STATISTISCHE METHODEN DER DATENANALYSE

Excersise 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

$$g\left(x_1, \ldots, x_m | \tau\right) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-\frac{x_i}{\tau}} \tag{1.1}$$

The logarithm of this density interpreted as a likelihood function is

$$\begin{aligned} l\left(\tau\right) = \ln g\left(x_1, \ldots, x_n | \tau\right) &= \sum_{i=1}^{n} \ln\left(\frac{1}{\tau} e^{-\frac{x_i}{\tau}}\right) && (1.2) \\ &= \sum_{i=1}^{n} - \ln\left(\tau\right) - \frac{1}{\tau} x_i && (1.3) \end{aligned}$$

Maximizing the chance to draw this particular sample:

$$\begin{aligned} \frac{\partial l}{\partial \tau} = \sum_{i=1}^{n} -\frac{1}{\tau} + \frac{1}{\tau^2} x_i \stackrel{!}{=} 0 && (1.4) \\ \Leftrightarrow - n + \sum_{i=1}^{n} \frac{1}{\tau} x_i = 0 && (1.5) \\ \Leftrightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} x_i && (1.6) \end{aligned}$$

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

$$\hat{\tau} = 1.581 \tag{1.7}$$

1.1.2 Showing that the Estimator is efficient

Showing $\hat{\tau}$ is unbiased

$$\begin{aligned} E_\tau\left[\frac{1}{n} \sum_{i=1}^{n} x_i\right] &= \frac{1}{n} \sum_{i=1}^{n} E\left[x_i\right] && (1.8) \\ &= \frac{1}{n} \sum_{i=1}^{n} \tau && (1.9) \\ &= \tau && (1.10) \end{aligned}$$

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

$$I_\tau = E\left[-\frac{\partial^2 \ln g\left(x_1,\ldots,x_2|\tau\right)}{\partial\tau^2}\right] \tag{1.11}$$
$$= E\left[-\sum_{i=1}^{n}\frac{1}{\tau^2} - \frac{2}{\tau^3}x_i\right] \tag{1.12}$$
$$= -\sum_{i=1}^{n}\frac{1}{\tau^2} - \frac{2}{\tau^3}E\left[x_i\right] \tag{1.13}$$
$$= -\sum_{i=1}^{n}\frac{1}{\tau^2} - \frac{2}{\tau^2} \tag{1.14}$$
$$= \sum_{i=1}^{n}\frac{1}{\tau^2} \tag{1.15}$$
$$= \frac{n}{\tau^2} \tag{1.16}$$

Comparing with estimator variance

$$V\left[\hat{\tau}\right] = V\left[\frac{1}{n^2}\sum_{i=1}^{n}V\left[x_i\right]\right] \tag{1.17}$$
$$= \frac{\tau^2}{n} \tag{1.18}$$

Conclusion, estimator is as efficient as it gets.

1.2 Laplace distribution

1.2.1 Expectation Value and Variance

The Laplace distribution density is

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

Expectation Value The expecation value of laplace distributed variable X is

$$E\left[X\right] = \int_{-\infty}^{\infty}\frac{x}{2s}\exp\left[-\frac{|x-m|}{s}\right]\,dx \tag{1.20}$$
$$= \frac{1}{2s}\int_{-\infty}^{\infty}\left(u+m\right)\exp\left[-\frac{|u|}{s}\right]\,du \tag{1.21}$$
$$= \frac{1}{2s}\int_{-\infty}^{\infty}u\exp\left[-\frac{|u|}{s}\right]\,du + m \tag{1.22}$$
$$= m \tag{1.23}$$

Variance

$$V[X] = E\left[(X-m)^2\right] \tag{1.24}$$
$$= \frac{1}{2s}\int_{-\infty}^{\infty}(x-m)^2\exp\left\{-\frac{|x-m|}{s}\right\}dx \tag{1.25}$$
$$= \frac{1}{2s}\int_{-\infty}^{\infty}u^2\exp\left\{-\frac{|u|}{s}\right\}du \tag{1.26}$$
$$= \frac{1}{s}\int_{0}^{\infty}u^2\exp\left\{-\frac{u}{s}\right\}du \tag{1.27}$$
$$= s^3\int_{0}^{\infty}v^2e^{-v}\,dv \tag{1.28}$$
$$= 2s^3 \tag{1.29}$$

1.2.2 Estimators for m and s

For a given sample the joint density is

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

The log likelihood function is

$$\ln g = \sum_{i=1}^{n}\ln\left(\frac{1}{2s}\exp\left\{-\frac{|x_i-m|}{s}\right\}\right) \tag{1.31}$$
$$= \sum_{i=1}^{n}\ln\left(\frac{1}{2s}\right)-\frac{|x_i-m|}{s} \tag{1.32}$$
$$= n\ln\left(\frac{1}{2s}\right)-\frac{1}{s}\sum_{i=1}^{n}|x_i-m| \tag{1.33}$$

The maximum likelihood estimator for m can now be found

$$\frac{\partial \ln g}{\partial s} = \frac{\partial}{\partial s}\left[n\ln\left(\frac{1}{2s}\right)-\frac{1}{s}\sum_{i=1}^{n}|x_i-m|\right] \tag{1.34}$$
$$= -\frac{n}{s}+\frac{1}{s^2}\sum_{i=1}^{n}|x_i-m| \overset{!}{=} 0 \tag{1.35}$$
$$\Rightarrow \hat{s} = \frac{1}{n}\sum_{i=1}^{n}|x_i-m| \tag{1.36}$$

The maximum likelihood estimator for s can also be found

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

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

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

1.3 Survey

The multimodal distribution is defined as

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

with the log density

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

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

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

Now the estimators for the voter shares can be found with

$$\frac{\partial \ln f}{\partial p_i} = \frac{n_i}{p_i} \tag{1.44}$$

2 Sheet

2.1 Bernoulli

2.1.1 Clopper and Pearson confidence interval

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

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

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

2.1.2 Approximation by normal distribution (bootstrap and robust)

Estimate p:

$$\hat{p} = \frac{k}{n} = \frac{121}{200} \tag{2.3}$$

With that estimate σ

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

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

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

Now the interval boundaries are for the bootstrap method:

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

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

and for the robust method

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

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

2.1.3 Agresti-Coull

$$G_1\left(k\right) \approx 0.585 \tag{2.11}$$

$$G_2\left(k\right) \approx 0.624 \tag{2.12}$$

2.2 Biased and unbiased Estimators for uniform distribution interval borders

The joint probability of the sample ist

$$g\left(X_1, \ldots, X_N | a, b\right) = \prod_n^N \frac{1}{b-a} \cdot I_{[a,b]}\left(X_n\right) \tag{2.13}$$

The ML estimator is

$$\hat{a} = \arg\max_a \prod_n^N \frac{1}{b-a} \cdot I_{[a,b]}\left(X_n\right) \tag{2.14}$$

This would not take a minimum if not for the constraint

$$\hat{a} \leq \min_n \{X_n\} \tag{2.15}$$

Therefore

$$\hat{a} = \min_n \{X_n\} \tag{2.16}$$

Similarly

$$\hat{b} \geq \max_n \{X_n\} \tag{2.17}$$

and

$$\hat{b} = \max_n \{X_n\} \tag{2.18}$$

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

$$\begin{align} F_{\hat{a}}(x) = P(\hat{a} \leq x) &= 1 - \prod_n P(X_n > x) \tag{2.19} \\ &= 1 - P^N(X > x) \tag{2.20} \\ &= 1 - \left(\frac{b-x}{b-a}\right)^N \tag{2.21} \end{align}$$

The density is

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

Now the expectation value is

$$\begin{align} E(\hat{a}) &= \int_a^b x \frac{n}{b-a}\left(\frac{b-x}{b-a}\right)^{N-1} dx \tag{2.23} \\ &= \left[-\left(\frac{b-x}{b-a}\right)^N x\right]_a^b + \int_a^b \left(\frac{b-x}{b-a}\right)^N dx \tag{2.24} \end{align}$$

Here is

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

and

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

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

$$= \frac{b-a}{N+1} \tag{2.28}$$

Together

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

Similarly for $\hat{b}$:

$$F_{\hat{b}}\left(x\right) = P\left(\hat{b} \leq x\right) \tag{2.30}$$

$$= \prod_n P\left(X_n \leq x\right) \tag{2.31}$$

$$= P^N\left(X \leq x\right) \tag{2.32}$$

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

The density is

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

The expectation value of $\hat{b}$ is

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

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

where

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

and

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

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

$$= \frac{b-a}{N+1} \tag{2.40}$$

Together

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

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

$$\hat{a}_c = \min_n \{X_n\} - \frac{b-a}{N+1} \tag{2.42}$$

$$\hat{b}_c = \max_n \{X_n\} + \frac{b-a}{N+1} \tag{2.43}$$

but a and b are not a priori known. Note that

$$\frac{E(\hat{a}+\hat{b})}{2} = \frac{E(\hat{a})+E(\hat{b})}{2} = \frac{a+b}{2} \tag{2.44}$$

is an estimator for the mean and unbiased.