

Matematisk Statistik och Diskret Matematik, MVE051/MSG810, VT22

Timo Vilkas



Chalmers - Göteborgs Universitet

Föreläsning 7
21 april 2022



Beskrivande Statistik
Inferentiell Statistik

Hittills har vi undersökt stokastiska variabler med en given fördelning. I praktiken är den exakta fördelningen ofta okänd. Istället undersöker man data, antar en viss form av fördelning och skattar motsvarande parametrar.

I statistik delen av kursen vill vi titta närmare på:

- 1 **Beskrivande Statistik:** organisera, karakterisera och sammanfatta data
- 2 **Inferentiell Statistik:** dra slutsatser (statistisk inferens) om en population genom att undersöka ett representativt urval (stickprov)



T. Vilkas

Matematisk statistik & diskret matematik, MVE051/MSG810

Grundläggande Definitioner

- **Population:** Samling av alla enheter som är intressanta för undersökningen.
- **Slumpmässigt urval** (eng. random sample): *Slumpmässigt valt* stickprov ur populationen, som undersöks för att samla information. Varje enhet i urvalet är en slumpvariabel med samma fördelning som hela populationen.
- **Data:** Informationen som samlas in i undersökningen (vanligtvis siffror, ibland kategorier)
- **Variabel:** Storleken som undersöks.

Grundläggande Definitioner

Vi skiljer på två olika typer av variabler:

- 1 **Kvantitativa** Variabler: värden ges som tal.
(t.ex.: vikt, längd, ålder av patienter osv.)
- 2 **Kvalitativa** Variabler: värden ges som kategorier.
(t.ex.: ögonfärg, blodgrupp, etc.)

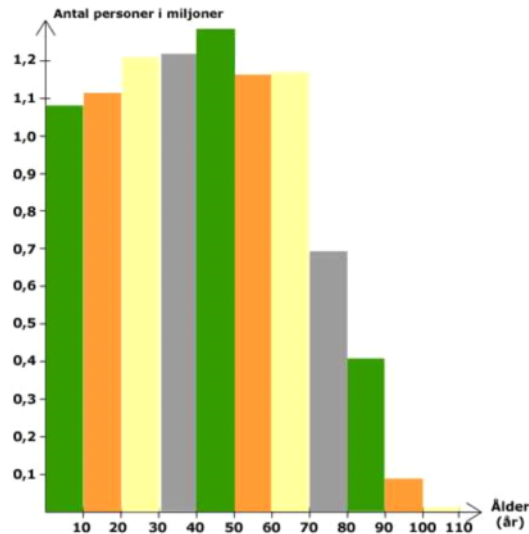
Vi skiljer på två olika typer av statistiska mått:

- 1 mått på centraltendens/**lägesmått** (medelvärde, median, kvantiler).
- 2 mått på **variation**/utspridning (varians, standardavvikelse).

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻

Som data har vi ett numeriskt värde för varje enhet i populationen: $x_1, x_2, \dots, x_N$. Dessa kan illustreras med hjälp av **relativa frekvenser**:



T. Vilkas

Matematisk statistik & diskret matematik, MVE051/MSG810

Fördelningens parametrar: medelvärde, median

- **Populationsmedelvärdet** ges av

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \mathbb{E}X \quad \text{för } X \text{ som är likformigt fördelad på } \{x_1, \dots, x_N\}.$$

- **Medianen** $\tilde{x}$ av värdemängden $\{x_1, x_2, \dots, x_N\}$ definieras som mellersta värde i mängden, närmare bestämt $\tilde{x} = x_m$ ifall $N = 2m - 1$ är udda och $\tilde{x} = \frac{x_m + x_{m+1}}{2}$ ifall $N = 2m$ är jämn.

T. Vilkas

Matematisk statistik & diskret matematik, MVE051/MSG810

Fördelningens parametrar: varians, variationsbredd

- Variansen av hela populationen ges av

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \mathbb{E}[(X - \mu)^2] = \text{var}(X) \text{ för } X, \mu \text{ som ovan.}$$

- **Standardavvikelsen** σ är som vanligt roten ur variansen.
- **Variationsbredd** (eng. *range*) är skillnaden mellan största och minsta värde, $x_L := \max\{x_1, \dots, x_N\}$ resp. $x_S := \min\{x_1, \dots, x_N\}$, $R = x_L - x_S$ (som blir $x_N - x_1$ ifall värden ges i stigande ordning).

OBS: Alla dessa parametrar är inte slump utan tal som framgår direkt ur datan och beskriver värdefördelningen.

T. Vilkas

Matematisk statistik & diskret matematik, MVE051/MSG810

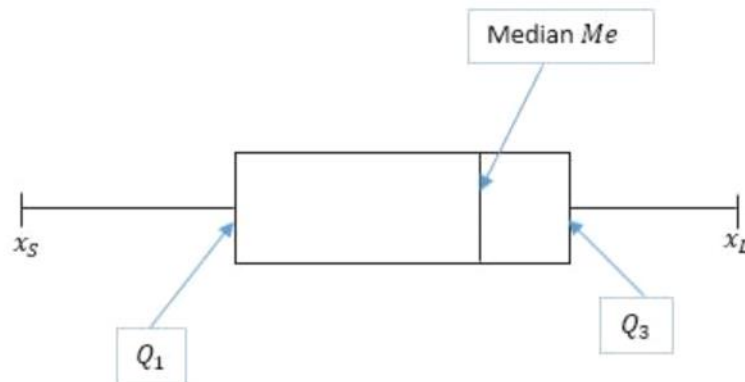
Percentiler/Kvartiler

- En **percentil** P anger det värde på en stokastisk variabel nedanför vilket en viss procent andel av observationerna av variabeln hamnar. För $p \in [0, 100]$ är p -percentilen P_p alltså värdet så att $p\%$ eller färre av observationerna är mindre än P_p och $(100 - p)\%$ eller färre är större än P_p .
- De percentiler som delar in datan i fyra delar kallas kvartiler (eng. *quartiles*): P_{25} eller Q_1 (**första kvartil**), P_{50} eller Q_2 (**andra kvartil** som motsvarar medianen) och P_{75} eller Q_3 (**tredje kvartil**).

T. Vilkas

Matematisk statistik & diskret matematik, MVE051/MSG810

Boxplot (eller Box-and-Whisker plot)



Lägesmått (eng. Location statistics): medelvärde, median

Låt $X_1, \dots, X_n$ vara ett slumpmässigt urval av utfall med samma fördelning som X .

- **Stickprovsmedelvärdet** (eng. *sample mean*) ges av:

$$\bar{X} := \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + \dots + X_n}{n}.$$

Detta är lätt att räkna fram dock mycket känsligt för extrema utfall (för lagom stor n).

- Precis som för hela populationen kan man bestämma medianen av ett stickprov (som också blir en slumpvariabel då). Den är mindre känslig för extrema utfall.

Spridningsmått (eng. measures of variability)

- **Stickprovsvariansen** (eng. sample variance) av $X_1, \dots, X_n$ defineras som

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

$S = \sqrt{S^2}$ kallas stickprovets standardavvikelse.

(Observera notationen: S för stickprovet, σ för populationen.)

- En enkel beräkning visar att stickprovsvariansen också kan skrivas som

$$S^2 = \frac{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}{n(n-1)}.$$

- Stickprovets variationsbredd (eng. sample range) defineras precis som ovan, fast nu för mängden av observationer.

Navigation icons

T. Vilkas

Matematisk statistik & diskret matematik, MVE051/MSG810

Exempel

Låt

18, 1, 20, 15, 12, 15, 14, 7, 11, 9, 6, 4

vara observationerna som utgör vårt stickprov. I stigande ordning har vi då värdemängden

1, 4, 6, 7, 9, 11, 12, 14, 15, 15, 18, 20

Medelvärde:

$$\bar{x} = \frac{20 + 18 + 15 + 15 + 14 + 12 + 11 + 9 + 7 + 6 + 4 + 1}{12} = 11$$

Median: Då $n = 12$ är jämn, ges medianen som medelvärdet av värde nummer 6 och 7: $Me = \frac{11+12}{2} = 11.5$.

Navigation icons

T. Vilkas

Matematisk statistik & diskret matematik, MVE051/MSG810

Exempel

Variationsbredd: $x_L = 20$, $x_S = 1$, alltså $R = 20 - 1 = 19$.

Varians och standardavvikelse:

$$s^2 = \frac{(20 - 11)^2 + (18 - 11)^2 + \dots + (-7)^2 + (-10)^2}{11} \approx 33.3$$

och

$$s = \sqrt{s^2} = 5.77$$

Kvartiler:

$$Q_1 = 6.5 \quad \text{och} \quad Q_3 = 15$$

Inferentiell Statistik

- Det finns två vanliga typer av **inferentiella statistika**: **skattning** och **hypotesprövning**.
- Det finns två vanliga typer av skattning: **punktskattning** och **intervallskattning** (konfidsensintervall: osäkerhetsintervall kring en punktskattning).
- En **parameter** är en statistisk storhet, som beskriver fördelningen i populationen (t.ex. μ , σ , etc.)
- I en skattning försöker man uppskatta/approximera (minst) en parameter θ med hjälp av ett representativt stickprov.

Punktskattning

- En **punktskattning** ger ett numeriskt värde som approximerar en populationsparameter θ .
- En regel för att beräkna en skattning av en given parameter baserad på ett stickprov av observerade data kallas en **estimator**. Estimatorer ges vanligtvis genom en formel, t.ex. är

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

en estimator för populationsmedelvärdet μ .

- Observera att en estimator är en *slumpvariabel*. Ett enskilt numeriskt värde, som fås genom att beräkna formeln för ett givet stickprov (utfall) kallas en (upp)skattning av parametern.
- En estimator av parametern θ betecknas med $\hat{\theta}$.

Väntevärdesriktig

- Det finns olika önskvärda egenskaper hos en estimator. En viktig sådan är att den är väntesvärderiktig (*eng. unbiased*).
- En estimator $\hat{\theta}$ kallas **väntevärdesriktig** om och endast om $\mathbb{E}[\hat{\theta}] = \theta$, där $\mathbb{E}[\hat{\theta}]$ är väntevärdet av $\hat{\theta}$ (som motsvarar genomsnittet av skattningarna av θ som framgår ur alla möjliga val för stickprovet av en given storlek ur populationen). "estimatorn (som är en slumpvariabel!) varierar kring rätt värde"
- I ord betyder det att genomsnittet av ett större antal skattningar kommer ligga nära parametern så länge vi skattar med en väntesvärderiktig estimator. (Stora talens lag)

Stickprovsmedelvärdet är väntevärdesriktigt

Låt $X_1, \dots, X_n$ vara ett stickprov av storlek n ur en fördelning med väntevärde μ , och låt $\bar{X}$ vara stickprovsmedelvärdet.

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}(X_1 + \dots + X_n)\right] \\ &\stackrel{\text{linj.}}{=} \frac{1}{n} [\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)] \\ &\stackrel{\mathbb{E}X_i = \mu}{=} \frac{1}{n} [n \cdot \mu] \\ &= \mu.\end{aligned}$$

$\bar{X}$ är alltså en väntesvärderikt estimator av parametern μ .

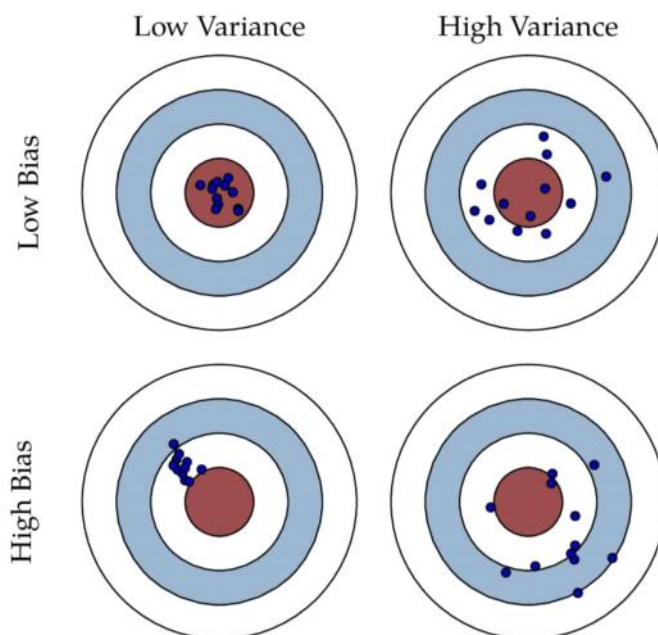
Konsistens

- En annan önskvärd egenskap är att estimatoren $\hat{\theta}$ har en liten varians när stickprovets storlek är tillräcklig (om den går mot 0 när $n \rightarrow \infty$ kallas estimatoren **konsistent**).
- Låt $\bar{X}$ som förr vara medelvärdet av ett stickprov av n stycken (oberoende) observationer ur en fördelning med väntevärde μ och varians σ^2 . Då gäller $\text{Var}[\bar{X}] = \frac{\sigma^2}{n}$. (räkna själv!)
- Följaktligen går $\text{Var}[\bar{X}]$ mot noll när man ökar stickprovets storlek tillräckligt mycket och $\bar{X}$ ligger då nära sitt väntevärde μ . $\rightarrow \bar{X}$ är en konsistent estimator för μ .
- $\frac{\sigma}{\sqrt{n}}$, som är standardavvikelsen av $\bar{X}$, kallas också **medelvärdets standardfel** (eng. *standard error of the mean*).

Stickprovsvariansen är väntevärdesriktig

Låt $X_1, \dots, X_n$ ett stickprov av storlek n ur en fördelning med väntevärde μ och varians σ^2

- En enkel beräkning visar att stickprovsvariansen $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ är en väntevärdesriktig estimator för σ^2 .
- Standardavvikelsen S är däremot **inte** väntevärdesriktig som estimator för σ .
- Anledningen varför man delar med $n - 1$ istället för n i formeln för S^2 är just att göra estimatorn väntevärdesriktig.



Stickprov ur en normalfördelning

Sats

Låt $X_1, \dots, X_n$ vara oberoende, normalfördelade slumpvariabler med motsvarande väntevärde $\mu_1, \dots, \mu_n$ och varians $\sigma_1^2, \dots, \sigma_n^2$. Då är också summan $X_1 + \dots + X_n$ normalfördelad med väntevärde $\mu = \sum_{i=1}^n \mu_i$ och varians $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

- Bevis: t.ex. med entydighet av momentgenererande funktionen
- Vi visste redan att $X \sim \mathcal{N}(\mu, \sigma^2)$ medför $cX \sim \mathcal{N}(c\mu, c^2\sigma^2)$ för godtyckligt $c \in \mathbb{R}$, jfr. föreläsning 3. $\rightarrow nX \sim \mathcal{N}(n\mu, n^2\sigma^2)$, till skillnad från ovan, där $X_1, \dots, X_n$ är oberoende!
- Om alltså $X_1, \dots, X_n$ är ett stickprov av storlek n ur en normalfördelning med väntevärde μ och varians σ^2 , så är $\bar{X}$ också normalfördelat med väntevärde μ och varians $\frac{\sigma^2}{n}$.

Enligt satsen är $(X_1 + \dots + X_n) \sim \mathcal{N}(n\mu, n\sigma^2)$ och enligt förel. 3: $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n) \sim \mathcal{N}\left(\frac{n\mu}{n}, \frac{n\sigma^2}{n^2}\right)$.