

# Statistik och Dataanalys I

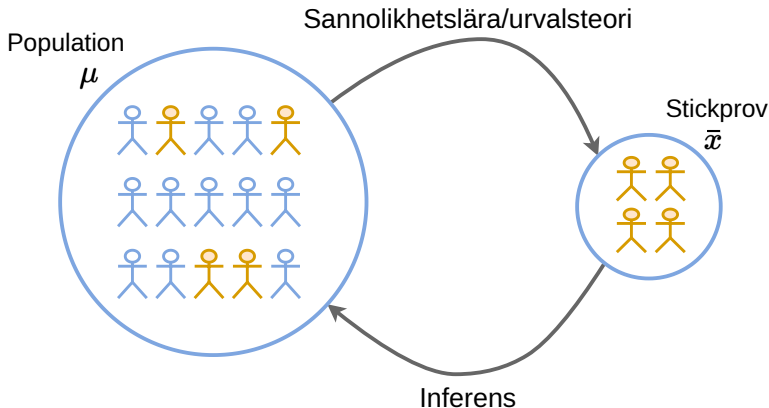
## Föreläsning 16 - Samplingfördelning och konfidensintervall för en andel

**Oskar Gustafsson**

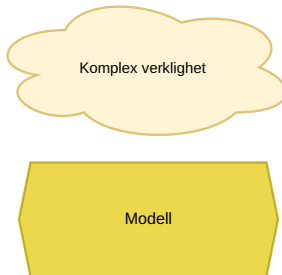
Statistiska institutionen  
Stockholms universitet

- **Sannolikhetsmodeller** och verkligheten.
- **Samplingfördelningen**.
- **Samplingfördelningen** för en **andel**.

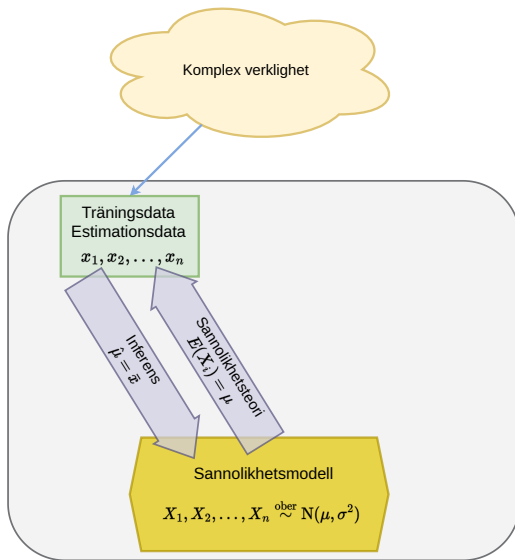
# Population och stickprov - ändliga populationer



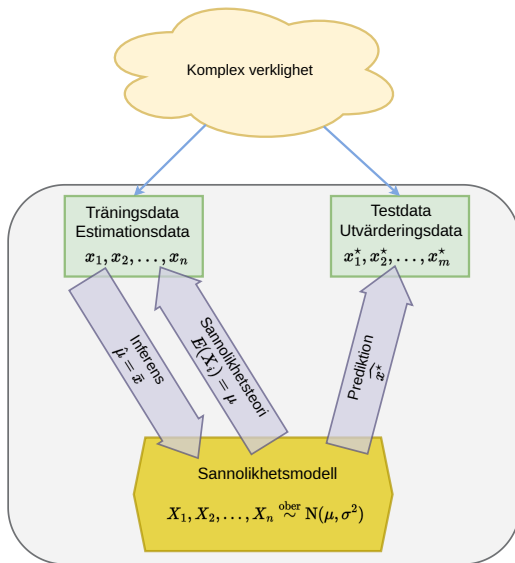
# Modeller som en förenkling av verkligheten



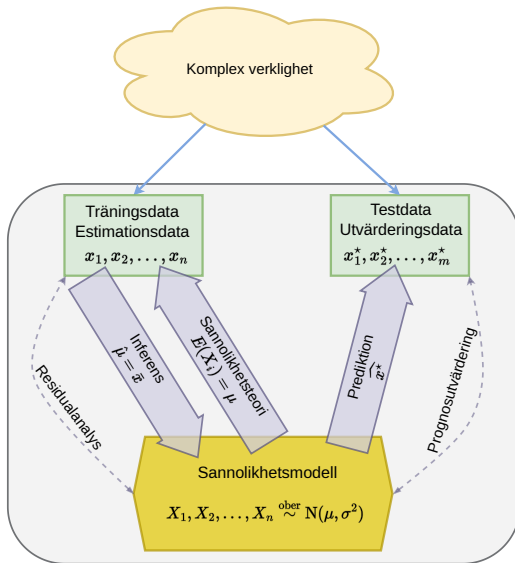
# Sannolikhetsmodeller och inferens



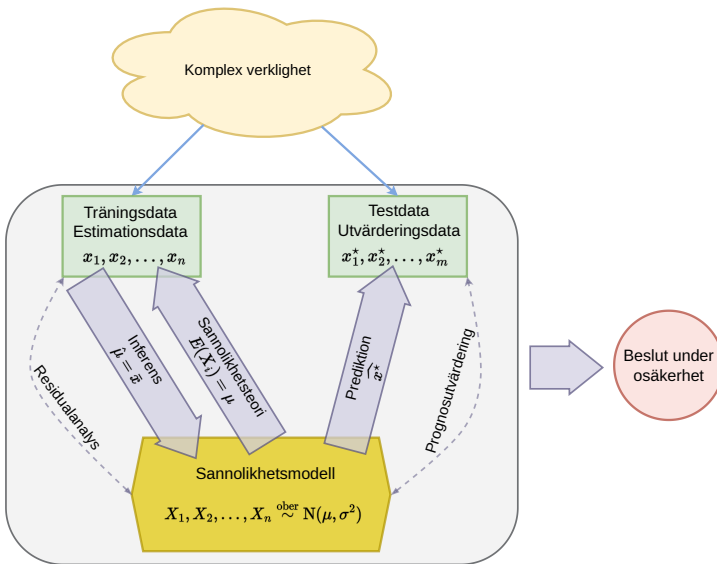
# Sannolikhetsmodeller möter verkligheten - prediktion



# Modellering är en iterativ process



# Slutmålet är ofta beslutsfattande i en osäker värld



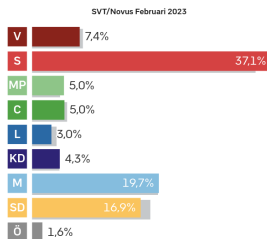


# Statistika, estimator och estimat

- **Statistika** - kvantitet som beräknas från ett stickprov.
- Andelen som röstar på socialdemokraterna i en valundersökning.
- Medelvärde  $\bar{x}$  av inkomster för personer i ett stickprov.
- Använder en statistika för **skatta en populationsparameter**.
- **Populationsväntevärdet**  $\mu$  skattas med **estimatorn**  $\bar{X}$ .
- För ett givet stickprov  $x_1, x_2, \dots, x_n$  får vi ett **estimat**  $\bar{x}$  av  $\mu$ .

# Väljarundersökningar

- Vilket parti skulle du rösta på om det var val idag?
- SVT/Novus. **Stickprov** med  $n = 3539$  personer.



- Kontaktade via telefon eller sms. Representativt? Bortfall?
- **Populationsparameter**: andelen S-röstare i populationen  $p$ .
- **Population**: röstberättigade i Sverige.
- **Estimator** för att skatta  $p$ : andelen S-röstare i stickprovet.
- **Estimat** i SVT/Novus undersökning:

$$\hat{p} = \frac{1313}{3539} \approx 0.371 \text{ dvs } 37.1\%$$

# Samplingfördelningen

- Men  $\hat{p} = 0.371$  är bara ett osäkert estimat från **ett** slumpmässigt valt stickprov.
- Om vi hade frågat 3539 **andra personer** hade vi säkerligen fått ett annat estimat.
- **Samplingfördelningen:**
  - ▶ fördelningen för en **estimator**
  - ▶ **över alla möjliga stickprov** av storleken  $n$ .
- **Statistiskt säkerställd** förändring från månaden innan?
- **Konfidsensintervall** för  $p$ : med 95% **säkerhet/konfidens** täcker intervallet  $(0.355, 0.387)$  den sanna andelen  $p$ .

# Samplingfördelningen - ändlig population

## Samplingfördelning för en andel - Ändlig population

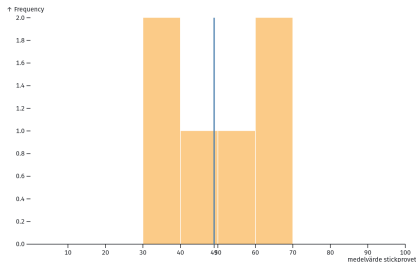
Pop.storlek:  $N$

Urvalstorlek:  $n$

Population: [24,79,41,52] med medelvärde 49.000

Det finns totalt  $\binom{4}{2} = 6$  möjliga stickprov av storlek  $n=2$ .

| Stickprov | Urval | Medelvärde |
|-----------|-------|------------|
| 1         | 24,79 | 51.5       |
| 2         | 24,41 | 32.5       |
| 3         | 24,52 | 38         |
| 4         | 79,41 | 60         |
| 5         | 79,52 | 65.5       |
| 6         | 41,52 | 46.5       |



# Samplingfördelningen för en andel

- **En andel är egentligen ett medelvärde** av binära variabler

$$X_i = \begin{cases} 1 & \text{om person } i \text{ röstar på } S \\ 0 & \text{om person } i \text{ inte röstar på } S \end{cases}$$

- **Populationsparameter** med  $N$  personer i populationen:

$$p = \frac{\sum_{i=1}^N x_i}{N}$$

- **Modell:**  $X_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$ .

- $\stackrel{\text{iid}}{\sim}$  betyder 'independent and identically distributed'.

**Oberoende och likafördelade.**

- **Sampling utan återläggning**  $\implies$  egentligen inte oberoende med samma sannolikhet  $p$ .
- Oberoende Bernoulli ändå OK modell, **om stickprovet är max 10% av populationen**. Korrektur ändlig population.

# Samplingfördelningen för en andel

- Vi skattar  $p$  med andelen i stickprovet

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

- Vilken fördelning har  $\hat{p}$ ?
- Eftersom  $X_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$  så vet vi att:

$$Y = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$$

$$E(Y) = np \quad \text{och} \quad SD(Y) = \sqrt{npq}$$

- $\hat{p} = \frac{1}{n} Y$ , så [F13 - skalning:  $E(aX) = aE(X)$  där  $a = \frac{1}{n}$ ]

$$E(\hat{p}) = \frac{1}{n} E(Y) = \frac{1}{n} np = p$$

och [F13 - skalning:  $SD(aX) = |a|SD(X)$ ]

$$SD(\hat{p}) = \left| \frac{1}{n} \right| \sqrt{npq} = \frac{1}{n} \sqrt{npq} = \sqrt{\frac{pq}{n}}$$

# Väntevärdesriktig estimator och stora talens lag

- Vi skattar  $p$  med andelen i stickprovet

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

- Vi vet nu att

$$E(\hat{p}) = p \quad \text{och} \quad SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

- Estimatorn  $\hat{p}$  är **väntevärdesriktig** för populationsandelen  $p$

$$E(\hat{p}) = p$$

- **Väntevärdesriktig = korrekt i genomsnitt**, sett över alla möjliga stickprov.

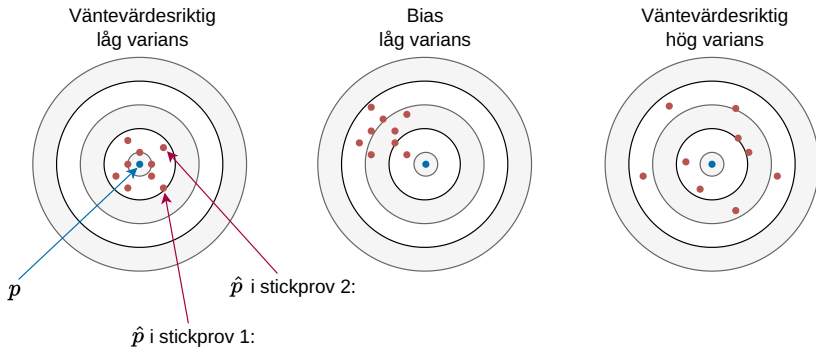
- **Bias**

$$\text{Bias}(\hat{p}) = E(\hat{p}) - p = 0$$

- $SD(\hat{p})$  minskar när stickprovstorleken  $n$  ökar.

- **Stora talens lag**:  $\hat{p}$  kommer vara nära  $p$  i stora stickprov.

# Väntevärdesriktighet, bias och varians





# Samplingfördelningen för $\hat{p}$ - normalapproximation

- Väntevärde och standardavvikelse för estimatoren  $\hat{p}$

$$E(\hat{p}) = p \quad \text{och} \quad SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

- **Normalapproximation**

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

- När är normalapproximationen tillräckligt bra?
  - ▶ stickprovsstorleken  $n \geq 30$  (centrala gränsvärdessatsen)
  - ▶  $np \geq 10$  och  $nq \geq 10$ .
  - ▶ **oberoendeantagandet** måste vara (hyfsat) uppfyllt.
  - ▶ stickprovet är **högst 10% av populationen**.

# Stora talens lag - andel

## Stora talens lag - normalapproximation av andel

Normalapproximation av samplingfördelningen för en andel

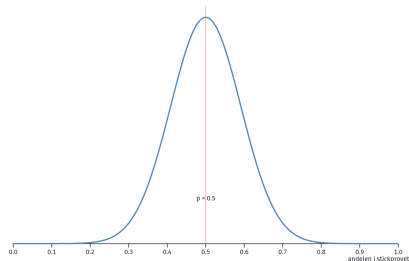
$n$   

$p$   

Stickprovet  $n = 30$  är större än 30. ✓

Antal förväntade lyckade  $np = 15.0$  är större än 10. ✓

Antal förväntade misslyckade  $nq = 15.0$  är större än 10. ✓



Normalapproximation av samplingfördelningen för en andel.

# Exempel - röstandel för S

- $\hat{p} = 0.371$ , men  $\hat{p}$  varierar från stickprov till stickprov.

- **Normalapproximation**

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

- **Standardavvikelse för estimator**

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

- Men vi vet inte  $p$ ! Lösning: sätt in skattning  $\hat{p}$  istället för  $p$ .

- **Standardfel för estimator**

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.371(1 - 0.371)}{3539}} \approx 0.0081$$

- Kan vi använda normalapproximation?

- ▶ stickprovsstorleken  $n = 3539 \geq 30$ . ✓
- ▶  $n\hat{p} = 3539 \cdot 0.371 = 1312.97 \geq 10$  ✓
- ▶  $n\hat{q} = 3539 \cdot (1 - 0.371) = 2226.03 \geq 10$  ✓
- ▶ **Oberoendeantagandet**. Sluppmässigt urval. ✓
- ▶ **Högst 10% av populationen**. Definitivt OK! ✓

# Normalapproximation för en andel

## ■ Normalapproximation

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

## ■ 68-95-99.7 regeln:

- ▶  $P(\hat{p} \text{ är högst en standardavvikelse från } p) = 0.683$
- ▶  $P(\hat{p} \text{ är högst två standardavvikelser från } p) = 0.954$
- ▶  $P(\hat{p} \text{ är högst tre standardavvikelser från } p) = 0.997$

## ■ Vanligt att vi vill ha “rundare” sannolikheter:

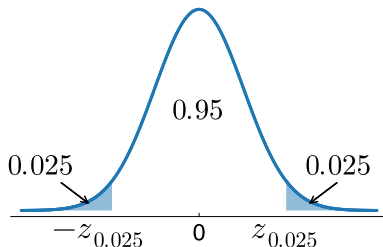
- ▶  $P(\hat{p} \text{ är högst 1.645 standardavvikelser från } p) = 0.90$
- ▶  $P(\hat{p} \text{ är högst 1.96 standardavvikelser från } p) = 0.95$
- ▶  $P(\hat{p} \text{ är högst 2.576 standardavvikelser från } p) = 0.99$

# Kritisk värde för 95%-igt konfidensintervall

## ■ Intervall med sannolikhet 0.95:

- ▶ Sannolikhetsmassa **utanför** intervallet:  $\alpha = 0.05$
- ▶ Sannolikhetsmassa **i vardera svans**:  $\alpha/2 = 0.025$
- ▶  $z_{\alpha/2} = z_{0.025} = 1.96$
- ▶ 2.5% av sannolikhetsmassan till *höger* om  $z_{0.025}$ .
- ▶ 97.5% av sannolikhetsmassan till *vänster* om  $z_{0.025}$ .
- ▶  $z_{0.025}$  från Z-tabell, det  $z$  där  $P(Z \leq z) = 1 - 0.025 = 0.975$

95%-igt intervall:  $z_{0.025} = 1.96$

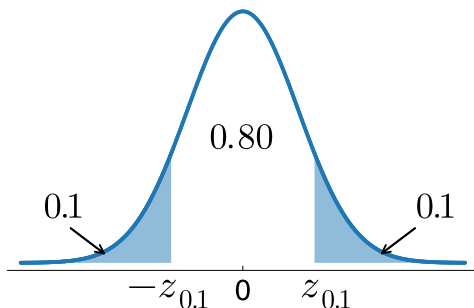


# Kritisk värde för 80%-igt konfidensintervall

## ■ Intervall med sannolikhet 0.8:

- ▶  $\alpha = 0.2$
- ▶  $z_{\alpha/2} = z_{0.1} = 1.282$
- ▶ “värdet som har 10% av sannolikhetsmassan till höger om sig i standard normalfördelning”

80%-igt intervall:  $z_{0.1} = 1.282$

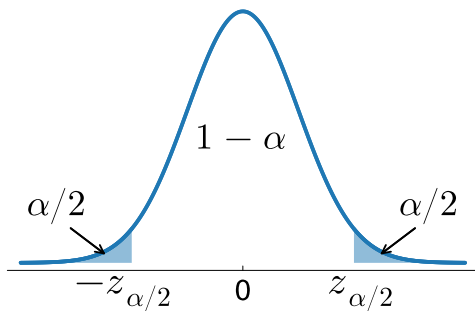


# Kritisk värde - konfidensintervall $1 - \alpha$ sannolikhet

■ **Kritiskt värde för ett intervall** med sannolikhet  $1 - \alpha$ :

- ▶ Sannolikhetsmassa **utanför** intervallet:  $\alpha$
- ▶ Sannolikhetsmassa **i vardera svans**:  $\alpha/2$
- ▶  $z_{\alpha/2}$  "har  $\alpha/2$  av sannolikhetsmassan till höger om sig i standard normalfördelning"

$1 - \alpha$  intervall



# 95%-igt konfidensintervall för en andel

## ■ Normalapproximation

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Approximativt 95%-igt konfidensintervall för andel  $p$

$$\hat{p} \pm z_{0.025} \cdot SE(\hat{p})$$

där

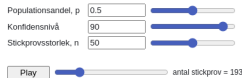
$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- Intervall från **givet stickprov** antingen täcker eller missar  $p$ .
- Ett 95%-igt konfidensintervall kommer innehålla populationsvärdet  $p$  **i 95% av alla möjliga stickprov**.
- “Den sanna andelen är i intervallet **med 95% säkerhet**”.

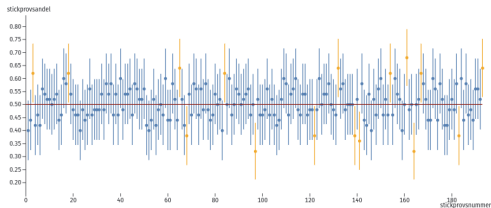


# Konfidsensintervall för en andel - interaktivt

## Konfidsensintervall för en andel - normalapproximation



Av totalt 193 stickprov innehöll 177 st (**91.710%**) av de 90%-iga konfidsensintervallen den sanna populationsandelen  $p = 0.5$ .



# Konfidsensintervall för en andel

## ■ Normalapproximation

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

**Approximativt  $(1-\alpha)\%$ -igt konfidsensintervall för andel  $p$**

$$\hat{p} \pm z_{\alpha/2} \cdot SE(\hat{p})$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

■ **Felmarginal** (eng. margin of error, ME):  $z_{\alpha/2} \cdot SE(\hat{p})$

■ Konfidsensintervall:

$$\text{Estimat} \pm \text{Felmarginal}$$

■ Trade-off: **högre konfidsensnivå  $\implies$  större felmarginal.**

■ SDM-boken:  $z^*$  istället för  $z_{\alpha/2}$ .

SDM-boken avrundar också ofta  $z_{0.025} = 1.96$  till  $z^* = 2$ .

## Exempel - röstandel för S

- **Estimat:**  $\hat{p} = 0.371$ .
- **Standardfel** för estimator

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.371(1 - 0.371)}{3539}} \approx 0.0081$$

- 95% konfidensintervall för andelen S-röstare i populationen:

$$\begin{aligned}\hat{p} \pm z_{0.025} \cdot SE(\hat{p}) \\ 0.371 \pm 1.96 \cdot 0.0081 \\ 0.371 \pm 0.015876\end{aligned}$$

vilket ger intervallet (0.355, 0.387).

- Intervallet (0.355, 0.387) innehåller andelen S-röstare,  $p$ , med 95% säkerhet. Men kom ihåg vad detta faktiskt betyder!
- Intervall som skapas med formeln  $\hat{p} \pm z_{0.025} \cdot SE(\hat{p})$  kommer innehålla  $p$  i 95% av alla möjliga stickprov från populationen.

# Konfidensintervall för andel i R

## ■ Socialdemokraternas väljarandel

```
> prop.test(x = 1313, n = 3539)
```

1-sample proportions test with continuity correction

```
data: 1313 out of 3539
X-squared = 235.02, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3551012 0.3871985
sample estimates:
      p
0.3710088
```

## ■ Överlevande Titanic

```
> library(sda123) # for titanic data with n = 886 passengers
> prop.test(~survived, data = titanic)
```

1-sample proportions test with continuity correction

```
data: titanic$survived [with success = 1]
X-squared = 46.002, df = 1, p-value = 1.181e-11
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3535443 0.4185986
sample estimates:
      p
0.3855693
```

Dessa slides skapades för kursen statistik och dataanalys 1 av Mattias Villani HT 2023, och har modifierats av Oscar Oelrich VT 2024, och Oskar Gustafsson för VT 2025.