Inferens 1, F2

Rolf Larsson

Uppsala Universitet

2 november 2022

Today

- 7.1: Statistical inference, introduction
- 7.2: Estimation
  - 7.2.1: Properties of estimates
  - 7.2.2: Asymptotic properties

In WMS: 8.1-4

Statistical inference, introduction

Example 1:

- In an opinion poll, 1000 randomly selected voters are asked about their political sympathies.
- Let X be the number of these voters who sympathize with the party P.
- Because the number of voters is so large, we may assume that $X \sim \mathrm{Bin}(1000, p)$.
- Say that 100 of the selected voters sympathize with the party P.
- Which is your guess on the value of p?

Statistical inference, introduction

http://www.fjellfotografen.se/Djur/Däggdjur/Älgar/Älgtjur i skog/@toj-03020

Example 2:

- A researcher measures the withers height (mankhöjd) of five randomly selected moose (älgtjurar).
- The researcher assumes that the withers height in cm is $N(\mu, \sigma^2)$.
- The measurements are
  200 185 210 190 190
- Which is your guess on the value of μ?
  190 195 197.5

Statistical inference, introduction

- *Probability*:

  Model with *known* parameters
  $\rightarrow$ (probabilities for) random samples.

- *Statistical inference*:

  Model with *unknown* parameters, observed random sample
  $\rightarrow$ make inference on (e.g. estimate) parameter values.

Statistical inference, introduction

Definition (7.1)

$x_1, x_2, ..., x_n$ is a *sample* (stickprov) from the random variable X with distribution F, if $x_1, x_2, ..., x_n$ are observations of $X_1, X_2, ..., X_n$, which all have distribution F.

If, in addition, the variables $X_1, X_2, ..., X_n$ are independent, we have a *random sample* (slumpmässigt stickprov) from X (from F).

Statistical inference, introduction

Example 1:

- In an opinion poll, 1000 randomly selected voters are asked about their political sympathies.
- Say that 100 of the selected voters sympathize with the party P.
- This is a random sample of size $n = 1$ where $x_1 = 100$ is an observation of the random variable $X_1 \sim \mathrm{Bin}(1000, p)$.
- Alternatively: A random sample $u_1, ..., u_n$, as observations of $U_1, ..., U_n$, $n = 1000$, where each $U_i \sim \mathrm{Be}(p)$ and where the U_i are independent. We observe

$$\sum_{i=1}^{n} u_i = 100.$$

Statistical inference, introduction

Example 2:

- A researcher measures the withers height (mankhöjd) of five randomly selected moose (älgtjurar).
- The researcher assumes that the withers height in cm is $N(\mu, \sigma^2)$, where it is known that $\sigma^2 = 100$.
- The measurements are
  200 185 210 190 190
- $(x_1, x_2, x_3, x_4, x_5) = (200, 185, 210, 190, 190)$ is a random sample from $X \sim N(\mu, 100)$, i.e.
- $x_1 = 200$ is an observation of the random variable $X_1 \sim N(\mu, 100)$,
  $x_2 = 185$ is an observation of the random variable $X_2 \sim N(\mu, 100)$,
  $x_3 = 210$ is an observation of the random variable $X_3 \sim N(\mu, 100)$,
  $x_4 = 190$ is an observation of the random variable $X_4 \sim N(\mu, 100)$,
  $x_5 = 190$ is an observation of the random variable $X_5 \sim N(\mu, 100)$,
  where X_1, X_2, X_3, X_4, X_5 are independent (and distributed as X).

Estimation

- Suppose we have one unknown parameter θ.
- Write

$$\mathbf{x} = (x_1, x_2, ..., x_n)$$
$$\mathbf{X} = (X_1, X_2, ..., X_n)$$

Definition (7.2)

An estimate (skattning) $\theta^* = \theta^*(\mathbf{x})$ is a function of the sample $\mathbf{x}$.

The estimate is an observation of the *estimator* $\theta^*(\mathbf{X})$.

Estimation

Example 2:

- $(x_1, x_2, x_3, x_4, x_5) = (200, 185, 210, 190, 190)$ is a random sample from $X \sim N(\mu, 100)$, i.e.
- $x_1 = 200$ is an observation of the random variable $X_1 \sim N(\mu, 100)$,

  $\vdots$

  $x_5 = 190$ is an observation of the random variable $X_5 \sim N(\mu, 100)$, where X_1, X_2, X_3, X_4, X_5 are independent and distributed as X.
- Estimate μ by the sample mean.
- Estimate

$$\mu^*(x_1, x_2, x_3, x_4, x_5) = \overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = 195$$

- Estimator

$$\mu^*(X_1, X_2, X_3, X_4, X_5) = \overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} \sim N(\mu, 20)$$

Estimation

Example 1:

- In an opinion poll, 1000 randomly selected voters are asked about their political sympathies.
- We have one observation $x = 100$ of the random variable $X \sim \mathrm{Bin}(1000, p)$.
- Estimate $p^* = x/1000 = 100/1000 = 0.1$.
- Estimator $p^*(X) = X/1000$.

Properties of estimates

- $\mathbf{X} = (X_1, ..., X_n)$
- Estimator $\theta^*(\mathbf{X})$
- $\theta^* - \theta = [E\{\theta^*(\mathbf{X})\} - \theta] + [\theta^* - E\{\theta^*(\mathbf{X})\}]$

  $= \text{systematic error} + \text{random error}$

Definition (7.3)

An estimate θ^* is said to be *unbiased* (väntevärdesriktig) if it satisfies $E\{\theta^*(\mathbf{X})\} = \theta$

i.e. it has no systematic error.

Properties of estimates

Example 2:

$$\mu^*(\mathbf{X}) = \overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} \sim N(\mu, 20)$$

Is $\mu^*(\mathbf{x})$ an unbiased estimate of μ?

Example 1:

$$p^*(X) = \frac{X}{1000}$$

where $X \sim \mathrm{Bin}(1000, p)$.
Is $p^*(\mathbf{x})$ an unbiased estimate of p?

Properties of estimates

If we have more than one unbiased estimate, which is the best one?

Definition (7.4)

If θ_1^* and θ_2^* are unbiased estimates of θ and

$$V\{\theta_1^*(\mathbf{X})\} \leq V\{\theta_2^*(\mathbf{X})\}$$

for all θ with strict inequality for some θ, we say that θ_1^* is *more efficient* than θ_2^*.

Properties of estimates

Example 2:

- $(x_1, x_2, x_3, x_4, x_5) = (200, 185, 210, 190, 190)$ is a random sample from $X \sim N(\mu, 100)$.
- Some possible estimates:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{200 + 185 + 210 + 190 + 190}{5} = 195,$$

$$\text{median}(x_1, x_2, x_3, x_4, x_5) = 190,$$

$$\text{midrange} = \frac{\min x_i + \max x_i}{2} = \frac{185 + 210}{2} = 197.5.$$

- It may be shown that all of these estimates are unbiased.
- Which one is most efficient?

Properties of estimates

Example 2:

- Suppose that $X_1, X_2, .., X_n$ are independent $N(\mu, \sigma^2)$, $n = 5$.
- It may be shown that
  - $V(\overline{X}) = \sigma^2/5 = 0.2\sigma^2$, (most efficient!)
  - $V(\text{median}) \approx 0.29\sigma^2$,
  - $V(\text{midrange}) \approx 0.23\sigma^2$.
- For n arbitrary,
  - $V(\overline{X}) = \sigma^2/n$,
  - $V(\text{median}) \approx (\pi/2)\sigma^2/n$, n large,
  - $V(\text{midrange}) \approx (\pi^2/24)\sigma^2/(\ln n)$, n large.
- For $n = 5$, the median (as a random variable), Y say, has density function

$$f(y) = 30 f_X(y) F_X(y)^2 \{1 - F_X(y)\}^2$$

and the minimum Y_1 and maximum Y_2 have the simultaneous density

$$f(y_1, y_2) = 20 f_X(y_1) f_X(y_2) \{F_X(y_2) - F_X(y_1)\}^3.$$

Properties of estimates

Example 2′:

- A researcher randomly selects four moose and observes the withers heights $x_1, ..., x_4$, assumed to be a random sample from $N(\mu, 100)$.
- Another researcher randomly selects five other moose and observes the withers heights $y_1, ..., y_5$, assumed to be a random sample from $N(\mu, 100)$.
- Based on this, the following two estimates of μ are proposed:

$$\mu_1^* = \frac{\overline{x} + \overline{y}}{2}$$

$$\mu_2^* = \frac{4 * \overline{x} + 5 * \overline{y}}{9}$$

1. Are the estimates unbiased?
2. If so, which one is most efficient?

Properties of estimates

Example 3 (stratification):
We are interested in the proportion p of Swedish citizens that last year have traveled by plane in connection with work. To find out, we want to take a sample consisting of $n = 1000$ people. Consider the following two ways to do this:

- Randomly draw 1000 people, of which x traveled by plane, and use the estimate
$$p_1^* = \frac{x}{n}.$$
- Randomly draw 500 men and 500 women, of which y and z, respectively, traveled by plane, and use the estimate
$$p_2^* = \frac{y+z}{n}.$$

1. Are the estimates unbiased?
2. If so, which one is most efficient?

Properties of estimates

How can we assign a numerical value to the dispersion of an estimate?

Definition (7.6)

The standard error (medelfelet) of the estimate θ^* is an estimate of the standard deviation $D\{\theta^*(\mathbf{X})\}$. It is denoted by $d\{\theta^*(\mathbf{x})\} = d(\theta^*)$.

Properties of estimates

Example 1:

- We have one observation $x = 100$ of the random variable $X \sim \mathrm{Bin}(1000, p)$.
- Estimate $p^* = x/1000 = 100/1000 = 0.1$.
- Calculate the standard error of this estimate.

  Answer: 0.0094

Properties of estimates

Example 2:

- $(x_1, x_2, x_3, x_4, x_5) = (200, 185, 210, 190, 190)$ is a random sample from $X \sim N(\mu, 100)$.
- Estimate

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{200 + 185 + 210 + 190 + 190}{5} = 195$$

- Calculate the standard error of this estimate.

  Answer: 4.47

Asymptotic properties

The accuracy of an estimate should improve as the sample size increases.

Definition (s.270)

The bias (väntevärdesfelet) for the estimate θ^* is defined as

$$B(\theta^*) = E\left\{\theta^*(\mathbf{X})\right\} - \theta.$$

Observe: An unbiased estimate has bias zero.

Definition (7.7)

If the bias $B(\theta_n^*)$ tends to zero as $n \to \infty$ for all θ, the estimate θ_n^* is said to be *asymptotically unbiased*.

Observe: An unbiased estimate is always asymptotically unbiased.

Asymptotic properties

Example 4:

- Let $x_1, ..., x_n$ be a random sample from $N(\mu, \sigma^2)$, where μ is unknown. The object is to estimate σ^2.
- The estimate $\sigma_n^{2*} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is *not* unbiased, but it is asymptotically unbiased.
- The estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is unbiased for σ^2. Why?
- Is s_n unbiased for σ?

Asymptotic properties

Definition (7.9)

The estimate θ_n^* is said to be *consistent* for θ if the corresponding estimator converges to θ in probability for all θ.

Definition (7.8)

The estimator $\theta_n^*(\mathbf{X})$ converges to θ in probability if for all $\varepsilon > 0$,

$$P\left(|\theta_n^*(\mathbf{X}) - \theta| > \varepsilon\right) \to 0$$

as $n \to \infty$ for all θ.

Asymptotic properties

Theorem (Sats 7.2)

If the estimate θ_n^* is asymptotically unbiased and $V\{\theta_n^*(\mathbf{X})\} \to 0$ as $n \to \infty$ for all θ, then it is consistent.

Asymptotic properties

Example 2'':

- Let $x_1, ..., x_n$ be a random sample from $N(\mu, \sigma^2)$, where σ^2 is known.
- Estimate μ by

$$\mu_n^* = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

1. Show that the estimate is unbiased.
2. Calculate the variance for the corresponding estimator.
3. Show that the estimate is consistent.

News of today

- Random sample
- Estimate (skattning)
- Estimator
- Unbiased (väntevärdesriktig)
- More efficient
- Standard error (medelfel)
- Asymptotically unbiased
- Consistent

Problem: 7.2.2