Högskolan i Halmstad

Centrala gränsvärdessatsen

Exempel Avrundningsfelet X är rektangulärfördelat på $(-\frac{1}{2}, \frac{1}{2})$.
Hur blir då det *totala* avrundningsfelet $X_1 + X_2 + \ldots + X_n$ fördelat?
Hur blir det *genomsnittliga* avsrundningsfelet $\frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ fördelat?
Dessa frågor ska vi besvara i detta avsnitt men låt oss börja från början.

Definition 1 *Antag att X har fördelningsfunktionen $F(x)$.*
Då kallas $X_1, X_2, \ldots, X_n$ **stickprov** *av X om*

$\begin{cases} 1. & \textit{Alla } X_i \textit{ har fördelningen } F \\ 2. & X_i \perp X_j \textit{ för alla } i \neq j \end{cases}$

Obs $\sum_{i=1}^n X_i$ och $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ är stokastiska variabler, och om $X_1, \ldots, X_n$ är ett stickprov på X så är $E(X_i) = E(X) = \mu$ och $V(X_i) = V(X) = \sigma^2$ vilket innebär att $E(\sum_{i=1}^n X_i) = n\mu$, $V(\sum_{i=1}^n X_i) = n\sigma^2$ och $E(\bar{X}) = \mu$, $V(\bar{X}) = \sigma^2/n$ (se avsnittet *Funktioner av stokastiska variabler*).

Sats	*Centrala gränsvärdessatsen (CGS)*
Om	$E(X) = \mu$, $V(X) = \sigma^2$, n stort
	$X_1, X_2, \ldots, X_n$ stickprov av X
så	$\sum_{i=1}^n X_i \overset{\text{appr.}}{\in} N(n\mu, n\sigma^2)$
	$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \overset{\text{appr.}}{\in} N(\mu, \frac{\sigma^2}{n})$

Detta är ett *oerhört* viktigt resultat inom sannolikhetsteorin! Det kan kanske låta som det inte säger så mycket mer än

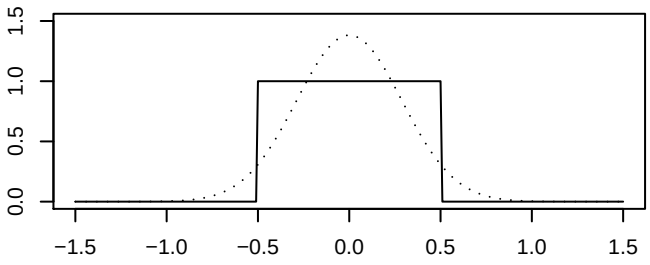
$$X_1, \ldots, X_n \text{ oberoende, } X_i \in N(\mu, \sigma^2),\ i = 1, \ldots, n \quad \Rightarrow$$

$$\Rightarrow \quad \bar{X} \in N(\mu, \tfrac{\sigma^2}{n}) \quad \text{och} \quad \textstyle\sum_{i=1}^n X_i \in N(n\mu, n\sigma^2)$$

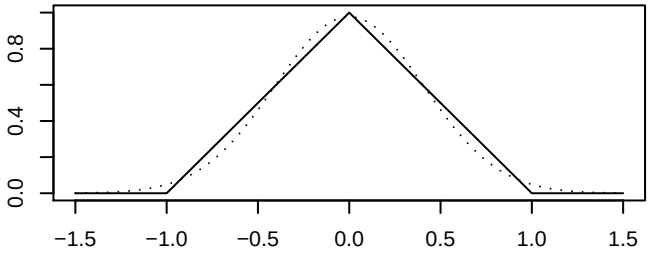
från föregående avsnitt, men detta är helt fel. Det fina med den centrala gränsvärdessatsen är att den inte antar något om fördelningen! Det enda som behövs är att variablerna är oberoende och har kända (ändliga) väntevärden och varianser. Nu finns det en uppsjö av lösare preciserade frågor vi (approximativt) kan besvara. Innan vi ger oss i kast med dem, låt oss se en illustration som troliggör innebörden av centrala gränsvärdessatsen.

Exempel (forts.) Antag att $X_1, X_2, X_3, \ldots$ är oberoende och rektangulärfördelade på $(-\frac{1}{2}, \frac{1}{2})$ (som avrundningsfelet). Då är skillnaden mellan fördelingingen för X_1, $X_1 + X_2$, $X_1 + X_2 + X_3$ och fördelingen för en motsvarande (dvs med samma väntevärde och varians) normalfördelad variabel illustrarad av att plotta deras respektive täthetsfunktioner i samma plot.

Plottar man alltså fördelningen för X_1 (de heldragna linjerna) och normalfördelningen (med samma väntevärde och varians, den punktade kurvan) ser man att dessa är ganska olika till sin karaktär.

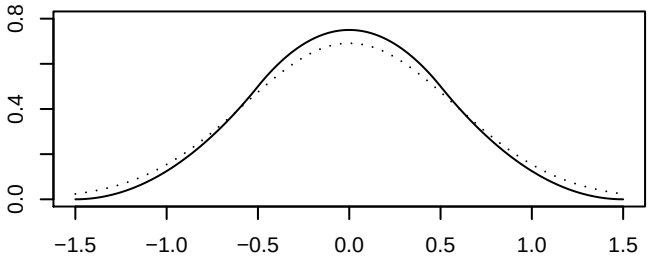


Om man istället betraktar fördelningen för $X_1 + X_2$ i jämförelse med motsvarande normalfördelning (dvs den med samma väntevärde och varians) får vi följande plot.



Fördelningen för summan av de två rektangulärfördelade variablerna är avsevärt mer lik normalfördelningen.

Om man går ännu ett steg och jämför fördelningen för $X_1+X_2+X_3$ med motsvarande normalfördelning ser vi att den förra nu även blivit av med sin kantighet och alltmer antar "klockformen".



Fortsätter vi med ännu fler plottar kommer budskapet bara vara att fördelningarna kommer närmare och närmare varandra – detta trots att den likformiga fördelningen är så olik normalfördelningen!

Svaret på frågan som ställdes inledningsvis är alltså att den exakta fördelningen är, i detta fall liksom i många andra, väldigt krånglig att beräkna (krångligare ju större n är) *men* enligt centrala gränsvärdessatsen är approximativt (med högre precision ju större n är!) $\bar{X} \in N(0, \frac{1}{12n})$ och $\sum_{i=1}^{n} X_i \in N(0, \frac{n}{12})$. □

Exempel I en demografibok finner Anna att livlängden i Sverige är fördelad med väntevärde 75 år och standardavvikelse 12 år och att det är vida känt att livlängden *ej* kan anses normalfördelad! Annas släkt består av 25 personer. Vad är approximativt sannolikheten att den genomsnittliga livlängden i släkten överstiger 76 år? Redovisa alla antaganden du gör.

Lösning: Låt $X_1, X_2, \ldots, X_{25}$ vara livlängderna av de 25 personerna i Annas släkt. Antag (med stöd av demografiboken) att $E(X_i) = 75$ och $V(X_i) = 12^2 = 144$ *och att alla livslängderna är obeoroende.* Då har vi enligt centrala gränsvärdessatsen att $\bar{X} = \sum_{i=1}^{25} \overset{\text{appr.}}{\in} N(75, 12/\sqrt{25}) = N(75, 2.4)$ varmed $P(\bar{X} > 76) = 1 - P(\bar{X} \leq 76) = 1 - P(\frac{\bar{X}-75}{2.4} \leq \frac{76-75}{2.4}) = 1 - \Phi(\frac{1}{2.4}) = 1 - 0.6628 = 0.3372$ □

Nyttan med centrala gränsvärdessatsen är mycket stor. Inte bara kan vi lösa denna typ av problem utan dessutom får vi approximationsregler för andra fördelningar. Eftersom en variabel som är binomalfördelad med n och p kan ses som summan av n oberoende variabler som är 1 med sannolikhet p och 0 med sannolikhet $1 - p$, så säger centrala gränsvärdessatsen att ju större n är, desto bättre approximeras sannolikhetsberäkningar för denna binomialfördelade variabel med en normalfördelad dito. På samma sätt har vi redan sett att summan av två variabler Poissonfördelade med λ_1 resp. λ_2 är Poissonfördelad med $\lambda_1 + \lambda_2$ varmed vi även kan approximera med normalfördelning då vi beräknar Poissonfördelningssannolikheter och λ är tillräckligt stort.

Sats
Om $X \in Bin(n, p)$ och $np(1-p) \geq 10$
så $X \overset{\text{appr.}}{\in} N(np, np(1-p))$

Sats
Om $X \in Poi(\lambda)$ och $\lambda \geq 15$
så $X \overset{\text{appr.}}{\in} N(\lambda, \lambda)$

Vid approximation av diskreta fördelningar med normalfördelning ska man dock göra **halvkorrektion**, dvs om X är en diskret variabel med väntevärde μ och varians σ^2 så är

$$P(X \leq x) \approx \Phi\left(\frac{x + \mathbf{0.5} - \mu}{\sigma}\right)$$

Exempel I en tillverkningsprocess tillverkas komponenter på löpande band. Varje komponent löper 10% risk att bli feltillverkad. Om vi nu slumpmässigt väljer 1000 enheter, vad är sannolikheten att högst 115 av dessa är feltillverkade?

Lösning: Låt oss först och främst försöka på "gammalt hederligt sätt" såsom vi lärde oss i avsnittet om diskreta variabler. Låt X vara antalet felaktigt tillverkade komponenter. Då är

$$
\begin{aligned}
P(X \leq 115) &= \sum_{x=0}^{115} P(X = x) \\
&= \left\{ \begin{array}{l} \text{binomialfördelning} \\ \text{med } n = 1000,\ p = 0.1 \end{array} \right\} \\
&= \sum_{x=0}^{115} \binom{1000}{x} 0.1^x 0.9^{1000-x} \\
&= \ldots \text{ jobbigt!! } \ldots \\
&(= \ 0.94655)
\end{aligned}
$$

Eftersom vi var tvungna att beräkna en summa med 115 termer är detta inget realistiskt sätt att lösa uppgiften. (Även om det finns tabeller också för binomialfördelningen löses denna uppgift inte med sådana eftersom x är så stort.)

Låt oss istället försöka med "normalapproximation": $np(1-p) = 1000 \cdot 0.1 \cdot 0.9 = 90 > 10$ så $X \overset{\text{appr.}}{\in} N(1000 \cdot 0.1, 90)$ varmed $P(X \leq 115) = \Phi(\frac{115-100}{\sqrt{90}}) = \Phi(1.58) = 0.9429$. Detta innebär ett fel på $|0.94655 - 0.9429| = 0.00365$ jämfört med det exakta svaret.

Låt oss slutligen kontrollera om det hade gått bättre om vi tillämpat halvkorrektion: $P(X \leq 115) = \Phi(\frac{115+0.5-100}{\sqrt{90}}) = \Phi(1.63) = 0.9484$ som endast är på avståndet $|0.94655 - 0.9484| = 0.00185$ från det exakta svaret, dvs hälften av vad vi fick utan halvkorrektion. □

Tack vare centrala gränsvärdessatsen har vi även en del andra resultat som vi återkommer till i senare avsnitt.