

SF1901: SANNOLIKHETSTEORI OCH
STATISTIK
FÖRELÄSNING 7
NORMALFÖRDELNING.
LINJÄRKOMBINATION AV OBEROENDE
NORMALFÖRDELADE S.V.
CENTRALA GRÄNSVÄRDESSATSEN.

Tatjana Pavlenko

13 april 2018



PLAN FÖR DAGENS FÖRELÄSNING

- ▶ Linjära kombinationer av s.v.: repetition (Kap. 5.5)
- ▶ Stora talens lag: repetition (Kap. 5.6)
- ▶ Standardiserad och allmän normalfördelning (Kap. 6.2-6.3)
- ▶ Linjärkombinationer av oberoende normalfördelade s.v. (Kap. 6.5)
- ▶ Centrala gränsvärdesatsen och normalfördelningsapproximation (Kap. 6.7)



STORA TALENS LAG (REP.)

- ▶ **Sats:** Låt X_1, X_2, \dots vara en följd av oberoende (likafördelade) s.v., alla med samma väntevärde $E(X_i) = \mu$ och std $D(X_i) = \sigma < \infty$.

Låt

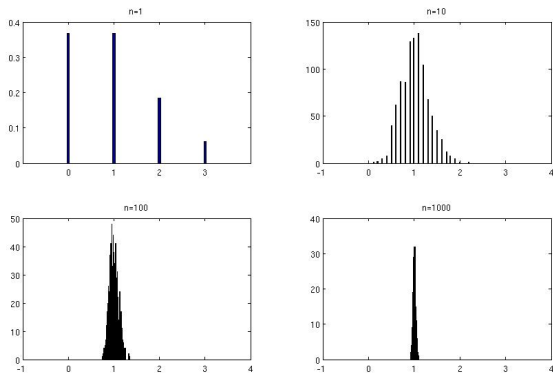
$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$$

vara medelvärdet av de n första variablerna. Då gäller, för godtyckligt $\varepsilon > 0$, att

$$P(|\bar{X}_n - \mu| > \varepsilon) \rightarrow 0 \quad \text{då} \quad n \rightarrow \infty.$$

- ▶ Detta säger att s.v \bar{X}_n , bestående av medelvärdet av de n s.v. X_1, \dots, X_n har en fördelning som koncentrerar sig runt $\mu = E(X_i)$.





FIGUR: Fördelningen för medelvärdet $\bar{X}_n = (X_1 + \dots + X_n)/n$, där de enskilda observationerna är $Po(1)$, för $n = 1, 10, 100$ och 1000 . Den diskreta fördelningen koncentreras alltmer runt värdet 1, dvs väntevärdet.

STORA TALENS LAG (REP.)

- ▶ *Tolkning:* medelvärde är en bra uppskattning av väntevärde!
- ▶ *Stora talens lag är en av grundstenarna inom empirisk vetenskap.* Om man gör många oberoende observationer av någon s.v., t ex hållfastheten hos en metall, blodtrycket hos patienter behandlade med en ny medicin, eller livslängden hos personer i ett försäkringskollektiv, då kommer medelvärdena av dessa observationer att ligga nära det *sanna* väntevärdet.
- ▶ *Om vi inte vet det sanna väntevärdet kan vi gissa, eller skatta väntevärdet med medelvärdet.* Detta förhållande är en viktig ingrediens i statistik delen av kursen.



STORA TALENS LAG (FRÅN FLS. 6).

- ▶ Stora talens lag är ett av de viktigaste resultaten inom sannolikhets teorin.
- ▶ Denna lag, först formulerad av den schweiziske matematikern Jacob Bernoulli (1654–1705), utsäger att *aritmetiska medelvärdet av flera oberoende s.v. med samma väntevärde μ ligger nära μ* , bara antalet är tillräckligt stort.





FIGUR: Jacob Bernoulli grundlade sannolikhetsläran med den postumt utgivna *Ars Conjectandi* (*Konsten att gissa* (1713)) där de stora talens lag presenteras för första gången.

NORMALFÖRDELNING

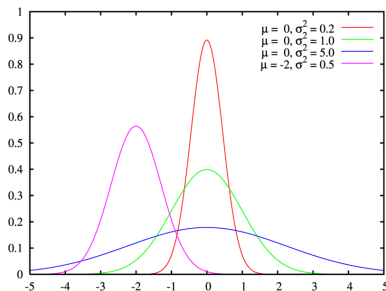
- ▶ Normalfördelningen, den viktigaste av alla fördelningar, är även kallad *Gauss-fördelningen* och *klockkurvan*. Benämningen Gauss-fördelning hänsyftar på den tyske matematikern Carl Friedrich Gauss (1777-1855).
- ▶ **Def:** En kontinuerlig s.v. X sägs vara *normalfördelad* med parametrar μ och σ , ($\sigma > 0$) om täthetsfunktionen ges av

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

- ▶ Beteckning: $X \in N(\mu, \sigma)$.
- ▶ Anm. Det finns en normalfördelning för varje par μ och $\sigma > 0$. Å andra sidan, givet μ och $\sigma > 0$ så är fördelningen helt preciserad och har täthet som ovan.



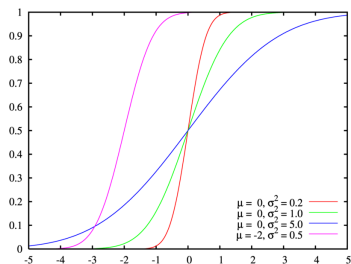
NORMALFÖRDELNING: TÄTHETSFUNKTION.



FIGUR: Täthetsfunktionen för en $N(\mu, \sigma)$ -fördelning för några olika värden på μ och σ .

- ▶ Effekten av att ändra μ : täthetensläge försjuts.
- ▶ Effekten av att ändra σ : fördelning blir mer koncentrerad när σ är liten, respektive mer utspridd när σ är stor.

NORMALFÖRDELNING: FÖRDELNINGSFUNKTION.



FIGUR: Fördelningsfunktionen för en $N(\mu, \sigma)$ -fördelning för några olika värden på μ och σ .

- ▶ Fördelningsfunktionen $F_X(x)$ ges av

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

- ▶ Denna integral har inget slutet uttryck. För givna μ , σ och x kan den beräknas numeriskt.



KTH Network

STANDARDISERAD NORMALFÖRDELNING.

- ▶ **Def:** En normalfördelad s.v. Z med parametrar $\mu = 0$ och $\sigma = 1$ sägs vara *standardiserad normalfördelad*, $Z \in N(0, 1)$.
- ▶ Täthetsfunktionen $f_Z(z)$ och fördelningsfunktionen $F_Z(z)$ för Z har egna beteckningar, $\phi(z)$ respektive $\Phi(z)$ och definieras av

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty,$$

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

- ▶ Vi ska se senare att det räcker att kunna räkna ut dessa funktioner för att kunna beräkna $F_X(x)$ för en godtycklig normalfördelning.



- ▶ $\phi(-z) = \phi(z)$, $\Phi(-z) = 1 - \Phi(z)$
- ▶ Areatolkning och numeriska exempel på tavlan.
- ▶ För $Z \in N(0, 1)$ gäller att

$$P(a < Z \leq b) = \Phi(b) - \Phi(a).$$

Eftersom fördelningen är kontinuerlig kan man byta \leq mot $<$ eller omvänt utan att sannolikheten ändras.

- ▶ Kvantiler för $N(0, 1)$ förekommer så ofta att dessa gets en egen beteckning, λ_α .
- ▶ **Def:** α -kvantilen, λ_α , för en standardiserad normalfördelning definieras som lösningen till $P(Z > \lambda_\alpha) = \alpha$. Men $P(Z > \lambda_\alpha) = 1 - \Phi(\lambda_\alpha)$ så λ_α löser tydligen

$$\Phi(\lambda_\alpha) = 1 - \alpha.$$



- ▶ Om $Z \in N(0, 1)$ så gäller att

$$E(Z) = 0, \quad D(Z) = 1.$$

Bevis: Eftersom $\phi(\cdot)$ är symmetrisk kring 0 så får man

$$E(Z) = \int_{-\infty}^{\infty} z\phi(z)dz = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.$$

- ▶ För att få $D(Z)$ använder vi Stas 5.6: $V(Z) = E(Z^2) - (E(Z))^2$.
Genom partiell integration fås

$$\begin{aligned} E(Z^2) &= \int_{-\infty}^{\infty} z^2\phi(z)dz = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= -z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0 + 1 = 1. \end{aligned}$$



- ▶ **Sats:** $X \in N(\mu, \sigma)$ om och endast om $Z = \frac{X-\mu}{\sigma} \in N(0, 1)$.
- ▶ Tolkning av μ och σ . Enligt sats får vi

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \mu,$$

$$V(X) = V(\sigma Z + \mu) = \sigma^2 E(Z) = \sigma^2,$$

dvs är parametrarna μ och σ *väntevärde* respektive *standardavvikelse* för $N(\mu, \sigma)$ -fördelad s.v.

- ▶ Vidare gäller att

$$f_X(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right), \quad F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

$$P(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

- ▶ Kort ex på tavlan!



- ▶ Användbarhet av resultat ovan.
- ▶ Exempel: Vid en industri produceras järnbalkar som väger 2000 kg. Vikten för en enskild balk kan beskrivas av en normalfördelning med $\mu = 2000$ kg och $\sigma = 2.3$ kg.
 - A) De balkarna som avviker med mer än 5 kg från avsedd vikt efterbehandlas för att få en vikt närmare 2000 kg. Bestäm andelen balkar som behöver efterbehandlas.
(Svar: 0.030)
 - B) Bestäm sannolikheten att en balk väger mindre än 1997 kg.
(Svar: 0.0968)
 - C) Bestäm sannolikheten att en balk väger mer än 2007 kg.
(Svar: 0.0012)

- ▶ I statistiska sammanhang vill man ofta ange symmetriska intervall så att $Z \in N(0, 1)$ ligger inom intervallet med en förutbestämd sannolikhet. Exempel: vilket tal $z > 0$ som gör att

$$P(-z \leq Z \leq z) = 0.95?$$

- ▶ Lösning till detta kan uttryckas med hjälp av kvantiler: z måste ges av $\lambda_{0.025} = 1.96$ och $-z$ blir $-\lambda_{0.025} = -1.96$. Bild med $\phi(z)$ på tavlan.
- ▶ **Allmänna förmeln:** för $Z \in N(0, 1)$ gäller att

$$P(-\lambda_{\alpha/2} \leq Z \leq \lambda_{\alpha/2}) = 1 - \alpha.$$



ALLMÄN NORMALFÖRDELNING (FORTS.)

- ▶ För en godtycklig $X \in N(\mu, \sigma)$ kan man uttrycka α -kvantilen x_α med hjälp av motsvarande kvantil λ_α för den standardiserade $Z \in N(0, 1)$: $P(Z > \lambda_\alpha) = \alpha$.
- ▶ Vi söker x_α som satisfierar $P(X > x_\alpha) = \alpha$.

$$P(X > x_\alpha) = P\left(\underbrace{\frac{X - \mu}{\sigma}}_{Z \in N(0,1)} > \underbrace{\frac{x_\alpha - \mu}{\sigma}}_{\lambda_\alpha}\right) = P(Z > \lambda_\alpha) = \alpha.$$

- ▶ Detta ger $x_\alpha = \mu + \sigma\lambda_\alpha$.
- ▶ Dessutom

$$P\left(-\lambda_{\alpha/2} \leq \frac{X - \mu}{\sigma} \leq \lambda_{\alpha/2}\right) = 1 - \alpha,$$

eller

$$P(\mu - \sigma\lambda_{\alpha/2} \leq X \leq \mu + \sigma\lambda_{\alpha/2}) = 1 - \alpha.$$



LINJÄR TRANSFORMATION AV NORMALFÖRDELNING

- ▶ En viktig egenskap hos normalfördelningen är att den *bevaras under linjära transformationer*.
- ▶ **Sats:** Om $X \in N(\mu, \sigma)$, så gäller att

$$Y = aX + b \in N(a\mu + b, |a|\sigma).$$

- ▶ **Sats:** Om X_1, X_2, \dots, X_n är oberoende och respektive $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2), \dots, N(\mu_n, \sigma_n)$ och konstanterna a_1, a_2, \dots, a_n, b är givna, så gäller att

$$\sum_{i=1}^n a_i X_i + b \in N\left(\sum_{i=1}^n a_i \mu_i + b, \sqrt{\sum_{i=1}^n a_i^2 \sigma_i^2}\right).$$

- ▶ Speciellt, om X_1, X_2, \dots, X_n är oberoende $N(\mu, \sigma)$ och $a_1 = a_2 = \dots = a_n = 1/n$ samt $b = 0$, så gäller att

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \in N(\mu, \sigma/\sqrt{n}).$$



CENTRALA GRÄNSVÄRDESSATSEN (CGS).

- ▶ Satsen är den viktigaste resultaten inom sannolikhetsteorin:

En summa av oberoende lika fördelade s.v. med godtycklig fördelning är ungefär normalfördelad, bara antalet komponenter i summan är tillräckligt stort.

- ▶ **Sats (CGS):** Låt X_1, \dots, X_n, \dots vara en oändlig följd av oberoende, likafördelade s.v. med väntevärde μ och standardavvikelse $0 < \sigma < \infty$. Sätt

$$Y_n = X_1 + \dots + X_n.$$

Då gäller för givna $a < b$ att

$$P\left(a < \frac{Y_n - n\mu}{\sqrt{n}\sigma} \leq b\right) \rightarrow \Phi(b) - \Phi(a) \quad \text{då } n \rightarrow \infty.$$

- ▶ CGS uttalar sig alltså om *fördelningen av Y_n då antalet n växer mot oändligheten*: Y_n är ungefär $N(n\mu, \sqrt{n}\sigma)$ -fördelad. Beteckning:

$$Y_n \in AsN(n\mu, \sqrt{n}\sigma).$$



CENTRALA GRÄNSVÄTDESSATSEN (FORTS.)

- ▶ Observera att $E(Y_n) = n\mu$ och $D(Y_n) = \sqrt{n}\sigma$. För varje givet n är

$$\frac{Y_n - n\mu}{\sqrt{n}\sigma}$$

en standardiserad s.v. Den har väntevärde lika med noll och standardavvikelse lika med 1 som en standardiserad normalfördelad s.v.

- ▶ Enligt CGS: när n går mot oändligheten kommer hela fördelningen för den angivna standardiserade s.v. att gå mot en *standardiserad normalfördelning*, dvs

$$\frac{Y_n - n\mu}{\sqrt{n}\sigma} \in \text{AsN}(0, 1).$$



- **Följdats:** För en oändlig följd av oberoende likafördelade s.v. X_1, \dots, X_n, \dots med $E(X_i) = \mu$ och $D(X_i) = \sigma$ ($0 < \sigma < \infty$) gäller att

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \in AsN \left(\mu, \frac{\sigma}{\sqrt{n}} \right) \quad \text{då } n \rightarrow \infty,$$

dvs aritmetisk medelvärde \bar{X}_n är *approximativt normalfördelat* för tillräckligt stort n .

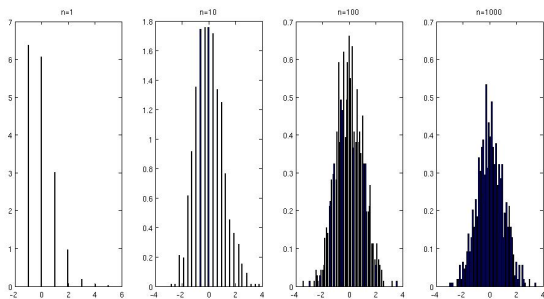
- Normalfördelningsapproximation. Enligt CGS: $\sum_{i=1}^n X_i \in AsN(n\mu, \sqrt{n}\sigma)$ och $\bar{X}_n \in AsN(\mu, \sigma/\sqrt{n})$. Detta ger approximationerna

$$P \left(a < \sum_{i=1}^n X_i \leq b \right) \approx \Phi \left(\frac{b - n\mu}{\sqrt{n}\sigma} \right) - \Phi \left(\frac{a - n\mu}{\sqrt{n}\sigma} \right),$$

$$P(c < \bar{X}_n \leq d) \approx \Phi \left(\frac{d - \mu}{\sigma/\sqrt{n}} \right) - \Phi \left(\frac{c - \mu}{\sigma/\sqrt{n}} \right).$$



CENTRALA GRÄNSVÄRDESSATSEN (FORTS.)



FIGUR: Fördelningen för $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ för $n = 1$, $n = 10$, $n = 100$ och $n = 1000$, där

$$\bar{X}_n = (X_1 + \dots + X_n) / n$$

och X_1, \dots, X_n, \dots är oberoende $Po(1)$ -variabler (så att $\mu = \sigma = 1$). Då $n \rightarrow \infty$ liknar sannolikhetsfunktionen alltmer den standardiserade normalfördelningstäthet.