Supplement til power-point presentasjonen i medisinsk statistikk, forelesning 7 januar 2013.
Skrevet av Stian Lydersen 16 januar 2013

Vi antar at vårt utvalg er et tilfeldig og representativt utvalg for en større (evt tenkt) populasjon. Ved estimering, hypotesetesting og konfidensintervall er spørsmålet: Hva kan vi si om populasjonen ut fra vårt utvalg? Nedenfor er dette belyst med fire eksempler: Uavhengige (uparede) grupper med dikotomt utfall (Eksempel 1) og kontinuerlig (skalavariabel) eller ordinalt utfall (Eksempel 3). Matchede (parede) data med dikotomt utfall (Eksempel 2) kontinuerlig (skalavariabel) eller ordinalt utfall (Eksempel 4).

Eksempel 1.
Eksempel med to behandlingsgrupper og dikotomt utfall (for eksempel suksess versus ikke suksess): Her har vi to binomiske fordelinger.

Behandling * Kvalmeklasse Krysstabell

			Kvalme		
			lite eller ingen	betydelig	Total
Behandling	Nei	Antall	18	12	30
		%	60,0%	40,0%	100,0%
	Ja	Antall	24	5	29
		%	82,8%	17,2%	100,0%
Total		Antall	42	17	59
		%	71,2%	28,8%	100,0%

Nullhypotesen er at effekten (suksess-sansynligheten) er den samme i begge behandlingsgruppene: $p_1 = p_2$. I vårt utvalg observerer vi totalt 71,2% suksesser. Dersom suksess-sansynlighheten var den samme, ville vi forvente 71,2% * 30 = 21,4 istedenfor 18 suksesser i kontrollgruppen. De 4 forventede antallene vist i tabellen nedenfor:

Behandling * Kvalme Crosstabulation

			Kvalme		
			lite eller ingen	betydelig	Total
Behandling	Nei	Count	18	12	30
		Expected Count	21,4	8,6	30,0
	Ja	Count	24	5	29
		Expected Count	20,6	8,4	29,0
Total		Count	42	17	59
		Expected Count	42,0	17,0	59,0

P-verdien er sannsynligheten for å få våre observasjoner eller noe mer ekstremt, dersom nullhypotesen var sann. ”Hvor mye” avviker våre observasjoner fra de forventede under nullhypotesen? Et fornuftig mål på dette er å ta differansen mellom det observerte O_i og det forventede E_i i hver av de celle nr *i*, kvadrere denne, og dele på E_i. Ved å summere disse for de 4 cellene fås Pearson’s kjikvadratobservator:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$
$$= \frac{(18 - 21,4)^2}{21,4} + \frac{(12 - 8,6)^2}{8,6} + \frac{(24 - 20,6)^2}{20,6} + \frac{(5 - 8,4)^2}{8,4} = 3,72$$

P-verdien er sannsynligheten for at denne observatoren er lik 3,72 eller større, gitt at nullhypotesen var sann. Denne finner man vha tabeller eller PC-program, og her får vi P-verdi=0,54 (Kjikvadratfordeling med 1 frihetsgrad). Siden denne er større enn 0,05, vil vi ikke forkaste nullhypotesen ved signifikansnivå 0,05: Denne studien gir ikke grunnlag for å påstå at behandling er bedre enn kontroll ved signifikansnivå 0,05.

Pearson’s kjikvadrattest er OK hvis tallene ikke er for ”små”: Forventet antall bør være minst 5 i hver celle. Her er minste forventede antall lik 8,4, så det er OK. Vær klar over at hvis tallene er mindre, kreves bruk av andre metoder (som ligger utenfor pensum her).

Konfidensintervall for differanse mellom to sannsynligheter $p_1 - p_2$: Wald-intervallet (OK bare ved store tall) og Agresti-Caffo intervallet (alltid OK) er vist i power point presentasjonen. Her er det viktig å være klar over at i eksempelet ovenfor er observasjonene uavhengige, dvs uparet. Det forventes ikke at man husker utledningen eller formlene for disse intervallene, men man må vite at man må velge en metode for uparede data i et eksempel som dette. Merk at i eksempelet på slide 53 er Agresti-Caffo intervallet konsistent med Pearson's kjikvadrattest: Det 95% Agresti-Caffo intervallet (-0.007 til 0.432) inneholder 0 som en mulig verdi for $p_1 - p_2$, samtidig som Pearson's kjikvadrat p=0.054 er over 0.05.

Eksempel 2:
Ligaarden et al (BMC, 2010) rapporterer et forsøk der 16 pasienter med irritabel tarm (Irritable Bowel Syndrome – IBS) behandles i 3 uker med probioticumet L.plantarum MF1298, og 3 uker med placebo. Dette kalles en overkrysningsstudie. Forsøket var dobbelt blindt, og for hver pasient ble det randomisert om vedkommende skulle ha behandling eller placebo i første periode. Resultatene er (forenklet):

Lindring i behandlingsperiode * Lindring i placeboperiode Crosstabulation

Count

		Lindring i placeboperiode		
		ingen	OK	Total
Lindring i behandlingsperiode	ingen	4	1	5
	OK	7	4	11
Total		11	5	16

Her fikk 5 av 16 pasienter suksess i behandlingsperioden, og 11 av 16 pasienter fikk suksess i placeboperioden. Estimert suksess-sannsynlighet er hhv 5/16= 0.31 og 11/16=0.69, og differansen er 0.31-0.69=-0.38. Merk at tabellen er satt opp på en annen måte enn i Eksempel

1. Her kan vi ikke anta uavhengie observasjoner som i Eksempel 1. Tvert i mot er observasjonene paret. Mer må man vite at dette krever metoder for parede data ved hypotesetesting og konfidensintervall (men estimater for suksessannsynligheter og deres differanse er som for uparede data). Den mest aktuelle metoden for parede binomiske sannsynligheteter heter McNemar's test. For spesielt interesserte kan det nevnes at i dette eksempelet fås McNemar's p-verdi lik 0.034, og 95% konfidensintervall (-0.62 til -0.03), altså en statistisk signifikant effekt i disfavør av behandling.

Eksempel 3:
Sammenlikning avforventningsverdi i mellom to grupper, se slide 54. Gruppene har antall observasjoner, forventningsverdi, varians hhv n_1, μ_1, σ_1^2 og n_2, μ_2, σ_2^2. Differansen mellom gjennomsnittene, $\bar{x}_1 - \bar{x}_2$, brukes som estimat på differensen mellom forventningsverdiene, $\mu_1 - \mu_2$. Standardavvik for enkeltobservasjoner i hhv gruppe 1 og 2 er σ_1 og σ_2.

Standardavvik for estimatet $\bar{x}_1 - \bar{x}_2$ er $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, som estimeres ved $SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$, der S_1 og S_2 er empirisk standardavvik for (enkelt-) observasjoner i hver av gruppene.
Standardavviket for estimatet, *SE*, kalles standardfeil. Dersom observasjonene er (tilnærmet) normalfordelt, vil $\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{SE}$ ha en sannsynlighetsfordeling som ligner på standard normalfordeling (standard normalfordeling har forventningsverdi 0 og varians 1), nærmere bestemt en Student's t-fordeling. Dette brukes til å beregne en p-verdi for nullhypotesen $\mu_1 = \mu_2$, og til å beregne konfidensintervall for $\mu_1 - \mu_2$.

I dette tilfelle er ikke data tilnærmet normalfordelt: Q-Q plottene viser at i begge gruppene er det en eller flere "ekstreme" observasjoner som er betydelig større enn forventet hvis data var normalfordelt (grafene på venstre side av slide 59). Vi bør derfor ikke bruke Student's t til å sammenlikne forventningsverdiene. Student's p-verdi er 0.102 eller 0.079 hvis en antar lik eller ulik varians i gruppene (slide 58), men disse bør ikke brukes her, da data avviker vesentlig fra normalfordelingen.

Derimot er en ikke-parametrisk test egnet til å teste om varighet av sykehusopphold er likt fordelt i de to gruppene, versus at det er forskjøvet mot høyere verdier i en av gruppene. En

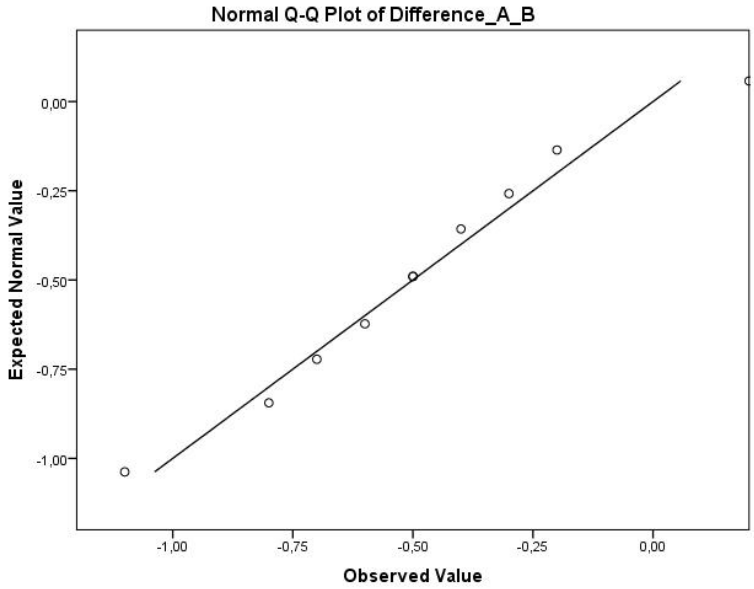
ikke-parametrisk test forutsetter ikke at data er tilnærmet normalfordelt, og forutsetter heller ingen annen såkalt parametrisk fordeing. Den aktuelle ikke-parametriske to-utvalgstesten kalles Wilcoxon-Mann-Whitney's test (noen bruker betegnelsen Mann-Whitney's test). Denne gir i dette eksempelet p=0.042 (slide 63). Altså statistisk signifikant lengre sykehusopphold i gruppe B, dersom en har valgt et signifikansnivå på 0.05 (fordi 0.042<0.05).

Eksempel 4:
Matchede par, kontinuering utfall. Slide 68: En skofabrikant vil sammenlikne slitasje for to materialer i skosåler. I alt 10 gutter får randomisert materiale A til venstre eller høyre fot, og materiale B til den andre foten. Observert slitasjedybde ble:

Case Summaries

	boy_no	wear_A	wear_B	Difference_A_B
1	1	13,2	14,0	-,8
2	2	8,1	8,8	-,7
3	3	10,8	11,3	-,5
4	4	14,1	14,3	-,2
5	5	10,8	11,9	-1,1
6	6	6,6	6,4	,2
7	7	9,4	9,8	-,4
8	8	10,8	11,4	-,6
9	9	8,8	9,3	-,5
10	10	13,2	13,5	-,3
Total N	10	10	10	10

Her må man bruke en paret analysemetode. En såkalt paret Student's t-test med tilhørende konfidensintervall baseres på differensen for hvert individ, se siste kolonne i tabellen ovenfor. Q-Q plottet for differensene blir som følger:



Her ligger verdiene nær en rett linje: Data er tilnærmet normalfordelt og vi kan trygt bruke Student's t-test. Resultatet blir:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	wear_A	10,580	10	2,4225	,7661
	wear_B	11,070	10	2,5272	,7992

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	wear_A - wear_B	-,4900	,3542	,1120	-,7434	-,2366	-4,375	9	,002

Estimert forskjell er -0.049, 95% konfidensintervall (-0.734 til -0.237), p=0.002. Altså en statistisk signifikant forskjell. Dersom vi (feilaktig) hadde brukt en (uparet) toutvalgs Student's t-test, ville vi fått estimert forskjell -0.049, 95% konfidensintervall (-2.82 til 1.84), p=0.66, altså ingen statistisk signifikant forskjell. Her er variasjonen mellom individene større enn variasjon mellom materialene (slide 68). Dermed er en matchet design, som var mulig her, en god design. I mange, kanskje de fleste, medisinske anvendelser er en matchet design imidlertid vanskelig eller umulig å gjennomføre.

Dersom data ikke var normalfordelt, ville en ikkeparametrisk test vært bedre. Den ville gitt p=0.008 (Wilcoxon's signed ranks test). Det er aldri direkte feil å bruke en ikkeparametrisk test selv om data er normalfordelt. Men Student's t-test har to fordeler fremfor en ikkeparametrisk test når data er normalfordelt:

1. Man får også et konfidensintervall for forskjellen.
2. Den har noe høyere statistisk styrke (evne til å påvise ne faktisk forskjell), dvs litt lavere sannsynlighet for type II feil, enn en ikke-parametrisk test. Dette gjelder særlig i små utvalg, som her.