

# Data: Stats & Itemsets

8/28

Given a random variable  $X$

and a set of  $n$  random variables  $\{S_1, \dots, S_n\}$

$\Rightarrow$  1. all  $S_i$  are independent

2. all  $S_i$  have same PDF (prob mass function) as  $X$

we call  $\{S_1, \dots, S_n\}$   $\forall_{S_i} \Pr(X=x) = \Pr(S_i=x) \forall x$

independent and identically distributed (iid)

a random sample of size  $n$  from  $X$

is an iid set of  $n$  random variables

$n$  observed values drawn from  $X$

$\{x_1, \dots, x_n\}$  as iid

we have

$$\begin{aligned} f(x_1, \dots, x_n) &= f(S_1=x_1, S_2=x_2, \dots, S_n=x_n) \quad \leftarrow \text{by (*) from prev notes} \\ &= f(S_1=x_1) \cdot f(S_2=x_2) \cdot \dots \cdot f(S_n=x_n) \\ &= f_X(x_1) \cdot \dots \cdot f_X(x_n) \quad \text{w/ } f_X \text{ is PDF for } X \\ &= \prod_{i=1}^n f_X(x_i) \end{aligned}$$

# Statistic

estimate a param of population w/ sample statistic

Let  $m \in \mathbb{Z}^+$

Let  $S_1, \dots, S_m$  be a random sample drawn from some multivariate rand variable  $X$

Consider unknow param  $\Theta$

statistic  $g: (S_1, \dots, S_m) \rightarrow \mathbb{R}$

$$\hat{\Theta} = g(S_1, \dots, S_m)$$

- estimate of population param (estimator)
- $\hat{\Theta}$  is also a random variable

point estimate: use value of stat to estimate pop param

$\Theta$ : unknow param

$g$ : function (estimator)

$\hat{\Theta}$ : is an estimate (a random var)

Eg. sepal length in iris dataset  $X_1$

5.9	6.9	6.6	4.6	6.0	4.7
5.0	5.0	5.7	5.0	7.2	5.9
5.4	5.0	5.7	5.8	5.1	5.6
4.8	7.1	5.7	5.3	5.7	5.7

←

unknown pop param mean  $\mu$

Observed values  $\vec{x} = \{x_i\}_{i=1}^n$  are random sample drawn from  $X_1$

Sample mean (e.g. average) for  $\mu$

$$g(\vec{x}) = \frac{1}{n} \sum_i x_i$$

point est for  $\mu$

$$\hat{\mu} = g(\vec{x}) = \frac{1}{24} (5.9 + 6.9 + \dots + 5.7) = \frac{135.4}{24} \approx 5.6$$

# Itemset mining

market basket analysis:

e.g.  $\{\text{milk, eggs}\}$   $\leftarrow$  item sets  
 $\{\text{syrup, waffles, butter}\}$   $\leftarrow$

given itemsets extract some association rules

$\{\text{milk, eggs}\} \rightarrow \{\text{cereal}\}$