SF1901 Sannolikhetsteori och statistik I

Johan Westerborn

Föreläsning 7
17 november 2016

Lite om kontrollskrivning och laborationer

- *Kontrollskrivningen*
  - omfattar Kap. 1–5 i boken, alltså Föreläsning 1–6. Hjälpmedel: endast miniräknare.
  - omfattar 5 uppgifter. För att erhålla bonus krävs helt korrekt svar till minst 3 av uppgifterna. Gamla kontrollskrivningar att öva på finns på hemsidan.
- Vad gäller *laborationerna* finns nu handledningen till Laboration 1 (bonusgivande, onsdag 25 november kl. 15–17) på hemsidan under ”aktuell information”.
- Då det är många studenter som följer kursen kommer laborationstillfället den 25 november att användas främst till *redovisning*.

Idag

Förra gången

Normalfördelningen och dess egenskaper (Kap. 6.3–6.4)

Linjärkombinationer av oberoende normalfördelade s.v. (Kap. 6.5)

Centrala gränsvärdessatsen (Kap. 6.7)

Denna kurs

- *Sannolikhetsteori:* Hur kan man ställa upp en matematisk modell för ett slumpmässigt försök? (lv. 1–3)
- *Inferensteori:* Vilka slutsatser kan man dra av ett givet datamaterial? (lv. 4–7)
- Sannolikhetsteorin utgör en grund för inferensteorin.

Idag

Förra gången

Normalfördelningen och dess egenskaper (Kap. 6.3–6.4)

Linjärkombinationer av oberoende normalfördelade s.v. (Kap. 6.5)

Centrala gränsvärdessatsen (Kap. 6.7)

Förra gången

- Väntevärden av funktion av flerdimensionell s.v.,
- kovarians och korrelation,
- väntevärde och varians för linjärkombinationer,
- stora talens lag.

Mer om kovarianser: bilinjäritet

- Man övertygar sig enkelt om att kovariansen är *bilinjär*, dvs. för godtyckliga linjärkombinationer $\sum_i a_i X_i$ och $\sum_j b_j Y_j$ av ev. beroende s.v. gäller att

$$\mathbb{C}\left(\sum_i a_i X_i, \sum_j b_j Y_j\right) = \sum_{i,j} a_i b_j \mathbb{C}(X_i, Y_j).$$

- Ofta praktiskt när man räknar!
- Speciellt gäller att (jfr. förra gången)

$$\mathbb{V}\left(\sum_i a_i X_i\right) = \mathbb{C}\left(\sum_i a_i X_i, \sum_j a_j X_j\right)$$
$$= \sum_{i,j} a_i a_j \mathbb{C}(X_i, X_j) = \sum_i a_i^2 \mathbb{V}(X_i) + 2\sum_{i<j} a_i a_j \mathbb{C}(X_i, X_j).$$

Exempel: summa och differens

- Låt X och Y vara s.v. sådana att $\mathbb{V}(X) = \mathbb{V}(Y)$. Beräkna $\mathbb{C}(X + Y, X - Y)$.

[svar: 0]

Idag

Förra gången

Normalfördelningen och dess egenskaper (Kap. 6.3–6.4)

Linjärkombinationer av oberoende normalfördelade s.v. (Kap. 6.5)

Centrala gränsvärdessatsen (Kap. 6.7)

Täthetsfunktion

Definition

Om en s.v. X har täthetsfunktionen

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R},$$

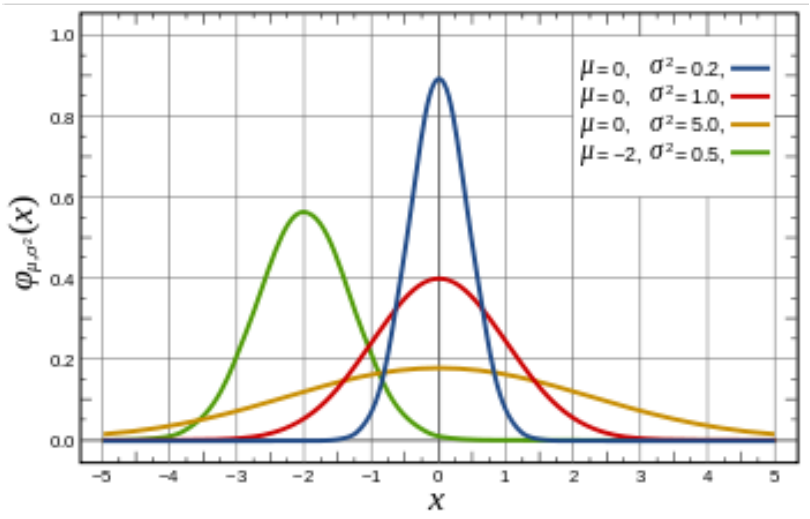
där μ och $\sigma > 0$ är givna tal, sägs X vara *normalfördelad* (kodbeteckning: $X \in \mathrm{N}(\mu, \sigma)$).

- Om $\mu = 0$ och $\sigma = 1$ sägs X vara *standardiserad normalfördelad*. Den standardiserade normalfördelningen är så viktig att man gett dess täthets- och fördelningsfunktion egna namn, φ resp. Φ:

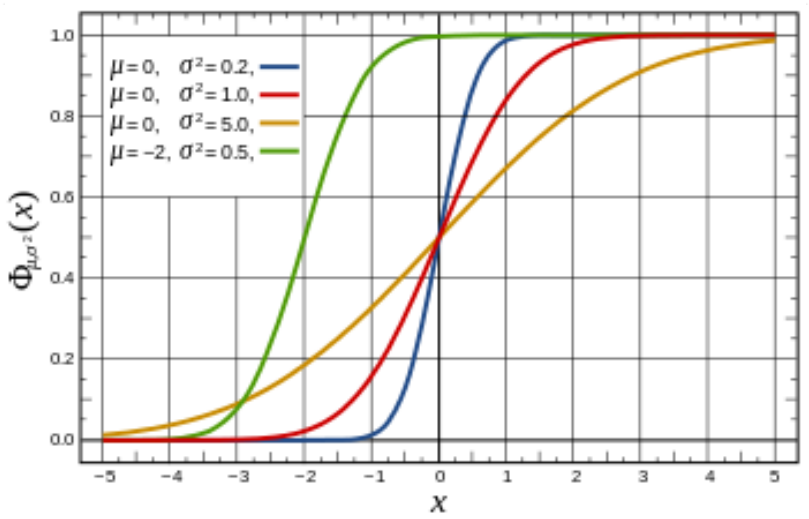
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{och} \quad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz.$$

Täthetsfunktion (forts.)

(a) Täthetsfunktioner



(b) Fördelningsfunktioner

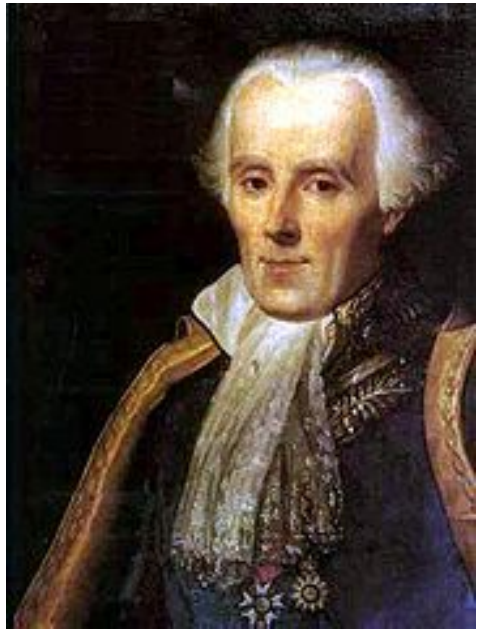


Figur: Täthets- och fördelningsfunktioner för N(μ, σ)-fördelningar med olika parametrar (μ, σ).

Gauß och Laplace

(a) C. F. Gauß (1777–1855)

(b) P. S. Laplace (1749–1827)



Exempel: lite tabellexercis

- Låt $X \in \mathrm{N}(0,1)$. Använd tabell för att beräkna
  (a) $\mathbb{P}(X \leq 0.75)$,
  (b) $\mathbb{P}(X \leq -0.75)$,
  (c) $x_{0.15}$, dvs. 15%-kvantilen för X.
- MATLAB-lösning:

```
>> normcdf([0.75 -0.75])

ans =

    0.7734    0.2266

>> norminv(1 - 0.15)

ans =

    1.0364
```

Väntevärde och varians för $\mathrm{N}(0,1)$

- Låt $X \in \mathrm{N}(0,1)$. Då φ är symmetrisk inses direkt att

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

- Vidare, då $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2)$ gäller att

$$\begin{aligned}\mathbb{V}(X) &= \int_{-\infty}^{\infty} x^2\varphi(x)\,dx = \int_{-\infty}^{\infty} x\left(x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right)\,dx \\ &= \underbrace{\left[-x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right]_{-\infty}^{\infty}}_{=0} + \underbrace{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx}_{=1} = 1.\end{aligned}$$

Allmän normalfördelning $\mathrm{N}(\mu, \sigma)$

Sats

$X \in N(\mu, \sigma)$ om och endast om $Y = (X - \mu)/\sigma \in N(0, 1)$.
Dessutom gäller att

$$f_X(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) \quad \textit{och} \quad F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}.$$

- Detta ger direkt att

$$\mathbb{E}(X) = E(\sigma Y + \mu) = \sigma E(Y) + \mu = \mu,$$
$$\mathbb{V}(X) = \mathbb{V}(\sigma Y + \mu) = \sigma^2 \mathbb{V}(Y) = \sigma^2,$$

dvs. parametrarna μ och σ är *väntevärde* resp. *standardavvikelse* för $\mathrm{N}(\mu, \sigma)$-fördelningen.

Exempel: lite mer tabellexercis

- Låt $X \in \mathrm{N}(6, 2)$ och beräkna $\mathbb{P}(7 < X < 9)$ med hjälp av tabell.
- MATLAB-lösning:

```
normcdf(9,6,2) - normcdf(7,6,2)

ans =

    0.2417
```

Normalfördelningens kvantiler

- Standardnormalfördelningens α-kvantil betecknas λ_α och kan hittas i tabell; om $X \in \mathrm{N}(0,1)$ gäller alltså att $\mathbb{P}(X > \lambda_\alpha) = \alpha$.
- Vidare, $\mathbb{P}(-\lambda_{\alpha/2} < X < \lambda_{\alpha/2}) = 1 - 2(\alpha/2) = 1 - \alpha$.
- Om $X \in \mathrm{N}(\mu, \sigma)$ är α-kvantilen $\mu + \lambda_\alpha \sigma$, ty

$$\alpha = \mathbb{P}\left(\frac{X-\mu}{\sigma} > \lambda_\alpha\right) = \mathbb{P}(X > \mu + \lambda_\alpha \sigma).$$

Dessutom,

$$1 - \alpha = \mathbb{P}\left(-\lambda_{\alpha/2} < \frac{X-\mu}{\sigma} < \lambda_{\alpha/2}\right)$$
$$= \mathbb{P}\left(\mu - \lambda_{\alpha/2}\sigma < X < \mu + \lambda_{\alpha/2}\sigma\right).$$

Idag

Förra gången

Normalfördelningen och dess egenskaper (Kap. 6.3–6.4)

Linjärkombinationer av oberoende normalfördelade s.v. (Kap. 6.5)

Centrala gränsvärdessatsen (Kap. 6.7)

Linjärkombinationer av oberoende normalfördelade s.v.

- Linjärkombinationer av normalfördelade s.v. är fortfarande normalfördelade!

Sats

Om $X_1 \in N(\mu_1, \sigma_1)$, $X_2 \in N(\mu_2, \sigma_2)$, $\ldots$, $X_n \in N(\mu_n, \sigma_n)$ *är oberoende gäller för alla konstanter* a_1, a_2, $\ldots$, a_n *och* b *att*

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

- Om X_1, X_2, $\ldots$, X_n är oberoende och $\mathrm{N}(\mu, \sigma)$-fördelade erhålls speciellt, för $a_1 = a_2 = \ldots = a_n = 1/n$,

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

Exempel: bomullsband

- En maskin klipper till bomullsband i bitar, vilkas längd (enhet: meter) uppvisar en slumpmässig variation som är $N(1, 0.05)$. Vid ett tillfälle vill man ha 10 bitar med en sammanlagd längd på 10 meter. Då gör man på ett av nedanstående två sätt:
  - (I) Tag ett band slumpmässigt med längden X_1 och klipp till ytterligare 9 precis lika långa bitar. Sammanlagda längden är då Y.
  - (II) Tag 10 bitar slumpmässigt med längder $X_1, X_2, \ldots, X_{10}$. Deras sammanlagda längd är då Z.

  Med vilken metod är sannolikheten störst att den sammanlagda längden ligger nära 10?

[svar: metod (II)]

Idag

Förra gången

Normalfördelningen och dess egenskaper (Kap. 6.3–6.4)

Linjärkombinationer av oberoende normalfördelade s.v. (Kap. 6.5)

Centrala gränsvärdessatsen (Kap. 6.7)

Centrala gränsvärdessatsen

- Följande är det viktigaste resultatet i matematisk statistik.

Sats (centrala gränsvärdessatsen)

Låt $X_1, X_2, \ldots$ vara en oändlig följd av likafördelade s.v. med väntevärde μ och standardavvikelse $\sigma > 0$ och sätt $Y_n = X_1 + \cdots + X_n$. Då gäller för alla $a < b$ att

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

- Med andra ord, när n är stort är Y_n ungefär $\mathrm{N}(n\mu, \sigma\sqrt{n})$-fördelad. Vi skriver detta $Y_n \in \mathrm{AsN}(n\mu, \sigma\sqrt{n})$.

Centrala gränsvärdessatsen (forts.)

- Notera att

$$Y_n \in \text{AsN}(n\mu, \sigma\sqrt{n}) \Rightarrow \bar{X} = \frac{Y_n}{n} \in \text{AsN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right),$$

  vilket är i linje med stora talens lag. Men centrala gränsvärdessatsen beskriver även *felet* mellan $\bar{X}$ och μ.
- Det märkliga med CGS:en är att den håller *oberoende av fördelningen hos X_i:na* sålänge väntevärde och varians är väldefinierade!
- Hur stort n som krävs för bra approximation beror på fördelningen hos X_i:na och i synnerhet på hur skev denna är.
- Många *mätfel* uppstår som summor av delfel av ungefär samma storlek och blir alltså approximativt normalfördelade.

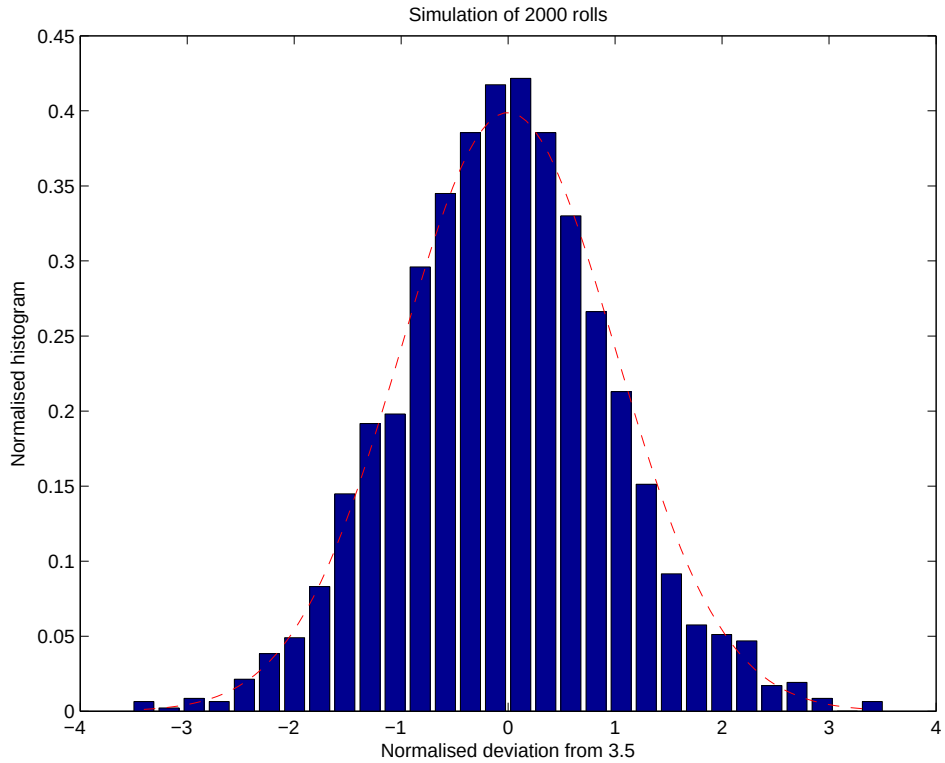
Exempel: upprepade kast med tärning

- För utfallet X av ett tärningskast gäller att $\mu = \mathbb{E}(X) = 3.5$ och $\sigma = \mathbb{D}(X) = 1.71$.
- Vi kastar en tärning upprepade gånger och betecknar utfallen $X_1, X_2, X_3, \ldots$
- Efter varje nytt kast skriver vi upp summan $Y_n = \sum_{i=1}^{n} X_i$.
- MATLAB-simulering:

```
X = randi(6,1,n);
S = cumsum(X);
means = S./(1:n);
```

- Slutligen beräknar vi den normerade summan $(Y_n - n\mu)/(\sigma\sqrt{n})$.

Exempel: upprepade kast med tärning (forts.)



Figur: Normerat histogram över 2000 simulerade normerade summor. Rödstreckad linje är täthetsfunktionen för $N(0, 1)$.

Centrala gränsvärdessatsen, bevisskiss*

- Idé: visa att täthetsfunktionen f_n för $(Y_n - n\mu)/(\sigma\sqrt{n})$ är lika med φ för stora n genom att visa att *Fouriertransformerna* av dessa funktioner är lika.
- Man visar tämligen enkelt (övning!) att Fouriertransformen av φ är $e^{-t^2/2}$.
- För varje täthet f_X med väntevärde 0 och varians 1 ges Fouriertransformen allmänt av

$$\begin{aligned}\mathcal{F}(t) &= \int_{-\infty}^{\infty} e^{itx} f_X(x)\,dx = \mathbb{E}\left(e^{itX}\right) \\ &\overset{\text{Taylor}}{=} 1 + \mathbb{E}(X)it - \frac{\mathbb{V}(X)}{2}t^2 + o(t^2) \\ &= 1 - \frac{1}{2}t^2 + o(t^2).\end{aligned}$$

*Överkurs för den intresserade studenten.

Centrala gränsvärdessatsen, bevisskiss (forts)*

- Trick: skriv

$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\frac{X_i - \mu}{\sigma}}_{\overset{\text{not.}}{=} \tilde{X}_i}.$$

- Nu erhålls, då $\mathbb{E}(\tilde{X}_i) = 0$ och $\mathbb{V}(\tilde{X}_i) = 1$.

$$\varphi_n(t) = \mathbb{E}\left(e^{itZ_n}\right) = \mathbb{E}\left(e^{i\frac{t}{\sqrt{n}}\sum_{i=1}^{n}\tilde{X}_i}\right) \overset{\text{ober.}}{=} \prod_{i=1}^{n} \mathbb{E}\left(e^{i\frac{t}{\sqrt{n}}\tilde{X}_i}\right)$$
$$= \left(1 - \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right)^n \xrightarrow{n\to\infty} e^{-t^2/2} = \varphi(t).$$

*Överkurs för den intresserade studenten.