Laboration 2

Simulering med Minitab

Datorn kan med stor framgång användas för att simulera slumpmässiga försök av olika typer. I den här laborationen kommer simulering att användas för att öka förståelsen av två viktiga begrepp inom statistiken. Dessa båda begrepp är **de stora talens lag** och **den centrala gränsvärdessatsen**.

Vad är simulering?

Med simulering menar vi här att låta datorn simulera slumpmässiga försök. Man kan på kort tid genomföra ett stort antal slumpmässiga försök vilkas resultat lagras i datakolumner för vidare bearbetning. Grunden för all simulering med dator är att datorn är utrustad med en slumptalsgenerator. I vårt fall kommer vi att använda oss av Minitab, som kommer att generera tal som kan betraktas som slumpmässiga. Egentligen är det en matematisk formel som genererar talen, men de betraktas som slumpmässiga. Men hjälp av Minitab kan vi generera slumptal från olika sannolikhetsfördelningar med enkla kommandon.

Nu till själva uppgifterna.

Uppgift 1. De stora talens lag

Hur många replikationer måste du göra för att få en stabil skattning?

Svaret på den frågan ger svaret på hur stort stickprov man behöver för att få en skattning som med stor säkerhet är väldigt nära det sanna parameter värdet. Undersök om samma antal gäller för olika värden på p.

Illustrera med spridningsdiagram och använd dem som ett empiriskt bevis för de stora talens dag.

För att studera de stora talens lag kommer vi att simulera fram ett antal Bernoulliförsök.

Så här går du till väga för respektive simulering:

Ett Bernoulliförsök är ett försök med endast två utfall t ex lyckat/misslyckat. Följande kommando används:

Calc➢Random Data➢Bernoulli…

Ange antalet försök samt i vilken kolumn (t ex C1) du vill lagra resultaten i. Du måste också ange sannolikheten för ett lyckat försök. Vid ex kast med ett symmetriskt mynt är sannolikheten för krona (ett lyckat försök) = 0,5.

I den angivna kolumnen finns nu ett antal ettor och nollor, där en etta står för ett lyckat försök och en nolla ett misslyckat försök. För att studera den successiva relativa frekvensen (lyckade försök) måste vi först utföra några enklare beräkningar. Dessa utför vi med Minitabs kalkylator.

Först kumulerar vi antalet lyckade försök genom att använda följande kommando:
Calc➢Calculator… Ange att resultatet ska placeras i kolumn C2 och använd funktionen *Partial Sums(C1)*.

För att få den successiva relativa frekvensen dividerar vi med antalet dittills gjorda försök. Vi skapar en ny kolumn (C3) med antalet gjorda försök successivt med kommandot:
Calc➢Make Patterned Data➢Simple Set of Numbers… Lägg in värden från 1 till n, där n=antal forsök.

Den successiva relativa frekvensen får vi genom:**Calc➢Calculator…** (Lägg resultatet i kolumn C4 och använd uttrycket C2/C3)

Nu kan vi studera den successiva relativa frekvensen genom att plotta den mot antalet gjorda försök.

Upprepa tills du funnit svaret!

Uppgift 2. Centrala gränsvärdessatsen

1. Stickprovsmedelvärdet $\overline{X}$

Centrala gränsvärdessatsen utnyttjas väldigt ofta vid statistisk slutledning. Ofta vill vi dra slutsatser angående en populations medelvärde utifrån ett stickprov. För att göra det använder vi oss ofta av normalfördelningen. Tyvärr är vår undersökningsvariabel inte alltid normalfördelad, men enligt CGS kommer stickprovsmedelvärdet att vara normalfördelad om vi har ett stort stickprov (fler än 30 observationer). Så oavsett vilken fördelning X följer så är $\overline{X}$ normalfördelad med väntevärdet $\mu_{\overline{X}} = \mu_X$ och standardavvikelsen $\sigma_{\overline{X}} = \sigma_X / \sqrt{n}$.

För att kontrollera att CGS stämmer studerar vi ett stort antal stickprovsmedelvärden. Först dras slumpmässigt ett antal stickprov från en variabel **X** som följer en fördelning som helst inte är normalfördelad. Vi kan givetvis låta **X** följa vilken sannolikhetsfördelning som helst, men genom att välja en fördelning som inte är symmetrisk (ex exponentialfördelningen) ser vi tydligast hur samplingfördelningen för $\overline{X}$ går mot en normalfördelning när stickprovsstorleken ökar.

Det kan vara lämpligt att studera 1000 stickprovsmedelvärden. Variera stickprovsstorleken. (Ex n=5, 15, 30)

Några exempel på osymmetriska sannolikhetsfördelningar är Chi-två-fördelningen och Exponentialfördelningen.

Vi kan låta Minitab generera slumptal som följer en viss fördelning genom kommandot:
Calc➢Random Data➢*Aktuell fördelning*…(Ex: Exponential med μ=5)

Ange antal rader (t ex 1000) som ska genereras samt vilka kolumner de ska läggas i (C1-C5 om varje stickprov ska ha n=5). Vi har nu ett antal rader och kolumner som följer den definierade fördelningen. Vi ser varje rad som ett stickprov. Det innebär att antalet kolumner motsvarar antalet observationer i ett stickprov.

Studera hur fördelningen för X ser ut genom att göra histogram över en av kolumnerna och ta fram medelvärde och standardavvikelse.

Sen är det stickprovsmedelvärdena som ska studeras. För att beräkna medelvärdet i varje stickprov (rad) och lagra dessa i en separat kolumn använder vi Minitabkommandot:
Calc➢Row Statistics…

Vi ska nu studera fördelningen för dessa medelvärden. Det gör vi lämpligast genom att illustrera med histogram, samt komplettera med beskrivande mått.

Kontrollera att standardavvikelsen för medelvärdena minskar i takt med att stickprovsstorleken ökar. Använd resultaten till ett empiriskt bevis för den centrala gränsvärdessatsen.

2. Stickprovsandelen $\hat{p}$ (Binomialfördelningen approximeras med Normalfördelningen)

Vi ska här studera hur fördelningen för stickprovsandelen tenderar att gå mot en normalfördelning då stickprovsstorleken ökar (n>40). För att studera detta tänker vi oss att följande gäller:

IQ, intelligenskvot, är ett statistiskt framtaget begrepp som ska mäta intelligensen. Ett IQ-test mäter enbart förmågan att känna igen mönster och logik. Emotionell och social intelligens kan inte mätas med hjälp av ett IQ-test. Genomsnittet för befolkningen ligger på runt 100. Ca 10 % av befolkningen har ett IQ som överstiger 120.

Vi är nu intresserade av att uppskatta andelen personer som har denna egenskap (IQ>120). Studerar vi ett slumpmässigt urval av personer borde stickprovsandelen, $\hat{p}$, variera kring 0,1. Om vi låter **X** vara antalet personer med ett IQ större än 120 och med OSU gör vi urval från en stor population så kommer slumpvariabeln **X** att följa en binomialfördelning med n=stickprovsstorleken och p=0,1.

Ex: Vill vi simulera fram 1000 stickprov, vardera innehållande 15 observationer gör vi det med kommandot:

Calc➢Random Data➢Binomial… Om vi vill ha 1000 stickprov ska vi ange 1000 rader, en målkolumn (t ex C1), samt att antal försök ska vara 15 och p=0,1.

Varje rad innehåller nu antalet personer med IQ>120 i varje stickprov. För att skatta andelen personer med hög IQ dividerar vi med antalet observationer i varje stickprov.

Calc➢Calculator… (Expression: C1/15; placera resultatet i C2)

I kolumn C2 har vi nu 1000 stycken stickprovsandelar. Fördelningen för dessa stickprovsandelar åskådliggörs lämpligast med histogram som kompletteras med beskrivande mått som medelvärde och standardavvikelse.

Upprepa nu proceduren ovan med t ex stickprovsstorlekarna 30 och 60 för att se hur stickprovsandelen går mot en normalfördelning. Kontrollera också hur standardavvikelsen minskar i takt med att stickprovsstorleken ökar. Jämför med förväntade värden för $\hat{p}$ och $se(\hat{p})$.

Konfidensintervall.

Här studerar vi hur konfidensintervallen faller ut. Vi utgår från exemplet med IQ.
Först bildar vi konfidensintervall för populationsandelen p för dessa 1000 stickprov. Därefter ska vi kontrollera hur stor andel av dessa konfidensintervall som innehåller den sanna populationsandelen (p=0,1).

Beräkna 90 % konfidensintervall för de 1000 stickproven enligt: $\hat{p} \pm 1{,}645 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Vi utnyttjar Minitab för att räkna ut konfidensintervallens gränser. Undre gränsen fås genom:

Calc➢Calculator… Ex: vid stickprovsstorleken n=60 om de skattade värdena finns i C2:
C2 - (1,645*SQRT(C2*(1-C2)/60)) - de beräknade värdena placeras i C3

Ta fram övre gränsen på motsvarande sätt och lägg den i en ny kolumn, t ex C4. Vi har nu två kolumner innehållande konfidensintervallens gränser. För att studera hur många av dessa intervall som innehåller det sanna värdet 0,1 kan vi göra på följande sätt:

Calc➢Calculator…Ex: C3<0.1 And C4>0.1, ange C5 som målkolumn.

Alla värden som nu ligger inom intervallets gränser kommer att få värdet 1 och de övriga får värdet 0 i C5. Studera andelen stickprov vars konfidensintervall innehöll det verkliga värdet p=0,1. Jämför resultatet med vad konfidensintervallet säger oss. Verkar det rimligt?