6/11 En liten oppsummering før vi starter med kursets høydepunkt: REGRESJON.

Oppsummering av hovedlinjene i kurset så langt:

Deskriptiv statistikk (Ch 1)

Kategoriske data
(2 eller flere kategorier / grupper)

tabeller: Frekvenser (antall)
%

Kontinuerlige data
↓
histogram
(boksplott)

Symmetrisk

$\bar{X}$, SD

Basert på gjennomsnitt

skjevt

median, Q_1, Q_3

Basert på rangeringer

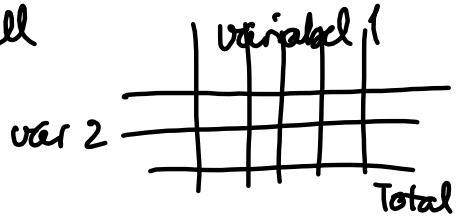
Bivariate sammenhenger (se neste side)

Neste: Regresjon

Bivariate sammenhenger: Sammenhenger mellom to ulike variabler

kategorisk mot kategorisk

Krysstabell



Ch 2.6

Hypotesetest for ikke pensum

H_0: Ingen sammenheng mellom de to variablene
Ingen forskjell på gruppene

Ch 8.2 og Ch 9

kategorisk mot kontinuerlig

2 kat
2 grupper

↓ (histogram) boksplott s. 97 Ch 1.3 "side by side boxplot"

Hypotesetest for H_0: ingen forskjell på gruppene (ingen sammenheng mellom variablene)

Basert på gjennomsnitt

to-utvalgs t-test
Ch 7.2

Wilcoxon Rank Sum test (Mann-Whitney-testen)
Ch 15.1

Basert på rangeringer av observasjoner

kategorisk mot kontinuerlig
>2 grupper

boxplot

Symmetrisk

Basert på gjennomsnitt

$\bar{x}$, SD i hver gruppe

skjevt

Basert på rangeringer

median/kvartiler i hver gruppe

ikke parvis

Hypotesetest for H_0: gruppene er like (null forskjell)

Ewens ANOVA

Kruskal-Wallis test

CLT etc : Teorien
Bygges opp omkring
normalfordeling og μ_1, μ_2

Parametriske tester
Parametriske metoder

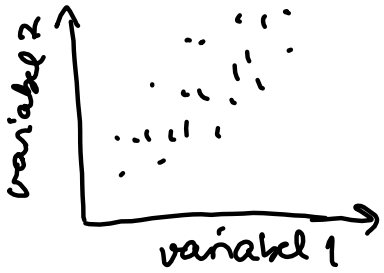
Basert på rangeringer,
ikke på normalfordeling og parametere
Kalles derfor

Ikke-parametriske tester

Ikke-parametriske metoder.

Kontinuerlig mot kontinuerlig

Scatter-plot

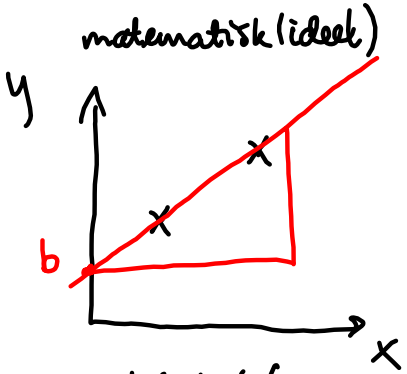


Oppsummerer sammenhengen mellom de to variablene med korrelasjonskoffisienten, r (Ch 2)

r er et tall mellom -1 og 1 Snl.no

HM: Hva med det vi lærte om at

$$y = ax + b$$



Stigningstallet a : Hvor mange enheter y endrer seg når x endrer seg en enhet

PUGG

Må estimere denne linja

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

Regresjonskoffisienten β_1 sier hvor mange enheter vi forventer at y skal endre seg når x endrer seg med en enhet

$\varepsilon_i \sim N(0, \sigma)$ uansett hva x er

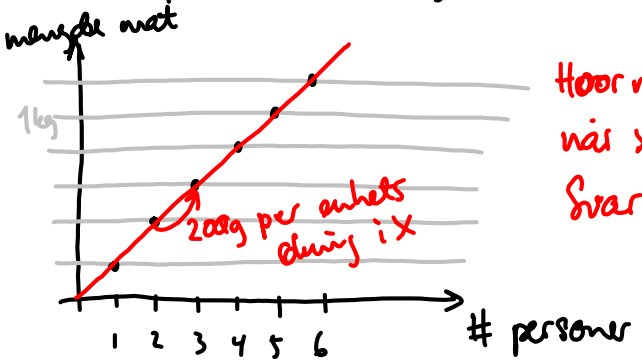
Lineær regresjon : introduksjon

funksjon $y = f(x)$: Det er en sammenheng mellom y og X

y = mengden mat (i gram) man lager til middag $= f($ # personer, aktivitet, hvor sulten, anledning ... $)$

Lineær funksjon (oppskrift)

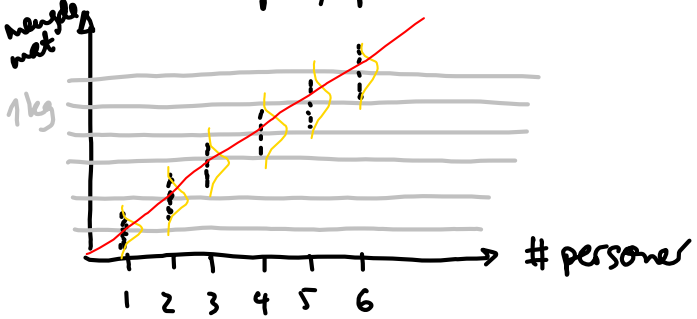
f.eks. kjøtt : ca 200g per person



Hvor mye øker y når x øker med én enhet?

Svar : 200g

Reell situasjon, fortsatt lineær sml :



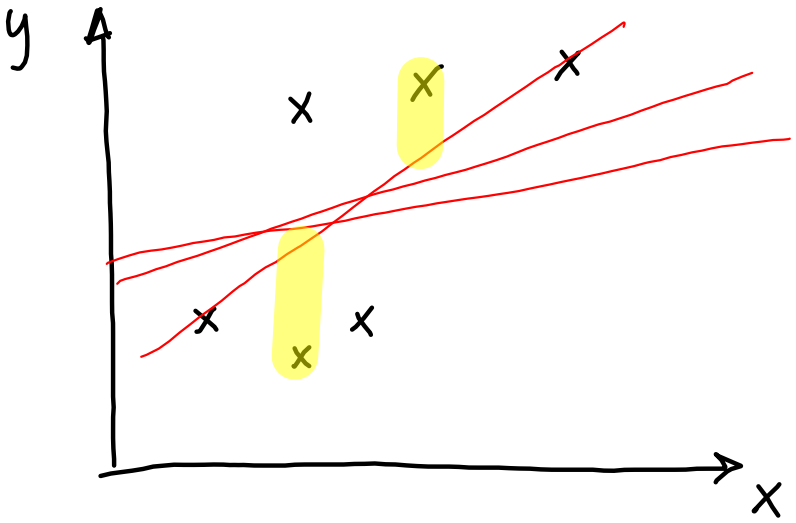
$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

$\beta_0 + \beta_1 \cdot x$: Hva vi forventer

$\varepsilon \sim N(0, \sigma)$ lik σ uansett hva x er.

ε: variasjonen

Estimering av regresjonslinja vha minste kvadraters metode
Method of least squares



Hvis $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,

så må vi finne de β_0 og β_1 som gjør at linja "passer best" til observasjonene.

I minste kvadraters metode betyr det å finne de β_0 og β_1 som minimerer den totale summerte kvadratavstanden til linja:

$$\sum_i \left(\underbrace{y_i - (\beta_0 + \beta_1 \cdot x_i)}_{\text{avstand fra linja}} \right)^2$$

de β_0 og β_1 som minimerer denne kvadratavstanden, kalles vi minste kvadraters estimater.

H_0: Ingen smh mellom x og y er det samme som $\beta_1 = 0$