Föreläsning 4: Normalfördelning och CGS

Johan Thim (johan.thim@liu.se)

January 9, 2022

Vi kommer nu fortsätta att studera normalfördelningen. Detta är en mycket viktig fördelning i tillämpningar och vi kommer i avsnittet med centrala gränsvärdessatsen (CGS) att belysa fenomenet att i princip alla summor av många stokastiska variabler tenderar att bli normalfördelade.

1 Normalfördelningen (forts)

Eftersom det är bökigt att hantera integraler av uttryck som innehåller $\exp(-x^2)$ så fokuserar vi istället på hur vi kan transformera problemet till $N(0,1)$-fallet och där helt enkelt använda tabell för att beräkna sannolikheter.

Standardisering av variabel

Om X är en stokastisk variabel med $E(X) = \mu$ och $V(X) = \sigma^2$, så är $Z = (X-\mu)/\sigma$ en stokastisk variabel med $E(Z) = 0$ och $V(Z) = 1$. Vi kallar Z för **standardiserad**.

Beviset för att $E(Z) = 0$ följer direkt från linjäriteten hos väntevärdet. Variansen kan ses genom följande argument:

$$\begin{aligned} V((X-\mu)/\sigma) &= E((X-\mu)^2/\sigma^2) - 0^2 = \frac{1}{\sigma^2}\left(E(X^2) - 2\mu E(X) + \mu^2\right) \\ &= \frac{1}{\sigma^2}\left(E(X^2) - E(X)^2\right) = \frac{\sigma^2}{\sigma^2} = 1. \end{aligned}$$

Standardiserad normalfördelning

Sats. Om $X \sim \mathrm{N}(0,1)$ så är $E(X) = 0$ och $V(X) = 1$.

Eftersom $e^{-x^2/2}$ går mot noll väldigt snabbt, så kommer

$$\int_{-\infty}^{\infty} |x\varphi(x)|\,dx < \infty.$$

Alltså måste

$$E(X) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\varphi(x)\,dx = 0$$

eftersom $x \mapsto x\varphi(x)$ är en udda funktion. På liknande sätt ser vi att

$$E(X^2) = \int_{-\infty}^{\infty} x^2\varphi(x)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\left(xe^{-x^2/2}\right)dx$$
$$= \frac{1}{\sqrt{2\pi}}\left(\left[x(-e^{-x^2/2})\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right) = 0 + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = 1,$$

där vi partialintegrerat och utnyttjat att $\varphi(x)$ är en täthetsfunktion så den sista integralen blir ett.

Ett korollarium av satsen ovan (gör ett variabelbyte i integralerna) är följande.

$$X \sim \mathbf{N}(\mu, \sigma)$$

Om $X \sim \mathrm{N}(\mu, \sigma)$ så är $E(X) = \mu$ och $V(X) = \sigma^2$.

Standardisering av normalfördelning

Sats. $X \sim \mathrm{N}(\mu, \sigma) \quad \Leftrightarrow \quad Z = \dfrac{X-\mu}{\sigma} \sim \mathrm{N}(0,1).$

Bevis. Antag att $Z \sim N(0,1)$. Eftersom

$$F_X(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

och $\Phi'(x) = \varphi(x)$ då φ är kontinuerlig, så följer det att

$$f_X(x) = F_X'(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbf{R},$$

vilket är precis hur vi definierat $N(\mu, \sigma)$ tidigare.

Omvänt, om $X \sim N(\mu, \sigma)$ så är

$$F_Z(z) = P(Z \le z) = P(\mu + \sigma Z < \mu + \sigma z) = F_X(\mu + \sigma z),$$

så

$$f_Z(z) = f_X(\mu + \sigma z)\sigma = \varphi(z),$$

dvs $Z \sim N(0,1)$. □

Exempel

Låt $X \sim \mathrm{N}(1,2)$. Bestäm $P(X \le 1)$, $P(X \le -1)$, $P(0 < X \le 1)$, samt $P(|X-2| < 3)$.

(i) $P(X \leq 1) = P\left(\frac{X-1}{2} \leq \frac{1-1}{2}\right) = \Phi(0) = \frac{1}{2}$.

(ii) $P(X \leq -1) = P\left(\frac{X-1}{2} \leq \frac{-1-1}{2}\right) = \Phi(-1) = 1 - \Phi(1) \approx 0.1587$.

(iii)

$$\begin{aligned} P(0 < X \leq 1) &= P\left(\frac{0-1}{2} < \frac{X-1}{2} \leq \frac{1-1}{2}\right) = \Phi(0) - \Phi(-1/2) \\ &= 1/2 - (1 - \Phi(1/2)) = -1/2 + \Phi(1/2) \approx 0.1915. \end{aligned}$$

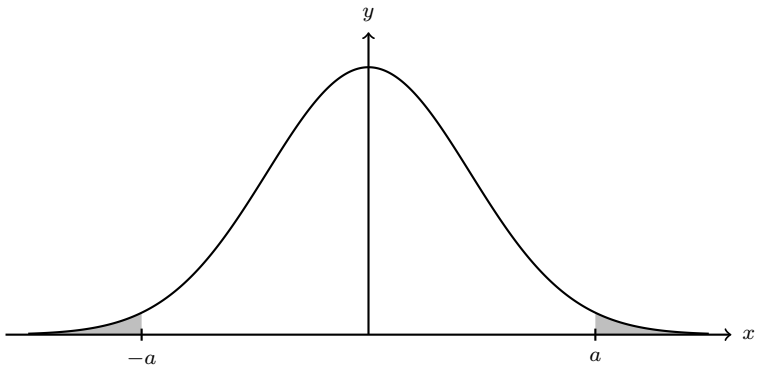
(iv)

$$\begin{aligned} P(|X-2| < 3) &= P\left(-3 < X - 2 < 3\right) = P\left(\frac{-1-1}{2} < \frac{X-1}{2} \leq \frac{5-1}{2}\right) \\ &= \Phi(2) - \Phi(-1) = \Phi(2) - 1 + \Phi(1) \approx 0.8186. \end{aligned}$$

Exempel

Låt $X \sim \mathrm{N}(0,1)$. Hitta ett tal a så att $P(|X| > a) = 0.05$.

Situationen ser ut som i bilden nedan. De skuggade områdena utgör tillsammans 5% av sannolikhetsmassan, och på grund av symmetri måste det vara 2.5% i varje ”svans”.



Om vi söker talet a, och vill använda funktionen $\Phi(x) = P(X \leq x)$, måste vi söka det tal som ger $\Phi(a) = 0.975$ (dvs de 2.5% i vänstra svansen tillsammans med de 95% som ligger i den stora kroppen). Detta gör vi genom att helt enkelt leta efter talet 0.975 i tabellen över $\Phi(x)$ värden. Där finner vi att $a = 1.96$ uppfyller kravet att $P(X \leq a) = 0.975$.

2 Linjärkombinationer av normalfördelade variabler

Det är inte på något sätt uppenbart att summan av två likafördelade variabler har samma fördelning. Oftast är det inte ens sant. Men, just normalfördelningen har precis denna trevliga egenskap!

Summa av normalfördelade variabler

Sats. Låt $X_1 \sim \mathrm{N}(\mu_1, \sigma_1)$ och $X_2 \sim \mathrm{N}(\mu_2, \sigma_2)$ vara oberoende och låt $a, b \in \mathbf{R}$. Då gäller att

$$aX_1 + bX_2 \sim \mathrm{N}(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}).$$

Beviset är inte trivialt. Att $E(aX_1 + bX_2) = a\mu_1 + b\mu_2$ och $V(aX_2 + bX_2) = a^2\sigma_1^2 + b^2\sigma_2^2$ är enkelt att se. Detta gäller oavsett vad variablerna har för slags fördelning (enbart egenskaper för väntevärde och varians). Det faktum att summan blir normalfördelad kräver ett djupare argument; se boken (man använder faltningssatsen).

Satsen ovan generaliserar direkt till flera variabler. Vi har följande användbara specialfall.

Summor och medelvärde

Sats. Låt $X_1, X_2, \ldots, X_n$ vara oberoende och $X_k \sim \mathrm{N}(\mu, \sigma)$ för $k = 1, 2, \ldots, n$. Då gäller följande:

$$X := \sum_{k=1}^{n} X_k \sim \mathrm{N}(n\mu, \sigma\sqrt{n}) \qquad \text{och} \qquad \overline{X} := \frac{1}{n}\sum_{k=1}^{n} X_k \sim \mathrm{N}(\mu, \sigma/\sqrt{n}).$$

Den sista likheten är intressant, då det innebär att ju fler ”likadana” variabler vi tar med i ett medelvärde, desto *mindre* blir variansen. Till exempel får vi alltså säkrare resultat ju fler mätningar vi gör (något som känns intuitivt korrekt). Det är dock mycket viktigt att variablerna är *oberoende*. Annars gäller inte satsen! Vi bildar aldrig heller några skillnader mellan varianser, utan det som gör att variansen minskar med antalet termer är faktorn $1/n$ i medelvärdet:

$$V(\overline{X}) = V\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2} V\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n} V(X_k) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

eftersom variablerna är oberoende och $V(X_k) = \sigma^2$ för alla k.

Exempel

Låt $X \sim \mathrm{N}(10, 3)$ och $Y \sim \mathrm{N}(21, 6)$ vara oberoende. Vad är sannolikheten att Y är mer än dubbelt så stor som X?

Lösning: Låt $W = Y - 2X \sim \mathrm{N}(21 - 20, \sqrt{6^2 + 4 \cdot 3^2}) = \mathrm{N}(1, \sqrt{72})$. Vi söker alltså

$$\begin{aligned} P(Y > 2X) &= P(Y - 2X > 0) = P(W > 0) = 1 - P(W \leq 0) = 1 - \Phi\left(\frac{0-1}{\sqrt{72}}\right) \\ &= 1 - (1 - \Phi(0.12)) = 0.5478. \end{aligned}$$

Exempel

Låt T vara livslängden för en viss sorts lysrör (enhet: månader) och $T \sim \mathrm{N}(20, \sqrt{8})$.

(i) Vad är sannolikheten att man klarar 43 månader om man har två (oberoende) lysrör och byter direkt det första går sönder?

(ii) Hur många lysrör måste man skaffa för att medellivslängden ska vara mer än 19 månader med sannolikhet 95%?

Lösning:

(i) Vi har två lysrör, T_1 och T_2. Vi söker sannolikheten att $T_1 + T_2 \geq 43$. Satsen ovan visar att $T_1 + T_2 \sim \mathrm{N}(40, 4)$, så

$$P(T_1 + T_2 \geq 43) = 1 - P(T_1 + T_2 < 43) = 1 - P\left(\frac{T_1 + T_2 - 40}{\sqrt{16}} < \frac{43 - 40}{\sqrt{16}}\right)$$
$$= 1 - \Phi(3/\sqrt{16}) \approx 1 - \Phi(0.75) \approx 0.2266.$$

(ii) Låt $\overline{T}$ vara medelvärdet av n stycken lysrör. Det följer att $\overline{T} \sim \mathrm{N}(20, \sqrt{8}/\sqrt{n})$. Vi vill att

$$0.95 = P(\overline{T} > 19) = P\left(\frac{\overline{T} - 20}{\sqrt{8/n}} > \frac{19 - 20}{\sqrt{8/n}}\right) = 1 - P(Z \leq -1/\sqrt{8/n})$$
$$= 1 - \Phi(-1/\sqrt{8/n}) = 1 - \Big(1 - \Phi(1/\sqrt{8/n})\Big) = \Phi(\sqrt{n/8}),$$

där $Z = \dfrac{\overline{T} - 20}{\sqrt{8/n}} \sim \mathrm{N}(0, 1)$. Ur tabell finner vi då att $\sqrt{n/8} = 1.645$, eller ekvivalent, att $n \approx 21.6$. Således behövs åtminstone 22 stycken lysrör.

3 De stora talens lag

Vi kommer nu betrakta en fundamental situation i sannolikhetslära. Vi låter $X_1, X_2, \ldots$ vara en oändlig följd av oberoende och likafördelade stokastiska variabler. Vi låter $E(X_i) = \mu$ och $V(X_i) = \sigma^2$ (så vi antar att variansen är ändlig just nu). I vanlig ordning definierar vi det aritmetiska medelvärdet $\overline{X}_n$ av de n första variablerna i följden som

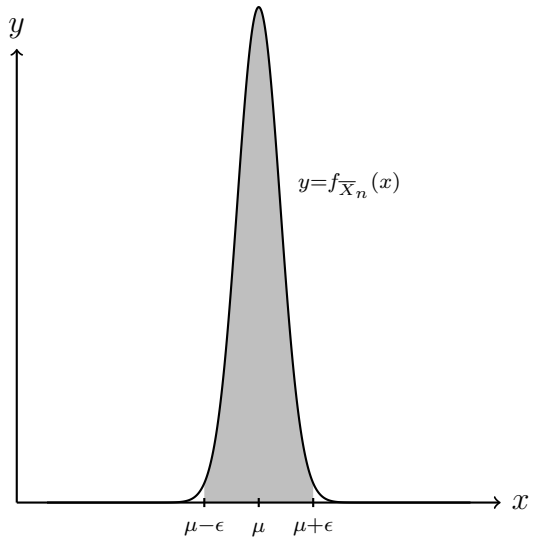
$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

De stora talens lag (svag formulering)

Sats. För varje $\epsilon > 0$ gäller att

$$P(|\overline{X}_n - \mu| < \epsilon) \to 1 \text{ då } n \to \infty.$$

En tolkning av satsen är att det aritmetiska medelvärdet av en följd oberoende och likafördelade variabler kommer att ha sin sannolikhetsmassa koncentrerad kring väntevärdet μ:



Variansen för $\overline{X}_n$ är som bekant

$$V(\overline{X}_n) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

eftersom variablerna är oberoende (och vi antagit ändlig varians). Så då $n \to \infty$ ser vi att variansen för medelvärdet går mot noll. Konvergensen i de stora talens lag förefaller alltså rimlig. Ett mer ordentligt bevis följer från kända olikheter, så låt oss formulera dessa.

Markovs olikhet

Sats. Om X är en icke-negativ stokastisk variabel med ändligt väntevärde så gäller att

$$P(X \geq a) \leq \frac{E(X)}{a}, \quad a > 0.$$

Bevis. För det kontinuerliga fallet med täthetsfunktion, eftersom $a > 0$ och $f_X(x) \geq 0$,

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \geq \int_{a}^{\infty} x\, f(x)\, dx \geq a \int_{a}^{\infty} f_X(x)\, dx = aP(X \geq a),$$

så följer att $P(X \geq a) \leq \dfrac{E(X)}{a}$. Det diskreta fallet hanteras analogt (gör det!) □

Tjebysjovs (Chebychevs) olikhet

Sats. Låt X vara en stokastisk variabel med ändligt väntevärde $E(X) = \mu$ och ändlig varians $V(X) = \sigma^2$, och låt $k > 0$. Då gäller att

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Bevis. Eftersom $(X-\mu)^2$ är en icke-negativ stokastisk variabel och $E(X-\mu)=0$, så gäller enligt Markovs olikhet att

$$P(|X-\mu| \geq k\sigma) = P((X-\mu)^2 \geq k^2\sigma^2) \leq \frac{E((X-\mu)^2)}{k^2\sigma^2}$$
$$= \frac{V(X-\mu)}{k^2\sigma^2} = \frac{1}{k^2},$$

där den näst sista likheten är Steiners sats. □

En följd av denna olikhet är att vi får en direkt uppskattning av hur mycket sannolikhetsmassa som finns i intervall av typen $(\mu - k\sigma,\ \mu + k\sigma)$. Vi kan till exempel se att det finns minst 50% av sannolikhetsmassan om $k = \sqrt{2}$, minst 75% om $k = 2$ och minst 96% om $k = 5$.

Vanligt missförstånd

Låt oss (oberoende) kasta en sex-sidig balanserad tärning 1800 gånger. Vi förväntar oss att medelvärdet ligger nära 3.5. Antag att medelvärdet blev 4.0. Betyder detta enligt satsen ovan att vi kommer att få fler resultat $1, 2, 3$ än $4, 5, 6$ om vi kastar tärningen 1800 gånger till? Svaret är nej. De olika kasten anses oberoende, och kan därför inte påverkas av tidigare utfall. Så hur kan då satsen ovan gälla? Faktum är att det inte behöver vara fler låga resultat vid kommande upprepningar, det räcker med att medelvärdet av de nya resultaten är mindre än 4.0 för att vi ska hamna närmare 3.5 *totalt* sett.

Det lönar sig alltså inte att satsa mer pengar bara för att man förlorat så många gånger på rad (om händelserna är oberoende, annars kan lite vad som helst inträffa!).

Bevis av de stora talens lag. I princip följer detta direkt av olikheterna ovan. Låt $\epsilon > 0$. Vi ser direkt att

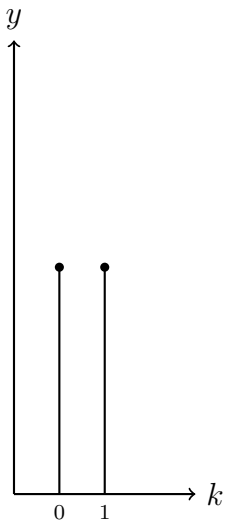
$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0,$$

då $n \to \infty$. □

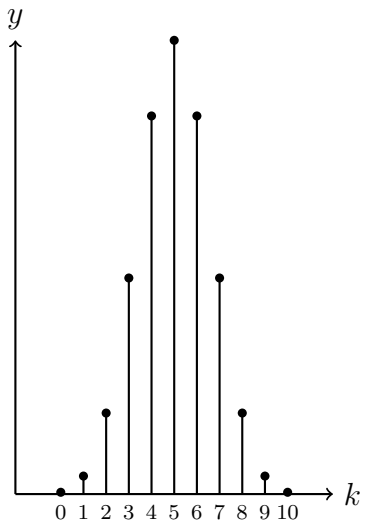
4 Centrala gränsvärdessatsen

Alla vägar leder till Rom. Eller åtminstone: alla fördelningar leder till normalfördelning? Faktum är att det är precis det den centrala gränsvärdessatsen säger: summan av ett stort antal oberoende och likafördelade stokastiska variabler är approximativt normalfördelad.

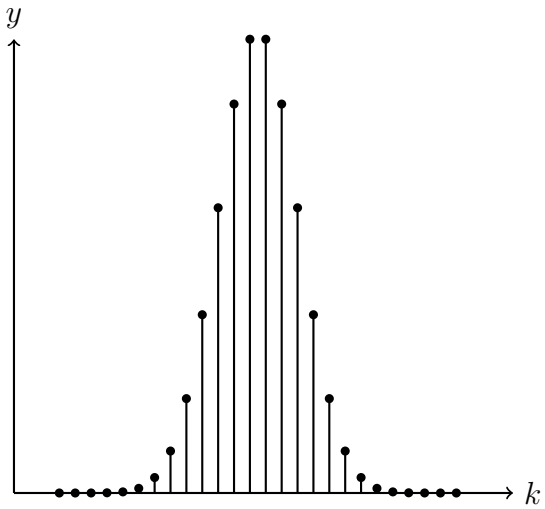
Vi betraktar ett exempel. Låt oss utföra det klassiska experimentet med slantsingling och räkna antalet X kronor vid ett visst antal, säg n, kast. Från tidigare exempel (inbrottstjuven) så vet vi att $X \sim \text{Bin}(n, p)$, där $p = 1/2$ om myntet är rättvist. En binomialfördelad variabel kan ses som en summa av oberoende Bernoulli-fördelade variabler X_k, en variabel för varje försök (slantsingling), där $X_k = 0$ om försök nr k "misslyckas" (klave), och $X_k = 1$ om försök k lyckas (krona). Alltså kan vi skriva $X = \sum_{k=1}^{n} X_k$. Varje X_k har sannolikhetsfunktionen $p_{X_k}(1) = p$ och $p_{X_k}(0) = 1 - p$. Med andra ord, binomialfördelningen kan ses som en summa av oberoende och likafördelade variabler. Om bara n är tillräckligt stort borde vi i så fall närma oss normalfördelningen. Hur stort? Vi skisserar några fall när n blir större och större och $p = 0.5$.



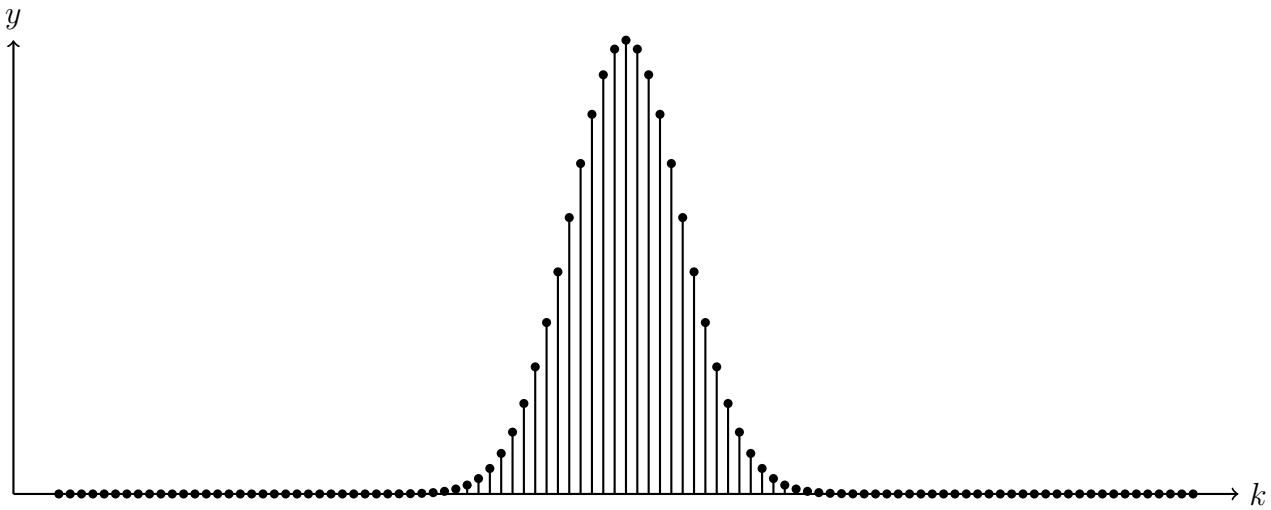
Med $n = 1$.



Med $n = 10$.



Med $n = 25$.



Med $n = 100$.

Här ser vi ganska tydligt att ju större n blir, desto mer lik blir sannolikhetsfördelning en normalfördelningskurva. Följande sats verkar alltså rimlig (åtminstone i Binomialfallet).

Centrala gränsvärdessatsen (CGS)

Sats. Låt $X_1, X_2, \ldots$ vara en oändlig följd av likafördelade och oberoende stokastiska variabler. Vidare, låt $E(X_k) = \mu$ och $V(X_k) = \sigma^2$ för $k = 1, 2, \ldots$. Då gäller att $Y_n = \sum_{k=1}^{n} X_k$ konvergerar i fördelning enligt

$$P\left(a < \frac{X - n\mu}{\sigma\sqrt{n}} < b\right) \to \Phi(b) - \Phi(a) \text{ då } n \to \infty$$

för alla $a, b \in \mathbf{R}$ med $a < b$. Vi säger att X är **asymptotiskt** normalfördelad.

Den kanske vanligaste situationen (åtminstone i denna kurs) kommer att vara att den summa vi är intresserade av är ett medelvärde, så vi formulerar detta specialfall separat.

Sats. Med samma förutsättningar som ovan så uppfyller medelvärdet $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ att

$$P\left(a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < b\right) \to \Phi(b) - \Phi(a) \text{ då } n \to \infty$$

för alla $a, b \in \mathbf{R}$ med $a < b$.

Beviset för satsen faller utanför ramen för denna kurs. Se, till exempel, Rick Durret: *Probability: Theory and Examples* eller Allan Gut: *An Intermediate Course in Probability.*

Så hur använder vi CGS?

Approximation via CGS

Med samma beteckningar och förutsättningar som ovan så är

$$P(X \le x) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right) \quad \text{och} \quad P(\overline{X} \le x) \approx \Phi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right), \quad x \in \mathbf{R},$$

om n är stort. Oftast brukar $n \ge 30$ duga, men skeva fördelningar kräver större n. Vi skriver $X \overset{\text{appr.}}{\sim} N(n\mu, \sqrt{n}\sigma)$ och $\overline{X} \overset{\text{appr.}}{\sim} N(\mu, \sigma/\sqrt{n})$; variablerna är **approximativt** normalfördelade.

Exempel

Vad är sannolikheten att summan av 50 stycken slumptal mellan 0 och 2 överstiger 53?

Lösning: Vi antar att slumptalen är likformigt fördelade, så varje slumptal $X_k \sim \text{Re}(0, 2)$, och att slumptalen är oberoende av varandra. Det råder likformig fördelning, så $E(X_k) = 1$ och $V(X_k) = 1/3$. Varför? Enkelt att se från definitionen:

$$E(X_k) = \int_0^2 x\frac{1}{2}\,dx = \left[\frac{x^2}{4}\right]_0^2 = 1$$

och

$$V(X_k) = \int_0^2 x^2\frac{1}{2}\,dx - 1^2 = \left[\frac{x^3}{6}\right]_0^2 - 1 = 1/3.$$

Så vi har en summa av 50 stycken likformigt fördelade variabler X_k med samma väntevärde och varians. Låt $X = \sum_{k=1}^{50} X_k$. CGS implicerar att $X \overset{\text{appr.}}{\sim} \text{N}(50, \sqrt{50/3})$. Alltså erhåller vi

$$P(X > 53) = 1 - P(X \le 53) \approx 1 - \Phi(3/\sqrt{50/3}) = 1 - \Phi(0.7348) \approx 0.2312.$$

Det är alltså ca 23% chans att summan överstiger 53.

Exempel

Antag att samtalstiderna till *1177* är oberoende och exponentialfördelade med väntevärde 15 minuter. Om en sjuksköterska förväntas svara på 28 samtal under ett åtta-timmars pass, vad är sannolikheten att hon lyckas?

Lösning: Låt $X_k \sim \text{Exp}(1/15)$ vara tiden för samtal k, $k = 1, 2, \ldots, 28$. Den totala tiden för 28 samtal ges av $X = \sum_{k=1}^{28} X_k$. Faktum är att man kan visa att X blir gamma-fördelad (se Blom et al.), men den fördelningen är ganska bökig att arbeta med. Vad säger CGS? Vi har kring 30 stycken samtal, så $X \overset{\text{appr.}}{\sim} \text{N}(28 \cdot 15, \sqrt{28} \cdot 15) = \text{N}(420, \sqrt{6300})$. Alltså är

$$P(X \leq 8 \cdot 60) \approx \Phi\left(\frac{480 - 420}{\sqrt{6300}}\right) \approx \Phi(0.76) = 0.7764.$$

Nästan 80% chans alltså! Hur bra stämmer då detta? Man kan härleda att X i själva verket har fördelningen $X \sim \Gamma(28, 1/15)$, så $P(X \leq 480) = 0.7838$ (MATLAB, `gamcdf`).

5 Approximation av binomialfördelning

Vi har redan stött på denna fördelning flera gånger. Situationen är att ett slumpförsök har två möjliga utfall, ett med sannolikhet p och det andra med $1 - p$. Vi upprepar försöket oberoende n gånger, och räknar antalet X gånger det första utfallet inträffar. Vi kallar X för binomialfördelad med parametrarna n och p, och skriver $X \sim \text{Bin}(n, p)$.

Binomialfördelning

Sats. Om $X \sim \text{Bin}(n, p)$, så är $E(X) = np$ och $V(X) = np(1-p)$. Vidare gäller att vi kan approximera $X \overset{\text{appr.}}{\sim} \text{N}(np, \sqrt{np(1-p)})$ om $np(1-p) \geq 10$.

Bevisskiss. Vi skriver X som en summa av n oberoende Bernoulli-variabler $X_k \sim \text{Be}(p)$, så väntevärdet $E(X) = \sum_{k=1}^{n} E(X_k) = np$, eftersom $E(X_k) = 0 \cdot (1-p) + 1 \cdot p = p$. På samma sätt, $V(X) = np(1-p)$ eftersom $V(X_k) = 0^2 \cdot (1-p) + 1^2 \cdot p - p^2 = p(1-p)$.

När det gäller approximationen till normalfördelning så följer detta av CGS. Att just kravet $np(1-p) \geq 10$ ger en bra approximation kräver en lite djupare analys av på vilket sätt sannolikheterna konvergerar.

Exempel

Vi sår 1000 stycken frön som har en grobarhet på 80% (sannolikheten att ett frö gror). Vad är sannolikheten att högst 180 stycken inte gror?

Lösning. Låt X vara antalet frön som inte gror. Vi antar att olika frön är oberoende av varandra. Då är $X \sim \text{Bin}(1000, 0.2)$. Eftersom

$$np(1-p) = 1000 \cdot 0.2 \cdot 0.8 = 160 \gg 10,$$

så är $X \overset{\text{appr.}}{\sim} \text{N}(200, \sqrt{160})$. Vi beräknar

$$P(X \leq 180) \approx \Phi((180 - 200)/\sqrt{160}) = \Phi(-1.5811) = 1 - \Phi(1.5811) = 0.057.$$

Det är alltså ca 6% sannolikhet. Verklig sannolikhet (MATLAB `binocdf(180,1000,0.2)`) är 6.02%.

5.1 Poissonapproximation

I vissa lägen så fungerar det dåligt med normalapproximationen (ofta då sannlikheten är nära 0 eller 1). Ibland kan då följande approximation fungera (se avsnittet i slutet av föreläsningen för argument kring varför detta är sant).

Approximation: Binomial till Poisson

Sats. Om $X \sim \text{Bin}(n, p)$ med $n \geq 10$ och $p \leq 0.1$, så är $X \overset{\text{appr.}}{\sim} \text{Po}(np)$.

Exempel

Sågaren Sverker sågar ut brädor som har normalfördelad längd $L \sim \text{N}(200, \sqrt{50})$ ($\sigma^2 = 50$), enhet: cm. Om Sverker en vacker dag sågar upp 300 brädor (oberoende av varandra), vad är sannolikheten att färre än 5 stycken är kortare än 185 cm?

Lösning: Vi räknar först ut sannolikheten p att en bräda är kortare än 185 cm:

$$p = P(L < 185) = P\left(\frac{L - 200}{\sqrt{50}} < \frac{-15}{\sqrt{50}}\right) = \Phi(-2.12) = 1 - \Phi(2.12) = 0.0170.$$

Låt X vara antalet brädor av 300 som är kortare än 185 cm. Det följer att $X \sim \text{Bin}(300, p)$. Sannolikheten p är alltså liten, och $300p(1-p) = 5.01$ är betydligt mindre än 10, så normalapproximation fungerar antagligen inte. Men Poissonapproximation borde fungera bra då $p \ll 0.1$ och $n = 300 \gg 10$. Alltså är $X \overset{\text{appr.}}{\sim} \text{Po}(300 \cdot 0.0170) = \text{Po}(5.10)$. Ur tabell (interpolation mellan Po(5.0) och Po(5.2)):

$$P(X \leq 4) \approx \frac{0.4405 + 0.4061}{2} = 0.4233.$$

Alltså ungefär 42% chans. Verklig sannolikhet blir 42.15%. Normalapproximation skulle i fallet ge 31%, vilket är alldeles för lågt.

6 Approximation av Poissonfördelning

Approximation av Poissonfördelning

Sats. Låt $X \sim \text{Po}(\mu)$ med $\mu \geq 15$. Då är $X \overset{\text{appr.}}{\sim} \text{N}(\mu, \sqrt{\mu})$ ($V(X) = \mu$).

Exempel

Låt X vara antal paket i en datakö under en sekund. Mätningar har visat att en vettig modell är $X \sim \text{Po}(250)$. Beräkna approximativt $P(X < 240)$.

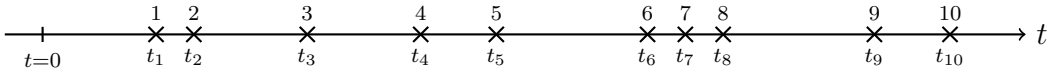
Lösning: Eftersom väntevärdet $\mu = 250 \gg 15$ så kan vi normalapproximera. Då blir

$$P(X < 240) = P(X \leq 239) \approx \Phi\left(\frac{239 - 250}{\sqrt{250}}\right) = \Phi(-0.6957) = 1 - \Phi(0.6957) = 0.2433.$$

Exakt värde: 0.2552.

7 $(\star)$Poissonfördelning

Antag att vi har en situation där händelser inträffar oberoende av varandra med en konstant intensitet λ, det vill säga, på t tidsenheter inträffar i genomsnitt λt händelser. Denna typ av situation brukar ofta moduleras med hjälp av *Poisson*-fördelningen. Om $X(t)$ är antalet händelser i tidsintervallet $[0, t]$, så säger vi att $X(t)$ är Poissonfördelad med väntevärde $\mu = \lambda t$.



Händelser (markerade med kryss och numrerade) i tidsintervallet $[0, t]$. Tiderna t_k är tidpunkten för händelsen k.

Poissonfördelning

Sats. Vi kallar X för Poissonfördelad med parametern μ, $X \sim \text{Po}(\mu)$, om sannolikhetsfunktionen ges av

$$p_X(k) = P(X = k) = \frac{\mu^k e^{-\mu}}{k!}, \quad k = 0, 1, 2, \ldots.$$

Variabeln X har $E(X) = V(X) = \mu$ (samma väntevärde som varians, parametern μ).

Hur hänger situationen ovan ihop med definitionen av p_X? Vi fixerar tiden t och delar in intervallet $[0, t]$ i n lika stora delar, där vi väljer n så stort att det högst finns en händelse i varje delintervall. Vi introducerar en sannolikhet p som är sannolikheten att ett visst delintervall innehåller en händelse. Det är samma p för alla delintervall och sambandet $np = \lambda t$ måste gälla. Eftersom händelserna är oberoende måste $X(t) \sim \text{Bin}(n, p)$. Egenskaper för binomialfördelningen medför att $E(X(t)) = np = \lambda t$.

Vi börjar med att betrakta fallet med noll händelser i intervallet $[0, t]$, det vill säga, händelsen att $X(t) = 0$:

$$P(X(t) = 0) = \binom{n}{0} p^0 (1-p)^{n-0} = \left(1 - \frac{\lambda t}{n}\right)^n = \exp\left(n \ln\left(1 - \frac{\lambda t}{n}\right)\right) \to e^{-\lambda t},$$

då $n \to \infty$. Här har vi utnyttjat standardgränsvärdet $s^{-1} \ln(1 + s) \to 1$ då $s \to 0$.

I det allmänna fallet kan vi visa att

$$P(X(t) = k) \to \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad n \to \infty.$$

För att se detta, låt k vara fix och betrakta

$$\begin{aligned} P(X(t) = k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{(n-k)!k!} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{(\lambda t)^k}{k!} \cdot \frac{\left(1 - \frac{\lambda t}{n}\right)^n}{\left(1 - \frac{\lambda t}{n}\right)^k} \to 1 \cdot \frac{(\lambda t)^k}{k!} \cdot \frac{e^{-\lambda t}}{1}, \quad n \to \infty. \end{aligned}$$

Vad detta innebär är att om $X_n \sim \text{Bin}(n, \lambda/n)$, så kommer $X_n \xrightarrow{\text{D}} X \sim \text{Po}(\lambda)$.

Detta ger oss även en användbar approximationssats för binomialfördelningen.

Approximation: Binomial till Poisson

Sats. Om $X \sim \text{Bin}(n, p)$ med $n \geq 10$ och $p \leq 0.1$, så är $X \overset{\text{appr.}}{\sim} \text{Po}(np)$.

Att p_X verkligen är en sannolikhetsfunktion följer från Maclaurinuteckling av e^x:

$$\sum_{k=0}^{\infty} p_X(k) = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} \cdot e^{\mu} = 1.$$

Låt oss även härleda väntevärde och varians. För väntevärdet:

$$\sum_{k=0}^{\infty} k p_X(k) = e^{-\mu} \left(0 + \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} \right) = \mu e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = \mu.$$

Variansen är lite bökigare. Vi kan inte direkt räkna ut $E(X^2)$, utan tar till omskrivningen

$$E(X^2) = E(X(X-1)) + E(X).$$

Alltså,

$$\begin{aligned} E(X(X-1)) &= \sum_{k=0}^{\infty} k(k-1) p_X(k) = e^{-\mu} \left(\sum_{k=2}^{\infty} \frac{\mu^k}{(k-2)!} \right) = \mu^2 e^{-\mu} \sum_{k=2}^{\infty} \frac{\mu^{k-2}}{(k-2)!} \\ &= \mu^2 e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = \mu^2, \end{aligned}$$

så $E(X^2) = \mu^2 + \mu$, vilket medför att $V(X) = E(X^2) - E(X)^2 = \mu$.

Addition av oberoende Poissonfördelade variabler

Sats. Låt $X \sim \text{Po}(\mu_1)$ och $Y \sim \text{Po}(\mu_2)$ vara oberoende. Då är $X + Y \sim \text{Po}(\mu_1 + \mu_2)$.

Satsen förefaller intuitivt att vara rimlig. Vi lägger helt enkelt ihop händelserna från två liknande processer, det förväntade antalet blir nu $\mu_1 + \mu_2$, och på grund av beteendet hos var och en tippar vi att summan fungerar på samma sätt. Formellt kan vi visa satsen medelst den så kallade faltningssatsen. Den simultana sannolikhetsfunktionen för (X, Y) ges av produkten $p_X(i)p_Y(j)$, och vi söker sannolikhetsfunktionen för $Z = X + Y$. Alltså,

$$p_Z(k) = P(X + Y = k) = \sum_{i+j=k} \sum p_X(i) p_Y(j) = \sum_{i=0}^{k} p_X(i) p_Y(k-i).$$

Dubbelsumman blir en enkelsumma eftersom vi bara summerar över ”diagonalen” (när k är fixt och $i + j = k$). Sen är $p_X(i) = 0$ då $i < 0$ och $p_Y(k-i) = 0$ då $i > k$ så det räcker att summera från $i = 0$ till $i = k$. Vidare,

$$\begin{aligned} p_Z(k) &= \sum_{i=0}^{k} e^{-\mu_1} \frac{\mu_1^i}{i!} e^{-\mu_2} \frac{\mu_2^{k-i}}{(k-i)!} = \frac{e^{-(\mu_1+\mu_2)}}{k!} \sum_{i=0}^{k} \frac{k!}{(k-i)!i!} \mu_1^i \mu_2^{k-i} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{k!} \sum_{i=0}^{k} \binom{k}{i} \mu_1^i \mu_2^{k-i} = \frac{e^{-(\mu_1+\mu_2)}}{k!} (\mu_1 + \mu_2)^k, \end{aligned}$$

där vi utnyttjat binomialsatsen i sista steget. Detta uttryck är inget annat än sannolikhetsfunktionen för en $\mathrm{Po}(\mu_1 + \mu_2)$-fördelad variabel, vilket var precis det vi ville visa!

Denna sats kan vi använda för att dela upp en $\mathrm{Po}(\mu)$ fördelad variabel i $\lfloor \mu \rfloor$ stycken oberoende variabler med väntevärde ett, och en liten svans (med längd $\mu - \lfloor \mu \rfloor$). På detta sätt kan man visa följande sats.

Approximation av Poissonfördelning

Sats. Låt $X \sim \mathrm{Po}(\mu)$ med $\mu \geq 15$. Då är $X \overset{\text{appr.}}{\sim} \mathrm{N}(\mu, \sqrt{\mu})$ $(V(X) = \mu)$.