



GÖTEBORGS
UNIVERSITET

MVE051/MSG810 - Statistik och diskret matematik (LP4 2022-2023)

Chalmers och Göteborgs universitet

Föreläsning 7 – Introduktion till statistik

March 30, 2023

Dagens föreläsning

- Introduktion till statistik
- Lägesmått
 - Medelvärde, median
- Spridningsmått
 - Percentiler, kvartiler, stickprovsvarians
- Histogram och boxplot
- Skattare
- Egenskaper hos skattare

Inledning

Vad gör man med statistik/Vad är statistik för någonting?

- Statistik är att samla in, klassificera, summera, organisera, analysera, förstå och tolka data.
- Med hjälp av statistik kan man bestämma vilken information som är relevant givet problemet.
- Vi kan bestämma om man kan lita på slutsatserna som man har dragit.

Terminologi

Population

Slumpmässigt urval

Stickprov

Data

Variabel

Datatyper

- numeriska (kvantitativa) data
 - t.ex. längden av en person, drickvattnets temperatur, antalet elever i en skolklass
 - delas upp i
 - diskreta, ofta antal och heltal
 - kontinuerliga, t.ex. längd, ålder
- kategoriska (kvalitativa) data
 - t.ex. kön (man/kvinna/annat), blodgrupp (A/B/AB/0)
 - delas upp i
 - ordinaldata: finns en naturlig ordning, t.ex. betygen U, G och VG
 - nominaldata: saknar ordning, t.ex. kön

Exempel (draghållfasthet)

Man har mätt draghållfastheten (tensile strength, enhet 1000 pounds per square inch), dvs. den maximala spänningbelastningen som ett material kan hantera, av ett material som används för att bygga flygplan och fått följande mätningar:

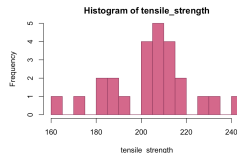
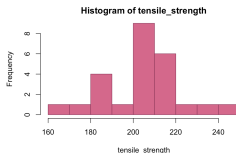
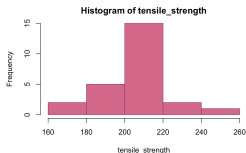
203.41 209.58 213.35 218.56 242.76 185.97 190.67 207.88 210.80 231.46
184.41 200.73 206.51 209.84 212.15 160.44 180.95 201.95 204.60 219.51
174.63 185.34 205.59 212.00 225.25

- Population: Allt material av typen som undersöks
- Datan: Siffrorna som samlades in
- Variabel: Uppmätt draghållfasthet hos materialet

Vad kan vi göra med data?

- Grafisk sammanställning
 - Histogram
 - Boxplot
 - ...
- Numerisk analys
 - Beräkning av lägesmått (median, medelvärde)
 - Beräkning av spridning (kvartiler, varians)
 - Skattning av parametrar

Histogram



Talet på y -axeln kan vara antingen antal n_i eller relativ frekvens n_i/n .

Medelvärde

Medelvärde

Medelvärdet av n observationer $x_1, x_2, \dots, x_n$ beräknas som

$$\frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Exempel (draghållfasthet)

Medelvärdet av draghållfastheten är

$$\frac{203.41 + 209.58 + \dots + 225.25}{25} = 203.93.$$

Median

Median

Om datamängden är i storleksordning, är medianen det mellersta värdet av observationerna. Om antalet observationer är jämnt är medianen medelvärdet av de två mellersta värdena.

Exempel (draghållfasthet)

160.44 174.63 180.95 184.41 185.34 185.97 190.67 200.73 201.95 203.41
204.60 205.59 206.51 207.88 209.58 209.84 210.80 212.00 212.15 213.35
218.56 219.51 225.25 231.46 242.76

Medianen blir alltså 206.51.

Percentiler och kvartiler

Percentiler

En q -percentil P_q är ett värde som är sådant att $q\%$ av observationer är lägre än detta värde (och $(100 - q)\%$ är större än detta värde).

Medianen kallas också för den 50:onde percentilen (hälften av värdena är lägre och hälften större än medianen).

Kvartiler

Tre vanliga percentiler är $Q_1 = P_{25}$, $Q_2 = P_{50}$ och $Q_3 = P_{75}$. Dessa kallas ofta för kvartiler.

Skattning av percentiler

Skattning av percentiler

Ett stickprov av storleken n delar populationen i $n + 1$ ungefär lika stora delar.

- Det minsta värdet kan användas som en skattning av $P_{100 \cdot 1 / (n+1)}$
- Det näst minsta värdet kan användas som en skattning av $P_{100 \cdot 2 / (n+1)}$
- ...

Exempel (draghållfasthet)

160.44 174.63 180.95 184.41 185.34 185.97 190.67 200.73 201.95 203.41
204.60 205.59 206.51 207.88 209.58 209.84 210.80 212.00 212.15 213.35
218.56 219.51 225.25 231.46 242.76

Vi uppskattar följande percentiler:

- $P_{100 \cdot 1 / 26} = P_{3.85} = 160.44$
- $P_{100 \cdot 2 / 26} = P_{7.69} = 174.63$
- $P_{100 \cdot 3 / 26} = P_{11.54} = 180.95$
- ...

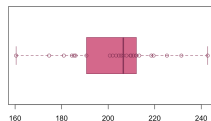
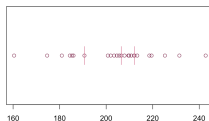
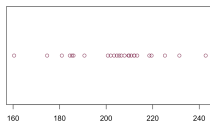
Skattning av kvartiler

Exempel (draghållfasthet)

$$\begin{cases} P_{23.1} = P_{100 \cdot 6 / 26} = 185.97 \\ P_{26.9} = P_{100 \cdot 7 / 26} = 190.67 \end{cases} \Rightarrow Q_1 \approx \frac{185.97 + 190.67}{2} = 188.32$$

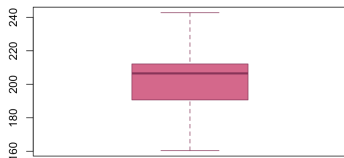
$$Q_2 = 206.51 \quad (\text{medianen})$$

På liknande sätt som vi hittade Q_1 kan man använda $P_{100 \cdot 19 / 26}$ och $P_{100 \cdot 20 / 26}$ för att beräkna $Q_3 \approx 212.75$.

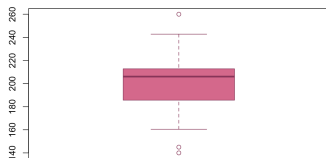


Boxplot

Boxplot



Boxplot över datan.



Boxplot över datan om vi lägger till uteliggarna 140, 145, 260.

En boxplot innehåller:

- En rektangel som går mellan Q_1 och Q_3 (innehåller 50% av observationerna).
- Ett stick i boxen vid Q_2 .
- Ett övre stick vid $\min(Q_2 + 1.5(Q_3 - Q_1), \max x_i)$.
- Ett undre stick vid $\max(Q_2 - 1.5(Q_3 - Q_1), \min x_i)$.
- Punkter vid varje datapunkt utanför intervallet $Q_2 \pm 1.5(Q_3 - Q_1)$

Spridningsmått

Stickprovsvariansen

Stickprovsvariansen s^2 av observationer $x_1, x_2, \dots, x_n$ ges av

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (x_i - \bar{x})^2 = \frac{n \sum x_i^2 - (\sum x_i)^2}{n(n-1)}$$

och standardavvikelsen s ges av

$$s = \sqrt{s^2}$$

Exempel (draghållfasthet)

I vårt exempel där medelvärdet var $\bar{x} = 203.93$ får vi

$$s^2 = \frac{1}{25-1} \left((160.44 - 203.93)^2 + \dots + (225.25 - 203.93)^2 \right) = 335.64$$

och

$$s = \sqrt{335.64} = 18.32.$$

Inferentiell statistik

Hittills har vi använt statistik för att beskriva data. I inferentiell statistik vill man istället använda statistik för att kunna dra slutsatser om populationen, ex. dess fördelning eller parametrar (ex. μ , σ , p , λ , etc.). Typiskt består inferentiell statistik av två delar:

1. Skattning av parametrar (gissa värdet på en parameter med hjälp av datan)
2. Beskrivning av osäkerhet (konfidsensintervall eller hypotestest)

Vilket mått när?

Medelvärde och stickprovsstandardavvikelsen i ett stickprov kan vara ganska känsligt för extrema observationer, medan medianen och kvartiler oftast är mer okänslig.

Punktskattning

Punktskattning

En punktskattning är ett tal $\hat{\theta}$ som man kan räkna ut givet data, vars syfte är att approximera en parameter θ .

Estimator

En estimator är en regel som anger hur man räknar ut en skattade från data. Till exempel är

$$\bar{X} = \frac{1}{n} \sum X_i$$

en estimator för medelvärdet μ .

Obs! En estimator är en slumpvariabel, och beror på vilket stickprov vi får (slump). Vi kommer senare i kursen att vara intresserade av vilket fördelning en estimator har för att kunna räkna ut konfidensintervall och göra hypotestest.

Punktskattning

Vad är en bra skattare?

Vi vill typiskt att en bra skattare gissar rätt i snitt, och att den oftast gissar någonting som är ganska bra.

- $\mathbb{E}[\hat{\theta}] = \theta$ (väntevärdesriktig) Om det här gäller så kommer skattaren att ligga närmre medelvärdet ju större stickprov vi har enligt de stora talens lag.
- $\text{Var}(\hat{\theta})$ liten, och helst $\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0$ (konsistent). $\sqrt{\text{Var}(\hat{\theta})}$ kallas för skattarens standardfel.

Obs! Det finns inte alltid en väntevärdesriktig skattare med liten varians, och ibland är det bättre att ha en skattare som inte är väntevärdesriktig men har liten varians.

Punktskattning

Exempel

Antag att $X_1, X_2, \dots, X_n$ är oberoende med $\mathbb{E}[X_1] = \dots = \mathbb{E}[X_n] = \mu$
 $\text{Var}(X_1) = \dots = \text{Var}(X_n) = \sigma^2$. Då är

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n} \sum X_i\right] = \frac{1}{n} \sum \mathbb{E}[X_i] = \frac{1}{n} \sum \mu = \frac{n\mu}{n} = \mu$$

och

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum \text{Var}(X_i) = \frac{1}{n^2} \sum \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Slutsats: $\bar{X}$ är en väntevärdesriktig och konsistent skattare av μ med medelfel $\sigma/\sqrt{n}$.

Punktskattning

Exempel

Man kan visa att

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

$\hat{\sigma}^2$ är alltså en väntevärdesriktig skattare av σ^2 . (Det var därför vi delade med $(n - 1)$ istället för n .) Man kan även visa att den är konsistent, och dess varians beror bland annat på det fjärde momentet av X .

Fördelning för en summa av normalfördelningar

Sats

Om $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$, $\dots$, $X_n \sim N(\mu_n, \sigma_n)$ är oberoende, så är

$$X_1 + \dots + X_n \sim N(\mu_1 + \dots + \mu_n, \sqrt{\sigma_1^2 + \dots + \sigma_n^2})$$

Följsats

Om $\mu_1 = \dots = \mu_n = \mu$ och $\sigma_1 = \dots = \sigma_n = \sigma$ så blir

$$X_1 + \dots + X_n \sim N(\mu n, \sigma \sqrt{n})$$

- Obs! Vi kan med tidigare formler kontrollera att medelvärdet och standardavvikelsen blir som ovan, men det är inte alls uppenbart att summan skulle bli normalfördelad, och det är inte sant för de flesta stokastiska variabler.
- Obs! Om inte oberoende så skulle vi ex. kunna ha $X_1 = \dots = X_n$, och då blir $X_1 + \dots + X_n \sim nX_1 \sim N(\mu n, \sigma n)$.

Punktskattning

Bevis

$$\begin{aligned}\mathbb{E}[e^{tX}] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{tx} e^{-(x-\mu)^2/(2\sigma^2)} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{\frac{-(x-(\mu+t\sigma^2))^2}{2\sigma^2} + \frac{2t\mu\sigma^2+t^2\sigma^4}{2\sigma^2}} dx \\ &= e^{t\mu+t^2\sigma^2/2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{\frac{-(x-(\mu+t\sigma^2))^2}{2\sigma^2}} dx = e^{t\mu+t^2\sigma^2/2}\end{aligned}$$

Därför är

$$\begin{aligned}\mathbb{E}[e^{t(X_1+X_2)}] &= \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}] = e^{t\mu_1+t^2\sigma_1^2/2} e^{t\mu_2+t^2\sigma_2^2/2} \\ &= e^{t(\mu_1+\mu_2)+t^2(\sigma_1^2+\sigma_2^2)/2}\end{aligned}$$

vilket är den momentgenererande funktionen för en stokastisk variabel med fördelning $N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Punkskattning

Fördelning av $\bar{X}$

I ett stickprov försöker vi se till att alla punkter är oberoende, och då gäller

$$\sum X_i \sim N(\mu n, \sigma \sqrt{n}) \Rightarrow \bar{X} = \frac{1}{n} \sum X_i \sim N(\mu, \sigma / \sqrt{n}).$$