F9 Normalfördelningen och CGS

Christian Tallberg

Avdelningen för Nationalekonomi/Statistik

Karlstads universitet

Normalfördelningen

- Formell definition: X är en normalfördelad variabel om täthetsfunktionen ges av:

$$f_x(x) = \frac{1}{\sigma sqrt(2\pi)} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ där } -\infty < x < \infty.$$

- Kortfattat skriver vi att $X \in N(\mu; \sigma)$.

Viktiga egenskaper:

- $E(X) = \mu$ där $-\infty < \mu < \infty$.
- $V(X) = \sigma^2$ där $0 < \sigma < \infty$.
- Utfallsrummet är hela reella talaxeln.
- De två parametrarna μ och σ (eller σ^2) bestämmer normalfördelningens utseende.
- Normalfördelningen är symmetrisk kring sitt väntevärde μ.

Kuriosa:

För en normalfördelning gäller att:

- Ca 68 % av fördelningen finns inom $\mu \pm \sigma$
- Ca 95 % av fördelningen finns inom $\mu \pm 2\sigma$
- Ca 99.7 % av fördelningen finns inom $\mu \pm 3\sigma$, dvs nästan hela fördelningen finns inom gränserna $\mu \pm 3\sigma$.

Varför är normalfördelningen viktig?

- Ofta beskrivs variabiliteten i populationer bra av en normalfördelning.
- Kan användas som approximation till vissa andra sannolikhetsfördelningar.
- Vid stora stickprov kan dessutom fördelningen för många estimatorer (skattningar av populationsparametrar) approximeras med normalfördelningen (*Centrala Gränsvärdessatsen*). (Vi återkommer till detta senare i kursen).
- Detta medför att många statistiska metoder är baserade på antagandet att data kommer från normalfördelade populationer eller att estimatorer är (approximativt) normalfördelade.

Beräkning av sannolikheter för N(0;1)

- Låt X vara en standardiserad normalfördelad variabel, dvs X är $N(0;1)$.
- *Täthetssfunktionen* för X, $f_x(x)$ betecknas
$$\rho(x).$$
- *Fördelningsfunktionen* för X, $P(X \leq x)$, betecknas
$$\Phi(x).$$
- Med hjälp av *tabell 4* i tabellsamlingen kan vi beräkna fördelningsfunktionens sannolikheter.

Exempel:

Variabeln X är $N(0;1)$. Beräkna följande sannolikheter:

1.
$$\begin{aligned} P(X \leq 1) &= \int_{-\infty}^{1} \rho(x)\,dx \\ &= 0.8413 \text{ enl tabell} \end{aligned}$$

2.
$$\begin{aligned} P(0 \leq X \leq 1) &= P(X \leq 1) - P(X \leq 0) \\ &= \Phi(1) - \Phi(0) \\ &= 0.8413 - 0.5 = 0.3413 \end{aligned}$$

3.
$$\begin{aligned} P(X > 0.73) &= 1 - \Phi(0.73) \\ &= 1 - 0.7673 = 0.2327 \end{aligned}$$

Alternativ 2:
$$\begin{aligned} P(X > 0.73) &= \Phi(-0.73) \\ &= 0.2327 \end{aligned}$$

4.
$$\begin{aligned} P(X > -1.59) &= \Phi(1.59) \\ &= 0.9441 \end{aligned}$$

5.
$$\begin{aligned} & P(-1.59 \leq X \leq 0.73) \\ &= \Phi(0.73) - \Phi(-1.59) \\ &= 0.7673 - 0.0559 \\ &= 0.7114 \end{aligned}$$

Beräkning av sannolikheter för godtycklig normalfördelning

- Endast $N(0;1)$, dvs den normerade (standardiserade) normalfördelningen finns i tabell.

- Om variabeln X är $N(\mu;\sigma)$, så är den standardiserade variablen

$$Y = \frac{X-\mu}{\sigma} \sim N(0;1).$$

Bevis:

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

Följande sannolikhet beräknas alltså som

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

Exempel:

Variabeln X är $N(170;10)$. Beräkna följande sannolikheter.

1.

$$\begin{aligned} P(X \leq 190) &= \Phi\left(\frac{190-170}{10}\right) \\ &= \Phi(2) = 0.97725 \end{aligned}$$

2.

$$\begin{aligned} P(X \leq 160) &= P\left(\frac{X-170}{10} \leq \frac{160-170}{10}\right) \\ &= P(Z \leq -1) = 0.1587 \end{aligned}$$

3.

$$\begin{aligned} & P(160 \leq X \leq 190) \\ = \; & P(X \leq 190) - P(X \leq 160) \\ = \; & 0.97725 - 0.1587 \\ = \; & 0.81855 \end{aligned}$$

4.

$$\begin{aligned} P(X \geq 190) &= 1 - P(X \leq 190) \\ &= 1 - 0.97725 = 0.0228 \end{aligned}$$

I tabellen kan vi för olika värden på sannolikheterna

$$P(Z \leq z)$$

och

$$P(|Z| \geq z)$$

finna tillhörande $z-$ värde.

Exempel:

- Hitta ett $x-$ värde sådant att vi har sannolikheten 0.01 i högra svansen, dvs ett $x-$ värde sådant att

$$\begin{aligned} P(X \geq x) &= 0.01 \\ 1 - \Phi(x) &= 0.01 \\ \Phi(x) &= 0.99. \end{aligned}$$

Sök på $\Phi(x) = 0.99$ vilket ger $x = 2.326$.

- Bestäm talet a så att $P(X \geq a) = 0.05$.

Lösning: Enligt tabellen är

$$\Phi(1.64) = 0.95$$

vilket medför att $1 - \Phi(1.64) = 0.05$.

$$\begin{aligned} P(X \geq a) &= 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) = 0.05 \\ &\Rightarrow \frac{a-\mu}{\sigma} = 1.64 \\ \frac{a-170}{10} &= 1.64 \\ a &= 1.64 \cdot 10 + 170 = 186.4 \end{aligned}$$

Linjärkombinationer av oberoende normalfördelade variabler

Tidigare undersökte vi egenskaper hos linjärkombinationer av stokastiska variabler, t ex väntevärden och varianser hos summor. Vad kan vi säga om *fördelningen* för linjärkombinationer av stokastiska variabler?

- Om $X \in N(\mu_x, \sigma_x)$ och $Y \in N\left(\mu_y, \sigma_y\right)$, där X och Y är oberoende, så gäller att

$$\begin{aligned} X + Y &\in N\left(\mu_x + \mu_y, sqrt\left(\sigma_x^2 + \sigma_y^2\right)\right) \\ X - Y &\in N\left(\mu_x - \mu_y, sqrt\left(\sigma_x^2 + \sigma_y^2\right)\right). \end{aligned}$$

Exempel: Johans och Marias utgifter under en månad kan antas variera som två stycken oberoende normalfördelade variabler med följande väntevärden och varianser (enhet: kronor):

$$\begin{aligned} X &\in N\left(6020; 30^2\right) \quad \text{(Johans utgifter)} \\ Y &\in N\left(5940; 25^2\right) \quad \text{(Marias utgifter)} \end{aligned}$$

a) Beräkna sannolikheten för att deras sammanlagda utgifter under en månad överstiger 12000 kronor.

b) Beräkna sannolikheten för att Marias utgifter under en månad överstiger Johans.

Lösning: $X =$ Johans utgifter, $Y =$ Marias utgifter

a) $(X+Y) =$ Sammanlagda utgifter

$$\begin{aligned} E(X+Y) &= E(X)+E(Y) \\ &= 6020+5940=11960 \\ V(X+Y) &= V(X)+V(Y)+2Cov(X;Y) \\ &= V(X)+V(Y) \ \text{(ty } X \text{ och } Y \text{ ober.)} \\ &= 30^2+25^2=1525 \end{aligned}$$

Vi har att

$$(X+Y) \in N(11960;1525)$$

Den sökta sannolikheten är då

$$\begin{aligned} P((X+Y)>12000) &= 1-\Phi\left(\frac{12000-11960}{sqrt(1525)}\right) \\ &= 1-\Phi(1.02) \\ &= 1-0.8461=0.1539 \end{aligned}$$

b) $(Y-X) =$ Skillnad i utgifter

$$\begin{aligned} E(Y-X) &= E(Y)-E(X) \\ &= 59400+6020=-80 \\ V(Y-X) &= V(Y)+V(X)-2Cov(X;Y) \\ &= V(Y)+V(X) \ \text{(ty } X \text{ och } Y \text{ ober.)} \\ &= 1525 \end{aligned}$$

Vi har att

$$(Y-X) \in N(-80;1525)$$

Den sökta sannolikheten är då

$$\begin{aligned} P(Y>X) &= P((Y-X)>0) \\ &= 1-\Phi\left(\frac{0-(-80)}{sqrt(1525)}\right) \\ &= 1-\Phi(2.05) \\ &= 1-\Phi(2.05) \\ &= 1-0.9798=0.0202 \end{aligned}$$

- Mer generellt: Om $X_1, X_2, ..., X_n$ är oberoende $N(\mu,\sigma)$, så gäller att

$$\begin{aligned} \sum_{i=1}^{n} X_i &\in N(n\mu, sqrt(n)\,\sigma) \\ \frac{\sum_{i=1}^{n} X_i - n\mu}{sqrt(n\sigma)} &\in N(0,1). \end{aligned}$$

Dessutom är fördelningen för det aritmetiska medelvärdet $\sum_{i=1}^{n} X_i/n$

$$\begin{aligned} \bar{X} &\in N\left(\mu; \frac{\sigma}{sqrt(n)}\right) \\ \frac{\bar{X}-\mu}{\frac{\sigma}{sqrt(n)}} &\in N(0;1). \end{aligned}$$

I situationer då man drar ett oberoende stickprov från en population, brukar fördelningen för stickprovsmedelvärdet kallas *samplingfördelningen*.

Centrala gränsvärdessatsen

Hur ser stickprovsmedelvärdets samplingfördelning ut om observationerna inte är normalfördelade? Det kan vi ofta inte säga något om med exakthet. Men vi kan uttala oss om hur samplingfördelningen för $\bar{X}$ ser ut *approximativt*, när vi har *stort stickprov*.

Centrala gränsvärdessatsen (CGS)
Summan av n oberoende slumpvariabler med samma fördelning är *ungefär normalfördelade* om n är tillräckligt stort.

- Vad menas med att n är tillräckligt stort? Vanlig tumregel: $n \geq 30$.

- CGS ger en förklaring till normalfördelningens användbarhet.

- Konsekvens av CGS: Summor och medelvärden vid stora stickprov är approximativt (eller asymptotiskt) normalfördelade, oavsett hur populationens fördelning ser ut. Dvs

$$\begin{aligned} \sum_{i=1}^{n} X_i &\in AsN\left(n\mu, sqrt\left(n\right)\sigma\right) \\ \frac{\sum_{i=1}^{n} X_i - n\mu}{sqrt\left(n\sigma\right)} &\in AsN\left(0,1\right) \end{aligned}$$

och

$$\begin{aligned} \bar{X} &\in AsN\left(\mu; \frac{\sigma}{sqrt\left(n\right)}\right) \\ \frac{\bar{X}-\mu}{\frac{\sigma}{sqrt(n)}} &\in AsN\left(0;1\right). \end{aligned}$$

- Det innebär att oavsett vad populationen har för fördelning kan vi beräkna sannolikheter som

$$P\left(\bar{X} \leq a\right) = \Phi\left(\frac{a-\mu}{\frac{\sigma}{sqrt(n)}}\right).$$

Exempel:

I en stor population är medelvärdet $\mu = 65$ och standardavvikelsen $\sigma = 8$. Man väljer helt slumpmässigt ut 100 "element". Vad är sannolikheten att man får ett stickprovsmedelvärde som understiger 64?

Fördelningen för X är okänd, men stickprovet är stort ($n = 100$). Det innebär att, $\bar{X} \sim AsN\left(65; \frac{8}{10}\right)$ enligt CGS, varför

$$\begin{aligned} P\left(\bar{X} < 64\right) &= \Phi\left(\frac{64-\mu}{\frac{\sigma}{sqrt(n)}}\right) = \Phi\left(\frac{64-65}{\frac{8}{10}}\right) \\ &= \Phi\left(-1.25\right) = 0.1056. \end{aligned}$$

Binomialfördelningen och dess släktingar

Om X är $Bin\left(n,p\right)$ gäller att

$$\begin{aligned} E\left(X\right) &= np \\ V\left(X\right) &= np\left(1-p\right) \\ SD\left(X\right) &= sqrt\left(np\left(1-p\right)\right). \end{aligned}$$

Binomialfördelningen kan approximeras med normalfördelningen om (tumregel enligt Blom) $np\left(1-p\right) > 10$. Dvs om

$$X \sim Bin\left(n;p\right)$$

är

$$\begin{aligned} X &\sim approx\ N\left(\mu;\sigma\right) \\ \text{där } \mu &= np \\ \text{och } \sigma &= sqrt\left(np\left(1-p\right)\right) \end{aligned}$$

- Anledningen till att man vill approximera fördelningar med en normalfördelning, är att man vill förenkla räknearbetet. Följande sannolikhet kan då beräknas som

$$\begin{aligned} P\left(a \leq X \leq b\right) &= P\left(X \leq b\right) - P\left(X < a\right) \\ &\approx \Phi\left(\frac{b-np}{sqrt\left(np\left(1-p\right)\right)}\right) \\ &\quad -\Phi\left(\frac{a-np}{sqrt\left(np\left(1-p\right)\right)}\right) \end{aligned}$$

Exempel:

Variabeln X är $Bin\,(44; 0.45)$. Beräkna $P\,(X \leq 26)$.

Vi har att $np\,(1-p) = 44 \cdot 0.45 \cdot 0.55 = 10.89 > 10$

Alltså kan vi approximera med normalförd.

$$\begin{aligned} E\,(X) &= \mu = np = 19.8 \\ V(X) &= \sigma^2 = np\,(1-p) = 44 \cdot 0.45 \cdot 0.55 = 10.89, \end{aligned}$$

dvs $X \sim approx\ N\,(19.8; 10.89)$

- $$\begin{aligned} P\,(X \leq 26) &= \Phi\left(\frac{26 - 19.8}{sqrt\,(10.89)}\right) \\ &= \Phi\,(1.88) = 0.9699 \end{aligned}$$

- Exakt enligt SPSS (dvs utan approximation)

$$P\,(X \leq 26) = 0.9786$$

Hypergeometriska fördelningen

Om X är $Hyp\,(N, n, p)$ gäller att

$$\begin{aligned} E\,(X) &= np \\ V\,(X) &= np\,(1-p)\,\frac{N-n}{N-1}, \end{aligned}$$

där $(N-n)\,/\,(N-1)$ kallas *korrektionsfaktorn för ändlig population*.

Exempel forts från ett par föreläsningar tidigare:

Om lamporna dras *med återläggning*,

a) vilken fördelning har X?

b) Vad är sannolikheten att högst en av de dragna lamporna är defekta?

c) Vad är sannolikheten att minst en av de dragna lamporna är defekta?

d) Beräkna väntevärde och varians.

Svar:

a) X är $Bin\,(5; 0.3)$

b)

$$P\,(X \leq 1) = 0.528 \text{ enl tab 2.}$$

c)

$$\begin{aligned} P\,(X \geq 1) &= 1 - P\,(X = 0) \\ &= 1 - 0.168 = 0.832 \text{ enl tab 2.} \end{aligned}$$

d)

$$\begin{aligned} E\,(X) &= n\pi = 5 \cdot 0.3 = 1.5 \\ V\,(X) &= 5 \cdot 0.3\,(1 - 0.3) = 1.05 \end{aligned}$$

Om X är $Hyp\,(n; S; N)$ och N är stort i förhållande till n (Tumregel enligt kurslitteraturen: $\frac{n}{N} < 0.05$) gäller att X är approximativt $Bin\,(n; \pi)$, och väntevärde och varians ges av:

$$\begin{aligned} E\,(X) &= n\pi \\ V\,(X) &\approx nn\pi\,(1 - \pi)\,, \end{aligned}$$

eftersom

$$\frac{N-n}{N-1} = \frac{N}{N-1} - \frac{n}{N-1} \approx 1 - 0 = 1.$$

Poissonfördelningen

Om X är $Po\,(\mu)$ gäller att

$$\begin{aligned} E\,(X) &= \mu \\ V\,(X) &= \mu. \end{aligned}$$

Poissonfördelningen kan approximeras med normalfördelningen om (tumregel enligt Blom) $\mu > 15$.