SENIORKURSUS

STATA OG BIOSTATISTIK

Aarhus Universitet juni 2011

Genopfriskning af statistik

Basale tankegange og begreber (i dag)
Sammenligninger (i morgen)
Sammenhænge (i overmorgen)

Brug af programpakken Stata (alle dage)

STATISTIK:

Metoder til indsamling, analyse og vurdering af data

Statistik
Deskriptive metoder – tabeller og figurer som belyser vigtige træk i data
Analytiske metoder – beregninger som søger at identificere og vurdere væsentlige aspekter af data

Statistisk inferens
Brug af statistiske metoder til at besvare (videnskabelige) spørgsmål ved hjælp af data som er underlagt tilfældig variation

Data betragtes som en ***stikprøve*** (sample) fra en population.

Den statistiske analyse af stikprøven har til formål at identificere og beskrive væsentlige forhold i denne population

De basale komponenter af en statistisk analyse

1. En statistisk model: Det videnskabelige problem "oversættes" til et statistisk problem.

Hvilke systematiske forskelle vil man tage hensyn?
Hvordan beskrives den tilfældige variation?

En statistisk analyse er altid baseret på en statistisk model, som skal afspejle undersøgelsens design og formalisere de antagelser som gøres vedrørende den systematiske og tilfældige variation.

Den statistiske model er en approksimation til "virkeligheden". Valg af en passende model er et kompromis mellem:

En **simpel model**, som giver en grov approksimation, men er nem at bruge

En **kompleks model**, som giver et præcist billede, men som er vanskelig – eller umulig – at bruge

Basale komponenter (2)

Den statistiske model vil typisk indeholde nogle ukendte **parametre**, som afspejler de aspekter af problemstillingen, og som man ønsker at kende eller tage hensyn til.

2. Estimation af relevante parametre og vurdering af hvor godt de er bestemt.

Hvad kan data fortælle om de ukendte parametre?

Hvor godt er vores skøn bestemt?

3. Test af hypoteser om modellens parametre

De spørgsmål som stilles til data kan ofte formuleres som spørgsmål vedrørende modellens parametre, typisk at de antager særlige værdier

Er de afvigelser vi ser mellem den estimerede værdi og den forventede værdi udtryk for en reel afvigelse eller afspejler den bare den tilfældige variation?

Basale komponenter (3)

Validiteten af analysens konklusioner afhænger af i hvor høj grad den statistiske model giver en rimelig beskrivelse af den systematiske og tilfældige variation, Variationskilderne afspejler design, dataindsamling/målemetoder og studiepopulation.

4. Modelkontrol – en vurdering af om analysens forudsætninger synes at være opfyldte – er derfor en vigtig del af en statistisk analyse.

Valide konklusioner forudsætter at den statistiske model giver en adækvat beskrivelse af hvordan data er indsamlet og inddrager de relevante systematiske og tilfældige variationskilder.

EKSEMPEL: Sammenligning af to behandlinger/lægemidler

Formål: Undersøge om en behandling er bedre end en anden

Design: Til en række "Eksperimentelle enheder" allokeres en af de to behandlinger. For hver enhed måles et kontinuert ***udfald*** som afspejler hvor virksom behandlingen er.

Analyse: For hver behandlingsgruppe beregnes ***gennemsnittet af udfaldene.*** De to gennemsnit sammenlignes for at afgøre hvilken behandling der er bedst

Mulige forklaringer på en forskel mellem de to gennemsnit

1. ***Den ene behandling er bedre end den anden***
2. ***Tilfældigheder***
3. ***"Bias" –dvs grupperne ikke er sammenlignelige. Andre faktorer som har betydning for udfaldet er forskelligt fordelt i de to grupper***

Spørgsmål: *Hvilken forklaring er den rigtige?*

Overordnet strategi: Forsøge at vise at den første forklaring er den rigtige ved at udelukke de to andre forklaringer

Bias som forklaring

Gennem studiets

design (randomisering og blinding etc) og
udførelse (compliance, drop-outs etc)

kan vi reducere (eliminere?) bias så denne forklaring ikke er sandsynlig

Om nødvendigt er det også muligt at reducere bias – og derved at forbedre sammenligneligheden - ved af foretage ***korrektioner i den statistiske analyse*** (f.eks. korrektion for baseline forskelle).

Tilfældig variation som forklaring

I den statistiske analyse estimerer man forskellen mellem de to behandlingers virkning og man evaluerer om tilfældigheder er en plausibel forklaring

Hvis undersøgelsen er
omhyggeligt planlagt og vel udført

og hvis den statistiske analyse viser
at tilfældig variation ***ikke*** er en plausibel forklaring

så kan man konkludere at den første forklaring synes at være den rigtige – dvs de to behandlinger virker forskelligt.

Den statistiske analyse giver også et **sikkerhedsinterval** for behandlingsforskellens størrelse
dvs de værdier af behandlingsforskellen, som er forenelige med de indsamlede data.

Eksempel: *Fiskeolie tilskud til gravide kvinder*

Baggrund: Lav fødselsvægt og/eller for tidlig fødsel forøger risikoen for en række komplikationer

Formål: At undersøge om tilskud af fiskeolie til gravide kvinder forlænger graviditeten og forøger fødselsvægten.
Her ses på et *sekundært endpoint*: **Diastolisk Blodtryk**

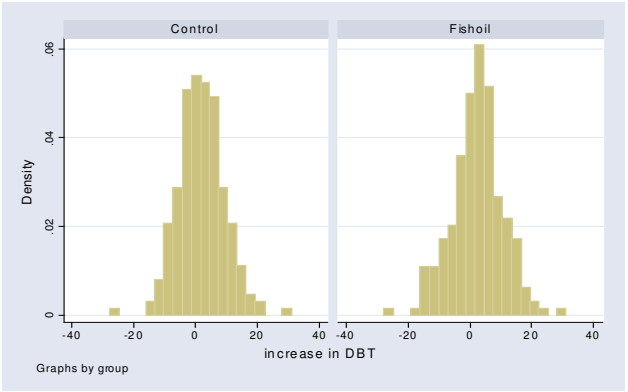
Design: Randomiseret, kontrolleret trial. 430 gravide kvinder blev inkluderet efter 30 ugers graviditet og randomiseret til enten at modtage et tilskud af fiskeolie eller et tilskud af anden olie.

Her ses på følgende data:

Ændring i diastolisk blodtryk = blodtryk i uge 37 – blodtryk i uge 30

Overvej:
Hvorfor korrigere for blodtryk i uge 30?
Og skal det i givet fald ske ved at se på ændringer?

Spørgsmål: Er der forskel på ændringen af det diastoliske blodtryk?
Hvad siger data?



Betydelig variation inden for hver gruppe. Beskeden forskel mellem grupperne.

Problemstilling: den statistisk terminologi
Et to-stikprøve problem for normaldelte observationer

Den statistisk analyse tager udgangspunkt i følgende **statistiske model**:

For hver behandlingsgruppe betragtes data som en stikprøve af uafhængige observationer fra en **normalfordelt population**.

En normal fordeling er fuldstændigt karakteriseret ved middelværdi og spredning (eller middelværdi og varians)

Vi antager

gruppe	middelværdi	varians
Kontrol	μ_1	σ^2
Fiskeolie	μ_2	σ^2

Forudsætninger

Uafhængige observationer – både indenfor og mellem stikprøver

Stikprøver fra to populationer med samme varians

I hver population kan den tilfældige variation beskrives med en normalfordeling

Er disse forudsætninger rimelige?

Estimation af modellens parametre

Populationsstørrelser estimeres med de tilsvarende størrelser i stikprøven

Kontrol Gennemsnitlig ændring $= \bar{x}_1 = 1.90\ mmHg$

Fiskeolie Gennemsnitlig ændring $= \bar{x}_2 = 2.19\ mmHg$

Skøn over fælles spredning $= s = 7.96\ mmHg$

Statistisk test: Sammenligning af to middelværdier

Den gennemsnitlige ændring af det diastoliske blodtryk er lidt højere i fiskeolie-gruppen. *Er det udtryk for en* ***systematisk forskel*** *eller skyldes det blot* ***tilfældigheder?***

Den basale strategi ved statistisk test af en hypotese:

Man bekræfter ikke hypoteser, man afviser hypoteser.

Her – og generelt i "superiority trials"
For at vise at der er en forskel forsøger man at afvise at de er ens.

Test af hypotesen H_0: $\mu_1 = \mu_2$

Hvis hypotesen er korrekt må den forskel, som ses i data, skyldes tilfældigheder. *Er dette rimeligt?*

Hvor store forskelle kan med rimelighed tilskrives tilfældig variation?

Behandlingsforskellen estimeres til

$$\bar{x}_2 - \bar{x}_1 = 2.19 - 1.90 = 0.29$$

Usikkerheden - induceret af den tilfældige variation i populationen - på dette skøn vil under de givne forudsætninger være

$$\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Under nulhypotesen (samme forventet ændring) kan usikkerheden af den standardiserede værdi

$$\frac{\bar{x}_2 - \bar{x}_1}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

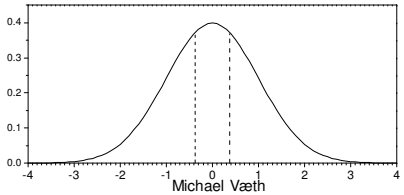
beskrives ved en standard normalfordeling (dvs middelværdi 0 og varians 1). Populationens spredning er ukendt. Bruges i stedet stikprøvens spredning skal den værdien vurderes i en t-fordeling.

t-fordelinger er kendte og tabellerede så vi kan vurdere om den aktuelle t-værdi er ekstrem

$$t = \frac{\bar{x}_2 - \bar{x}_1}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{2.19-1.90}{7.96\sqrt{\frac{1}{213}+\frac{1}{217}}} = \frac{0.29}{0.77} = 0.38$$

Sandsynligheden for at få en værdi mere ekstrem end den aktuelt observerede, også kaldet ***p-værdien***, er 0.70.

Er behandlingerne lige virksomme vil man i 7 ud af 10 forsøg af denne type se en forskel som er mindst lige så stor som den aktuelle forskel.



Traditionelt vælger man at afvise tilfældigheder som en mulig forklaring hvis p-værdien er mindre end 5% (signifikansniveauet)

Konklusion: Vi kan ikke afvise at behandlingerne virker ens og at forskellen blot skyldes tilfældigheder

MEN vi har ***ikke bevist*** at behandlingerne virker på samme måde for der er også andre værdier end 0 som heller ikke kan afvises.

Et **95% sikkerhedsinterval** (eller konfidensinterval) indeholder de værdier af forskellen som ikke kan afvises på niveau 5%

Vi får følgende 95% sikkerhedsinterval for forskellen mellem den forventede ændring i de to grupper

$$0.29 - t_{0.975}\,0.77 \le \mu_2 - \mu_1 \le 0.29 + t_{0.975}\,0.77$$
$$-1.22 \le \mu_2 - \mu_1 \le 1.80$$

forskel mellem gennemsnit

97.5% percentil i t-fordeling

spredning på forskel mellem gennemsnit

Sikkerhedsintervaller og statistiske test

I eksemplet testede vi hypotesen $\mu_1 = \mu_2$, dvs samme ændring af det diastoliske blodtryk i de to grupper. Hypotesen kunne ikke afvises, p = 0.70, og derfor indeholder er 95% sikkerhedsinterval for den forventede forskel værdien 0.

Et **smalt** sikkerhedsinterval for en parameter, her forskellen $\delta = \mu_1 - \mu_2$, betyder at den estimerede forskel er godt bestemt. Undersøgelsen indeholder meget information om denne parameter.

Omvendt afspejler et **bredt** sikkerhedsinterval at parameteren er dårligt bestemt. De værdier som ikke kan afvises, kan være meget forskellige.

Et sikkerhedsintervals bredde vil typisk aftage som $1/\sqrt{n}$, hvor *n* er stikprøvens størrelse.

Brug (misbrug?) af sikkerhedsintervaller

Det Nationale Indikatorprojekt

- kvalitetsudvikling i det danske sundhedsvæsen

I Det Nationale Indikatorprojekt forsøger man at vurdere "kvalitet i behandlingen" ved at opgøre om række forhold omkring behandlingerne lever op til nogle fastlagte standarder.

Resultaterne er tilgængelige på nettet, se f.eks.

http://www.nip.dk/
https://www.sundhed.dk/Artikel.aspx?id=49473.1
https://www.sundhed.dk/Profil.aspx?id=28739.1

Det Nationale Indikatorprojekt (fortsat)

National rapport for borgere om kvaliteten i behandlingen af lungekræft 2008

Nedenfor følger resultater for en række af de mest målbare kvalitetsresultater i behandlingen af patienter med lungekræft.

Resultaterne opgøres i forhold til følgende indikatorer:

- Overlevelse – 1, 2 og 5 år (Indikator 1a-1c)
- Overlevelse, opererede patienter - 30 dage, 1, 2 og 5 år (Indikator 2a - 2d)
- Varighed af undersøgelse (udredning) og tid indtil behandling (Indikator 3a-3e)
- Hyppigheden af operation med helbredende sigte (Indikator 5)

Tærskelværdier
For hver indikator er der beskrevet det resultat, som afdelingerne/regionerne mindst skal opnå for at leve op til kvalitetskravet. Læs mere om indikatorerne og tærskelværdier i "Hvordan måler vi kvalitet i behandlingen af lungekræft?", som kan tilgås på www.sundhed.dk/find.aspx?id=2144070606162839

Eksempel: ***Udredning og ventetid inden behandling*** for patienter med lungecancer:

Kvalitetskrav: ventetiden må højst være 4 uger

Standard: kvalitetskravet overholdes for mindst 85% af patienterne.

OVERVEJ: Hvordan vurderer man om standarden er opfyldt? Er det en test situation? Hvilken hypotese vil man afvise?

Behandling af lungekræft 2008

Indikator 2d. 5-års overlevelse

Overlevelse, opererede patienter efter 5 år (Indikator 2d)		
Tærskel: **Mindst 40 %** af opererede patienter bør være i live 5 år efter operationen		
	Tærskelværdi nået?	Procentdelen af patienter, der er i live efter 5 år
		2003
Patienter opereret i Region Hovedstaden	Ja*	36
Patienter opereret i Region Syddanmark	Ja*	34
Patienter opereret i Region Midtjylland	Ja*	35
Patienter opereret i Region Nordjylland	Nej	21
Landsresultat	Nej	34
*Der er taget højde for den statistiske usikkerhed, derfor er tærskelværdien nået		

Hvad foregår her?

Behandling af lungekræft 2008

Indikator 3a. Varighed af udredning

Varighed af undersøgelse (Indikator 3a)

Tærskel: **Mindst 85 %** af patienter henvist til sygehuset med lungekræft bør være undersøgt (udredt) inden for 4 uger efter, at udredningen er påbegyndt

	Tærskelværdi nået?	Procentdelen af patienter, der er undersøgt inden for 4 uger		
		2008	2007	2006
Patienter udredt i Region Hovedstaden	Ja*	83	73	73
Patienter udredt i Region Sjælland	Nej	81	75	74
Patienter udredt i Region Syddanmark	Ja	88	87	82
Patienter udredt i Region Midtjylland	Nej	79	56	45
Patienter udredt i Region Nordjylland	Nej	79	59	57
Landsresultat	Nej	83	72	68

*Der er taget højde for den statistiske usikkerhed, derfor er tærskelværdien nået

Hvad foregår her?

Tidligere brugte man et andet format:

Indikator 2: Udredningstid og ventetid inden behandling

***Indikator 2.a**: Andel af patienter, hvor udredningstiden (højst 4 uger) overholdes.*

Standarden er, at udredningstiden overholdes for 90 % af patienterne.

Indikatorværdier og standardopfyldelse

Afdeling	Standard opfyldt? Ja/Nej	Antal patient-forløb	Antal patientforløb som opfylder kravet	Andel i % af patientforløbene som opfylder kravet (95 % CI)		Mang-lende oplys-ninger
Frederikshavn-Skagen Sygehus, Medicinsk afdeling	**Ja**	**2**	**1**	**50**	**(1-99)**	**1**
Sygehus Himmerland, Medicinsk afd., Farsø	**Ja**	**10**	**8**	**80**	**(44-97)**	**0**
Hjørring-Brønderslev Sygehus, Lungemedicinsk afdeling	**Nej**	63	39	62	(49-74)	1
Aalborg Sygehus, Lungemedicinsk afdeling	**Nej**	57	25	44	(31-58)	1
Nordjyllands Amt	**Nej**	132	73	55	(46-64)	3
Landsresultat		2328	1546	66	(64-68)	17

ER KONKLUSIONERNE FORNUFTIGE??? GIVER DET MENING??

Estimation – lidt mere

I simple situationer bruges de ***"naturlige" estimater***:

Stikprøvens gennemsnit bruges som estimat for populationens middelværdi

Den relative hyppighed bruges som estimat for tilsvarende sandsynlighed

I mere komplicerede situationer er der ingen "naturlige" estimater.

Der findes ***generelle metoder*** til beregning af "optimale" estimater, dvs estimater, som bedst muligt udnytter den information der er i data. Disse metoder giver også skøn over estimaternes usikkerhed

Stata – og andre statistiske programpakker - vil typisk bruge disse metoder til beregningerne.

Statistiske test – lidt mere (1)

Mange statistiske test har formen

$$T = \frac{\text{Beregnet værdi - Forventet værdi}}{\text{Usikkerhed på beregnet værdi}}$$

Den beregnede værdi vil typisk være et estimat eller en passende teststørrelse (eksempel = rangsummen i den ene stikprøve)

Usikkerheden måles som standard fejlen (s.e.) for den beregnede størrelse.

Teststørrelser af denne form vil i store stikprøver typisk kunne vurderes i en standard normalfordeling (følger af et matematisk resultat, en såkaldt central grænseværdi sætning) hvis hypotesen er korrekt.

Alternativt kan man – også i store stikprøver – vurdere T^2 i en chi-i-anden fordeling.

Statistiske test – lidt mere (2)

Er den beregnede værdi et estimat, dvs en teststørrelse på formen

$$T = \frac{estimat - forventet\ værdi}{s.e.(estimat)}$$

kan dette resultat bruges til at beregne ***approksimative*** 95% sikkerhedsintervaller. Disse intervaller har formen

$$estimat \pm 1.96 \cdot s.e.(estimat)$$

Sådanne sikkerhedsintervaller vil være valide i store stikprøver, men kan være mindre gode i små stikprøver. Dette problem kan ofte reduceres ved at beregne sikkerhedsintervallet for en passende transformation af estimatet – og så regne tilbage.

Er estimatet et simpelt gennemsnit fås

$$gennemsnit \pm 1.96 \cdot s.e.m.$$

som et approksimativt 95% sikkerhedsinterval for middelværdien

SD eller SE?

Intervallet

$$gennemsnit \pm 1.96 \cdot s.e.m.$$

fortæller således noget om hvor ***populationens middelværdi*** med ligger.

Intervallet

$$gennemsnit \pm 1.96 \cdot s.d.$$

fortæller noget om hvor de ***enkelte observationer*** typisk ligger. Kan observationernes variation beskrives med en normalfordeling ville dette interval omfatte (cirka) 95% af observationerne.

Hvilket interval skal man bruge i figurer og tabeller?

Svaret afhænger af hvad man ønsker at vise.

Statistiske test – nogle generelle betragtninger (1)

For at evaluere en hypoteses rimelighed beregnes en test størrelse, som følger en kendt fordeling, hvis hypotesen er sand. Det er derfor muligt at beregne sandsynligheden for at få en værdi, som er mindst lige så ekstrem som den observerede værdi af test størrelsen. ***Denne sandsynlighed kaldes p-værdien***.

P-værdien beskriver i hvilken grad data understøtter hypotesen. Resultatet af det statistiske test klassificeres ofte som "*statistisk signifikant*" eller "*ikke statistisk signifikant*" alt efter om p-værdien et mindre eller større end ***signifikansniveauet***, kaldet **α**, og sædvanligvis lig med 5%.

Den hypotese som testes kaldes ofte nulhypotesen og repræsenterer sædvanligvis en forenkling af den statistiske model: "ingen forskel" eller "ingen sammenhæng".

Statistiske test – nogle generelle betragtninger (2)

I en ***beslutningsteoretisk formulering*** af hypotese testning er en nul hypotese er enten ***sand*** eller ***falsk*** og der skal træffes et valg – en beslutning – på basis af de indsamlede data

Når hypotese testning betragtes som en beslutningsproblem er der to fejl som kan begås

Fejl af Type 1: En sand nulhypotese forkastes

Fejl af Type 2: En falsk nulhypotese accepteres (dvs forkastes ikke)

Signifikansniveauet angiver den højest acceptable **risiko for fejl af type 1**. Traditionelt testes en nulhypotese mod en **alternativ hypotese** som specificere en række andre mulige værdier af parameteren, f.eks.

$$H_0 : \mu = 0 \quad \text{versus} \quad H_A : \mu \neq 0$$

Statistiske test – nogle generelle betragtninger (3)

Risikoen for fejl af type 2 afhænger af hvilken af disse alternative værdier, der er den sande værdi.

Styrken (power) af et statistisk test er 1 minus risikoen for type 2 fejl og afhænger derfor **også** af det betragtede alternativ.

En undersøgelse har derfor ikke én risiko for type 2 fejl eller én statistisk styrke, men mange.

Styrkeovervejelser baseret på et klinisk relevant alternativ bruges sædvanligvis til beregning af ***den nødvendige størrelse af studiet***.

Stata kan udføre sådanne beregninger ved hjælp at kommandoen **sampsi**

Når data er indsamlet er sikkerhedsintervaller den mest relevante måde at beskrive resultaternes usikkerhed.

STATA – analyserne af ændring af diastolisk blodtryk

Data findes i en Stata-fil som hedder **fishoil.dta.**

Filen indeholder følgende variable

grp	behandlingsgruppe, 1 for "control", 2 for "fish oil"
group	a string variabel with med gruppernes navne
difsys	ændring i systolisk blodtryk fra uge 30 til uge 37
difdia	ændring i diastolisk blodtryk fra uge 30 til 37

Histogrammerne laves med kommandoen

```
histogram difdia , by(group)
```

Normalfordelingsantagelsen kan checkes grafisk (Q-Q plots) ved

```
qnorm difdia if grp==1, title("Control")
qnorm difdia if grp==2, title("Fish oil")
```


Antagelsen om **samme variation** i de to grupper kan vurderes ved at teste hypotesen $H: \sigma_1^2 = \sigma_2^2$

Dette gøre i Stata med kommandoen ***sdtest***

```
sdtest difdia , by(grp)
```

Denne hypotese kan ikke afvises (p = 0.12)

Sammenligning af den forventede ændring i de to behandlingsgrupper foretages i Stata med kommandoen

```
ttest difdia , by(grp)
```

Output indeholder (stort set) alle relevante oplysninger, bl.a. 95% sikkerhedsinterval for forskellen og p-værdi for test af hypotesen og samme ændring i de to grupper.

Stata's output fra kommandoen ttest:

```
------------------------------------------------------------------------------
   Group     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 control |   213    1.901408    .5158685    7.528853    .8845197    2.918297
fish oil |   217    2.193548    .5678467    8.364904     1.074318    3.312778
---------+--------------------------------------------------------------------
combined |   430    2.048837    .3835675    7.953826     1.294932    2.802743
---------+--------------------------------------------------------------------
    diff |       -.2921399    .7679341               -1.801531    1.217252
------------------------------------------------------------------------------
Degrees of freedom: 428

                 Ho: mean(control) - mean(fish oil) = diff = 0
     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
      t =  -0.3804                 t =  -0.3804                  t =  -0.3804
  P < t =   0.3519             P > |t| =   0.7038              P > t =   0.6481
```

Kommandoen summarize giver basale oplysninger om en variabel

```
sum difdia if grp==1
```

```
Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
  difdia |       213    1.901408    7.528853       -28         29
```

Flere detaljer fås ved at skrive

```
sum difdia if grp==1 , detail
```

Basale oplysninger for gruppe 1 og gruppe 2 hver for sig:

```
bysort grp: sum difdia
```

15. juni 2011 Michael Væth 33