



Statistikens grunder, 15p dagtid

HT 2012

F7 forts. Kap 6

- *Lite repetition + lite utveckling av*
 - *Stokastisk variabel*
 - *Diskreta och kontinuerliga sv*
 - *Frekvensfunktion (diskr.),
Täthetsfunktion (kont.)
Fördelningsfunktion (diskr. + kont.)*
- *Sedan*
 - *Väntevärden och varianser*
 - *Simultana fördelningar (två sv)*

Stokastiska variabler

Experiment definierade genom ett utfallsrum Ω , bestående av möjliga händelser $A_1, A_2, A_3, \dots$

Varje möjlig händelse tilldelas en sannolikhet $P(A_i)$ som hade vissa egenskaper (axiomen!).

Nu koncentrerar vi oss på experi-ment och utfall som kan uttryckas som, eller översättas till, **reella tal**.

Stokastiska variabler

Ex) Antalet möss som insjuknar efter exponering av ngt smittoämne.

Ex) Kroppslängden hos en slumpmässigt utvald person boende i Sverige.

Ex) Hur gör vi med experimentet "regnar" / "regnar inte"?

Stokastiska variabler

En stokastisk variabel X antar ett siffervärde slumpmässigt ($X = x$).

Möjliga värden på x bestäms av utfallsrummet

Sannolikhetsmodellen för X kan beskrivas enkelt enligt:

- Utfallsrum Ω_x : $x = \dots$
- Sannolikheter för alla delmängder A av Ω_x :

$P(X = x)$, $P(X \leq x)$, $P(X \neq 2)$ osv.

Funktionerna för en s.v.

Diskreta sv:

- Frekvensfunktionen: $f(x) = P(X = x)$
- Fördelningsfunktion: $F(x) = P(X \leq x)$
 - Båda ger sannolikheter

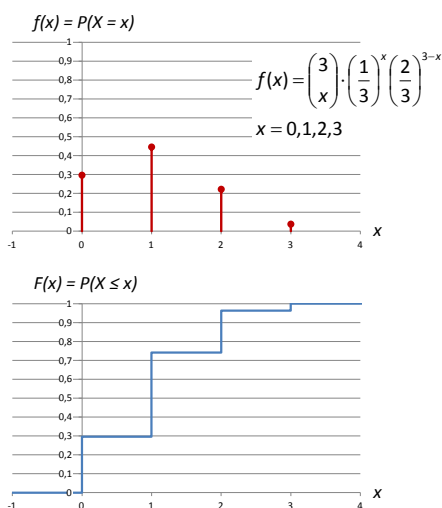
$$F(x) = \sum_{k=0}^x f(k)$$

Kontinuerliga sv:

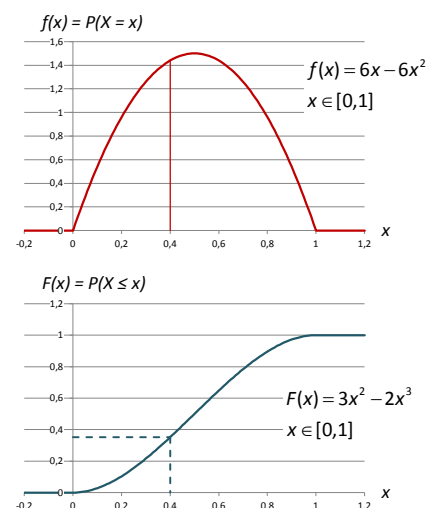
- Täthetsfunktionen: $f(x) = \text{"tätet"}$
 - Inte en sannolikhet
- Fördelningsfunktion: $F(x) = P(X \leq x)$
 - En sannolikhet

$$F(x) = \int_{-\infty}^x f(t) dt$$

Diskret s.v.



Kontinuerlig s.v.



Poängen med funktionen

En stokastisk variabel definieras helt genom utfallsrummet och frekvens- alt. täthetsfunktion *eller* genom fördelningsfunktionen.

Med dessa funktioner kan vi beräkna alla de eftersökta sannolikheterna men även andra egenskaper för X .

Kommentar:

Man säger ofta " X är ...-fördelad" eller "fördelningen är ..." osv.

Med **fördelning** menas då **modellen** för X vilket implicit inkluderar utfallsrummet.

Poängen med funktionen

Olika fördelningar (funktioner) används för att beskriva olika situationer.

Läget eller medelvärdet är större för X än för Y .

X kanske varierar mer än Y i sina observationer, **variansen** är större.

Allt detta kan beräknas genom att vi känner fördelningen.

Exempel 1

Två olika s.v.

- $f(x) = 2^x e^{-2}/x!$, $\Omega_x = 0, 1, 2, 3, \dots$
- $g(y) = 5^y e^{-5}/y!$, $\Omega_y = 0, 1, 2, 3, \dots$

Hur jämför sig dessa fördelningar med varann?

Kommer den ena typiskt vara större än andra eller hamnar observationerna ungefär på samma ställe?

Är den ena mer eller mindre utspridd än den andra?

Vi behöver sammanfattande mått som ger en bild av fördelningarna!

Exempel 2

Två olika set med empiriska observationer

	x_i	y_i
• Hur jämför sig dessa fördelningar med varann?	32	38
	11	95
• Är den ena typiskt vara större än andra?	45	19
	61	83
	72	18
• Är den ena mer eller mindre utspridd än den andra?	101	16
	48	39
	64	26
	60	26
	20	83

Vi behöver sammanfattande mått som ger en bild av fördelningarna!

Väntevärde 1

Antag att ett experiment med utfallsrum $\Omega_x = \{0, 1, \dots, C\}$, upprepas n gånger.

Det aritmetiska medelvärdet kan då skrivas som

$$\bar{x} = \sum_{x=0}^C \left(x \cdot \frac{n(x)}{n} \right)$$

där $n(x)$ är antalet gånger det blev x .

Om n är tillräckligt stort, så kan vi misstänka att de relativa frekvenserna $n(x)/n$ kommer att stabiliseras runt de "sanna" sannolikheterna $f(x)$.

Väntevärde 2

Låt X vara en diskret s.v. med utfallsrum $\{x_1, x_2, \dots, x_K\}$.

Väntevärdet för en s.v. X betecknas med $E(X)$ och definieras

$$E(X) = \sum_{i=1}^K x_i f(x_i)$$

För en kontinuerlig s.v. med utfallsrum på intervallet (a, b) definieras väntevärdet

$$E(X) = \int_a^b x f(x) dx$$

Väntevärde 3

Medelvärdet för ett observerat data-material eller **väntevärdet** för en fördelning ger en indikation om var observationerna kommer att hamna, dvs. **läget**.

Olika fördelningar har olika lägen se t.ex. Figur 6.5.

*Det finns även alternativa **lägesmått** inom statistiken såsom median, vägda medelvärden, geometriska och harmoniska medelvärden, typvärde.*

Varians 1

När vi upprepar experimentet är observationerna oftast utspridda över flera olika utfall. **Variansen** och **standardavvikelsen** är mått på hur stor spridningen är.

Olika fördelningar är olika mycket utspridda, se t.ex. Figur 6.6.

*Det finns alternativa **spridningsmått** inom statistiken såsom kvartilavstånd, variationsvidd (max-min), Mean Absolute Deviation (MAD).*

Varians 2

Ett sätt att mäta spridning är att mäta avståndet till ett centralmått, t.ex. väntevärdet och ta ett genomsnittligt avstånd (vägt mot sannolikheten):

$$\sum_{i=1}^K (x_i - \bar{x}) f(x_i)$$

Dessvärre blir den här summan alltid noll och duger inte som mått (kan man inse det?)

Men de **kvadrerade** avstånden är alltid ≥ 0 och då blir summan alltid ≥ 0

Varians 3

Låt X vara en diskret s.v. med utfallsrum $\{x_1, x_2, \dots, x_K\}$.

Variansen för en s.v. X betecknas med $E(X)$ och definieras

$$V(X) = \sum_{i=1}^K [x_i - E(X)]^2 f(x_i)$$

För en kontinuerlig s.v. med utfallsrum på intervallet (a, b) definieras variansen

$$V(X) = \int_a^b [x - E(X)]^2 f(x) dx$$

Transformationer

Ibland har man behov av att **transformera** en s.v. Vad händer med väntevärde och varians om?

- alla löner får ett påslag på 100 kr?
Väntevärdet adderas med 100 (väl?). Hur gör vi med variansen?
- genomsnittlig lön i SEK omräknas till EUR?
Väntevärdet multipliceras med valutakursen (väl?). Hur gör vi med variansen?

Räkneregler

Antag att c är konstant (ett tal som aldrig ändras).

$$E(c) = c$$

$$V(c) = 0$$

$$E(c+X) = c + E(X)$$

$$V(c+X) = V(X)$$

$$E(cX) = c \cdot E(X)$$

$$V(cX) = c^2 V(X)$$

Varians 4

Alternativa men helt ekvivalenta sätt att räkna variansen är

$$V(X) = \sum_{i=1}^K x_i^2 f(x_i) - E(X)^2$$

resp.

$$V(X) = \int_a^b x^2 f(x) dx - E(X)^2$$

Standardavvikelse och engelskan

Standardavvikelse definieras sedan helt enkelt som kvadratroten ur variansen, dvs.

$$SD(X) = \sqrt{V(X)}$$

Engelska översättningar:

Väntevärde	<i>Expected value</i>
Varians	<i>Variance</i>
Standard avvikelse	<i>Standard deviation</i>
Frekvensfunktion	<i>Frequency function</i>
Sannolikhetsfunktion	<i>Probability (mass) function</i>
Täthetsfunktion	<i>Density function</i>
Fördelningsfunktion	<i>Distribution function</i>

Övning

Vi arrangerar ett lotteri.
Vi säljer 100 lotter för 2kr styck.
Endast en lott vinner, 100kr.

Fråga: Vad är den förväntade vinsten?

Definiera den diskreta s.v.:

X = vinst i lotteriet

Identifiera utfallsrummet:

$$\Omega_x = \{-2, 98\}$$

Övning, forts.

Ange sannolikheterna:

$$f(-2) = 99/100, f(98) = 1/100$$

Beräkna väntevärdet:

$$\begin{aligned} E(X) &= -2 \cdot 99/100 + 98 \cdot 1/100 \\ &= (-198 + 98)/100 = -1 \end{aligned}$$

Tolkning: I det långa loppet går vi i genomsnitt back en krona.

Variansen:

$$\begin{aligned} V(X) &= (-2+1)^2 \cdot 99/100 + (98+1)^2 \cdot 1/100 = \\ &= 99/100 + 99^2/100 = 99 \end{aligned}$$

Övning, forts.

Obs! Varje gång vi spelar så förväntas vi gå back 1 kr.

Spelar vi tusen gånger så är den **förväntade totala vinsten** vad då?

Total vinst efter n gånger: nX

$$E(nX) = n \cdot E(X) = n \cdot (-1) = -n$$

Variansen:

$$V(nX) = n^2 \cdot V(X) = 99n^2$$

Två s.v.

Ibland (ganska ofta m.a.o) har man behov av att studera två eller flera s.v. **simultant**.

Man samlar nästan alltid in mer uppgifter än bara ett variabelvärde.

- Hur jämför sig olika s.v. med avseende på läge och spridning?
- Finns det tecken på att de **samvarierar**?
- Kan det finnas **samband** mellan två stokastiska variabler?

Simultana fördelningar 1

Låt X och Y vara två diskreta s.v.

Låt X beskrivas av $f_x(x)$ och Ω_x och Y av $f_y(y)$ och Ω_y

Var och en uppfyller $f_x(x)$ resp. $f_y(y)$ alla de krav vi ställer på frekvens- alt. täthetsfunktioner.

Vi definierar sedan den **simultana frekvensfunktionen** som

$$f_{xy}(x, y) = P(X = x \cap Y = y)$$

Simultana fördelningar 2

Notera att vi använder index för att skilja funktionerna och utfallsrummen åt.

Den **simultana fördelningsfunktionen** definieras på motsvarande sätt:

$$F_{xy}(x, y) = \sum_{i \leq x} \sum_{j \leq y} f_{xy}(i, j)$$

Simultana fördelningar 3

För den simultana frekvensfunktionen och för den simultana fördelningsfunktionen gäller samma regler som tidigare. Vilka?

1. $f(x,y) \geq 0$ för alla par (x,y)
2. $\sum_x \sum_y f(x,y) = 1$ (summera alla sannolikheter i det gemensamma utfallsrummet Ω_{xy})
3. Sannolikheten för disjunkta händelser adderas

Exempel

Antag följande frekvensfunktion:

$x \setminus y$	1	2	3
1	0,05	0,10	0,20
2	0,40	0,15	0,10

Beräkna

- $f_{xy}(2,2)$
 - $F_{xy}(2,2)$
 - $f_x(2)$
 - $P(X=2 | Y=2) = P(X=2 \cap Y=2) / P(Y=2)$
 $= 0,15 / 0,25 = 3/5$
 - Är X och Y oberoende?
- Och inget annat!*

Repetitionsövning

Låt A och B var två händelser i et experiment.

1. Antag att A och B är disjunkta
 - a) Är A och B oberoende?
 - b) Vad är $P(A|B)$?
2. Antag att $A \subset B \subset \Omega$
 - a) Är A och B oberoende?
 - b) Vilken är större, $P(A|B)$ el. $P(A)$?

Tips! Rita ett Venndiagram

F8 forts. Kap 6

- Lite repetition + lite utveckling av
 - Väntevärden och varianser
 - Simultana fördelningar (två sv)
- Sedan ...
 - Marginalfördelningar
 - Betingade fördelningar
 - Betingade väntevärden
 - Kovarians, linjärt samband

Repetition

Väntevärde

Låt X vara en diskret s.v. med utfallsrum $\{x_1, x_2, \dots, x_K\}$.

Väntevärdet för en s.v. X betecknas med $E(X)$ och definieras

$$E(X) = \sum_{i=1}^K x_i f(x_i)$$

För en kontinuerlig s.v. med utfallsrum på intervallet (a, b) definieras väntevärdet

$$E(X) = \int_a^b x f(x) dx$$

Exempel

Temperaturen i en bastu kan ställas in med ett reglage med 5 lägen. Vilket läge det blir kan ses som en slumpmässig händelse som beror på vilken person som råkar ställa in den.

Sannolikhetsfördelningen för X = temperaturen har det visat sig kan beskrivas av följande tablå:

x	50	60	70	80	90
$F(x)$	0,05	0,25	0,70	0,90	1,00
	0,05	0,20	0,45	0,20	0,10
	2,5	12,0	31,5	16,0	9,0

Beräkna väntevärdet $E(X)$: **= 71**

Repetition

Varians

Låt X vara en diskret s.v. med utfallsrum $\{x_1, x_2, \dots, x_K\}$.

Variansen för en s.v. X betecknas med $V(X)$ och definieras

$$V(X) = \sum_{i=1}^K [x_i - E(X)]^2 f(x_i)$$

För en kontinuerlig s.v. med utfallsrum på intervallet (a, b) definieras variansen

$$V(X) = \int_a^b [x - E(X)]^2 f(x) dx$$

Exempel, forts.

I bastuexemplet, beräkna variansen $V(X)$ och standardavvikelsen $SD(X)$

x	50	60	70	80	90
$f(x)$	0,05	0,20	0,45	0,20	0,10
$x - E(X)$	-21	-11	-1	9	19
$[x - E(X)]^2$	441	121	1	81	361
$[x - E(X)]^2 f(x)$	22,05	24,20	0,45	16,20	36,10
x^2	2500	3600	4900	6400	8100
$x^2 f(x)$	125	720	2205	1280	810

$$V(X) = \sum_{i=1}^K [x_i - E(X)]^2 f(x_i) = 99$$

$$V(X) = \sum_{i=1}^K x_i^2 f(x_i) - E(X)^2 = 5140 - 71^2$$

Exempel, forts.

En amerikan är intresserad av resultaten men vill ha siffrorna i Fahrenheit.

Omräkning sker enligt formeln:

$$Y = \frac{9}{5}X + 32$$

Väntevärde och varians i Fahrenheit:

$$E(Y) = E\left(\frac{9}{5}X + 32\right) = \frac{9}{5}E(X) + 32 = 159,8$$

$$V(Y) = V\left(\frac{9}{5}X + 32\right) = \left(\frac{9}{5}\right)^2 V(X) = 320,76$$

$$SD(Y) = \sqrt{V(Y)} \approx 17,91$$

Repetition

Simultana fördelningar

Låt X och Y vara två diskreta s.v.

Vi definierar sedan den **simultana frekvensfunktionen** som

$$f_{XY}(x, y) = P(X = x \cap Y = y)$$

För kontinuerliga s.v. definieras på samma sätt den **simultana täthetsfunktionen** för att beräkna sannolikheter

$$P(X \in A \cap Y \in B) = \int_{x \in A} \int_{y \in B} f_{XY}(x, y) dx dy$$

Exempel 1

- X och Y är diskreta s.v. där

$$f_{XY}(x, y) = x/40,$$

$$\Omega_X = \Omega_Y = \{1, 2, 3, 4\}$$

$$f_{XY}(2, 1) = 2/40 = 1/20 = 0,05$$

$$f_{XY}(2, 3) = 2/40 = 1/20 = 0,05$$

- Låt X och Y är kontinuerliga s.v. där

$$f_{XY}(x, y) = 6e^{-3x-2y}, \quad x \text{ och } y \in (0, \infty)$$

Utfallsrummet Ω_{XY}

Låt X och Y vara två s.v. med en simultan frekvensfunktion (alt. täthetsfunktion) $f_{XY}(x, y)$.

Var för sig kan X och Y ha sina resp. utfallsrum, säg Ω_X resp. Ω_Y

Men det **simultana utfallsrummet** Ω_{XY} behöver inte nödvändigtvis vara "**produkten**" av dessa.

Dvs. det kan finnas kombinationer av utfall för X och Y som inte kan inträffa.

Exempel 2

Antag att X och Y är diskreta och att $f_{XY}(x,y)$ kan beskrivas enligt:

$x \setminus y$	1	2	3	4
1	0,1	0	0	0
2	0,1	0,1	0	0
3	0,1	0,1	0,1	0
4	0,1	0,1	0,1	0,1

Var för sig ser man att X och Y har samma utfallsrum $\{1,2,3,4\}$ men att flera kombinationer har noll i sannolikhet att inträffa, dvs. de ingår inte i Ω_{XY} .
T.ex. $f_{XY}(4,3) = 0$

Marginalfördelningar

Ofta är man intresserad av att titta på en av de två s.v. och se hur den är fördelad för sig.

Vi beräknar då **marginalfördelningarna** enligt:

$$f_X(x) = \sum_{j \in \Omega_Y} f_{XY}(x, y)$$

$$f_Y(y) = \sum_{x \in \Omega_X} f_{XY}(x, y)$$

Dvs. för att få den ena summerar man över den andra.

Exempel 2, forts.

Summera sannolikheterna radvis resp. kolumnvis:

$x \setminus y$	1	2	3	4	Σ
1	0,1	0	0	0	0,1
2	0,1	0,1	0	0	0,2
3	0,1	0,1	0,1	0	0,3
4	0,1	0,1	0,1	0,1	0,4
Σ	0,4	0,3	0,2	0,1	1

På nedersta raden har vi marginalfördelningen för Y och i sista kolumnen den för X .

Exempel 1, forts.

- X och Y är diskreta s.v. där

$$f_{XY}(x,y) = x/40,$$

$$\Omega_X = \Omega_Y = \{1,2,3,4\}$$

$$f_X(x) = \sum_{y=1}^4 \frac{x}{40} = \frac{x}{40} \sum_{y=1}^4 1 = \frac{x}{40} \cdot 4 = \frac{x}{10}$$

$$f_Y(y) = \sum_{x=1}^4 \frac{x}{40} = \frac{1}{40} \sum_{x=1}^4 x = \frac{1}{40} \cdot 10 = \frac{1}{4}$$

Exempel 1, forts.

- $f_{XY}(x, y) = 6e^{-3x-2y}$,
 x och $y \in (0, \infty)$

$$f_X(x) = 3e^{-3x},$$

$$x \in (0, \infty)$$

$$f_Y(y) = 2e^{-2y},$$

$$y \in (0, \infty)$$

Betingade fördelningar

På samma sätt som vi kunde betinga på en händelse när vi beräknade sannolikheten för en annan händelse kan vi manipulera en simultanfördelning för att få en **betingad fördelning**.

$$f_{X|Y}(x|y) = P(X = x \mid Y = y) =$$

$$= \frac{P(X = x \cap Y = y)}{P(Y = y)} = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Exempel 2, forts.

Ta fram de betingade fördelningarna för X givet att $Y = 3$ och för Y givet att $X = 3$:

$x \setminus y$	1	2	3	4	Σ
1	0,1	0	0	0	0,1
2	0,1	0,1	0	0	0,2
3	0,1	0,1	0,1	0	0,3
4	0,1	0,1	0,1	0,1	0,4
Σ	0,4	0,3	0,2	0,1	1

$$f_{X|Y}(x|3) = 0,10/0,20 = 1/2 \text{ för } x \in \{3,4\}$$

$$f_{Y|X}(y|3) = 0,10/0,30 = 1/3 \text{ för } y \in \{1,2,3\}$$

Exempel 1, forts.

- X och Y är diskreta s.v. där

$$f_{XY}(x, y) = x/40,$$

$$\Omega_X = \Omega_Y = \{1, 2, 3, 4\}$$

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{x/40}{1/4} = \frac{x}{10}$$

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(X)} = \frac{x/40}{x/10} = \frac{1}{4}$$

Repetition

Oberoende

Två s.v. är oberoende om

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

Och omvänt, om A och B är oberoende så är $P(A | B) = P(A)$.

Om A och B är oberoende så inses att följande gäller:

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

Oberoende s.v.

Två s.v. är oberoende om

$$f_{X|Y}(x|y) = f_X(x)$$

$$f_{Y|X}(y|x) = f_Y(y)$$

Om X och Y är oberoende så inses att följande gäller:

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$$

Betingade väntevärden

När man har den betingade fördelningen så kan vi ta fram det betingade väntevärdet och den betingade variansen.

Betrakta $f_{X|Y}(x|y)$ resp. $f_{Y|X}(y|x)$ som en vanlig frekvens- alt. täthetsfunktion och räkna på som "vanligt".

Obs! Kontrollera att du har koll på det betingade utfallsrummet!

Exempel 2, forts.

Vi hade $f_{X|Y}(x|3) = 1/2$ för $x \in \{3,4\}$

$x \setminus y$	1	2	3	4	Σ
1	0,1	0	0	0	0,1
2	0,1	0,1	0	0	0,2
3	0,1	0,1	0,1	0	0,3
4	0,1	0,1	0,1	0,1	0,4
Σ	0,4	0,3	0,2	0,1	1

$$E(X|Y=3) = \frac{1}{2} \cdot 3 + \frac{1}{2} \cdot 4 = 3,5$$

$$E(X^2|Y=3) = \frac{1}{2} \cdot 3^2 + \frac{1}{2} \cdot 4^2 = 12,5$$

$$V(X|Y=3) = E(X^2|Y=3) - E(X|Y=3)^2 = 12,5 - 12,25 = 0,25$$

En räkneregel till

Antag att X och Y är s.v. och a , b och c är konstanter.

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

$$V(aX + bY + c) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$$

Om X och Y är oberoende gäller

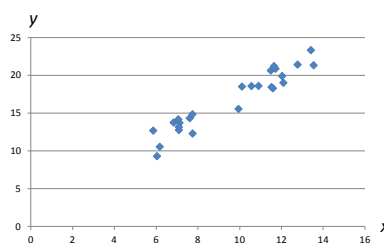
$$V(aX + bY + c) = a^2V(X) + b^2V(Y)$$

Mäta samband

I praktiken kan man observera att X och Y tycks **samvariera**.

När X ökar så ökar även Y och när Y minskar så minskar även X .

Ibland kan det vara omvänt, dvs. att när den ena ökar så minskar den andra.



Kovarians 1

Ett mått för beskriva hur X och Y samvarierar. Betrakta en s.v.

$$C = (X - \mu_X) \cdot (Y - \mu_Y)$$

där

$$\mu_X = E(X) \quad \text{och} \quad \mu_Y = E(Y)$$

Se resonemanget i avsnitt 6.7 sid 28 i Nyquist. Vad kan vi förvänta oss för beteende hos C för olika avstånd mellan X och Y och deras resp. väntevärden?

Kovarians 2

Vi definierar kovariansen mellan X och Y som väntevärdet av C :

$$\begin{aligned} Cov(X, Y) &= E(C) = \\ &= E[(X - \mu_X) \cdot (Y - \mu_Y)] = \\ &= E[(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y)] \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y \end{aligned}$$

där

$$\mu_X = E(X) \quad \text{och} \quad \mu_Y = E(Y)$$

Kovarians 3

Väntevärdet av produkten XY måste räknas fram:

$$E(XY) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot f_{xy}(x, y)$$

Exempel 2, forts.

$x \setminus y$	1	2	3	4	Σ
1	0,1	0	0	0	0,1
2	0,1	0,1	0	0	0,2
3	0,1	0,1	0,1	0	0,3
4	0,1	0,1	0,1	0,1	0,4
Σ	0,4	0,3	0,2	0,1	1

$$E(XY) = (1 \cdot 1 + 2 \cdot 1 + 2 \cdot 2 + 3 \cdot 1 + 3 \cdot 2 + 3 \cdot 3 + 4 \cdot 1 + 4 \cdot 2 + 4 \cdot 3 + 4 \cdot 4) \cdot 0,1 = 6,5$$

$$E(X) = 1 \cdot 0,1 + 2 \cdot 0,2 + 3 \cdot 0,3 + 4 \cdot 0,4 = 3$$

$$E(Y) = 1 \cdot 0,4 + 2 \cdot 0,3 + 3 \cdot 0,2 + 4 \cdot 0,1 = 2$$

$$\text{Cov}(X, Y) = 6,5 - 3 \cdot 2 = 0,5 \quad \leftarrow \text{Positivt tal Tolka!}$$

F9 forts. Kap 6 + 7

- Lite repetition + lite utveckling av
 - Kovarianser och linjära samband
- Sedan ...
 - Korrelation
- Och sedan ...
 - Några viktiga familjer av diskreta fördelningar

Repetition

Kovarians

Kovariansen mellan X och Y definieras som väntevärdet:

$$\begin{aligned} \text{Cov}(X, Y) &= \\ &= E[(X - \mu_X) \cdot (Y - \mu_Y)] = \\ &= E[(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y)] \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y \end{aligned}$$

där

$$\mu_X = E(X) \quad \text{och} \quad \mu_Y = E(Y)$$

Repetition

Kovarians

Väntevärdet av produkten XY måste räknas fram:

$$E(XY) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} xy \cdot f_{xy}(x, y)$$

$x \setminus y$	1	2	3	4	Σ
1	0,1	0	0	0	0,1
2	0,1	0,1	0	0	0,2
3	0,1	0,1	0,1	0	0,3
4	0,1	0,1	0,1	0,1	0,4
Σ	0,4	0,3	0,2	0,1	1

$$E(XY) = (1 \cdot 1 + 2 \cdot 1 + 2 \cdot 2 + 3 \cdot 1 + 3 \cdot 2 + 3 \cdot 3 + 4 \cdot 1 + 4 \cdot 2 + 4 \cdot 3 + 4 \cdot 4) \cdot 0,1 = 6,5$$

Korrelation

Kovariansen är ett mått på det **linjära sambandet** mellan två s.v. säg X och Y , hur de samvarierar.

Men värdet beror på med vilken skala man har mätt X resp. Y .

Om både X och Y antar stora (absoluta) värden kommer kovariansen också vara "stor".

Precis som deras resp. varianser kommer att vara "stora".

Exempel

Antag två par av s.v. (X, Y) och (S, T)

Antag att

$$\text{Cov}(X, Y) = 2 \text{ och } \text{Cov}(S, T) = 10$$

Det är inte meningsfullt att jämföra dessa rakt av utan att samtidigt ta hänsyn till deras respektive varianser.

Korrelation

Korrelationen:

ett **standardiserat mått** på det **linjära sambandet** mellan två s.v. säg X och Y , dvs. hur de **samvarierar linjärt**.

Definition:

$$\rho_{XY} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}}$$

Exempel, forts.

Antag två par av s.v. (X, Y) och (S, T)

$$\text{Cov}(X, Y) = 2, \quad V(X) = V(Y) = 2$$

- Verifiera att $\rho_{XY} = 2/2 = 1$

$$\text{Cov}(S, T) = 10, \quad V(S) = 16, \quad V(T) = 64$$

- Verifiera att $\rho_{ST} = 10/32 = 0,625$

Det är ett starkare linjärt samband mellan X och Y än mellan S och T .

Repetition

En räkneregel till

Antag att X och Y är s.v. och a , b och c är konstanter.

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

$$V(aX + bY + c) = a^2V(X) + b^2V(Y)$$

om de är oberoende. Annars gäller

$$V(aX + bY + c) = a^2V(X) + b^2V(Y) + 2ab\text{Cov}(X, Y)$$

Alltså ...

... om

X och Y oberoende $\Rightarrow \text{Cov}(X, Y) = 0$

Men det omvända gäller **inte** nödvändigtvis

$\text{Cov}(X, Y) = 0 \nRightarrow X$ och Y oberoende

Finns många exempel där X och Y är beroende men där $\text{Cov}(X, Y) = 0$.

Linjära kombinationer

Antag två s.v. X och Y . och två konstanter a och b . Skapelser av typen

$$Z = aX + bY$$

Kallas en **linjär kombination** av X och Y .

Om vi säg fixerar värdet på den ena av X och Y , säg $Y = y$, och skriver $bY = c$ så kan sambandet mellan X och Z skrivas

$$Z = aX + c$$

dvs. **ekvationen för en rät linje**.

($Z = aX + bY$ är ett plan i det 3-dimensionella rummet)

Kap 7. Några vanliga diskreta fördelningar

Bernoulli fördelningen:

Ett Bernoulli försök definieras som ett experiment med två möjliga utfall:

"Lyckat" = 1

"Ej lyckat" = 0

Sannolikhetsfunktion:

$$P(\text{"lyckat"}) = P(Z = 1) = p$$

$$P(\text{"ej lyckat"}) = P(Z = 0) = 1 - p$$

Bernoulli

Alt. om Z är Bernoulli fördelad kan vi skriva sannolikhetsfunktion:

$$f_Z(z) = p^z(1-p)^{1-z} \quad \text{Parameter}$$

för $z = 0, 1$ och där $0 < p < 1$.

Väntevärde och varians:

$$E(Z) = 0 \cdot p^0(1-p)^{1-0} + 1 \cdot p^1(1-p)^{1-1} = p$$

$$E(Z^2) = 0^2 \cdot p^0(1-p)^{1-0} + 1^2 \cdot p^1(1-p)^{1-1} = p$$

$$V(Z) = E(Z^2) - E(Z)^2 = p - p^2 = p(1-p)$$

Familjer av fördelningar

En fördel med att beskriva en s.v. med en **standardfördelning** (el. familj av fördeln.) är att vi kan beskriva egenskaperna utan att i förväg ha bestämt alla detaljerna.

(Alltså, utöver att de är vettiga modeller för att beskriva ganska vanlig situationer)

Genom att beskriva dem i allmänna termer med hjälp av s.k. **parametrar** så kan vi på förhand säga något om dem, t.ex. väntevärde och varians.

Väntevärde och varians blir då (enkla) funktioner av dessa parametrar

Binomialfördelningen 1

Antag en sekvens av n stycken **oberoende** Bernoulli försök,

$$(Z_1, Z_2, \dots, Z_n)$$

Definiera $X = \sum_i Z_i$ = "antalet lyckade utfall".

Utfallsrummet: av n försök kan man få $0, 1, \dots, n$ lyckade.

Vad är sannolikheten att få $X = x$?

Binomialfördelningen 2

Eftersom Bernoulliförsöken är oberoende kan vi skriva sannolikheten för varje given sekvens $(z_1, z_2, \dots, z_n)$

$$f_Z(z_1) \cdot f_Z(z_2) \cdot \dots \cdot f_Z(z_n)$$

Eller, om det nu blev x stycken ettor och därmed $n-x$ nollor

$$(p^1(1-p)^{1-1})^x \cdot (p^0(1-p)^{1-0})^{n-x} \\ = p^x \cdot (1-p)^{n-x}$$

(det spelar ju ingen roll i vilken ordning vi multiplicerar)

Binomialfördelningen 3

Men hur många sekvenser har exakt x stycken ettor och $(n-x)$ nollor?

Vi ska välja ut x stycken experiment av n , utan återläggning, där det blev en 1:a.

Kan göras på hur många sätt?

$$\binom{n}{x} = \binom{n}{n-x} = \frac{n!}{x!(n-x)!}$$

Och eftersom alla möjliga sekvenser är sinsemellan disjunkta adderas sannolikheterna.

Binomialfördelningen 4

Vi skriver $X \sim \text{Bin}(n, p)$ el. $X \in \text{Bin}(n, p)$
 Frekvensfunktion:

$$f_X(x) = \binom{n}{x} \cdot p^x (1-p)^{n-x}$$

för $x = 0, 1, \dots, n$ och där $0 < p < 1$ och n är ett heltal ≥ 0 .

Väntevärde: $E(X) = np$

Varians: $V(X) = np(1-p)$

(verkar det vettigt?)

Binomialfördelningen 4

Istället för att beräkna sannolikheterna används ofta en färdig tabell, typiskt med $P(X \leq x)$.

Men sådan tar typiskt inte med alla värden på p , typiskt upp till $p = 0,50$ i steg om $0,05$.

Skapa då $Y = \text{"antal ej lyckade"} = n - X$
 Y är då också binomialfördelad men med sannolikhet $1 - p$ isf p .

$$P(X \leq x) = P(Y \geq n-x) = 1 - P(Y \leq n-x-1)$$

Exempel

Låt $X \sim \text{Bin}(10; 0,7)$

Beräkna $P(X \leq 4) = P(X < 5)$

Definiera: $Y = 10 - X$; $X = 10 - Y$

$Y \sim \text{Bin}(10; 0,3)$

$P(X \leq 4) = P(Y \geq 10-4) = P(Y \geq 6) =$

$= 1 - P(Y < 6) = 1 - P(Y \leq 5)$

$\approx$ *[avläst i tabell]*

$= 1 - 0,95265 = \mathbf{0,04735}$

Multinomialfördelning

Utökning av binomialfördelningen.

Istället för två möjliga utfall har man säg C stycken (disjunkta) möjliga utfall.

Y_j = antal försök med utfall j

Sannolikheten för utfall j i ett enskilt försök är p_j och $j = 1, 2, \dots, C$

$$\sum_{j=1}^C Y_j = n \qquad \sum_{j=1}^C p_j = 1$$

Obs! Fel på sidan 11

Poissonfördelningen 1

Låt $X \sim \text{Bin}(n, p)$ och låt $n \rightarrow \infty$ och $p \rightarrow 0$ på ett sådant sätt att $np = \lambda$, en konstant, dvs. $p = \lambda/n$.

När $n \rightarrow \infty$ blir binomialfördelningen en **Poissonfördelning**:

$$X \sim \text{Po}(\lambda) \qquad f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

för $x = 0, 1, 2, \dots$ och där $\lambda > 0$.

Väntevärde: $E(X) = \lambda$

Varians: $V(X) = \lambda$

Approximera Bin med Po

Resonemanget var att om n var stort och p litet så "blev" det en Poissonfördelning av en Binomial.

- En Poisson är enklare att använda beräkningsmässigt.
- Eller om Binomialtabellen inte har större värden än säg $n = 19$

Alltså, om $n (\geq 20)$ är stort och p litet ($\leq 0,1$) sätt $\lambda = np$ och använd Poisson.

Exempel

Låt $X \sim \text{Bin}(100; 0,05)$

Beräkna $P(X \leq 4)$

Sätt $\lambda = np = 5$ och använd $Po(5)$

$$P(X \leq 4) = P(X=0) + P(X=1) + P(X=2) \\ + P(X=3) + P(X=4)$$

$$= \frac{5^0}{0!} e^{-5} + \frac{5^1}{1!} e^{-5} + \frac{5^2}{2!} e^{-5} + \frac{5^3}{3!} e^{-5} + \frac{5^4}{4!} e^{-5}$$

$$= e^{-5} \left(1 + 5 + \frac{5^2}{2} + \frac{5^3}{6} + \frac{5^4}{24} \right) \approx 0,4405$$

**Obs! Räkna exakt med Binomial ger 0,4360;
ett fel på 0,0045 eller ca 1 %**