F8 BINOMIALFÖRDELNINGEN, BESKRIVANDE STATISTIK

Kap. 7.1–7.4

Linnéa Gustafsson
linneag2@kth.se

- Binomialfördelningen och dess släktingar
- Approximationer (till normalfördelningen)

KAP. 7

Hypergeometriska fördelningen

Anta att vi har N enheter där andelen enheter med egenskap A är p.

Vi drar n st enheter utan återläggning. Låt X vara antalet enheter med egenskap A som vi får.

Då gäller att $X \in Hyp(N, n, p)$

$$p_X(k) = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}$$ se F.S. §3

Binomialfördelningen

"med återläggning", men måste inte vara egentlig återläggning

Anta att händelsen A inträffar med samma sannolikhet $P(A) = p$ i varje försök.

Anta att X är antal lyckade försök då A inträffar.

Då gäller att $X \in Bin(n, p)$

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$ se F.S. §3

Varför vill man det? Bin-p_X enklare än Hyp-p_X

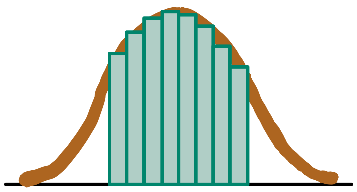
När kan vi approximera $Hyp(N, n, p)$ till $Bin(n, p)$?

Om vi har att $\frac{n}{N} \leq 0.1$ gäller att $Hyp(N, n, p) \sim Bin(n, p)$ se F.S. §6

För $Bin(n,p)$:

Approximationen $Bin(n,p) \sim N(\overbrace{np}^{E(X)}, \overbrace{\sqrt{np(1-p)}}^{D(X)})$

se F.S. §6



Anta att $X \in Bin(n,p)$ där X är antalet lyckade försök.

Vi kan skriva $X = I_1 + I_2 + \dots I_n$ där $I_k = 0$ när försöket misslyckas och $I_k = 1$ när försöket lyckas.

- $P(I_k = 0) = 1 - p$
- $P(I_k = 1) = p$
- $E(I_k) = \sum_{\text{alla } i} i\, p_{I_k}(i) = 0 \cdot (1-p) + 1 \cdot p = p$
- $E(X) = E[I_1 + \dots + I_n] = n \cdot p$
- $V(I_k) = E[I_k^2] - E^2(I_k) = \left[E[I_k^2] = \sum i^2 p_{I_k}(i) = 0^2(1-p) + 1^2 \cdot p = p \right] = p - p^2 = p(1-p)$
- $V(X) = V(I_1 + \dots I_n) = [\text{ober.}] = \sum V(I_k) = np(1-p)$

$\boxed{Bin(np) \sim N(np, \sqrt{np(1-p)})}$ om $np(1-p) \geq 10$ se F.S. §6

Detta är egentligen ett CGS-villkor ty det gäller ju om n tillräckligt stort att $X = I_1 + I_2 + \dots I_n$ är approximativt N-fördelad ty I_k:na är ju ober. och likafördelade

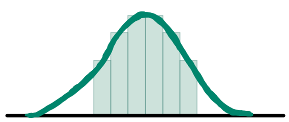
Faktorn $p(1-p)$ är ett symmetrivillkor. Ju mer symmetrisk en fördelning är ju mindre behöver n vara för att CGS ska gälla.

Ex Om $p = 1 - p = 0.5$ räcker det att $n = 40$

Om $p = \frac{1}{1000}$ måste n vara $> 10\,000$

Halvkorrektion

När vi approximerar en diskret fördelning till en kontinuerlig fördelning så går vi från att summera arean av stolpar med basen 1 och höjden $p_X(x)$ genom att integrera täthetsfunktionen $\int_a^b f_X(x)\,dx$.



I det diskreta fallet är t.ex. $P(120 \leq X \leq 180) =$
$= P(119 < X < 181)$

Vilka gränser ska väljas i $\int f_X(x)\,dx$?

Kompromissen blir då $P(119.5 < X < 180.5) = \int_{119.5}^{180.5} f_X(x)\,dx$

Detta kallas halvkorrektion.

Vi behöver inte använda detta
men det blir bättre om man gör det!

Alla tre typer av gränser är godkända

Tentauppgift 2017-08-14, 2

Uppgift 2

Den trendiga badbutiken Poolen och Plurret vill i slutet av säsongen bli av med de 2 000 badbyxorna som är kvar av årets kollektion. Därför anordnar man en realisation på de kvarvarande badbyxorna. Av erfarenhet vet man att antalet badbyxor som en kund köper på sådana realisationer dels är oberoende av hur många badbyxor andra kunder köper, dels kan betraktas som en stokastisk variabel X med sannolikhetsfunktionen

$$p_X(k) = \begin{cases} 0.2 & \text{om} \quad k = 0, \\ 0.5 & \text{om} \quad k = 1, \\ 0.2 & \text{om} \quad k = 2, \\ 0.1 & \text{om} \quad k = 3, \\ 0 & \text{om} \quad k > 3. \end{cases}$$

Beräkna med lämplig och välmotiverad approximation det minsta antalet kunder som måste komma till butiken om sannolikheten ska vara 90% att butiken ska få samtliga kvarvarande badbyxor sålda.

X_i = antal badbyxor kund i köper

$p_{X_i}(0) = 0.2$
$p_{X_i}(1) = 0.5$
$p_{X_i}(2) = 0.2$
$p_{X_i}(3) = 0.1$
$p_{X_i}(i) = 0$ för $i > 3$

Beräkna med lämplig och välmotiverad approximation det minsta värdet på n där n är antalet kunder så att sannolikheten att de 2000 badbyxorna säljs är minst 90 %.

Lösning

Y är totala antalet badbyxor som säljs.

$Y = X_1 + X_2 + \dots + X_n$

X_i:na är många, ober. och likafördelade $\Rightarrow$ $Y \sim N(n\mu, \sqrt{n}\sigma)$ enl. CGS se F.S. §5

$$\mu = E(X_i), \quad \sigma = D(X_i)$$

Sökt: Det minsta n så att $P(Y \geq 2000) = 0.90$

$$\left.\begin{aligned} \mu = E(X_i) &= 0\cdot 0.2 + 1\cdot 0.5 + 2\cdot 0.2 + 3\cdot 0.1 = 1.2 \\ E[X_i^2] &= 0^2\cdot 0.2 + 1^2\cdot 0.5 + 2^2\cdot 0.2 + 3^2\cdot 0.1 = 2.2 \end{aligned}\right\} \Rightarrow V(X_i) = 2.2 - 1.2^2 = 0.76$$

$$Y \sim N(n\cdot 1.2, \sqrt{0.76}\sqrt{n})$$

Eftersom Y är diskret stokastisk variabel från början kan vi lika gärna räkna ut
$P(Y > 1999) = 0.90$

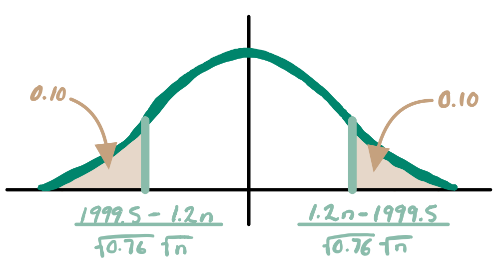
Halvkorrektion $\Rightarrow$ vi räknar ut $P(Y > 1999.5) = 0.90$

Vi approximerar till N-fördelning m.h.a. CGS och gör om till $N(0,1)$

$$P\Bigg(\underbrace{\frac{Y - n\cdot 1.2}{\sqrt{0.76}\sqrt{n}}}_{Z \sim N(0,1)} > \frac{1999.5 - n\cdot 1.2}{\sqrt{0.76}\sqrt{n}}\Bigg) = 0.90$$

$$P\left(Z > \frac{1999.5 - n\cdot 1.2}{\sqrt{0.76}\sqrt{n}}\right) = 0.90$$

Gränsen söks, sannolikheten känd $\Rightarrow$ tab. 2, $P(Z > \lambda_\alpha) = \alpha$



Symmetri i figuren

$$\Rightarrow \frac{1.2n - 1999.5}{\sqrt{0.76}\sqrt{n}} = \lambda_{0.10} = \underbrace{1.2816}_{\text{tab. 2}}$$

Sätt $\sqrt{n} = m$ $\Rightarrow$ $1.2m^2 - 1999.5 = 1.2816\sqrt{0.76}\, m$ $\Rightarrow$ $m^2 - \frac{1.2816\sqrt{0.76}\, m}{1.2} - \frac{1999.5}{1.2} = 0$

$\Rightarrow$ $m = 0.46553 \pm 40.82238$, m måste vara positiv $\Rightarrow$ $n = m^2 = 1704.69$

Svar: Det måste komma minst 1705 kunder.

Poissonfördelningen

Om vi har ett visst intervall där en viss typ av händelse inträffar med samma sannolikhet i varje punkt och X är antalet händelser som inträffar på detta intervall så gäller att

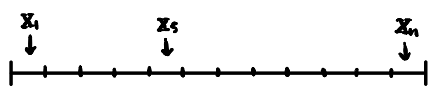
$X \in Po(\mu)$

> **Viktig sats som ej står i F.S.**
>
> Om $X \in Po(\mu_1)$
> och $Y \in Po(\mu_2)$
> och X och Y ober.
> $\Rightarrow X + Y \in Po(\mu_1 + \mu_2)$

N-approximation med Poissonfördelningen

Enligt F.S. §6 gäller att $Po(\mu) \sim N(\mu, \sqrt{\mu})$ om $\mu \geq 15$

__Motivering:__



Dela upp intervallet i n st delintervall

$X_i \in Po(\mu_i)$

$Y = X_1 + X_2 + \dots + X_n$, om X_i:na är ober. $\underbrace{\Rightarrow}_{\text{satsen ovan}} Y \in Po(\mu_1 + \mu_2 + \dots + \mu_n) = Po(\mu)$

Även här kan vi se villkoret som ett CGS-villkor, ty om n tillräckligt stort får vi ett $\mu > 15$

Approximationen $Bin(n,p) \sim Po(np)$

Se F.S. §6 Helst vill man

$$\begin{tikzcd} \text{Hyp} \arrow[r] & \text{Bin} \arrow[r] \arrow[d] & N \\ & Po \arrow[ur] & \end{tikzcd}$$

Om detta inte går: testa Po istället

Bin-fördelningen går att approximera till Po-fördelningen, om $p \leq 0.1$ gäller att:

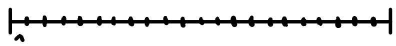
$Bin(n,p) \sim Po(np)$ se F.S. §6

__Bevis:__ Anta att vi har att X är antal händelser som inträffar av en viss typ på intervallet I.

I:

$X \in Po(\mu)$

Dela upp intervallet i n st delintervall där n är stort:

I: 

Varje delintervall är så litet att det bara inträffar högst en händelse där med sannolikhet p

Då blir totala antalet händelser binomialfördelat Bin(n, p)

Vi visar nu att binomialfördelningens sannolikhetsfunktion övergår i Poissonfördelningens sannolikhets-funktion då $p \to 0$, $n \to \infty$ och $np \to \mu$

Börja med Bin-fördelningen:

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k} = \left[\text{ska bli } \frac{\mu^k}{k!}e^{-\mu}\right] =$$

$$= \frac{1}{k!} n(n-1)(n-2)\dots(n-k+1)\frac{1}{n^k}\left(1-\frac{\mu}{n}\right)^{n-k}\mu^k = \frac{\mu^k}{k!}\cdot\underbrace{\frac{n}{n}\cdot\frac{n-1}{n}\cdot\ldots\cdot\frac{n-k+1}{n}}_{\to 1 \text{ ty } n>>k}\cdot\overbrace{\left(1-\frac{\mu}{n}\right)^n}^{\to e^{-\mu} \text{ ty } n\to\infty}\cdot\underbrace{\frac{1}{\left(1-\frac{\mu}{n}\right)^k}}_{\to 1 \text{ ty } n>>k} =$$

$$= \frac{\mu^k}{k!}e^{-\mu} \quad \text{d.v.s.} \quad p_X(k) = \frac{\mu^k}{k!}e^{-\mu} \qquad \underline{\underline{\text{V.S.V}}}$$

Enl. §6 gäller att då $p < 0.1$: $\text{Bin}(n, p) \sim \text{Po}(np)$

Poissonfördelningen egentligen binomialfördelning i botten!