

Föreläsning 7: Punktskattningar

Matematisk statistik

David Bolin
Chalmers University of Technology
September 24, 2018



Marginalfördelningar

Givet en diskret tvådimensionell slumpvariabel (X, Y) definieras marginalfördelningarna för X och Y som

$$f_X(i) = \sum_{j=-\infty}^{\infty} f_{X,Y}(i, j)$$
$$f_Y(j) = \sum_{i=-\infty}^{\infty} f_{X,Y}(i, j)$$

I det kontinuerliga fallet definieras marginalfördelningarna för X och Y på motsvarande sätt som

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

Tvådimensionella fördelningar

Definition

En tvådimensionell slumpvariabel (X, Y) tillordnar två numeriska värden till varje utfall i Ω .

För diskreta variabler har vi sannolikhetsfunktionen

$$f_{X,Y}(i, j) = P(X = i, Y = j).$$

Där $f_{X,Y}(i, j) \geq 0$ och $\sum_{i,j} f_{X,Y}(i, j) = 1$. För kontinuerliga variabler har vi en täthetsfunktion $f_{X,Y}(x, y)$ som är sådan att

- 1 $f_{X,Y}(x, y) \geq 0$,
- 2 $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$, och
- 3 $P(a \leq X \leq b \text{ och } c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dx dy$.

Väntevärde

Givet en tvådimensionell slumpvariabel (X, Y) definieras väntevärdena för X och Y som

$$E(X) = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} i f_{X,Y}(i, j), & \text{om } X \text{ diskret,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy, & \text{om } X \text{ kontinuerlig,} \end{cases}$$

och

$$E(Y) = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} j f_{X,Y}(i, j), & \text{om } Y \text{ diskret,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y) dx dy, & \text{om } Y \text{ kontinuerlig.} \end{cases}$$

Notera att vi också kan formulera väntevärdena med marginalfördelningarna för X och Y . Till exempel

$$E(X) = \sum_{i,j} i f_{X,Y}(i, j) = \sum_{i=-\infty}^{\infty} i \sum_{j=-\infty}^{\infty} f_{X,Y}(i, j) = \sum_{i=-\infty}^{\infty} i f_X(i)$$

Betingade fördelningar

Den betingade fördelningen för X givet $Y = y$ definieras som

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

givet att $f_Y(y) > 0$. På motsvarande sätt har vi den betingade fördelningen för Y givet $X = x$:

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

Oberoende slumpvariabler

Två slumpvariabler X och Y sägs vara oberoende om deras täthetsfunktion (eller sannolikhetsfunktion) kan skrivas som produkten av marginalfördelningarna. Det vill säga

$$f_{X,Y}(i, j) = f_X(i)f_Y(j)$$

Repetition

David Bolin

Kovarians

Kovarians

Kovariansen mellan X och Y definieras som $C(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ där $\mu_X = E(X)$ och $\mu_Y = E(Y)$.

- Enligt definitionen kan vi skriva kovariansen som

$$C(X, Y) = \begin{cases} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} (i - \mu_X)(j - \mu_Y) f_{X,Y}(i, j), & \text{diskret} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y), & \text{kont.} \end{cases}$$

- Notera att vi har att $C(X, X) = V(X)$.
- $C(X, Y)$ kan beräknas som $C(X, Y) = E(XY) - E(X)E(Y)$.
- Om X och Y oberoende: $C(X, Y) = 0$, $E(XY) = E(X)E(Y)$.
- För två slumpvariabler X och Y , och två tal a och b har vi

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

Repetition

David Bolin

Korrelation och oberoende

Korrelation

Den så kallade korrelationskoefficienten definieras som

$$\rho(X, Y) = \frac{C(X, Y)}{\sqrt{V(X)V(Y)}}.$$

- Detta mått beskriver linjär samvariation mellan X och Y .
- Vi har att $-1 \leq \rho \leq 1$.
- Vi säger att X och Y är okorrelerade om $\rho(X, Y) = 0$.

Vi har följande samband mellan beroende och korrelation:

- Om X och Y är oberoende så är de okorrelerade.
- Om X och Y är okorrelerade så behöver de inte vara oberoende.

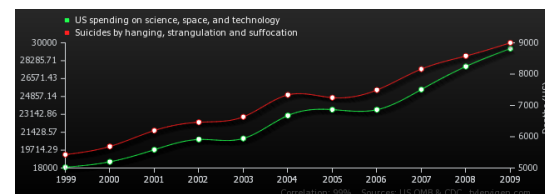
Dessa samband är naturliga eftersom två slumpvariabler är oberoende om det inte finns någon samvariation alls mellan dem, medan de är okorrelerade om det saknas *linjär* samvariation.

Repetition

David Bolin

Korrelation och kausitet

- Korrelation säger inget om orsakssamband (kausalitet)!
- Vi kan också ha att all samvariation kan förklaras av en tredje icke uppmätt variabel.
- Två exempel:
 - Dagar med hög glassförsäljning tenderar att ha fler drunkningsolyckor. Dags att förbjuda glass?
 - US spending on science, space, and tech. correlates ($\rho = 0.99$) with Suicides by hanging, strangulation and suffocation¹

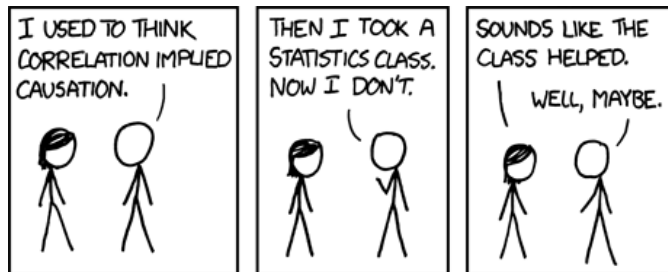


¹Källa: http://www.tylervigen.com/view_correlation?id=1597

Repetition

David Bolin

- Kausala frågor handlar om mekanismer som genererar datan eller prediktioner om vad som kommer ske efter någon åtgärd.
- Tex: Förändras antalet drunkingsolyckor om vi förbjuder glass?
- Det finns mängder av kausala påståenden i nyheterna!
 - "Ät frukost om du vill minska din risk för kranskärslsjukdomar"
- Vi måste vara försiktiga med kausala påståenden:
 - Om personer som äter frukost tenderar ha färre kranskärslsjukdomar, borde vi därför äta frukost?
- Vi måste veta hur datan samlas in för att svara på kausala frågor! Vi återkommer till detta senare.



Statistikteori

Antag att vi har fått observationer $x_1, \dots, x_n$ från en viss fördelning. Givet dessa observationer, vad kan vi säga om parametrarna i fördelningen?

Vi kommer titta på tre olika frågeställningar:

punktskattning Vad är ett troligt värde för parametrarna?

konfidensintervall Hur säker är denna skattning? För att

karaktisera osäkerheten vill vi bilda ett intervall som med hög sannolikhet ska täcka det sanna värdet på parametern.

hypotesprövning Undersök hypoteser kring parametrarna, till exempel kan vi vilja veta om $\mu > 0$ i en normalfördelning.

Senare kommer vi också titta på jämförelser mellan olika stickprov.

🎧 Minskad risk för stroke hos ammande mödrar

2:24 min [Min lista](#) [8](#)

Publicerat kl 06.00

Kvinnor som har ammat verkar ha en minskad risk att drabbas av stroke senare i livet, visar en observationsstudie från USA.

I Studien har data från mer än 80 000 kvinnor analyserats. Resultaten visar att amning verkar ha en skyddande effekt mot stroke, men vad som ligger bakom den minskade risken för stroke vet man ännu inte.

– Man kan spekulera i att det är andra faktorer än amningen, att amningen är en markör för att man kanske lever lite sundare. Rör på sig mer, kollar sitt blodtryck oftare, säger Erik Lundström, som är överläkare i neurologi och universitetslektor i Uppsala.

Men det skulle också kunna ha med hormonerna att göra, säger han.

– Vi vet ju att amningen gör att man får lägre risk för hjärt-kärlsjukdomar. Det finns ett slags samband där, men vi vet inte varför.

Viktoria Elestrand är ute och går med sin en vecka gamla bebis Björn, och hon har valt att amma.

– Det är väldigt intressant att veta att amningen kan ha sådana hälsoeffekter så klart. Men anknytningen till barnet, alla positiva effekter som det kan ha för en fyrfödd var anledningen till att jag valde att amma, säger hon.

Stickprov

Vi kallar observationerna vi baserar vår analys på för ett stickprov.

Definition: Stickprov

Ett stickprov av storlek n är n oberoende observationer av en slumpvariabel X . Vi kan alltså skriva stickprovet som observationer av n slumpvariabler $X_1, \dots, X_n$ där alla X_i är oberoende och likafördelade (alltså har samma fördelning).

Notation

Vi använder små bokstäver för mätningarna $(x_1, \dots, x_n)$, dessa är alltså numeriska värden. Vi använder stora bokstäver för motsvarande slumpvariabler $(X_1, \dots, X_n)$.

Populationer

- Ofta tolkar vi stickprovet som en dragning ur en stor population som vi vill få kunskap om. Begreppet population kan innebära lite olika saker, vi kan skilja på två typfall:
 - En väl avgränsad mängd saker, där vi i teorin räkna upp alla ingående element i en lista.
 - Något lite mer abstrakt utan någon konkret avgränsning. Till exempel kan vi ha en process av vilken vi gör upprepade mätningar. Populationen kan då ses som alla möjliga slumpmässiga försök av denna typ.
- Vår definition av stickprov täcker det andra fallet och kan oftast ses som en bra approximation i det första fallet om n är litet i förhållande till storleken på populationen.
- Det är viktigt att vårt stickprov är *representativt* för populationen, vilket betyder att det är någorlunda typiskt för populationen. Om detta inte är fallet kan vi få systematiska fel i våra skattningar.