Föreläsningsanteckningar Sannolikhet, statistik och risk 2015

Johan Jonasson *†‡

Maj 2015

1 Slump och sannolikhet

Finns det äkta slump eller är allt vi betraktar som slump egentligen bara ofullständig information? I kvantvärlden verkar det finnas äkta slump och den kan också användas i avancerad fysikaliska experiment. I de allra flesta tillämpningar handlar det dock om ofullständig information och/eller att även om all information fanns, skulle det vara omöjligt att processa alla dessa data. Tänk till exempel på vad för information och databearbetning som skulle krävas för att förutsäga vädret om ett år; det skulle förmodligen inte vara möjligt att göra ens i teorin med mindre än att bara låta naturen ha sin gång.

Andra exempel på frågor som knappast (just nu i alla fall) kan besvaras exakt utan osäkerhet är:

- Hur kommer tärningen att landa?
- Hur kommer svenska folket att rösta i nästa riksdagsval?
- Härstammar dinosaurier från fåglar?
- Finns det liv på andra planeter?
- Hur många rätt kommer jag att få på tipset nästa vecka?

Vad är sannolikhet? Det finns två vanliga tolkningar. Den ena är i term av relativa frekvenser; sannolikheten för en händelse är den gränsvärdet av den relativa

*Chalmers University of Technology
†Göteborg University
‡jonasson@chalmers.se

frekvensen av antalet försök då händelsen i fråga inträffar, vid en oändlig följd av oberoende upprepningar av försöket. Som exempel, slå en tärning n gånger och låt X_n vara antalet sexor på de n kasten. Om A är händelsen att ett slag ger en sexa, skulle vi tro att

$$\mathbb{P}(A) = \lim_n \frac{X_n}{n}.$$

Ett problem är förstås att en sådan definition kräver att försöket är upprepbart och att gränsvärdet existerar, och hur skulle vi kunna garantera något sådant?

Den andra tolkningen är att sannolikheten för A är graden av hur mycket vi tror på att A kommer att inträffa. Ett spelbolag till exempel använder sin tro (baserad på erfarenhet och gissningar) hur en match mellan Barcelona och Chelsea kommer att sluta, när oddsen ska fastställas. Detta är ett fall där det inte ens i princip är möjligt att upprepa experimentet. Eller för ett ännu mer extremt exempel: om jordens medeltemperatur stiger med 4 grader till år 2100, jämfört med 1990, vad får detta för konsekvenser?

Den matematiska teorin för sannolikheter gör ingen tolkning av vad sannolikheter "egentligen" är, utan sätter istället upp regler för hur de måste uppföra sig. Teorin uttalar sig, precis som all matematiska modeller, alltså inte om hur realistisk en given modell är i en given situation.

I sannolikhetsteorin betraktar man ett *slumpförsök*. Till slumpförsöket finns ett *utfallsrum*, betecknat S, som är mängden av allt som kan hända. Det är i en given situation inte givet hur man ska välja S, utan beror av vad man är intresserad av. Exempelvis är man i en slantsinglingssituation förmodligen intresserad av vilken sida av myntet som kommer upp, och tar då $S = \{H, T\}$. Men det kan ju faktiskt vara så att vi snarare är intresserade av åt vilket håll kungens näsa pekar och tar då kanske $S = [0, 2\pi)$.

Ett element $u \in S$ kallas för ett *utfall* och en delmängd $A \subset S$ kallas för en *händelse*. Eftersom det ofta inte är klart vad S ska vara och ibland heller inte så viktigt att precis tala om vad S och därmed en händelse A är som mängder, utan beskriver bara händelser "i ord".

Exempel 1.1. Tärningsslag. Om vi tar $S = \{1, 2, 3, 4, 5, 6\}$ är händelsen A, som i ord är "jämnt utfall", formellt $A = \{2, 4, 6\}$.

Exempel 1.2. Singla slant tre gånger och betrakta antalet H. Då kan vi ta $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$, men vi kan också välja $S = \{0, 1, 2, 3\}$. Om A är "minst två H, blir A i det första fallet $A = \{HHH, HHT, HTH, THH\}$ och i det andra $A = \{2, 3\}$.

Definition 1.3. En funktion $\mathbb{P} : \mathcal{P}(S) \to [0, 1]$ kallas för ett sannolikhetsmått om

$\mathbb{P}(S) = 1$ och för alla disjunkta händelser $A_1, A_2, \ldots$ gäller att

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Exempel 1.4. Välj ett tal på måfå i $[0, 1]$. Här har vi ta $S = [0, 1]$ och $\mathbb{P}(A) =$ längden av A. Händelserna A = "talet mellan $1/2$ och 1 och $B =$ "talet är högst $2/3$, blir $A = [1/2, 1]$ och $B = [0, 2/3]$. Händelsen "talet ligger mellan $1/2$ och $2/3$ blir $A \cap B = [1/2, 2/3]$.

Proposition 1.5. *För ett sannolikhetsmått $\mathbb{P}$ gäller att*

(a) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,

(b) $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B$,

(c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,

(d) $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.

Bevis. Eftersom A och A^c är disjunkta gäller att $1 = \mathbb{P}(A) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$ och (a) följer. Skriv nu $A = (A \setminus B) \cup (A \cap B)$ och få att $\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)$ och (b) följer. Eftersom $A \cup B$ kan skrivas som den disjunkta unionen $A \cup (B \setminus A)$, följer (c) av (b). Om nu $A \subseteq B$, gäller att $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$. □

Exempel 1.6. Risken för regn på lördag är 50 %, liksom på söndag. Risken för regn båda dessa dagar är 35 %. Vad är chansen till uppehåll hela helgen?

Låt A vara "regn på lördag" och B vara "regn på söndag. Vi söker

$$\mathbb{P}((A \cup B)^c = 1 - \mathbb{P}(A \cup B) = 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A \cap B) = 1 - 0.5 - 0.5 + 0.35.$$

Del (c) i propositionen är ett specialfall av den s.k. inklusion-exklusionsformeln. För tre händelser A, B och C säger den att

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap B) + Pro(A \cap B \cap C).$$

Kontinuitet hos sannolikheter. Antag att $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ och skriv $A = \bigcup_{n=1}^{\infty} A_n$. Då gäller att

$$\mathbb{P}(A) = \lim_n \mathbb{P}(A_n).$$

Vidare gäller att om $B_1 \supseteq B_2 \supseteq \ldots$ och $B = \bigcap_{n=1}^{\infty} B_n$, är

$$\mathbb{P}(B) = \lim_n \mathbb{P}(B_n).$$

Bevis. Låt $A_0 = \emptyset$ och skriv A som den disjunkta unionen $A = \bigcup_{n=1}^{\infty}(A_n \setminus A_{n-1}$. Då

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) - \mathbb{P}(A_{n-1}) = \lim_N \sum_{n=1}^{N} \mathbb{P}(A_n) - \mathbb{P}(A_{n-1}) = \lim_N \mathbb{P}(A_N).$$

Andra delen analogt. □

I fall då utfallsrummet S är ändligt eller uppräkneligt, kan vi skriva $S = \{u_1, u_2, \ldots\}$. Om $p_1, p_2, \ldots$ är ickenegativa tal sådana att $\sum_{n=1}^{\infty} p_n = 1$, kan vi definiera ett sannolikhetsmått genom att ta

$$\mathbb{P}(A) = \sum_{i:u_i \in A} p_i.$$

Exempel 1.7. Antag att vi är intresserade av antalet slantsinglingar som krävs tills vi får H för första gången. Då kan vi ta $S = \{1, 2, \ldots$ och ett rimligt sannolkhetsmåt skulle kunna ges genom att ta $p_1 = 1/2$, $p_2 = 1/4$, $p_3 = 1/8$, etc.

2 Det klassiska sannolikhetsmåttet

Ett viktigt specialfall är situationer där S är ändligt, $S = \{u_1, \ldots, u_n\}$ och $p_i = 1/n$ för alla i. Då blir $\mathbb{P}(A) = \#A/\|$, dvs antalet utfall i A delat med det totala antalet utfall. Detta sannolikhetsmått kallas för det *klassiska* sannolikhetsmåttet. Detta är en mycket vanlig situation, så det finns all anledning att klara av att beräkna antalet utfall i en given händelse.

Exempel 2.1. Om man kastar en tärning tre gånger, vad är sannolikheten att få exakt en sexa? Det finns $6^3 = 216$ olika utfall och det är rimligt att anta att de är lika sannolika. Det finns $1 \cdot 5 \cdot 5 + 5 \cdot 1 \cdot 5 + 5 \cdot 5 \cdot 1 = 75$ olika utfall som ger precis en sexa. Den sökta sannolikheten är alltså $75/216 = 25/72$.

Den s.k. *multiplikationsprincipen* är ett axiom, som säger att om r st experiment utförs i tur och ordning och de individuella experimenten har $n_1, n_2, \ldots, n_r$ olika utfall, så är det totala antalet utfall för alla experimenten tillsammans $n_1 n_2 \ldots n_r$.

Exempel 2.2. *Födelsedagsproblemet.* I en skolklass finns n elever. Vad är sannolikheten att de alla har olika födelsedagar? Om vi antar att för ett på måfå valt barns födelsedag kan vara vilken dag som helst på året med samma sannolikhet, har vi en situation med 365^n olika utfall, av vilka $365 \cdot 364 \cdot \ldots \cdot (365 - n + 1)$ ger olika födelsedagar. Den sökta sannolikheten blir alltså $365 \cdot 364 \cdot \ldots \cdot (365-n+1)/365^n$. Detta blir ca 0.5 då $n = 23$.

I exemplet nyss, utnyttjade vi den princip som säger att ur en mängd med n element, kan man välja k element på $(n)_k = n(n-1)\ldots(n-k+1)$ olika sätt, om man tar hänsyn till i vilken ordning elementen väljs. Detta följer direkt av multiplikationsprincipen. Om man inte tar hänsyn till ordningen på hur elementen väljs, blir antalet val

$$\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!}.$$

Med andra ord, $\binom{n}{k}$ är antalet delmängder med k element till en mängd med n element.

Exempel 2.3. Om man på måfå väljer tre kort ur en kortlek, vadär sannolikheten att man inte får någon spader? Korten kan väljas på $\binom{52}{3}$ olika sätt. Antalet sätt att välja de tre korten så att det inte bir någon spader är $\binom{39}{3}$. Den sökta sannolikheten är alltså $\binom{39}{3}/\binom{52}{3} = 703/1700$.

De nyss nämnda principerna gäller val utan återläggning. Om man istället väljer med återläggning, dvs man valeljer elementen ett i taget, och varje gång ett element väljs, noterar man vilket det var och stoppar sedan tillbaka i mängden. Antalet val blir då n^k om man tar hänsyn till i vilken ordning elementen valdes. Om man istället inte tar hänsyn till ordningen i vilken elementen veljs, blir antalet val detsamma som antalet ickenegativa heltalslösningar till ekvationen

$$x_1 + \ldots + x_n = k.$$

Man kan inse att detta antal är

$$\binom{n-1+k}{n-1} = \binom{n-1+k}{k}.$$

Exempelvis är antalet ickenegativa heltalslösningar till $x_1 + \ldots + x_5 = 3$ lika med $\binom{7}{4} = 35$.

3 Betingad sannolikhet och oberoende händelser

Exempel 3.1. Två tärningar, en blå och en gul, slås. Låt A vara händelsen att den blåa tärningen visar sexa och B att den gula tärningen visar sexa. Det är rimligt

att anta att alla 36 utfall är lika sannolika. Eftersom A och B vardera innehåller 6 utfall och $A \cap B$ endast innehåller utfallet $(6,6)$, får vi att $\mathbb{P}(A) = \mathbb{P}(B) = 1/6$ och $\mathbb{P}(A \cap B) = 1/36$. Vi ser att $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Denna likhet kan vi tolka som att A och B är *oberoende*: chansen att få både A och B om man vet att A inträffar är densamma som sannolikheten att att få B utan denna information.

Definition 3.2. Man säger att A och B är oberoende om $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

I allmänhet kan informationen att A inträffar i heogsta grad påverka chansen att få B. Exempelvis är sannolkiheten att det regnar den 25 augusti nästa år betydligt större om vi vet det kommer att regna den 24 augusti än utan den informationen. Man säger att den *betingade* sannolikheten för regn den 25/8 *givet* regn den 24/8 är större än sannolikheten för regn den 25/8 utan någon information om den 24/8.

Definition 3.3. Den betingade sannolikheten för B givet A, skrivs $\mathbb{P}(B|A)$ och ges av

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

Definitionen förutsätter förstås att $\mathbb{P}(A) > 0$. Om $\mathbb{P}(A) > 0$ gäller att $\mathbb{P}(B|A) = \mathbb{P}(B)$ om och endast om A och B är oberoende. Observera att $\mathbb{P}(B|A)$ i allmänhet inte är lika med $\mathbb{P}(A|B)$ och tolkningen av de två är olika: i det en fallet får vi infomation om att A inträffar medan i det andra får vi information om att B inträffer och dessa situationer kan vara helt olika.

Exempel 3.4. Vi slår en gul och en blå tärning. Låt A vara händlesen tt den gula tärningen visar sexa och B att den blåa tärningen visar sexa. Låt vidare C vara händelsen att summan av de två tarningskasten är 7 och D att summan är 8.

Vi såg ovan att A och B är oberoende. Eftersom C bastår av sex olka utfall får vi $\mathbb{P}(C) = 6/36 = 1/6$. Vidare är $A \cap C$ den händelse som består av det enda utfallet $(6,1)$ och har således sannolikhet $1/36$. Detta betyder att A och C är oberoende. På samma sätt ser vi att B och C är oberoende. Alltså är A, B och C alla parvis oberoende. Är det rimligt att säga att de tre handelserna då är oberoende? Nej, knappast; det är ju omöjligt för alla tre att inträffa samtidigt, så $\mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$, vilket vi också borde kräva.

Handelsen D består av fem utfall, så $\mathbb{P}(D) = 5/36$. Vi har $A \cap D = \{(6,2)\}$, så $\mathbb{P}(A \cap D) = 1/36 > \mathbb{P}(A)\mathbb{P}(D)$, så A och D är inte oberoende.

Proposition 3.5. *OM A och B är oberoende ä även A och B^c oberoende.*

Bevis. Eftersom A och B är oberoende gäller

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c)$$

där oberoendet utnyttjas i den andra likheten. □

Proposition 3.6. *Fixera B och låt $\mathbb{Q}(A) = \mathbb{P}(A|B)$ för alla A. Då är $\mathbb{Q}$ ett sannolikhetsmått.*

Bevis. Eftersom $\mathbb{Q}(A) = \mathbb{P}(A \cap B)/\mathbb{P}(B) \in [0,1]$ för alla A och blir 1 då $A = S$, räcker det att visa uppräknelig additivitet. Låt $A_1, A_2, \ldots$ vara disjunkta händelser då är $A_1 \cap B, A_2 \cap B, \ldots$ disjunkta, så

$$\mathbb{Q}(\bigcup_{n=1}^{\infty} A_n) = \frac{\mathbb{P}(\bigcup_{n=1}^{\infty} A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \mathbb{Q}(A_n).$$

□

En tankebild man kan ha när man tänker på betingade sannolikheter är att när man betingar på en händelse B, stryker man helt enkelt den del av utfallsrummet som utgörs av B^c utan att inbördes ändra sannolikheterna på den del som utgörs av B.

Exempel 3.7. I en hatt ligger tre kort, kort A, B och C. Kort A är rött på bägge sidor, kort B är rött på ena sidan och svart på den andra, medan kort C är svart på bägge sidor. Man blundar och tar ett kort på måfå och lägger det på ett bord med ena sidan upp. När man sedan öppnar ögonen ser man att den sida på kortet som ligger upp är röd. Vad är sannolikheten att den andra sidan också är röd?

I det försök väljs en av de sex sidorna på måfå. Låt oss kalla utfallen, A1, A2, B1, B2, C1, C2 där B1 står för att den röda sidan på kort B är den som syns. Om E är händelsen att man ser en röd sida och F är handelsen att den isda som inte syns är röd, har vi $E = \{A1, A2, B1\}$ och $F = \{A1, A2, B2\}$ så

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)} = \frac{\mathbb{P}(\{A1, A2\})}{\mathbb{P}(\{A1, A2, B1\})} = \frac{2}{3}.$$

I ett exempel ovanför såg vi att för att kalla fler än två händelser oberoende, så räcker det inte med att kräva parvis oberoende. Den allmänna definitionen är följande.

Definition 3.8. Händelserna $A_1, A_2, A_3, \ldots$ sägs vara oberoende om det för alla ändliga indexmängder $i_1, i_2, \ldots, i_n$ gäller att

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}) = \prod_{k=1}^{n} \mathbb{P}(A_{i_k}).$$

I en del exempel har vi, efter att ha räknat, kunnat se att vissa händelser är oberoende. I själva verket är ofta ett antagande om oberoende en del av modellen man arbetar med. Ta till exempel en följd av upprepade tärningskast. Här är det fysikaliskt rimligt att anta att olika tärningskast inte påverkar varandra, varför händelser som har med olika tärningskast att göra är oberoende. Utfallsrummet blir mängden av alla följder $(x_1, x_2, \ldots)$ där varje x_n är ett heltal mellan 1 och 6 och händelsen "sexa i kast nr n" blir mängden av alla sådana följder för vilka $x_n = 6$. Om vi låter B_n vara händelsen att den första sexan kommer på kast nummer n, får vi

$$\mathbb{P}(B_n) = \mathbb{P}(A_1^c \cap A_2^c \cap \ldots \cap A_{n-1}^c \cap A_n) = (\frac{5}{6})^{n-1}\frac{1}{6}.$$

Vi kan också se att chansen att aldrig få en sexa är sannolikheten för händelsen $A_1^c \cap A_2^c \cap \ldots$ som enligt kontinuitet hos sannolikheter har sannolikhet $\lim_n (5/6)^n = 0$.

I många situationer är det lättare att förstå vad vissa betingade sannolikheter är, snarare än de sannolikheter man är intresserad av. Då kan man ha god hjälp av den *totala sannolikhetslagen*. Låt $B_1, \ldots, B_n$ vara en partition av S, dvs disjunkta mängder vars union är S. Då gäller för alla händelser A att

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

Ett specialfall är

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

Exempel 3.9. Antag att 0.1% av befolkningen i Sverige någon gång drabbas av lungcancer. Bland rökare är det 0.4% som drabbas. Om rökarna utgör 20% av befolkningen, vad är då risken för en ickerökare att drabbas?

Välj en person på måfå. Låt A vara händelsen att den valda personen är rökare och B händelsen att personen drabbas av lungcancer. I problemställningen får vi veta att $\mathbb{P}(A) = 0.2$, $\mathbb{P}(B) = 0.001$ och $\mathbb{P}(B|A) = 0.004$. Vi söker $\mathbb{P}(B|A^c)$. Enligt totala sannolikhetslagen är

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c).$$

Lös ut $\mathbb{P}(B|A^c)$ och få

$$\mathbb{P}(B|A^c) = \frac{\mathbb{P}(B) - \mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(A^c)} = \frac{0.001 - 0.2 \cdot 0.004}{0.8} = 0.00025 = 0.025\%.$$

Vad är sannolikheten att en på måfå vald lungcancerpatient är rökare? Vi söker nu $\mathbb{P}(A|B)$. Vi har

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = 0.8.$$

I slutet på det nyss avslutade exemplet, vände vi på betingning. Vi fick

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}.$$

Mer generellt, om $A_1, \ldots, A_n$ är en partition av S,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{k=1}^{n} \mathbb{P}(B|A_k)\mathbb{P}(A_k)}.$$

Denna formel är känd som *Bayes formel.*

Exempel 3.10. I en hatt ligger tre mynt: ett rättvist mynt, ett som ger klave med sannolikhet 0.3 och ett som ger klave med sannolikhet 0.8. Man väljer att mynt på måfå och singlär det tre gånger. Om det blir exakt en klave, vad är då den betingade sannolikheten att mynt nr i valdes?

Låt A vara händelsen att det blir exakt en klave och B_i att mynt nummer i väljs, $i = 1, 2, 3$. Det är rimligt att anta att givet det valda myntet, är de tre kasten oberoende. Om p_i är sannolikheten att mynt nr i ger klave vid ett kast, har vi

$$\mathbb{P}(A|B_i) = 3p_i(1 - p_i)^2.$$

Detta ger $\mathbb{P}(A|B_1) = 3/8 = 0.375$, $\mathbb{P}(A|B_2) = 0.441$ och $\mathbb{P}(A|B_3) = 0.096$ och enligt totala sannolikhetslagen,

$$\mathbb{P}(A) = \frac{1}{3}(0.375 + 0.441 + 0.096) = 0.304.$$

Enligt Bayes formel blir

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)}.$$

Detta blir 0.41 för $i = 1$, 0.48 för $i = 2$ och 0.11 för $i = 3$.

Exempel 3.11. Sjukdomen S. Erfarenheten visar att bland de patienter som visar symptom på den allvarliga sjukdomen S och därmed skickas på provtagning, så är det 1 % som verkligen har sjukdomen. Testet är sådant att för en som verkligen har sjukdomen, kommer testet att vara positivt med sannolikhet 0.95 (effektivitet). Om en person inte har sjukdomen, kommer testet att bli negativt med sannolikhet 0.9 (specificitet). För en given patient som skickats på prov och fått ett positivt provsvar, hur stor är sannolikheten att hon verkligen har sjukdomen?

Låt A vara händelsen att patienten har sjukdomen och T vara händelsen att testet ger positivt svar. Vi vill finna $\mathbb{P}(A|T)$. Enligt Bayes

$$\mathbb{P}(A|T) = \frac{\mathbb{P}(T|A)\mathbb{P}(A)}{\mathbb{P}(T|A)\mathbb{P}(A) + \mathbb{P}(T|A^c)\mathbb{P}(A^c)} = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.1 \cdot 0.99} = 0.095.$$

Vi ser alltså att även vid ett positivt provsvar, finns det goda chanser att vara frisk. Vi kan också se att det är viktigt att bara den som uppvisar tecken på sjukdomen skickar på test. Antag t.ex. att sjukdomen förekommer hos 0.01% av befolkningen och en på måfå vald person testar positivt. Då får vi byta ut 0.01 i räkningen ovan med 0.0001 och får att den betingade sannolikheen att vara sjuk givet ett positivt test endast är 0.00095. Att införa ett allmänt test (s.k. screening) kommer allttså att leda till stora mängder av falsklarm.

Ett kanep som ofta kan anvaendas tillsammans med otala sannolikhetslagen ae rekursivt taenkande. Laot oss se pao detta med ett exempel.

Exempel 3.12. Detta exempel aer kaent som the Gambler's ruin problem. Anna och Bo har tillsammans n kronor, varav Anna startar med a kr. Spelet ges av slantsinglingar, och om en slantsingling ger klave ger Bo 1 kr till Anna och vid utfall krona ger Anna 1 kr till Bo. Vi inser att foerr eller senare kommer hela den samlade foermoegenhet att hamna ho en av spelarna. Laot p_a vara sannolkheten att Anna vinner om hon startar med a kr. Genom att betinga pao resultatet av den foersta slantsinglingen faor vi

$$p_a = \frac{1}{2}p_{a-1} + \frac{1}{2}p_{a+1}$$

med randvillkor $p_0 = 0$, $p_n = 1$. Detta ger $p_2 = 2p_1$, $p_3 = 3p_1$ och allmaent $p_k = kp_1$. Eftersom $p_n = 1$, faor vi $p_1 = 1/n$ och allmaent att $p_a = a/n$.

4 Stokastiska variabler

En stokastisk variabel är ett slumptal. Eftersom vi modellerar slumpförsök med ett utfallsrum S där det är slumpen som bestämmer vilket utfall $u \in S$ som realiseras, är det lämpligt att se ett slumptal, som ett rellt tal vars värde beror på u.

Definition 4.1. En stokastisk variabel är en funktion $X : S \to \mathbb{R}$.

Exempel 4.2. I en fruktskål ligger ett äpple, en apelsin och en banan. Vi blundar och väljer an frukt på måfå. Då är det naturligt att ta $S = \{$äpple,apelsin,banan$\}$ och vi kan till exempel ha $X(u)$ =vikten av u, $Y(u)$ =energiinnehållet i u och $Z(u)$ =sockermängden i u, exempelvis $X(\text{banan}) = 160$ gram och $Z(\text{apelsin}) = 10$ gram etc.

Exempel 4.3. Välj en punkt slumpmässigt i enhetsskivan $D = \{(x, y \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$, (till exempel genom att kasta pil på en piltavla.) Möjliga stokastiska variabler är till exempel, $X(u) = |u|$, $Y(u) = \pi|u|^2$, $Z(u) =$ poäng i kastet.

För en sv X, kan vi före försöket yttra oss om sannolikheter hos utsagor om X. (Efter försöket vet vi forstås exakt vad X blev.) Talen $\mathbb{P}(X \in B) = \mathbb{P}(\{u \in S : X(u) \in B\})$ för alla $B \subseteq \mathbb{R}$, sägs utgöra X:s *fordelning*. För att kunna beräkna sannolikheter för alla intressanta utsagor om X, räcker det att känna till X:s *fördelningsfunktion*, som anges av

$$F_X(x) = \mathbb{P}(X \leq x),\ x \in \mathbb{R}.$$

Obeservera att F_X är ickeavtagande och högerkontinuerlig (enligt kontinuitet hos sannolikheter; visa gärna detta). Om man vet F_X får man till exempel

$$\mathbb{P}(X \in [a, b]) = F_X(b) - F_X(a-).$$

Exempel 4.4. Singla slant två gånger. Låt X vara antalet klavar. Om vi tar utfallrummet som $S = \{HH, HT, TH, TT\}$ blir $X(HH) = 1$, $X(TH) = X(HT) = 1$ och $X(TT) = 0$. Fördelningsfunktionen ges av

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}, & 0 \leq x < 1 \\ \frac{3}{4}, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

5 Diskreta stokastiska variabler

Låt $X : S \to \mathbb{R}$ vara en stokastisk variabel. Om värdemängen V_X är ändlig eller uppräknelig, kallas X för en *diskret stokastisk variabel.* För en sådan sv ges *frekvensfunktionen* av

$$p_X(x) = \mathbb{P}(X = x),\ x \in V_X.$$

Till exempel för en rättvis slantsingling kan vi ha $X(H) = 1$, $X(T) = 0$ och får då frekvensfuntionen $p_X(0) = p_X = 1/2$.

Det är uppenbart att om man känner frekvensfuntionen så kan man beräkna alla sannolikheter för utsagor om X. Exempelvis blir

$$F_X(x) \sum_{v \in V_X : v \leq x} p_X(v),$$

och åt andra hållet,

$$p_X(x) = F_X(x) - F_X(x-).$$

Exempel 5.1. Slå en tärning tills den första sexan kommer. Om X är antalet slag som krävs, blir $V_X = \{1, 2, 3, \ldots\}$ och

$$p_x(k) = \frac{1}{6}(\frac{5}{6})^{k-1}.$$

Exempel 5.2. Låt X vara antal klavar vid fyra oberoende och rättvisa slantsinglingar. Vi får

$$p_X(k) = \binom{4}{k}(\frac{1}{2})^4$$

så $p_X(0) = p_X(4) = 1/16$, $p_X(1) = p_X(3) = 1/4$ och $p_X(2) = 3/8$.

6 Kontinuerliga stokastiska variabler

Låt X vara en sv vars värdemängd är överuppräknelig. Då är således X inte diskret. Då *kan* X istället vara kontinuerlig.

Definition 6.1. En sv variabel kallas kontinuerlig om det finns en funktion f_X, kallad X:s *sannolikhetstäthet*, sådan att

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

för alla $x \in \mathbb{R}$.

Man inser att X är kontinuerlig om och endast om F_X är styckvis deriverbar och $F_X' = f_X$ i alla punkter där f_X äe kontinuerlig. Om X är kontinuerlig får vi

$$\mathbb{P}(X \in B) = \int_B f_X(t)dt$$

för alla $B \subseteq \mathbb{R}$. Exempelvis gäller

$$\mathbb{P}(X = b) = \int_b^b f_X(t)dt = 0$$

för alla b, så för kontinuerliga sv behöver man inte vara noga med om man använder strikta eller ickestrikta olikheter i sina utsagor om X. Till exempel gäller

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_x(t)dt.$$

En funktion $f : \mathbb{R} \to \mathbb{R}$ är en möjlig täthet för en kontinuerlig sv om och endast om $f \geq 0$ och $\int_{-\infty}^{\infty} f(x)dx = 1$.

Tätheten till en sv är inte en sannolikhet utan just en sannolikhetstäthet. En intuitiv tolkning är att om $\epsilon > 0$ ar mycket litet, är

$$\mathbb{P}(X \in (x - \epsilon/2, x + \epsilon/2)) \approx \epsilon f_X(x).$$

En kontinuerlig sv kallas för *likformigt fördelad* på $[a, b]$ om

$$f_X(x) = \frac{1}{b-a},\ x \in [a, b].$$

En förkortad skrivform för detta är

$$X \sim \text{likf}[a, b].$$

Man ser direkt att om $X \sim \text{likf}[a, b]$, är

$$F_X(x) = \frac{x-a}{b-a},\ x \in [a, b].$$

Ett speciall fall är då $[a, b] = [0, 1]$, då både f_X och F_X blir konstant lika med 1 på $[0, 1]$. Vi ser att om $X \sim \text{likf}[a, b]$, är

$$\mathbb{P}(X \in (c, d)) = \frac{d-c}{b-a} = \frac{\ell(c, d)}{\ell(a, b)}$$

om $a \leq c < d \leq b$, där ℓ står för längden av en delmängd av $\mathbb{R}$.

Om man har en sv X som är likformigt fördelad på $[0, 1]$ är det lätt att transformera denna till en sv som är likf$[a, b]$ genom att ta $Y = a + (b - a)X$. Detta inses av

$$F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}((b-a)X \leq x - a) = \mathbb{P}(X \leq \frac{x-a}{b-a}) = \frac{x-a}{b-a}$$

dvs Y har önskad fördelingsfunktion.

Den vanligaste typen av (pseudo)slumptal som genereras i en dator är likf$[0, 1]$.

7 Funktioner av stokastiska variabler

Låt $X : S \to \mathbb{R}$ vara en sv och $g : \mathbb{R} \to \mathbb{R}$ vara en funktion. Då blir även sammansättningen $g \circ X = g(X)$ en avbildning från S till $\mathbb{R}$, dvs en sv. Skriv $Y = g(X)$. Eftersom V_Y inte är ett större rum än V_X, blir Y diskret då X är diskret. Om X är kontinuerlig, kan Y vara av vilken typ som helst, berönde på hur funktionen g är definierad. Antag t.ex. att $X \sim \text{likf}[0, 1]$. Om

$$g(x) = \begin{cases} 1, & x \geq 2/3 \\ 0, & x < 2/3 \end{cases}$$

blir Y diskret med $\mathbb{P}(Y=1)=1/3$ och $\mathbb{P}(Y=0)=2/3$. Å andra sidan om t.ex. $g(x)=x$ blir ju Y också likformigt fördelad på $[0,1]$.

Om g är sådan att Y är kontinuerlig, finns det ingen generell formel som ger tätheten för Y, utan denna får beräknas från fall till fall. Oftast är det en god strategi att börja med att försöka bestämma fördelningsfunktionen för Y.

Exempel 7.1. Antag att X är likformig på $[0,1]$ och att Y är arean av en rektangel med sidorna X och $1-X$, dvs $Y=X(1-X)$. Då är

$$F_Y(y)=\mathbb{P}(X(1-X)\le y)=\mathbb{P}(X^2-X+y\ge 0)=\mathbb{P}\left(X\ge\frac{1}{2}+\sqrt{\frac{1}{4}-y}\right)$$

$$+\mathbb{P}\left(X\le\frac{1}{2}-\sqrt{\frac{1}{4}-y}\right)=1-2\sqrt{\frac{1}{4}-y}$$

för $y\in[0,1/4]$.

Om g är strängt växande eller strängt avtagande existerar inversen g^{-1} och det går att göra en generell beräkning av tätheten av $Y=g(X)$ (som alltså är kontinuerlig i dessa fall). Antag att g är strängt växande. Då blir

$$F_Y(y)=\mathbb{P}(g(X)\le y)=\mathbb{P}(X\le g^{-1}(y))=F_X(g^{-1}(y)).$$

Genom att ta derivatan får vi att tätheten för Y ges av

$$f_Y(y)=f_X(g^{-1}(y))(g^{-1})'(y).$$

En liknande beräkning då g är strängt avtagande ger den generella formeln då g är inverterbar som

$$f_Y(y)=f_X(g^{-1}(y))|(g^{-1})'(y)|.$$

Exempel 7.2. Antag att X är likformig på $[0,1]$ och att $Y=X^3$. På $[0,1$ är $g(x)=x^3$ strikt växande med invers $g^{-1}(x)=x^{1/3}$, så

$$f_Y(y)=\frac{1}{3}x^{-2/3},\ x\in(0,1).$$

8 Väntevärden

Man säger att en följd $X_1,X_2,\ldots$ av sv är oberoende om alla handelser som är utsagor om olika variabler är oberoende. Formellt betyder detta att $X_1,X_2\ldots$ kallas oberoende om det för alla mängder $A_1,A_2,\ldots$ av reella tal gäller att händelserna $\{X_1\in A_1\},\{X_2\in A_2\},\ldots$ är oberoende.

Antag nu att $X_1, X_2, \ldots$ är oberoende och alla har samma diskreta fördelning som X, och tar sina värden i värdemängden $V_X = \{x_1, x_2, \ldots, x_k\}$. För alla $x \in V$ och $n = 1, 2, 3, \ldots$, $N_n(x)$ vara antalet av $X_1, \ldots, X_n$ som antar värdet x. Man förvänter sig att om n är mycket stort, så blir $N_n(x)/n \approx p(x)$. Detta innebär att man förväntar sig att medelvärdet av $X_1, \ldots, X_n$ blir approximativt $\sum_{x \in V} x p_X(x)$. Eftersom man förväntar sig att relativa frekvenser konvergerar mot sannolikheter, bör denna approximation bli allt bättre ju större n. Vi kan alltså se $\sum_x x p_X(x)$ som ett slag idealt medelvärde i det långa loppet. Detta ideala medelvärde kallas för *väntevärdet* av den sv som har den aktuella fördelningen.

Definition 8.1. Låt X vara en diskret sv med frekvensfunktion p_X. Väntevärdet för X ges av

$$\mathbb{E}[X] = \sum_{x \in V_X} x p_X(x).$$

Om X är en kontinuerlig sv med täthet f_X ges väntevärdet av

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Definitionen förutsätter att summan respektive integralen är väldefinierad, vilket kräver att $\sum_{x \in V_X} |x| p_X(x) < \infty$ respektive $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.

Exempel 8.2. Spela på svart i Roulette. Antag att $p_X(-1) = 19/37$ och $p_X(1) = 18/37$. Då är $\mathbb{E}[X] = (-1)19/37 + 1 \cdot 18/37 = -1/37$. Man går alltså back i genomsnitt $1/37$ marker för varje satsad marker.

Exempel 8.3. Låt X vara antalet klavar vid fyra oberoende rättvisa slantsinglingar. Vi såg tidigare att $p(0) = p(4) = 1/16$, $p(1) = p(3) = 1/4$, $p(2) = 3/8$. Vi får alltså

$$E[X] = \frac{1}{16}(0 + 4) + \frac{1}{4}(1 + 3) + \frac{3}{8}2 = 2.$$

I det sista exemplet ovan, var frekvensfunktionen symmetrisk och väntevärdet blev just symmetripunkten. Mer generellt är väntevärdet *tyngdpunkten* for frekvensfunktionen. I det kontinuerliga fallet blir väntevärdet tyngdpunkten för tätheten.

Exempel 8.4. Väntevärdet av ett tärningsslag. Låt X vara värdet av ett tärningsslag. Då är frekvensfunktonen $p_X(x) = 1/6$, $x = 1, 2, 3, 4, 5, 6$. Av symmetri ser vi direkt att $\mathbb{E}[X] = 3.5$. Mer generellt om $p_X(x) = 1/n$, $x = 1, 2, \ldots, n$, har vi $\mathbb{E}[X] = (n + 1)/2$.

Exempel 8.5. Om $X \sim$ likf$[a, b]$, är tätheten symmetrisk och vi ser att $\mathbb{E}[X] = (a + b)/2$.

Exempel 8.6. Låt $X \sim$ likf$[0,1]$. och låt A vara arean av en kvadrat med sidan X, dvs $A = X^2$. Fördelningen av A får vi via fördelningsfunktionen:

$$F_A(a) = \mathbb{P}(X \leq \sqrt{a}) = \sqrt{a}.$$

Tätheten ges alltså av derivatan $f_A(a) = 1/2\sqrt{a}$. Därmed är väntevärdet

$$\mathbb{E}[A] = \frac{1}{2}\int_0^1 x\,dx/\sqrt{x} = \frac{1}{3}.$$

Om X är kontinuerlig och endast antar positiva värden, finns det en användbar formel för väntevärdet, baserad på fördelningsfunktionen.

Proposition 8.7. *Antag att $V_X \subseteq [0,\infty)$. Om X är kontinuerlig gäller att*

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))dx = \int_0^\infty \mathbb{P}(X > x)dx.$$

Om X är diskret och $V_X \subseteq \{1,2,\ldots\}$ gäller

$$\mathbb{E}[X] = \sum_{n=1}^\infty \mathbb{P}(X \geq n) = \sum_{n=0}^\infty \mathbb{P}(X > n).$$

Bevisen fås genom att byta integrationsordning eller summationsordning. I det kontinuerliga fallet ser det ut som följer

$$\int_0^\infty \mathbb{P}(X > x)dx = \int_0^\infty \int_x^\infty f_X(t)dt\,dx = \int_0^i yf_X(t)\int_0^t dx\,dt = \int_0^\infty tf_X(t)dt = \mathbb{E}[X].$$

Exempel 8.8. Se på förra exemplet. Där var $F_A(a) = sqrta$, så $\mathbb{E}[A]\int_0^1(1 - sqrta)da = 1/3$.

Exempel 8.9. Låt X vara antal tärningsslag tills första sexan kommer. Då gäller $\mathbb{P}(X > x) = (5/6)^x$ så $\mathbb{E}[X] = \sum_{x=0}^\infty (5/6)^x = 6$.

Låt oss nu för en tredje gång titta på exemplet $A = X^2$, $X \sim$ likf$[0,1]$. Vi hade

$$\mathbb{E}[X^2] = \mathbb{E}[A] = \int_0^1 x\,f_A(x)dx = \frac{1}{2}\int_0^1 \sqrt{x}dx.$$

Genom att göra substitutionen $x = t^2$ blir detta

$$\mathbb{E}[X^2] = \int_0^1 t^2 dt.$$

Vi hade alltså $A = g(X) = X^2$ och fick $\mathbb{E}[g(X)] = \int_0^\infty g(t)f_X(t)dt$. Finns det ett mönster här? Ja faktiskt. Detta resultat är känt som *den omedvetne statistikerns lag*.

Sats 8.10. *Låt $g : \mathbb{R} \to \mathbb{R}$. Om X är kontinuerlig gäller att*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Om X är diskret:

$$\mathbb{E}[g(X)] = \sum_{x \in V_X} g(x) f_X(x).$$

Bevis. Vi tar det diskreta beviset och lämnar det kontinuerliga fallet som övning. Sätt $Y = g(X)$. Då är

$$\mathbb{E}[Y] = \sum_{y \in V_Y} y\mathbb{P}(Y = y) = \sum_{y \in V_Y} \sum_{x \in V_X : g(x) = y} y\mathbb{P}(X = x) = \sum_{x \in V_X} g(x)\mathbb{P}(X = x).$$

□

I ett av exemplen ovan såg vi att om $X \sim$ likf$[0, 1]$ och $g(x) = x^2$, så var $\mathbb{E}[g(X)] = 1/3$, medan $\mathbb{E}[X] = 1/2$, så att $g(\mathbb{E}[X]) = 1/4$. Det är alltså *inte* i allmänhat sant att $\mathbb{E}[g(X)] = g(\mathbb{E}[X])$, vilket man kanske skulle kunna lockas att tro.

Exempel 8.11. Låt X vara likformig på $[0, 1]$ och

$$g(x) = \begin{cases} 1, & x \geq 2/3 \\ 0, & x < 2/3 \end{cases}$$

Här är

$$\mathbb{E}[g(X)] = \int_0^1 g(x) f_X(x) dx = \int_{2/3}^1 dx = \frac{1}{3}.$$

Vi noterar att $\mathbb{E}[X] = 1/2$, så $g(\mathbb{E}[X]) = 0$.

Exempel 8.12. Låt g vara som i exemplet ovanför och låt X ha täthet $f(x) = 2x$, $0 \leq x \leq 1$. Vi får

$$\mathbb{E}[g(X)] = \int_0^1 g(x) f(x) dx = \int_{2/3}^1 2x\, dx = \frac{2}{3}(1 - (2/3)^3) = \frac{38}{81}.$$

8.1 Oändliga väntvärden

Kom ihåg att för att väntevärdet av den kontinuerliga sv:en X, kräver vi att $\int |x| f_X(x) dx < \infty$, med mostsvarande summeringskrav för diskreta sv. Detta kräver vi för att vi annars skulle kunna råka ut för att vi skulle få

$$\mathbb{E}[X] = \int_0^\infty x f(x) dx - \int_{-\infty}^0 (-x) f(x) dx = \infty - \infty = ?$$

Om nu $X \geq 0$, finns inte den andra termen och det möter inga problem att om den första termen är ∞, helt enkelt definiera $\mathbb{E}[X] = \infty$. Kort sagt: Om $X \geq 0$ låter vi $\mathbb{E}[X] = \int_0^\infty x f_X(x) dx$ oavsett om detta blir ändligt eller oändligt.

Exempel 8.13. S:t Petersburgsparadoxen/Dubbling vid slantsingling. Antag att vi sätter 1 kr på att en slantsingling ska bli klave. Om vi förlorar, dubblar vi insatsen till 2 kr till nästa kast. Om vi förlorar igen, dubblar vi till 4 kr etc. Med denna strategi vet vi att vi förr eller senare kommer att vinna, och när vi gör det kommer vi att vinna tillbaka alla pengar vi satsat plus en krona. Lå X vara antalet förlorade kast innan man vinner. Då är $\mathbb{P}(X \geq k) = 2^k$, så $\mathbb{E}[X] = \sum_{k=1}^\infty 2^{-k} = 1$. Låt nu Y vara antal förlorade kronor före första vinsten. Då är $Y = 2^X - 1$ så vi får

$$\mathbb{E}[Y] = \sum_{k=1}^{\infty} (2^k - 1)\mathbb{P}(X = k) = \sum_{k=1}^{\infty} (2^k - 1)2^{-k-1} = \infty.$$

8.2 Varians

Väntevärdet av X är alltså tyngdpunkten hos X:s täthet/frekvensfunktion. Det säger vad medelvärdet av upprepade oberoende observationer av X närmar sig, men säger ingenting om hur mycket X tenderar att avvika från detta. Ett mått på denna avvikelse är *variansen* av X.

Definition 8.14. Låt X vara med ändligt väntevärde μ. Då ges variansen av X av

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[(X - \mu)^2].$$

Standardavvikelsen ges av

$$\mathbb{S}\mathrm{td}[X] = \sqrt{\mathbb{V}\mathrm{ar}[X]}.$$

Notera att variansen kan vara oändlig. Man kan fråga sig varför man inte använder sig av $\mathbb{E}[|X - \mu|]$ som mått på avvikelse. Svaret är att det rent matematisk är betydligt enklare med kvadratavvikelsen.

Genom att utveckla kvadraten i definitionen av varians, följer att

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - \mu^2.$$

Denna likhet går ibalnd under namnet *Steiners formel* och sager alltså att

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Exempel 8.15. Låt $X \sim \text{likf}[a, b]$. Vi har redan sett att $\mathbb{E}[X] = (a + b)/2$. Vidare är

$$\mathbb{E}[X^2] = \frac{1}{b-a} \int_a^b x^2 dx.$$

enligt Steiners formel är alltså

$$\mathbb{V}\text{ar}[X] = \frac{1}{b-a} \int_a^b x^2 dx - \frac{(a+b)^2}{4} = \ldots = \frac{(b-a)^2}{12}.$$

Standardavvikelsen blir därmed $(b - a)/\sqrt{12}$.

Följande resultat gäller för alla sv X och konstanter a och b

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Detta är uppenbart från definitionerna av väntevärde och har redan använts, till exempel till beviset av Steiners formel nyss. Det gäller också att

$$\mathbb{V}\text{ar}[aX + b] = a^2\mathbb{V}\text{ar}[X].$$

Detta inses av foleljande, där $\mu = \mathbb{E}[X]$,

$$\mathbb{V}\text{ar}[aX+b] = \mathbb{E}[((aX+b)-(a\mu+b))^2] = \mathbb{E}[a^2(X-\mu)^2] = a^2\mathbb{E}[(X-\mu)^2] = a^2\mathbb{V}\text{ar}[X].$$

Följande två resultat är av mycket stort teoretiskt intresse.

Markovs olikhet. Antag att $X \geq 0$. Då gäller att

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Bevis. Låt

$$g(x) = \begin{cases} 0, & x < a \\ 1, & x \geq a \end{cases}$$

Då är $g(X) \leq X$, så $\mathbb{E}[g(X)] \leq \mathbb{E}[X]$ (övörtyga dig om att du kan visa att $Y \leq Z \Rightarrow \mathbb{E}[Y] \leq \mathbb{E}[Z]$). Men $\mathbb{E}[g(X)] = a\mathbb{P}(X \geq a)$, så $\mathbb{E}[X] \geq a\mathbb{P}(X \geq a)$, som önskat. □

Chebyshevs olikhet. För alla sv X med ändligt väntevärde μ och alla $\epsilon > 0$, gäller att

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\mathbb{V}\text{ar}[X]}{\epsilon^2}.$$

Bevis.

$$= \mathbb{P}(|X - \mu| \geq \epsilon) = \mathbb{P}((X - \mu)^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{\epsilon^2} = \frac{\mathbb{V}\text{ar}[X]}{\epsilon^2},$$

där olikheten följer av Markovs olikhet. □

Exempel 8.16. IQ. Om intelligenskvoten hos den svenska befolkningen i genomnitt är 100, med en standardavvikelse på 15, ge en övre skattning av andelen av befolkningen som har IQ på över 200.

Låt X vara IQ hos en på måfå vald svensk person. Enligt Chebyshevs olikhet får vi

$$\mathbb{P}(X \geq 200) \leq \mathbb{P}(|X - 100| > 100) \leq \frac{\mathbb{V}\mathrm{ar}[X]}{100^2} = \frac{15^2}{100^2} = 0.0225.$$

(I själva verket är denna andel betydligt mindre än så. Vi kommer att komma tillbaka till dettta exempel.)

9 Speciella fördelningar

9.1 Indikatorer

Om $A \subseteq S$ är en händelse, kallas den stokastiska variabeln I_A given av

$$I_A(u) = \begin{cases} 1, & u \in A \\ 0, & u \notin A \end{cases}$$

för *indikatorfunktionen* för A. Det är uppenbart att $\mathbb{E}[I_A] = \mathbb{E}[I_A^2] = p$ och därmed också att $\mathbb{V}\mathrm{ar}[I_A] = p(1-p)$.

9.2 Binomialfördelning

Detta har vi redan sett i en del exempel. Antag att ett slumpförsök upprepas n gånger, oberoende upprepningar. Om A är en händelse med sannolikhet p till ett av dessa slumpförsök och X är antalet ggr som A inträffar bland dessa n försök, säger vi att X är *binomialfördelad* med parametrar n och p. Ett annat sätt att uttrycka detta är följande. En viss slant visar klave med sannolikhet p vid ett kast. Antag att vi singlar en slant n ggr och låter X vara antal kast som ger klave. Det är uppenbart att vi kan skriva $X = \sum_{i=1}^{n} I_{B_i}$ där $B_1, \ldots, B_n$ är oberoende och B_i är handelsen att vi får klave vid kast nummer i. På kortform skriver man $X \sim \mathrm{Bin}(n, p)$. Notera att $X \sim \mathrm{Bin}(n, p) \Rightarrow n - X \sim \mathrm{Bin}(n, 1-p)$.

Den exakta frekvensfunktionen ges av

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Låt oss beräkna $\mathbb{E}[X]$:

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np$$

där den sista likheten följer av binomialsatsen. Man kan också visa att

$$\mathbb{V}\mathrm{ar}[X] = np(1-p).$$

Vi återkommer tll det.

Exempel 9.1. Antag att lag A och lag B möts i bäst av sju matcher. Styrkeförhållandet mellan lagen är så att A har 60% chans att vinna en given match. Vad är sannolkheten att A vinner matchserien?

Låt X vara antalet matcher som B vinner. Vi kan anta att alla sju matcher alltid spelas, även om det i själva verket är så att man slutar så fort något lag vunnit fyra matcher. Då är $X \sim \mathrm{Bin}(7, 0.4)$. Vi söker $\mathbb{P}(X \leq 3)$.

$$\mathbb{P}(X \leq 3) = \sum_{k=0}^{3} \mathbb{P}(X = k) = \sum_{k=0}^{3} \binom{7}{k} 0.4^k 0.6^{7-k} \approx 0.71.$$

Svaret är alltså att det är cirka 71% chans att A vinner machserien.

9.3 Gemetrisk fördelning

Betrakta en följd av oberoende slantsinglingar, sådana att varje kast ger klave med sannolikhet p. Om vi sätter X till antal kast till och med den första klaven, får X värderum $\{1, 2, 3, \ldots\}$ och frekvensfunktionen

$$p_X(n) = p(1-p)^{n-1}$$

eftersom att få klave för första gången i kast nummer n kräver att kast $1, \ldots, n-1$ alla ger krona och kast nr n ger klave. En sv med denna fördelning kallas *geometriskt* fördelad med parameter p, på kortform $X \sim \mathrm{Geo}(p)$.

Exempel 9.2. Som en fortsättning på det förra exemplet, låt Y vara antalet matcher man behöver spela innan B vinner sin första match. Då är $Y \sim \mathrm{Geo}(0.4)$ och vi har till exmpel att $\mathbb{P}(Y = 3) = 0.4 \cdot 0.6^3 = 0.144$.

Väntevärdet för den geometriska fördelningen ges av

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} np(1-p)^{n-1} = -p\frac{d}{dp}\sum_{n=0}^{\infty}(1-p)^n = -p\frac{d}{dp}\frac{1}{p} = \frac{1}{p}.$$

Man kan också visa att

$$\mathbb{V}\mathrm{ar}[X] = \frac{1-p}{p^2}.$$

Denna beräkning sparas till senare.

Exempel 9.3. Spela 100 rader på Lotto varje veacka tills du får sju rätt för första gången. Antalet veckor, X, man måste spela innan man vinner är geometriskt fördelad med parameter p, där p är sannolikheten att få sju rätt en given vecka. Vi har

$$p = \frac{100}{\binom{39}{7}}.$$

Detta ger bl.a. att

$$\mathbb{E}[X] = \frac{\binom{39}{7}}{100} \approx 153809.$$

I genomsnitt får man alltå vänta i 153809 veckor, dvs ca 2950 år.

9.4 Poissonfördelning

Denna fördelning är tätt sammankopplad med exponentialfördelning, som vi strax ska gå igenom. Man säger att X är Poissonfördelad med parameter λ (ett positivt tal), om

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, 2, \ldots$$

och skriver på kortform $X \sim \text{Poi}(\lambda)$. En naturlig tolkning återkommer vi med senare. Ett viktigt faktum är att om p är mycket litet och n mycket stort, gäller att $X \sim \text{Bin}(n, p)$ medför att X är approximativt Poissonfördelad med parameter np. Från detta anar vi att $\mathbb{E}[X] = \lambda$. Mycke riktigt, med hjälp av Taylors fomel:

$$\mathbb{E}[X] = e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

Den typiska Poissonfördelningen uppstår när man betraktar antalet av någon typ av ovanliga händelser under en viss idsperiod, t.ex.

- antal jordbävningar över 7 på Richterskalan per år i varlden,
- antal trafikolyckor per månad i Västra Götaland,
- antal lungcancerfall per år i Stockholm,
- antal hundägare man möter under sin kvällspromenad i joggingspåret,

etc.

Exempel 9.4. På en enslig skogsväg passerar i genomsnitt 5.3 fordon per dag. Vad är sannolikheten att högst två forden passerar en given dag? Här verkar det rimligt

att anta att antal fordon, X, är Possonfördeld och parametern måste vara $\lambda = 5.3$. Svaret är alltså

$$\mathbb{P}(X \leq 2) = e^{-5.3}(1 + 5.3 + \frac{5.3^2}{2}) \approx 0.102.$$

För att beräkna variansen för Poissonfördelningen beräknar vi

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1) + X] = e^{-\lambda}\left(\sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right) = \lambda^2 + \lambda$$

på samma sätt som för väntevärdet ovan. Alltaso blir

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda.$$

Exempel 9.5. Sverige har ca 10 miljoner invånare och ca 300 dödsfall per år i trafiken. Om en människa lever i 100 år, såvida hon inte omkommer i entrafikolycka innan dess, vad är risken att en given person kommer att dö i just en trafikolycka.

Antalet dödsolyckor X per person och 100 år är rimligt att anta är Poissonfördelad med parameter $100 \cdot 300/10^7 = 0.003$. Vi får $\mathbb{P}(X = 0) = e^{-0.003} \approx 0.997$. Risken att dö just i en trafikolycka är alltså approximativt 0.3%.

Nu går vi över till att se på några speciella kontinuerliga fördelningar.

9.5 Exponentialfördelning

Antag att X är en kontinuerlig sv med täthet

$$f_X(x) = \lambda e^{-\lambda x},\ x \geq 0.$$

Då sägs X vara *exponentialfördelad* med parameter λ ($\lambda > 0$), kortform $X \sim \exp(\lambda)$. Sådana stokastiska variabler uppkommer typiskt som livslängden hos ting som inte åldras. Exempelvis

- livslängd hos elektriska komponenter,
- tid till sonderfall av en radioaktiv atomkärna,
- tid från nu till nästa jordbävning.

Vi har, via partiell integration,

$$\mathbb{E}[X] = \int_0^\infty \lambda x e^{-\lambda x} dx = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Vidare är, enligt den omedvetne statistikerns lag,

$$\mathbb{E}[X^2] = \int_0^\infty \lambda x^2 e^{-\lambda x} dx = \frac{2}{\lambda^2}.$$

Därmed bli variansen

$$\mathbb{V}\mathrm{ar}[X] = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}.$$

Exempel 9.6. Livsländen hos en LED-lampa anges till 10 år, vilket betyder att väntevärdet av livslängden är 10 år. Vad är sannolikheten att en given lampa håller i minst tio år?

Det är rimligt att anta att livslängden X är exponentialfördelad med väntevärde $1/\lambda = 10$, dvs $\lambda = 1/10$. Vi får

$$\mathbb{P}(X \geq 10) = \mathbb{P}(X \geq 1/\lambda) = \lambda \int_{1/\lambda}^\infty e^{-\lambda x} dx = e^{-1} \approx 0.37.$$

Det är alltså så att av lamporna ifråga är det en andel på ca 37 procent som verkligen håller i minst tio år.

Skriv $G(x) = 1 - F_X(x) = \mathbb{P}(X > x)$. Denna kallas för X:s överlevnadsfunktion. Om $X \sim \exp(\lambda)$, får vi

$$G(x) = \int_x^\infty \lambda e^{-\lambda x} dx = e^{-\lambda x}$$

(så $F_X(x) = 1 - e^{-\lambda x}$.) Det följer speciellt att $G(x + y) = G(x)G(y)$. Detta ger

$$\mathbb{P}(X > x + y | X > y) = \frac{G(x+y)}{G(y)} = G(x) = \mathbb{P}(X > x).$$

Detta är den s.k. *glömskeegenskapen* hos exponentialfödelningen; det som har exponentialfördelad livslängd åldras inte. Finns det andra fördelnngar som har denna egenskap, dvs egenskapen $G(x + y) = G(x)G(y)$ för alla x, y? Om denna likhet gäller så följer att

$$\frac{G(x+y) - G(x)}{y} = G(x) \frac{G(y) - G(0)}{y}.$$

Genom att låta $y \to 0$, följer att $G'(x) = CG(x)$, där $C = G'(0)$. Den unika lösningen med $G(0) = 1$ för denna differentialekvation är $G(x) = e^{-Cx}$, dvs en exponentialfördelning. Svaret är alltså att exponentialfördelningen är den enda fördelning som har glömskeegenskapen.

Exempel 9.7. Om man är observant ser man att exponentialfördelningen är mycket lik den geometriska fördelningen, med skillnaden att den ena är kontinuerlig och den andra diskret. Om $X \sim \exp(\lambda)$ och $Y = \lceil X \rceil$, gäller

$$\mathbb{P}(Y = k) = \int_{k-1}^{k} \lambda e^{-\lambda x} dx = e^{-\lambda(k-1)} - e^{-\lambda k} = (1 - e^{-\lambda})(e^{-\lambda})^{k-1}$$

dvs Y är geometriskt fördelad med parameter $1 - e^{-\lambda}$.

Kopplingen, som redan nämnts, mellan exponentialfördelningen och Poissonfördelningen, går via den s.k. *Poissonprocessen*. Antag att tidpunkterna $T_1, T_2, \ldots$ är de tidpunkter då vi registrerar en *impuls*, dvs att en viss händelse inträffar, t.ex. en trafikolycka. Antag vidare att T_1,$T_2 - T_1, T_3 - T_2$ är oberoende och alla exponentialfördelade med parameter λ. Då sägs tidpunkterna $T_1, T_2, \ldots$ utgöra en Poissonprocess med intensitet λ.

Om $X(t)$ är antal impulser före tiden t, dvs $X(t) = \#\{n : T_n \leq t\}$, gäller att $X(t)$ är Poissonfördelad med parameter λt. Vi återkommer med ett bevis av detta.

9.6 Normalfördelningen

Den stokastiska variabeln X sägs vara *normalfördelad* med parametrar μ och σ^2, på kortform $X \sim \mathrm{N}(\mu, \sigma^2)$, om den här tätheten

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

Denna täthet har den karaktäristiska klockformen och är symmetrisk kring μ, av vilket det följer att $\mathbb{E}[X] = \mu$. Normalfördelningen är mycket vanlig, tack vare den centrala gränsvärdessatsen (kommer senare).

Om $Z \sim \mathrm{N}(0, 1)$, kallar man Z för standard-normalfördelad och Z har då alltså täthet

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Vi har, som redan påpekats, att $\mathbb{E}[Z] = 0$. Vidare är

$$\mathbb{V}\mathrm{ar}[Z] = \mathbb{E}[Z^2] = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$$

där den andra likheten följer av partiell integration.

Proposition 9.8. *Om Z är standard-normalfördelad, är $X = \sigma Z + \mu$ normalfördelad med parametrar μ och σ^2. Vice versa om $X \sim N(\mu, \sigma^2)$ är $(X - \mu)/\sigma$ standard-normalfördelad.*

Bevis.

$$\mathbb{P}(X \leq x) = \mathbb{P}(X \leq (x-\mu)/\sigma) = \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \int_{-\infty}^{u} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

som önskat, via substitutionen $u = \sigma x + \mu$. Andra delen helt analogt. □

Standardbeteckning för täthaten i standard-normalfördelningen är

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Motsvarande fördelningsfunktion betecknas med Φ, dvs

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

Som konsekvens av detta och propositionen nyss, gäller för $X \sim \mathrm{N}(\mu, \sigma^2)$ att

$$F_X(x) = \Phi(\frac{x - \mu}{\sigma}).$$

Enligt propositionen ovan, är det bara funktionen Φ man behöver veta för att beräkna alla sannolikheter som har att göra med normalfördelade sv. Funktionen Φ finns i tabeller och som funktion i matematisk mjukvara av alla former. Ofta är man intresserad av Φ^{-1}, dvs av s.k. *percentiler* i standard-normalfördelningen. (Generellt gäller att om X har fördelningsfunktion F och $F(x) = \alpha$, dvs $x = F^{-1}(\alpha)$, kallas x för α-percentilen för X eller för F.) Några intressanta och vanligt använda percentiler för Φ är

- $\Phi^{-1}(0.95) \approx 1.64$,
- $\Phi^{-1}(0.975) \approx 1.96$,
- $\Phi^{-1}(0.995) \approx 2.58$,
- $\Phi^{-1}(0.9995) \approx 3.29$.

Om $Z \sim \mathrm{N}(0, 1)$ är alltså t.ex. $\mathbb{P}(Z \leq 1.96) = 0.975$ och, eftersom $\Phi(-x) = 1 - \Phi(x)$, $\mathbb{P}(|Z| \leq 1.96) = 0.95$.

Exempel 9.9. IQ igen. Antag att IQ hos en på måfå vald svensk är normalfördelad med vantevärde 100 och standardavvikelse 15. Hur stor andel av befolkningen har då en IQ på över 200?

Låt X vara IQ hos vår på måfå valda person. Vi har tydligen att $X \sim \mathrm{N}(100, 15^2)$. Då blir

$$\mathbb{P}(X \geq 200) = 1 - \Phi((200 - 100)/15) = 1 - \Phi(6.67) \approx 1.3 \cdot 10^{-11}.$$

Detta värkar alltså vara mycket osannolikt att någen så intelligent svensk någonsin kommer att existera. Notera dock att när storheter antas vara normalfördelade är detta nästan alltid en approximation och speciellt kommer beräkningar som ger mycket små sannolikheter sällan att vara goda approximationer.

Exempel 9.10. Antag att X är vikten av en burk Abbas fiskbullar angiven till 250 gram. Vad är risken att man får tag på en burk med mindre än 245 gram om vikten är normalfördelad med väntevärde 250 och standardavvikelse 3 gram? Svaret är

$$\mathbb{P}(X \leq 245) = \Phi((245 - 250)/3) = 1 - \Phi(5/3) \approx 0.048$$

enligt tabell.

10 Felintensiteter

Kom ihag att överlevnadsfunktionen för en stokastisk variabel $T \geq 0$ definierades som $G(t) = \mathbb{P}(T > t)$. Om X är kontinuerlig och har täthet f så definieras *felintensiteten* eller *dödsintensiteten* för T som

$$r(t) = \frac{f(t)}{G(t)}.$$

En tolkning av felintensiteten är att

$$\mathbb{P}(T \in (t + \Delta t) | T > t) \approx r(t)\Delta t$$

då ΔT är mycket litet. Vi har redan sett att om T är exponentialfördelad med parameter λ, så är $r(t)$ konstant lika med λ. Att ha konstant felintensitet är ekvivalent med att ha glömskeegenskapen, dvs T är då livslängden för något som inte åldras och heller inte ”föryngras”. Om T istället är livslängd för något som åldras, t.ex. människor och djur, kommer $r(t)$ att vara växande i t, medan om T till exempel är kötiden i vissa kösystem, kan $r(t)$ mycket väl vara avtagande i t.

Om man känner $r(t)$, kan man beräkna $G(t)$ från detta via

$$r(t) = -\frac{d}{dt} G(t)$$

så att

$$G(t) = e^{-\int_0^t r(s)ds},$$

dvs felintensiteten bestämmer fördelningen för T

Exempel 10.1. Låt X vara likformig på $[0, 1]$. Då blir

$$r(t) = \frac{f(t)}{G(t)} = \frac{1}{1-t},\ t \in (0, 1).$$

Vi ser att i detta fall är $r(t)$ växande i t.

Exempel 10.2. Antag att $G(t) = 1/(1+t)^2$. Då blir $f(t) = 2/(1+t)^3$ och därmed

$$r(t) = \frac{2}{1+t}$$

som alltså i detta fall blir avtagande i t.

Om T har egenskapen att $r(t)$ är strikt avtagande i t, så säger man att T har en *tungsvansad* fördelning.

11 Bivariata och multivariata foerdelningar

Lao S vara ett utfallsrum och $X, Y : S \to \mathbb{R}$ vara tvao stokastiska variabler. Paret (X, Y) kallar vi för en (taodimensionell) stokastisk vektor. Naer man skriver $\mathbb{P}(X \in A, Y \in B)$ menar man sannoliheten att baode $\{X \in A\}$ och $\{Y \in B\}$ intraeffar, dvs

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(\{u \in S : X(u) \in A\} \cap \{u \in S : Y(u) \in B\}).$$

Den bivariata eller gemensamma fördelingsfunktionen för (X, Y) ges av

$$F(x, y) = F_{(X,Y)}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

Denna karaktaeriserar hela fördelningen för (X, Y), exempelvis faor man

$$\mathbb{P}(X \in [a, b], Y \in [c, d]) = F(b, d) - F(b, c) - F(a, d) + F(a, c).$$

Observera att

$$F_X(x) = \mathbb{P}(X \leq x, Y < \infty) = F(x, \infty) = \lim_{y \to \infty} F(x, y)$$

och $F_Y(y) = F(\infty, y)$. Dessa tvao kallas för (fördelningsfunktionerna för) *marginalfördelningarna* för X respektive Y.

11.1 Diskreta stokastiska vektorer

Om $V_{(X,Y)}$ är ändlig eller uppräknelig, kallar man (X, Y) för en diskret stokastisk vektor. Notera att

$$V_X = \{x : \exists y : (x, y) \in V_{(X,Y)}\}$$

och analogt för V_Y och att $V_{(X,Y)} \subseteq V_X \times V_Y$.

Den bivariata frekvensfunktionen ges av

$$p(x, y) = p_{(X,Y)}(x, y) = \mathbb{P}(X = x, Y = y),\ (x, y) \in V_{(X,Y)}.$$

Några uppenbara samband är

- $F(a, b) = \sum_{(x,y)\in V_{(X,Y)}: x\leq a, y\leq b} p(x, y)$,
- $p_X(x) = \sum_{y\in V_Y} p(x, y)$,
- $p_Y(y) = \sum_{x\in V_X} p(x, y)$.

Exempel 11.1. Slå en blå och en gul tärning. Låt X vara antalet ögon som den gula tärningen visar och låt Y vara summan antalet ögon som de två tärningarna visar. Då blir $V_X = \{1, 2, 3, 4, 5, 6\}$, $V_Y = \{2, 3, \ldots, 12\}$ och $V_{(X,Y)} = \{(x, y) \in V_X \times V_Y : x + 1 \leq y \leq x + 6\}$. Om tärningarna är symmetriska får vi

$$p(x, y) = \frac{1}{36},\ (x, y) \in V_{(X,Y)}.$$

11.2 Kontinuerliga stokastiska vektorer

Den stokastiska vektorn (X, Y) sags vara kontinuerlig om det finns en funktion $f = f_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}$, sådan att

$$F(x, y) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(x, y)dy\, dx,\ a, b \in \mathbb{R}.$$

Man kallar f för den bivariata eller gemensamma tathetsfunktionen. Generellt gäller då för alla $B \subseteq \mathbb{R}^2$ att

$$\mathbb{P}((X, Y) \in B) = \iint_B f(x, y)dx\, dy.$$

Det följer speciellt att

$$\mathbb{P}(X \in A) = \mathbb{P}((X, Y) \in A \times \mathbb{R}) = \int_A\int_{-\infty}^{\infty} f(x, y)dy\, dx$$

av vliket man bland annat ser att *marginalfördelningen* fö X har täthet

$$f_X(x) \int_{-iy}^{\infty} f(x,y)dy$$

och analogt

$$f_y(y) = \int_{-}^{\infty} f(x,y)dx.$$

Exempel 11.2. Antag att (X,Y) har täthet $c(x+3y)$, $0 \leq x,y \leq 1$. Vad är konstanten c och vad är marginalfördelningarna för X och Y?

Vi har att

$$1 = \iint_{\mathbb{R}^2} f(x,y)dx\,dy = c\int_0^1 \int_0^1 (x+3y)dy\,dx = 2c$$

så $c = 1/2$. Marginalfördelningen för X har täthet

$$f_X(x) = \int_0^1 \frac{1}{2}(x+3y)dy = \frac{1}{2}x + \frac{3}{4}$$

och för Y,

$$f_Y(y) = \int_0^1 \frac{1}{2}(x+3y)dx = \frac{1}{4} + \frac{3}{2}y.$$

Exempel 11.3. Välj en punkt på måfå i enhetsskivan $D = \{(x,y) : x^2+y^2 \leq 1\}$. Detta betyder per definition att täthäten f är konstant på D. För att $\iint_D f(x,y)dy\,dx = 1$ krävs då att

$$f(x,y) = \frac{1}{\pi}.$$

Marginalfördelningarna för X och Y är av symmetri desamma och ges av

$$f_X(x) = \frac{1}{\pi}\int_{-\sqrt{1-x^2}}^{\sqrt{1+x^2}} dy = \frac{2}{\pi}\sqrt{1-x^2},\ x \in (0,1).$$

Vi noterar att X och Y kan vara kontinuerliga sv utan att (X,Y) är kontinuerlig, låt t.ex. X vara vilken kontinuerlig sv som helst och låt $X = Y$.

12 Funktioner av flera stokastiska variabler

I detta avsnitt jobbar vi med funktioner av två sv. Alla resultat har uppenbara utvidgningar till fler än två variabler. Antag att X och Y, är två stokastiska variabler (dvs (X,Y) är en stokastisk vektor) och att $g : \mathbb{R}^2 \to \mathbb{R}$ är en funktion. Vad är fördelningen för den stokastiska variabeln $g(X,Y)$? Preis som när det gäller funktioner för en sv, finns det ingen generell formel, utan an får försöka avgöra detta från fall till fall.

Exempel 12.1. Antag att (X, Y) är likformigt fördelad på enhestkvadraten $[0, 1] \times [0, 1]$, vilket betyder att de har den gemensamma tätheten $f(x, y) = 1$, $0 \leq x, y \leq 1$. Låt A vara arean av rektangeln med sidor X och Y, dvs $A = XY$. Då är

$$F_A(a) = \mathbb{P}(XY \leq a) = \mathbb{P}((X, Y) \in B_a)$$

där $B_a = \{(x, y) : 0 \leq x, y \leq 1, xy \leq a$. Alltså

$$F_A(a) = \int_0^1 \int_0^{\min(1,a/x)} dy\, dx = \int_0^1 \min(1, a/x)dx = a - \ln(a).$$

enom att derivera ser vi att tätheten för A blir

$$f_A(a) = -\ln(a),\ 0 \leq a \leq 1.$$

Även för funktioner av två eller flera sv finns det en version av den omdevetne statistikerns lag.

Sats 12.2. *Låt $g : \mathbb{R}^2 \to \mathbb{R}$. Om (X, Y) är en diskret stokastisk vektor med bivariat frekvensfunktion p gäller att*

$$\mathbb{E}[g(X, Y)] = \sum_{(x,y)\in V_{(X,Y)}} g(x, y)p(x, y).$$

Om (X, Y) är kontinuerlig med bivariat täthet f gäller

$$\mathbb{E}[g(X, Y)] = \iint_{\mathbb{R}^2} g(x, y)f(x, y)dx\, dy.$$

Bevis. Låt oss göra det diskreta fallet. Då är

$$\mathbb{E}[g(X, Y) = \sum_{z\in g(V_{(X,Y)})} z\mathbb{P}(g(X, Y) = z)$$

$$= \sum_{z\in g(V_{(X,Y)})} z \sum_{(x,y)\in V_{(X,Y)}:g(x,y)=z} p(x, y) = \sum_{(x,y)\in V_{(X,Y)}} g(x, y)p(x, y).$$

□

En direkt följd av detta är att väntevärdet av en summa är summan av väntevärdena; sätt bara $g(x, y) = x + y$.

Sats 12.3. *Låt (X, Y) vara en diskret eller kontinuerlig stokastisk vektor. Då är*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Exempel 12.4. Enklare sätt att beräkna väntevärdet i binomialfördelningen. Låt $X \sim \text{Bin}(n, p)$. Då kan vi tänka på X som antalet klavar vid n oberoende kast, där varje kast ger klave med sannolikhet p. Vi kan då skriva

$$X = \sum_{i=1}^{n} I_{B_i}$$

där B_i är händelsen att kast nummer i ger klave. Vi vet att $\mathbb{E}[I_{B_i}] = p$ för alla i, så genom att summera ser vi att $\mathbb{E}[X] = np$.

13 Betingade fördelningar

Låt (X, Y) vara en diskret stokastisk vekotor. Då definierar vi

$$p_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x).$$

för alla $(x, y) \in V_{(X,Y)}$. Per definition av betingad sannolikhet blir detta

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)}.$$

Exempel 13.1. Låt (X, Y) ha den bivariata frekvensfunktionen $p(0, 0) = 0.1$, $p(0, 1) = 0.2$, $p(1, 0) = 0.25$, $p(1, 1) = 0.45$. Marginalfördelningarna blir $p_x(1) = 1 - p_X(0) = 0.7$ och $p_Y(1) = 1 - p_Y(0) = 0.65$. Då är $p_{Y|X}(1|0) = p(0, 1)/p_X(0) = 0.2/0.3 = 2/3$ och därmed $p_{Y|X}(0|0) = 1/3$. Vi får också till exempel $p_{X|Y}(1|1) = p(1, 1)/p_Y(1) = 0.45/0.65 = 9/13$.

Enligt totala sannolikhetslagen gäller

$$p_Y(y) = \sum_{x \in V_X} p_{Y|X}(y|x)p_X(x).$$

Exempel 13.2. Vad är sannolikheten att man får summan 9 vid slag av två tärningar?

Låt X vara värdet som tärning 1 visar och låt Y vara summan. Då är

$$p_Y(9) = \sum_{x=1}^{6} p_{Y|X}(9|x)p_X(x) = \frac{1}{6}(0 + 0 + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}) = \frac{1}{9}.$$

Kom ihåg att man säger att X och Y är oberoende om $\{X \in A\}$ och $\{Y \in B\}$ är oberoende för alla A och B, dvs om

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

för alla A och B.

Sats 13.3. *Om (X, Y) är diskret gäller att X och Y är oberoende om och endast om $p(x, y) = p_X(x)p_Y(y)$ för alla $x \in V_X$ och $y \in V_Y$.*

Bevis. Endast om-delen är trivial, så antag att $p(x, y) = p_X(x)p_Y(y)$ för alla (x, y). Då är

$$\mathbb{P}(X \in A, Y \in B) = \sum_{x\in A}\sum_{y\in B} p(x, y) = \sum_{x\in A}\sum_{y\in B} p_X(x)p_Y(y)$$

$$= \sum_{x\in A} p_X(x) \sum_{y\in B} p_Y(y) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

□

Exempel 13.4. Låt X och Y vara oberoende och båda med frekvensfunktion $p_X(j) = p_Y(j) = j/6$, $j = 1, 2, 3$. Vad blir den gemensamma frekvensfunktionen? Vad blir $\mathbb{P}(X = j|X = Y)$?

Eftersom X och Y är oberoende blir

$$p(i, j) = \frac{ij}{36}$$

for $i, j \in \{1, 2, 3\}$. Vi får också

$$\mathbb{P}(X = j|X = Y) = \frac{\mathbb{P}(X = j, Y = j)}{\mathbb{P}(X = Y)} = \frac{j^2/36}{1^2/36 + 2^2/36 + 3^2/36} = \frac{j^2}{14}.$$

Om (X, Y) är kontinuerlig, är det inte lika klart hur man ska definiera fördelningen för Y givet $X = x$ eftersom den senare är en handelse med sannolikhet 0. Man om vi gör en överslagsräkning får vi

$$\mathbb{P}(Y \in B|X \in (x, x + \Delta x)) = \frac{\mathbb{P}(Y \in B, X \in (x, x + \Delta x))}{\mathbb{P}(X \in (x, x + \Delta x))}$$

$$\approx \frac{\int_B \int_x^{x+\Delta x} f(x, y)dx\, dy}{\Delta x f_X(x)} \approx \frac{\int_B f(x, y)dy}{f_X(x)}$$

$$= \int_B \frac{f(x, y)}{f_X(x)}dy.$$

Alltså verkar det rimligt att helt enkelt definiera den *betingade tätheten* av Y givet $X = x$ som

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}.$$

Vi får då

$$\mathbb{P}(Y \in B|X = x) = \int_B f_{Y|X}(y|x)dy.$$

Observera att eftersom $f(x, y) = f_{Y|X}(y|x)f_X(x)$ blir

$$\mathbb{P}(Y \in B) = \int_B \int_{-\infty}^{\infty} f(x, y)dx\, dy = \int_{-\infty}^{\infty} f_X(x) \int_B f_{Y|X}(y|x)dy\, dx$$

$$= \int_{-\infty}^{\infty} \mathbb{P}(Y \in B|X = x)f_X(x)dx.$$

Detta är en kontinuerlig variant av den totala sannolikhetslagen; vi beräknar sannolikheter av utsagor om Y genom att betinga på X och integrerar istället för att summera. Notera också att vi har följande variant av den totala sannolikhetslagen.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx.$$

Vi ser att X och Y är oberönde om och endast om $f(x, y) = f_X(x)f_Y(y)$ för alla (x, y). Detta är ekvivalent med att $f_{X|Y}(x|y) = f_X(x)$ för alla (x, y).

Exempel 13.5. Antag att (X, Y) är likfirmigt fördelad på enhetsskivan $D = \{(x, y); x^2 + y^2 \leq 1\}$. I ett tidigare exempel såg vi att $f(x, y) = 1/\pi$ och $f_X(x) = (2/\pi)\sqrt{1 - x^2}$ för $(x, y) \in D$. Vi får då

$$f_{Y|X}(y|x) = \frac{1/\pi}{(2/\pi)\sqrt{1 - x^2}} = \frac{1}{2\sqrt{1 - x^2}}, -\sqrt{1 - x^2} \leq y \leq \sqrt{1 - x^2}$$

dvs givet $X = x$ är Y likformigt fördelad på intervallet $[-\sqrt{1 - x^2}, \sqrt{1 - x^2}]$.

Exempel 13.6. Välj X likformigt på $[0, 1]$ och sedan Y likformigt på $[0, X]$. Vad blir den bivariata tätheten?

Har är ett fall där $f_{Y|X}(y|x)$ är lättare att förstå än $f(x, y)$. Vi får

$$f(x, y) = f_X(x)f_{Y|X}(y|x) = \frac{1}{x}, 0 \leq y \leq x \leq 1.$$

Vad är sannolikheten att $Y \leq X^2$?

$$\mathbb{P}(Y \leq X^2) = \int_0^1 \mathbb{P}(Y \leq x^2|X = x)f_X(x)dx = \int_0^1 \int_0^{x^2} \frac{1}{x}dy\, dx$$

$$= \int_0^1 xdx = \frac{1}{2}.$$

14 Oberoende stokastiska variabler och stora talens lag

Proposition 14.1. *Om X och Y är oberoende stokastiska variabler, gäller*

(a) $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$,

(b) $\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]$.

Bevis. Antag att X och Y är kontinuerliga. Det diskreta fallet är analogt. Det gäller enligt den omedvetne statistikerns lag att

$$\mathbb{E}[XY] = \iint xyf(x,y)dx\,dy = \iint xyf_X(x)f_Y(y)dx,\,dy$$

$$= \int xf_X(x)dx \int yf_Y(y)dy = \mathbb{E}[X]\mathbb{E}[Y].$$

För del (b), använd (a) till att få

$$\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{E}[(X+Y-\mathbb{E}[X]-\mathbb{E}[Y])^2]$$

$$= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$

$$= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] + 2\mathbb{E}[X-\mathbb{E}[X]]\mathbb{E}[Y-\mathbb{E}[Y]]$$

$$= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y].$$

□

Naturligtvis generaliserar sig dessa resultat till att om $X_1, X_2, \ldots, X_n$ är oberoende, så gäller

$$\mathbb{E}[X_1X_2\cdots X_n] = \mathbb{E}[X_1\mathbb{E}[X_2]\cdots\mathbb{E}[X_n]$$

och

$$\mathbb{V}\mathrm{ar}[X_1+\ldots+X_n] = \mathbb{V}\mathrm{ar}[X_1]+\ldots+\mathbb{V}\mathrm{ar}[X_n].$$

Exempel 14.2. Antag att $X \sim \mathrm{Bin}(n,p)$. Då kunde man skriva $X = \sum_{i=1}^{n} I_{B_i}$ där de n indikatorfunktionerna är oberoende. Vi ser att

$$\mathbb{V}\mathrm{ar}[X] = n\mathbb{V}\mathrm{ar}[I_{B_1}] = np(1-p),$$

ty $\mathbb{V}\mathrm{ar}[I_{B_1}] = p - p^2 = p(1-p)$.

Exempel 14.3. Låt $X_1, X_2, \ldots, X_n$ vara oberoende och likafördelade med varians σ^2 och väntevärde μ. Låt $\overline{X}$ vara medelvärdet

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Då är $\mathbb{E}[\overline{X}] = \mu$ och

$$\mathbb{V}\mathrm{ar}[\overline{X}] = \frac{1}{n^2} n \mathbb{V}\mathrm{ar}[X_1] = \frac{\sigma^2}{n}.$$

Från det sista exemplet följer

Sats 14.4. Stora talens lag (den svaga formen).
Låt $X_1, X_2, \ldots$ vara obeoroende och likafördelade med väntevärde μ och varians $\sigma^2 < \infty$. Skriv $\overline{X}_n$ för medelvärdet av de n första X_i:na. Det gäller för alla $\epsilon > 0$ att

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \to 0$$

då $n \to \infty$.

Bevis. Enligt Chebyshevs olikhet är

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = \mathbb{P}((\overline{X}_n - \mu)^2 > \epsilon^2) \leq \frac{\mathbb{V}\mathrm{ar}[\overline{X}_n]}{\epsilon^2}$$

$$= \frac{\sigma^2}{n\epsilon^2} \to 0.$$

□

15 Betingade väntevärden

Om (X, Y) är ett par av diskreta sv, definierar vi $\mathbb{E}[Y|X = x]$ som väntevärdet av den sv som har samma fördelning som Y givet $X = x$, dvs

$$\mathbb{E}[Y|X = x] = \sum_{y \in V_Y} y\mathbb{P}(Y = y|X = x) = \sum_{y \in V_Y} y p_{Y|X}(y|x).$$

Om (X, Y) är kontinuerlig definierar vi

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy.$$

Följande resultat är den totala sannolikhetslagen för betingade väntevärden.

Proposition 15.1. *Om (X, Y) är diskret gäller*

$$\mathbb{E}[Y] = \sum_{x \in V_X} \mathbb{E}[Y|X = x]\mathbb{P}(X = x).$$

Om (X, Y) är kontinuerlig gäller

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x] f_X(x) dx.$$

Bevis. Vi gör det kontinuerliga fallet. Då har vi att integralen i högerledet är

$$\int_{-\infty}^{\infty} \mathbb{E}[Y|X = x] f_X(x) dx = \int \int y f_{Y|X}(y|x) dy \, f_X(x) dx = \int y \int f(x, y) dx \, dy$$

$$\int y f_Y(y) dy = \mathbb{E}[Y].$$

□

Exempel 15.2. Låt X vara likformig på $[0, 1]$ och sedan Y vara likformig på $[0, X]$. Vi såg tidigare att $f(x, y) = \frac{1}{x}$, $0 \leq y \leq x \leq 1$, så

$$f_Y(y) = \int f(x, y) dy = \int_y^1 \frac{1}{x} dx = -\ln(y).$$

Därför blir

$$\mathbb{E}[Y] = -\int_0^1 y \ln(y) dy = \frac{1}{2} \int_0^1 y dy = \frac{1}{4}.$$

Enklare blir det dock med totala sannolikhetslagen, ty då får vi att $\mathbb{E}[Y|X = x] = \frac{x}{2}$, så

$$\mathbb{E}[Y] = \frac{1}{2} \int_0^1 x dx = \frac{1}{4}.$$

Vi definierar nu $\mathbb{E}[Y|X]$ som den stokastiska variabel som har värdet $\mathbb{E}[Y|X = x]$ då $X = x$. Med andra ord är $\mathbb{E}[Y|X = x]$ en funktion av X; om vi låter $g(x) = \mathbb{E}[Y|X = x]$ så är $\mathbb{E}[Y|X] = g(X)$. Antag att (X, Y) är kontinuerlig. Då har vi

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[g(X)] = \int g(x) f_X(x) dx$$

$$= \int \mathbb{E}[Y|X = x] f_X(x) dx = \mathbb{E}[Y].$$

Detta funkar lika bra i det diskreta fallet, så vi har det generella resultatet

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y].$$

Exempel 15.3. Antag att $X \sim \text{Geo}(p)$ och sedan $Y \sim \text{Bin}(X, r)$. Då är

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Xr] = \frac{r}{p}.$$

Exempel 15.4. Låt N vara en ickenegativ heltalsvärd sv, oberoende av de stokastiska variablerna $X_1, X_2, \ldots$, daer $\mathbb{E}[X_i] = \mu$ för alla i. Då får vi

$$\mathbb{E}[\sum_{i=1}^{N} X_i] = \mathbb{E}[\mathbb{E}[\sum_{i=1}^{N} X_i|N]] = \mathbb{E}[N\mu] = \mu\mathbb{E}[N].$$

16 Kovarians och korrelation

Om X och Y aer två sv, aer $\mathbb{C}\text{ov}(X, Y)$ ett mått på hur X och Y samverkar. Definitionen aer att

$$\mathbb{C}\text{ov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Analogt med Steiners formel for varians, finns formeln

$$\mathbb{C}\text{ov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Om $\mathbb{C}\text{ov}(X, Y) > 0$ brukar man saega att X och Y aer positivt korrelerade och om $\mathbb{C}\text{ov}(X, Y) < 0$ talar man om negativt korrelation. Vi ser att om X och Y aer oberoende, aer $\mathbb{C}\text{ov}(X, Y) = 0$. Det aer dock *inte sant* att $\mathbb{C}\text{ov}(X, Y) = 0$ medför oberoende. Låt t.ex. ϕ vara likformig på $[0, 2\pi]$ och ta $X = \cos\phi$ och $Y = \sin\phi$. Då aer $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ och

$$\mathbb{E}[XY] = \int_0^{2\pi} \cos t \sin t dt = 0.$$

Alltså aer $\mathbb{C}\text{ov}(X, Y) = 0$, men det två stokastiska variablerna aer uppenbart inte oberoende.

En uppenbar egenskap hos kovariansen aer att den aer bilinjaer, dvs

$$\mathbb{C}\text{ov}(X, a_1Y_1 + a_2Y_2) = a_1\mathbb{C}\text{ov}(X, Y_1) + a_2\mathbb{C}\text{ov}(X, Y_2).$$

Andra uppenbara observationer aer att $\mathbb{V}\text{ar}(X) = \mathbb{C}\text{ov}(X, X)$ och att $\mathbb{C}\text{ov}(X, Y) = \mathbb{C}\text{ov}(Y, X)$. Att laegga till konstanter till X eller Y föraendrar inte kovariansen: $\mathbb{C}\text{ov}(X + a, Y + b) = \mathbb{C}\text{ov}(X, Y)$. Följande anvaendbara formel följer av dessa enkla observationer

$$\mathbb{V}\text{ar}[X + Y] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y] + 2\mathbb{C}\text{ov}(X, Y).$$

Vi observerar att mittermen aer 0 då X och Y aer oberoende.

Exempel 16.1. Låt X och Y vara oberoende och likformiga på $[0,1]$. Låt $Z = XY$ och $W = X + Y$. Vi får

$$\mathbb{C}\mathrm{ov}(W,Z) = \mathbb{E}[XY(X+Y)] - \mathbb{E}[XY]\mathbb{E}[X+Y] = \mathbb{E}[X^2Y] + \mathbb{E}[XY^2] - \mathbb{E}[XY]$$
$$= 2\mathbb{E}[X^2]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Det aer ingen overraskning att dessa två sv aer positivt korrelerade.

Sats 16.2. Schwarz olikhet. *För alla stokastiska variabler X och Y gaeller att* $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$.

Bevis. För alla $t \in \mathbb{R}$ gaeller att

$$0 \leq \mathbb{E}[(X - tY)^2] = \mathbb{E}[X^2] - 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2].$$

Minimering m.a.p. t ger att $t = \mathbb{E}[XY]/\mathbb{E}[Y^2]$ och saetter man in detta i uttrycket får man

$$0 \leq \mathbb{E}[X^2] - 2\frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} + \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} = \mathbb{E}[X^2] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}$$

som önskat. □

En direkt konsekvens av Schwarz olikhet tillaempad på $X - \mathbb{E}[X]$ och $Y - \mathbb{E}[Y]$ ger

$$\mathbb{C}\mathrm{ov}(X,Y)^2 \leq \mathbb{V}\mathrm{ar}[X]\mathbb{V}\mathrm{ar}[Y].$$

Definition 16.3. Korrelationsköfficienten för X och Y ges av

$$\rho(X,Y) = \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\sqrt{\mathbb{V}\mathrm{ar}[X]\mathbb{V}\mathrm{ar}[Y]}}.$$

Enligt vad vi just såg gaeller att $-1 \leq \rho(X,Y) \leq 1$. Det galler också att om X och Y aer oberoende, medför detta att $\rho(X,Y) = 0$. Man kan också visa att $\rho(X,Y) = 1$ om och endast om $Y = aX + b$ för något $a > 0$ och $\rho(X,Y) = -1$ om och endast om $Y = aX + b$ för något $a < 0$.

Om $\rho(X,Y) = 0$, kallas X och Y okorrelerade. Vi har tidigare sett att detta inte nödvaendigtvis medför att X och Y aer oberoende.

Exempel 16.4. Låt oss fortsaetta på det förra exemplet. Vi hade att $\mathbb{C}\mathrm{ov}(W,Z) = 1/12$. Vi har också att $\mathbb{V}\mathrm{ar}[W] = 2\mathbb{V}\mathrm{ar}[X] = 1/6$ och

$$\mathbb{V}\mathrm{ar}[Z] = \mathbb{E}[X^2Y^2] - \mathbb{E}[XY]^2 = \mathbb{E}[X^2]\mathbb{E}[Y^2] - (\mathbb{E}[X]\mathbb{E}[Y])^2 = \frac{1}{9} - \frac{1}{16} = \frac{7}{144}.$$

Således blir

$$\rho(X,Y) = \frac{1/12}{\sqrt{(1/6)\cdot(7/144)}} = \sqrt{6/7}.$$

En liten korrelationsköfficient betyder som vi sett inte att beroendet mellan X och Y aer svagt (aeven som så oftast aer fallet). Daeremot betyder en till beloppet stor köfficient ett starkt beroende. Ett saett att illustrera detta aer att se på vad som aer baesta linjaera prediktor av Y givet X: vilken linjaer funktion $aX+b$ aer baesta prediktor av Y i meningen att $\mathbb{E}[(Y-(ax+b))^2]$ minimeras? Antag att både X och Y har vaentevaerde 0 och varians 1. Då aer $\mathbb{C}\mathrm{ov}(X,Y) = \mathbb{E}[XY] = 1$, så

$$\mathbb{E}[Y^2] - 2\mathbb{E}[Y(aX+b)] + \mathbb{E}[(aX+b)^2] = 1 - 2\rho a + a^2 + b^2$$

vilket minimeras för $a = \rho$ och $b = 0$. Baest linjaera prediktorn aer alltså $Y = \rho X$. Tillaempar vi nu detta på $(X-\mu_X)/\sigma_X$ och $(Y-\mu_Y)/\sigma_Y$ i det generella fallet med vaentevaerden μ_X och μ_Y och varianser σ_X^2 och σ_Y^2, får vi att den baesta linjaera prediktorn av Y givet X aer

$$\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X).$$

17 Summor av oberoende stokastiska variabler

Låt X och Y vara oberoende och $Z = X+Y$. Vad aer fördelningen för Z? Antag att X och Y är kontinuerliga. Då gaeller enligt totala sannolikhetslagen,

$$F_Z(x) = \mathbb{P}(X+Y \le x) = \int_{-\infty}^{\infty} \mathbb{P}(t+Y \le x | X = t) f_X(t)dt$$

$$= \int_{-\infty}^{\infty} F_Y(x-t) f_X(t)dt$$

Genom att derivera innanför integraltecknet (vilket man kan visa aer OK under enkla villkor på f_X och f_Y) följer

$$f_Z(x) = \int_{-\infty}^{\infty} f_Y(x-t) f_X(t).$$

Högersidan, som alltså aer taetheten av $X+Y$, kallas också för *faltningen* av f_X och f_Y (från tyskans ”faltung”. På engelska aer ordet ”convolution”.) Notera att om både X och Y är ickenegativa, är faltningen

$$f_Z(x) = \int_0^x f_Y(x-t) f_X(t)dt.$$

Om X och Y är diskreta får vi på samma sätt för alla $k \in V_{X+Y}$ att

$$p_{X+Y}(k) = \sum_{x: x\in V_X, k-x\in V_Y} p_Y(k-x)p_X(x).$$

Exempel 17.1. Låt X och Y vara värdena av två tärningsslag. Då får vi t.ex. att

$$p_{X+Y}(9) = \sum_{j=1}^{6} p_Y(9-j)p_X(j) = \frac{4}{36} = \frac{1}{9}.$$

Exempel 17.2. Låt $X \sim \text{Poi}(\lambda_1)$ och $Y \sim \text{Poi}(\lambda_2)$ vara oberoende. Eftersom dessa ofta modellerar antalen av några ovanliga handelser och är oberoende, blir summan antalet ovanliga händelser av de två sorterna tillsammans. Detta borde betyda att $X + Y$ också är Poisson, med parameter $\lambda_1 + \lambda_2$. Vi kollar

$$p_{X+Y}(k) = \sum_{j=0}^{k} p_Y(k-j)p_X(j) = \sum_{j=0}^{k} e^{-\lambda_2}\frac{\lambda_2^{k-j}}{(k-j)!}e^{-\lambda_1}\frac{\lambda_1^j}{j!}$$

$$= e^{-(\lambda_1+\lambda_2)}\frac{1}{k!}\sum_{j=0}^{k}\frac{k!}{(k-j)!j!}\lambda_1^j\lambda_2^{k-j} = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!}$$

precis som önskat, där den sista likheten följer av binomialsatsen.

Exempel 17.3. Om X och Y är oberoende och normalfördelade med vaentevaerdesparametrar parametrar μ_1 repsektive μ_2 och variansparametrar σ_1^2 respektive σ_2^2 gäller att $X + Y$ are oberoende med parametrar $\mu_1 + \mu_2$ $\sigma_1^2 + \sigma_2^2$. Att kontrollera detta i det allmänna fallet blir en bökig räkning, men låt oss göra det i fallet $\mu_1 = \mu_2 = 0$ och $\sigma_1^2 = \sigma_2^2 = 1$. Vi får då via kvadratkomplettering i exponenten, att

$$f_{X+Y}(x) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x-t)^2/2}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-(t-1/2)^2-x^2/4}dt$$

$$= \frac{1}{\sqrt{4\pi}}e^{-x^2/4}\int_{-\infty}^{\infty}\frac{1}{\sqrt{\pi}}e^{-(t-1/2)^2}dt = \frac{1}{\sqrt{4\pi}}e^{-x^2/4}$$

eftersom uttrycket under den sista integralen är tätheten för en $\text{N}(1/2, 1/2)$-fördelad sv.

18 Poissonprocessen

Vissa ”incidenter”, eller *impulser* inträffar då och då i tiden. Det kan handla om jördbävningar, bilar som passerar på en enslig väg, mål i en fotbolsmatch, etc. Denna typ av processer är ofta rimliga att modellera med *Poissonprocessen*. En Poissonprocess definieras av att de tidpunkter $\tau_1, \tau_2, \ldots$ då impulser inträffar är

sådana att $T_1 = \tau_1$, $T_2 = \tau_2 - \tau_1$, $T_3 = \tau_3 - \tau_2 \ldots$ är oberoende och exponentialfördelade med samma parameter λ. Denna parameter kallas för Poissonprocessens *intensitet*. Man kan visa att för en sådan process gäller att om $X(s,t)$ är antalet incidenter i tidsintervallet (s,t), så är antalet impulser i disjunkta intervall oberoende och $X(s,t)$ är Poissonfördelad med parameter $\lambda(t-s)$. Tiden τ_n då den n:te impulsen inträffar, är Gammafördelad med parametrar n och λ. För att visa detta, notera först att påståendet är sant för $n = 1$ och använd induktion och totala sannolikhetslagen till att visa att

$$\mathbb{P}(\tau_n > x) = e^{-\lambda x} \sum_{j=0}^{n-1} \frac{(\lambda x)^j}{j!}$$

som önskat. Detta finns som uppgift i Matlabhäftet. Av detta kan man direkt visa att $X(t) = X(0,t)$ är Poissonfördelad, ty

$$\mathbb{P}(X(t) = k) = \mathbb{P}(\tau_{k+1} > t) - \mathbb{P}(\tau_k > t) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}.$$

Ett annat viktigt faktum är *sammanvägningsegenskapen*: Om $X(t)$ och $Y(t)$ är oberoende Poissonprocesser med intensitet λ_1 respektive λ_2, är $X(t) + Y(t)$ en Poissonprocess med intensitet $\lambda_1 + \lambda_2$. Detta följar av att summan av två oberoende Poisson sv är en Poisson sv, vilket vi nyligen sett. Notera att detta är ekvivalent med det faktum att om $X \sim \exp(\lambda_1)$ och $Y \sim \exp(\lambda_2)$ är oberoende, så är $\min(X,Y) \sim \exp(\lambda_1 + \lambda_2)$. Detta påstående kan också lätt bevisas direkt. Gör gärna det.

Dessutom gäller *uttunningsegenskapen*: Om $X(t)$ är in Poissonprocess med intensitet λ och $Y(t)$ ges av att medräkna varje impuls för X med sannolikhet p, oberoende för de olika impulserna. Då är $Y(t)$ en Poissonprocess med intensitet λp.

19 Momentgenererande funktion och centrala gränsvärdessatsen

Man definierar den *momentgenererande funktionen* (mgf) till en sv X, som

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Till exempel bli mgf till kontinuerliga fördelningar

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx.$$

I det diskreta fallet får vi

$$M_X(t) = \sum_{x \in V_X} e^{tx} p_X(x).$$

Vi noterar likheten med Laplacetransfomen; om $X \geq 0$ är $M_X(t) = \mathcal{L}_{f_X}(-t)$. I ljuset av den observationen är det ingen överraskning att om man känner till $M_X(t)$ för alla s i en omgivning till 0, så känner man också till hela X:s fördelning.

Proposition 19.1. *Om X och Y är oberoende, är*

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Bevis. Tack vare oberoendet gäller

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = M_X(t) M_Y(t).$$

□

Man brukar kalla $\mathbb{E}[X^n]$ för det n:te *momentet* till X. Detta förklarar varifrån den momentgenererande funktionen fått sitt namn. Se nämligen på

$$M'_X(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}[X e^{tX}]$$

förutsatt att vi får derivera innenför vantevärdet, vilket man kan visa går bra att göra. Vi ser att om vi tar $t = 0$, får vi

$$M'_X(0) = \mathbb{E}[X].$$

Om man fortsätter att derivera ser man att generellt gäller

$$M_X^{(n)}(0) = \mathbb{E}[X^n].$$

En fördel med mgf är att det ibland går betydligt lättare att finna vad en fördelning är via den än på annat sätt.

Exempel 19.2. Låt $X \sim \mathrm{N}(\mu, \sigma^2)$. Då är, med hjälp av kvadratkomplettering,

$$M_X(t) = \mathbb{E}[e^{tX}] = \int e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

$$= e^{\mu t + \sigma^2 t^2/2} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-(\mu+\sigma^2 t))^2}{\sigma^2}} dx$$

$$= e^{\mu t + \sigma^2 t^2/2}$$

där den sista likheten följer av att uttrycket i integralen är tätheten för N($\mu + \sigma^2 t, \sigma^2$).

Om nu X och Y är oberoende och normalfördelade med parametrar μ_X, σ_X^2, μ_Y och σ_Y^2 ser vi att

$$M_{X+Y}(t) = e^{\mu_X t + \sigma_X^2 t^2/2} e^{\mu_Y t + \sigma_Y^2 t^2/2} = e^{(\mu_X + \mu_Y)t + (\sigma_X^2 + \sigma_Y^2)t^2/2}$$

varur det följer att $X + Y$ är normalfördelad med väntevärde $\mu_X + \mu_Y$ och varians $\sigma_X^2 + \sigma_Y^2$.

En viktig generalisering av det faktum att mgf bestämmer fördelningen är att om $X_1, X_2, \ldots$ och X är kontinuerliga sv galler att om $M_{X_n}(t) \to M_X(t)$ för alla t medför att $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ för alla x.

Här följer ett av sannolikhetsteorins viktigaste resultat.

Sats 19.3. Centrala gransvärdessatsen.

Låt $X_1, X_2, X_3, \ldots$ vara oberoende och likafördelade med $\mu = \mathbb{E}[X_1]$ och $\sigma^2 = \mathbb{V}\mathrm{ar}[X_1]$. Skriv $S_n = \sum_{i=1}^n X_i$. Då gäller för alla x att

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x)$$

då $n \to \infty$.

Med andra ord: en summa av många oberoende likafördelade sv är ungefär normalfördelad med parametrar $n\mu$ och $n\sigma^2$, *oavsett* vilken fördelning våra sv har i övrigt. Mgf låter oss ge en bevisskiss. För att förenkla, låt oss anta att $\mu = 0$ och $\sigma^2 = 1$. Vi vill visa att mgf för $S_n/\sqrt{n}$ går mot $M_{N(0,1)}(t) = e^{t^2/2}$. Det gäller att

$$M_{S_n/\sqrt{n}}(t) = \mathbb{E}[e^{tS_n/\sqrt{n}}] = M_{S_n}(t/\sqrt{n}) = M_X(t/\sqrt{n})^n.$$

Taylorutveckla $M_X(s)$ kring 0 (förutsatt att detta går bra) och få

$$M_X(s) \approx M_X(0) + sM_X'(0) + \frac{1}{2}s^2 M_X''(0)$$

$$= 1 + s\mu + \frac{1}{2}s^2(\mu^2 + \sigma^2) = 1 + \frac{1}{2}s^2.$$

Detta betyder att

$$M_{s_n/\sqrt{n}}(t) \approx (1 + \frac{t^2}{2n})^n \to e^{t^2/2}$$

som önskat.

Exempel 19.4. Antag att det per år i världen inträffar i genomsnitt 100 ”stora” jordbävningar (över en viss magnitud). Vad är en approximativ sannolikhet att det ett givet år ske mer än 110 sådana jordbävningar?

Det är rimligt att anta att de stora jordbävningarna kommer som en Poissonprocess, dvs tiderna $T_1, T_2, \ldots$ mellan dem är oberoende och exp(100)-fördelade. Antalet jordbavningar på ett år är då Poissonfördelat med parameter 100 och vi vill veta sannolikheten att det blir fler än 110 av dem. Detta blir en bökig uträkning. Ett alternativ är att observera att $T = T_1 + \ldots + T_{110}$ är approximativt normalfördelad. Eftersom de enskilda T_i:na har väntevärde $1/100$ och varians $1/10000$ blir T cirka normalfördeled med väntevärde 1.1 och varians 0.011. Alltså blir

$$\mathbb{P}(T < 1) = \mathbb{P}(\frac{T - 1.1}{\sqrt{0.011}} < \frac{1 - 1.1}{\sqrt{0.011}}) \approx \Phi(-\frac{0.1}{\sqrt{0.011}}) \approx 0.17.$$

Exempel 19.5. Slå tärning 700 gånger. Vad är chansen att få minst 100 sexor?

Om X är antalet sexor, är $X \sim \text{Bin}(700, 1/6)$, så

$$\mathbb{P}(X \geq 100) = \sum_{k=100}^{700} \binom{700}{k} (\frac{1}{6})^k (\frac{5}{6})^{700-k}$$

vilket är en jobbig uträkning. Men

$$X = I_1 + I_2 + \ldots + I_{700}$$

där I_k är indikatorn att man får sexa i kast nummer k. Dessa indikatorer är oberoende och har vantevärde $1/6$ och varians $(1/6)(5/6) = 5/36$. Enligt centrala gränsvärdessatsen är X alltså approximativt $\text{N}(700/6, 3500/36) = \text{N}(350/3, 875/9)$. Detta ger

$$\mathbb{P}(X \geq 100) = \mathbb{P}(\frac{X - 350/3}{\sqrt{875/9}} \geq \frac{100 - 350/3}{\sqrt{875/9}})$$

$$= 1 - \Phi(-\frac{50}{\sqrt{875}}) = \Phi(\frac{50}{\sqrt{875}}) \approx 0.955.$$

I det nyss avslutade exemplet såg vi att för stort n är $\text{Bin}(n, p)$-fördelningen ungefär normal. Genom att resonera som i exemplet dessförinnan, ser man att då λ är stort är Poissonfördelningen också ungefär normal.

Centrala gränsvärdessatsen fungerar i många fall även utan antagandet att de stokastiska variablerna är likafördelade. Man kan också ibland släppa villkoret på oberoende, men då med en justerad variansparameter. Sammantaget kan man säga att alla storheter som är summor av många små bidrag tenderar att vara approximativt normalfördelade. Det är då viktigt att komma ihåg att det är just approximativa normalfördelningar vi får och att approximationen ofta är dålig långt ute i svansarna. Med andra ord, en mycket stor eller liten sannolikhet uträknad via centrala gransvärdessatsen är ofta en dålig approximation.

20 Statistikteori

Lite löst, kan man säga att skillnaden mellan sannolikhetsteori och statistikteori är att i den förra beräknar man sannolikheter utifrån givna parametrar, medan man i det senare fallet jobbar ”baklänges” i det att man utifrån observationer försöker uppskatta parametrarna.

Definition 20.1. Om $X_1, X_2, \ldots, X_n$ är oberoende och alla fördelade som X, kallas $(X_1, \ldots, X_n)$ för ett stickprov på X (eller på X:s fördelning).

Antag att X:s fördelning beror av en parameter θ, t.ex. $X = 0$ med sannolikhet $1 - \theta$ och 1 med sannolikhet θ eller $X \sim \text{Poi}(\theta)$, eller ...

En funktion $\hat{\theta} = \hat{\theta}_n$ av $X_1, \ldots, X_n$ kallas för en *punktskattning* av θ. Observera att $\hat{\theta}$ är en stokastisk variabel, men efter man fått värdet på $X_1, \ldots, X_n$, dvs när stickprovet *realiserats*, får man förstås ett givet värde, ett *estimat* av θ.

Definition 20.2. Om $\mathbb{E}[\hat{\theta}] = \theta$ för alla θ, kallas $\hat{\theta}$ för en *väntvärdesriktig* (vvr) skattning av θ.

Definition 20.3. Om $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ då $n \to \infty$, kallas $\hat{\theta}_n$ för en *konsistent* skattning av θ.

Proposition 20.4. *Om $\hat{\theta}_n$ är vvr och $\mathbb{V}\text{ar}[\hat{\theta}_n] \to 0$, är $\hat{\theta}_n$ konsistent.*

Beviset är en direkt tillämpning av Chebyshevs olikhet.

Exempel 20.5. Antag att X har väntevärde μ och varians $\sigma^2 < \infty$ (dvs σ^2 är eventuellt okänd men åtminstone ändlig) och att $X_1, \ldots, X_n$ är ett stickprov på X. Låt

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

vara medelvärdet av observationerna. Då är $\mathbb{E}[\overline{X}] = \mu$ och $\mathbb{V}\text{ar}[\overline{X}] = \sigma^2/n \to 0$. Det foljer att medelvärdet är en konsistent skattning av väntevärdet för alla X med ändlig varians.

Om nu $\hat{\theta}$ är en konsistent skattning, är allt bra då? Nej, man ska snarare se konsistens om ett minimikrav på en bra skattning och vvar som önskvärt.

Definition 20.6. Om hth och $\overline{\theta}$ är två punktskattningar av θ och det för alla θ gäller att $\mathbb{V}\text{ar}[\hat{\theta}] < \mathbb{V}\text{ar}[\overline{\theta}]$, kallas $\hat{\theta}$ *mer effektiv* än $\overline{\theta}$.

Se gärna på Sats 6 i boken om Fisherinformationen.

21 Skattning av varians

Låt $X_1, \ldots, X_n$ vara ett stickprov på X med $\mathbb{E}[X] = \mu$ och $\mathbb{V}\mathrm{ar}[X] = \sigma^2$, båda okända Man brukar skatta σ^2 med

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

(Om μ är känd, använd skattningen $(1/n)(X_i - \mu)^2$.) Varför $n-1$ i nämnaren? Det beror på att $\overline{X}$ i sig är en skattning, vilket gör att kvadratsumman tenderar att bli mindre.

Proposition 21.1. *Skattningen s^2 är vvr för σ^2. Om också $\mathbb{E}[X^4] < \infty$ är den även konsistent.*

Bevis. Det är lätt att se att

$$s^2 = \frac{1}{n-1} (\sum_{i=1}^{n} X_i^2 - n\overline{X}^2).$$

Vi har $\mathbb{E}[\overline{X}^2] = \sigma^2/n + \mu^2$, så

$$\mathbb{E}[s^2] = \frac{1}{n-1} (n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)) = \sigma^2.$$

□

22 Konfidensintervall

Låt $X_1, \ldots, X_n$ vara ett stickprov och T_1 och T_2 två funktioner av data sådana att

$$\mathbb{P}(T_1 \leq \theta \leq T_2) = q.$$

Då kallas $[T_1, T_2]$ för ett *konfidensintervall* för θ med *konfidensgrad* q. Man skriver ofta

$$T_1 \leq \theta \leq T_2 \ (q)$$

Man ska observera att sannolikheten att intervallet täcker θ gäller *innan* data realiserats. Om man fått $T_1 = t_1$ och $T_2 = t_2$ är påståendet $t_1 \leq \theta \leq t_2$ antingen sant eller falskt, men vi vet inte vilket (θ är ju en parameter och ingen stokastisk variabel).

Exempel 22.1. Låt $X \sim \text{likf}[0, \theta]$. Vi har $\mu = \theta/2$, så $2X$ är en vvr skattning av θ. Men å andra sidan vet vi ju att $\theta \geq X_{(n)} = \max_i X_i$. Vi har

$$\mathbb{P}(X_{(n)} \leq x) = (\frac{x}{\theta})^n$$

så $X_{(n)}$ har täthet nx_{n-1}/θ^n, så

$$\mathbb{E}[X_{(n)}] = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$

Därmed är även $\frac{n+1}{n} X_{(n)}$ en vvr skattning av θ. Man ser på samma sätt att

$$\mathbb{E}[X_{(n)}^2] = \frac{n}{n+2}\theta^2.$$

Därmed är

$$\mathbb{V}\text{ar}[X_{(n)}] = (\frac{n}{n+2} - \frac{n^2}{(n+1)^2})\theta^2 \approx \frac{1}{n^2}\theta^2.$$

Jämfört med $2\overline{X}$ som har en varians som är av storleksordning θ^2/n är detta betydligt bättre.

Låt oss göra ett 95% konfidensintervall för θ baserat på $X_{(n)}$, dvs finna T_1 och T_2 som funktioner av $X_{(n)}$ sådana att $\mathbb{P}(T_1 \leq \theta \leq T_2) = 0.95$. Då kan vi till exempel välja så att $\mathbb{P}(T_1 > \theta) = \mathbb{P}(T_2 < \theta) = 0.025$. Vi har $P(X_{(n)} \leq x) = (x/\theta)^n$, vilket är 0.025 då $x = 0.025^{1/n}\theta$ och 0.975 då $x = 0.975^{1/n}\theta$. Vi får

$$\frac{X_{(n)}}{0.975^{1/n}} \leq \theta \leq \frac{X_{(n)}}{0.025^{1/n}} \ (95\%).$$

I boken har vi ett fall där $n = 10$ och $X_{(n)} = 8.69$. Med denna information er övriga data ointressanta. Vi får

$$8.71 \leq \theta \leq 12.57$$

med konfidensgrad 95%.

23 Konfidensintervall för μ i normalfördelningen

Antag att $X_1, \ldots, X_n$ är ett stickprov på en $\text{N}(\mu, \sigma^2)$-fördelning med känd varians men okänt μ. Man kan visa att den mest effektiva skattningen av μ är $\overline{X}$, så det verkar bra att basera sitt konfidensintervall på $\overline{X}$. Vi utnyttjar att $\sqrt{n}(\overline{X} - \mu)/\sigma$ är standardnormal, så vi får som symmetriskt konfidensintervall

$$-z \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z$$

med sannolikhet q om $z = \Phi^{-1}((1+q)/2)$. Genom att möblera om uttrycket ser vi att

$$\mu = \overline{X} \pm z\frac{\sigma}{\sqrt{n}}\ (q).$$

Om man vill ha ett t.ex. nedåt begränsat konfidensintervall, tas $z = \Phi^{-1}(q)$ och vi får

$$\mu \geq \overline{X} - z\frac{\sigma}{\sqrt{n}}.$$

och ett uppåt begränsat konfidensintervall med konfidensgrad q ges av

$$\mu \leq \overline{X} + z\frac{\sigma}{\sqrt{n}}.$$

Nu är det ju i de allra flesta situationer att även σ^2 är okänd. Då ersätter vi σ^2 med s^2 och betraktar kvoten $\sqrt{n}(\overline{X} - \mu)/s$ som nu inte är normalfördelad utan i stället har en s.k. t-fördelning med $n-1$ frihetsgrader. Denna fördelning finns i tabeller. Den är liksom standardnormalfördelningen symmetrisk kring 0, så ett symmetriskt konfidensintervall blir

$$\mu = \overline{X} \pm z\frac{s}{\sqrt{n}}\ (q)$$

där $z = F^{-1}_{t_{n-1}}((1+q)/2)$. Ett nedåt begränsat konfidensintervall fås som

$$\mu = \overline{X} - z\frac{s}{\sqrt{n}}\ (q)$$

där z nu är $F^{-1}_{t_{n-1}}(q)$ och motsvarande för ett uppåt begränsat intervall.

Exempel 23.1. Man mäter nätspänningen i ett vägguttag vid sju olika tillfällen och fick följande (i volt)

$$230.7\ 226.9\ 228.8\ 232.2\ 227.3\ 227.0\ 229.1$$

Avvikelserna från 230 volt är uppbyggda av många små saker utanför vår kontroll, så det är rimligt att anta att mätdata var normalfördelade med okänt väntevärde och okänd varians, μ respektive σ^2. Vi har $\overline{X} = 228.9$, $s^2 = 2.014^2$ och $n = 7$. Vi slår upp i tabell för t_6-fördelningen att $z = F^{-1}_{t_6}(0.975) = 2.46$, så ett symmetriskt konfidensintervall av konfidensgrad 95% är

$$\mu = 228.9 \pm 2.46\frac{2.014}{\sqrt{7}} \approx 228.9 \pm 1.9.$$

Om vi hade antagit att vi kände till att $\sigma^2 = 2^2 = 4$ skulle vi få

$$\mu = 228.9 \pm 1.96\frac{2}{\sqrt{7}} = 228.9 \pm 1.5.$$

Vi anar att då $n \to \infty$ gäller att $F_{t_n}^{-1}(a) \to \Phi^{-1}(a)$. Man kan ta som tumregel att då $n \geq 100$ gäller likhet här.

Det är naturligtvis också intressant att göra skattningar och konfidensintervall för σ^2. Detta kan man göra uitfrån det faktum att $(n-1)s^2/\sigma^2$ är χ^2_{n-1}-fördelad. Här ska vi dock nöja oss med att punktskatta σ^2 med s^2.

24 Prediktion i normalfördelningen

Antag att σ^2 är känd, μ okänd och att vi har stickprovet $X_1, \ldots, X_n$ och vill förutsäga vad nästa observation, $Y = X_{n+1}$, blir. Nu är ju $Y - \overline{X}$ normalfördelad med vantevärde 0 och varians $\sigma^2 + \sigma^2/n$, dvs

$$\frac{Y - \overline{X}}{\sigma(1 + 1/n)} \sim \text{N}(0, 1).$$

Om vi låter $z = \Phi^{-1}((1+q)/2)$ får vi ett symmetriskt prediktionsintervall

$$Y = \overline{X} \pm z\sigma\sqrt{1 + \frac{1}{n}}$$

med prediktionsgrad q. Som exempel kan vi titta på det nyss avslutade exemplet där vi kan anta att $\sigma = 2$ och får

$$Y = \overline{X} \pm 1.96 \cdot 2 \cdot \sqrt{8/7} = 228.9 \pm 4.2.$$

Nedan följer några reflektioner kring bildande av konfidensintervall och prediktion i normalfördelningen.

- Normalfördelningsantagandet är *alltid* fel, bara approximativt rätt. Sämst stämmer antagandet i svansarna.
- Tack vara centrala gränsvärdessatsen är $\overline{X}$ ”mer normal” än data sjalva. Även detta är dock sämst i svansarna, så ju mer exterma sannolikheter vi uttalar oss om, desto sämre approximation.
- Om antagandet om oberoende är fel kan konfidensgraden i sjalva verket bli en helt annan än den vi tror.
- Det finns tester, s.k. goodness-of-fit-tester, för att kontrollera normalfördelningsantagandet. Man kan också kolla med normalfördelningsplot, se Matlabhäftet.

25 Konfidensintervall för p i binomialfördelningen

Detta är precis det problem man ställs inför med opinionsundersökningar. Se på följande exempel. Bland 10000 tillfrågade säger sig 8.4 % stödja partiet A. Vad kan vi säga om A:s verkliga stöd? Vi har data $X_1, \ldots, X_n$ där X_k är indikatorn att person nr k stödjer A. Vi baserar oss på $\overline{X}$, som blir den relativa frekvensen av folk som säger sig stödja A. Eftersom $n = 10000$ är $\overline{X}$ ungefär normal med vv p och varians $p(1-p)/n$. Om vi nu sätter $z = \Phi^{-1}((1+q)/2)$ får vi det symmeriska konfidensintervallet

$$p = \overline{X} \pm z\sqrt{p(1-p)/n}.$$

Nu dyker ju det okända p upp även i högerledet. Därför brukar man ersätta p med $\overline{X}$ här, så att det symmetriska konfidensintervallet av approximativ konfidensgrad q blir

$$p = \overline{X} \pm z\sqrt{\overline{X}(1-\overline{X})/n}.$$

I vårt exempel får vi $\overline{X} = 0.084$, $n = 10000$ och $z = 1.96$ (för ett 95% intervall), så estimatet blir

$$p = 0.084 \pm 1.96\sqrt{0.084 \cdot 0.916/10000} = 0.084 \pm 0.0054.$$

Konfidensintervallet på 95 % nivå är alltså $8.4\% \pm 0.54\%$.

Några reflektioner kring detta är

- Inga problem i svansarna!
- Mycket bra som skattning av stödet just nu bland dem som svarar och talar sanning.
- Svarsbortfall, lögner och tidsvariation är stora problem i opinionsundersökningar.

26 ML-skattning

ML står för ”maximum likelihood” och är en mycket viktig generell princip för att finna punktskattningar av okända parametrar. Situationen är att vi har observationer $(X_1, \ldots, X_n)$ och att fördelningen för denna vektor beror av en (endimensionell eller flerdimensionell) parameter θ. Principen går ut på att man ser frekvensfunktionen/täthetsfunktionen för data som en funktion av θ och skattar θ med det värde som maximerar denna för de observerade data. Med andra ord, antag att tätheten för $(X_1, \ldots, X_n)$ är $f_\theta(x_1, \ldots, x_n)$ och att vi observerat $X_1 = x_1, \ldots, X_n = x_n$. Då ser vi $f_\theta(x_1, \ldots, x_n)$ som en funktion av θ. Vi skriver

$$L(\theta; x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n)$$

och maximerar denna funktion. Beteckningen L står för "likelihood" (trolighet på svenska, men vi använder den engelska benämningen); eftersom θ är en okänd parameter och ingen stokastisk variabel, vill man använda ett annat ord än frekvensfunktion eller sannolikhetstäthet. Vi skriver $\hat{\theta}$ för det parametervärde som maximerar L och kallar $\hat{\theta}$ för ML-skattningen av θ.

Exempel 26.1. Antag att $X \sim \text{Bin}(3, \theta)$ där θ är okänd, Vi observerar $X = 1$. Vad är ML-skattningen av p?

Frekvensfunktionen är

$$p_\theta(x) = \binom{3}{x}\theta^x(1-\theta)^{3-x}.$$

Med andra ord

$$L(\theta; x) = \binom{3}{x}\theta^x(1-\theta)^{3-x}.$$

I vårt fall har vi $x = 1$, så vi vill maximera

$$L(\theta; 1) = 3\theta(1-\theta)^2.$$

Vi sätter

$$\frac{\partial}{\partial\theta}L(\theta; 1) = 3(1-\theta)^2 - 6\theta(1-\theta) = 0$$

vilket har den intressanta lösningen $\theta = 1/3$. Vi får alltå $\hat{\theta} = 1/3$.

Det nyss givna exemplet visar att även i ett förhållandevis enkelt fall blir beräkningen för att maximera likelihood en aning arbetskrävande. Vad man brukar göra för att göra det hela lite enklare är att istället maximera $\log L$; eftersom logaritmen är en växande funktion är det θ som maximerar $\log L$ samma som det som maximerar L. Låt oss göra detta i det mer allmänna fallet då $X \sim \text{Bin}(n, \theta)$ och vi observerar $X = x$. Vi har då

$$L(\theta; x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

så

$$\ln L = \ln\binom{n}{x} + x\ln(\theta) + (n-x)\ln(1-\theta).$$

Derivatan av detta m.a.p. θ är $x/\theta - (n-x)/(1-\theta)$ som blir 0 då $\theta = x/n$. Vi får alltså $\hat{\theta} = x/n$.

Exempel 26.2. Antag att X är geometriskt fördelad med paramater θ och att vi observerar $X = x$. Vi har då

$$L(\theta; x) = \theta(1-\theta)^{x-1}$$

och

$$\ln L = (x-1)\ln(1-\theta) + \ln(\theta) = 0$$

då $\theta = 1/x$. ML-skattningen av θ är alltså

$$\hat{\theta} = \frac{1}{x}.$$

Exempel 26.3. Antag att $X_1, \ldots, X_n$ är oberoende och Poissonfördelade med parameter θ. Då är

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} (e^{-\theta} \frac{\theta^{x_i}}{x_i!})$$

$$= e^{-n\theta} \frac{\theta^{\sum_i x_i}}{x_1! \cdots x_n!}$$

Därför blir

$$\log L = -n\theta + (\sum_i x_i)\ln(\theta) - \log(x_1! \cdots x_n!)$$

Derivera och sätt till 0 och få

$$-n + \frac{\sum_i x_i}{\theta} = 0$$

vilket ger lösningen

$$\hat{\theta} = \frac{1}{\overline{x}}.$$

Exempel 26.4. Låt $X_1, \ldots, X_n$ vara ett stickprov på en normalfördelning med okänt väntevärde μ och okänd varians σ^2. Här är θ tvådimensionell, $\theta = (\mu, \sigma)$. Tätheten är

$$L(\mu, \sigma^2; x_1, \ldots, x_n) = \prod_{i=1}^{n} (\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}})$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2}\sum_i \frac{(x_i-\mu)^2}{\sigma^2}}.$$

Logaritmera och få

$$\ln L = -n\ln(\sigma) - \frac{1}{2}\sum_i \frac{(x_i-\mu)^2}{\sigma^2} - \frac{n}{2}\ln(2\pi).$$

Den partiella derivatan m.a.p. μ är $\sum_i (x_i - \mu)/\sigma^2$ som blir 0 precis då $\mu = \overline{x}$. Sätt in detta i den partiella derivatan för σ och få

$$\frac{\sum_i (x_i - \overline{x})^2}{\sigma^3} - \frac{n}{\sigma} = 0$$

som blir 0 för $\sigma^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2$. ML-skattningarna blir alltså

$$\hat{\mu} = \overline{x},\ \hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2.$$

Vi noterar att ML-skattningen av σ^2 inte blir s^2 utan $(n-1)s^2/n$.

27 Statistiska tester

Principen för statistiska tester illusteras kanske bäst genom att vi har ett illustrerande exempel att arbeta med. Antag att vi har ett mynt som vi misstänker inte är rättvist vid slantsingling, dvs vi misstänker att sannolikheten att ett kast ger klave inte är 0.5. Vi vill testa detta genom att göra ett antal slantsinglingar. Filosofin är då den att vi utgår från att det vi vill motbevisa är sant. Detta antagande kallas för testets *nollhypotes*, förkortat H_0. Om vi låter p stå för den sanna sannolikheten att myntet ger klave, är alltså H_0 hypotesen att $p = 1/2$. Vi skriver

$$H_0 : p = \frac{1}{2}.$$

Vi utgår alltså från att H_0 är sann, för att sedan se om mätdata tyder på att detta antagande är orimligt. För att avgöra vad det betyder att mätdata tyder på att nollhypotesen är orimlig, ställer vi upp en *alternativhypotes*, H_A, som fångar vad vi i själva verket tror. I det aktuella fallet kan vi t.ex. ha

$$H_A : p \neq 1/2$$

ifall vi bara misstänker att myntet är osymmetriskt, men inte har någon uppfattning om vilken avvikelsen är. Om det å andra sidan är så att vi misstanker att myntet ger klave för ofta, skulle det kunna vara bättre att ta

$$H_A : p > 1/2.$$

Vilket vi väljer har betydelse för vilka utfall av mätdata som vi ska betrakta som visande att H_0 är orimlig, eftersom vi för detta vill att data ska tyda mer på H_A än på H_0. Med $H_A : p \neq 1/2$ tyder alla data som avviker kraftigt fräon hälften klavar på H_A snarare än H_0, medan vi med $H_A : p > 1/2$ endast skulle se utfall med klart fler än hälften klavar som tydande på H_A framför H_0.

Det generella tillvägagångssättet är att skapa en lämplig funktion, T, av mätdata, en s.k. *teststatistika* eller *testfunktion*. Denna ska vald så att vi utifrån värdet på T ska kunna se om H_0 är rimlig eller om vi snarare ska förkasta H_0 till förmån för

H_A. Konkret gör vi så att vi väljer ett litet tal α, ofta 0.05 eller 0.01 eller 0.001, kallad testets *signifikansnivå* och en mängd C av möjliga värden på T, sådan att om $T \in C$ så stöder detta H_A framför H_0, och så att *om* H_0 *är sann*, gäller att $\mathbb{P}(T \in C) \leq \alpha$. Vi skriver ofta $\mathbb{P}_{H_0}(T \in C) \leq \alpha$ för [1] att markera att denna sannolikhet gäller under förutsättning att H_0 är sann. Om nu det nu visar sig att $T \in C$, så *förkastar* vi H_0 till förmån för H_A på signifikansnivå α. Om $T \notin C$ *accepterar* vi H_0.

Observera asymmetrin mellan H_0 och H_A. H_0 är den neutrala hypotesen, som vi accepterar om inte mätdata tyder starkt på H_A framför H_0. Om mätdata blir sådana att H_0 förkastas till förmån för H_A, ser vi detta som ett statistiskt belägg för H_A, ju lägre signifikansnivå desto starkare belägg. Om vi tvingas acceptera H_0 betyder detta å andra sidan inte alls att vi har belägg för H_0, bara att vi inte hade tillräckligt starka data för att förkasta den.

Man ska observera att eftersom alla datamängder kommer att inehålla någon form av mönster, så är det mycket viktigt att den statistiska analysen är utförd (sånär som på att stoppa in siffror i formlerna) innan mätdata samlats in eller åtminståne innan någon sett dem. Man får alltså *inte* tjuvkika på data för att utifrån det avgöra vad som skulle vara ett lämpligt test att göra. (Vad som däremot är helt i sin ordning är att utföra en pilotstudie och utifrån denna avgöra vilka hypoteser som ska testas. Testet måste dock i så fall utföras på en ny mängd av data.)

Att välja T och mängden C, *förkastelsemangden* eller *förkastelseområdet*, är ingen exakt vetenskap eftersom det ofta finns många olika sätt på vilket data skulle kunna tyda på H_A före H_0. Ofta är det dock i den givna situationen ganska tydligt vad som är det mest naturliga att göra. Låt oss nu återvända till exemplet med slantsingling. Säg att vi kastar myntet n gånger och som data får vi X, antalet gånger vi får klave. Om H_0 är sann, dvs $p = 1/2$ förväntar vi oss att få ca $n/1$ klavar och stora avvikelser från detta tyder på $H_A : p \neq 1/2$. Med denna alternativhypotes har vi inte på förhand någn uppfattning om vilken typ av avvikelser vi kommer att gå, så det är naturligt att förkasta både för extremt höga och för extremt låga värden på X. Här kan vi alltså ta $T(X) = X$ och välja $C = [0, a) \cup (b, n]$ där a och b är valda så att $\mathbb{P}(X > a) \leq \alpha/2$ och $\mathbb{P}_{H_0}(X < b) \leq \alpha/2$ (med så nära likhet som möjligt). Av symmetrin framgår att vi kan ta a och b lika långt från $n/2$, dvs vi kan välja x så att

$$\mathbb{P}_{H_0}(|X - \frac{n}{2}| > x) \leq \alpha.$$

[1] Egentligen vill vi ha $\mathbb{P}_{H_0}(T \in C) = \alpha$, men om mätdata är diskreta kan det vara omöjligt att finns C så att det blir exakt likhet. Då väljer vi så nära likhet som möjligt.

Detta betyder att vi väljer x så att

$$\sum_{i=n/2-x}^{n/2+x} \binom{n}{i} 2^{-n} \geq 1-\alpha.$$

Denna summa är lite besvärlig att handskas med, men om n är stort är $(x - n/2)/\sqrt{n/4}$ ungefär standardnormalfördelad, så om vi tar $z = \Phi^{-1}(1-\alpha/2)$ får vi

$$\mathbb{P}(X \in \frac{n}{2} \pm z\frac{\sqrt{n}}{2}) \approx 1-\alpha$$

och kan alltså förkasta H_0 till förmån för $H_A : p \neq 1/2$ om $|X - n/2| > z\sqrt{n}/2$. För att vara mer konkret, antag att $n = 100$ och $\alpha = 0.05$. Då är $z \approx 1.96$, så vi förkastar H_0 till förmån för H_A om $|Z - 50| > 1.96 \cdot 10/2$ dvs om $|Z - 50| > 10$. Med $n = 10000$ får vi atället att vi ska förkasta H_0 om $|Z - 5000| > 100$.

Om vi istället jobbar med alternativhypotesen $H_A : p > 1/2$ förkastar vi bara om X är stor. Vi väljer då $z = \Phi^{-1}(1-\alpha)$ och förkastar H_0 till förmån för $H_A : p > 1/2$ på signifikansnivå α om $Z - n/2 > z\sqrt{n}/2$. Med $\alpha = 0.05$ blir $z \approx 1.64$, så med $n = 100$ förkastar vi om $Z > 50 + 1.64 \cdot 10/2$, dvs om $Z > 58$. Med $n = 10000$ förkastar vi om $Z - 5000 > 1.64 \cdot 100/2$, dvs $Z > 5082$. Vi ser att genom att använda en ensidig alternativhypotes, får vi större möjlighet att detektera en avvikelse från $p = 1/2$ om den går åt det håll vi tror, men å andra sidan förlorar vi helt möjligheten att upptäcka en avvikelse åt andra hållet.

Det sätt på vilket man väljer teststatistikan T är att man väljer en funktion som skattar den okända parametern θ bra. Ofta är det ML-skattningen av θ som används, men även andra val förekommer.

Exempel 27.1. Antag att $X_1, \ldots, X_n$ är ett stickprov på en $\mathrm{N}(\mu, \sigma^2)$-fördelning där μ och σ är okända. Vi vill testa

$$H_0 : \mu = \mu_0$$

mot

$$H_A : \mu \neq \mu_0.$$

ML-skattningen av μ är $\overline{X}$ så vi baserar testet på denna funktion. Under H_0, dvs under antagandet att $\mu = \mu_0$, gäller att

$$\sqrt{n}\frac{\overline{X} - \mu_0}{s} \sim t_{n-1}.$$

Med $z = F_{t_{n-1}}^{-1}(1-\alpha/2)$ fås alltså att

$$\mathbb{P}_{H_0}(\overline{X} \in \mu_0 \pm z\frac{s}{\sqrt{n}}) = 1-\alpha.$$

Om $|\overline{X} - \mu_0| > zs/\sqrt{n}$ förkastas H_0 till förmån för H_A.

Exempel 27.2. Se på det förra exemplet och låt oss goara det hela konkret. Antag att vi misstänker att konserverna med fiskbullar som anges innehålla 200 gram i själva verket inte gör det. För att kolla detta väger vi 10 stycken på måfå valda burkar. Mätdata blir då $X_1, \ldots, X_{10}$ där vi antar att detta är ett stickprov på en $\mathrm{N}(\mu.\sigma^2)$-fördelning och vill testa $H_0 : \mu = 200$ mot $H_A : \mu \neq 200$. Vi förkastar på signifikansnivå 0.05 om $|\overline{X} - 200| > zs/\sqrt{10}$ där $z = F_{t_9}^{-1}(0.975) = 2.26$, dvs om $\overline{X}$ avviker från 200 med mer än $0.71s$.

Men vänta litet, är det inte egentligen så att vi misstänker att $\mu < 200$? Jo, så kan vara fallet och då gör vi stället ett ensidigt test. Vi tar $z = F_{t_9}^{-1}(0.95) = 1.83$ och förkastar H_0 till förmån för H_A om $\overline{X} < 200 - 0.58s$.

27.1 Korrespondensen mellan test och konfidensintervall

Antag att θ är en endimensionell parameter. Då är konfidensintervall för θ av konfidensgrad $1 - \alpha$ och tester av $\theta = \theta_0$ på signifikansnivå α två sidor av samma mynt.

För att inse att man kan göra ett test av $H_0 : \theta = \theta_0$ mot $H_A : \theta \neq \theta_0$ utifrån ett konfidensintervall, notera att det faktum att man har tillgång till ett konfidensintervall av konfidensgrad $1 - \alpha$, betyder att man har två statistikor T_1 och T_2 sådana att $\mathbb{P}(T_1 \leq \theta \leq T_2) = 1 - \alpha$ oavsett vad det sanna värdet på θ är. Förkasta nu H_0 till förmån för H_A om $\theta_0 \notin [T_1, T_2]$. Testet får då signifikansnivå α eftersom $\mathbb{P}_{H_0}(\theta_0 \notin [T_1, T_2]) = \alpha$.

Å andra sidan, om man för varje θ_0 har tillgång till ett test av $\theta = \theta_0$ på signifikansnivå α, dvs om man har en teststatistika T och en förkastelsemangd $C(\theta_0)$ sådan att man förkastar $\theta = \theta_0$ om $T \in C(\theta_0)$, gäller att $P_{\theta_0}(T \in C(\theta_0)) = \alpha$. Låt nu konfidensintervallet vara $I = \{\theta_0 : T \notin C(\theta_0)\}$. Då gäller för alla θ_0 att $P_{\theta_0}(\theta_0 \in I) = 1 - \alpha$, dvs konfidensintervallet får konfidensgrad $1 - \alpha$.

28 Att jämföra två stickprov

Antag att man vill testa om en viss medicin får folk att må bättre. Man gör då en *dubbelblind studie* där n personer lottas till att få medicin och m personer till att få placebo. Efter en tid mäter man hur försöksdeltagarna mår, t.ex. med en skattningsskala då man skattar sitt välmående på en skala mellan 0 och 100. Det är en rimlig modell att anta att stickprovet $X_1, \ldots, X_n$ från gruppen med medicin och stickprovet $Y_1, \ldots, Y_m$ från placebogruppen är oberoende och normalfördelade med väntevärden μ_1 respektive μ_2 och samma varians σ^2. (Antagandet om lika

varians i stickproven är naturligtvis oftast inte exakt sann, men ändå rimligt i fallet då $\mu_1 = \mu_2$, vilket är just den hypotes som ska testas.)

Man vill testa

$$H_0 : \mu_1 = \mu_2$$

ensidigt mot

$$H_A : \mu_1 > \mu_2$$

eller tvåsidigt mot

$$H_A : \mu_1 \neq \mu_2.$$

Som teststatistika väljer man ML-skattningen av $\mu_1 - \mu_2$, som man lätt ser är $\overline{X} - \overline{Y}$. In enstickprovsfallet skattas σ^2 med s^2. Här har vi tillgång till två stickprov och vill utnyttja alla data till en så bra skattning av σ^2 som möjligt. Man använder då den *poolade stickprovsvariansen*

$$s_P^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

Här är förstås s_X^2 och s_Y^2 de två vanliga stickpovsvarianserna. Man kan visa att om H_0 är sann gäller att

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{s_P\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

Ett symmetriskt konfidensintervall för $\mu_1 - \mu_2$ med konfidensgrad $1 - \alpha$ blir då

$$\mu_1 - \mu_2 = \overline{X} - \overline{Y} \pm z_{1-\alpha/2} s_P \sqrt{\frac{1}{n} + \frac{1}{m}}$$

där $z_a = F^{-1}_{t_{n+m-2}}(a)$. Ett nedåt begränsat konfidensintervall blir

$$\mu_1 - \mu_2 \geq \overline{X} - \overline{Y} - z_{1-\alpha} s_P \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Test av $H_0 : \mu_1 = \mu_2$ mot $H_A : \mu_1 \neq \mu_2$ på signifikansnivå α förkastar H_0 till förmån för H_A om

$$|\overline{X} - \overline{Y}| \geq z_{1-\alpha/2} s_P \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Man förkastar H_0 till förmån för $H_A : \mu_1 > \mu_2$ om

$$\overline{X} - \overline{Y} \geq z_{1-\alpha} s_P \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Exempel 28.1. För att ta ett konkret fall, antag att $n = 27$, $\overline{X} = 73.1$, $s_X^2 = 103.6$, $m = 19$, $\overline{Y} = 64.3$ och $s_Y^2 = 91.4$. Då är

$$s_P^2 = \frac{26 \cdot 103.6 + 18 \cdot 91.4}{44} = 98.6.$$

För konfidensgrad 0.99 eller signifikansnivå 0.01 för att ensidigt kofidensintervall/test tar vi $z_{0.99} = F_{t_{44}}^{-1}(0.99) = 2.414$. Ett nedåt begränsat konfidensintervall med konfidensgrad 99% blir

$$\mu_1 - \mu_2 \geq 73.1 - 64.3 - 2.414\sqrt{98.6}\sqrt{\frac{1}{27} + \frac{1}{19}}$$

$$= 8.8 - 7.2 = 1.6$$

Detta betyder också att vi kan förkasta H_0 till förmån för alternativhypotesen $\mu_1 > \mu_2$ på signifikansnivå 1%.

I situationer där antagandet att de två stickprovens varianser är lika inte är rimligt, finns det approximativa tester och konfidensintervallsmetoder, men detta berör vi inte här. Man kan dock observera att om $\sigma_1 \neq \sigma_2$, men båda är *kända*, gäller att

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2}} \sim N(0, 1).$$

Det här fungerar också bra då σ_1 och σ_2 är okända men m och n är stora, såg över 100, eftersom man då kan anta at $\sigma_1^2 = s_X^2$ och $\sigma_2^2 = s_Y^2$.

29 P-värden och styrkor

Vid ett test av en nollhypotes H_0 mot en alternativhypotes H_A ges ofta resultatet i termer av det s.k. p-värdet. Definitionen av p-värdet är den minsta signifikansnivå på vilken H_0 kan förkastas till förmån för H_A med de mätdata man har. Det följer att p-värdet understiger α om och endast om H_0 förkastas till förmån för H_A på signifikansnivå α.

Exempel 29.1. Se på det senaste exemplet. Där hade vi det nedåt begränsade konfidensintervallet

$$\mu_1 - \mu_2 \geq \overline{X} - \overline{Y} - zs_P\sqrt{\frac{1}{n} + \frac{1}{m}}$$

vilket också svarar mot det ensidiga testet $H_0 : \mu_1 = \mu_2$ mot $H_A : \mu_1 > \mu_2$. Högerledet blir 0, dvs man är precis på gränsen att förkasta H_0, då

$$z = \frac{\overline{X} - \overline{Y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Med de konkreta mätdata vi hade, går gränsen vid $z = 2.96$. Eftersom vi finner i en t-fördelningstabell eller via en t-fördelningskalkylator (finns på www) att $F_{t_{44}}(2.96) = 0.995$ betyder det att p-värdet är $1 - 0.995 = 0.005$. Testets p-värde är alltså ca 0.5%.

Exempel 29.2. Beräkna p-värdet för enstickprovstestet $H_0 : \mu = 200$ mot $H_A : \mu \neq 200$ för $n = 10$, $\overline{X} = 196.0$ och $s^2 = 26.0$. Vi hade det tvåsidiga konfidensintervallet

$$\mu = \overline{X} \pm z\frac{s}{\sqrt{n}}$$

vilket har talet 200 precis på gränsen då $200 = 196.0 + z\sqrt{26.0/10}$, dvs då $z = 1.538$. I tabell eller med kalkylator för t-fördelningen ser vi att $F_{t_9}(1.838) \approx 0.92$ vilket är $1 - \alpha/2$ för $\alpha = 0.16$. Det tvåsidiga testets p-värde är alltså ca 0.16. Vi kan alltså inte förkasta p på t.ex. 5% signifikansnivå.

Om nollhypotesen $H_0 : \theta = \theta_0$ är falsk betyder detta $\theta = \theta_1$ för något $\theta_1 \neq \theta_0$. Vad är sannolikheten att ett test upptäcker detta? Med andra ord; om θ_1 är det korrekta parametervärdet, vad är sannolikheten att H_0 kommer att förkastas. Detta beror, förutom på vilken alternativhypotesen är och vad testets signifikansnivå är, på θ_1. Om θ_1 är mycket nära θ_0 är det förstås mycket svårare att förkasta än om skillnaden är stor.

Definition 29.3. Om ett test av $H_0 : \theta = \theta_0$ använder teststatistikan T och här förkastelsemängd $C(\theta_0)$, ges *styrkan* av testet för θ_1 av

$$g(\theta_1) = \mathbb{P}_{\theta_1}(T \in C(\theta_0)).$$

Styrkor är mycket viktiga i kliniska prövningar. Man brukar då bestämma sig för en ”minsta klinisk effekt” vilket vi här kan tolka som ett minsta tal s sådant att endast om $|\theta_1 - \theta_0| \geq s$ anser vi skillnaden som intressant. Man designar sedan sin studie så att $\min\{g(\theta_1) : |\theta_1 - \theta_0|\}$ blir tillräckligt stor. Med ”tillräckligt stor” brukar man då mena att kostnaden för att höja styrkan ytterligare börjar överstiga den eventuella förlusten av att tvingas acceptera en falsk nollhypotes, vilket i detta fall brukar betyda att en i själva verket effektiv medicin inte kommer ut på marknaden.

Exempel 29.4. Ett mynt ger klave med sannolikhet p vid slantsingling. För att testa $H_0 : p = 0.5$ mot $H_A : p \neq 0.5$ singlar vi slanten 100 gånger och förkastar H_0 till förmån för H_A på signifikansnivå 0.05 om $|X - 50| \geq 10$. Om det korrekta värdet på p är 0.7, vad är testets styrka? Vi har

$$g(0.7) = \mathbb{P}_{0.7}(|X - 50| \geq 10) = \mathbb{P}_{0.7}(X \geq 60) + \mathbb{P}_{0.7}(X \leq 40).$$

Detta kan beräknas exakt med binomialfördelningen eftersom $X \sim \text{Bin}(100, p)$, men låt oss gör an normalapproximation och få

$$\mathbb{P}_{0.7}(X \geq 60) = \mathbb{P}_{0.7}(\frac{X - 70}{\sqrt{100 \cdot 0.7 \cdot 0.3}} \geq \frac{60 - 70}{\sqrt{100 \cdot 0.7 \cdot 0.3}})$$

$$\approx 1 - \Phi(-\frac{10}{\sqrt{21}}) \approx 0.985.$$

Sannolikheten att understiga 40 klavar med $p = 0.7$ är försumbar i sammanhanget, så vi ser att styrkan är ca 98.5 %, dvs ett p på 0.7 är det mycket stor chans att vi upptäcker. Notera dock att med ett korrekt p på 0.6 är styrkan bara något över 50 %.

30 Om multipeltestning och tjuvkikande

Antag att vi testar 20 olika nollhypoteser på signifikansnivå 0.05. Om alla nollhypoteserna är sanna, kommer på denna signifikansnivå i genomsnitt en av dem att förkastas. Med andra ord: var tjugonde sann nollhypotes kommer felaktigt att förkastas. Vad är sensmoralen av detta? Jo, att man ska inte göra statistiska tester på måfå, utan endast när man har goda skäl att tro att nollhypotesen är falsk. Goda skäl kan utgöras av rimliga argument för att nollhypotesen är falsk eller av tidigare studier eller pilotstudier.

Vi har redan tidigare nämnt problemen med att utföra tester på data som man redan tjuvkikat på. Om man i data finner ett intressant mönster, visar detta inget annat än på det faktum att alla datamaterial kommer att inehålla någon form av skenbara samband. Om vi vill utföra ett test på grundval av ett mönster som vi sett i data, måste vi ha nya data.

31 Linjär regression

I världen runt omkring oss är det mycket vanligt med linjära samband, till exempel mellan strömstyrka och spänning, mellan avstånd till en galax och dess rödförskjutning i spektrum, mellan längden hos en mor och längden hos hennes

dotter etc. Ibland är dessa samband perfekta, ibland bara ungefärliga i den mening att insamlade data inte kommer att ligga exakt langs en rät linje. Här är vi intresserade av det senare fallet (och i verkligheten finns det nästan inga andra fall).

Modellen vi ska jobba med är att vi har parvisa datapunkter $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ där vi antar att Y_k:na beror linjärt av x_k:na. Med detta menas att det finns två konstanter a och b sadana att

$$Y_k = a + bx_k + \epsilon_k$$

där ϵ_k:na är oberönde och normalfördelade stokastiska variabler med väntevärde 0 och okänd varians σ^2. Modellen innehåller alltså tre okända parametrar, a, b och σ^2. Talen $x_1, \ldots, x_k$ kan vara slumpmässiga eller fixa och i vilket fall ser vi dem som givna och kallar dem ibland för *ställvariabler*. Det slumpmässiga finns i de stokastiska variablerna $Y_1, \ldots, Y_n$ som följaktligen ibland kallas för *svarsvariabler*. Observera att $Y_k \sim \mathrm{N}(a + bx_k, \sigma^2)$.

Vad blir ML-skattningarna av parametrarna? Vi har

$$L(a, b, \sigma; y_1, \ldots, y_n) = f_{(Y_1,\ldots,Y_n)}(y_1, \ldots, y_n)$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2}\frac{\sum_k (y_k - a - bx_k)^2}{\sigma^2}}.$$

Med avseenda på (a, b) blir

$$\ln L = K - \frac{1}{2\sigma^2} \sum_k (y_k - a - bx_k)^2$$

(där K är en konstant), vilket maximeras om a och b väljs så att kvadratsumman $\sum_k (y_k - a - bx_k)^2$ minimeras. Genom att sätta gradienten till 0, få vi ekvationssystemet

$$\sum_k (y_k - a - bx_k) = 0$$

$$\sum_k x_k (y_k - a - bx_k) = 0.$$

Det är inte svårt att lösa detta och få

$$\hat{b} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

där de två kvadratsummorna S_{xy} och S_{xx} ges av

$$S_{xy} = \sum_k (y_k - \overline{y})(x_k - \overline{x})$$

och

$$S_{xx} = \sum_k (x_k - \overline{x})^2.$$

Man kan visa (men vi gör inte det här) att

$$\hat{a} \sim \mathrm{N}(a, \frac{\sigma^2 \sum_{k=1}^n x_k^2}{nS_{xx}})$$

$$\hat{b} \sim \mathrm{N}(b, \frac{\sigma^2}{S_{xx}}).$$

Om σ^2 är känd kan man göra tester och konfidensintervall för a och b baserat på dessa, men vanligen är ju σ^2 okänd och skattas då med

$$s^2 = \frac{1}{n-2} \sum_{k=1}^n (Y_k - \hat{a} - \hat{b}x_k)^2.$$

Denna är en väntevärdesriktig skattning av σ^2 och man kan visa att teststatistikorna T_a och T_b båda är t-fördelade med $n-2$ frihetsgrader. Här är

$$T_a = \frac{\hat{a} - a}{s\sqrt{\sum_k x_k^2 / S_{xx}}}$$

och

$$T_b = \frac{\hat{b} - b}{s\sqrt{1/S_{xx}}}.$$

Exempel 31.1. Antag att vi observerar hur långa ett antal mammor och deras äldsta är. Observationerna är

Mammor x_k (cm)	170.1	161.3	177.1	167.0	168.6	162.2
Döttrar y_k (cm)	173.4	162.6	170.5	162.8	171.2	165.5

Här har vi alltså $n = 6$ och

$$S_{xy} = \sum_k (x_k - \overline{x})(y_k - \overline{y}) = 91.323$$

$$S_{xx} = \sum_k (x_k - \overline{x})^2 = 166.628.$$

Detta ger

$$\hat{b} = 0.548$$

och

$$\hat{a} = 75.74.$$

Den linjära regressionslinjen blir alltså

$$y \approx 75.7 + 0.55x.$$

Vi har också

$$s^2 = \frac{1}{4}\sum_k (y_k - \hat{a} - \hat{b}x_k)^2 = 14.34.$$

Detta ger ett symmetriskt konfidensintervall för b med konfidensgrad 95 % som

$$b = \hat{b} \pm F_{t_4}^{-1}(0.975)\frac{s}{\sqrt{S_{xx}}} \approx 0.55 \pm 0.81.$$

Detta betyder att om vi testar $H_0 : b = 0$ mot $H_A : b \neq 0$ (vilket är den vanligaste hypotesen att pröva) på 5% signifikansnivå, kan vi inte förkasta H_0.

Vad är testets p-värde? Om vi löser $0 = 0.55 \pm F_{t_4}^{-1}(1-\alpha/2)\frac{s}{\sqrt{S_x}}$ får vi $F_{t_4}^{-1}(1-\alpha/2) = 0.55\sqrt{166.628}/\sqrt{14.34} \approx 1.87$. Detta ger $1 - \alpha/2 = 0.9326$, dvs $\alpha \approx 0.135$. P-värdet är alltså ca 13.5%.

32 Prediktion i linjär regression

Antag att vi har ställvariabelvärde x och vill förutsäga $Y = a + bx + \epsilon$. Låt $D = Y - \hat{a} - \hat{b}x$. Vi kan skriva om $D = Y - \overline{Y} + \hat{b}(x - \overline{x})$ och ser därmed att D, som uppenbarligen har väntevärde 0, har varians $\sigma^2 + \sigma^2/n + \mathbb{V}\mathrm{ar}(\hat{b})(x-\overline{x})^2 = (1 + 1/n)\sigma^2 + (x-\overline{x})^2\sigma^2/S_{xx}$. Detta medför att teststatistikan

$$T = \frac{D}{s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

Ett prediktionsintervall med prediktionsgrad $1 - \alpha$ blir alltså

$$Y = \hat{a} + \hat{b}x \pm F_{t_{n-2}}^{-1}(1-\alpha/2)s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{X})^2}{S_{xx}}}.$$

I mor-dotter-exemplet ovan får vi att ett prediktionsintervall med prediktionsgrad 95% till en mor av längd 165 cm, ges av

$$166.2 \pm 2.776\sqrt{14.3}\sqrt{\frac{7}{6} + \frac{(165 - 167.7)^2}{166.6}} = 166.2 \pm 11.7.$$

33 Bayesiansk statistik

Hittills har vi tänkt på okända parametrar som fixerade tal, om än just okända. Men annars brukar vi ju använda slump för att modellera osäkerhet, så varför inte tänka på okända parametrar som stokastiska variabler, ett slumpförsök utfört av naturen och som vi inte vet utfallet av? Ofta är det naturligt och vägvinnande att göra just så, ibland inte.

Exempel 33.1. En slant singlas 10 gånger. Vi vet att myntet i fråga antingen är mynt A, som ger klave med sannolikhet $1/3$ eller mynt B, som ger klave med sannolikhet $2/3$. Låt X var antalet observerade klavar. Då är $X \sim \text{Bin}(10, \theta)$ där θ antingen är $1/3$ eller $2/3$, okänt vilket. Det är då naturlig att se θ som en stokastisk variabel med $\mathbb{P}(\theta = 2/3) = \mathbb{P}(\theta = 1/3) = 1/2$. Antag att vi observerar sex klavar. Bayes formel ger då

$$\mathbb{P}(\theta = \frac{1}{3}|X = 6) = \frac{\mathbb{P}(X = 6|\theta = \frac{1}{3})\mathbb{P}(\theta = \frac{1}{3})}{\mathbb{P}(X = 6|\theta = \frac{2}{3})\mathbb{P}(\theta = 2/3) + \mathbb{P}(X = 6|\theta = \frac{1}{3})\mathbb{P}(\theta = \frac{1}{3})}$$

$$= \frac{\frac{1}{2}\binom{10}{6}(\frac{1}{3})^6(\frac{2}{3})^4}{\frac{1}{2}\binom{10}{6}(\frac{1}{3})^6(\frac{2}{3})^4 + \frac{1}{2}\binom{10}{6}(\frac{2}{3})^6(\frac{1}{3})^4}$$

$$= \frac{(\frac{1}{3})^2}{(\frac{1}{3})^2 + (\frac{2}{3})^2} = \frac{1}{5}.$$

Antagandet att $\mathbb{P}(\theta = 1/3) = 1/2$ kallas för θ:s *à-priori-fördelning*. Denna är som vi ser subjektiv. Den betingade sannolikheten $\mathbb{P}(\theta = 1/3|X = 6) = 1/5$ kallas för θ:s *à-posteriori-fördelning*. Denna är objektiv om man accepterar a-priori-fördelningen. Resultatet av den är dock mycket beroende av a-priori-fördelningen. Antag till exempel att slanten är vald av Lisa, som gärna väljar slantar med stor sannolikhet att visa klave, så att en a-priori-fördelning med $\mathbb{P}(\theta = 1/3) = 1/4$ är mer rimlig. Då får vi

$$\mathbb{P}(\theta = \frac{1}{3}|X = 6) = \frac{\frac{3}{4}(\frac{1}{3})^2}{\frac{3}{4}(\frac{1}{3})^2 + \frac{1}{4}(\frac{2}{3})^2} = \frac{3}{7}.$$

Exempel 33.2. Vi observerar en $\exp(\theta)$-fördelad stokastisk variabel $X = 0.713$. Om vi vet att $0 < \theta < 1$, vad är rimligt att tro om θ efter att ha sett detta resultat? Antag att vi a-priori inte har någon speciell uppfattning om theta. Då kan vi ta $f_\theta(t) = 1, 0 < t < 1$. Vi får

$$f_{\theta|X}(t|x) = \frac{f_{X|\theta}(x|t)f_\theta(t)}{\int_0^1 f_{X|\theta}(x|s)f_\theta(s)ds}$$

$$= Kte^{-xt} = Kte^{-0.713t}$$

där K är en normaliserande konstant. Å andra sidan, med $f_\theta(t) = 2t$, fås $f_{\theta|X}(t|x) = K't^2e^{-0.713t}$.

Man kan göra en allmän beräkning. Antag att vi a-priori har en täthet eller frekvensfunktion f_θ och att den betingade tätheten eller frekvensfunktionen $f_{X|\theta}(x|t)$ är känd. A-posteriori-tätheten/frekvensfunktionen blir då enligt Bayes formel

$$f_{\theta|X}(t|x) = \frac{f_{X|\theta}(x|t)f_\theta(t)}{\int_{-\infty}^{\infty} f_{X|\theta}(x|s)f_\theta(s)ds} = Cf_{X|\theta}(x|t)f_\theta(t).$$

Om man också vill ha en a-posteriori punktskattning av θ, brukar man ta $\hat{\theta}$ som väntevärdet i a-posteriori-fördelningen.

Exempel 33.3. En slant visar klave med sannolikhet θ. Den singlas två gånger och ger ingen klave. Vad är din skattning av θ. Om vi ser θ som fix, är ML-skttningen av θ

$$\hat{\theta} = \frac{X}{n}$$

som i detta fall blir 0. Är det verkligen rimligt att helt döma ut chansen att få klave bara baserat på två kast? En Bayesiansk anstats ter sig rimligare. Antag att $f_\theta(t) = 1$, $0 < t < 1$. Då blir

$$f_{\theta|X}(t|x) = Cf_{X|\theta}(x|t)f_\theta(t) = C\binom{2}{x}t^x(1-t)^{2-x}.$$

I vårt fall blir $f_{\theta|X}(t|0) = C(1-t)^2 = 3(1-t)^2$. Som skattning tar vi väntevärdet i denna fördelning, som är

$$\hat{\theta} = 3\int_0^1 t(1-t)^2dt = \frac{1}{4}.$$

Det största problememt med Bayesiansk statistik är att ge en objektiv a-priori-fördelning. I vissa fall går detta lättare än annars, till exempel om man kan basera sin a-priori-fördelning på erfarenhet av tidigare data.

Exempel 33.4. Speech recognition. Ett meddelande bestående av tre nollor eller ettor skickas. Varje tecken kodas dock fel med sannolikhet 0.1 oberönde av de andra två tecknen. Om vi låter X vara det sanna meddelandet och Y det vi läser av, vad blir sannolkhetsfördelningen för X givet Y? Antag att de åtta olika meddelandena förekommer enlig följande proportioner

Meddelande	000	001	010	011	100	101	110	111
Sannolikhet	0.05	0.10	0.08	0.12	0.14	0.21	0.17	0.13

Då får vi exempelvis

$$\mathbb{P}(X = 110|Y = 110) = \frac{\mathbb{P}(Y = 110|X = 110)\mathbb{P}(X = 110)}{\sum_m \mathbb{P}(Y = 110|X = m)\mathbb{P}(X = m)}$$

$$= \frac{0.9^3 \cdot 0.17}{0.9^3 \cdot 0.17 + 0.9^2 \cdot 0.1(0.13 + 0.14 + 0.08) + 0.9 \cdot 0.1^2(0.21 + 0.12 + 0.05) + 0.1^3 \cdot 0.1}$$

$$= 0.796.$$

Vi får också exempelvis

$$\mathbb{P}(X = 111|Y = 110) = 0.068.$$

Detta exempel illustrerar en mycket vanlig situation, som förekommer till exempel i ansiktsigenkänning, oljeletande, spel på fotboll, etc.

34 Markovkedjor

En Markovkedja (i diskret tid) är en följd $X_0, X_1, X_2, \ldots$ av stokastiska variabler som tar sina värden i något ändligt rum S, kallat för Markovkedjans *tillståndsrum*, och som uppfyller att det för alla $i, j \in S$ finns tal p_{ij} sådana att för alla $t = 1, 2, 3, \ldots$ och alla $i_0, i_1, \ldots, i_t, j \in S$ gäller att

$$\mathbb{P}(X_{t+1} = j|X_t = i_t, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{t+1} = j|X_t = i) = p_{i_t j}.$$

Matrisen P som på rad i och i kolonn har elementet p_{ij} kallas för Markovkedjans *övergångsmatris* och talen p_{ij} för Markovkedjans övergångssannolikheter. Det gäller alltså per definition att $p_{ij} = \mathbb{P}(X_{t+1} = j|X_t = i)$ och att denna sannolikhet inte alls beror av var kedjan varit före tid t (dvs inte alls beror av $X_0, \ldots, X_{t-1}$) utan endast var den är vid tid t. Enligt totala sannolikhetslagen gäller att

$$\mathbb{P}(X_{t+1} = j) = \sum_{i \in S} \mathbb{P}(X_t = i)p_{ij}.$$

Om vi skriver $\mathbf{p}_t$ för vektorn $[\mathbb{P}(X_t = i)]_{i \in S}$, blir detta på matrisform

$$\mathbf{p}_{t+1} = \mathbf{p}_t P.$$

Med hjälp av induktion följer att

$$\mathbf{p}_t = \mathbf{p}_0 P^t.$$

Exempel 34.1. Antag att

$$P = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.5 & 0.4 & 0.1 \end{bmatrix}.$$

Då får vi till exempel, med $\mathbf{p}_0 = [1\ 0\ 0]$,

$$\mathbf{p}_1 = [1\ 0\ 0]P = [0.7\ 0.2\ 0.1]$$

och

$$\mathbf{p}_5 = [1\ 0\ 0]P^5 \approx [0.554\ 0.289\ 0.157]$$

och

$$\mathbf{p}_{10} = [1\ 0\ 0]P^{10} \approx [0.553\ 0.289\ 0.158].$$

I exemplet ser det ut som att $\mathbf{p}_t$ konvergerar mot en fix vektor då $t \to \infty$. Kan detta vara en generell sanning? Ja, under mycket milda villkor. Vi ska nu titta på en version av detta resultat.

Definition 34.2. Om vekorn π är sådan $\pi_i > 0$ för alla $i \in S$ och $\sum_{i \in S} \pi_i = 1$ och

$$\pi P = \pi,$$

kallas π för en invariant fördelning för Markovkedjan med övergångsmatris P.

Observera att om π är en invariant fördelning, så är π en vänsteregenvektor till P med egenvardet 1.

Sats 34.3. *Om P är sådan att $p_{ij} > 0$ för alla $i, j \in S$ så gäller att det finns en unik invariant fördelning π, och att för alla $i \in S$ och alla $\mathbf{p}_0$ gäller*

$$\lim_t \mathbb{P}(X_t = i) = \pi_i.$$

Bevis. Låt $\epsilon = \min_{i,j} p_{ij}$. Enligt förutsättning är $\epsilon > 0$. Låt $Y_0, Y_1, \ldots$ vara en Markovkedja med samma övergångsmatris som $\{X_t\}$ med egenskapen att Y_0 har fördelning π, där π är en invariant fördelning. Eftersom π är invariant gäller $\mathbb{P}(Y_t = i) = \pi_i$ för alla i och alla t. Mer specifikt låter vi Y_t:s rörelse ges av $Y_{t+1} = X_{t+1}$ om $Y_t = X_t$ och Y_{t+1} väljs oberoende av alla X_s i annat fall. Då har $\{Y_t\}$ mycket riktigt samma övergångsmatris som $\{X_t\}$ och så fort $X_t = Y_t$ kommer $X_s = Y_s$ för alla $s \geq t$. Dessutom gäller att

$$\mathbb{P}(X_t \neq Y_t) \leq (1 - \epsilon)^t$$

så

$$|\mathbb{P}(X_t = i) - \pi_i| = |\mathbb{P}(X_t = i) - \mathbb{P}(Y_t = i)| \leq \mathbb{P}(X_t \neq Y_t) \leq (1 - \epsilon)^t \to 0$$

och π måste vara unik. □

Om $\mathbb{P}(X_t = i) \to \pi_i$ då $t \to \infty$ för alla i, sager man att π är en *stationär fördelning*. Villkoret $p_{ij} > 0$ är ganska restriktivt, men kan mildras betydligt. Det räcker till exempel med att för alla $i, j \in S$ gäller att $\mathbb{P}(X_{t+s} = j | X_t = i) > 0$ för något s.

En viktig klass av Markovkedjor är dem som kallas *reversibla*. Intiuitivt: om man tittar på en stationär Markovkedja baklänges i tiden, får man alltid en ny Markovkedja. Om denna visar sig ha samma övergångsmatris som originalet, kallar man sin Markovkedja reversibel. Den formella definitionen av att vara reversibel är att det finns en vektor π sådan att det för alla $i, j \in S$ gäller at

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

Man inser lätt att π då måste vara en stationär fördelning. Att beräkna an stationär fördelning av detta slag är betydligt enklare än att lösa det ekvationssystem av n variabler som uppstår i det allmänna fallet.