Uppsala Universitet
Matematiska Institutionen
I. Kaj, R. Müller, Y. Wang

Repetition och viktiga formler från modul 1

1 Statistiska Mått

Syftet med olika statistiska mått är att tydligt presentera och organisera empiriska data med hjälp av nyckeltal (t.ex. medelvärde, etc.), grafer och tabeller. Låt oss anta att vi har data/observationer $x_1, \ldots, x_n$ och $y_1, \ldots, y_n$.

1.1 Lägesmått

Läs kapitel 2.2 i boken för förklaringar och exempel till de olika mått!

Aritmetiskt medelvärde (arithmetic mean):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + \ldots + x_n\right)$$

Vägt medelvärde (weighted mean): Antar man har m grupper med medelvärde $\bar{x}_1, \ldots, \bar{x}_m$ och i varje grupp finns $n_1, \ldots, n_m$ observationer respektive. Då är

$$\bar{x}_V = \frac{n_1\bar{x}_1 + \ldots + n_m\bar{x}_m}{n_1 + \ldots + n_m}.$$

Geometriskt medelvärde (geometric mean):

$$x_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \left(x_1 \cdot \ldots \cdot x_n\right)^{1/n}$$

Median (median): Värdet för ett ordnat datamaterial som delar materialet i två lika stora delar så att precis hälften av observationerna är mindre och andra hälften större än (eller lika med) detta värde. Om antalet observationer är udda är medianen exakt lika med värdet i mitten, vid jämnt antal är den medelvärdet av de två mellersta värdena.

Typvärde (mode): Värdet som förekommer oftast i datamaterialet.

1.2 Spridningsmått

(Stickprovs-)varians (sample variance):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(Stickprovs-)standardavvikelse ((sample) standard deviation):

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Variationskoefficient (coefficient of variation):

$$\frac{s}{\bar{x}}$$

Variationsbredd (range):

$$R = x_{\max} - x_{\min}$$

1.3 Samvariation

Korrelationskoefficient (correlation coefficient):

$$R_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \qquad \text{Pearson's}$$

där

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Spearmans korrelationskoefficient (Spearman's correlation coefficient): Sortera x och y värdena från 1 till n och definiera d_i som differensen i rang mellan x_i och y_i, dvs $d_i = \text{rang}(x_i) - \text{rang}(y_i)$. Sedan är

$$r_S = 1 - 6\,\frac{\sum_{i=1}^{n} d_i^2}{n^3 - n}.$$

1.4 Visualisering av statistiska mått

Stolpdiagram (bar graph): Användbar för diskret datamaterial. För varje klass rita en linje som representerar antal observationer i motsvarande klass.

Histogram (histogram): Dela in kontinuerliga data i ett lämpligt antal klasser. För varje klass rita en linje som antingen representerar antal observationer eller skalad relativ frekvens i motsvarande klass.

Lådagram (boxplot): Rita ett horisontalt streck för median värdet och en box omkring det så att 25% av observationerna ligger under den undre gränsen (25 percentil) och 75% under den övre gränsen (75 percentil). På varje sida rita en linje som är $1,5$ gånger boxens längd lång. Om den minsta eller största observation ligger inför den linje förkortar vi linjen. Om det finns observationer utanför denna linjer ("outliers") så ritar vi dem speciellt.

Sambandsdiagram (scatterplot): Rita observationspar $(x_1, y_1), \ldots, (x_n, y_n)$ i ett koordinatsystem.

Läs gärna mer om visualiseringar och deras interpretation i boken eller online!