Statistik 1 för biologer, logopeder och psykologer

Föreläsningar, del 3

Innehåll

1. Grunderna i sannolikhetslära
   - Grundbegrepp
   - Egenskaper och räkneregler
2. Fördelningar
   - Binomialfördelningen
   - Normalfördelningen

Innehåll

1. Grunderna i sannolikhetslära
   - Grundbegrepp
   - Egenskaper och räkneregler
2. Fördelningar
   - Binomialfördelningen
   - Normalfördelningen

Satistik och sannolikhetslära

- Statistik handlar om att utvinna information från data.
- I praktiken inhehåller de data man analyserar oftast någon form av variabilitet eller osäkerhet.
- Denna variabilitet strävar man efter att kvantifiera med hjälp av sannolikhetslära.
- Sannolikhetslära är således nödvändigt för att förstå statistiska analysmetoder.

Slumpmässiga försök

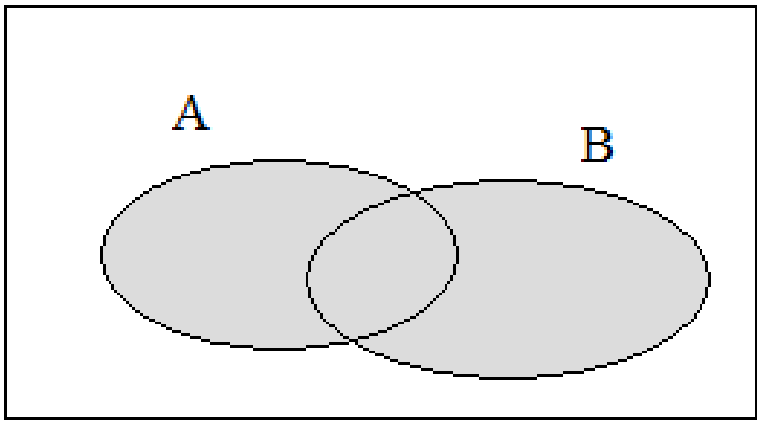
- Ett Slumpmässigt försök är en företeelse som kan upprepas under likartade förhållanden.
- Resultatet kan inte anges i förväg även om man många gånger tidigare utfört samma försök.
- T.ex. att kasta tärning kan ses som ett slumpmässigt försök.

Utfall och händelser

- Ett utfall är resultatet av ett slumpmässigt försök.
- Alla tänkbara utfall tillsammans kallas utfallsrummet.
  - T.ex. i tärningskast är utfallsrummet $\{1, 2, 3, 4, 5, 6\}$.
- En händelse är en samling ufall.
  - T.ex. "vi får ett udda tal" $= \{1, 3, 5\}$
  - Händelser betecknas ofta med stora bokstäver $A, B, C, \ldots$
  - Ibland kan det vara nyttigt att visualisera kombinationer av händelser med hjälp av sk. Venn diagram.

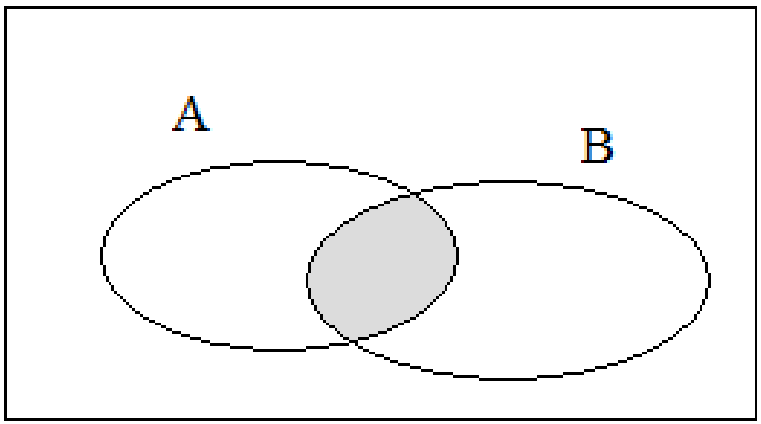
Kombinationer av händelser

$A \cup B =$ ”A eller B”



Kombinationer av händelser

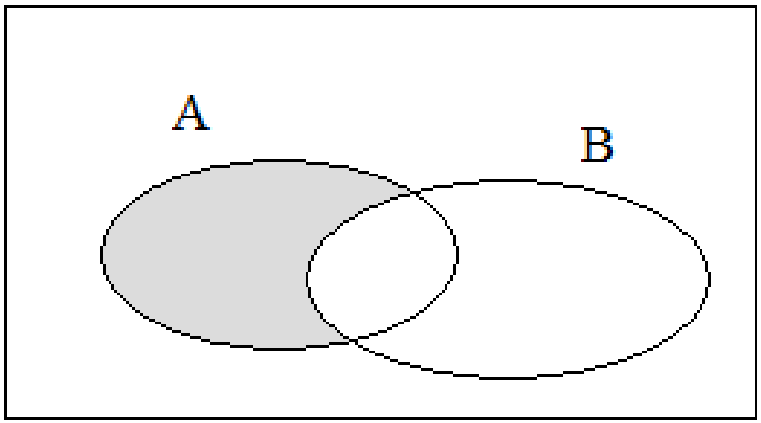
$A \cap B =$ ”A och B”



- Om A och B inte alls skär varandra säger vi att händelserna är *oförenliga*.

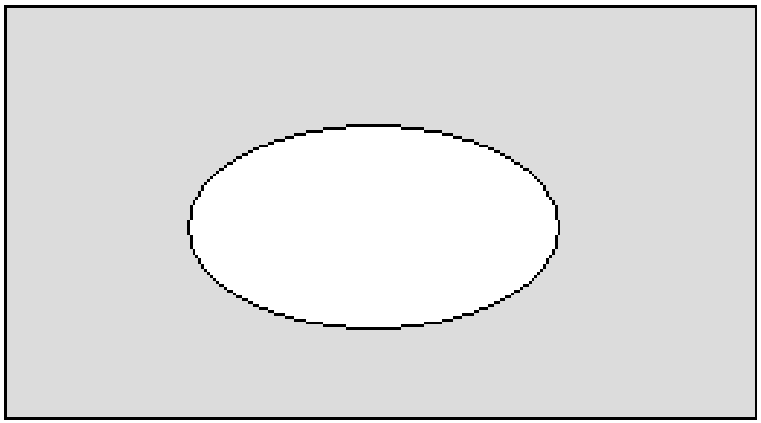
Kombinationer av händelser

$A \backslash B =$ ”A men inte B”



Kombinationer av händelser

$A^c =$ ”inte A”



Sannolikhet

- Sannolikhet är ett tal mellan 0 och 1 som förknippas med en händelse eller kombination av händelser.
  - 0 betyder att händelsen är *omöjlig*.
  - 1 betyder att händelsen är *säker*.
- Sannolikheten för händelsen A betecknas $P(A)$.
- Sannolikheten för händelsen A eller B betecknas $P(A \cup B)$.
- Alternativt kan man explicit skriva ut $P(A$ inträffar), $P(A$ eller $B)\ldots$

Tolkning av sannolikhet

Sannolikhet kan tolkas på flera olika sätt:

- Enligt den klassiska tolkningen definieras sannolikhet som
  $P(A) = \frac{\text{"antalet för } A \text{ gynnsamma utfall"}}{\text{"totala antalet utfall"}}$.
- Den empiriska (eller *frekventistiska*) tolkningen av sannolikhet är
  $P(A) =$ den relativa frekvensen för A i ett stort antal försök.
- Det finns även en subjektiv tolkning där
  $P(A) =$ subjektiv uppfattning om hur trolig händelsen A är.

Komplement.

- Om händelsen E består av hela utfallsrummet får vi $P(E) = 1$.
- Sannolikheten för komplementhändelsen till A, dvs. händelsen $A^c =$ ”inte A” är $P(A^c) = 1 - P(A)$.
- Sannolikheten för komplementet till utfallsrummet är $P(E^c) = 1 - P(E) = 0$.
  - Utfallsrummets komplement E^c (den omöjliga händelsen) betecknas ofta $\emptyset$.

Additionssatsen

- Sannolikheten för att händelsen A eller B inträffar är

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

- Ordet ”eller” förknippas i sannolikhetskalkyl med *addition*.
- För att inte räkna de gemensama utfallen för A och B två gånger subtraherar vi $P(A \cap B)$ från summan av sannolikheterna $P(A)$ och $P(B)$.

Additionssatsen

Exempel.

Av ett tillverkat parti enheter har 2% fel vikt, 4% fel färg och 1% både fel vikt och färg. Vi räknar sannolikheten att en slumpmässigt vald enhet har antingen fel vikt eller fel färg (eller båda)?

Låt $A =$ "enheten har fel vikt" och $B =$ "enheten har fel färg". De angivna sannolikheterna är då $P(A) = 0.02$, $P(B) = 0.04$ och $P(A \cap B) = 0.01$. Från additionssatsen får vi

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.02 + 0.04 - 0.01 = 0.05\,.$$

Oförenliga händelser

- Om händelserna A och B är oförenliga utesluter de varandra (endast en åt gången av dem kan inträffa).
- Sannolikheten för "A och B" blir då $P(A \cap B) = 0$.
- Additionssatsen för oförenliga händelser förenklas till

$$P(A \text{ eller } B) = P(A \cup B) = P(A) + P(B).$$

- T.ex. sannolikheten att få resultatet 2 eller 3 vid *ett* tärningskast är

$$P(\text{vi får 2}) + P(\text{vi får 3}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

eftersom resultaten utesluter varandra.

Betingad sannolikhet

- Den betingade sannolikheten för A givet B betyder att vi begränsar oss till att endast betrakta sådana utfall som tillhör händelsen B.
- Vi kan också tänka oss den betingade sannolikheten för A givet B som sannolikheten för A då vi vet att B har inträffat.
- Sannolikheten för A påverkas av vår kunskap om händelsen B.

Betingad sannolikhet

- Den betingade sannolikheten för A givet B definieras som

$$P(A|\,B) = \frac{P(A \cap B)}{P(B)}\,,$$

där vi antar att $P(B) > 0$.

- Den betingade sannolikheten kan även tolkas som:

$$P(A|\,B) = \frac{\text{"antalet för } A \text{ gynnsamma utfall i } B\text{"}}{\text{"totala antalet utfall i } B\text{"}}\,.$$

Betingad sannolikhet

Exempel.

Vi drar slumpmässigt ett kort ur en kortlek på 52 kort. Vi frågar oss först vad sannoikheten är att det dragna kortet är en kung. Vi betecknar $A =$ "kung". Eftersom det finns sammanlagt 4 kungar i en kortlek blir sannolikheten då

$$P(A) = 4/52 \approx 0.08 \, .$$

Låt oss vidare anta att vi vet att det dragna kortet är ett bildkort. Vi betecknar $B =$ "bildkort". Vad sannolikheten nu att det dragna kortet är en kung? Eftersom det finns sammanlagt 12 bildkort i en kortlek av vilka 4 är kungar blir sannolikheten

$$P(A|B) = 4/12 \approx 0.33 \, .$$

Sannolikheten ovan är den *betingade sannolikheten* för att få en kung givet att vi har dragit ett bildkort.

Multiplikationssatsen

- Sannolikheten för att händelserna A och B inträffar är

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A).$$

- Ordet ”och” förknippas i sannolikhetskalkyl med *multiplikation*.
- Betydelsen av den betingade sannolikheten i formeln blir tydligare om vi tänker oss att A och B inträffar i följd: först inträffar A och då vi vet att A har inträffat inträffar B.

Exempel.

Vi har en skål med 3 vita bollar och 5 röda bollar, dvs totalt 8 bollar. Vi räknar sannolikheten för att då vi slumpmässigt drar två bollar ur skålen är båda röda. Vi har då händelserna

$$A = \text{"vi drar en röd boll"}$$
$$B|A = \text{"vi drar en röd givet att vi redan dragit en röd boll"}$$

som ger sannolikheten

$$P(A)P(B|A) = \frac{5}{8} \cdot \frac{4}{7} = \frac{20}{56} \approx 0.36 \ .$$

Oberoende händelser

- Två händelser A och B är oberoende (betecknas ofta $A \perp\!\!\!\perp B$) om händelserna inte på något sätt påverkar varandra.
- För de oberoende händelserna A och B gäller att

$$P(A \cap B) = P(A)P(B)\,,$$

jfr. den allmänna multiplikationssatsen!

- Sannolikheten för B påverkas inte av vår kunskap om händelsen A:

$$P(B|A) = P(B)$$

dvs. sannolikheten för B förblir densamma oberoende om vi betingar med A eller inte.

Exempel.

Vi har händelserna

$$A = \text{"vi får krona då vi singlar slant"}$$
$$B = \text{"vi får resultatet 4 i tärningskast"}.$$

Det är uppenbart att händelserna är oberoende och sannolikheten för "A och B" blir således

$$P(A \cap B) = P(A)P(B) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12} \approx 0.08 \ .$$

Total sannolikhet och Bayes sats.

- Den totala sannolikheten anger sannolikheten för en händelse som kan inträffa på ett antal alternativa sätt.
- Med hjälp av Bayes sats räknar man ut sannolikheterna för de enskilda alternativa sätten då vi vet att händelsen har inträffat.
- Vi belyser begreppen genom ett exempel:
  `http://web.abo.fi/fak/mnf/mate/kurser/statistik1/TotSann&Bayes.pdf`

Innehåll

1 Grunderna i sannolikhetslära
- Grundbegrepp
- Egenskaper och räkneregler

2 Fördelningar
- Binomialfördelningen
- Normalfördelningen

Slumpvariabler

- En variabel som för varje utfall av ett slumpmässigt försök antar ett reelt tal kallas slumpvariabel.
- Slumpvariabler betecknas ofta med stora bokstäver från slutet av alfabetet: $X, Y, Z, \ldots$

Exempel.

Vi singlar två slantar och observerar antalet kronor. I stället för att definiera händelserna $A =$ "vi får 0 kronor", $B =$ "vi får 1 krona", $C =$ "vi får 2 kronor" kan vi definiera en slumpvariabel X som kan anta värdena 0, 1, 2.

- En *diskret* slumpvariabel kan anta ett uppräkneligt antal distinkta värden, medan en *kontinuerlig* slumpvariabel kan anta ett oändligt antal värden i ett givet intervall.

Sannolikhetsfördelning

- Sannolikhetsfördelningen för en slumpvariabel anger med vilken sannolikhet variabeln antar olika värden.

Exempel.

Låt den diskreta slumpvariablen X beteckna antalet kronor då vi singlar två slantar. Vi får då följande sannolikhetsfördelning för variablen:

$$P(X = 0) = 0.25$$
$$P(X = 1) = 0.5$$
$$P(X = 2) = 0.25\ .$$

Väntevärde och varians

- Vi har tidigare sett att empiriska fördelningar av datamaterial kan beskrivas med hjälp av central- och spridningsmått.
- Motsvarande mått används även för att beskriva sannolikhetsfördelningar för slumpavriabler.
- Det genomsnittliga värdet av slumpvariabeln X kallas väntevärde och betecknas $E(X)$ eller μ.
  - Väntevärdet motsvarar medelvärdet av en slumpvariabel då vi upprepar ett slumpmässigt försök oändligt många gånger.
- Variansen, som betecknas $Var(X)$ eller σ^2, är ett mått på hur utspridd fördelningen är kring väntevärdet.
  - Variansen definieras som $Var(X) = E[(X - \mu)^2]$.

Väntevärde

Väntevärdet för en diskret slumpvariabel fås genom att räkna en viktad summa av variabelns alla värden, där vikterna utgörs av sannolikheterna för värdena.

Exempel.

Väntevärdet för slumpvariabeln X som betecknar antalet kronor då vi singlar två slantar är

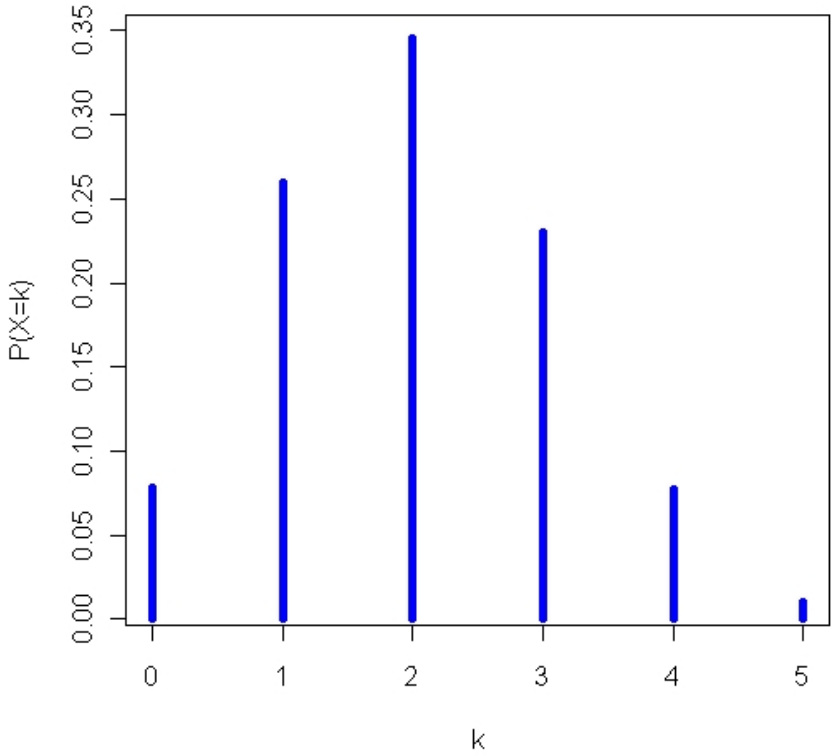
$$\begin{aligned} E(X) &= P(X=0)\cdot 0 + P(X=1)\cdot 1 + P(X=2)\cdot 2 \\ &= 0.25\cdot 0 + 0.5\cdot 1 + 0.25\cdot 2 = 1\ . \end{aligned}$$

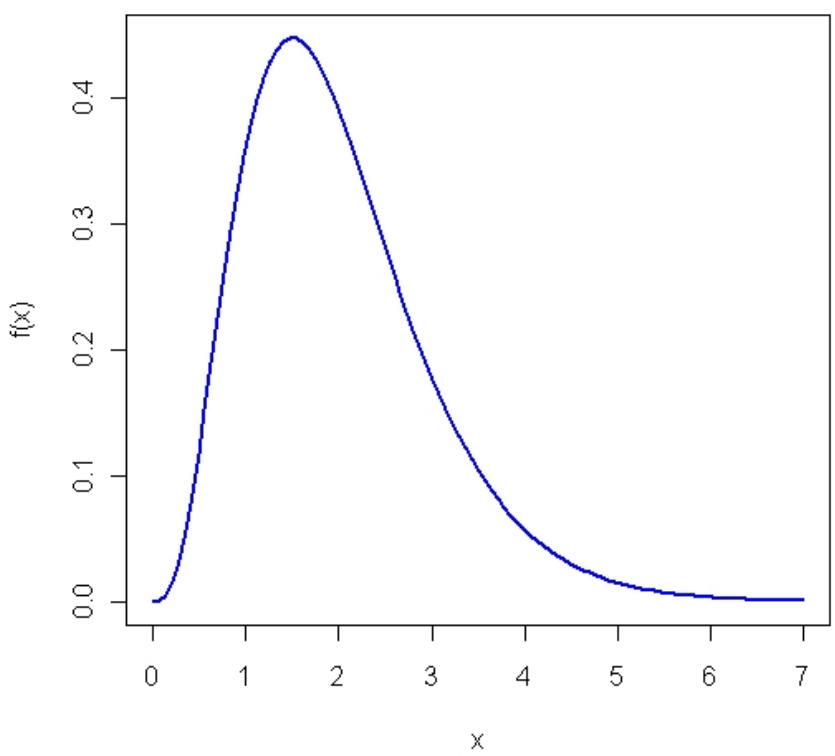
Om vi m.a.o. skulle upprepa försöket oändligt många gånger skulle medelvärdet av antalet kronor per försök bli 1.

Sannolikhets- och täthetsfunktion

- I stället för att räkna upp sannolikheter för olika värden av en slumpvariabel är det ofta mera praktiskt att beskriva en fördelning i fom av en *funktion* av slumpvariabeln.
- En sannolikhetsfunktion anger sannolikheten för ett givet värde av en *diskret* slumpvariabel.
- För en *kontinuerlig* slumpvariabel är sannolikheten för enskilda värden 0, varför man i stället använder en sk. täthetsfunktion som konstrueras så att
  - mer sannolika värden får högre täthet (inte samma som sannolikhet!)
  - ytan mellan täthetsfunktionen och x-axeln blir 1.

Sannolikhets- och täthetsfunktion





Fördelningsfunktion

- En fördelningsfunktion anger för både diskreta och kontinuerliga slumpvariabler sannolikheten att få ett värde som är mindre än eller lika med ett visst värde x

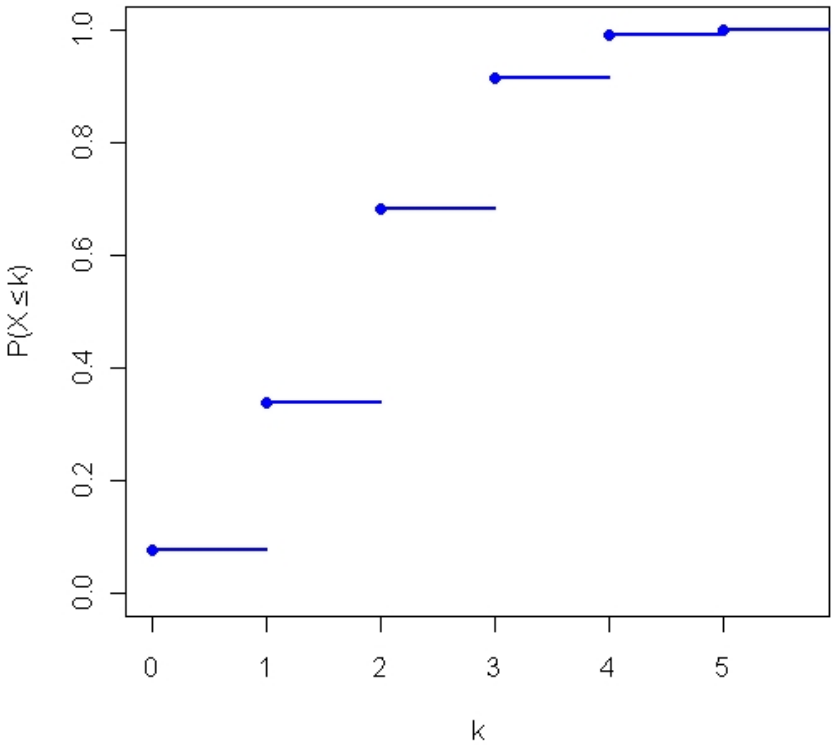
$$F(x) = P(X \leq x)\,.$$

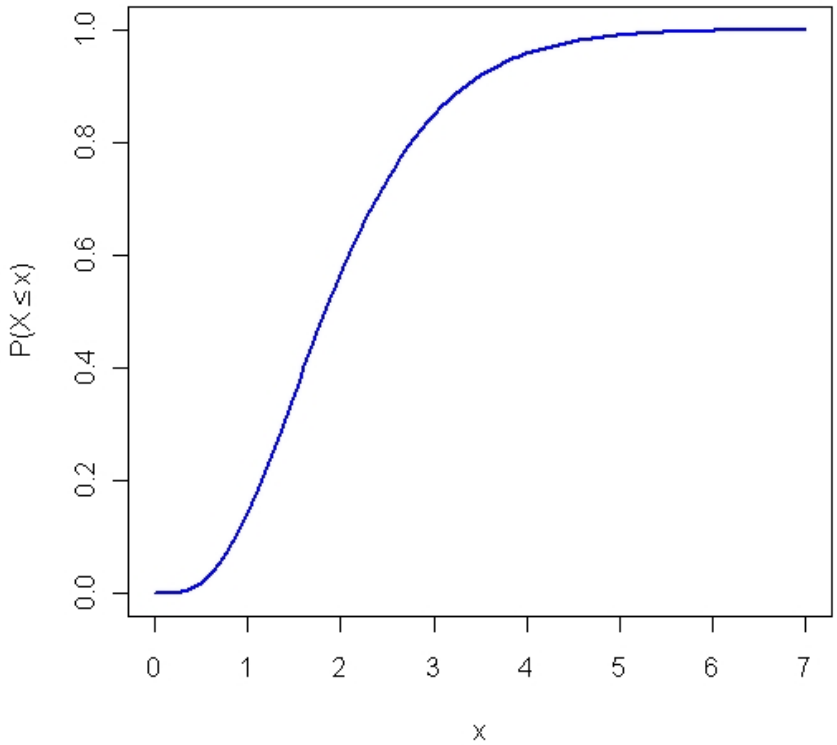
- Sannolikheten för att en slumpvariabel ska anta värden större än a och mindre eller lika med b kan beräknas med:

$$P(a < X \leq b) = F(b) - F(a)\,.$$

Fördelningsfunktion

Diskret och kontinuerlig fördelningsfunktion.





Fördelningsfunktion

- Med *diskreta* slumpvariabler bör man se upp med att få gränserna rätt då man gör uträkningar med fördelningsfunktionen.

Exempel.

Låt oss anta att X kan få värdena 0, 1, 2, 3. Vi får då

$$P(1 < X < 3) = F(2) - F(1) = P(X = 2)$$
$$P(1 \leq X < 3) = F(2) - F(0)$$
$$P(1 < X \leq 3) = F(3) - F(1)$$
$$P(1 \leq X \leq 3) = F(3) - F(0)\,.$$

- Då det gäller *kontinuerliga* variabler gör vi ingen skillnad mellan "$<$" och "$\leq$".
- Om alltså X i exemplet ovan hade varit kontinuerlig skulle samtliga sannolikheter ha räknats $F(3) - F(1)$.

Bernoulliförsök

- Vi utför ett försök som kan resultera i endast två olika utfall A och A^c.
- Försök med endast två möjliga utfall som utgör varandras komplement kallas ofta Bernoulliförsök.
- Vi betecknar framöver $P(A) = p$ och följaktligen $P(A^c) = 1 - P(A) = 1 - p$.

Binomialfördelningen

Binomialfördelningen kan karakteriseras på följande sätt:

- Ett Bernoulliförsök upprepas så att
  - antalet upprepningar är n
  - sannolikheten $P(A) = p$ hålls konstant över alla upprepningar
  - försöken upprepas oberoende från varandra.
- Vi definierar en slumpvariabel X som anger antalet försök som resulterar i A.
- Slumpvariabeln X följer då en binomialfördelning med parametrarna n och p, vilket betecknas

$$X \sim Bin(n, p) \ .$$

Definitioner

- *Sannolikhetsfunktionen* för en binomialfördelad slumpvariabel definieras som

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \ ,$$

där k är antalet försök där A inträffar, n är totala antalet försök och p är sannolikheten för att A inträffar i ett försök.

- Binomialkoefficienten $\binom{n}{k}$ anger på hur många sätt man kan ordna en följd av n försök där A inträffar i k av försöken.

Definitioner

- *Fördelningsfunktionen* för en binomialfördelad slumpvariabel X definieras som

$$F(k) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} .$$

- Väntevärdet för binomialfördelningen är $E(X) = np$.
- Variansen för binomialfördelingen är $Var(X) = np(1-p)$.

Binomialfördelningen

Exempel.

Låt oss anta att andelen ljushåriga i en stad är 30%. Om vi slumpmässigt plockar 5 personer ur befolkningen, vad är sannolikheten att vi får minst 2 och högst 4 ljushåriga i vårt stickprov? Vi låter splumvariabeln X beteckna antalet ljushåriga. Vi får då

$$P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4) =$$

$$\binom{5}{2} 0.3^2 \cdot 0.7^3 + \binom{5}{3} 0.3^3 \cdot 0.7^2 + \binom{5}{4} 0.3^4 \cdot 0.7 =$$

$$0.3087 + 0.1323 + 0.02835 = 0.46935 .$$

Alternativt kan vi använda oss av fördelningsfunktionen. Från en tabell eller med hjälp av ett statistiskt programpaket kan vi direkt läsa ut värden för $F(4) = P(X \leq 4)$ och $F(1) = P(X \leq 1)$ och får då

$$P(2 \leq X \leq 4) = F(4) - F(1) = 0.99757 - 0.52822 = 0.46935 .$$

Binomialfördelningen

Exempel.

Väntevärdet för slumpvariabeln X i föregående exempel är

$$E(X) = n \cdot p = 5 \cdot 0.3 = 1.5 \ .$$

Vi kan tolka det som att vi i ett mycket stort antal stickprov i medeltal skulle få 1.5 ljushåriga per stickprov.

Variansen för X är

$$Var(X) = n \cdot p \cdot (1 - p) = 5 \cdot 0.3 \cdot 0.7 = 1.05.$$

Normalfördelningen

Normalfördelningen har en mycket central plats inom statistik. Detta är bl.a. för att

- många (men långt ifrån alla!) variabler följer en normal fördelning
- icke-normalfördelade variabler kan ibland transformeras så att de följer en normalfördelning
- många statistiska mått (t.ex. medelvärdet för stora stickprov) följer en normalfördelning.

Egenskaper

- Normalfördelningen är en kontinuerlig sannolikhetsfördelning med en symmetrisk och ”klockformad” täthetsfunktion.
- Fördelningen bestäms helt av väntevärdet μ och standardavvikelsen σ.
  - μ anger var toppen av kurvan befinner sig.
  - σ anger hur koncentrerad kurvan är kring μ.
- Att en slumpvariabel X följer en normalfördelning med parametrarna μ och σ betecknas

$$X \sim N(\mu, \sigma) \, .$$

Standardiserad normalfördelning

- Eftersom det för varje tänkbart värdepar (μ, σ) finns en normalfördelning finns det oändligt många normalfördelningar.
- Alla normalfördelningar kan standardiseras till en sk. standardiserad normalfördelning som har väntevärdet 0 och standardavvikelsen 1, dvs. $N(0, 1)$.
- Om $X \sim N(\mu, \sigma)$ får vi genom standardisering en ny variabel

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

- Fördelingsfunktionen för en slumpvariabel Z som följer en standardiserad normalfördelning betecknas $\Phi(z) = P(Z \leq z)$.

Exempel.

Vid användning av en viss mätmetod antas de erhållna värdena vara normlafördelade med väntevärdet 28.0 och standardavikelsen 0.25, dvs. $N(28.0, 0.25)$. Vi frågar oss nu vad sannolikheten är att ett mätvärde ligger mellan 27.5 och 28.5. Uträkningen blir som följande:

$$\begin{aligned} P(27.5 < X \leq 28.5) &= P(\frac{27.5 - 28.0}{0.25} < Z \leq \frac{28.5 - 28.5}{0.25}) \\ &= \Phi(2) - \Phi(-2) \\ &= 0.9772 - 0.0228 = 0.954 \ . \end{aligned}$$

Om man använder en tabell där endast icke-negativa värden är tabulerade kan man ersätta $\Phi(-2)$ med $1 - \Phi(2)$.

I statistik ofta använda sannolikheter

- För en standardiserad normalfördelning $N(0,1)$ gäller att
  - 95% av alla värden ligger mellan -1.96 och 1.96:

$$P(-1.96 < Z \leq 1.96) = 0.95$$

  - 99% av alla värden ligger mellan -2.58 och 2.58:

$$P(-2.58 < Z \leq 2.58) = 0.99$$

  - 99.9% av alla värden ligger mellan -3.29 och 3.29:

$$P(-3.29 < Z \leq 3.29) = 0.999\ .$$

I statistik ofta använda sannolikheter

- På motsvarande sätt gäller för en allmän normalfördelning $N(\mu, \sigma)$ att

$$P(\mu - 1.96 \cdot \sigma < X \leq \mu + 1.96 \cdot \sigma) = 0.95$$
$$P(\mu - 2.58 \cdot \sigma < X \leq \mu + 2.58 \cdot \sigma) = 0.99$$
$$P(\mu - 3.29 \cdot \sigma < X \leq \mu + 3.29 \cdot \sigma) = 0.999 \ .$$

Andra fördelningar

Andra i statistiska sammanhang ofta förekommande fördelningar är

- χ^2-fördelningen (”*khi i kvadrat*” -fördelningen)
- t-fördelningen
- F-fördelningen .