Stockholm University

Statistics and Data Analysis
Formula sheet

Descriptive: One variable

Mean: $\bar{x} = \frac{1}{N}\sum_U x_i$

Mode: Most frequently occurring value.

Variance: $S_x'^2 = \frac{1}{N}\sum_U (x_i - \bar{x})^2$ 总体方差. 真实.

$S_x^2 = \frac{1}{N-1}\sum_U (x_i - \bar{x})^2$ 样本方差: Bessel 校正

If x is a dummy variable with mean $\bar{x} = P_x$, then $S_x'^2 = P_x(1-P_x)$ and $S_x^2 = \frac{N}{N-1}P_x(1-P_x)$.

Standard deviation: 标准差

$S_x' = \sqrt{S_x'^2}$ and $S_x = \sqrt{S_{x,U}^2}$

Skewness: $Sk_x = \frac{\frac{1}{N}\sum_U (x_i-\bar{x})^3}{S_x^3}$ 偏度:

右偏 (>0): 收入分布

左偏 (<0): 考试成绩

Let $x_{(1)}, x_{(2)}, \cdots, x_{(N)}$ be the ordered population.

若为整数则直接得, 否则取2个比例

pth percentile: Let $c = (N-1)p+1$, a be the integer part of c and b be the decimal part of c. The pth percentile is $\breve{x}_p = (1-b)x_{(a)} + bx_{(a+1)}$.

First quartile: $\breve{x}_{0.25}$

Median (second quartile): $\breve{x}_{0.50}$

Third quartile: $\breve{x}_{0.75}$

Range: $\text{range}_x = x_{(N)} - x_{(1)}$

Interquartile range: $\text{IQR}_x = \breve{x}_{0.75} - \breve{x}_{0.25}$

Result: Let a and b be constants, x be a variable and $y_i = b(x+a)$ then

$\bar{y} = b(\bar{x}+a)$	$\dot{y} = b(\dot{x}+a)$
$\breve{y}_p = b(\breve{x}_p + a)$	$Sk_y = Sk_x$
$\text{range}_y = b\,\text{range}_x$	$\text{IQR}_y = b\,\text{IQR}_x$
$S_y = b\,S_x$	$S_y^2 = b^2\,S_x^2$

Standardization:

Let $z = \frac{x-\bar{x}}{S_x}$ then $\bar{z} = 0$ and $S_z = 0$.

?

Descriptive: Two vars.

Correlation:

$r_{xy} = \frac{\sum_U (x_i-\bar{x})(y_i-\bar{y})}{\left(\sum_U (x_i-\bar{x})^2 \sum_U (y_i-\bar{y})^2\right)^{1/2}}$.

排名次序 不要求正态, 线性, 仅测量单调关系.

$\rho=-1$ 负单调 $\rho=1$ 正单调

Spearman's correlation:

减轻异常值影响.

$r^s_{xy,U} \equiv \frac{\sum_U (R(x_i)-\bar{R}_U)(R(y_i)-\bar{R}_U)}{\left(\sum_U (R(x_i)-\bar{R}_U)^2 \sum_U (R(y_i)-\bar{R}_U)^2\right)^{1/2}}$ $\in[-1,1]$

with $\bar{R}_U = (N+1)/2$ and $R(x)$ and $R(y)$ the ranks of x and y

x的排名

Kendall's correlation: 一致对+1. 不一致-1

$r^k_{xy,U} \equiv \frac{2}{n(n-1)}\sum_{i<j} \text{sgn}(x_j - x_i)\text{sgn}(y_j - y_i)$

$\in[-1,1]$

Simple Linear regression

Fitted line: $\hat{y} = b_0 + b_1\,x$ with

$b_1 = r_{xy}\frac{S_y}{S_x} = \frac{\sum_U (x_i-\bar{x})(y_i-\bar{y})}{\sum_U (x_i-\bar{x})^2}$

$b_0 = \bar{y} - b_1\bar{x}$.

Fitted values: $\hat{y}_i = b_0 + b_1\,x_i$.

Residuals: $e_i = y_i - \hat{y}_i$.

Sum of squares: $SST = SSE + SSR$

$SST = \sum_s (y_i - \bar{y}_s)^2$ 数据总平方和, 固定

$SSE = \sum_s (y_i - \hat{y}_i)^2$ 误差平方和, 越小越好

$SSR = \sum_s (\hat{y}_i - \bar{y}_s)^2$ 回归平方和, 越大越好.

Coefficient of determination:

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$.

Prediction for y given x_0: $\hat{y}_0 = b_0 + b_1 x_0$

Multiple Linear regression

Fitted line: $\hat{y} = b_0 + b_1 x_1 + \cdots + b_K x_K$.

Fitted values: $\hat{y}_i = b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki}$.

Residuals: $e_i = y_i - \hat{y}_i$.

Sum of squares: Same as above.

Coefficient of determination: Same.

Adjusted coef. of determination:

?

$\bar{R}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}$.