# Statistical Inference

## Lecture 4: Multiple Random Variables

MING GAO

DASE @ ECNU
(for course related communications)
mgao@dase.ecnu.edu.cn

Mar. 21, 2018

# Outline

Joint and Marginal Distributions

Conditional Distribution and Independence

Bivariate Transformations

Hierarchical Models and Mixture Distributions

Hierarchical Models and Mixture Distributions

Covariance and Correlation

Multivariate Distributions

Inequalities
- Numerical Inequalities
- Functional Inequalities

Take-aways

# Random vector

## Definition

An n-dimensional random vector is a function from a sample space $\Omega$ into $\mathcal{R}^n$, $n-$dimensional Euclidean space.

# Random vector

## Definition

An n-dimensional random vector is a function from a sample space $\Omega$ into $\mathcal{R}^n$, $n-$dimensional Euclidean space.

## Example

Consider the experiment of tossing two fair dices. Let

$X =$ sum of the two dices, and $Y = |$difference of the two dices$|$.

- For the sample point $(3,3)$, $X = 6$ and $Y = 0$.
- For the sample point $(4,1)$ or $(1,4)$, $X = 5$ and $Y = 3$.
- Since each of the 36 sample points in $\Omega$ is equally likely, thus

$$P(X = 5 \wedge Y = 3) = \frac{1}{18}.$$

# Joint PMF

## Definition

Let $(X, Y)$ be a discrete bivariate random vector. Then the function $f(x, y)$ from $\mathcal{R}^2$ into $\mathcal{R}$ defined by

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

is called the joint probability mass function or joint pmf of $(X, Y)$. If it is necessary, the notation $f_{X,Y}(x, y)$ will be used.

# Joint PMF

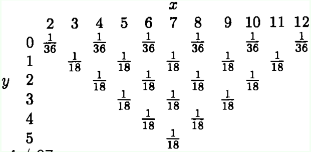
## Definition

Let $(X, Y)$ be a discrete bivariate random vector. Then the function $f(x, y)$ from $\mathcal{R}^2$ into $\mathcal{R}$ defined by

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

is called the joint probability mass function or joint pmf of $(X, Y)$. If it is necessary, the notation $f_{X,Y}(x, y)$ will be used.

## Example



| | | | | | | | $x$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| | 0 | $\frac{1}{36}$ | | $\frac{1}{36}$ | | $\frac{1}{36}$ | | $\frac{1}{36}$ | | $\frac{1}{36}$ | | $\frac{1}{36}$ |
| | 1 | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | $\frac{1}{18}$ | |
| $y$ | 2 | | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | |
| | 3 | | | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | | |
| | 4 | | | | | $\frac{1}{18}$ | | $\frac{1}{18}$ | | | | |
| | 5 | | | | | | $\frac{1}{18}$ | | | | | |

There are 21 possible values of $(X, Y)$. Two of these values, $f(5, 3) = \frac{1}{18}$ and $f(6, 0) = \frac{1}{36}$.

# Probability calculation

The joint pmf can be used to compute the probability of any event defined in terms of $(X, Y)$. Let $A$ be any subset of $\mathcal{R}^2$. Then

$$P((X, Y) \in A) = \sum_{(x,y)\in A} f(x, y).$$

# Probability calculation

The joint pmf can be used to compute the probability of any event defined in terms of $(X, Y)$. Let $A$ be any subset of $\mathcal{R}^2$. Then

$$P((X,Y) \in A) = \sum_{(x,y)\in A} f(x,y).$$

## Example

Let $A = \{(x,y)|x = 7 \text{ and } y \leq 4.\}$
Thus

$$P(A) = P(X = 7, Y \leq 4) = f(7,1) + f(7,3) = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

## Expectation

Expectations of functions of random vectors are computed just as with univariate r.v.s. Let $g(x, y)$ be a real-valued function defined for all possible values $(x, y)$ of the discrete random vector $(X, Y)$. Then $g(X, Y)$ is itself a random variable and its expected value $E(g(X, Y))$ is given by

$$E(g(X, Y)) = \sum_{(x,y)\in\mathcal{R}^2} g(x, y) f(x, y).$$

# Expectation

Expectations of functions of random vectors are computed just as with univariate r.v.s. Let $g(x, y)$ be a real-valued function defined for all possible values $(x, y)$ of the discrete random vector $(X, Y)$. Then $g(X, Y)$ is itself a random variable and its expected value $E(g(X, Y))$ is given by

$$E(g(X, Y)) = \sum_{(x,y)\in\mathcal{R}^2} g(x, y) f(x, y).$$

**Question:**

For the above given $(X, Y)$, what is the average value of $XY$?

# Expectation

Expectations of functions of random vectors are computed just as with univariate r.v.s. Let $g(x, y)$ be a real-valued function defined for all possible values $(x, y)$ of the discrete random vector $(X, Y)$. Then $g(X, Y)$ is itself a random variable and its expected value $E(g(X, Y))$ is given by

$$E(g(X,Y)) = \sum_{(x,y)\in\mathcal{R}^2} g(x,y)f(x,y).$$

**Question:**
For the above given $(X, Y)$, what is the average value of $XY$?
**Answer:** Letting $g(x, y) = xy$, we compute $E(XY) = E(g(X, Y))$. Thus,

$$E(XY) = 2 \times 0 \times \frac{1}{36} + \cdots + 7 \times 5 \times \frac{1}{18} = 13\frac{11}{18}.$$

# Properties of joint pmf

## Properties

- For any $(x, y)$, $f(x, y) \geq 0$ since $f(x, y)$ is a probability.
- Since $(X, Y)$ is certain to be in $\mathcal{R}^2$

$$\sum_{(x,y)\in\mathcal{R}^2} f(x, y) = P((X, Y) \in \mathcal{R}^2) = 1.$$

- It turns out that any nonnegative function from $\mathcal{R}^2$ to $\mathcal{R}$ that is nonzero for at most a countable number of $(x, y)$ pairs and sums to 1 is the joint pmf for some bivariate discrete random vector $(X, Y)$.

# Marginal pmf

Theorem

Let $(X, Y)$ be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then the marginal pmfs of $X$ and $Y$, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by

$$f_X(x) = \sum_{y \in \mathcal{R}} f_{X,Y}(x, y), f_Y(y) = \sum_{x \in \mathcal{R}} f_{X,Y}(x, y).$$

Proof.

For any $x \in \mathcal{R}$, let $A_x = \{(x, y) | y \in \mathcal{R}\}$. That is, $A_x$ is the line in the plane with first coordinate equal to $x$. Then, for any $x \in \mathcal{R}$:

$$\begin{aligned} f_X(x) &= P(X = x) = P(X = x, -\infty < Y < \infty) = P((X, Y) \in A_x) \\ &= \sum_{(x,y) \in A_x} f_{X,Y}(x, y) = \sum_{y \in \mathcal{R}} f_{X,Y}(x, y). \end{aligned}$$

## Example

Given the above joint pmf, we can compute the marginal pmf of $Y$.

$$\begin{aligned} f_Y(0) &= f_{X,Y}(2,0) + f_{X,Y}(4,0) + f_{X,Y}(6,0) \\ &\quad + f_{X,Y}(8,0) + f_{X,Y}(10,0) + f_{X,Y}(12,0) = \frac{1}{6} \end{aligned}$$

## Example

Given the above joint pmf, we can compute the marginal pmf of $Y$.

$$\begin{aligned} f_Y(0) = {} & f_{X,Y}(2,0) + f_{X,Y}(4,0) + f_{X,Y}(6,0) \\ & + f_{X,Y}(8,0) + f_{X,Y}(10,0) + f_{X,Y}(12,0) = \frac{1}{6} \end{aligned}$$

Similarly, we have $f_Y(1) = \frac{5}{18}$, $f_Y(2) = \frac{2}{9}$, $f_Y(3) = \frac{1}{6}$, $f_Y(4) = \frac{1}{9}$, and $f_Y(1) = \frac{1}{18}$.

# Example

Given the above joint pmf, we can compute the marginal pmf of $Y$.

$$\begin{aligned} f_Y(0) &= f_{X,Y}(2,0) + f_{X,Y}(4,0) + f_{X,Y}(6,0) \\ &\quad + f_{X,Y}(8,0) + f_{X,Y}(10,0) + f_{X,Y}(12,0) = \frac{1}{6} \end{aligned}$$

Similarly, we have $f_Y(1) = \frac{5}{18}$, $f_Y(2) = \frac{2}{9}$, $f_Y(3) = \frac{1}{6}$, $f_Y(4) = \frac{1}{9}$, and $f_Y(1) = \frac{1}{18}$.
Note that $\sum_{k=0}^{5} f_Y(k) = 1$, as it must, since these are the only six possible values of $Y$.

# Joint PDF

## Definition

A function $f(x,y)$ from $\mathcal{R}^2$ into $\mathcal{R}$ is called a joint probability density function or joint pdf of the continuous bivariate random vector $(X,Y)$ if, for every $A \in \mathcal{R}^2$

$$P((X,Y) \in A) = \int\int_A f(x,y)dxdy.$$

- If $g(x,y)$ be a real-valued function, then the expected values of $g(X,Y)$ is defined to be

$$E(g(X,Y)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)f(x,y)dxdy.$$

- The marginal probability density functions of $X$ and $Y$ are

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

# Example

Define a joint pdf by

$$f(x, y) = \begin{cases} 6xy^2, & 0 < x < 1 \text{ and } 0 < y < 1; \\ 0, & \text{otherwise.} \end{cases}$$

It is indeed a joint pdf, since

# Example

Define a joint pdf by

$$f(x,y) = \begin{cases} 6xy^2, & 0 < x < 1 \text{ and } 0 < y < 1; \\ 0, & \text{otherwise.} \end{cases}$$

It is indeed a joint pdf, since

- $f(x,y) \geq 0$ for all $(x,y)$ in the defined range;

# Example

Define a joint pdf by

$$f(x,y)=\begin{cases} 6xy^2, & 0<x<1 \text{ and } 0<y<1; \\ 0, & \text{otherwise.} \end{cases}$$

It is indeed a joint pdf, since

- $f(x,y)\geq 0$ for all $(x,y)$ in the defined range;
- $$\begin{aligned}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy &= \int_0^1\int_0^1 6xy^2dxdy \\ &= \int_0^1 3x^2y^2|_0^1dy = \int_0^1 3y^2dy = y^3|_0^1 = 1.\end{aligned}$$

# Calculating probability I

Now consider calculating a probability such as $P(X+Y \geq 1)$. Letting $A = \{(x,y) | x+y \geq 1\}$, i.e., $P((X,Y) \in A)$.

# Calculating probability I

Now consider calculating a probability such as $P(X + Y \geq 1)$. Letting $A = \{(x, y)|x + y \geq 1\}$, i.e., $P((X, Y) \in A)$.

$$\begin{aligned} A &= \{(x, y)|x + y \geq 1, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y)|x \geq 1 - y, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y)|1 - y \leq x < 1, 0 < y < 1\} \end{aligned}$$

# Calculating probability I

Now consider calculating a probability such as $P(X + Y \geq 1)$. Letting $A = \{(x, y)|x + y \geq 1\}$, i.e., $P((X, Y) \in A)$.

$$\begin{aligned} A &= \{(x, y)|x + y \geq 1, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y)|x \geq 1 - y, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y)|1 - y \leq x < 1, 0 < y < 1\} \end{aligned}$$

Thus,

$$P((X, Y) \in A) = \int\int_A f(x, y)dxdy = \int_0^1 \int_{1-y}^1 6xy^2 dxdy = \frac{9}{10}.$$

# Calculating marginal pdf

To calculate $f_X(x)$, we note that for $x \geq 1$ or $x \leq 0$, $f(x, y) = 0$. Thus for $x \geq 1$ or $x \leq 0$, we have

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = 0.$$

# Calculating marginal pdf

To calculate $f_X(x)$, we note that for $x \geq 1$ or $x \leq 0$, $f(x, y) = 0$. Thus for $x \geq 1$ or $x \leq 0$, we have

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = 0.$$

For $0 < x < 1$, we have

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_0^1 6xy^2 dy = 2xy^3|_0^1 = 2x.$$

## Calculating marginal pdf

To calculate $f_X(x)$, we note that for $x \geq 1$ or $x \leq 0$, $f(x, y) = 0$. Thus for $x \geq 1$ or $x \leq 0$, we have

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = 0.$$

For $0 < x < 1$, we have

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_0^1 6xy^2 dy = 2xy^3|_0^1 = 2x.$$

Similarly, we can calculate

$$f_X(x) = \begin{cases} 3y^2, & 0 < y < 1; \\ 0, & \text{otherwise.} \end{cases} .$$

# Calculating probability II

Let $f(x, y) = e^{-y}, 0 < x < y < \infty$, and $A = \{(x, y)|x+y \geq 1\}$.

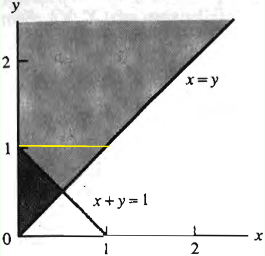
# Calculating probability II

Let $f(x, y) = e^{-y}, 0 < x < y < \infty$, and $A = \{(x, y)|x+y \geq 1\}$. Notice that region $A$ is an unbounded region with three sides given by the lines $y = x$, $x + y = 1$ and $x = 0$. To integrate over this region, we would have to break the region into at least two parts to write this appropriate limits of integration.

# Calculating probability II

Let $f(x,y) = e^{-y}, 0 < x < y < \infty$, and $A = \{(x,y)|x+y \geq 1\}$. Notice that region $A$ is an unbounded region with three sides given by the lines $y = x$, $x + y = 1$ and $x = 0$. To integrate over this region, we would have to break the region into at least two parts to write this appropriate limits of integration.

Thus $P((X,Y) \in A)$ can be calculated as



$$\begin{aligned} P(X+Y \geq 1) &= 1 - P(X+Y<1) \\ &= 1 - \int_0^{\frac{1}{2}} \int_x^{1-x} e^{-y} dydx \\ &= 1 - \int_0^{\frac{1}{2}} (e^{-x} - e^{-(1-x)})dx \\ &= 2e^{-\frac{1}{2}} - e^{-1} \end{aligned}$$

## Joint cdf

The joint probability distribution of $(X, Y)$ can be completely described with the joint cdf rather than with the joint pmf or joint pdf.

# Joint cdf

The joint probability distribution of $(X, Y)$ can be completely described with the joint cdf rather than with the joint pmf or joint pdf.

The joint cdf is the function $F(x, y)$ defined by

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

# Joint cdf

The joint probability distribution of $(X, Y)$ can be completely described with the joint cdf rather than with the joint pmf or joint pdf.

The joint cdf is the function $F(x, y)$ defined by

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

- The joint cdf is usually not very handy for discrete cases;

# Joint cdf

The joint probability distribution of $(X, Y)$ can be completely described with the joint cdf rather than with the joint pmf or joint pdf.

The joint cdf is the function $F(x, y)$ defined by

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

- The joint cdf is usually not very handy for discrete cases;
- For continuous bivariate random vector,

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

# Joint cdf

The joint probability distribution of $(X, Y)$ can be completely described with the joint cdf rather than with the joint pmf or joint pdf.

The joint cdf is the function $F(x, y)$ defined by

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

- The joint cdf is usually not very handy for discrete cases;
- For continuous bivariate random vector,

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

- $$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y).$$

# Conditional pmf

Let $(X, Y)$ be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any $x$ such that $P(X = x) = f_X(x) > 0$, the conditional pmf of $Y$ given that $X = x$ is the function of $y$ denoted by $f(y|x)$ and defined by

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

# Conditional pmf

Let $(X, Y)$ be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any $x$ such that $P(X = x) = f_X(x) > 0$, the conditional pmf of $Y$ given that $X = x$ is the function of $y$ denoted by $f(y|x)$ and defined by

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

For any $y$ such that $P(Y = y) = f_Y(y) > 0$, the conditional pmf of $X$ given that $Y = y$ is the function of $x$ denoted by $f(x|y)$ and defined by

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

## Example

Define the joint pmf of $(X, Y)$ by

$$f(0,10) = f(0,20) = \frac{2}{18}, f(1,10) = f(1,30) = \frac{3}{18}$$
$$f(1,20) = \frac{4}{18}, f(2,30) = \frac{4}{18}.$$

## Example

Define the joint pmf of $(X, Y)$ by

$$f(0,10) = f(0,20) = \frac{2}{18}, f(1,10) = f(1,30) = \frac{3}{18}$$

$$f(1,20) = \frac{4}{18}, f(2,30) = \frac{4}{18}.$$

First, the marginal pmf of $X$ is

$$f_X(0) = f(0,10) + f(0,20) = \frac{4}{18}, f_X(2) = f(2,30) = \frac{4}{18}$$

$$f_X(1) = f(1,10) + f(1,20) + f(1,30) = \frac{10}{18}$$

## Example

Define the joint pmf of $(X, Y)$ by

$$f(0,10) = f(0,20) = \frac{2}{18}, f(1,10) = f(1,30) = \frac{3}{18}$$
$$f(1,20) = \frac{4}{18}, f(2,30) = \frac{4}{18}.$$

First, the marginal pmf of $X$ is

$$f_X(0) = f(0,10) + f(0,20) = \frac{4}{18}, f_X(2) = f(2,30) = \frac{4}{18}$$
$$f_X(1) = f(1,10) + f(1,20) + f(1,30) = \frac{10}{18}$$

For $x = 0$,

$$f_X(10|0) = \frac{f(0,10)}{f_X(0)} = \frac{1}{2}, f_X(20|0) = \frac{f(0,20)}{f_X(0)} = \frac{1}{2}$$

# Example Cont'd

For $x = 1$,

$$f_X(10|1) = \frac{f(1,10)}{f_X(1)} = \frac{3}{10}$$
$$f_X(20|1) = \frac{f(1,20)}{f_X(1)} = \frac{4}{10}$$
$$f_X(30|1) = \frac{f(1,30)}{f_X(1)} = \frac{3}{10}$$

# Example Cont'd

For $x = 1$,

$$f_X(10|1) = \frac{f(1,10)}{f_X(1)} = \frac{3}{10}$$
$$f_X(20|1) = \frac{f(1,20)}{f_X(1)} = \frac{4}{10}$$
$$f_X(30|1) = \frac{f(1,30)}{f_X(1)} = \frac{3}{10}$$

For $x = 2$,

$$f_X(30|2) = \frac{f(2,30)}{f_X(2)} = 1$$

# Conditional pdf

Let $(X, Y)$ be a continuous bivariate random vector with joint pmf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any $x$ such that $P(X = x) = f_X(x) > 0$, the conditional pdf of $Y$ given that $X = x$ is the function of $y$ denoted by $f(y|x)$ and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

# Conditional pdf

Let $(X, Y)$ be a continuous bivariate random vector with joint pmf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any $x$ such that $P(X = x) = f_X(x) > 0$, the conditional pdf of $Y$ given that $X = x$ is the function of $y$ denoted by $f(y|x)$ and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

For any $y$ such that $P(Y = y) = f_Y(y) > 0$, the conditional pdf of $X$ given that $Y = y$ is the function of $x$ denoted by $f(x|y)$ and defined by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

# Calculating conditional pdf

Let $f(x,y) = e^{-y}, 0 < x < y < \infty$, and $A = \{(x,y)|x+y \geq 1\}$. We need to compute the conditional pdf of $Y$ given $X = x$.

## Calculating conditional pdf

Let $f(x,y) = e^{-y}, 0 < x < y < \infty$, and $A = \{(x,y)|x+y \geq 1\}$. We need to compute the conditional pdf of $Y$ given $X = x$. The marginal pdf of $X$ is computed as

- For $x \leq 0$, $f_X(x) = 0$ since $f(x,y) = 0$;
- For $x > 0$,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{x}^{\infty} e^{-y}dy = e^{-x}.$$

Thus, the conditional pdf of $Y$ given $X = x$ can be

- $f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)}$, if $y > x$;
- $f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{0}{e^{-x}} = 0$, if $y \leq x$;

## Conditional expectation

If $g(Y)$ is a function of $Y$, then the conditional expected value of $g(Y)$ given that $X = x$ is denoted by $E(g(Y|x))$ and is defined by

$$E(g(Y|x)) = \sum_y g(y)f(y|x)$$

$$E(g(Y|x)) = \int_{-\infty}^{\infty} g(y)f(y|x)dy$$

# Conditional expectation

If $g(Y)$ is a function of $Y$, then the conditional expected value of $g(Y)$ given that $X = x$ is denoted by $E(g(Y|x))$ and is defined by

$$E(g(Y|x)) = \sum_y g(y)f(y|x)$$

$$E(g(Y|x)) = \int_{-\infty}^{\infty} g(y)f(y|x)dy$$

- The conditional expected value has all of the properties of the usual expected value;
- $E(Y|x)$ provides the best guess at $Y$ based on knowledge of $X$.

# Calculating conditional expectation and variance

Given above example, the conditional expected value of $Y$ given $X = x$ can be calculated as

$$E(Y|X = x) = \int_{x}^{\infty} ye^{-(y-x)}dy = 1 + x.$$

## Calculating conditional expectation and variance

Given above example, the conditional expected value of $Y$ given $X = x$ can be calculated as

$$E(Y|X = x) = \int_x^\infty ye^{-(y-x)}dy = 1 + x.$$

The conditional variance can be computed as

$$\begin{aligned} Var(Y|X = x) &= E(Y^2|x) - (E(Y|x))^2 \\ &= \int_x^\infty y^2e^{-(y-x)}dy - (\int_x^\infty ye^{-(y-x)}dy)^2 = 1 \end{aligned}$$

# Calculating conditional expectation and variance

Given above example, the conditional expected value of $Y$ given $X = x$ can be calculated as

$$E(Y|X = x) = \int_x^\infty ye^{-(y-x)}dy = 1 + x.$$

The conditional variance can be computed as

$$\begin{aligned} Var(Y|X = x) &= E(Y^2|x) - (E(Y|x))^2 \\ &= \int_x^\infty y^2e^{-(y-x)}dy - (\int_x^\infty ye^{-(y-x)}dy)^2 = 1 \end{aligned}$$

Note that the marginal distribution of $Y$ is $gamma(2, 1)$, which has $Var(Y) = 2$. Given the knowledge that $X = x$, the variability in $Y$ is considerably reduced.

## Independent r.v.s

Let $(X, Y)$ be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then $X$ and $Y$ are called independent r.v.s if, for any $x \in \mathcal{R}$ and $y \in \mathcal{R}$

$$f(x, y) = f_X(x) f_Y(y).$$

# Independent r.v.s

Let $(X, Y)$ be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then $X$ and $Y$ are called independent r.v.s if, for any $x \in \mathcal{R}$ and $y \in \mathcal{R}$

$$f(x, y) = f_X(x) f_Y(y).$$

- If $X$ and $Y$ are independent, the conditional pdf of $Y$ given $X = x$ is

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y).$$

## Independent r.v.s

Let $(X, Y)$ be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then $X$ and $Y$ are called independent r.v.s if, for any $x \in \mathcal{R}$ and $y \in \mathcal{R}$

$$f(x, y) = f_X(x) f_Y(y).$$

- If $X$ and $Y$ are independent, the conditional pdf of $Y$ given $X = x$ is

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y).$$

- For any $A \subset \mathcal{R}$ and $x \in \mathcal{R}$,

$$P(Y \in A|x) = \int_A f(y|x) dy = \int_A f_Y(y) dy = P(Y \in A).$$

# Checking independent I

Define the joint pmf of $(X, Y)$ by

$$f(10,1) = f(20,1) = f(20,2) = \frac{1}{10}$$

$$f(10,2) = f(10,3) = \frac{1}{5}, f(20,3) = \frac{3}{10}.$$

# Checking independent I

Define the joint pmf of $(X, Y)$ by

$$f(10,1) = f(20,1) = f(20,2) = \frac{1}{10}$$
$$f(10,2) = f(10,3) = \frac{1}{5}, f(20,3) = \frac{3}{10}.$$

The marginal pmfs are

$$f_X(10) = f_X(20) = \frac{1}{2}$$
$$f_Y(1) = \frac{1}{5}, f_Y(2) = \frac{3}{10}, f_Y(3) = \frac{1}{2}$$

## Checking independent I

Define the joint pmf of $(X, Y)$ by

$$f(10,1) = f(20,1) = f(20,2) = \frac{1}{10}$$
$$f(10,2) = f(10,3) = \frac{1}{5}, f(20,3) = \frac{3}{10}.$$

The marginal pmfs are

$$f_X(10) = f_X(20) = \frac{1}{2}$$
$$f_Y(1) = \frac{1}{5}, f_Y(2) = \frac{3}{10}, f_Y(3) = \frac{1}{2}$$

Thus, the r.v.s $X$ and $Y$ are not independent since

$$f(10,3) = \frac{1}{5} \neq \frac{1}{2}\frac{1}{2} = f_X(10)f_Y(3).$$

# Lemma for independent r.v.s

Let $(X, Y)$ be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then $X$ and $Y$ are independent r.v.s if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathcal{R}$ and $y \in \mathcal{R}$

$$f(x, y) = g(x)h(y).$$

Proof.

## Lemma for independent r.v.s

Let $(X, Y)$ be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then $X$ and $Y$ are independent r.v.s if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathcal{R}$ and $y \in \mathcal{R}$

$$f(x, y) = g(x)h(y).$$

Proof.

$\Rightarrow$: Easily to prove based on the definition.

# Lemma for independent r.v.s

Let $(X, Y)$ be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then $X$ and $Y$ are independent r.v.s if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathcal{R}$ and $y \in \mathcal{R}$

$$f(x, y) = g(x)h(y).$$

Proof.

$\Rightarrow$: Easily to prove based on the definition.

$\Leftarrow$: Let $f(x, y) = g(x)h(y)$. We define

$$\int_{-\infty}^{\infty} g(x)dx = c$$
$$\int_{-\infty}^{\infty} h(y)dy = d$$

□

# Proof Cont'd

$$cd = (\int_{-\infty}^{\infty} g(x)dx)(\int_{-\infty}^{\infty} h(y)dy) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)dxdy$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

Furthermore, the marginal pdfs are given by

$$f_X(x) = \int_{-\infty}^{\infty} g(x)h(y)dy = dg(x), f_Y(y) = \int_{-\infty}^{\infty} g(x)h(y)dx = ch(y)$$

Thus we have

$$f(x,y) = g(x)h(y) = g(x)h(y)cd = f_X(x)f_Y(y)$$

That is $X$ and $Y$ are independent.

# Checking independent II

Consider the joint pdf $f(x,y) = \frac{1}{384}x^2y^4e^{-y-\frac{x}{2}}, x > 0$ and $y > 0$.

**Question:** Please confirm whether r.v.s $X$ and $Y$ are independent.

# Checking independent II

Consider the joint pdf $f(x,y) = \frac{1}{384}x^2y^4e^{-y-\frac{x}{2}}, x > 0$ and $y > 0$.

**Question:** Please confirm whether r.v.s $X$ and $Y$ are independent.

**Answer:** If we define

$$g(x) = \begin{cases} x^2e^{-\frac{1}{2}}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

$$h(y) = \begin{cases} \frac{1}{384}y^4e^{-y}, & y > 0; \\ 0, & \text{otherwise.} \end{cases}$$

# Checking independent II

Consider the joint pdf $f(x,y) = \frac{1}{384}x^2y^4e^{-y-\frac{x}{2}}, x > 0$ and $y > 0$.

**Question:** Please confirm whether r.v.s $X$ and $Y$ are independent.

**Answer:** If we define

$$g(x) = \begin{cases} x^2e^{-\frac{1}{2}}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

$$h(y) = \begin{cases} \frac{1}{384}y^4e^{-y}, & y > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Then $f(x,y) = g(x)h(y)$ for all $x \in \mathcal{R}$ and $y \in \mathcal{R}$. In terms of the lemma, we conclude that $X$ and $Y$ are independent r.v.s.

# Checking independent II

Consider the joint pdf $f(x,y) = \frac{1}{384}x^2y^4e^{-y-\frac{x}{2}}, x > 0$ and $y > 0$.

**Question:** Please confirm whether r.v.s $X$ and $Y$ are independent.

**Answer:** If we define

$$g(x) = \begin{cases} x^2e^{-\frac{1}{2}}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

$$h(y) = \begin{cases} \frac{1}{384}y^4e^{-y}, & y > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Then $f(x,y) = g(x)h(y)$ for all $x \in \mathcal{R}$ and $y \in \mathcal{R}$. In terms of the lemma, we conclude that $X$ and $Y$ are independent r.v.s. Note that we no not have to compute the marginal pdfs.

## Theorem for independent r.v.s

Let $X$ and $Y$ are independent r.v.s

- For any $A \subset \mathcal{R}$ and $B \subset \mathcal{R}$,

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

  i.e., the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events;

# Theorem for independent r.v.s

Let $X$ and $Y$ are independent r.v.s

- For any $A \subset \mathcal{R}$ and $B \subset \mathcal{R}$,

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

  i.e., the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events;
- Let $g(x)$ and $h(y)$ be functions only of $x$ and $y$, respectively, then

  $$E(g(X)h(Y)) = (E(g(X)))(E(h(Y))).$$

# Theorem for independent r.v.s

Let $X$ and $Y$ are independent r.v.s

- For any $A \subset \mathcal{R}$ and $B \subset \mathcal{R}$,

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

  i.e., the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events;

- Let $g(x)$ and $h(y)$ be functions only of $x$ and $y$, respectively, then

  $$E(g(X)h(Y)) = (E(g(X)))(E(h(Y))).$$

- The moment generating function of the r.v. $Z = X + Y$ is given by

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

## Expectation of independent r.v.s

Let $X$ and $Y$ are independent *exponential*$(1)$ r.v.s

# Expectation of independent r.v.s

Let $X$ and $Y$ are independent *exponential*(1) r.v.s

- $$\begin{aligned} P(X \geq 4, Y < 3) &= P(X \geq 4)P(Y < 3) \\ &= e^{-4}(1 - e^{-3}) \end{aligned}$$

# Expectation of independent r.v.s

Let $X$ and $Y$ are independent *exponential*(1) r.v.s

- $$\begin{aligned} P(X \geq 4, Y < 3) &= P(X \geq 4)P(Y < 3) \\ &= e^{-4}(1 - e^{-3}) \end{aligned}$$

- Letting $g(x) = x^2$ and $h(y) = y$, we see that

$$\begin{aligned} E(X^2 Y) &= (E(X^2))(E(Y)) \\ &= (Var(X) + (E(X))^2)E(Y) \\ &= (1 + 1^2)1 = 2. \end{aligned}$$

# MGF of a sum of normal variables

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent r.v.s. Then, the mgfs of $X$ and $Y$ are

$$M_X(t) = exp^{\mu_1 t + \sigma_1^2 t^2/2}$$
$$M_Y(t) = exp^{\mu_2 t + \sigma_2^2 t^2/2}$$

In terms of the theorem, the mgf of $Z = X + Y$ is

$$M_Z(t) = M_X(t)M_Y(t) = exp^{(\mu_1+\mu_2)t+(\sigma_1^2+\sigma_2^2)t^2/2}.$$

## MGF of a sum of normal variables

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent r.v.s. Then, the mgfs of $X$ and $Y$ are

$$M_X(t) = exp^{\mu_1 t + \sigma_1^2 t^2/2}$$
$$M_Y(t) = exp^{\mu_2 t + \sigma_2^2 t^2/2}$$

In terms of the theorem, the mgf of $Z = X + Y$ is

$$M_Z(t) = M_X(t)M_Y(t) = exp^{(\mu_1+\mu_2)t+(\sigma_1^2+\sigma_2^2)t^2/2}.$$

### Theorem

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent r.v.s. Then, the r.v. $Z = X+Y$ has a $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ distribution.

# Distribution of bivariate function

Let $(X, Y)$ be a bivariate random vector with a known probability distribution. Now consider a new bivariate random vector $(U, V)$ defined by $U = g_1(X, Y)$ and $V = g_2(X, Y)$, where $g_i(x, y)$ is some specified function.

# Distribution of bivariate function

Let $(X, Y)$ be a bivariate random vector with a known probability distribution. Now consider a new bivariate random vector $(U, V)$ defined by $U = g_1(X, Y)$ and $V = g_2(X, Y)$, where $g_i(x, y)$ is some specified function.
If $B \subset \mathcal{R}^2$ if and only if $(X, Y) \in A$, where

$$A = \{(x, y)|(g_1(x, y), g_2(x, y)) \in B\}.$$

# Distribution of bivariate function

Let $(X, Y)$ be a bivariate random vector with a known probability distribution. Now consider a new bivariate random vector $(U, V)$ defined by $U = g_1(X, Y)$ and $V = g_2(X, Y)$, where $g_i(x, y)$ is some specified function.
If $B \subset \mathcal{R}^2$ if and only if $(X, Y) \in A$, where

$$A = \{(x, y) | (g_1(x, y), g_2(x, y)) \in B\}.$$

Thus

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

i.e., the probability distribution of $(U, V)$ is completely determined by the probability distribution of $(X, Y)$.

# Transformation of discrete r.v.s

If $(X, Y)$ is discrete bivariate random vector, then there is only a countable set of values for which the joint pmf of $(X, y)$ is positive. Call this set $A$.

# Transformation of discrete r.v.s

If $(X, Y)$ is discrete bivariate random vector, then there is only a countable set of values for which the joint pmf of $(X, y)$ is positive. Call this set $A$.

Define the set

$$B = \{(u, v) | u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in A\}.$$

# Transformation of discrete r.v.s

If $(X, Y)$ is discrete bivariate random vector, then there is only a countable set of values for which the joint pmf of $(X, y)$ is positive. Call this set $A$.
Define the set

$$B = \{(u, v) | u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in A\}.$$

Then $B$ is the countable set of possible values for the discrete random vector $(U, V)$. And if, for any $(u, v) \in B$, we define

$$A_{uv} = \{(x, y) \in A | u = g_1(x, y) \text{ and } v = g_2(x, y)\}.$$

# Transformation of discrete r.v.s

If $(X, Y)$ is discrete bivariate random vector, then there is only a countable set of values for which the joint pmf of $(X, y)$ is positive. Call this set $A$.

Define the set

$$B = \{(u, v) | u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in A\}.$$

Then $B$ is the countable set of possible values for the discrete random vector $(U, V)$. And if, for any $(u, v) \in B$, we define

$$A_{uv} = \{(x, y) \in A | u = g_1(x, y) \text{ and } v = g_2(x, y)\}.$$

Then the joint pmf of $(U, V)$ can be computed as

$$\begin{aligned} f_{U,V}(u, v) &= P(U = u, V = v) = P((X, Y) \in A_{uv}) \\ &= \sum_{(x,y) \in A_{uv}} f_{X,Y}(x, y). \end{aligned}$$

# Distribution of the sum of Poisson variables

Let $X$ and $Y$ are independent Poisson r.v.s with parameters $\theta_1$ and $\theta_2$, respectively. Thus the joint pmf of $(X, Y)$ is

$$f_{X,Y}(x, y) = \frac{\theta_1^x e^{-\theta_1}}{x!} \frac{\theta_2^y e^{-\theta_2}}{y!}, x \in N, y \in N$$

## Distribution of the sum of Poisson variables

Let $X$ and $Y$ are independent Poisson r.v.s with parameters $\theta_1$ and $\theta_2$, respectively. Thus the joint pmf of $(X, Y)$ is

$$f_{X,Y}(x,y) = \frac{\theta_1^x e^{-\theta_1}}{x!} \frac{\theta_2^y e^{-\theta_2}}{y!}, x \in N, y \in N$$

Now define $U = X + Y$ and $V = Y$. That is, $g_1(x,y) = x + y$ and $g_2(x,y) = y$. Thus,

$$\begin{aligned} A &= \{(x,y) | x \in N, y \in N\} \\ B &= \{(u,v) | v \in N, u \geq v, u \in N\}. \end{aligned}$$

# Distribution of the sum of Poisson variables

Let $X$ and $Y$ are independent Poission r.v.s with parameters $\theta_1$ and $\theta_2$, respectively. Thus the joint pmf of $(X, Y)$ is

$$f_{X,Y}(x,y) = \frac{\theta_1^x e^{-\theta_1}}{x!}\frac{\theta_2^y e^{-\theta_2}}{y!}, x \in N, y \in N$$

Now define $U = X + Y$ and $V = Y$. That is, $g_1(x,y) = x + y$ and $g_2(x,y) = y$. Thus,

$$\begin{aligned} A &= \{(x,y)|x \in N, y \in N\} \\ B &= \{(u,v)|v \in N, u \geq v, u \in N\}. \end{aligned}$$

$$f_{U,V}(u,v) = f_{X,Y}(u-v,v) = \frac{\theta_1^{u-v} e^{-\theta_1}}{(u-v)!}\frac{\theta_2^v e^{-\theta_2}}{v!}$$

# Distribution of the sum of Poisson variables Cont'd

In this example it is interesting to compute the marginal pmf of $U$. Thus

$$
\begin{aligned}
f_U(u) &= \sum_{v=0}^{u} \frac{\theta_1^{u-v} e^{-\theta_1}}{(u-v)!} \frac{\theta_2^{v} e^{-\theta_2}}{v!} \\
&= e^{-(\theta_1+\theta_2)} \sum_{v=0}^{u} \frac{\theta_1^{u-v}}{(u-v)!} \frac{\theta_2^{v}}{v!} \\
&= \frac{e^{-(\theta_1+\theta_2)}}{u!} \sum_{v=0}^{u} \binom{u}{v} \theta_1^{u-v} \theta_2^{v} \\
&= \frac{e^{-(\theta_1+\theta_2)}}{u!} (\theta_1+\theta_2)^u \\
&= \frac{(\theta_1+\theta_2)^u}{u!} e^{-(\theta_1+\theta_2)}
\end{aligned}
$$

# Distribution of the sum of Poisson variables Cont'd

In this example it is interesting to compute the marginal pmf of $U$. Thus

$$\begin{aligned} f_U(u) &= \sum_{v=0}^{u} \frac{\theta_1^{u-v} e^{-\theta_1}}{(u-v)!} \frac{\theta_2^{v} e^{-\theta_2}}{v!} \\ &= e^{-(\theta_1+\theta_2)} \sum_{v=0}^{u} \frac{\theta_1^{u-v}}{(u-v)!} \frac{\theta_2^{v}}{v!} \\ &= \frac{e^{-(\theta_1+\theta_2)}}{u!} \sum_{v=0}^{u} \binom{u}{v} \theta_1^{u-v} \theta_2^{v} \\ &= \frac{e^{-(\theta_1+\theta_2)}}{u!} (\theta_1+\theta_2)^u \\ &= \frac{(\theta_1+\theta_2)^u}{u!} e^{-(\theta_1+\theta_2)} \end{aligned}$$

**Theorem**

Let $X$ and $Y$ are independent Poission r.v.s with parameters $\theta_1$ and $\theta_2$, respectively. Thus $X + Y \sim Poisson(\theta_1 + \theta_2)$.

# Transformation of continuous r.v.s

If $(X, Y)$ is a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y)$, then the joint pdf of $U, V$ can be expressed in terms of $f_{X,Y}(x, y)$.

# Transformation of continuous r.v.s

If $(X, Y)$ is a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y)$, then the joint pdf of $U, V$ can be expressed in terms of $f_{X,Y}(x, y)$.
Define the sets

$$A = \{(x, y) | f_{X,Y}(x, y) > 0\}$$
$$B = \{(u, v) | u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in A\}$$

# Transformation of continuous r.v.s

If $(X, Y)$ is a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y)$, then the joint pdf of $U, V$ can be expressed in terms of $f_{X,Y}(x, y)$.
Define the sets

$$A = \{(x, y) | f_{X,Y}(x, y) > 0\}$$
$$B = \{(u, v) | u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in A\}$$

For the simplest version of this result we assume that the transformation $u = g_1(x, y)$ and $v = g_2(x, y)$ defines a one-to-one transformation of $A$ onto $B$.

# Transformation of continuous r.v.s

If $(X, Y)$ is a continuous bivariate random vector with joint pdf $f_{X,Y}(x, y)$, then the joint pdf of $U, V$ can be expressed in terms of $f_{X,Y}(x, y)$.
Define the sets

$$\begin{aligned} A &= \{(x, y) | f_{X,Y}(x, y) > 0\} \\ B &= \{(u, v) | u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in A\} \end{aligned}$$

For the simplest version of this result we assume that the transformation $u = g_1(x, y)$ and $v = g_2(x, y)$ defines a one-to-one transformation of $A$ onto $B$.
For such a one-to-one, onto transformation, we can obtain a reverse transformation by $x = h_1(u, v)$ and $y = h_2(u, v)$. The role played by a derivative in the univariate case is now played by a quantity called the Jacobian of the transformation.

# Transformation of continuous r.v.s

We further define the Jacobian determinant of the transformation as

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u},$$

where $\frac{\partial x}{\partial u} = \frac{\partial h_1(u,v)}{\partial u}, \frac{\partial y}{\partial v} = \frac{\partial h_2(u,v)}{\partial v}, \frac{\partial x}{\partial v} = \frac{\partial h_1(u,v)}{\partial v}, \frac{\partial y}{\partial u} = \frac{\partial h_2(u,v)}{\partial u}$. The joint pdf of $(U, V)$ is 0 outside the set $B$ and on the set $B$ is given by

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v))|J|,$$

where $|J|$ is the absolute value of $J$.

# Transformation of continuous r.v.s

We further define the Jacobian determinant of the transformation as

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u},$$

where $\frac{\partial x}{\partial u} = \frac{\partial h_1(u,v)}{\partial u}, \frac{\partial y}{\partial v} = \frac{\partial h_2(u,v)}{\partial v}, \frac{\partial x}{\partial v} = \frac{\partial h_1(u,v)}{\partial v}, \frac{\partial y}{\partial u} = \frac{\partial h_2(u,v)}{\partial u}$.
The joint pdf of $(U, V)$ is 0 outside the set $B$ and on the set $B$ is given by

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v))|J|,$$

where $|J|$ is the absolute value of $J$.
Note that it is sometimes just as difficult to determine the set $B$ and verify that the transformation is one-to-one as it is to substitute into the formula.

# Sum and difference of normal variables

Let $X$ and $Y$ are independent, standard normal r.v.s. Consider the transformation $U = X + Y$ and $V = X - Y$, thus we have

$$g_1(x, y) = x + y, g_2(x, y) = x - y$$

$$h_1(u, v) = \frac{u + v}{2}, h_2(u, v) = \frac{u - v}{2}.$$

## Sum and difference of normal variables

Let $X$ and $Y$ are independent, standard normal r.v.s. Consider the transformation $U = X + Y$ and $V = X - Y$, thus we have

$$g_1(x,y) = x + y, g_2(x,y) = x - y$$
$$h_1(u,v) = \frac{u+v}{2}, h_2(u,v) = \frac{u-v}{2}.$$

Furthermore,

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

## Sum and difference of normal variables

Let $X$ and $Y$ are independent, standard normal r.v.s. Consider the transformation $U = X + Y$ and $V = X - Y$, thus we have

$$g_1(x,y) = x + y, g_2(x,y) = x - y$$
$$h_1(u,v) = \frac{u+v}{2}, h_2(u,v) = \frac{u-v}{2}.$$

Furthermore,

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

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v))|J|$$
$$= \frac{1}{4\pi} e^{-((u+v)/2)^2} e^{-((u-v)/2)^2} = (\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-u^2/4})(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-v^2/4})$$

# Analysis

- The joint pdf has factored into a function of $u$ and a function of $v$. By the above lemma, $U$ and $V$ are independent.
- $U \sim N(0, 2)$ and $V \sim N(0, 2)$.
- This important fact, that sums and differences of independent normal r.v.s are independent normal r.v.s, is true regardless of the means of $X$ and $Y$, so long as $Var(X) = Var(Y)$.

# Analysis

- The joint pdf has factored into a function of $u$ and a function of $v$. By the above lemma, $U$ and $V$ are independent.
- $U \sim N(0, 2)$ and $V \sim N(0, 2)$.
- This important fact, that sums and differences of independent normal r.v.s are independent normal r.v.s, is true regardless of the means of $X$ and $Y$, so long as $Var(X) = Var(Y)$.

## Theorem

Let $X$ and $Y$ be independent r.v.s. Let $g(x)$ be a function only of $x$ and $h(y)$ be a function only of $y$. Then the r.v.s $U = g(X)$ and $V = h(Y)$ are independent.

# Distribution of the ratio of normal variables

Let $X$ and $Y$ be independent $N(0,1)$ r.v.s. Consider the transformation $U = \frac{X}{Y}$ and $V = |Y|$.

# Distribution of the ratio of normal variables

Let $X$ and $Y$ be independent $N(0,1)$ r.v.s. Consider the transformation $U = \frac{X}{Y}$ and $V = |Y|$.

Note that this transformation is not one-to-one since the points $(x, y)$ and $(-x, -y)$ are both mapped into the same $(u, v)$ point.

# Distribution of the ratio of normal variables

Let $X$ and $Y$ be independent $N(0,1)$ r.v.s. Consider the transformation $U = \frac{X}{Y}$ and $V = |Y|$.
Note that this transformation is not one-to-one since the points $(x,y)$ and $(-x,-y)$ are both mapped into the same $(u,v)$ point.
Let

$$A_1 = \{(x,y) : y > 0\}, A_2 = \{(x,y) : y < 0\}, A_0 = \{(x,y) : y = 0\}.$$

# Distribution of the ratio of normal variables

Let $X$ and $Y$ be independent $N(0,1)$ r.v.s. Consider the transformation $U = \frac{X}{Y}$ and $V = |Y|$.

Note that this transformation is not one-to-one since the points $(x, y)$ and $(-x, -y)$ are both mapped into the same $(u, v)$ point.

Let

$$A_1 = \{(x, y) : y > 0\}, A_2 = \{(x, y) : y < 0\}, A_0 = \{(x, y) : y = 0\}.$$

Thus, $B = \{(u, v) : v > 0\}$ is the image of both $A_1$ and $A_2$ under the transformation.

# Distribution of the ratio of normal variables

Let $X$ and $Y$ be independent $N(0,1)$ r.v.s. Consider the transformation $U = \frac{X}{Y}$ and $V = |Y|$.
Note that this transformation is not one-to-one since the points $(x, y)$ and $(-x, -y)$ are both mapped into the same $(u, v)$ point.
Let

$$A_1 = \{(x, y) : y > 0\}, A_2 = \{(x, y) : y < 0\}, A_0 = \{(x, y) : y = 0\}.$$

Thus, $B = \{(u, v) : v > 0\}$ is the image of both $A_1$ and $A_2$ under the transformation.
The inverse transformation from $B$ to $A_1$ and $B$ to $A_2$ are given by

$$x = h_{11}(u, v) = uv, y = h_{21}(u, v) = v$$
$$x = h_{12}(u, v) = -uv, y = h_{22}(u, v) = -v$$

## Distribution of the ratio of normal variables Cont'd

Note that the Jacobians from the two inverses are $J_1 = J_2 = v$, and the joint pdf is

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

# Distribution of the ratio of normal variables Cont'd

Note that the Jacobians from the two inverses are $J_1 = J_2 = v$, and the joint pdf is

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

Thus, we obtain

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi} e^{-\frac{(uv)^2}{2}} e^{-\frac{v^2}{2}} |v| + \frac{1}{2\pi} e^{-\frac{(-uv)^2}{2}} e^{-\frac{(-v)^2}{2}} |v| \\ &= \frac{v}{\pi} e^{-\frac{(u^2+1)^2 v^2}{2}}, -\infty < u, v < \infty \end{aligned}$$

# Distribution of the ratio of normal variables Cont'd

Note that the Jacobians from the two inverses are $J_1 = J_2 = v$, and the joint pdf is

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

Thus, we obtain

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi} e^{-\frac{(uv)^2}{2}} e^{-\frac{v^2}{2}} |v| + \frac{1}{2\pi} e^{-\frac{(-uv)^2}{2}} e^{-\frac{(-v)^2}{2}} |v| \\ &= \frac{v}{\pi} e^{-\frac{(u^2+1)^2 v^2}{2}}, -\infty < u, v < \infty \end{aligned}$$

From this the marginal pdf of $U$ can be computed to be

$$f_U(u) = \int_0^\infty \frac{v}{\pi} e^{-\frac{(u^2+1)^2 v^2}{2}} dv = \frac{1}{2\pi} \int_0^\infty e^{-\frac{(u^2+1)^2 z}{2}} dz = \frac{1}{\pi(u^2+1)}.$$

So we see that the ratio of two independent standard normal r.v.s is a Cauchy r.v.

# Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?

# Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?

The "large number" of eggs laid is a r.v., often taken to be $Poisson(\lambda)$. Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials.

# Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?

The “large number” of eggs laid is a r.v., often taken to be $\textit{Poisson}(\lambda)$. Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials. Let

$$X = \text{ number of survivors}, Y = \text{ number of eggs laid},$$

Thus, we have a hierarchical model as

$$\begin{aligned} X|Y &\sim \textit{Binomial}(Y, p), \\ Y &\sim \textit{Poisson}(\lambda). \end{aligned}$$

# Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?
The "large number" of eggs laid is a r.v., often taken to be $\mathit{Poisson}(\lambda)$. Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials. Let

$$X = \text{ number of survivors}, Y = \text{ number of eggs laid},$$

Thus, we have a hierarchical model as

$$X|Y \sim \mathit{Binomial}(Y, p),$$
$$Y \sim \mathit{Poisson}(\lambda).$$

Recall that we use notation such as $X|Y \sim \mathit{Binomial}(Y, p)$ to mean that the conditional distribution of $X$ given $Y = y$ is $\mathit{Binomial}(y, p)$.

# Binomial-Poisson hierarchy Cont'd

$$
\begin{aligned}
P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) = \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \\
&= \sum_{y=0}^{\infty} \left[\binom{y}{x} p^x (1-p)^{(y-x)}\right]\left[\frac{\lambda^y e^{-\lambda}}{y!}\right] \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{(y-x)}}{(y-x)!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} = \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} = \frac{(\lambda p)^x}{x!} e^{\lambda p}.
\end{aligned}
$$

# Binomial-Poisson hierarchy Cont'd

$$
\begin{aligned}
P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) = \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \\
&= \sum_{y=0}^{\infty} \left[\binom{y}{x} p^x (1-p)^{(y-x)}\right]\left[\frac{\lambda^y e^{-\lambda}}{y!}\right] \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{(y-x)}}{(y-x)!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} = \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} = \frac{(\lambda p)^x}{x!} e^{\lambda p}.
\end{aligned}
$$

Thus, any marginal inference on $X$ is with respect to a $Poisson(\lambda p)$ distribution, with $Y$ playing not part at all.

# Binomial-Poisson hierarchy Cont'd

$$
\begin{aligned}
P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) = \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \\
&= \sum_{y=0}^{\infty} \left[\binom{y}{x} p^x (1-p)^{(y-x)}\right]\left[\frac{\lambda^y e^{-\lambda}}{y!}\right] \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{(y-x)}}{(y-x)!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^{t}}{t!} = \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} = \frac{(\lambda p)^x}{x!} e^{\lambda p}.
\end{aligned}
$$

Thus, any marginal inference on $X$ is with respect to a $Poisson(\lambda p)$ distribution, with $Y$ playing not part at all.

The answer to the original question is now easy to compute $E(X) = \lambda p$.

# Theorem for expectation of conditional expectation

If $X$ and $Y$ are any two r.v.s, then

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

provided that the expectations exist.

Proof.

Let $f(x, y)$ denote that joint pdf of $X$ and $Y$. By definition, we have

$$E(X) = \int\int xf(x,y)dxdy = \int\Big[\int xf(x|y)dx\Big]f_Y(y)dy.$$

Thus, we have

$$E(X) = \int E(X|y)f_Y(y)dy = E(E(X|Y)).$$

Replace integrals by sums to prove the discrete case. □

# Mixture distribution

From the above theorem, we can easily compute the expected number of survivors

$$E(X) = E(E(X|Y)) = E(pY) = p\lambda.$$

# Mixture distribution

From the above theorem, we can easily compute the expected number of survivors

$$E(X) = E(E(X|Y)) = E(pY) = p\lambda.$$

Definition

A r.v. $X$ is said to have a mixture distribution if the distribution of $X$ depends on quantity that also has a distribution.

# Mixture distribution

From the above theorem, we can easily compute the expected number of survivors

$$E(X) = E(E(X|Y)) = E(pY) = p\lambda.$$

Definition

A r.v. $X$ is said to have a mixture distribution if the distribution of $X$ depends on quantity that also has a distribution.

In the above example, the $Poisson(\lambda p)$ distribution is a mixture distribution since it is the result of combining a $Binomial(Y, p)$ with $Y \sim Poisson(\lambda)$.

# Mixture distribution

From the above theorem, we can easily compute the expected number of survivors

$$E(X) = E(E(X|Y)) = E(pY) = p\lambda.$$

### Definition

A r.v. $X$ is said to have a mixture distribution if the distribution of $X$ depends on quantity that also has a distribution.

In the above example, the $Poisson(\lambda p)$ distribution is a mixture distribution since it is the result of combining a $Binomial(Y, p)$ with $Y \sim Poisson(\lambda)$.

In general, we can say that hierarchical models lead to mixture distributions.

## Example generalization

Instead of one mother insect, there are a large number of mothers and one mother is chosen at random. We are still interested in knowing the average number of survivors, but is is no longer clear that the number of eggs laid follows the same Poisson distribution for each mother.

# Example generalization

Instead of one mother insect, there are a large number of mothers and one mother is chosen at random. We are still interested in knowing the average number of survivors, but is is no longer clear that the number of eggs laid follows the same Poisson distribution for each mother.

The following three-stage hierarchy may be more appropriate. Let

$$X = \text{ number of survivors}, X \sim \mathit{binomial}(Y, p)$$
$$Y|\Lambda \sim \mathit{Poisson}(\Lambda), \Lambda \sim \mathit{exponential}(\beta),$$

Thus, the expectation of $X$ can easily be calculated as

$$E(X) = E(E(X|Y)) = E(pY) = E(E(pY|\Lambda)) = E(p\Lambda) = p\beta.$$

# Rethinking the three-stage model

Note that this three-stage model can also be thought of as a two-stage hierarchy by combining the last two stages.

# Rethinking the three-stage model

Note that this three-stage model can also be thought of as a two-stage hierarchy by combining the last two stages. If $Y|\Lambda \sim$ *Poisson*$(\Lambda)$ and $\Lambda \sim$ *exponential*$(\beta)$, then

$$
\begin{aligned}
&P(Y=y) = P(Y=y, 0<\Lambda<\infty) = \int_0^\infty f(y,\lambda)d\lambda \\
&= \int_0^\infty f(y|\lambda)f(\lambda)d\lambda = \int_0^\infty \left[\frac{e^{-\lambda}\lambda^y}{y!}\right]\frac{1}{\beta}e^{-\frac{\lambda}{\beta}}d\lambda \\
&= \frac{1}{\beta y!}\int_0^\infty \lambda^y e^{-\lambda(1+\beta^{-1})}d\lambda = \frac{1}{\beta y!}\Gamma(y+1)\Big(\frac{1}{1+\beta^{-1}}\Big)^{y+1} \\
&= \frac{1}{1+\beta}\Big(\frac{1}{1+\beta^{-1}}\Big)^{y+1}.
\end{aligned}
$$

It forms a negative binomial pmf. Therefore, our three-stage hierarchy is equivalent to the two-stage hierarchy

## Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?

# Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?

The "large number" of eggs laid is a r.v., often taken to be $\mathit{Poisson}(\lambda)$. Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials.

## Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?
The "large number" of eggs laid is a r.v., often taken to be *Poisson*$(\lambda)$. Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials. Let

$$X = \text{ number of survivors}, Y = \text{ number of eggs laid},$$

Thus, we have a hierarchical model as

$$\begin{gathered} X|Y \sim \textit{Binomial}(Y, p), \\ Y \sim \textit{Poisson}(\lambda). \end{gathered}$$

## Binomial-Poisson hierarchy

An insect lays a large number of eggs, each surviving with probability $p$. On the average, how many eggs will survive?

The "large number" of eggs laid is a r.v., often taken to be $\textit{Poisson}(\lambda)$. Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials. Let

$$X = \text{ number of survivors}, Y = \text{ number of eggs laid},$$

Thus, we have a hierarchical model as

$$X|Y \sim \textit{Binomial}(Y,p),$$
$$Y \sim \textit{Poisson}(\lambda).$$

Recall that we use notation such as $X|Y \sim \textit{Binomial}(Y,p)$ to mean that the conditional distribution of $X$ given $Y = y$ is $\textit{Binomial}(y,p)$.

## Binomial-Poisson hierarchy Cont'd

$$
\begin{aligned}
P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) = \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \\
&= \sum_{y=0}^{\infty} \left[\binom{y}{x} p^x (1-p)^{(y-x)}\right]\left[\frac{\lambda^y e^{-\lambda}}{y!}\right] \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{(y-x)}}{(y-x)!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} = \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} = \frac{(\lambda p)^x}{x!} e^{\lambda p}.
\end{aligned}
$$

## Binomial-Poisson hierarchy Cont'd

$$
\begin{aligned}
P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) = \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \\
&= \sum_{y=0}^{\infty} \left[\binom{y}{x} p^x (1-p)^{(y-x)}\right]\left[\frac{\lambda^y e^{-\lambda}}{y!}\right] \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{(y-x)}}{(y-x)!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} = \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} = \frac{(\lambda p)^x}{x!} e^{\lambda p}.
\end{aligned}
$$

Thus, any marginal inference on $X$ is with respect to a $Poisson(\lambda p)$ distribution, with $Y$ playing not part at all.

# Binomial-Poisson hierarchy Cont'd

$$
\begin{aligned}
P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) = \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \\
&= \sum_{y=0}^{\infty} \left[\binom{y}{x} p^x (1-p)^{(y-x)}\right]\left[\frac{\lambda^y e^{-\lambda}}{y!}\right] \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{(y-x)}}{(y-x)!} \\
&= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^{t}}{t!} = \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} = \frac{(\lambda p)^x}{x!} e^{\lambda p}.
\end{aligned}
$$

Thus, any marginal inference on $X$ is with respect to a $Poisson(\lambda p)$ distribution, with $Y$ playing not part at all.

The answer to the original question is now easy to compute $E(X) = \lambda p$.

## Beta-binomial hierarchy

One generalization of the binomial distribution is to allow the success probability to vary according to a distribution. A standard model for this situation is

$$\begin{aligned} X|P &\sim \mathit{Binomial}(P), \\ P &\sim \beta(\alpha, \beta). \end{aligned}$$

# Beta-binomial hierarchy

One generalization of the binomial distribution is to allow the success probability to vary according to a distribution. A standard model for this situation is

$$X|P \sim \mathit{Binomial}(P),$$
$$P \sim \beta(\alpha, \beta).$$

By iterating the expectation, we calculate the mean of $X$ asThus, any marginal inference on $X$ as

$$E(X) = E(E(X|P)) = E(n|P) = \frac{n\alpha}{\alpha + \beta}.$$

# Conditional variance identity

Theorem

For any two r.v.s

$$Var(X) = E(Var(X|Y)) + Var(E(X|Y)),$$

provided that the expectations exist.

Proof.
By definition, we have

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E([X - E(X|Y) + E(X|Y) - E(X)]^2) \\ 0 &= E([X - E(X|Y)][E(X|Y) - E(X)]) \\ E([X - E(X|Y)]^2) &= E(E\{[X - E(X|Y)]^2|Y\}) = E(Var(X|Y)) \\ E([E(X|Y) - E(X)]^2) &= Var(E(X|Y)) \end{aligned}$$

□

# Beta-binomial hierarchy Cont'd

To calculate the variance of $X$, we have from

$$Var(X) = Var(E(X|P)) + E(Var(X|P))$$

Note that $E(X|P) = nP$ and $Var(X|P) = nP(1-P)$, where $P \sim beta(\alpha, \beta)$,

# Beta-binomial hierarchy Cont'd

To calculate the variance of $X$, we have from

$$Var(X) = Var(E(X|P)) + E(Var(X|P))$$

Note that $E(X|P) = nP$ and $Var(X|P) = nP(1-P)$, where $P \sim beta(\alpha, \beta)$,

$$Var(E(X|P)) = Var(nP) = n^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

$$\begin{aligned} E(Var(X|P)) = nE(P(1-P)) &= \frac{n\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p(1-p)p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= n\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} = \frac{n\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}. \end{aligned}$$

# Beta-binomial hierarchy Cont'd

To calculate the variance of $X$, we have from

$$Var(X) = Var(E(X|P)) + E(Var(X|P))$$

Note that $E(X|P) = nP$ and $Var(X|P) = nP(1-P)$, where $P \sim beta(\alpha, \beta)$,

$$Var(E(X|P)) = Var(nP) = n^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

$$\begin{aligned} E(Var(X|P)) = nE(P(1-P)) &= \frac{n\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p(1-p)p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= n\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} = \frac{n\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}. \end{aligned}$$

Thus we have

$$Var(X) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

# Dirichlet-multinomial hierarchy

Suppose we have a dice of $K$ sides. We toss the dice and the probability of landing on side $k$ is $p(t = k|f) = f_i$. We throw the dice $N$ times and obtain a set of results $s = \{s_1, s_2, \cdots, s_N\}$. The joint probability is

# Dirichlet-multinomial hierarchy

Suppose we have a dice of $K$ sides. We toss the dice and the probability of landing on side $k$ is $p(t = k|f) = f_i$. We throw the dice $N$ times and obtain a set of results $s = \{s_1, s_2, \cdots, s_N\}$. The joint probability is

$$p(s|f) = \prod_{n=1}^{N} p(s_n|f) = f_1^{n_1} f_2^{n_2} \cdots f_K^{n_K} = \prod_{i=1}^{K} f_i^{n_i},$$

where $n_i$ is the number of $i-$th slides.

# Dirichlet-multinomial hierarchy

Suppose we have a dice of $K$ sides. We toss the dice and the probability of landing on side $k$ is $p(t = k|f) = f_i$. We throw the dice $N$ times and obtain a set of results $s = \{s_1, s_2, \cdots, s_N\}$. The joint probability is

$$p(s|f) = \prod_{n=1}^{N} p(s_n|f) = f_1^{n_1} f_2^{n_2} \cdots f_K^{n_K} = \prod_{i=1}^{K} f_i^{n_i},$$

where $n_i$ is the number of $i-$th slides.

Suppose that $f$ is a Dirichlet distribution with $\alpha$ as hyper-parameter. Then we express the probability of $f$ as

$$Dir(f|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} f_k^{\alpha_k - 1}.$$

## Example Cont'd

If we want to estimate the parameter $f$ based on the observation of $s$, then we can express $f$ in the following manner

# Example Cont'd

If we want to estimate the parameter $f$ based on the observation of $s$, then we can express $f$ in the following manner

$$
\begin{aligned}
p(f|s,\alpha) &= \frac{p(s|f,\alpha)p(f|\alpha)}{\int_0^1 p(s|f,\alpha)p(f|\alpha)df} \\
&= \frac{\prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k - 1}}{\int_0^1 \prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k - 1} df} \\
&= \frac{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{n_k + \alpha_k - 1}}{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int_0^1 \prod_{k=1}^K f_k^{n_k + \alpha_k - 1} df} \\
&= \frac{\Gamma(\sum_{k=1}^K (n_k + \alpha_k))}{\prod_{k=1}^K \Gamma(n_k + \alpha_k)} \prod_{k=1}^K f_k^{n_k + \alpha_k - 1}
\end{aligned}
$$

# Example Cont'd

If we want to estimate the parameter $f$ based on the observation of $s$, then we can express $f$ in the following manner

$$p(f|s,\alpha) = \frac{p(s|f,\alpha)p(f|\alpha)}{\int_0^1 p(s|f,\alpha)p(f|\alpha)df}$$

$$= \frac{\prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k-1}}{\int_0^1 \prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k-1} df}$$

$$= \frac{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{n_k+\alpha_k-1}}{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int_0^1 \prod_{k=1}^K f_k^{n_k+\alpha_k-1} df}$$

$$= \frac{\Gamma(\sum_{k=1}^K (n_k+\alpha_k))}{\prod_{k=1}^K \Gamma(n_k+\alpha_k)} \prod_{k=1}^K f_k^{n_k+\alpha_k-1}$$

Notice that after estimating $f$ based on $s$ observations, $f$ is still a Dirichlet distribution with parameter $\alpha + \mathbf{n}$, where $\mathbf{n} = (n_1, n_2, \cdots, n_k)$. This property is known as conjugate priors. Based on this property, estimating the parameters $f_i$ after observing $N$ trials is a simple counting procedure.

# Covariance and correlation

In this section, we discuss two numerical measures of the strength of a relationship between two r.v.s, the covariance and correlation.

## Covariance and correlation

In this section, we discuss two numerical measures of the strength of a relationship between two r.v.s, the covariance and correlation.

The covariance and correlation of $X$ and $Y$ are the numbers defined by

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)),$$
$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

where the value of $\rho_{XY}$ is also called the correlation coefficient.

## Covariance and correlation

In this section, we discuss two numerical measures of the strength of a relationship between two r.v.s, the covariance and correlation.

The covariance and correlation of $X$ and $Y$ are the numbers defined by

$$Cov(X,Y) = E((X-\mu_X)(Y-\mu_Y)),$$
$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

where the value of $\rho_{XY}$ is also called the correlation coefficient.

- The large values of $X$ tend to be observed with large values of $Y$ and small values of $X$ with small values of $Y$, then $Cov(X,Y)$ with be positive.

# Covariance and correlation

In this section, we discuss two numerical measures of the strength of a relationship between two r.v.s, the covariance and correlation.

The covariance and correlation of $X$ and $Y$ are the numbers defined by

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)),$$

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

where the value of $\rho_{XY}$ is also called the correlation coefficient.

- The large values of $X$ tend to be observed with large values of $Y$ and small values of $X$ with small values of $Y$, then $Cov(X, Y)$ with be positive.
- Thus the sign of $Cov(X, Y)$ gives information regarding the relationship between $X$ and $Y$.

## Theorem

For any r.v.s $X$ and $Y$,

$$Cov(X, Y) = E(XY) - \mu_X \mu_Y$$

## Theorem

For any r.v.s $X$ and $Y$,

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y$$

Proof.

$$\begin{aligned} Cov(X,Y) &= E((X-\mu_X)(Y-\mu_Y)) \\ &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y \end{aligned}$$

□

The correlation is always between $-1$ and 1, with the values $-1$ and 1 indicating a perfect linear relationship between $X$ and $Y$.

# Example of correlation

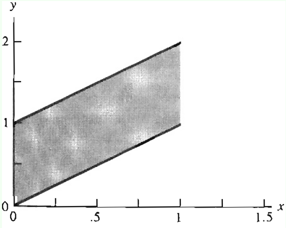
Let the joint pdf of $(X, Y)$ be
$f(x, y) = 1, 0 < x < 1, x < y < x+1.$

## Example of correlation

Let the joint pdf of $(X, Y)$ be $f(x, y) = 1, 0 < x < 1, x < y < x+1$. The marginal distribution of $X$ is *uniform*$(0, 1)$ so $\mu_X = \frac{1}{2}$ and $\sigma_X^2 = \frac{1}{12}$.

# Example of correlation

Let the joint pdf of $(X, Y)$ be $f(x, y) = 1, 0 < x < 1, x < y < x+1$.
The marginal distribution of $X$ is *uniform*$(0, 1)$ so $\mu_X = \frac{1}{2}$ and $\sigma_X^2 = \frac{1}{12}$.
The marginal distribution of $Y$ is $f_Y(y) = y, 0 < y < 1$ and $f_Y(y) = 2 - y, 1 \le y < 2$ so $\mu_Y = 1$ and $\sigma_Y^2 = \frac{1}{6}$.



$$E(XY) = \int_0^1 \int_x^{x+1} xydxdy = \int_0^1 \frac{1}{2}xy^2|_x^{x+1} dx = \int_0^1 (x^2 + \frac{1}{2}x)dx = \frac{7}{12}.$$

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{7}{12} - \frac{1}{2} \times 1}{\sqrt{\frac{1}{12}\frac{1}{6}}} = \frac{1}{\sqrt{2}}.$$

## Theorem

If r.v.s $X$ and $Y$ are independent r.v.s, then $Cov(X, Y) = 0$ and $\rho_{XY} = 0$.

For $X \sim f(x-\theta)$, symmetric around 0 with $E(X) = \theta$, and $Y$ is the indicator function $Y = I(|X-\theta| < 2)$, then $X$ and $Y$ are obviously not independent. However,

$$E(XY) = \int_{-\infty}^{\infty} xI(|X-\theta| < 2) f(x-\theta)dx = \int_{-2}^{2} (t+\theta) f(t)dt$$

$$= \theta \int_{-2}^{2} f(t)dt = E(X)E(Y), (\int_{-2}^{2} tf(t)dt = 0)$$

Thus, it is easy to find uncorrelated, dependent r.v.s.

## Theorem

If r.v.s $X$ and $Y$ are independent r.v.s, then $Cov(X, Y) = 0$ and $\rho_{XY} = 0$.

Proof.

Since $X$ and $Y$ are independent, we have $E(XY) = E(X)E(Y)$. Thus

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0$$

$$\rho_{XY} = 0$$

□

## Theorem

If r.v.s $X$ and $Y$ are independent r.v.s, then $Cov(X, Y) = 0$ and $\rho_{XY} = 0$.

Proof.

Since $X$ and $Y$ are independent, we have $E(XY) = E(X)E(Y)$. Thus

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0$$

$$\rho_{XY} = 0$$

For $X \sim f(x - \theta)$, symmetric around 0 with $E(X) = \theta$, and $Y$ is the indicator function $Y = I(|X - \theta| < 2)$, then $X$ and $Y$ are obviously not independent. However,

$$E(XY) = \int_{-\infty}^{\infty} xI(|X - \theta| < 2)f(x - \theta)dx = \int_{-2}^{2} (t + \theta)f(t)dt$$

$$= \theta \int_{-2}^{2} f(t)dt = E(X)E(Y), (\int_{-2}^{2} tf(t)dt = 0)$$

# Theorem

If r.v.s $X$ and $Y$ are independent r.v.s, then $Cov(X,Y)=0$ and $\rho_{XY}=0$.

Proof.

Since $X$ and $Y$ are independent, we have $E(XY)=E(X)E(Y)$. Thus

$$Cov(X,Y)=E(XY)-E(X)E(Y)=0$$

$$\rho_{XY}=0$$

For $X\sim f(x-\theta)$, symmetric around 0 with $E(X)=\theta$, and $Y$ is the indicator function $Y=I(|X-\theta|<2)$, then $X$ and $Y$ are obviously not independent. However,

$$E(XY)=\int_{-\infty}^{\infty}xI(|X-\theta|<2)f(x-\theta)dx=\int_{-2}^{2}(t+\theta)f(t)dt$$

$$=\theta\int_{-2}^{2}f(t)dt=E(X)E(Y),(\int_{-2}^{2}tf(t)dt=0)$$

Thus, it is easy to find uncorrelated, dependent r.v.s.

## Theorem

If $X$ and $Y$ are any two r.v.s, $a$ and $b$ are any two constants, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y).$$

If $X$ and $Y$ are independent r.v.s, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

Proof.

The mean of $aX + bY$ is $E(aX + bY) = a\mu_X + b\mu_Y$. Thus,

$$\begin{aligned} &Var(aX + bY) = E((aX + bY) - (a\mu_X + b\mu_Y))^2 \\ &= E((aX - a\mu_X) + (bY - b\mu_Y))^2 \\ &= E(a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)) \\ &= a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y) \end{aligned}$$

## Theorem

If $X$ and $Y$ are any two r.v.s,

a. $-1 \leq \rho_{XY} \leq 1$.

b. $|\rho_{XY}| = 1$ if and only if there exist numbers $a \neq 0$ and $b$ such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$; and if $\rho_{XY} = -1$, then $a < 0$.

Proof.

Consider the function $h(t)$ defined by

$$h(t) = E((X - \mu_X)t + (Y - \mu_Y))^2.$$

## Theorem

If $X$ and $Y$ are any two r.v.s,

a. $-1 \leq \rho_{XY} \leq 1$.

b. $|\rho_{XY}| = 1$ if and only if there exist numbers $a \neq 0$ and $b$ such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$; and if $\rho_{XY} = -1$, then $a < 0$.

Proof.

Consider the function $h(t)$ defined by

$$h(t) = E((X - \mu_X)t + (Y - \mu_Y))^2.$$

Expanding this expression, we obtain

$$\begin{aligned} h(t) &= t^2 E(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2t(X - \mu_X)(Y - \mu_Y) \\ &= t^2 \sigma_X^2 + 2tCov(X, Y) + \sigma_Y^2. \end{aligned}$$

□

# Proof Cont'd

$$\Delta = (2Cov(X, Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0.$$

# Proof Cont'd

$$\Delta = (2Cov(X, Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0.$$

This is equivalent to

$$-\sigma_X\sigma_Y \leq Cov(X, Y) \leq \sigma_X\sigma_Y, \text{ i.e., } -1 \leq \rho_{XY} = \frac{Cov(X, Y)}{\sigma_X\sigma_Y} \leq 1.$$

## Proof Cont'd

$$\Delta = (2Cov(X,Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0.$$

This is equivalent to

$$-\sigma_X\sigma_Y \leq Cov(X,Y) \leq \sigma_X\sigma_Y, \text{ i.e., } -1 \leq \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y} \leq 1.$$

$|\rho_{XY}| = 1$ if and only if $h(t)$ has a single root. But since $((X-\mu_X)t + (Y-\mu_Y))^2 \geq 0$, the expected value $h(t) = E((X-\mu_X)t + (Y-\mu_Y))^2 = 0$ if and only if

$$P\big(((X-\mu_X)t + (Y-\mu_Y))^2 = 0\big) = 1.$$

# Proof Cont'd

$$\Delta = (2Cov(X,Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0.$$

This is equivalent to

$$-\sigma_X\sigma_Y \leq Cov(X,Y) \leq \sigma_X\sigma_Y, \text{ i.e., } -1 \leq \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y} \leq 1.$$

$|\rho_{XY}| = 1$ if and only if $h(t)$ has a single root. But since $((X-\mu_X)t + (Y-\mu_Y))^2 \geq 0$, the expected value $h(t) = E((X-\mu_X)t + (Y-\mu_Y))^2 = 0$ if and only if

$$P\big(((X-\mu_X)t + (Y-\mu_Y))^2 = 0\big) = 1.$$

This is equivalent to

$$P\big((X-\mu_X)t + (Y-\mu_Y) = 0\big) = 1$$

# Proof Cont'd

This is $P(Y = aX + b) = 1$ with $a = -t$ and $b = \mu_X t + \mu_Y$, where $t$ is the root of $h(t)$. Using the quadratic formula, we see that this root is $t = -\frac{Cov(X,Y)}{\sigma_X^2}$. Thus $a = -t$ has the same sign as $\rho_{XY}$, proving the final assertion.

# Proof Cont'd

This is $P(Y = aX + b) = 1$ with $a = -t$ and $b = \mu_X t + \mu_Y$, where $t$ is the root of $h(t)$. Using the quadratic formula, we see that this root is $t = -\frac{Cov(X,Y)}{\sigma_X^2}$. Thus $a = -t$ has the same sign as $\rho_{XY}$, proving the final assertion.

If there is a line $y = ax + b$ $(a \neq 0)$, such that the values of $(X, Y)$ have a high probability of being near this line, then the correlation between $X$ and $Y$ will be near 1 or $-1$.

# Proof Cont'd

This is $P(Y = aX + b) = 1$ with $a = -t$ and $b = \mu_X t + \mu_Y$, where $t$ is the root of $h(t)$. Using the quadratic formula, we see that this root is $t = -\frac{Cov(X,Y)}{\sigma_X^2}$. Thus $a = -t$ has the same sign as $\rho_{XY}$, proving the final assertion.

If there is a line $y = ax + b$ $(a \neq 0)$, such that the values of $(X, Y)$ have a high probability of being near this line, then the correlation between $X$ and $Y$ will be near 1 or $-1$.

But if no such line exists, the correlation will be near 0. This is an intuitive notion of the linear relationship that is being measured by correlation.

## Example

Let $X$ have a *uniform*$(-1,1)$ distribution and $Z$ have a *uniform*$(0,\frac{1}{10})$ distribution. Suppose $X$ and $Z$ are independent. Let $Y = X^2 + Z$ and consider the random vector $(X,Y)$. The conditional distribution of $Y$ given $X = x$ is *uniform*$(x^2, x^2+\frac{1}{10})$. The joint pdf of $(X,Y)$ is

$$f(x,y) = 5, -1 < x < 1, x^2 < y < x^2 + \frac{1}{10}.$$

## Example

Let $X$ have a *uniform*$(-1,1)$ distribution and $Z$ have a *uniform*$(0,\frac{1}{10})$ distribution. Suppose $X$ and $Z$ are independent. Let $Y = X^2 + Z$ and consider the random vector $(X,Y)$. The conditional distribution of $Y$ given $X = x$ is *uniform*$(x^2, x^2+\frac{1}{10})$. The joint pdf of $(X,Y)$ is

$$f(x,y) = 5, -1 < x < 1, x^2 < y < x^2 + \frac{1}{10}.$$

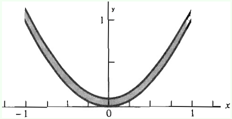
There is a strong relationship between $X$ and $Y$, as indicated by the conditional distribution of $Y$ given $X = x$.

# Example

Let $X$ have a *uniform*$(-1,1)$ distribution and $Z$ have a *uniform*$(0,\frac{1}{10})$ distribution. Suppose $X$ and $Z$ are independent. Let $Y = X^2 + Z$ and consider the random vector $(X,Y)$. The conditional distribution of $Y$ given $X = x$ is *uniform*$(x^2, x^2+\frac{1}{10})$. The joint pdf of $(X,Y)$ is

$$f(x,y) = 5, -1 < x < 1, x^2 < y < x^2 + \frac{1}{10}.$$

There is a strong relationship between $X$ and $Y$, as indicated by the conditional distribution of $Y$ given $X = x$.



In fact, $E(X) = E(X^3) = 0$, since $X$ and $Z$ are independent, $E(XZ) = E(X)E(Z)$.

$Cov(X,Y) = E(X(X^2+Z)) - E(X)(E(X^2+Z)) = 0, \rho_{XY} = 0.$

# Bivariate normal pdf

Let $\mu_X, \mu_Y \in \mathcal{R}$, $\sigma_X, \sigma_Y \in \mathcal{R}^+$ and $\rho \in [-1, 1]$ be five real numbers. The bivariate normal pdf with means $\mu_X$ and $\mu_Y$, variances $\sigma_X^2$ and $\sigma_Y^2$, and correlation $\rho$ is the bivariate pdf given by

$$f(x,y) = (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}$$
$$\cdot\, exp^{\left(-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\right)\right)}$$

# Bivariate normal pdf

Let $\mu_X, \mu_Y \in \mathcal{R}$, $\sigma_X, \sigma_Y \in \mathcal{R}^+$ and $\rho \in [-1, 1]$ be five real numbers. The bivariate normal pdf with means $\mu_X$ and $\mu_Y$, variances $\sigma_X^2$ and $\sigma_Y^2$, and correlation $\rho$ is the bivariate pdf given by

$$
\begin{aligned}
f(x, y) = {} & (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1} \\
& \cdot exp^{\left(-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\right)\right)}
\end{aligned}
$$

- The marginal distribution of $X$ is $N(\mu_X, \sigma_X^2)$;

## Bivariate normal pdf

Let $\mu_X, \mu_Y \in \mathcal{R}$, $\sigma_X, \sigma_Y \in \mathcal{R}^+$ and $\rho \in [-1, 1]$ be five real numbers. The bivariate normal pdf with means $\mu_X$ and $\mu_Y$, variances $\sigma_X^2$ and $\sigma_Y^2$, and correlation $\rho$ is the bivariate pdf given by

$$f(x,y) = (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1} \cdot exp^{\left(-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\right)\right)}$$

- The marginal distribution of $X$ is $N(\mu_X, \sigma_X^2)$;
- The marginal distribution of $Y$ is $N(\mu_Y, \sigma_Y^2)$;

## Bivariate normal pdf

Let $\mu_X, \mu_Y \in \mathcal{R}$, $\sigma_X, \sigma_Y \in \mathcal{R}^+$ and $\rho \in [-1, 1]$ be five real numbers. The bivariate normal pdf with means $\mu_X$ and $\mu_Y$, variances $\sigma_X^2$ and $\sigma_Y^2$, and correlation $\rho$ is the bivariate pdf given by

$$\begin{aligned} f(x,y) &= (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1} \\ &\cdot exp^{\left(-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\right)\right)} \end{aligned}$$

- The marginal distribution of $X$ is $N(\mu_X, \sigma_X^2)$;
- The marginal distribution of $Y$ is $N(\mu_Y, \sigma_Y^2)$;
- The correlation between $X$ and $Y$ is $\rho_{XY} = \rho$;

## Bivariate normal pdf

Let $\mu_X, \mu_Y \in \mathcal{R}$, $\sigma_X, \sigma_Y \in \mathcal{R}^+$ and $\rho \in [-1, 1]$ be five real numbers. The bivariate normal pdf with means $\mu_X$ and $\mu_Y$, variances $\sigma_X^2$ and $\sigma_Y^2$, and correlation $\rho$ is the bivariate pdf given by

$$f(x, y) = (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}$$
$$\cdot \, exp^{\left(-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\right)\right)}$$

- The marginal distribution of $X$ is $N(\mu_X, \sigma_X^2)$;
- The marginal distribution of $Y$ is $N(\mu_Y, \sigma_Y^2)$;
- The correlation between $X$ and $Y$ is $\rho_{XY} = \rho$;
- For any constants $a$ and $b$, the distribution of $aX + bY$ is $N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$.

# Multivariate distributions

We will use boldface letters to denote multiple variates. Thus, we write $\mathbf{X}$ to denote the r.v.s $\mathbf{X}_1, \cdots, \mathbf{X}_n$ and $\mathbf{x}$ to denote the sample $x_1, \cdots, x_n$.

The random vector $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$ has a sample space that is a subset of $\mathcal{R}^n$.

- If $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is a discrete random vector, then the joint pmf of $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is the function defined by
$$f(\mathbf{x}) = f(x_1, \cdots, x_n) = P(\mathbf{X}_1 = x_1, \cdots, \mathbf{X}_n = x_n)$$
for any $A \subset \mathcal{R}^n, P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x})$.

# Multivariate distributions

We will use boldface letters to denote multiple variates. Thus, we write $\mathbf{X}$ to denote the r.v.s $\mathbf{X}_1, \cdots, \mathbf{X}_n$ and $\mathbf{x}$ to denote the sample $x_1, \cdots, x_n$.

The random vector $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$ has a sample space that is a subset of $\mathcal{R}^n$.

- If $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is a discrete random vector, then the joint pmf of $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is the function defined by
$$f(\mathbf{x}) = f(x_1, \cdots, x_n) = P(\mathbf{X}_1 = x_1, \cdots, \mathbf{X}_n = x_n)$$
$$\text{for any } A \subset \mathcal{R}^n, P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}).$$
- If $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is a continuous random vector, then the joint pdf of $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is the function defined by
$$\text{for any } A \subset \mathcal{R}^n, P(\mathbf{X} \in A) = \int \cdots \int_A f(\mathbf{x}) d\mathbf{x}.$$

# Multivariate distributions Cont'd

Let $g(\mathbf{x}) = g(x_1, \cdots, x_n)$ be a real-valued function defined on the sample space of $\mathbf{X}$. Then the expected value of $g(\mathbf{X})$ is

$$E(g(\mathbf{X})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

$$E(g(\mathbf{X})) = \sum_{\mathbf{x} \in \mathcal{R}^n} g(\mathbf{x}) f(\mathbf{x})$$

# Multivariate distributions Cont'd

Let $g(\mathbf{x}) = g(x_1, \cdots, x_n)$ be a real-valued function defined on the sample space of $\mathbf{X}$. Then the expected value of $g(\mathbf{X})$ is

$$E(g(\mathbf{X})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

$$E(g(\mathbf{X})) = \sum_{\mathbf{x} \in \mathcal{R}^n} g(\mathbf{x}) f(\mathbf{x})$$

Let $(\mathbf{X}_1, \cdots, \mathbf{X}_k)$ be the first $k$ coordinates of $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$, is given by the pdf or pmf

$$f(x_1, \cdots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_k) dx_{k+1} \cdots dx_n$$

$$f(x_1, \cdots, x_k) = \sum_{(x_{k+1}, \cdots, x_n) \in \mathcal{R}^{n-k}} f(x_1, \cdots, x_k)$$

# Multinomial distribution

## Multinomial theory

Let $n$ and $m$ be positive integers, and $A$ be the set of vectors $\mathbf{x} = (x_1, \cdots, x_n)$ such that each $x_i$ is a nonnegative integer and $\sum_{i=1}^{n} x_i = m$, then for any real numbers $p_1, \cdots, p_n$

$$(p_1 + \cdots + p_n)^m = \sum_{\mathbf{x} \in A} \frac{m!}{x_1! \cdots X_n!} p_1^{x_1} \cdots p_n^{x_n}.$$

# Multinomial distribution

## Multinomial theory

Let $n$ and $m$ be positive integers, and $A$ be the set of vectors $\mathbf{x} = (x_1, \cdots, x_n)$ such that each $x_i$ is a nonnegative integer and $\sum_{i=1}^n x_i = m$, then for any real numbers $p_1, \cdots, p_n$

$$(p_1 + \cdots + p_n)^m = \sum_{\mathbf{x} \in A} \frac{m!}{x_1! \cdots X_n!} p_1^{x_1} \cdots p_n^{x_n}.$$

Let $n$ and $m$ be positive integers and $p_1, \cdots, p_n$ be numbers satisfying $0 \leq p_i \leq 1, i = 1, \cdots, n$, and $\sum_{i=1}^n p_i = 1$. Then $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$ has a *multinomial distribution* with $m$ trials and cell probabilities $p_1, \cdots, p_n$ if the joint pmf of $\mathbf{X}$ is small

$$f(x_1, \cdots, x_n) = \frac{m!}{x_1! \cdots X_n!} p_1^{x_1} \cdots p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!}$$

# Marginal pdf of multinomial distribution

$$
\begin{aligned}
f(x_n) &= \sum_{(x_1,\cdots,x_{n-1})\in B} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n} \\
&= \sum_{(x_1,\cdots,x_{n-1})\in B} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n} \frac{(m-x_n)!(1-p_n)^{m-x_n}}{(m-x_n)!(1-p_n)^{m-x_n}} \\
&= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n} \sum \frac{(m-x_n)!}{x_1!\cdots x_{n-1}!} \prod_{i=1}^{n-1} \Big(\frac{p_i}{1-p_n}\Big)^{x_i} \\
&= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n}
\end{aligned}
$$

# Marginal pdf of multinomial distribution

$$
\begin{aligned}
f(x_n) &= \sum_{(x_1,\cdots,x_{n-1})\in B} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n} \\
&= \sum_{(x_1,\cdots,x_{n-1})\in B} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n} \frac{(m-x_n)!(1-p_n)^{m-x_n}}{(m-x_n)!(1-p_n)^{m-x_n}} \\
&= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n} \sum \frac{(m-x_n)!}{x_1!\cdots x_{n-1}!} \prod_{i=1}^{n-1} (\frac{p_i}{1-p_n})^{x_i} \\
&= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n}
\end{aligned}
$$

Hence, the marginal distribution of $X_n$ is $\mathit{binomial}(m, p_n)$.

## Marginal pdf of multinomial distribution

$$\begin{aligned}
f(x_n) &= \sum_{(x_1,\cdots,x_{n-1})\in B} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n} \\
&= \sum_{(x_1,\cdots,x_{n-1})\in B} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n} \frac{(m-x_n)!(1-p_n)^{m-x_n}}{(m-x_n)!(1-p_n)^{m-x_n}} \\
&= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n} \sum \frac{(m-x_n)!}{x_1!\cdots x_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right)^{x_i} \\
&= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n}
\end{aligned}$$

Hence, the marginal distribution of $X_n$ is *binomial*$(m, p_n)$. Similar arguments show that each of the other coordinates is marginally binomially distributed.

# Mutually independent random vectors

Let $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ be random vectors with joint pdf or pmf $f(\mathbf{x}_1, \cdots, \mathbf{x}_n)$. Let $f_{\mathbf{X}_i}(\mathbf{x}_i)$ denote the marginal pdf or pmf of $\mathbf{X}_i$. Then $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ are called mutually independent random vectors if, for every $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$$f(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i).$$

If $\mathbf{X}_i$ are all one-dimensional, then $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ are called mutually independent random variables.

# Conditional pdf of multinomial distribution

$$f(x_1, \cdots, x_{n-1}|x_n) = \frac{f(x_1, \cdots, x_n)}{f(x_n)}$$

$$= \frac{\frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}}{\frac{m!}{x_n!(m-x_n)!} p_n^{x_n} (1-p_n)^{m-x_n}} = \frac{(m-x_n)!}{x_1! \cdots x_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right)^{x_i}$$

# Conditional pdf of multinomial distribution

$$f(x_1,\cdots,x_{n-1}|x_n)=\frac{f(x_1,\cdots,x_n)}{f(x_n)}$$

$$=\frac{\frac{m!}{x_1!\cdots x_n!}p_1^{x_1}\cdots p_n^{x_n}}{\frac{m!}{x_n!(m-x_n)!}p_n^{x_n}(1-p_n)^{m-x_n}}=\frac{(m-x_n)!}{x_1!\cdots x_{n-1}!}\prod_{i=1}^{n-1}\left(\frac{p_i}{1-p_n}\right)^{x_i}$$

- This is the pmf of a multinomial distribution with $m-x_n$ trials and cell probabilities $\frac{p_1}{1-p_n},\cdots,\frac{p_{n-1}}{1-p_n}$.

# Conditional pdf of multinomial distribution

$$f(x_1, \cdots, x_{n-1}|x_n) = \frac{f(x_1, \cdots, x_n)}{f(x_n)}$$

$$= \frac{\frac{m!}{x_1!\cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}}{\frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n}} = \frac{(m-x_n)!}{x_1!\cdots x_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right)^{x_i}$$

- This is the pmf of a multinomial distribution with $m - x_n$ trials and cell probabilities $\frac{p_1}{1-p_n}, \cdots, \frac{p_{n-1}}{1-p_n}$.
- The conditional distribution of any subset of the coordinates of $X_1, \cdots, X_n$ given the values of the rest of the coordinates is a multinomial distribution.

## Conditional pdf of multinomial distribution

$$f(x_1, \cdots, x_{n-1}|x_n) = \frac{f(x_1, \cdots, x_n)}{f(x_n)}$$

$$= \frac{\frac{m!}{x_1!\cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}}{\frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n}} = \frac{(m-x_n)!}{x_1! \cdots x_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right)^{x_i}$$

- This is the pmf of a multinomial distribution with $m - x_n$ trials and cell probabilities $\frac{p_1}{1-p_n}, \cdots, \frac{p_{n-1}}{1-p_n}$.
- The conditional distribution of any subset of the coordinates of $X_1, \cdots, X_n$ given the values of the rest of the coordinates is a multinomial distribution.
- We see from the conditional distribution that the coordinates of the vector $X_1, \cdots, X_n$ are related. It turns out that all of the pairwise covariances are negative and are given by $Cov(X_i, X_j) = E[(X_i - p_i)(X_j - p_j)] = -mp_ip_j$.

# Mgf of mutually independent random variables

Let $(X_1, \cdots, X_n)$ be mutually independent r.v.s.

# Mgf of mutually independent random variables

Let $(X_1, \cdots, X_n)$ be mutually independent r.v.s.

- Let $g_1, \cdots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of $x_i, i = 1, \cdots, n$. Then

$$E(\prod_{i=1}^{n} g_i(X_i)) = \prod_{i=1}^{n} E(g_i(X_i)).$$

# Mgf of mutually independent random variables

Let $(X_1, \cdots, X_n)$ be mutually independent r.v.s.

- Let $g_1, \cdots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of $x_i, i = 1, \cdots, n$. Then

$$E(\prod_{i=1}^{n} g_i(X_i)) = \prod_{i=1}^{n} E(g_i(X_i)).$$

- Let $M_{X_1}(t), \cdots, M_{X_n}(t)$ be mgfs, and $Z = X_1 + \cdots + X_n$. Then the mgf of $Z$ is

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

# Mgf of mutually independent random variables Cont'd

### Corollary

Let $(X_1, \cdots, X_n)$ be mutually independent r.v.s. Let $M_{X_1}(t), \cdots, M_{X_n}(t)$ be mgfs. Let $a_i$ and $b_i$ be fixed constants, and $Z = \sum_{i=1}^{n}(a_i X_i + b_i)$. Then the mgf of $Z$ is

$$M_Z(t) = \left(e^{t(\sum b_i)}\right) \prod_{i=1}^{n} M_{X_i}(t).$$

# Mgf of mutually independent random variables Cont'd

## Corollary

Let $(X_1, \cdots, X_n)$ be mutually independent r.v.s. Let $M_{X_1}(t), \cdots, M_{X_n}(t)$ be mgfs. Let $a_i$ and $b_i$ be fixed constants, and $Z = \sum_{i=1}^n (a_i X_i + b_i)$. Then the mgf of $Z$ is

$$M_Z(t) = \left(e^{t(\sum b_i)}\right) \prod_{i=1}^n M_{X_i}(t).$$

Let $(X_1, \cdots, X_n)$ be mutually independent r.v.s, and the distribution of $X_i$ is $Gamma(\alpha_i, \beta)$ with mgf $M(t) = (1 - \beta t)^{\alpha_i}$. Thus, the mgf of $Z = X_1 + \cdots + X_n$ is

$$M_Z(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n (1 - \beta t)^{\alpha_i} = (1 - \beta t)^{-(\sum_{i=1}^n \alpha_i)}.$$

This is the mgf of a $Gamma(\sum_{i=1}^n \alpha_i, \beta)$ distribution.

# Linear combination of independent normal r.v.s

Let $(X_1, \cdots, X_n)$ be mutually independent r.v.s. with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $a_i$ and $b_i$ be fixed constants,

$$Z = \sum_{i=1}^{n}(a_i X_i + b_i) \sim N(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2\sigma_i^2).$$

- A linear combination of independent normal r.v.s is normally distributed.
- It can be proved by the above corollary.

## Generalization

Let $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ be random vectors. Then $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are mutually independent random vectors if and only if there exist functions $g_i(\mathbf{x}_i)$ such that the joint pdf or pmf of $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ can be written as

$$f(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \prod_{i=1}^{n} g_i(\mathbf{x}_i).$$

# Generalization

Let $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ be random vectors. Then $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are mutually independent random vectors if and only if there exist functions $g_i(\mathbf{x}_i)$ such that the joint pdf or pmf of $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ can be written as

$$f(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \prod_{i=1}^{n} g_i(\mathbf{x}_i).$$

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be random vectors. Let $g_i(\mathbf{x}_i)$ be a function only of $\mathbf{x}_i$. Then the random variables $U_i = g_i(\mathbf{X}_i)$ are mutually independent

# Tail bounds

### Question

Consider the experiment of tossing a fair coin $n$ times. What is the probability that the number of heads exceeds $\frac{3n}{4}$.

### Notes

The tail bounds of a r.v. $X$ are concerned with the probability that it deviates significantly from its expected value $E(X)$ on a run of the experiment

# Markov inequality

## Markov inequality

If $X$ is any r.v. and $0 < a < +\infty$, then

$$P(X > a) \leq \frac{E(X)}{a} \text{ or } P(X > aE(X)) \leq \frac{1}{a}$$

Proof.

$$P(X > a) = \int_{X>a} dx \leq \int \frac{X}{a} dx = \frac{E(X)}{a}.$$

□

## Example

$$P(X > \frac{3n}{4}) \leq \frac{n/2}{3n/4} = \frac{2}{3}$$

## Chebyshev's inequality

If r.v. $X$ is a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,

$$P(g(X) \geq r) \leq \frac{E(g(X))}{r}.$$

Proof.

$$\begin{aligned} E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx &\geq \int_{x:g(x) \geq r} g(x) f_X(x) dx \\ \geq r \int_{x:g(x) \geq r} f_X(x) dx &= rP(g(X) \geq r) \end{aligned}$$

Rearranging now produces the desired inequality. □

# Widespread used Chebyshev's inequality

Let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$, where $\mu = E(X)$ and $\sigma^2 = Var(X)$. For convenience write $r = t^2$. Then

$$P(\frac{(x-\mu)^2}{\sigma^2} \geq t^2) \leq \frac{E(\frac{(x-\mu)^2}{\sigma^2})}{t^2} = \frac{1}{t^2}.$$

## Widespread used Chebyshev's inequality

Let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$, where $\mu = E(X)$ and $\sigma^2 = Var(X)$. For convenience write $r = t^2$. Then

$$P(\frac{(x-\mu)^2}{\sigma^2} \geq t^2) \leq \frac{E(\frac{(x-\mu)^2}{\sigma^2})}{t^2} = \frac{1}{t^2}.$$

- i.e., $P(|x-\mu| \geq t\sigma) \leq \frac{1}{t^2}$ and $P(|x-\mu| \leq t\sigma) \geq 1 - \frac{1}{t^2}$.

## Widespread used Chebyshev's inequality

Let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$, where $\mu = E(X)$ and $\sigma^2 = Var(X)$. For convenience write $r = t^2$. Then

$$P(\frac{(x-\mu)^2}{\sigma^2} \geq t^2) \leq \frac{E(\frac{(x-\mu)^2}{\sigma^2})}{t^2} = \frac{1}{t^2}.$$

- i.e., $P(|x-\mu| \geq t\sigma) \leq \frac{1}{t^2}$ and $P(|x-\mu| \leq t\sigma) \geq 1 - \frac{1}{t^2}$.
- For example, tossing a fair coin $n$ times.

$$P(X > \frac{3n}{4}) < P(|X - \frac{n}{2}| > \frac{n}{4}) \leq \frac{Var(X)}{(\frac{n}{4})^2} = \frac{4}{n}.$$

# Widespread used Chebyshev's inequality

Let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$, where $\mu = E(X)$ and $\sigma^2 = Var(X)$. For convenience write $r = t^2$. Then

$$P(\frac{(x-\mu)^2}{\sigma^2} \geq t^2) \leq \frac{E(\frac{(x-\mu)^2}{\sigma^2})}{t^2} = \frac{1}{t^2}.$$

- i.e., $P(|x-\mu| \geq t\sigma) \leq \frac{1}{t^2}$ and $P(|x-\mu| \leq t\sigma) \geq 1 - \frac{1}{t^2}$.
- For example, tossing a fair coin $n$ times.

$$P(X > \frac{3n}{4}) < P(|X - \frac{n}{2}| > \frac{n}{4}) \leq \frac{Var(X)}{(\frac{n}{4})^2} = \frac{4}{n}.$$

## Widespread used Chebyshev's inequality

Let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$, where $\mu = E(X)$ and $\sigma^2 = Var(X)$. For convenience write $r = t^2$. Then

$$P(\frac{(x-\mu)^2}{\sigma^2} \geq t^2) \leq \frac{E(\frac{(x-\mu)^2}{\sigma^2})}{t^2} = \frac{1}{t^2}.$$

- i.e., $P(|x-\mu| \geq t\sigma) \leq \frac{1}{t^2}$ and $P(|x-\mu| \leq t\sigma) \geq 1 - \frac{1}{t^2}$.
- For example, tossing a fair coin $n$ times.

$$P(X > \frac{3n}{4}) < P(|X - \frac{n}{2}| > \frac{n}{4}) \leq \frac{Var(X)}{(\frac{n}{4})^2} = \frac{4}{n}.$$

- Many other probability inequalities exist similar in spirit to Chebyshev's inequality, e.g.,

$$P(X \geq a) \leq \frac{M_X(t)}{e^{at}}.$$

# Chernoff bound

## Deriving Chernoff bound

Let $X_i$ be a sequence of independent r.v.s with $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. r.v. $X = \sum_{i=1}^{n} X_i$.

- $P(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$, where $\mu = \sum_{i=1}^{n} p_i$
- $P(X < (1-\delta)\mu) < \exp\left(-\mu\delta^2/2\right)$

Proof.

For $t > 0$,

$$\begin{aligned} P(X < (1-\delta)\mu) &= P\big(\exp(-tX) > \exp(-t(1-\delta)\mu)\big) \\ &< \frac{\prod_{i=1}^{n} E(\exp(-tX_i))}{\exp(-t(1-\delta)\mu)}. \end{aligned}$$

□

# Proof of Chernoff bound Cont.d

Note that $1 - x < e^{-x}$ if $x > 0$,

$$\begin{aligned}\prod_{i=1}^{n} E(\exp(-tX_i)) = \prod_{i=1}^{n}(p_i e^{-t} + (1-p_i)) &= \prod_{i=1}^{n}(1 - p_i(1-e^{-t})) \\ &< \prod_{i=1}^{n}\exp(p_i(e^{-t}-1)) = \exp(\mu(e^{-t}-1)).\end{aligned}$$

## Proof of Chernoff bound Cont.d

Note that $1 - x < e^{-x}$ if $x > 0$,

$$\begin{aligned}\prod_{i=1}^{n} E(\exp(-tX_i)) = \prod_{i=1}^{n}(p_i e^{-t} + (1-p_i)) = \prod_{i=1}^{n}(1 - p_i(1-e^{-t})) \\ < \prod_{i=1}^{n} \exp(p_i(e^{-t}-1)) = \exp(\mu(e^{-t}-1)).\end{aligned}$$

That is

$$P(X < (1-\delta)\mu) < \frac{\exp(\mu(e^{-t}-1))}{\exp(-t(1-\delta)\mu)} = \exp(\mu(e^{(-t)} + t - t\delta - 1))$$

# Proof of Chernoff bound Cont.d

Note that $1 - x < e^{-x}$ if $x > 0$,

$$\prod_{i=1}^{n} E(\exp(-tX_i)) = \prod_{i=1}^{n}(p_i e^{-t} + (1 - p_i)) = \prod_{i=1}^{n}(1 - p_i(1 - e^{-t}))$$

$$< \prod_{i=1}^{n} \exp(p_i(e^{-t} - 1)) = \exp(\mu(e^{-t} - 1)).$$

That is

$$P(X < (1-\delta)\mu) < \frac{\exp(\mu(e^{-t} - 1))}{\exp(-t(1-\delta)\mu)} = \exp(\mu(e^{(-t)} + t - t\delta - 1))$$

Now its time to choose $t$ to make the bound as tight as possible. Taking the derivative of $\mu(e^{(-t)} + t - t\delta - 1)$ and setting $-e^{(-t)} + 1 - \delta = 0$. We have $t = \ln(1/1 - \delta)$,

## Proof of Chernoff bound Cont.d

Note that $1 - x < e^{-x}$ if $x > 0$,

$$\prod_{i=1}^{n} E(\exp(-tX_i)) = \prod_{i=1}^{n}(p_i e^{-t} + (1 - p_i)) = \prod_{i=1}^{n}(1 - p_i(1 - e^{-t}))$$
$$< \prod_{i=1}^{n} \exp(p_i(e^{-t} - 1)) = \exp(\mu(e^{-t} - 1)).$$

That is

$$P(X < (1-\delta)\mu) < \frac{\exp(\mu(e^{-t} - 1))}{\exp(-t(1-\delta)\mu)} = \exp(\mu(e^{(-t)} + t - t\delta - 1))$$

Now its time to choose $t$ to make the bound as tight as possible. Taking the derivative of $\mu(e^{(-t)}+t-t\delta-1)$ and setting $-e^{(-t)}+1-\delta = 0$. We have $t = \ln(1/1-\delta)$,

$$P(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}.$$

# Proof of second statement

To get the simpler form of the bound, we need to get rid of the clumsy term $(1-\delta)^{(1-\delta)}$.

## Proof of second statement

To get the simpler form of the bound, we need to get rid of the clumsy term $(1-\delta)^{(1-\delta)}$. Note that

$$(1-\delta)\ln(1-\delta) = (1-\delta)(\sum_{i=1} -\frac{\delta^i}{i}) > -\delta + \frac{\delta^2}{2}$$

Thus, we have

$$(1-\delta)^{(1-\delta)} > \exp(-\delta + \frac{\delta^2}{2})$$

Furthermore,

$$\begin{aligned} P(X < (1-\delta)\mu) &< \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \\ &< \left(\frac{e^{-\delta}}{e^{(-\delta+\frac{\delta^2}{2})}}\right)^{\mu} = \exp(-\mu\delta^2/2). \end{aligned}$$

# Chernoff bound (Upper tail)

**Theorem**

Let $X_i$ be a sequence of independent r.v.s with $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. r.v. $X = \sum_{i=1}^{n} X_i$ and $\mu = \sum_{i=1}^{n} p_i$.

- $P(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$
- $P(X > (1+\delta)\mu) < \exp\left(-\mu\delta^2/4\right)$

# Chernoff bound (Upper tail)

## Theorem

Let $X_i$ be a sequence of independent r.v.s with $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. r.v. $X = \sum_{i=1}^{n} X_i$ and $\mu = \sum_{i=1}^{n} p_i$.

- $P(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$
- $P(X > (1+\delta)\mu) < \exp\left(-\mu\delta^2/4\right)$

## Example

Let $X$ be # heads in $n$ tosses of a fair coin, then $\mu = \frac{n}{2}$ and $\delta = \frac{1}{2}$, we have

$$P(X > \frac{3n}{4}) = P(X > (1+\frac{1}{2})\frac{n}{2}) < \exp\left(-\frac{n}{2}\delta^2/4\right) = \exp\left(-n/32\right)$$

If we toss the coin 1000 times, the probability is less than $\exp\left(-125/4\right)$.

# Hoeffding inequality

Let $X_1, X_2, \cdots, X_n$ be i.i.d. observations such that $E(X_i) = \mu$ and $a \leq X_i \leq b$. Then, for any $\epsilon > 0$,

$$P(|\overline{X} - \mu| > \epsilon) < 2\exp\left(-2n\epsilon^2/(b-a)^2\right)$$

## Example

If $X_1, X_2, \cdots, X_n \sim Bernoulli(p)$

- In terms of Hoeffding inequality, we have

$$P(|\overline{X} - p| > \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$$

- If $p = 0.5$,

$$P(\overline{X} - 0.5 > \frac{1}{4}) < P(|\overline{X} - 0.5| > \frac{1}{4}) \leq 2\exp\left(-8n\right).$$

# Outline

Joint and Marginal Distributions

Conditional Distribution and Independence

Bivariate Transformations

Hierarchical Models and Mixture Distributions

Hierarchical Models and Mixture Distributions

Covariance and Correlation

Multivariate Distributions

Inequalities

- Numerical Inequalities
- Functional Inequalities

Take-aways

## Lemma

Let $a$ and $b$ be any positive numbers, and let $p$ and $q$ be any positive numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\frac{1}{p} a^p + \frac{1}{q} b^q \geq ab,$$

with equality if and only if $a^p = b^q$.

## Lemma

Let $a$ and $b$ be any positive numbers, and let $p$ and $q$ be any positive numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\frac{1}{p} a^p + \frac{1}{q} b^q \geq ab,$$

with equality if and only if $a^p = b^q$.

Proof.

Fix $b$, and consider the function

$$g(a) = \frac{1}{p} a^p + \frac{1}{q} b^q - ab.$$

## Lemma

Let $a$ and $b$ be any positive numbers, and let $p$ and $q$ be any positive numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab,$$

with equality if and only if $a^p = b^q$.

Proof.

Fix $b$, and consider the function

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab.$$

To minimize $g(a)$, differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \Rightarrow a^{p-1} - b = 0 \Rightarrow b = a^{p-1}.$$

# Proof cont'd

A check of the second derivative will establish that this is indeed a minimum. Note that $(p-1)q = p$, the value of the function at the minimum is

$$\frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p = 0.$$

Since the minimum is unique, equality holds only if $a^{p-1} = b$, which is equivalent to $a^p = b^q$.

# Proof cont'd

A check of the second derivative will establish that this is indeed a minimum. Note that $(p-1)q = p$, the value of the function at the minimum is

$$\frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p = 0.$$

Since the minimum is unique, equality holds only if $a^{p-1} = b$, which is equivalent to $a^p = b^q$.

The inequalities in this subsection, although often stated in terms of expectations, rely mainly on properties of numbers.

# Proof cont'd

A check of the second derivative will establish that this is indeed a minimum. Note that $(p-1)q = p$, the value of the function at the minimum is

$$\frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p = 0.$$

Since the minimum is unique, equality holds only if $a^{p-1} = b$, which is equivalent to $a^p = b^q$.

The inequalities in this subsection, although often stated in terms of expectations, rely mainly on properties of numbers. In fact, they are all based on the following simple lemma.

# Hölder's inequality

Let $X$ and $Y$ be any two r.v.s, and let $p$ and $q$ satisfy $\frac{1}{p}+\frac{1}{q}=1$. Then

$$|E(XY)| \leq E|XY| \leq (E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}.$$

# Hölder's inequality

Let $X$ and $Y$ be any two r.v.s, and let $p$ and $q$ satisfy $\frac{1}{p}+\frac{1}{q}=1$. Then

$$|E(XY)| \leq E|XY| \leq (E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}.$$

Proof.

The first inequality follows from $-|XY| \leq XY \leq |XY|$. To prove the second inequality, define

$$a = \frac{|X|}{(E|X|^p)^{\frac{1}{p}}} \text{ and } b = \frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}.$$

## Hölder's inequality

Let $X$ and $Y$ be any two r.v.s, and let $p$ and $q$ satisfy $\frac{1}{p}+\frac{1}{q}=1$. Then

$$|E(XY)| \le E|XY| \le (E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}.$$

Proof.

The first inequality follows from $-|XY| \le XY \le |XY|$. To prove the second inequality, define

$$a = \frac{|X|}{(E|X|^p)^{\frac{1}{p}}} \text{ and } b = \frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}.$$

Applying the above lemma,

$$\frac{1}{p}\frac{|X|^p}{(E|X|^p)} + \frac{1}{q}\frac{|Y|^q}{(E|Y|^q)} \ge \frac{|XY|}{(E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}}.$$

# Hölder's inequality

Let $X$ and $Y$ be any two r.v.s, and let $p$ and $q$ satisfy $\frac{1}{p}+\frac{1}{q}=1$. Then

$$|E(XY)| \leq E|XY| \leq (E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}.$$

Proof.

The first inequality follows from $-|XY| \leq XY \leq |XY|$. To prove the second inequality, define

$$a = \frac{|X|}{(E|X|^p)^{\frac{1}{p}}} \text{ and } b = \frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}.$$

Applying the above lemma,

$$\frac{1}{p}\frac{|X|^p}{(E|X|^p)} + \frac{1}{q}\frac{|Y|^q}{(E|Y|^q)} \geq \frac{|XY|}{(E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}}.$$

Now take expectations of both sides. The expectation of the left-hand side is 1, and rearrangement gives the conclusion. □

## Cauchy-Schwarz inequality

For any two r.v.s $X$ and $Y$

$$|E(XY)| \leq E|XY| \leq (E|X|^2)^{\frac{1}{2}}(E|Y|^2)^{\frac{1}{2}}.$$

Perhaps the most famous special case of Hölder's inequality is that for which $p = q = 2$.

## Cauchy-Schwarz inequality

For any two r.v.s $X$ and $Y$

$$|E(XY)| \leq E|XY| \leq (E|X|^2)^{\frac{1}{2}}(E|Y|^2)^{\frac{1}{2}}.$$

Perhaps the most famous special case of Hölder's inequality is that for which $p = q = 2$.

Example: covariance inequality

If $X$ and $Y$ have means $\mu_X$ and $\mu_Y$, and variances $\sigma_X^2$ and $\sigma_Y^2$, respectively, we can apply the Cauchy-Schwarz inequality to get

$$E|(X - \mu_X)(Y - \mu_Y)| \leq \{E(X - \mu_X)^2\}^{\frac{1}{2}}\{E(Y - \mu_Y)^2\}^{\frac{1}{2}}.$$

Squaring both sides and using statistical notation, we have

$$(Cov(X, Y))^2 \leq \sigma_X^2\sigma_Y^2.$$

## Special cases of Hölder's inequality

If we set $Y = 1$, we get

$$E|X| \leq (E|X|^p)^{\frac{1}{p}}, 1 < p < \infty.$$

# Special cases of Hölder's inequality

If we set $Y = 1$, we get

$$E|X| \leq (E|X|^p)^{\frac{1}{p}}, 1 < p < \infty.$$

For $1 < r < p$, if we replace $|X|$ by $|X|^r$, we obtain

$$E|X|^r \leq (E|X|^{pr})^{\frac{1}{p}}, 1 < p < \infty.$$

# Special cases of Hölder's inequality

If we set $Y = 1$, we get

$$E|X| \leq (E|X|^p)^{\frac{1}{p}}, 1 < p < \infty.$$

For $1 < r < p$, if we replace $|X|$ by $|X|^r$, we obtain

$$E|X|^r \leq (E|X|^{pr})^{\frac{1}{p}}, 1 < p < \infty.$$

Now write $s = pr$ (note that $s > r$) and rearrange terms to get

$$\{E|X|^r\}^{\frac{1}{r}} \leq (E|X|^s)^{\frac{1}{s}}, 1 < r < s < \infty.$$

which is known as Liapounov's inequality.

# Minkowski's inequality

Let $X$ and $Y$ be any two r.v.s. Then for $1 \leq p < \infty$,

$$\{E|X+Y|^p\}^{\frac{1}{p}} \leq \{E|X|^p\}^{\frac{1}{p}} + \{E|Y|^p\}^{\frac{1}{p}}.$$

## Minkowski's inequality

Let $X$ and $Y$ be any two r.v.s. Then for $1 \le p < \infty$,

$$\{E|X+Y|^p\}^{\frac{1}{p}} \le \{E|X|^p\}^{\frac{1}{p}} + \{E|Y|^p\}^{\frac{1}{p}}.$$

**Proof:**

$$\begin{aligned} E|X+Y|^p &= E\big(|X+Y||X+Y|^{p-1}\big) \\ &\le E\big(|X||X+Y|^{p-1}\big) + E\big(|Y||X+Y|^{p-1}\big), \end{aligned}$$

where we have used the fact that $|X+Y| \le |X| + |Y|$.

## Minkowski's inequality

Let $X$ and $Y$ be any two r.v.s. Then for $1 \le p < \infty$,

$$\{E|X+Y|^p\}^{\frac{1}{p}} \le \{E|X|^p\}^{\frac{1}{p}} + \{E|Y|^p\}^{\frac{1}{p}}.$$

**Proof:**

$$\begin{aligned} E|X+Y|^p &= E\big(|X+Y||X+Y|^{p-1}\big) \\ &\le E\big(|X||X+Y|^{p-1}\big) + E\big(|Y||X+Y|^{p-1}\big), \end{aligned}$$

where we have used the fact that $|X+Y| \le |X| + |Y|$. Now apply Hölder's inequality to each expectation on the right-hand side of above inequality to get

$$E|X+Y|^p \le \{E|X|^p\}^{\frac{1}{p}}\{E|X+Y|^{q(p-1)}\}^{\frac{1}{q}} + \{E|Y|^p\}^{\frac{1}{p}}\{E|X+Y|^{q(p-1)}\}^{\frac{1}{q}},$$

Now divide through by $\{E|X+Y|^{q(p-1)}\}^{\frac{1}{q}}$, noting that $q(p-1) = p$ and $1 - \frac{1}{q} = \frac{1}{p}$, we obtain the conclusion.

# A new version of Hölder's inequality

For numbers $a_i$ and $b_i, i = 1, 2, \cdots, n$, the inequality

$$\sum_{i=1}^{n} |a_i b_i| \leq \big(\sum_{i=1}^{n} a_i^p\big)^{\frac{1}{p}} \big(\sum_{i=1}^{n} b_i^q\big)^{\frac{1}{q}}, \frac{1}{p} + \frac{1}{q} = 1.$$

# A new version of Hölder's inequality

For numbers $a_i$ and $b_i, i = 1, 2, \cdots, n$, the inequality

$$\sum_{i=1}^{n} |a_i b_i| \leq \big(\sum_{i=1}^{n} a_i^p\big)^{\frac{1}{p}} \big(\sum_{i=1}^{n} b_i^q\big)^{\frac{1}{q}}, \frac{1}{p} + \frac{1}{q} = 1.$$

To establish the conclusion occurs when $b_i = 1, p = q = 2$. We then have

$$\frac{1}{n}\big(\sum_{i=1}^{n} |a_i|\big)^2 \leq \sum_{i=1}^{n} a_i^2.$$

# Outline

Joint and Marginal Distributions

Conditional Distribution and Independence

Bivariate Transformations

Hierarchical Models and Mixture Distributions

Hierarchical Models and Mixture Distributions

Covariance and Correlation

Multivariate Distributions

Inequalities

- Numerical Inequalities
- Functional Inequalities

Take-aways

# Convex inequality

A function $g(x)$ is convex if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y),$$

for all $x$ and $y$, and $0 < \lambda < 1$. The function $g(x)$ is concave if $-g(x)$ is convex.

Informally, we can think of convex functions as functions that "hold water"-that is, they are bowl-shaped ($g(x) = x^2$ is convex), while concave functions "spill water"($g(x) = \log x$ is concave).

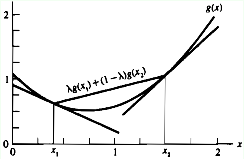
# Convex inequality

A function $g(x)$ is convex if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y),$$

for all $x$ and $y$, and $0 < \lambda < 1$. The function $g(x)$ is concave if $-g(x)$ is convex.

Informally, we can think of convex functions as functions that "hold water"-that is, they are bowl-shaped ($g(x) = x^2$ is convex), while concave functions "spill water"($g(x) = \log x$ is concave).



More formally, convex functions lie below lines connecting any two points. As $\lambda$ from 0 to 1, $\lambda g(x_1) + (1-\lambda)g(x_2)$ defines a line connecting $g(x_1)$ and $g(x_2)$. This line lies above $g(x)$ if $g(x)$ is convex.

## Jensen's inequality

For any r.v. $X$, if $g(x)$ is a convex, then

$$E(g(X)) \geq g(E(X)).$$

Equality holds if and only if, for every line $a + bx$ that a tangent to $g(x)$ at $x = E(X)$, $P(g(X) = a + bX) = 1$.

## Jensen's inequality

For any r.v. $X$, if $g(x)$ is a convex, then

$$E(g(X)) \geq g(E(X)).$$

Equality holds if and only if, for every line $a + bx$ that a tangent to $g(x)$ at $x = E(X)$, $P(g(X) = a + bX) = 1$.

Proof.

To establish the inequality, let $l(x)$ be a tangent line to $g(x)$ at the point $g(E(X))$.

## Jensen's inequality

For any r.v. $X$, if $g(x)$ is a convex, then

$$E(g(X)) \geq g(E(X)).$$

Equality holds if and only if, for every line $a+bx$ that a tangent to $g(x)$ at $x = E(X)$, $P(g(X) = a + bX) = 1$.

Proof.

To establish the inequality, let $l(x)$ be a tangent line to $g(x)$ at the point $g(E(X))$. Write $l(x) = a + bx$ for some $a$ and $b$.

## Jensen's inequality

For any r.v. $X$, if $g(x)$ is a convex, then

$$E(g(X)) \geq g(E(X)).$$

Equality holds if and only if, for every line $a+bx$ that a tangent to $g(x)$ at $x = E(X)$, $P(g(X) = a + bX) = 1$.

Proof.

To establish the inequality, let $l(x)$ be a tangent line to $g(x)$ at the point $g(E(X))$. Write $l(x) = a + bx$ for some $a$ and $b$. Now, by the convexity of $g$ we have $g(x) \geq a + bx$. Since expectations preserve inequalities,

$$E(g(X)) \geq E(a + bX) = a + bE(X) = l(E(X)) = g(E(X)).$$

□

# Jensen's inequality

For any r.v. $X$, if $g(x)$ is a convex, then

$$E(g(X)) \geq g(E(X)).$$

Equality holds if and only if, for every line $a + bx$ that a tangent to $g(x)$ at $x = E(X)$, $P(g(X) = a + bX) = 1$.

Proof.

To establish the inequality, let $l(x)$ be a tangent line to $g(x)$ at the point $g(E(X))$. Write $l(x) = a + bx$ for some $a$ and $b$. Now, by the convexity of $g$ we have $g(x) \geq a + bx$. Since expectations preserve inequalities,

$$E(g(X)) \geq E(a + bX) = a + bE(X) = l(E(X)) = g(E(X)).$$

One immediate application of Jensen's Inequality shows that $E(X^2) \geq (E(X))^2$, since $g(x) = x^2$ is convex. □

# An inequality for means

Jensen's inequality can be used to prove an inequality between three different kinds of means. If $a_1, \cdots, a_n$ are positive numbers, define

$$\begin{aligned} a_A &= \frac{1}{n}(a_1 + a_2 + \cdots + a_n), \text{(arithmetic mean)} \\ a_G &= \big(a_1 \cdot a_2 \cdot \cdots \cdot a_n\big)^{\frac{1}{n}}, \text{(geometric mean)} \\ a_H &= \frac{1}{\frac{1}{n}(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n})}. \text{(harmonic mean)} \end{aligned}$$

An inequality relating these means is

$$a_H \le a_G \le a_A.$$

To apply Jensen's inequality, let $X$ be a r.v. with range $a_1, \cdots, a_n$ and $P(X = a_i) = \frac{1}{n}, i = 1, \cdots, n$.

# An inequality for means Cont'd

Since $\log x$ is a concave function, Jensen's inequality shows that $E(\log X) \leq \log E(X)$; hence

$$\log a_G = \frac{1}{n}\sum_{i=1}^{n} \log a_i = E(\log X) \leq \log E(X) = \log a_A,$$

So $a_G \leq a_A$.
Now again use the fact that $\log x$ is concave to get

$$\log \frac{1}{a_H} = \log \frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_i} = \log E(\frac{1}{X}) \geq E(\log \frac{1}{X}) = -E(\log X).$$

Since $E(\log X) = \log a_G$, it then follows that $\log \frac{1}{a_H} \geq \log \frac{1}{a_G}$, or $a_G \geq a_H$.

## Covariance inequality

If $X$ is a r.v. with finite mean $\mu$ and $g(x)$ is a nondecreasing function, then $E(g(X)(X-\mu)) \geq 0$. Since

$$E(g(X)(X-\mu)) = E(g(X)(X-\mu)[I_{(-\infty,0)}(X-\mu) + I_{(0,\infty)}(X-\mu)])$$
$$\geq E(g(\mu)(X-\mu)I_{(-\infty,0)}(X-\mu)) + E(g(\mu)(X-\mu)I_{(0,\infty)}(X-\mu))$$
$$= g(\mu)E(X-\mu) = 0.$$

### Theorem

If $X$ is a r.v., $g(x)$ and $h(x)$ are any functions s.t. $E(g(X))$, $E(h(X))$, and $E(g(X)h(X))$ exist.

## Covariance inequality

If $X$ is a r.v. with finite mean $\mu$ and $g(x)$ is a nondecreasing function, then $E(g(X)(X-\mu)) \geq 0$. Since

$$E(g(X)(X-\mu)) = E(g(X)(X-\mu)[I_{(-\infty,0)}(X-\mu) + I_{(0,\infty)}(X-\mu)])$$
$$\geq E(g(\mu)(X-\mu)I_{(-\infty,0)}(X-\mu)) + E(g(\mu)(X-\mu)I_{(0,\infty)}(X-\mu))$$
$$= g(\mu)E(X-\mu) = 0.$$

### Theorem

If $X$ is a r.v., $g(x)$ and $h(x)$ are any functions s.t. $E(g(X))$, $E(h(X))$, and $E(g(X)h(X))$ exist.

- If $g(x)$ is nondecreasing and $h(x)$ is nonincreasing, then

$$E(g(X)h(X)) \leq E(g(X))E(h(X)).$$

# Covariance inequality

If $X$ is a r.v. with finite mean $\mu$ and $g(x)$ is a nondecreasing function, then $E(g(X)(X-\mu)) \geq 0$. Since

$$E(g(X)(X-\mu)) = E(g(X)(X-\mu)[I_{(-\infty,0)}(X-\mu) + I_{(0,\infty)}(X-\mu)])$$
$$\geq E(g(\mu)(X-\mu)I_{(-\infty,0)}(X-\mu)) + E(g(\mu)(X-\mu)I_{(0,\infty)}(X-\mu))$$
$$= g(\mu)E(X-\mu) = 0.$$

### Theorem

If $X$ is a r.v., $g(x)$ and $h(x)$ are any functions s.t. $E(g(X))$, $E(h(X))$, and $E(g(X)h(X))$ exist.

- If $g(x)$ is nondecreasing and $h(x)$ is nonincreasing, then
$$E(g(X)h(X)) \leq E(g(X))E(h(X)).$$
- If $g(x)$ and $h(x)$ are nondecreasing or nonincreasing, then
$$E(g(X)h(X)) \geq E(g(X))E(h(X)).$$

# Take-aways

## Conclusions

- Joint and marginal distributions
- Continuous distributions
- Independence
- Bivariate transformation
- Hierarchical models and mixture distributions
- Multivariate distribution
- Inequalities