ECON 2130

HG
Mai 2011

Supplement til siste forelesning 2011 vår

På siste forelesning fikk jeg beklageligvis ikke tid til å ta noen eksempler på testing og bruk av p-verdier. Derfor dette supplementet med 4 eksempler på test-diskusjoner.

Eksempel 1 - Er gjennomsnittshøyden for kvinner i Norge økende?

Det er velkjent at gjennomsnittshøyden for mennesker i Europa har økt jevnlig over tid under 19-hundretallet. Er denne tendensen fortsatt til stede? Siden vi har samlet inn data for mødres og døtres høyder, kan vi for eksempel spørre om dette datamaterialet tyder på en fortsatt økende tendens blant kvinner i Norge når det gjelder høyde.

La for datter nr. *i* i utvalget X_i betegne mors høyde og Y_i datters høyde for $i = 1, 2, \ldots, n$, der $n = 88$ - idet vi slår sammen data fra 2010 og 2011. Vi antar uavhengighet og felles fordeling for X_i-ene og tilsvarende for Y_i-ene. Skriv forventningene som $\mu_M = E(X_i)$ og $\mu_D = E(Y_i)$ for døtrene. En økende tendens kan i denne modellen uttrykkes som at $\mu_D > \mu_M$, eller, mao, $\mu_D - \mu_M > 0$. At det ikke er noen økende tendens uttrykkes ved $\mu_D - \mu_M = 0$. Det er naturlig nå å tolke problemstillingen som et hypoteseprøvingsproblem der hypotesen $\mu_D - \mu_M > 0$ plasseres i alternativet[1]:

$$H_0 : \mu_D - \mu_M \leq 0 \quad \text{mot} \quad H_1 : \mu_D - \mu_M > 0$$

Som mål på hvor mye evidens det er i data for påstanden $H_1 : \mu_D - \mu_M > 0$, vil vi beregne p-verdien basert på data og en passende test. Desto lavere p-verdien er, desto mer evidens er det i data for H_1.

La for par nr. *i* høydeforskjellen mellom mor og datter være, $D_i = Y_i - X_i$. Det er nå rimelig å anta (som vil være vår modell) at $D_1, D_2, \ldots, D_n$ er ~ uid (uavhengige og identisk fordelt) med

[1] Vi kunne like gjerne brukt likhetstegn i H_0 istedenfor $\leq$, som virker mer logisk i og med at muligheten $\mu_D - \mu_M < 0$ neppe er aktuell. Nå viser det seg imidlertid at samme test med samme egenskaper, samme konklusjon og samme p-verdi kan benyttes enten H_0 formuleres med $=$ eller med $\leq$. Grunnen til at vi bruker $\leq$ er at hypotesene nå har samme form som er behandlet i kapittel 6 i Løvås og i testoversikten på nettet. Det gjør altså ikke noe forskjell i praksis om H_0 omfatter noen uinteressante mulige verdier for $\mu_D - \mu_M$.

felles forventning, $E(D_i) = \theta = E(Y_i - X_i) \overset{\text{Regel 4.12}}{=} E(Y_i) - E(X_i) = \mu_D - \mu_M$ (der vi har innført parameteren θ for $\mu_D - \mu_M$), og $\text{var}(D_i) = \sigma_D^2$. Både $\theta = \mu_D - \mu_M$ og σ_D^2 anses som ukjente parametre, og problemet er å teste

$$H_0 : \theta \le 0 \quad \text{mot} \quad H_1 : \theta > 0$$

Vi ser at vi er i situasjon 3 i tabell 2 i testoversikten på nettet og at testobservatoren er

$$Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}, \text{ der } \theta_0 = 0, \quad \hat{\theta} = \bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, \text{ og } SE(\hat{\theta}) = \frac{S_D}{\sqrt{n}}, \text{ der } S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2}$$

Dermed $Z = \dfrac{\bar{D}}{S_D/\sqrt{n}} = \dfrac{\bar{D}}{S_D}\sqrt{n}$. Det er klart (jfr begrunnelsen for tabell 1 i testoversikten) at vi skal forkaste H_0 for tilstrekkelig store observerte verdier av Z (dvs Z_{obs}, også kalt z_o i testoversikten), og at den kritiske verdien basert på signifikansnivået α, er z_α (den øvre α-kvantilen i N(0, 1)-fordelingen).

Testkriteriet for en (tilnærmet) α-nivå test blir således

(1) Forkast H_0 hvis $Z > z_\alpha$

Dette kriteriet er ekvivalent[2] med testkriteriet

(2) Forkast H_0 hvis $\hat{\alpha} < \alpha$

der $\hat{\alpha}$ er p-verdien som beregnes ut fra data som $P_{\theta=\theta_0}(Z > z_o)$ der z_0 betegner den observerte verdien av Z som er det tallet vi får når vi setter data inn i Z (også kalt Z_{obs} noen steder). Siden Z er (tilnærmet) $N(0,1)$-fordelt når $\theta = \theta_0 = 0$, kan vi bruke tabell D3 bak i Løvås til å finne $\hat{\alpha}$.

Beregning: Ut fra data finner vi (sjekk selv med data på http://folk.uio.no/haraldg/) ved Excel de observerte verdiene,

$$\bar{D}_{obs} = 0.8295, \quad S_{D,obs} = 6.4237,$$

som gir observert verdi av Z: $\quad z_0 = \dfrac{0.8295}{6.4237} \cdot \sqrt{88} = 1.2114$

Av dette finner vi p-verdien: $\hat{\alpha} = P_{\theta=0}(Z > 1.21) \overset{\text{Tabell D3 i Løvås}}{\approx} 1 - 0.8864 = 0.1131.$

Diskusjon: Ut fra kriteriet (2) vil en p-verdi på 11.3% tilsi at vi ikke kan forkaste H_0 om vi bruker signifikansnivå for eksempel 5% eller 10%. Hadde vi akseptert et signifikansnivå på 12% (som ville vært uvanlig), ville vi kunnet forkaste H_0. Konvensjonelt anses en p-verdi på 11.3% vanligvis et for tynt evidensgrunnlag for å påstå $H_1 : \theta > 0$. Resultatet anses ikke-signifikant.

---

[2] Følger av definisjonen av $\hat{\alpha}$ som det minste nivået som leder til forkastning av H_0.

På den annen side ligger 11.3% såpass nær signifikansområdet at årsaken til ikke-forkastning godt kunne være at θ er bare litt større enn 0 slik at H_1 bare så vidt er sann – noe som slett ikke er utenkelig over bare en generasjons tidsforskjell. Det ville kreve langt flere observasjoner enn i vårt materiale for å kunne avsløre en θ som er positiv men nær 0. Hvis for eksempel den sanne verdien av θ var 0.5 cm (dvs. at gjennomsnittshøyden har økt med 0.5 cm fra mødre- til dattergenerasjonen) , så ville sannsynligheten for å oppdage dette (dvs forkaste H_0) med en 5% nivå test ikke være større enn ca 18% [3]. For å ha en rimelig sjanse å avsløre det i dette tilfelle, ville vi derfor trenge langt flere observasjoner enn våre 88 observasjonspar. En naturlig konklusjon på analysen i en eventuell rapport ville derfor være å si at det ikke er informasjon nok i data til å kunne påstå at det har vært en økning av gjennomsnittshøyden sammen med en anbefaling å hente inn flere observasjoner for å få avklart problemet.

Eksempel 2 - Om T-testing

Grunnen til at vi kunne bruke en *Z*-test i eksempel 1 (der vi regner som om populasjons-standardavviket, σ_D, for D_i var kjent lik estimatet, $S_{D,obs}$), er at vi hadde så mange observasjoner som 88. Hvis vi hadde hatt færre observasjoner, for eksempel bare $n=10$ istedenfor 88, ville ikke lenger *Z*-testen være velbegrunnet siden tilleggs-usikkerheten som oppstår ved å erstatte σ_D med estimatet $S_{D,obs}$, ikke lenger er neglisjerbar. Vi kan i dette tilfellet bruke en T-test i stedet, dersom det er rimelig å forutsette i tillegg til forutsetningene i eksempel 1 at enkeltobservasjonene, D_i, er normalfordelte, $D_i \sim N(\theta, \sigma_D)$, $i=1,2,\ldots,n$. La problemet uttrykt ved H_0, H_1 være det samme som i eksempel 1. Testobservatoren vil nå være den samme som *Z* i eksempel 1 (bare at det nå er mer vanlig å kalle den *T*),

$$T = Z = \frac{\hat{\theta}-\theta_0}{SE(\hat{\theta})} = \frac{\overline{D}}{S_D/\sqrt{n}} = \frac{\overline{D}}{S_D}\sqrt{n},$$

men fordelingen for *T* er nå t_{n-1}-fordelt (istedenfor $N(0,1)$-fordelt) hvis $\theta=\theta_0$, noe som gjør at den kritiske verdien blir endret til $t_{n-1,\alpha}$ (istedenfor z_α). Den kritiske verdien, $c=t_{n-1,\alpha}$, er nemlig løsningen av ligningen $P_{\theta=\theta_0}(\text{forkast } H_0) = P_{\theta=\theta_0}(T>c) = \alpha$ med hensyn på *c*. Denne testen er oppsummert under situasjon 2 i tabell 2 i testoversikten.

For eksempel, hvis vi ønsker nivå 5% ($\alpha=0.05$) og $n=10$, blir *Z*-testen som i (1) ovenfor fortsatt

(3) Forkast H_0 hvis $Z > z_{0.05} = 1.645$

---

[3] Dette er en styrkeberegning som du kan sjekke selv: Anta den sanne verdien av θ er $\theta=0.5$ og at vi velger nivå 5%. Da er den kritiske verdien $z_{0.05}=1.645$, slik at H_0 forkastes hvis $Z>1.645$. Merk at hvis $\theta=0.5$, er ikke lenger *Z* ~ N(0, 1). Derimot er $W = \frac{\overline{D}-0.5}{S_D}\sqrt{n} \sim N(0,1)$, og $Z = W + \frac{0.5}{S_D}\sqrt{n}$. Dermed

$$P_{\theta=0.5}(\text{forkaste } H_0) = P(Z>1.645) = P\left(W > 1.645 - \frac{0.5}{S_D}\sqrt{88}\right) \approx P(W>0.91) = 0.18,$$

der vi, for enkelthets skyld, har erstattet S_D med estimat-verdien 6.4237.

mens den mer korrekte T-testen blir (der altså *Z* og *T* er like)

(4) Forkast H_0 hvis $T > t_{9,\,0.05} = 1.833$ (fra tabell D5 bak i Løvås).

Som tabell 2 i testoversikten viser, kan p-verdien beregnes ved $\hat{\alpha} = P_{\theta=\theta_0}(T > t_o)$, der t_o betegner den observerte verdien av *T* beregnet ut fra data. Hvis vi i tillegg til modellen i eksempel 1 antar at D_i er normalfordelt (som er en rimelig forutsetning i dette tilfellet), er testobservatoren $T = Z \sim t_{87}$-fordelt hvis $\theta = \theta_0$. *t*-fordelingen kan beregnes ved TDIST-funksjonen i Excel. Bruker vi den og den observert verdien av *T*, $t_0 = z_0 = 1.2114$, får vi den mer korrekte p-verdien:

$$\hat{\alpha} = P_{\theta=\theta_0}(T > t_o) = P_{\theta=\theta_0}(T > 1.2114) = 0.115$$

som er praktisk talt det samme som den tilnærmete p-verdien, 0.113, vi fikk i eksempel 1 basert på N(0, 1)-fordelingen. Den illustrerer at tilnærmelsen brukt i eksempel 1 er fullt tilfredsstillende når vi har så mange observasjoner som 88. For sammenligningens skyld anta at vi hadde oppnådd samme observert verdi av *T* (1.2114), men bare basert på $n = 10$ observasjoner. Da ville den mer korrekte p-verdien være basert på t_9-fordelingen som ifølge TDIST i Excel blir:

$$\hat{\alpha} = P_{\theta=\theta_0}(T > t_o) = P_{\theta=\theta_0}(T > 1.2114) = 0.128$$

mens den tilnærmete p-verdien fra eksempel 1 fortsatt ville være 0.113. Her er forskjellen ikke lenger neglisjerbar.

Til eksamen har vi ikke tilgang på Excel. Hvis vi trenger å beregne en p-verdi basert på en t-fordeling, kan vi bare bruke den begrensete informasjonen gitt i tabell D5 bak i Løvås. I eksempel 6.26 (side 250) gir Løvås eksempel på en slik bestemmelse uten forklaring, og flere studenter har lurt på hvordan han kommer fram til svaret. Problemet er å undersøke om data (8 observasjoner) tyder på at forventet vekt på hamburgere er mindre enn påstått vekt 100g. Hvis X_i betegner vekt av hamburger *i* i utvalget, benyttes modellen: $X_1, X_2, \ldots, X_n \sim uid$ ($n = 8$) og normalfordelt, $X_i \sim N(\mu, \sigma)$. Vi skal teste $H_0 : \mu \geq 100$ mot $H_1 : \mu < 100$. Vi er nå i situasjon 2 i tabell 2 i testoversikten på nettet med alternativ 2 i tabell 1. I det generelle opplegget blir $\theta = \mu, \quad \theta_0 = \mu_0 = 100, \quad \hat{\theta} = \hat{\mu} = \bar{X}$, og $SE(\hat{\theta}) = \dfrac{S}{\sqrt{n}}$. Testobservatoren

$$T = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} = \frac{\bar{X} - 100}{S}\sqrt{8}$$

er t-fordelt med 7 frihetsgrader hvis den sanne verdien av μ er $\mu = \theta_0 = 100$. En α-nivå test består i å forkaste H_0 hvis *T* blir tilstrekkelig liten, slik at testkriteriet er

Forkast H_0 hvis $T < -t_{n-1,\,\alpha}$

der den kritiske verdien, $-t_{n-1,\,\alpha}$, er løsningen av ligningen
$P_{\mu=100}(\text{forkast } H_0) = P_{\mu=100}(T < c) = \alpha$ med hensyn på *c*. P-verdien beregnes ved $P_{\mu=100}(T < t_o)$ der t_o er den observerte verdien av *T*. Løvås finner $t_o = -2.43$, slik at p-verdien blir

$\hat{\alpha} = P_{\mu=100}(T < -2.43)$ der $T \sim t_7$-fordelt hvis $\mu = 100$

Denne anslår Løvås til ≈ 0.02 uten forklaring. Han har da brukt informasjonen om t_7-fordelingen i tabell D5 bak i Løvås. Tabellen gir $t_{7,\alpha}$ som oppfyller $P(T > t_{7,\alpha}) = \alpha$ for noen utvalgte α. Dermed kan vi plukke ut følgende sannsynligheter fra t_7-fordelingen:

$P(T > t)$	0.25	0.1	0.05	0.025	0.01	0.005
t	0.711	1.415	1.895	2.365	2.998	3.499

Siden t-fordelingen er symmetrisk om 0, blir p-verdien, $\hat{\alpha} = P(T < -2.43) = P(T > 2.43)$. Tabellen gir $P(T > 2.365) = 0.025$ og $P(T > 2.998) = 0.01$, som tilsier at $P(T > 2.43)$ ligger mellom 0.01 og 0.025 – dvs. i nærheten av 0.02 tatt på øyemål som er Løvås' anslag. (Merk at lineær interpolasjon ville antakelig gi et bedre anslag, men er ikke pensum slik at "øyemålsmetoden" er fullt akseptabel til eksamen.)

En p-verdi på 2% vil normalt tolkes som sterk evidens for H_1 siden valgte signifikansnivåer helt ned mot 2% ville lede til forkastning av H_0.

Merknad. I regresjonsmodellen som er behandlet i pensum vil testing av en av parametrene i forventningen til *Y* være en *t*-test med det samme opplegget som *t*-testing av μ.

Testobservatoren har samme form som før, $T = \dfrac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$. Har man derfor regnet ut de observerte verdiene av $\hat{\theta}$ og $SE(\hat{\theta})$, har man nok informasjon til å kunne gjennomføre testen. Den eneste forskjellen er at man bruker t-fordeling med $n-2$ frihetsgrader istedenfor $n-1$. (Se regneeksempel i regresjon II notatet (eller i Løvås kapittel 7)). Det kan nevnes at ifølge videregående teori blir antall frihetsgrader bestemt av antall ukjente parametre i forventningen til responsen. Dette antallet trekkes fra *n*. I regresjonsmodellen er det to ukjente parametre i forventningen, mens i μ-modellen er det bare en ukjent parameter i forventningen (μ).

Eksempel 3 P-verdier er nyttige

P-verdien, $\hat{\alpha}$, er definert som det minste signifikansnivået som leder til forkastning av H_0. $\hat{\alpha}$ er bestemt av data og testen som benyttes. Hvis det er bare en kritisk verdi i testen, kan $\hat{\alpha}$ beregnes som den α som gjør at den kritiske verdien faller sammen med den observerte verdien av testobservatoren. Hvis testen har et valgt signifikansnivå på α, er testen ekvivalent med testkriteriet: "Forkast H_0 hvis $\hat{\alpha} < \alpha$". Dette betyr at hvis vi får oppgitt p-verdien for en test, kan vi gjennomføre testen uten å kjenne hverken til testobservator eller kritisk verdi. Det er nok å sjekke om p-verdien er mindre eller ikke det nivået som vi har valgt (α). Hvis p-verdien er mindre enn α forkaster vi H_0.

For eksempel: I eksempel 2 påsto jeg at forutsetningen at differensen D_i fra eksempel 1 er normalfordelt, er en rimelig forutsetning. Det er mange tester for å sjekke statistisk realismen

av forutsetningen. Disse testene er ofte litt kompliserte og langt utenfor pensum å kjenne til. Imidlertid er det pensum å forstå betyningen av en p-verdi, slik at om vi få oppgitt en p-verdi basert på en slik test, kan vi likevel til en viss grad vurdere realismen av forutsetningen. Excel har ikke implementert noen slike tester – så vidt jeg vet – men programmet STATA (som brukes i Stat 2 – kurset) har flere. For eksempel bruker vi Shapiro-Wilk testen i STATA, som er anerkjent som en god test, på de 88 - D_i -observasjonene i eksempel 1, viser utskriften at p-verdien er 0.772. Nullhypotesen (H_0) er at D_i -ene er trukket fra en eller annen normalfordeling mens alternativet (H_1) er at de er trukket fra en fordeling som ikke er normal. Denne p-verdien ligger langt unna signifikante verdier - slik at vi kan konkludere:

Det er ikke noe evidens i data som tyder på at forutsetningen om normalfordeling er urealistisk.

Andre normalfordelingstester i STATA viser seg å gi lignende resultater.

Eksempel 4 Tosidig testing i en binomisk modell

Anta vi er interessert i å sjekke om en gitt terning er rettferdig med hensyn på å produsere seksere. Vi kaster terningen $n = 100$ ganger og registrerer antall seksere vi får. La X være antall seksere vi får. En åpenbart (hvorfor?) rimelig modell er:

$$X \sim \text{bin}(n, p)$$

der p er sannsynligheten for å få sekser i et enkelt kast, og $n = 100$. Hvis terningen er rettferdig m.h.p. seksere, er $p = 1/6$. Dette[4] vil utgjøre vår nullhypotese, H_0 . Hvis $p \neq 1/6$ anser vi terningen ikke rettferdig m.h.p. seksere. Vi skal altså teste

$$H_0 : p = p_0 \text{ mot } H_1 : p \neq p_0$$

der $p_0 = 1/6$. Vi er nå i en situasjon beskrevet i tabell 3 i testoversikten kombinert med alternativ 3 i tabell 1. I det generelle opplegget er

$$\theta = p, \qquad \theta_0 = p_0 = 1/6, \qquad \hat{\theta} = \hat{p} = \frac{X}{n} \text{ og } SE(\hat{\theta}) = \sqrt{\frac{p_0(1-p_0)}{n}} \text{ hvis}^5 \ p = p_0$$

Betingelsen for å kunne benytte normaltilnærmelsen, $\text{var}(X) \geq 5$ er klart oppfylt siden $\text{var}(X) = 100p(1-p)$ er godt over 5 for p i nærheten av 1/6. Testobservatoren er

$$Z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} = \frac{X/n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{n\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

[4] Ved tosidige problemer plasseres alltid likhetsalternativet ($\theta = \theta_0$) i H_0 . Vi har altså ved tosidige problemer ikke det samme dilemmaet som oppstår i ensidige problemer om hvilken av de to hypotesene som skal utgjøre H_0 .

[5] Jamfør merknad 4 etter tabell 3 i testoversikten.

som er (tilnærmet) standard normalfordelt, $Z \overset{\text{tilnærmet}}{\sim} N(0,1)$, hvis $p = p_0$. Velger vi signifikansnivå 5%, får vi de to kritiske verdiene, $\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.96$, fra $N(0,1)$-fordelingen. Testkriteriet for vår 5%-nivå test blir dermed (jamfør tabell 1 i testoversikten)

Forkast H_0 hvis $Z < -1.96$ eller $Z > 1.96$

eller

Forkast H_0 hvis $\dfrac{X - np_0}{\sqrt{np_0(1-p_0)}} < -1.96$ eller $\dfrac{X - np_0}{\sqrt{np_0(1-p_0)}} > 1.96$

For å få en mer praktisk anvendelig forkastningsregel i dette tilfellet, kan det være en ide å overføre kriteriet til et kriterium for X direkte. Kriteriet er klart ekvivalent med

Forkast H_0 hvis $X < np_0 - 1.96 \cdot \sqrt{np_0(1-p_0)}$ eller $X > np_0 + 1.96 \cdot \sqrt{np_0(1-p_0)}$

som ved innsetting av $n = 100$ og $p_0 = 1/6$ gir

Forkast H_0 hvis $X < 9.36$ eller $X > 23.97$

eller, siden X bare kan ta hele tall som verdier,

(*) Forkast H_0 hvis $X \le 9$ eller $X \ge 24$

Vi har nå overført testen på en enkel form. Vi har fått en regel som sier at X-verdier blant tallene $10,11,12,\ldots,23$ er forenelig med hypotesen at terningen er rettferdig m.h.p. seksere, mens X-verdier utenfor disse gir sterk evidens (med nivå tilnærmet 5%) for at terningen ikke er rettferdig.

Det nominelle nivået vi har brukt, $\alpha = 0.05$, er pga diverse tilnærmelser og tilpasninger, bare tilnærmet. Siden Excel kan beregne binomiske sannsynligheter, kan vi nå bestemme signifikansnivået for testen i (*) mer eksakt. La α_s betegne det sanne nivået for testen. Ved hjelp av BINOMDIST-funksjonen i Excel finner vi (sjekk selv):

$$\alpha_s = P_{p=1/6}(\text{forkaste } H_0) = P_{p=1/6}(X \le 9) + P_{p=1/6}(X \ge 24) = P_{p=1/6}(X \le 9) + 1 - P_{p=1/6}(X \le 23) =$$
$$= 0.021 + 1 - 0.962 = 0.059$$

Vi ser at det nominelle nivået gir en rimelig god tilnærming til det sanne nivået i dette tilfellet.