

HG  
Revidert mars 2013

## Notat til kapittel 5 i Løvås

### Regler om normalfordelingen

Kjennskap til reglene for normalfordelingen er grunnleggende for den statistiske analysen i kapittel 6 i Løvås, og studentene må kunne beherske disse skikkelig i dette kurset. Reglene er stort sett gitt i kapittel 5 i Løvås, men litt ufullstendig og usammenhengende. Jeg vil derfor i dette notatet samle og utfylle de viktigste reglene som må kunne, samt gi noen eksempler på bruk av dem. Reglene er stort sett gitt uten bevis (som er relativt tekniske).

#### Definisjon:

Vi skriver kort  $X \sim N(\mu, \sigma)$  for en modell der  $X$  er normalfordelt med forventning,  $E(X) = \mu$ , og standardavvik,  $\sigma = \sqrt{\text{var}(X)} = \text{SD}(X)$ . Den spesielle normalfordelingen  $N(0, 1)$ , kalles *standard normalfordeling* og Løvås bruker symbolikken  $G(z) = P(Z \leq z)$  for den kumulative fordelingsfunksjonen hvis  $Z \sim N(0, 1)$ .

*Eksempel:* Hvis du for eksempel i en oppgave får oppgitt at  $X \sim N(-1, 2)$ , så følger spesielt at  $E(X) = -1$  og  $\text{var}(X) = 4$ .

**Merknad.** Du trenger ikke å kjenne formelen for sannsynlighetstettheten for  $X \sim N(\mu, \sigma)$  i dette kurset, men for fullstendighetens skyld skriver jeg den opp her

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad \text{for } -\infty < x < \infty \quad (\text{ikke pensum})$$

Det du må vite om denne fordelingen (utover det som står i definisjonen) er:

- $X$  er kontinuerlig fordelt slik at  $P(X = x) = 0$  for alle  $x$ .
- Fordelingen (grafene til  $f(x)$ ) er entoppet, klokkeformet med maksimumsverdi for  $x = \mu$ . Og den er symmetrisk rundt  $x = \mu$  slik at den delen av  $f(x)$  som ligger til venstre for  $x = \mu$  er et nøyaktig speilbilde av delen som ligger til høyre for  $x = \mu$  (jfr. figur 5.14 i Løvås). Dette impliserer bl.a. at, hvis  $a$  er et tall  $> 0$ , er  $P(X \leq \mu - a) = P(X \geq \mu + a)$  (illustrer på en figur!).

**Regel R1** (står ikke i Løvås, men brukes implisitt flere steder):

Hvis  $X \sim N(\mu, \sigma)$  og  $Y = a + bX$ , der  $a$  og  $b$  er konstanter, er også  $Y$  normalfordelt

$$Y \sim N(E(Y), SD(Y)) = N(a + b\mu, |b|\sigma)$$

(Merk at uttrykkene i  $N$  til høyre følger av regler for forventning og varians i kapittel 4.)

*Eksempler som følger av regel R1 (sjekk):*

- Hvis  $X \sim N(-1, 2)$ , er (i)  $Y = 1 - X \sim N(E(Y), SD(Y)) = N(2, 2)$ .  
(ii)  $\frac{X}{2} \sim N\left(-\frac{1}{2}, 1\right)$  og (iii)  $1000 + 3X \sim N(997, 6)$ .
- Hvis  $X \sim N(\mu, \sigma)$ , er  $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$  (sett  $a = -\frac{\mu}{\sigma}$  og  $b = \frac{1}{\sigma}$  i R1). La  $G(x)$  være den kumulative fordelingsfunksjonen for  $Z$ , dvs  $G(x) = P(Z \leq x)$ , som er tabulert i tabell D.3 i Løvås. Av dette får vi **regel 5.14** i Løvås:

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = G\left(\frac{x - \mu}{\sigma}\right)$$

- Hvis  $X \sim N(-1, 2)$ , er, siden  $P(X = 0) = 0$ ,  
 $P(X < 0) = P(X \leq 0) = G\left(\frac{0 + 1}{2}\right) = G\left(\frac{1}{2}\right) = 0,6915$  i følge tabell D.3 i Løvås.

**Merknad.** Sannsynligheter i en normalfordeling kan også regnes ut i Excel: Bruk NORM.DIST funksjonen i Excel 2010 (heter NORMDIST i Excel 2007). Prøv det for  $P(X < 0)$  i siste eksempel og sammenlign med tabell D.3.

**Regel R2 . Summer av normalfordelte variable (regel 5.17 i Løvås presisert)**

La  $X_1, X_2, \dots, X_n$  være uavhengige og normalfordelte variable slik at  $X_i \sim N(\mu_i, \sigma_i)$  for  $i = 1, 2, \dots, n$ . ( $X_i$ -ene behøver altså ikke å være identisk fordelte). La  $a_1, a_2, \dots, a_n$  være vilkårlige konstanter. Da er  $Y = a_1X_1 + a_2X_2 + \dots + a_nX_n$  også normalfordelt:

$$Y \sim N\left(E(Y), \sqrt{\text{var}(Y)}\right) = N\left(a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n, \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2}\right)$$

der uttrykkene i  $N$  til høyre følger av regler om forventning og varians i kapittel 4.

Av **regel R2** følger direkte en viktig regel om gjennomsnitt av normalfordelte variable (brukt flere ganger i kurset, men ikke satt opp eksplisitt som en regel i Løvås).

**Regel R3. Gjennomsnittet av normalfordelte variable er (eksakt) normalfordelt.**

La  $X_1, X_2, \dots, X_n$  være uavhengige og *identisk* normalfordelte variable slik at  $X_i \sim N(\mu, \sigma)$  for  $i = 1, 2, \dots, n$ . Da er  $Y = \bar{X} = (X_1 + X_2 + \dots + X_n)/n$  også normalfordelt:

$$Y \sim N\left(E(Y), \sqrt{\text{var}(Y)}\right) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

**Bevis:** R3 følger direkte av R2: La i R2,  $\mu_1 = \mu_2 = \dots = \mu_n = \mu$ ,  $\sigma_1 = \sigma_2 = \dots = \sigma_n = \sigma$  (siden alle fordelingene for  $X_1, X_2, \dots, X_n$  er like, må alle forventninger være like og alle standardavvik være like). La  $a_1 = a_2 = \dots = a_n = \frac{1}{n}$ . Da ser vi at  $Y = \bar{X}$  omfattes av regel R2, og vi kan slutte at  $Y$  er normalfordelt. Parameterverdiene i den aktuelle normalfordelingen er gitt ved  $E(Y)$  og  $\sqrt{\text{var}(Y)}$  som er funnet før i kapittel 4:  $E(Y) = \mu$  og  $\sqrt{\text{var}(Y)} = \frac{\sigma}{\sqrt{n}}$ .

Alternativt, kunne man få fram disse parameterverdiene fra formlene i siste uttrykk i regel R2:

$$a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n = \frac{1}{n}\mu + \frac{1}{n}\mu + \dots + \frac{1}{n}\mu = n \cdot \frac{1}{n}\mu = \mu$$

$$\sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2} = \sqrt{\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \dots + \frac{1}{n^2}\sigma^2} = \sqrt{n \cdot \frac{1}{n^2}\sigma^2} = \sqrt{\frac{1}{n}\sigma^2} = \frac{\sigma}{\sqrt{n}}$$

**Bevis slutt.**

*Eksempler på bruk av R1-R3:*

- Anta  $X$  og  $Z$  er uavhengige og normalfordelte der  $X \sim N(-1, 2)$  og  $Z \sim N(0, 1)$ . Finn  $P(Y < 0)$  der  $Y = X - Z$ .

*Løsning:* I følge regel R2 er  $Y$  normalfordelt (sett for eksempel  $X_1 = X$ ,  $X_2 = Z$ ,  $a_1 = 1$ ,  $a_2 = -1$  og  $n = 2$ ).

Parameterverdiene finner vi som  $E(Y) = -1 - 0 = -1$  og

$$\sqrt{\text{var}(Y)} \stackrel{\text{regel 4.17}}{=} \sqrt{\text{var}(X) + \text{var}(Z)} = \sqrt{4 + 1} = \sqrt{5}$$

Dermed er  $Y \sim N(-1, \sqrt{5})$ , og det andre eksemplet etter regel R1 gir:

$$P(Y < 0) = P(Y \leq 0) = G\left(\frac{0 - (-1)}{\sqrt{5}}\right) = G(0, 45\dots) = 0,6736 \text{ (altså litt}$$

mindre enn sannsynligheten funnet ovenfor (= 0,6915) for at  $X$  selv

får negativ verdi. Mao. å trekke fra (eller legge til)  $Z$  på  $X$ -en er som å tilføre ”støy” på  $X$ ).

- Anta  $X_1, X_2, \dots, X_n$  er uavhengige og *identisk* normalfordelte med  $X_i \sim N(-1, 2)$  for  $i = 1, 2, \dots, n$ . For en vilkårlig  $n$  har vi fra R3 at gjennomsnittet er normalfordelt,  $\bar{X} \sim N\left(-1, \frac{2}{\sqrt{n}}\right)$ . Siden spredingen i denne fordelingen avtar med  $n$ , er det rimelig å forvente at sannsynligheten for at  $\bar{X}$  skal få negative verdier øker med  $n$ . Dette bekreftes av tabell 1 under. Som før bruker vi det andre eksemplet etter R1 og finner:

$$P(\bar{X} < 0) = G\left(\frac{0 - (-1)}{2/\sqrt{n}}\right) = G\left(\frac{\sqrt{n}}{2}\right)$$

hvorav, ifølge tabell D.3 i Løvås (sjekk tallene!),

**Tabell 1** Gjelder hvis hver  $X_i$  er  $N(-1, 2)$  fordelt

$n$	$\frac{\sqrt{n}}{2}$	$P(\bar{X} < 0)$
1	0,50	0,6915
5	1,12...	0,8686
10	1,58...	0,9429
20	2,24...	0,9875
30	2,74...	0,9969
50	3,54...	> 0,9990

Egenskapen til gjennomsnittet som er beskrevet i regel R3 bygger på forutsetningen at enkelt-observasjonene er normalfordelte. Dette er tilsynelatende en sterk forutsetning som bare unntaksvis er realistisk i praksis. Imidlertid, hvis antall observasjoner ( $n$ ) ikke er for lite (vi bør ha  $n \geq 20$  som en tommelfingerregel), vil konklusjonen at gjennomsnittet er normalfordelt fortsatt gjelde tilnærmet *uansett* hvilken fordeling enkeltobservasjonene er trukket fra. Dette er det berømte sentralgrenseteoremet (kjent siden 16-17-hundretallet). Løvås serverer sentralgrenseteoremet i to versjoner, regel 5.18 og 5.19, samlet i regel R4 nedenfor.

Dette betyr i eksemplet ovenfor at, selv om *vi ikke vet noe mer* om fordelingen til  $X_i$  utover at  $E(X_i) = -1$  og  $\sqrt{\text{var}(X_i)} = 2$ , så kan vi likevel slutte at  $\bar{X}$  er tilnærmet  $N\left(-1, \frac{2}{\sqrt{n}}\right)$  fordelt når  $n \geq 20$ , slik at

$$P(\bar{X} < 0) \approx G\left(\frac{0 - (-1)}{2/\sqrt{n}}\right) = G\left(\frac{\sqrt{n}}{2}\right)$$

(merk at den første likheten er erstattet med  $\approx$ ) og tabell 1 kan i hvert fall delvis fylles ut:

**Tabell 2.**  $P(\bar{X} < 0)$  når hver  $X_i$  har en vilkårlig fordeling med forventning  $-1$  og standardavvik  $2$ .

$n$	$\frac{\sqrt{n}}{2}$	$P(\bar{X} < 0)$
1	0,50	-----
5	1,12...	-----
10	1,58...	-----
20	2,24...	$\approx 0,9875$
30	2,74...	$\approx 0,9969$
50	3,54...	$> 0,9990$

Merk at vi her ikke kan angi noen verdier for  $P(\bar{X} < 0)$  i tilfellene  $n = 1, 5$  og  $10$  uten at vi vet mer om fordelingen til enkeltobservasjonene. Det er også verdt å merke seg at tilnærmelsen til normalfordelingen blir bedre og bedre dess større  $n$  er.

#### Regel R4 Sentralgrenseteoremet

La  $X_1, X_2, \dots, X_n$  være uavhengige variabler fra samme sannsynlighetsfordeling (ikke nødvendigvis normalfordeling!) med forventning,  $E(X_i) = \mu$ , og standardavvik,  $\sqrt{\text{var}(X_i)} = \sigma$ . Hvis  $n$  er "stor" ( $n \geq 20$  anses vanligvis tilstrekkelig), gjelder

$$(a) \text{ (regel 5.18) } \bar{X} \stackrel{\text{tilnærmet}}{\sim} N\left(E(\bar{X}), \sqrt{\text{var}(\bar{X})}\right) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$(b) \text{ (regel 5.19) } Y = X_1 + \dots + X_n \stackrel{\text{tilnærmet}}{\sim} N\left(E(Y), \sqrt{\text{var}(Y)}\right) = N\left(n\mu, \sqrt{n} \sigma\right)$$

**Merk** at (b) (regel 5.19) strengt tatt er overflødig når vi har tilgang til (a) og regel R1. For, hvis  $\bar{X}$  er tilnærmet  $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$  fordelt, følger (hvorfor?) av regel R1 at  $Y = n\bar{X}$  også er tilnærmet normalfordelt:

$$Y \stackrel{\text{tilnærmet}}{\sim} N\left(E(n\bar{X}), \sqrt{\text{var}(n\bar{X})}\right) = N\left(nE(\bar{X}), n\sqrt{\text{var}(\bar{X})}\right) = N\left(n\mu, n\frac{\sigma}{\sqrt{n}}\right) = N\left(n\mu, \sqrt{n} \sigma\right).$$

I tillegg til disse reglene er regel 5.20 i Løvås viktig som viser at normalfordelingen ofte (ikke alltid) kan brukes som tilnærming til binomiske, hypergeometriske og poisson-fordelinger.

Jeg skriver ikke opp den regelen her, men oppfordrer studentene til å øve seg på å bruke den. Spesielt viktig i kapittel 6. Forhåpentligvis vil presiseringen av reglene ovenfor gjøre dette lettere.