Lecture Recording

- **Note: These lectures will be recorded and posted onto the IMPRS website**

- Dear participants,

- We will record all lectures on "*Making sense of data: introduction to statistics for gravitational wave astronomy*", including possible Q&A after the presentation, and we will make the recordings publicly available on the IMPRS lecture website at:

  - https://imprs-gw-lectures.aei.mpg.de/2021-making-sense-of-data/

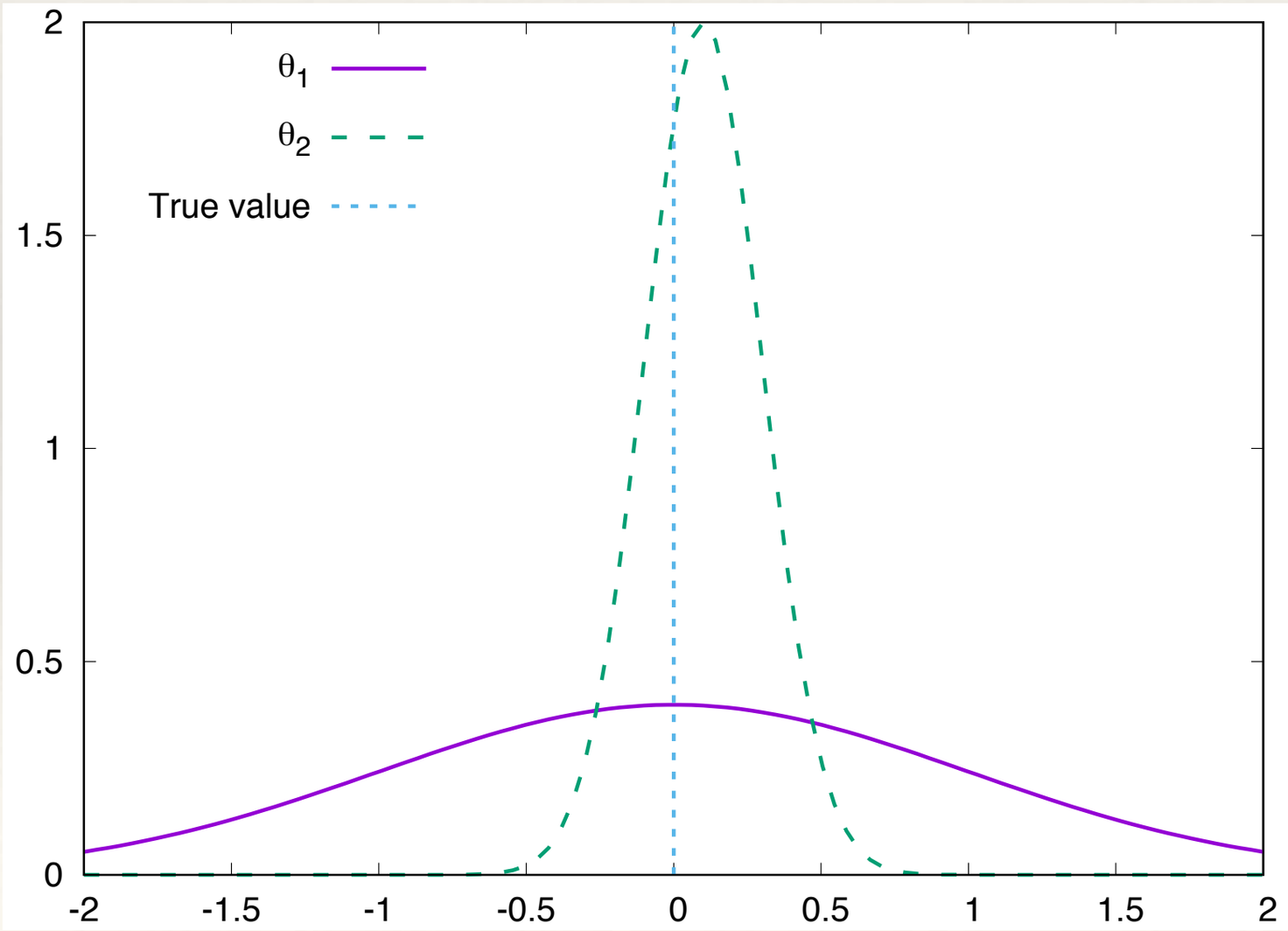
- By participating in this Zoom meeting, you are giving your explicit consent to the recording of the lecture and the publication of the recording on the course website.

Making sense of data: introduction to statistics for gravitational wave astronomy

Lecture 2: statistics and estimators

AEI IMPRS Lecture Course

Jonathan Gair jgair@aei.mpg.de



Frequentist Statistics

- Many measurable quantities are random variables. **Inference** describes the process of learning the probability distribution of the random variable from observations.

- In **parametric inference** the form of the distribution is assumed and inference reduces to making statements about the parameters of the distribution.

- In **frequentist statistics** the parameters are assumed to be **fixed** but **unknown**. Statements, e.g., about **significance** or **confidence**, are about **repetitions of the same experiment** with the parameters fixed.

- Central to frequentist statistics are the notions of **likelihood**, **statistics**, and **estimators**.

Likelihood

- The **likelihood** of an event *E* governed by some probability distribution determined by a set of parameters $\vec{\theta}$ is $\mathbb{P}(E|\vec{\theta})$, regarded as a function of $\vec{\theta}$.
- The likelihood, usually denoted $L(\vec{\theta};\mathbf{x})$ is functionally the same quantity as the pdf, but the latter is a function of x for fixed parameters, while the former is considered a function of the parameters for fixed (observed) x.
- It is often convenient to work with the **log-likelihood**, denoted $l(\theta;\mathbf{x})$

$$l(\theta;\mathbf{x}) = \ln[L(\theta;\mathbf{x})] = \ln[p(\mathbf{x}|\,\theta)]$$

- One interpretation of the likelihood is the relative plausibility of two different values of the parameters, given the observed data. This is expressed by

$$\frac{L(\vec{\theta}_1;\mathbf{x})}{L(\vec{\theta}_2;\mathbf{x})} \qquad \text{or} \qquad l(\vec{\theta}_1;\mathbf{x}) - l(\vec{\theta}_2;|\mathbf{x})$$

Likelihood

- Typically we will observe more than one random variable and so will be interested in the **joint likelihood**. If the RVs are independent we usually have

$$L(\theta; \mathbf{x}) = \prod_{j=1}^{n} p(x_j|\,\theta) \qquad \Rightarrow \qquad l(\theta; \mathbf{x}) = \sum_{j=1}^{n} l(x_j|\,\theta)$$

- **Example**: Poisson distribution. We observe $\{x_1, \ldots, x_n\}$, n IID observations from a Poisson distribution with parameter λ. Writing $n\bar{x} = \sum_{j=1}^{n} x_j$

$$L(\theta; \mathbf{x}) = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_j x_j!} \qquad (\lambda > 0)$$

$$l(\lambda; \mathbf{x}) = \log\left(L(\lambda; \mathbf{x})\right] = -n\lambda + n\bar{x}\ln\lambda - \ln(\prod_j x_j!)$$

- You have to be a little careful with rounding of continuous RVs when the rounding error is comparable to the variability in the data (see notes).

Maximum Likelihood

- The **score** is the derivative of the log-likelihood, also regarded as a function of parameters

$$\frac{\partial l}{\partial \theta_i}$$

- Point(s) where the score vanishes define the **maximum likelihood**

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{x}) = \arg\max_{\theta \in \Theta} \ell(\theta; \mathbf{x})$$

- This is a function of the observed data only and is an **estimator**. It has various nice properties which we will discuss later.

Statistics and estimators

- A **statistic** is any function, **t**(**Y**), of a random variable. It is a function of the observed values of the data only, not the (unknown) parameters of the distribution.

- An **estimator** is any statistic used to estimate the value of parameter. Typically the observed data would be a set of realisations of IID random variables, X_1, …., X_N and an estimator is some function $\widehat{\theta}(X_1, \ldots, X_n)$ used to infer values of the parameters of the underlying pdf.

- **Examples**

  - **maximum likelihood estimator**

  - **sample mean** (used to estimate mean)

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

  - **sample variance** (used to estimate variance)

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Sufficient statistics

- Any function of the data is a statistic and any statistic could be an estimator, so how do we find *good* statistics?
- For some probability distributions there is a lower-dimensional vector that contains the same information about the parameters of the distribution as the full data x. These are **sufficient statistics**.
- A statistic is **sufficient** for parameters $\vec{\theta}$ if the distribution of X given S does not depend on $\vec{\theta}$, i.e., $p_{\mathbf{X}|S}(\mathbf{X}|s,\vec{\theta})$ does not depend on $\vec{\theta}$.
- The full set of observations X is always sufficient, but often there are sufficient statistics of much lower dimensionality.
- Sufficient statistics lead to a reduction in the size of the data. Statistics achieving the greatest reduction are called **minimal sufficient**.

Sufficient statistics

- **Example: Bernoulli trials**

  - Consider a sequence of trials which yield "success" with probability p

$$p_{\mathbf{X}}(\mathbf{x}|\,p) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{\sum x_j}(1-p)^{n-\sum x_j}$$

  - The sum statistic $S = X_1 + \ldots + X_n$ follows a Binomial distribution

$$p_S(s|\,p) = \left(\begin{array}{c} n \\ s \end{array} \right) p^s(1-p)^{n-s} \quad (s = 0, 1, \ \ldots \ , n)$$

- The pdf of X given S can be found to be

$$\begin{aligned} p_{\mathbf{X}|s}(\mathbf{x}|s) &= \frac{\mathbb{P}(X_1 = x_1, \ \ldots \ , X_n = x_n, X_1 + \ \cdots \ + X_n = s|\,\theta)}{\mathbb{P}(X_1 + \ \cdots \ + X_n = s)} \\ &= \begin{cases} \frac{p_{\mathbf{X}}(\mathbf{x}|\,p)}{p_S(s|\,p)} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases} \\ &= \begin{cases} \left(\begin{array}{c} n \\ s \end{array} \right)^{-1} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases} \end{aligned}$$

- which does not depend on p, so S is sufficient for p.

Sufficient statistics

- Sufficient statistics can be recognised by looking at the likelihood. This is formalised by the *Neyman Factorisation Theorem*

Theorem 1. *(Neyman Factorization Theorem). Let* $\mathbf{X} = (X_1, \ldots, X_n) \sim p(\mathbf{x}|\,\vec{\theta})$. *Then, statistic* $s = s(X_1, \ldots, X_n)$ *is sufficient for* θ *iff there exist functions* h *of* $\mathbf{x}$ *and* g *of* $(s, \vec{\theta})$ *such that*

$$p(\mathbf{x} \mid \vec{\theta}) = L(\vec{\theta}; \mathbf{x}) = g(s(\mathbf{x}), \vec{\theta}) h(\mathbf{x}) \quad \forall \vec{\theta} \in \Theta,\ \mathbf{x} \in \mathcal{X}$$

- **Example: Poisson distribution**
  - The likelihood factorises $p(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} = \mathrm{e}^{-n\lambda} \lambda^{s} \times \frac{1}{\prod_{i=1}^{n} x_i!}$
  - where $s = x_1 + \ldots x_n$. We recognise s as a sufficient statistic, which can be verified using

$$p_{X|s}(X|s) = \begin{cases} \frac{p_{\mathbf{X}}(\mathbf{x}|\,\lambda)}{p_S(s|\,\lambda)} = \frac{e^{-n\lambda} \lambda^{\sum x_j} (\prod_j x_j!)^{-1}}{\frac{e^{-n\lambda}(n\lambda)^s}{s!}} = \frac{n^{-s} s!}{\prod_j x_j!} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases}$$

Sufficient statistics

- **Example**: gravitational wave data analysis
  - The usual likelihood for observed gravitational wave data takes the form

$$p(\mathbf{d}|\vec{\theta}) \propto \exp\left[-\frac{1}{2}\left(\mathbf{d}-\mathbf{h}(\vec{\theta})|\mathbf{d}-\mathbf{h}(\vec{\theta})\right)\right]$$

  - where

$$(a|b) = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)}\,\mathrm{d}f$$

  - For many waveform families it is possible to find a *reduced basis* that can be used to reconstruct all the waveforms in the family

$$h(t;\vec{\theta}) = \sum_{i=1}^{M} a_i(\vec{\theta})h_i(t)$$

  - The overlaps of the basis waveforms with the data, $S_i = (\mathbf{h}_i|\mathbf{b})$, are sufficient statistics for inferring the waveform parameters.

Exponential families

- Distributions taking particular forms have nice properties. In particular those that belong to an **exponential family**.
- An exponential family is any family of distributions of the form

$$p(x|\,\theta) = \exp\left\{\sum_{j=1}^{K} A_j(x)B_j(\vec{\theta}) + C(\vec{\theta}) + D(x)\right\} \;\; \forall x, \vec{\theta}$$

- where $\{A_j; j = 1\ldots, K\}, \{B_j; j = 1\ldots, K\}, C, D$ are real-valued functions.
- Given a set of IID observations *{x₁, …, xₙ}* from this distribution, the set

$$\{\textstyle\sum_{j=1}^{n} A_i(x_j) \;:\; 1 \leq \; i \; \leq \; K\}$$

- of statistics are sufficient for $\vec{\theta}$ and are called the *natural statistics* of the family.
- Any distribution that depends on a *K*-dimensional parameter and has a *K*-dimensional minimal sufficient statistic is a member of the exponential family.

Exponential families

- Examples of exponential families

$$\text{Pois}(\lambda): \qquad p(x|\,\lambda) = \frac{\mathrm{e}^{-\lambda}\theta^x}{x!} = \exp[(\ln\lambda)x - \lambda - \ln(x!)]$$

$$\text{Bin}(n,p): \qquad p(x|\,p) = \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x} = \exp\left[\ln\left(\frac{p}{1-p}\right)x + n\ln(1-p) + \ln\left(\begin{array}{c} n \\ x \end{array}\right)\right]$$

$$N(\mu,\sigma^2): \qquad p(x|\,\mu,\sigma) = \exp\left\{\mu\sigma^{-2}x - \frac{1}{2}\sigma^{-2}x^2 - \left(\frac{1}{2}\mu^2\sigma^{-2} + \ln\sigma + \frac{1}{2}\ln(2\pi)\right)\right\}$$

$$\mathcal{E}(\lambda): \qquad p(x|\,\lambda) = \lambda\mathrm{e}^{-\lambda x} = \exp(-\lambda x + \ln\lambda)$$

$$N(\mu_0,\sigma^2)(\sigma \text{ unknown}): \qquad p(x|\,\mu_0,\sigma) = \exp[-\frac{1}{2\sigma^2}(x-\mu_0)^2 - \ln\sigma - \frac{1}{2}\ln(2\pi)]$$

Estimators: unbiasedness

- The **bias** of an estimator of a parameter measures the difference between the mean value and the value of the parameter being estimated.

$$\text{bias}(\hat{\theta}) = \mathbf{b}(\theta) = \mathbb{E}\left(\hat{\theta}\right) - \theta$$

- An estimator is **unbiased** if the bias is zero

Definition 1. *$\widehat{\theta}$ (r.v.) is an unbiased estimator of θ iff*

$$\mathbb{E}(\widehat{\theta}) = \theta.$$

- Estimators may also be **asymptotically unbiased**

Definition 2. *$\widehat{\theta}$ (r.v.) is asymptotically unbiased estimator of θ iff*

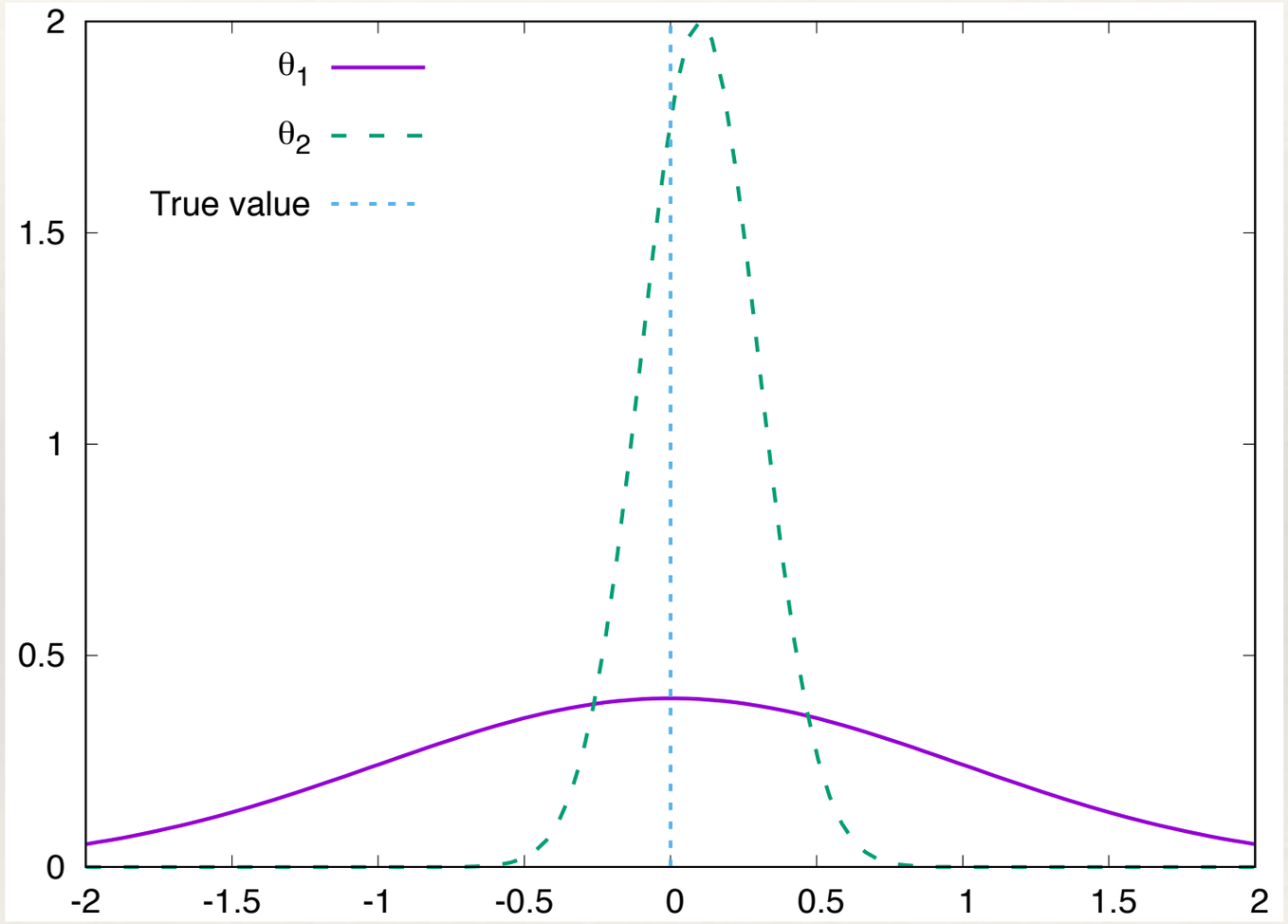
$$\mathbb{E}(\widehat{\theta}) \to \theta \quad as \quad n \to \infty.$$

Estimators: unbiasedness

- Historically, a lot of weight was placed on unbiasedness. Now, minimising **mean square error** is considered more important.

$$\mathrm{MSE}(\widehat{\theta}) \quad = \quad \mathbb{E}[(\widehat{\theta} - \theta)^2] = \mathrm{var}(\widehat{\theta}) + [\mathrm{bias}(\widehat{\theta})]^2$$

- MSE can often be reduced by trading bias for variance. An estimator with a larger bias, but smaller variance, can be preferable.



Estimators: consistency

- An estimator is **consistent** if it becomes increasingly concentrated around the true value as the number of observations increases.

Definition 3. *$\widehat{\theta}$ is a (weakly) consistent estimator for θ if*

$$\mathbb{P}(|\ \widehat{\theta} - \theta\ | > \epsilon) \to 0 \qquad \text{as } n \to \infty$$

for any $\epsilon > 0$.

- From Markov inequality

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

- we can deduce

$$\mathbb{P}[|\ \widehat{\theta} - \theta\ | > \epsilon] \quad \leq \quad \frac{1}{\epsilon^2}\mathbb{E}(\widehat{\theta} - \theta)^2$$

- The term on the right hand side is the mean square error, which is the sum of variance and bias-squared. Therefore, if the bias and variance of an estimator tend to zero asymptotically, the estimator will be (weakly) consistent.

Estimators: efficiency

- The **efficiency** of an estimator is the ratio of the minimum possible variance to the variance of the estimator.
- An unbiased estimator with efficiency of 1 is called **efficient** or a **minimum variance unbiased estimator** (MVUE).
- Efficiency can also be defined asymptotically. An estimator whose efficiency tends to 1 as the number of observations tends to infinity is **asymptotically efficient**.
- The **(asymptotic) relative efficiency** of two estimators is the reciprocal of the ratio of their variances (as the sample size tends to infinity)

$$\lim_{n\to\infty} \frac{Var(\widehat{\theta}_1)}{Var(\widehat{\theta}_2)}$$

Cramer-Rao bound

- The notion of efficiency requires knowledge of the smallest achievable variance of an estimator. This is provided by the **Cramer-Rao bound**.
- In the univariate case we first define **Fisher's Information Matrix**

$$I_\theta = \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right]$$

- where l is the log-likelihood, the derivative is evaluated at the true parameter values, and the expectation value is taken with respect to the pdf for the same parameter value. The second equality follows under certain conditions (see next slide).
- The Cramer-Rao inequality states that, for a random sample $X_1, \ldots, X_n$ from a probability distribution with pdf $p(x|\theta)$ and some estimator $\widehat{\theta}$ with bias $b(\theta)$

$$\mathrm{var}(\widehat{\theta}) \geq \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{I_\theta}$$

Cramer-Rao bound: assumptions

- The proof of the Cramer-Rao inequality relies on certain **regularity conditions**

1. $\forall \theta_1, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$, $p(x \mid \theta_1) \neq p(x \mid \theta_2)$ [identifiability].
2. $\forall \theta \in \Theta$, $p(x \mid \theta)$ have common support.
3. Θ is an open set.
4. $\exists \partial p(x \mid \theta)/\partial \theta$.
5. $\mathbb{E}\left(\partial \log p(\mathbf{X}|\theta)/\partial \theta\right)^2 < \infty$.

Cramer-Rao bound: attainability

Lemma 3. *The Cramér-Rao lower bound is attainable iff there exists a function $f(x)$ of x only, and functions $a(\theta)$, $c(\theta)$ of θ only such that*

$$\frac{\partial l}{\partial \theta} = \frac{(f(x) - a(\theta))}{c(\theta)},$$

in which case $\widehat{\theta} = f(x)$ attains it. The expectation value $\mathbb{E}_\theta \hat{\theta} = a(\theta)$ and $\mathrm{d}a/\mathrm{d}\theta = c(\theta) I_\theta$.

Corollary 1. *There is an unbiased estimator that attains the Cramér-Rao lower bound iff there exists a function $g(x)$ of x only such that*

$$\frac{\partial l}{\partial \theta} = I_\theta (g(x) - \theta),$$

in which case the unbiased estimator $\widehat{\theta} = g(x)$ attains it.

Cramer-Rao bound: example

- Consider $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 known. The log-likelihood is

$$l = \log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

- We can compute

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \qquad \frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2},$$

- and hence obtain the Fisher matrix

$$I_\theta = \mathbb{E}\left[-\frac{\partial^2 l}{\partial \mu^2}\right] = \frac{n}{\sigma^2}$$

- We know that $\mathrm{Var}(\bar{X}) = \sigma^2/n$ and hence it achieves this lower bound and is efficient. We could also deduce this from the earlier lemma (Lemma 3) by noticing

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum(X_i - \mu) = \frac{n}{\sigma^2}(\bar{X} - \mu)$$

Cramer-Rao bound: counterexample

- Now consider $X_1, X_2, \ldots, X_n \sim U[0, \theta]$ with likelihood

$$L(\theta; \mathbf{x}) = \begin{cases} \frac{1}{\theta^n} & 0 \leq x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n)} \leq \theta \\ 0 & \text{elsewhere} \end{cases}$$

- here $x_{(i)}$ denotes the i'th element in the ordered sequence of observations and is called an *order statistic*. We can compute the Fisher matrix

$$\frac{\partial l}{\partial \theta} = -\frac{n}{\theta} \qquad \Rightarrow \qquad I_\theta = \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = \frac{n^2}{\theta^2}$$

- But if we consider the estimator $X_{(n)}$ we find

$$\text{var}\left[\frac{n+1}{n} X_{(n)}\right] = \frac{\theta^2}{n(n+2)} < I_\theta^{-1}$$

- So the Cramer-Rao bound is violated. This is because one of the regularity conditions (common support) is violated in this case.

Cramer-Rao bound: multivariate

- In the multivariate case we generalise the definition of the Fisher matrix to

$$[\mathbf{I}_\theta]_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right] = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right]$$

- For a multi-dimensional statistic $\mathbf{T}$, we introduce an **expectation vector** $\mathbf{m} = \mathbb{E}(\mathbf{T})$.

- The multivariate Cramer-Rao bound is then

$$\mathrm{cov}(t_i, t_j) \geq \frac{\partial m_i}{\partial \theta_k} [\mathbf{I}_\theta]^{-1}_{kl} \frac{\partial m_j}{\partial \theta_l}$$

Maximum likelihood estimators

- The MLE may or may not be **unbiased**, but it is always **asymptotically unbiased** and **asymptotically efficient**. In fact it is asymptotically Normal

Lemma 5. *Let $X_1, \ldots, X_n \sim p(x \mid \theta)$ IID, $\theta \in \Theta \subset \mathbb{R}^K$. Under the regularity conditions of Cramer-Rao inequality, the MLE asymptotically satisfies*

$$\widehat{\theta} \sim N_K(\theta, I_\theta^{-1}) \quad n \to \infty,$$

- In fact, if any unbiased estimator exists that attains the Cramer-Rao bound, it has to be the MLE.

Lemma 6. *Suppose there exists an unbiased estimator $\tilde{\theta}$ that attains Cramer-Rao lower bound, and suppose that MLE $\hat{\theta}$ is the solution of $\frac{\partial \ell}{\partial \theta} = 0$. Then, $\tilde{\theta} = \hat{\theta}$.*

- This, and the fact the MLE can be computed for any distribution, are reasons why the MLE is the mostly widely used frequentist estimator.

MLE Example

- Consider n IID samples from an exponential distribution with pdf

$$p(x|\lambda) = \lambda \mathrm{e}^{-\lambda x}$$

- The likelihood is

$$p(\mathbf{x}|\lambda) = \lambda^n \mathrm{e}^{-\lambda \sum x_i}$$

- giving the MLE

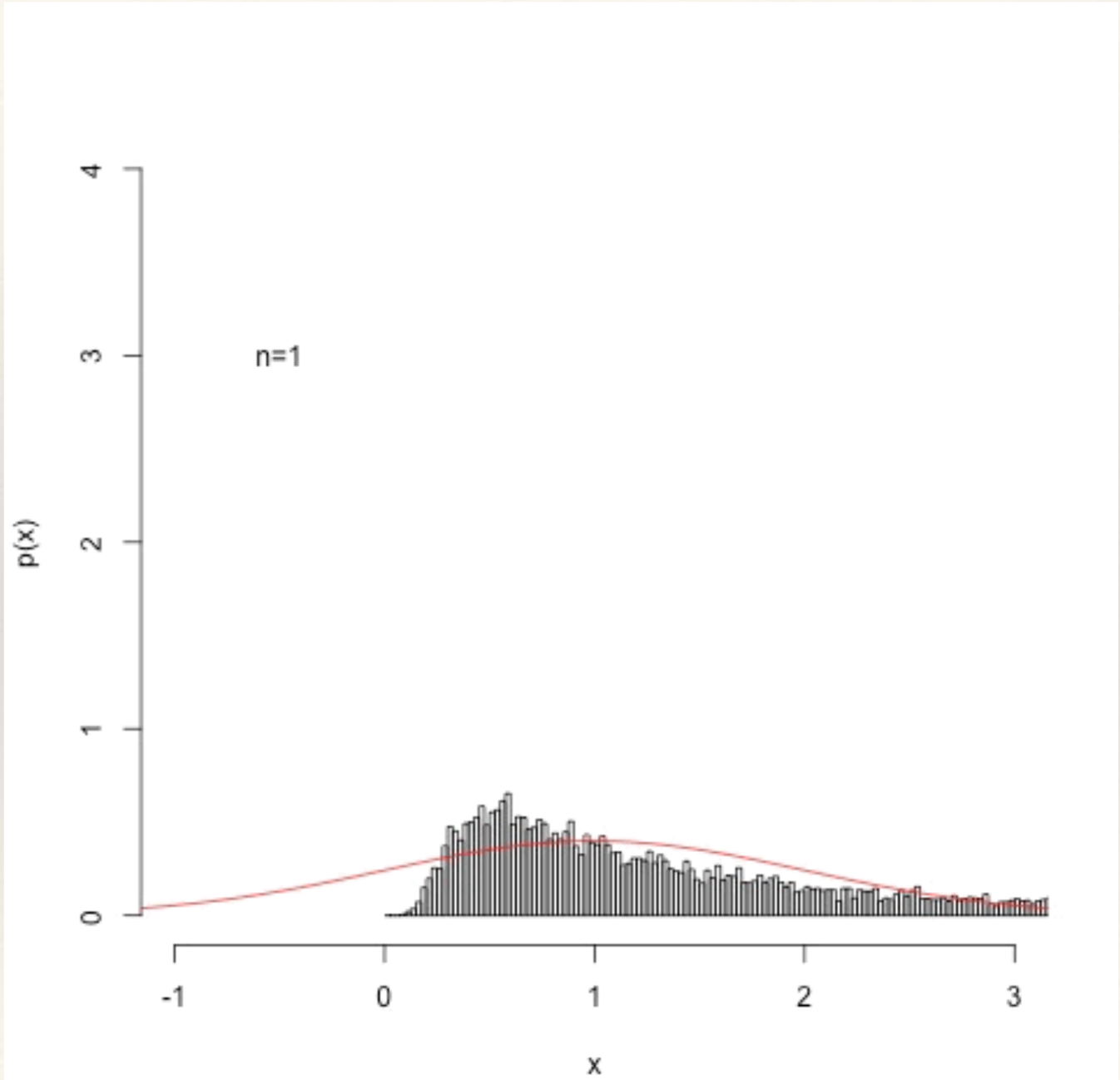
$$\hat{\lambda}_{\mathrm{MLE}} = \frac{n}{\sum x_i}$$

- The mean and variance are

$$\mathbb{E}(\hat{\lambda}) = \frac{n\lambda}{(n-1)} \qquad \mathrm{var}(\hat{\lambda}) = \frac{n^2\lambda^2}{(n-1)^2(n-2)}$$

- The Fisher matrix is

$$I_\lambda = \frac{n}{\lambda^2}$$



Confidence Regions

- We are not only interested in a point estimate of a parameter but an estimate of the uncertainty in that value. This is characterised by a **confidence region or interval**.
- The boundaries of a confidence region are random variables. The construction of a confidence interval is a procedure, which, when repeated, will contain the true parameter values a certain fraction (the **confidence level**) of the time. Formally

$$S_\alpha(\mathbf{X}) \text{ is a } (1-\alpha) \textbf{ confidence region } \text{for } \psi \text{ if}$$

$$\mathbb{P}(S_\alpha(\mathbf{X}) \ni \psi; \psi, \lambda) = 1 - \alpha \quad \forall \psi, \lambda.$$

- Confidence regions can be constructed from **pivotal quantities**, quantities constructed from data and parameter values that have a common distribution.
- **Example:** Normal distribution: the quantity $\sqrt{n}(\bar{x} - \mu)/\sqrt{\sum (x_i - \bar{x})^2} \sim t_{n-1}$
  - giving a confidence interval for the mean

$$\bar{x} - \frac{1}{\sqrt{n}}\sqrt{\sum (x_i - \bar{x})^2}\; t_{\frac{\alpha}{2}} < \mu < \bar{x} + \frac{1}{\sqrt{n}}\sqrt{\sum (x_i - \bar{x})^2}\; t_{\frac{\alpha}{2}}$$