-1} - \frac{\hat{b} - \hat{a}}{N-1} \quad (2.51)$$

$$= \frac{\hat{a}N - \hat{b}}{N-1} \quad (2.52)$$

$$\hat{b}_c = \hat{b} + \frac{N+1}{N-1} \frac{\hat{b} - \hat{a}}{N+1} \quad (2.53)$$

$$= \frac{\hat{b}(N-1)}{N-1} + \frac{\hat{b} - \hat{a}}{N-1} \quad (2.54)$$

$$= \frac{\hat{b}N - \hat{a}}{N-1} \quad (2.55)$$

and we are done.