# MSF100
# Statistical Inference Principles - Lecture 1

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

January 16, 2012

## 1 Introduction

- **Parameter** value $\theta \in \Theta$, where $\Theta$ is the parameter space.
- **Sample** $X_1, \ldots, X_n$ independently drawn according to pdf $f_i(x|theta)$ which may be different for the different $X_i$.
- **Observed** data $x_1, \ldots, x_n$

I will use notation $\tilde{X}, X$ or $X^n$ to denote the whole sample, and $\tilde{x}, x, x^n$ for the set of observed values. Note, $X$ and $x$ can also refer to a single random variable or observation, but hopefully the distinction will be clear from the context.

**Parametric model**: $F = \{f(\tilde{x}|\theta), \theta \in \Theta\}$
If $\theta$ is also allowed to have a distribution $\pi(\theta)$ we are *Bayesian*, o/w *Frequentist*.
In both cases, observed data is an instance of something that could be repeated under identical conditions.

- **Statistics** computed from the observed data $\tilde{x}$ are used for **inference** about the parameter $\theta$ (or its credible values if you're Bayesian).
- $\tilde{x} \to T(\tilde{x}) \to \hat{\theta} \to [\hat{\theta} \pm 1.96 se(\hat{\theta})]$ covers true $\theta$ with probability .95 (in 95% of repeated experiments) A lot goes into getting $\hat{\theta}$, its standard error and the appropriate form of the confidence interval as we shall see.
- If you're Bayesian, inherent uncertainty/varability in the world, prior belief $\pi(\theta)$ observed data used to *update* the prior to posterior $\pi(\theta|\tilde{x})$ from which credible region (smallest interval with mass 95%) can be constructed.

## 2 Sufficiency

Before we can get this far, we go over some basics.

The sample $\tilde{X}$ (in Frequentist settings) carries all the information about $\theta$ but constitute, perhaps, an inefficient way of summarizing this information. A **statistic** $T(\tilde{X})$ is a form of *data reduction*.

<u>Sufficiency Principle:</u> $T(\tilde{X})$ is sufficient for $\theta$ if any information in $\tilde{X}$ about $\theta$ depends on $\tilde{X}$ only through $T(\tilde{X})$.
It follows that the conditional distribution $X|T(X)$ does no depend on $\theta$.

Under the sufficiency principle, $T(\tilde{X})$ is such that no information about $\theta$ has been lost by the data reduction.

How can we see that a $T(\tilde{X})$ is sufficient?

- compute the conditional distribution $P_\theta(X = x|T(X) = t)$ and check it is the same for all $\theta$ (Puh!)
- **Factorization theorem**

<u>Factorization Theorem</u>: For $f(x|\theta)$ pdf of $X$, $T(X)$ is sufficient iff there are functions $g(t|\theta)$ and $h(x)$ for all $x$ and values of $\theta$ such that

$$f(\tilde{x}|\theta) = g(T(\tilde{x})|\theta)h(\tilde{x})$$

Example: $X_1, \dots, X_n$ with $f_{X_i}(x|\theta) = e^{i\theta - x}1\{x \geq i\theta\}$.
Then $f(\tilde{x}|\theta) = \prod_{i=1}^n e^{i\theta}e^{-x_i}1\{x_i/i \geq \theta\} = \prod_{i=1}^n e^{i\theta}e^{-x_i}1\{\min_i x_i/i \geq \theta\} =$
$e^{-\sum x_i}e^{\sum i\theta}1\{T(\tilde{x}) \geq \theta\}$, where $T(\tilde{x}) = \min_i x_i = x_{(1)}$

Note, in the multiple parameter case, we may need more than one summary statistic $T(X) = (T_1(X), T_2(X), \dots, T_r(X))$ for $\theta = (\theta_1, \dots, \theta_s)$. Typically, $r = s$ but we may have $r > s$ in some cases.

# 3 Minimal sufficiency

There may be many different $T$ that are sufficient for $\theta$. We call $T$ a minimally sufficient statistic (MSS) if for any other sufficient statistic $T'$, $T$ is a function of $T'$. The MSS $T$ is the coarsest possible data summary that doesn't constitute information loss about $\theta$.

<u>Minimally sufficient statistic</u>: $T$ is MSS if any other sufficient statistic $T'$, $T$ is a function of $T'$.

Using the above definition is not that easy, but a theorem by Lehmann and Scheffe' makes it very easy indeed to check for MSS.

<u>Lehmann Scheffe'</u>: Suppose there is a $T(\tilde{x})$ such that for $\tilde{x}$ and $\tilde{y}$ (observations) $\frac{f(\tilde{x}|\theta)}{f(\tilde{y}|\theta)}$ is constant, independent of $\theta$ iff $T(\tilde{x}) = T(\tilde{y})$. Then $T(\tilde{x})$ is MSS for $\theta$.

Example: A hierarchical model. First draw $N$ from a distribution $P(N = n) = p_n$, $\sum_k p_k = 1$. Observing $N = n$, draw $n$ samples from $Be(\theta)$ and let the sum of these be denoted by $X$.

$$\frac{f(\tilde{x}, n|\theta)}{f(\tilde{y}, n'|\theta)} = \frac{f(\tilde{x}|n, \theta)p_n}{f(\tilde{y}|n', \theta)p_{n'}} = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}p_n}{\binom{n'}{y}\theta^y(1-\theta)^{n'-y}p_{n'}} = \theta^{x-y}(1-\theta)^{n-n'-x+y}\frac{\binom{n}{x}p_n}{\binom{n'}{y}p_{n'}}$$

In the last expression, the ratio term does not depend on $\theta$. The first two terms do not depend on $\theta$ if $x = y$ and $n = n'$ and so $X, N$ are the MSS for $\theta$ in this case.

Look at the above example carefully? Is it not strange that $N$ should features as a MSS for $\theta$? $N$ does not depend on $\theta$. That is $P(N = n) = p_n$ is not a function of $\theta$.

# 4 Ancillary statistics

A statistic $S(X)$ that does not depend on $\theta$ (i.e. P(S(X)=s) not a function of $\theta$, or its pdf is not a function of $\theta$ is called an **Ancillary statistic**.

Ancillary Statistic: If $S(X)$ is such that its distribution does not depend on $\theta$ we call it ancillary. (Ancillary: supplemental or extra).

So, in the above example $N$ is ancillary, and alone does not carry any information about $\theta$. However, together with $X$ it provides sufficient information for $\theta$. This tells us we have to be careful about making snap judgements about what part of the data is needed for estimating $\theta$.

It's even more complicated that this. We can have $T$, a sufficient statistic, and $S$, an ancillary, and have $T$ and $S$ *dependent* on each other! This feels rather counterintuitive since $T$ is the part of the data related to $\theta$ and $S$ is not. However, for "nice" distribution families our intuition tends to be correct: $S$ and $T$ are not dependent (more later).

Examples:

- $X \sim N(\mu, \sigma^2)$. $\frac{f(x|\mu,\sigma^2)}{f(y|\mu,\sigma^2)} = exp[\frac{1}{2\sigma^2}(-n(\bar{x}^2 - \bar{y}^2) + 2n\mu(\bar{x} - \bar{y}) - (n-1)(s_x^2 - s_y^2))]$
  This ratio does not depend on $\mu$ and $\sigma^2$ if $\bar{x} = \bar{y}$, $s_x^2 = s_y^2$ (or equivalently $\sum x = \sum y$ and $\sum x^2 = \sum y^2$).
  MSS: $(\bar{X}, S_X^2)$ or $(\sum X, \sum X^2)$
  One can show that $S^2$ does not depend on $\mu$ (in fact is distributed $\sigma^2/(n-1)\chi^2_{n-1}$) and is independent of $\mu$.

- n samples, $X \sim U[\theta, \theta+1]$. $f(x|\theta) = 1\{\theta < x_i < \theta + 1\} = 1\{\max_i x_i - 1 < \theta < \min_i x_i\}$.
  MSS is obtained from ratio $f(x|\theta)/f(y|\theta)$ and is thus $(X_{(1)}, X_{(n)})$.
  However, any change of variables (one-to-one map) is also a MSS: e.g. $(X_{(n)} - X_{(1)}, (X_{(1)} + X_{(n)})/2) = (R, M)$, the range and mean.
  One can show that the range, $R$ is ancillary (intuitive also since the support of $f$ is fixed length 1). So, alone, the range does not carry any information about $\theta$, but together with the mean value, $M$, it provides information.

- $\tilde{X}$ consists of 2 samples from $f(x|\theta) = 1/3$ for $x = \theta, \theta+1, \theta+2$, $\theta$ an integer.
  Consider all possibilities. There are 9 cases, each equally likely.

| $x_1/x_2$ | $\theta$ | $\theta+1$ | $\theta+2$ |
|---|---|---|---|
| $\theta$ | $R=0, M=\theta$ | $R=1, M=\theta+.5$ | $R=2, M=\theta+1$ |
| $\theta+1$ | $R=1, M=\theta+.5$ | $R=0, M=\theta$ | $R=1, M=\theta+1.5$ |
| $\theta+2$ | $R=2, M=\theta+1$ | $R=1, M=\theta+1.5$ | $R=0, M=\theta+2$ |

  The distribution for $R$ does not depend on $\theta$ and so is ancillary.
  Let's say we are given an observed value for $M = m$, $m$ integer valued. Then, we know that $\theta = m, m-1$ or $m-2$ with relative probability $1/5, 3/5, 1/5$. However, if we also find out that $R = 2$, then we know that $\theta = m-1$. So, while $R$ is ancillary, together with $M$ it resolved the value of $\theta$.

# 5 Completeness, Exponential families

Often, $T$ and $S$ *are* independent and so estimation can focus on $\theta$ (but as we shall see later, interval estimation or testing will perhaps use $S$ nonetheless).

When are $T$ and $S$ independent? When $T$ is MSS and *complete*.

Completeness: $f(t|\theta)$ is a complete family of pdf's for statistic $T(X)$. If $E_\theta[g(T)] = 0$ for all $\theta$ implies $g(T) = 0$ for all $\theta$, then $T(X)$ is a complete statistic (and vice versa).

Basu's theorem: If $T(X)$ is MSS and complete, then $T(X)$ is independent of every ancillary statistic $S(X)$.

Proof: If $S(X)$ is ancillary, then $P(S(x) = s)$ is not a function of $\theta$.
Also, $P(S(X) = s|T(X) = t)$ is not a function of $\theta$ by the definition that $T$ is a sufficient statistic.
We need only show that $P(S(X) = s|T(X) = t) = P(S(X) = s)$ for all $t$ (showing $S$ and $T$ are independent).

We have
(a) $P(S(X) = s) = \sum_{t \in T} P(S(X) = s|T(X) = t)P_\theta(T(X) = t)$ (definition of marginal probability)
(b) but also $sum_{t \in T} P_\theta(T(X) = t) = 1$ and so $P(S(X) = s) = \sum_{t \in T} P(S(X) = s)P_\theta(T(X) = t)$
(c) Define $g(t) = P(S(X) = s|T(X) = t) - P(S(X) = s)$ and take (a)-(b): we have

$$0 = E_\theta g(T) = \sum_{t \in T} g(t)P_\theta(T(X) = t), for\ all\ \theta$$

but since $T$ is complete this means that $g(t) = 0$ and so $P(S(X) = s|T(X) = t) - P(S(X) = s)$. QED.

It is not very easy to show completeness. Is it important? It is a technicality that ensures uniqueness in estimation. Luckily, it's established for $f(x|\theta)$ in the exponential family distribution.

Exponential family: If $X_i$ iid $f(x|\theta)$ where

$$f(x|\theta) = h(x)c(\theta)exp[w(\theta)t(x)]$$

then $T(\tilde{X}) = \sum_{i=1}^n t(X_i)$ is complete.
For multiple parameters, $\theta = (\theta_1, \ldots, \theta_r)$:

$$f(x|\theta) = h(x)c(\theta)exp[\sum_{j=1}^k w_j(\theta)t_k(x)]$$

Examples

- $X \sim N(\mu, \sigma^2)$, assume $\sigma^2$ is fixed.
  Then
  $$f(x|\mu, \sigma^2) = exp(\mu x/\sigma^2)exp(-\mu^2/2\sigma^2 + x^2/2\sigma^2 + \log(2\pi sigma^2)/2)$$
  and so $t(x) = x$, $T(\tilde{X}) = \sum X_i$.

- 2 parameter case.
  $$f(x|\mu, \sigma^2) = exp(\frac{\mu}{\sigma^2}x - \frac{1}{\sigma^2}x^2)exp(\frac{-\mu}{2\sigma^2} + \log(2\pi\sigma^2))$$
  and so $t_1(x) = x, t_2(x) = x^2$. $T_1(X) = \sum_i X_i, T_2(X) = \sum_i X_i^2$.
  1-1 map: alternative $(T_1(X) = \bar{X}, T_2(X) = S^2)$.
  Note, $S^2$ is ancillary for $\mu$ and from the above we thus have that $S$ and $T(X) = \bar{X}$ are independent.

- $f(x|\theta) = (\theta/2)^{|x|}(1-\theta)^{(1-|x|)}$, $x \in -1, 0, 1$, $0 \leq \theta \leq 1$.
  $X$ is a sufficient statistic since it is the data. However, it is not complete.
  $E[g(X)] = g(-1)\theta/2 + g(0)(1-\theta) + g(1)\theta/2 = 0$ requires $g(0) = 0, g(1) = -g(-1)$ which does not mean that $g(t) = 0$ for all values $t$.
  However, change of variables to $Y = |X|$, $f(y|\theta) = \theta^y(1-\theta)^{1-y}$, for $y \in 0, 1$.
  $E[g(Y)] = g(1)\theta + g(0)(1-\theta) = 0$ only if $g(0) = g(1) = 0$ so $|X|$ is complete.
  Note also that $f(x|\theta) = exp[|x|log(\theta/(2(1-\theta)))](1-\theta)$, an exponential family with $t(x) = |x|$ so $T(X) = |X|$ complete.

- $f(x|\theta) = \theta x^{\theta-1}$, $0 < x < 1$ and $\theta > 0$.
  $f(\tilde{x}|\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n (\prod_{i=1}^{n} x_i)^{\theta-1}$ and so $\prod_i X_i$ is sufficient (not $\sum_i X_i$ as we often expect).
  $f(x|\theta) = exp[(\theta - 1)log(x) + log(\theta)]$ is an exponential family, $t(x) = log(x)$ and so $T(X) = \sum_i log(X_i) = log(\prod_i X_i)$ and as log is 1-1 in $0 < x < 1$ it follows that $\prod_i X_i$ is complete.