Lecture 1: Review of Basic Probability

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

March 02, 2020

Outline

Outline

Set

A *set* is a collection of objects. Given two sets A, B, key concepts include

- *empty set*: $\emptyset$
- A is a *subset* of B: $A \subseteq B$
- *union* of A and B: $A \cup B$
- *intersection* of A and B: $A \cap B$
- *complement* of A: A^c
- *De Morgan's laws*:

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Translation Between English & Sets

English	Sets
Events and occurrences	
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
something must happen	$s_{\text{actual}} \in S$
New events from old events	
A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$
Relationships between events	
A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \emptyset$
$A_1, \ldots, A_n$ are a partition of S	$A_1 \cup \cdots \cup A_n = S, A_i \cap A_j = \emptyset$ for $i \neq j$

General Definition of Probability

Definition

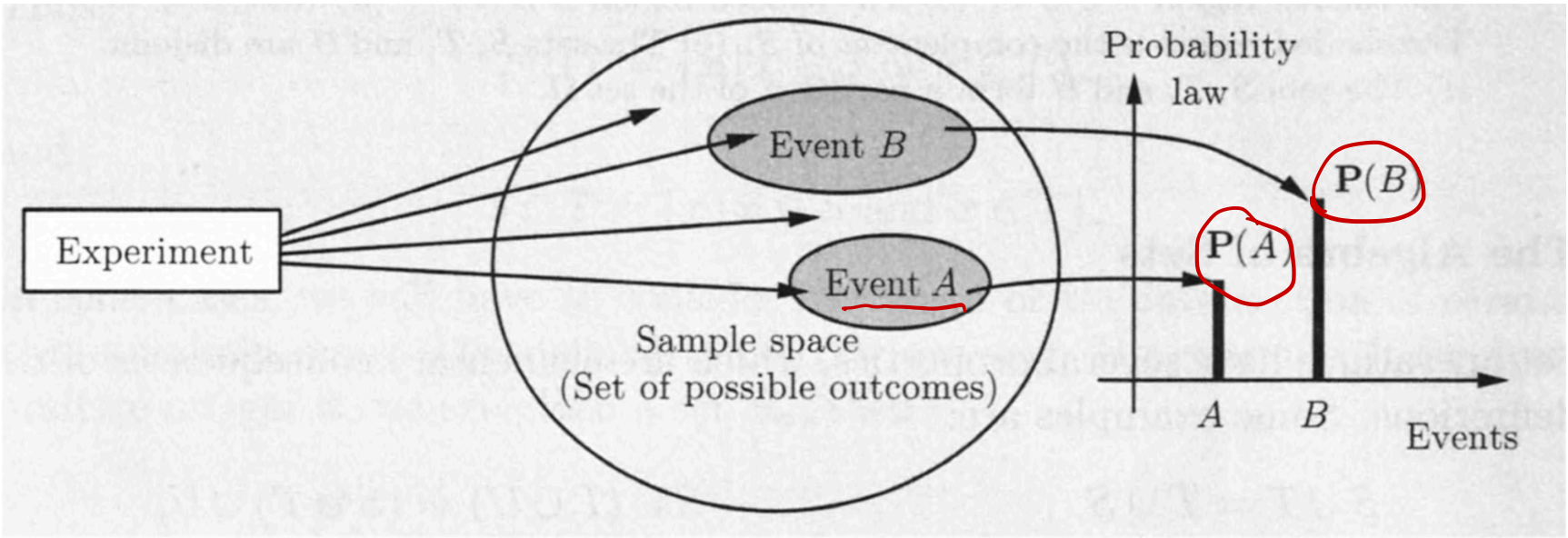
A *probability space* consists of a *sample space* S and a *probability function* P which takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output. The function P must satisfy the following axioms:

1. $P(\emptyset) = 0$, $P(S) = 1$.
2. If $A_1, A_2, \ldots$ are disjoint events, then

$$P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$$

(Saying that these events are disjoint means that they are mutually exclusive: $A_i \cap A_j = \emptyset$ for $i \neq j$.

Probability Space



Interpretation of Probability

- *The frequentist view*: probability represents a long-run frequency over a large number of repetitions of an experiment.
- If we say a coin has probability $1/2$ of Heads, that means the coin would land Heads 50% of the time if we tossed it over and over and over.

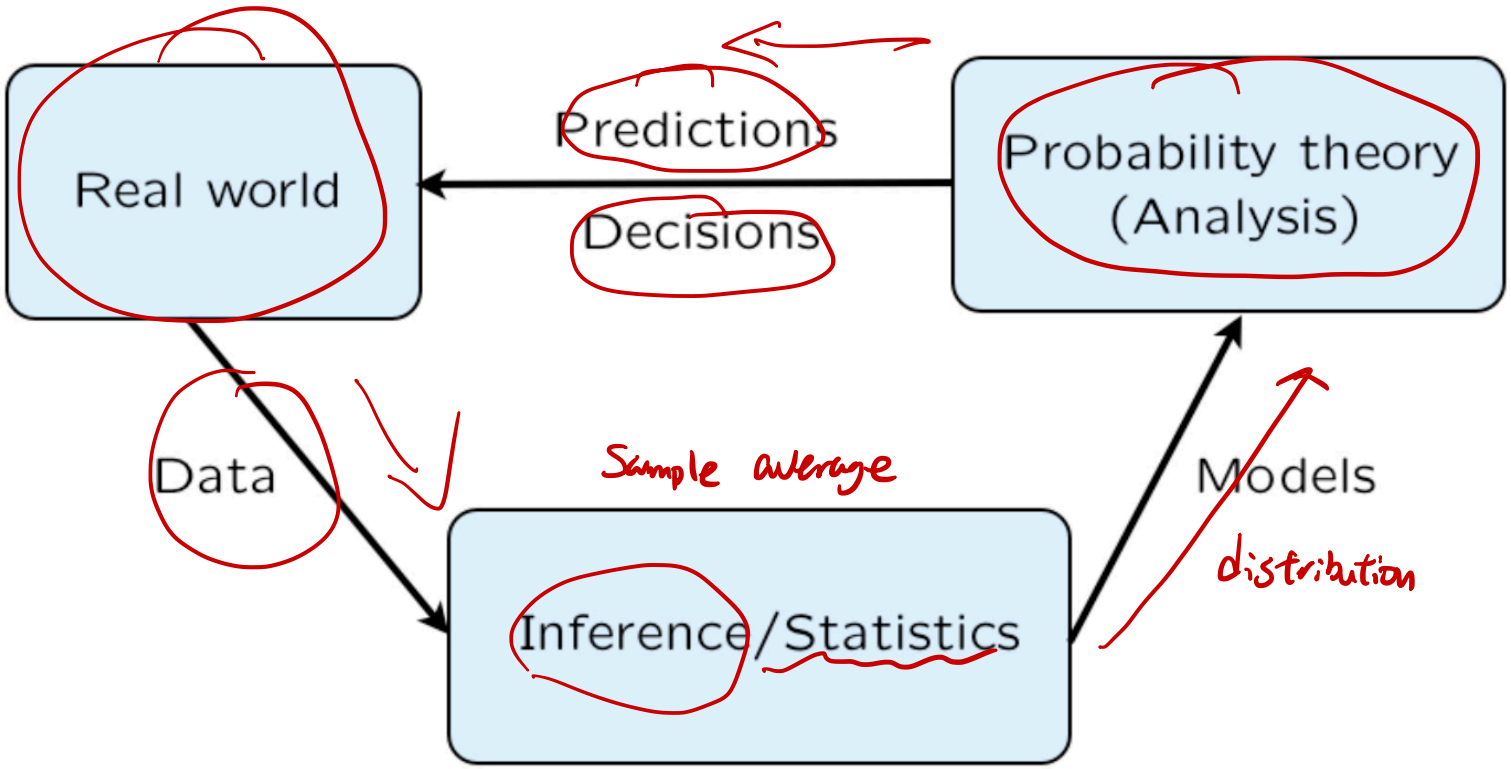
Interpretation of Probability

- *The Bayesian view*: probability represents a degree of belief about the event in question.
- So we can assign probabilities to hypotheses like "candidate A will win the election" or "the defendant is guilty" even if it isn't possible to repeat the same election or the same crime over and over again.

The Role of Probability & Statistics

A framework for analyzing phenomena with uncertain outcomes:

- Rules for consistent reasoning
- Used for predictions and decisions



Outline

Thinking Conditionally

- New data & information may affect our uncertainties
- Conditional probability: how to update our belief?
- All probabilities are conditional! (explicit/implicit background info or assumption)

Thinking Conditionally

- Conditioning :a powerful problem-solving strategy
- Reducing a complicated probability problem to a bunch of simpler conditional probability problems
- First-step analysis: obtain recursive solution to multi-stage problems
- **Conditioning is the soul of statistics!**

Definition of Conditional Probability

Definition

If A and B are events with $P(B) > 0$, then the *conditional probability* of A given B, denoted by $P(A|B)$, is defined as

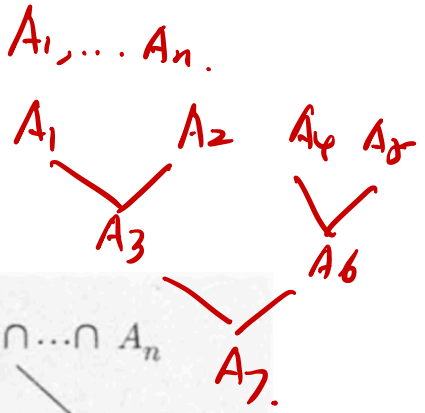
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

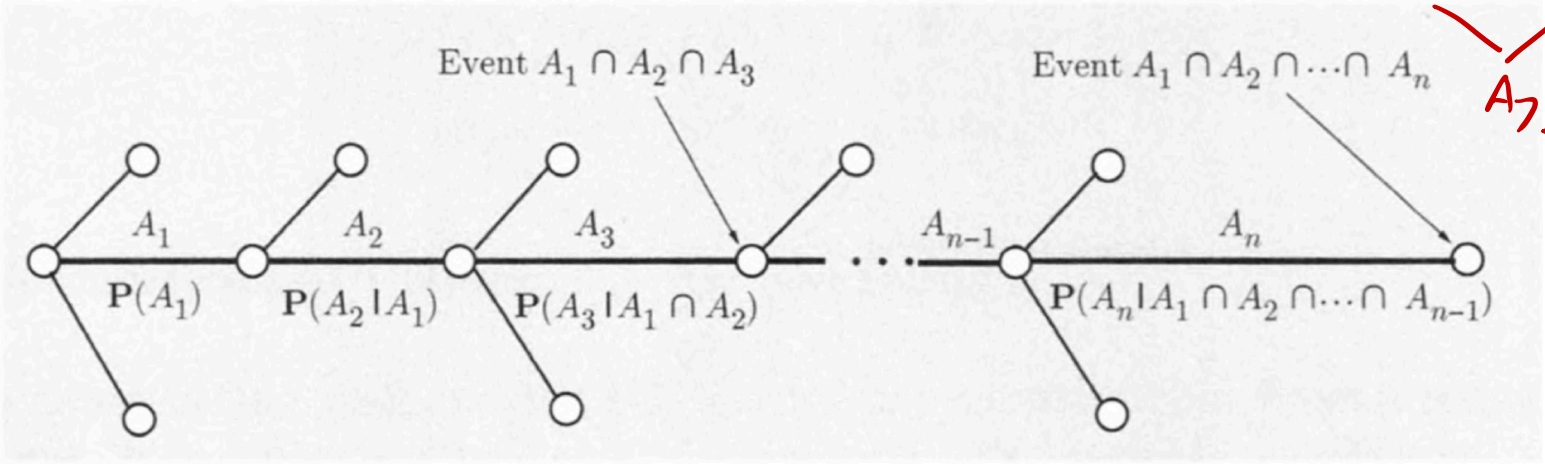
- $P(A)$: *prior probability* of A.
- $P(A|B)$: *posterior probability* of A.

B : new data / information

Chain Rule

Bayes Network. Graph.





Theorem

For any events $A_1, \ldots, A_n$ with positive probabilities,

$$P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \ldots, A_{n-1}).$$

$\log P(A_1, \ldots A_n) = \log P(A_1) + \log P(A_2|A_1) + \cdots + \log P(A_n|A_1, \ldots, A_{n-1})$

Distributed Computing.

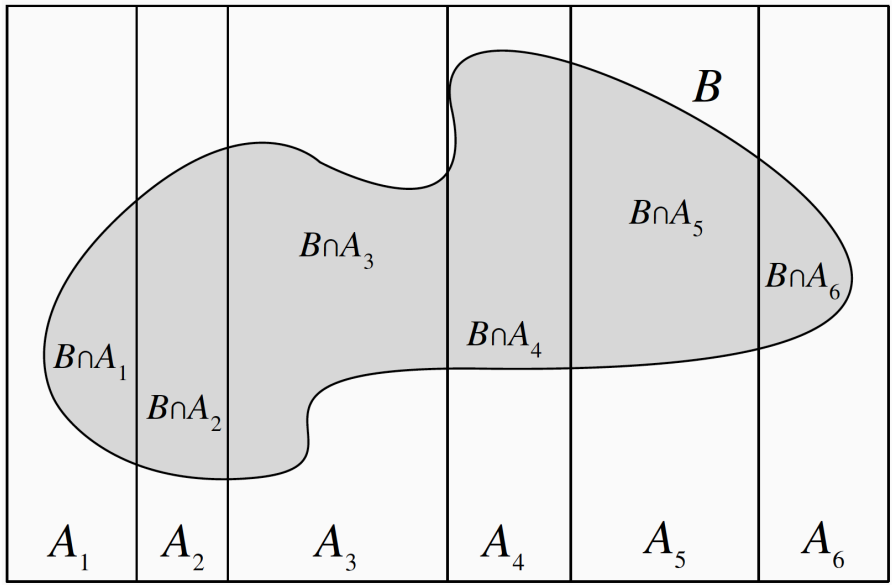
Bayes' Rule

Theorem

For any events A and B with positive probabilities,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

The Law of Total Probability (LOTP)



Theorem

Let $A_1, \ldots, A_n$ be a partition of the sample space S (i.e., the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then

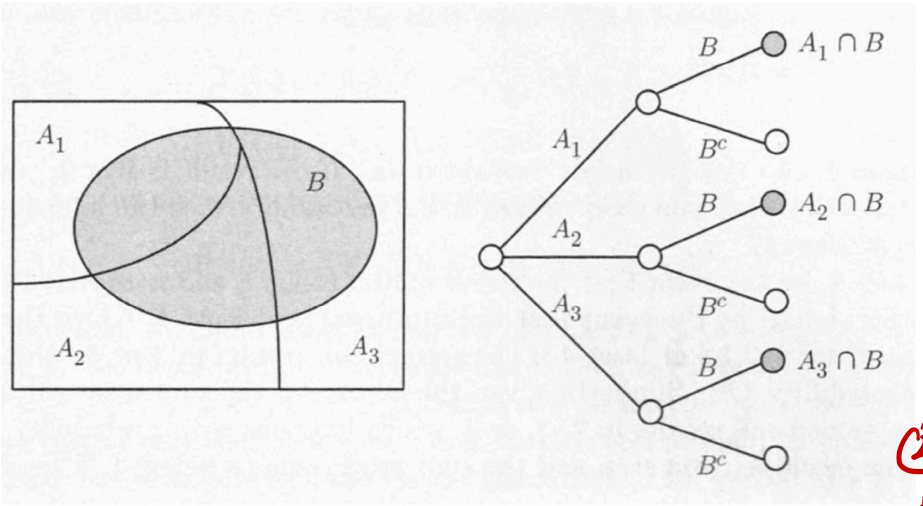
$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Inference & Bayes' Rule

③ false alarm

Neyman-Pearson Theorem.

trade-off.



① $A_1, \ldots A_n$: Alternate Hypothesis.

$P(A_1) \ldots P(A_n)$. prior belief

obtain new data/info. B.

Update our belief on A_i.

$P(A_1|B), \ldots P(A_n|B)$

② Decision rule

$i^* = \arg\max_i P(A_i|B)$

MAP (Maximum A Posterior) Probability.

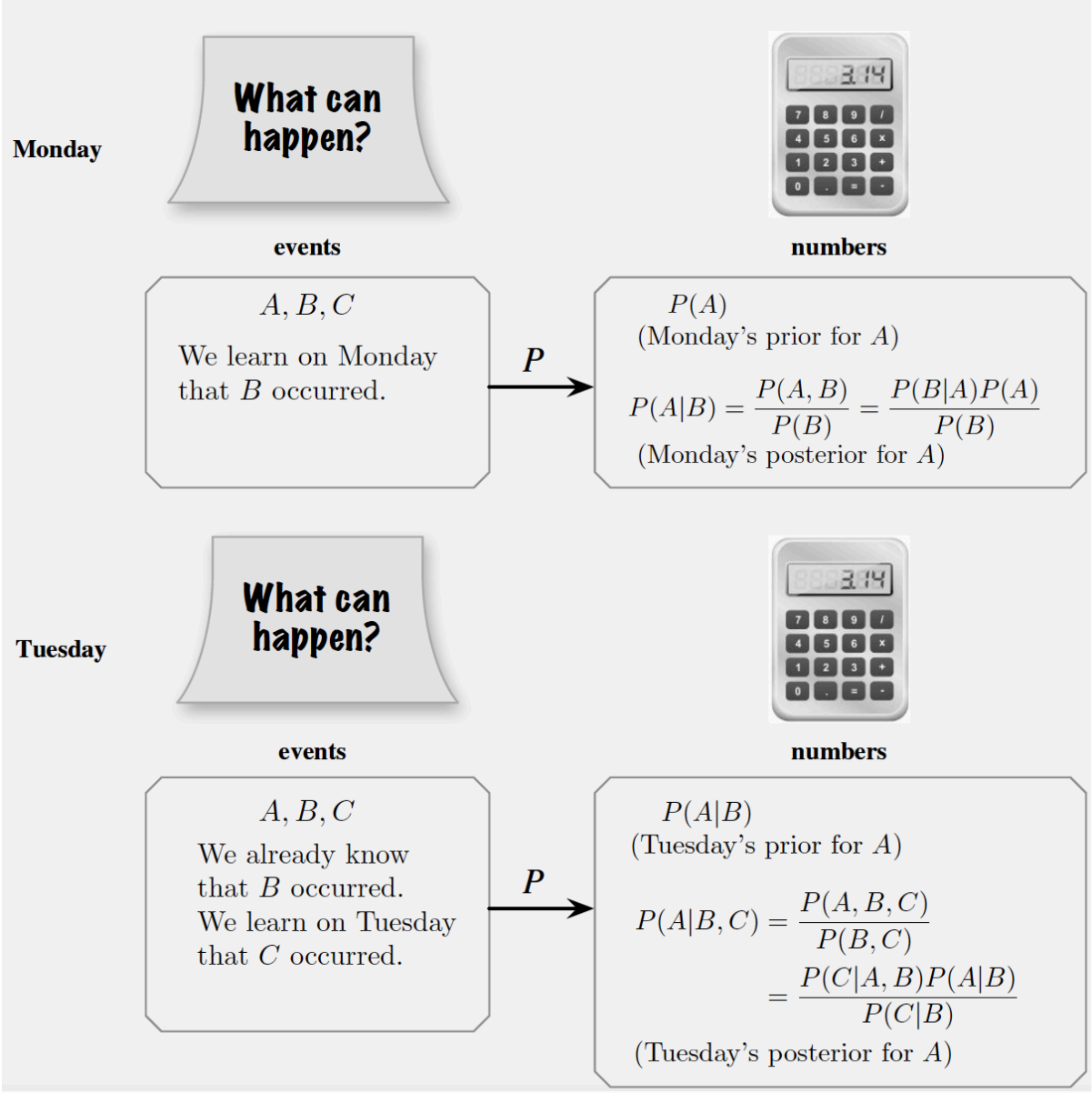
Theorem

Let $A_1, \ldots, A_n$ be a partition of the sample space S (i.e., the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then for any event B such that $P(B) > 0$, we have

Rule.

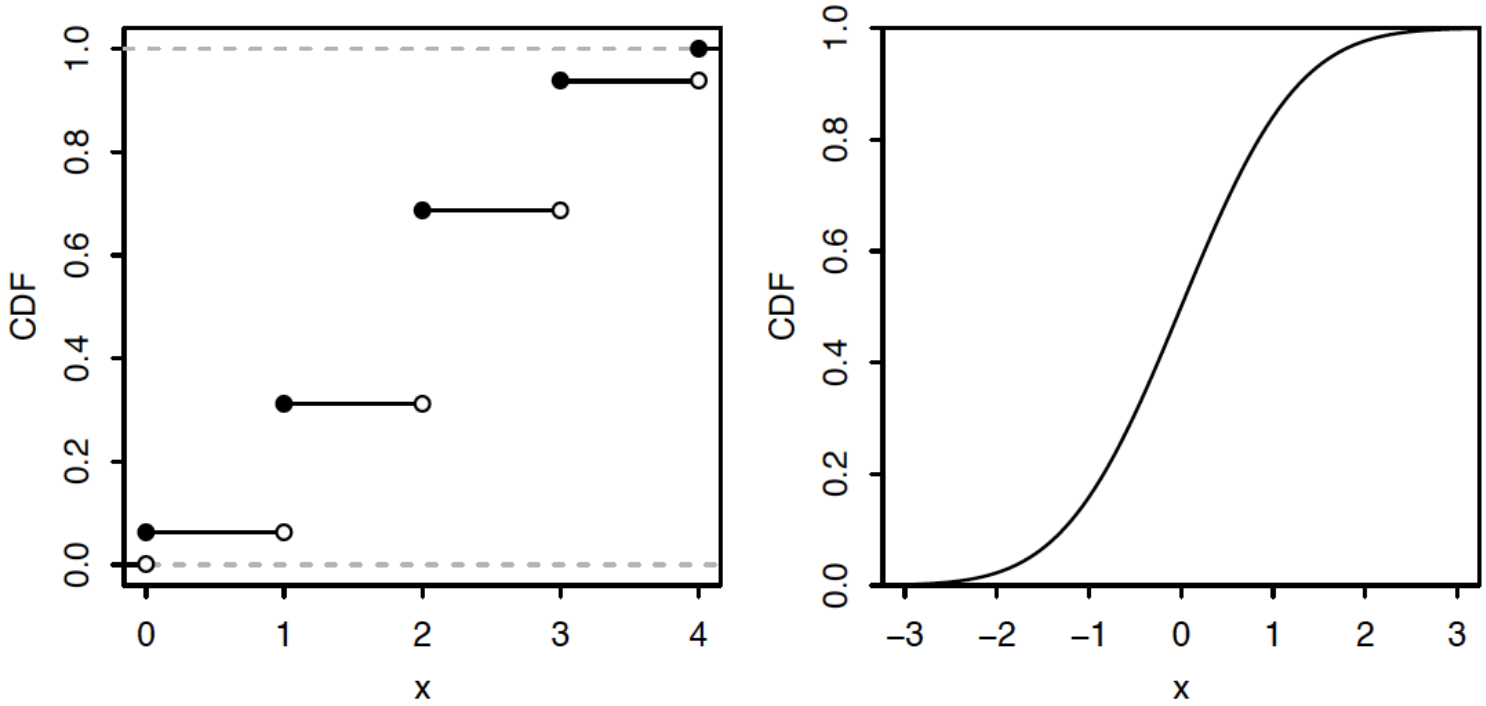
$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}.$$

Example: Conditional Probability & Inference



Outline

Discrete vs. Continuous



Continuous Random Variables

Definition

An r.v. has a *continuous distribution* if its CDF is differentiable. We also allow there to be endpoints (or finitely many points) where the CDF is continuous but not differentiable, as long as the CDF is differentiable everywhere else. A *continuous random variable* is a random variable with a continuous distribution.

Probability Density Function

Definition

For a continuous r.v. X with CDF F , the *probability density function* (PDF) of X is the derivative f of the CDF, given by $f(x) = F'(x)$. The *support* of X, and of its distribution, is the set of all x where $f(x) > 0$.

Illustration of PDF

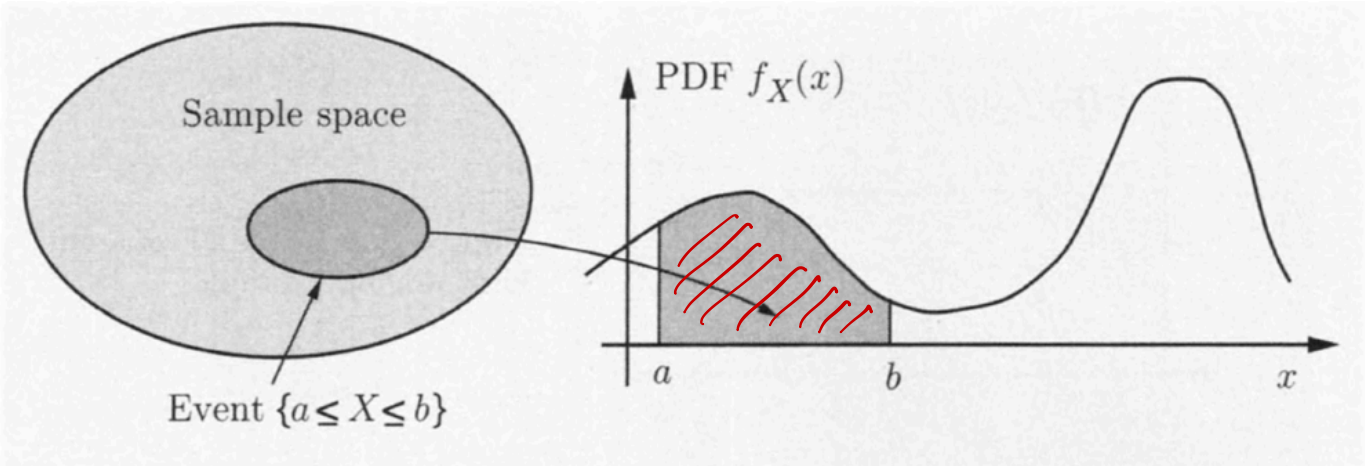
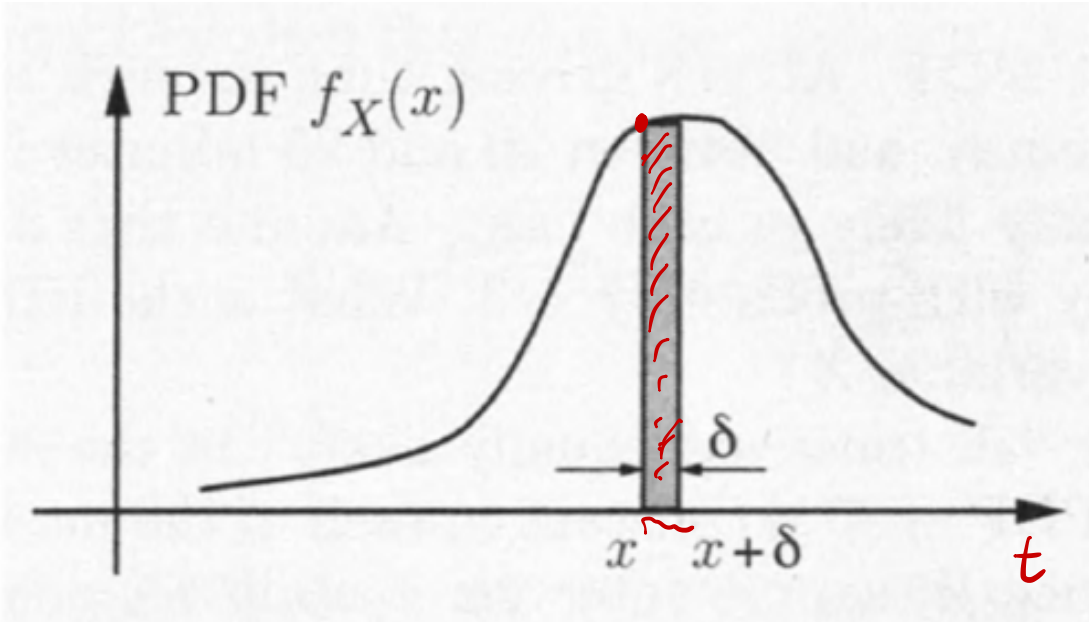


Illustration of PDF

$$P(x \le X \le x+\delta) = \int_x^{x+\delta} f_X(t)\,dt \approx f_X(x)\cdot\delta$$



PDF vs. PMF

① PDF f : $f(x) \neq$ probability $\left(\int_A f(x)dx\right)$

$f(x) > 1$ could be

PMF f : Prob: $\in [0,1]$

② For a continuous r.v. X $P(X=a)=0$, $\forall a \in R$

discrete r.v. X. $P(X=a)>0$ if $a \in$ support set of X.

PDF to CDF

Theorem

Let X be a continuous r.v. with PDF f. Then the CDF of X is given by

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Including or Excluding Endpoints

① For Continuous r.v. X

$$P(a<X<b) = P(a<X\le b) = P(a\le X<b) = P(a\le X\le b)$$

(a,b) $(a,b]$ $[a,b)$ $[a,b]$

Since $P(X=a) = P(X=b) = 0$

② if X, Y Continuous r.v., X, Y are independent.

$P(X=Y) = 0$.

?

$\forall c \in R$, $P(X=c, Y=c)$

$= P(X=c)P(Y=c)$

$= 0$

Valid PDFs

$$\frac{f(x)}{\int_{-\infty}^{\infty} f(x)\,dx}.$$ renormalization.

Theorem

The PDF f of a continuous r.v. must satisfy the following two criteria:

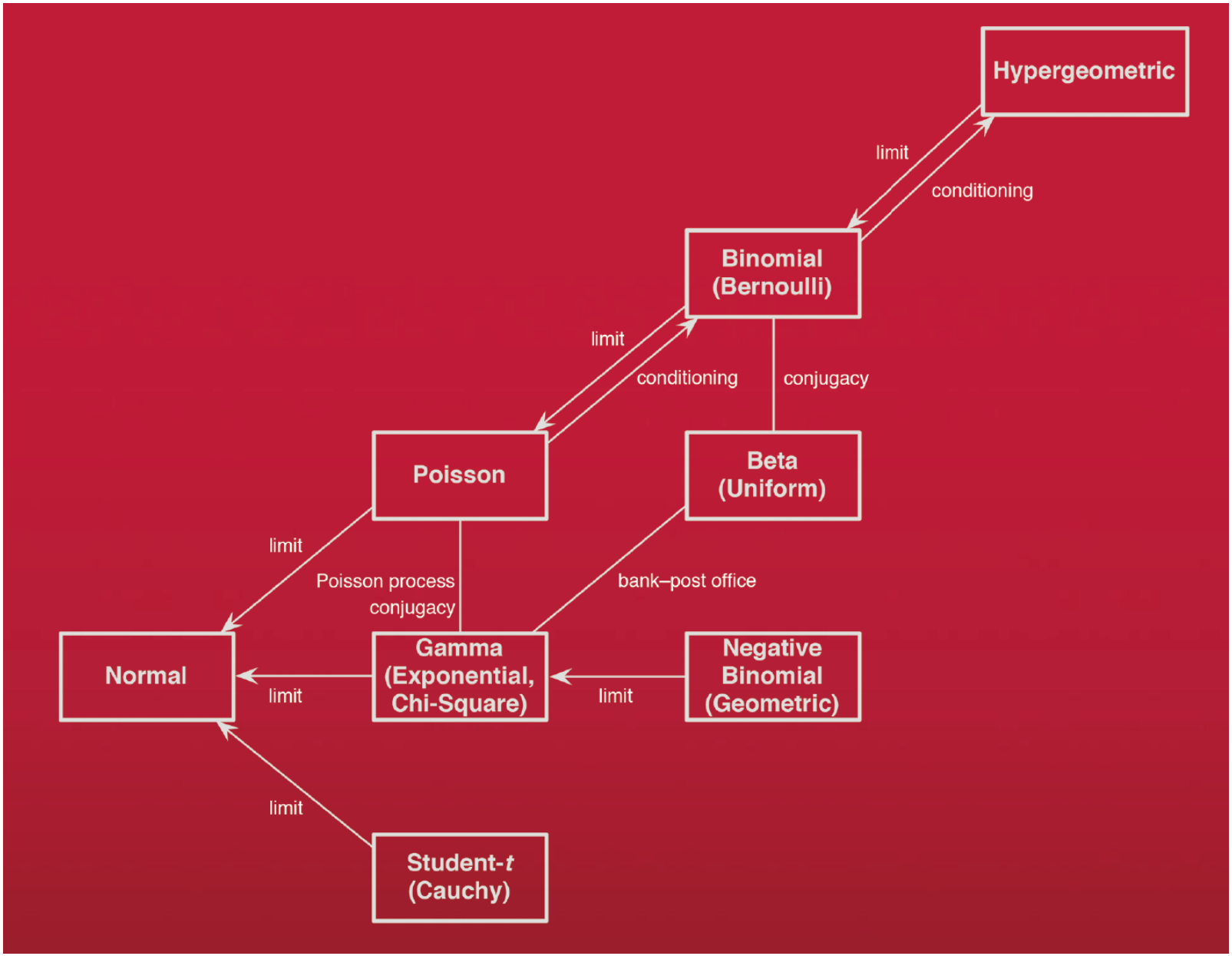
- *Nonnegative:* $f(x) \geq 0$;
- *Integrates to 1:* $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Typical Distributions

Name	Param.	PMF or PDF	Mean	Variance
Bernoulli	p	$P(X=1)=p, P(X=0)=q$	p	pq
Binomial	n,p	$\binom{n}{k}p^k q^{n-k}$, for $k \in \{0,1,\dots,n\}$	np	npq
FS	p	pq^{k-1}, for $k \in \{1,2,\dots\}$	$1/p$	q/p^2
Geom	p	pq^k, for $k \in \{0,1,2,\dots\}$	q/p	q/p^2
NBinom	r,p	$\binom{r+n-1}{r-1}p^r q^n, n \in \{0,1,2,\dots\}$	rq/p	rq/p^2
HGeom	w,b,n	$\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$, for $k \in \{0,1,\dots,n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1})n\frac{\mu}{n}(1-\frac{\mu}{n})$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!}$, for $k \in \{0,1,2,\dots\}$	λ	λ
Uniform	$a<b$	$\frac{1}{b-a}$, for $x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ,σ^2	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2
Log-Normal	μ,σ^2	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}, x>0$	$\theta = e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$
Expo	λ	$\lambda e^{-\lambda x}$, for $x>0$	$1/\lambda$	$1/\lambda^2$
Gamma	a,λ	$\Gamma(a)^{-1}(\lambda x)^a e^{-\lambda x}x^{-1}$, for $x>0$	a/λ	a/λ^2
Beta	a,b	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$, for $0<x<1$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{a+b+1}$
Chi-Square	n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$, for $x>0$	n	$2n$
Student-t	n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$

Relationship Among Distributions

Passive knowledge.
Active knowledge.



Total Variation Distance

Kullback-Leibler (KL) divergence
Jensen-Shannon (JS) divergence
Earth-mover (EM) distance or Wasserstein-1

ML, DL, AI.
GAN/WGAN...

Target distribution: π
Approximation distribution:
Behavior. (μ)
Off-policy learning.

- Distance measure between two probability distributions

Definition

The **total variation distance** between two distributions μ and ν on a countable set Ω is

$$\begin{aligned} d_{TV}(\mu,\nu) &= \| \mu - \nu \|_{TV} \\ &= \max_{A \subset \Omega} |\mu(A) - \nu(A)| \\ &= \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|. \end{aligned}$$

≤ 1

The Law of Small Numbers

Law of Rare Events.

Theorem

Given independent random variables $Y_1, \cdots, Y_n$ such that for any $1 \leq m \leq n$, $\mathbb{P}(Y_m = 1) = p_m$ and $\mathbb{P}(Y_m = 0) = 1 - p_m$. Let $S_n = Y_1 + \cdots + Y_n$. Suppose

$$\sum_{m=1}^{n} p_m \to \lambda \in (0, \infty) \quad \text{as } n \to \infty,$$

and

$$\max_{1 \leq m \leq n} p_m \to 0 \quad \text{as } n \to \infty,$$

then

$$d_{TV}(S_n, Poi(\lambda)) \to 0 \quad \text{as } n \to \infty.$$

Sample Mean

Monte Carlo method.

$\mu \approx \bar{X}_n$

unknown

Definition

Let $X_1, ..., X_n$ be i.i.d. random variables with finite mean μ and finite variance σ^2. The *sample mean* $\bar{X}_n$ is defined as follows:

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

The sample mean $\bar{X}_n$ is itself an r.v. with mean μ and variance σ^2/n.

Strong Law of Large Numbers SLLN

Theorem

The sample mean $\bar{X}_n$ converges to the true mean μ pointwise as $n \to \infty$, with probability 1. In other words, the event $\bar{X}_n \to \mu$ has probability 1.

Weak Law of Large Numbers WLLN.

Theorem

For all $\epsilon > 0$, $P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \to 0$ as $n \to \infty$. (This form of convergence is called convergence in probability).

Central Limit Theorem

Theorem

As $n \to \infty$,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to \mathcal{N}(0,1) \text{ in distribution.}$$

In words, the CDF of the left-hand side approaches the CDF of the standard Normal distribution.

Outline

Generating Functions

- Three kinds of generating functions
  - Probability Generating Functions (PGF): related to Z-transform
  - Moment Generating Function (MGF): related to Laplace transform
  - Characteristic Functions (CF): related to Fourier transform

Probability Generating Function

Definition

The *probability generating function* (PGF) of a nonnegative integer-valued r.v. X with PMF $p_k = P(X = k)$ is the generating function of the PMF. By LOTUS, this is

$$E\left(t^X\right) = \sum_{k=0}^{\infty} p_k t^k.$$

The PGF converges to a value in $[-1, 1]$ for all t in $[-1, 1]$ since $\sum_{k=0}^{\infty} p_k = 1$ and $\left|p_k t^k\right| \leq p_k$ for $|t| \leq 1$.

Moment Generating Function

Definition

The *moment generating function* (MGF) of an r.v. X is $M(t) = E(e^{tX})$, as a function of t, if this is finite on some open interval $(-a, a)$ containing 0. Otherwise we say the MGF of X does not exist.

Characteristic Function

Definition

The characteristic function of a random variable X is the function $\phi : \mathbb{R} \to \mathbb{C}$ defined by

$$\phi(t) = E(e^{itX}), i = \sqrt{-1}.$$

Differences

- **Probability generating functions(PGF)**: handling non-negative integral random variables
- **Moment generating functions(MGF)**: handling general random variables
- Some integrals of MGF may not be finite
- **Characteristic Function**: equally useful with MGF and guarantee finiteness

Applications of Generating Functions

- An easy way of calculating the moments of a distribution
- Powerful tools for addressing certain counting and combinatorial problems
- An easy way of characterizing the distribution of the sum of independent random variables
- Tools for dealing with the distribution of the sum of a random number of independent random variables.

Applications of Generating Functions

- Play a central role in the study of branching processes
- Provide a bridge between complex analysis and probability
- Play a key role in large deviations theory, that is, in studying the asymptotic of tail probabilities of the form $P(X \geq c)$, when c is a large number

  Outlier Analysis.
- Powerful tools for proving limit theorems, such as laws of large numbers and the central limit theorem

Outline

Multivariate Distribution

- Joint distribution provides complete information about how multiple r.v.s interact in high-dimensional space
- Marginal distribution is the individual distribution of each r.v.
- Conditional distribution is the updated distribution for some r.v.s after observing other r.v.s

Joint CDF

Definition

The *joint CDF* of r.v.s X and Y is the function $F_{X,Y}$ given by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y).$$

The joint CDF of n r.v.s is defined analogously.

Joint PMF

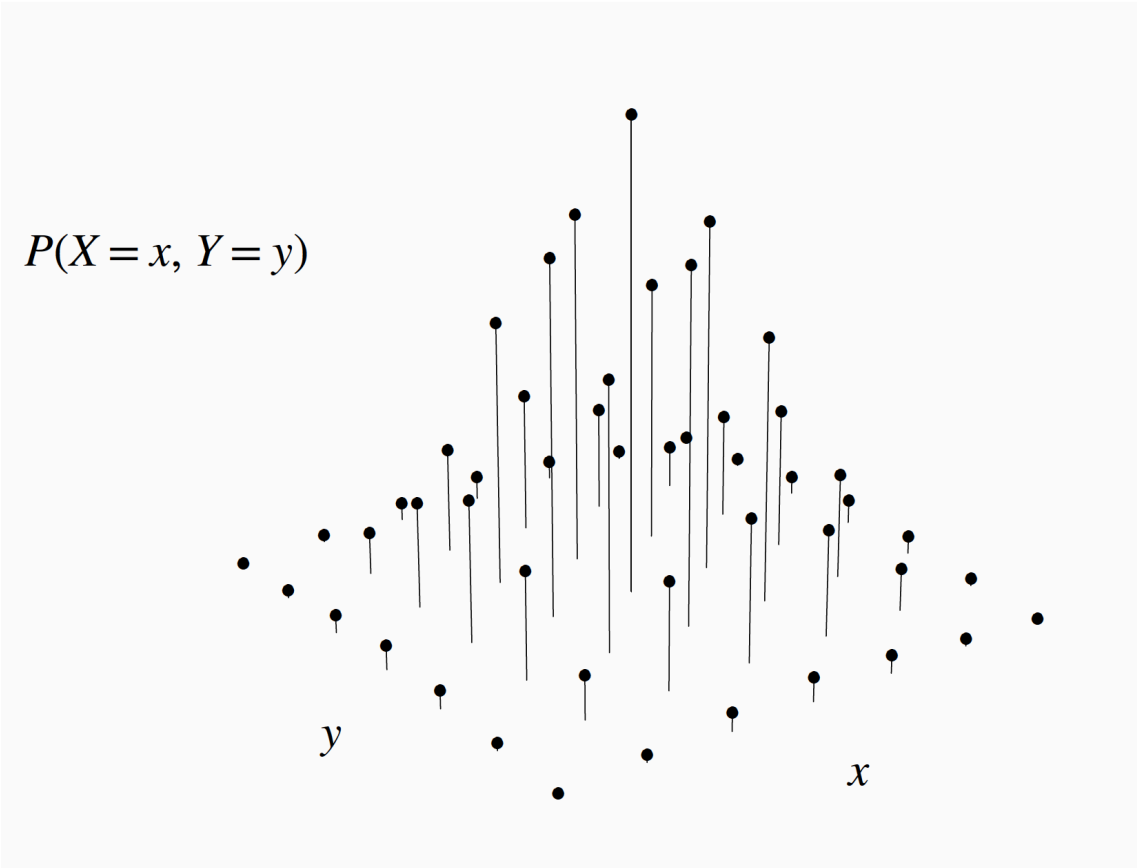
Definition

The joint PMF of discrete r.v.s X and Y is the function $p_{X,Y}$ given by

$$p_{X,Y}(x,y) = P(X = x, Y = y).$$

The joint PMF of n discrete r.v.s is defined analogously.

Joint PMF



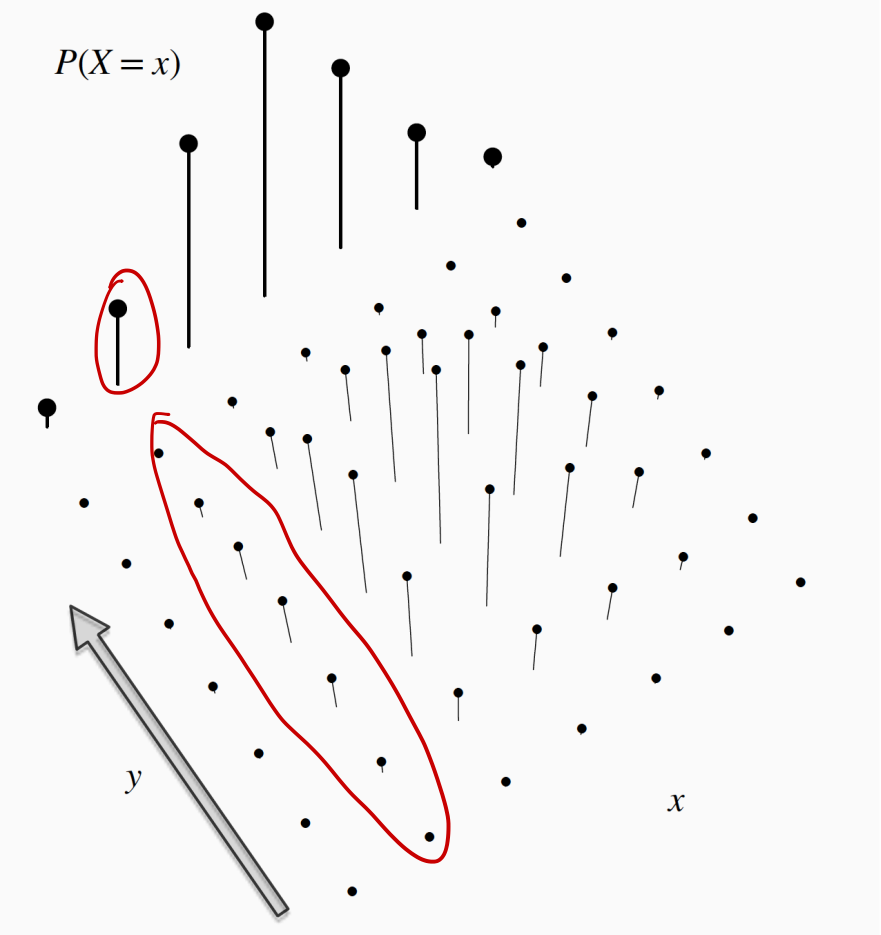
Marginal PMF

Definition

For discrete r.v.s X and Y, the *marginal PMF* of X is

$$P(X = x) = \sum_y P(X = x, Y = y).$$

Marginal PMF



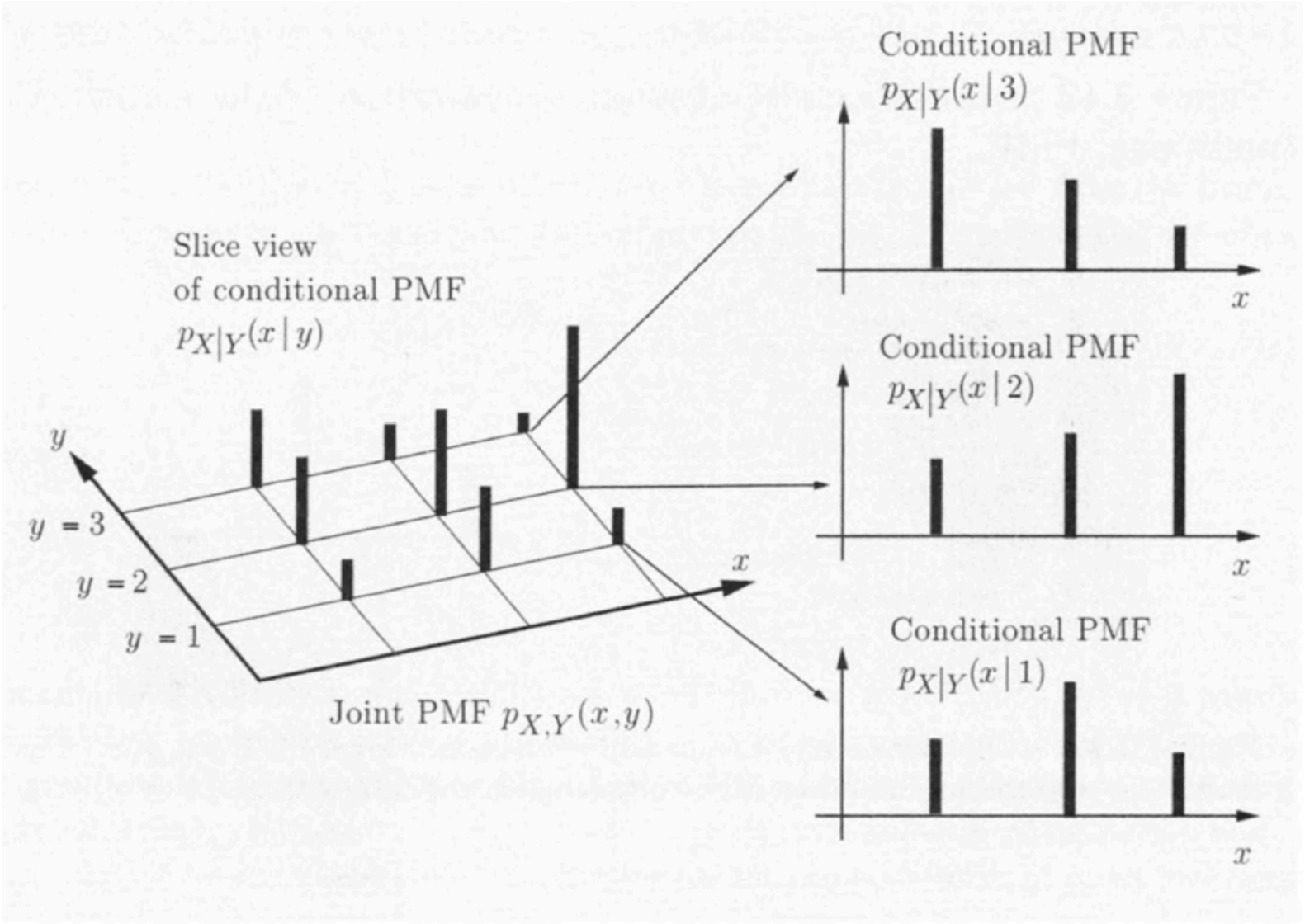
Conditional PMF

Definition

For discrete r.v.s X and Y , the *conditional PMF* of X given $Y = y$ is

$$P_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

Conditional PMF



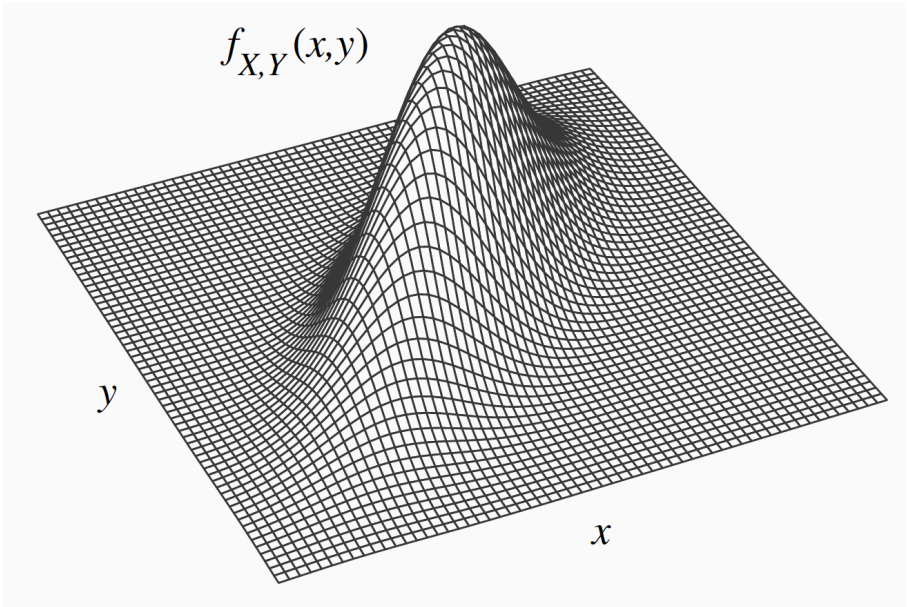
Joint PDF

Definition

If X and Y are continuous with joint CDF $F_{X,Y}$, their joint PDF is the derivative of the *joint CDF* with respect to x and y:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Joint PDF



Marginal PDF

Definition

For continuous r.v.s X and Y with joint PDF $f_{X,Y}$, the *marginal PDF* of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

This is the PDF of X, viewing X individually rather than jointly with Y.

Conditional PDF

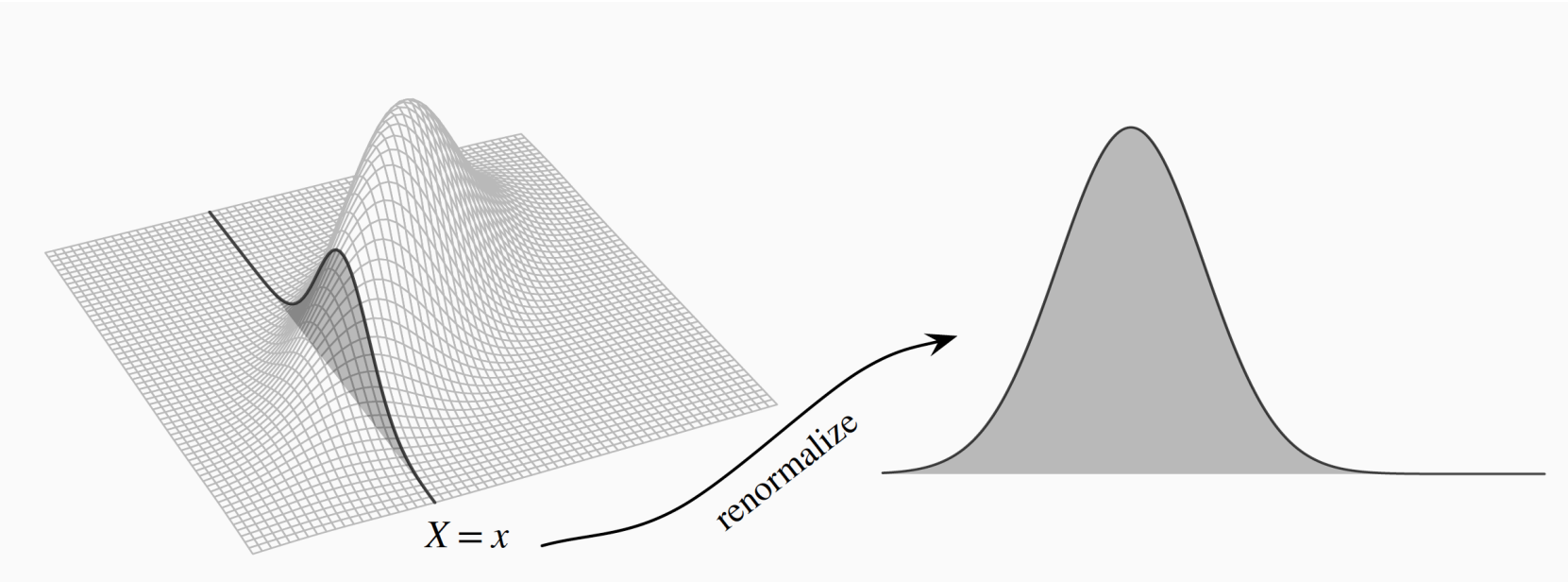
a valid PDF.

Definition

For continuous r.v.s X and Y with joint PDF $f_{X,Y}$, the *conditional PDF* of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Conditional PDF



Technique Issue

- What is the meaning of conditioning on zero-probability event $X = x$ for a continuous r.v. X.
- We are actually conditioning on the event that X falls within a small interval of x: $X \in (x - \epsilon, x + \epsilon)$ and then taking a limit as $\epsilon \to 0$.

Continuous form of Bayes' rule and LOTP

Theorem

For continuous r.v.s X and Y,

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) \, dy$$

General Bayes' Rule

① : $P(Y=y|X=x)$

$$P(Y=y|X\in(x-\varepsilon,x+\varepsilon)) = \frac{P(Y=y)\,P(X\in(x-\varepsilon,x+\varepsilon)|Y=y)}{P(X\in(x-\varepsilon,x+\varepsilon))}$$

$$\approx \frac{P(Y=y)\cdot f_X(x|Y=y)\cdot 2\varepsilon}{f_X(x)\cdot 2\varepsilon}$$

	Y discrete	Y continuous
X discrete	$P(Y=y\|X=x)=\frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	② $f_Y(y\|X=x)=\frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X continuous	① $P(Y=y\|X=x)=\frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x)=\frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

$\varepsilon \to 0$

$$\Rightarrow \text{LHS. } P(Y=y|X=x) = \frac{P(Y=y)\cdot f_X(x|Y=y)}{f_X(x)}$$

General LOTP

	Y **discrete**	Y **continuous**
X **discrete**	$P(X=x)=\sum_y P(X=x\|Y=y)P(Y=y)$	$P(X=x)=\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$f_X(x)=\sum_y f_X(x\|Y=y)P(Y=y)$	$f_X(x)=\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

$$\frac{P(X\in(x-\varepsilon,x+\varepsilon))}{2\varepsilon} \qquad \frac{P(X\in(x-\varepsilon,x+\varepsilon)\,|\,Y=y)}{2\varepsilon}$$

$\varepsilon \to 0$.

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.

Change of Variables

Sometimes. $Y=g(X)$ is given, $X=g^{-1}(Y)$ hard to compute.

since $\left|\frac{\partial X}{\partial y}\right| = \left|\frac{\partial y}{\partial x}\right|^{-1}$

$\Rightarrow f_{Y|y} = f_X(x) \cdot \frac{1}{\left|\frac{\partial y}{\partial x}\right|}$.

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**

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

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

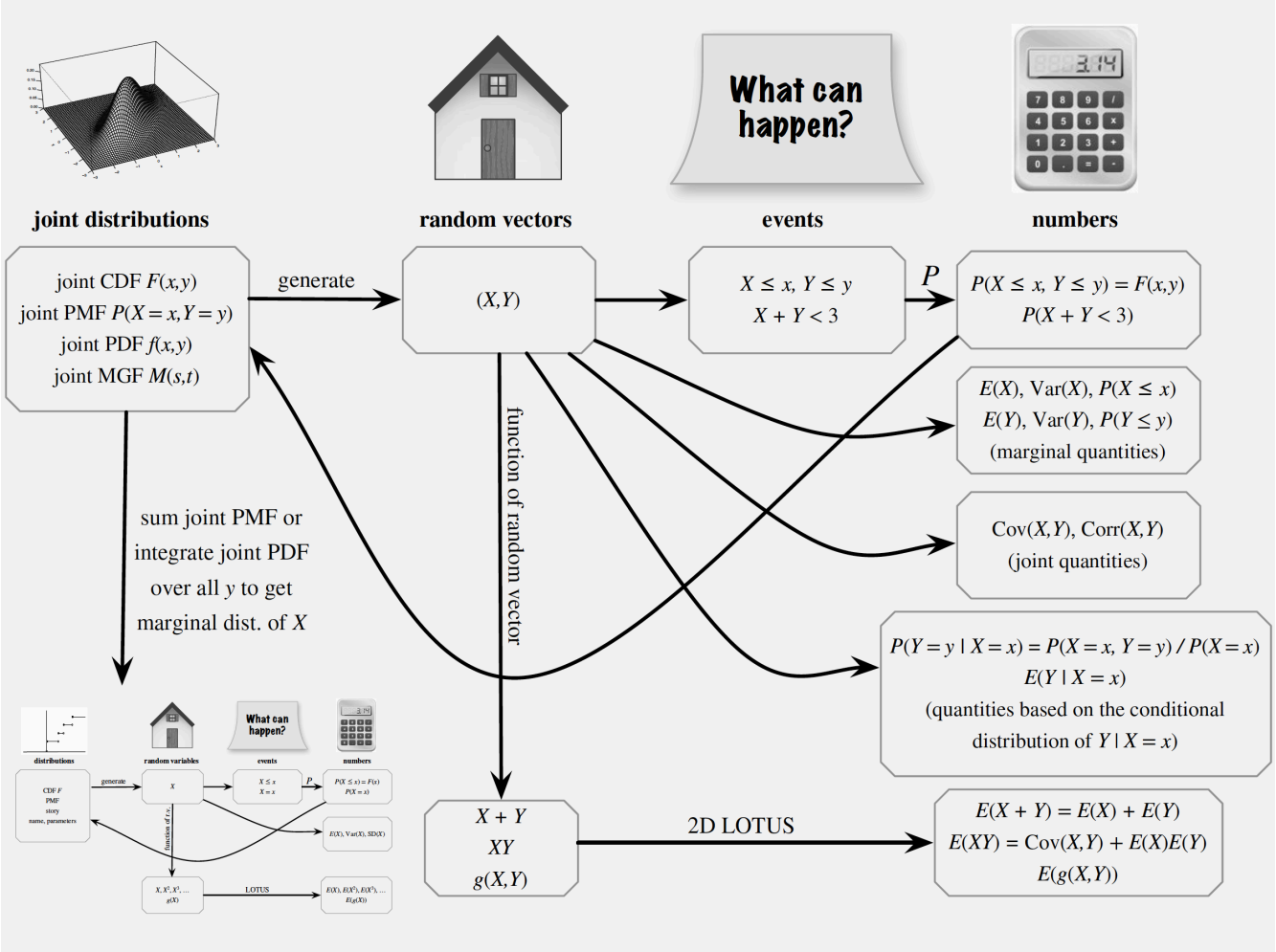
$|\cdot|$ absolute value of the determinant of the matrix.

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

Summary: Discrete & Continuous

	Two discrete r.v.s	Two continuous r.v.s
Joint CDF	$F_{X,Y}(x,y) = P(X \le x, Y \le y)$	$F_{X,Y}(x,y) = P(X \le x, Y \le y)$
Joint PMF/PDF	$P(X = x, Y = y)$ • Joint PMF is nonnegative and sums to 1: $\sum_x \sum_y P(X = x, Y = y) = 1$.	$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$ • Joint PDF is nonnegative and integrates to 1: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$. • To get probability, integrate joint PDF over region of interest.
Marginal PMF/PDF	$P(X = x) = \sum_y P(X = x, Y = y)$ $= \sum_y P(X = x \mid Y = y)P(Y = y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ $= \int_{-\infty}^{\infty} f_{X\mid Y}(x\mid y) f_Y(y)dy$
Conditional PMF/PDF	$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$ $= \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$	$f_{Y\mid X}(y\mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ $= \frac{f_{X\mid Y}(x\mid y) f_Y(y)}{f_X(x)}$
Independence	$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$ $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x and y.	$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$ $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all x and y.
	$P(Y = y \mid X = x) = P(Y = y)$ for all x and y, $P(X = x) > 0$.	$f_{Y\mid X}(y\mid x) = f_Y(y)$ for all x and y, $f_X(x) > 0$.
LOTUS	$E(g(X,Y)) = \sum_x \sum_y g(x,y)P(X = x, Y = y)$	$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dxdy$

Summary: Multivariate Distribution



Outline

Order Statistics

$X_1, \ldots X_n$.

order statistics $X_{(1)} = \min(X_1, \ldots, X_n)$

$X_{(2)}$: the second smallest of $X_1, \ldots, X_n$

$\vdots$

$X_{(n)} = \max(X_1, \ldots X_n)$

$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$

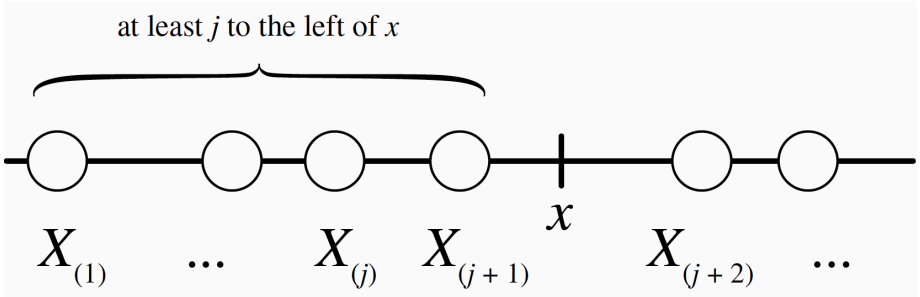
CDF of Order Statistics

Theorem

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

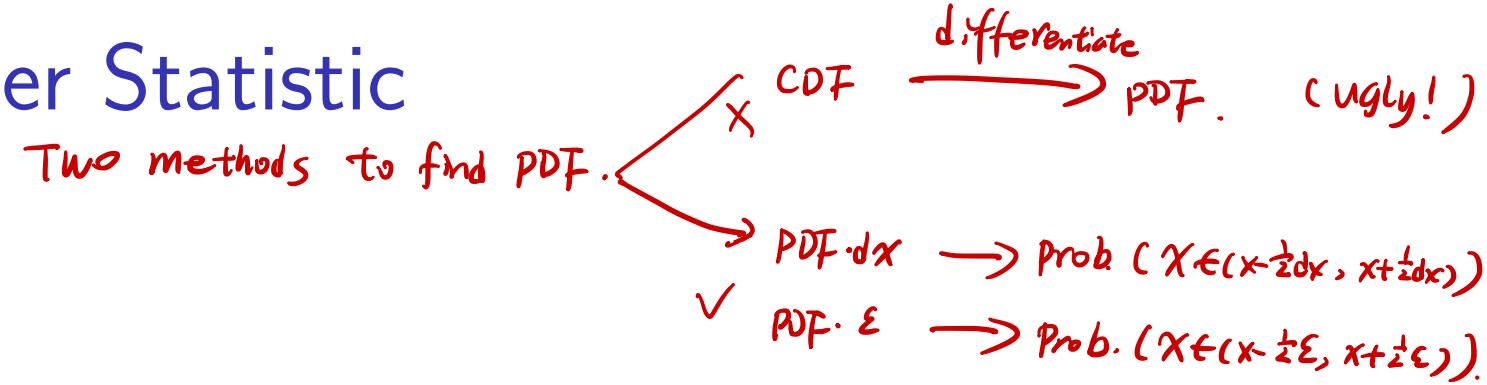
$$P\left(X_{(j)} \leq x\right) = \sum_{k=j}^{n} \binom{n}{k} F(x)^k \left(1 - F(x)\right)^{n-k}.$$

Proof



Model.!

PDF of Order Statistic

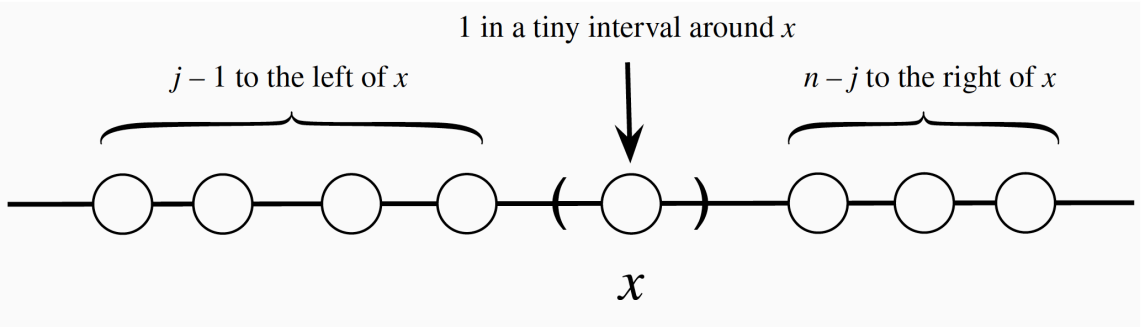


Theorem

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

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

Proof


1 in a tiny interval around x
$j-1$ to the left of x
$n-j$ to the right of x
x

Model.!

Joint PDF

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. continuous r.v.s with PDF f. Then the joint PDF of all order statistics is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i), x_1 < x_2 < \cdots < x_n$$

Example: Order Statistics of Uniforms

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

$U_{(j)}$: order statistics

PDF $\Rightarrow$ $f_{U_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j}$

CDF $\Rightarrow$ $F_{U_{(j)}}(x) = \sum_{m=j}^{n} \binom{n}{m} x^m (1-x)^{n-m} = \int_0^x f_{U_{(j)}}(t)\,dt$

$= \frac{n!}{(j-1)!(n-j)!} \int_0^x t^{j-1}(1-t)^{n-j}\,dt$

Example: Order Statistics of Uniforms

$$1 - F_{U_{(j)}}(x) = 1 - \int_0^x f_{U_{(j)}}(t)\,dt$$

$$\Rightarrow \quad \sum_{m=0}^{j-1} \binom{n}{m} x^m (1-x)^{n-m} = \int_x^1 f_{U_{(j)}}(t)\,dt$$

$$= \frac{n!}{(j-1)!(n-j)!} \int_x^1 t^{j-1}(1-t)^{n-j}\,dt$$

Let $x = p$, $k = j-1$, $\Rightarrow$ the following equation

Related Identity

$U_1, \dots U_n$ i.i.d. unif(0,1)

P

Perform n independent Bernoulli trials.

$\{U_i < p\} \triangleq$ success event (left of p)

Prob $(U_i < p) = p$

Theorem

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

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

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

$LHS = P(N \le k)$ = P(at most k of $U_1, \dots U_n$ are left of p)

$= P(U_{(k+1)} > p) = \int_p^1 f_{U_{(k+1)}}(x)dx$

$= RHS$

Proof

① $P(N>k) = P(N \geq k+1) = P(U_{(k+1)} \leq p)$

② $U_{(j)} \sim \text{Beta}(j, n-j+1)$

Outline

Beta Distribution

1°. Continuous distribution on (0,1)

2°. $a=b=1$, $f(x)=1$, $0<x<1$, $\text{Unif}(0,1)=\text{Beta}(1,1)$

3°. If the unknown parameter $\in(0,1)$, its prior distribution: Beta.

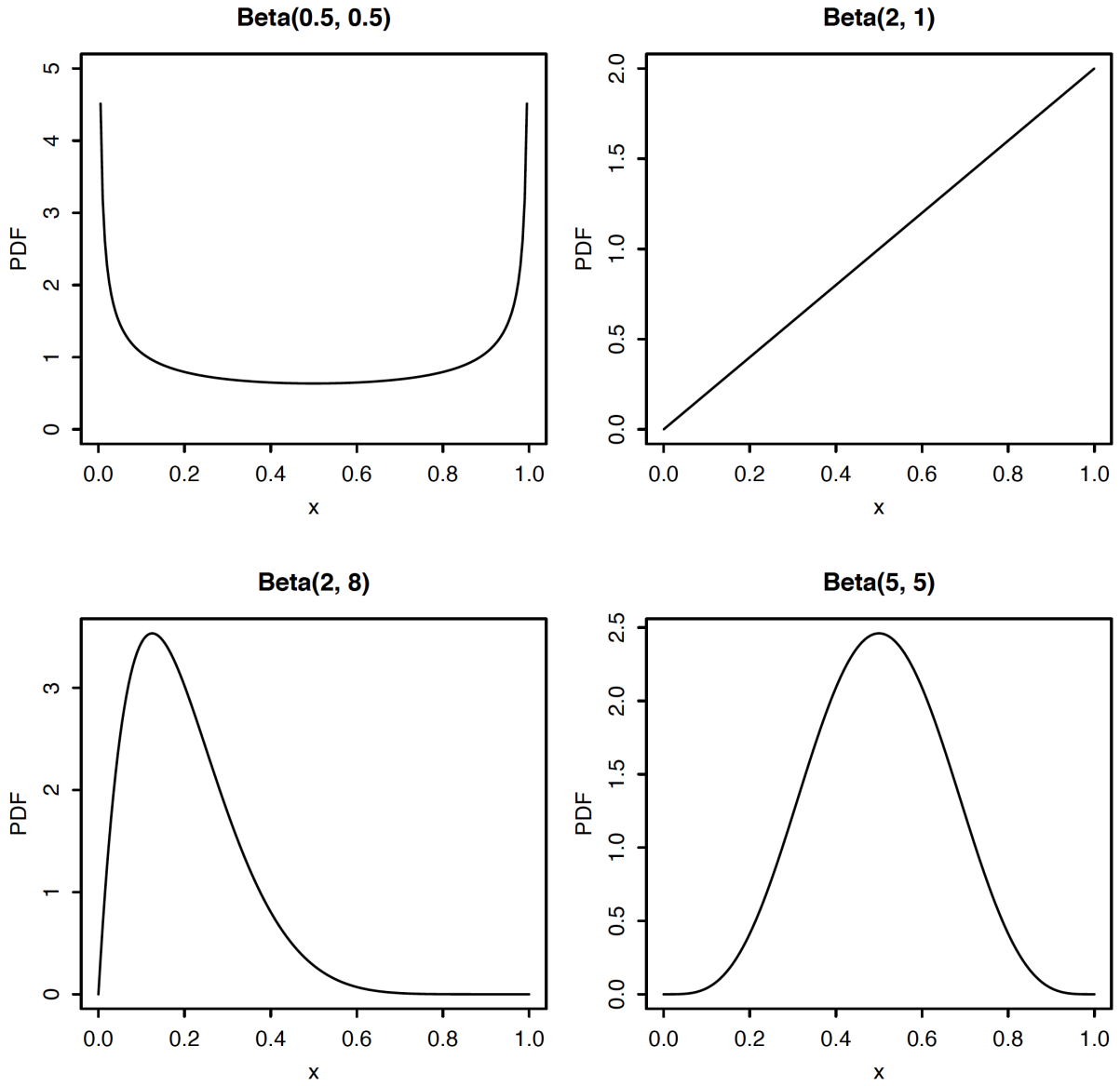
Definition

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

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

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

PDF of Beta Distribution



Beta Integral

When a and b are positive integers

1°. $B(a,b) = \dfrac{(a-1)!\,(b-1)!}{(a+b-1)!}$

2°. if $X \sim \text{Beta}(a,b)$, $E(X) = \dfrac{a}{a+b}$

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

Story: Bayes' billiards

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

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

Model.

Story: Beta-Binomial Conjugacy

1^0. n tosses; K of n tosses are landing heads.

2^0. $\hat{p}$ (estimate of p) $= \frac{k}{n}$. MLE.

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

$n=3,\ k=3,\ \hat{p}=1$,

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.

3^0. $\hat{p}$ depend on sample size n. $\hat{p} = f(n)$ $n \to \infty$

$f(n) \to 1$.

Bayesian Inference

prior conjugacy.

- 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.

↓ hard to compute.

Story: Beta-Binomial Conjugacy

① P : r.v. choose a prior distribution $\in [0,1]$

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

X : # of heads in n tosses of the coin.

$X \mid P=p \sim \text{Bin}(n,p)$

② $f(p)$: prior PDF of P.

$f(p \mid X=k)$: posterior PDF of P. (observation: k heads out of n tosses)

③ $f(p \mid X=k) = \dfrac{Pr(X=k \mid P) f(p)}{Pr(X=k)} = \dfrac{\binom{n}{p} p^k (1-p)^{n-k} \cdot \frac{1}{B(a,b)} p^{a-1} (1-p)^{b-1}}{Pr(X=k)}$

$Pr(X=k) = \int_0^1 Pr(X=k \mid P) f(p)\, dp = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} f(p)\, dp.$

Very complex !

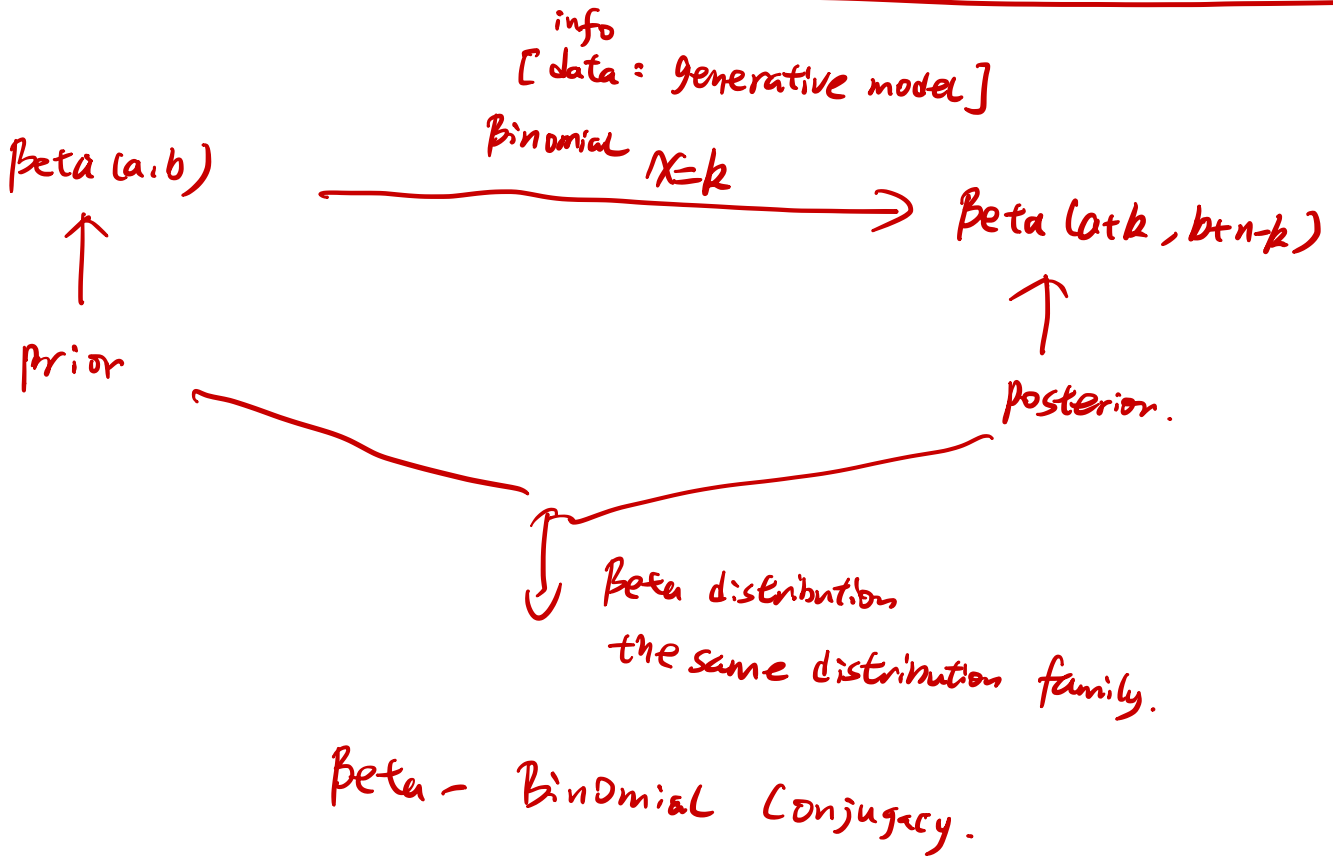
④. $f(p \mid X=k)$ is a function of P. (everything that does not depend on p is a constant)

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

Story: Beta-Binomial Conjugacy

$P \mid X=k \sim \text{Beta}(a+k, b+n-k)$

Posterior distribution of P (after observation of $X=k$) is still a Beta distribution



Beta - Binomial Conjugacy.

Story: Beta-Binomial Conjugacy

Story: Beta-Binomial Conjugacy

$\text{Beta}(a,b) \xrightarrow{X=k} \text{Beta}(a+k, b+n-k) \xrightarrow{Y=m} \text{Beta}(a+k+m, b+n-k+n-m)$

- 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

$p \sim \text{Beta}(a,b)$ $\quad E(p) = \frac{a}{a+b}$

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

$\Rightarrow E(p|X=k) = \frac{a+k}{a+b+n}$

- Infer the value of p (probability of coin lands heads)
- Observed k heads out of n tosses of the coin
- Mean: $\frac{k}{n}$ MLE.
- 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)

$n \to \infty$, 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

Monte Carlo Computing

1°. given a function $\phi : R^n \to R$. . then if $p(\cdot)$ denotes a PDF support over R^n, $\Rightarrow$ the multi-dimensional integral

$$\int_{R^n} \phi(x)dx = \int_{R^n} \frac{\phi(x)}{p(x)} p(x)dx = E_p\left\{\frac{\phi(x)}{p(x)}\right\}$$

E_p : expectation w.r.t. PDF p.

$$\hat{\wedge} \frac{1}{N} \sum_{k=1}^{N} \frac{\phi(x_k)}{p(x_k)}$$

2°. Simulating i.i.d samples $\{x_k\}$, $k=1,2,..N$. from the PDF $p(\cdot)$.

Classical M.C. Sample average. as an approximation.

$$\frac{1}{N} \sum_{k=1}^{N} \frac{\phi(x_k)}{p(x_k)} \quad , \; x_k \sim p(\cdot)$$

3°. MCMC . x_k . Can be generated. with a Markov chain.

Stationary distribution $p(\cdot)$.

Simulation of Random Varibles

- Assuming an algorithm is available for generating Unif(0, 1) random numbers
- Two elementary methods for simulating random variables
  - Inverse-transform method
  - The acceptance rejection method

Inverse Transform Method

- Given a Unif(0, 1) r.v., we can construct an r.v. with any continuous distribution we want.
- Conversely, given an r.v. with an arbitrary continuous distribution, we can create a Unif(0, 1) r.v.
- Other names:
  - probability integral transform
  - inverse transform sampling
  - the quantile transformation
  - the fundamental theorem of simulation

Inverse Transform Method

Theorem

Let F be a CDF which is a continuous function and strictly increasing on the support of the distribution. This ensures that the inverse function F^{-1} exists, as a function from $(0,1)$ to $\mathbb{R}$. We then have the following results.

1. *Let $U \sim \text{Unif}(0,1)$ and $X = F^{-1}(U)$. Then X is an r.v. with CDF F.*
2. *Let X be an r.v. with CDF F. Then $F(X) \sim \text{Unif}(0,1)$.*

Proof: Universality of the Uniform

① $U \sim \text{unif}(0,1)$ $X = F^{-1}(U)$, for all $x \in R$.

CDF of X: $F_X(x) = P(X \le x) = P[F^{-1}(U) \le x] = P[U \le F(x)] = F(x)$

$0 \le F(x) \le 1$

$\Rightarrow$ X is an r.v. CDF F.

② $Y = F(X) \in [0,1]$. For $y \in R$.

$y \le 0$, $P(Y \le y) = 0$

$y \ge 1$, $P(Y \le y) = 1$

$y \in (0,1)$ $P(Y \le y) = P(F(X) \le y) = P(X \le F^{-1}(y))$

$= F[F^{-1}(y)] = y$

$\Rightarrow$ $Y \sim \text{unif}(0,1)$

Inverse Transform Method

Algorithm 2.3.1: Inverse-Transform Method

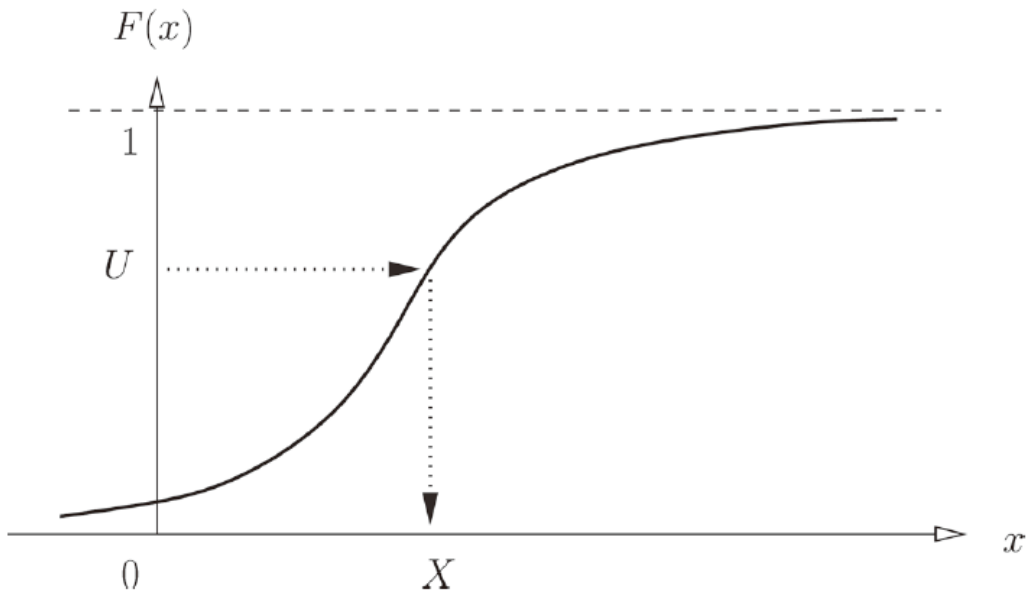
input : Cumulative distribution function F.

output: Random variable X distributed according to F.

1 Generate U from $\mathsf{U}(0,1)$.

2 $X \leftarrow F^{-1}(U)$

3 **return** X



Example: Universality with Logistic

1^0. Logistic CDF. $F(x) = \frac{e^x}{1+e^x}$, $x \in R$

2^0. Given $U \sim \text{unif}(0,1)$ $\quad F^{-1}(x) = \log \frac{x}{1-x}$

$\Rightarrow F^{-1}(U) = \log\left(\frac{U}{1-U}\right) \sim$ logistic

Verify: $\forall x \in R$

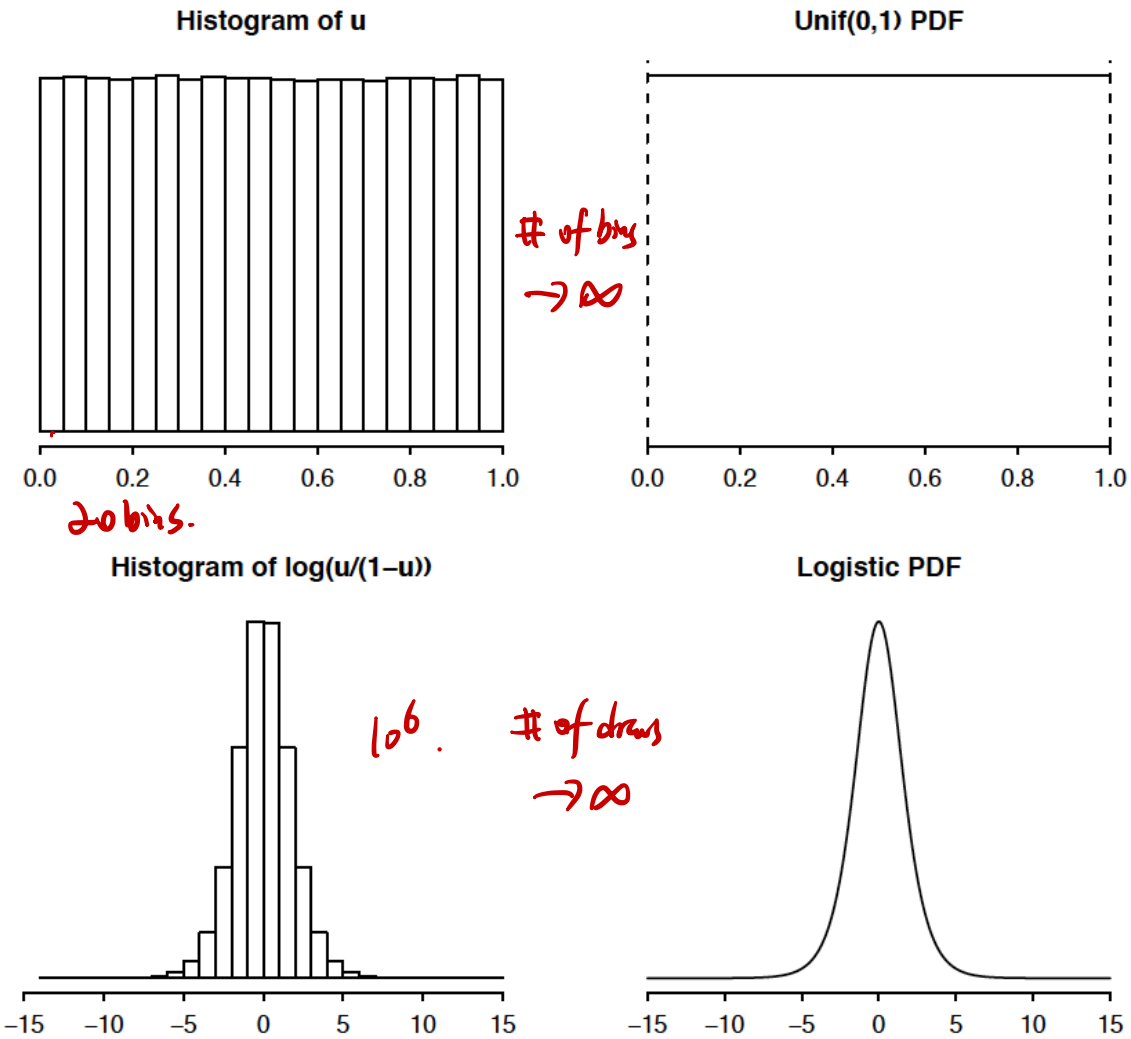
$$P\left(\log\left(\frac{U}{1-U}\right) \le x\right) = P\left(\frac{U}{1-U} \le e^x\right)$$

$$= P\left(U \le \frac{e^x}{1+e^x}\right) = \frac{e^x}{1+e^x} = F(x)$$

Histogram

- Introduced by Karl Pearson
- A graphical representation of the distribution of numerical data
- An estimate of the probability distribution (density estimation) of a continuous variable
- To construct a histogram, the first step is to "bin" the range of values: divide the entire range of values into a series of intervals and then count how many values fall into each interval.
- The bins are usually specified as consecutive, non-overlapping intervals of a variable.

Histogram & PDF



Example: Universality with Rayleigh

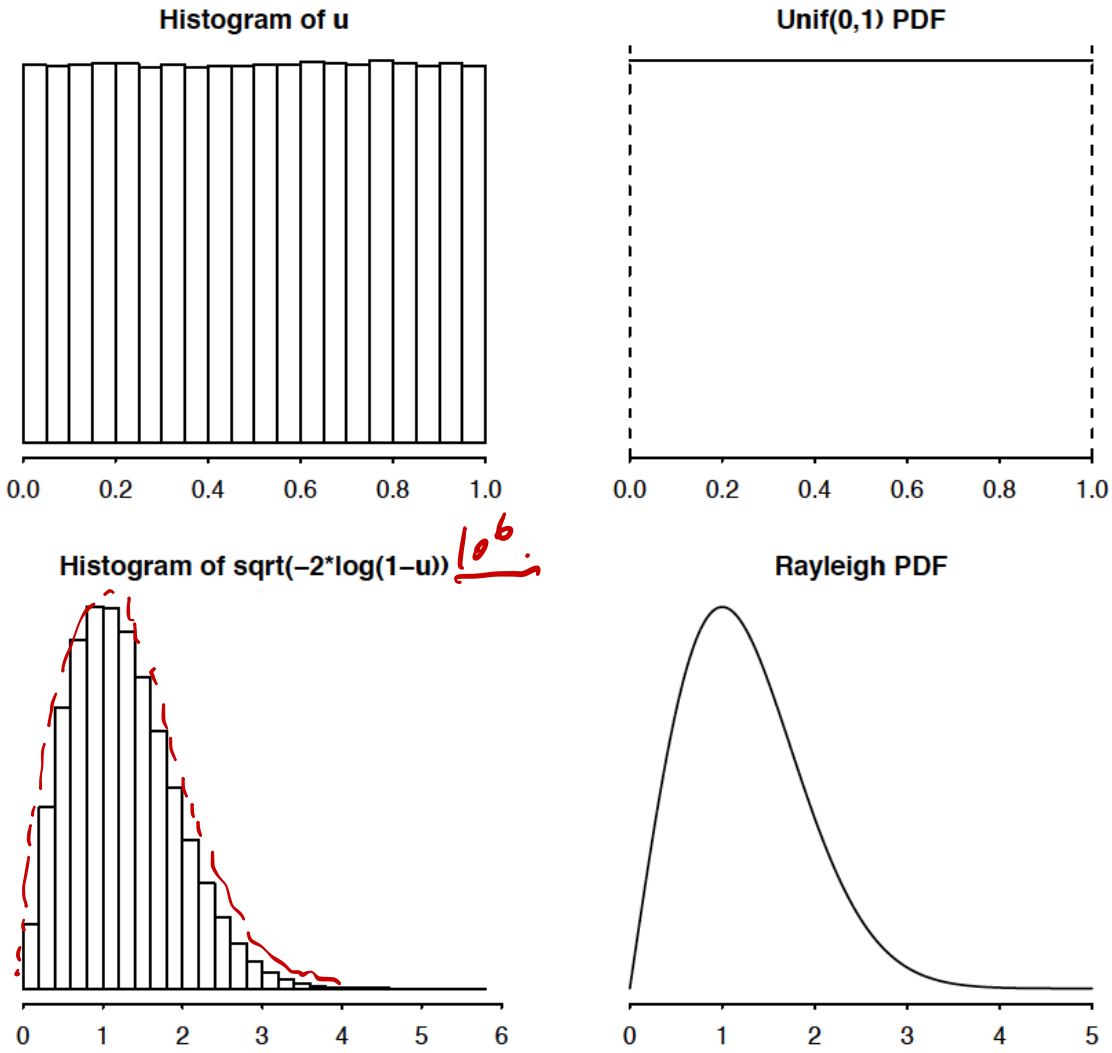
① $F(x) \overset{CDF}{=} 1 - e^{-\frac{1}{2}x^2}$ $x > 0$

$F^{-1}(x) = \sqrt{-2\log(1-x)}$ $0 < x < 1$

② Given $U \sim \text{unif}(0,1)$

$F^{-1}(U) = \sqrt{-2\log(1-U)}$ $\sim$ Rayleigh.

Histogram & PDF



Inverse Transform Method for Discrete Distribution

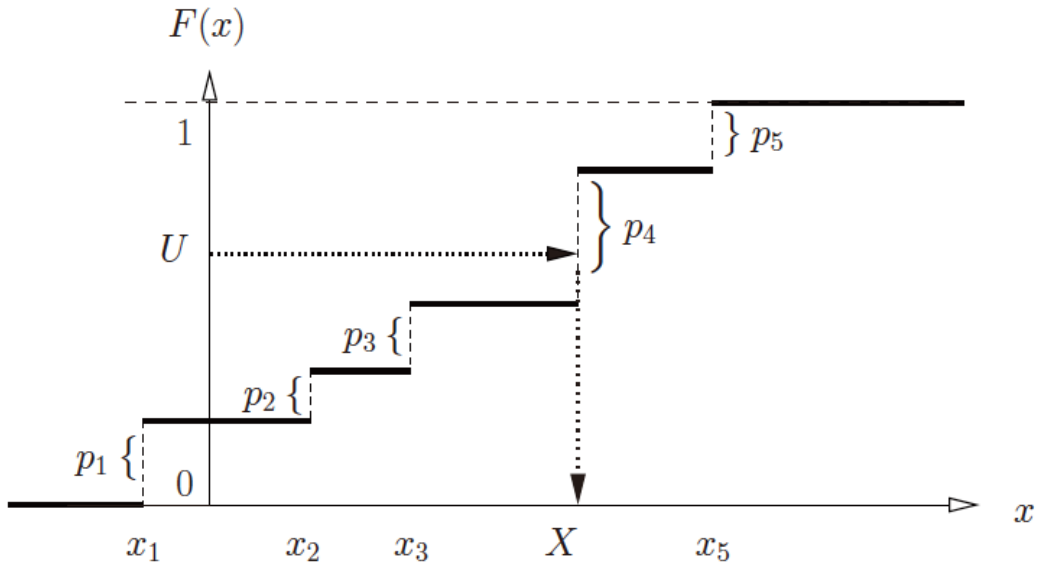


Figure 2.2: Inverse-transform method for a discrete random variable.

Hence the algorithm for generating a random variable from F can be written as follows:

Algorithm 2.3.2: Inverse-Transform Method for a Discrete Distribution

input : Discrete cumulative distribution function F.
output: Discrete random variable X distributed according to F.

1 Generate $U \sim \mathsf{U}(0,1)$.
2 Find the smallest positive integer, k, such that $U \leqslant F(x_k)$. Let $X \leftarrow x_k$.
3 **return** X

Box-Muller

① $(U,T) \to (X,Y)$ Jacobian matrix.

② Joint PDF of r.v.s. U and T

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

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

Let $U \sim \mathrm{Unif}(0, 2\pi)$, and let $T \sim \mathrm{Expo}(1)$ 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?

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

$$|\ \downarrow\ | = 1 \quad \Rightarrow \quad \left|\frac{\partial(x,y)}{\partial(u,t)}\right| = 1$$

Solution

④ $f_{X,Y}(x,y) = f_{U,T}(u,t) \cdot \frac{1}{\left|\frac{\partial(x,y)}{\partial(u,t)}\right|}$

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

$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$ $\sim N(0,1) \cdot N(0,1)$.

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

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

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

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

⑤ . X and Y are independent , $X, Y \sim N(0,1)$

Bivariate Normal Joint PDF

① X, Y i.i.d. r.v.s $N(0,1)$

② $Z = X$

$W = \rho X + \sqrt{1-\rho^2}\, Y \quad -1<\rho<1$

(Z, W) is a bivariate Normal with $\mathrm{Corr}(Z,W)=\rho$.

(Z, W) marginally $N(0,1)$

1°. $Z = X$, $W = \rho X + \sqrt{1-\rho^2}\, Y$ $\Rightarrow$ $X = Z$, $Y = \frac{1}{\sqrt{1-\rho^2}} W - \frac{\rho}{\sqrt{1-\rho^2}} Z$ $\Rightarrow \left|\frac{\partial(x,y)}{\partial(z,w)}\right| = \frac{1}{\sqrt{1-\rho^2}}$

2°. $f_{Z,W}(z,w) = f_{X,Y}(x,y) \cdot \left|\frac{\partial(x,y)}{\partial(z,w)}\right| = f_X(x) \cdot f_Y(y) \cdot \frac{1}{\sqrt{1-\rho^2}}$

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

3°. $Z, W \sim N(0,1)$

Bivariate Normal Joint PDF

Box Muller Method

$X \sim Exp(\lambda)$

$F(x) = 1 - e^{-\lambda x}, x \geq 0$

$F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$.

$U \sim unif(0,1)$

$1-U \sim unif(0,1)$

Algorithm 2.4.2: Normal Random Variable Generation: Box–Muller Approach

$= -\frac{1}{\lambda}\ln U$ still expo(λ)

output: Independent standard normal random variables X and Y.

1 Generate two independent random variables, U_1 and U_2, from $\mathrm{U}(0,1)$.

2 $X \leftarrow (-2\ln U_1)^{1/2}\cos(2\pi U_2)$

3 $Y \leftarrow (-2\ln U_1)^{1/2}\sin(2\pi U_2)$

4 **return** X, Y

Acceptance Rejection Method

- Suppose one can generate samples (relatively easily) from PDF q
- How can random samples be simulated from PDF p

Algorithm 27 Acceptance Rejection algorithm

Let c denote a constant such that $c \geq \sup_\zeta \frac{p(\zeta)}{q(\zeta)}$. Then:

Step 1: Generate $y \sim q$.
Step 2: Generate $u \sim U[0,1]$.
Step 3: If $u < \frac{p(y)}{cq(y)}$, set $x = y$.
Otherwise go back to step 1.

Acceptance Rejection Method

PDF of Y $q(y)$

PDF of U $1.\ 0 \le u \le 1$

1°. $P(X \le \xi) = P\left(Y \le \xi \mid U \le \frac{p(Y)}{Cq(Y)}\right)$

$$= \frac{P\left(Y \le \xi,\ U \le \frac{p(Y)}{Cq(Y)}\right)}{P\left(U \le \frac{p(Y)}{Cq(Y)}\right)} = \frac{\text{Prob. of } Y \le \xi \text{ and accept}}{\text{Prob. of accept}}$$

$$= \frac{\int_{-\infty}^{\xi} \int_{0}^{\frac{p(y)}{Cq(y)}} 1 \cdot du \cdot q(y)\,dy}{\int_{-\infty}^{\infty} \int_{0}^{\frac{p(y)}{Cq(y)}} 1 \cdot du \cdot q(y)\,dy} = \frac{\frac{1}{C}\int_{-\infty}^{\xi} p(y)\,dy}{\frac{1}{C}\int_{-\infty}^{\infty} p(y)\,dy}$$

2°. If p is a valid PDF, $\int_{-\infty}^{\infty} p(y)\,dy = 1 \Rightarrow P(X \le \xi) = \int_{-\infty}^{\xi} p(y)\,dy$

$X \sim p$.

Acceptance Rejection Method

- The inverse transform method operates on the distribution function
- The acceptance rejection method operates on the density functions
  - The algorithm is self-normalizing in the sense that the PDF $p(\cdot)$ does not need to be normalized.
  - $c \geq 1$

$c \geq \sup_z \frac{p(z)}{q(z)}$ $\quad \forall z,\ \frac{p(z)}{q(z)} \leq c$

$\Rightarrow p(z) \leq c\,q(z)$

$\Rightarrow \int_{-\infty}^{\infty} p(z)dz \leq \left(\int_{-\infty}^{\infty} q(z)dz\right) \cdot c$

$\Rightarrow 1 \leq c$

Example: Normal Distribution

1°. Expo(1). PDF. $q(x) = e^{-x}$, $x \geq 0$

2°. $|Z| \sim \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, $x > 0$

$= p(x)$

$Z \sim N(0,1)$ $\quad |Z|$ $\quad 3^\circ$. $\sup_x \frac{p(x)}{q(x)} = \sup_x \sqrt{\frac{2}{\pi}} e^{x - \frac{1}{2}x^2} = \sqrt{\frac{2e}{\pi}}$

- N(0,1) can be generated from Expo(1)

4°. $C = \sqrt{\frac{2e}{\pi}}$.

Example: Normal Distribution

Step 1 : generate exp. $Y \sim q$. (inverse-transform)

Step 2 : $U \sim \text{unif}(0,1)$

Step 3 : $U \le \frac{p(y)}{c\,q(y)} \Leftrightarrow U < \frac{\frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}}{\sqrt{\frac{2e}{\pi}}\, e^{-y}} = e^{-\frac{1}{2}(y-1)^2}$

$X = y$ if $U \le e^{-\frac{1}{2}(y-1)^2}$

---

obtain $X \sim |Z|$.

For $Z \sim N(0,1)$ $\quad Z = \begin{cases} X & \text{w.p. } 0.5 \\ -X & \text{w.p. } 0.5 \end{cases}$

Variance Reduction in Simulation

For M.C.

① N i.i.d. samples of x_k, $k=1,2,..N$. from PDF of p.

$$\frac{1}{N}\sum_{k=1}^{N} C(x_k) \longrightarrow E_p\{C(x)\} = \int_A C(x)p(x)dx \quad \text{w.p.1.}$$

↓

unbiased estimation

② The variance of estimate could be very large.

Variance Reduction in Simulation

- Popular Schemes including
  - Importance sampling
  - Conditioning
  - Stratified sampling

Importance Sampling

① Let $p(x)$ denote a target distribution, for any density $q(x)$.

$\frac{p(x)}{q(x)}$ is finite for all $x \in A$.

$$E_p\{c(x)\} = \int_A c(x)p(x)dx = \int_A c(x)\frac{p(x)}{q(x)} \cdot q(x)dx = E_q\left[c(x)\frac{p(x)}{q(x)}\right]$$

② q : importance distribution.

$$\hat{c}_N = \frac{1}{N}\sum_{k=1}^{N} c(x_k)\frac{p(x_k)}{q(x_k)} \qquad x_k \sim q.$$

By SLLN. $\hat{c}_N \to E_p\{c(x)\}$ w.p. 1.

By CLT, $\lim_{N\to\infty} \sqrt{N}\left(\hat{c}_N - E_p\{c(x)\}\right) \xrightarrow{d} N(0, \mathrm{Var}_q(c(x)))$

$$\mathrm{Var}_q(c(x)) = \int_A c^2(x)\frac{p^2(x)}{q(x)}dx - E_p^2\{c(x)\}$$

Importance Sampling

$p(x) \propto e^{\beta\phi(x)}$

$p(x) = \frac{e^{\beta\phi(x)}}{C}$

$C = \sum_{x \in \mathcal{X}} e^{\beta\phi(x)}$

$\mathcal{X}$ is large.

C is hard to compute

③ Self-normalized importance sampling.

$p(x)$ is known only up to a normalization constant.

④ Define weight $w(x) = \frac{p(x)}{q(x)}$.

$$\widehat{C_N}' = \frac{\frac{1}{N}\cdot\sum_{k=1}^{N} C(x_k)\, w(x_k)}{\frac{1}{N}\cdot\sum_{k=1}^{N} w(x_k)}, \quad x_k \sim \mathcal{E}$$

$$\hat{C}_N = \frac{1}{N}\sum_{k=1}^{N} C(x_k)\, w(x_k)$$

[$E_q\{w(x)\} = \int_A \frac{p(x)}{q(x)} \cdot q(x)\,dx$

$= \int_A p(x)\,dx$

$= 1$]

By SLLN, $\frac{1}{N}\sum_{k=1}^{N} C(x_k)\, w(x_k) \xrightarrow{w.p.1} E_p\{C(x)\}$. , $\frac{1}{N}\sum_{k=1}^{N} w(x_k) \xrightarrow{w.p.1} 1$

$\Rightarrow \widehat{C_N}' \to E_p\{C(x)\}$. w.p.1.

Importance Sampling

⑥. tilted densities. (Esscher transform)

$$q(x) = \frac{e^{tx} p(x)}{\int e^{tx} p(x) dx}, \quad t \in R.$$

1°. if $p \sim N(\mu, \sigma^2)$ $\rightarrow$ $q \sim N(\mu + \sigma^2 t, \sigma^2)$

2°. if $p \sim Expo(\lambda)$ $\rightarrow$ $q \sim Expo(\lambda - t)$

Example of Importance Sampling

[$E(I(X_k>\alpha)) = P(X_k>\alpha)$]

① Standard Monte Carlo estimator.

N i.i.d. samples. $\hat{c} = \frac{1}{N}\sum_{k=1}^{N} I(X_k>\alpha)$, $X_k \sim N(0,1)$

② Choose the importance density $q = N(\mu,1)$ ($t=\mu$)

$$\hat{c}_{IS} = \frac{1}{N}\sum_{k=1}^{N} I(X_k>\alpha)\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}X_k^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X_k-\mu)^2}} = \frac{1}{N}\sum_{k=1}^{N} I(X_k>\alpha)e^{\frac{1}{2}\mu^2-\mu X_k}$$

Suppose for a fixed real number α, the aim is to evaluate $c = \mathbb{P}(x > \alpha)$, where $x \sim N(0,1)$.

$X_k \sim N(\mu,1)$ iid.

$\mu=\alpha$ as a reasonable choice.

Example of Importance Sampling

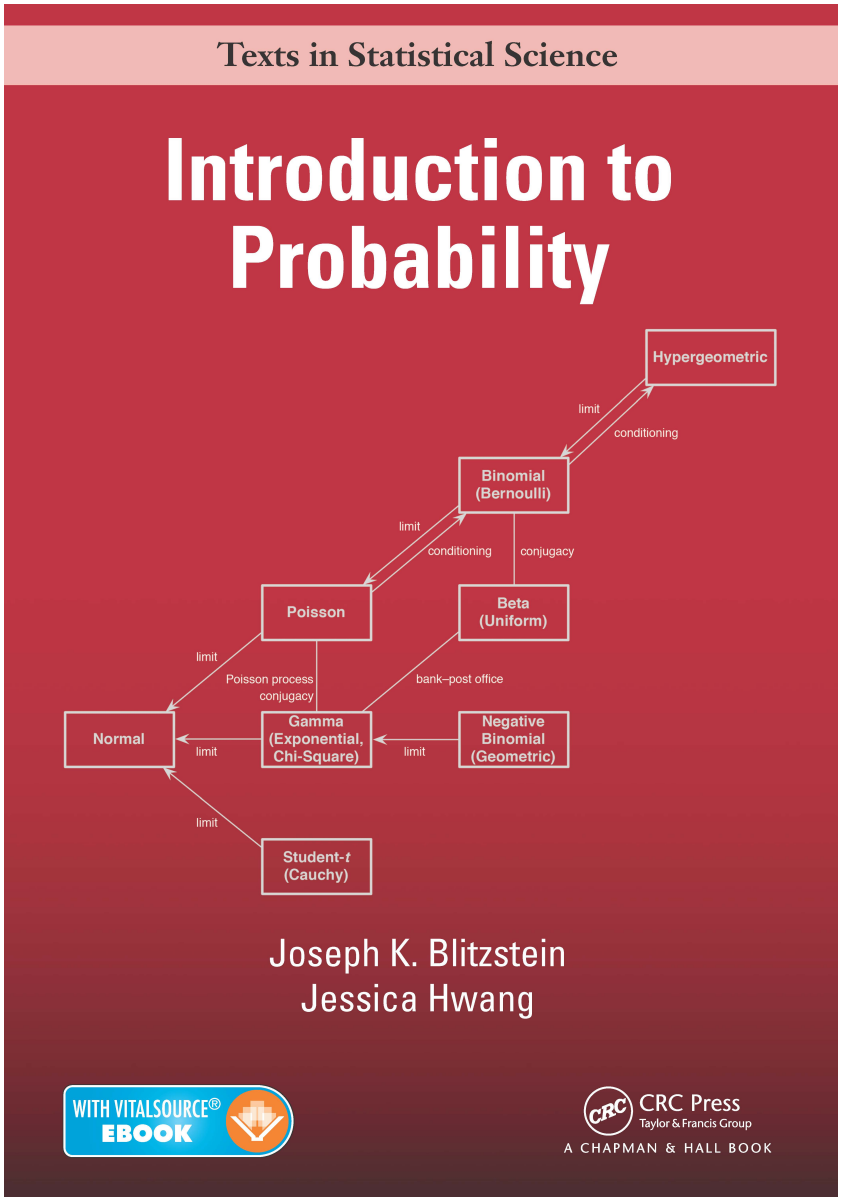
$N = 50000$; $\hat{c} = 0$

$\hat{c}_{IS} = 6.28 \times 10^{-16}$ $(H=d=8)$

Suppose for a fixed real number $\alpha = 8$, the aim is to evaluate $c = \mathbb{P}(x > \alpha)$, where $x \sim N(0, 1)$.

Outline

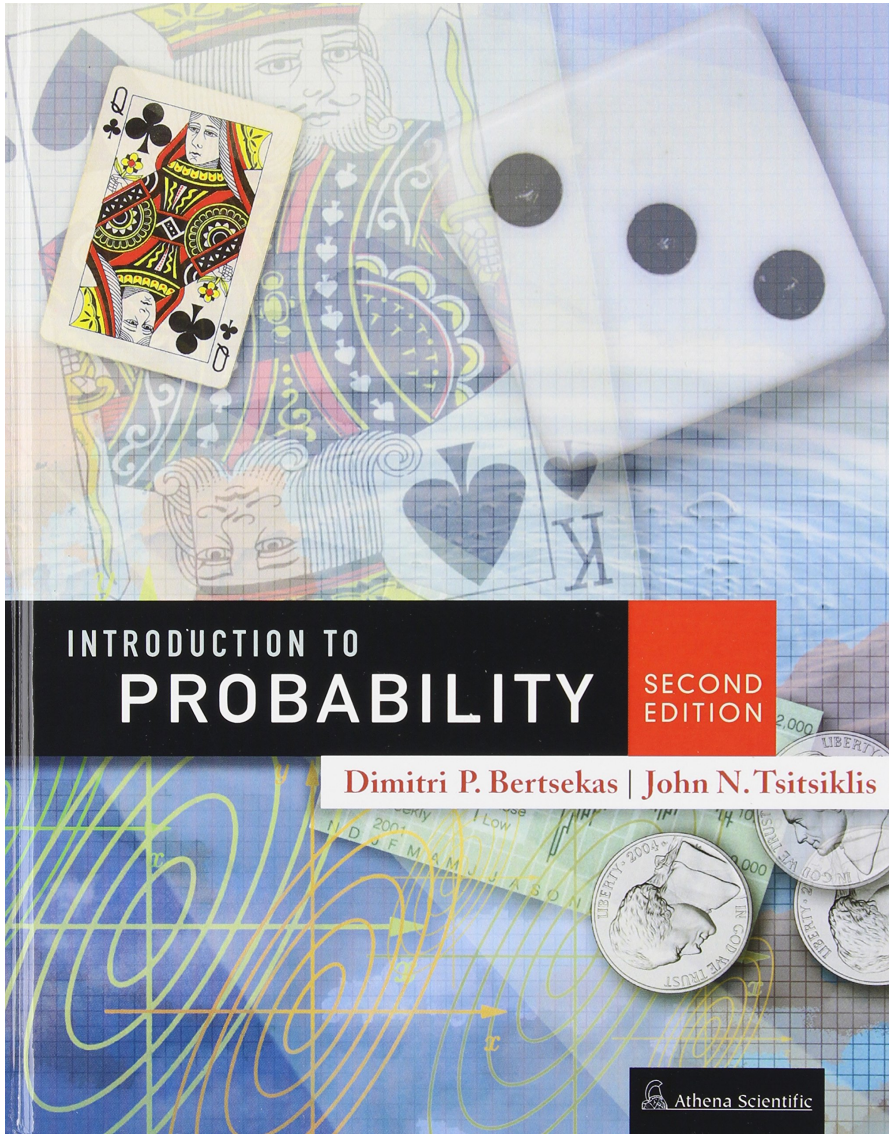
Reference for Probability & Statistics



BH

- Introduction to Probability
- Chapman & Hall/CRC, 2014.
- Chapman & Hall/CRC, 2019.

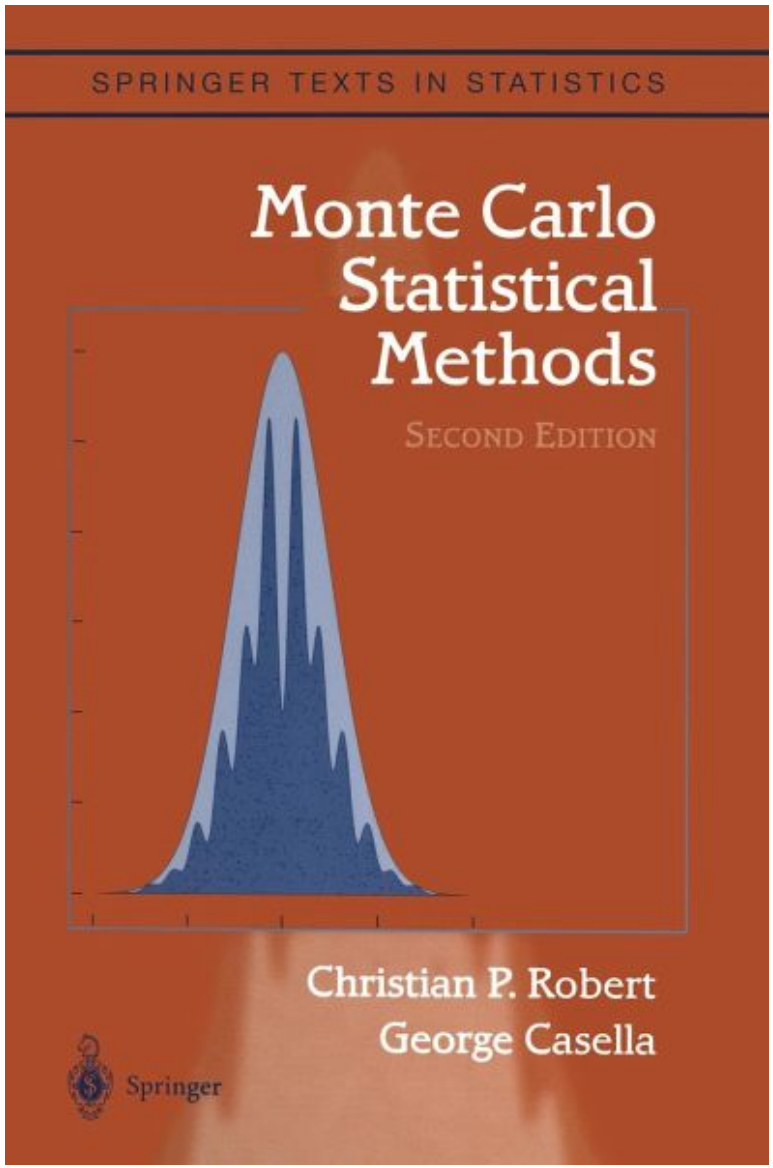
Reference for Probability & Statistics



BT

- Introduction to Probability (2nd Edition)
- Athena Scientific, 2008.

Reference for Monte Carlo Methods



RC

- Monte Carlo Statistical Methods (2nd Edition)
- Springer, 2004.