University of Southern Denmark

Eksamensnoter

Statistik

Af:
Jes Grydholdt Jepsen
jejep12@student.sdu.dk

Dato: 26. januar 2015

Indhold

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

1 Emne 1 – Stokastisk variabel og fordeling (Random variable and distributions) PART 1

Sandsynlighed er også kaldet for chance eller risiko, for eksempel hvis man har en ordentligt bunke blandede kort, så kan der være 65% chance for at det går op og 35% risiko for at den ikke gør.

Et andet eksempel er, at med en seks–sidet terning er det 1/6 chance for at slå en sekser og at slå to seksere med to terninger er der 1/36 chance for at det sker. I denne beregning er der gjort en underforstået forudsætning, nemlig at hvert terningkast er **uafhængigt** af hinanden. I statistik kan sandsynlighederne ganges sammen, hvis de er uafhængige af hinanden.

Udfald (sample points) er en række alternative muligheder der eksisterer for den sandsynlighed man taler om. Alle udfald udgør **udfaldsrummet (sample space)**, betegnet U og delmængder af udfaldsrummet kaldes **hændelser (events)**.

For eksempel med en seks–sidet terning er der 6 mulige udfald, så udfaldsrummet vil være

$$U = [1,2,3,4,5,6]$$

og en delmængde kan være den hændelse A, at terningen viser et lige antal øjne, således

$$A = [2,4,6]$$

For hvert terningkast noteres et faktisk observeret udfald, for eksempel 3 øjne, og sådan observeret udfald kaldes **en observation**.

Antal øjne	Antal kast (absolutte hyppighed)	Relativt hyppighed (frekvens)
1	9	0,18
2	10	0,20
3	11	0,22
4	9	0,18
5	4	0,08
6	7	0,14
I alt	50	1,00

I eksemplet er der 50 **delforsøg** med en terning, hvor det totale antal betegnes n.

- Den absolutte hyppighed er er hvor mange gange hver hændelse forekommer
- Den relative hyppighed (frekvens) fåes ved at dividere den absolutte hyppighed med n

Når ordet hyppighed bruges er det altid den relative hyppighed der menes, og er vist ved symbolet

$$H[udfald/hændelser]$$

således at hvis udfaldet i forekommer a_i ud af n gange, så er hyppigheden

$$H[i] = \frac{a_i}{n}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

Så hyppighedsbegrebet fortæller hvor ofte en hændelse er observeret, i forhold til hvor mange gange den kunne være observeret.

Når man i en eller anden forbindelse siger noget om, hvor ofte man forventer en hændelse vil indtræffe i fremtiden, så knytter man et sandsynlighedstal til hændelsen. Sandsynlighed er det teoretiske modstykke til hyppighed. Sandsynligheden for hændelsen A betegnes med

$$P[A]$$

Sandsynlighed er en slags idealhyppighed, der fremkommer som grænseværdi ved en uendelig gentagelse af den pågældende type delforsøg. Både sandsynlighed og hyppighed er et tal mellem 0 og 1, så deres grænser ligger

$$0 \leq P[A] \leq 1$$

Forstået, at hvis en hændelse umuligt kan indtræffe, så er dens sandsynlighed 0 og omvendt 1, hvis en hændelse er sikker at forekomme hver gang.

- En hændelses hyppighed er summen af hyppighederne for de udfald, den omfatter. Da udfaldsrummet omfatter alle udfald, så derfor vil udfaldsrummet altid have hyppigheden og sandsynligheden på 1. Dette er en god kontrol til at tjekke om alle hyppigheder er fundet.

Eksemplet med terningen er en hyppighedsfordeling (observeret fordeling), fordi hyppigheden er angivet for hvert af de mulige udfald. Lader man i en observeret fordeling af uafhængige observationer gå imod uendelig, så går hyppigheden af det enkelte udfald imod sandsynligheden for dette udfald. Grænsefordelingen for den observerede fordeling kaldes en sandsynlighedsfordeling (eller teoretisk fordeling). En fordeling er således er til hvert udfald eller enhver hændelse et udfaldsrum der knytter dennes sandsynlighed eller hyppighed.

En observation kan være kvalitativ eller kvantitativ;

- **Kvalitativ**; med ord, for eksempel

$$U[mand,kvinde]$$

- **Kvantitativ**; med tal, lige som med terningekastet.

I det kvantitative tilfælde skelner man mellem, om udfaldsrummet i den tilsvarende teoretiske fordeling er **diskret** eller **kontinuert**

- **Diskret**; når der tælles ting, for eksempel elever i en klasse.
- **Kontinuert**; omfatter alle værdier i et vist interval, for eksempel en persons alder angivet i år (år er et interval af dage).

Observationer af kontinuert varierende egenskaber medfører, at udfaldsrummet for den teoretiske fordeling opdeles i grupper, hvor hver gruppe svarer til et udfald i den observerede fordeling. Klassegrænser er den grænse der er mellem to grupper i den teoretiske fordelings udfaldsrum.

Histogram og pindediagrammer; en fordeling kan præsenteres ved tegninger — en fremstillingsmåde, der for de fleste er mere anskuelig og lettere at huske i hovedtræk end andre måder. Eksemplet nedenfor viser, hvordan en fordeling lettere kan overskues ved brug af enten et histogram eller pindediagram, fremfor en tabel.

- Drejer det sig om observationer af en diskret egenskab, så er et pindediagram bedst at bruge til at repræsentere fordelingen. Hvis det derimod har en kontinuer egenskab, så er et histogram bedst at bruge.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

1.1 Kumuleret fordelingsfunktion

Også kaldet for **sumfunktionen**, F, og er defineret ved

$$F_{x_1} = P[x \leq x_1]$$

som for x_1 er den kumulerede sandsynlighed. En **sumkurve** kan grafisk vise dette (side 60). Sumkurven angiver, hvor stor en del af sandsynligheden, der ligger til venstre for x_1. Det kan ønskes at finde den stokastiske variabel der deler udfaldsrummet i to dele, hvoraf det til venstre har sandsynligheden y og det til højre har sandsynligheden $y - 1$.

1.1.1 Stokastisk variabel

Når der tales om sandsynlighedsfordeling, for eksempel med

- Et udfaldsrum med tilhørende sandsynligheder, for eksempel med terningerne der har et udfaldsrum

  [1,2,3,4,5,6]

  og med sandsynlighed for 1,2,3,4,5 og 6 i forbindelse med et bestemt terningkast.

Her er det naturligt at give størrelsen "antal øjne"et matematisk symbolnavn, x. En sådan størrelse, som kan antage værdier svarende til de forskellige udfald i udfaldsrummet, og som til hver af sine værdier har knyttet en sandsynlighed, kaldes **en stokastisk variabel**.

- En funktion af en stokastisk variabel er selv en stokastisk variabel, for eksempel er "kvadratet på antal øjne"en stokastisk variabel, som antager at 9 har samme sandsynlighed som 3.

I praksis er det umuligt at finde en sandsynlighedsfordeling ved en grænseovergang fra en hyppighedsfordeling, men derimod kan man idealisere et det med forholdsvis enkle egenskaber. Dette er også kaldet opstilling af en **statistisk model**.

Sumfunktionen er håndterlig ved kontinuerte og diskrete udfaldsrum. Det samme gælder også for intervalsandsynligheder, som der kan findes ved differenser mellem sumfunktionens værdier i intervallets endepunkter.

Sandsynlighedstæthedsfunktionen (probability density function) for den stokastiske variabel, x, er betegnet ved

$$p[x]$$

Denne funktion af x har den egenskab, at sandsynligheden for at observere en værdi i intervallet fra x_1 til x_2 afbilledes som arealet kurven, $y = p[x]$, og abscisseaksen (x–aksen). Udtrykket for dette er

$$Px_1 \leq x \leq x_2 = \int_{x_1}^{x_2} px \, dx = F(x_2) - F(x_1)$$

Sandsynligheden for at den stokastiske variabel ligger i udfaldsrummet, er 1, så er relationen

$$\int p[x] \, dx = 1$$

idet der integreres over hele udfaldsrummet. Ved diskrete fordelinger får relationen formen

$$\sum_i P[i] = 1$$

idet der summeres op over alle udfald i, som findes i udfaldsrummet.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

1.2 Statistisk population og stikprøve

Ved hjælp af begrebet statistisk population harman forsøgt at danne sig et overblik over, hvad tilfældig variation og sandsynlighed vil sige. For eksempel med en seks–sidet terning kan populationen af terningekast ses som en uendelig stor kasse med sedler, hvor der står antal øjne på. Almen er en population i statistisk forstand en uendelig stor mængde af abstrakte genstande, som er mærket med de udfald der kan forekomme. Hvis der skulle slåes 100 gange med en terning, så ville det svare til at der blev trukket 100 sedler op af kassen og noteret. I denne situation er der taget en **stikprøve** (sample) af størrelsen 100 fra den abstrakte population. Heri ligger ikke andet end, at der er udført 100 ens forsøg, som er stokastisk uafhængige. Et udfalds sandsynlighed er altså ifølge denne abstraktion dets andel af populationen, mens udfaldets observerede hyppighed er dets andel af den foreliggende stikprøve. Hvis der forekommer stikprøver, der ikke er repræsentative, så har man et **skævt (biased) udvalg**.

1.3 Fordelings karakteristika

I praksis kan det ofte være svært at overskue en detaljeret fordeling, hvad enten det er et histogram eller en teoretisk kurve, så derfor foretrækker man en resumerende beskrivelse ved hjælp af to eller tre karakteriserende talstørrelser. Det er nærtliggende at resumere en fordeling ved at angive udvalgte sumfunktioner, for eksempel ved at lave en sammenfatning ved hjælp af 10%–, 50%– og 90%–sumfunktioner.

Centrale mål. i anvendelse af statistik ønsker man næsten altid at vide, hvor store observationerne typisk er; også sagt med andre ord, hvor på tallinjen ligger fordelingens centrum?

Der er tre måder at beskrive det centrale mål på:

- **Medianen (median);** den midterste observation, 50%–sumfunktion eller den observation med sandsynligheden 0,5
- **Modus (mode);** den hyppigst forekommende (mest sandsynlige) udfald
- **Middel (mean);** rent formelmæssigt skelnes der mellem to middel–værdier; observeret og teoretisk middel.
  - **Observeret middel eller gennemsnit (average);** i en observeret fordeling defineres ved, at summen af observationernes afvigelse herfra, regnet med fortegn, er lig nul.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_1 = \frac{\text{Sum af observationer}}{\text{Antal observationer}}$$

  - **Teoretisk middel eller forventet værdi (Expectation, derfor $E[]$);** den forventede værdi af x og kan også anskues som det gennemsnitlige x i den population, hvorfra x er hentet. I en sandsynlighedsfordeling er den defineret ved, at de forventede observationers afvigelser herfra, regnet med fortegn, er lig nul.
    * En kontinuert sandsynlighedsfordeling

$$\mu = E[x] = \int x \cdot p[x]\, dx$$

hvor der integreres over alle mulige værdier af den stokastiske variabel x.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

* Diskret sandsynlighedsfordeling

$$\mu = E[x] = \sum x \cdot P[x]$$

hvor der summeres op over alle mulige værdier af den stokastiske variabel x.

- En normal fordeling er symetrisk omkring modus, men fordelingen kan også være højre– eller venstreskæve.

1.4 Variabilitetsmål

Man vil gerne kunne beskrive, hvor meget en fordeling breder sig ud over tallinjen;

- Differensen mellem 90%– og 10%–sumfunktionen, dvs. bredden af det interval, hvor de typiske 80% kan findes.
- Middelafvigelsen fra $E[x]$ til $\bar{x}$.
  - En kontinuert sandsynlighedsfordeling

$$\int |x - E[x]|p[x]\,dx$$

hvor der integreres over alle mulige værdier af den stokastiske variabel x.
  - En diskret sandsynlighedsfordeling

$$\sum |x - E[x]|P[x]$$

hvor der summeres over alle mulige værdier af den stokastiske variabel x.
  - En observeret fordeling er variabilitetsmålet

$$\sum |x - \bar{x}|H[x]$$

hvor der summeres op over alle de x, der er observeret. Hx er hyppigheden for hver observeret x.

Det mest almindelige variabilitetsmål er **spredning (standardafvigelse, standard deviation)**. Standardafvigelsen er kvadratroden af **variansen (variance)**, som er defineret som middelværdien af de kvadrede afvigelser fra fordelingens middel. **Variansen** på x er dermed

- Kontinuert

$$\sigma^2 = Var[x] = \int (x - E[x])^2 p[x]\,dx$$

- Diskret

$$\sigma^2 = Var[x] = \sum (x - E[x])^2 P[x]$$

og **spredningen** er

$$\sigma[x] = \sqrt{Var[x]}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

For observerede fordelinger fås på tilsvarende måde den gennemsnitlige kvadratiske afvigelse fra $\bar{x}$ således

$$\sigma^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

og kan udtrykkes

$$\sum (x - \bar{x})^2 H[x]$$

hvor der summeres op over de x, der er observeret. Den **observerede varians** betegnes s^2 og er givet ved

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

og den **observerede spredning**, s, er naturligvis kvadratroden af den observerede varians.

1.5 Fordeling

Der er forskellige fordelingstyper, og fordelinger indenfor samme type baseres på en fælles tankekonstruktion; de adskiller sig blot ved, at en eller flere talstørrelser, der indgår i den fælles tankekonstruktion, har forskellige parametre. **Parametre** er dermed variable, hvis enkelte værdier udpeger en fordeling ud af en stor mængde fordeling med visse fælles egenskaber. Der er tre fordelingstyper;

1.5.1 Normalfordeling

Normalfordelingen, kaldet $N(m,s)$, er beskrevet ved den velkendte klokkeformede kurve. Parameteren m er middelværdien for fordelingen, og kurven er fuldstændig symmetrisk herom. Den anden parameter s er spredningen. Til beregninger af sandsynligheder i normalfordelingen refereres altid til standard normalfordelingen, der har middelværdi 0 og spredning 1. Man betegner en standard normalfordelt stokastiske variabel med bogstavet Z. Det skrives

$$Z \sim N(0,1)$$

Til at beregne $P(Z < z)$, kan man bruge en z–tabel over $N(0,1)$. Man kan omvendt finde værdien af z ud fra en kendt sandsynlighed.

For at omregne fra en vilkårlig normalfordeling,

$$X \sim N(m, s)$$

til en standard normalfordeling, benytter man formlen

$$Z = \frac{X - m}{s}$$

1.5.2 Binomialfordeling

En binomial stokastisk variabel beskriver antallet af gange, en bestemt hændelse forekommer i en serie af n eksperimenter, hvor der hver gang kun er to mulige udfald. Traditionelt kalder man i hvert eksperiment den søgte hændelse for succes og den anden hændelse for fiasko. Der skal gælde følgende 4 betingelser

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

1. De n eksperimenter er identiske.
2. Udfaldet i ethvert af eksperimenterne er uafhængigt af udfaldet i ethvert af de andre eksperimenter.
3. Der er kun to mulige udfald af hvert eksperiment.
4. Sandsynligheden for succes, p, er den samme for alle eksperimenter.

Der er to parametre, n og p i denne fordeling. Man skriver

$$X \sim Bin(n,p)$$

Sandsynligheden for en enkelt hændelse er givet ved

$$P(X = x) = \binom{n}{x} p_x(1 - p^{n-x})$$

I en binomialfordeling med parametrene n og p gælder der, at middelværdien

$$m = np$$

og spredningen

$$s = np(1 - p)$$

Hvis n er passende stor (f.eks. $n > 20$) kan man approximere binomialfordelingen med en normalfordelingen ved at bruge

$$m = np$$

og

$$s^2 = np(1 - p)$$

1.5.3 Poissonfordeling

Yderlig er der tre fordelingstyper, som spiller en rolle i analyse af observationer.

1.5.4 Student's t–fordeling

Den centrale grænseværdisætning siger, at gennemsnittet i en stikprøve tilnærmelsesvis er normalfordelt med middelværdi μ og spredning $\sigma/\sqrt{n}$. Dette udtryk indeholder 2 ubekendte, nemlig m og s. I praksis er man nødt til at estimere σ med s.

Dette estimat giver en ekstra usikkerhed, så man har ikke længere en normalfordeling men derimod en t-fordeling, der er bredere og lavere end normalfordelingen. Jo mindre n er, jo færre observationer er der i stikprøven, jo bredere er t-fordelingen. Fordelingens frihedsgrader er $n - 1$. Den stokastiske variabel er givet ved

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Sandsynlighederne i en t–fordeling kan findes, ligesom med normalfordeling, i en t–tabel.

1.5.5 χ^2–fordeling

1.5.6 F–fordeling

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

2 Emne 1 – Stokastisk variabel og fordeling (Random variable and distributions) PART 2

En stokastisk hændelse er en variabel, hvis observerede værdi kan ses som et udfald af et stokastisk eksperiment. En stokastisk variabel er en funktion af en stokastisk hændelse. D.v.s. at en stokastisk variabel altid er et tal. Man betegner en stokastisk variabel mod store bogstaver.

Hændelse	**Type af data**	**Stokastisk variabel**
En persons køn	Bogstaver eller nominaltal	nominaltal
Antal ankomster til en skadestue	Diskrete data/ordinaltal	diskrete tal
En løveunges vægt	Kontinuert data/ intervaltal	kontinuerte tal

De frekvenser, der er knyttet til alle mulige værdier af den stokastiske variabel i populationen, kaldes sandsynlighedsfordelingen af den stokastiske variabel.

2.1 Diskrete

Diskrete stokastiske variable har et tælleligt antal udfald. Her kan man definere en sandsynlighed for en enkelt hændelse. Dette kan skrives P(X=a), der læses "sandsynligheden for at den stokastiske variabel X antager værdien a". Hver sandsynlighed kan angives ved en tabel, et matematisk udtryk eller et stolpediagram.

Kaldes sandsynlighederne for de enkelte hændelser pi gælder der at

1. $0 \leq p_i \leq 1$
2. $\sum_{i=1}^{n} p_i = 1$

2.2 Kontinuerte

Kontinuerte stokastiske variable har fordelinger, der angives ved et matematisk udtryk og en tilhørende kurve. Her findes sandsynligheden for, at variablen antager en værdi i et interval, som arealet under kurven indenfor dette interval. Det skrives

$$P(a \leq f(x) \leq b)$$

Sandsynligheden for, at den stokastiske variabel antager én bestemt værdi, er 0.

Kaldes fordelingsfunktionen f(x), gælder der at

1. $0 \leq f(x) \leq 1$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

3 Emne 2 – Parameter estimering og confidence intervals

3.1 Central Limit Theorem og fordeling af teststørrelsen

En af de mest fundamentale koncepter i statistik, og den siger: Hvis man fra en population med en vilkårlig fordeling tager stikprøver af størrelsen n, så vil gennemsnittene være tilnærmelsesvis normalfordelte. Jo større n er, jo nærmere kommer man en normalfordeling. Desuden vil variansen af denne normalfordeling falde, jo større n er. Kan bruges på alle fordelinger, med et defineret gennemsnit og varians. Kan både være en kontinuer eller diskret fordeling. Eksemplet tager udgangspunkt i en diskret fordeling, som vist på figuren, og der prøves af med forskellige stikprøve størrelser, for eksempel

$$n = 4$$

så en stikprøve kan se ud sådan

$$s_1 = [1,1,3,6]$$

og gennemsnittet for denne stikprøve er

$$\bar{x_1} = 2{,}75$$

Tager vi to stikprøve mere

$$s_2 = [3,4,3,1]$$

og

$$s_3 = [1,1,6,6]$$

med gennemsnittet

$$\bar{x_2} = 2{,}75$$

og

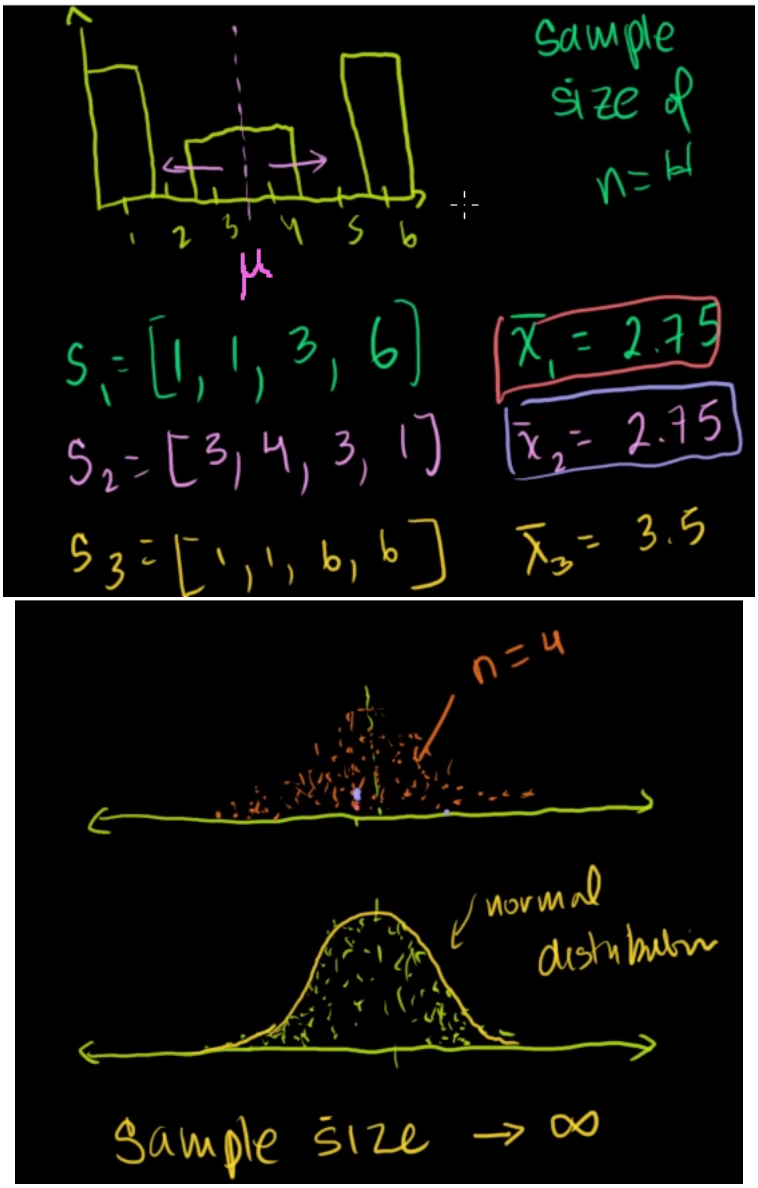
$$\bar{x_3} = 3{,}5$$

Hvis vi fortsætter med at tage disse stikprøver (for eksempel 10.000 stikprøver), regne deres gennemsnit ud og plotte dem, så vil vi få noget der tilnærmelsesvis ligner en normalfordeling. Større stikprøve størrelse, jo bedre tilnærmelse får man af en normalfordeling. Gennemsnittet vil være det samme for de forskellige stikprøve størrelser, men standard afvigelsen vil blive mindre, jo større stikprøve størrelse man bruger. Med en stikprøve størrelse

$$s \longrightarrow \infty$$

vil man få en perfekt normalfordeling.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK



Alt dette er kaldet for **fordeling af teststørrelsen**. Teststørrelsen fordeling har det samme gennemsnit, som den originale fordeling.

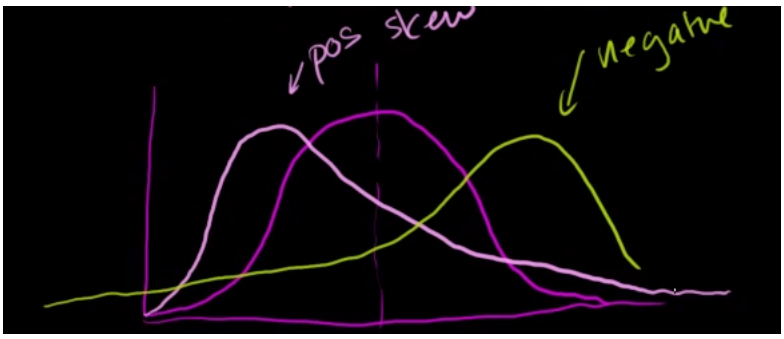
- Der bliver taget n stikprøver af sandsynlighedsfordelingen
- Gennemsnittet af stikprøverne bliver plottet og danner en normalfordeling

En ideal normalfordeling har lige store arealer på højre– og venstresiden af gennemsnittet, men hvis

- der er mest på højresiden, så er den positiv skæv
- der er mest på venstresiden, så er den negativ skæv

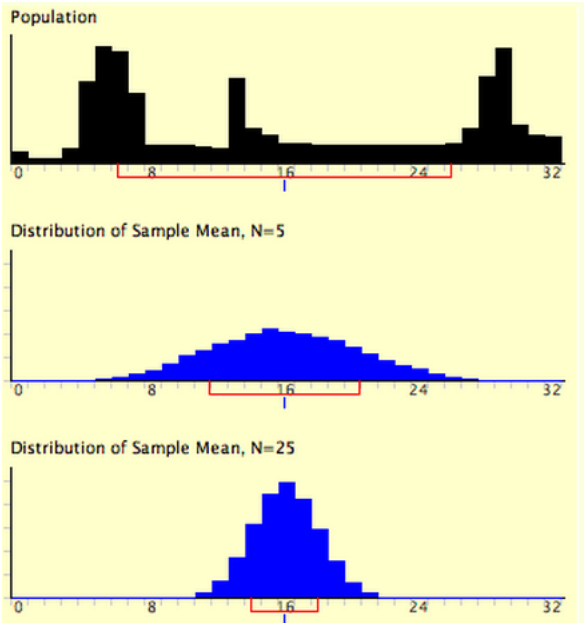
En ideal normalfordeling er ikke skæv.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK



Kurtosis: positiv kurtosis, så er der federe tail og mere spids peak.

Figuren nedenfor viser, hvordan man med større stikprøve størrelse, får en normalfordeling med en skævhed og kurtosis tættere på nul.



FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

3.2 Standard Error for gennemsnittet

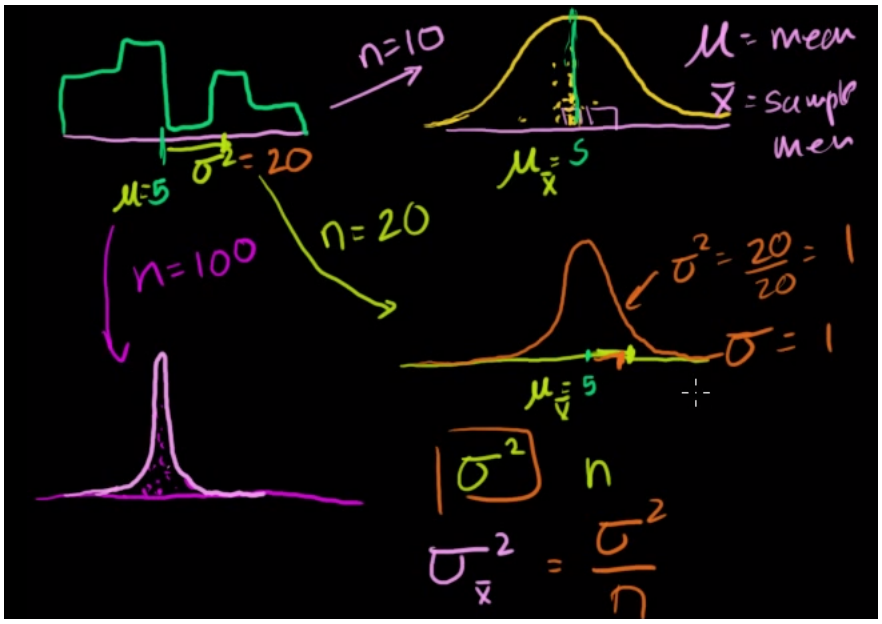
Hvis man kender standard afvigelsen (eller variansen) og stikprøvestørrelsen, så kan standard afvigelsen for hver fordeling af teststørrelsen tilnærmelsesvis findes. Variansen for teststørrelse fordelingen er

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

og standard afvigelsen er

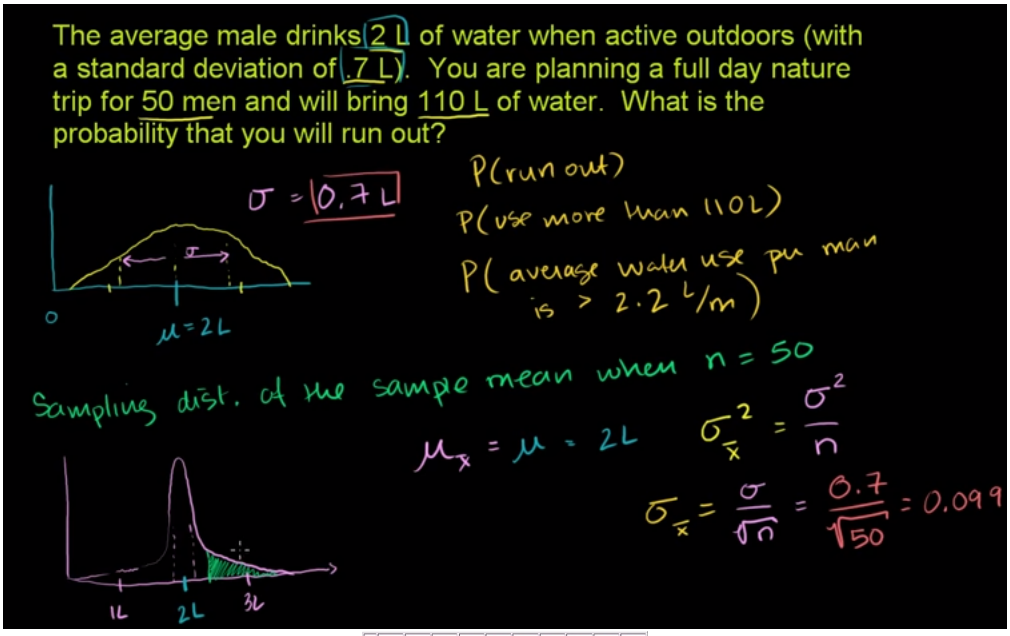
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Eksemplet nedenfor viser forskellige fordelinger af teststørrelser, og ud fra formlen kan det ses, at jo højere stikprøvestørrelsen er, jo lavere standard afvigelsen/varians vil man have i fordelingerne af teststørrelserne.

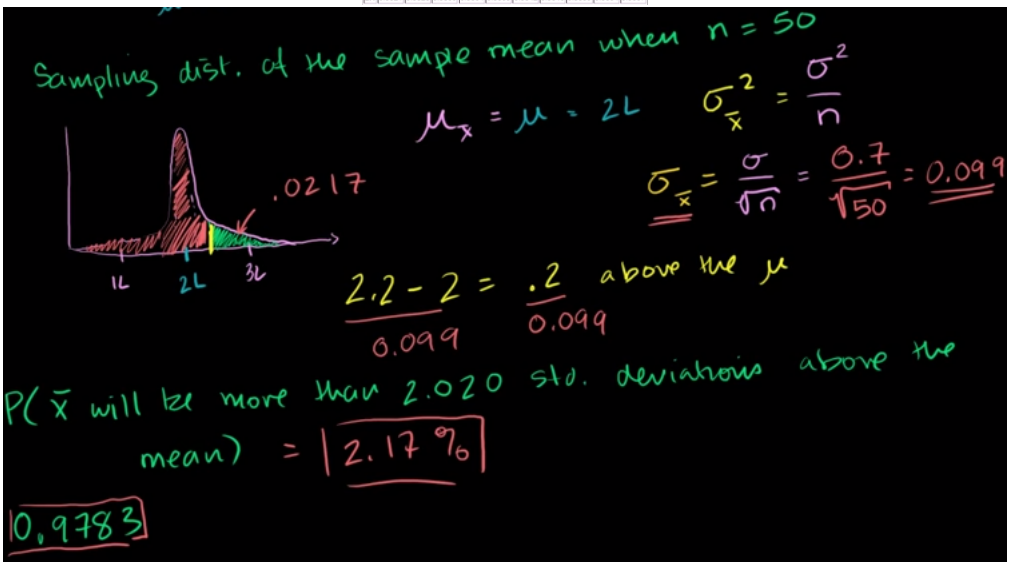


FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

3.3 Eksempel på fordeling af teststørrelsen



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857



FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

3.4 Interval estimation (Confidence Intervals) (Ch.8)

Den centrale grænseværdi; Hvis man fra en population med en vilkårlig fordeling tager stikprøver af størrelsen n, så vil gennemsnittene være tilnærmelsesvis normalfordelte. Jo større n er, jo nærmere kommer man en normalfordeling. Desuden vil variansen af denne normalfordeling falde, jo større n er,

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Hvis man tager stikprøver fra en standard normalfordeling N(0,1), ved man, at der Eksempel: der bliver taget en stikprøve på 36 æbler fra en høst på over 200.000 æbler. Gennemsnitsvægten er stikprøven er 112 gram (med en standard afvigelsen på 40 gram). Hvad er sandsynligheden for at gennemsnitsvægten for alle æbler ligger enden for 100 og 124 gram?

Vores population har en fordeling, med et gennemsnit, μ, og en standard afvigelse, σ. Ud fra vores stikprøve vil vi få en fordeling af stikprøverne, der vil har det samme gennemsnit, μ_x, og en standard afvigelse næsten lig med populationens standard afvigelse divideret med kvadratroden af vores antal stikprøver, således

$$\mu_{\bar{x}} = \mu$$

og

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

hvor $n = 36$. Men vi kender endnu ikke μ.

Der bliver spurgt om, hvad sandsynligheden er, at gennemsnitsvægten ligger inden for ± 12 gram, således

$$P(\mu \text{ er indenfor 12 gram af vores } \bar{x})$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

Dette kan også skrives som

$$P(\bar{x} \text{ er indenfor 12 gram af vores } \mu)$$

Ved at formulere det sådan, så kan vores stikprøve fordeling bruges. Vi har et ukendt gennemsnit i stikprøve fordelingen, som er den sammen som populationens gennemsnit, og fordi de to er ens, så kan vi sige

$$P(\bar{x} \text{ er indenfor 12 gram af vores } \mu_{\bar{x}})$$

så hvis man kan finde ud af hvor mange standard afvigelser det er væk fra stikprøvens gennemsnit, $\mu_{\bar{x}}$, så kan z–tabellen bruges til at finde ud af hvad sandsynligheden er. Der er et problem, for vi ved ikke hvad den faktiske standard afvigelse for fordelingen er, vi ved kun at det er

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

så vi laver en tilnærmelse af σ, for at løse problemet, således at vores standard afvigelse for stikprøverne er tilnærmelsesvis lig med populationens standard afvigelse

$$\sigma \approx s = 40\, gram$$

og nu kan vi regne sandsynligheden ud, ved først at regne standard afvigelsen ud

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{36}} = \frac{\sigma}{6} \approx \frac{s}{6} = \frac{40}{6} = 6{,}67$$

så vores standard afvigelse for stikprøve fordelingen er 6,67 og for at finde sandsynligheden for at alle æblers vægt ligger mellem 100 og 124 gram, så divideres de 12 gram med 6,67, således

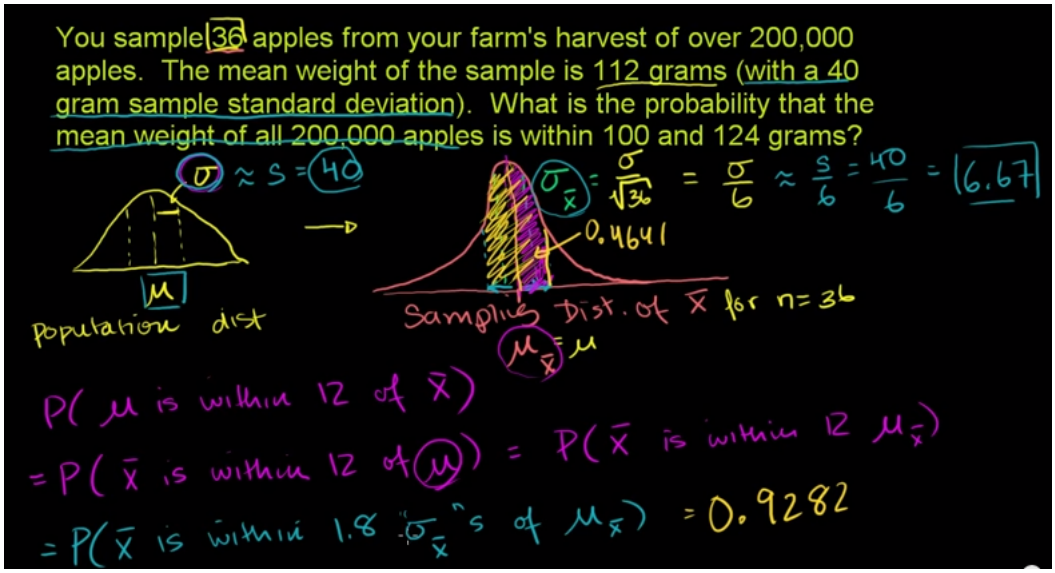
$$12/6{,}67 = 1{,}8$$

så på vores stikprøve fordeling er det 1,8 $\sigma_{\bar{x}}$ væk for gennemsnittet. Ved at kigge i z–tabellen, så kan vi finde værdien 0.9641, men der skal trækkes 0.5 fra, for at få kun den ene halvdel med. Fordi det er en normal fordeling, så er den symetrisk omkring gennemsnittet, μ, og vi kan gange den med to, for at få vores sandsynlighed, således

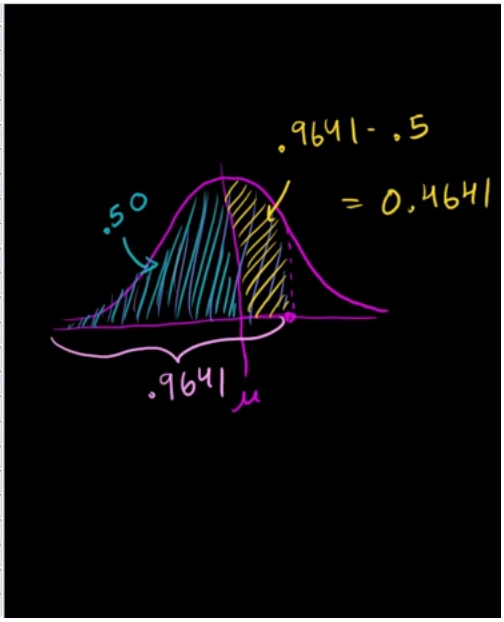
$$(0.9641 - 0.5) \cdot 2 = 0.9282$$

så der er 92.82% chance for at den rent faktiske gennemsnit ligger indenfor 100 og 124 gram.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916



FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

4 Emne 3 – Hypotesetestning

4.1 En population

4.1.1 En– og tosidede hypotesetests

Når man slår plat og krone, forventer man at få lige mange af hver. Man har altså en hypotese, som man ikke udtrykker klart, men som kan formuleres således: P(plat) = P(krone). Erfaring siger, at man ikke kan forvente præcis lige mange af hver. Der er en vis variation, og engang imellem kan man være meget uheldig, og andre gange kan man være meget heldig. Spørgsmålet er så, hvor uheldige man skal være, før man kan tillade sig at sætte spørgsmålstegn ved møntens ægthed. D.v.s, hvor skævt skal resultatet være, før man forkaster hypotesen og vælger en alternativ hypotese.

Hvis man skal indføre en ny metode eller behandling, er man i reglen kun interesseret i, om den nye metode er bedre end den gamle. Hvis den nye er dårligere end den gamle, vil man selvfølgelig ikke have den, og hvis den kun er lige så god som den gamle, kan det nok ikke betale sig at indføre en ny metode. Man vil altså teste ensidigt:

$$\begin{aligned} H_0 &: \quad \text{Effekt af ny} = \text{Effekt af gammel} \\ H_1 &: \quad \text{Effekt af ny} > \text{Effekt af gammel} \end{aligned}$$

Hvis man derimod ønsker at teste f.eks. skoleformens indflydelse på rygning blandt skoleelever, kan man argumentere for, at skoleformerne kan virke begge veje, således at hypoteserne bør være:

$$\begin{aligned} H_0 &: \quad \text{Rygning blandt folkeskoleelever} = \text{Rygning blandt efterskoleelever} \\ H_1 &: \quad \text{Rygning blandt folkeskoleelever} > \text{Rygning blandt efterskoleelever} \end{aligned}$$

Den tosidede test giver en dobbelt så stor P-værdi som den ensidede test. P beregnes som sandsynligheden for at få noget, der er mere ekstremt end den fundne værdi. I den tosidige test er ekstreme værdier både større og mindre værdier. I den ensidige test ovenfor, er ekstreme værdier kun de store værdier for den ny effekt. Hvis man får en meget mindre effekt af den ny metode, skal man helt undlade at udføre testen. Hvis man gør det, vil man få en meget stor P-værdi.

De mest enkle hypoteser sammenligner en eller anden parameter med et fast tal. Hvis man f.eks. laver lineær regression, tester man, om der er en stigende eller faldende sammenhæng mellem de to variable, og hypoteserne er:

$$\begin{aligned} H_0 &: \quad \text{Hældning} = 0 \\ H_1 &: \quad \text{Hældning} \neq 0 \end{aligned}$$

Teststørrelsen er

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Hvis P–værdien er under 0,05, så afvises H_0.

4.1.2 Type I og II fejl

Når man laver en hypotesetest, kan man opskrive følgende tabel:

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

	H_0 er sand	H_0 er ikke sand
H_0 forkastes	Type I fejl	OK
H_0 forkastes ikke	OK	Type II fejl

Når man tester en hypotese, vil der altid være en sandsynlighed for, at man forkaster hypotesen, selv om den er sand. Dette kaldes en Type I fejl, eller a i tabellen forneden. Sandsynligheden for en Type I fejl er den, man vælger som signifikantsniveau, altså a. Sandsynligheden for en Type II fejl kaldes b. Styrken defineres som 1–b, og er sandsynligheden for, at H0 forkastes, når den ikke er sand.

	H_0 er sand	H_0 er ikke sand
H_0 forkastes	α	$1-\beta$
H_0 forkastes ikke	$1-\alpha$	β

4.1.3 Tosidig hypotese om gennemsnittet

Når man har en enkelt population, og ønsker at opstille en hypotese om middelværdien, skal man sammenligne middelværdien med en fast værdi, kaldet m0. Man kan skrive sin nulhypotese og alternative hypotese således:

$$\begin{aligned} H_0 &: \quad \mu = m_0 \\ H_1 &: \quad \mu \neq m_0 \end{aligned}$$

Konfidensintervaller omkring gennemsnittet: Når man laver et tosidig t-test med signifikansniveau α, kan man, i stedet for at se på P–værdien, se på det interval, hvor H_0 ikke forkastes. Ud fra t–værdien beregnes et

4.1.4 Ensidig hypotese om gennemsnittet

Hvis man har en fast overbevisning om, at det, man undersøger, har et gennemsnit, der er enten større eller mindre end det faste tal, kan man lave en ensidig test.

$$\begin{aligned} H_0 &: \quad \mu = m_0 \\ H_1 &: \quad \mu < m_0 \end{aligned}$$

eller

$$\begin{aligned} H_0 &: \quad \mu = m_0 \\ H_1 &: \quad \mu > m_0 \end{aligned}$$

Som eksempel, kunne det være at undersøge om der er mere fedt end angivet i en bestemt slags chips? Hypoteserne, der skal testes er nu

$$\begin{aligned} H_0 &: \quad \mu = 34g \\ H_1 &: \quad \mu > 34g \end{aligned}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

4.2 Eksempel med en population

4.3 Tosidede hypotesetests

Først opstilles nul–hypotesen

$$H_0: \qquad \text{Ingen effekt}$$

med andre ord, så vil gennemsnittet forblive det samme, dvs

$$\mu = 1.2s$$

Den alternative hypotese er så

$$H_1: \qquad \text{Har en effekt}$$

så det vil sige

$$\mu \neq 1.2s$$

Først starter vi med at antage at H_0 er sandt, og ved at regne sandsynligheden for det, så kan det vurderes om det er sandt eller ej. Hvis sandsynligheden er lavere end 0.05, så antager man at det ikke er sandt.

For at teste dette, så laver vi en stikprøve fordeling af det data vi har. Vi får så en normalfordeling og vi kan derfor antage at

$$\mu_{\bar{x}} = \mu = 1.2$$

Standard afvigelsen er så

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{100}} = 0.05$$

så

$$\hat{\sigma x} = 0.05$$

(Hat for at vise det er en approksimation)

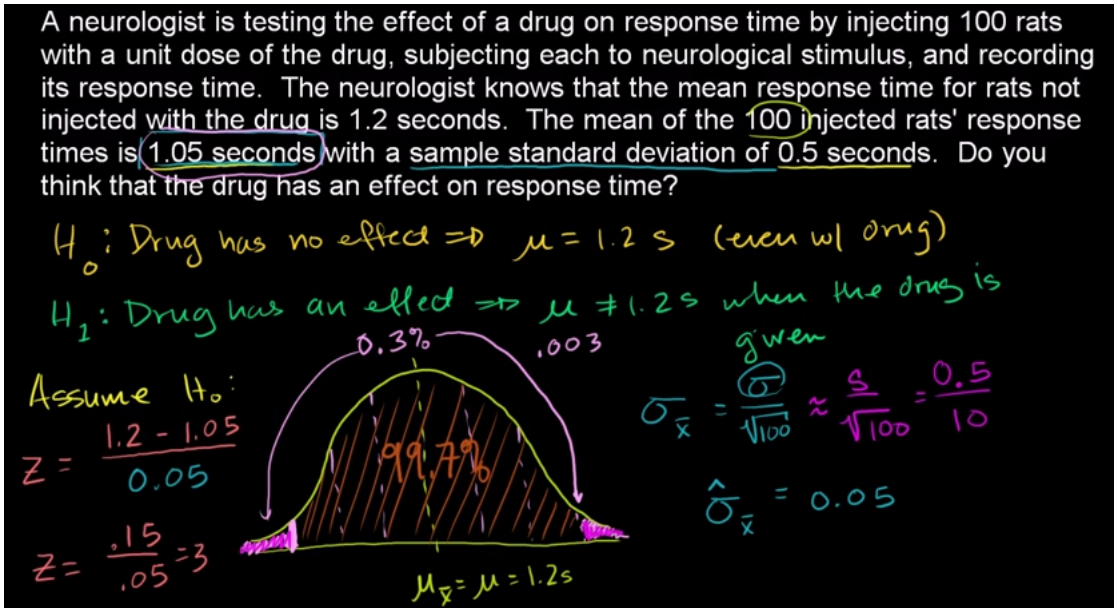
Så regnes teststørrelsen

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.2 - 1.05}{0.05} = 3$$

Så det er tre standard afvigelser væk fra gennemsnittet, μ. Hvad er sandsynligheden så for at få et resultat der er længere væk fra gennemsnittet end tre standard afvigelser? (Både i den negative og positive retning)

Ved at kigge i en z–tabel, så vil vi kunne se, at tre standard afvigelser ligger indenfor 99.7% og derved er det kun 0.3% der ligger udenfor, så sandsynligheden for at det ikke har nogen effekt er 0.003 og vi kan derfor afvise denne hypotese (fordi den er under 0.05).

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK



4.3.1 Ensidede hypotesetests

Vi kan foretage en hypotesetest på den samme opgave, men nu fokuserer vi kun på den ene af enderne; enten er har den en positiv (sænker tiden) eller negativ effekt (øger tiden).

Vi har stadig den samme nul–hypotese

$$H_0: \qquad \text{Ingen effekt}$$

med andre ord, så vil gennemsnittet forblive det samme, dvs

$$\mu = 1.2s$$

Den alternative hypotese er så

$$H_1: \qquad \text{Sænker responstiden}$$

så det vil sige

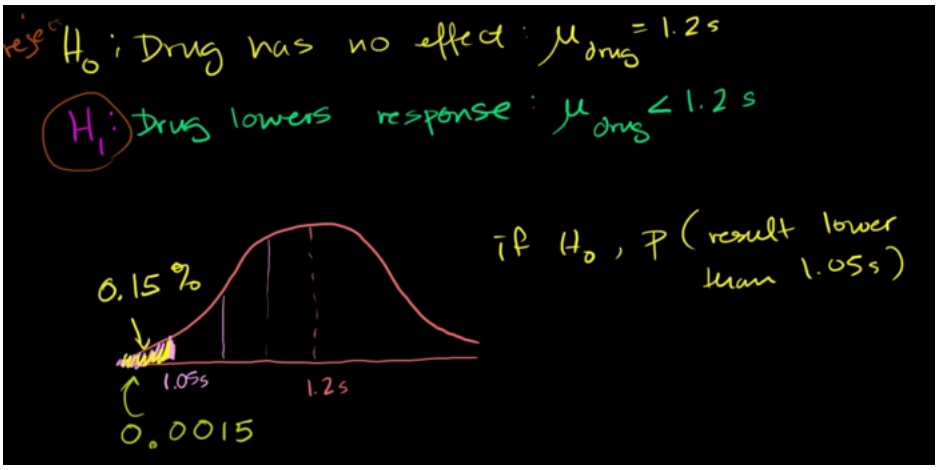
$$\mu < 1.2s$$

Igen antager vi at H_0 er sand og kigger på hvad sandsynligheden er for $\mu = 1.05s$. Så hvis H_0 er sandt, så

$$P(\text{Resultat lavere end 1.05s}) = 0.003/2 = 0.0015$$

Dette er meget usandsynligt, så vi afviser vores H_0 hypotese.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK



4.4 To populationer

Ofte har man flere grupper, som man ønsker at sammenligne. Det simpleste tilfælde er at undersøge to ting af gangen.

Der findes groft sagt to udgangspunkter: Enten ønsker man at sammenligne to naturligt forekommende grupper, eller også ønsker man at danne to grupper, som man så gør noget forskelligt ved. Disse grupper skal dannes så tilfældigt som muligt.

For eksempel, I en undersøgelse af brystkræftoperationer bliver patienterne fordelt tilfældigt i to grupper. Den ene gruppe får kun fjernet knuden, den anden gruppe får fjernet hele brystet. Patienterne er nødvendigvis vidende om behandlingens art, og resultatet kan være påvirket af patientens eget syn på behandlingen.

I nogen tilfælde kan man danne grupper, hvor objekterne er parvis ens - eller endog samme objekt i to situationer.

Når man skal sammenligne grupperne, kan man sammenligne gennemsnittene eller medianerne. For at sammenligne gennemsnittene, skal grupperne have ens varians. Men det kan være interessant i sig selv at undersøge om varianserne er ens.

To sovemediciner giver begge en gennemsnitssøvn på 8 timer. Søvnen ved den ene varierer mellem 7 og 9 timer, søvnen ved den anden varierer mellem 2 og 18 timer. De fleste vil nok foretrække den første medicin.

Hvis man har data fra to uafhængige populationer, kan man opstille hypoteserne

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 \\ H_1 &: \quad \mu_1 \neq \mu_2 \end{aligned}$$

eller

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 \\ H_1 &: \quad \mu_1 < \mu_2 \end{aligned}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

eller

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 \\ H_1 &: \quad \mu_1 > \mu_2 \end{aligned}$$

Teststørrelsen er

$$z_0 = \frac{\bar{x_1} - \bar{x_1} - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

men vi er interesseret $\delta_0 = 0$, så

$$z_0 = \frac{\bar{x_1} - \bar{x_1} - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

For at fortsætte må man antage, at de to populationer er normalfordelte og har samme varians. (Det kræver en test. Denne varians skal estimeres. Det er ikke lovligt at tage det fælles estimat for de samlede data, da dette estimat anvender det fælles gennemsnit. Dette ville betyde, at man antager et fælles gennemsnit til at vise, om middelværdierne er ens. Man må i stedet beregne hver stikprøves varianser ud fra de faktiske gennemsnit. Man har:

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

da vi antager at $\sigma_1^2 = \sigma_2^2$

4.4.1 Parvise tests

For at lave t-testene i sidste afsnit antog man, at der var tale om to uafhængige populationer. Man kan også have eksempler, hvor populationerne ikke er uafhængige, idet observationerne i de to grupper er parrede. Der kan dreje sig om de samme patienter før og efter en behandling, eller individer, hvor man måler på højre og venstre hånd. Andre gange kan man stille forsøget således op, at man først går gennem den ret besværlige fase, hvor man parrer sine forsøgspersoner, så de er ens m.h.t. nogle relevante faktorer. Derefter fordeler man hvert par tilfældigt på de to behandlinger, som man ønsker at undersøge. Fordelen ved denne procedure er, at man fjerner en hel del af den variation, der naturligt er imellem forsøgspersoner.

Beregningsmæssigt betyder dette, at man ser på forskellen mellem de to målinger indenfor hvert par, og i stedet for at regne med en varians i hver af de to grupper kan man nøjes med at se på variansen af denne forskel.

Hvis man beregner forskellen mellem observationerne indenfor hvert par, kan man fortsætte beregningerne, som om man testede i en population.

4.4.2 Hypoteser om varianser

For at teste en hypotese af formen

$$\begin{aligned} H_0 &: \quad \sigma_1 = \sigma_2 \\ H_1 &: \quad \sigma_1 \neq \sigma_2 \end{aligned}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

bruger man en varians ratio test. Man beregner en af disse to brøker:

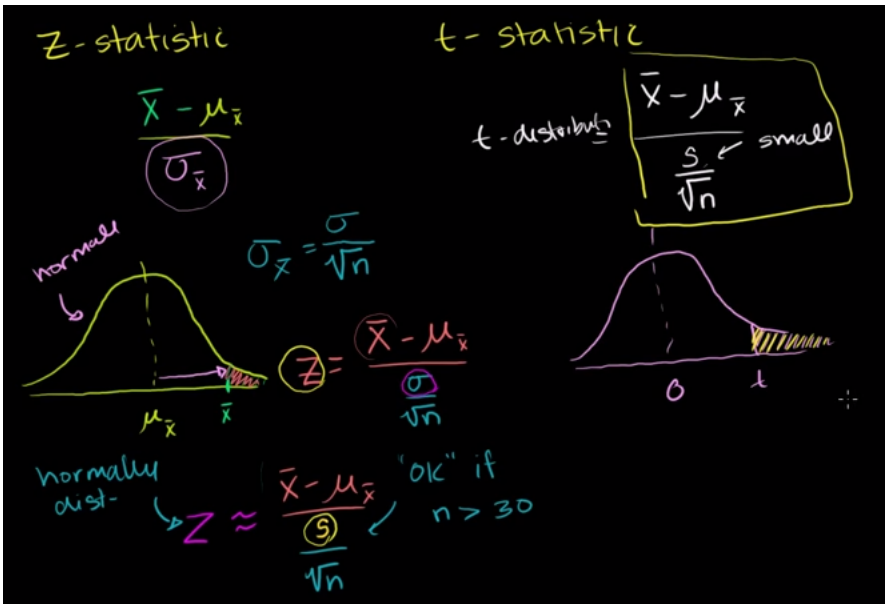
$$F = \frac{s_1^2}{s_2^2} \qquad eller \qquad \frac{s_2^2}{s_1^2}$$

Teststørrelserne har en F-fordelingKapiletskift. Klik to gange for at komme tilbage med hhv. (n_1,n_2) og (n_2,n_1) frihedsgrader. Hvis man skal bruge tabelopslag, beregner man den af de to brøker, der giver $F > 1$, og forkaster for $F > F_{tabelværdi}$.

Hvis $P < 0{,}05$, forkastes H0. Varianserne er altså ikke ens. Hvis man fortsætter og tester for, om de har ens gennemsnit, forkaster man ikke - men resultatet er i bund og grund uinteressant.

4.5 Z–statistik og T–statistik

Hvis s er lille ($n < 30$ stikprøver), så skal der bruges samme fremgangsmåde til udregning af teststørrelsen, men i stedet for at bruge en z–tabel, så bruger man en t–tabel.



FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

5 Emne 4 – Regression analyse

5.1 Simpel regression

I alle de tidligere nævnte tests har man for hvert objekt kun målt 1 variabel. Derfor kaldes det også univariabel analyse. Hvis man måler flere ting, kaldes det multivariabel analyse. Den simpleste model er lineær regression, hvor man har en uafhængig og en afhængig variabel, for eksempel vægt som funktion af højden.

Formålet med regressionen er, at kunne forudsige noget om den afhængige variabel, hvis man kender den uafhængige. Der kan være andre variable, der også forudsiger noget om den afhængige variabel, og man kan derfor også lave modeller med flere uafhængige variable.

Man kan også have variable, der varierer på en bestemt måde i forhold til hinanden, uden at der er tale om et direkte afhængighedsforhold. Der vil ofte være tale om en afhængighed af en helt tredie (fjerde, femte...) variabel. Der er så tale om korrelation.

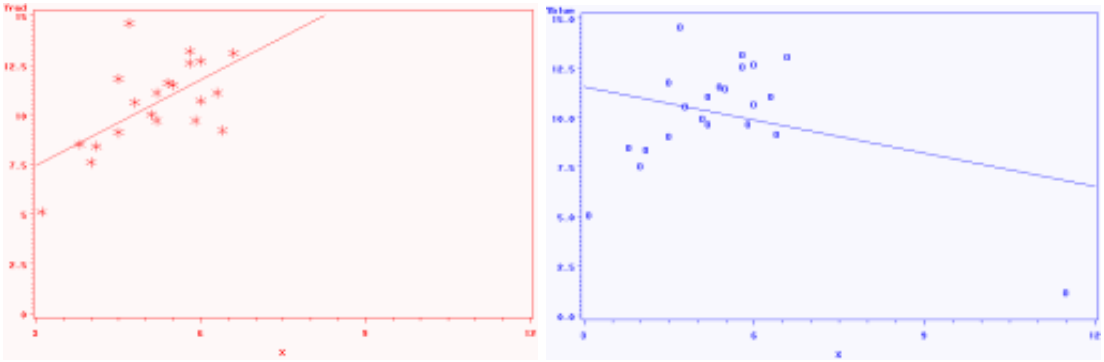
5.2 Den lineære model

Den lineære regressionsligning er givet ved

$$Y = \beta_0 + \beta_1 x + \epsilon$$

hvor β_0 og β_1 er henholdsvis skæringen på y–aksen og hældningen. ϵ er den tilfældige afvigelse, da de enkelte observationspunkterne meget sjældent ligger lige på linien.

Når man laver lineær regression skal man være meget opmærksom på observationer, der ligger langt fra de andre, såkaldte ekstreme værdier. De kan have stor indflydelse på regressionsliniens hældning. Man bør undersøge om de skyldes en tastefejl, en målefejl eller et objekt, der skal tages ud af undersøgelsen. Den røde figur viser en lineær regression og den blå figur er tilføjet en ekstreme værdi langt til højre for de andre punkter. Her er det tydeligt at se, hvad det har af betydning med de ekstreme værdier.



FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

5.2.1 Estimation af β_0 og β_1

Regressionslinen beregnes således, at man minimerer summen af kvadraterne på den lodrette afstand mellem punkterne og linien, og resultatet af minimeringen giver estimatet på β_1 som

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\sum_{i=1}^n x_i}{n}}$$

og estimatet på β_0 får ved

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

hvor $\bar{y} = (1/n)\sum_{i=1}^n y_i$ og $\bar{x} = (1/n)\sum_{i=1}^n x_i$.

5.2.2 Sum af kvadrater i regression

Den totale sum af kvadrater har $n-1$ frihedsgrader og beregnes ved

$$SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$$

Regressionens (modellens) afvigelse fra gennemsnittet har 1 frihedsgrad og beregnes ved

$$SS_M = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Forskellen på SS_T og SS_M angiver, hvor meget punkterne varierer omkring den fundne linie. For hvert x kaldes afstanden mellem den observerede og den beregnede y-værdi for residualen. Den sidste sum af kvadrater kaldes derfor enten SSR eller SSE (for error). Her benyttes SSE ligesom i de tidligere variansanalyser.

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

der har $n-2$ frihedsgrader.

Estimatet (unbiased) for variansen er

$$\sigma^2 = \frac{SS_E}{n-2}$$

5.2.3 Hypotesetest

Hypoteserne i en lineær regression kan skrives

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_c \\ H_1 &: \quad \beta_1 \neq \beta_c \end{aligned}$$

hvor β_c er den konstant vi ønsker at undersøge om hældningen er lig med. Man tester kun, om der er en stigende eller faldende sammenhæng mellem x og y. Man tester altså ikke, om denne sammenhæng er lineær.

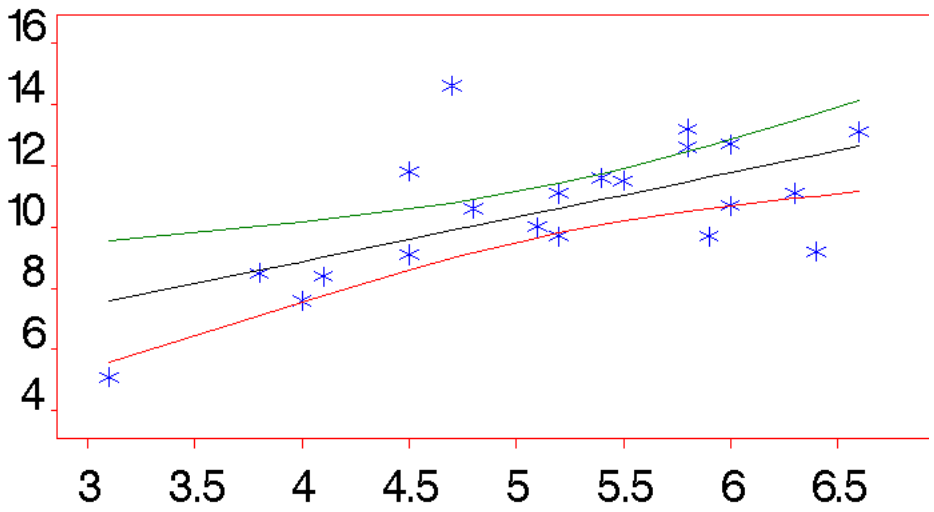
FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

5.2.4 Konfidensintervaller

Konfidensinterval for hældnignen: Til estimatet for β_1 kan man beregne et konfidensinterval. Man kan så tegne de to linier, hvis hældning er h.h.v. øverste og nederste konfidensgrænse for β_1. Det vil blive to linier, der roterer omkring punktet $\bar{y}$,$\bar{x}$. Man vil se, at jo længere man bevæger sig væk fra fixpunktet, jo større er usikkerheden. Konfidensintervallet for hældningen β_1 er givet ved

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma^2}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma^2}}{S_{xx}}}$$

Konfidensinterval for det estimerede y: β_0 er beregnet ud fra $\beta 1$ og punktet $\bar{y}$,$\bar{x}$. Da dette punkt også er et estimat, er der en usikkerhed herpå, og dermed er der en usikkerhed på β_0. Så man kan forestille sig, at konfidensgrænserne på den estimerede linie fås ved dels at vippe linien ligesom ovenfor og dels ved at skubbe linien lidt op og ned. Resultatet bliver et konfidensbånd omkring linien. Konfidensintervallet for



5.3 Korrelation

Den lineære regression kræver, at den ene variabel kan siges at være afhængig af den anden. Hvis man derimod har to ligeværdige variable, der afhænger af hinanden, så må man nøjes med at lave en lineær korrelation. Korrelationskoefficienten beregnes som

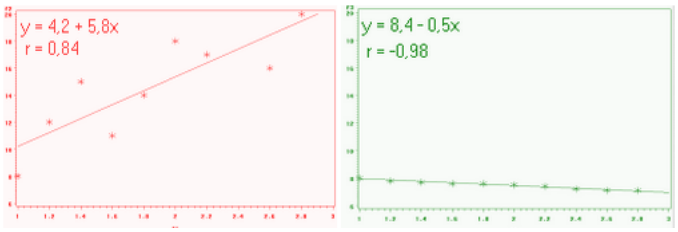
$$R = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}}$$

Hvis r er positiv, vil de to variable stige sammen, og det vil typisk vise sig ved, at punkterne ligger i en sværm, der ligger på skrå opad til højre, mens en negativ korrelation betyder, at den ene variabel stiger, når den anden falder, så sværmen af punkter går nedad mod højre. Der er dog ingen garanti for, at punkterne ligger på linie, se næste afsnit.

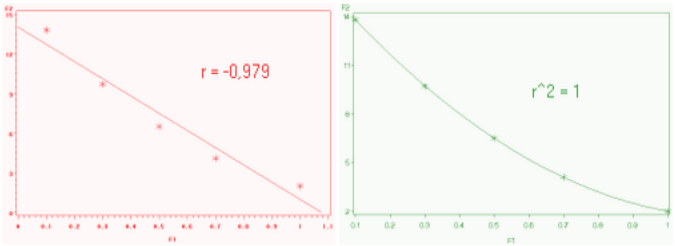
Der er to meget udbredte misforståelser om r (eller r2). Selv om eksemplerne her formuleres om lineær regression, gælder det samme i korrelationsanalyse, hvor man godt nok ikke må lave en ret linie, men hvor man alligevel har en ide om fordelingen at data.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

1. Korrelationskoefficienten måler ikke størrelsen af liniens hældning



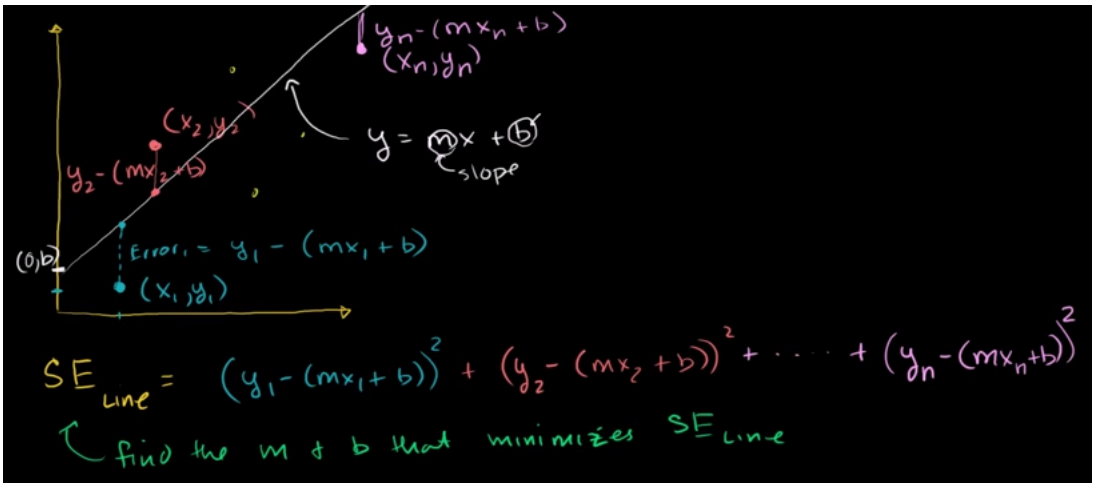
2. Korrelationskoefficienten måler ikke, hvor godt en ret linie passer



5.4 Eksempel

5.4.1 Squared Error

Tegn en masse punkter i koordinatsystemet og en linje der har den mindste fejl (den vertikale afstand op til regressions linjen)



Prøv med tre punkter; (1,2) (2,1) (4,3)

Find så

- Gennemsnittet af x'erne

$$\bar{x} = \frac{1+2+4}{3} = 7/3$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

- Gennemsnittet af y'erne

$$\bar{y} = \frac{2+1+3}{3} = 2$$

- Gennemsnittet af punkterne

$$\bar{xy} = \frac{1 \cdot 2 + 2 \cdot 1 + 4 \cdot 3}{3} = 16/3$$

- Gennemsnittet af de kvadrede x'er

$$\bar{x^2} = \frac{1^2 + 2^2 + 4^2}{3} = 21/3 = 7$$

Med udregning, så får vi en hældning på

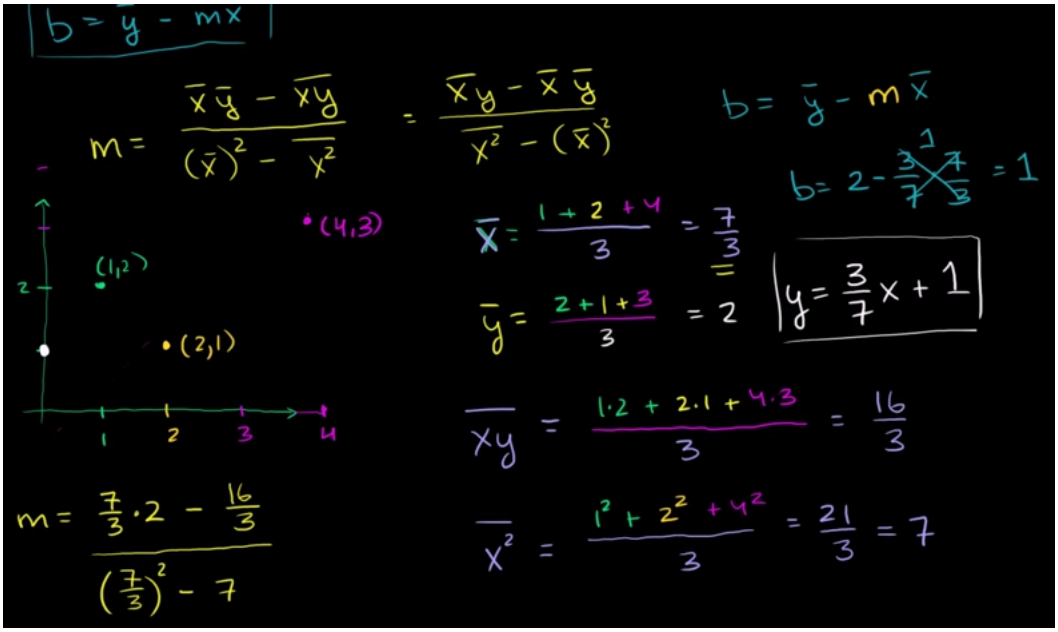
$$\beta_1 = 3/7$$

og vi kan nu finde skæringen på y–aksen, således

$$\beta_0 = 2 - \frac{3}{7} \cdot \frac{7}{3} = 1$$

Så det færdige udtryk for vores regressions linje er

$$y = \frac{3}{7}x + 1$$



FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

5.4.2 R–Squared (Coefficient of Determination)

Efter at have fundet regressions ligningen, så kan det være en fordel at kigge på, hvor godt linjen passer på datasættet.

$$SE_{line} = (y_1 - (\beta_0 + \beta_1 x_1))^2 + (y_2 - (\beta_0 + \beta_1 x_2))^2 + \cdots + (y_n - (\beta_0 + \beta_1 x_n))^2$$

Med andre ord, hvor meget (procent) af den totale variation i y er beskrevet af variationen i x?

Den totale variation i y (Squared Error of $\bar{y}$) er givet ved

$$SE_{\bar{y}}(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$$

Hvor meget af den totale variation er **ikke** beskrevet af regressions linjen?

$$\frac{SE_{line}}{SE_{\bar{y}}}$$

så for at finde r^2

$$r^2 = 1 - \frac{SE_{line}}{SE_{\bar{y}}}$$

hvis SE_{line} er er lavt, så er det et godt fit, og derfor tæt på 1.

5.5 Andre regression

Den polynomielle regression af grad n er givet ved ligningen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \cdots + \beta_i + x_i^n + \epsilon$$

Hvor ϵ er fejlen, og summen af ϵ'erne er lig nul.
Den multipel regression med k regressions variable er givet ved

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k + x_k + \epsilon$$

Hvor ϵ er fejlen, og summen af ϵ'erne er lig nul. De uafhængige variable skal være indbyrdes uafhængige og ukorrelerede.

Estimatet (unbiased) af variansen for en multipel regression er

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p}$$

Den overordnede hypotese er

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \quad \beta_j \neq 0, \text{ for mindst et } j \end{aligned}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

6 Emne 5 – Analyse af variance

6.1 Analysis Of Variance – Envejs (ANOVA 1 (Ch.13))

Hvis man har flere end to populationer, er det ikke praktisk at sammenligne dem to og to, da det giver en meget stor Type I fejl. Selv om man bruger en α på 0,05 hver gang, vil den samlede Type I fejl blive meget større. Hvis der er 4 populationer, er der 6 sammenligninger, og sandsynligheden for ikke at forkaste nogen af de 6, hvis H_0 er sand, er 0,95 = 0,74. Derved bliver sandsynligheden for mindst 1 Type I fejl = 1 - 0,95 = 0,26. I stedet bør man bruge en test, hvor den samlede sandsynlighed for Type I fejl er det valgte a.

ANOVA 1: I første omgang nøjes man med at teste den noget simple hypotese

$$H_0 \quad : \quad \tau_1 = \tau_2 = \cdots = \tau_k$$
$$H_1 \quad : \quad \text{Ikke alle } tau\text{'erne er ens}$$

Man tester hypotesen ved at se på forskellige former for variation. Derfor har testen fået navnet Analysis Of Variance.

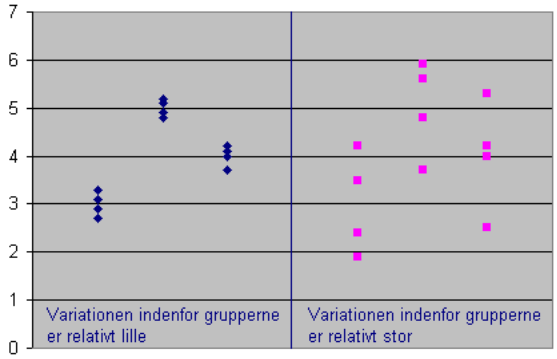
I stedet for at sige, at man har forskellige grupper, kan man sige, at man har en faktor med forskellige niveauer (deraf navnet envejs variansanalyse). Den variabel, der angiver disse niveauer, kaldes den uafhængige variabel. De målinger, man foretager, afhænger af hvilket niveau, man er på, og de kaldes derfor den afhængige variabel.

6.1.1 Antagelser

Når man laver en variansanalyse, antager man, at observationerne kommer fra uafhængige, normalfordelte populationer med ens varians. Testen er meget pålidelig selv ved store afvigelser fra normalfordelingen, bare stikprøverne ikke er for små. Ligeledes kan testen tåle store afvigelser indenfor varianserne, bare stikprøvestørrelserne fra de forskellige populationer er ens eller næsten ens.

Hvis man kigger nærmere på t-testen i to populationer, vil man se, at man i brøken sammenligner afstanden mellem de to gennemsnit (i tælleren) med spredningen (i nævneren). Man får en numerisk stor t-værdi, når afstanden mellem gennemsnittene er stor i forhold til spredningen, og en stor t-værdi betyder, at man forkaster hypotesen om ens gennemsnit.

Når man vil teste hypotesen H_0 mod H_1, så sammenligner man variationen indenfor grupperne med variationen mellem grupperne. Hvis variationen indenfor grupperne er lille i forhold til variationen mellem grupperne, forkaster man hypotesen om ens gennemsnit.



FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

Det ser umiddelbart ud til, at der er forskel på de tre grupper til venstre, mens der måske ikke er signifikant forskel på grupperne til højre.

6.1.2 Modellen

Envejs variansanalyse og t-test anvendes til eksperimenter med et meget simpelt design: 1 - 2 - mange grupper, der kun skal testes for forskellighed. Eksperimenter kan laves med meget mere indviklet design, og man kan teste for mange andre sammenhænge end blot forskellighed (tænk bare på lineær regression). Derfor er der brug for at beskrive design og sammenhænge med en matematisk model. Den model, der kan opskrive for både envejs variansanalyse og t-test, er

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

hvor Y_{ij} er den j'te observation i det i'te niveau af faktoren i den i'te gruppe, $\mu_i = \mu + \tau_i$ er gennemsnittet af det i'te observation og ϵ_{ij} er en eksperimentel fejl.

6.1.3 Summer af kvadrater (Sums of Squares – SS_T)

Størrelsen SS_E, der er variationen indenfor grupperne, kaldes også error sums of squares eller residual sums of squares, SS_R. Ud fra modellen beregnes her variationen på den normalfordelte eksperimentelle fejl

$$SS_E = \text{Indenfor grupperne SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y_{i\bullet}})^2$$

med frihedsgraderne

$$a(n-1)$$

Variationen mellem grupperne (gruppe sums of squares) er i modellen forklaret vha. a'erne. Da denne ene faktor kan kaldes A, benævnes variationen SS_A. Den beregnes ved

$$SS_A = \text{Mellem gurpperne SS} = \sum_{i=1}^{k} n_i(\bar{y_{i\bullet}} - \bar{y_{\bullet\bullet}})^2$$

med frihedsgraderne

$$a-1$$

Til sidst er der den totale variation, der også er summen af de to andre

$$SS_T = SS_E + SS_A = \sum$$

med frihedsgraderne

$$an-1$$

Estimatet indenfor grupperne er givet ved

$$MS_E = \text{Error mean square} = \frac{SS_E}{(a(n-1))}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

Fra SSA beregner man mellem-gruppe-variansen, der altså sammenligner gruppernes gennemsnit med det fælles gennemsnit, således

$$MS_A = \frac{SS_A}{a-1}$$

Under H_0 er MS_A også et estimat på s^2, da

$$\sum_{i=1}^{a} \tau_i = 0$$

Teststørrelsen bliver herefter

$$F_0 = \frac{MS_A}{MS_E}$$

Hvis H_0 er sand, burde F være tæt på 1, da MS_E og MS_A estimerer samme størrelse. Hvis MS_A derimod er stor i forhold til MS_E, kan det skyldes, at *tau*'erne ikke er 0, og derved bliver F stor, og så forkastes hypotesen om ens middelværdi. Hvis H_1 er sand, så MS_A et estimat på s^2 plus et positiv udtryk der inkorporerer variation som følge af systematisk forskel i gennemsnittene.

6.1.4 ANOVA test

Testen skrives op på følgende måde:

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Square	F_0
Model	SS_A	a-1	MS_A	$\frac{MS_A}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

Hvis $k = 2$ giver ANOVA testen samme resultat som t-testen, idet teststørrelserne opfylder at $F = t^2$, og fordelingerne opfylder at $F(1,N-2) = t(N-2)^2$.

To typer ANOVA: Man skelner mellem to typer eksperimenter

- I "fast faktor"modellen er de forskellige grupper faste, d.v.s. de er valgt, fordi netop de grupper har interesse. Det kan være nogle bestemte typer af behandling, som man vil sammenligne, og konklusionen af testen vil kun omhandle de anvendte behandlinger.
- I "tilfældig faktor"modellen er de forskellige grupper valgt tilfældigt ud fra en større mængde af mulige emner. Det kan f.eks. være 4 sygehuse i mindre danske byer. Man ønsker at drage en konklusion om den større gruppe, så hypotesen bør formuleres "Der er ikke forskel på ..... i sygehusene i mindre danske byer".

Så længe man kun har 1 faktor (1 gruppeinddeling), udføres ANOVA testen nøjagtig ens i de to modeller.

6.2 Analysis Of Variance – Tovejs (ANOVA 2 (Ch.14))

Envejs variansanalyse (ANOVA 1) blev introduceret som en udvidelse af t–testen til at analysere flere end to grupper. Disse grupper må nødvendigvis have et overordnet faktor. De enkelte grupper kaldes så niveauer af denne faktor. Et eksempel kunne være

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

Faktor	**Niveauer**
Køn	Mand, kvinde
Diæt	Kontrol(anbefalet energifordeling), høj fedt%, høj kulhydrat%

Hvis nu man undersøger vægtændring (den afhængige variabel) og ser på effekten af de forskellig diæter, så kunne det jo være interessant at se, om effekten er den samme hos mænd og kvinder. Derfor vil man gerne se på de to faktorer samtidig, og man får brug for tovejs variansanalyse. Man kunne tilføje en tredie faktor, f.eks. aldersgruppe, og så bruge trevejs variansanalyse. Her skal kun gennemgås tovejs variansanalyse.

6.2.1 Antagelser

Når man laver en variansanalyse, antager man, at observationerne kommer fra uafhængige, normalfordelte populationer med ens varians. Testen er meget pålidelig selv ved store afvigelser fra normalfordelingen, bare stikprøverne ikke er for små. Ligeledes kan testen tåle store afvigelser indenfor varianserne, bare stikprøvestørrelserne fra de forskellige populationer er ens eller næsten ens.

Balanceret krydsdesign: Hvis hvert niveau af den ene faktor optræder i kombination med hvert niveau af den anden faktor, sige man, at de to faktorer er krydsede. Man stiller sine data op i en tabel, hvor den ene faktor er repræsenteret i søjlerne og den anden faktor er repræsenteret af rækkerne. Hver niveau kombination kaldes en celle. Antallet af observationer i den i'te række og j'te celle kaldes n_{ij}. Hvis der er lige mange observationer i hver celle, har man et balanceret design.

Køn \Diæt	Normal	Fedtrig	Kulhydratrig
Mænd	$n_{11} = n$ = n; ‘x11	n12 = n; ‘x12	n13 = n; ‘x13
Kvinder	n21 = n; ‘x11	n22 = n; ‘x11	n23 = n; ‘x11

Den afhængige variabel er vægtændring.

Der kan opstille tre nul–hypoteser

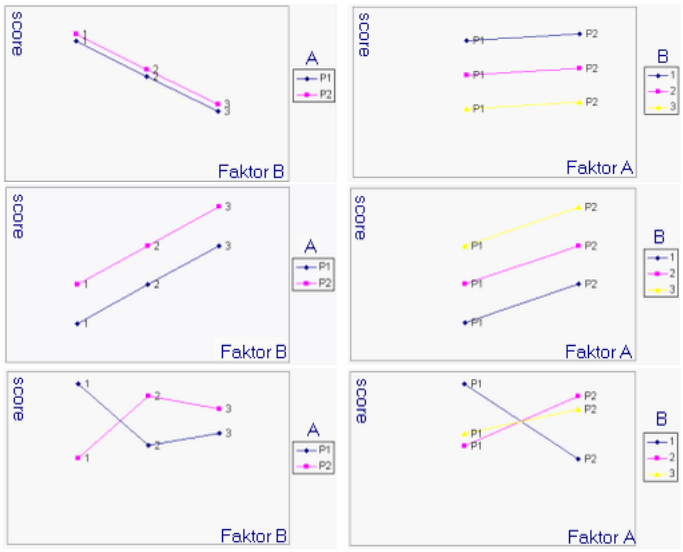
H_0	:	Der er ingen interaktion mellem køn og diæt
H_0	:	Køn har ingen effekt på vægtændringen
H_0	:	Diæt har ingen effekt på vægtændringen

Hypotesen om interaktion er med vilje sat øverst, selv om den i tabellerne står nederst. Man bør nemlig først teste for interaktionen. Hvis H0 ikke forkastes, d.v.s. at der ikke er interaktion, så går man videre og tester, om der er nogen effekt af de to faktorer. Hvis man til gengæld forkaster og altså finder en interaktion, så anbefales det, at man stopper der og ikke tester, om der er en effekt af de enkelte faktorer (kaldet main effekt eller hovedeffekten), da disse tests så ikke er gyldige (se eksempel 5 i næste afsnit). Der kan dog være situationer, hvor man ønsker at teste for hovedeffekterne, men man må så gøre sig selv og andre klart, hvordan hovedeffekterne skal forstås i netop denne situation.

For at lave en skitse af effekterne afsætter man cellernes gennemsnit i et koordinatsystem, hvor niveauerne af den ene faktor er afsat på x–aksen. For hvert niveau af den anden faktor, forbinder man de tilhørende gennemsnit. Man behøver kun at lave én figur. Nedenfor er begge mulige skitser

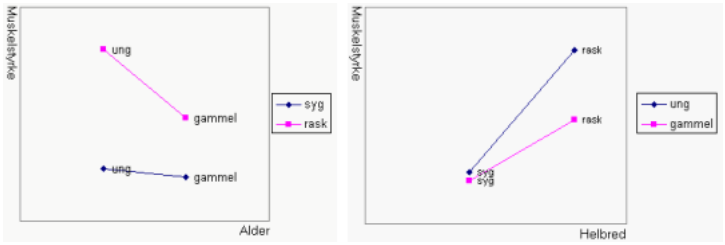
FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

med for at fremme forståelsen. Der er vist forskellige scenarier, når faktor A har niveauerne P1 og P2, mens faktor B har niveauerne 1, 2 og 3, og man har målt en point scoring



Når der ikke er interaktion, er linierne nogenlunde parallelle, og en eventuel effekt af faktorerne viser sig ved, at linierne parallelforskydes væk fra hinanden. Man skal selvfølgelig teste for at se, om effekterne er signifikante. Interaktion viser sig ved, at linierne ikke er parallelle - men man skal selvfølgelig teste, om denne interaktion er signifikant. Hvordan skal interaktionen så forstås?

Antag, at man måler muskelstyrke hos unge og gamle, hvoraf nogle er raske og andre har en sygdom, der svækker musklerne.



På figurerne ses klart en interaktion. På venstre figur er der en tydelig effekt af alder hos de raske, hvorimod der næsten ingen effekt er af alder hos de syge. Ser man i stedet på højre figur, kan man se, at der er en effekt af sygdommen, *men denne effekt er ikke lige stor hos unge og gamle.* Hvis testen viser, at denne interaktion er signifikant, så er der jo ingen grund til at teste for en separat effekt af alder eller af sygdom, for disse effekt er jo netop indbygget i interaktionen, jvf. det, der er skrevet med kursiv.

6.2.2 Modellen

Modellen for det balancerede krydsdesign er

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

hvor Y_{ijk} er den k'te observation på det i'te niveau af faktor A og det j'te niveau af faktor B, μ er gennemsnittet af det hele, τ_i er effekten af niveau i af faktor A, β_j er effekten af niveau j af faktor B, $(\tau\beta)_{ij}$ er interaktionen mellem niveau i af faktor A med niveau j af faktor B og ϵ_{ijk} er en normalfordelt eksperimentel fejl.

Hypoteserne kan skrives som

$$\begin{array}{lll} H_0 & : & \tau_1 = \tau_2 = \cdots = \tau_a = 0 \quad \text{Ingen main effekt for faktor A} \\ H_1 & : & \text{Mindste en } \tau_i \neq 0 \\ H_0 & : & \beta_1 = \beta_2 = \cdots = \beta_b = 0 \quad \text{Ingen main effekt for faktor B} \\ H_1 & : & \text{Mindste en } \beta_j \neq 0 \\ H_0 & : & (\tau\beta)_{11} = (\tau\beta)_{12} = \cdots = (\tau\beta)_{ab} = 0 \quad \text{Ingen interaktion} \\ H_1 & : & \text{Mindste en } (\tau\beta)_{ij} \neq 0 \end{array}$$

6.2.3 Summer af kvadrater

Den totale variation fås ved formlen

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2$$

med frihedsgraderne $a - 1$.

Som i envejs ANOVA kan SS_T deles i to portioner, SS_E, der er variationen indenfor grupperne og SS_M, der er variationen mellem grupperne og derfor den variation, der skal beskrives ved modellen

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij\bullet})^2$$

med frihedsgrader $ab(n-1)$.

$$SS_M = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$$

med frihedsgrader $ab - 1$.

For at lave tovejs ANOVA skal man dele variationen mellem celler op i tre komponenter: et bidrag fra faktor A, et bidrag fra faktor B og endelig et bidrag fra interaktionen.

Variationen mellem niveauerne af faktor A beregnes ved

$$SS_A = an\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$$

med frihedsgraden $a - 1$.

Tilsvarende fås variationen mellem niveauerne af faktor B ved

$$SS_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$$

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

med frihedsgraden $b-1$.

Endelig fås variationen fra interaktionen ved $SS_I = SS_M - SS_A - SS_B$ med frihedsgraden $(a-1)(b-1)$.

Alt i alt får man så at

$$SS_T = SS_A + SS_B + SS_I + SS_E$$

med frihedsgraden $abn-1$.

Teststørrelsen for

- Faktor A
$$F_0 = \frac{MS_A}{MS_E}$$
- Faktor B
$$\frac{MS_B}{MS_E}$$
- Interaktion mellem A og B
$$\frac{MS_I}{MS_E}$$

hvor

$$MS_A = \frac{SS_A}{a-1}$$
$$MS_B = \frac{SS_B}{b-1}$$
$$MS_I = \frac{SS_I}{(a-1)(b-1)}$$
$$MS_E = \frac{SS_E}{ab(n-1)}$$

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Square	F_0
Faktor A	SS_A	a-1	$MS_A = \frac{SS_A}{a-1}$	$\frac{MS_A}{MS_E}$
Faktor B	SS_B	b-1	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
Interaktion	SS_I	(a-1)(b-1)	$MS_I = \frac{SS_I}{(a-1)(b-1)}$	$\frac{MS_I}{MS_E}$
Error	SS_E	$ab(n-1)$	$MS_E = \frac{SS_E}{ab(n-1)}$	
Total	SS_T	$abn-1$		

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK

6.3 Eksempel

Datasæt; tre forskellige slags lægemiddel til en gruppe mennesker

$$\begin{array}{lll} 1 & : & 3\ 2\ 1 \\ 2 & : & 5\ 3\ 4 \\ 3 & : & 5\ 6\ 7 \end{array}$$

Det samlede gennemsnit

$$\bar{x} = \frac{3+2+1+5+3+4+5+6+7}{9} = 4$$

giver det samme som at tage de enkeltes gennemsnit og så gennemsnittet af det

$$\begin{array}{lll} \bar{x_1} & : & \dfrac{3+2+1}{3} = 2 \\ \bar{x_2} & : & \dfrac{5+3+4}{3} = 4 \\ \bar{x_3} & : & \dfrac{5+6+7}{3} = 6 \end{array}$$

Total sum of Squares

$$SS_T = (3-4)^2+(2-4)^2+(1-4)^2+(5-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2+(7-4)^2 = 30$$

SS indenfor grupperne er givet ved

$$SS_E = (3-2)^2+(2-2)^2+(1-2)^2+(5-4)^2+(3-4)^2+(4-4)^2+(5-6)^2+(6-6)^2+(7-6)^2 = 6$$

I stedet for at tage afstanden fra det samlede gennemsnit, så tages der nu afstand for hver gruppes gennemsnit, for at finde hvor stor variationen er.

SS mellem gurpperne er givet ved

$$SS_A = 3 \cdot (2+4)^2 + 3 \cdot (4+4)^2 + 3 \cdot (6+4)^2 = 24$$

hvilket også giver god mening, for

$$SS_T = SS_E + SS_A$$

Hypoteser

$$\begin{array}{lll} H_0 & : & \text{Lægemidlerne har ikke nogen effekt} \quad \mu_1 = \mu_2 = \mu_3 \\ H_1 & : & \text{Lægemidlerne har en effekt} \end{array}$$

Gå ud fra at H_0 er sand og udregn teststørrelsen

$$F_0 = \frac{MS_A}{MS_E} = \frac{12}{1} = 12$$

Rimeligt højt tal. Aflæs i F–tabel med en $\alpha = 0.10$ og med DoF 2 og 6, så finder vi en

$$F_c = 3{,}46$$

og vi kan så afvise H_0 fordi vores teststørrelse er langt større en vores critical F.

FACULTY OF ENGINEERING
UNIVERSITY OF SOUTHERN DENMARK.DK