Samplingsfördelning (ar)

En population har en parameter vars värde är okänt. Vi vill (hitta) gissa värdet utan att hålla hela populationen.

Ex1: Coca cola shall lansera den nya colan CCX som utmanare till Pepsi max. Hur stor del av populationen föredrar CCX framför Pepsi max?

(population = alla i sverige/USA/världen, parameter p = proportionen som föredrar CCX).

Ex2: Pigens frallor har normalfördelad vikt med parametrar μ o σ^2. Vad är dessa? (kvalitetskontroll).

Hur shall vi gissa p, μ, σ^2?

1/ vi samlar in data (stickprov)

2/ Data sammanfattes i en teststatistika

Fördelningen av teststatistikan är avgörande för hur bra vår gissning blir.

Ex1 (forts): CC frågar 1000 personer. Låt

$$X_k = \begin{cases} 1 & \text{om person } k \text{ föredrar CCX} \\ 0 & \text{" " " " Pepsi} \end{cases}$$

Våra data är $x_1, x_2, \ldots, x_{1000}$.

För att gissa p tar vi $\frac{x_1+x_2+\cdots+x_{1000}}{1000} = \bar{x}$

dvs andelen svarande som föredrog CCX.

Är $p = \bar{x}$? Nej!! Är p nära $\bar{x}$? Förhoppningsvis!

Ex 2: Påsen väger 500 trallor. Vikterna kallas $x_1, x_2, \ldots, x_{500}$ (data).

$\bar{x} = \frac{x_1 + \cdots + x_{500}}{500}$ är teststatistikan för μ

$s^2 = \frac{1}{499} \sum_{k=1}^{500} (x_k - \bar{x})^2$ " " σ^2.

OBS!!! Det finns ett extremt viktigt före/efter (mätning) perspektiv.

1/ När modellen sätts upp (före mätning) är vikterna i ex 2 slumpvariabler $\left(\overline{X}_k \sim N(\mu, \sigma^2)\right)$.

2/ Efter att mätningen genomförts är det numeriska data, (x_k).

Ex 1: Innan mätning genomförts har vi

$\overline{X} = \frac{\overline{X}_1 + \cdots + \overline{X}_{1000}}{1000}$ som är en slumpvariabel.

Efter mätningen genomförts har vi

$$\bar{x} = \frac{x_1 + \cdots + x_{1000}}{1000}$$ som är ett numerisk värde.

Om väntevärdet på teststatistikan = värdet på sökta parametern så är teststatistikan väntevärdesriktig (VVR) (unbiased)

Ex 1: $$E[\bar{X}] = E\left[\frac{X_1 + \cdots + X_{1000}}{1000}\right] = \frac{E[X_1] + \cdots + E[X_{1000}]}{1000}$$

$= E[X_1] = \mu$ så VVR!

Ex 2: Även $$E[s^2(X)] = E\left[\frac{1}{499}\sum_{k=1}^{500}(X_k - \bar{X})^2\right] = \sigma^2$$ så VVR

↑ lång räkning.

Låt oss fokusera på $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$.

1/ Som ovan gäller att $E[\bar{X}] = E[X_1] = \mu$

2/ vi har att $Var(\bar{X}) = \frac{Var(X_1)}{n} = \frac{\sigma^2}{n}$.

Med notationen $\sigma_{\bar{X}}^2 = Var(\bar{X})$ har vi att

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

OBS! Om $\sigma_{\bar{X}}$ är "stor" har vi stor spridning. Därmed är sann. stor att vår gissning missar målet rejält.

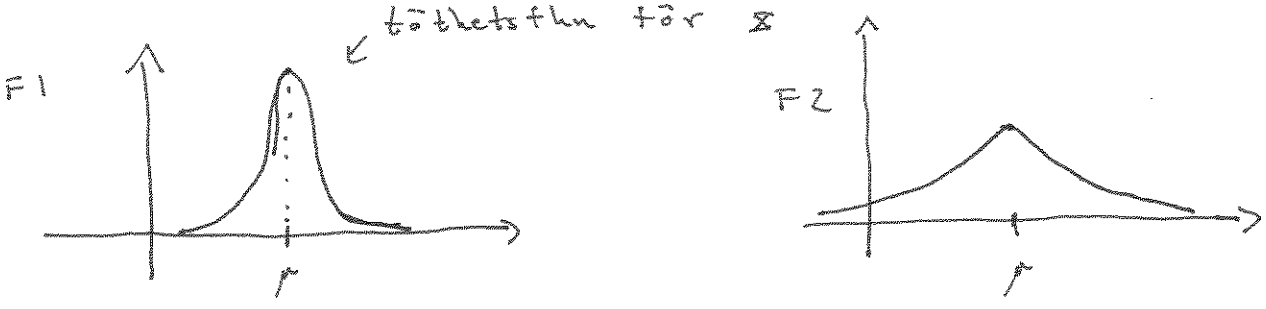
Sats 1: Om $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$ är $N(\mu, \sigma^2)$ (och oberoende) $\Rightarrow$ $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Sats 2: Om $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$ har någon fördelning (vilken som helst!!!!) med $\mu = E[\overline{X}_i]$, $\sigma^2 = \mathrm{Var}(\overline{X}_i)$ gäller att

$$\overline{X} \overset{\text{approximativt}}{\approx} N\left(\mu, \frac{\sigma^2}{n}\right).$$

Anm: Sats 2 kallas för CLT (Central Limit Theorem) och är ~~en~~ statistikens viktigaste resultat. Om n är stort blir datans ursprung (nästan) irrelevant!

Varför är fördelningen hos $\overline{X}$ viktig?



I F1 är det troligt att vi hamnar nära μ

I F2 " " " " " långt ifrån μ.
