Fasit endelig eksamen STK1000

Høsten 2011

Oppgave 1

(*a*) $X \sim \text{bin}(150, 0.08)$, $\mu_X = 150 \cdot 0.08 = 12$, $\sigma_X = \sqrt{150 \cdot 0.08 \cdot 0.92} = 3.32$

(*b*) Tilnærmet $X \sim N(12, 3.32)$. $P(X \geq 15) = P(Z \geq 0.90) = 1 - P(Z < 0.90) = 1 - 0.8159 = 0.1841$ (her er $Z \sim N(0,1)$). Normalfordelingen regnes som god tilnærming til $\text{bin}(n,p)$ når $np > 10$ og $n(1-p) > 10$. Oppfyllt her.

Oppgave 2

(*a*) Parret fordi to påfølgende fredager i samme måned er mer like enn tilfeldige fredager. Ved å bruke differansen mellom 6. og 13. i samme måned reduserer vi effekten av sesongvariasjoner i biltrafikken. Se s. 410 i boken. Antakelser: samme antakelser som for vanlig ett-utvalgs-t-test skal her gjelde for differansene. Bør ha uavhengige, tilnærmet normalfordelte differanser. La X = trafikk den 6. − trafikk den 13., dvs. differanse innen hvert par. Antar X tilnærmet $\sim N(\mu, \sigma)$. Parameteren av interesse er da forventet differanse μ. Hvis $\mu = 0$ er det ingen forskjell.

(*b*) $H_0 : \mu = 0$ mot $H_a : \mu > 0$. Vi skal finne ut om folk holder seg hjemme den 13. Derfor velger vi $H_a : \mu > 0$, som betyr at det er forventet færre biler ute den 13. enn den 6. Estimat: $\hat{\mu} = \bar{x} = 2022$ (Mean for Difference i utskriften). Estimert standardavvik til $\bar{x}$ er $s/\sqrt{5} = 1539/\sqrt{5} = 688$ (s finnes som StDev for Difference i utskriften).

(*c*) Testobservator $t = \frac{\bar{x}-0}{s/\sqrt{5}} = \frac{2022}{688} = 2.94$ som stemmer med utskriften. Vi har 5-1=4 frihetsgrader. P-verdien finnes som $P(T > 2.94)$ der T er t-fordelt med 4 frihetsgrader. Fra tabellen finner vi at 0.02<P-verdi<0.025, som stemmer med P=0.021 fra utskriften. Denne P-verdien er mindre enn 0.05, og vi forkaster H_0. Forventet trafikk på M25 er mindre på fredag den 13. enn på fredagen før (P=0.021).

Oppgave 3

(*a*) Data (x_i, y_i), $i = 1, \ldots, 54$, for PCB-innhold og skalltykkelse. Modell:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

der ϵ_i er uavhengige og normalfordelte $N(0, \sigma)$, $i = 1, \ldots, 54$. Parametre er β_0, β_1 og σ. Estimater $b_0 = 0.377$, $b_1 = -0.000309$ og $s = 0.05925$. For $x^* = 200$ får vi $\hat{\mu}_y = 0.31483$.

(*b*) $H_0 : \beta_1 = 0$ mot $H_a : \beta_1 \neq 0$. Testobservator $t = b_1/\mathrm{SE}_{b_1} = -2.43$. P-verdi fra utskrift 0.019, som er mindre enn 0.05. Med signifikansnivå 0.05 får vi forkastning av H_0 og konkluderer med at det er en signifikant lineær sammenheng mellom PCB og eggeskalltykkelse.

(*c*) r^2 angir andelen av variasjonen i y som kan forklares av regresjonen av y på x. Med $r^2 = 0.102$ kan bare ca. 10% av variasjonen i skalltykkelse forklares av den lineære sammenhengen med PCB-innhold. Dette gir oss svært usikre prediksjoner.

(*d*) Har: California, $N(\mu_1, \sigma)$: $n_1 = 54$, $\bar{x}_1 = 203.04$, $s_1 = 64.02$ og Florida, $N(\mu_2, \sigma)$: $n_2 = 48$, $\bar{x}_2 = 176.51$, $s_2 = 59.89$. To-utvalgs-t med lik varians gir konfidensintervall for $\mu_1 - \mu_2$ som

$$\bar{x}_1 - \bar{x}_2 \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

der s_p finnes fra formel s. 442 i boken. Finner $s_p = 62.1131$. For t-fordeling med $n_1 + n_2 - 2 = 100$ frihetsgrader og 99% konfidens finner vi $t^* = 2.626$. Innsatt får vi et 99% konfidensintervall for $\mu_1 - \mu_2$ som (-5.83,58.89).

(*e*) Et 99% konfidensintervall for $\mu_1 - \mu_2$ kan brukes til å teste

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{mot} \quad H_a : \mu_1 - \mu_2 \neq 0$$

på nivå 0.01. Siden $\mu_1 - \mu_2 = 0$ er inneholdt i intervallet, kan vi ikke forkaste H_0.