Sannolikhetslära och Statistik
Föreläsning 7

Mattias Villani

Avdelningen för Statistik och Maskininlärning
Institutionen för datavetenskap
Linköpings universitet

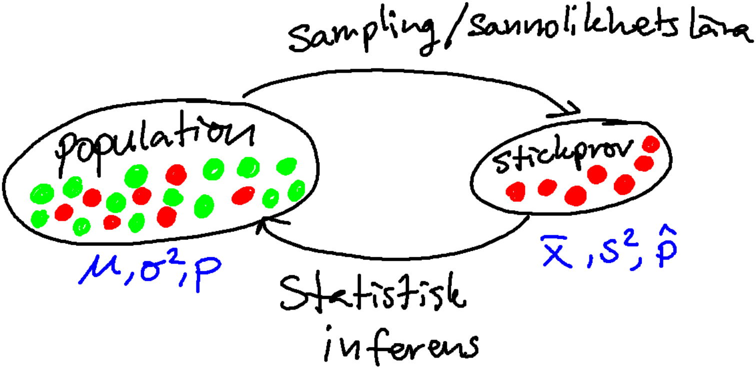
Översikt

- 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 en bild 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.
  - $\bar{X}$, medelvärdet. s^2, stickprovsvariansen, 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

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

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

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

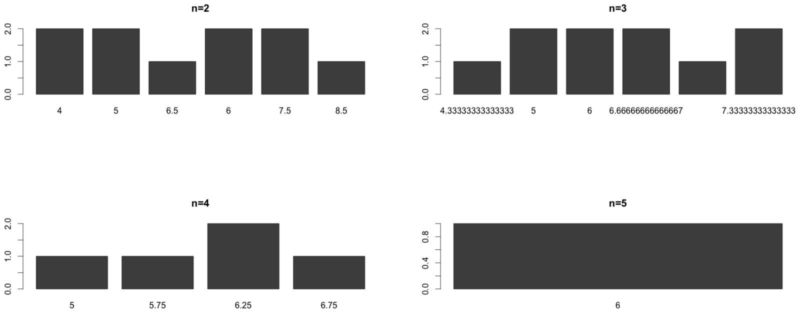
- **Bias**:

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

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.
- Ex. Population: $\{3, 5, 5, 7, 10\}$. $\theta = \frac{3+5+5+7+10}{5} = 6$.
- Stickprov av storleken n, utan återläggning.
- Samplingfördelning för $\bar{X}$.
- Ex. $n = 3$.
  - 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 $\bar{X}$



Medelvärdesestimatorn

- Medelvärde: $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$
- **Väntevärdesriktig** för $\mu = E(X)$

$$\mathbb{E}\bar{X} = \mu$$

- **Enkelt slumpmässigt urval**: **samplingdesign** där enheter väljs **oberoende** av varandra från populationen och med **lika sannolikheter**.
- **iid** (**i**ndependent **i**dentically **d**istributed). sv. **oberoende likafördelade**.
- **Samplingvarians**, eller **standardfel**, för $\bar{X}$ om $X_1, \ldots, X_n$ är **iid** med väntevärde μ och varians σ^2:

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

Linjärkombinationer av normalvariabler är normal

- 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 fel variabler. Speciellt gäller för $X_1, ..., X \overset{iid}{\sim} N(\mu, \sigma^2)$
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Konsistens och CLT

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

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

- **Sats**: för ett iid stickprov är $\bar{X}$ är konsistent för $\mu = \mathbb{E}X$.
- **Bevis**: Chebyshevs olikhet:

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

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

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

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

Median och kvantiler

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

$$P(X > M) \leq 0.5$$
$$P(X < M) \leq 0.5$$

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

$$\hat{M} = \begin{cases} \left(\frac{n+1}{2}\right) \text{:te minsta observationen} \\ \text{medelvärdet av } \left(\frac{n}{2}\right) \text{:te minsta observation och } \left(\frac{n+2}{2}\right) \text{:te observat} \end{cases}$$

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

$$P(X > c) \leq p$$
$$P(X < c) \leq 1 - p$$

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

Stickprovsvariansen

- Populationsvarians: $\sigma^2 = \mathbb{E}\left(X-\mu\right)^2$. Hur skatta σ^2 från stickprov?
- **Stickprovsvariansen**

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

- 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 (se Remark på sid. 220 för bevis):

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

$$\mathbb{E}s^2 = \frac{\sum_{i=1}^{n} \mathbb{E}X_i^2 - n\mathbb{E}\left(\bar{X}^2\right)}{n-1}$$

där $\mathbb{E}X_i^2 = Var(X_i) + \mu^2 = \sigma^2 + \mu^2$ och $\mathbb{E}\left(\bar{X}^2\right) = Var(\bar{X}) + \left(\mathbb{E}\bar{X}\right)^2 = \frac{\sigma^2}{n} + \mu^2$.
Så

$$\sum_{i=1}^{n} \mathbb{E}X_i^2 - n\mathbb{E}\left(\bar{X}^2\right) = n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2(n-1)$$