# Data: Probabilistic View (Pt 2)

8/26

Assume that each numic attrib $X$ is a **random variable**

$\Omega$ domain of $X$ (e.g. all outcomes of an experiment)
aka **sample space**

$X: \Omega \to \mathbb{R}$
(when $\Omega$ is numeric $X$ is often assumed to be identity function)

**discrete random variable** - only takes a finite (or countably inf) set of values

**cts random variables** - taking any value in a range

**prob mass function** (PMF)

$X$ is discrete

$f(x) = P(X = x) \quad x \in \mathbb{R}$

props for $f$:

- $f(x) \geq 0 \quad \forall x$ (non-neg)
- $\sum_x f(x) = 1$ (sum to 1)

E.g. sepal length $X_1$

| | | | | | |
|---|---|---|---|---|---|
| 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 |

$n = 24$

cts random variable

$X_1(v) = v$

range: $[4.?, 7.1]$

as discrete (e.g. short vs. long)

$$A(v) = \begin{cases} 0 & v < 6 \\ 1 & v \geq 6 \end{cases}$$

range: $\{0, 1\}$

$f(0) = P(A=0) = \frac{19}{24} \simeq .79 = p$

$f(1) = P(A=1) = \frac{5}{24} \simeq .21 = 1-p$

## prob density function (PDF)

$X$ cts $\Rightarrow$ $P(X=x)=0$ $\forall\, x\in\mathbb{R}$

generally we consider intervals $[a,b]\subset\mathbb{R}$

$$P(X\in[a,b])=\int_a^b f(x)\,dx$$

props

- non-neg $f(x)\geq 0$ $\forall\, x\in\mathbb{R}$
- integrate to 1 $\int_{-\infty}^{\infty} f(x)\,dx = 1$

## Cumulative distribution function (CDF)

$F:\mathbb{R}\to[0,1]$ prob of observing a value of at most $x$

$$F(x)=P(X\leq x)$$

discrete!

$$F(x)=P(X\leq x)=\sum_{u\leq x} f(u)$$

cts

$$F(x)=P(X\leq x)=\int_{-\infty}^{x} f(u)\,du$$

Multivariate random variables

$\vec{X} = (X_1, X_2, \cdots, X_d)^T$

w/ $\Omega$ domain

$\vec{X} : \Omega \to \mathbb{R}^d$

for $\vec{x} \in \mathbb{R}^d$ $\vec{x} = (x_1, \cdots, x_d)^T$

- w/ all $X_j$ numeric $\Rightarrow$ $\vec{X}$ is identity function (e.g. rows of the data matrix is an outcome in sample space)
- w/ not all numeric $\Rightarrow$ $X_j$ map to a value in range $(X_j)$

All $X_j$ are discrete $\Rightarrow$ $\vec{X}$ is jointly discrete

w/ joint prob mass function

$f(\vec{x}) = P(\vec{X} = \vec{x})$

$\Leftrightarrow$ $f(x_1, \cdots, x_d) = P(X_1 = x_1, X_2 = x_2, \cdots, X_d = x_d)$

w/ $f(\vec{x}) \geq 0$

$\sum f(\vec{x}) = 1 = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_d} f(x_1, \cdots, x_d)$

All $X_i$ cts $\Rightarrow \vec{X}$ is **jointly cts**

w/ **joint prob density function**

$\omega \subseteq \mathbb{R}^d$

$$P(\vec{X} \in \omega) = \int \cdots \int_{\vec{x} \in \omega} f(\vec{x})\, d\vec{x}$$

$$\Leftrightarrow P((X_1, \cdots, X_d)^T \in \omega) = \int \cdots \int_{(x_1, x_2, \cdots, x_d)^T \in \omega} f(x_1, x_2, \cdots, x_d)\, dx_1 dx_2 \cdots dx_d$$

w/ prop $f(\vec{x}) \geq 0$

$$\int_{\mathbb{R}^d} f(\vec{x})\, d\vec{x} = 1$$

**joint CDF**

$$F(\vec{x}) = P(\vec{X} \leq \vec{x})$$

$$= P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_d \leq x_d)$$

$X_1, \dots, X_d$ independent random variables

- $\forall\ \omega_1, \dots, \omega_d \subset \mathbb{R}$

$$P(X_1 \in \omega_1, X_2 \in \omega_2, \dots, X_d \in \omega_d) = \prod_{i=1}^{d} P(X_i \in \omega_i)$$

e.g. $d=2$

$\forall\ \omega_1, \omega_2 \in \mathbb{R}^2$

$$P(X_1 \in \omega_1, X_2 \in \omega_2) = P(X_1 \in \omega_1)\, P(X_2 \in \omega_2)$$

when $X_i$ are ind

Let $F_i$ be the CDF of $X_i$
$f_i$ be the PDF of $X_i$

$$F(\vec{x}) = F(x_1, \dots, x_d) = \prod_i F_i(x_i)$$

$$f(\vec{x}) = f(x_1, \dots, x_d) = \prod_i f_i(x_i) \quad (*)$$

# Random sampling & Stats

**population**: universe of all entities in a study (e.g. grad students in US.)

- **param of population** (e.g. heights of grad students in U.S.)

- infer pop params

  w/ random sample from pop
  compute **statistic**
  as an
  **estimate** of population param