NTNU

Det skapende universitet

HYPOTESETESTING
Del 1 av 3: Testing basert på ett utvalg

Gunnar Taraldsen
Institutt for nevromedisin

8. oktober 2008

Hypotesetesting del 1 av 3

— Del 1 tilsvarer pensum fra læreboken (Rosner, 2006, Kapittel 7:1-4,7,10,12-13), og omhandler tester basert på utvalg fra 1 populasjon.

— Del 2 omhandler tester basert på utvalg fra 2 populasjoner (Kapittel 8).

— Del 3 drøfter begrepene teststyrke og utvalgsstørrelse (Kapittel 7.5-6, Geir Jacobsen).

NTNU
Det skapende universitet

Hypotesetesting del 1 av 3

— Del 1 tilsvarer pensum fra læreboken (Rosner, 2006, Kapittel 7:1-4,7,10,12-13), og omhandler tester basert på utvalg fra 1 populasjon.
— Del 2 omhandler tester basert på utvalg fra 2 populasjoner (Kapittel 8).
— Del 3 drøfter begrepene teststyrke og utvalgsstørrelse (Kapittel 7.5-6, Geir Jacobsen).

Fremstillingen er basert på læreboken (Rosner, 2006), samt
— En tilsvarende presentasjon fra i fjor (Lydersen, 2007).
— Noen norske lærebøker (Aalen, 2006; Løvås, 2008).
— Undertegnedes tidligere studier (Lehmann and Romano, 2005; Taraldsen, 1997).

NTNU
Det skapende universitet

Hypotesetesting

"We find the defendant not guilty but not all that innocent, either."

Hypotesetesting: Generelle begrep

En hypotesetest består blant annet av følgende elementer

— En veldefinert nullhypotese H_0.

— En alternativ hypotese H_1.

— En regel som avgjør om H_0 forkastes ut fra de gitte observasjonene.

— En statistisk modell.

NTNU
Det skapende universitet

Binomisk test (Aalen, 2006, s.98-102)

La X være antall pasienter som foretrekker en ny type medisin mot migrene fremfor et tradisjonelt medikament. I et randomisert og dobbeltblindet forsøk kan det argumenteres for at X er binomisk fordelt med suksessansynlighet p. La $H_0 : p = p_0$ og $H_1 : p \neq p_0$ hvor $p_0 = 0.5$. Anta at et forsøk med $n = 8$ personer gir $x = 7$. En beregning gir p-verdien $\Pr^{p_0}(X \geq 7 \text{ eller } X \leq 1) = 7\%$, så H_0 forkastes ikke ved et 5% nivå.

NTNU
Det skapende universitet

Binomisk test (Aalen, 2006, s.98-102)

La X være antall pasienter som foretrekker en ny type medisin mot migrene fremfor et tradisjonelt medikament. I et randomisert og dobbeltblindet forsøk kan det argumenteres for at X er binomisk fordelt med suksessansynlighet p. La $H_0 : p = p_0$ og $H_1 : p \neq p_0$ hvor $p_0 = 0.5$. Anta at et forsøk med $n = 8$ personer gir $x = 7$. En beregning gir p-verdien $\Pr^{p_0}(X \geq 7 \text{ eller } X \leq 1) = 7\%$, så H_0 forkastes ikke ved et 5% nivå.

Generell regel for en α-nivå test (for eksempel $\alpha = 5\%$):

- Forkast H_0 dersom p-verdien er mindre eller lik α.
- p-verdien er gitt som sannsynligheten for å observere noe like eller mer ekstremt enn det som er observert når H_0 er sann.

NTNU
Det skapende universitet

Hypotesetesting: Generelle begrep

Beslutningsregelen i en hypotesetest er definert av

- En teststatistikk (testobservator) W som er en funksjon av observasjonene.
- Et forkastningsområde (kritisk område) R_w.
- Nullhypotesen H_0 forkastes dersom $w \in R_w$.

NTNU
Det skapende universitet

Hypotesetesting: Generelle begrep

Beslutningsregelen i en hypotesetest er definert av

— En teststatistikk (testobservator) W som er en funksjon av observasjonene.

— Et forkastningsområde (kritisk område) R_w.

— Nullhypotesen H_0 forkastes dersom $w \in R_w$.

Et eksempel er gitt ved $w =$ p-verdien, og $R_w = [0, \alpha]$. Null hypotesen forkastes ved lave p-verdier. Det fremgår at p-verdien er en normalisert testobservator, og derfor foretrekkes denne i praksis.

NTNU
Det skapende universitet

Egenskaper til en hypotesetest

Sannheten

Beslutning

	H_0	H_1
Behold H_0	OK	Type II feil
Forkast H_0 (påstå H_1)	Type I feil	OK

NTNU

Det skapende universitet

Egenskaper til en hypotesetest

		Sannheten	
		H_0	H_1
Beslutning	Behold H_0	OK	Type II feil
	Forkast H_0 (påstå H_1)	Type I feil	OK

Definisjon: En hypotesetest er en α-nivå test dersom sannsynligheten for type I feil er mindre eller lik α. Styrken til en test er $1 - \beta$ hvor β er sannsynligheten for type II feil.

NTNU
Det skapende universitet

Type I feil er mest alvorlig

Type I feil sees som mest alvorlig, og derfor inngår sannsynligheten for type I feil i definisjonen av en α-nivå test. Blant de mulige α-nivå testene så fortrekkes den testen som har størst styrke.

NTNU
Det skapende universitet

Type I feil er mest alvorlig

Type I feil sees som mest alvorlig, og derfor inngår sannsynligheten for type I feil i definisjonen av en α-nivå test. Blant de mulige α-nivå testene så fortrekkes den testen som har størst styrke.

NTNU
Det skapende universitet

Diagnostiske tester er hypotesetester

Spesifisiteten (Rosner, 2006, s.58) er sannsynligheten for at symptomet (gruppe av symptomer) ikke er der gitt at personen ikke har sykdommen. Sensitiviteten til et symptom er sannsynligheten for at symptomet er der gitt at personen har sykdommen.

NTNU
Det skapende universitet

Diagnostiske tester er hypotesetester

Spesifisiteten (Rosner, 2006, s.58) er sannsynligheten for at symptomet (gruppe av symptomer) ikke er der gitt at personen ikke har sykdommen. Sensitiviteten til et symptom er sannsynligheten for at symptomet er der gitt at personen har sykdommen.

— H_0: Pasienten har ikke sykdommen. H_1: Pasienten har sykdommen.

— Regel: Forkast H_0 dersom symptomet er der.

— For en test med gitt spesifisitet (= 1 - testnivå), så fortrekkes en test med stor sensitivitet (= styrke).

— Falsk positiv = Type I feil. Falsk negativ = Type II feil.

NTNU
Det skapende universitet

Konfidensintervall gir hypotesetest

La $[t_L, t_U]$ være et $(1-\alpha)$-nivå konfidensintervall for en parameter τ.

— $H_0 : \tau = \tau_0$ og $H_1 : \tau \neq \tau_0$.

— Regel: Forkast H_0 dersom $\tau_0 \notin [t_L, t_U]$.

NTNU
Det skapende universitet

Konfidensintervall gir hypotesetest

La $[t_L, t_U]$ være et $(1-\alpha)$-nivå konfidensintervall for en parameter τ.

- $H_0 : \tau = \tau_0$ og $H_1 : \tau \neq \tau_0$.
- Regel: Forkast H_0 dersom $\tau_0 \notin [t_L, t_U]$.

Dette gir en α-nivå test, fordi

$$\Pr^{\tau_0}(\tau_0 \notin [T_L, T_U]) = 1 - \Pr^{\tau_0}(\tau_0 \in [T_L, T_U]) \leq \alpha \tag{1}$$

En kan også gå den motsatte veien, dvs en kan utlede konfidensintervall fra en (familie) hypotesetester.

NTNU
Det skapende universitet

Eksempel på t-test

Har barn av mødre med lav sosioøkonomisk status (SØS) annen forventet fødselsvekt enn andre? Vi samler inn vektdata fra $n = 100$ etterfølgende terminfødsler fra et sykehus i et område med lav SØS. Resultatet er oppsummert ved empirisk middel $\overline{x} = 115$oz og varians $s = 24$oz. To hypoteser

— H_0: Forventet fødselsvekt μ ved dette sykehuset er den samme som totalt i USA, dvs lik $\mu_0 = 120\text{oz} = 3400\text{g}$.

— $H_1 : \mu \neq \mu_0$.

NTNU
Det skapende universitet

Eksempel på t-test

- Type I feil: Konkluder at forventet fødselsvekt ved SØS sykehuset er ulik gjennomsnittet, når den i virkeligheten er den samme.
- Type II feil: Behold hypotesen om at forventet fødselsvekt ved SØS sykehuset er lik gjennomsnittet, når den i virkeligheten er ulik.

NTNU
Det skapende universitet

Eksempel på t-test

— Et 95%-konfidensintervall er gitt ved

$$\overline{x} \pm t_{n-1,1-\alpha/2} s/\sqrt{n} = (115 \pm 1.98 \cdot \frac{24}{\sqrt{100}})\text{oz} \tag{2}$$

dvs $[\mu_L, \mu_U] = [110.2, 119.8]$oz.

— Konklusjon: $H_0 : \mu = 120$oz forkastes i testen med nivå $\alpha = 5\%$.

NTNU
Det skapende universitet

Eksempel på t-test

Utledningen av konfidensintervallet som ble benyttet bygger på en antagelse om at $T = (\overline{X} - \mu)/(S/\sqrt{n})$ er student t-fordelt med $n - 1$ frihetsgrader når $\mu = \mu_0$.

— Den observerte verdien til T er $t = (\bar{x} - \mu_0)/(s/\sqrt{n}) = -2.08$.

— Sannsynligheten for lik eller mer ekstreme observasjoner er $\Pr^{\mu_0,\sigma}(|T| \geq 2.08) = 4\%$, så p-verdien er 4%.

— Som tidligere betyr dette at H_0 forkastes ved et 5% nivå, men i en test med nivå 1% forkastes ikke H_0.

— Hvorfor er p-verdien å foretrekke fremfor t-verdien som testobservator? Begrunn dette selv!

NTNU

Det skapende universitet

Rapportering i hypotesetesting

— **Rapporter p-verdien** fremfor for eksempel t-verdien fordi p-verdien er en normalisert test statistikk.

— **Rapporter konfidensintervallet** fordi dette gir nøyaktigheten til estimatet av effekten en ser på. I følge *ISO Guide to the Expression of Uncertainty in Measurement* (GUM) skal også standard feil u rapporteres. I foregående eksempel er $u = s/\sqrt{n}$.

— Et statistisk signifikant resultat behøver ikke å være vitenskapelig signifikant.

NTNU
Det skapende universitet

Statistisk signifikans medfører ikke medisinsk relevans

Følgende sitat fra læreboken utdyper dette (Rosner, 2006, p.220)

> *In writing up the results of a study, a distinction between scientific and statistical significance should be made, because the two terms do not necessarily coincide. The results of a study can be statistically significant but still not be scientifically important. Conversely, some statistically nonsignificant results can be scientifically important, encouraging researchers to perform lager studies.*

NTNU
Det skapende universitet

Statistisk signifikans medfører ikke medisinsk relevans

Følgende sitat fra læreboken utdyper dette (Rosner, 2006, p.220)

> *In writing up the results of a study, a distinction between scientific and statistical significance should be made, because the two terms do not necessarily coincide. The results of a study can be statistically significant but still not be scientifically important. Conversely, some statistically nonsignificant results can be scientifically important, encouraging researchers to perform lager studies.*

Moral: **Oppgi konfidensintervall i tillegg til p-verdi!**

NTNU
Det skapende universitet

Statistisk signifikans medfører ikke medisinsk relevans

Følgende sitat fra Vancouver-retningslinjene (1978, IV.A.6.c. Statistics, `http://www.icmje.org`) utdyper også dette

> *When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important information about effect size.*

NTNU
Det skapende universitet

Statistisk signifikans medfører ikke medisinsk relevans

Følgende sitat fra Vancouver-retningslinjene (1978, IV.A.6.c. Statistics, `http://www.icmje.org`) utdyper også dette

> *When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important information about effect size.*

Moral: **Oppgi konfidensintervall i tillegg til p-verdi!**

NTNU
Det skapende universitet

Referanser

Aalen, O. (Ed.) (2006). *Statistiske metoder i medisin og helsefag*. Gyldendal.

Lehmann, E. and J. Romano (2005). *Testing statistical hypotheses*. Springer.

Løvås, G. (2008). *Statistikk for universiteter og høgskoler*. Universitetsforlaget.

Lydersen, S. (2007, Foredrag 10. oktober). Hypotesetest, ett utvalg. *KLMED8004, Medisinsk statistikk Del I*.

Rosner, B. (2006). *Fundamentals of biostatistics*. Thomson.

Taraldsen, G. (1997). *Grunnkurs i statistikk og sannsynlighetsteori*.

NTNU
Det skapende universitet