ACM 116: Lectures 3–4

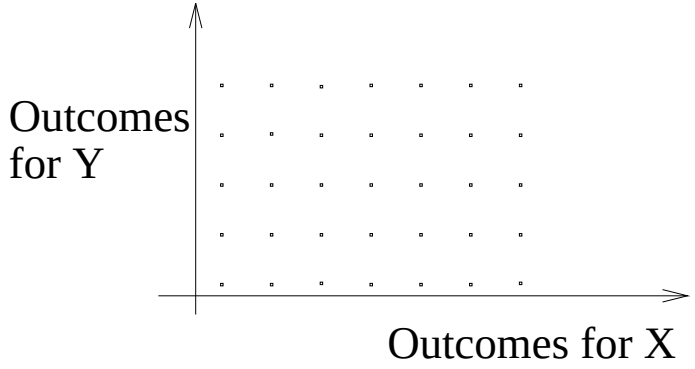
- Joint distributions
- The multivariate normal distribution
- Conditional distributions
- Independent random variables
- Conditional distributions and Monte Carlo: Rejection sampling
- Variance of a random variable
- Covariance
- Conditional Expectation

Joint distributions

Two random variables X and Y.

Interested in their joint outcome.

$$(X, Y) = (x, y) \Leftrightarrow X = x \quad \text{and} \quad Y = y.$$



Example : Server on the web

X : # customers hitting your server in the next hour

Y : # customers in the next 10 minutes

or # customers who will purchase an item in the next hour

Joint frequency function

X, Y discrete r.v.'s taking on values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ resp.

Joint frequency (distribution) function :

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

Marginal probabilities

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

Why? Addition rule

Joint frequency function

Several random variables : similar story

$X_1, \ldots X_m$ r.v.'s defined on the same sample space

$$p(x_1, \ldots x_m) = P(X_1 = x_1, \ldots, X_m = x_m)$$

$$p_{X_1}(x_1) = \sum_{x_2, \ldots, x_m} p(x_1, x_2, \ldots, x_m)$$

$$p_{X_1, X_2}(x_1, x_2) = \sum_{x_3, \ldots, x_m} p(x_1, x_2, x_3, \ldots, x_m)$$

Example: the multinomial distribution

- n *fixed* independent experiments resulting in one of r possible outcomes.
- Probabilities $p_1, \ldots, p_r$
- X_i, $1 \leq i \leq r$, # outcomes of type i.

Joint distribution of the X_i's ?

$$P(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \ldots x_r!} p_1^{x_1} \ldots p_r^{x_r}$$

Why?

Example : multinomial distribution

- Outcome of experiments such that $(\boldsymbol{X_1 = x_1, \ldots, X_r = x_r})$ has prob. $\boldsymbol{p_1^{x_1} \ldots p_r^{x_r}}$ (all equally likely). Indeed,

$$\begin{aligned} \text{First } x_1 &\rightarrow \text{Type 1} \Rightarrow p_1^{x_1} \\ \text{Next } x_2 &\rightarrow \text{Type 2} \Rightarrow p_2^{x_2} \\ &\ldots \\ \text{Last } x_r &\rightarrow \text{Type r} \Rightarrow p_r^{x_r} \end{aligned}$$

- How many configurations?

$$\binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{x_r}{x_r} = \frac{n!}{x_1!\ldots x_r!}$$

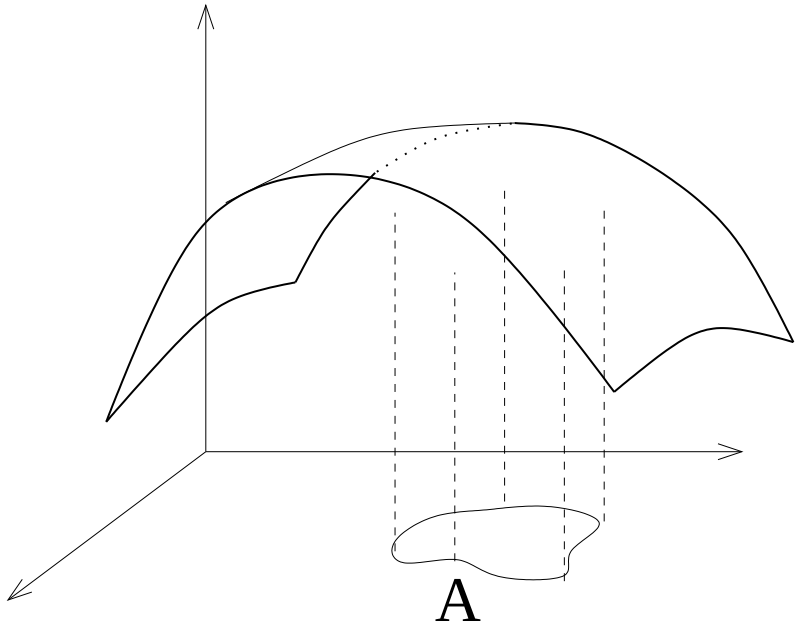
Joint distributions: the continuous case

X and Y continuous r.v.'s

Joint density function $p(x, y)$ s.t.

$$P((x, y) \in A) = \int_A f(x, y)\, dxdy$$

for any 'reasonable' set A.

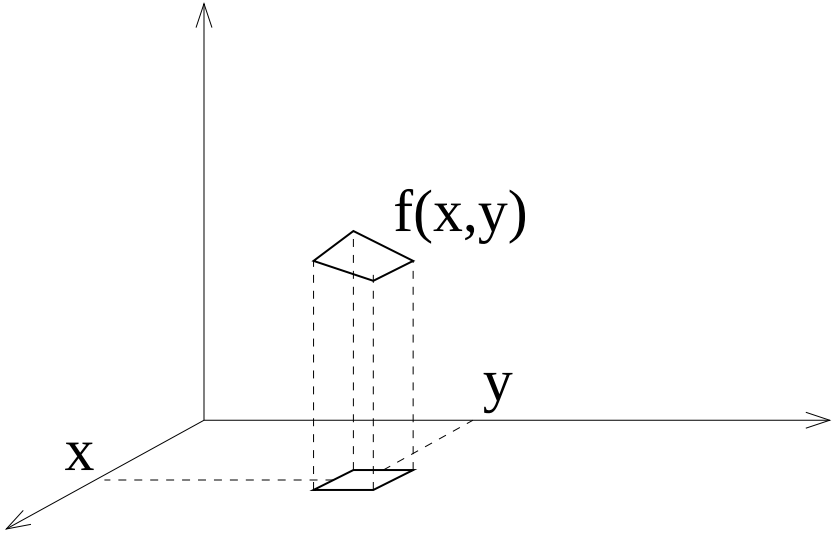


$P((X, Y) \in A)$ = this volume

Joint distributions: the continuous case

Interpretation

$$P(x \leq X \leq x + dx, y \leq Y \leq y + dy) \sim f(x, y)\, dxdy$$

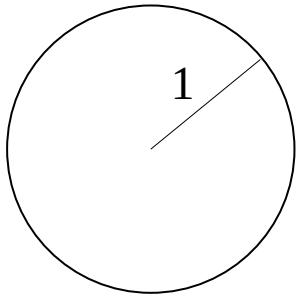


Marginal density of X :

$$f_X(x) = \int f_{XY}(x, y)\, dy$$

Joint distributions: the continuous case

Example Disk of radius 1.



$$f(x,y) = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

R: distance from point to origin. Density of R ?

Joint distributions: the continuous case

R: distance from point to origin. Density of R ?

$$P(R \le r) = \frac{\pi r^2}{\pi} = r^2$$

$$\Rightarrow f_R(r) = \begin{cases} 2r & \text{if } 0 \le r \le 1 \\ 0 & \text{otherwise} \end{cases}$$

X : first coordinate. Density of X ?

$$f_X(x) = \int f_{XY}(x,y)\,dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\,dy$$

$$= \frac{2}{\pi}\sqrt{1-x^2}$$

$$\Rightarrow f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{if } -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Bivariate normal density

5 parameters : $-\infty < \mu_x, \mu_y < \infty, \qquad 0 < \sigma_x, \sigma_y < \infty, \qquad -1 \leq \rho \leq 1$

Extensively used in modeling

Galto, mid 19th century : modeling of the heights of fathers and sons, e.g.

$$X = \text{father's height}$$
$$Y = \text{son's height}$$

Density

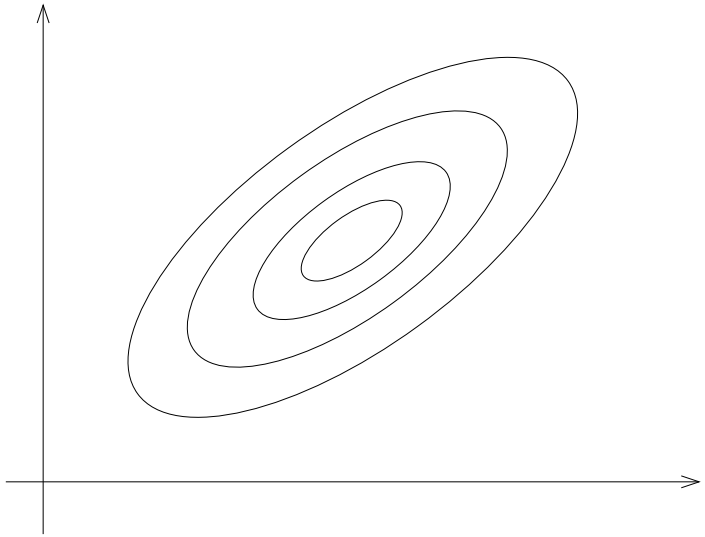
$\underline{x} = (x, y), \qquad \underline{\mu} = (\mu_x, \mu_y)$ is the vector of means

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$ is the covariance matrix

$$f_{XY}(x, y) = f_{\underline{X}}(\underline{x}) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})}$$

Bivariate normal density

Level sets = ellipses



Interesting calculation : marginals

$$X \sim N(\mu_x, \sigma_x^2)$$

$$Y \sim N(\mu_y, \sigma_y^2)$$

Independent random variables

Two random variables are independent if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for 'any' subsets A and B.

- Discrete r.v.'s : Independence iff

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

- Continuous r.v.'s : Independence iff

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

(Complicated)

Example

Node of a communication network

If two packets of info arrive within time τ of each other they collide and then have to be retransmitted.

Times of arrival $\sim$ independent $\boldsymbol{U}[\mathbf{0}, \boldsymbol{T}]$.

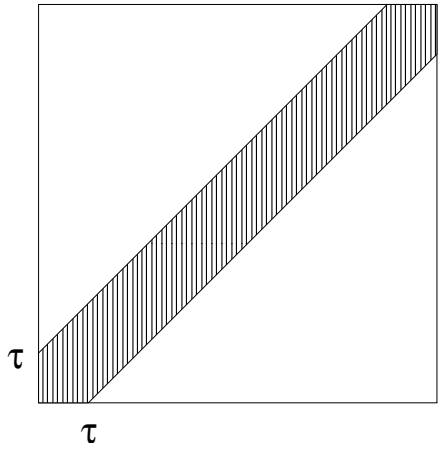
What is the probability that they collide ?

Independent random variables

$T_1, T_2 \sim$ independent $U[0, T]$.

Joint distribution :

$$f(t_1, t_2) = \frac{1}{T^2} \quad \sim \text{uniform distribution over the square}$$



$$\begin{aligned} P(\text{Collision}) &= 1 - (1 - \frac{\tau}{T})^2 \\ &= 2\frac{\tau}{T} - (\frac{\tau}{T})^2 \\ &= \frac{\tau}{T}(2 - \frac{\tau}{T}) \end{aligned}$$

Conditional distributions

Discrete case : X and Y joint discrete r.v.'s

Conditional prob. of $X = x_i$ given $Y = y_j$ is

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

Notation : $p_{X|Y}$

Continuous case : X and Y jointly continuous r.v.'s

Conditional density of Y given X is defined to be

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

NB :

$$P(y \leq Y \leq y + dy | x \leq X \leq x + dx) = \frac{f_{XY}(x, y)\, dxdy}{f_X(x)\, dx}$$
$$= \frac{f_{XY}(x, y)\, dy}{f_X(x)}$$

Conditional distributions

- Independence $\Rightarrow f_{Y|X} = f_Y$
- Joint density = conditional $\times$ marginal

$$f_{XY}(x, y) = f_{Y|X}(y|x) f_X(x)$$

- Law of total probability

$$f_Y(y) = \int f_{Y|X}(y|x) f_X(x)\, dx$$

NB : $p_{X|Y}(x|Y = y)$ is a distribution :

- $p_{X|Y}(x|Y = y) \geq 0$
- $\sum_{\text{all } x} p_{X|Y}(x|Y = y) = 1$

i.e. conditional distribution.

Example

Imperfect particle detector

Detects each incoming particle w.p. p only.

- N true number of particles. Poisson distribution

$$P(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}$$

- X: detected number of particles

What is the distribution of the detected number of particles?

Conditional distributions

$$p_{X|N=n} \sim \text{Binomial}(n, p)$$

$$\begin{aligned}
P(X = k) &= \sum_{n=0}^{\infty} P(N = n) P(X = k | N = n) \\
&= \sum_{n \geq k} \binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n e^{-\lambda}}{n!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} \sum_{n \geq k} \lambda^{n-k} \frac{(1-p)^{n-k}}{(n-k)!} \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda} e^{\lambda(1-p)} \\
&= e^{-\lambda p} \frac{(\lambda p)^k}{k!} \qquad \sim \text{Poi}(\lambda p)
\end{aligned}$$

Conditional distributions

Example: Bivariate normal density

$\boldsymbol{X}$ and $\boldsymbol{Y}$ bivariate normal distribution

Distribution of $\boldsymbol{Y}|\boldsymbol{X} = x$ is normal too :

$$\boldsymbol{N}(\boldsymbol{\mu_y} + \boldsymbol{\rho}\frac{\boldsymbol{\sigma_y}}{\boldsymbol{\sigma_x}}(x - \boldsymbol{\mu_x}), (\mathbf{1} - \boldsymbol{\rho^2})\boldsymbol{\sigma_y^2})$$

Linear regression !

Sampling & Monte Carlo

Density function f we wish to sample from.

- f is nonzero on an interval $[a, b]$ and zero outside the interval (a and b may be infinite).
- Let M be a function such that $M(x) \geq f(x)$ on $[a, b]$ and let

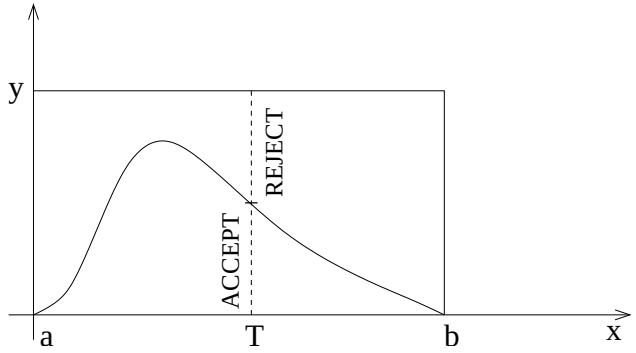
$$m(x) = \frac{M(x)}{\int_a^b M(x)dx}$$

The idea is to choose M so that it is easy to generate random variables from m. If $[a, b]$ is finite, m can be chosen to be the uniform distribution.

Rejection sampling algorithm.

- *Step 1:* Generate T with the density m.
- *Step 2:* Generate U, uniform on $[0, 1]$ and independent of T.

$$\text{If } M(T) \times U \leq f(T) \quad \rightarrow \quad \text{accept, } X = T.$$
$$\text{Otherwise} \quad \rightarrow \quad \text{reject, goto step 1}$$



Remark. High efficiency if algorithm accepts with high probability, i.e. M close to f.

Why does this work?

To show

$$P(X \in A) = \int_A f(t)\,dt$$

$$P(X \in A) = P(T \in A \mid \text{Accept}) = \frac{P(T \in A \text{ and Accept})}{P(\text{Accept})}$$

Condition on $T = t$ ($I = \int_a^b M(t)dt$)

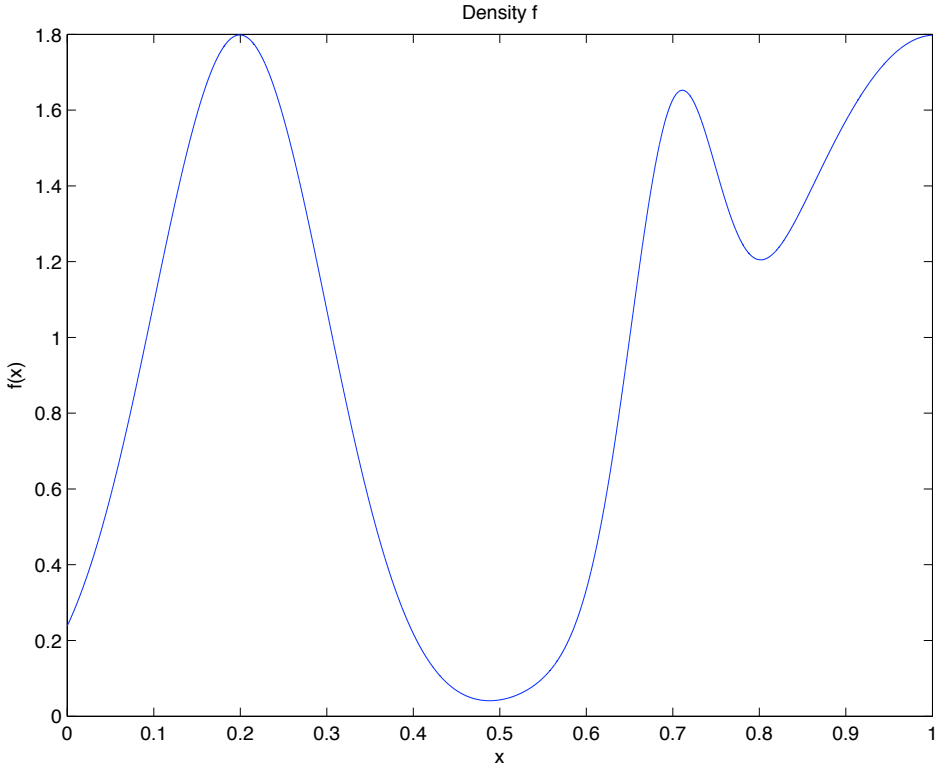
$$\begin{aligned} P(T \in A \text{ and Accept}) &= \int_a^b P(T \in A \text{ and Accept}|T = t) f_T(t)\,dt \\ &= \int_a^b P(U \le f(t)/M(t) \text{ and } t \in A) m(t)\,dt \\ &= \int_A \frac{f(t)}{M(t)} m(t)\,dt = \frac{1}{I}\int_A f(t)\,dt \end{aligned}$$

Similarly

$$P(\text{Accept}) = \int_a^b P(\text{Accept}|T = t) m(t)\,dt = \int_a^b \frac{f(t)}{M(t)} m(t)\,dt = \frac{1}{I}.$$

Example

Suppose we want to sample from a density whose graph is shown below.



In this case we let $M(T)$ be the maximum of f over the interval $[0, 1]$, namely

$$M(x) = \max(f), \quad 0 \leq x \leq 1$$

so that m is the uniform density over the interval $[0, 1]$.

Implementation

```
N = 1000;
t = (1:N)/N;
M = max(f(t));

n = 2000;
x = rejection_sampling(n,@f,M);
```

This routine will sample n iid samples from the density f

```
function x = rejection_sampling(n,fun,M)


x = [];
for k = 1:n,
  OK = 0;
  while not(OK);
    T = rand(1); % Generate T
    U = rand(1); % Generate U
    if (M*U <= fun(T)),
      OK = 1;
      x = [x T];
    end
  end
end
```

Visualization of the results

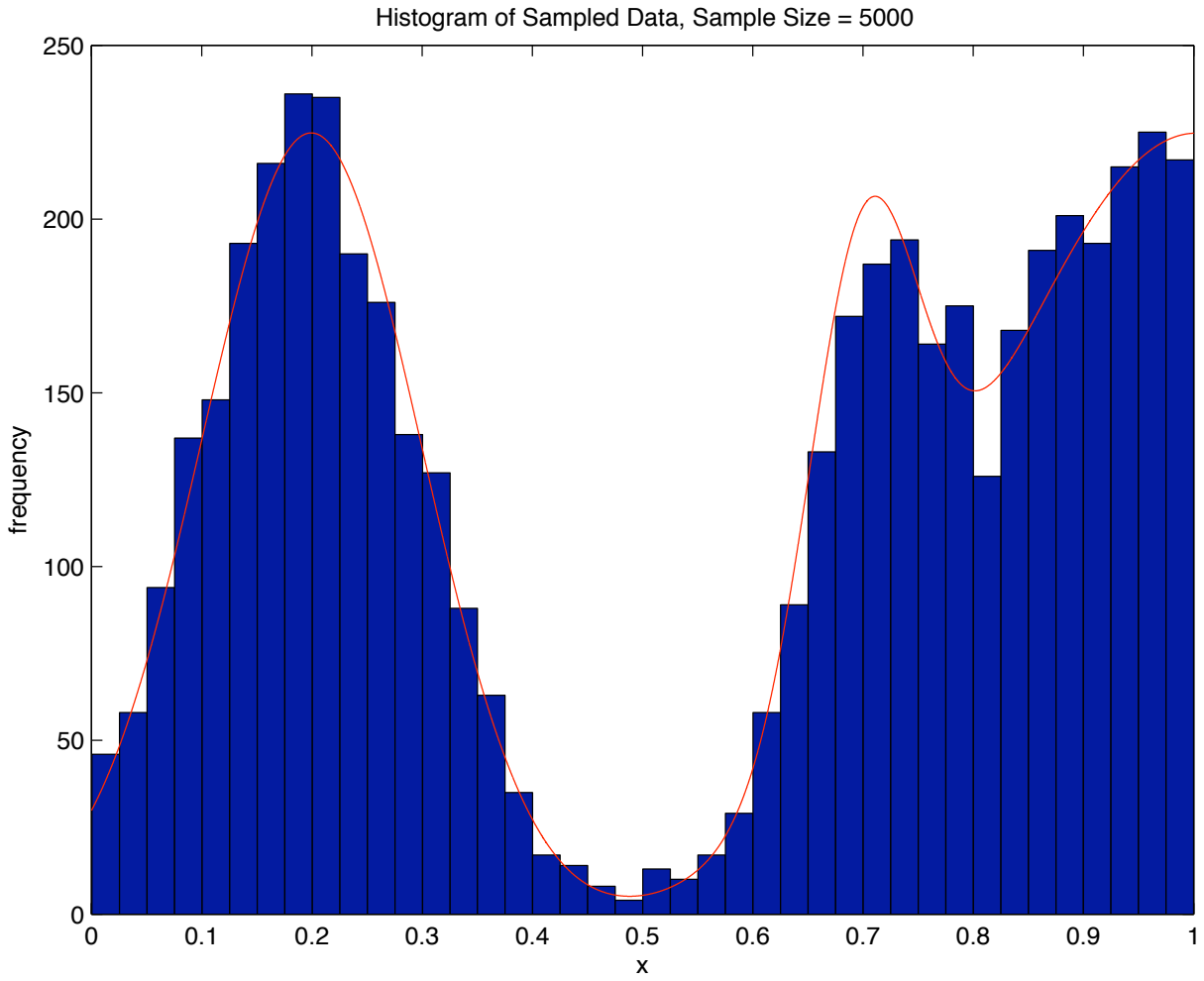


Figure 1: Histogram of the Sampled Data

Variance and Standard Deviation

- The **Expected value** can be viewed as an indication of the central value of the density or frequency function.
- The **variance** or **standard deviation** gives an indication about the dispersion of the probability distribution.
- The **variance** of random variable X is the mean square deviation from $E[X] = \mu$

$$\text{Var}(X) = E[(X - \mu)^2] = E[(X - E[X])^2].$$

The **standard deviation** is the square root of the variance

$$SD(X) = \sqrt{\text{Var}(X)}.$$

- We often write $\mu = E[X]$ and $\sigma^2 = \text{Var}(X)$ so that $\sigma = SD(X)$.
- We may also calculate the variance as follows:

$$\text{Var}(X) = E[X^2] - (E[X])^2.$$

Examples

- Bernouilli: $X \sim \text{Ber}(p)$

$$\text{Var}(X) = p(1-p)$$

- Normal: $X \sim N(\mu, \sigma^2)$

$$\begin{aligned} \text{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int (x-\mu)^2 e^{-(x-\mu)^2/(2\sigma^2)}\, dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int u^2 e^{-u^2/2}\, du \\ &= \sigma^2. \end{aligned}$$

- Remark: $Y = a + bX \Rightarrow \text{Var}(Y) = b^2\text{Var}(X)$
- Example: X standard normal $\mu = 0$, $\sigma = 1$

$$Y = \mu + \sigma X \sim N(\mu, \sigma^2)$$

Then

$$\text{Var}(Y) = \sigma^2\text{Var}(X) = \sigma^2.$$

Covariance and correlation

- The **variance** is a measure of the variability of a r.v.
- The **covariance** is a measure of the joint variability of two r.v's.
- The **covariance** of two jointly distributed r.v's (X, Y) is given by

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[(X - \mu_X)(Y - \mu_Y)].$$

- Another formula

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y].$$

Interpretation

- Positive covariance
  - When X is larger than its mean, Y tends to be larger than its mean as well.
  - When X is smaller than its mean, Y tends to be smaller than its mean as well.
- Negative covariance
  - When X is larger than its mean, Y tends to be smaller than its mean.
  - When X is smaller than its mean, Y tends to be larger than its mean.

Properties

1. $\text{Cov}(X, X) = \text{Var}(X)$
2. X and Y two r.v's:

$$\text{Var}(X + Y) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y)$$

3. If X and Y are independent, then

$$\text{Cov}(X, Y) = 0 (= E[XY] - E[X]E[Y]).$$

4. Let $X_1, X_2, ...X_m$ be m r.v's and $Y_1, Y_2, ...Y_n$ be n r.v's, let $S_1 = \sum_{i=1}^{m} X_i$ and $S_2 = \sum_{i=1}^{n} Y_i$. Then

$$\text{Var}(S_1) = \text{Var}(\sum_{i=1}^{m} X_i) = \sum_{i=1}^{m} \text{Var}(X_i) + 2 \sum_{i<j} \text{Cov}(X_i, X_j),$$

and

$$\text{Cov}(S_1, S_2) = \sum_{i=1}^{m} \sum_{j=1}^{n} \text{Cov}(X_i, Y_j).$$

5. Suppose $\boldsymbol{X_1}, \boldsymbol{X_2}, ...\boldsymbol{X_m}$ are independent:

$$\text{Var}(\boldsymbol{X_1} + \ldots + \boldsymbol{X_m}) = \text{Var}(\boldsymbol{X_1}) + \ldots + \text{Var}(\boldsymbol{X_m}).$$

Example

The variance of the binomial $X \sim \text{Bin}(n, p)$.

X is the sum of n independent Bernouilli's

$$X = I_1 + \ldots + I_n$$

and therefore

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(I_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

Correlation

The **correlation** between two r.v's X and Y is defined by

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

- The correlation is a dimensionless quantity
- The correlation ρ between X and Y obeys

$$-1 \leq \rho \leq 1.$$

- The correlation is a measure of the strength of the linear relationship existing between X and Y.

$$\rho \pm 1 \Leftrightarrow aX + b = Y.$$

Remark: (X, Y) bivariate normal: the correlation between X and Y is ρ and

$$\text{Cov}(X, Y) = \rho \sigma_X \sigma_Y$$

Conditional expectation

The **conditional expectation** of Y given $X = x$ is the mean of the conditional distribution of Y given $X = x$.

For example, in the discrete case we have

$$E[Y|X = x] = \sum_y y P_{Y|X}(y|x)$$

and more generally the conditional expectation of $h(Y)$ given $X = x$ is

$$E[h(Y)|X = x] = \sum_y h(y) P_{Y|X}(y|x).$$

Example

- $X \sim$ Binomial(n, p).
- Set $m \leq n$ and let Y be the number of successes in the first m trials.

What are the conditional distribution and mean of Y given $X = x$?

- Conditional distribution:

$$P(Y = y|X = x) = \frac{\binom{m}{y}\binom{n-m}{x-y}}{\binom{n}{x}}.$$

- Conditional mean:

$$Y = I_1 + \ldots I_m$$

and

$$\begin{aligned} E(Y|X = x) &= E(I_1|X = x) + \ldots + E(I_m|X = x) \\ &= P(I_1|X = x) + \ldots + P(I_m|X = x) \\ &= x/n + \ldots + x/n = \frac{m}{n}x \end{aligned}$$

Conditional expectation as a random variable

- Assuming that the conditional expectation of $\boldsymbol{Y}$ given $\boldsymbol{X = x}$ exists for every $\boldsymbol{x}$, it is a well defined function of $\boldsymbol{X}$ and, hence, a random variable. (The expectation of $\boldsymbol{Y}$ given $\boldsymbol{X = x}$ is a function depending on $\boldsymbol{x}$: $\boldsymbol{E[Y|X = x] = g(x)}$. $\boldsymbol{E[Y|X] = g(X)}$ is r.v.)
- Example:

$$\boldsymbol{E(Y|X) = \frac{m}{n} X.}$$

- This random variable has an expectation, and a variance (proviso absolute convergence, etc.) Its expectation is

$$\boldsymbol{E[E(Y|X)]}$$

Iterated Expectation

Theorem

$$E(Y) = E[E(Y|X)]$$

Interpretation: the expectation of Y can be calculated by first conditioning on X, finding $E(Y|X)$ and then averaging this quantity with over X:

$$E[Y] = \sum_x E[Y|X = x]P_X(x)$$

where

$$E[Y|X = x] = \sum_y yP_{Y|X}(y|x).$$

Example: $E(Y|X) = \frac{m}{n}X$ and $E(X) = np$ give

$$E(Y) = \frac{m}{n}E(X) = mp.$$

Proof

We need to show

$$\boldsymbol{E(Y) = \sum_x E(Y|X=x)P(X=x)}$$

where

$$\boldsymbol{E(Y|X=x) = \sum_y y P_{Y|X}(y|x).}$$

We have

$$\begin{aligned}\sum_x \boldsymbol{E(Y|X=x)P_X(x)} &= \sum_y y \sum_x \boldsymbol{P_{Y|X}(y|x)P_X(x)} \\ &= \sum_y \boldsymbol{y P_Y(y) = E(Y).}\end{aligned}$$

Random Sums

Consider sums of the type

$$T = \sum_{i=1}^{N} X_i,$$

where

- N is a r.v. with finite expectation $E(N) < \infty$,
- the X_i's are independent of N, with comon mean $E[X_i] = E[X]$, $\forall i$.

Such sums arise in a variety of applications

- Insurance companies might receive N claims in a given period of time and the amounts of the individual claims may be modeled as r.v's $X_1, X_2, \ldots$.
- N is the number of jobs in a single server queue and X_i the service time for the ith job. T is the time to serve all the jobs in the queue.

$$E(T) = \sum_{\text{all } n} E(T|N = n)P(N = n)$$

and

$$E(T|N = n) = E\left(\sum_{i=1}^{n} X_i\right) = nE(X).$$

i.e.

$$E(T) = \sum_{n} nE(X)P(N = n) = E(N) \cdot E(X).$$

In other words, the average time to complete n jobs when n is random is the average value of N times the average time amount to complete a job.

The Moment Generating Function

The moment generating function (mgf) of a random variable X is given by

$$M(t) = E[e^{tX}]$$

(In the continuous case, $M(t) = \int e^{tx} f(x)\, dx$, Laplace transform of f)

The mgf may not exist

- If X is Cauchy distributed ($f(x) = (\pi(1 + x^2))^{-1}$), the mgf does not exist for any $t \neq 0$.
- The mgf of a normal random variable exists for all t.

Important property: if the mgf exists for t in an open interval containing zero, it uniquely determines the prob. distribution.

Why the Name MGF?

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2} + \ldots + \frac{t^rX^r}{r!} + \ldots$$

$$E(e^{tX}) = 1 + tE(X) + \frac{t^2E(X^2)}{2} + \ldots + \frac{t^rE(X^r)}{r!} + \ldots$$

Suppose that the mgf exists in an open interval containing 0. Then

$$M(0) = 1, M'(0) = E(X), \ldots, M^{(r)}(0) = E(X^r).$$

May be useful to compute all the moments of a distribution.

Example

Mgf of the Poisson distribution

$$\begin{aligned} M(t) &= \sum_{k\geq 0} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k\geq 0} e^{-\lambda} \frac{(e^t\lambda)^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^t}. \end{aligned}$$

This gives

$$M'(t) = \lambda e^t M(t), \quad M''(t) = \lambda e^t M(t) + \lambda^2 e^{2t} M(t).$$

It follows that if X is Poisson

$$E(X) = \lambda, \quad E(X^2) = \lambda^2 + \lambda \ \Rightarrow \text{Var}(X) = \lambda.$$

Mgf of Sums of Independent Random Variables

Very useful property: If X and Y are RV's with mgf's M_X and M_Y and $S = X + Y$, then

$$M_Z(t) = M_X(t)M_Y(t)$$

on the common interval where both mgf's exist.

Why?

$$M_S(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t).$$

Obvious extensions to sums of random variables.

Example

The sum of two independent Poisson rv's is a Poisson rv.

$X \sim$ Poi(λ) and $Y \sim$ Poi(μ), then $X + Y \sim$ Poi$(\lambda + \mu)$

$$M_{X+Y}(t) = e^{-\lambda} e^{\lambda e^t} e^{-\mu} e^{\mu e^t} = e^{-(\lambda+\mu)} e^{(\lambda+\mu)e^t}$$

but this is the mgf of a Poisson rv with parameter $\lambda + \mu$.