Types of data

- Categorical
  - > nominal (colore, hobby)
  - > ordinal (Level of education, marks at school)
- Quantutive
  - > numerical (blood pressure, car accidents)

How to describe categorical data?

Categorical nominal

We can calculate:

mode - most frequent element

frequency

Categorical ordinal

$x_1, \dots, x_n$

Order statistic:

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

Ranks : $r_1, \dots, r_n$

if $x_{(1)} < \dots < x_{(n)}$ then rank of $r(x_{(k)}) = k$

ties in data

$$X_{(1)} < X_{(2)} = X_{(3)} < X_{(4)}$$

$$r_2 = r_3 = 2{,}5$$

if we have more tie elements we calculate mean

We have n ranks:

$$\sum_{i=1}^{n} r_i = \frac{n(n+1)}{2}$$

We can calculate for ordered data:

- mode
- quantiles
- sample mean (careful with interpretation)

Dependence measure

X, Y random vars. measures on n objects

$x_1, \dots, x_n$
$y_1, \dots, y_n$

e.g. do tall people have also bigger weight?

Motivation of correlation

Correlation measures the degree of linear association between two numeric variables

Pearson correlation coefficient (linear correlation coefficient)

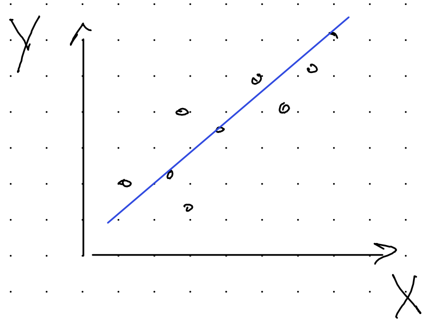
$$r(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}$$

Linear correlation: Pearson correlation

Props:

- When X and Y are independent
  $r(X,Y)=0$
- $-1 \leq r(X,Y) \leq 1$
- When $Y = aX+b$, $a,b \in \mathbb{R}$ and $a>0$
  then $r(X,Y)=1$
  > and $r(X,Y)=-1$ when $a<0$

Correlation plot



Statistical significance of Pearson's correlation

CC measure is descriptive statistic (not inferancial statistic measure)

$H_0 : r = 0$ vs $H_1 : r \neq 0$ (two tailed hypothes.)

If $p < \alpha = 0{,}05$ reject H_0
correlation is stat-y significant

Classification (not ML)

Often we want to classify numerical or ordinal data into classes

Rule of thumb number of classes $k \approx \sqrt{n}$

Rank correlations

For ordinal data we can't use liear correlation

monotone dependence

Assume k values $x_i^* \sim X$ ℓ values $y_j^* \sim Y$

denote ranks X by $r_1 \dots r_k$

Y by $q_1 \dots q_\ell$

Denote the number of pairs in the sample where $X = x_i^*$ and $Y = y_j^*$ by n_{ij}

$$D = \sum_{i=1}^{k} \sum_{j=1}^{\ell} (r_i - q_j)^2 n_{ij}$$

if all ellements x_i y_j are different $k = \ell = n$

if $D = 0$ then X and Y increasing with each other

IF Y depends on X decreasingly then D has max value

Rank correlation

Def