TDAB01 Sannolikhetslära och Statistik

Jose M. Peña
IDA, Linköpings Universitet

Föreläsning 7

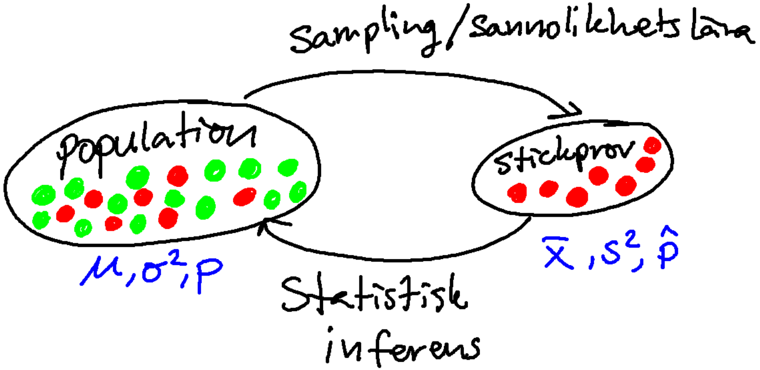
Översikt

- **Population, parametrar, stickprov och statistik**
- **Deskriptiv statistik**
- **Introduktion till parameterestimation och samplingfördelningar**
- **Grafiska metoder - demo**

Grundläggande begrepp

- **Population** = **alla** enheter av intresse.
  - Sveriges befolkning.
  - Alla möjliga handskrivna siffror.
  - Alla producerade enheter vid en fabrik.
- **Parameter** = numerisk beskrivning av **populationen**.
  - Genomsnittsinkomst (μ) eller inkomstspridning (σ^2).
  - Medelintensitet i gråskala vid en viss pixel i bilderna av en 8:a.
  - Andelen trasiga produkter.
- **Stickprov** (eng. sample) = en **delmängd** av observerade enheter från populationen.
  - 1000 slumpmässigt valda personer.
  - 1000 handskrivna siffror (0-9) av olika personer i olika åldrar.
  - 10 utvalda lådor med produkter.
- **Statistika** (eng. statistic) = sammanfattande funktion av **stickprovet**.
  - Medelvärdet $\bar{X}$, stickprovsvariansen s^2, eller andelen trasiga produkter $\hat{p}$.

Sannolikhetslära och statistisk inferens



Estimator

- **Populationsparameter:** θ. Okänd. **Inferens**/**learning**: Lära sig om θ från data.
- $\hat{\theta}$ är en **estimator** av θ. För ett givet stickprov får vi ett **estimat** (ett värde) av $\hat{\theta}$ som representerar vår **bästa "gissning"** av θ baserat på information i stickprovet.
- Exempel: $\theta = p$, sannolikheten i en sekvens Bernoulliförsök:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \text{andelen lyckade}$$

- $\hat{p}$ är **rätt i genomsnitt** sett över alla möjliga stickprov av storleken n

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} \mathbb{E}(X_i)}{n} = \frac{\sum_{i=1}^{n} p}{n} = \frac{np}{n} = p$$

- En estimator $\hat{\theta}$ av θ är **väntevärdesriktig** (eng. unbiased) om

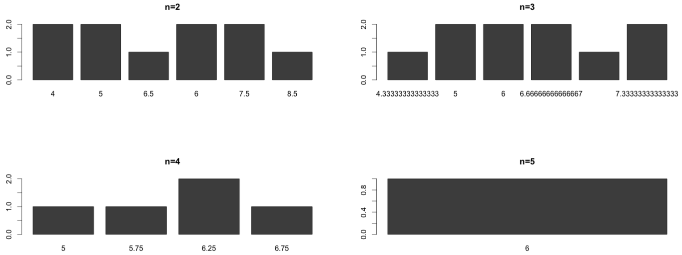
$$\mathbb{E}(\hat{\theta}) = \theta$$

- **Bias**:

$$Bias(\hat{\theta}) = \mathbb{E}\left(\hat{\theta}\right) - \theta$$

Samplingfördelning

- Men hur fel kan det bli i ett givet stickprov ?
- **Samplingfördelning** för $\hat{\theta}$ beskriver hur $\hat{\theta}$ kan variera från stickprov till stickprov.
- Exempel: Population $\{3, 5, 5, 7, 10\}$. $\theta = \frac{3+5+5+7+10}{5} = 6$.
- Stickprov av storleken $n = 3$:
  - Stickprov 1: $\{3, 5, 5\}$ med $\bar{x} = 4.333$.
  - Stickprov 2: $\{3, 5, 7\}$ med $\bar{x} = 5.000$.
  - $\vdots$
  - Stickprov 10: $\{5, 7, 10\}$ med $\bar{x} = 7.333$.
- Samplingfördelning för $\bar{X}$ med $n = 2, 3, 4, 5$:



Medelvärdesestimatorn

- Medelvärde: $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$
- Medelvärdet är en väntevärdesriktig estimator av $\mu = E(X)$, dvs $\mathbb{E}(\bar{X}) = \mu$.
- **Enkelt slumpmässigt urval** eller **oberoende likafördelade dragningar** (eng. independent and identically distributed eller **iid** samples): Samplingdesign där enheter väljs **oberoende** av varandra från populationen och med **lika sannolikheter**.
- **Samplingvarians** eller **standardfel** för $\bar{X}$ om $X_1, \ldots, X_n$ är **iid** med väntevärde μ och varians σ^2:

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Konsistens och centralagränsvärdessatsen

- $\bar{X}$ är en **konsistent** estimator av μ om samplingfördelningen blir alltmer koncentrerad kring μ när stickprovsstorleken n ökar.
- Formellt är estimatorn $\hat{\theta}$ konsistent för θ om, för alla $\varepsilon > 0$

$$\boldsymbol{P}\left\{\left|\hat{\theta} - \theta\right| > \varepsilon\right\} \to 0 \text{ när } n \to \infty$$

- **Sats**: För ett iid stickprov är $\bar{X}$ en konsistent estimator av $\mu = \mathbb{E}(X)$.
- **Bevis** via Chebyshevs olikhet:

$$\boldsymbol{P}\left\{\left|\bar{X} - \mu\right| > \varepsilon\right\} \leq \frac{Var(\bar{X})}{\varepsilon^2} = \frac{\sigma^2/n}{\varepsilon^2} \to 0 \text{ när } n \to \infty.$$

- **Centralagränsvärdessatsen** säger att samplingfördelningen för $\bar{X}$ är approximativt $N\left(\mu, \sigma^2/n\right)$ för stora n (tumregel: $n > 30$).
- Formellt, cdf:en för

$$Z = \frac{\bar{X} - \mathbb{E}(\bar{X})}{Std(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

konvergerar mot cdf:en för en standard normal $N(0, 1)$.

Konsistens och centralagränsvärdessatsen

- **Sats**: Om $X \sim N(\mu_X, \sigma_X^2)$ och $Y \sim N(\mu_Y, \sigma_Y^2)$ är oberoende så gäller att
$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$
- Om X och Y är beroende är $aX + bY$ fortfarande normalfördelad, men med annan varians.
- Detta resultat gäller även för flera variabler. Speciellt gäller för $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ att
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$
Då, i denna fall är $\bar{X}$ inte approximativt normalfördelad (pga CLT) utan **exakt** normalfördelad.

Median och kvantiler

- Medelvärdet är känsligt till extrema mätvärden, s k **outliers**.
- **Medianen** M är mer **robust**:

$$\boldsymbol{P}\left(X < M\right) \leq 0.5$$
$$\boldsymbol{P}(X > M) \leq 0.5$$

- Median = hälften av sannolikhetsmassan till vänster, hälften till höger.
- **Samplemedianen**:

$$\hat{M} = \begin{cases} \left(\frac{n+1}{2}\right)\text{:te minsta observationen} & \text{om } n \text{ udda} \\ \text{medelvärdet av } \left(\frac{n}{2}\right)\text{:te och } \left(\frac{n+2}{2}\right)\text{:te observationerna} & \text{om } n \text{ jämnt} \end{cases}$$

- Generalisering av median: p-**kvantil** är ett tal c som löser

$$\boldsymbol{P}\left(X < c\right) \leq p$$
$$\boldsymbol{P}(X > c) \leq 1 - p$$

- **Percentiler**: 5%, 37% etc. **Kvartiler**: 25%, 50%, 75%.
- R kod: `qnorm(p=0.05,mean=1,sd =2)` returnerar -2.289707.

Stickprovsvariansen

- Populationsvarians: $\sigma^2 = \mathbb{E}[(X-\mu)^2]$. Hur skattar man σ^2 från ett stickprov?
- **Stickprovsvariansen**:
$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\boldsymbol{n-1}}$$
- s^2 verkar vara en naturlig estimator, men **varför division med $n-1$?**
- **Svar**: Därför att bara med $n-1$ får man $\mathbb{E}(s^2) = \sigma^2$.
- **Bevis**: Vi kan skriva om s^2 som
$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2\bar{X}X_i)}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 + n\bar{X}^2 - 2n\bar{X}^2}{n-1}$$
$$= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$
Då
$$\mathbb{E}(s^2) = \frac{\sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\mathbb{E}\left(\bar{X}^2\right)}{n-1}$$
Dessutom
$$Var(X_i) = \sigma^2 = \mathbb{E}(X_i^2) - \mu^2 \text{ och } Var(\bar{X}) = \frac{\sigma^2}{n} = \mathbb{E}\left(\bar{X}^2\right) - \mathbb{E}(\bar{X})^2 = \mathbb{E}\left(\bar{X}^2\right) - \mu^2$$
Så
$$\sum_{i=1}^{n}\mathbb{E}(X_i)^2 - n\mathbb{E}\left(\bar{X}^2\right) = n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2(n-1)$$

Grafiska metoder - demo

- Se `SS7GraferDemo.R`.

Översikt

- **Population, parametrar, stickprov och statistik**
- **Deskriptiv statistik**
- **Introduktion till parameterestimation och samplingfördelningar**
- **Grafiska metoder - demo**