STK-IN1050 - Statistikk for informatikere

16. Hypotesetesting

Matematisk institutt

2026-04-13

Statistiske signifikanstester

Statistiske signifikanstester brukes når vi ønsker å ta stilling til et utsagn om en parameter i en populasjon.

- Utsagnet kaller vi hypotesen.
- Vi bruker observerte data til å teste hypotesen om populasjonen.

Overordnet prosedyre:

- Beregn sannsynlighet for observert utfall (eller noe mer ekstremt) av observator gitt antatt hypotese.
- Hvis sannsynlighet liten, forkast hypotese.

Hypotesetest: oversikt

De viktigste stegene i en hypotesetest er:

1. Formuler nullhypotesen H_0 og den alternative hypotesen H_a.
2. Bestem signifikansnivået α.
3. Velg en testobservator som har en kjent/estimerbar fordeling under H_0.
4. Beregn p-verdien: sannsynligheten for å observere noe like eller mer ekstremt enn det vi har observert dersom H_0 er sann.
5. Trekk en konklusjon.

Eksempel: Den mistenkelige mynten

Vi har en mynt og lurer på om dette er en rettferdig mynt.

Nullhypothesen er som før:

H_0: *Mynten er rettferdig — hvert kast gir kron med sannsynlighet* 0.5*, uavhengig av alle andre kast*

Vi har tidligere gjennomført en Monte Carlo-test for å sjekke nullhypotesen basert på observert data.

Vi kan også vurdere den samme hypotesen ved å bruke egenskaper til fordelinger vi har lært tidligere i kurset.

Eksempel: Den mistenkelige mynten

La X være antall kron i $n = 10$ myntkast. Da er $X \sim Bin(n = 10, p)$.

Nullhypotesen om en rettferdig mynt svarer til

H_0: $p = 0.5$

Nullmodellen er altså:

x	0	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

Vurderingene fra siste uke svarer til en ensidig alternativ hypotese

H_a: $p > 0.5$

Vi setter signifikansnivået til $\alpha = 0.05$.

Eksempel: Den mistenkelige mynten

Nullhypotesen handler om verdien til den ukjente parameteren p.

En naturlig estimator for p er $\hat{p} = X/n$, så vi velger den som testobservator:

$$T = \frac{X}{n}.$$

Under H_0 har T en kjent fordeling:

x	0	1	2	3	4	5	6	7	8	9	10
t	0	1/10	2/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	1
$P(X = x)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
$P(T = t)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001

Eksempel: Den mistenkelige mynten

Alle disse vurderingene skjer *før* vi ser data. Vi observerer nå 8 kron i 10 kast,

$$t = \frac{8}{10}.$$

For å finne *p*-verdien, finner vi *sannsynligheten for å observere noe like eller mer ekstremt enn det vi har observert dersom H_0 er sann*:

t	0	1/10	2/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	1
$P(T = t)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
$H_a : p > 0.5$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	**0.044**	**0.010**	**0.001**

P-verdien er: $P(T \geq t) = 0.044 + 0.01 + 0.001 = 0.055 > \alpha = 0.05$

$\Rightarrow$ vi har ikke grunnlag til å forkaste nullhypotesen.

Eksempel: Den mistenkelige mynten

Vi kunne også ha valgt en tosidig alternativ hypotese:

H_a: $p \neq 0.5$

P-verdien er fortsatt *sannsynligheten for å observere noe like eller mer ekstremt enn det vi har observert dersom H_0 er sann*, men vi må nå se i bette retninger (dvs. vurdere både *usedvanlig mange* kron og *usedvanlig få* kron):

t	0	1/10	2/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	1
$P(T = t)$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
$H_a : p > 0.5$	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	**0.044**	**0.010**	**0.001**
$H_a : p \neq 0.5$	**0.001**	**0.010**	**0.044**	0.117	0.205	0.246	0.205	0.117	**0.044**	**0.010**	**0.001**

P-verdien er nå: $2 \times P(T \geq t) = 2 \times 0.055 = 0.11 > \alpha = 0.05$

$\Rightarrow$ vi har ikke grunnlag til å forkaste nullhypotesen.

Nullhypotese H_0 og alternativ hypotese H_a

Nullhypotese H_0:

- Utgangshypotese
- Typisk nullhypotese: Det er ingen effekt/forskjell
- Signifikanstesten er designet for å angi bevisstyrke mot H_0

Alternativ hypotese H_a:

- Det vi ønsker å konkludere med dersom dataene ikke er i samsvar med H_0
- Typisk alternativ hypotese: “Det er en effekt”, eller “Det er en forskjell”

Triks: Fordi H_a uttrykker effekten vi skal undersøke om eksisterer eller ikke, kan det være praktisk å starte med å formulere H_a og deretter sette opp H_0 som utsagnet om at den ønskede effekten ikke er tilstede.

Hypoteser: generelle prinsipper

Hypoteser er alltid utsagn/antakelser om populasjoner (eller modeller), ikke et spesielt utfall. Derfor må H_0 og H_a alltid formuleres ut fra de ukjente populasjonsparameterne.

Når vi skal formulere H_a, må vi ta stilling til om den skal være ensidig (f.eks. $H_a : p > 0.5$) eller tosidig (f.eks. $H_a : p \neq 0.5$). Det uttrykker om parameteren er forskjellig fra nullhypoteseverdien i en bestemt retning eller ikke.

Merk: Metoden er ikke gyldig om man ser på dataene for å formulere H_a.

Med mindre du har god begrunnelse for retning på effekten gjennom forhåndskunnskap, velg tosidig H_a.

Signifikansnivå og beslutning

Hvordan konkludere om vi har observert en statistisk signifikant forskjell?

Vi må på forhånd ha bestemt signifikansnivå α:

- Typisk: $\alpha = 0.05$ (eller 0.01)
- α er grenseverdi: Vi forkaster H_0 når p-verdi $\leq \alpha$

$\Rightarrow$ Det er ikke grunnlag for å forkaste når p-verdi $> \alpha$.

Signifikant betyr at bevisene mot nullhypotesen nådde standarden satt av α for datasettet som ble analysert.

NB: Det er ingen skarp grense mellom signifikant og ikke-signifikant, men det er økende grad av bevis mot H_0 når p-verdien blir mindre.

$\Rightarrow$ Rapporter alltid p-verdien!

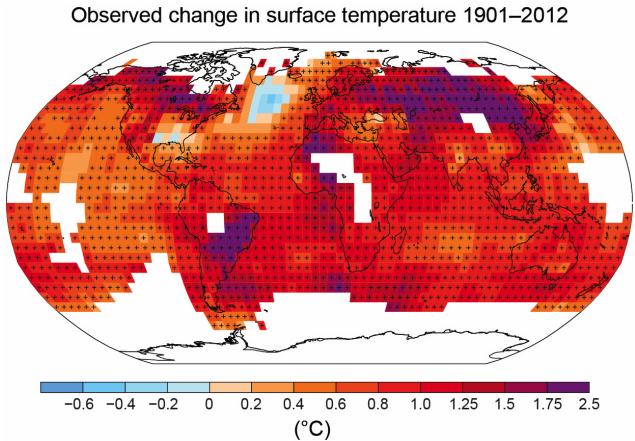
Feilrate

En hypotesetest kan føre til feil konklusjon på to forskjellige måter:

- **Type I-feil**: Vi forkaster H_0 selv om H_0 faktisk er sann.
  (Falsk positiv / "vi tror det er en effekt, men det er det ikke".)
- **Type II-feil**: Vi forkaster **ikke** H_0 selv om H_a faktisk er sann.
  (Falsk negativ / "det er en effekt, men vi oppdager den ikke".)

Ved $\alpha = 0.05$ er det 5% sjanse for å gjøre Type I-feil hvis alt annet stemmer, men hvis vi setter α ned, øker sjansen for Type II-feil.

Multippel testing: Ved veldig mange test på en gang, kan 5% feilrate bli mange ⇒ metodene må justeres i forhold til antall test



Kilde: https://www.ipcc.ch/report/ar5/wg1/summary-for-policymakers/figspm-01/

Enkelt eksempel: Normal forventning med kjent varians

Anta vi har

$$X \sim N(\mu, \sigma^2).$$

Vi observerer data $x_1, \ldots, x_n$ og ønsker å sjekke om vi kan konkludere basert på dataene at $\mu \neq \mu_0$. Vi setter opp en test på nivå α med hypoteser:

- $H_0 : \mu = \mu_0$
- $H_a : \mu \neq \mu_0$

Vi bruker den informasjonen vi har observert: $\overline{x}$ (og antar at σ er kjent). Vi standardiserer

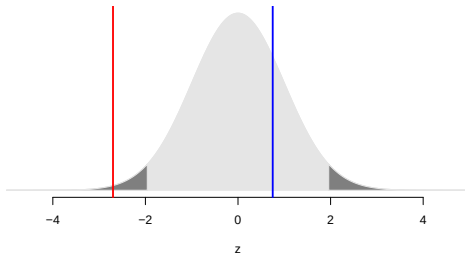
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

Under nullhypotesen gjelder at

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Enkelt eksempel: Normal forventning med kjent varians

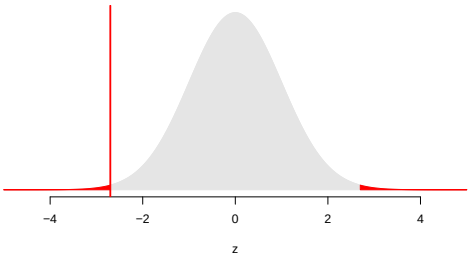
Under nullhypotesen gjelder at $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n}) \sim N(0, 1)$.

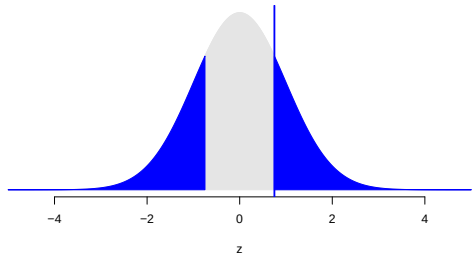


Under nullhypotesen er det liten sjanse for å observere den **røde**, mens det er stor sjanse for å observere den **blå**.

Enkelt eksempel: Normal forventning med kjent varians

For å ha litt mer informasjon, så regner vi ut *p*-verdien: Hvor sannsynlig er det å observere noe like eller mer ekstremt enn $\bar{x}$ under nullhypotesen?

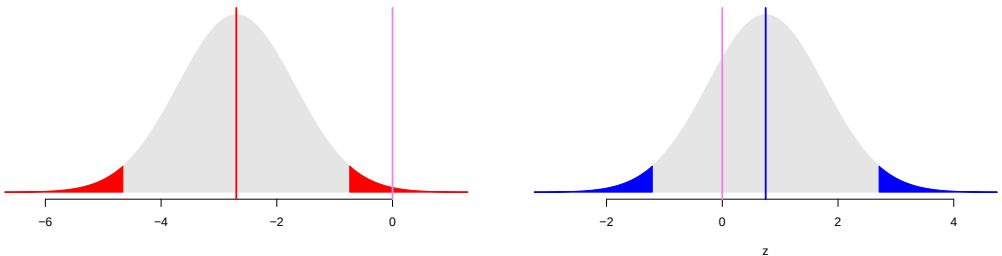




P-verdien for den **røde** er 0.007 mens den er 0.45 for den **blå**.

Enkelt eksempel: Normal forventning med kjent varians

Husk at et $(1-\alpha)100\%$ konfidensintervall for μ er gitt ved $\overline{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$. Vi kan også beregne disse konfidensintervallene (her for $\alpha = 0.05$) og sjekke om μ_0 ligger innenfor:



$\Rightarrow \mu_0$ ligger utenfor den **røde** men innenfor den **blå**.

Oppsummering: Tosidige test og konfidensintervall

Konfidensintervall for μ med konfidensnivå C:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Verdier av μ utenfor intervall er ikke kompatible med data.
- Mulig test-prosedyre: Forkaste H_0 hvis μ_0 ikke er i konfidensintervallet.

Denne prosedyren er ekvivalent med signifikanstest:

- En tosidig signifikanstest med nivå α som forkaster $H_0 : \mu = \mu_0$ er ekvivalent med at μ_0 faller utenfor konfidensintervallet for μ med nivå $C = (1 - \alpha)100\%$.

Noen ganger er det enklere å konstruere konfidensintervaller.

Standardisert testobservator

Vi har her brukt en standardisert testobservator av type:

$$z = \frac{\text{Estimat} - \text{Parameterverdi under } H_0}{\text{(Estimert) standardavvik}}$$

Dette er typisk form på testobservator for forventning, hvor det avgjørende (og krevende) er å ha et bra (nok) estimat for standardavviket.

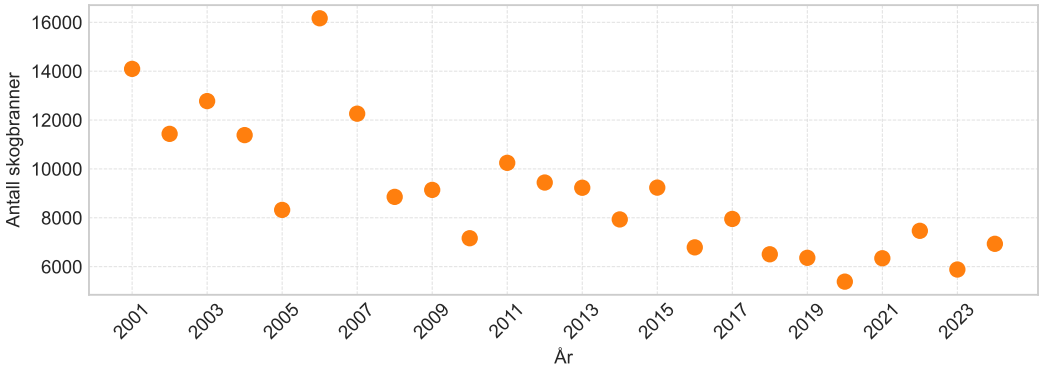
Hypotesetest for forventning i tre forskjellige situasjoner

- Vi har en populasjon og ønsker å si noe om forventningen
- Vi har to populasjoner og ønsker å vite om de har samme forventning
- Vi har to populasjoner hvor observasjonene danner par med en enhet i hver gruppe (matched par-design)

Vi sammenligner to prosedyrer:

- ***t*-test**: Sentralgrenseteoremet sier at et gjennomsnitt er ca. normalfordelt $\Rightarrow$ det standardiserte gjennomsnittet er dermed ca. *t*-fordelt
- **Permutasjonstest**: Minimale antakelser om form på fordeling til testobservator

1) Var gjennomsnittlig årlig antall skogbranner i USA i perioden 2011–2024 systematisk forskjellig fra langtidsnivået for 2001–2024?



Antall skogbranner forårsaket av lynnedslag i USA (Kilde: National Interagency Fire Center).

Oppsett for én-utvalgs t-test

Vi har observert $x_1, \ldots, x_{14}$ med gjennomsnitt $\bar{x} = 7551$ og empirisk standardavvik $s_x = 1499$ fra en populasjon med forventning μ.

Vi ønsker å teste

$$H_0 : \mu = 9055 \quad \text{mot} \quad H_a : \mu \neq 9055,$$

hvor vi velger $\alpha = 0.05$. Testobservatoren er gitt ved

$$t = \frac{\bar{x} - 9055}{s_x/\sqrt{n}} = \frac{7551 - 9055}{1499/\sqrt{14}} = -3.75$$

og den har t-fordeling med $n - 1 = 13$ frihetsgrader.

P-verdien er gitt ved $2P(T > |t|) = 2P(T > 3.75)$ hvor $T \sim t(13)$.

Beregning av *p*-verdi

```python
from scipy.stats import t

d_vec = np.array([10249, 9443, 9230, 7933,
  9235, 6790, 7953, 6507, 6362, 5387, 6344,
  7467, 5883, 6935], dtype=float)
n = len(d_vec)
s = np.std(d_vec, ddof=1)/np.sqrt(n)
t_est = (np.mean(d_vec) - 9055)/s
p = 2 * (1- t.cdf(abs(t_est), n-1))

print(f"t = {t_est:.2f}")
print(f"Two-sided p-value: {p:.6f}")
```

```
t = -3.75
Two-sided p-value: 0.002412
```

```python
import numpy as np
from scipy import stats

d_vec = np.array([10249, 9443, 9230, 7933,
  9235, 6790, 7953, 6507, 6362, 5387, 6344,
  7467, 5883, 6935], dtype=float)

t_est, p = stats.ttest_1samp(d_vec, popmean=9055)

print(f"t = {t_est:.2f}")
print(f"Two-sided p-value: {p:.6f}")
```

```
t = -3.75
Two-sided p-value: 0.002412
```

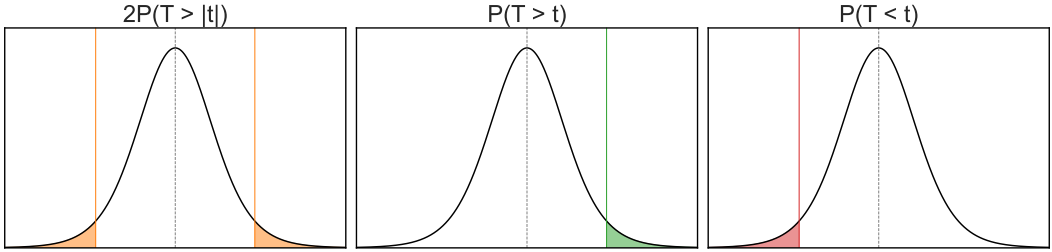
$\Rightarrow$ Vi forkaster nullhypotesen og konkluderer at $\mu \neq 9055$.

Forutsetninger for én-utvalgs *t*-test

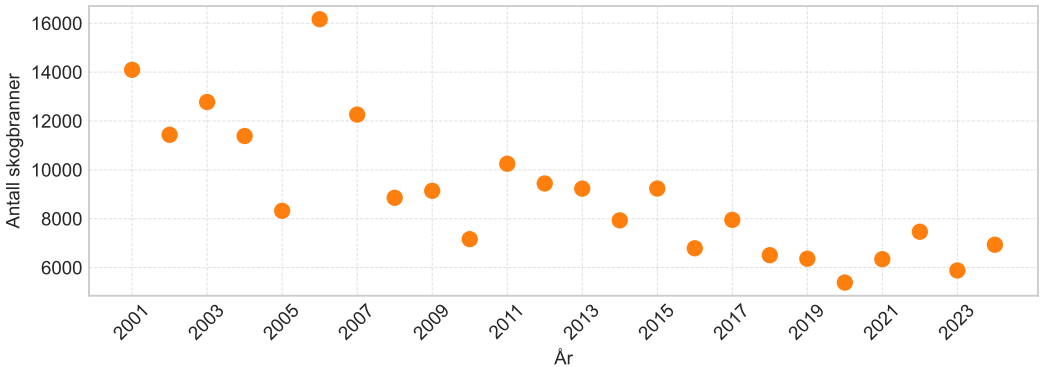
- Data kan ses som et *enkelt tilfeldig utvalg*
- Populasjonsfordelingen er omtrent normal:
  - For små utvalg ($n < 15$) kreves data som er nær normalfordelt
  - For store utvalg ($n > 40$) tåles større avvik fra normalitet
- Ingen ekstreme outliers

Vi har brukt tosidig alternativ hypotese. Valget av alternativ hypotese endrer beregningen av *p*-verdien slik:

$$H_a : \mu \neq \mu_0 \quad \text{gir} \quad p = 2P(T > |t|)$$
$$H_a : \mu > \mu_0 \quad \text{gir} \quad p = P(T > t)$$
$$H_a : \mu < \mu_0 \quad \text{gir} \quad p = P(T < t)$$



2) Endret det gjennomsnittlige årlige antallet skogbranner seg mellom perioden 2001–2010 og perioden 2011–2024?



Antall skogbranner forårsaket av lynnedslag i USA (Kilde: National Interagency Fire Center).

Oppsett for to-utvalgs *t*-test

Vi har to populasjoner, 1 og 2, og ønsker å teste

$$H_0 : \mu_1 = \mu_2 \quad \text{mot} \quad H_a : \mu_1 \neq \mu_2,$$

som også kan skrives

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{mot} \quad H_a : \mu_1 - \mu_2 \neq 0.$$

Betegn med $\overline{X}_1$ og $\overline{Y}_2$ gjennomsnittene i de to gruppene. Da er testobservatoren gitt ved

$$t = \frac{\overline{X}_1 - \overline{Y}_2}{\text{estimert standardfeil av } (\overline{X}_1 - \overline{Y}_2)}.$$

To-utvalgs *t*-test med lik (pooled) varians

- Observasjonene er uavhengige både innenfor og mellom utvalgene
- Begge populasjoner er tilnærmet normalfordelte
- De to populasjonene har samme varians: $\sigma_1^2 = \sigma_2^2$

Da bruker vi

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

og testobservator

$$t = \frac{\bar{X}_1 - \bar{Y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\big(\text{df} = n_1 + n_2 - 2\big).$$

To-utvalgs *t*-test med ulik varians (Welch *t*-test)

- Observasjonene er uavhengige både innenfor og mellom utvalgene
- Begge populasjoner er tilnærmet normalfordelte
- De to populasjonene har ulik varians: $\sigma_1^2 \neq \sigma_2^2$

Da bruker vi testobservator

$$t = \frac{\bar{X}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t\big(\text{df} \approx \min\{n_1 - 1, n_2 - 1\}\big).$$

Vi bruker vanligvis denne versjonen, med mindre vi har informasjon som sier at variansene bør være like.

Beregning av *p*-verdi

```python
import numpy as np
from scipy import stats

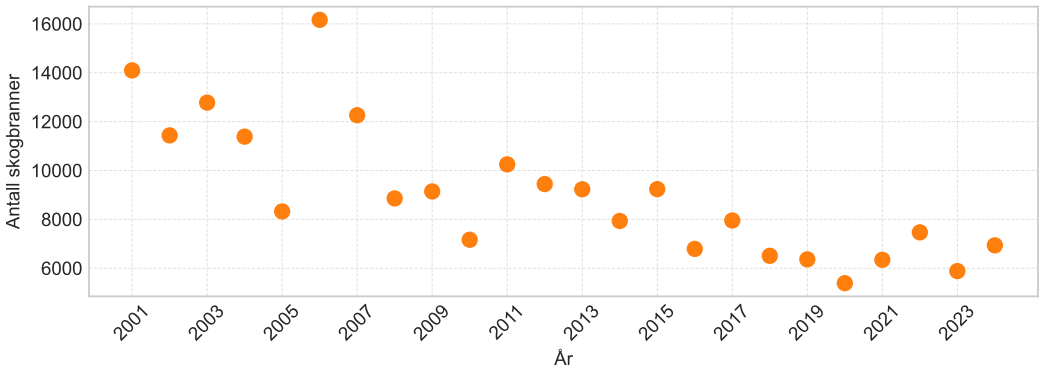
x = lightning_fires[0:10]
y = lightning_fires[10:24]

# 1) Equal-variance (pooled) two-sample t-test
t_equal, p_equal = stats.ttest_ind(x, y, equal_var=True)
print(f"Equal-variance t-test: t = {t_equal:.3f}, p-value = {p_equal:.3f}")

# 2) Unequal-variance (Welch) two-sample t-test
t_welch, p_welch = stats.ttest_ind(x, y, equal_var=False)
print(f"Welch t-test:          t = {t_welch:.3f}, p-value = {p_welch:.3f}")
```

```
Equal-variance t-test: t = 4.084, p-value = 0.000
Welch t-test:          t = 3.704, p-value = 0.003
```

Vi kan også spørre mer generelt: *2b) Endret fordelingen til årlig antall skogbranner seg mellom 2001–2010 og 2011–2024?*



Antall skogbranner forårsaket av lynnedslag i USA (Kilde: National Interagency Fire Center).

Oppsett: Permutasjonstest for to grupper

Dette gir en hypotesetest som handler om hele fordelingen:

H_0 : De to gruppene kommer fra samme fordeling

H_a : De to gruppene kommer fra forskjellige fordelinger

Under H_0 kommer alle observasjonene fra samme fordeling og etiketten tildelt hver observasjon er vilkårlig.

Vi velger en testobservator som beskriver egenskaper til fordelingene, f.eks. forskjell i gjennomsnitt eller median til de to gruppene.

Vi setter opp en permutasjonstest slik som beskrevet i antibiotika eksemplet tidligere, hvor testobservator for de observerte dataene sammenlignes mot tilsvarende verdier for versjoner av dataene med vilkårlige gruppeetiketter.

Beregning av *p*-verdi (se kap. 10.6 i bok for kodedetaljer)

```
x = lightning_fires[0:10]
y = lightning_fires[10:24]

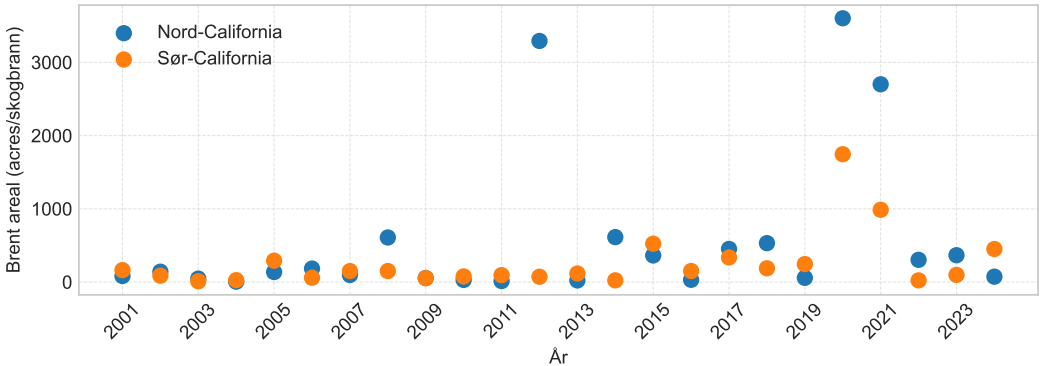
obs_mean, p_mean = permutation_test_mean(x, y, n_perm=20000, random_state=0)
obs_median, p_median = permutation_test_median(x, y, n_perm=20000, random_state=0)

print(f"Mean-based:   statistic = {obs_mean:.3f}, p-value = {p_mean:.5f}")
print(f"Median-based: statistic = {obs_median:.3f}, p-value = {p_median:.5f}")
```

```
Mean-based:   statistic = 3608.714, p-value = 0.00030
Median-based: statistic = 4208.500, p-value = 0.00020
```

$\Longrightarrow$ Vi forkaster nullhypotesen og konkluderer at de to gruppene kommer fra forskjellige fordelinger.

3) Er de gjennomsnittlige brente arealene i Nord- og Sør-California systematisk forskjellige, eller skyldes forskjellene bare tilfeldig variasjon fra år til år?



Gjennomsnittlig brent areal per skogbrann (Kilde: National Interagency Fire Center).

Matched par-design

En spesiell situasjon oppstår når hvert datapunkt i én gruppe naturlig kan pares med ett datapunkt i den andre gruppen (paret utvalg).

Skogbranner og skader avhenger sterkt av klima (tørke, hetebølger), som ofte påvirker hele California samtidig.

Vi bruker derfor en differanse-tilnærming:

- Vi ser på ett datasett med årlige forskjeller mellom regionene.
- Hvis en ekstrem tørke påvirker begge regioner likt et år, vil differansen fjerne denne felles effekten.

Datene blir da en enkel vektor med tall:

```
[-80, 54, 36, -20, -155, 125, -53, 458, 2, -48, -81, 3220,
 -94, 590, -160, -120, 116, 344, -188, 1856, 1713, 281, 269, -377]
```

Standard én-utvalgs *t*-test

Vi er tilbake i en sitasjon med et utvalg og hypoteser:

$$H_0 : \mu_D = 0 \quad \text{mot} \quad H_a : \mu_D \neq 0,$$

hvor vi setter $\alpha = 0.05$ og antar

- parstrukturen er riktig
- parene er uavhengige
- differansene er tilnærmet normalfordelte

Vi utfører en standard *t*-test og kan ikke forkaste nullhypotesen:

```
t_stat, p_value = stats.ttest_1samp(d_vec, popmean=0)
```

```
t = 1.92
Two-sided p-value: 0.0678
```

Permutasjonstest

Vi kan alternativt gjennomføre en permutasjonstest med hypoteser:

H_0 : Fordelingen er symmetrisk rundt 0
H_a : Fordelingen har et senterpunkt ulikt 0.

En naturlig testobservator er den gjennomsnittlige differansen

$$\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i.$$

Permutasjonstest

Under H_0 er det helt tilfeldig om D_i har et positivt eller negativt fortegn.

Nullfordelingen til testobservatoren kan derfor beregnes ved å endre fortegn på differansene på alle mulige måter:

$$\frac{1}{n}\sum_{i=1}^{n} k_i D_i,$$

der hver k_i er enten 1 eller -1.

```python
import numpy as np

def paired_permutation_test(x, y, n_perm=10000, random_state=None):
    rng = np.random.default_rng(random_state)
    d = x - y
    obs = np.mean(d)

    count = 0
    for _ in range(n_perm):
        signs = rng.choice([-1, 1], size=len(d))
        perm_mean = np.mean(signs * d)
        if abs(perm_mean) >= abs(obs):
            count += 1

    p_value = count / n_perm
    return obs, p_value

obs_diff, pval = paired_permutation_test(nca_avg, sca_avg, n_perm=10000, random_state=0)
print(f"Observed mean difference: {obs_diff:.4f}, p-value: {pval:.4f}")
```

```
Observed mean difference: 320.2984, p-value: 0.0544
```

Bruk og misbruk av tester

Det er ikke innafor å teste hypoteser med data brukt til å formulere dem!

Det er nesten for enkelt å utføre en statistisk hypotesetest med Python:

1. Å tolke og å forstå resultatet er vanskeligere
   – Testen er kun gyldig under visse forutsetninger
   – Konfidensnivå: Ingen klar grense; vanlig og nyttig å rapportere *p*-verdi
2. Ofte gjennomføres mange ulike tester
   – *p*-verdi relatert til å gjøre én test
   – Ved mange tester må justeringer gjøres for å kontrollere på falske positive.

Statistisk signifikant vs praktisk relevant

Å forkaste H_0 betyr at vi har observert en statistisk signifikant forskjell.

Ikke vektlegg statistisk signifikans alene, se også på de faktiske resultatene. God idé å rapportere konfidensintervall for parameteren vi undersøker i tillegg.

- På den ene siden: Kan ha forkastet nullhypotesen selv om effekten er ubetydelig liten, hvis n er stor.
- På den andre siden: Å ikke kunne forkaste betyr ikke at H_0 er sann; dette er spesielt viktig å huske når n er liten.

Statistisk signifikans betyr ikke at det er praktisk relevant!

Eksempel: korrelasjon

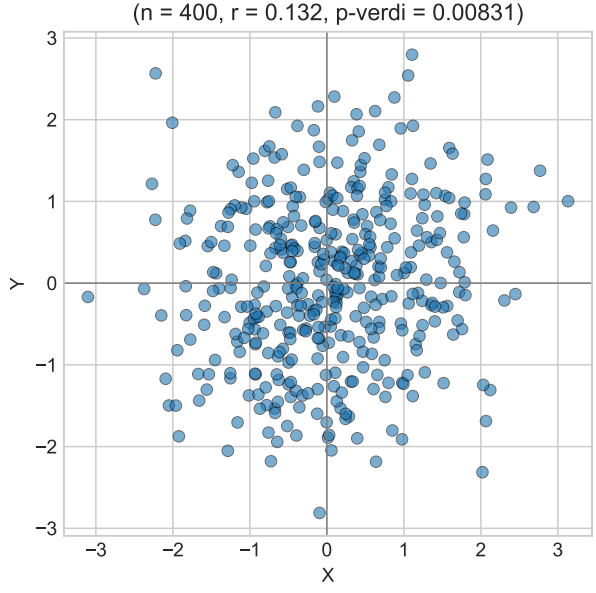
Anta at vi har to variabler X og Y hvor vi ønsker å teste

$$H_0: \text{ Ingen korrelajson, } \rho = 0 \quad \text{mot} \quad H_a: \rho \neq 0,$$

Vi har 400 observasjoner og observerer empirisk korrelasjon $r = 0.132$.

Vi får p-verdi på 0.00831 og testen er statistisk signifikant med $\alpha = 0.05$.

Klar indikasjon på sammenheng mellom X og Y, men er det praktisk relevant?


(n = 400, r = 0.132, p-verdi = 0.00831)
Y
X
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Fra data til parameter

Statistisk inferens:
Fra observerte data om utvalg til konklusjon om populasjonsparameter.

To viktige metoder:

- Konfidensintervall
- Signifikanstester

Basert på utvalgsfordelinga til observator:

- Krever sannsynlighetsmodell for dataene
- Statistisk inferens baserer seg på at dataene kommer fra et tilfeldig utvalg eller et randomisert eksperiment

Konfidensintervall

- Intervall av plausible verdier for en ukjent parameter
- Konstruert slik at det dekker sannheten i en andel $(1 - \alpha)$ av gjentatte forsøk
- Formel avhenger av modellantakelser:
  - Normalmodell:
    - kjent σ: $\overline{x} \pm z_{\alpha/2}\, \sigma/\sqrt{n}$
    - ukjent σ: $\overline{x} \pm t_{n-1,\alpha/2}\, s/\sqrt{n}$
  - Andeler (binomisk modell, normaltilnærming):
    - $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
  - Wald-intervaller:
    - generell form: $\hat{\theta} \pm z_{\alpha/2}\, \widehat{\text{se}}(\hat{\theta})$

Bootstrap-metoden

- En data-dreven måte å måle usikkerhet på, uten å anta en bestemt sann fordeling.
- Lar oss imitere prosessen med å skaffe nye datasett, slik at vi kan anslå variasjonen til estimatet uten å generere flere virkelige utvalg.
- I stedet for å hente gjentatte, uavhengige datasett fra populasjonen, konstruerer vi nye datasett ved
  - parametrisk bootstrap: antar en parametrisk modell for dataene
  - ikke-parametrisk bootstrap: gjentatt trekking av observasjoner fra det opprinnelige datasettet med tilbakelegging

Hypotesetesting

Hvordan gjennomføre en statistisk signifikanstest?

1. Formuler H_0 og H_a
2. Beregn testobservator
3. Finn p-verdi
4. Formuler en konklusjon

NB: Vi bruker som regel datamaskin til å finne p-verdien, men trenger å ha generell forståelse til å selv kunne

- Formulere H_0 og H_a
- Tolke p-verdien
- Bedømme om forutsetningene for å bruke testen er oppfylt