# Korrelasjon og lineær regresjon, litt om resultatpresentasjon

4. Mai 2005

Tron Anders Moger

# Forelesningen om t-tester:

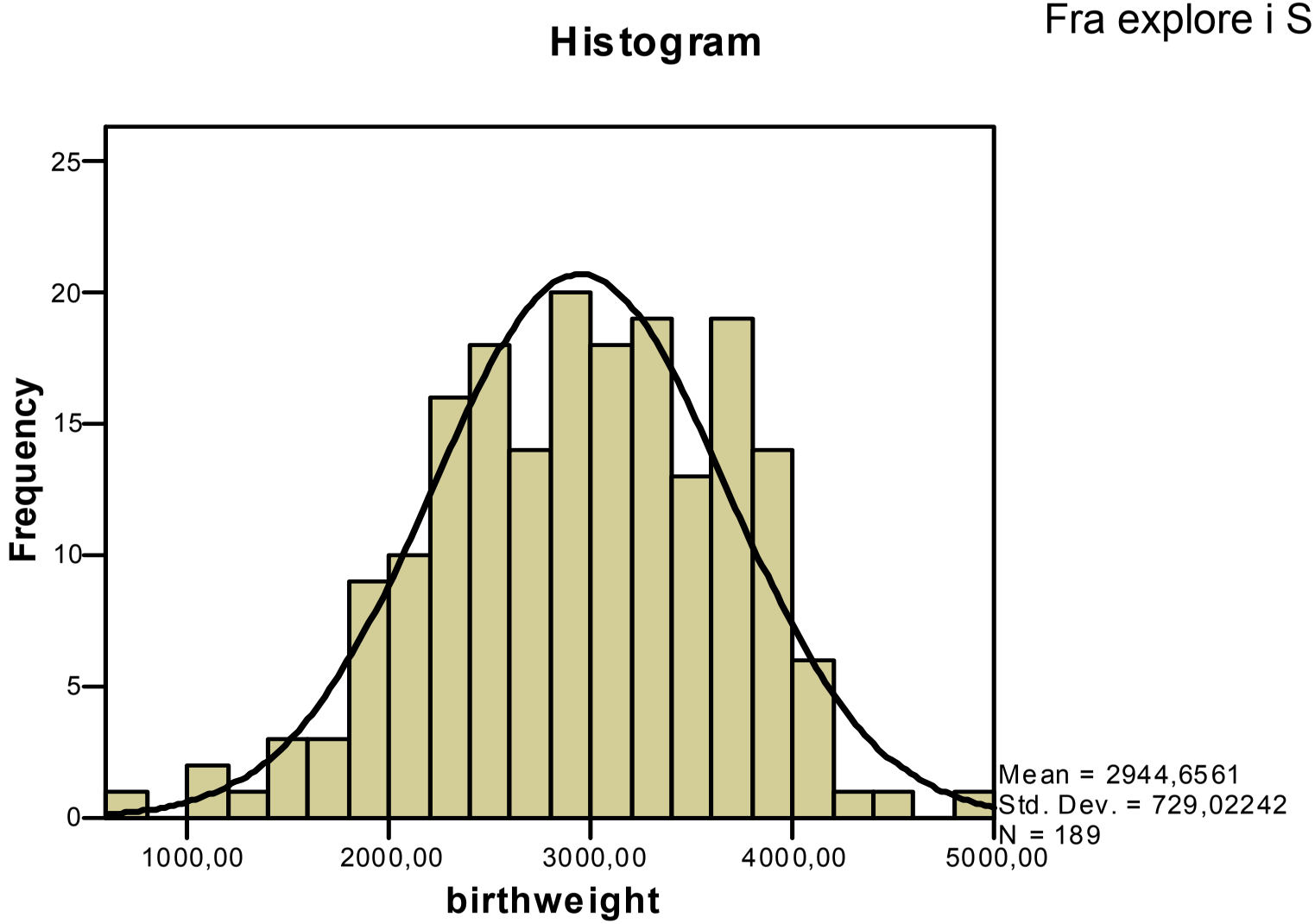
- Så på kontinuerlige utfall som var normalfordelte
- Brukte t-tester for å undersøke om det var signifikant forskjell mellom grupper
- Hvor stor sammenheng er det mellom to variabler? Korrelasjon
- Hva hvis man vil undersøke hvordan utfallet henger sammen med flere variabler? Lineær regresjon

# Data-eksempel i denne forelesningen:

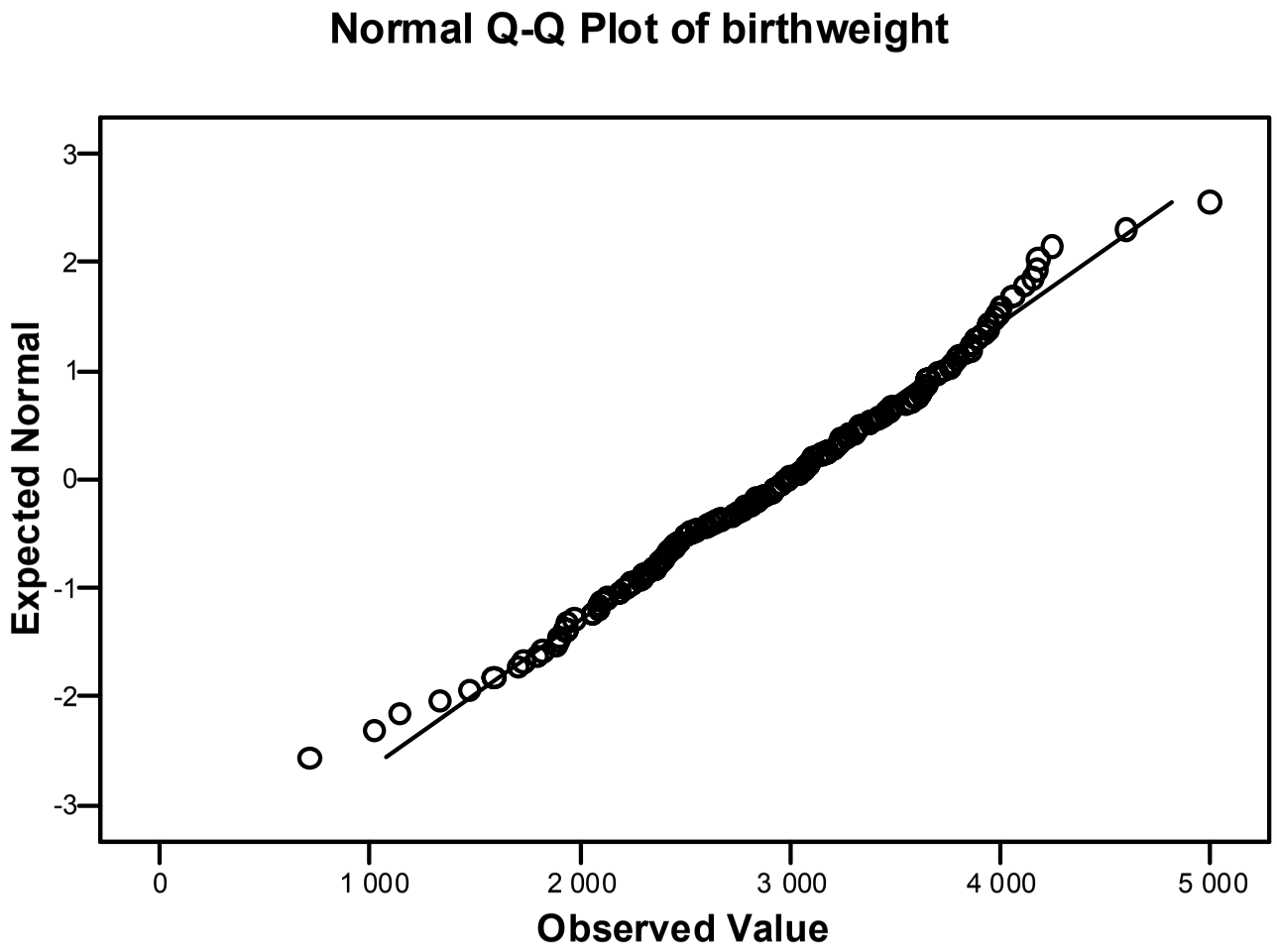
- Fødselsvekt og røyking, fra Massachusetts
- Barn av 189 kvinner
- Lav fødselsvekt er en risikofaktor
- Påvirker mors røykestatus fødselsvekten?
- Også sammenheng mellom fødselsvekt og andre forhold: Høyt blodtrykk, mors alder, mors vekt, rase osv.

# Er fødselsvekten normalfordelt?

Fra explore i SPSS



# Q-Q plott (kryss av for normality plots with tests under plots):



# Tester:

Nullhypotesen er at dataene er normalfordelt. Stor p-verdi indikerer normalfordeling. For store utvalg, vil p-verdien ofte være lav. Legg mer vekt på de grafiske metodene

**Tests of Normality**

| | Kolmogorov-Smirnov(a) | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| birthweight | ,043 | 189 | ,200(*) | ,992 | 189 | ,438 |

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

# Pearsons korrelasjonskoeffisient r

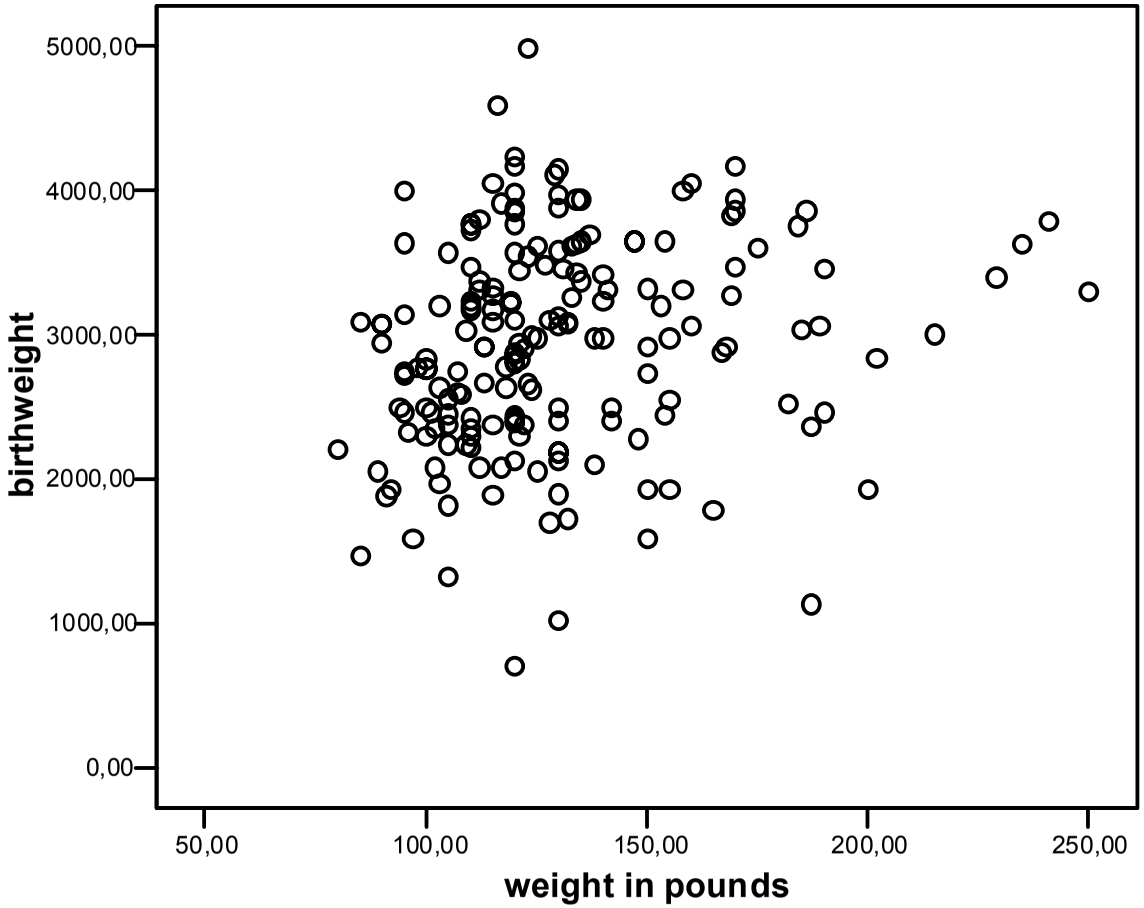
- Måler lineær sammenheng mellom to variabler
- r=1: Alle datapunkter ligger på en rett linje med positivt stigningstall
- r=-1: Alle datapunkter ligger på en rett linje med negativt stigningstall
- r=0: Ingen sammenheng
- Forklart varians $r^2$ måler effekten av å bruke regresjon
- $r^2$ nær 1 betyr at observasjonene ligger nær linjen, $r^2$ nær 0 betyr at punktene ligger langt fra linjen

# Pearsons korrelasjonskoeffisient i SPSS:

- Analyze->Correlate->bivariate

  Kryss av på Pearson
- Tester om r er forskjellig fra 0
- Nullhypotesen er at r=0
- Variablene må være normalfordelte
- Uavhengighet mellom observasjonene

# Eksempel:


5000,00
4000,00
3000,00
2000,00
1000,00
0,00
birthweight
50,00
100,00
150,00
200,00
250,00
weight in pounds

# Hvis man ikke har lineær sammenheng eller normalfordelte data: Spearmans korrelasjonskoeffisient, $r_s$

- Måler alle typer monotone sammenhenger, ikke bare lineære
- Ingen forutsetninger om fordeling
- $r_s$ er mellom -1 og 1, som Pearsons korrelasjon
- I SPSS: Analyze->Correlate.>bivariate

  Kryss av på Spearman
- Tester også her om $r_s$ er forskjellig fra 0

# Regresjonsanalyse

- Ønsker å tilpasse en linje som er så nær datapunktene som mulig
- Eksempelet: Fødselsvekt=konstant+B*mors vekt
- I SPSS: Analyze->regression->linear

  Utfallsvariabelen (her fødselsvekt) velges som dependent, og variablene vi tror kan predikere utfallet (her mors vekt) velges som independent
- Klikk på statistics og be om konfidensintervaller for regresjonskoeffisientene

# Sjekk av forutsetninger

- Residualene, dvs. avstanden fra hvert punkt til linja skal være normalfordelte
- I SPSS:
  - I linear regression klikk på statistics. Under residuals klikk på casewise diagnostics, og du får ”outliere” større enn 3 eller mindre enn -3.
  - I linear regression klikk også på Plots. Under standardized residuals plots merker du av for Histogram og Normal probability plot. Videre velg *Zresid som y-variabel og *Zpred som x-variabel

# Fortolkning:

- Har tilpasset linja

  Fødselsvekt=2369.672+4.429*mors vekt
- Hvis mors vekt øker med 20 pund, hvor mye øker barnets vekt (hvis man tror på modellen)?

  4.429*20=89 gram
- Hvilken vekt forventer man at et barn født av en kvinne som veier 150 pund har?

  2369.672+4.429*150=2591 gram

# Sammenheng mellom kontinuerlig utfall og flere variabler: Multippel lineær regresjon

- Fordelen med regresjon kontra korrelasjon er at man kan justere for flere andre variabler samtidig
- Legger inn mors røykestatus i tillegg i modellen.
- Tilpasser

  fødselsvekt=konstant+$B_1$*mors vekt+$B_2$*røykestatus
- I SPSS: Legger inn både mors vekt og røykestatus som independents

# Fortolkning:

- Har tilpasset linja

  Fødselsvekt=2500.174+4.238*mors vekt-270.013*røykestatus
- Hvis mor røyker (og vekten er konstant), hva skjer med barnets vekt (hvis man tror på modellen)?

  -270.013*1= -270 gram
- Hvilken vekt forventer man at et barn født av en kvinne som veier 150 pund og røyker har?

  2500.174+4.238*150-270.013*1=2866 gram

# Resultatpresentasjon

- Tittel
- Kort sammendrag
- Introduksjon (hvorfor gjorde du dette?)
- Metoder (hva har du gjort?)
- Resultater (hva fant du?)
- Diskusjon (hva betyr dette?)

  I en muntlig presentasjon kommer sammendraget i form av en kort oppsummering helt til slutt. Anbefaler også en setning om målet med prosjektet på lysarket etter tittelen.

# Metoder

- Beskriv klart hva som er gjort. Skal kunne forstås og repeteres av andre.
- Design
  - Observasjonelle studier (kontrollgruppe?, matching?, retrospektiv, tverrsnitt eller prospektiv?)
  - Kontrollerte kliniske studier (Def. Av behandlings-regimer, randomisering, blinding)

# Metoder, forts.

- Formål, hovedhypoteser
- Inklusjons- og eksklusjonskriterier: Antall forsøkspersoner, hvilke er ekskludert, hvilke er inkludert, og hvorfor
- Forsøkspersonene og hvordan de ble valgt
- Typen observasjoner, observasjonsmetodikk

# Statistiske metoder

- Hvilke metoder er brukt?
- I hvilke situasjoner er de ulike metodene brukt?
- Er noen av variablene kategorisert, eller noen kategorier kollapset?
- Signifikansnivå, ensidig eller tosidig testing?
- Statistikkprogram

# Resultater

- Deskriptiv beskrivelse av dataene (bakgrunnsvariabler: kjønn, alder osv. og hovedvariabler)
- Beskriv eventuelle avvik fra opprinnelig design (drop-out)
- Observasjonelle studier: Beskrive non-respondere? Representativitet?
- Tilfredsstiller dataene antagelsene som de statistiske analysene krever?
- Resultater av de statistiske analysene

# Diskusjon

- Pass på at resultatene tolkes riktig
- Tolk resultatene i lyset av funn fra tidligere studier
- Styrker/svakheter ved studien
  - Ved designet?
  - Ved gjennomføringen?
  - Har du gjort svært mange tester?
  - Styrkebetraktninger
- Generaliserbarhet?
- Kort oppsummering, videre arbeid

# Statistiske feil kan oppstå under:

- Planlegging
- Design
- Utføring
- Analyse
- Presentasjon
- Tolkning/forståelse

# Presentasjon av resultater

- I presentasjonen er det tre statistiske mål som er viktig å få frem:
  - Effektmålet (gjennomsnitt, relativ risiko, regresjonskoeffisient
  - Konfidensintervall for effektmålet
  - P-verdien for effektmålet

# Numerisk presisjon

- Rådata: Som regel nok med en eller ingen desimaler (46% kvinner, gj.snittlig vekt 65.5 kg)
- P-verdier: vanlig med 2 desimaler. Oppgi p-verdien, ikke $p>0.05$, $p<0.05$ eller p=NS
- P-verdier mindre enn 0.01: $p<0.01$
- P-verdier mindre enn 0.001: $p<0.001$
- Altman 16.3.5 og 8.10

# 1. Kontinuerlig utfallsvariabel

| Nr. | Beskrivelse | Navn |
|---|---|---|
| • 1 | Id-nummer | ID |
| • 2 | Lav fødselsvekt (1=BWT≤2500g, 0=BWT>2500g) | LOW |
| • 3 | Mors alder (år) | AGE |
| • 4 | Mors vekt (pund) | LWT |
| • 5 | Røykestatus (1=Røyker, 0=ikke-røyker) | SMK |
| • 6 | Hypertensjon (1=ja, 0=nei) | HT |
| • 7 | Rase (1=hvit, 2=svart, 3=annen) | RAC |
| • 8 | Fødselsvekt (g) | BWT |

# Deskriptiv presentasjon av data:

- Utvalgte variabeler:

| | Barnets vekt (g) | Mors alder (år) | Mors vekt (pund) |
|---|---|---|---|
| N | 189 | 189 | 189 |
| Gjennomsnitt (SD*) | 2944.7 (729.0) | 23.2 (5.3) | 129.8 (30.6) |
| Range | 709-4990 | 14-45 | 80-250 |

*SD=Standardavvik

- Skjevfordelte data:

| | Barnets vekt (g) | Mors alder (år) | Mors vekt (pund) |
|---|---|---|---|
| N | 189 | 189 | 189 |
| Median | 2977.0 | 23.0 | 121.0 |
| Q1-Q3* | 2412.0-3481.0 | 19.0-26.0 | 110.0-140.5 |

*Q1=1. kvartil og Q2=3. kvartil

# Presentasjon av analyseresultater:

- To-utvalgs t-test, barn av røykere og ikke-røykere:

| | Antall | Fødselsvekt (g)* | P-verdi |
|---|---|---|---|
| Ikke-røykere | 115 | 3055.0 (2916.0,3193.9) | |
| Røykere | 74 | 2773.1 (2620.3,2926.2) | 0.01 |

*Gjennomsnitt og 95% konfidensintervall i parentes

- Multippel lineær regresjon, fødselsvekt som utfall:

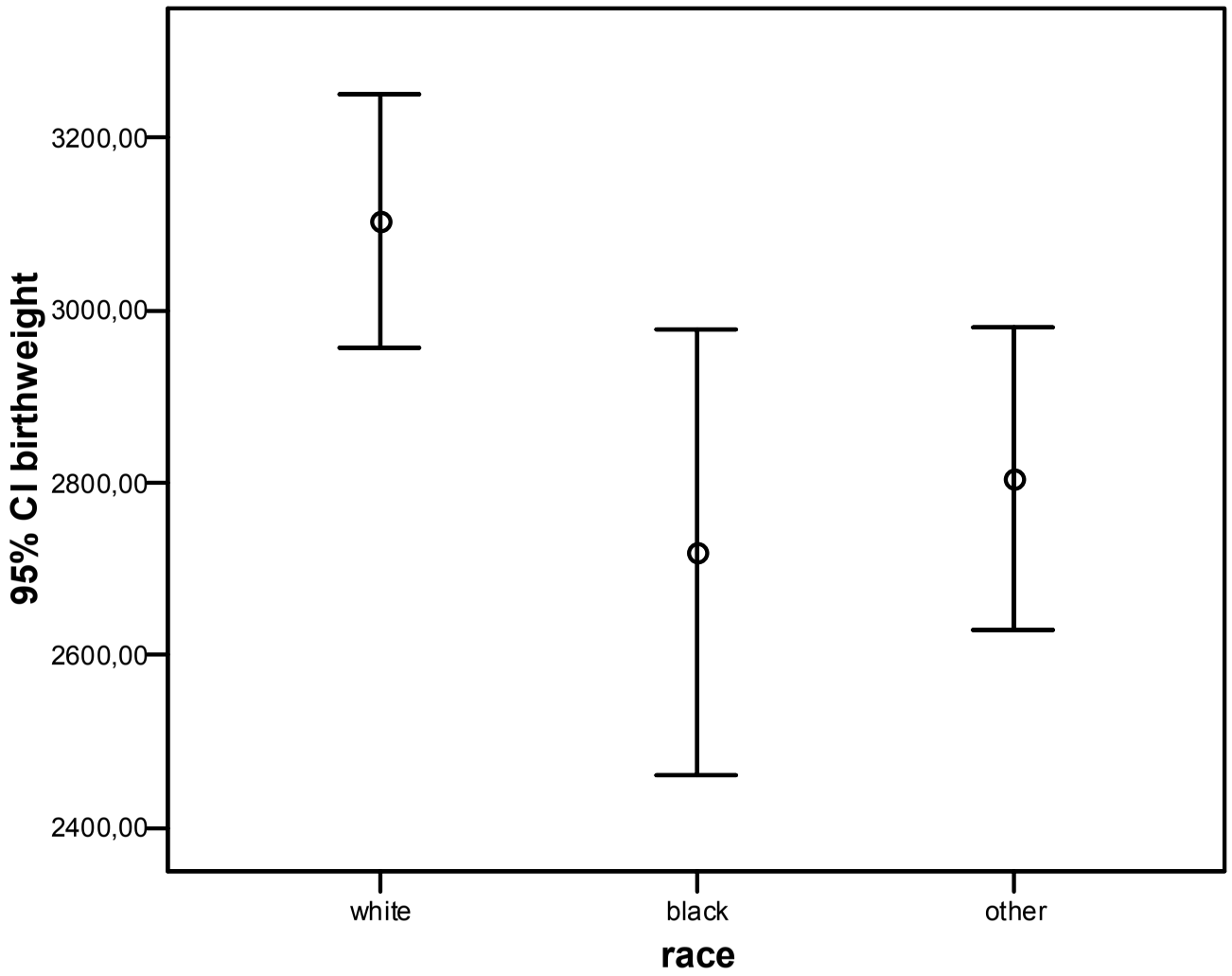
| Variabel | Ujustert effekt | 95% KI | P-verdi | Justert effekt | 95% KI | P-verdi |
|---|---|---|---|---|---|---|
| Mors vekt | 4.43 | (1.05,7.81) | 0.01 | 4.24 | (0.91,7.57) | 0.01 |
| Røykestatus* | -281,7 | (-492.7,-70.7) | 0.01 | -270.0 | (-478.3,-61.7) | 0.01 |

*Røykere i forhold til ikke-røykere

# Presentasjon av figurer

- Kan ta med et histogram for utfallsvariabelen, scatterplott som viser sammenhengen med andre kontinuerlige variabler osv.
- Analysert fødselsvekt og rase med enveis variansanalyse, signifikant forskjell (p=0.01). Figur som illustrerer dette på neste slide

  (I SPSS: Graph-error bar-simple, Variable: BWT, Category axis: RAC)

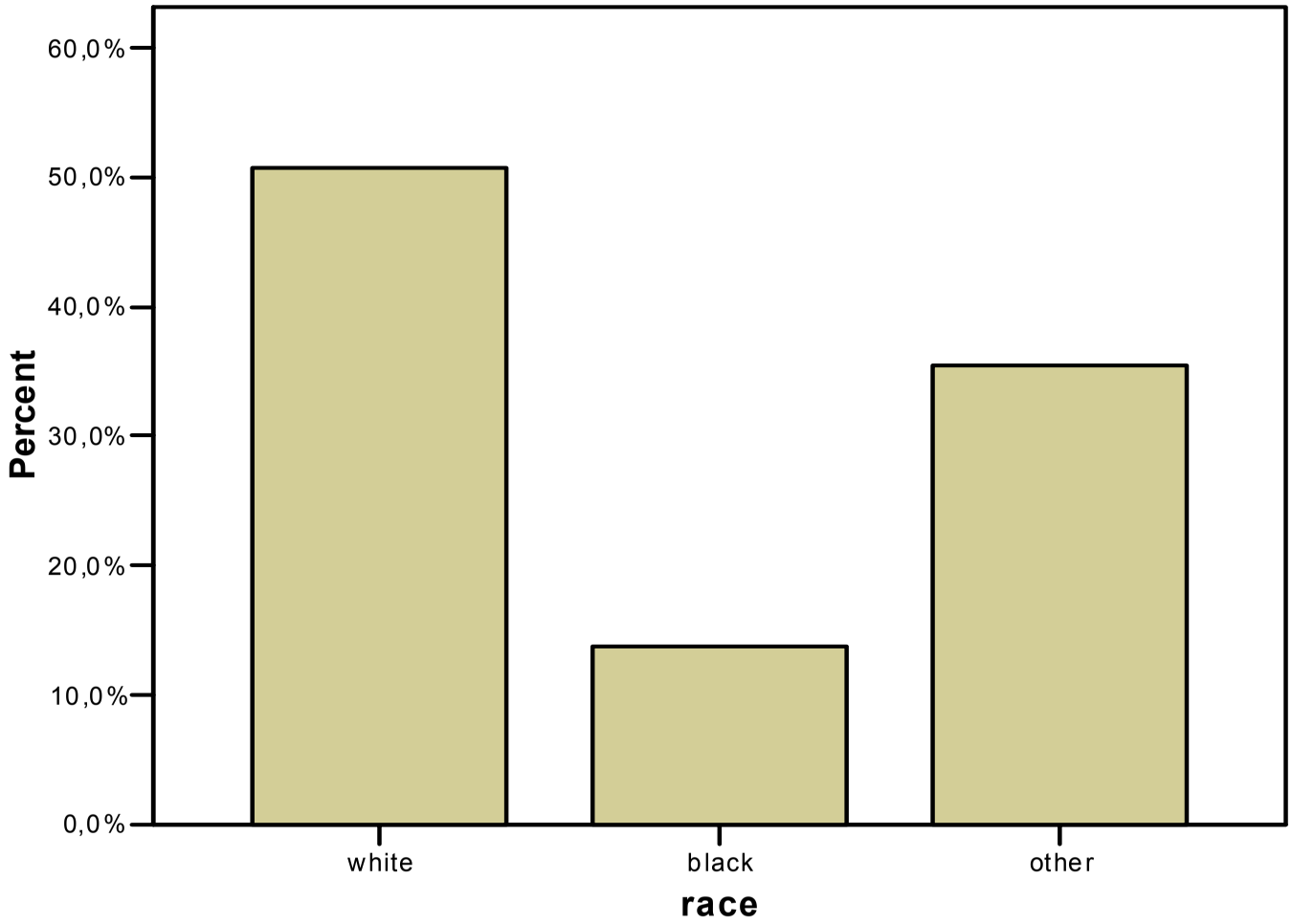

95% CI birthweight
3200,00
3000,00
2800,00
2600,00
2400,00
white
black
other
race

# 2. Kategorisk utfall

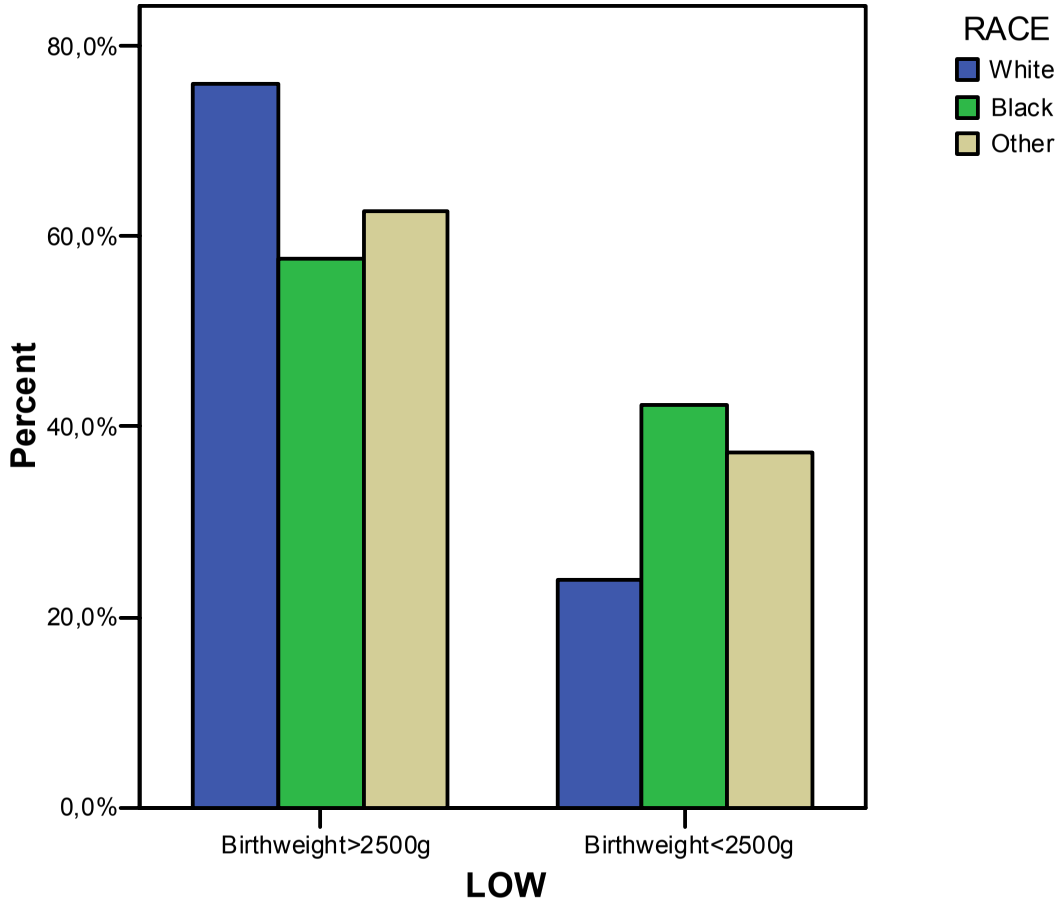
- Oppgi andelen i de forskjellige gruppene i en deskriptiv tabell
- Crosstabs: La utfallsvariabelen være radvariabel og kryss av på prosenter i columns i cell
- Tabeller for sammenhengen mellom lav fødselsvekt og rase, og lav fødselsvekt og røykestatus følger på neste slide

# Figurer:

- Prosentvis fordeling for rase (Graph->Bar->Simple. Bar represent: % of cases. Category axis: Race):



Signifikant forskjell på røykevaner mellom rasene (p<0.001, kji-kvadrat test). Figuren illustrerer dette:




Graph->bar
->clustered