Datorövning 2 - 732G01/732G40

HT2017

Introduktion

Denna datorövning behandlar olika sannolikhetsfördelningar och hur man kan använda Minitab för att underlätta beräkningar.

Övningen ska göras i Minitab. Minitab finns på datorerna i PC1-5 som vissa är bokade för dessa pass men Minitab finns också att ladda ner via Studentportalen både till Windows och Mac-datorer. Följ instruktionerna ni hittar där.

Uppgift 1

Grobarheten hos en viss typ av frön är 60% och vi planterar tre frön under samma förutsättningar. Vi kan då betrakta antalet frön som gror som $X \sim bin(n = 3, \pi = 0.6)$ eftersom delförsöken, de olika fröna, är oberoende av varandra och varje frö kan antingen gro eller inte, dvs. Bernoulli-försök.

a)

Börja med att skriva in värdena $0 - 3$ i kolumnen **C1** och namnge kolumnen **x**. Denna kolumn benämner alla utfall som X kan anta, dvs. hur många av de tre frön som planteras som också gror.

Beräkna sannolikheterna för varje utfall med en miniräknare och skriv in resultaten i kolumnen **C2** som benämns med **p(x)**.

Visualisera denna sannolikhetsfördelning med ett stapeldiagram. Notera att vi nu har höjden på stapeln som värden i en kolumn istället för att behöva låta Minitab räkna antalet förekomster av olika utfall. Detta innebär att i *Graph -> Bar Charts* måste vi ändra *Bars represent* till **Values from a table** och sedan *Simple*. Lägg in kolumnen med sannolikheterna, **p(x)**, i *Graph variables* och kolumnen med utfall, **x**, i *Categorical variable*.

b)

Anta nu att vi planterar 10 frön under samma förutsättningar. Då kommer $X \sim bin(n = 10, \pi = 0.6)$ vilket innebär ganska många beräkningar för att få fram hela sannolikhetsfördelningen.

För att vi ska kunna använda Minitabs funktioner för att beräkna denna fördelning måste vi skapa en kolumn med alla utfall. Skriv in värdena $0 - 10$ i **C3** och ge den ett lämpligt namn. (OBS! Minitab tillåter inte att två kolumner heter samma sak.)

När alla utfall finns i **C3** gå till *Calc -> Probability Distributions -> Binomial*. Välj att vi vill räkna ut utfallens *Probability*, *Number of trials* till 10, och *Event probability* till 0.6 (Minitab använder , som decimalseparator). I *Input column* skriver vi kolumnen med de utfall vi har, alltså **C3**, och för att spara alla sannolikheter skriver vi en ny tom kolumn, t.ex. **C4**, i *Optional storage*.

Använd den skapade fördelningen för att svara på följande frågor.

- Vad är sannolikheten att tre frön gror, $Pr(X = 3)$?
- Vad är sannolikheten att minst åtta frön gror, $Pr(X \geq 8)$?

c)

Vi kan även direkt visualisera olika binomialfördelningar genom att gå till *Graph -> Probability Distribution Plots -> View Single*, välja **Binomial** under *Distribution* och skriva in värdena för *Number of trials* och *Event probability*.

Visualisera binomialfördelningen från steg 1 med denna metod och jämför med diagrammet som vi skapade tidigare.

Extra

Minitabs funktion som används i steg 2 fungerar att använda på alla diskreta fördelningar. Testa själv att beräkna, med hjälp av Minitab, och tolka i ord $Pr(X = 2)$ för följande variabler. Tänk på vad de olika fördelningarna beskriver.

- $X \sim hyp(n = 5, \pi = 0.3, N = 30)$
- $X \sim poi(\mu = 10)$
- $X \sim geo(\pi = 0.1)$

Dessa fördelningar kan även visualiseras med funktionen som användes i steg 3. Visualisera ett diagram för respektive fördelning.

Uppgift 2

På liknande sätt kan vi beräkna kontinuerliga sannolikheter.

Efter att ha mätt automatiskt sågade meterlånga lister vid ett sågverk kan informationen sammanfattas med att längden, X, är normalfördelad med väntevärde 100 cm och standardavvikelse 2.5. Det vill säga:

$$X \sim N(\mu = 100, \sigma = 2.5)$$

a)

Sågverket kastar alla lister som är mindre än 95 och större än 105 cm för de inte uppfyller längdkraven. Hur stor andel av alla lister som produceras kommer kastas? Vi vill alltså beräkna:

$$1 - Pr(95 < X < 105)$$

Från föreläsning 3 har vi fått en teknik för att räkna ut sannolikheter av ett intervall. Vi kan alltså skriva om uttrycket ovan som:

$$1 - (Pr(X < 105) - Pr(X < 95))$$

Om vi skulle räkna ut denna sannolikhet för hand skulle vi behövt standardisera fördelningen till den standard normal som finns i tabellsamlingen, men eftersom vi har nu tillgång till en programvara som kan göra avancerade beräkningar mycket snabbare än oss själva tar vi vara på denna resurs.

Lägg in de eftertraktade X, 95 och 105, i en kolumn, t.ex. **C5**. Gå sedan in till *Calc -> Probability Distributions -> Normal Distribution* och skriv in medelvärdet, 100, i *Mean* och standardavvikelsen, 2.5, i *Standard deviation*. Dessa inställningar beskriver den normalfördelning som vi vill beräkna sannolikheter på. Markera nu kolumn **C5** i *Input column* och spara sannolikheterna genom att skriva in **C6** i *Optional storage*.

Nu har vi fått ut de sannolikheter vi behöver för att räkna ut ovanstående uttryck. Gör detta och besvara frågeställningen.

b)

Vi kan även använda samma metod som i a) för att beräkna vid vilket x-värde som en viss sannolikhet framkommer.

Sågverket är intresserade av att veta vilket mått 20% av deras brädor är större än, det vill säga $Pr(X > x) = 0.20$. De flesta tabeller och denna funktion i Minitab beräknar dock bara sannolikheter för $Pr(X < x)$, vilket innebär att vi måste skriva om sannolikheten till:

$$1 - Pr(X < x) = 0.20$$
$$Pr(X < x) = 1 - 0.20 = 0.80$$

Under *Calc -> Probability Distributions -> Normal Distribution* skriver vi återigen in parametrarna för den normalfördelning vi vill genomföra beräkningar på. En ändring vi måste göra för att hitta ett x-värde är att markera *Inverse cumulative probability*.

Eftersom vi enbart är intresserade av en sannolikhet markerar vi även *Input constant* och skriver in den sannolikhet vi vill beräkna, 0.80, och trycker *OK*.

Vilket mått besvarar frågan?

c)

På samma sätt som de diskreta fördelningarna kan visualiseras med ett diagram, kan vi göra detsamma för de kontinuerliga.

I uppgift a) ville vi beräkna $Pr(95 < X < 105)$ och detta ska vi nu visualisera. Under *Graph -> Probability Distribution Plot -> View Probability* väljer vi *Normal* som *Distribution* och anger parametervärden för *Mean* och *Standard deviation*.

Under fliken *Shaded Area* markerar vi först att vi ska ange *X Value* och *Middle* eftersom vi har x-värden i ett intervall. Skriv in 95 i *X value 1* och 105 i *X value 2* och klicka på *OK*.

Diagrammet som nu skapats visar sannolikheten att slumpvariabeln X antar värden mellan 95 och 105.

Extra

I fliken *Shaded Area* finns det fler val vi kan ange för att få andra sorters markerade regioner och därigenom visualisera andra sannolikheter. Försök att skapa diagram som visualiserar följande sannolikheter:

- $Pr(X > 105)$
- $Pr(X < 93)$
- $Pr(X > x) = 0.20$ (Ledning: Markera *Probability* istället för *X Value*)

Uppgift 3

Innan vi börjar med denna uppgift skapar vi ett nytt arbetsblad (Worksheet) genom *File -> New -> Minitab Worksheet*.

a)

Enligt Centrala gränsvärdessatsen (CGS) skall en summa av slumpvariabler bli ungefär normalfördelad om antalet variabler i summan är tillräckligt stort. Vidare gäller att dessa variabler skall vara av samma sort, man brukar säga likafördelade, och inte bero av varandra.

Det enklaste exemplet på detta är att man gjort ett urval om n observationer. Var och en av dessa är som regel oberoende tagna. Detta gäller om populationen är oändligt stor eller åtminstone mycket stor. Varje enskild observation är ju när den skall göras ett "oskrivet kort" och detta brukar modelleras med att det värde man får är en observation av en slumpvariabel, som gäller enbart just för denna observation.

Antag t.ex. att vi skall göra ett urval av n personer bosatta i Sverige och undersöka hur många syskon de har. För varje utvald person är antalet syskon en slumpvariabel och det innebär att vi har totalt n slumpvariabler i vårt urval. Innan vi har frågat respektive person om antalet syskon vet vi ju inte hur många de är och det gör denna storhet till en slumpvariabel.

Om vi nu vill göra en bedömning av det totala antalet angivna syskon i vårt urval kan vi skriva detta som $\sum_{i=1}^{n} X_i$ där $X_1, ..., X_n$ är antalet syskon hos var och en av de n personerna.

Denna summa är nu enligt CGS ungefär normalfördelad med väntevärde $n * \mu$ och standardavvikelse $\sigma * \sqrt{n}$ där μ och σ är medeltal och standardavvikelse för antalet syskon i hela populationen, dvs bland antalet bosatta i Sverige, om n är tillräckligt stor. (Vi bryr oss i detta fall inte om det faktum att två eller flera personer i populationen kan vara syskon och därmed ha lika många syskon, vilket egentligen komplicerar det hela men kan bedömas vara ett mindre problem eftersom populationen är så stor.)

Vidare gäller att urvalsmedeltalet av antalet syskon, dvs

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

blir ungefär normalfördelad med väntevärde μ och standardavvikelse $\frac{\sigma}{\sqrt{n}}$.

Man kan (och ska) naturligtvis lita på dessa resultat, eftersom det handlar om ganska lång tids forskning och matematiskt ovedersägliga resultat, men det är ändå nyttigt att studera hur bra denna approximation är och vad ett "stort n" kan vara.

b)

Börja med att mata in värdena 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 och 10 i kolumn **C1**. Antag att detta är det antal syskon som kan förekomma i en viss population, dvs ingen i populationen har fler än 10 syskon, och det finns de som inte har några syskon alls. Anta att antalet syskon följer fördelningen i tabell 1.

Med hjälp av denna tabell över utfall och dess sannolikheter, räkna för hand ut medeltalet, μ, och standardavvikelsen, σ, i populationen av antalet syskon och skriv upp dessa värden någonstans.

Lägg nu proportionerna som decimaltal i kolumnen **C2**, dvs. mata in värdena 0,16, 0,35 etc. i **C2**.

Antag nu att vi skall göra ett urval om 10 personer från populationen och bestämma hur många syskon var och en av dessa har. Via slumpvariabelbegreppet kan detta utföras genom att slumpmässigt generera 10 observationer från den slumpvariabel som antar värdena i **C1** med sannolikheter motsvarande värdena i **C2**.

Antal syskon	Frekvens i populationen (%)
0	16.0
1	35.0
2	29.0
3	10.0
4	6.0
5	2.0
6	0.5
7	0.5
8	0.4
9	0.4
10	0.2

Table 1: Tabell över antalet syskon

Vi skapar urvalet genom att gå till *Calc -> Random Data -> Discrete* och skriver in 10 rader som *Number of rows of data to generate*, **C3** i *Store in column(s)* och **C1** och **C2** i *Values in* respektive *Probabilities in*.

Kommandot innebär att vi slumpar 10 observationer från kolumnen **C1** och lägger dessa i **C3** och att slumpningen görs så att varje värde dras med en sannolikhet som motsvarar värdet i **C2**. Ni bör därför få observationer i **C3** som till större delen är något av värdena 0, 1, 2, 3 och 4, eftersom dessa värden har betydligt högre sannolikheter än de övriga (motsvarar högre frekvenser i populationen).

Beräkna medelvärdet av detta urval genom att skapa beskrivande statistik över **C3**. Stämmer detta medelvärde någorlunda överens med medeltalet i populationen? Borde det göra det?

c)

För att se hur väl CGS stämmer måste vi på något sätt uppskatta samplingfördelningen hos detta urvalsmedeltal och då krävs att vi upprepar urvalsförfarandet ett stort antal gånger. Kunde vi till exempel skapa 10000 urval av detta slag borde motsvarande urvalsmedeltal ge en hyfsad bild över hur ett urvalsmedeltal kan variera.

Nu är det ganska arbetskrävande att upprepa ovanstående procedur 10000 gånger varför det finns ett datamaterial som detta redan är gjort. Datamaterialet **syskon.xls** importeras genom *File -> Open Worksheet* och innehåller kolumner av medelvärden från alla 10000 urval om olika storlek som genomförts. Kolumnerna heter vilken stickprovsstorlek som använts.

n = 10

Gör ett histogram över **C4**. Ser histogrammet ut att motsvara en normalfördelning? Beräkna vidare medelvärdet och standardavvikelsen av värdena i **C4**. Verkar medelvärdet överensstämma någorlunda med populationsmedeltalet? Verkar standardavvikelsen överensstämma någorlunda med $\sigma/\sqrt{10}$? Teorin säger ju att dessa överensstämmelser skall finnas, och detta gäller oavsett om populationen är normalfördelad eller ej, vilket antalet syskon inte är.

Större n

Skapa ett histogram och beräkna medelvärdet och standardavvikelsen för $n = 30$, $n = 50$ och $n = 100$, det vill säga det som gjordes ovan och besvara frågorna.

d)

Jämför fördelningsform, medelvärde, standardavvikelse med den teoretiska normalfördelningen. Försök säga något om från och med vilken urvalsstorlek CGS verkar fungera.