

Statistik och Dataanalys I

Föreläsning 19 - Hypotestest

Mattias Villani



Statistiska institutionen
Stockholms universitet



mattiasvillani.com



[@matvil](https://twitter.com/matvil)



[@matvil](https://mastodon.social/@matvil)



[mattiasvillani](https://github.com/mattiasvillani)

- Hypotesttest för en andel
- Hypotesttest för ett väntevärde

Exempel: trasiga mobilskärmar

- Ett företag producerar skärmar till mobiltelefoner.
- 15% av skärmarna får pixeldefekter och måste kasseras.
- Ny teknik. Stickprov: $n = 160$ skärmar. 14 var defekta.
- Bör företaget köpa in den nya tekniken?
- **Modell** för nya tekniken: $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$.
- Skattning: $\hat{p} = \frac{14}{160} = 0.0875$
- Verkar bättre, men kan vara **slumpen i detta stickprov**.
- Hur sannolikt är det att få $\hat{p} = 0.0875$ om $p = 0.15$?

Konfidsensintervall för andelen trasiga skärmar

- **Samplingfördelning** (check: $n\hat{p} = 14 \geq 10$, $n\hat{q} = 146 \geq 10$)

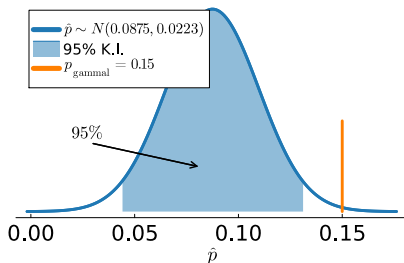
$$\hat{p} \stackrel{\text{approx}}{\sim} N(p, SD(\hat{p}))$$

- $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$ skattas med **standardfelet** $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.

- **95% konfidsensintervall för p**

$$\hat{p} \pm z_{0.025} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.0875 \pm 1.96 \cdot \sqrt{\frac{0.0875 \cdot (1 - 0.0875)}{160}} \approx (0.049, 0.139)$$

Skattad samplingfördelning för \hat{p}



Hypotestest för andelen trasiga skärmar

- Företaget vill fatta beslut: köpa ny teknik eller inte?
- **Nollhypotes** (H_0): ny teknik lika bra som gamla.
- **Alternativhypotes** (H_A): ny teknik **inte** lika bra som gamla.

$$H_0 : p = 0.15$$

$$H_0 : p = p_0$$

$$H_A : p \neq 0.15$$

$$H_A : p \neq p_0$$

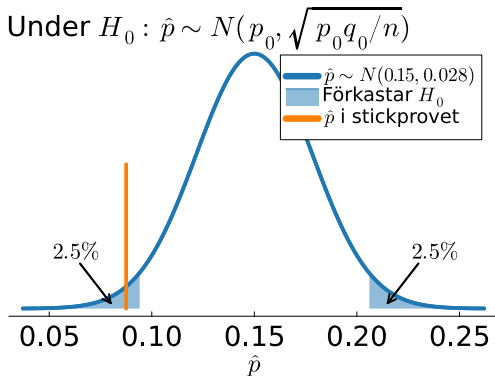
- Hur sannolikt är $\hat{p} = 0.0875$ i stickprov om $p = 0.15$?
- Samplingfördelning om H_0 är sann

$$\hat{p} \stackrel{\text{approx}}{\sim} N\left(p_0, \sqrt{\frac{p_0 q_0}{n}}\right)$$

- **Antag att nollhypotesen är sann**, dvs $p = 0.15$

$$\text{Under } H_0 : \hat{p} \stackrel{\text{approx}}{\sim} N\left(0.15, \sqrt{\frac{0.15 \cdot 0.85}{160}}\right) = N(0.15, 0.028)$$

Hypotestest för andelen trasiga skärmar



- Ett stickprov med $\hat{p} = 0.0875$ är osannolikt **om H_0 är sann** ($p = 0.15$). Vi tror därför inte på H_0 .

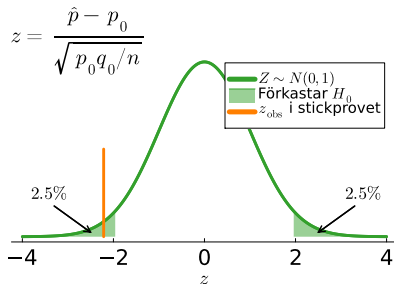
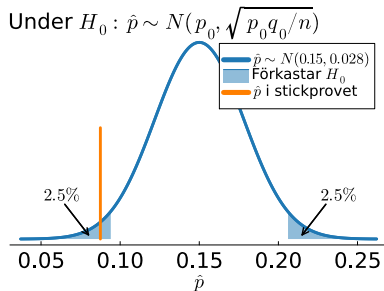
Hypotestest för andelen trasiga skärmar

■ Samplingfördelning under H_0

$$\hat{p} \stackrel{\text{approx}}{\sim} N\left(p_0, \sqrt{\frac{p_0 q_0}{n}}\right)$$

■ Standardiserad samplingfördelning under H_0

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim N(0, 1)$$



Hypotestest för andelen trasiga skärmar

■ Teststatistika under H_0

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim N(0, 1)$$

■ Observerad teststatistika i stickprovet

$$z_{\text{obs}} = \frac{0.0875 - 0.15}{\sqrt{\frac{0.15 \cdot 0.85}{160}}} = -2.214$$

■ Kritiskt värde från $N(0, 1)$

$$z_{\text{crit}} = z_{0.025} = 1.96$$

$$\text{■ } |z_{\text{obs}}| = 2.214 \geq z_{\text{crit}} = 1.96$$

\implies **förkastar** H_0 på 5% signifikansnivå.

- Vi använder absolutbeloppet $|z_{\text{obs}}|$ eftersom vi förkastar i båda svansarna. **Dubbelsidigt test.**

Hypotestest för andelen trasiga skärmar

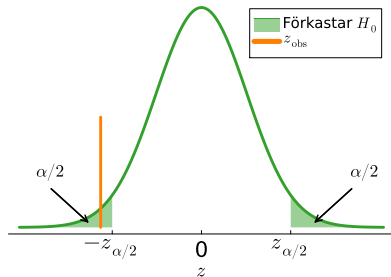
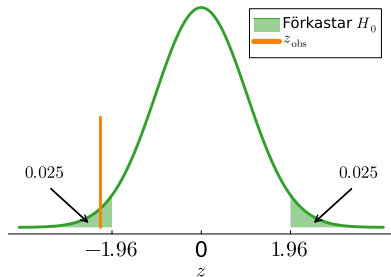
- Observerad teststatistika i stickprovet

$$z_{\text{obs}} = -2.214$$

- Kritiskt värde från $N(0, 1)$

$$z_{\text{crit}} = z_{0.025} = 1.96$$

- $|z_{\text{obs}}| \geq z_{\text{crit}} \Rightarrow$ **förkastar** H_0 på 5% signifikansnivå.



Alternativ approach: p -värde

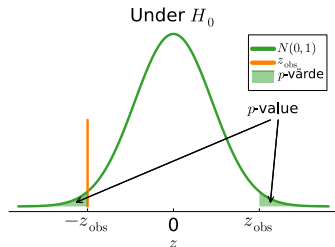
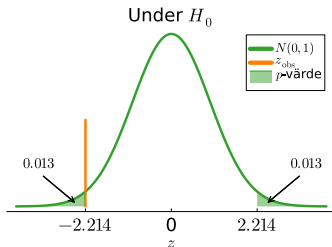
- p -värde: sannolikhet observera z_{obs} (eller värre) om H_0 sann:

$$p\text{-värde} = P(|Z| \geq |z_{\text{obs}}| \mid H_0 \text{ är sann})$$

- $p\text{-värde} < 0.05 \implies$ vi förkastar H_0 på 5% signifikansnivå.
- $p\text{-värde} \geq 0.05 \implies$ vi förkastar inte H_0 på 5% signifikansnivå.
- Från Z-tabell (eller `pnorm(-2.214)`)

$$P(Z \leq z_{\text{obs}}) = P(Z \leq -2.214) \approx 0.013$$

- p -värdet är $2 \cdot 0.013 = 0.026$.



K.I. för ett väntevärde - internethastighet

- Min internethastighet (i Mbit/sekund) under fem dagar:

15.77, 20.5, 8.26, 14.37, 21.09

- Mitt bredbandsbolag: du får 20 Mbit/sekund i genomsnitt.
- Jag: hold my beer ...
- **Modell:** $X_1, X_2, \dots, X_5 \sim N(\mu, \sigma)$ [bortse från negativa]
- Antag: enligt Bredbandskollen är $\sigma = 5$.
- **95% konfidensintervall**

$$\bar{x} \pm z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}$$

$$15.998 \pm 1.96 \cdot \frac{5}{\sqrt{5}}$$

$$(11.615, 20.381)$$

Hypotestest för ett väntevärde - känd varians

■ Hypoteser

$$H_0 : \mu = 20$$

$$H_A : \mu \neq 20$$

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0$$

■ Teststatistiska

$$Z = \frac{\bar{X} - \mu_0}{SD(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

■ Om H_0 sann: $Z \sim N(0, 1)$.

■ Internethastighet

$$z_{\text{obs}} = \frac{15.998 - 20}{\frac{5}{\sqrt{5}}} \approx -1.790$$

■ p-värde

$$2 \cdot P(Z \leq -1.790) \approx 2 \cdot 0.037 = 0.074$$

■ p-värde $> 0.05 \Rightarrow$ kan inte förkasta nollhypotesen på 5% signifikansnivå.

K.I. för ett väntevärde - skattad varians

- Antag nu att σ inte är känd och skattas med

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

- Internetdata:

$$s = \sqrt{\frac{(15.77 - 15.998)^2 + \dots + (21.09 - 15.998)^2}{4}} = 5.2147$$

- 95% konfidensintervall**

$$\begin{aligned} \bar{x} \pm t_{0.025,4} \cdot \frac{s}{\sqrt{n}} \\ 15.998 \pm 2.776 \cdot \frac{5.2147}{\sqrt{5}} \\ (9.523, 22.472) \end{aligned}$$

- Bredare intervall när variansen måste skattas.

Hypotesttest för ett väntevärde - skattad varians

■ Teststatistiska

$$T = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Om H_0 sann: $T \sim t_{n-1}$, **student-t med $n - 1$ frihetsgrader**.
- Teststatistiska för internethastighet

$$t_{\text{obs}} = \frac{15.998 - 20}{\frac{5.2147}{\sqrt{5}}} \approx -1.716$$

- **p-värde** från t_4 -fördelningen

$$2 \cdot P(T \leq -1.716) \approx 2 \cdot 0.081 = 0.162$$

- p-värde större än 0.05 \Rightarrow kan inte förkasta nollhypotesen på 5% signifikansnivå.
- Det är nu ännu mer troligt att få $\bar{X} = 15.998$ även om H_0 är sann, dvs om $\mu = 20$.

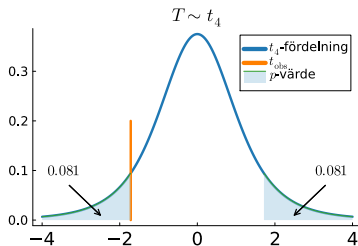
Hypotesttest för ett väntevärde - skattad varians

■ Teststatistiska

$$t_{\text{obs}} \approx -1.716$$

■ p-värde från t_4 -fördelningen

$$2 \cdot P(T \leq -1.716) \approx 2 \cdot 0.081 = 0.162$$



Ensidigt hypotesttest för ett väntevärde

- Egentligen vill jag nog göra ett **ensidigt test** med hypoteser

$$H_0 : \mu \geq 20$$

$$H_0 : \mu \geq \mu_0$$

$$H_A : \mu < 20$$

$$H_A : \mu < \mu_0$$

- Samma **teststatistiska**

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Om H_0 sann: $T \sim t_{n-1}$, **student-t med $n - 1$ frihetsgrader**.
- Teststatistiska

$$t_{\text{obs}} \approx -1.716$$

- **p-värde** från t_4 -fördelningen [inte gånger 2 pga ensidigt test]

$$P(T \leq -1.716) \approx 0.081$$

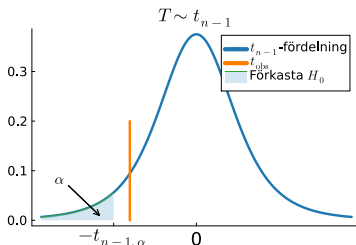
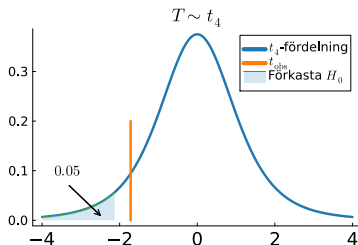
- Eftersom p-värde är större än 0.05 kan jag inte förkasta nollhypotesen på 5% signifikansnivå.

Ensidigt hypotesttest för ett väntevärde

- Beslut-variant med kritiskt värde (jfr tidigare $t_{0.025,4} = 2.776$)

$$t_{0.05,4} = 2.132$$

- Eftersom $t_{\text{obs}} = -1.716 > -2.132$ så förkastar vi inte H_0 på 5% signifikansnivå.



Hypotesttest - fatta principen bakom!

- **Hypotesttest andel.** Teststatistiska

$$Z = \frac{\hat{p} - p_0}{SE(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

- **Hypotesttest väntevärde.** Teststatistiska

$$T = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- **Allmänt**

$$\frac{\text{Estimatet} - \text{Parametern under } H_0}{\text{Standardfel Estimator under } H_0}$$

- Är estimatet \bar{x} tillräckligt långt från μ_0 , jämfört med den naturliga samplingvariation vi har för \bar{X} om H_0 är sann?
I så fall kommer data nog inte från H_0 . Förkasta H_0 !