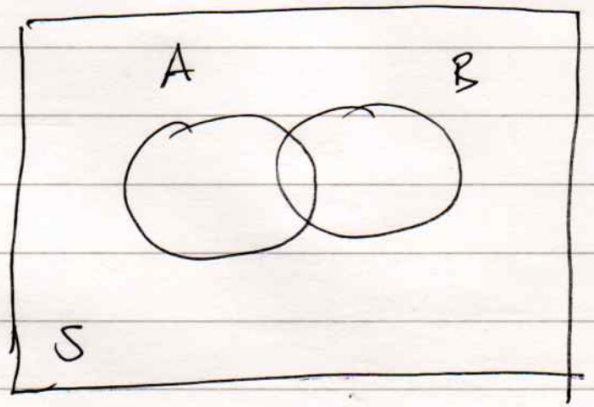
# Statistical Data Analysis
# Discussion notes – week 1

- Kolmogorov axioms
- Examples with marginal, conditional pdfs
- pdf/cdf with delta functions for discrete data
- Bayes' theorem to find number of false positives

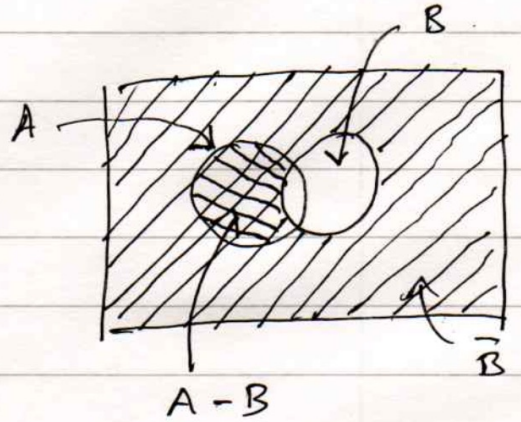
1) Example w/ Kolmogorov axioms:

Show $P(A \cup B) = P(A) + P(B) - P(A \cap B)$




Recall

$A - B \equiv A \cap \bar{B}$




$A \cup B = (A - A \cap B) \cup B$

disjoint

$\Rightarrow P(A \cup B) = P(A - A \cap B) + P(B)$ (1)

also $A = (A - A \cap B) \cup A \cap B$ (disjoint)

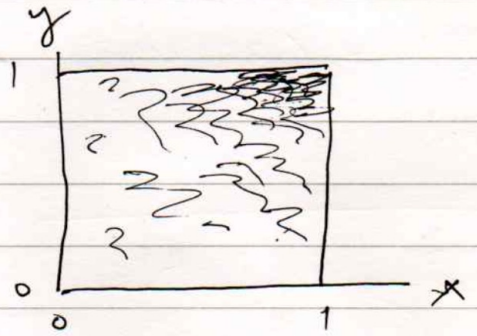
$\Rightarrow P(A) = P(A - A \cap B) + P(A \cap B)$ (2)

Use (1) & (2) to eliminate $P(A - A \cap B)$

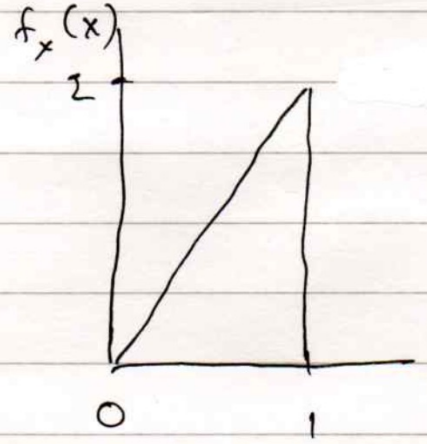
$\Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2) Example w/ joint, marginal, conditional pdfs

Consider $f(x,y) = 4xy$ $\quad 0 \le x \le 1$, $\quad 0 \le y \le 1$




- Marginal pdf $\quad f_x(x) = \int_0^1 4xy\, dy$




$= 2x, \quad 0 \le x \le 1$

By symmetry,

$f_y(y) = 2y, \quad 0 \le y \le 1$

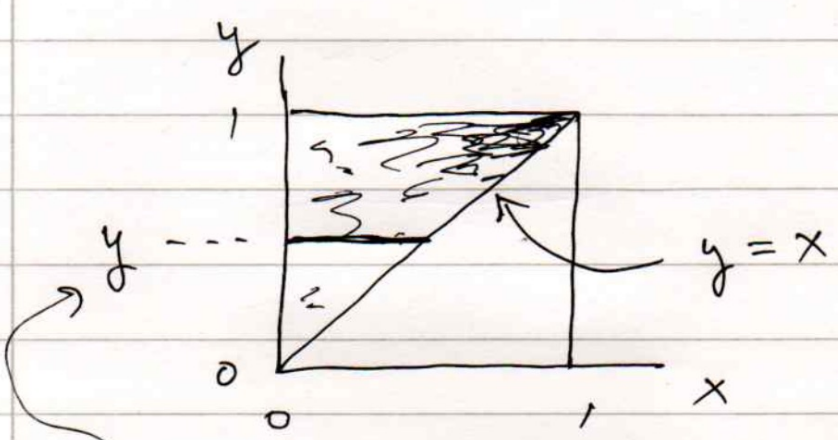
- Conditional pdf

$$f(x|y) = \frac{f(x,y)}{f_y(y)} = \frac{4xy}{2y}$$

$$= 2x, \quad 0 \le x \le 1$$

$\Rightarrow$ $x$ & $y$ independent ( $f(x|y)$ indep. of $y$ )

3) $f(x,y) = \begin{cases} 8xy & 0 \le x \le 1,\ x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$




$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^y 8xy\,dx$$

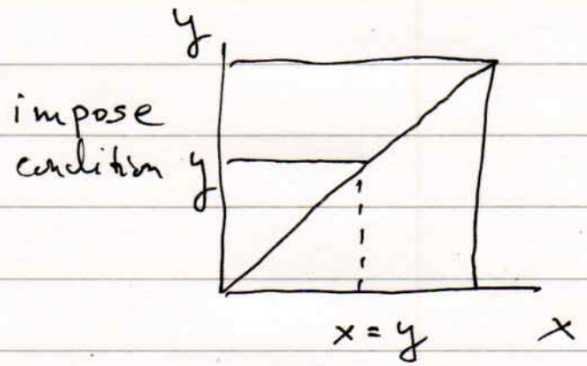
$$= 4y^3, \quad 0 \le y \le 1$$

$$f_x(x) = \int_x^1 8xy\,dy = 4xy^2\Big|_x^1$$

$$= 4x(1-x^2), \quad 0 \le x \le 1$$

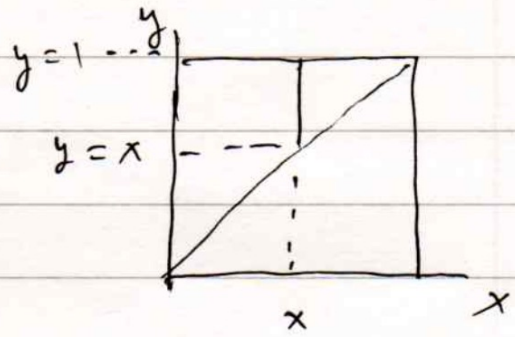
$$f(x|y) = \frac{f(x,y)}{f_y(y)} = \frac{8xy}{4y^3}$$

$$= \frac{2x}{y^2}, \quad 0 \le x \le y$$




$\neq f(x,y) \Rightarrow x$ and $y$ not independent

$$f(y|x) = \frac{8xy}{4x(1-x^2)} = \frac{2y}{1-x^2}, \quad x \le y \le 1$$




↑ impose cond.

Check Bayes' theorem

$$f(x|y) \stackrel{?}{=} \frac{f(y|x)\, f_x(x)}{f_y(y)}$$

$$\frac{2x}{y^2} \stackrel{?}{=} \frac{2y}{1-x^2} \cdot \frac{4x(1-x^2)}{4y^3}$$

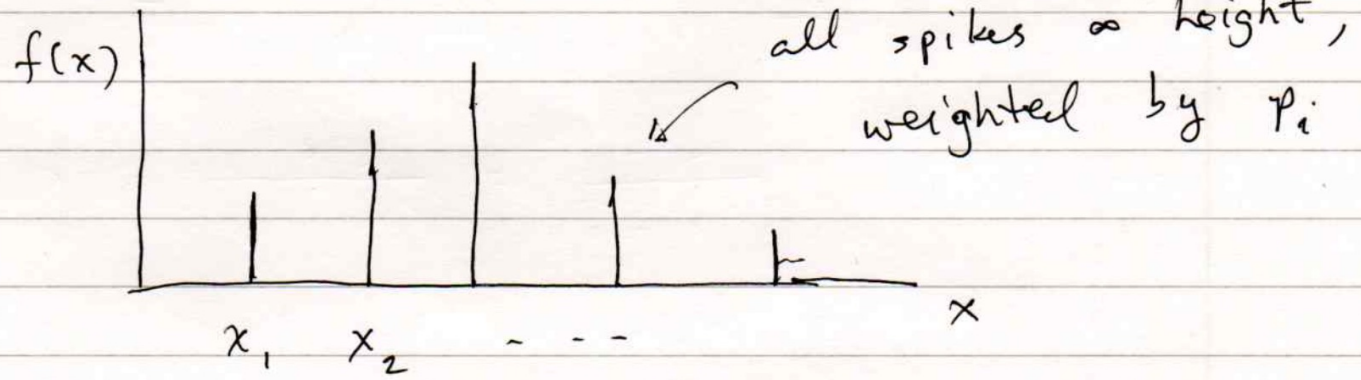
$$= \frac{2x}{y^2} \quad \checkmark$$

4) $\delta$ function $\longleftrightarrow$ pdf / cdf

Suppose discrete r.v. $x_i$, $i = 1, 2, \ldots$

$P(x_i) = p_i$ ( prob. mass func. )

$\Rightarrow$ pdf $f(x) = \sum_i p_i \, \delta(x - x_i)$



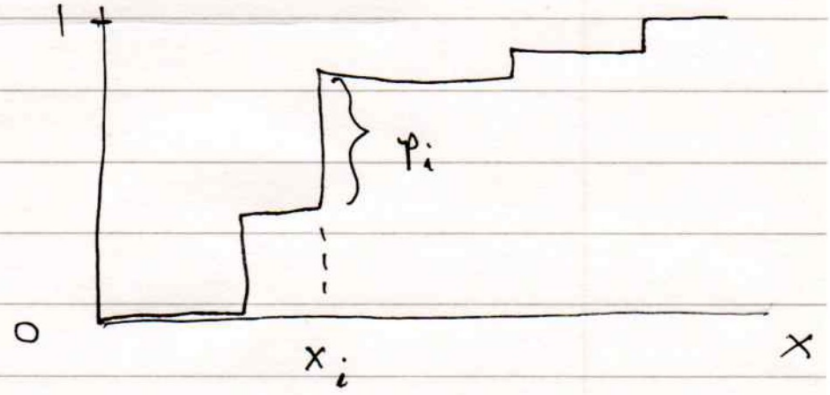
For cumulative dist., use

$$\int_a^b \delta(x - x_i)\, dx = \begin{cases} 1 & a < x_i < b \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow F(x) = \int_{-\infty}^{x} f(x')\, dx'$$

$$= \sum_i^1 p_i \int_{-\infty}^{x} \delta(x' - x_i)\, dx'$$

$$= \sum_{\{i:\, x_i \le x\}}^{1} p_i$$



5) Estimate # of false positives

Given $P(+|D) = 0.90$ ($D$: has disease) } for some med. test

$P(-|\bar{D}) = 0.93$ ($\bar{D}$: doesn't have disease)

Out of a sample of $N = 1000$ individuals,

$n_+ = 250$ have + test result.

Estimate # of <u>false positives</u> in sample

$$n_+|_{\bar{D}} \approx P(+|\bar{D}) \cdot N_{\bar{D}}$$

($N_{\bar{D}}$: # w/ no disease in sample)

$$P(+|\bar{D}) = 1 - P(-|\bar{D}) = 0.07$$

$\pi_D$ ← prior prob to have D

$$N_{\bar{D}} \approx N \pi_{\bar{D}} = N(1 - \pi_D)$$

$$P(+) = P(+|D)\pi_D + P(+|\bar{D})\pi_{\bar{D}} \approx \frac{n_+}{N}$$

$\pi_{\bar{D}} \leftarrow (1-\pi_D)$

$$\Rightarrow \pi_D = \frac{P(+) - P(+|\bar{D})}{P(+|D) - P(+|\bar{D})}$$

$$\Rightarrow \pi_D \approx \frac{\frac{250}{1000} - 0.07}{0.90 - 0.07} = 0.22 \Rightarrow \pi_{\bar{D}} \approx 0.78$$

$$\Rightarrow \underbrace{n_{+|\bar{D}}}_{\#\text{ of false positives}} = P(+|\bar{D}) N \pi_{\bar{D}} = 0.07 \times 1000 \times 0.78$$

$$= 54.6$$