Kapitel 5

Multivariata sannolikhetsfördelningar

Multivariata sannolikhetsfördelningar

En slumpvariabel som, när slumpförsöket utförs, antar *exakt* ett värde sägs vara en *univariat* slumpvariabel.

En slumpvariabel som, när slumpförsöket utförs, antar *exakt* två värden sägs vara en *bivariat* slumpvariabel. Vi betecknar (X_1,X_2) och när försöket utförs fås $(X_1=x_1, X_2=x_2)$.

Mer generellt fås en sk *k-variat* slumpvariabel som betecknas $(X_1,X_2,\ldots,X_k)$ och när försöket utförs fås $(X_1=x_1, X_2=x_2,\ldots,X_k=x_k)$

Vi pratar mer allmänt om *multivariata slumpvariabler* och *multivariata sannolikhetsfördelningar*.

Simultan sannolikhetsfunktion

Den *simultana* sannolikhetsfunktionen för två (univariata) diskreta slumpvariabler X_1 och X_2 ges av

$$p(x_1, x_2) = \Pr(X_1 = x_1, X_2 = x_2)$$

för alla (x_1, x_2) som kan förekomma.

För en simultan sannolikhetsfunktion gäller att

$$\mathbf{1}.\ 0 \leq p(x_1, x_2) \leq 1, \ \text{för alla} \ (x_1, x_2)$$

Uppgift 5.3

Av nio anställda i ett företag är fyra gifta, tre aldrig gifta och två skilda. Tre av de anställda ska befordras. Låt

$$\begin{aligned} Y_1 &= \text{Antal gifta bland de befordrade} \\ Y_2 &= \text{Antal av de aldrig gifta bland de befordrade} \end{aligned}$$

Den simultana sannolikhetsfunktionen för Y_1 och Y_2 ges av

$$p(y_1, y_2) = \frac{\binom{4}{y_1}\binom{3}{y_2}\binom{2}{3-y_1-y_2}}{\binom{9}{3}},$$

$$y_1 = 0, 1, 2, 3,$$

Simultan fördelningsfunktion

Den *simultana* fördelningsfunktionen för två (univariata) diskreta slumpvariabler X_1 och X_2 ges av

$$F(x_1, x_2) = \Pr(X_1 \leq x_1, X_2 \leq x_2)$$

Om det för två (univariata) *kontinuerliga* slumpvariabler X_1 och X_2 existerar en funktion $f(x_1,x_2)$ sådan att den simultana fördelningsfunktionen kan skrivas som

$$F(x_1, x_2) = \Pr(X_1 \leq x_1, X_2 \leq x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2)\, dt_2 dt_1$$

sägs $f(x_1,x_2)$ vara den simultana *täthetsfunktionen* för X_1 och X_2.

Egenskaper hos bivariata kontinuerliga slumpvariabler

För den *simultana* fördelningsfunktionen $F(x_1,x_2)$ för två (univariata) kontinuerliga slumpvariabler X_1 och X_2 gäller att

$$\mathbf{1}. F(-\infty,-\infty) = F(-\infty,x_2) = F(x_1,-\infty) = 0$$

$$\mathbf{2}. F(\infty,\infty) = 1$$

$$\mathbf{3}. F(x_1,x_2) \text{ är en växande funktion}$$

För den *simultana* täthetsfunktionen $f(x_1,x_2)$ för två (univariata) kontinuerliga slumpvariabler X_1 och X_2 gäller att

$$\mathbf{1}. f(x_1,x_2) \geq 0 \text{ för alla } (x_1,x_2)$$

$$\mathbf{2}. \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2)\,dx_1 dx_2 = 1$$

Uppgift 5.15b

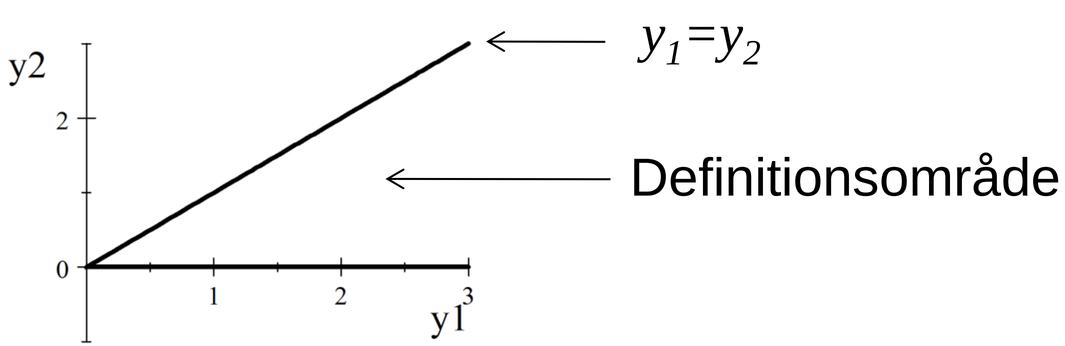
Vi studerar nu kö- och kassatid i en snabbmatsrestaurang. Låt

Y_1 = Total servicetid
Y_2 = Kötid

Den simultana tätheten ges av

$$f(y_1, y_2) = e^{-y_1}, \ 0 \leq y_2 \leq y_1 < \infty$$

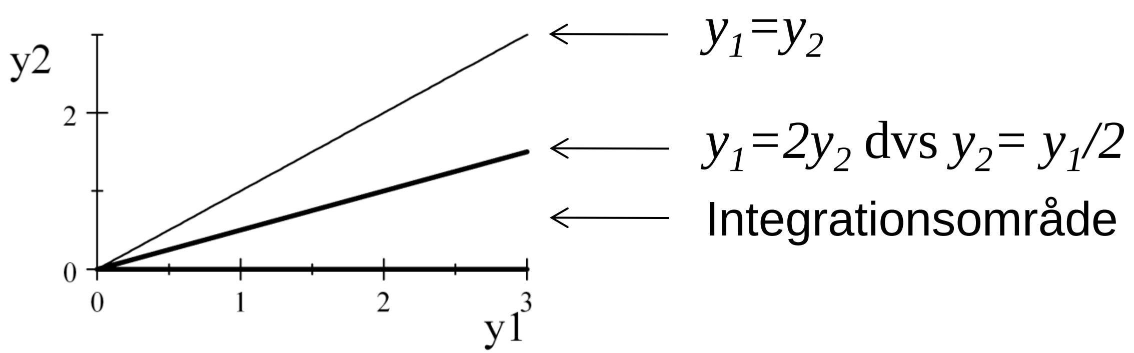
Vi börjar med att åskådliggöra definitionsområdet grafiskt.



UPPSALA UNIVERSITET

Uppgift 5.15b

Vi söker nu $Pr(Y_1 \geq 2Y_2)$ och för att finna integrationsgränserna börjar vi med att åskådliggöra integrationsområdet grafiskt.



Detta innebär att den sökta sannolikheten kan beräknas via

$$\Pr\left(Y_1 \geq 2Y_2\right)$$

Den avslutande integralen blir 1 eftersom det är väntevärdet för en Exp(1).

Marginalfördelningar

Låt de diskreta slumpvariablerna X_1 och X_2 ha simultan sannolikhetsfunktion $p(x_1,x_2)$. Vi kan då alltid finna (den univariata) *sannolikhetsfunktionen* för X_1 respektive X_2 via

$$p_1(x_1) = \sum_{\text{alla } x_2} p(x_1, x_2), \quad \text{och} \quad p_2(x_2) = \sum_{\text{alla } x_1} p(x_1, x_2)$$

Låt de kontinuerliga slumpvariablerna X_1 och X_2 ha simultan täthetsfunktion $f(x_1,x_2)$. Vi kan då alltid finna (den univariata) *täthetsfunktionen* för X_1 respektive X_2 via

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2, \quad \text{och} \quad f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1$$

Uppgift 5.21a

Utifrån Uppgift 5.3 följer att marginalfördelningen för Y_1 ges av

$$p_{Y_1}(y_1)$$

Sannolikhetsfunktion
För Hyp($3-y_1$,3,5).

Således följer (som väntat) att Y_1 är Hyp(3,4,9).

UPPSALA UNIVERSITET

Uppgift 5.33a

Bestäm marginaltätheterna för Y_1 och Y_2 i Uppgift 5.15.
För fixt värde y_1 följer att $0 \leq y_2 \leq y_1$ varför det följer att

$$f_1(y_1)$$

vilket innebär att Y_1 är *Ga(2,1)*. På motsvarande sätt följer att

$$f_2(y_2)$$

vilket innebär att Y_2 är *Ga(1,1)*, eller *Exp(1)*.

Betingade fördelningar

Låt de diskreta slumpvariablerna X_1 och X_2 ha simultan sannolikhetsfunktion $p(x_1,x_2)$ *samt* marginalfördelningarna $p_1(x_1)$ och $p_2(x_2)$. Den *betingade* sannolikhetsfunktionen för X_1 givet $X_2=x_2$ ges då via

$$p\left(x_1 \mid x_2\right) = \Pr\left(X_1 = x_1 \mid X_2 = x_2\right)$$

Låt de kontinuerliga slumpvariablerna X_1 och X_2 ha simultan täthetsfunktion $f(x_1,x_2)$ *samt* marginalfördelningarna $f_1(x_1)$ och $f_2(x_2)$. Den *betingade* täthetsfunktionen för X_1 givet $X_2=x_2$ ges då via

$$f\left(x_1 \mid x_2\right) = \frac{f\left(x_1, x_2\right)}{f_2\left(x_2\right)}$$

Uppgift 5.33bc

Bestäm den betingade fördelningen för Y_1 givet att $Y_2 = y_2$. För fixt värde y_2 följer att $0 \leq y_2 \leq y_1$ varför det följer att

$$f(y_1 \mid y_2)$$

På motsvarande sätt följer att den betingade fördelningen för Y_2 givet att $Y_1 = y_1$ ges av

$$f(y_2 \mid y_1)$$

Uppenbarligen gäller att $Y_2 \mid Y_1 = y_1$ är $U(0,y_1)$.

Oberoende slumpvariabler

Låt de diskreta slumpvariablerna X_1 och X_2 ha simultan sannolikhetsfunktion $p(x_1,x_2)$ *samt* (marginal)sannolikhetsfunktionerna $p_1(x_1)$ och $p_2(x_2)$.
X_1 och X_2 är *oberoende* om och endast om

$$p(x_1, x_2) = p_1(x_1)\, p_2(x_2), \quad \text{för alla } (x_1, x_2)$$

Låt de kontinuerliga slumpvariablerna X_1 och X_2 ha simultan täthetsfunktion $f(x_1,x_2)$ *samt* marginaltätheterna $f_1(x_1)$ och $f_2(x_2)$.
X_1 och X_2 är *oberoende* om och endast om

$$f(x_1, x_2) = f_1(x_1)\, f_2(x_2), \quad \text{för alla } (x_1, x_2)$$

Oberoende slumpvariabler

Det är ofta besvärligt att via definitionen undersöka huruvida två slumpvariabler X_1 och X_2 är oberoende. Då har man glädje av följande resultat

<u>X_1 och X_2 är oberoende om och endast om</u>

1. …definitionsområdet är *rektangulärt*, dvs det finns konstanter a, b, c och d sådana att $a \leq x_1 \leq b$, $c \leq x_2 \leq d$, <u>och</u>

2. …den simultana funktionen kan skrivas som en produkt av två funktioner $g(x_1)$ och $h(x_2)$ där g är en funktion som enbart beror på x_1 och h är en funktion som enbart beror på x_2.

UPPSALA UNIVERSITET

Uppgift 5.62

Person A singlar en slant tills han för första gången får krona. Sedan singlar en annan person, B, samma slant tills hon för första gången får krona.

Antag att sannolikheten att få krona (i en viss singling) är *p*.

Bestäm sannolikheten att A och B singlar slanten lika många gånger.

Låt Y_1 och Y_2 representera det antal singlingar A respektive B behöver.

Vidare är det rimligt att förutsätta att Y_1 och Y_2 är oberoende.

Uppgift 5.62

Sannolikheten att de båda behöver singla slanten ett visst givet, *k*, antal gånger ges av

$$p\left(k, k\right)$$

Vi kräver dock inte ett visst givet antal kast utan enbart att det är lika många för A och B. Den sökta sannolikheten blir därför

$$\Pr\left(Y_1 = Y_2\right)$$

Uppgift 5.68

En buss går <u>en</u> gång och vi som passagerare känner till den aktuella timmen, dvs vi vet när timmen börjar och slutar. Låt

Y_1 = *Bussens ankomsttid*

Y_2 = *Passagerarens ankomsttid*

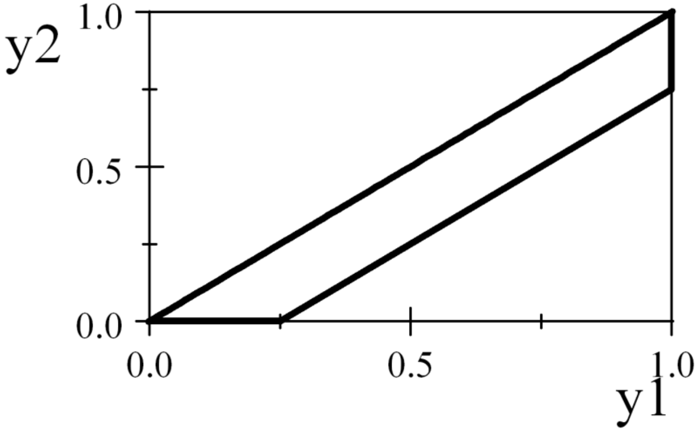
Utifrån förutsättningarna följer att Y_1 och Y_2 är *oberoende* och *U(0,1)*-fördelade slumpvariabler vilket innebär att den simultana täthetsfunktionen ges av

$$f(y_1, y_2) = 1,\ 0 \leq y_1 \leq 1,\ 0 \leq y_2 \leq 1$$

Uppgift 5.68

Om vi betecknar den sökta sannolikheten *P* följer att

$$P = \Pr\left(y_2 \leq Y_1 \leq y_2 + 1/4\right)$$



Vi börjar med att beskriva definitions- och integrationsområdet grafiskt.

Vi ser att integrationsområdet måste delas upp i två delar vilket ger oss uttrycket

$$P = \int_0^{1/4}\int_0^{y_1} dy_2 dy_1 + \int_{1/4}^{1}\int_{y_1-1/4}^{y_1} dy_2 dy_1$$

Väntevärden för funktioner av slumpvariabler

Vi är ofta intresserade av att studera funktioner av slumpvariabler.

Låt X_1 och X_2 ha simultan sannolikhetsfunktion $p(x_1,x_2)$ (eller simultan täthetsfunktion $f(x_1,x_2)$) och betrakta funktionen $g(X_1,X_2)$. Väntevärdet för denna funktion är

$$E\left[g\left(X_1,X_2\right)\right] = \sum_{\text{alla } x_1}\sum_{\text{alla } x_2} g\left(x_1,x_2\right)p\left(x_1,x_2\right)$$

Väntevärden för funktioner av slumpvariabler

1. Om en funktion är komplicerad kan den eventuellt brytas ner i mer lätthanterliga delar. Antag att

$$g(X_1, X_2) = c_1 g_1(X_1, X_2) + c_2 g_2(X_1, X_2)$$

Då följer att

$$E[g(X_1, X_2)] = c_1 E[g_1(X_1, X_2)] + c_2 E[g_2(X_1, X_2)]$$

2. Antag att X_1 och X_2 är oberoende och att $g(X_1)$ och $h(X_2)$ är funktioner som enbart beror på X_1 respektive X_2. Då gäller att

$$E[g(X_1)h(X_2)]$$

Uppgift 5.81

Låt Y_1 och Y_2 representera livslängden (i 100-tals timmar) för komponenter av typ I respektive II. Den simultana tätheten är

$$f\left(y_1, y_2\right) = \frac{1}{8} y_1 e^{-(y_1+y_2)/2}, \; y_1 \geq 0, \; y_2 \geq 0$$

Vi är nu intresserade av att mäta den *relativa effektiviteten*, vilket är kvoten Y_2/Y_1, och beräkna väntevärdet för denna.

Eftersom integrationsområdet är rektangulärt samtidigt som

$$f\left(y_1, y_2\right) = \frac{1}{8} y_1 e^{-(y_1+y_2)/2}$$

följer att Y_1 och Y_2 är oberoende.

Uppgift 5.81

Det sökta väntevärdet blir därmed

$$E\left(\frac{Y_2}{Y_1}\right)$$

Täthetsfunktion
för Exp(2).

Väntevärde
för Exp(2).

Kovarians och korrelation

För att mäta graden av (linjärt) samband mellan två slumpvariabler X_1 och X_2 används *kovarians* och *korrelation*.

$$\sigma_{X_1X_2} = Cov\left(X_1, X_2\right) = E\left[\left(X_1 - \mu_1\right)\left(X_2 - \mu_2\right)\right]$$

Vid manuella beräkningar av kovariansen används med fördel att

$$Cov\left(X_1, X_2\right) = E\left(X_1X_2\right) - E\left(X_1\right)E\left(X_2\right)$$

Då X_1 och X_2 är *oberoende* finns inget samband varför $Cov(X_1,X_2)=0$.

Kovariansen är skalberoende. För att få ett linjärt sambandsmått som inte är skalberoende beräknas *korrelationskoefficienten*.

$$\rho_{X_1,X_2} = \frac{Cov\left(X_1, X_2\right)}{\sqrt{V\left(X_1\right)}\sqrt{V\left(X_2\right)}} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}$$

Kovarians: Bevis för beräkningsformel

Räkneregler för väntevärden ger att

$$Cov\left(X_1, X_2\right)$$

Väntevärden för linjärkombinationer av slumpvariabler

Låt $a_1, a_2, \ldots, a_n$ vara konstanter, $Y_1, Y_2, \ldots, Y_n$ vara slump-variabler med väntevärden $\mu_1, \mu_2, \ldots, \mu_n$. Då sägs

$$U = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n = \sum_{i=1}^{n} a_iY_i$$

vara en linjärkombination av $Y_1, Y_2, \ldots, Y_n$. För U gäller att

$$E(U) = \sum_{i=1}^{n} a_iE(Y_i)$$

Trinomialfördelningen

Betrakta en urna med vita och svarta bollar där andelen vita bollar är p. Då vi *med* återläggning drar n bollar ur urnan och låter X räkna hur många av bollarna som blev vita gäller att X är $Bi(n,p)$.

Låt nu urnan innehålla vita, *röda* och svarta bollar där andelen vita och röda bollar är p_1 respektive p_2. Då vi *med* återläggning drar n bollar ur urnan och låter X_1 räkna vita bollar och X_2 räkna röda bollar fås den simultana sannolikhetsfunktionen för X_1 och X_2 till

$$p(x_1, x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}, \ 0 \leq x_1 + x_2 \leq n, 0 \leq p_1 + p_2 \leq 1$$

Då vi på detta sätt simultant studerar två binomialfördelade slumpvariabler uppstår den sk *trinomialfördelningen*.

Multinomialfördelningen

Genom att även räkna de svarta bollarna skapar vi indirekt i binomial- och trinomialfördelningen ännu en slumpvariabel.

Binomialfallet: $X_2 \quad = \quad \text{Antal svarta bollar i urvalet} = n - X_1$

Trinomialfallet:

Drag med återläggning *n* bollar ur en urna som innehåller bollar i *k* olika färger och låt $X_1, X_2, \ldots, X_k$ räkna hur många bollar i respektive färg som kommer med i urvalet.

$$p\left(x_1, x_2, \ldots, x_k\right) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_h}, \ \sum x_i = n, \ \sum p_i = 1$$

Bivariat normalfördelning

Låt Y_1 och Y_2 ha den simultana täthetsfunktionen

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}Q\right\}$$

där

$$Q = \frac{1}{(1-\rho^2)}\left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right]$$

Vidare gäller att $-\infty \leq y_1 \leq \infty$, $-\infty \leq y_2 \leq \infty$. Detta sägs vara en bivariat normalfördelning med parametrar μ_1, σ_1, μ_2, σ_2 och ρ

Det går att visa att Y_1 är $N(\mu_1, \sigma_1)$ och Y_2 är $N(\mu_2, \sigma_2)$. Vidare gäller att ρ är korrelationskoefficienten mellan Y_1 och Y_2.

Hierarkiska modeller

Låt X vara en slumpvariabel vars sannolikhetsfördelning beror på en parameter θ, dvs vi känner den *betingade* fördelningen $f(x|\theta)$

Antag vidare att även θ är en slumpvariabel med fördelning $f_2(\theta)$.

I dessa sammanhang pratar vi om en *hierarkisk modell*.

Vi är nu intresserade av att finna (den obetingade) marginalfördelningen för X. För att göra detta går vi tillväga i två steg.

1. Först bestämmer vi den simultana fördelningen via $f(x,\theta)= f(x|\theta)f_2(\theta)$.

2. Sedan finner vi marginalfördelningen, $f_1(x)$, för X genom att integrera (eller summera) bort θ.

Uppgift 5.42 (avsnitt 5.3)

Antal defekter på en yard av ett visst tyg, *Y*, kan ses som en poissonfördelad slumpvariabel med intensitet λ.

Dock gäller att denna intensitet λ inte är känd utan själv en slumpvariabel. Vår modell anger att λ är *Exp(1)*. Alltså gäller att

$$p\left(y \mid \lambda\right) = \frac{\lambda^y e^{-\lambda}}{y!}, \; y = 0, 1, 2, \ldots$$

Vi söker nu marginalfördelningen, *p(y)*, för *Y*. Utifrån denna kan vi sedan beräkna väntevärde och varians för *Y*.

Uppgift 5.42 (avsnitt 5.3)

$p_Y(y)$

En antydan till täthet för *Ga(y+1, 1/2)*.

Täthetsfunktion för *Ga(y+1, 1/2)*.

Alltså gäller att *Y+1* är *Ge(1/2)* vilket betyder att *E(Y)=1* och *V(Y)=2*.

Betingade väntevärden

Vi är intresserade av om $E(X_1)$ och $V(X_1)$ kan beräknas utifrån det vi känner till, dvs $f(x_1 \mid x_2)$ och $f_2(x_2)$, så att vi inte behöver bestämma marginalfördelningen $f_1(x_1)$.

En betingad sannolikhetsfördelning uppfyller alla krav på en sannolikhetsfördelning och således kan vi beräkna väntevärden. Vi får

$$E\left(g\left(X_1\right) \mid X_2 = x_2\right) = \begin{cases} \int_{-\infty}^{\infty} g\left(x_1\right) f\left(x_1 \mid x_2\right) dx_1, & \text{i det kontinuerliga fallet} \\ \\ \sum_{\text{alla } x_1} g\left(x_1\right) p\left(x_1 \mid x_2\right), & \text{i det diskreta fallet} \end{cases}$$

Betingade väntevärden

Låt oss studera det betingade väntevärdet $E(X_1 | X_2=x_2)$. Detta betingade väntevärde kan ses som en funktion av x_2, dvs

$$h(x_2) = E(X_1 \mid X_2 = x_2)$$

Om vi låter x_2 variera över alla sina värden kan vi således se det som en funktion av slumpvariabeln X_2.

$$h(X_2) = E(X_1 \mid X_2)$$

En funktion av slumpvariabeln X_2 är i sig en slumpvariabel och intressanta saker händer när vi beräknar väntevärde och varians för $h(X_2)$.

Betingade väntevärden

Vad gäller då för $E[h(X_2)]$? I det diskreta fallet får vi att

$$E\left[E\left(X_1 \mid X_2\right)\right]$$

Betingade väntevärden

Låt X_1 och X_2 vara två slumpvariabler. Då gäller alltså att

$$E(X_1) = E\left[E(X_1 \mid X_2)\right]$$

I högerledet gäller det inre väntevärdet den betingade fördelningen $f(x_1 \mid x_2)$. Medan det yttre väntevärdet gäller $f_2(x_2)$.

Vidare gäller att

$$V(X_1) = E\left[V(X_1 \mid X_2)\right] + V\left[E(X_1 \mid X_2)\right]$$

Uppgift 5.136 a

Betrakta

Y = Antal defekta enheter per yard

λ = Den intensitet med vilken defekta enheter uppkommer

Enligt förutsättningarna gäller att *Y|λ* är *Po(λ)* och *λ* är *Exp(1)*.

Egenskaper hos *Po(λ)* ger att

$$E\left(Y \mid \lambda\right) = \lambda$$

vilket innebär att

$$E\left(Y\right)$$

Uppgift 5.136 b

Eftersom egenskaper hos *Po(λ)* vidare ger *V(Y|λ)*= *λ* följer att

$$E\left[V\left(Y \mid \lambda\right)\right]$$

I a-uppgiften fann vi att *E(Y|λ)*= *λ* varför vi därmed får att

$$V\left[E\left(Y \mid \lambda\right)\right]$$

Slutligen följer därmed att

$$V\left(Y\right)$$