Nawaz Rahman
DS21Göteborg
R programmering och statistik

1. Att städa data är viktigt eftersom det förbättrar datakvalitén vilket i sin tur ökar den totala produktiviteten. Ett exempel är att ta bort *NA* värden. Städa data kan innebära att ta bort, uppdatera, korrigera och konsolidera data.
   Två funktioner för att städa data är exempelvis:
   - Seperate (). Används för att separera en dataframe kolumn i flera kolumner.
   - Pivot_longer (). Denna funktion "vidgar" data, ökar antalet kolumner och minskar antalet rader.

2. En Dataframe är det mest populära sättet att lagra data i R, och det är också den mest använda datastrukturen för dataanalys. En Dataframe är bara en uppsättning vektorer av samma längd. Längden på varje element i listan är lika med antalet rader, och varje element i listan är en kolumn. Som en konsekvens kan distinkta typer av objekt lagras i varje kolumn av dataramar (dvs numerisk, tecken, faktor). En Dataframe kan ses som ett Excel-kalkylblad med kolumner med olika typer av data och rader av liknande längd.

3. Funktioner är användbara när du vill utföra en viss uppgift flera gånger. En funktion accepterar inmatningsargument och producerar utdata genom att utföra giltiga R-kommandon som finns inuti funktionen.
   f = function(argument) {
   statement
   }

4. Skillnaden mellan for-loop och while-loop är att i for-loop är antalet iterationer som ska göras redan känt och används för att erhålla ett visst resultat medan kommandot i while-loop körs tills ett visst villkor uppnås och påståendet bevisas att vara FALSE.

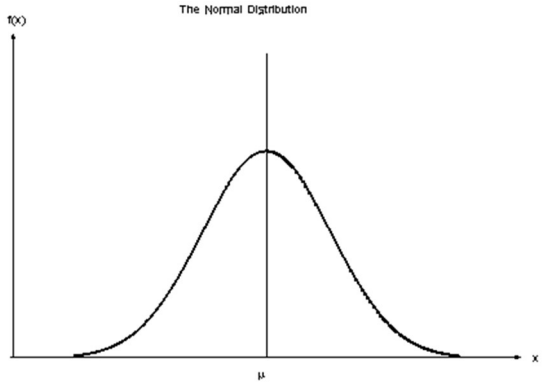
5. If-satser säger till R att köra en kodrad. If-satser utvärderar om en condition är TURE eller FALSE. Om condition returnerar TRUE kommer programmet att köra all kod mellan hakparenteserna {} men om det returnerar FALSE kommer ingen kod att exekveras. Det passar att använda if-satser när man vill se om ens kod uppfyller en condition eller inte.
   If (condition) {
   code block
   }

6. R-squared (R^2) är ett passformsmått för linjära regressionsmodeller. Denna statistik indikerar procentandelen av variansen i den beroende variabeln som de oberoende variablerna förklarar kollektivt. R-squared mäter styrkan i sambandet mellan din modell och den beroende variabeln på en lämplig skala från 0 – 100 % (0-1).

7. När du vill veta om en händelse är statistiskt signifikant och om den uppnår en viss grad av signifikans använder du dig av teststatistik. Man undersöker om en eller flera kopplingar mellan variablerna beror på något annat än slumpen. För att avgöra om datan har statistisk

signifikans, använder analytiker ofta statistisk hypotestestning.

8. En normalfördelad population innebär att populationen kretsar gemensamt runt ett genomsnittsvärde.



9. Konfidensintervallet är en uppsättning siffror som omfattar ett populationsvärde med en hög grad av säkerhet. När populationens medelvärde ligger mellan två intervall anges det vanligtvis i procent. Konfidensintervall är en enkel teknik för att avgöra om ditt urval korrekt representerar populationen du undersöker. För att beräkna konfidensintervallet, beräkna först provets medelvärde och "standard error".

10. Styrkan hos en koppling mellan två variabler mäts med korrelationskoefficienter. En korrelation mellan variabler innebär att när en variabels värde ändras, tenderar den andra att förändras på samma sätt. Vi kan använda värdet på en variabel för att prediktera värdet på den andra variabeln om vi förstår det sambandet. Längd och vikt hänger till exempel ihop, när höjden stiger, ökar också vikten. Som ett resultat, om vi ser någon som är onormalt lång, kan vi anta att han också är extremt tung.

11. I ett slumpmässigt urval från en population är en outlier ett värde som är onormalt långt ifrån andra värden. När en datapunkt klassificeras som en anomali är det upp till analytikern att ta reda på vad som är ovanligt – och vad man ska göra med det. Ett sätt att hantera outliers är att exkludera dem.

Källor: Föreläsningsmaterial och egna anteckningar.