Econ 2130
HG mars 2017

Supplement til forelesningen 27. februar

Illustrasjon av regel 5.19 om sentralgrenseteoremet og litt om heltallskorreksjon (som i eksempel 5.20).

Regel 5.19 sier at summer, $Y = X_1 + X_2 + \cdots + X_n$, av uavhengige og identisk fordelte (uid) variable, $X_1, X_2, \ldots, X_n$, er tilnærmet normalfordelt, $Y \overset{\text{tilnærmet}}{\sim} N\left(E(Y), \sqrt{\text{Var}(Y)}\right)$, når *n* ikke er for liten (tommelfingerregel $n \geq 20$) . Dette gjelder *uansett* hvilken fordeling enkeltvariablene (X_i) har! Om fordelingen til X_i er det nok å vite hva forventningen ($\mu = E(X_i)$) er og hva standardavviket ($\sigma = \text{SD}(X_i) = \sqrt{\text{Var}(X_i)}$) er. I så fall kjenner vi også forventningen og standardavviket for *Y*: Regel 4.12 og 4.17 gir nemlig at

$$E(Y) = E\left(X_1 + X_2 + \cdots + X_n\right) = \mu + \mu + \cdots + \mu = n\mu$$

$$\sqrt{\text{Var}(Y)} = \sqrt{\text{Var}\left(X_1 + X_2 + \cdots + X_n\right)} = \sqrt{\sigma^2 + \sigma^2 + \cdots + \sigma^2} = \sqrt{n\sigma^2} = \sigma\sqrt{n}$$

Dermed har vi at $Y \overset{\text{tilnærmet}}{\sim} N\left(n\mu, \sigma\sqrt{n}\right)$ når $n \geq 20$ [1]. Tilnærmelsen brukes til å beregne kumulative sannsynligheter for *Y* tilnærmet:

$$(1) \qquad P(Y \leq y) = P\left(\frac{Y - n\mu}{\sigma\sqrt{n}} \leq \frac{y - n\mu}{\sigma\sqrt{n}}\right) \approx P\left(Z \leq \frac{y - n\mu}{\sigma\sqrt{n}}\right) = G\left(\frac{y - n\mu}{\sigma\sqrt{n}}\right)$$

der $Z \sim N(0,1)$ med kumulativ fordelingsfunksjon, $G(z) = P(Z \leq z)$, som er tabulert i tabell E3. i boka.

Dette teoremet er særdeles nyttig i praksis siden den eksakte fordelingen til *Y* ofte er meget komplisert og vanskelig å beregne.

---

[1] Av dette følger direkte den tilsvarende regelen 5.18 om gjennomsnitt, nemlig at $\overline{X} \overset{\text{tilnærmet}}{\sim} N\left(E(\overline{X}), \sqrt{\text{Var}(\overline{X})}\right) = N\left(\mu, \sigma/\sqrt{n}\right)$ når $n \geq 20$. Dette skyldes regelen gitt på forelesningen 23. februar om normalfordelingen som sier at en lineær-kombinasjon av en normalfordelt variabel må også være normalfordelt, som impliserer at hvis *Y* er (tilnærmet) normalfordelt, må også en konstant ganger *Y* være det, hvorav $\overline{X} = (1/n)Y \sim$ tilnærmet normalfordelt med $E\left(\overline{X}\right) = (1/n)E(Y)$ og $\text{Var}\left(\overline{X}\right) = (1/n)^2 \text{Var}(Y)$.

Som illustrasjon vil vi se på et par tilfeller der denne tilnærmelsen virker dårlig og et par tilfeller der den virker bra. La oss se nærmere på eksempelet vi diskuterte på forelesningen om de statistiske egenskapene til sum antall øyne ved flere kast med en rettferdig terning. På forelesningen diskuterte vi gjennomsnittlig antall øyne, mens vi her skal se på sum antall øyne.

La X_i være antall øyne vi får i kast nr. *i* med terningen, og $Y = X_1 + X_2 + \cdots + X_n$ er sum antall øyne for *n* kast. Alle X_i-ene har samme fordeling beskrevet i tabell 1 og er uavhengige.

Tabell 1 Fordeling for X_i

x	1	2	3	4	5	6
$P(X_i = x)$	1/6	1/6	1/6	1/6	1/6	1/6

Sjekk selv at forventning, varians og standardavvik er gitt ved

$$\mu = E(X_i) = 3.5, \quad \sigma^2 = \mathrm{Var}(X_i) = 2.9167, \text{ og } \sigma = \sqrt{\mathrm{Var}(X_i)} = 1.7078$$

1. Tilfellet *n* = 1 kast

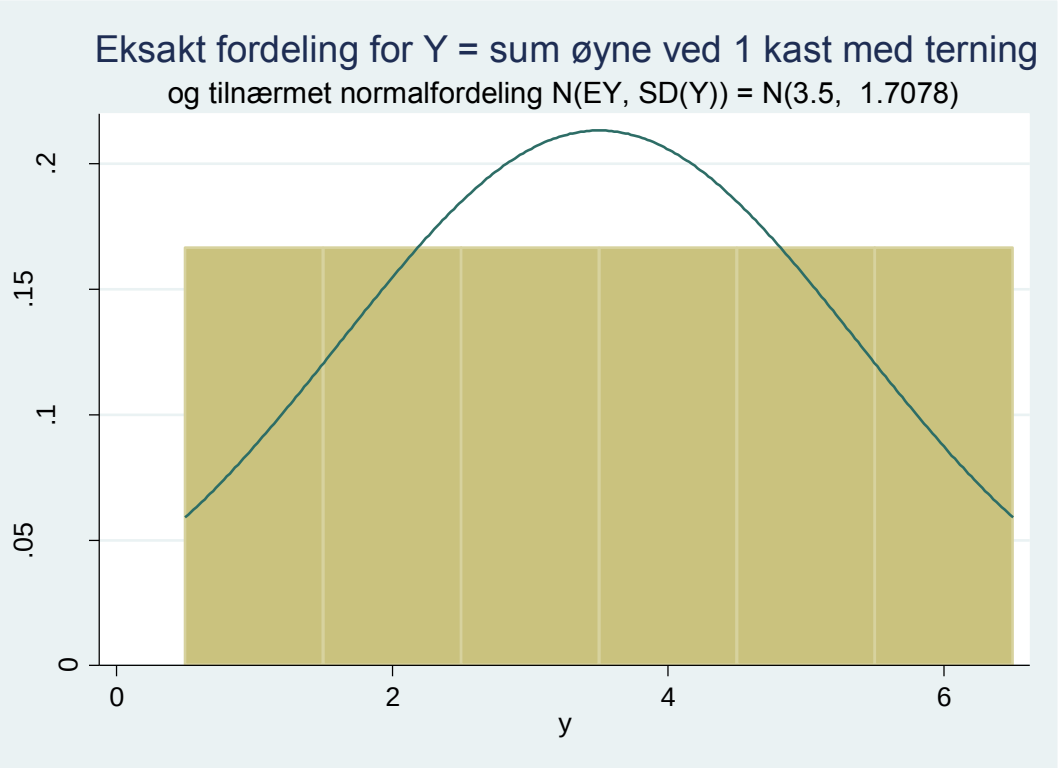
La oss først se på tilfellet med bare ett kast ($n = 1$), slik at $Y = X_1$. Dette er et tilfelle der normaltilnærmelsen antakelig ikke fungerer bra. La oss likevel prøve å tilnærme fordelingen for *Y* (som er gitt i tabell 1) med en normalfordeling. Det er mange normalfordelinger å velge blant, men, som antydet i regel 5.19, den normalfordelingen som vanligvis anses som gir den beste tilnærmelsen, er den som har samme forventning og standardavvik som *Y*, som i dette tilfellet blir

$$E(Y) = E(X_1) = 3.5, \qquad \mathrm{SD}(Y) = \mathrm{SD}(X_1) = 1.7078$$

Den beste tilnærmingen er derfor normalfordelingen $N(3.5, 1.7078)$. I figur 1 har jeg plottet både den eksakte fordelingen for *Y* fra tabell 1 sammen med den beste normale tilnærmingstettheten. Det er vanskelig å få til et slikt dobbeltplott i Excel, så jeg brukte i stedet STATA som bl.a. brukes i Statistikk 2.[2]

[2] Plottet er litt misvisende i og med at normaltettheten egentlig fortsetter på begge sider av intervallet mellom 1 og 6.

Figur 1



De eksakte sannsynlighetene fra tabell 1 framkommer som flateinnholdet av søylene i histogrammet. Merk at flateinnholdet av en søyle også er lik høyden på søylen siden lengden av grunnlinjen i søylen er lik 1. Anta vi er interessert i å se hvor god tilnærmelse normalfordelingen gir for

$$P(Y \le 2) = 1/6 + 1/6 = 1/3 = 0.333...$$

I histogrammet er denne (eksakte) sannsynligheten lik flateinnholdet av de to første søylene til sammen. I normalfordelingen framkommer den tilsvarende sannsynligheten som flateinnholdet under tetthetsfunksjonen opp til 2. Vi ser imidlertid av figuren at denne beregningen "mister" halvparten av siste søyle som er over intervallet 1.5 til 2.5. En bedre tilnærmelse ville være å ta flateinnholdet under normaltettheten opp til 2.5 i stedet. Det er dette som kalles *heltallskorreksjon* (Løvås side 201), som er aktuelt når man forsøker å tilnærme en diskret fordeling for en stokastisk variabel som bare kan ta hele tall som mulige verdier. Merk at begivenhetene $(Y \le 2)$ og $(Y \le 2.5)$ er logisk ekvivalente[3] og derfor like sannsynlige, $P(Y \le 2) = P(Y \le 2.5)$, siden Y kun kan ta hele tall som verdier.

[3] Hvis den ene begivenheten inntreffer så må den andre inntreffe og omvendt.

Med heltallskorreksjon blir derfor tilnærmelsen i (1) generelt seende ut som

(2)

$$P(Y \le y) = P(Y \le y + 0.5) = P\left(\frac{Y - n\mu}{\sigma\sqrt{n}} \le \frac{y + 0.5 - n\mu}{\sigma\sqrt{n}} \right) \approx P\left(Z \le \frac{y + 0.5 - n\mu}{\sigma\sqrt{n}} \right) = G\left(\frac{y + 0.5 - n\mu}{\sigma\sqrt{n}} \right)$$

der y er et helt tall og $Z \sim N(0, 1)$.

Bruker vi (2), får vi

$$P(Y \le 2) = P(Y \le 2.5) \approx G\left(\frac{2.5 - 1 \cdot (3.5)}{(1.7078) \cdot \sqrt{1}} \right) = G(-0.59) \overset{\text{Tabell E3.}}{=} 0.2776$$

som er betydelig forskjellig fra den eksakte verdien $0.333\ldots$ (men ikke så altfor galt). Ved bruk av (1) uten heltallskorreksjon får vi (sjekk selv) tilnærmelsen $G(-0.88) = 0.1894$ som er betydelig verre.

2. Tilfellet *n* = 2 kast

Her er $Y = X_1 + X_2$, og, siden X_1, X_2 er uavhengige og identisk fordelte (uid), har de samme forventning og varians, og vi får av regel 4.12 og 4.17 i Løvås

$$E(Y) = 2E(X_1) = 2(3.5) = 7, \quad \text{Var}(Y) = 2 \cdot \text{Var}(X_1) = 5.8333 \text{ og}$$

$$\text{SD}(Y) = \sqrt{\text{Var}(Y)} = 2.4152$$

Hvis vi vil tilnærme fordelingen til Y med en normalfordeling, bør vi altså bruke $N(7,\ 2.4152)$-fordelingen.

Vi trenger også den eksakte fordelingen til Y for å kunne sammenligne. Dette er ikke så vanskelig i dette tilfellet. Det er $6^2 = 36$ mulige kombinasjoner av verdier for paret (X_1, X_2) som alle er like sannsynlige $(1/36)$. De mulige verdiene for $Y = X_1 + X_2$ er $2, 3, 4, \ldots, 11$, og 12.

Tabell 2 viser hvilke kombinasjoner som gir en gitt verdi av Y. For eksempel ser vi at begivenheten $(Y = 8)$ inntreffer for 5 forskjellige kombinasjoner, slik at $P(Y = 8) = 5/36$.

Tabell 2 Verdier av *Y* for forskjellige kombinasjoner av X_1 og X_2.

		X_2					
		1	2	3	4	5	6
X_1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

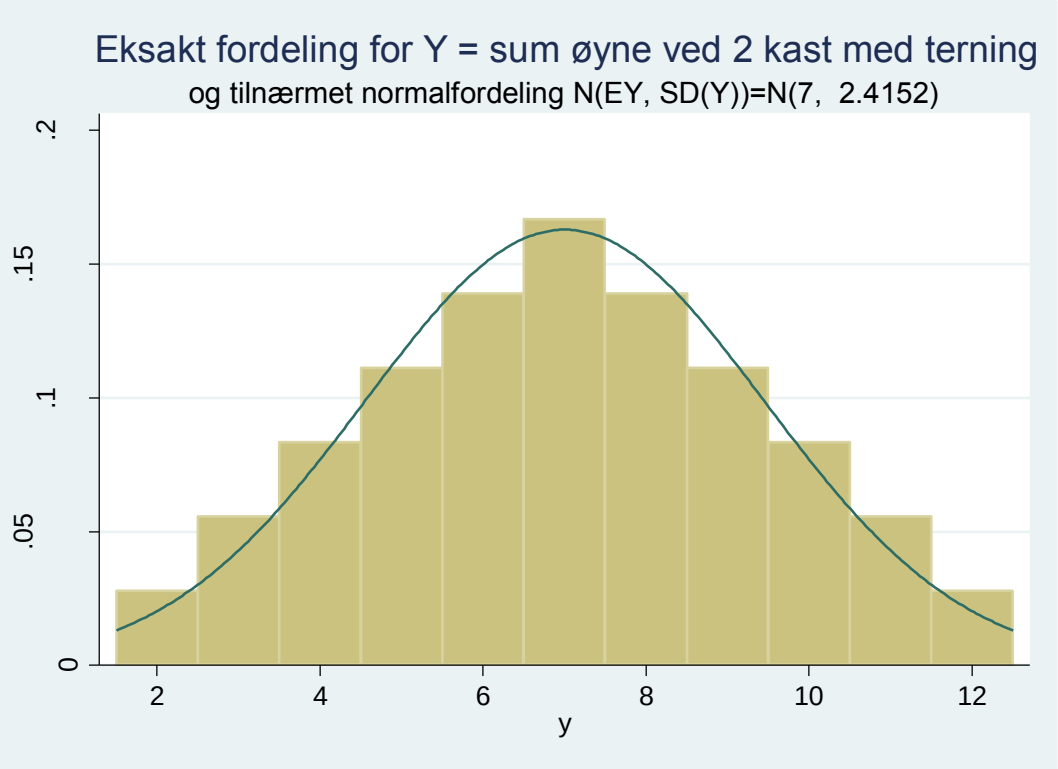
Den eksakte fordelingen for den diskrete variabelen *Y* blir derfor som gitt i tabell 3.

Tabell 3 Eksakt fordeling for sum øyne, *Y*, ved to kast.

y	2	3	4	5	6	7	8	9	10	11	12
$P(Y = y)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

I figur 2 har jeg plottet denne sammen med den beste normal-tilnærmelsen

Figur 2



Regneeksempel:

Eksakt blir etter tabell 3

$$P(Y \le 4) = \frac{1+2+3}{36} = \frac{1}{6} = 0.1666...$$

Med normaltilnærmelsen, $Y \overset{\text{tilnærmet}}{\sim} N(7,\ 2.4152)$, får vi:

Med heltallskorreksjon som i (2)

$$P(Y \le 4) = P(Y \le 4.5) = P\left(\frac{Y-7}{2.4152} \le \frac{4.5-7}{2.4152}\right) \approx P\left(Z \le \frac{4.5-7}{2.4152}\right) = G(-1.04) = 0.1492,$$

altså en feil på ca 0.016, som ikke er så verst.

Uten heltallskorreksjon (1) får vi

$$P(Y \le 4) \approx P\left(Z \le \frac{4-7}{2.4152}\right) = G(-1.24) = 0.1075,$$

altså en feil på ca 0.060.

3. Tilfellet *n* = 5 kast

Her er $Y = X_1 + X_2 + X_3 + X_4 + X_5$, og, siden $X_1, X_2, \ldots, X_5$ er uavhengige og identisk fordelte (uid), har de samme forventning og varians, og vi får av regel 4.12 og 4.17 i Løvås

$$E(Y) = 5E(X_1) = 5(3.5) = 17.5, \quad \text{Var}(Y) = 5 \cdot \text{Var}(X_1) = 14.5833 \text{ og}$$

$$\text{SD}(Y) = \sqrt{\text{Var}(Y)} = 3.8188$$

Hvis vi vil tilnærme fordelingen til Y med en normalfordeling, bør vi altså bruke $N(17.5,\ 3.8188)$-fordelingen.

Vi trenger også den eksakte fordelingen til Y for å kunne sammenligne. Dette er litt verre nå. Det er $6^5 = 7776$ mulige kombinasjoner av verdier for $(X_1, X_2, \ldots, X_5)$ som alle er like sannsynlige $(1/7776)$. De mulige verdiene for $Y = X_1 + X_2 + \cdots + X_5$ er $5, 6, 7, \ldots, 29, 30$.

Å gå igjennom alle disse for å finne ut hvor mange som gir en gitt verdi av Y er kjedelig å gjøre manuelt, så jeg laget et lite program i GAUSS (et kraftig og elegant

programmeringsspråk som flere på instituttet benytter) som løste oppgaven for meg[4]. Resultatet er gitt i tabell 4.

Tabell 4 Antall kombinasjoner av $X_1, X_2, \ldots, X_5$ som gir gitte verdier av summen *Y*.

y	5	6	7	8	9	10	11	12	13	14	15	16	17
Antall kombin. med $Y = y$	1	5	15	35	70	126	205	305	420	540	651	735	780
Antall kombin. med $Y \le y$	1	6	21	56	126	252	457	762	1182	1722	2373	3108	3888
y	18	19	20	21	22	23	24	25	26	27	28	29	30
Antall kombin. med $Y = y$	780	735	651	540	420	305	205	126	70	35	15	5	1
Antall kombin. med $Y \le y$	4668	5403	6054	6594	7014	7319	7524	7650	7720	7755	7770	7775	7776

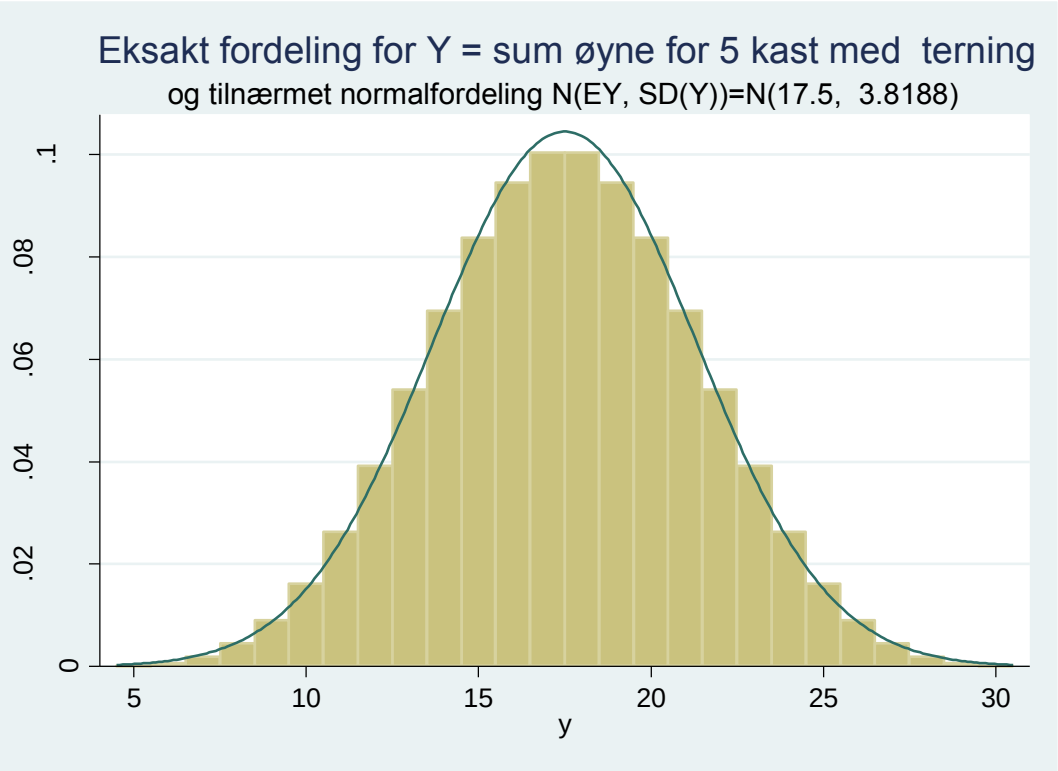
Sannsynligheter for *Y* får vi ved å dele tallene i tabell 4 med $6^5 = 7776$. La oss for eksempel se på sannsynligheten for at $P(Y \le 15)$. I følge tabellen er det 2373 kombinasjoner av X_i-ene som har sum ≤ 15. Siden alle kombinasjoner er like sannsynlige, blir den eksakte sannsynligheten

$$P(Y \le 15) = \frac{2373}{7776} = 0.30517...$$

I figur 3 har jeg plottet både den eksakte fordelingen for *Y* og den normalfordelingstettheten som passer best i henhold til regel 5.19.

[4] Det er mulig at dette kan gjøres i Excel, men jeg tror ikke det er lett. I stedenfor å kaste bort tiden på å prøve å finne på noe lurt i Excel, brukte jeg heller GAUSS med en gang der programmeringen ikke var vanskelig.

Figur 3



Vi ser at normaltilnærmelsen begynner å bli bedre.

Med normaltilnærmelsen, $Y \overset{\text{tilnærmet}}{\sim} N(17.5,\ 3.8188)$, får vi:

Med heltallskorreksjon som i (2):

$$P(Y \le 15) = P(Y \le 15.5) = P\left(\frac{Y-17.5}{3.8188} \le \frac{15.5-17.5}{3.8188}\right) \approx P\left(Z \le \frac{15.5-17.5}{3.8188}\right) = G(-0.52) = 0.3015,$$

altså en feil på ca 0.004 som er ganske bra.

Uten heltallskorreksjon (1) får vi:

$$P(Y \le 15) \approx P\left(Z \le \frac{15-17.5}{3.8188}\right) = G(-0.65) = 0.2578,$$

altså en feil på ca 0.047 som ikke er så bra.

4. Tilfellet *n* = 10 kast

Her er $Y = X_1 + X_2 + \cdots + X_{10}$, og, siden $X_1, X_2, \ldots, X_{10}$ er uavhengige og identisk fordelte (uid), har de samme forventning og varians, og vi får av regel 4.12 og 4.17 i Løvås

$$E(Y) = 10E(X_1) = 10(3.5) = 35, \quad \text{Var}(Y) = 10 \cdot \text{Var}(X_1) = 29.1667 \text{ og}$$

$$\text{SD}(Y) = \sqrt{\text{Var}(Y)} = 5.4006$$

Hvis vi vil tilnærme fordelingen til Y med en normalfordeling, bør vi altså bruke $N(35,\ 5.4006)$-fordelingen.

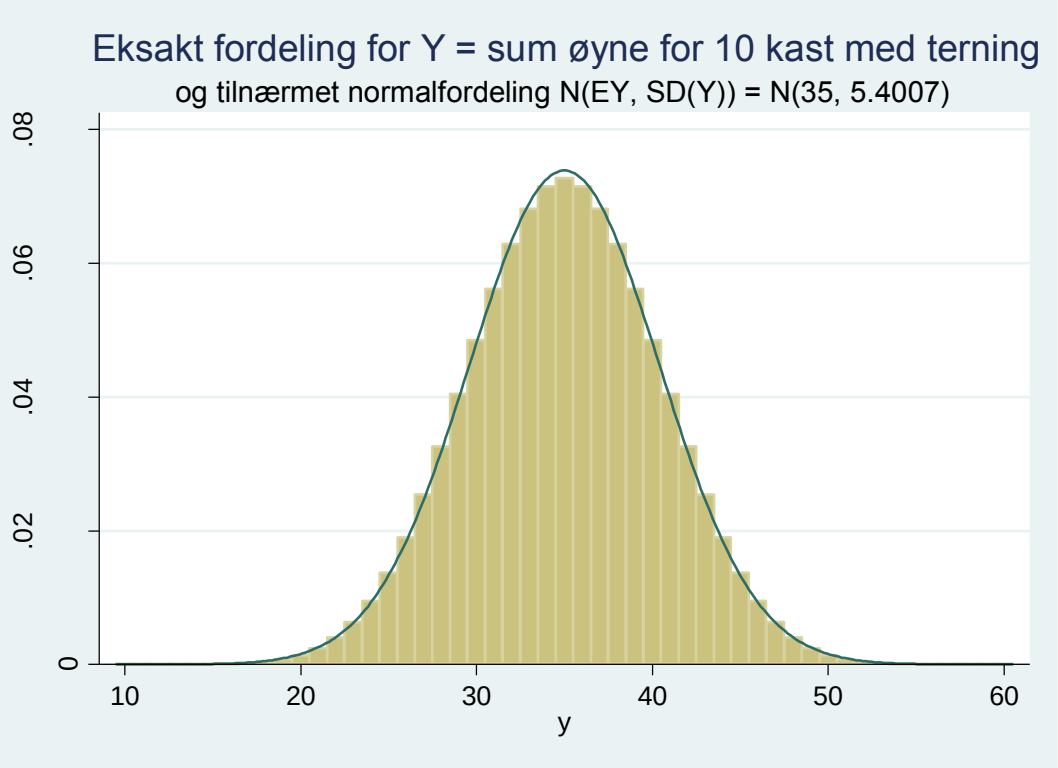
Vi trenger også den eksakte fordelingen til Y for å kunne sammenligne. Dette mye verre nå. Det er $6^{10} =$ 60 466 176 mulige kombinasjoner av verdier for $(X_1, X_2, \ldots, X_{10})$ som alle er like sannsynlige ($1/6^{10}$). De mulige verdiene for $Y = X_1 + X_2 + \cdots + X_{10}$ er $10, 11, 12, \ldots, 59, 60$.

Som regneeksempel skal vi se på $P(Y \le 30)$. Jeg lot GAUSS-programmet gå igjennom disse 60.5 millionene kombinasjoner (det tok laptop-en min ca 30 sekunder (!)) og laget en tabell som tabell 4 (ikke rapportert her). Ifølge den tabellen var det 12 393 645 kombinasjoner som hadde sum ≤ 30. Den eksakte sannsynligheten blir derfor

$$P(Y \le 30) = \frac{12\,393\,645}{6^{10}} = 0.20497...$$

I figur 4 har jeg plottet både den eksakte fordelingen for Y og den normalfordelingstettheten som passer best i henhold til regel 5.19.

Figur 4



Vi ser at normaltilnærmelsen har blitt enda bedre.

Med normaltilnærmelsen, $Y \overset{\text{tilnærmet}}{\sim} N(35,\ 5.4007)$, får vi:

Med heltallskorreksjon som i (2):

$$P(Y \le 30) = P(Y \le 30.5) = P\left(\frac{Y-35}{5.4007} \le \frac{30.5-35}{5.4007}\right) \approx P\left(Z \le \frac{30.5-35}{5.4007}\right) = G(-0.83) \overset{\text{Tabell E3}}{=} 0.2033,$$

altså en feil på ca 0.0016, som er ganske bra og bedre enn for $n = 5$.

Uten heltallskorreksjon (1) får vi:

$$P(Y \le 30) \approx P\left(Z \le \frac{30-35}{5.4007}\right) = G(-0.93) = 0.1762,$$

altså en feil på ca 0.028, som viser at heltallskorreksjon fortsatt lønner seg.

4. Tilfellet *n* = 20 kast

Her er $Y = X_1 + X_2 + \cdots + X_{20}$, og, siden $X_1, X_2, \ldots, X_{20}$ er uavhengige og identisk fordelte (uid), har de samme forventning og varians, og vi får av regel 4.12 og 4.17 i Løvås

$$E(Y) = 20E(X_1) = 20(3.5) = 70, \quad \text{Var}(Y) = 20 \cdot \text{Var}(X_1) = 58.3333 \text{ og}$$

$$\text{SD}(Y) = \sqrt{\text{Var}(Y)} = 7.6376$$

Hvis vi vil tilnærme fordelingen til *Y* med en normalfordeling, bør vi altså bruke $N(70,\ 7.6376)$ -fordelingen.

Å finne den eksakte fordelingen for *Y* for sammenligning blir svært mye vanskeligere nå. Det er $6^{20} = \left(6^{10}\right)^2 = (60\ 466\ 176)^2$ mulige kombinasjoner av verdier for $(X_1, X_2, \ldots, X_{20})$ som alle er like sannsynlige ($1/6^{20}$). De mulige verdiene for $Y = X_1 + X_2 + \cdots + X_{20}$ er $20, 21, 22, \ldots, 119, 120$.

Som regneeksempel skal vi se på $P(Y \le 60)$. Her kommer nok det fine GAUSS-programmet mitt til kort. Hvis vi regner med ca et halvt minutt på å gå igjennom 60.5 millioner kombinasjoner med min laptop, vil det ta ca 60.5 millioner halvminutter å gå gjennom 6^{20} kombinasjoner, dvs ca 504 000 timer som svarer til ca 58 år. Nå er det sikkert mulig å utvikle smarte formler[5] og algoritmer for å redusere beregningstiden til et praktisk nivå for akkurat denne situasjonen, men jeg gjorde ikke noe forsøk på det. Grunnen til det er ganske enkelt at denne oppgaven er komplett overflødig når formålet er å beregne sannsynligheter for *Y*. Vi har nemlig sentralgrenseteoremet som formulert i regel 5.19, og beregningene ovenfor som viser at vi kan regne (med kun neglisjerbar tap av realisme) at

$$Y \sim N(70,\ 7.6376) \text{ fordelt}$$

Med denne normaltilnærmelsen får vi:

Med heltallskorreksjon som i (2)

$$P(Y \le 60) = P(Y \le 60.5) = P\left(\frac{Y - 70}{7.6376} \le \frac{60.5 - 70}{7.6376}\right) \approx P\left(Z \le \frac{60.5 - 70}{7.6376}\right) = G(-1.24) \overset{\text{Tabell E3}}{=} 0.1075$$

Vi kan regne at feilen ligger godt under feilen (0.002) som vi hadde når $n = 10$.

Uten heltallskorreksjon (1) får vi

$$P(Y \le 60) \approx P\left(Z \le \frac{60 - 70}{7.6376}\right) = G(-1.31) = 0.0951$$

[5] som faktisk finnes i videregående sannsynlighetsteori…

Vi ser at forskjellen på beregningen med og uten heltallskorreksjon er redusert til ca 0.012, slik at poenget med heltallskorreksjon har nesten blitt borte. Når *n* blir enda større, vil forbedringen som oppnås med heltallskorreksjon forsvinne etter hvert. Dette poenget er relevant ved forståelse av regel 5.20 som sier at vi kan bruke tilsvarende normalfordelingstilnærmelser for binomiske, hypergeometriske og poissonfordelte stokastiske variable. Det er først og fremst i grenseområdet for *n* der normaltilnærmelsen begynner å bli effektiv, at heltallskorreksjon har noe for seg. For større *n* (eller varians, σ^2, i regel 5.20) er den overflødig.

Sluttmerknad. I dette eksemplet viste normaltilnærmelsen seg å gi akseptable resultater selv for så liten *n* som 5. Dette skyldes først og fremst symmetrien og formen på utgangsfordelingen i tabell 1. For andre fordelinger, for eksempel skjeve og flertoppete fordelinger, vil *n* måtte være større før normaltilnærmelsen skal være tilfredstillende. En mengde av simuleringer og beregninger ligger bak tommelfingerregelen, $n \geq 20$ i Løvås. Denne tommelfingerregelen burde gi akseptable sannsynlighetsberegninger i basert på normalfordelingen de fleste situasjoner man kan havne i, og, ikke minst, i situasjoner der man vet lite eller ingenting om fordelingen til enkeltvariablene, X_i.