TMA4245 Statistikk
Eksamen 9. desember 2013

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a) Define the following events:

- A: Getting an ace as your first card
- B: Getting a blackjack

Since the suit of a card does not matter in this game and we have a situation with replacement, we can consider the deck to have 13 cards with just one ace, king, queen, jack and ten. Then, dividing the number of favorable by the number of possible hands, disregarding the order in which the cards are dealt, gives

$$P(B) = \frac{4 \cdot 1 \cdot 2!}{13^2} = \frac{8}{13^2} = \underline{\underline{0.04734}},$$

and

$$P(B|A) = 4/13 = \underline{\underline{0.3077}}.$$

When $P(A) = 0.1$, $P(B) = 0.06$, and $P(A|B) = 0.4$, we have that

$$P(B|A) = P(A|B) \cdot \frac{P(B)}{P(A)} = 0.4 \cdot \frac{0.06}{0.1} = \underline{\underline{0.24}}.$$

b) This situation satisfies the conditions for a Bernoulli process: each game results in either a loss or a win for Lars, the probability of a win is constant, and the games are independent. The number of the trial in a Bernoulli process on which the first success occurs, follows a geometric distribution. Thus, the distribution of X is

$$f(x) = g(x; p) = p(1-p)^{x-1} = \underline{\underline{0.3 \cdot (0.7)^{x-1}}},$$

where $p = 0.3$ is his chance of winning any one game.

Lars's bet in a game is 2^{x-1}, where x is the number of the game—i.e., $2^0 = 1$ for the first game, $2^1 = 2$ for the second game, $2^2 = 4$ for the third game, $2^3 = 8$ for the fourth game and $2^4 = 16$ for the fifth game, if he has not won before that. Since his winnings are twice his bet, $W = 2 \cdot 2^{X-1} = 2^X$. Therefore, W is a function of X and

$$\begin{aligned} E(W) &= \sum_{x=1}^{5} 2^x f(x) = p \sum_{x=1}^{5} 2^x (1-p)^{x-1} \\ &= 0.3 \cdot (2 + 2^2 \cdot 0.7 + 2^3 \cdot 0.7^2 + 2^4 \cdot 0.7^3 + 2^5 \cdot 0.7^4) = \underline{\underline{6.567}}. \end{aligned}$$

So Lars's expected winnings are 6 dollars and 57 cents.

There are two ways of finding $E(Y)$. The most elegant solution involves realizing that if Lars wins, his winnings will be one dollar more than his previous losses. E.g., that $4-(1+2)=8-(1+2+4)=1$, or in general:

$$\sum_{i=1}^{n} 2^{i-1} = 2^n - 1.$$

(This formula is given in a more general form at the top of page 112 in *Rottmann*.) Then we see that

$$\begin{aligned} E(Y) &= 1 \cdot P(\text{Lars wins}) - \sum_{i=1}^{5} 2^{i-1} \cdot P(\text{Lars loses 5 times}) \\ &= (1-(1-p)^5) - (2^5-1)\cdot(1-p)^5 = (1-0.7^5) - 31 \cdot 0.7^5 = \underline{\underline{-4.4}}. \end{aligned}$$

The more straightforward approach to finding $E(Y)$ is to compute the expected value of Lars's bets and subtract that from $E(W)$:

$$E(Y) = E(W) - \left(\sum_{x=1}^{5} P(X=x) \sum_{j=1}^{x} 2^{j-1} \right) - P(X>5) \sum_{j=1}^{5} 2^{j-1},$$

where

$$\begin{aligned} &\sum_{x=1}^{5} P(X=x) \sum_{j=1}^{x} 2^{j-1} \\ &= 1 \cdot f(1) + (1+2)\cdot f(2) + (1+2+4) f(3) + (1+2+4+8) f(4) + (1+2+4+8+16) f(5) \\ &= p + 3p(1-p) + 7p(1-p)^2 + 15p(1-p)^3 + 31p(1-p)^4 \\ &= 0.3 \cdot (1 + 3\cdot 0.7 + 7\cdot 0.7^2 + 15 \cdot 0.7^3 + 31 \cdot 0.7^4) = 5.73, \end{aligned}$$

and

$$\begin{aligned} P(X>5) &= 1 - P(X \le 5) = 1 - \sum_{x=1}^{5} p(1-p)^{x-1} \\ &= 1 - 0.3 \cdot (1 + 0.7 + 0.7^2 + 0.7^3 + 0.7^4) = 1 - 0.83 = 0.17, \end{aligned}$$

and

$$\sum_{j=1}^{4} 2^{j-1} = 1 + 2 + 2^2 + 2^3 + 2^4 = 1 + 2 + 4 + 8 + 16 = 31,$$

so that

$$E(Y) = 6.567 - 5.73 - 0.17 \cdot 31 = \underline{\underline{-4.4}}.$$

Thus, Lars can expect to walk out of the casino 4 dollar and 40 cents poorer.

Oppgave 2

a) Expected value in the binomial distribution is $E(X) = np = 20 \cdot 0.8 = \underline{\underline{16}}$.

$P(X > 16) = 1 - P(X \leq 16) = 1 - 0.59 = \underline{\underline{0.41}}$.

$P(X = 20|X > 16) = \frac{P(X=20)}{P(X>16)} = \frac{0.8^{20}}{0.41} = \frac{0.0115}{0.41} = \underline{\underline{0.028}}$.

b) $H_0 : p = 0.8$, $H_1 : p > 0.8$.

Normal approximation means $Z = \frac{X-E(X)}{\sqrt{\text{Var}(X)}} = \frac{X-np}{\sqrt{np(1-p)}} = \frac{X-20\cdot 0.8}{\sqrt{20\cdot 0.8\cdot 0.2}}$ is assumed Gaussian.

We reject H_0 when the observed fraction of success is significantly large.

Setting $\alpha = 0.1$ the rejection value is $z_{0.1} = \underline{1.28}$. We get $X = 18$ and $Z = \frac{18-16}{\sqrt{3.2}} = \underline{1.12}$. Since $Z < z_{0.1}$, the observed number is not significantly large. $\underline{\underline{\text{We do not reject } H_0.}}$

Half correction could also be used in this exercise, and gives more accurate results. In that case we get $Z = \frac{17.5-16}{\sqrt{3.2}} = 0.8385$. Full score is given to both methods.

c) The hypothesis test is rejected when the observation is extreme. In particular, we reject if the observation is larger than c, and the critical level c is determined as the smallest c that satisfies $P(X > c|p = 0.8) \leq \alpha$. For various c we get: $P(X > 19|p = 0.8) = 0.011$, $P(X > 18|p = 0.8) = 0.07$, $P(X > 17|p = 0.8) = 0.21$. At significance level $\alpha = 0.1$ we reject when $X > 18$.

Based on the observed value $X = 18$, the P value is

$$P(X \geq 18|p = 0.8) = 1 - P(X \leq 17|p = 0.8) = \underline{\underline{0.21}}.$$

The power is the probability of rejection, given that the parameter is $H_1 : p = 0.9$:

$P(\text{reject}|p = 0.9) = P(X > 18|p = 0.9) = 1 - P(X \leq 18|p = 0.9) = 1 - 0.61 = \underline{\underline{0.39}}$.

Oppgave 3

a) Variansen til utvalgsgjennomsnittet er

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Sannsynlighetstetthetsfunksjonen til normalfordelingen er gitt på s. 25 i *Tabeller og formler i statistikk* som

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

slik at vi har

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\cdot\frac{0}{\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma}e^0 = \frac{1}{\sqrt{2\pi}\sigma}.$$

Dette gir at

$$\mathrm{Var}(\tilde{X}) = \frac{1}{4n\big(f(\mu)\big)^2} = \frac{1}{4n\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^2} = \frac{\pi\sigma^2}{2n} = \frac{\pi}{2}\mathrm{Var}(\bar{X}),$$

hvilket skulle vises.

Når vi skal velge mellom to estimatorer som begge er forventningsrette, velger vi alltid den med minst varians. Siden $\frac{\pi}{2} \approx 1.57 > 1$ har vi $\mathrm{Var}(\tilde{X}) > \mathrm{Var}(\bar{X})$, som betyr at vi foretrekker å bruke $\bar{X}$ som estimator for μ.

b) På grunn av de to tydelige outlierne på oppsiden, kommer medianen $\tilde{X}$ til å være mindre enn utvalgsgjennomsnittet $\bar{X}$ (for disse dataene er $\tilde{X} = 171.0$ mens $\bar{X} = 175.3$).

Vi har antatt at rekruttenes høyder er normalfordelte. Utfra histogrammet ser det ut til at gjennomsnittet ligger rundt 170 cm. I så fall er sannsynligheten for at to av de tretti datapunktene er større enn 235 cm neglisjerbar, så de ekstreme verdiene til disse to datapunktene skyldes antakelig en feil hos rekrutten som fylte inn dataene i regnearket – ikke spesielt usannsynlig, gitt det gulnede papiret og falmede blekket. Siden utvalgsgjennomsnittet er følsomt for outliere, mens utvalgsmedianen ikke er det, gir medianen et bedre estimat enn gjennomsnittet i dette tilfellet.

Anmerkning vedrørende dataene
Datasettet i denne oppgaven er naturligvis fiktivt. Histogrammet er laget for 28 datapunkt trukket tilfeldig fra en normalfordeling med forventningsverdi 166 cm (litt lavere enn gjennomsnittshøyden for 1878, som er 169.5 cm) og standardavvik 7 cm, og med to outliere på 239 cm og 251 cm (høyden til verdens høyeste mann). Når $X \sim N(166, 7^2)$ så er $P(X \geq 239) = 9 \cdot 10^{-26}$.

Oppgave 4

a)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} = \frac{\sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i}$$
$$= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2} = \frac{2\,752\,667 - \frac{1}{24}\cdot 925 \cdot 71\,194}{35\,727 - \frac{1}{24}\cdot 925^2} = \underline{\underline{115}}.$$

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n}\sum_{i=1}^n x_i = \frac{71\,194}{24} - \frac{115 \cdot 925}{24} = \underline{\underline{-1465}}.$$

We have $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/(\sum_{i=1}^n (x_i - \bar{x})^2)$. Then $T = \frac{\hat{\beta}_1 - \beta_1}{s_b}$, where

$$s_b = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{s}{\sqrt{\sum_{i=1}^n x_i^2 - (1/n)(\sum_{i=1}^n x_i)^2}}.$$

Based on the t-distribution $P(-t_{n-2,\alpha/2} < T < t_{n-2,\alpha/2}) = 1-\alpha$. This gives:

$$P(\hat{\beta}_1 - t_{n-2,\alpha/2}s_b < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2}s_b) = 1-\alpha.$$

Using $t_{22,0.025} = 2.07$ we get: $(115 \pm 2.07 \cdot 194/\sqrt{35727 - (1/24)(925)^2})$, $= (115 \pm 46) = \underline{\underline{(69, 161)}}$.

b) Prediction at week 40:

$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 40 = -1465 + 115 \cdot 40 = \underline{\underline{3135}}$.

$\hat{Y}_0 - Y_0$ is Gaussian distributed. Because of unbiased estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we have mean $E(\hat{Y}_0) = \beta_0 + \beta_1 40$. We also have $E(Y_0) = \beta_0 + \beta_1 40$. This means $E(\hat{Y}_0 - Y_0) = 0$.

We next use that $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{Y} + \hat{\beta}_1(x_0 - \bar{x})$. The mean of the data is $\bar{x} = (1/24)(925) = 38.5$.

The variance is

$$\mathrm{Var}(\hat{Y}_0 - Y_0) = \mathrm{Var}(\hat{Y}_0) + \mathrm{Var}(Y_0)$$

$$= \mathrm{Var}(\bar{Y}) + (40-\bar{x})^2\mathrm{Var}(\hat{\beta}_1) + \mathrm{Var}(Y_0) = \sigma^2/n + \sigma^2\frac{(40-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sigma^2$$

We estimate σ^2 with s^2 and we get

$$s_0 = s\sqrt{1 + \frac{1}{n} + \frac{(40-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} = 194\sqrt{1 + \frac{1}{24} + \frac{(40-38.5)^2}{35727 - (1/24)(925)^2}} = 201$$

and get a statistic with a T-distribution: $\frac{\hat{Y}_0 - Y_0}{s_0} \sim t(n-2)$. Then,

$$P(-t_{n-2,\alpha/2} < \frac{\hat{Y}_0 - Y_0}{s_0} < t_{n-2,\alpha/2}) = 1-\alpha,$$

and this gives:

$$P(\hat{Y}_o - t_{n-2,\alpha/2} \cdot s_0 < Y_0 < \hat{Y}_0 + t_{n-2,\alpha/2} \cdot s_0) = 1-\alpha.$$

The significance level $\alpha = 0.1$ means we set $t_{22,0.95} = 1.72$. This gives prediction interval $3135 \pm 1.72 \cdot 201 \approx \underline{\underline{(2\,789, 3\,481)}}$.

The interval at week 42 has the same form, but now we have a term $(42-38.5)^2$ instead of $(40-38.5)^2$ in the variance.

This gives only a small increase in the width since $\frac{(42-38.5)^2}{35727-(1/24)(925)^2} = 0.16$ and $\frac{(40-38.5)^2}{35727-(1/24)(925)^2} = 0.0296$ are both relatively small compared with 1. We get width at week 40:

$$2 \cdot 1.72 \cdot 194 \cdot \sqrt{1 + (1/24) + 0.03} = \underline{\underline{690}},$$

and width at week 42:

$$2 \cdot 1.72 \cdot 194 \cdot \sqrt{1 + (1/24) + 0.16} = \underline{\underline{732}}.$$

The prediction interval is narrowest at the mean of the week data, i.e., at week 39.

c) From the plot we clearly see that boys are heavier than girls on average. The increase with gestational weeks seem similar, and for this reason we fit the same slope. The error level seems to be about the same for boys and girls, and there is no reason to use a different variance.

The model would predict a non-zero intercept term and different weight for boys and girls at 0 weeks, which is unntatural. Perhaps a model with different slopes and intercept 0 would be more appropriate. Then again the linear regression model is valid only within the region where data is available.

$$SSE = \sum_{i=1}^{n_b}(y_i - \beta_b - \beta_1 x_i)^2 + \sum_{i=n_b+1}^{n}(y_i - \beta_g - \beta_1 x_i)^2$$

We differentiating this expression and set the derivative equal to 0.

$$\frac{dSSE}{d\beta_b} = -\sum_{i=1}^{n_b}(y_i - \beta_b - \beta_1 x_i) = n_b\beta_b + \beta_1\sum_{i=1}^{n_b}x_i + \sum_{i=1}^{n_b}y_i = 0$$

$$\frac{dSSE}{d\beta_g} = -\sum_{i=n_b+1}^{n}(y_i - \beta_g - \beta_1 x_i) = n_g\beta_g + \beta_1\sum_{i=n_b+1}^{n}x_i + \sum_{i=n_b+1}^{n}y_i = 0$$

From these two first expressions we have:

$$\hat{\beta}_b = (1/n_b)\sum_{i=1}^{n_b}y_i - \hat{\beta}_b(1/n_b)\sum_{i=1}^{n_b}x_i$$

$$\hat{\beta}_g = (1/n_g)\sum_{i=n_b+1}^{n}y_i - \hat{\beta}_1(1/n_g)\sum_{i=n_b+1}^{n}x_i$$

Here, $n_g = n - n_b = 12$.

We have to insert these into the expression for β_1 to solve for the slope:

$$\frac{dSSE}{d\beta_1} = -\sum_{i=1}^{n_b}x_i(y_i - \beta_b - \beta_1 x_i) - \sum_{i=n_b+1}^{n}x_i(y_i - \beta_g - \beta_1 x_i)$$

$$\frac{dSSE}{d\beta_1} = -\sum_{i=1}^{n}x_i y_i + \beta_b\sum_{i=1}^{n_b}x_i + \beta_g\sum_{i=n_b+1}^{n}x_i + \beta_1\sum_{i=1}^{n}x_i^2$$

Now, inserting for the solution to the slopes we get:

$$\frac{dSSE}{d\beta_1} = -a + \beta_1 b = 0, \hat{\beta}_1 = a/b$$

where

$$a = \sum_{i=1}^{n}x_i y_i - (1/n_b)\sum_{i=1}^{n_b}y_i\sum_{i=1}^{n_b}x_i - (1/n_g)\sum_{i=n_b+1}^{n}y_i\sum_{i=n_b+1}^{n}x_i$$

$$b = \sum_{i=1}^{n} x_i^2 - (1/n_b)(\sum_{i=1}^{n_b} x_i)^2 - (1/n_g)(\sum_{i=n_b+1}^{n} x_i)^2$$

Inserting numbers we have: $a = 2\,752\,667 - (1/12) \cdot 36\,258 \cdot 460 - (1/12) \cdot 34\,936 \cdot 465 = 9\,007$, $b = 35\,737 - (1/12) \cdot 460^2 - (1/12) \cdot 465^2 = 75$.

Finally, $\hat{\beta}_1 = 9\,007/75 = \underline{\underline{120.22}}$.

$\hat{\beta}_b = 36\,258/12 - 120.22 \cdot (1/12) \cdot 460 = -\underline{\underline{1\,587}}$.

$\hat{\beta}_g = 34\,936/12 - 120.22 \cdot (1/12) \cdot 465 = -\underline{\underline{1\,747}}$.