Matematisk statistik V
Föreläsningsanteckningar
Tommy Norberg
16 mars 2005

F5-F6: Ch 5 Probability and sampling distributions

Nyckelord: experiment, händelse, Venn-diagram, och, eller samt komplement, disjunkta händelser, sannolikhetsaxiomen, additionsregler, betingad sannolikhet, oberoende händelser, totala sannolikhetslagen, Bayes formel, diskreta och kontinuerliga stokastiska variabler, sannolikhetsfördelning, väntevärde, varians, standardavvikelse, kvantil, statistika, standard- eller typfel, centrala gränsvärdessatsen

5.1 Slumpmässiga försök (s 184-188)

Terminologi (s 184)

Ett *slumpmässigt försök* är den ram vi lägger runt våra sannolikhetsberäkningar. Vi observerar *variabler* men det vi räknar ut sannolikheter för är händelser. En *händelse* ("event") är något som antingen inträffar eller inte inträffar då vi utför ett slumpmässigt försök. Ofta använder vi ordet *utfall* ("sample point, outcome") för resultatet av ett slumpmässigt försök. Mängden av alla möjliga utfall kallar vi för försökets *utfallsrum* ("sample space"). Det är inget annat än variabelns värdemängd.

Tänk dig att du utför ett slumpmässigt försök som består i att observera en variabel x. När du gjort försöket har du erhållit ett utfall som vi också kallar x. Tänker du göra fler försök, så är det naturligt att låta x_1 beteckna det första försöksresultatet, x_2 det andra, etc. Låt X vara motsvarande stokastiska variabel.

Variabeln x kan egentligen vara av vilken matematisk typ som helst, men vi ska tänka oss den reellvärd. Utfallsrummet är då en delmängd av eller hela R. Låt $A \subseteq R$. Utsagan $X \in A$ är en händelse, ty efter det att försöket har utförts vet vi om försöksresultatet (dess utfall) x tillhör A eller ej. Alla händelser kan kopplas ihop med en mängd A på detta sätt. Ofta studeras

endast en variabel. Då är den ofta underförstådd och man skriver A för händelsen och säger att A inträffar istället för $X \in A$.

Tänk dig nu att vi ska utföra ett slumpmässigt försök. Det kan vara allt ifrån att man observerar hur många ägg ett visst svalpar lägger i boet till en avancerad mätning av mängden ozon i luften vi andas..

Utfallsrummet S är mängden av alla möjliga försöksresultat. Elementen x i S kallas för *utfall*. Så x är variabeln vi mäter och till den hör en stokastisk variabel X. Delmängderna $A, B, \ldots$ till S kallas för *händelser*. Händelsen A utläses "A inträffar". Den kan också skrivas $X \in A$ eller $x \in A$.

Sannolikheten att A inträffar skrives

$$P(A) \text{ eller } P(X \in A) \text{ eller } P(x \in A).$$

Låt A, B vara händelser. Då betecknar

$A \cap B$ händelsen "A och B"

$A \cup B$ händelsen "A eller B"

A^c eller A' händelsen "icke-A"

Dessutom definierar vi

$A \setminus B = A \cap B^c$ (utläses A minus B eller A men ej B)

$A \triangle B = A \setminus B \cup B \setminus A$ (symmetriska differensen)

Observera att $A \triangle B$ inträffar om, och endast om, exakt en av händelserna A och B inträffar.

Händelserna A och B sägs vara *ömsesidigt uteslutande* eller *disjunkta*, om

$$A \cap B = \emptyset$$

Händelserna $A_1, \ldots, A_n$ *partitionerar* en händelse B, om

$$B \subseteq \bigcup_{i=1}^{n} A_i$$

och

$$A_i \cap A_j = \emptyset \quad \text{då} \quad i \neq j$$

Istället för partitionerar kan man säga *delar in i disjunkta eller ömsesidigt uteslutande delar*.

Exempel 1 Förorenad mark. Låt y vara koncentrationen av någon giftig tungmetall. Låt x vara resultatet av en mätning av y.

Marken betraktas som förorenad om $y \geq c$, där c är ett av myndigheterna specificerat kritiskt värde. Denna händelse betecknas C.

Låt d vara den s k detekteringsnivån (obs att typiskt är $d \neq c$). Händelsen $x \geq d$ betecknas D.

Rita ett träddiagram över vad som kan hända.

Rita in händelserna C och D i ett koordinatsystem. Observera därvid att

$$C = \{y \geq 0 : y \geq c\}$$

$$D = \{x \geq 0 : x \geq d\}$$

Utfallsrummet för båda variablerna är $R_+ = [0, \infty)$.

I ett Venn-diagram tänker man sig att utfallsrummet är en rektangel. I rektangeln kan man illustrera händelser genom att rita in motsvarande mängder.

Rita in händelserna C och D i ett Venndiagram.

5.2 Sannolikheter (s 189-193)

Sannolikhetsaxiomen (s 190-191)

Låt A, B vara två godtyckliga händelser i utfallsrummet S. Då

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. $P(A \cup B) = P(A) + P(B)$ om A, B är ömsesidigt uteslutande

Sannolikhetsräkningar (s 191-193)

Sats 1: $P(A \setminus B) = P(A) - P(A \cap B)$

Följd 1: $P(\emptyset) = 0$

Följd 2: $P(A^c) = 1 - P(A)$

Följd 3: $A \subseteq B \Rightarrow P(A) \leq P(B)$

Sats 2: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Följd 1: $P(A \cup B) \leq P(A) + P(B)$

Bevis:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

ty händelserna $A \cap B$ och $A \cap B^c$ är ömsesidigt uteslutande (disjunkta). Vi får

$$P(A \setminus B) = P(A \cap B^c) = P(A) - P(A \cap B)$$

(sats 1). Vidare,

$$P(\emptyset) = P(A \setminus A) = P(A) - P(A \cap A) = 0$$

ty $A \cap A = A$ (följd 1). Vidare,

$$P(A^c) = P(S \setminus A) = P(S) - P(S \cap A) = 1 - P(A)$$

ty $S \cap A = A$ och $P(S) = 1$ enligt axiom 2 (följd 2). Om $A \subseteq B$, så

$$0 \leq P(B \setminus A) = P(B) - P(B \cap A) = P(B) - P(A)$$

(följd 3). Notera nu att $A \setminus B$ och B är ömsesidigt uteslutande och att $A \setminus B \cup B = A \cup B$. Detta ger

$$P(A \cup B) = P(A \setminus B) + P(B)$$

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

(sats 2). Till sist ser vi att

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

(följd 1). *QED*

Exempel 2 Förorenad mark (forts)
Antag $P(C) = 0.8$, $P(D) = 0.74$ och $P(C \cap D) = 0.72$. Beräkna

1. $P(C \cup D)$
2. $P(C \setminus D)$
3. $P(D \setminus C)$
4. $P(C \triangle D)$

Vi får

$$P(C \cup D) = P(C) + P(D) - P(C \cap D)$$
$$= 0.8 + 0.74 - 0.72 = 0.82$$

Vidare

$$P(C \setminus D) = P(C) - P(C \cap D) = 0.8 - 0.72 = 0.08$$

och

$$P(D \setminus C) = P(D) - P(C \cap D) = 0.74 - 0.72 = 0.02$$

Så falsklarmsannolikheten är

$$P(C \triangle D) = P(C \setminus D \cup D \setminus C) = 0.08 + 0.02 = 0.1$$

5.3 Betingade sannolikheter och oberoende händelser (s 194-199)

Betingade sannolikheter (s 195-196)

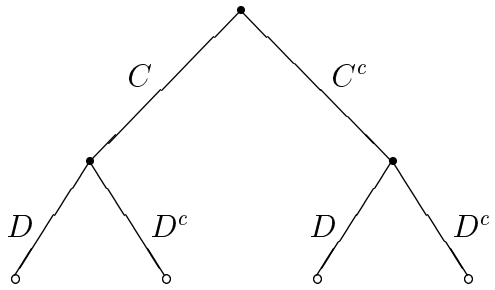
Den *betingade sannolikheten för A givet B*, är

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

Observera att

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

Exempel 3 Förorenad mark (forts) Nedanstående träddiagram ger en bra överblick över problematiken:



Rita in sannolikheter och räkna ut och rita in de betingade sannolikheterna i trädet:

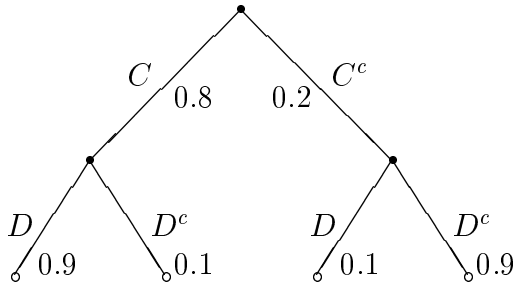
$$P(D|C) = \frac{P(C \cap D)}{P(C)} = \frac{0.72}{0.8} = 0.9$$

$$P(D'|C) = 1 - 0.9 = 0.1$$

$$P(D|C') = \frac{P(D \cap C')}{P(C')} = \frac{0.02}{0.2} = 0.1$$

$$P(D'|C') = 1 - 0.1 = 0.9$$

Vi skriver in dessa sannolikheter i trädet och erhåller:



M.h.a. trädsannolikheterna är det nu lätt att räkna ut

$$P(C \cap D) = P(C)P(D|C) = 0.8 \cdot 0.9 = 0.72$$

$$P(C \setminus D) = P(C \cap D') = P(C)P(D'|C) = 0.8 \cdot 0.1 = 0.08$$

$$P(D \setminus C) = P(C' \cap D) = P(C')P(D|C') = 0.2 \cdot 0.1 = 0.02$$

$$P(C' \cap D') = P(C')P(D'|C') = 0.2 \cdot 0.9 = 0.18$$

Observera att summan av dessa sannolikheter är 1. Vi kan nu igen beräkna falsklarmssannolikheten

$$P(C \triangle D) = P(C \setminus D \cup D \setminus C) = 0.08 + 0.02 = 0.10$$

Oberoende händelser (s 196-198)

Händelserna A och B sägs vara *oberoende*, om

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

Observera att A, B är oberoende om, och endast om,

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

Combining several concepts (s 198-199)

Exempel 4 Förorenad mark (forts)

Beräkna m h a trädet

$$P(D) = P(C \cap D) + P(C' \cap D)$$

$$= P(C)P(D|C) + P(C')P(D|C')$$

$$= 0.72 + 0.02 = 0.74$$

Lagen om total sannolikhet: Låt $A_1, \ldots, A_n$ partitionera utfallsrummet S. Då gäller

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

Exempel 5 Förorenad mark (forts)

Beräkna

$$P(C|D) = \frac{P(C \cap D)}{P(D)}$$

$$= \frac{0.72}{0.74} \approx 0.973$$

Observera att

$$P(D) = P(C)P(D|C) + P(C')P(D|C')$$

(händelserna C och C' partitionerar ju hela utfallsrummet) och att

$$P(C \cap D) = P(C)P(D|C)$$

Härur följer

$$P(C|D) = \frac{P(C)P(D|C)}{P(C)P(D|C) + P(C')P(D|C')}$$

Detta är ett specialfall av Bayes formel. Det allmänna fallet visas analogt.

Bayes formel: Under samma förutsättningar som i lagen om total sannolikhet, gäller

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

5.4 Stokastiska variabler (s 202-211)

Från och med nu, ska vi försöka att mer konsekvent arbeta med *stokastiska variabler* för att betona det faktum att i våra variablers numeriska utfall finns en opredikterbar slumpkomponent.

Stokastiska variabler har (sannolikhets-) fördelningar som beskrivs av tätheter i det kontinuerliga fallet och massfunktioner i det diskreta fallet. Oftast använder vi stora bokstäver t ex X som beteckning för den stokastiska variabeln när vi tidigare betecknat variabeln x. Fast läroboken verkar konsekvent använda små bokstäver.

Väntevärde, varians, percentiler, e dyl, räknas ut som förut m h a tätheten eller massfunktioner. Jag påminner om att många som skriver $E[X]$ istället för bokens μ_x för X:s väntevärde och $\mathrm{Var}[X]$ istället för σ_x^2.

Exempel 5.10 (s 207) Låt T vara livstiden för en produkt. Funktionen

$$R(t) = P(T > t)$$

kallas för produktens *överlevnadsfunktion*. Väntevärdet

$$\mu_T = E[T]$$

kallas ofta för "mean time to failure" (MTTF) eller för"mean time between failures" (MTBF). För en exponentialfördelad variabel gäller att överlevnadsfunktionen är

$$R(t) = \int_t^\infty \lambda e^{-\lambda x}\, dx = \cdots = e^{-\lambda t}$$

och att MTTF är

$$E[T] = \int_0^\infty x\lambda e^{-\lambda x}\, dx = \cdots = \frac{1}{\lambda}$$

Exempel 6 Gränsvärdet

$$h(t) = \lim_{h\to 0} \frac{P(T \le t + h | T > t)}{h} = \frac{f(t)}{R(t)}$$

kallas för *felintensiteten*. För en exponentialfördelad variabel gäller att felintensiteten är konstant, ty

$$h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

Man kan visa att om felintensiteten är konstant, så måste variabeln vara exponentialfördelad.

Exempel 5.11 (s 208) är ett exempel på hur man kan räkna med sannolikheter för stokastiska variabler. Titta själv på detta.

5.5-5.6 Stickprovsvariabler (statistikor) (s 214-227)

Låt $x_1, \ldots, x_n$ vara ett *stickprov* på (den stokastiska) variabeln x. Då är alltså $x_1, \ldots, x_n$ oberoende observationer av x.

Definition 1 (s 214) Reellvärda funktioner av ett stickprov kallar vi *stickprovsvariabler* (statistikor).

Stickprovsvariabler används för att dra slutsatser om x:s fördelning.

Exempel 7 Proportionen observationer i ett intervall, medelvärdet, medianen, standardavvikelsen, IQR, undre och övre kvartilen, alla övriga kvantiler (percentiler).

Definition 2 (s 215) Stickprovsvariabler är stokastiska variabler, så de har fördelningar. På engelska säger man ofta "sampling distribution" för att göra tydligt att det är fördelningen för en stickprovsvariabel som avses.

När vi räknar teoretiskt på stickprov är det praktiskt att använda stora bokstäver för de stokastiska variablerna och små för deras utfall. Vi skriver alltså i fortsättningen $X_1, \ldots, X_n$ för stickprovet och medelvärdet betecknar vi $\bar{X}$. Vi låter S^2 beteckna stickprovets varians och s^2 motsv. empiriska (uppmätta, experimentella) värde.

Exempel 5.14 (s 215) är ett bra exempel på att stickprovsvariabler har fördelningar...

Ofta är det svårt att teoretiskt härleda en viss stickprovsvariabels fördelning. Men vissa egenskaper kan härledas ur faktumet att observationerna är oberoende och likafördelade.

Vi fokuserar intresset på medelvärdet och variansen.

Sats 1 (s 220) Låt $X_1, \ldots, X_n$ vara ett stickprov av den stokastiska variabeln X med väntevärde μ och varians σ^2. Då gäller för medelvärdet $\bar{X}$ att

$$\begin{aligned} E[\bar{X}] &= \mu \\ \mathrm{Var}[\bar{X}] &= \frac{\sigma^2}{n} \end{aligned}$$

Således är

$$\sigma_{\bar{X}} = \sqrt{\mathrm{Var}[\bar{X}]} = \frac{\sigma}{\sqrt{n}}$$

Denna storhet kallas i engelskspråkig litteratur för "standard error". På svenska skulle man kunna säga standardfel eller standardosäkerhet.

Sats 2 (s 221) Om X är normalfördelad med parametrar μ och σ, så är $\bar{X}$ normalfördelad med parametrar μ och $\sigma/\sqrt{n}$.

Sats 3 Centrala gränsvärdessatsen (s 222) Om X ej är normalfördelad, så är ändå fördelningen för medelvärdet $\bar{X}$ är approximativt normalfördelad med parametrar μ och $\sigma/\sqrt{n}$ förutsatt att antalet observationer n är tillräckligt stort.

Anta nu att vårt stickprov $x_1, \ldots, x_n$ består av $\{0, 1\}$-värda variabler. Vi kan tänka oss att vi observerar huruvida en viss händelse inträffar eller ej, och vi låter p vara händelsens sannolikhet. Från diskussionen om Binomialfördelningen vet vi att frekvensen

$$f = X_1 + \ldots + X_n$$

är binomialfördelad med parametrar n och p. Obs att

$$\bar{X} = \frac{f}{n}$$

Relativa frekvensen f/n är alltså ett medelvärde.

Sats 4 (s 225) Om X antar värdet 1 med sannolikheten p och 0 med sannolikheten $1 - p$, så gäller

$$\begin{aligned} \mu &= E[X] = p \\ \sigma^2 &= \mathrm{Var}[X] = p(1-p) \end{aligned}$$

Bevis Vi har att

$$E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

och, eftersom $X^2 = X$, att

$$E[X^2] = p$$

Således gäller att

$$\mathrm{Var}[X] = p - p^2 = p(1-p)$$

QED

Sats 5 (s 226) För relativa frekvensen f/n gäller att väntevärdet är

$$E[f/n] = p$$

och att variansen är

$$\mathrm{Var}[f/n] = \frac{p(1-p)}{n}$$

Centrala gränsvärdessatsen gäller för relativa frekvenser (de är ju medelvärden). Alltså

Sats 6 Då antalet observationer n är stort, så är relativa frekvensen f/n approximativt normalfördelad med parametrar p och $\sqrt{p(1-p)/n}$.

Exempel 8 EC inbjuder till vadslagning och säger: Jag ska singla det här myntet 100 gånger. Det är helt symmetriskt. Titta gärna på det. Om du vill att vi ska singla ett annat mynt, så är det helt ok för mig. Jag satsar 10 kr om du satsar 2 kr emot, på att jag får 40 - 60 klave. Antar du vadet?

Vi analyserar situationen m.h.a. centrala gränsvärdessatsen (cgs). Låt X vara antalet erhållna klave. Då gäller att X är binomialfördelad med parametrar $n = 100$ och $p = 0.5$. Vi har ju ingen alls anledning att tro att myntet är osymmetriskt och ett helt symmetriskt mynt borde i snitt landa lika ofta med klave som med krona upp.

[0]PS. EC skulle kunna stå för Emila eller Emili Chalmers, men EC kan också betyda något helt annat, t.ex. Ezra Cewärd.

Enligt cgs är X/n approximativt normalfördelad med väntevärdet $\mu = 0.5$ och standardavvikelsen $\sigma = \sqrt{0.25/100} = 0.5\sqrt{1/100} = 0.05$ ($n = 100$ är tillräckligt stort för att cgs ska verka). Vi kan nu räkna ut att

$$P(\text{EC vinner}) = P(40 \leq X \leq 60) = P(0.4 \leq X/n \leq 0.6)$$

$$= P\left(\frac{0.4 - 0.5}{0.05} \leq Z \leq \frac{0.60 - 0.5}{0.05}\right) = P(-2 \leq Z \leq 2) \approx 0.95$$

där (ty) $Z = (X/n - 0.5)/0.05$ är approximativt $\mathrm{N}(0, 1)$.

Den vinst EC kan förvänta sig (och också ungefär erhåller i snitt om vadet antas många gånger) är

$$\approx 2 \cdot 0.95 - 10 \cdot 0.05 = 1.9 - 0.5 = 1.4$$

I snitt vinner alltså EC 1 kr och 40 öre varje gång vadet antas. Detta är en synnerligen god affär för EC. Jag skulle inte antagit vadet.

Hemövning 1 Om du satsar 2 kr, hur mycket tycker du att EC ska satsa för att vadet ska bli rättvist?