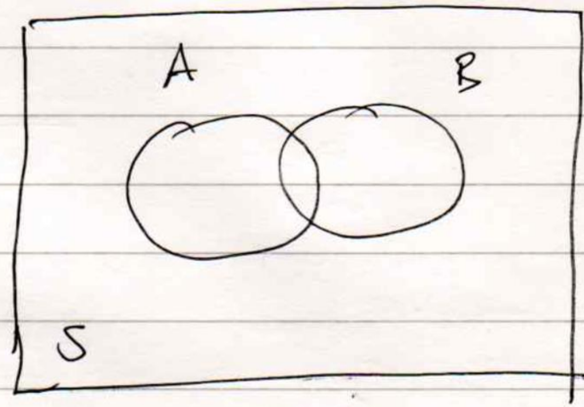
Statistical Data Analysis
Discussion notes – week 1

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

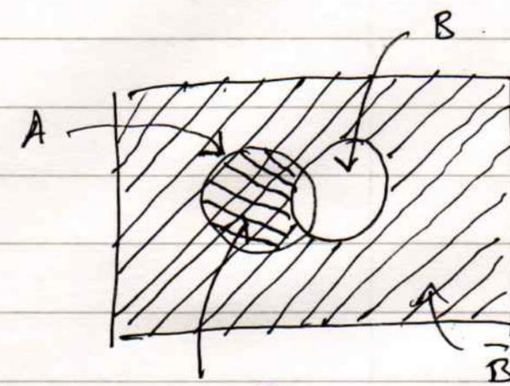
1) Example w/ Kolmogorov axioms:

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

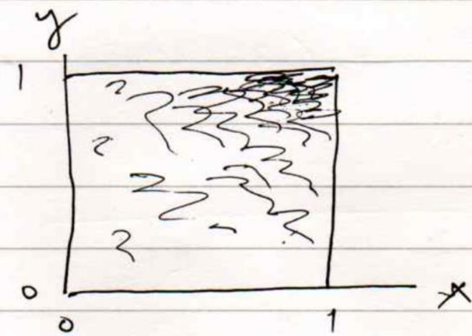
Use ① & ② to eliminate $P(A - A \cap B)$

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

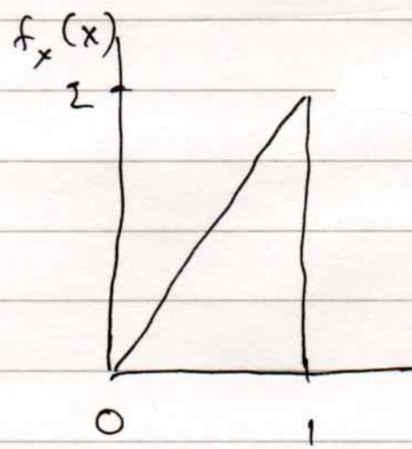
2) Example w/ joint, marginal, conditional pdfs

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

$0 \leq y \leq 1$



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



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

By symmetry,

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

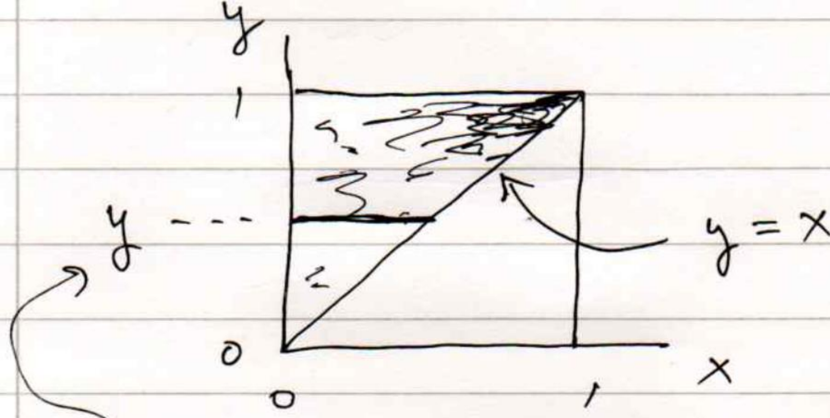
- Conditional pdf

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

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

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

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



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

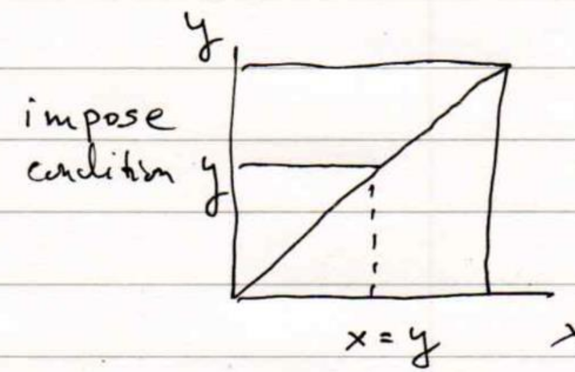
$$= 4y^3, \quad 0 \le y \le 1$$

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

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

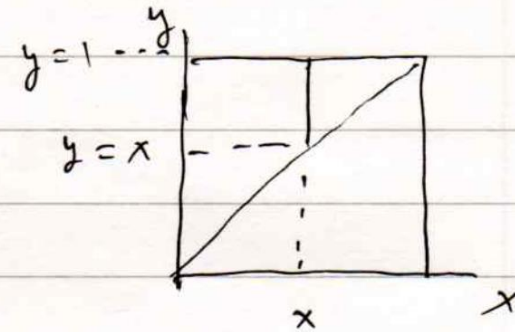
$$f(x|y) = \frac{f(x,y)}{f_y(y)} = \frac{8xy}{4y^3}$$

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



$$\ne f_x(x) \;\Rightarrow\; x, y \text{ \underline{not} independent}$$

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



↑ impose cond.

Check Bayes' theorem

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

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

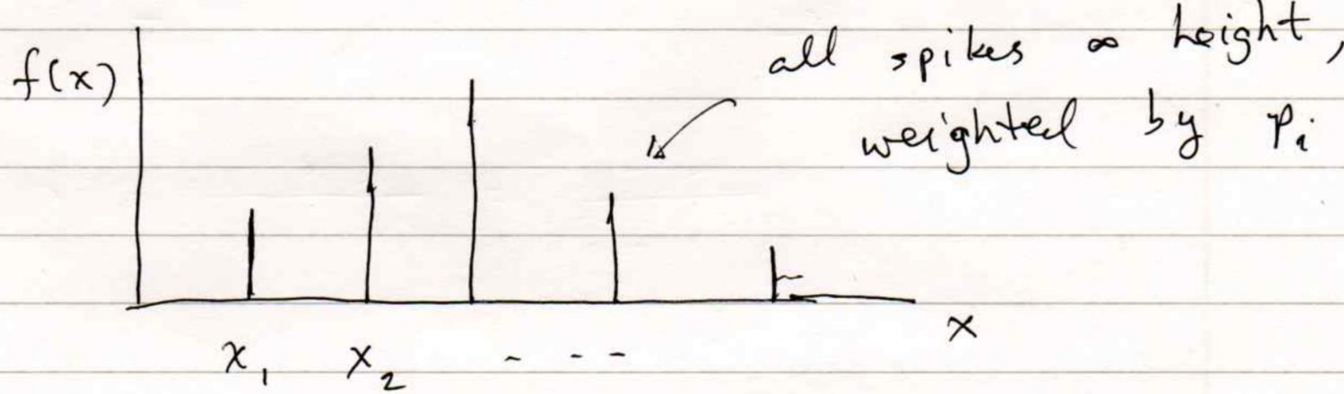
$$= \frac{2x}{y^2} \quad \checkmark$$

4) δ function $\longleftrightarrow$ pdf / cdf

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

$$P(x_i) = p_i \quad \text{(prob. mass func.)}$$

$$\Rightarrow \text{pdf} \quad f(x) = \sum_i p_i \, \delta(x - x_i)$$



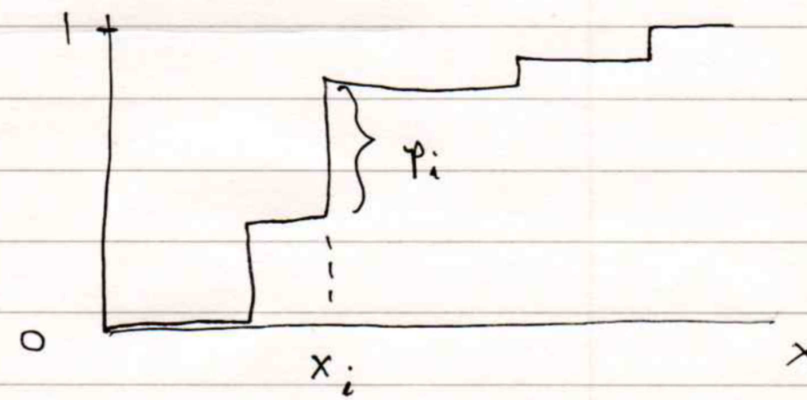
For cumulative dist., use

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

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

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

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



6

5) Estimate # of false positives

$$P(+ \mid D) = 0.90 \quad \text{(has disease)}$$

Given

$$\left.\begin{aligned} P(+ \mid D) &= 0.90 \\ P(- \mid \bar{D}) &= 0.93 \end{aligned}\right\} \text{ for some med. test}$$

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

\# w/ no disease in sample

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

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

prior prob to have D

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

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

$(1-\pi_D)$

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

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

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

$$= 54.6$$

Variable transformation w/ δ function

cf. D. Gillespie, Am. J. Phys. 51 (1983) 520,

Harrison Prosper INFN School of Statistics lectures on probability (2024)

Theorem: $\vec{x} \sim f(\vec{x})$ $\quad \vec{x} = (x_1, \ldots, x_n)$

proof

$\vec{y} = \vec{a}(\vec{x})$ $\quad \vec{y} = (y_1, \ldots, y_m)$

pdf $\longrightarrow$ $\boxed{g(\vec{y}) = \int d^n\vec{x} \, f(\vec{x}) \, \delta^{(m)}(\vec{y} - \vec{a}(\vec{x}))}$

Example: $f(x,y)$ $\qquad z = xy$

$$g(z) = \iint dx\, dy\, f(x,y)\, \delta(\underbrace{z - xy})$$

$\hookrightarrow h(y) \equiv z - xy$

Use $\delta(h(y)) = \dfrac{\delta(y - y_0)}{|h'(y_0)|}$, y_0 = zero of h

$$h(y_0) = z - x y_0 = 0 \Rightarrow y_0 = \frac{z}{x}, \quad h'(y_0) = -x$$

$$g(z) = \int dx \int dy\, f(x,y)\, \frac{\delta(y - \frac{z}{x})}{|-x|}$$

$$= \int f\left(x, \frac{z}{x}\right) \frac{dx}{x}$$

same as before

Example of transformation of variables with Mellin convolution

Consider the joint pdf

$$f(x,y) = 1 \quad , \quad 0 \le x \le 1 \; , \; 0 \le y \le 1$$

a) Find pdf of $z = xy$

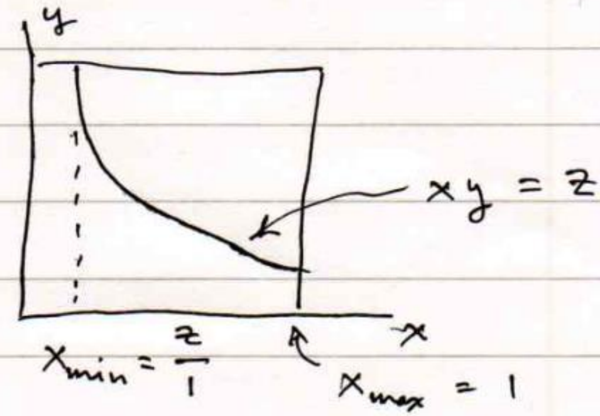
In lectures we found

$$g(z) = \int f\left(x, \frac{z}{x}\right) \frac{dx}{x} \qquad \text{(Mellin convolution)}$$

$$= \int_{x_{min}}^{x_{max}} 1 \cdot \frac{dx}{x}$$

$f(x,y)$ is nonzero for $0 \le x \le 1$, $0 \le y \le 1$

$\Rightarrow 0 \le \frac{z}{x} \le 1 \quad \Rightarrow \quad 0 \le z \le x$



$\Rightarrow x_{min} = z$

$x_{max} = 1$

$$\Rightarrow g(z) = \int_z^1 \frac{dx}{x} = \ln x \Big|_z^1$$

$$= -\ln z \, , \quad 0 < z \le 1$$

b) Alternative method – let

$$z = xy \qquad \Rightarrow \qquad x = u$$

$$u = x \qquad\qquad\qquad y = \frac{z}{u}$$

Jacobian is

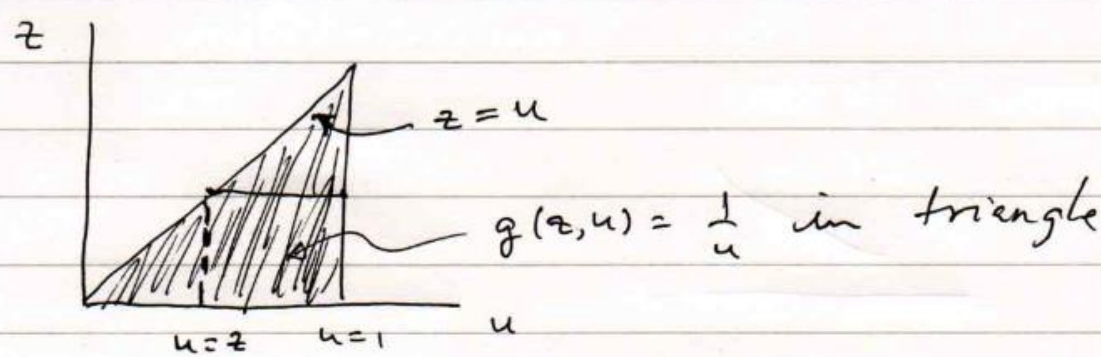
$$J = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial u} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \frac{1}{u} & -\frac{z}{u^2} \end{vmatrix} = -\frac{1}{u}$$

$$g(z,u) = |J|\, f(x(z,u), y(z,u))$$

$$= \frac{1}{u}, \quad 0 \le u \le 1, \quad 0 \le z \le u$$

Because: $0 \le x \le 1 \Rightarrow 0 \le u \le 1$

$0 \le y \le 1 \Rightarrow 0 \le \frac{z}{u} \le 1 \Rightarrow 0 \le z \le u$



$$g_z(z) = \int g(z,u)\, du = \int_z^1 \frac{du}{u} = -\ln z, \quad 0 < z \le 1$$

$\uparrow u \ge z$ (see above)