Statistisk Dataanalyse

~~Normalfordelingen~~ Tilfældige værdier.

En tilfældig værdi giver bedre mulighed for at forudsige og forstå udfald.

Et eksempel er at 20% ~~kan ikke~~ køber ikke bøgerne. 55% tekstbogen, 25% begge bøger. Så

$$20 \cdot 0 + 55 \cdot 1 + 25 \cdot 2 = 105 \text{ bøger}$$

En variabel eller process med et numerisk udfald kaldes en "tilfældig værdi", denne repræsenteres ved X. De forskellige udfald nummereres med x_i for $i = 1, 2, 3$. Vi kan opskrives

i	1	2	3	total
x_i	\$0	\$137	\$170	
$P(X = x_i)$	0.2	0.55	0.25	1.00

Her med X for boghandlens ~~om~~ omsætning per studerende.

We call the expected outcome of

$$0 \cdot 0.2 + 137 \cdot 0.55 + 170 \cdot 0.25 = 117.85$$

Is called the expected value. ~~This is~~ Dette skrives som $E(X)$. $E(X)$ er ~~??~~ den gennemsnitlige omsætning per studerende.

Variansen af X er: $\mathrm{Var}(X) = \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j) = \sigma^2$

Standard afvigelsen er $sd = \sqrt{\mathrm{Var}(X)} = \sigma$

Lineære kombinationer af tilfældige værdier

Eks. ~~sa~~ en samling af hvor meget tid det tager for en mand at tage på arbejde

$$W = X_1 + X_2 + \dots + X_5$$
$$E(W) = E(X_1 + X_2 + \dots + X_5)$$
$$E(W) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5)$$
$$= 18 + 18 + 18 + 18 + 18 = 90 \text{ min.}$$

Den totale tid er summen af individuelle tider. En lineær kombination er:

$$aX + bY$$

for to tilfældige variabler X, Y og to koefficienter a, b. Så for eksemplet ovenover der er W:

$$W = 1X_1 + 1X_2 + 1X_3 + 1X_4 + 1X_5$$

Varians i lineære kombinationer af randomiserede variabler:

Variansen i en lineær kombination

$$Var(aX + bY) = a^2 \cdot Var(X) + b^2 \cdot Var(Y)$$

for uafhængige variabler

Eksempel med aktier $6000 i CAT $2000 XOM

Vi har:

	Gennemsnit ($\bar{x}$)	Standard Af	Varians
CAT	0.0204	0.0757	0.0057
XOM	0.0025	0.0455	0.0021

$$Var(6000 \times X + 2000 \cdot Y) = 6000^2 \cdot Var(X) + 2000^2 \cdot Var(Y)$$
$$= 36\,000\,000 \times 0.0057 + 4\,000\,000 \cdot 0.0021$$
$$= 213\,000$$
$$\sqrt{213\,000} = \$463$$ afvigelse

Med et månedligt overskud på gennemsnitligt $124 er en spredning på $463 rigtig meget.

Kontinuerte numeriske værdier

Dette handler om når output ikke er diskret. Når man ser et histogram med meget små intervaller da repræsenterer den en 'probability density function'. Eller en densitets distribution. En densitet har altid et areal under kurven på 1.
Vi kan finde andelen af folk med en højde mellem 180 og 185 ved:

$$\frac{\text{antal: } 180-185}{\text{total antal}}$$

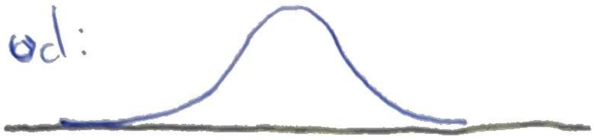
vi kan også:

P(højde givet mellem 180 og 185)

Normalfordelingen.

Normalfordelingen er universelt kendt i statistik.

Den ser således ud:



Parametrene for normalfordelingen er gennemsnitsværdien og spredningen. Disse kalder vi for μ og σ
Vi kan skrive en normalfordeling som ~~N~~ $N(\mu,\sigma)$.
Standardfordelingen er $N(0,1)$.

Standardisering med Z-værdier.

Nogle gange vil vi gerne kunne sammenligne ~~til~~ data, der ligner hinanden fx. testscore i forskellige skalaer.

En Z-score er det antal spredninger et givent udfald falder væk (over/under) fra gennemsnittet. Man kan på den måde sammenligne. Fx hvis en score er 1 spredning over gennemsnittet, da er $Z = 1$.
Derfor findes Z ved:

tallet der sammenlignes med

$$Z = \frac{x - \mu}{\sigma}$$

gennemsnit

spredningen

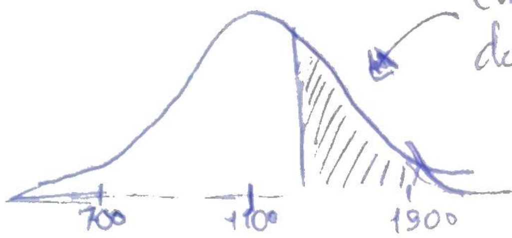
Hvis den absolutte værdi af Z_1 til observationen x_1 er større end Z_2, da kaldes den at være mere usædvanlig. (Mindre sandsynlig for at det sker).

At finde haler: Det samme som at finde i hvilken procentdel man er i.
Dette gøres i R. pnorm(x_1, mean = μ, sd = σ)

Et eksempel på normalsandsynlighed.

En tilfældig testtager vil have 1190 på sin test hvad er chancen? Normalfordelingen er $N(\mu = 1100, \sigma = 200)$

for at man får mindst 1190



Chancen for at havne i det her område.

Z-scoren er: $Z = \frac{x-\mu}{\sigma} = \frac{1190-1100}{200} = \frac{90}{200} = 0.45$

Arealet for $Z = 0.45$ er 0.6736. Arealet under hele kurven er $= 1$ derfor

$$1 - 0.6736 = 0.3264$$

Testdeltagerens chance for at få mindst 1190 er 0.3264.

0.6736 kan findes ved pnorm(Z).

68 - 95 - 99.7 reglen.

I normal fordelt data falder der sjældent noget data ~~...~~ 4-5 gange væk fra spredningen. Sandsynligheden for at noget data er mere end 4 gange væk fra spredningen er $\frac{1}{15000}$

Dvs: 68% af data er indenfor $\pm 1\sigma$
95% af data er indenfor $\pm 2\sigma$
99,7% af data er indenfor $\pm 3\sigma$

Geometriske distributioner

Bernoulli distribution.

~~En Bernoulli random variabel~~

Hvis ~~et~~ individuelt forsøg ~~t~~ kun har to mulige udfald. Da er det Bernoulli random variabel.

En stikprøve ~~andel~~ $\hat{p}$ ~~er gennemsnittet~~ er gennemsnittet af positive resultater fra stikprøven.

$$\hat{p} = \frac{\#\, success}{\#\, prøver}$$

Bernoulli random variabel: hvis X er en variabel der tager en værdi 1 for chancen for success og 0 med chancen $1-p$. Da er X en random variabel med $\mu = p$ $\sigma = \sqrt{p(1-p)}$

gennemsnit. spredning

Den geometriske distribution:

Den bruges til at beskrive hvormange prøver det kræves for ~~å~~ at opnå success.
Således hvis $p = 0.7$ (chancen for success. Chancen or fiasko er 0.3. Dan kan vi finde chancen for at skal tage to prøver for at få success

$$(0.3)(0.7) = 0.21$$

ved 3. $(0.3)(0.3)(0.7) = 0.063$.

Dette kræver at de randomiserede er uafhængige og identiske distribuerede. Med identisk menes der at hændelserne har samme chance for at ske.
Chancen aftager eksponentielt.

Det tager i gennemsnit $1/p$ forsøg for at få success under den geometriske distribution.

Binomialfordelingen.

Denne beskriver antallet af successer i et begrænset antal stikprøver. Dette er forskelligt fra den geometriske-fordeling. Denne beskriver hvor længe vi skal vente for at opnå success.

Binomialfordelingen beskriver hvad chancen er for at have k successer i n uafhængige Bernoulliprøver med en chance for success p.

den endelige chance er:

$$[\# \text{ scenarier}] \cdot P(\text{chancen } \cancel{\text{per}} \text{ for ét scenarie})$$

Vi bruger reglen ~~for~~ Multiplikationsreglen for uafhængige hændelser,

$$p^k(1-p)^{n-k}$$

Antallet mulige måder at man kan have k successer i n prøver er $\binom{n}{k}$ (n choose k)

Derfor er chancen for at observere præcist k successer i n uafhængige stikprøver, hvor sandsynligheden for success i en prøve er p:

$$\binom{n}{k} p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k(1-p)^{n-k}$$

Binomial?

1. uafhængighed i prøver
2. n faste prøver
3. outcome er enten success eller fiasko
4. chancen for success p er den samme for hver prøve

Normal fordelingen ~~og~~ som approksimation af binomial fordelingen.

Når stikprøvens størrelse (n) er stor er formlen til binomial besværlig at bruge. Det er nemmere at udregne sandsynlighederne ~~for~~ når man anvender normal modellen. Derfor når

$$\mu = np \geq 10 \qquad \sigma = \sqrt{np(1-p} \geq 10$$

Men når vi bruger normal fordelingen, skal vi ikke anvende den på små intervaller.

Poisson distributionen.

Poisson distributionen er god til at estimere antallet hændelser i en stor population over tid.
Her taler man om en rate ~~efter~~ dvs. antal hændelser per tid. Ved at bruge raten kan vi beskrive sandsynlighed for at observere k hændelser for en bestemt tidsenhed.
Antallet af observationer → raten kaldes λ el. μ

$$P(\text{observér } k \text{ hændelser}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Gennemsnit = λ, ~~var~~ spredningen = $\sqrt{\lambda}$

Grundlæggende interferens

Her skal vi forstå usikkerheden i estimater.
- Punktestimater og stikprøvevarians.
- Konfidensintervaller ~~og~~ for en stikprøve.
- Hypotesetests.

Punktestimat og fejl. I fx en meningsmåling af hvor mange der er for en 45%. De 45% er punktestimatet. Når vi taler om hele populationen omtales det om parameteret af interesse. Stikprøve usikkerhed beskriver hvor meget $\underbrace{\text{estimatet}}_{\text{stikprøven}}$ varierer fra en anden stikprøve.

Bias beskriver en systematisk tendens til at over eller underestimere den sande population. Bias kan minimeres igennem hvordan vi indsamler data.

Stikprøvedistributionen er der ~~hvor~~ hvor punktestimate kommer fra. Vi kommer ikke til at møde denne distribution.

Den centrale grænseværdi sætning. Når observationerne er uafhængige og store nok, vil stikprøve omfanget (sample proportion) $\hat{p}$, vil følge en normalfordeling. Da er:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

For at den centrale grænseværdi sætning skal holde skal stikprøven være stor nok så $np \geq 10$ og $n(1-p) \geq 10$

Emner i et eksperiment er uafhængige, hvis de er tildelt tilfældige grupper. ~~Hvis en observation~~ Hvis observationerne er fra én randomiseret stikprøve, er de uafhængige.

Fejlmargen

For et konfidensinterval er fejlmargen:

$$z^* \cdot SE.$$

Hypotese tests for en proportion.

Null hypotesen H_0 er et perspektiv der skal testes.

Den alternative hypotese H_A repræsenterer et alternativt perspektiv der er under vurdering.
Vi skal være skeptiske.

H_0 er et "no difference" perspektiv. Fx gør det en forskel på om almindelige mennesker er bedre ~~end~~ end uddannede.
H_A er ~~et~~ nyt eller stærkere perspektiv.