Lecture 7: Transformations

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

April 20, 2023

Outline

Outline

Change of Variables in One Dimension

Theorem

Let X be a continuous r.v. with PDF f_X, and let $Y = g(X)$, where g is differentiable and strictly increasing (or strictly decreasing). Then the PDF of Y is given by

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|,$$

where $x = g^{-1}(y)$. The support of Y is all $g(x)$ with x in the support of X.

Proof

① W.l.o.G... Let g be a strictly increasing function.

② We consider the CDF of Y.

$\forall\, y \in \mathbb{R}$.

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) \quad (y = g(x) \Rightarrow x = g^{-1}(y),\ x \in \mathbb{R})$$

$$= P(X \le g^{-1}(y))$$

$$= F_X(g^{-1}(y)) = F_X(x).$$

$\Rightarrow$ Chain rule, PDF of Y.

$$f_Y(y) = F_Y'(y) = \frac{\partial F_X(x)}{\partial x} \cdot \frac{dx}{dy}.$$

③ if $g \downarrow$, $f_Y(y) = f_X(x)\left(-\frac{dx}{dy}\right) = f_X(x) \cdot \frac{dx}{dy}$.

④ $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$

Example: Log-Normal PDF

① $X = \log Y$, $X \sim N(0,1)$, PDF of Y.

$Y = g(X) = e^X$. [$y = e^x > 0$. $\Rightarrow x = \log y$

$\Rightarrow \frac{dx}{dy} = \frac{1}{y}$].

② $f_Y(y) = f_X(x) \left|\frac{dx}{dy}\right|$ $(y>0)$.

$= f_X(x) \cdot \frac{1}{y} = f_X(\log y) \cdot \frac{1}{y}$.

③ $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

$\Rightarrow f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\log y)^2} \cdot \frac{1}{y}$, $y > 0$

Change of Variables

Theorem

Let $\mathbf{X} = (X_1, ..., X_n)$ be a continuous random vector with joint PDF $f_{\mathbf{X}}(x)$, and let $\mathbf{Y} = g(\mathbf{X})$ where g is an invertible function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Let $y = g(\mathbf{x})$ and suppose that all the partial derivatives $\frac{\partial x_i}{\partial y_j}$ exists and are continuous, so we can form the ***Jacobian matrix***

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

$$\left|\frac{\partial x}{\partial y}\right| = \frac{1}{\left|\frac{\partial y}{\partial x}\right|}$$

Also assume that the determinant of the Jacobian matrix is never 0. Then the joint PDF of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(x)\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|$$

$|\cdot|$ $\quad \left|\det\left(\frac{\partial x}{\partial y}\right)\right|$

Jacobian or not

① Discrete r.v.s $X, Y > 0$, $Y = X^3$

No Jacobian. $P(Y=y) = P(X=y^{\frac{1}{3}})$

② Continuous r.v.s. $X, Y > 0$, $Y = X^3$

$X = Y^{\frac{1}{3}}$. ($x = y^{\frac{1}{3}}$ $\frac{dx}{dy} = \frac{1}{3} y^{-\frac{2}{3}}$)

$y > 0$, $f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$

$= f_X(y^{\frac{1}{3}}) \cdot \frac{1}{3} y^{-\frac{2}{3}}$.

Box-Muller

① $(X, Y) = g(U, T)$: $\frac{\partial(u,t)}{\partial(x,y)} \times \frac{\partial(x,y)}{\partial(u,t)} = 1$ ✓

$x = \sqrt{2t}\cos u$

$y = \sqrt{2t}\sin u$

$x^2+y^2 = 2t$

$t = \frac{1}{2}(x^2+y^2)$

$u = \arctan\frac{x}{\sqrt{x^2+y^2}}$

$$f_{X,Y}(x,y) = f_{U,T}(u,t) \cdot \left|\frac{\partial(x,y)}{\partial(u,t)}\right|$$

② $f_{U,T}(u,t) = f_U(u) \cdot f_T(t) = \frac{1}{2\pi} \cdot e^{-t}$ $u \in (0, 2\pi)$, $t > 0$

Let $U \sim \text{Unif}(0, 2\pi)$, and let $T \sim \text{Expo}(1)$ be independent of U. Define $X = \sqrt{2T}\cos U$ and $Y = \sqrt{2T}\sin U$. Find the joint PDF of (X, Y). Are they independent? What are their marginal distributions?

③ Jacobian. $\frac{\partial(x,y)}{\partial(u,t)} = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial t} \end{pmatrix} = \begin{pmatrix} -\sqrt{2t}\sin u & \frac{1}{\sqrt{2t}}\cos u \\ \sqrt{2t}\cos u & \frac{1}{\sqrt{2t}}\sin u \end{pmatrix}$

$\det\left(\frac{\partial(x,y)}{\partial(u,t)}\right) = -\sin^2(u) - \cos^2(u) = -1$ $\quad |-1| = 1$

④ $f_{X,Y}(x,y) = \frac{1}{2\pi}e^{-t} \cdot 1 = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}$.

Solution

(5) $f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$, $x, y \in \mathbb{R}$

$h(x) \cdot g(y)$.

$\Rightarrow$ X and Y are independent.

$X \sim N(0,1)$

$Y \sim N(0,1)$

Bivariate Normal Joint PDF

$f_{Z,W}(z,w)$ ~ Bivariate Normal, $-1<\rho<1$, $\mathrm{Corr}(Z,W)=\rho$

1°. $X, Y \sim$ i.i.d. $N(0,1)$

$$\begin{cases} Z = X \\ W = \rho X + \sqrt{1-\rho^2}\, Y \end{cases}$$

2°. $(Z, W) = g(X, Y)$

$$f_{Z,W}(z,w) = f_{X,Y}(x,y) \left| \frac{\partial(x,y)}{\partial(z,w)} \right|$$

3°. Jacobian $\begin{cases} z = x \\ w = \rho x + \sqrt{1-\rho^2}\, y \end{cases} \Rightarrow \begin{cases} x = z \\ y = \frac{1}{\sqrt{1-\rho^2}} w - \frac{\rho}{\sqrt{1-\rho^2}} z \end{cases}$

$$\frac{\partial(x,y)}{\partial(z,w)} = \begin{pmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{pmatrix}$$

$$\det\left(\frac{\partial(x,y)}{\partial(z,w)}\right) = \frac{1}{\sqrt{1-\rho^2}}.$$

Bivariate Normal Joint PDF

$$\begin{cases} x = z \\ y = \frac{1}{\sqrt{1-\rho^2}} w - \frac{\rho}{\sqrt{1-\rho^2}} z \end{cases}$$

$$4^{\circ}. \quad f_{z,w}(z,w) = f_{x,y}(x,y) \cdot \frac{1}{\sqrt{1-\rho^2}}$$

$$= f_X(x) \cdot f_Y(y) \cdot \frac{1}{\sqrt{1-\rho^2}}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \cdot \frac{1}{\sqrt{1-\rho^2}}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}(x^2+y^2)}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left[z^2 + \left(\frac{1}{\sqrt{1-\rho^2}} w - \frac{\rho}{\sqrt{1-\rho^2}} z\right)^2\right]}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[z^2 + w^2 - 2\rho z w\right]}, \quad z, w \in \mathbb{R}.$$

$$\frac{z^2 + w^2 - 2\rho z w}{1-\rho^2} = c$$

$$\Downarrow \ ?$$

$$\frac{z^2}{a^2} + \frac{w^2}{b^2} = 1$$

Outline

Convolution Sums and Integrals

Theorem

If X and Y are independent discrete r.v.s, then the PMF of their sum $T = X + Y$ is

$$P(T = t) = \sum_x P(Y = t - x)\, P(X = x)$$

$$= \sum_y P(X = t - y)\, P(Y = y)$$

If X and Y are independent continuous r.v.s, then the PDF of their sum $T = X + Y$ is

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t - x)\, f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} f_X(t - y)\, f_Y(y)\, dy$$

Proof ①. For continuous r.v.s X, Y.

$$F_T(t) = P(X+Y \le t) \overset{\text{LoTP}}{=} \int_{-\infty}^{\infty} P(X+Y \le t \mid X=x) \cdot f_X(x)\,dx$$

$$= \int_{-\infty}^{\infty} P(Y \le t-x \mid X=x)\, f_X(x)\,dx \quad (X, Y \text{ independent}) \quad = \int_{-\infty}^{\infty} P(Y \le t-x) \cdot f_X(x)\,dx$$

$$= \int_{-\infty}^{\infty} F_Y(t-x) f_X(x)\,dx$$

Diff w.r.t. t

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t-x) f_X(x)\,dx$$

$\int_{b(t)}^{a(t)} F_Y(t-x) f_X(x)\,dx$?

$T = \frac{X}{Y}$

$T = X \cdot Y$

Exponential Convolution

② $T = X+Y$, $V = X$

$(T, V) = g(X, Y)$

$t = x+y \Rightarrow \begin{cases} x = v \\ y = t-v \end{cases}$, $J = \begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$

$\frac{\partial(x,y)}{\partial(T,V)}$

$\Rightarrow f_{T,V}(t,v) = f_{X,Y}(x,y) \cdot |J| = f_X(x) \cdot f_Y(y) \cdot 1 = f_X(v) \cdot f_Y(t-v) \cdot 1$

Let $X, Y \overset{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$. Find the distribution of $T = X + Y$.

$\Rightarrow f_T(t) = \int_{-\infty}^{\infty} f_{T,V}(t,v)\,dv = \int_{-\infty}^{\infty} f_X(v) f_Y(t-v)\,dv$

$= \int_{-\infty}^{\infty} f_X(x) f_Y(t-x)\,dx$

$\forall t \geq 0 \quad f_T(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x)\,dx$

$= \int_0^t \lambda e^{-\lambda(t-x)} \cdot \lambda e^{-\lambda x}\,dx = \lambda^2 t e^{-\lambda t}$

Outline

Order Statistics

Definition

For r.v.s $X_1, X_2, ..., X_n$, the *order statistics* are the random variables $X_{(1)}, ..., X_{(n)}$, where

$$X_{(1)} = \min(X_1, ..., X_n),$$
$X_{(2)}$ is the second-smallest of $X_1, ..., X_n$,
$$\vdots$$
$X_{(n-1)}$ is the second-largest of $X_1, ..., X_n$,
$$X_{(n)} = \max(X_1, ..., X_n)$$

Note that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ by definition. We call $X_{(j)}$ the jth order statistic.

Order Statistics

$X_1, \dots X_n$ are i.i.d. continuous r.v.s. CDF F. PDF f.

1°. $X_{(n)} = \max(X_1, \dots, X_n)$

$\forall x \in \mathbb{R}$, $F_{X_{(n)}}(x) = P(X_{(n)} \le x) = P(\max(X_1, \dots X_n) \le x)$

$= P(X_1 \le x, \dots X_n \le x) = P(X_1 \le x) \cdots P(X_n \le x) = [F(x)]^n$.

$\Rightarrow f_{X_{(n)}}(x) = n[F(x)]^{n-1} \cdot f(x)$.

2°. $X_{(1)} = \min(X_1, \dots, X_n)$

$\forall x \in \mathbb{R}$, $F_{X_{(1)}}(x) = P(X_{(1)} \le x) = 1 - P(X_{(1)} > x)$

$= 1 - P(\min(X_1, \dots X_n) > x) = 1 - \underbrace{P(X_1 > x)}_{1-F(x)} \cdots P(X_n > x)$

$= 1 - [1 - F(x)]^n$

$\Rightarrow f_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} f(x)$.

CDF of Order Statistics

$N \sim Bin(n, F(x))$

$P(N=k) = \binom{n}{k} \cdot F(x)^k \cdot (1-F(x))^{n-k}$

$F(x) = P(X \le x)$, $x \in (0,1)$

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with CDF F. Then the CDF of the jth order statistic $X_{(j)}$ is

$$P\left(X_{(j)} \le x\right) = \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1-F(x))^{n-k}.$$

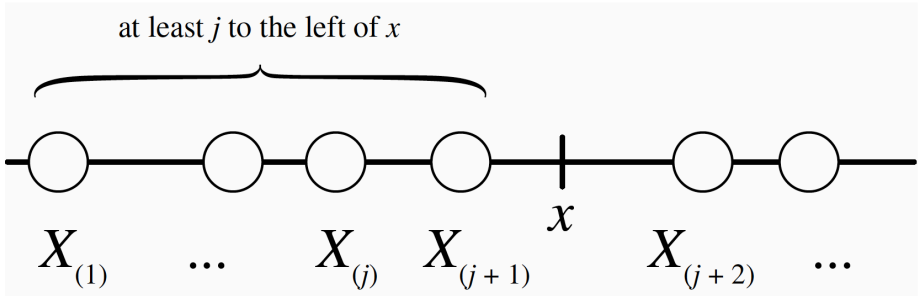
$\sum_{k=j}^{n} \cdot P(N=k) = P(N \ge j)$

Proof

① Given $x \in \mathbb{R}$, construct n independent Bernoulli trials.

For each trial i, the result is successful if r.v. X_i lands to the left of x; such successful prob. $P(X_i < x) = F(x)$.

② Define a new r.v. N : # of X_i that land to the left of x.



$N \sim \text{Bin}(n, F(x))$.

③ $P(X_{(j)} \le x) = P(\text{"at least } j \text{ of } X_1, \dots X_n \text{ land to the left of } x\text{"})$

$= P(N \ge j) = \sum_{k=j}^{n} P(N=k)$

$= \sum_{k=j}^{n} \binom{n}{k} \cdot F(x)^k \cdot [1-F(x)]^{n-k}$.

Proof

PDF of Order Statistic

Two methods of the pdf: CDF $\xrightarrow{\text{diff}}$ PDF.

PDF $\cdot\, dx \rightarrow$ prob. $(x-\frac{1}{2}dx,\ x+\frac{1}{2}dx)$

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with CDF F and PDF f. Then the marginal PDF of the jth order statistic $X_{(j)}$ is

$$f_{X_{(j)}}(x) = n \binom{n-1}{j-1} f(x) F(x)^{j-1} (1-F(x))^{n-j}.$$

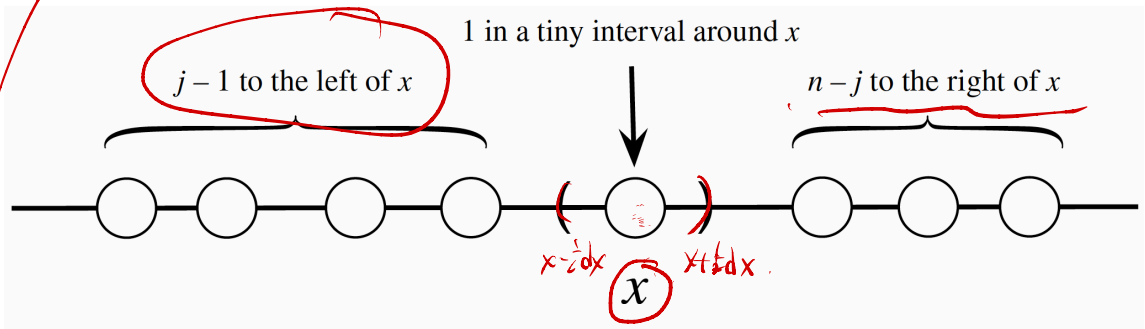
$f_{X_{(j)}}(x) \cdot dx \sim$ prob. $(X_{(j)} \in (x-\frac{1}{2}dx,\ x+\frac{1}{2}dx))$

Proof

① prob. $\left(X_{(j)} \in (x-\frac{1}{2}dx, x+\frac{1}{2}dx) \right) \simeq f_{X_{(j)}}(x)\,dx$

② 1^0. choose one of $X_1, \ldots, X_n$, falls into the interval $(x-\frac{1}{2}dx, x+\frac{1}{2}dx)$

n choices; each with prob. $f(x)\cdot dx$.



2^0. choose $j-1$ out of $n-1$ remaining r.v.s., put on the left of x.

$N' \sim Bin(n-1, F(x))$;

$P(N'=j-1) = \binom{n-1}{j-1} \cdot [F(x)]^{j-1} \cdot (1-F(x))^{n-j}$

③ $f_{X_{(j)}}(x)\cdot dx \overset{\lim_{dx\to 0}}{\simeq} n f(x)\,dx \cdot \binom{n-1}{j-1} [F(x)]^{j-1} (1-F(x))^{n-j}$

$f_{X_{(j)}}(x) = n f(x) \cdot \binom{n-1}{j-1} \cdot [F(x)]^{j-1} \cdot (1-F(x))^{n-j}$

Example: Order Statistics of Uniforms

1° $U_1, \dots, U_n \sim$ i.i.d. $\mathrm{Unif}(0,1)$. U.

CDF. $F_U(x) = x \quad (0 \le x \le 1)$

PDF $f_U(x) = 1 \quad (0 \le x \le 1)$

2° $U_{(1)}, \dots U_{(n)}$ order statistics.

$\forall\, 0 \le x \le 1$, $f_{U_{(j)}}(x) = n \binom{n-1}{j-1} f_U(x) [F_U(x)]^{j-1} [1-F_U(x)]^{n-j}$

$= n \binom{n-1}{j-1} \cdot 1 \cdot x^{j-1} (1-x)^{n-j}$

$= \frac{n!}{(n-j)!\,(j-1)!} \cdot x^{j-1} (1-x)^{n-j}$.

Example: Order Statistics of Uniforms

Related Identity

Application : RL (Thompson Sampling)

① $U_1, \dots, U_n \sim$ i.i.d. $\mathrm{unif}(0,1)$.

n independent Bernoulli trials. "$U_i < p$". success.

$\Pr(U_i < p) = p$.

Theorem

For $0 < p < 1$, and nonnegative integer k, we have

$$\sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} = \frac{n!}{k!(n-k-1)!} \int_p^1 x^k (1-x)^{n-k-1} dx$$

② N : # of successful trials. $N \sim \mathrm{Bin}(n, p)$.

③ LHS : $\Pr(N \le k)$ = Pr("at most k of $U_1, \dots, U_n$ are left of p").

$= \sum_{j=0}^{k} P(N=j)$ $= \Pr(U_{(k+1)} > p) = \int_p^1 f_{U_{(k+1)}}(x) \cdot dx$

Proof

Outline

Beta Distribution

1° $a=b=1$, $f(x)=\frac{1}{\beta(a,b)}=\frac{1}{\beta(1,1)}$ is a constant.

$\int_0^1 f(x)dx=1 \Rightarrow \frac{1}{\beta(1,1)}=1 \Rightarrow f(x)=1$, $0<x<1$

Definition

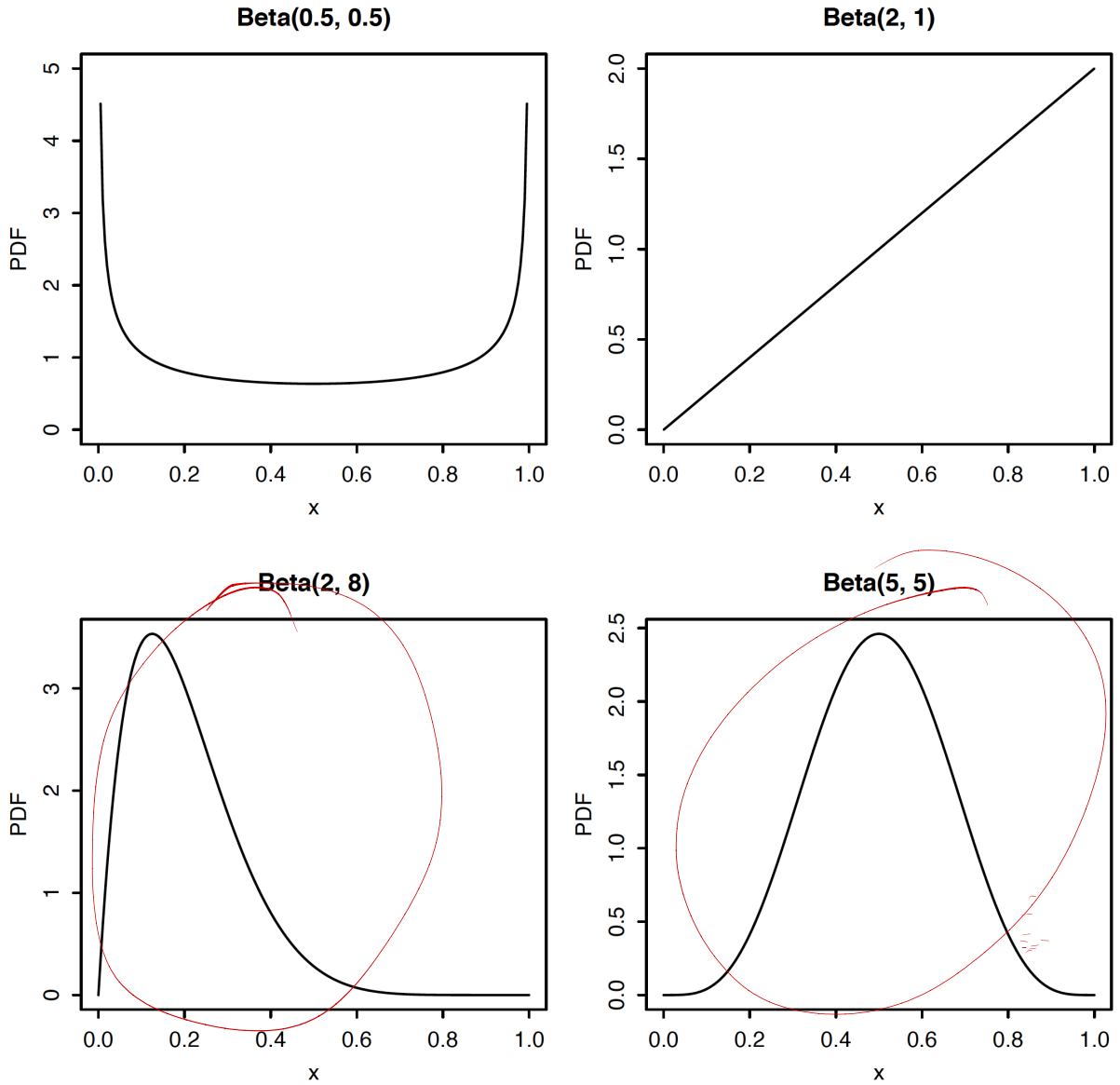
An r.v. X is said to have the *Beta distribution* with parameters a and b, $a>0$ and $b>0$, if its PDF is

$f(x) \propto x^{a-1}(1-x)^{b-1}$

$$f(x)=\frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1},\ 0<x<1,$$

where the constant $\beta(a,b)$ is chosen to make the PDF integrate to 1. We write this as $X \sim \mathrm{Beta}(a,b)$. Beta distribution is a generalization of uniform distribution.

PDF of Beta Distribution



Beta Integral

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx.$$

Story: Bayes' billiards

② $P(Y=k \mid X=x) = \binom{n}{k} x^k (1-x)^{n-k}$

③ $\int_0^1 P(Y=k \mid X=x) \cdot f_X(x)\, dx \overset{\text{LOTP}}{=} P(Y=k)$

Show *without using calculus* that for any integers k and n with $0 \leq k \leq n$,

$$\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}.$$

① $X \sim \text{Bin}(n, x)$ $P(X=k)$

① $Y \sim \text{Bin}(n, X)$, X is a r.v. $\sim \text{Unif}(0,1)$

$Y \mid X=x \sim \text{Bin}(n, x)$ $f_X(x) = 1$

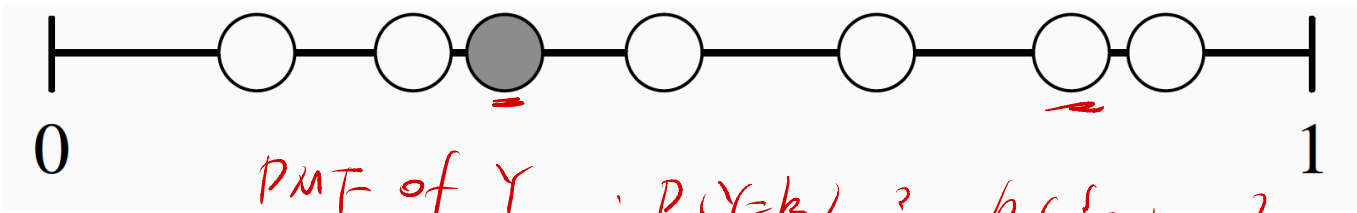
Solution

Story 1: $n+1$ balls: n white balls, 1 gray ball.

Randomly put each ball $\longrightarrow [0,1]$ s.t. postion of balls $\sim$ i.i.d. $unif(0,1)$

Y : # of white balls to the left of the gray ball.

Y : discrete r.v. $\in \{0,1,\dots n\}$



PMF of Y : $P(Y=k)$? $k \in \{0,1,\dots n\}$

① Conditioning on the postion of gray ball B, $\sim unif(0,1)$

② $f_B(p) = 1 = 1$, $0<p<1$

③ Conditioning on $B=p$, $Y|B=p \sim Bin(n,p)$

④ $P(Y=k) \overset{LOTP}{=} \int_0^1 P(Y=k|B=p) f_B(p)\,dp = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k}\,dp$

Solution

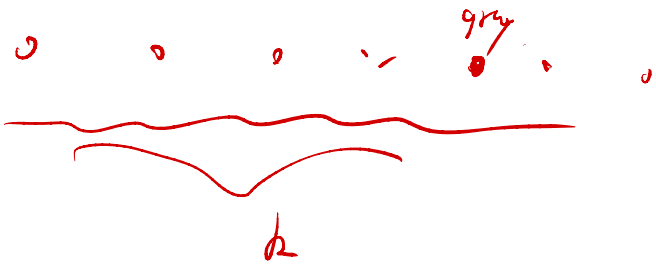
Story 2 :

① $n+1$ white balls $\longrightarrow [0,1]$.

position $\sim$ unif $(0,1)$ iid.

② Choose one white ball at random (equal prob.) and paint it gray

③ Y : # of white balls to the left of the gray ball.



$P(Y=k) = ?$

$\frac{1}{n+1}$. ✓

Solution

$$\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \, dp = \frac{1}{n+1}$$

$k = a-1$

$n-k = b-1$

$a, b > 0$ integer.

$a = k+1$; $b = n-k+1$; $n = a+b-2$;

$p = x$

$$\Rightarrow \int_0^1 \binom{a+b-2}{a-1} x^{a-1} (1-x)^{b-1} \, dx = \frac{1}{a+b-1}$$

$$\Rightarrow \int_0^1 x^{a-1} (1-x)^{b-1} \, dx = \frac{1}{\binom{a+b-2}{a-1} \cdot (a+b-1)}$$

$$= \frac{(a-1)!\,(b-1)!}{(a+b-1)!} = B(a,b)$$

Story: Beta-Binomial Conjugacy

$\hat{p} = \frac{k}{n}$; n tosses. k lands Heads.

$n = 10^{10}$; $k = \frac{1}{2} \times 10^{10}$. $\hat{p} = \frac{1}{2}$. ✓

We have a coin that lands Heads with probability p, but we don't know what p is. Our goal is to infer the value of p after observing the outcomes of n tosses of the coin. The larger that n is, the more accurately we should be able to estimate p.

$n = 5$, $k = 5$; $\hat{p} = 1$?

Bayesian Inference

- Treats all unknown quantities as random variables.
- In the Bayesian approach, we would treat the unknown probability p as a random variable and give p a distribution.
- This is called a **prior distribution**, and it reflects our uncertainty about the true value of p before observing the coin tosses.
- After the experiment is performed and the data are gathered, the prior distribution is updated using Bayes' rule; this yields the **posterior distribution**, which reflects our new beliefs about p.

Story: Beta-Binomial Conjugacy

① p: unknown, model it as a r.v $\in [0,1]$

prior distribution. $p \sim \text{Beta}(a,b)$.

② Data model: n tosses of coin.

X: # of heads.

$X \mid p = p \sim \text{Bin}(n,p)$ (p: r.v, real number)

③ $f(p)$: prior distribution PDF of p.

$f(p \mid X=k)$: posterior distribution pdf of p.

Story: Beta-Binomial Conjugacy

$p \sim \text{Beta}(a,b)$

$$f(p|X=k) = \frac{Pr(X=k|p)}{Pr(X=k)} \cdot f(p) = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{Pr(X=k)} \cdot \frac{1}{B(a,b)} \cdot p^{a-1}(1-p)^{b-1}$$

$$Pr(X=k) \overset{LOTP}{=} \int_0^1 Pr(X=k|p) f(p)\, dp = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \frac{1}{B(a,b)} \cdot p^{a-1}(1-p)^{b-1} dp$$

④ $f(p|X=k)$ is a function of p. [every item that does not dependent on p can be regarded as a constant]

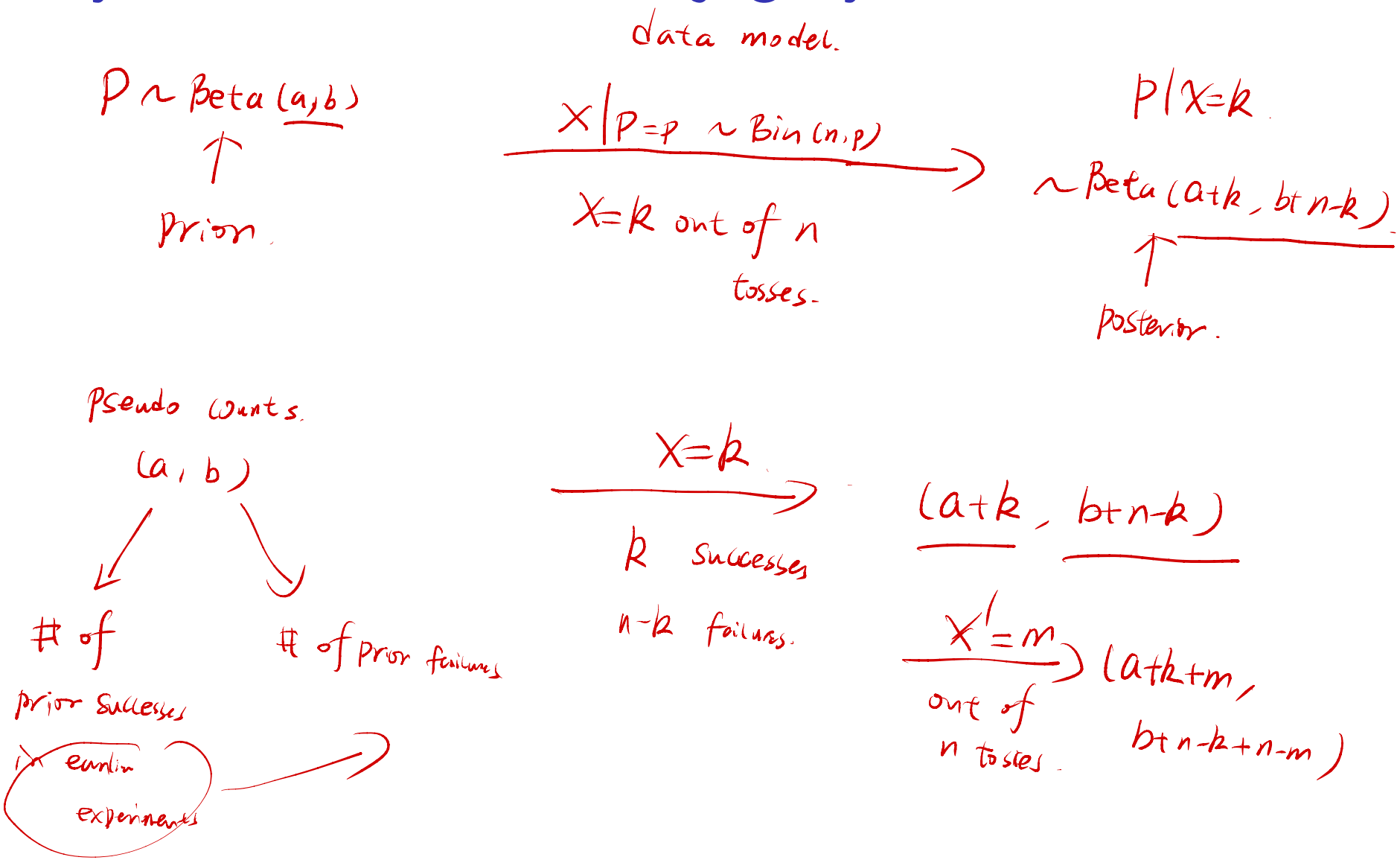
$$f(p|X=k) = C \cdot p^k (1-p)^{n-k} \cdot p^{a-1} \cdot (1-p)^{b-1} = C \cdot p^{k+a-1} (1-p)^{n-k+b-1}$$

$$\propto p^{k+a-1} \cdot (1-p)^{n-k+b-1}$$

$\text{Beta}(k+a, n-k+b)$

$$\Rightarrow p|X=k \sim \text{Beta}(a+k, b+n-k)$$

Story: Beta-Binomial Conjugacy



Story: Beta-Binomial Conjugacy

- Furthermore, notice the very simple formula for updating the distribution of p.
- We just add the number of observed successes, k, to the first parameter of the Beta distribution.
- We also add the number of observed failures, $n - k$, to the second parameter of the Beta distribution.
- So a and b have a concrete interpretation in this context:
  - a as the number of prior successes in earlier experiments
  - b as the number of prior failures in earlier experiments
  - a, b: pseudo counts

Mean vs. Bayesian Average

$Y \sim \text{Beta}(a, b)$

$E(Y) = \frac{a}{a+b}$

- Infer the value of p (probability of coin lands heads)
- Observed k heads out of n tosses of the coin
- Mean: $\frac{k}{n}$
- Bayesian Average: $E(p|X = k) = \frac{a+k}{a+b+n}$
- Suppose the prior distribution is Unif(0,1): $a = 1, b = 1$
- Bayesian Average: $\frac{k+1}{n+2}$
- When $k = n$, we have: 1 (mean) vs. $\frac{n+1}{n+2}$ (Bayesian average)

$n = 2, 3, 4$

$n \to \infty \to 1$.

Story: Beta-Binomial Conjugacy

If we have a Beta prior distribution on p and data that are conditionally Binomial given p, then when going from prior to posterior, we don't leave the family of Beta distributions. We say that **the Beta is the conjugate prior of the Binomial**.

Outline

Gamma Function

Gamma Distribution (Motivation)

prior distribution of unknown parameters

$(0, \infty)$

Definition

The gamma function Γ is defined by

$$\Gamma(a) = \int_0^\infty x^a e^{-x} \frac{dx}{x},$$

for real numbers $a > 0$.

$$1 = \int_0^\infty \underbrace{\frac{1}{\Gamma(a)} x^{a-1} e^{-x}}_{f(x)} dx = \int_0^\infty f(x)\, dx$$

pdf of Gamma $(a, 1)$

Property of Gamma Function

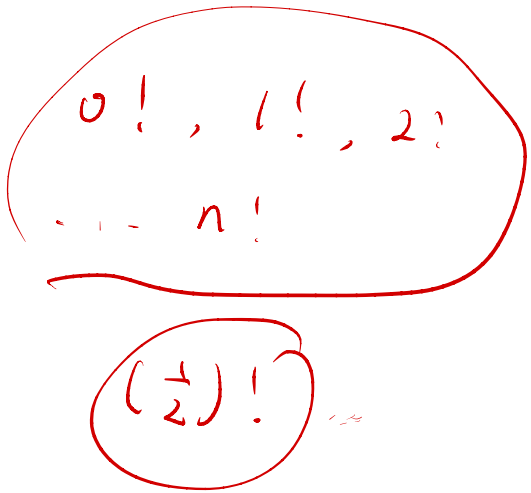
- $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$.
- $\Gamma(n) = (n-1)!$ if n is a positive integer.

$\Gamma(n+1) = n\Gamma(n)$

$\Gamma(1) = \int_0^\infty 1 \cdot e^{-x} dx = 1$

$\Gamma(2) = 1$

$\Gamma(3) = 2$, ... $\Gamma(n) = (n-1)!$



Gamma Distribution

$a=1$, $f(y)=\frac{1}{\Gamma(1)}\cdot\lambda y\cdot e^{-\lambda y}\cdot\frac{1}{y}$

$=\lambda e^{-\lambda y}$, $y>0$.

$\lambda>0$

$\text{Expo}(\lambda)$

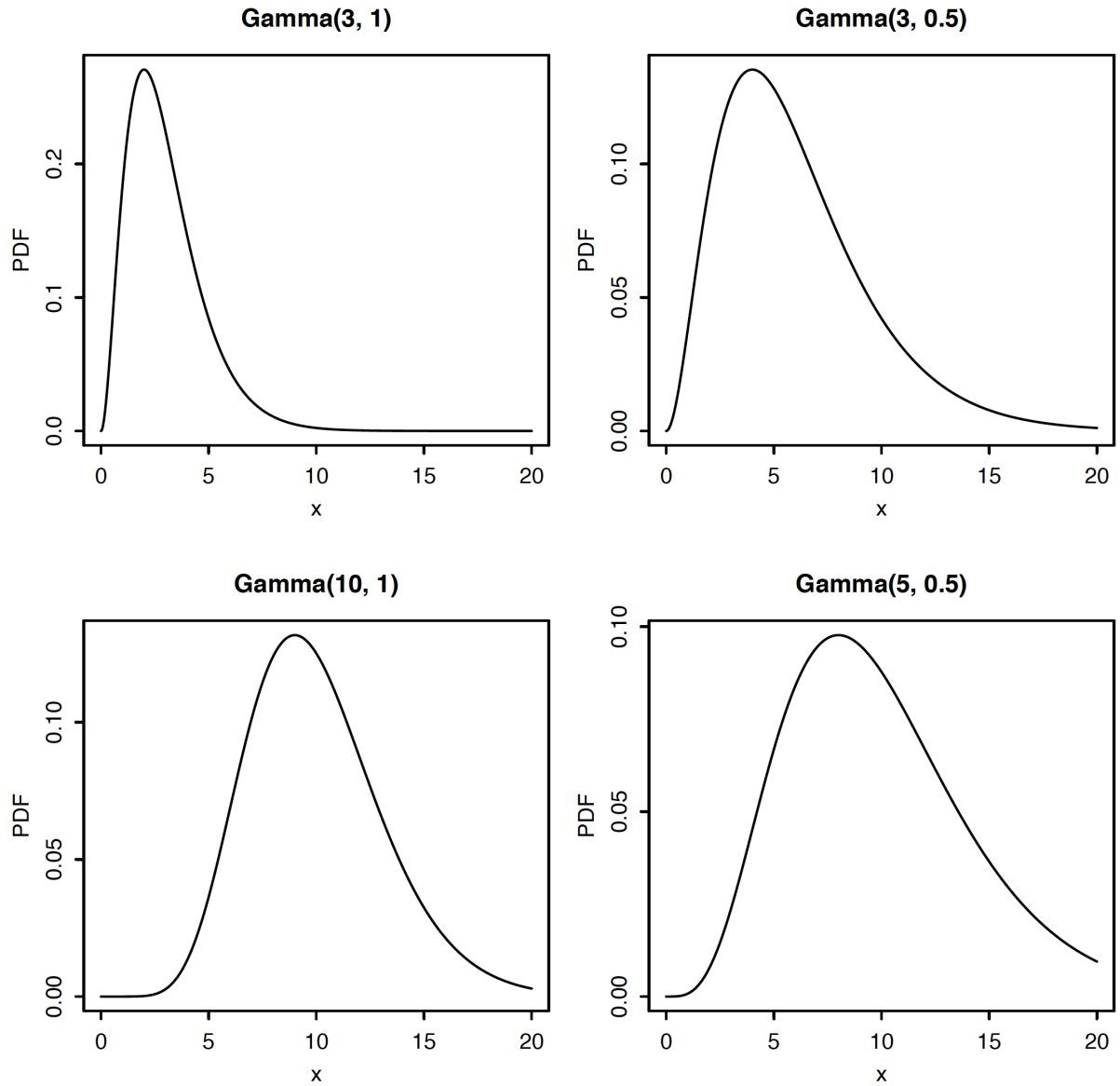
Definition

An r.v. Y is said to have the *Gamma distribution* with parameters a and λ, $a>0$ and $\lambda>0$, if its PDF is

$$f(y)=\frac{1}{\Gamma(a)}(\lambda y)^a e^{-\lambda y}\frac{1}{y},\quad y>0.$$

We write $Y\sim\text{Gamma}(a,\lambda)$. Gamma distribution is a generalization of the exponential distribution.

PDF of Gamma Distribution



Moments of Gamma Distribution

$Y \sim \text{Gamma}(a, \lambda)$

$E(Y) = \frac{a}{\lambda}$;

$\text{Var}(Y) = \frac{a}{\lambda^2}$;

if $\lambda = 1$;

$E(Y) = \text{Var}(Y) = a$

Pois

Poisson - Gamma Duality

if $X \sim \text{pois}(\lambda)$, PMF $P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$

if $X \sim \text{Gamma}(k+1, 1)$, PDF $f(x) = \frac{x^k e^{-x}}{k!}$

Gamma: Convolution of Exponential

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Expo}(\lambda)$. Then

$$X_1 + \cdots + X_n \sim \mathrm{Gamma}(n, \lambda).$$

$M_{X(t)}$

MGF $=$ MGF ✓

Proof

Beta-Gamma Connection

When we add independent Gamma r.v.s X and Y with the same rate λ, the total $X+Y$ has a Gamma distribution, the fraction $\frac{X}{X+Y}$ has a Beta distribution, and the total is independent of the fraction.

Story: Bank–post Office

While running errands, you need to go to the bank, then to the post office. Let $X \sim \text{Gamma}(a, \lambda)$ be your waiting time in line at the bank, and let $Y \sim \text{Gamma}(b, \lambda)$ be your waiting time in line at the post office (with the same λ for both). Assume X and Y are independent. What is the joint distribution of $T = X + Y$ (your total wait at the bank and post office) and $W = \frac{X}{X+Y}$ (the fraction of your waiting time spent at the bank)?

Story: Bank–post Office

$T = X + Y$; $W = \frac{X}{X+Y}$.

① $t>0$, $t = x+y$; $w>0$, $w = \frac{x}{x+y}$ $\Rightarrow$ $x = tw$, $y = t(1-w)$

$\Rightarrow \frac{\partial(x,y)}{\partial(t,w)} = \begin{pmatrix} w & t \\ 1-w & -t \end{pmatrix}$

$\det\left(\frac{\partial(x,y)}{\partial(t,w)}\right) = -t < 0$. $\quad | \; | = t > 0$.

② $f_{T,W}(t,w) = f_{X,Y}(x,y) \left|\frac{\partial(x,y)}{\partial(t,w)}\right| = f_X(x) f_Y(y) \cdot t$.

$(|\cdot| = |\det(\cdot)|)$

$= \frac{1}{\Gamma(a)} (\lambda x)^a e^{-\lambda x} \cdot \frac{1}{x} \cdot \frac{1}{\Gamma(b)} \cdot (\lambda y)^b e^{-\lambda y} \cdot \frac{1}{y} \cdot t \quad \begin{pmatrix} x = tw \\ y = t(1-w) \end{pmatrix}$

$= \frac{1}{\Gamma(a)} \cdot \frac{1}{\Gamma(b)} \cdot w^{a-1} (1-w)^{b-1} \cdot (\lambda t)^{a+b} e^{-\lambda t} \cdot \frac{1}{t}$

$= \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1}}_{h(w)} \cdot \underbrace{\frac{1}{\Gamma(a+b)} \cdot (\lambda t)^{a+b} e^{-\lambda t} \cdot \frac{1}{t}}_{g(t)}$

Story: Bank–post Office

(3) T, W independent.

$T \sim Gamma(a+b, \lambda)$

$W \sim Beta(a, b)$.

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} w^{a-1}(1-w)^{b-1}$$

$$\frac{1}{B(a,b)} w^{a-1}(1-w)^{b-1}$$

(4) $\Rightarrow$ $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Story: Bank–post Office

(5). $W \sim \text{Beta}(a, b)$, $a>0$, $b>0$. $E[W]$?

$T = X+Y$; $W = \frac{X}{X+Y}$. $X \sim \text{Gamma}(a, \lambda)$, $E(X) = \frac{a}{\lambda}$

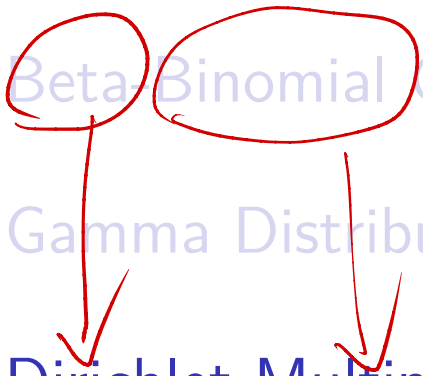
$Y \sim \text{Gamma}(b, \lambda)$ $E(Y) = \frac{b}{\lambda}$

T, W are independent. $E[T \cdot W] = E[T] \cdot E[W]$

$$\Rightarrow E[W] = \frac{E[T \cdot W]}{E[T]} = \frac{E[X]}{E[X+Y]} = \frac{E[X]}{E[X]+E[Y]}$$

$$= \frac{\frac{a}{\lambda}}{\frac{a}{\lambda}+\frac{b}{\lambda}} = \frac{a}{a+b}, \quad a>0, b>0$$

Outline



Story of Multinomial Distribution

Each of n objects is independently placed into one of k categories. An object is placed into category j with probability p_j, where the p_j are nonnegative and $\sum_{j=1}^{k} p_j = 1$. Let X_1 be the number of objects in category 1, X_2 the number of objects in category 2, etc., so that $X_1 + \ldots + X_k = n$. Then $X = (X_1, \ldots, X_k)$ is said to have the Multinomial distribution with parameters n and $\mathbf{p} = (p_1, \ldots, p_k)$. We write this as $\mathbf{X} \sim Mult_k(n, \mathbf{p})$.

Multinomial Joint PMF

Theorem

If $\mathbf{X} \sim \mathrm{Mult}_k(n, \mathbf{p})$, *then the joint PMF of* $\mathbf{X}$ *is*

$$P(X_1 = n_1, ..., X_k = n_k) = \frac{n!}{n_1! n_2! ... n_k!} \cdot p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$$

for $n_1, ..., n_k$ *satisfying* $n_1 + \cdots + n_k = n$.

Dirichlet Distribution

The Dirichlet distribution is parameterized by a vector α of positive real numbers. The PDF is:

$$f(p_1, p_2, \ldots, p_k; \alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

where $p_1 + \ldots + p_k = 1$ and $0 < p_i < 1$.

Story: Dirichlet-Multinomial Conjugacy

① Unknown $P = (p_1, \dots p_k)$.

$P = (p_1, \dots p_k)$.

Prior distribution: Dirichlet distribution.

pdf $f(p) = C \cdot p_1^{\alpha_1 - 1} \cdot p_2^{\alpha_2 - 1} \cdots p_k^{\alpha_k - 1}$.

② Data model: $X = (X_1, \dots X_k)$ $\quad X_1 + \cdots + X_k = n$

$X | P = p \sim \text{Multi}_k(n, p)$.

Define $\vec{n} = (n_1, \dots n_k)$. $\quad \left(X = \vec{n} \iff \begin{matrix} X_1 = n_1 \\ X_2 = n_2 \\ \cdots X_k = n_k \end{matrix} \right)$

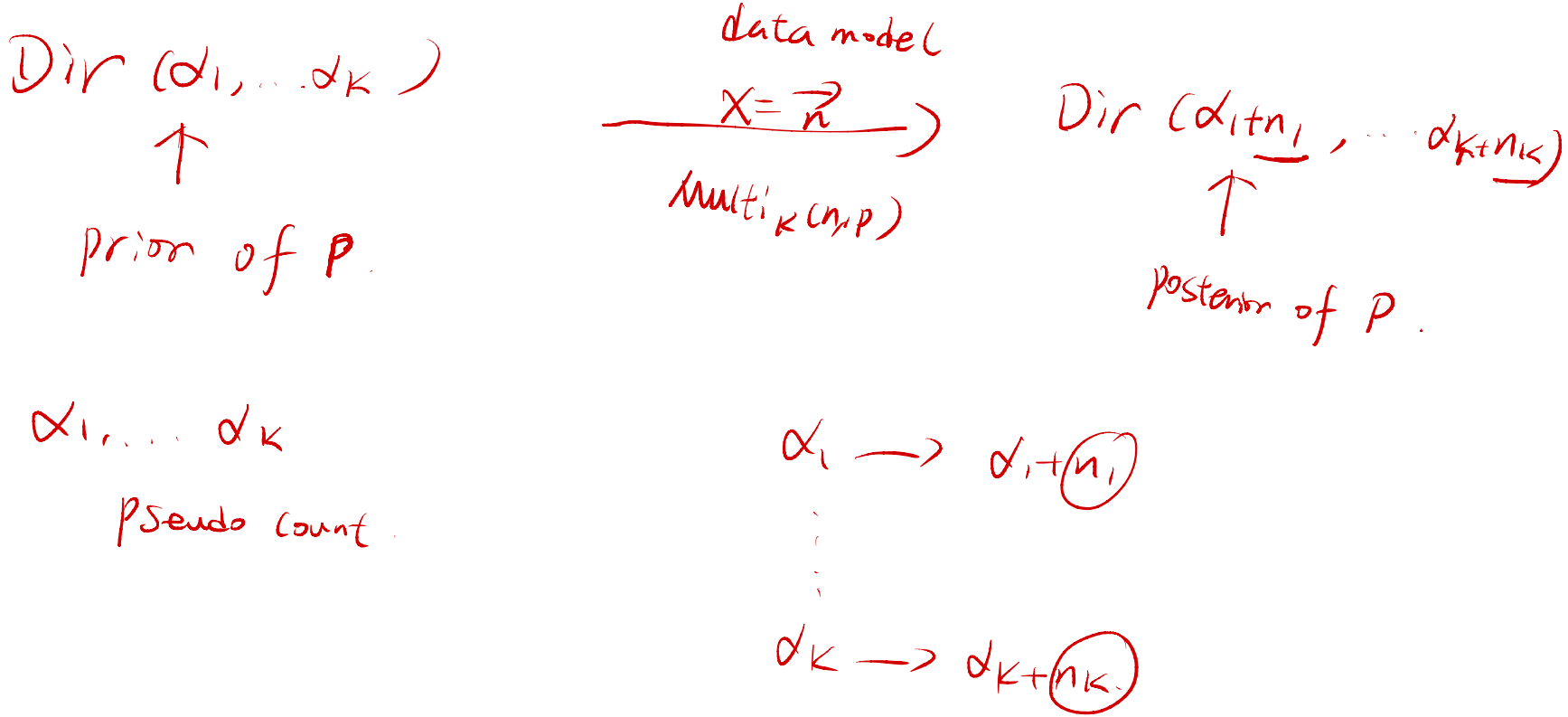
③ $f(p | X = \vec{n}) = \dfrac{Pr(X = \vec{n} | p)}{Pr(X = \vec{n})} \cdot f(p) = \dfrac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k} \cdot C \cdot p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}$

$Pr(X = \vec{n})$

$\propto p_1^{n_1 + \alpha_1 - 1} \cdots p_k^{n_k + \alpha_k - 1}$

Still Dirichlet distribution.

Story: Dirichlet-Multinomial Conjugacy



Story: Dirichlet-Multinomial Conjugacy

If we have a Dirichlet prior distribution on $\mathbf{p}$ and data that are conditionally Multinomial given $\mathbf{p}$, then when going from prior to posterior, we don't leave the family of Dirichlet distributions. We say that the Dirichlet is the conjugate prior of the Multinomial.

Outline

Rating System

- Consumers rely on the collective intelligence of other consumers: rating
- A common metric: 5 star rating
- Requirement: many ratings are needed to make this system work
- Quality of rating system depends on
  - average number of stars
  - average number of reviews

Which One to Choose?

- 1. Presto Coffee Pot - average rating of 5 (1 review).
- 2. Cuisinart Brew Central - average rating of 4.1 (78 reviews).

Example: Movie Ranking

- Data Set : http://grouplens.org/datasets/movielens/
- Top Ten Movies

Top 10 Movies chosen by Mean

title	count	mean
Aiqing wansui (1994)	1	5
They Made Me a Criminal (1939)	1	5
Great Day in Harlem, A (1994)	1	5
Saint of Fort Washington, The (1993)	2	5
Entertaining Angels: The Dorothy Day Story (1996)	1	5
Someone Else's America (1995)	1	5
Star Kid (1997)	3	5
Santa with Muscles (1996)	2	5
Prefontaine (1997)	3	5
Marlene Dietrich: Shadow and Light (1996)	1	5

Tool: Bayesian Estimation

- Mean of star reviews with a limited number of observations
- Useful for recommender services and other predictive algorithms that use preference space measures like star reviews.

Joint Distribution

- To use Bayesian estimation to compute the posterior probability for star ratings, we must use a joint distribution.
- We are not estimating the distribution of some scalar value X but, rather, the joint distributions of the probability estimate of whether or not the reviewer will give the movie a 1, 2, 3, 4, or 5 star rating (not just a simple thumbs up or down).
- In this case, the random variable is a categorical distribution because it can take some value within 1,2,3,4,5 with probabilities as follows:

$$p_1 + p_2 + p_3 + p_4 + p_5 = 1$$

Multinomial Distribution

- We can compute our posterior probability with N observations for five categories with corresponding numbers K_1, K_2, K_3, K_4, K_5 as follows:

$$Pr(O|p_1, p_2, p_3, p_4, p_5) \propto p_1^{K_1} p_2^{K_2} p_3^{K_3} p_4^{K_4} p_5^{K_5}$$

  where $K_1 + \ldots + K_5 = N$.
- This is a multinomial distribution.

Dirichlet Distribution: Prior

- If we include our prior as a distribution of the exact same form in the proportionality equation (e.g. a Dirichlet distribution with parameter α^0), then

$$Pr(p_1, p_2, p_3, p_4, p_5|O) \propto \prod_{j=1}^{5} p_j^{K_j+\alpha_j^0-1}$$

- This is another Dirichlet distribution with another parameter α^1:

$$\alpha_j^1 = K_j + \alpha_j^0, \forall j$$

Expected Average

- What is the expected value of the average rating given a posterior in the shape of our Dirichlet distribution?
- The expected value of the average rating based on the posterior is then computed for our star ratings as follows:

$$E(p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5|O) = \sum_{i=1}^{5} iE(p_i|O)$$

- Using our Dirichlet distribution we can compute the probability of a star value given our observations as the ratio of the Dirichlet parameter for that star to the sum of the Dirichlet parameters:

$$E(p_i|O) = \frac{\alpha_i^1}{\sum_{j=1}^{5} \alpha_j^1}$$

Intra-Item: Bayesian Average Rating

$$\text{Bayes Average Rating} = \frac{\sum_{i=1}^{5} i\alpha_i^0 + \sum_{i=1}^{5} iK_i}{N + \sum_{i=1}^{5} \alpha_i^0}$$

- N: the number of reviews
- $\sum_{i=1}^{5} iK_i$: sum of all review scores
- $\sum_{i=1}^{5} \alpha_i^0$: prior(given) number of reviews
- $\sum_{i=1}^{5} i\alpha_i^0$: prior sum of all review scores

Intra-Item: Bayesian Average Rating

$$\text{Bayes Average Rating} = \frac{C \cdot m + \sum(\text{ratings})}{C + N}$$

- N: the number of reviews
- m: a prior for the average of review scores
- C: a prior for the number of reviews

Example: Movie Ranking

- Data Set : http://grouplens.org/datasets/movielens/
- Top Ten Movies

Case 1: $m = 3.25$ & $C = 50$

title	bayes	count	mean
One Flew Over the Cuckoo's Nest (1975)	4.125796	264	4.291667
Raiders of the Lost Ark (1981)	4.145745	420	4.252381
Rear Window (1954)	4.167954	209	4.387560
The Silence of the Lambs (1991)	4.171591	390	4.289744
The Godfather (1972)	4.171706	413	4.283293
The Usual Suspects (1995)	4.206625	267	4.385768
Casablanca (1942)	4.250853	243	4.456790
The Shawshank Redemption (1994)	4.265766	283	4.445230
Star Wars (1977)	4.270932	583	4.358491
Schindler's List (1993)	4.291667	298	4.466443

Case 2: $m = 2$ & $C = 6$

title	count	bayes	mean
One Flew Over the Cuckoo's Nest (1975)	264	4.244526	4.291667
The Godfather (1972)	413	4.252955	4.283293
The Silence of the Lambs (1991)	390	4.257500	4.289744
Star Wars (1977)	583	4.335582	4.358491
The Usual Suspects (1995)	267	4.335740	4.385768
The Wrong Trousers (1993)	118	4.351562	4.466102
A Close Shave (1995)	112	4.368852	4.491071
The Shawshank Redemption (1994)	283	4.395904	4.445230
Casablanca (1942)	243	4.399209	4.456790
Schindler's List (1993)	298	4.418831	4.466443

Inter-Items: **Pseudo** Bayesian Average Rating

$$\bar{m}_i = \frac{C_i \cdot m_i + \sum(\text{ratings})}{C_i + N}$$

- $\bar{m}_i$: bayesian average rating for item i
- N: the number of reviews for all items
- m_i: average of review scores for item i
- C_i: the number of reviews for item i

Example: Bayesian Changes Order

MacBook	No. Ratings	Ave. Rating	Rank	Bayesian Rating	Bayesian Rank
MB991LL	10	4.920	1	4.436	2
MB403LL	15	4.667	2	4.433	3
MB402LL	228	4.535	3	4.459	1
MC204LL	150	4.310	4	4.401	5
MB061LL	124	4.298	5	4.402	4

Reverse Engineering Amazon

- Bayesian adjustment
- Recency of view
- Reputation score
  - www.amazon.com/review/top-reviewers-classic

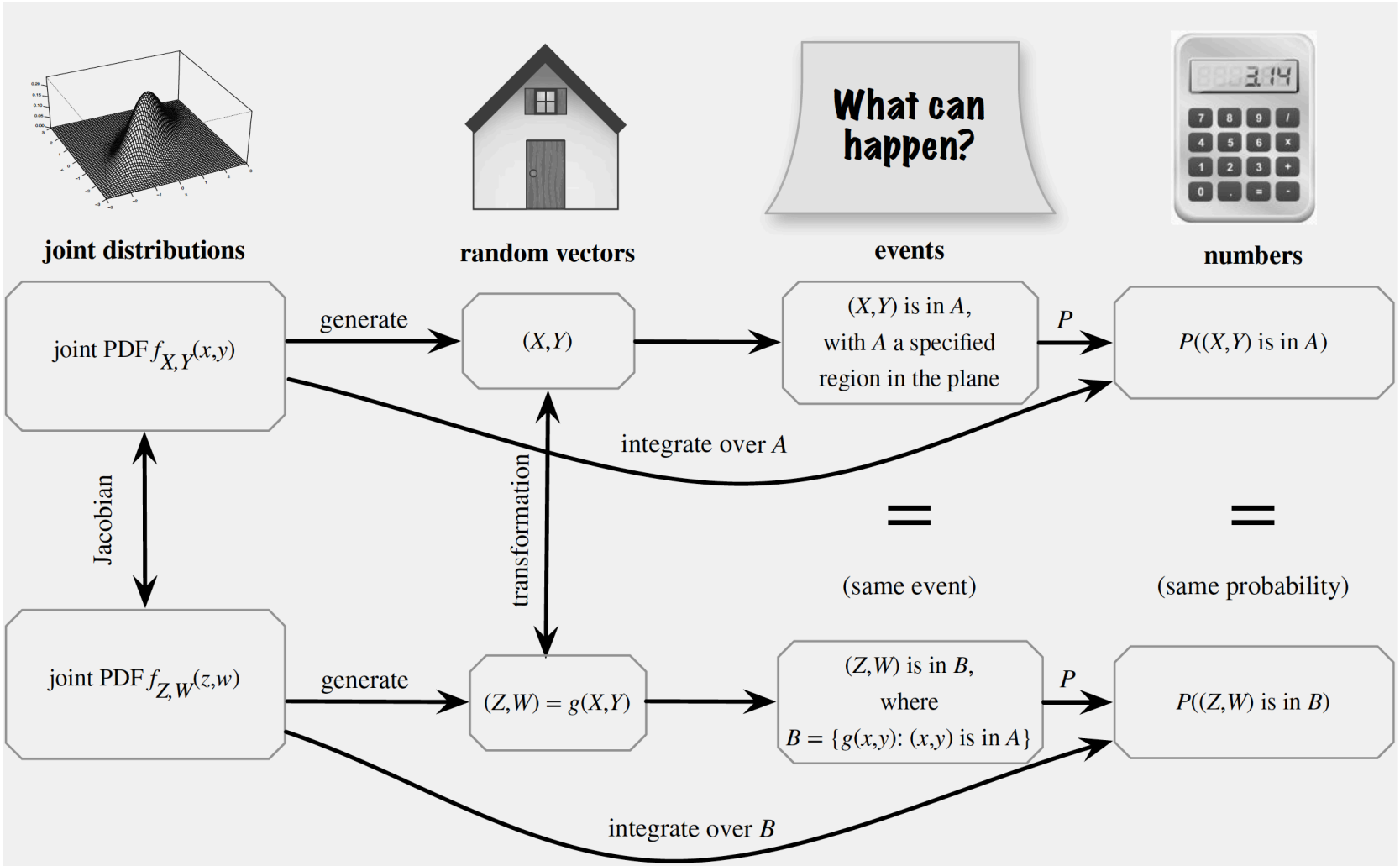
Key Factors

- Bayesian ranking
- Too few or too outdated reviews penalized
- Very high quality reviews help a lot

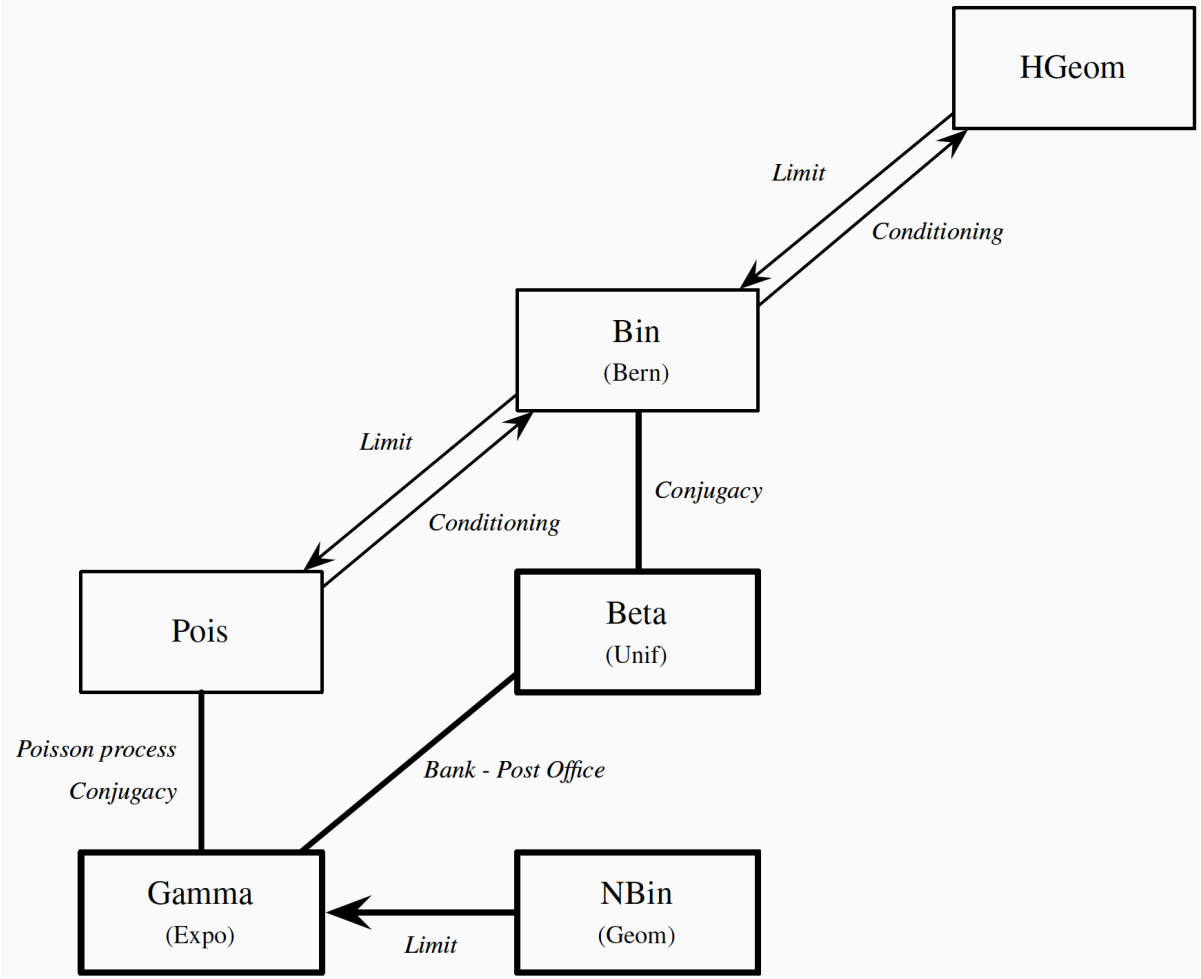
Summary

- Average ratings scalarize a vector and ranks
- Number of ratings should matter, Bayesian ranking does that
- Other statistical methods help too

Summary 1: Transformation



Summary 2


HGeom
Limit
Conditioning
Bin
(Bern)
Limit
Conditioning
Conjugacy
Pois
Beta
(Unif)
Poisson process
Conjugacy
Bank - Post Office
Gamma
(Expo)
Limit
NBin
(Geom)

References

- Chapter 8 of **BH**
- Chapter 4 of **BT**