# Intro til hypotesetesting – Analyse av kontinuerlige data 21. april 2005

Tron Anders Moger

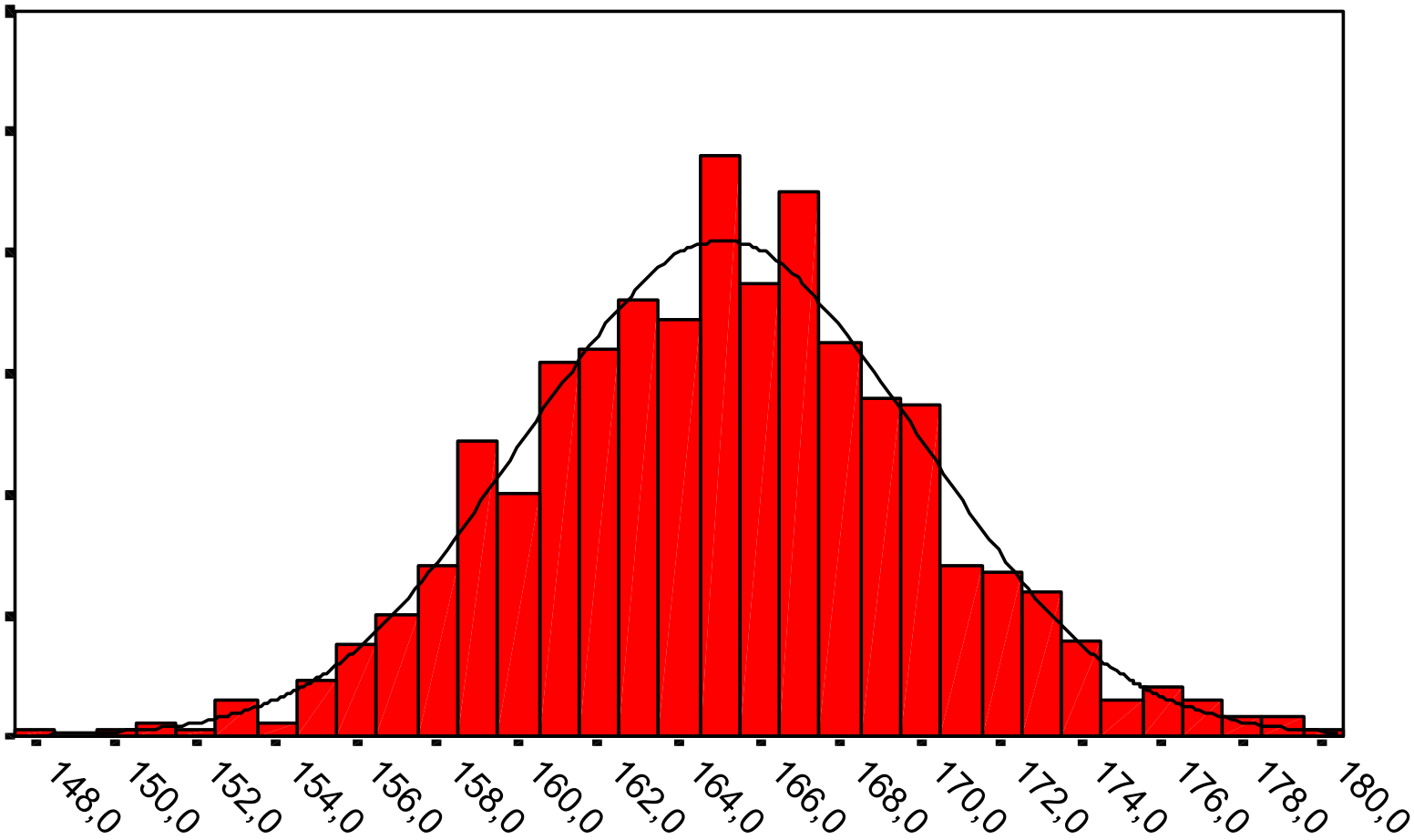
Seksjon for medisinsk statistikk, UIO

# Repetisjon fra i går: Normalfordelingen

- Variasjon i målinger kan ofte beskrives med normal-fordelingen
- Fordelingen er symmetrisk og har to parametere:
  - Forventning: $\mu$
  - Standardavvik (Standard deviation): $\sigma$
- Disse angir tyngdepunkt og spredning i fordelingen
- Viktig: Et intervall som er 1.96 standardavvik på hver side av forventningen dekker 95% av fordelingen

Normal distribution

mean = 165, std. deviation=5



# Standardfeil (standard error)

- La oss si at du har samlet et utvalg på *n* observasjoner for å beregne et gjennomsnitt
- Samler du flere utvalg av samme størrelse får du ikke samme gjennomsnitt
- Standardavviket til gjennomsnittet kalles standardfeil og estimeres med $\frac{s}{\sqrt{n}}$
- Hvis observasjonene er normalfordelt, kan du være 95% sikker på at det sanne gjennomsnittet (forventningen) ligger i intervallet gj.snitt+/-1.96*standardfeil

# Konfidensintervall i SPSS

- *Analyze- Descriptive Statistics - Explore*

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| WEIGHT | Mean | | 70,7000 | 2,6548 |
| | 95% Confidence Interval for Mean | Lower Bound | 65,1435 | |
| | | Upper Bound | 76,2565 | |
| | 5% Trimmed Mean | | 70,5000 | |
| | Median | | 69,0000 | |
| | Variance | | 140,958 | |
| | Std. Deviation | | 11,8726 | |
| | Minimum | | 50,00 | |
| | Maximum | | 95,00 | |
| | Range | | 45,00 | |
| | Interquartile Range | | 10,0000 | |
| | Skewness | | ,427 | ,512 |
| | Kurtosis | | ,263 | ,992 |

# t-fordelingen

- Få observasjoner: Dårlig estimat for standardavviket
- Konfidensintervall basert på normalfordelingen vil være for smalt
- Betyr at et 95% konfidensintervall IKKE vil dekke 95% av den sanne fordelingen
- Må bruke t-fordelingen isteden: Er også symmetrisk, men har større standardavvik

# Eksempel: Gjennomsnittlig daglig energiinntak i kJ hos 11 friske kvinner målt over 10 dager

- | SUBJECT | INTAKE |
  |---|---|
  | 1 | 5260 |
  | 2 | 5470 |
  | 3 | 5640 |
  | 4 | 6180 |
  | 5 | 6390 |
  | 6 | 6515 |
  | 7 | 6805 |
  | 8 | 7515 |
  | 9 | 7515 |
  | 10 | 8230 |
  | 11 | 8770 |

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| INTAKE | Mean | | 6753,64 | 344,36 |
| | 95% Confidence Interval for Mean | Lower Bound | 5986,35 | |
| | | Upper Bound | 7520,93 | |
| | 5% Trimmed Mean | | 6724,60 | |
| | Median | | 6515,00 | |
| | Variance | | 1304445 | |
| | Std. Deviation | | 1142,12 | |
| | Minimum | | 5260 | |
| | Maximum | | 8770 | |
| | Range | | 3510 | |
| | Interquartile Range | | 1875,00 | |
| | Skewness | | ,428 | ,661 |
| | Kurtosis | | -,793 | 1,279 |

# Ønsker å teste om kvinnenes energi-inntak er likt det anbefalte energi-inntaket (ett-utvalgs t-test):

- Anbefalt energi-inntak: 7725kJ
- Nullhypotese $H_o$: Energi-inntaket er 7725kJ
- Alternativ hypotese $H_1$: Energi-inntaket er forskjellig fra 7725 kJ

$$\text{Test} = \frac{\text{observert gj.snitt - forventet gj.snitt}}{\text{standardfeil til observert gj.snitt}}$$

$$= \frac{6753.6 - 7725}{1142.1/\sqrt{11}} = -2.821$$

- Denne størrelsen er t-fordelt med 10 frihetsgrader (antall personer -1)
- En stor positiv eller negativ verdi av testen indikerer at nullhypotesen er gal
- Hvordan vite om man skal forkaste $H_o$ eller ikke?

# To VIKTIGE begreper

- *P-verdi*: Sannsynligheten for at det observerte gjennomsnittet er likt det forventede gjennomsnittet etter at testen er gjort
- Hvis *p-verdien* er lavere enn *signifikansnivået* forkastes nullhypotesen
- *Signifikansnivået* gir den øvre sannsynligheten for å forkaste nullhypotesen hvis den er sann
- Vanlige signifikansnivåer: 5% eller 1%
- Signifikansnivå 5% betyr at det er 5% sannsynlighet for å forkaste nullhypotesen når den er sann
- Medfører at man er minst 95% sikker på å ikke forkaste nullhypotesen hvis den er sann
- Eller: Man vil være minst 95% sikker på at man ikke forkaster nullhypotesen på grunn av tilfeldigheter
- Dette er begreper som benyttes i alle statistiske tester, ikke bare for t-tester

I SPSS: *Analyze - Compare means - One-sample t test*

Test variable*: intake*

Test value*: 7725*

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| INTAKE | 11 | 6753,64 | 1142,12 | 344,36 |

**One-Sample Test**

| | Test Value = 7725 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Mean Difference | Lower | Upper |
| INTAKE | -2,821 | 10 | ,018 | -971,36 | -1738,65 | -204,07 |

# Hva blir konklusjonen av testen?

- Vi samlet inn data på energi-inntaket til 11 kvinner for å studere om dette var signifikant forskjellig fra det anbefalte energi-inntaket på 7725kJ.
- I gjennomsnitt lå energi-inntaket til kvinnene 971kJ lavere enn det anbefalte inntaket.
- 95% konfidensintervall for differansen er (-1739, -205).
- Dette betyr at vi kan være 95% sikre på at den *sanne* differansen (ergo differansen vi hadde funnet hvis vi hadde data på alle kvinner istedenfor kun 11) ligger i dette intervallet.
- Differansen er signifikant på 5%-signifikansnivå, p-verdien er 0.02.

# MERK FØLGENDE:

- Hvis konfidensintervallet hadde omsluttet verdien 0, hadde p-verdien vært over 5%
- Det er altså en 1-1 sammenheng mellom p-verdier og konfidensintervall.
- Den gjennomsnittlige differansen mellom målingene og den forventede verdien er *effektmålet* i studien
- I en rapport oppgir man vanligvis effektmål, 95% konfidensintervall for effektmålet, samt p-verdien i konklusjonen

# Eksempel: Energiintak i kJ hos 11 friske kvinner målt før menstruasjon og etter menstruasjon

| SUBJECT | PREMENST | POSTMENS |
|---|---|---|
| 1 | 5260.0 | 3910.0 |
| 2 | 5470.0 | 4220.0 |
| 3 | 5640.0 | 3885.0 |
| 4 | 6180.0 | 5160.0 |
| 5 | 6390.0 | 5645.0 |
| 6 | 6515.0 | 4680.0 |
| 7 | 6805.0 | 5265.0 |
| 8 | 7515.0 | 5975.0 |
| 9 | 7515.0 | 6790.0 |
| 10 | 8230.0 | 6900.0 |
| 11 | 8770.0 | 7335.0 |

Number of cases read: 11 Number of cases listed: 11

# To målinger på samme individ: Parret t-test

- Hvis du har to like målinger per individ i et utvalg, sier vi at de er parrede
- Målingene vil være avhengige
- Eksempel: To blodtrykksmålinger av samme mann på en uke, vil ligge nær hverandre
- Baserer testen på differansen mellom målingene, reduserer dataene til ett utvalg
- $H_0$: Pre - og post menstruelle energiinntak er like

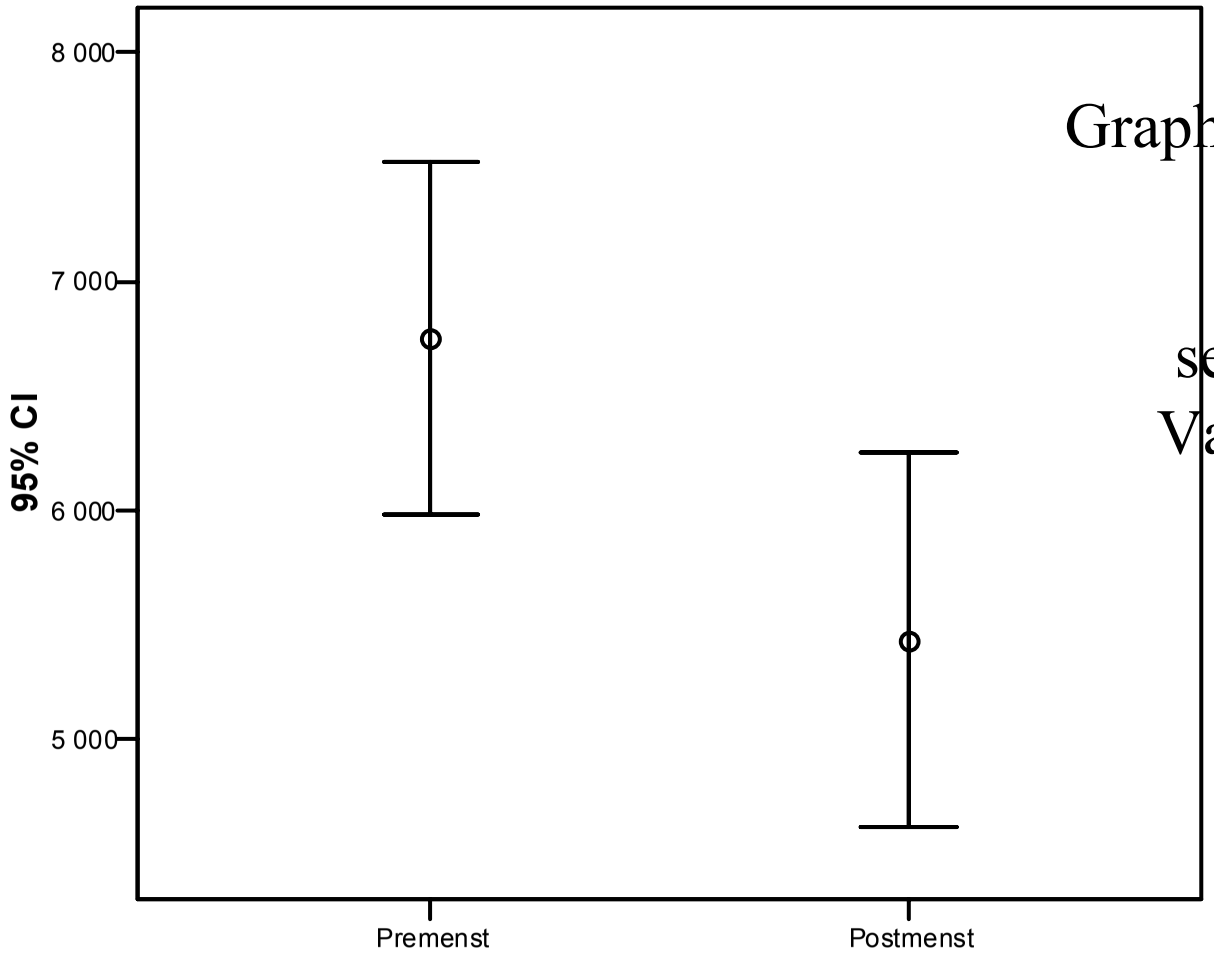
# Konfidensintervall og p-verdier for paret t-test i SPSS

- *Analyze - Compare Means - Paired-Samples T Test.*
- Klikk på de to variabelene du vil teste, og flytt dem over til høyre

# Konklusjonen av testen blir:

- Vi gjennomførte en studie for å undersøke om energi-inntaket hos kvinner forandrer seg før og etter menstruasjon.
- Vi samlet inn data på 11 kvinner, og i gjennomsnitt var energi-inntaket 1320kJ høyere etter menstruasjon sammenlignet med før.
- 95% konfidensintervall for forskjellen er (1074kJ, 1566kJ).
- Forskjellen er signifikant på 5%-nivå, p-verdien er $<0.01$.

# Grafisk fremstilling




I SPSS:
Graphs->error bar->simple
Kryss av på
summaries of
separate variables
Variable: Premenst
og postmenst

| ID | GROUP | ENERGY |
|---|---|---|
| 1 | 0 | 6.13 |
| 2 | 0 | 7.05 |
| .... | | |
| 12 | 0 | 10.15 |
| 13 | 0 | 10.88 |
| 14 | 1 | 8.79 |
| 15 | 1 | 9.19 |
| .... | | |
| 21 | 1 | 11.85 |
| 22 | 1 | 12.79 |

Number of cases read: 22
Number of cases listed: 22

Eksempel:
24 timers energiforbruk (MJ/dag) i 2 grupper av kvinner: 13 tynne og 9 overvektige

# To-utvalgs t-test

- Har her to uavhengige grupper: De som er tynne og de som er overvektige
- Vil teste om det er forskjellig energiforbruk
- $H_0$: energiforbruket er det samme for tynne og overvektige
- I SPSS: *Analyze - Compare Means - Independent-Samples T Test*
- Flytt Energy til "Test-variable"
- Flytt Group til "Grouping variable"
  Trykk på "Define Groups" og skriv 0 og 1 for de to gruppene

**Group Statistics**

| | GROUP | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| ENERGY | lean | 13 | 8.0662 | 1.2381 | .3434 |
| | obese | 9 | 10.2978 | 1.3979 | .4660 |

**Independent Samples Test**

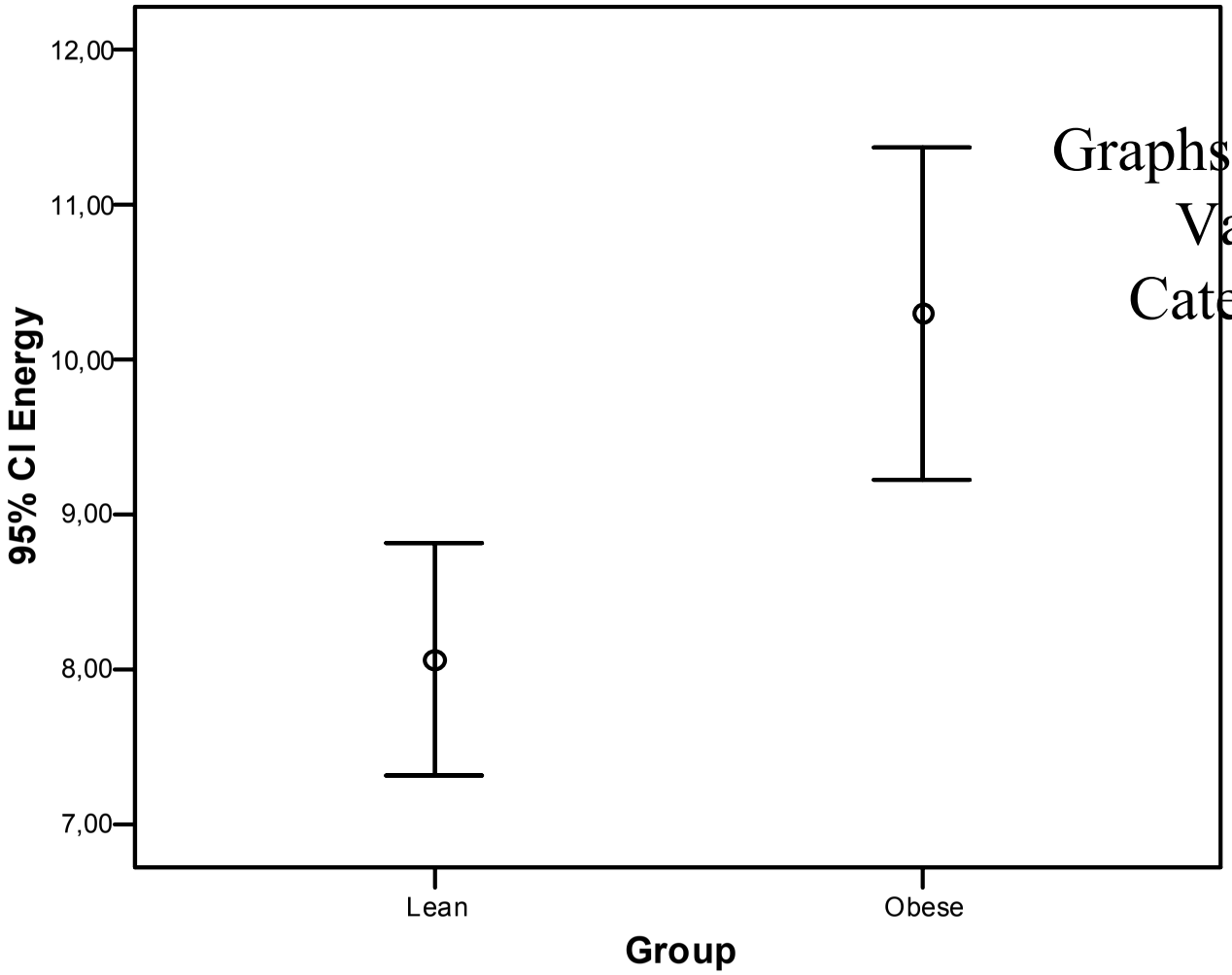
| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| ENERGY | Equal variances assumed | 1.002 | .329 | -3.946 | 20 | .001 | -2.2316 | .5656 | -3.4115 | -1.0518 |
| | Equal variances not assumed | | | -3.856 | 15.919 | .001 | -2.2316 | .5788 | -3.4592 | -1.0041 |

Over 0.05: Les første linje (Equal variances assumed)
Ellers: Les andre linje (Equal variances not assumed)

# Konklusjon

- Gjennomsnittlig energiforbruk for tynne og overvektige kvinner var henholdsvis 8.1MJ/dag og 10.3MJ/dag
- Den gjennomsnittlige forskjellen mellom gruppene var -2.2MJ/dag, 95% KI (-3.4,-1.1)
- Forskjellen mellom gruppene var signifikant på 5%-nivå, med p-verdi <0.01.
- Hva hvis vi har mer enn 2 uavhengige grupper? Enveis variansanalyse, ANOVA

# Grafisk fremstilling




I SPSS:
Graphs->error bar->simple
Variable: Energy
Category axis: Group

# Styrkeberegninger

- Alle eksemplene bruker små datasett
- Hvis man har mye data, vil usikkerheten bli mindre, og konfidensintervallene smalere
- Dette gjør at det er lettere å få signifikante forskjeller/resultater
- Ikke bare positivt: Med svært mye data vil alle nullhypoteser kunne forkastes. Dette betyr at vi kan få signifikante funn som ikke har klinisk relevans
- Vanlig å gjøre styrkeberegninger før studiestart, dvs man bestemmer seg for hvor stor forskjell må være for å være klinisk relevant, så beregner man hvor mye data man trenger for å kunne påvise denne forskjellen

# Antagelser

1. *Uavhengighet*: Alle observasjoner er uavhengige. Oppnåes ved tilfeldig trekning av individer; for parret t-test får vi uavhengighet ved å se på *differansen* mellom målingene.
2. *Normalfordelte data* (Sjekk: histogram, tester, Q-Q plott imorgen)
3. *Lik varians eller standardavvik i gruppene*

# Hva hvis antagelsene ikke holder?

- *Bruk ikke-parametriske metoder*

  Par data: Sign test eller Wilcoxon
  To-utvalg: Mann-Whitney

  *Eller*

- Transformer dataene (log-transformasjon)

# Parametriske metoder vi har sett:

- Effektmål: $\mu$ estimert ved gjennomsnitt
- Estimering:
  - Konfidensintervall for $\mu$
  - Konfidensintervall for $\mu_1$ -$\mu_2$

  (basert på gjennomsnitt og standardavvik)
- Testing:
  - Ett-utvalgs t-test
  - To-utvalgs t-test
- Metodene baserer seg på at dataene er normalfordelte (evt. gjennomsnittet normalfordelt)

# Ikke-parametriske metoder:

- Effektmål: Median, siden den ikke påvirkes av noen få ekstreme observasjoner i motsetning til gjennomsnittet
- Estimering
  - Vanlig å oppgi 25%-fraktilen og 75%-fraktilen som et ”konfidensintervall” for medianen
- Testing:
  - Parrede data: Sign test og Wilcoxon signed rank test
  - To uavhengige utvalg: Mann-Whitney test/Wilcoxon rank-sum test
- Gjør (nesten) ingen antagelser om spesielle fordelinger

# Ikke-parametriske tester

- De fleste testene baserer seg på rang-summer, og ikke de observerte verdiene. Summer av ranger antas tilnærmet normalfordelt, så vi kan bruke normaltilnærming for testobservatoren
- Ved to eller flere like verdier, gis de en gjennomsnittsrang

# Parrede data (ett utvalg)

- Eksempel: Daglig energi-inntak i kJ hos de 11 kvinnene

SUBJECT PREMENST POSTMENS

| | | |
|---|---|---|
| 1 | 5260.0 | 3910.0 |
| 2 | 5470.0 | 4220.0 |
| 3 | 5640.0 | 3885.0 |
| 4 | 6180.0 | 5160.0 |
| 5 | 6390.0 | 5645.0 |
| 6 | 6515.0 | 4680.0 |
| 7 | 6805.0 | 5265.0 |
| 8 | 7515.0 | 5975.0 |
| 9 | 7515.0 | 6790.0 |
| 10 | 8230.0 | 6900.0 |
| 11 | 8770.0 | 7335.0 |

Number of cases read: 11 Number of cases listed: 11

# Sign test

- Nullhypotesen er at differansen er 0
- Antar uavhengige observasjoner, ingen fordelingsantakelse
- Under nullhypotesen vil vi forvente at like mange differanser er over og under 0
- Testen teller opp antall positive differanser
- P-verdi fra binomisk fordeling når n er liten, normalfordeling når n stor
- SPSS: Analyze->nonparametric tests->two related samples. Kryss av sign under test

# To uavhengige utvalg: Wilcoxon/Mann-Whitney

- Nullhypotesen er her at fordelingene i de to gruppene/utvalgene er like
- For t-test: Forventningene like
- Antagelser: Uavhengighet innen og mellom grupper, lik fordeling i begge grupper
- Rangerer alle observasjonene som om de kom fra samme gruppe
- P-verdien beregnes fra tabeller eller normalfordelingen

| ID | GROUP | ENERGY |
|---|---|---|
| 1 | 0 | 6.13 |
| 2 | 0 | 7.05 |
| .... | | |
| 12 | 0 | 10.15 |
| 13 | 0 | 10.88 |
| 14 | 1 | 8.79 |
| 15 | 1 | 9.19 |
| .... | | |
| 21 | 1 | 11.85 |
| 22 | 1 | 12.79 |

Number of cases read: 22
Number of cases listed: 22

14

# I SPSS:

- Analyze->nonparametric tests->two independent samples tests
- Test type: Mann-Whitney U

# Oppsummering, ikke-parametriske tester

- Fordeler: Ingen strenge antakelser om fordeling
- Ulemper: For normalfordelte data vil ikke-parametriske tester ha dårligere styrke, dvs. hvis utvalgene er små, vil man ikke finne noen signifikante forskjeller
- Får ikke noe annet enn p-verdier fra ikke-parametriske tester