Math 4501 - Probability and Statistics II

6.7 - Sufficient Statistics

Relation with maximum likelihood estimators

Theorem

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution with pmf/pdf $f(x;\theta)$, with $\theta \in \Omega$. ← unknown parameter

If a sufficient statistic $Y = u(X_1, X_2, \ldots, X_n)$ for θ exists and if a maximum likelihood estimator $\hat{\theta}$ of θ also exists uniquely, then $\hat{\theta}$ is a function of $Y = u(X_1, X_2, \ldots, X_n)$.

REMARK: in this theorem $\hat{\theta} = v(Y)$ + v is invertible → $\hat{\theta}$ is also sufficient

↳ INTERPRETATION: MLE are based on sufficient statistic and (under mild conditions) MLE are sufficient too!

Proof sketch: $x_1, \ldots, x_n$ random sample from distribution $f(x;\theta)$, $\theta \in \Omega$

$Y = u(x_1, \ldots, x_n)$ is sufficient iff

FISHER-NEYMAN FACTORIZATION THM

$$f_{joint}(x_1, \ldots, x_n; \theta) = \phi(y, \theta) \cdot h(x_1, \ldots, x_n)$$

ϕ depends on $x_1, \ldots, x_n$ only through

h does not depend on θ

BUT, then the likelihood function is

$$L(\theta) = \phi(y, \theta) \cdot h(x_1, \ldots, x_n)$$ ← does not depend on θ

Thus, to maximize $L(\theta)$ w.r.t to θ, we need to maximize $\phi(y, \theta)$ w.r.t θ

When we maximize ϕ wrt θ, we obtain $\hat{\theta} = \hat{\theta}(Y)$, which by assumption is unique $\Rightarrow$ MLE is a function of the sufficient statistic!!

More than one unknown parameter

The notions and results discussed above admit a natural extension to the case of two or more unknown parameters.

We illustrate such extension by discussing the case of two unknown parameters.

Definition (Jointly sufficient statistics)

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf/pmf $f(\cdot; \theta_1, \theta_2)$ depending on two parameters $(\theta_1, \theta_2) \in \Omega$.

The statistics $Y_1 = u_1(X_1, \ldots, X_n)$ and $Y_2 = u_2(X_1, \ldots, X_n)$ are said to be *jointly sufficient* for θ_1 and θ_2 if the conditional distribution of $X_1, \ldots, X_n$ given $Y_1 = y_1$ and $Y_2 = y_2$ does not depend on (θ_1, θ_2) for any values y_1 of Y_1 and y_2 of Y_2.

Same interpretation as in the one-parameter case: once the sufficient statistics are given, there is no additional information about the parameters left in the remaining (conditional) distribution.

$P(X_1 = x_1, \ldots, X_n = x_n \mid Y_1 = y_1, Y_2 = y_2)$ does not depend on θ_1 and θ_2

Theorem (Fisher-Neyman Factorization Theorem: two parameters)

Let $X_1, X_2, \ldots, X_n$ denote random variables with joint pdf/pmf $f(x_1, x_2, \ldots, x_n; \theta_1, \theta_2)$ depending on the parameters θ_1 and θ_2.

The statistics $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ and $Y_2 = u_2(X_1, X_2, \ldots, X_n)$ are jointly sufficient for θ_1 and θ_2 if and only if

$$f(x_1, \ldots, x_n; \theta_1, \theta_2) = \phi(u_1(x_1, \ldots, x_n), u_2(x_1, \ldots, x_n); \theta_1, \theta_2) h(x_1, \ldots, x_n),$$

does not depend on θ_1 or θ_2

where ϕ depends on $x_1, \ldots, x_n$ only through $u_1(x_1, \ldots, x_n)$ and $u_2(x_1, \ldots, x_n)$ while $h(x_1, \ldots, x_n)$ does not depend upon θ_1 or θ_2.

Note: as in the one-parameter case, it is often easier to check sufficiency using the Factorization Theorem than it is using the definition.

Remark

Let Y_1 and Y_2 be jointly sufficient for θ_1 and θ_2, and let

$$(Z_1, Z_2) = \nu(Y_1, Y_2)$$

be a function of Y_1 and Y_2 not involving θ_1 and θ_2 that has a single-valued inverse.

Then Z_1 and Z_2 are also jointly sufficient statistics for θ_1 and θ_2.

Example

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta_1, \theta_2)$, $\theta_1 \in \mathbb{R}$, $\theta_2 > 0$.

Show

$$Y_1 = \sum_{i=1}^{n} X_i \quad \text{and} \quad Y_2 = \sum_{i=1}^{n} X_i^2$$

are jointly sufficient statistics for θ_1 and θ_2.

Conclude that

$$\bar{X} = \frac{Y_1}{n} \quad \text{and} \quad S^2 = \frac{Y_2 - \frac{1}{n} Y_1^2}{n-1}$$

are also jointly sufficient statistics for θ_1 and θ_2

μ σ^2

pdf of $N(\theta_1, \theta_2)$ is

$$f(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x-\theta_1)^2}{2\theta_2}}$$

$\bar{X}$ is the MLE for μ → ← S^2 is a function of MLE for σ^2

To show that Y_1 and Y_2 are jointly sufficient for θ_1 and θ_2, we will use the Fisher-Neyman factorization theorem:

$$f_{\text{joint}}(x_1, \ldots, x_n; \theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} e^{-(x_i-\theta_1)^2/2\theta_2}$$

$$= \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \exp\left(\sum_{i=1}^{n} -\frac{(x_i-\theta_1)^2}{2\theta_2}\right) =$$

$$= \left(2\pi\theta_2\right)^{-n/2} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right)$$

$$= \left(2\pi\theta_2\right)^{-n/2} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}\left(x_i^2 - 2\theta_1 x_i + \theta_1^2\right)\right)$$

$$= \left(2\pi\theta_2\right)^{-n/2} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n} x_i^2 + \frac{\theta_1}{\theta_2}\sum_{i=1}^{n} x_i - \frac{n\theta_1^2}{2\theta_2}\right)$$

$$= \underbrace{\left(2\pi\theta_2\right)^{-n/2} \exp\left(\frac{\theta_1}{\theta_2}\underbrace{\sum_{i=1}^{n} x_i}_{Y_1} - \frac{1}{2\theta_2}\underbrace{\sum_{i=1}^{n} x_i^2}_{Y_2} - \frac{n\theta_1^2}{2\theta_2}\right)}_{2\,\phi(y_1,y_2,\theta_1,\theta_2)} \cdot \underbrace{1}_{h_2(x_1,\ldots,x_n)}$$

CONCLUSION:

$$f_{joint}(x_1,\dots,x_n;\theta_1,\theta_2) = \phi(y_1,y_2,\theta_1,\theta_2)\cdot h(x_1,\dots,x_n)$$

with $$\phi(y_1,y_2,\theta_1,\theta_2) = (2\pi\theta_2)^{-n/2}\cdot \exp\left(\frac{\theta_1}{\theta_2}y_1 - \frac{1}{2\theta_2}y_2 - \frac{n\theta_1^2}{2\theta_2}\right)$$

$$y_1 = \sum_{i=1}^{n} x_i \quad , \quad y_2 = \sum_{i=1}^{n} x_i^2$$

$$h(x_1,\dots,x_n) = 1$$

which means that

$$Y_1 = \sum_{i=1}^{n} X_i \quad \text{and} \quad Y_2 = \sum_{i=1}^{n} X_i^2$$

are jointly sufficient for θ_1 and θ_2

To see that $\bar{X}$ and s^2 are also sufficient, it is enough to check that $\bar{X}$ and s^2 may be written as an invertible function of Y_1 and Y_2

that is, there exists an invertible function v such that

$$(\bar{X}, s^2) = v(Y_1, Y_2)$$

Note that:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} Y_1$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{X}\underbrace{\sum_{i=1}^{n} X_i}_{n\bar{X}} + n\bar{X}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \underbrace{2n\bar{X}^2}_{n\bar{X}^2}\right)$$

$$= \frac{\overbrace{\sum_{i=1}^{n} X_i^2}^{Y_2} - \frac{1}{n}\left(\overbrace{\sum_{i=1}^{n} X_i}^{Y_1}\right)^2}{n-1}$$

CONCLUSION:

$\bar{X} = \frac{1}{n} Y_1$

$S^2 = \dfrac{Y_2 - \frac{1}{n}(Y_1)^2}{n-1}$

inverse is

$Y_1 = n\bar{X}$

$Y_2 = (n-1)S^2 + n(\bar{X})^2$

$\Rightarrow$ $\bar{X}$ and S^2 are jointly sufficient statistics for μ and σ^2

that is $(\bar{X}, S^2) = v(Y_1, Y_2)$,

where $v(y_1, y_2) = \left(\frac{1}{n} y_1, \frac{y_2 - y_1^2/n}{n-1}\right)$

is invertible

because we can solve the system

$x_1 = \frac{1}{n} y_1$

$x_2 = \frac{y_2 - y_1^2/n}{n-1}$

for y_1 and y_2 in term x_1 and x_2

and the solution is unique

Math 4501 - Probability and Statistics II

7.1 - Confidence intervals for means

Overview

We will see how to determine interval estimates for the unknown mean of a certain distribution.

We will require that the actual unknown mean falls within such intervals with a specified likelihood, representing the level of confidence in our interval estimate:

- we call such interval estimates *confidence intervals*.

exact results

We will construct confidence intervals for the mean of:

- a normal distribution with known variance
- a normal distribution with unknown variance

approximations

We will construct approximate confidence intervals for the mean of:

- an unknown distribution, using a large sample (known and unknown variance)
- an unknown distribution, using a smaller sample ← smaller but not necessarily small!

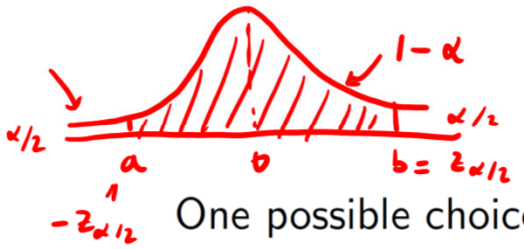
Normal distribution with known variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$ with known variance σ^2.

We will use $\bar{X}$, the unbiased estimator of the distribution mean μ, to determine an interval estimate for μ, called a *confidence interval for* μ.

We have previously seen that $\bar{X}$ is $N(\mu, \sigma^2/n)$ and so $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is $N(0,1)$.

Take a small number $\alpha \in (0,1)$, and look for two numbers a and b such that



$$P\left(a \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le b\right) = 1 - \alpha.$$

One possible choice is to take $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$ to get

$$P\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1 - \alpha$$

Solve the inequality in

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

for μ to get

$$-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

multiply by $\frac{\sigma}{\sqrt{n}}$

$$-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{X}-\mu \leq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

add $-\bar{X}$ to all terms

$$-\bar{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq -\mu \leq -\bar{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

multiply by -1 (invert inequalities)

$$\bar{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Since the probability of the first inequality is $1-\alpha$, so is the probability of the last.

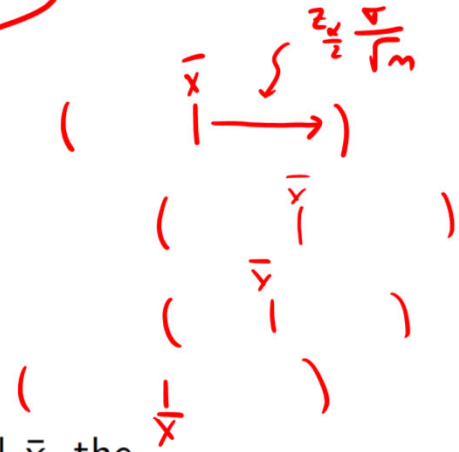
Thus, we have

$$P\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1-\alpha\ ,$$

and we conclude that the probability of the random interval

$$\Rightarrow \quad \left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

including the unknown mean μ is $1-\alpha$



Once the sample is observed and the sample mean computed to equal $\bar{x}$, the interval

$$\rightarrow \quad \left[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})\right]$$

becomes known.

Since the probability that the random interval

$$\left[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})\right]$$

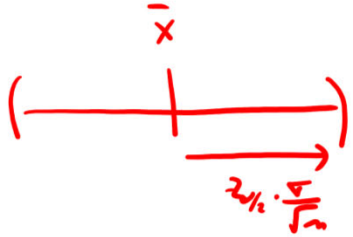
covers μ before the sample is drawn is equal to $1 - \alpha$, we call the computed interval a $100(1 - \alpha)\%$ *confidence interval for the unknown mean* μ

The number $100(1 - \alpha)\%$, or just, $1 - \alpha$, is called the *confidence coefficient*.

Notes:

1) The confidence interval for μ:

$$\left[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})\right]$$



is centered at the point estimate $\bar{x}$ and has length $2z_{\alpha/2}(\sigma/\sqrt{n})$.

2) As n increases, $z_{\alpha/2}(\sigma/\sqrt{n})$ decreases, resulting in a shorter confidence interval with the same confidence coefficient $1-\alpha$.

- A shorter confidence interval gives a more precise estimate of μ, regardless of the confidence we have in the estimate of μ.
- There is a potential cost (time, money, effort) in increasing the sample size, even when more observations may be collected.

3) For a fixed sample size n, the length of the confidence interval can also be shortened by decreasing the confidence coefficient $1-\alpha$.

- If this is done, we achieve a shorter confidence interval at the expense of losing some confidence.

Example

Let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer.

normal distribution with known variance

Assume that the distribution of X is $N(\mu, 1296)$. A random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $\bar{x} = 1478$ hours.

Determine a 95% confidence interval for μ.

Key reasoning on how to get CI:

σ^2 is known

$\Rightarrow X_1, \dots, X_n \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$\Rightarrow P\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1 - \alpha$

Note that:

1. sample is taken from a normal distribution
2. variance is known $\sigma^2 = 1296 \Rightarrow \sigma = \sqrt{1296} = 36$

The 95% confidence interval for μ is

solve for n

Endpoints are $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$$\left[\bar{x} - z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right)\right] =$$

$$= \left[1478 - 1.96\left(\frac{36}{\sqrt{27}}\right), 1478 + 1.96\left(\frac{36}{\sqrt{27}}\right)\right] = [1464.42, 1491.58] .$$