# Statistik och Dataanalys I

## Föreläsning 21 - Hypotestest och jämföra grupper

**Oscar Oelrich**

Statistiska institutionen
Stockholms universitet

# Översikt

- Mera hypotestest: fel av typ I och II
- Jämföra två populationer - oberoende stickprov
- Jämföra två populationer - parade data

# Praktisk vs Statistisk signifikans

- **Teststatistiska**

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

- Stora stickprov: även små skillnader $\bar{X} - \mu_0$ blir **signifikanta**.
- Det betyder inte alltid att de är **praktiskt betydelsefulla**.
- Studie 1:
  - $\bar{x} = 1,\ \mu_0 = 0,\ s = 2,\ n = 10.$

$$t_{\text{obs}} = \frac{1 - 0}{\frac{2}{\sqrt{10}}} = 1.58$$

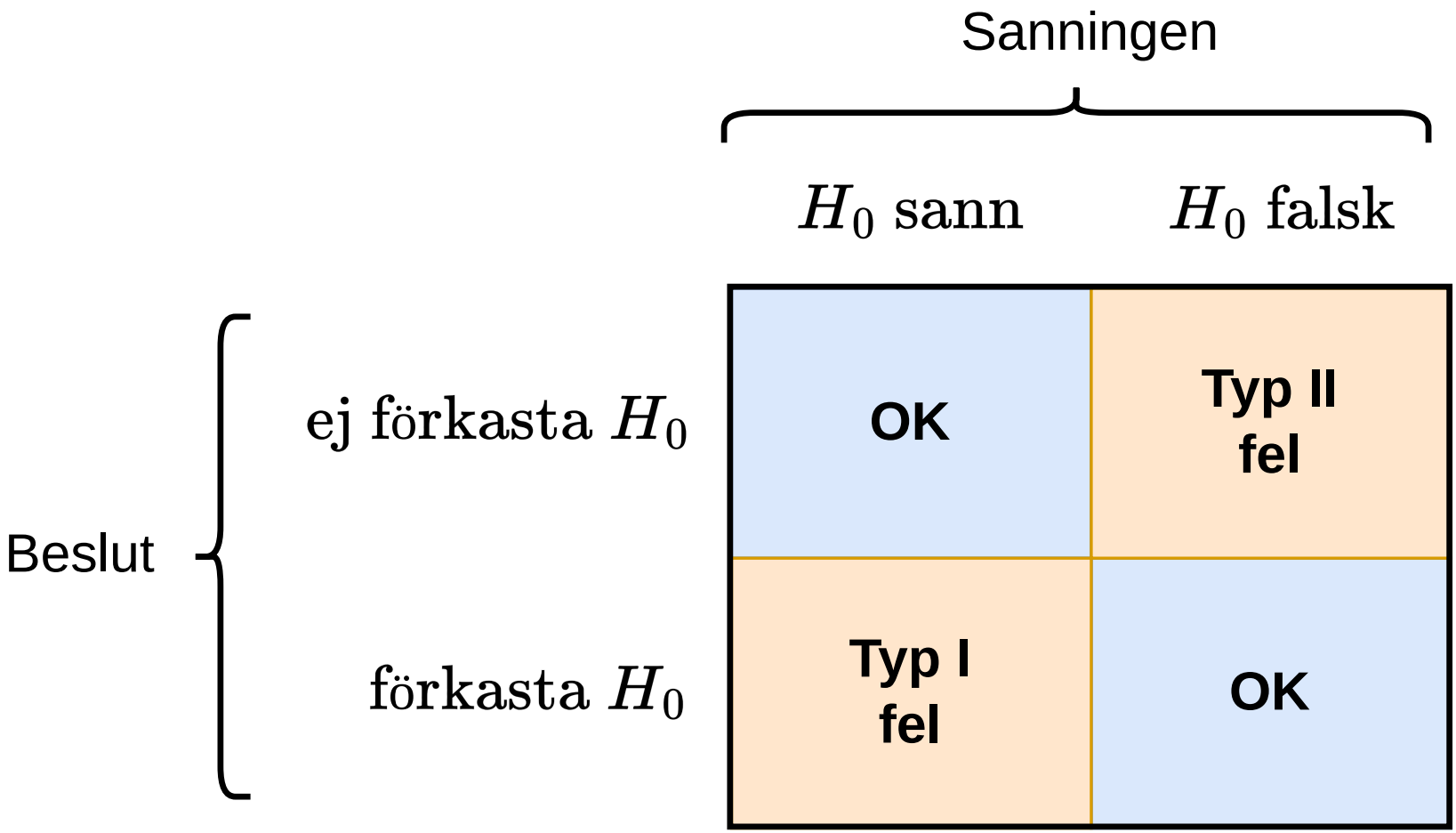
- Studie 2:
  - $\bar{x} = 0.05,\ \mu_0 = 0,\ s = 2,\ n = 10000.$

$$t_{\text{obs}} = \frac{0.05 - 0}{\frac{2}{\sqrt{10000}}} = 2.5$$

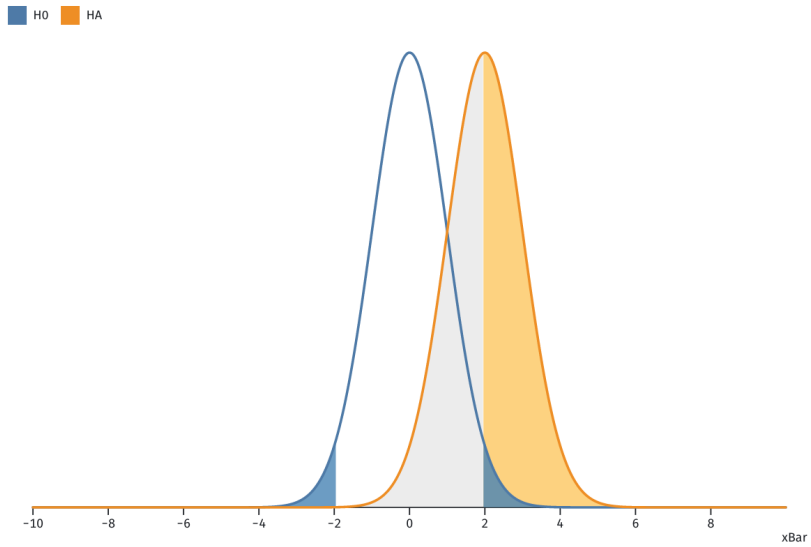
# Fel av typ I och II

- **Fel av typ I**:
  - $\alpha = P(\text{förkasta } H_0 \mid H_0 \text{ sann})$.
  - Bestäms av kritiska värdet.
- **Fel av typ II**:
  - $\beta = P(\text{inte förkasta } H_0 \mid H_A \text{ sann})$.
  - Beror på kritiska värdet och **värdet på $\mu$ under $H_A$**.
- **Testets styrka**
  - $1 - \beta = P(\text{förkasta } H_0 \mid H_A \text{ sann})$

# Fel av typ I och II

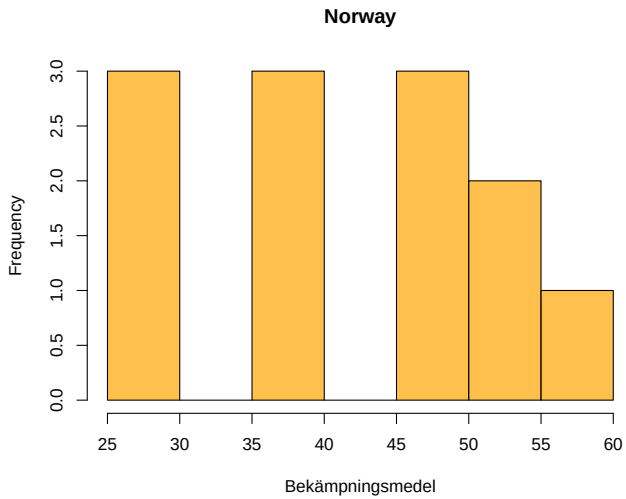


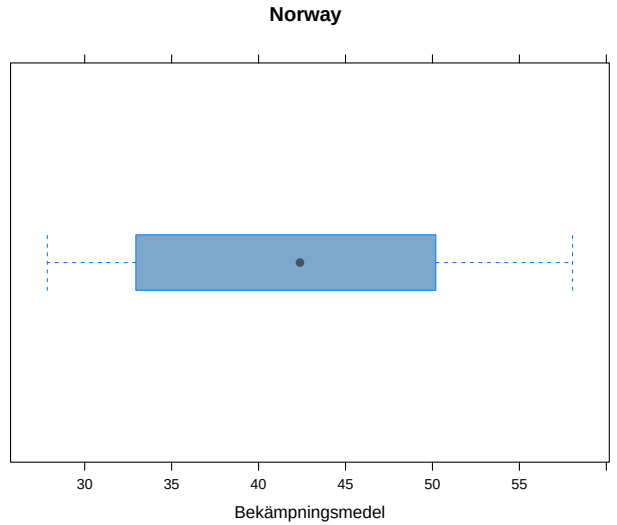
# Fel av typ I och II - interaktivt



# Jämföra grupper - bekämpningsmedel i lax

- `Total.pestocide` i $153$ laxar vid 8 olika platser.
- Grupp 1: Eastern Canada med $n = 24$ laxar. $N(\mu_1, \sigma_1)$.
  - $\bar{x}_1 = 33.572$
  - $s_1 = 7.671$
- Grupp 2: Norge med $n = 12$ laxar. $N(\mu_2, \sigma_2)$.
  - $\bar{x}_2 = 41.763$
  - $s_2 = 10.373$
- Är $\mu_1 = \mu_2$?



# Jämföra grupper - konfidensintervall för $\mu_1 - \mu_2$

- Grupp 1: $n_1$ observationer från populationen $N(\mu_1, \sigma_1)$.
  - Medelvärde $\bar{x}_1$ och standardavvikelse $s_1$.
- Grupp 2: $n_2$ observationer från populationen $N(\mu_2, \sigma_2)$.
  - Medelvärde $\bar{x}_2$ och standardavvikelse $s_2$.
- **Oberoende** observationer **inom och mellan grupperna**.

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- **Standardfel**

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- **Konfidensintervall för differensen** $\mu_1 - \mu_2$

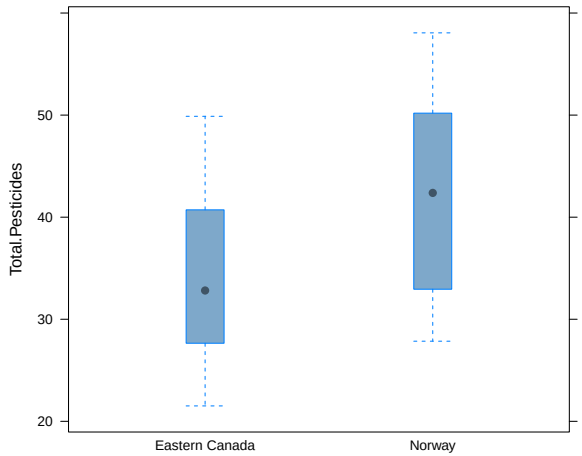
$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.025,\mathrm{df}} \cdot SE(\bar{x}_1 - \bar{x}_2)$$

- Frihetsgraderna $\mathrm{df}$ har en komplicerad formel.

# Antal frihetsgrader för differensen

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

# Jämföra grupper - bekämpningsmedel i lax




```
> # Using the t.test function to compute C.I. and do test H0: mu1 = mu2
> t.test(Total.Pesticides ~ Location, data = salmonTwoPop)

        Welch Two Sample t-test

data:  Total.Pesticides by Location
t = -2.424, df = 17.223, p-value = 0.02663
alternative hypothesis: true difference in means between group Eastern Canada and g
95 percent confidence interval:
 -15.314121  -1.068879
sample estimates:
mean in group Eastern Canada         mean in group Norway
                    33.57167                     41.76317
```

# Jämföra grupper - test för $\mu_1 = \mu_2$

- Hypotestest för skillnaden mellan två gruppers väntevärden:

$$H_0 : \mu_1 = \mu_2 \qquad H_0 : \mu_1 - \mu_2 = d_0$$
$$H_A : \mu_1 \neq \mu_2 \qquad H_A : \mu_1 - \mu_2 \neq d_0$$

- Recall: Allmän formel teststatistika

$$\frac{\text{Estimat} - \text{parameter under } H_0}{\text{Standardfel estimator under } H_0}$$

- **Teststatistika**

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\text{df}} \text{ under } H_0$$

- Förkastar $H_0$ om $|t_{obs}| > t_{crit}$.
- $p$-värde från $t_{\text{df}}$.

# Jämföra parade grupper

- **Parade data**. Ex **mäter samma $n$ personer** vid två tillfällen:
- Mätningar vid tidpunkt 1: $N(\mu_1, \sigma_1)$.
- Mätningar vid tidpunkt 2: $N(\mu_2, \sigma_2)$.
- **Beroende** stickprov, med lika många observationer.
- Skapa **differenser** av mätningarna: $d_i = x_{1i} - x_{2i}$.
- **Differenserna ska vara oberoende** mellan personer.
- **Standardfel**

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$$

- **95%-igt konfidensintervall**

$$\bar{d} \pm t_{0.025}, n-1 \cdot \frac{s_d}{\sqrt{n}}$$

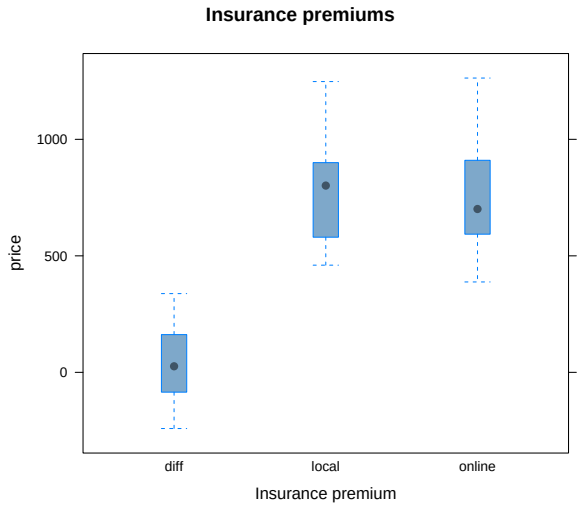
- **Hypotestest**

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

$$T = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} \sim t_{n-1}$$

# Local vs online insurance sales

| Person | Local | Online | PriceDiff |
|---|---|---|---|
| 1 | 550 | 388 | 162 |
| 2 | 856 | 593 | 263 |
| 3 | 460 | 497 | -37 |
| 4 | 1248 | 910 | 338 |
| 5 | 580 | 665 | -85 |
| 6 | 1022 | 1263 | -241 |
| 7 | 773 | 703 | 70 |
| 8 | 830 | 789 | 41 |
| 9 | 900 | 1001 | -101 |
| 10 | 710 | 699 | 11 |
| medelvärde | 792.900 | 750.800 | **42.100** |
| standardfel | 235.431 | 254.956 | **174.964** |




```
> t.test(price ~ place, data = df, paired = TRUE)

        Paired t-test

data:  price by place
t = 0.76091, df = 9, p-value = 0.4662
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -83.06154 167.26154
sample estimates:
mean difference
           42.1
```

# Credits

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