2 Frequentist statistics

In the last section we discussed the notion of a random variable. When observing phenomena in nature or performing experiments we would like to deduce the distribution of the random variable, i.e., the probability distribution from which realisations of that random variable are drawn. In **parametric inference** we assume that the distribution of the random variable takes a particular form, i.e., it belongs to a known family of probability distributions. All of the distributions that were described in the previous section are characterised by one or more parameters and so inference about the form of the distribution reduces to inference about the values of those parameters.

In *frequentist statistics* we assume that the parameters characterising the distribution are *fixed* but *unknown*. Statements about the parameters, for example *significance* and *confidence* are statements about multiple repetitions of the same observation, with the parameters fixed. Key frequentist concepts are *statistics*, *estimators* and *likelihood*.

A **statistic** is a random variable or random vector $T = t(\mathbf{X})$ which is a function of $\mathbf{X}$ but does not depend on the parameters of the distribution, θ. Its realised value is $t = t(\mathbf{x})$. In other words a statistic is a function of observed data only, not the unknown parameters.

An **estimator** is a statistic used to estimate the value of a parameter. Typically the random vector would be a set of IID random variables, $X_1, \ldots, X_n$ with pdf $p(x|\,\theta)$. A function $\widehat{\theta}(X_1, \ldots, X_n)$ of $X_1, \ldots, X_n$ used to infer the parameter values is called an **estimator** of θ; note that $\widehat{\theta}$ is a random variable with a sampling distribution in this latter context. The value of the estimator at the observed data $\widehat{\theta}(x_1, \ldots, x_n)$ is called an **estimate** of θ.

A statistic might also be used to provide an upper or lower limit for a *confidence interval* on the value of a parameter, or to evaluate the validity of a hypothesis in *hypothesis testing*.

2.1 Likelihood

Likelihood is central to the theory of frequentist parametric inference.

If an event E has probability which is a specified function of parameters $\vec{\theta}$, then the likelihood of E is $\mathbb{P}(E|\,\vec{\theta})$, regarded as a function of $\vec{\theta}$.

The likelihood, denoted $L(\vec{\theta}; \mathbf{x})$, is functionally the same as the pdf of the data generating process, the difference is that the likelihood is regarded as a function of the parameters $\vec{\theta}$ while the pdf is regarded as a function of the observed data, $\mathbf{x}$. It is often convenient to work with the **log likelihood**

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

Another useful quantity is the **score**

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

which is a vector that is also regarded as a function of $\vec{\theta}$ with the data fixed at the observed values.

One interpretation of likelihood is that, given data $\mathbf{x}$, the relative plausibility of or support for different values $\vec{\theta}_1$, $\vec{\theta}_2$ of $\vec{\theta}$ is expressed by

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

As a result, inferences are unchanged if $L(\vec{\theta}|\mathbf{x})$ is multiplied by a positive constant (possibly depending on $\mathbf{x}$).

Typically we will be interested in cases where we observe more than one independent realisation of the random variable. For discrete random variables the combined likelihood is then the product of the likelihoods of each observed event.

Example: Poisson distribution

We observe a set $\{x_1, \ldots, x_n\}$, of n IID observations from a Poisson distribution with parameter λ. Denoting $n\bar{x} = \sum_{j=1}^{n} x_j$ the likelihood is

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

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

For continuous random variables the joint likelihood can usually be written as

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

or just $p(\mathbf{x}|\,\theta)$ for a vector $\mathbf{x}$ of random variables that are not IID. One case where this does not necessarily hold is when measurements are imperfect. Typically we cannot observe a quantity with infinite precision, but inevitably round to the nearest measurement unit. Observations of continuous random variables therefore typically involve grouping measurements into bins.

Suppose random variables $X_1, \ldots, X_n$ are IID with cumulative distribution function $P(x|\,\vec{\theta})$ and we observe that there are $n_1, \ldots, n_k$ observations in each of the k intervals $(a_0, a_1], \ldots, (a_{k-1}, a_k]$, where $-\infty \le a_0 < a_1 < \ldots < a_k \le \infty$ and $\mathbb{P}(a_0 < X_j \le a_k) = 1$. The distribution of $(N_1, \ldots N_k)$ is Multinomial with parameters $(n, p_1(\vec{\theta}), \ldots p_k(\vec{\theta}))$ with

$$p_r(\vec{\theta}) = \mathbb{P}(a_{r-1} < X_j \le a_r|\,\vec{\theta}) = P(a_r|\,\vec{\theta}) - P(a_{r-1}|\,\vec{\theta}),$$

and the likelihood is given by (3). For example, with common distribution $N(\mu, \sigma^2)$ we have

$$p_r(\mu, \sigma^2) = \Phi\left(\frac{a_r - \mu}{\sigma}\right) - \Phi\left(\frac{a_{r-1} - \mu}{\sigma}\right).$$

If observations of the IID random variables are made with a resolution (or maximum grouping error)of $\pm\frac{1}{2}h$, then we are effectively in the above situation, and a recorded value x represents a value in the range $x \pm \frac{1}{2}h$. Assuming that the grouping error is small, the likelihood is

$$\prod_{j=1}^{n}\{P(x_j + \frac{1}{2}h|\,\theta) - P(x_j - \frac{1}{2}h|\,\theta)\}. \tag{45}$$

If $p(x|\,\theta)$ does not vary too rapidly in each interval $(x_j - \frac{1}{2}h, x_j + \frac{1}{2}h)$ then (45) can be approximated by

$$\prod_{j=1}^{n}\{hp(x_j|\,\theta)\},$$

or, ignoring the constant h^n,

$$L(\theta;\mathbf{x}) \simeq \prod_{j=1}^{n} p(x_j|\,\theta).$$

which is the result we wrote down when there was no grouping error. However, this argument can fail, as illustrated in the two examples below.

Examples where this approximation fails

- Single observation from $N(\mu,\sigma^2)$

$$L(\mu,\sigma|x) = \Phi\left\{\frac{x+\frac{1}{2}h-\mu}{\sigma}\right\} - \Phi\left\{\frac{x-\frac{1}{2}h-\mu}{\sigma}\right\} \quad (46)$$

$$\simeq \frac{h\ \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}{\sqrt{2\pi}\sigma} \quad (47)$$

  if $\sigma > h$. If $\mu = x$ and $\sigma \to 0$, (46)$\to 1$ but (47)$\to \infty$.

- Uniform distribution on $[0,\theta]$, $U(0,\theta)$
  If $X_1, \ldots, X_n$ are IID with pdf given by

$$p(x|\,\theta) = \begin{cases} \frac{1}{\theta} & (0 < x \le \theta) \\ 0 & \text{otherwise} \end{cases}$$

  then

$$p(\mathbf{x}|\,\theta) = \begin{cases} \frac{1}{\theta^n} & (0 < x_{(n)} \le \theta) \\ 0 & \text{otherwise} \end{cases}$$

  where $x_(i)$ denotes the i'th element in the ordered sequence of $\{x_i\}$. The likelihood is

$$L(\theta;\mathbf{x}) \simeq \begin{cases} 0 & (\theta < x_{(n)}) \\ \frac{1}{\theta^n} & (\theta \ge x_{(n)}) \end{cases} \quad (48)$$

  Taking account of a grouping error of $\pm\frac{1}{2}h$, the probability assigned to $(x_j-\frac{1}{2}h, x_j+\frac{1}{2}h)$ is

$$\begin{cases} \frac{h}{\theta} & (x_j + \frac{1}{2}h < \theta) \\ \frac{\theta - x_j + \frac{1}{2}h}{\theta} & (x_j - \frac{1}{2}h \le \theta < x_j + \frac{1}{2}h) \end{cases}$$

  and, if $h \le x_{(n)} - x_{(n-1)}$,

$$L(\theta;\mathbf{x}) \propto \begin{cases} 0 & (\theta < x_{(n)} - \frac{1}{2}h) \\ \frac{\left[(\theta - x_{(n)} + \frac{1}{2}h)/h\right]^a}{\theta^n} & (x_{(n)} - \frac{1}{2}h \le \theta < x_{(n)} + \frac{1}{2}h) \\ \frac{1}{\theta^n} & (\theta > x_{(n)} + \frac{1}{2}h) \end{cases} \quad (49)$$

  where a is the number of observations equal to $x_{(n)}$. The continuous likelihood (Eq. (48)) and the likelihood accounting for grouping error (Eq. (49)) are shown in Figure 1.

  Ignoring grouping, $x_{(n)}$ is the ML estimator and has variance of order n^{-2}; with grouping the asymptotic variance is the usual $O(n^{-1})$.

To summarise: if the precision of observing the data (h) is much smaller than the variability of the data (e.g. than the standard deviation) then it is fine to use the approximation of the likelihood by the density. However, if the precision h is comparable with the variability, in order to estimate the unknown parameters reliably, one has to use the discrete version of the likelihood.

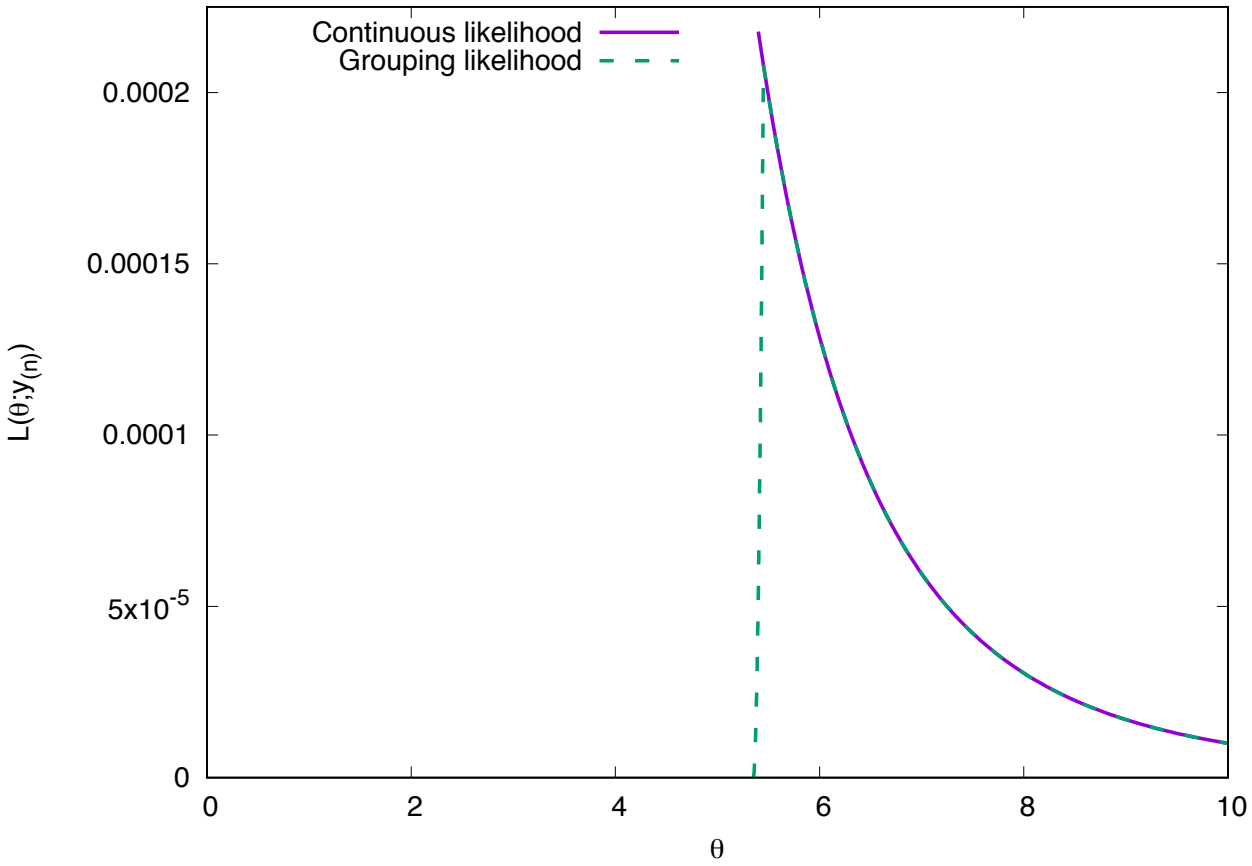


Figure 1: The continuous likelihood for the parameter, θ, of the uniform distribution, as given in Eq. (48), based on $n = 5$ observations with maximum observed value $x_{(n)} = 5.4$ (solid purple line). Also shown is the likelihood including grouping error, as given in Eq. (49), assuming that results are rounded to one decimal place, $h = 0.1$, and there are $a = 2$ observations equal to 5.4 (dashed green line).

2.2 Sufficient statistics

If a parametric form is assumed for the distribution of X, then there may exist a lower dimensional function of the vector of observations $\mathbf{x}$ that contains the same information on the value of $\vec{\theta}$ as vector $\mathbf{x}$. Such a function is called a **sufficient statistic**.

2.3 Definition

Suppose a random vector $\mathbf{X}$ has distribution function in a parametric family $\{P(\mathbf{y}|\,\theta); \theta \in \Theta\}$ and realized value $\mathbf{y}$. A statistic (recall this just means a function of observed data only) is said to be **sufficient** for $\vec{\theta}$ if the distribution of $\mathbf{X}$ given S does not depend on $\vec{\theta}$, i.e. $p_{\mathbf{X}|S}(\mathbf{X}|s, \vec{\theta})$ does not depend on $\vec{\theta}$. Note that

(i) if S is sufficient for $\vec{\theta}$, so is any one-to-one function of S.

(ii) $\mathbf{X}$ is trivially sufficient.

Examples

- Bernoulli trials : $X_1, \ \ldots \ , X_n$ take values 0 or 1 independently with probabilities $1-p$ and p; n is fixed.

$$p_{\mathbf{X}}(\mathbf{x}|\,p) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{\sum x_j}(1-p)^{n-\sum x_j} \tag{50}$$

If $S = X_1 + \ \cdots \ + X_n$, then S has the Binomial p.d.f.

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

and the p.d.f. of $\mathbf{X}$ given S is

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

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

For example, in the case when $n = 3$ the conditional p.d.f of $\mathbf{x} = (x_1, x_2, x_3)$ given $s = \sum x_i$ is as follows:

Sample	$s = \sum x_i$			
(y_1, y_2, y_3)	0	1	2	3
(0 0 0)	1	0	0	0
(1 0 0)	0	$\frac{1}{3}$	0	0
(0 1 0)	0	$\frac{1}{3}$	0	0
(0 0 1)	0	$\frac{1}{3}$	0	0
(1 1 0)	0	0	$\frac{1}{3}$	0
(1 0 1)	0	0	$\frac{1}{3}$	0
(0 1 1)	0	0	$\frac{1}{3}$	0
(1 1 1)	0	0	0	1

- Pois(λ) , $S = X_1 + \cdots + X_n$ has distribution Pois($n\lambda$) and p.d.f.

$$p_S(s|\,\lambda) = \frac{e^{-n\lambda}(n\lambda)^s}{s!},$$

so the distribution of $\mathbf{X}$ given s has p.d.f.

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

which does not depend on λ (it is a multinomial distribution), so S is sufficient for λ.

Interpretation of sufficiency: If S is sufficient for $\vec{\theta}$, we can argue that $\mathbf{x}$ contains no information on $\vec{\theta}$ beyond what is contained in the value s of S, i.e. all the information in $\mathbf{X}$ about $\vec{\theta}$ is contained in s. This suggests that inferences about the value of $\vec{\theta}$ should be based on the value of s. The rest of the information in $\mathbf{y}$ is still relevant to testing the correctness of the assumed parametric family, e.g., by a residual analysis. Sufficiency leads to replacing $\mathbf{x}$ by s and hence to a reduction in the data, so there is an advantage in using statistical models and designs which lead to sufficient statistics of low dimensionality.

2.4 Recognizing sufficient statistics: Neyman Factorization 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*

$$p(\mathbf{x} \mid \vec{\theta}) = L(\vec{\theta}; \mathbf{x}) = g(s(\mathbf{x}), \vec{\theta})h(\mathbf{x}) \quad \forall \vec{\theta} \in \Theta,\ \mathbf{x} \in \mathcal{X} \tag{51}$$

Proof. Proof (discrete case only).

If s is sufficient, then the conditional p.d.f. $p_{\mathbf{X}|S}(\mathbf{x}|s)$ does not depend on $\vec{\theta}$ and we can take $h(\mathbf{x})$ to be $p_{\mathbf{X}|S}(\mathbf{x}|s)$ and $g(s;\theta)$ to be $f_S(s|\,\theta)$. Then

$$\begin{aligned} L(\vec{\theta}; \mathbf{x}) = p_{\mathbf{X}}(\mathbf{x}|\,\vec{\theta}) &= \mathbb{P}(\mathbf{X} = \mathbf{x}|\,\vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x}\,\&\,S = s(\mathbf{x}) \mid \vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}), \vec{\theta})\,\mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}))\,\mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \text{ [since } S \text{ is sufficient]} \\ &= h(\mathbf{x})g(s(\mathbf{x}), \vec{\theta}). \end{aligned}$$

Conversely, if (51) holds, then for any given s there is a subset A_s of $\mathcal{X}$ in which $s(\mathbf{x}) = s$; for $\mathbf{x}$ in A_s

$$\mathbb{P}(\mathbf{X} = \mathbf{x}|S = s, \vec{\theta}) = \frac{f_{\mathbf{X}}(\mathbf{y}|\,\vec{\theta})}{\sum_{\mathbf{z}\in A_s} f_{\mathbf{X}}(\mathbf{z}|\,\vec{\theta})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{z}\in A_s} h(\mathbf{z})},$$

while for $\mathbf{x} \notin A_s$ $\mathbb{P}(\mathbf{X} = \mathbf{x}|S = s, \vec{\theta}) = 0$. Thus the conditional distribution does not depend on $\vec{\theta}$, i.e. S is sufficient for $\vec{\theta}$.

□

Note: the statistic $s(\mathbf{x})$ divides the sample space $\mathcal{X}$ into equivalence classes A_s (one for each value of s). This partitioning of $\mathcal{X}$ is unchanged if s is replaced by any one-to-one function of s.

Examples

- Bernoulli trials
  $$L(p; \mathbf{y}) = p^{\sum x_j}(1-p)^{n-\sum x_j},$$
  so if $s(\mathbf{x}) = \sum x_j$, we could take $h(\mathbf{x}) = 1$, $g(s,p) = p^s(1-p)^{n-s}$
  [or, alternatively, we could take $h(\mathbf{x}) = \begin{pmatrix} n \\ s \end{pmatrix}^{-1}$, $g(s,p) = \begin{pmatrix} n \\ s \end{pmatrix} p^s(1-p)^{n-s}$].
- Pois(λ), with $s = \sum x_i$ we have the factorization
  $$L(\lambda; \mathbf{x}) = (\prod x_j!)^{-1} \cdot e^{-n\lambda}\lambda^s$$
- The Gamma distribution $\Gamma(\alpha, \lambda)$
  $$p_{\mathbf{X}}(\mathbf{x}|\,\alpha, \lambda) = \prod_{j=1}^{n}\left[\frac{\lambda^\alpha x_j^{\alpha-1}e^{-\lambda x_j}}{\Gamma(\alpha)}\right] = \frac{\lambda^{n\alpha}(\prod_j x_j)^{\alpha-1}e^{-\lambda\sum x_j}}{\{\Gamma(\alpha)\}^n} = 1 \cdot \frac{\lambda^{n\alpha}(s_2)^{\alpha-1}e^{-\lambda s_1}}{\{\Gamma(\alpha)\}^n}$$
  Therefore, $(s_1, s_2) = (\sum x_j, \prod x_j)$ is sufficient for (α, λ).
- In a gravitational wave context, reduced order models are used to form a basis for the space of waveforms. Given a set $\{h_i(t)\}$ of basis functions that describe a waveform model, the set $\{(\mathbf{d}|\mathbf{h}_i)\}$ of overlaps of the basis functions with the data are sufficient statistics for deducing the waveform parameters.

2.5 Minimal sufficiency

(Non-trivial) sufficiency leads to a reduction in the data; sufficient statistics achieving the greatest reduction are called **minimal sufficient**, i.e. a minimal sufficient statistic is a function of all other sufficient statistics.

While such statistics are usually obvious, a general method for finding them is implied from the following lemma.

Lemma 1. *Consider the following partition of the sample space of* $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$*:* $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ *belong to the same class of the partition if and only if* $L(\vec{\theta}; \mathbf{x})/L(\vec{\theta}; \mathbf{y})$ *does not depend on* $\vec{\theta}$*.*

Then, any statistic defining this partition is minimal sufficient.

Example

- Weibull distribution: $\{X_1, \ldots, X_n\}$ IID from Weibull with pdf

$$p(y|\,\alpha,\lambda) = \alpha\lambda^\alpha x^{\alpha-1}\exp[-(\lambda x)^\alpha] \quad (x>0;\alpha,\lambda>0)$$

Then

$$L(\alpha,\lambda;\mathbf{x}) = \alpha^n\lambda^{n\alpha}(\prod_{j=1}^n x_j)^{\alpha-1}\exp(-\lambda^\alpha\sum x_j^\alpha)$$

For $L(\alpha,\lambda;\mathbf{z})/L(\alpha,\lambda;\mathbf{y})$ not to depend on α,λ, the z_j must be some permutation of the x_j, but no other reduction in the data retains sufficiency, i.e. the order statistics $x_{(1)} \le \ldots \le x_{(n)}$ are minimal sufficient.

2.6 Exponential families of distributions

A family of distributions indexed by a multivariate parameter $\vec{\theta} \in \Theta \subset \mathbb{R}^p$, is an **exponential family** iff for some real-valued functions $\{A_j; j = 1\ldots,K\}, \{B_j; j = 1\ldots,K\}, C, D$ the pdf has the form

$$p(x|\,\theta) = \exp\left\{\sum_{j=1}^K A_j(x)B_j(\vec{\theta}) + C(\vec{\theta}) + D(x)\right\} \ \forall x, \vec{\theta} \tag{52}$$

Given observations $\{x_1,\ldots,x_n\}$, the set of K statistics $\{\sum_{j=1}^n A_i(x_j) : 1 \le i \le K\}$ are sufficient for $\vec{\theta}$ and they are called the *natural statistics* of the exponential family

In fact, for a K-dimensional parameter $\vec{\theta}$, the minimal sufficient statistic vector is also K-dimensional only for the distributions from the exponential family (under certain regularity conditions, which are the same as those that apply for the validity of the Cramer-Rao inequality described below).

Example. $N(\mu,\sigma^2)$:

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

and $B_1(\mu,\sigma) = \mu\sigma^{-2}$, $B_2(\mu,\sigma) = -\frac{1}{2}\sigma^{-2}$, $A_1(x) = x$, $A_2(x) = x^2$. The vector $S = (\sum_i x_i, \sum_i x_i^2)$ based on sample $(x_1,\ldots,x_n)$ is sufficient for $\vec{\theta} = (\mu,\sigma)$.

2.7 Estimators

Recall that an estimator is a statistic (i.e., a function of data only) that is used to obtain an estimate of one or more parameters of the underlying distribution. Often we consider *point estimators* which are single valued functions $\widehat{\theta}(X_1,\ldots,X_n)$ of $X_1,\ldots,X_n$.

Examples of point estimators:

1. if $\theta = \mathbb{E}(X)$, we can take $\widehat{\theta}$ to be mean, median, mode of the empirical distribution;

2. moment estimators, including the **sample mean**

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

and the **sample variance**

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

3. MLE - maximum likelihood estimator, which minimizes the *score*.

Typically there will be several possible estimators of a parameter θ. To choose between estimators we will define various desirable properties: *unbiasedness*, *consistency* and *efficiency*. *Admissibility* and *sufficiency* are also desirable properties but we won't discuss these here. Sufficiency of an estimator is closely related to sufficiency of a statistic. Robustness and ease of computation are not considered in this course, but may be important in practical applications.

2.7.1 Unbiasedness

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

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

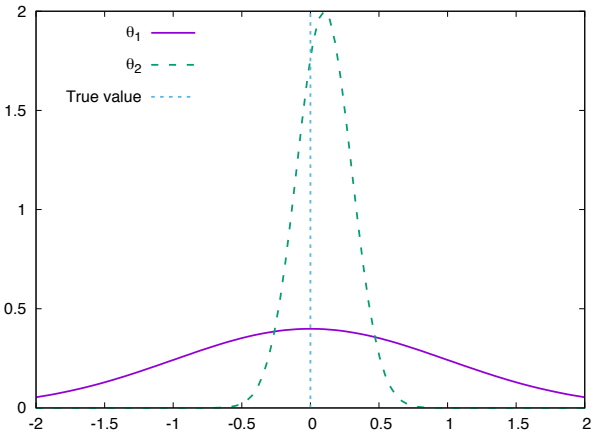
If $\mathbb{E}(\widehat{\theta}) \neq \theta$ then $\widehat{\theta}$ is a biased estimator and we define the bias function of $\widehat{\theta}$ as

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

As an example, suppose θ is a population mean, then the sample mean $\bar{X}$ is unbiased. Also, X_1 (first observation in sample) is unbiased, and if the distribution is symmetric so is the sample median.

There are often several unbiased estimators to choose from, but which is best?

Unbiasedness is not necessarily required for all estimation problems, e.g.,



$\widehat{\theta}_1$ (with wide density) and
$\widehat{\theta}_2$ (with narrow density)
are estimators of θ;
$\widehat{\theta}_1$ is unbiased;
$\widehat{\theta}_2$ is biased;
but $\widehat{\theta}_2$ may be preferred because it is less likely to be a long way from θ.

Biased estimators may be preferred to unbiased estimators in some circumstances. A good property is asymptotic unbiasedness.

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

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

2.7.2 Consistency

As sample size is increased the sampling pdf of any reasonable estimator should become more closely concentrated about θ.

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

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

for any $\epsilon > 0$.

For a particular problem, it may be difficult to verify consistency from this definition, however, a sufficient (not necessary) condition for consistency is given in the lemma below.

Lemma 2. *If* var $(\widehat{\theta}) \to 0$ *and* bias$(\widehat{\theta}) \to 0$ *as $n \to \infty$, then $\widehat{\theta}$ is (weakly) consistent.*

Definition 4. *The mean square error of an estimator $\widehat{\theta}$ is defined as*

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

Mean squared error consists of two terms: variance of $\widehat{\theta}$ and its squared bias.
The *Markov inequality* states that, for a non-negative random variable X and $a > 0$

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

which can be proved straightforwardly

$$\mathbb{E}(X) = \int_0^\infty xp(x)\mathrm{d}x = \int_0^a xp(x)\mathrm{d}x + \int_a^\infty xp(x)\mathrm{d}x \geq \int_a^\infty xp(x)\mathrm{d}x \geq a\int_a^\infty p(x)\mathrm{d}x = a\mathbb{P}(X \geq a).$$

Setting $X = (\hat{\theta} - \theta)^2$ and $a = \epsilon^2$ we find

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

The term on the right had side is the mean square error. If both bias and variance tend to zero asymptotically, the mean square error tends to zero and therefore the left hand side must tend to zero. Hence we have proven Lemma 2.

Examples

1. Estimation of the mean of a normal distribution: using the sample mean $\bar{X}$ or median or just the value of X_1 (first observation in sample) are all unbiased estimators and have variances $\frac{\sigma^2}{n}$, $\alpha\frac{\sigma^2}{n}$ (α is a constant > 1) and σ^2. Therefore the first two are consistent. However, it is evident that X_1 is not consistent as its distribution does not change with sample size.

2. The Cauchy distribution with scale 1 and pdf $p(x|\,\theta) = \pi^{-1}[1+(x-x_0)^2]^{-1}$. In this case, the sample mean $\bar{X}$ has the same distribution as any single X_i, thus $\mathbb{P}[|\ \bar{X} - x_0\ | > \epsilon]$ is the same for any n. This does not tend to zero as $n \to \infty$, and so $\bar{X}$ is not (weakly) consistent. (However, the sample median is a consistent estimator of x_0.)

2.8 Efficiency

Definition 5. *The* ***efficiency*** *of an unbiased estimator* $(\widehat{\theta})$ *is the ratio of the minimum possible variance to var*$(\widehat{\theta})$.

Definition 6. *An unbiased estimator with efficiency equal to 1 is called* ***efficient*** *or a* ***minimum variance unbiased estimator (MVUE)****.*

We can also define asymptotic efficiency of an (asymptotically) unbiased estimator $(\widehat{\theta})$ is the limit of the ratio of the minimum possible variance to var$(\widehat{\theta})$ as sample size $n \to \infty$.

Definition 7. *An estimator with asymptotic efficiency equal to 1 is called* ***asymptotically efficient****.*

We can compare the efficiency of two estimators in the following way.

Definition 8. *The* ***(asymptotic) relative efficiency*** *of two unbiased estimators* $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$ *is the reciprocal of the ratio of their variances, as sample size* $\to \infty$*:* $\lim_{n\to\infty} \frac{Var(\widehat{\theta}_1)}{Var(\widehat{\theta}_2)}$.

The definition of asymptotic relative efficiency can also be extended to asymptotically unbiased estimators. These definitions are all fine, but they rely on knowing what the smallest possible variance is. Under certain assumptions we can obtain this from the Cramér-Rao inequality.

2.8.1 Cramér-Rao lower bound (inequality)

The theorem below (Cramér-Rao inequality) provides a lower bound on the variance of an estimator. When this lower bound is attainable for unbiased estimators, it can be used in the definition of efficiency.

Regularity conditions for the Cramér-Rao inequality.

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. $\mathbb{E}\left(\partial \log p(\mathbf{X}|\theta)/\partial\theta\right)^2 < \infty$.

Here $I(\theta) = \mathbb{E}\left(\frac{\partial \log f(\mathbf{X}|\theta)}{\partial\theta}\right)^2$ is the Fisher information matrix.

Theorem 2. *(Cramér-Rao inequality) Let $X_1, \ldots, X_n$ denote a random sample from $p(x|\,\theta)$, and suppose that $\widehat{\theta}$ is an estimator for θ. Then, subject to the above regularity conditions,*

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

where

$$b(\theta) = \mathrm{bias}(\widehat{\theta}) \quad \text{and} \quad I_\theta = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial\theta}\right)^2\right].$$

Comments

1. For unbiased $\widehat{\theta}$, the lower bound simplifies to $\mathrm{var}(\widehat{\theta}) \geq I_\theta^{-1}$.
2. I_θ is called Fisher's information about θ contained in the observations.
3. Regularity conditions are needed to change the order of differentiation and integration in the proof given below.
4. The result can be extended to estimators of functions of θ.

Proof of Theorem 2.

$$\begin{aligned}\mathbb{E}[\widehat{\theta}] &= \int \ldots \int \widehat{\theta}(x_1, \ldots, x_n) \left\{\prod_{i=1}^{n} p(x_i|\,\theta)\right\} d\mathbf{x} \\ &= \int \ldots \int \widehat{\theta}(x_1, x_2, \ldots, x_n) L(\theta; \mathbf{x}) d\mathbf{x}\end{aligned}$$

$\int \ldots \int$ is a multiple integral with respect to $\mathbf{x}$=$(x_1, x_2, \ldots, x_n)$.

From the definition of bias we have

$$\theta + b = \mathbb{E}(\widehat{\theta}) = \int \ldots \int \widehat{\theta} L(\theta; \mathbf{x}) d\mathbf{x}.$$

Differentiating both sides with respect to θ gives (using regularity conditions)

$$1 + \frac{\partial b}{\partial \theta} = \int \dots \int \widehat{\theta} \frac{\partial L}{\partial \theta} d\mathbf{x}$$

since $\widehat{\theta}$ does not depend on θ. Since $l = \ln(L)$ we have

$$\frac{\partial l}{\partial \theta} = \frac{\partial \ln(L)}{\partial \theta} = \frac{1}{L}\frac{\partial L}{\partial \theta}, \quad \text{and thus} \quad \frac{\partial L}{\partial \theta} = L\frac{\partial l}{\partial \theta}.$$

Thus

$$1 + \frac{\partial b}{\partial \theta} = \int \dots \int \widehat{\theta} \frac{\partial l}{\partial \theta} L d\mathbf{x} = \mathbb{E}\left(\widehat{\theta}\frac{\partial l}{\partial \theta}\right).$$

Now use the result that for any two r.v.s $U$and V,

$$\{\mathrm{cov}(U,V)\}^2 \leq \mathrm{var}(U)\mathrm{var}(V)$$

and let

$$U = \widehat{\theta}, \text{ and } V = \partial l/\partial\theta.$$

Then

$$\begin{aligned}
\mathbb{E}[V] &= \int \dots \int \frac{\partial l}{\partial \theta} L d\mathbf{x} = \int \dots \int \frac{\partial L}{\partial \theta} d\mathbf{x} \\
&= \frac{\partial}{\partial \theta}\left(\int \dots \int L \, d\mathbf{x}\right) \quad \text{(using regularity conditions)} \\
&= \frac{\partial}{\partial \theta}(1) = 0.
\end{aligned}$$

Hence

$$\mathrm{cov}(U,V) = \mathbb{E}(UV) = 1 + \frac{\partial b}{\partial \theta}.$$

Similarly

$$\mathrm{var}(V) = \mathbb{E}(V^2) = \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = I_\theta \quad \text{(by definition of } I_\theta)$$

and since $\mathrm{var}(U) = \mathrm{var}(\widehat{\theta})$ we obtain the Cramér-Rao lower bound as

$$\mathrm{var}(\widehat{\theta}) \geq \frac{\{\mathrm{cov}(U,V)\}^2}{\mathrm{var}(V)} = \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{I_\theta}.$$

□

The Cramér-Rao lower bound will only be useful if it is attainable or at least nearly attainable.

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.

Lemma 4. *Under the same regularity conditions as for the Cramér-Rao lower bound*

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

Example
$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, $\quad \sigma^2$ known.
Likelihood for μ

$$L(\mu; \mathbf{x}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

log likelihood for μ

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

Thus we have

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

and

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

The lower bound for unbiased estimators is $I_\theta^{-1} = \frac{\sigma^2}{n}$. However,

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n},$$

so $\bar{X}$ attains its lower bound. No other unbiased estimator can have smaller variance than $\bar{X}$. Therefore $\bar{X}$ is MVUE.

Alternatively, we can use Lemma 3, and

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

Therefore the bound is attainable.

Regularity conditions are essential to be able to use the lower bound. Consider the uniform distribution case $X_1, X_2, \ldots, X_n \sim U[0, \theta]$

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

In the range where L is differentiable $l = -n \log \theta$

$$\frac{\partial l}{\partial \theta} = -\frac{n}{\theta} \quad \text{and} \quad \frac{\partial^2 l}{\partial \theta^2} = \frac{n}{\theta^2}.$$

Thus

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

but

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

Therefore the lower bound should be $\frac{\theta^2}{n^2}$, but

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

The lower bound is violated because the regularity conditions don't hold. In particular the second condition is violated, since the support of the distribution depends on θ.

The derivation and examples above were all for a one dimensional parameter. The corresponding result for the multiple parameter case is

$$\operatorname{cov}(t_i, t_j) \geq \frac{\partial m_i}{\partial \theta_k} [\mathbf{I}_\theta]^{-1}_{kl} \frac{\partial m_j}{\partial \theta_l}, \qquad [\mathbf{I}_\theta]_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j}\right],$$

where $\mathbf{t}$ is the realised value of some multi-dimensional statistic $\mathbf{T}$ and $\mathbf{m} = \vec{\theta} + \mathbf{b} = \mathbb{E}(\mathbf{T})$.

2.9 Rao-Blackwell Theorem

The Rao-Blackwell theorem gives a method of improving an unbiased estimator, and involves conditioning on a sufficient statistic.

Theorem 3. *(Rao-Blackwell theorem). Let X_1, $X_2, \ldots, X_n$ be a random sample of observations from a distribution with pdf $p(x|\,\theta)$. Suppose that S is a sufficient statistic for θ and that $\widehat{\theta}$ is any unbiased estimator for θ. Define $\widehat{\theta}_S = \mathbb{E}[\widehat{\theta} \mid S]$. Then*

(a) $\widehat{\theta}_S$ is a function of S only;

(b) $\mathbb{E}[\widehat{\theta}_S] = \theta$;

(c) $\operatorname{var} \widehat{\theta}_S \leq \operatorname{var} \widehat{\theta}$.

2.10 Maximum likelihood estimators

Definition 9. *The maximum likelihood estimator (MLE) is defined by* $\widehat{\theta} = \arg\max_{\theta\in\Theta} L(\theta;\mathbf{x}) = \arg\max_{\theta\in\Theta} \ell(\theta;\mathbf{x})$.

If $\exists \partial\ell/\partial\theta_j$ and Θ is open, then the MLE $\widehat{\theta}$ satisfies $\partial\ell/\partial\theta_j(\widehat{\theta}) = 0$, $j = 1,\ldots,K$, $\theta \in \Theta \subset \mathbb{R}^K$.

The MLE can be biased or unbiased but it is asymptotically unbiased and efficient and it is also consistent. In fact the following lemma holds.

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 particular, $\mathbb{E}(\widehat{\theta}) \to \theta$ *and for* $K = 1$, $Var(\widehat{\theta})/I_\theta^{-1} \to 1$ *as* $n \to \infty$.

If there exists an unbiased efficient estimator this 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,* $\tilde{\theta} = \hat{\theta}$.

Proof. $\tilde{\theta}$ is unbiased and attains Cramer-Rao lower bound, hence, by the corollary to Lemma 3, $\frac{\partial\ell}{\partial\theta} = I_\theta(\tilde{\theta} - \theta)$. Then, the only solution of $\frac{\partial\ell}{\partial\theta} = 0$ is $\tilde{\theta}$, that is, $\tilde{\theta} = \hat{\theta}$. □

Thus, (under the regularity conditions of Cramer-Rao inequality) if the Cramer-Rao lower bound is attainable, the MLE attains it, thus in this case the MLE is efficient. If the bound is unattainable, then the MLE is asymptotically efficient.

2.11 Confidence intervals and regions

Point estimators provide single estimated values for parameters, but we usually also need an estimate of the uncertainty in those estimated values. These are characterised by **confidence intervals**. A confidence interval is a random variable since the ends of the interval are typically determined as a function of the observed data. The interval has the property that over many realisations of the same experiment, the intervals constructed randomly by this procedure will contain the true value of the parameter a certain fraction of the time.

Formally a set $S_\alpha(\mathbf{X})$ is a $(1-\alpha)$ **confidence region** for ψ if

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

Thus, $S_\alpha(\mathbf{X})$ is a random set of ψ-values which includes the true value with probability $1-\alpha$. If more than one value of α is considered, we usually require

$$S_{\alpha_1}(\mathbf{x}) \supset S_{\alpha_2}(\mathbf{x}) \text{ if } \alpha_1 < \alpha_2. \tag{53}$$

e.g. a 99% region contains the 95% region.

If ψ is a scalar and $S_\alpha(\mathbf{x})$ has the form $\{\psi : t^\alpha \geq \psi\}$ for some statistic t^α, then t^α is a $(1-\alpha)$ **upper confidence limit** for ψ.

If ψ is a scalar and $S_\alpha(\mathbf{x})$ has the form $\{\psi : s^\alpha \leq \psi\}$ for some statistic s^α, then s^α is a α **lower confidence limit** for ψ.

If $S_\alpha(\mathbf{x}) = \{\psi : a_\alpha(\mathbf{x}) \leq \psi \leq b_\alpha(\mathbf{x})\}$, it is a **two-sided confidence interval**.

A two-sided confidence interval is called **equitailed** if $a_\alpha(\mathbf{x})$ is the $\alpha/2$ lower confidence limit and $b_\alpha(\mathbf{x})$ is the $1-\alpha/2$ upper confidence limit.

A **high density confidence region** is $\{\theta \in \Theta : p(\mathbf{x}|\theta) \geq K_\alpha\}$ where the constant K_α is determined by the condition $\mathbb{P}\{p(\mathbf{X}|\theta) \geq K_\alpha\} = 1-\alpha$.

Confidence intervals/regions for estimators can be constructed by identifying **pivotal quantities**. A pivotal quantity $U = u(\mathbf{X}, \psi)$ is a scalar function of $\mathbf{X}$ and ψ with the same distribution for all ψ and λ. If u_α is the upper α point of this distribution, then

$$\mathbb{P}(u(\mathbf{X}, \psi) \leq u_\alpha) = 1 - \alpha,$$

so that the set $\{\psi : u(\mathbf{x}, \psi) \leq u_\alpha\}$ defines a $(1-\alpha)$ confidence region for ψ.

If ψ is a scalar and $u(mathbfx, \psi)$ is monotone in ψ, this yields a one-sided interval. In this case we may also define two-sided intervals by $\{\psi : u_{\alpha_L} \leq u(\mathbf{x}, \psi) \leq u_{\alpha_U}\}$ with $\alpha_U - \alpha_L = 1 - \alpha$.

Examples of pivotal quantities

- $\mathcal{E}(\lambda)$: $2\theta \sum X_j$ which has distribution $\chi^2(2n)$;
- $N(\mu, \sigma^2)$, inference about μ with σ unknown: $\sqrt{n}(\bar{x} - \mu)/s$ which has distribution $t(n-1)$;
- Ratio of two Normal variances: $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ which has distribution $F(n_1-1, n_2-1)$.