# Introduction to Probability

**George Voutsadakis**[1]

[1]Mathematics and Computer Science
Lake Superior State University

LSSU Math 308

Subsection 1

## Introduction

# Expected Value Revisited

- Recall the definition of the **expected value** of a random variable $X$:
  - If $X$ is a discrete random variable with probability mass function $p(x)$, it is defined by
$$E[X] = \sum_x xp(x).$$
  - If $X$ is a continuous random variable with probability density function $f(x)$, it is defined by
$$E[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

# Bounds for Expected Value

- $E[X]$ is a weighted average of the possible values of $X$.
- Thus, if $X$ must lie between $a$ and $b$, then so must its expected value:

$$\text{If } P\{a \leq X \leq b\} = 1, \text{ then } a \leq E[X] \leq b.$$

- To verify this, suppose that $X$ is a discrete random variable for which $P\{a \leq X \leq b\} = 1$.

  This implies that $p(x) = 0$ for all $x$ outside of the interval $[a, b]$.

  Therefore,

$$\begin{aligned} E[X] &= \sum_{x:p(x)>0} xp(x) \geq \sum_{x:p(x)>0} ap(x) \\ &= a \sum_{x:p(x)>0} p(x) = a. \end{aligned}$$

  In the same manner, it can be shown that $E[X] \leq b$.
- The proof in the continuous case is similar.

Subsection 2

## Expectation of Sums of Random Variables

# Expectation of a Function of Random Variables

## Proposition

Suppose that $X$ and $Y$ are random variables and $g$ is a function of two variables. If $X$ and $Y$ have a joint probability mass function $p(x, y)$, then

$$E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y).$$

If $X$ and $Y$ have a joint probability density function $f(x, y)$, then

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

- We give a proof under the hypotheses that:
  - The random variables $X$ and $Y$ are jointly continuous with joint density function $f(x, y)$;
  - $g(X, Y)$ is a nonnegative random variable.

# Expectation of a Function of Random Variables (Cont'd)

- Because $g(X, Y) \geq 0$, we have, by a previous lemma, that

$$E[g(X, Y)] = \int_0^\infty P\{g(X, Y) > t\}dt.$$

Write $P\{g(X, Y) > t\} = \iint_{(x,y):g(x,y)>t} f(x, y)dydx$ :

$$E[g(X, Y)] = \int_0^\infty \iint_{(x,y):g(x,y)>t} f(x, y)dydxdt.$$

Interchange the order of integration:

$$\begin{aligned} E[g(X, Y)] &= \int_x \int_y \int_{t=0}^{g(x,y)} f(x, y)dtdydx \\ &= \int_x \int_y g(x, y)f(x, y)dydx. \end{aligned}$$

The result is proven for $g(X, Y)$ a nonnegative random variable.

The general case then follows as in the one-dimensional case.

# Example

- An accident occurs at a point $X$ that is uniformly distributed on a road of length $L$.

  At the time of the accident, an ambulance is at a location $Y$ that is also uniformly distributed on the road.

  Assuming that $X$ and $Y$ are independent, find the expected distance between the ambulance and the point of the accident.

  We need to compute $E[|X - Y|]$.

  The joint density function of $X$ and $Y$ is

$$f(x, y) = \frac{1}{L^2}, \quad 0 < x < L, \ 0 < y < L.$$

  Thus, we get

$$E[|X - Y|] = \frac{1}{L^2} \int_0^L \int_0^L |x - y| dy dx.$$

# Example (Cont'd)

- Now,

$$\begin{aligned}\int_0^L |x-y|dy &= \int_0^x (x-y)dy + \int_x^L (y-x)dy \\ &= \left(xy - \frac{1}{2}y^2\right)\Big|_{y=0}^{y=x} + \left(\frac{1}{2}y^2 - xy\right)\Big|_{y=x}^{y=L} \\ &= x^2 - \frac{x^2}{2} + \frac{L^2}{2} - xL - \frac{x^2}{2} + x^2 \\ &= \frac{L^2}{2} + x^2 - xL.\end{aligned}$$

Therefore,

$$E[|X-Y|] = \frac{1}{L^2}\int_0^L (\frac{L^2}{2} + x^2 - xL)dx = \frac{1}{L^2}\left(\frac{L^2x}{2} + \frac{x^3}{3} - \frac{Lx^2}{2}\right)\Big|_0^L = \frac{L}{3}.$$

# Expectation of a Sum of Random Variables

- Suppose that $E[X]$ and $E[Y]$ are both finite.
- Let $g(X, Y) = X + Y$.
- Then, in the continuous case,

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dydx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dxdy \\ &= \int_{-\infty}^{\infty}xf_X(x)dx + \int_{-\infty}^{\infty}yf_Y(y)dy \\ &= E[X] + E[Y]. \end{aligned}$$

- The same result holds in general.
- Thus, whenever $E[X]$ and $E[Y]$ are finite,

$$E[X+Y] = E[X] + E[Y].$$

# Example

- Suppose that, for random variables $X$ and $Y$, $X \geq Y$.

  That is, for any outcome of the probability experiment, the value of the random variable $X$ is greater than or equal to the value of the random variable $Y$.

  But $x \geq y$ is equivalent to the inequality $X - Y \geq 0$.

  Therefore, $E[X - Y] \geq 0$, i.e., $E[X] - E[Y] \geq 0$.

  Equivalently,

$$E[X] \geq E[Y].$$

# Expectation of a Sum of Random Variables

- Using the equation for the expectation of the sum of two random variables, we may show by a simple induction proof that if $E[X_i]$ is finite for all $i = 1, \ldots, n$, then

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n].$$

- This is an extremely useful formula whose utility will now be illustrated by a series of examples.

# Example: The Sample Mean

- Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having distribution function $F$ and expected value $\mu$.

  Such a sequence of random variables is said to constitute a **sample** from the distribution $F$.

  The quantity $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ is called the **sample mean**.

  Compute $E[\overline{X}]$.

$$\begin{aligned} E[\overline{X}] &= E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu \quad (\text{since } E[X_i] \equiv \mu) \\ &= \mu \end{aligned}$$

  That is, the expected value of the sample mean is $\mu$, the mean of the distribution.

- When the distribution mean $\mu$ is unknown, the sample mean is often used in statistics to estimate it.

# Example: Boole's Inequality

- Let $A_1, \ldots, A_n$ denote events.
  Define the indicator variables $X_i$, $i = 1, \ldots, n$, by

$$X_i = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

  Let $X = \sum_{i=1}^{n} X_i$.
  So $X$ denotes the number of the events $A_i$ that occur.
  Finally, let $Y = \begin{cases} 1, & \text{if } X \geq 1 \\ 0, & \text{otherwise} \end{cases}$.
  So $Y$ is equal to 1 if at least one of the $A_i$ occurs and is 0 otherwise.
  It is immediate that $X \geq Y$. So $E[X] \geq E[Y]$.
  We have

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P(A_i), \\ E[Y] &= P\{\text{at least one of the } A_i \text{ occur}\} = P\left(\bigcup_{i=1}^{n} A_i\right). \end{aligned}$$

  So we obtain $P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$.

# Example: Expectation of a Binomial Random Variable

- Let $X$ be a binomial random variable with parameters $n$ and $p$.

  Recall that such a random variable represents the number of successes in $n$ independent trials when each trial has probability $p$ of being a success.

  Let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{if the } i\text{th trial is a failure} \end{cases}$$

  Then

$$X = X_1 + X_2 + \cdots + X_n.$$

  $X_i$ is Bernoulli with expectation $E[X_i] = 1(p) + 0(1-p)$.

  Thus,

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = np.$$

# Example: Expected Number of Matches

- Suppose that $N$ people throw their hats into the center of a room. The hats are mixed up, and each person randomly selects one. Find the expected number of people that select their own hat.
  Let $X$ denote the number of matches.
  Then $X = X_1 + X_2 + \cdots + X_N$, where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th person selects his own hat} \\ 0, & \text{otherwise} \end{cases}$$

  Since, for each $i$, the $i$th person is equally likely to select any of the $N$ hats,

$$E[X_i] = P\{X_i = 1\} = \frac{1}{N}.$$

  Thus,

$$E[X] = E[X_1] + \cdots + E[X_N] = \frac{1}{N} \cdot N = 1.$$

# Example: Coupon-Collecting Problems

- Suppose that there are $N$ different types of coupons, and each time one obtains a coupon, it is equally likely to be any one of the $N$ types.

  Find the expected number of coupons one need amass before obtaining a complete set of at least one of each type.

  Let $X$ denote the number of coupons collected before a complete set is attained.

  Let $X_i$, $i = 0, 1, \ldots, N-1$ be the number of additional coupons that need be obtained after $i$ distinct types have been collected in order to obtain another distinct type.

  Then

$$X = X_0 + X_1 + \cdots + X_{N-1}.$$

# Example: Coupon-Collecting Problems (Cont'd)

- When $i$ distinct types of coupons have already been collected, a new coupon obtained will be of a distinct type with probability $\frac{N-i}{N}$.
  Therefore, for $k \geq 1$, $P\{X_i = k\} = \frac{N-i}{N}\left(\frac{i}{N}\right)^{k-1}$.
  Recall from infinite series that for $|x| < 1$:

$$1 + 2x + 3x^2 + \cdots = (x + x^2 + \cdots)' = \left(\frac{x}{1-x}\right)' = \frac{1}{(1-x)^2}.$$

  So we have

$$E[X_i] = \frac{N-i}{n}\left[1 + 2\frac{i}{N} + 3\left(\frac{i}{N}\right)^2 + \cdots\right] = \frac{N-i}{N}\frac{1}{(1-\frac{i}{N})^2} = \frac{N}{N-i}.$$

  This implies that

$$\begin{aligned} E[X] &= 1 + \tfrac{N}{N-1} + \tfrac{N}{N-2} + \cdots + \tfrac{N}{1} \\ &= N\left[1 + \cdots + \tfrac{1}{N-1} + \tfrac{1}{N}\right]. \end{aligned}$$

# Example

- Ten hunters are waiting for ducks to fly by.
  When a flock of ducks flies by, the hunters fire at the same time, but each chooses his target at random, independently of the others.
  Suppose each hunter independently hits his target with probability $p$.
  Compute the expected number of ducks that escape unhurt when a flock of size 10 flies overhead.
  Let $X_i$ equal 1 if the $i$th duck escapes unhurt and 0 otherwise.
  The expected number of ducks to escape can be expressed as

$$E[X_1 + \cdots + X_{10}] = E[X_1] + \cdots + E[X_{10}].$$

  Each of the hunters will, independently, hit the $i$th duck with probability $\frac{p}{10}$. Thus $E[X_i] = P\{X_i = 1\} = (1 - \frac{p}{10})^{10}$.
  Hence,

$$E[X] = 10\left(1 - \frac{p}{10}\right)^{10}.$$

# Example: Expected Number of Runs

- Suppose that a sequence of $n$ 1's and $m$ 0's is randomly permuted so that each of the $\frac{(n+m)!}{n!m!}$ possible arrangements is equally likely.

  Any consecutive string of 1's is said to constitute a **run of** 1's.
  - E.g., suppose $n = 6$, $m = 4$, and the ordering is $1, 1, 1, 0, 1, 1, 0, 0, 1, 0$. There are 3 runs of 1's.

  We are interested in computing the mean number of such runs.

  To compute this quantity, let

$$I_i = \begin{cases} 1, & \text{if a run of 1's starts at the } i\text{th position} \\ 0, & \text{otherwise} \end{cases}$$

  Then, $R(1)$, the number of runs of 1, is given by $R(1) = \sum_{i=1}^{n+m} I_i$.

  Thus, $E[R(1)] = \sum_{i=1}^{n+m} E[I_i]$.

# Example: Expected Number of Runs (Cont'd)

- We have $E[I_1] = P\{\text{“1” in position 1}\} = \frac{\frac{(n+m-1)!}{(n-1)!m!}}{\frac{(n+m)!}{n!m!}} = \frac{n}{n+m}$.

  Moreover, for $1 < i \leq n+m$,

$$\begin{aligned} E[I_i] &= P\{\text{“0” in position } i-1, \text{ “1” in position } i\} \\ &= \frac{m}{n+m}\frac{n}{n+m-1}. \end{aligned}$$

  Hence,

$$E[R(1)] = \frac{n}{n+m} + (n+m-1)\frac{nm}{(n+m)(n+m-1)}.$$

  Similarly, $E[R(0)]$, the expected number of runs of 0's, is

$$E[R(0)] = \frac{m}{n+m} + \frac{nm}{n+m}.$$

  The expected number of runs of either type is

$$E[R(1)+R(0)] = 1 + \frac{2nm}{n+m}.$$

# Example: A Random Walk in the Plane

- Consider a particle initially located at a given point in the plane, and suppose that it undergoes a sequence of steps of fixed length, but in a completely random direction.

  Specifically, suppose that the new position after each step is one unit of distance from the previous position and at an angle of orientation from the previous position that is uniformly distributed over $(0, 2\pi)$.

  Compute the expected square of the distance from the origin after $n$ steps.

  Let $(X_i, Y_i)$ denote the change in position at the $i$th step, $i = 1, \ldots, n$, in rectangular coordinates.

  Then

$$X_i = \cos\theta_i, \quad Y_i = \sin\theta_i,$$

  where $\theta_i$, $i = 1, \ldots, n$, are, by assumption, independent uniform $(0, 2\pi)$ random variables.

# Example: A Random Walk in the Plane (Cont'd)

- The position after $n$ steps has coordinates $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i)$. Thus, $D^2$, the square of the distance from the origin, is given by

$$\begin{aligned} D^2 &= (\textstyle\sum_{i=1}^{n} X_i)^2 + (\sum_{i=1}^{n} Y_i)^2 \\ &= \textstyle\sum_{i=1}^{n}(X_i^2 + Y_i^2) + \sum\sum_{i \neq j}(X_i X_j + Y_i Y_j) \\ &= n + \textstyle\sum\sum_{i \neq j}(\cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j). \end{aligned}$$

Note that

$$\begin{aligned} 2\pi E[\cos\theta_i] &= \int_0^{2\pi} \cos u du = \sin 2\pi - \sin 0 = 0; \\ 2\pi E[\sin\theta_i] &= \int_0^{2\pi} \sin u du = \cos 0 - \cos 2\pi = 0. \end{aligned}$$

Thus, using the independence of $\theta_i$ and $\theta_j$ when $i \neq j$, we get

$$E[D^2] = n.$$

# The Probability of a Union of Events

- Let $A_1, \ldots, A_n$ be events.
  Define the indicator variables $X_i$, $i = 1, \ldots, n$, by

$$X_i = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

  Note that

$$1 - \prod_{i=1}^{n}(1 - X_i) = \begin{cases} 1, & \text{if } \bigcup A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

  Hence, $E[1 - \prod_{i=1}^{n}(1 - X_i)] = P(\bigcup_{i=1}^{n} A_i)$.
  Expanding the left side of the preceding formula yields

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) &= E[\sum_{i=1}^{n} X_i + \sum\sum_{i<j} X_i X_j + \sum\sum\sum_{i<j<k} X_i X_j X_k \\ &\quad - \cdots (-1)^{n+1} X_1 \cdots X_n]. \end{aligned}$$

# The Probability of a Union of Events (Cont'd)

- However,

$$X_{i_1} X_{i_2} \cdots X_{i_k} = \begin{cases} 1, & \text{if } A_{i_1} A_{i_2} \cdots A_{i_k} \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

So $E[X_{i_1} \cdots X_{i_k}] = P(A_{i_1} \cdots A_{i_k})$.

Thus, the preceding equation is just a statement of the well-known formula for the union of events:

$$\begin{aligned} P(\bigcup A_i) &= \sum_i P(A_i) - \sum\sum_{i<j} P(A_i A_j) + \sum\sum\sum_{i<j<k} P(A_i A_j A_k) \\ &\quad - \cdots + (-1)^{n+1} P(A_1 \cdots A_n). \end{aligned}$$

# Sum of Infinitely Many Variables

- Consider an infinite collection of random variables $X_i$, $i \geq 1$, each having a finite expectation.
- It is not necessarily true that $E[\sum_{i=1}^{\infty} X_i] = \sum_{i=1}^{\infty} E[X_i]$.
- Note that $\sum_{i=1}^{\infty} X_i = \lim_{n\to\infty} \sum_{i=1}^{n} X_i$.
- Thus,

$$\begin{aligned} E[\sum_{i=1}^{\infty} X_i] &= E[\lim_{n\to\infty} \sum_{i=1}^{n} X_i] \\ &\stackrel{?}{=} \lim_{n\to\infty} E[\sum_{i=1}^{n} X_i] \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{\infty} E[X_i]. \end{aligned}$$

- Hence, the equation valid whenever we are justified in interchanging the expectation and limit operations.
- In general, this interchange is not justified, but it is valid in two important special cases:
  1. The $X_i$ are all nonnegative random variables.
  2. $\sum_{i=1}^{\infty} E[|X_i|] < \infty$.

# Example

- Consider any nonnegative, integer-valued random variable $X$.

  For $i \geq 1$, we define $X_i = \begin{cases} 1, & \text{if } X \geq i \\ 0, & \text{if } X < i \end{cases}$

  Then

$$\begin{aligned} \sum_{i=1}^{\infty} X_i &= \sum_{i=1}^{X} X_i + \sum_{i=X+1}^{\infty} X_i \\ &= \sum_{i=1}^{X} 1 + \sum_{i=X+1}^{\infty} 0 = X. \end{aligned}$$

  Hence, since the $X_i$ are all nonnegative, we obtain

$$E[X] = \sum_{i=1}^{\infty} E(X_i) = \sum_{i=1}^{\infty} P\{X \geq i\}.$$

# Example

- Suppose that $n$ elements, call them $1, 2, \ldots, n$, must be stored in a computer in the form of an ordered list.

  Each unit of time, a request will be made for one of these elements.

  Each $i$ is requested, independently of the past, with probability $P(i)$, $i \geq 1$, $\sum_i P(i) = 1$.

  Assuming that these probabilities are known, what ordering minimizes the average position in the line of the element requested?

  Suppose that the elements are numbered so that

$$P(1) \geq P(2) \geq \cdots \geq P(n).$$

  To show that $1, 2, \ldots, n$ is the optimal ordering, let $X$ denote the position of the requested element.

# Example (Cont'd)

- Consider any ordering, say,

$$O = i_1, i_2, \ldots, i_n.$$

$$P_O\{X \geq k\} = \sum_{j=k}^{n} P(i_j) \geq \sum_{j=k}^{n} P(j) = P_{1,2,\ldots,n}\{X \geq k\}.$$

Sum over $k$, using the equation of the preceding example:

$$E_O[X] \geq E_{1,2,\ldots,n}[X].$$

Therefore, ordering the elements in decreasing order of the probability that they are requested minimizes the expected position of the element requested.

Subsection 3

## Moments of the Number of Events that Occur

# Number of Events that Occur

- In the previous section we studied several problems of the form:

  For given events $A_1, \ldots, A_n$, find $E[X]$, where $X$ is the number of these events that occur.

- The solution involved defining an indicator variable $I_i$ for event $A_i$ such that

$$I_i = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

- We observed that, then, we get

$$X = \sum_{i=1}^{n} I_i.$$

- Thus, we obtained the result

$$E[X] = E\left[\sum_{i=1}^{n} I_i\right] = \sum_{i=1}^{n} E[I_i] = \sum_{i=1}^{n} P(A_i).$$

# Number of Pairs of Events that Occur

- Suppose we are interested in the number of pairs of events that occur.
- $I_i I_j$ equals 1 if both $A_i$ and $A_j$ occur, and 0 otherwise.

  Thus, the number of pairs is equal to $\sum_{i<j} I_i I_j$.
- $X$ is the number of events that occur.

  So the number of pairs of events that occur is $\binom{X}{2}$.
- Consequently, $\binom{X}{2} = \sum_{i<j} I_i I_j$, where there are $\binom{n}{2}$ terms in the summation.
- Taking expectations yields

$$E\left[\binom{X}{2}\right] = \sum_{i<j} E[I_i I_j] = \sum_{i<j} P(A_i A_j).$$

- Equivalently, $E\left[\frac{X(X-1)}{2}\right] = \sum_{i<j} P(A_i A_j)$.

# Number of Pairs of Events that Occur (Cont'd)

- The equality $E\left[\frac{X(X-1)}{2}\right] = \sum_{i<j} P(A_i A_j)$ gives that

$$E[X^2] - E[X] = 2\sum_{i<j} P(A_i A_j).$$

- We can then compute $E[X^2]$, and thus $\mathrm{Var}(X) = E[X^2] - (E[X])^2$.
- Moreover, by considering the number of distinct subsets of $k$ events that all occur, we see that

$$\binom{X}{k} = \sum_{i_1<i_2<\cdots<i_k} I_{i_1} I_{i_2} \cdots I_{i_k}.$$

- Taking expectations gives the identity

$$E\left[\binom{X}{k}\right] = \sum_{i_1<i_2<\cdots<i_k} E[I_{i_1} I_{i_2} \cdots I_{i_k}] = \sum_{i_1<i_2<\cdots<i_k} P(A_{i_1} A_{i_2} \cdots A_{i_k}).$$

# Moments of Binomial Random Variables

- Consider $n$ independent trials, with each trial being a success with probability $p$.
- Let $A_i$ be the event that trial $i$ is a success.
- When $i \neq j$, $P(A_iA_j) = p^2$.
- Consequently, we obtain

$$E\left[\binom{X}{2}\right] = \sum_{i<j} p^2 = \binom{n}{2} p^2;$$
$$E[X(X-1)] = n(n-1)p^2;$$
$$E[X^2] - E[X] = n(n-1)p^2.$$

- Now, $E[X] = \sum_{i=1}^{n} P(A_i) = np$.
- So, from the preceding equation

$$\text{Var}(X) = E[X^2] - (E[X])^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

# Moments of Binomial Random Variables (Cont'd)

- Noting that $P(A_{i_1} A_{i_2} \cdots A_{i_k}) = p^k$, we obtain that

$$E\left[\binom{X}{k}\right] = \sum_{i_1 < i_2 < \cdots < i_k} p^k = \binom{n}{k} p^k.$$

- Equivalently,

$$E[X(X-1)\cdots(X-k+1)] = n(n-1)\cdots(n-k+1)p^k.$$

- The successive values $E[X^k]$, $k \geq 3$, can be recursively obtained from this identity.
- For instance, with $k = 3$, it yields

$$E[X(X-1)(X-2)] = n(n-1)(n-2)p^3;$$
$$E[X^3 - 3X^2 + 2X] = n(n-1)(n-2)p^3;$$
$$\begin{aligned} E[X^3] &= 3E[X^2] - 2E[X] + n(n-1)(n-2)p^3 \\ &= 3n(n-1)p^2 + np + n(n-1)(n-2)p^3. \end{aligned}$$

# Moments of Hypergeometric Random Variables

- Suppose an urn contains $N$ balls, of which $m$ are white.
- $n$ balls are randomly selected.
- Let $A_i$ be the event that the $i$th ball selected is white.
- Then the number $X$ of white balls selected is equal to the number of the events $A_1, \ldots, A_n$ that occur.
- Because the $i$th ball selected is equally likely to be any of the $N$ balls, of which $m$ are white, $P(A_i) = \frac{m}{N}$.
- Consequently, we get $E[X] = \sum_{i=1}^{n} P(A_i) = \frac{nm}{N}$.
- Moreover, $P(A_iA_j) = P(A_i)P(A_j|A_i) = \frac{m}{N}\frac{m-1}{N-1}$.
- Hence, we obtain

$$E\left[\binom{X}{2}\right] = \sum_{i<j} \frac{m(m-1)}{N(N-1)} = \binom{n}{2}\frac{m(m-1)}{N(N-1)};$$

$$E[X(X-1)] = n(n-1)\frac{m(m-1)}{N(N-1)};$$

$$E[X^2] = n(n-1)\frac{m(m-1)}{N(N-1)} + E[X].$$

# Moments of Hypergeometric Random Variables (Cont'd)

- This formula yields the variance of the hypergeometric, namely,

$$\begin{aligned}\mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\ &= n(n-1)\frac{m(m-1)}{N(N-1)} + \frac{nm}{N} - \frac{n^2m^2}{N^2} \\ &= \frac{mn}{N}\left[\frac{(n-1)(m-1)}{N-1} + 1 - \frac{mn}{N}\right].\end{aligned}$$

- For higher moments of $X$, we have

$$P(A_{i_1}A_{i_2}\cdots A_{i_k}) = \frac{m(m-1)\cdots(m-k+1)}{N(N-1)\cdots(N-k+1)}.$$

- So we get

$$E\left[\binom{X}{k}\right] = \binom{n}{k}\frac{m(m-1)\cdots(m-k+1)}{N(N-1)\cdots(N-k+1)};$$

$$E[X(X-1)\cdots(X-k+1)] = n(n-1)\cdots(n-k+1)\frac{m(m-1)\cdots(m-k+1)}{N(N-1)\cdots(N-k+1)}.$$

# Example: Moments in the Match Problem

- For $i = 1, \ldots, N$, let $A_i$ be the event that person $i$ selects his or her own hat in the match problem.
- Then
$$P(A_iA_j) = P(A_i)P(A_j|A_i) = \frac{1}{N}\frac{1}{N-1}.$$
This follows, since, conditional on person $i$ selecting her own hat, the hat selected by person $j$ is equally likely to be any of the other $N-1$ hats, of which one is his own.
- Let $X$ be the number of people who select their own hat.
- Then
$$E\left[\binom{X}{2}\right] = \sum_{i<j} \frac{1}{N(N-1)} = \binom{N}{2}\frac{1}{N(N-1)};$$
$$E[X(X-1)] = 1;$$
$$E[X^2] = 1 + E[X].$$

# Example: Moments in the Match Problem (Cont'd)

- Note that $E[X] = \sum_{i=1}^{N} P(A_i) = \sum_{i=1}^{N} \frac{1}{N} = 1$.
- So we get

$$\text{Var}(X) = E[X^2] - (E[X])^2 = 1 + E[X] - (E[X])^2 = 1.$$

- Hence, both the mean and variance of the number of matches is 1.
- For higher moments, we observe that

$$P(A_{i_1} A_{i_2} \cdots A_{i_k}) = \frac{1}{N(N-1)\cdots(N-k+1)}.$$

- So

$$E\left[\binom{X}{k}\right] = \binom{N}{k} \frac{1}{N(N-1)\cdots(N-k+1)}.$$

- Equivalently

$$E[X(X-1)\cdots(X-k+1)] = 1.$$

# Example: Another Coupon-Collecting Problem

- Suppose that there are $N$ distinct types of coupons.

  Independently of past types collected, each new one obtained is of type $j$ with probability $p_j$, where $\sum_{j=1}^{N} p_j = 1$.

  Find the expected value and variance of the number of different types of coupons that appear among the first $n$ collected.

  We will work with the number of uncollected types.

  - Let $Y$ equal the number of different types of coupons collected.
  - Let $X = N - Y$ denote the number of uncollected types.

  Let $A_i$ be the event that there are no type $i$ coupons in the collection.

  Then $X$ is equal to the number of the events $A_1, \ldots, A_N$ that occur.

# Example: Another Coupon-Collecting Problem (Cont'd)

- The types of the successive coupons collected are independent. Moreover, each new coupon is not type $i$ with probability $1 - p_i$. Thus, we get

$$P(A_i) = (1 - p_i)^n.$$

Hence,

$$E[X] = \sum_{i=1}^{N} (1 - p_i)^n.$$

Therefore,

$$E[Y] = N - E[X] = N - \sum_{i=1}^{N} (1 - p_i)^n.$$

# Example: Another Coupon-Collecting Problem (Cont'd)

- Each of the $n$ coupons collected is neither a type $i$ nor a type $j$ coupon with probability $1 - p_i - p_j$.
  Thus, we have

$$P(A_iA_j) = (1 - p_i - p_j)^n, \quad i \neq j.$$

We now get

$$E[X(X-1)] = 2\sum_{i<j} P(A_iA_j) = 2\sum_{i<j}(1 - p_i - p_j)^n;$$
$$E[X^2] = 2\sum_{i<j}(1 - p_i - p_j)^n + E[X].$$

Hence, we obtain

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(X) \\ &= E[X^2] - (E[X])^2 \\ &= 2\sum_{i<j}(1 - p_i - p_j)^n + \sum_{i=1}^{N}(1 - p_i)^n \\ &\qquad - (\sum_{i=1}^{N}(1 - p_i)^n)^2. \end{aligned}$$

Subsection 4

## Covariance, Variance of Sums and Correlations

# Expectation of Product of Independent Variables

## Proposition

If $X$ and $Y$ are independent, then, for any functions $h$ and $g$,

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

- Assume $X$ and $Y$ are jointly continuous with joint density $f(x, y)$. Then

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx \\ &= E[h(Y)]E[g(X)]. \end{aligned}$$

The proof in the discrete case is similar.

# Covariance

## Definition

The **covariance** between $X$ and $Y$, denoted by $\text{Cov}(X, Y)$, is defined by

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

- Upon expanding the right side of the preceding definition, we see that

$$\begin{aligned} \text{Cov}(X, Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

# Covariance and Independence

- If $X$ and $Y$ are independent, then, by the preceding proposition, $\text{Cov}(X, Y) = 0$.
- The converse is not true.

  Example: Let $X$ be a random variable such that

$$P\{X = 0\} = P\{X = 1\} = P\{X = -1\} = \frac{1}{3}.$$

  Define

$$Y = \begin{cases} 0, & \text{if } X \neq 0 \\ 1, & \text{if } X = 0 \end{cases}$$

  We have $XY = 0$. So $E[XY] = 0$. Also, $E[X] = 0$.

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

  However, $X$ and $Y$ are clearly not independent.

# Properties of Covariance

## Proposition

(i) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$;

(ii) $\text{Cov}(X, X) = \text{Var}(X)$;

(iii) $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$;

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

- Note that

$$\begin{aligned}
\text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\
&= E[YX] - E[Y]E[X] = \text{Cov}(Y, X); \\
\text{Cov}(X, X) &= E[X^2] - E[X]E[X] = \text{Var}(X); \\
\text{Cov}(aX, Y) &= E[aXY] - E[aX]E[Y] = aE[XY] - aE[X]E[Y] \\
&= a(E[XY] - E[X]E[Y]) = a\text{Cov}(X, Y).
\end{aligned}$$

# Properties of Covariance (Cont'd)

- To prove $\text{Cov}(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$, let $\mu_i = E[X_i]$ and $\nu_j = E[Y_j]$.
  Then
$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mu_i, \quad E\left[\sum_{j=1}^{m} Y_j\right] = \sum_{j=1}^{m} \nu_j.$$
  Now we get
$$\begin{aligned}
&\text{Cov}(\textstyle\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) \\
&= E[(\textstyle\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i)(\sum_{j=1}^{m} Y_j - \sum_{j=1}^{m} \nu_j)] \\
&= E[\textstyle\sum_{i=1}^{n} (X_i - \mu_i) \sum_{j=1}^{m} (Y_j - \nu_j)] \\
&= E[\textstyle\sum_{i=1}^{n} \sum_{j=1}^{m} (X_i - \mu_i)(Y_j - \nu_j)] \\
&= \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m} E[(X_i - \mu_i)(Y_j - \nu_j)],
\end{aligned}$$
  where the last equality follows because the expected value of a sum of random variables is equal to the sum of the expected values.

# Variance and Covariance

- From parts (ii) and (iv) of the proposition, upon taking $Y_j = X_j$, $j = 1, \ldots, n$, we get

$$\begin{aligned} \mathrm{Var}(\textstyle\sum_{i=1}^n X_i) &= \mathrm{Cov}(\textstyle\sum_{i=1}^n X_i, \sum_{j=1}^n X_j) \\ &= \textstyle\sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}(X_i, X_j) \\ &= \textstyle\sum_{i=1}^n \mathrm{Var}(X_i) + \sum\sum_{i \neq j} \mathrm{Cov}(X_i, X_j). \end{aligned}$$

- Since each pair of indices $i, j, i \neq j$, appears twice in the double summation, the preceding formula is equivalent to

$$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

- If $X_1, \ldots, X_n$ are pairwise independent,

$$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

# Example

- Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having expected value $\mu$ and variance $\sigma^2$.
- Let $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ be the sample mean.
- The differences between the individual data and the sample mean $X_i - \overline{X}$, $i = 1, \ldots, n$, are called **deviations**.
- The random variable

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

  is called the **sample variance**.
- We are interested in computing:
  (a) $\text{Var}(\overline{X})$;
  (b) $E[S^2]$.

# Example (Cont'd)

(a)

$$\begin{aligned}
\text{Var}(\overline{X}) &= \text{Var}(\textstyle\sum_{i=1}^{n} \frac{X_i}{n}) = (\frac{1}{n})^2 \text{Var}(\sum_{i=1}^{n} X_i) \\
&= (\tfrac{1}{n})^2 \textstyle\sum_{i=1}^{n} \text{Var}(X_i) \text{ (by independence)} \\
&= \tfrac{1}{n^2} n\sigma^2 = \tfrac{\sigma^2}{n}.
\end{aligned}$$

(b) We start with the following algebraic identity:

$$\begin{aligned}
(n-1)S^2 &= \textstyle\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\
&= \textstyle\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i=1}^{n}(\overline{X} - \mu)^2 \\
&\qquad - 2(\overline{X} - \mu)\textstyle\sum_{i=1}^{n}(X_i - \mu) \\
&= \textstyle\sum_{i=1}^{n}(X_i - \mu)^2 + n(\overline{X} - \mu)^2 - 2(\overline{X} - \mu)n(\overline{X} - \mu) \\
&= \textstyle\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2.
\end{aligned}$$

# Example (Cont'd)

- We showed $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2$.

  Taking expectations yields

$$\begin{aligned}(n-1)E[S^2] &= \sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\overline{X} - \mu)^2] \\ &= \sum_{i=1}^{n} \text{Var}(X_i) - n\text{Var}(\overline{X}) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2,\end{aligned}$$

  where:

  - The final equality made use of Part (a);
  - The one preceding used $E[\overline{X}] = \mu$, seen previously.

  Dividing through by $n-1$ shows that the expected value of the sample variance is the distribution variance $\sigma^2$.

# Variance of a Binomial Random Variable

- Compute the variance of a binomial random variable $X$ with parameters $n$ and $p$.

  Such a random variable represents the number of successes in $n$ independent trials when each trial has probability $p$ of success.

  Thus, $X = X_1 + \cdots + X_n$, where the $X_i$ are independent Bernoulli random variables such that

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{otherwise} \end{cases}$$

  Hence, we obtain $\text{Var}(X) = \text{Var}(X_1) + \cdots + \text{Var}(X_n)$.

  But

$$\begin{aligned} \text{Var}(X_i) &= E[X_i^2] - (E[X_i])^2 \\ &= E[X_i] - (E[X_i])^2 \text{ (since } X_i^2 = X_i) \\ &= p - p^2. \end{aligned}$$

  Thus, $\text{Var}(X) = np(1-p)$.

# Correlation

- The **correlation** of two random variables $X$ and $Y$, denoted by $\rho(X, Y)$, is defined, as long as $\mathrm{Var}(X)\mathrm{Var}(Y)$ is positive, by

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

- We show that $-1 \leq \rho(X, Y) \leq 1$.

  Let $\sigma_x^2$ and $\sigma_y^2$ be the variances of $X$ and $Y$, respectively.

  Then

$$\begin{aligned} 0 &\leq \mathrm{Var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}\right) \\ &= \frac{\mathrm{Var}(X)}{\sigma_x^2} + \frac{\mathrm{Var}(Y)}{\sigma_y^2} + \frac{2\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y} \\ &= 2[1 + \rho(X, Y)]. \end{aligned}$$

  This implies that $-1 \leq \rho(X, Y)$.

# Correlation (Cont'd)

- Moreover

$$\begin{aligned} 0 &\leq \text{Var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}\right) \\ &= \frac{\text{Var}(X)}{\sigma_x^2} + \frac{\text{Var}(Y)}{(-\sigma_y)^2} - \frac{2\text{Cov}(X,Y)}{\sigma_x \sigma_y} \\ &= 2[1 - \rho(X,Y)]. \end{aligned}$$

This implies that $\rho(X, Y) \leq 1$.

# Correlation and Linearity

- Since $\text{Var}(Z) = 0$ implies that $Z$ is constant with probability 1 (to be proven in the next chapter), it follows from the proof of the inequality above that:
  - $\rho(X, Y) = 1$ implies that $Y = a + bX$, where $b = \frac{\sigma_y}{\sigma_x} > 0$;
  - $\rho(X, Y) = -1$ implies that $Y = a + bX$, where $b = -\frac{\sigma_y}{\sigma_x} < 0$.
- The reverse is also true:

  If $Y = a + bX$, then $\rho(X, Y)$ is either $+1$ or $-1$, depending on the sign of $b$.
- The correlation coefficient is a measure of the degree of linearity between $X$ and $Y$:
  - A value of $\rho(X, Y)$ near $+1$ or $-1$ indicates a high degree of linearity between $X$ and $Y$;
  - A value near 0 indicates that such linearity is absent;
  - A positive value of $\rho(X, Y)$ indicates that $Y$ tends to increase when $X$ does;
  - A negative value indicates that $Y$ tends to decrease when $X$ increases.
- If $\rho(X, Y) = 0$, then $X$ and $Y$ are said to be **uncorrelated**.

# Example

- Let $I_A$ and $I_B$ be indicator variables for the events $A$ and $B$:

$$I_A = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise} \end{cases}, \quad I_B = \begin{cases} 1, & \text{if } B \text{ occurs} \\ 0, & \text{otherwise} \end{cases}.$$

Then

$$\begin{aligned} E[I_A] &= P(A); \\ E[I_B] &= P(B); \\ E[I_A I_B] &= P(AB). \end{aligned}$$

So $\text{Cov}(I_A, I_B) = P(AB) - P(A)P(B) = P(B)[P(A|B) - P(A)]$.

This shows that the indicator variables for $A$ and $B$ are:

- Positively correlated if $P(A|B)$ is greater than $P(A)$;
- Uncorrelated if $P(A|B) = P(A)$;
- Negatively correlated if $P(A|B)$ is less than $P(A)$.

# Example

- Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having variance $\sigma^2$. Show that $\text{Cov}(X_i - \overline{X}, \overline{X}) = 0$.
  We have

$$\begin{aligned}
\text{Cov}(X_i - \overline{X}, \overline{X}) &= \text{Cov}(X_i, \overline{X}) - \text{Cov}(\overline{X}, \overline{X}) \\
&= \text{Cov}(X_i, \tfrac{1}{n}\textstyle\sum_{j=1}^{n} X_j) - \text{Var}(\overline{X}) \\
&= \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \text{Cov}(X_i, X_j) - \tfrac{\sigma^2}{n} \\
&= \tfrac{\sigma^2}{n} - \tfrac{\sigma^2}{n} = 0.
\end{aligned}$$

  The next-to-last equality uses a previous result.
  The final equality follows by

$$\text{Cov}(X_i, X_j) = \begin{cases} 0, & \text{if } j \neq i \text{ by independence} \\ \sigma^2, & \text{if } j = i \text{ since } \text{Var}(X_i) = \sigma^2 \end{cases}$$

- Although $\overline{X}$ and the deviation $X_i - \overline{X}$ are uncorrelated, they are not, in general, independent.

# Example

- Consider $m$ independent trials, each of which results in any of $r$ possible outcomes with probabilities $P_1, P_2, \ldots, P_r$, $\sum_1^r P_i = 1$.
- If we let $N_i$, $i = 1, \ldots, r$, denote the number of the $m$ trials that result in outcome $i$, then $N_1, N_2, \ldots, N_r$ have the multinomial distribution

$$P\{N_1 = n_1, N_2 = n_2, \ldots, N_r = n_r\} = \frac{m!}{n_1! n_2! \cdots n_r!} P_1^{n_1} P_2^{n_2} \cdots P_r^{n_r}, \quad \sum_{i=1}^r n_i = m.$$

- For $i \neq j$, we expect that when $N_i$ is large, $N_j$ would tend to be small.
- Hence, it is intuitive that they should be negatively correlated.
- We compute their covariance by using a previous proposition and the representation $N_i = \sum_{k=1}^m I_i(k)$ and $N_j = \sum_{k=1}^m I_j(k)$, where

$$I_i(k) = \begin{cases} 1, & \text{if trial } k \text{ results in outcome } i \\ 0, & \text{otherwise} \end{cases},$$

$$I_j(k) = \begin{cases} 1, & \text{if trial } k \text{ results in outcome } j \\ 0, & \text{otherwise} \end{cases}.$$

# Example (Cont'd)

- We have
$$\mathrm{Cov}(N_i, N_j) = \sum_{\ell=1}^{m}\sum_{k=1}^{m} \mathrm{Cov}(I_i(k), I_j(\ell)).$$
- Now, on the one hand, when $k \neq \ell$, $\mathrm{Cov}(I_i(k), I_j(\ell)) = 0$ since the outcome of trial $k$ is independent of the outcome of trial $\ell$.
- On the other hand,
$$\begin{aligned}\mathrm{Cov}(I_i(\ell), I_j(\ell)) &= E[I_i(\ell)I_j(\ell)] - E[I_i(\ell)]E[I_j(\ell)] \\ &= 0 - P_iP_j = -P_iP_j.\end{aligned}$$
This equation uses the fact that $I_i(\ell)I_j(\ell) = 0$, since trial $\ell$ cannot result in both outcome $i$ and outcome $j$.
- Hence, we obtain $\mathrm{Cov}(N_i, N_j) = -mP_iP_j$.
- This is in accord with the intuition that $N_i$ and $N_j$ are negatively correlated.

Subsection 5

## Conditional Expectation

# Conditional Expectation

## Definition

Recall that if $X$ and $Y$ are jointly discrete random variables, then the **conditional probability mass function of $X$, given that $Y = y$**, is defined, for all $y$ such that $P\{Y = y\} > 0$, by

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

We therefore define, the **conditional expectation of $X$ given that $Y = y$**, for all values of $y$ such that $p_Y(y) > 0$, by

$$\begin{aligned} E[X|Y = y] &= \sum_x xP\{X = x|Y = y\} \\ &= \sum_x xp_{X|Y}(x|y). \end{aligned}$$

# Example

- If $X$ and $Y$ are independent binomial random variables with identical parameters $n$ and $p$, calculate the conditional expected value of $X$ given that $X+Y=m$.

  We first calculate the conditional probability mass function of $X$ given that $X+Y=m$. For $k \leq \min(n,m)$,

$$\begin{aligned}
P\{X=k|X+Y=m\} &= \frac{P\{X=k, X+Y=m\}}{P\{X+Y=m\}} \\
&= \frac{P\{X=k, Y=m-k\}}{P\{X+Y=m\}} \\
&= \frac{P\{X=k\}P\{Y=m-k\}}{P\{X+Y=m\}} \\
&= \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{m-k}p^{m-k}(1-p)^{n-m+k}}{\binom{2n}{m}p^m(1-p)^{2n-m}} \\
&= \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}.
\end{aligned}$$

  We used that $X+Y$ is binomial with parameters $2n$ and $p$.

# Example (Cont'd)

- Now for $0 \leq i \leq \min(m, n)$, we get:

$$i\frac{\binom{n}{i}\binom{n}{m-i}}{\binom{2n}{m}} = i\frac{\frac{n}{i}\binom{n-1}{i-1}\binom{n}{m-i}}{\frac{2n}{m}\binom{2n-1}{m-1}} = \frac{m}{2}\frac{\binom{n-1}{i-1}\binom{n}{m-i}}{\binom{2n-1}{m-1}}.$$

Therefore, we get

$$\begin{aligned} E[X|X+Y=m] &= \sum_{i=0}^{m} i p_{X|X+Y}(i|m) \\ &= \sum_{i=0}^{m} i\frac{\binom{n}{i}\binom{n}{m-i}}{\binom{2n}{m}} \\ &= \frac{m}{2}\sum_{i=0}^{m}\frac{\binom{n-1}{i-1}\binom{n}{m-i}}{\binom{2n-1}{m-1}} \\ &= \frac{m}{2}. \end{aligned}$$

# Conditional Expectation: Continuous Case

- If $X$ and $Y$ are jointly continuous with a joint probability density function $f(x, y)$, then the **conditional probability density of $X$, given that $Y = y$**, is defined, for all values of $y$ such that $f_Y(y) > 0$, by

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

- In this case, provided that $f_Y(y) > 0$, we define the **conditional expectation of $X$, given that $Y = y$**, by

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx.$$

# Example

- Suppose that the joint density of $X$ and $Y$ is given by

$$f(x,y) = \frac{e^{-x/y}e^{-y}}{y}, \quad 0 < x < \infty, 0 < y < \infty.$$

Compute $E[X|Y=y]$.

We start by computing the conditional density

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\int_{-\infty}^{\infty} f(x,y)dx} \\ &= \frac{(1/y)e^{-x/y}e^{-y}}{\int_0^{\infty}(1/y)e^{-x/y}e^{-y}dx} = \frac{(1/y)e^{-x/y}}{\int_0^{\infty}(1/y)e^{-x/y}dx} \\ &= \frac{(1/y)e^{-x/y}}{[-e^{-x/y}]_{x=0}^{x=\infty}} = \frac{1}{y}e^{-x/y}. \end{aligned}$$

Thus,

$$E[X|Y=y] = \int_0^{\infty} \frac{x}{y}e^{-x/y}dx \overset{\text{By-Parts}}{=} y.$$

# Properties of Conditional Expectation

- Just as conditional probabilities satisfy all of the properties of ordinary probabilities, so do conditional expectations satisfy the properties of ordinary expectations.
- For instance, we have

$$E[g(X)|Y=y] = \begin{cases} \sum_x g(x)p_{X|Y}(x|y), & \text{discrete case} \\ \int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx, & \text{continuous case} \end{cases}$$

- We also have

$$E\left[\sum_{i=1}^{n} X_i|Y=y\right] = \sum_{i=1}^{n} E[X_i|Y=y].$$

- Conditional expectation given that $Y=y$ can be thought of as being an ordinary expectation on a reduced sample space consisting only of outcomes for which $Y=y$.

# Computing Expectations by Conditioning

- Denote by $E[X|Y]$ that function of the random variable $Y$ whose value at $Y = y$ is $E[X|Y = y]$.
- Note that $E[X|Y]$ is itself a random variable.

**Proposition**

$E[X] = E[E[X|Y]]$.

- If $Y$ is a discrete random variable, then the Proposition states that

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\}.$$

- If $Y$ is continuous with density $f_Y(y)$, then the Proposition states

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)dy.$$

# Computing Expectations by Conditioning (Cont'd)

- We give a proof in the case where $X$ and $Y$ are both discrete.
- We must show that $E[X] = \sum_y E[X|Y=y]P\{Y=y\}$.

  Now, the right-hand side can be written as

$$\begin{aligned}
\sum_y E[X|Y=y]P\{Y=y\} &= \sum_y \sum_x xP\{X=x|Y=y\}P\{Y=y\} \\
&= \sum_y \sum_x x\frac{P\{X=x,Y=y\}}{P\{Y=y\}}P\{Y=y\} \\
&= \sum_y \sum_x xP\{X=x,Y=y\} \\
&= \sum_x x\sum_y P\{X=x,Y=y\} \\
&= \sum_x xP\{X=x\} \\
&= E[X].
\end{aligned}$$

# Example

- A miner is trapped in a mine containing 3 doors.
  - The first door leads to a tunnel that will take him to safety after 3 hours of travel.
  - The second door leads to a tunnel that will return him to the mine after 5 hours of travel.
  - The third door leads to a tunnel that will return him to the mine after 7 hours.

If the miner is at all times equally likely to choose any one of the doors, what is the expected length of time until he reaches safety?

Let $X$ denote the amount of time until the miner reaches safety.

Let $Y$ denote the door he initially chooses.

$$\begin{aligned} E[X] &= E[X|Y=1]P\{Y=1\} + E[X|Y=2]P\{Y=2\} \\ &\quad + E[X|Y=3]P\{Y=3\} \\ &= \tfrac{1}{3}(E[X|Y=1] + E[X|Y=2] + E[X|Y=3]). \end{aligned}$$

# Example (Cont'd)

- However,

$$\begin{aligned} E[X|Y=1] &= 3; \\ E[X|Y=2] &= 5+E[X]; \\ E[X|Y=3] &= 7+E[X]. \end{aligned}$$

To understand why this equation is correct, consider, for instance, $E[X|Y=2]$ and reason as follows:

If the miner chooses the second door, he spends 5 hours in the tunnel and then returns to his cell.
But once he returns to his cell, the problem is as before.
Thus his expected additional time until safety is just $E[X]$.
Hence, $E[X|Y=2] = 5+E[X]$.

Hence,

$$E[X] = \frac{1}{3}(3+5+E[X]+7+E[X]) \;\Rightarrow\; E[X] = 15.$$

# Sum of a Random number of Random Variables

- Suppose that the number of people entering a department store on a given day is a random variable with mean 50.

  Suppose further that the amounts of money spent by these customers are independent random variables having a common mean of \$8.

  Finally, suppose that the amount of money spent by a customer is also independent of the total number of customers who enter the store.

  What is the expected amount spent in the store on a given day?

  Let $N$ denote the number of customers that enter the store.

  Let $X_i$ the amount spent by the $i$th such customer.

  Then the total amount spent can be expressed as $\sum_{i=1}^{N} X_i$.

  Now,

$$E[\sum_{i=1}^{N} X_i] = E[E[\sum_{i=1}^{N} X_i | N]].$$

# Sum of a Random number of Random Variables (Cont'd)

- But

$$\begin{aligned} E\left[\sum_{i=1}^{N} X_i | N = n\right] &= E\left[\sum_{i=1}^{n} X_i | N = n\right] \\ &= E\left[\sum_{i=1}^{n} X_i\right] \\ &\quad \text{(by independence of the } X_i \text{ and } N\text{)} \\ &= nE[X]. \text{ (where } E[X] = E[X_i]\text{)} \end{aligned}$$

This implies that

$$E\left[\sum_{i=1}^{N} X_i | N\right] = NE[X].$$

Thus,

$$E\left[\sum_{i=1}^{N} X_i\right] = E[NE[X]] = E[N]E[X].$$

Hence, in our example, the expected amount of money spent in the store is $50 \times \$8 = \$400$.

# Example

- The game of craps is begun by rolling an ordinary pair of dice.
  - If the sum of the dice is 2, 3 or 12, the player loses.
  - If it is 7 or 11, the player wins.
  - If it is any other number $i$, the player continues to roll the dice until the sum is either 7 or $i$.
    - If it is 7, the player loses;
    - if it is i, the player wins.

Let $R$ denote the number of rolls of the dice in a game of craps.

Find $E[R]$.

# Example (Cont'd)

If we let $P_i$ denote the probability that the sum of the dice is $i$, then

$$P_i = P_{14-i} = \frac{i-1}{36}, \quad i = 2, \ldots, 7.$$

To compute $E[R]$, we condition on $S$, the initial sum, giving

$$E[R] = \sum_{i=2}^{12} E[R|S=i]P_i.$$

However,

$$E[R|S=i] = \begin{cases} 1, & \text{if } i = 2, 3, 7, 11, 12 \\ 1 + \frac{1}{P_i + P_7}, & \text{otherwise} \end{cases}$$

The preceding equation follows because:

- If the sum is a value $i$ that does not end the game, then the dice will continue to be rolled until the sum is either $i$ or 7;
- The latter happens with probability $P_i + P_7$.

# Example (Cont'd)

- Therefore,

$$
\begin{aligned}
E[R] &= 1 + \sum_{i=4}^{6} \frac{P_i}{P_i + P_7} + \sum_{i=8}^{10} \frac{P_i}{P_i + P_7} \\
&= 1 + \frac{P_4}{P_4 + P_7} + \frac{P_5}{P_5 + P_7} + \frac{P_6}{P_6 + P_7} \\
&\qquad + \frac{P_8}{P_8 + P_7} + \frac{P_9}{P_9 + P_7} + \frac{P_{10}}{P_{10} + P_7} \\
&= 1 + 2\left(\frac{3}{3+6} + \frac{4}{4+6} + \frac{5}{5+6}\right) \\
&= 1 + 2\left(\frac{3}{9} + \frac{4}{10} + \frac{5}{11}\right) \\
&= 3.376.
\end{aligned}
$$

# Example

- Recall the bivariate normal joint density function of the random variables $X$ and $Y$, given by

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}.$$

- We will now show that $\rho$ is the correlation between $X$ and $Y$.
- We have seen in a previous example that $\mu_x = E[X]$, $\sigma_x^2 = \text{Var}(X)$, and $\mu_y = E[Y]$, $\sigma_y^2 = \text{Var}(Y)$.
- Consequently,

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_x\sigma_y} = \frac{E[XY] - \mu_x\mu_y}{\sigma_x\sigma_y}.$$

# Example (Cont'd)

- To determine $E[XY]$, we condition on $Y$, i.e., we use $E[XY] = E[E[XY|Y]]$.
- Recall from a previous example that the conditional distribution of $X$ given that $Y = y$ is normal with mean

$$E[X|Y = y] = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y).$$

Thus,

$$\begin{aligned} E[XY|Y = y] &= E[Xy|Y = y] \\ &= yE[X|Y = y] \\ &= y[\mu_x + \rho\tfrac{\sigma_x}{\sigma_y}(y - \mu_y)] \\ &= y\mu_x + \rho\tfrac{\sigma_x}{\sigma_y}(y^2 - \mu_y y). \end{aligned}$$

# Example (Cont'd)

- Consequently,

$$E[XY|Y] = Y\mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y^2 - \mu_y Y).$$

- This implies that

$$\begin{aligned} E[XY] &= E[Y\mu_x + \rho\tfrac{\sigma_x}{\sigma_y}(Y^2 - \mu_y Y)] \\ &= \mu_x E[Y] + \rho\tfrac{\sigma_x}{\sigma_y}E[Y^2 - \mu_y Y] \\ &= \mu_x\mu_y + \rho\tfrac{\sigma_x}{\sigma_y}(E[Y^2] - \mu_y^2) \\ &= \mu_x\mu_y + \rho\tfrac{\sigma_x}{\sigma_y}\mathrm{Var}(Y) \\ &= \mu_x\mu_y + \rho\sigma_x\sigma_y. \end{aligned}$$

- Therefore,

$$\mathrm{Corr}(X, Y) = \frac{E[XY] - \mu_x\mu_y}{\sigma_x\sigma_y} = \frac{\rho\sigma_x\sigma_y}{\sigma_x\sigma_y} = \rho.$$

# Example

- Suppose each of $n$ independent trials results in one of the outcomes $1, \ldots, k$, with respective probabilities $p_1, \ldots, p_k$, $\sum_{i=1}^{k} p_i = 1$.
- Let $N_i$ be the number of trials that result in outcome $i$, $i = 1, \ldots, k$.
- For $i \neq j$, find: $E[N_j | N_i > 0]$;
- Let $I = \begin{cases} 0, & \text{if } N_i = 0 \\ 1, & \text{if } N_i > 0 \end{cases}$
- Then

$$\begin{aligned} E[N_j] &= E[N_j | I = 0] P\{I = 0\} + E[N_j | I = 1] P\{I = 1\}. \\ E[N_j] &= E[N_j | N_i = 0] P\{N_i = 0\} + E[N_j | N_i > 0] P\{N_i > 0\}. \end{aligned}$$

# Example (Cont'd)

- The unconditional distribution of $N_j$ is binomial with parameters $n, p_j$.
- Given that $N_i = r$, each of the $n - r$ trials that do not result in outcome $i$ will, independently, result in outcome $j$ with probability $P(j|\text{not } i) = \frac{p_j}{1-p_i}$.
- Consequently, the conditional distribution of $N_j$, given that $N_i = r$, is binomial with parameters $n - r, \frac{p_j}{1-p_i}$.
- But $P\{N_i = 0\} = (1 - p_i)^n$.
- Thus, the preceding equation yields

$$np_j = n\frac{p_j}{1-p_i}(1-p_i)^n + E[N_j|N_i > 0](1-(1-p_i)^n).$$

- This gives

$$E[N_j|N_i > 0] = np_j\frac{1-(1-p_i)^{n-1}}{1-(1-p_i)^n}.$$

# Variance of the Geometric Distribution

- Independent trials, each resulting in a success with probability $p$, are successively performed.
  Let $N$ be the time of the first success. Find $\mathrm{Var}(N)$.
  Set $Y = \begin{cases} 1, & \text{if the first trial results in a success} \\ 0, & \text{otherwise} \end{cases}$
  We have $\mathrm{Var}(N) = E[N^2] - (E[N])^2$.
  To calculate $E[N^2]$, we condition on $Y$: $E[N^2] = E[E[N^2|Y]]$.
  However,
  $$\begin{aligned} E[N^2|Y=1] &= 1; \\ E[N^2|Y=0] &= E[(1+N)^2]. \end{aligned}$$
  For these two equations:
  - If the first trial results in a success, then $N = 1$ and so $N^2 = 1$.
  - If the first trial results in a failure, then the total number of trials necessary for the first success will have the same distribution as 1 plus the necessary number of additional trials.
  - The latter has same distribution as $N$, so $E[N^2|Y=0] = E[(1+N)^2]$.

# Variance of the Geometric Distribution (Cont'd)

- Hence,

$$\begin{aligned} E[N^2] &= E[N^2|Y=1]P\{Y=1\} + E[N^2|Y=0]P\{Y=0\} \\ &= p + (1-p)E[(1+N)^2] \\ &= 1 + (1-p)E[2N+N^2]. \end{aligned}$$

However, as was shown in a previous example $E[N] = \frac{1}{p}$.

Therefore,

$$E[N^2] = 1 + \frac{2(1-p)}{p} + (1-p)E[N^2];$$

$$E[N^2] = \frac{2-p}{p^2}.$$

Consequently,

$$\mathrm{Var}(N) = E[N^2] - (E[N])^2 = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}.$$

# Example

- Consider a gambling situation in which there are $r$ players, with player $i$ initially having $n_i$ units, $n_i > 0$, $i = 1, \ldots, r$.
- At each stage, two of the players are chosen to play a game, with the winner of the game receiving 1 unit from the loser.
- Any player whose fortune drops to 0 is eliminated, and this continues until a single player has all $n = \sum_{i=1}^{r} n_i$ units, with that player designated as the victor.
- Assume that:
  - The results of successive games are independent;
  - Each game is equally likely to be won by either of its two players.
- Find the average number of stages until one of the players has all $n$ units.

# Example (Cont'd)

- Suppose first that there are only 2 players.
  - Player 1 initially has $j$ units;
  - Player 2 initially has $n - j$ units.
- Let $X_j$ denote the number of stages that will be played.
- Let $A_j$ be the additional number of stages needed beyond the first.
- Let $m_j = E[X_j]$.
- Then, for $j = 1, \ldots, n-1$, $X_j = 1 + A_j$.
- Taking expectations gives $m_j = 1 + E[A_j]$.
- Conditioning on the result of the first stage then yields

$$m_j = 1 + E[A_j|1 \text{ wins first stage}]\frac{1}{2} + E[A_j|2 \text{ wins first stage}]\frac{1}{2}.$$

# Example (Cont'd)

- Now, if player 1 wins at the first stage, then the situation from that point on is exactly the same as in a problem which supposes that:
  - Player 1 starts with $j+1$ units;
  - Player 2 starts with $n-(j+1)$ units.
- Consequently,

$$E[A_j|1 \text{ wins first stage}] = m_{j+1}, \ E[A_j|2 \text{ wins first stage}] = m_{j-1};$$
$$m_j = 1 + \tfrac{1}{2}m_{j+1} + \tfrac{1}{2}m_{j-1};$$
$$m_{j+1} = 2m_j - m_{j-1} - 2, \ j = 1, \ldots, n-1.$$

- Using that $m_0 = 0$, the preceding equation yields

$$\begin{aligned} m_2 &= 2m_1 - 2; \\ m_3 &= 2m_2 - m_1 - 2 = 3m_1 - 6 = 3(m_1 - 2); \\ m_4 &= 2m_3 - m_2 - 2 = 4m_1 - 12 = 4(m_1 - 3). \end{aligned}$$

- This suggests that

$$m_i = i(m_1 - i + 1), \quad i = 1, \ldots, n.$$

# Example (Cont'd)

- To prove the preceding equality, we use mathematical induction.
- The equation is true for $i = 1, 2$.
- Take as the induction hypothesis that it is true whenever $i \leq j < n$.
- Now we must prove that it is true for $j + 1$.
- Using the previously obtain equation yields

$$\begin{aligned} m_{j+1} &= 2m_j - m_{j-1} - 2 \\ &= 2j(m_1 - j + 1) - (j-1)(m_1 - j + 2) - 2 \\ &\quad \text{(induction hypothesis)} \\ &= (j+1)m_1 - 2j^2 + 2j + j^2 - 3j + 2 - 2 \\ &= (j+1)m_1 - j^2 - j = (j+1)(m_1 - j). \end{aligned}$$

- Letting $i = n$, and using that $m_n = 0$, now yields that $m_1 = n - 1$.
- Again using the equation proved above gives the result

$$m_i = i(n - i).$$

# Example (Cont'd)

- We return to the problem involving $r$ players with initial amounts $n_i$, $i = 1, \ldots, r$, $\sum_{i=1}^{r} n_i = n$.
- Let $X$ denote the number of stages needed to obtain a victor.
- Let $X_i$ denote the number of stages involving player $i$.
- From the point of view of player $i$, starting with $n_i$, he will continue to play stages, independently being equally likely to win or lose each one, until his fortune is either $n$ or 0.
- Thus, the number of stages he plays is exactly the same as when he has a single opponent with an initial fortune of $n - n_i$.
- Consequently, by the preceding result it follows that

$$E[X_i] = n_i(n - n_i).$$

- So

$$E\left[\sum_{i=1}^{r} X_i\right] = \sum_{i=1}^{r} n_i(n - n_i) = n^2 - \sum_{i=1}^{r} n_i^2.$$

# Example (Cont'd)

- But because each stage involves two players,

$$X = \frac{1}{2}\sum_{i=1}^{r} X_i.$$

- Taking expectations now yields

$$E[X] = \frac{1}{2}\left(n^2 - \sum_{i=1}^{r} n_i^2\right).$$

- This argument shows that the mean number of stages does not depend on the manner in which the teams are selected at each stage.
- However, the same is not true for the distribution of the number of stages:
  - Suppose $r = 3$, $n_1 = n_2 = 1$, and $n_3 = 2$.
    - If Players 1 and 2 are chosen in the first stage, then it will take at least three stages to determine a winner.
    - If Player 3 is in the first stage, then it is possible for there to be only two stages.

# Example

- Let $U_1, U_2, \ldots$ be a sequence of independent uniform $(0,1)$ random variables.

  Find $E[N]$ when $N = \min\{n : \sum_{i=1}^{n} U_i > 1\}$.

  We will find $E[N]$ by obtaining a more general result.

  For $x \in [0,1]$, let $N(x) = \min\{n : \sum_{i=1}^{n} U_i > x\}$.

  Set $m(x) = E[N(x)]$.

  $N(x)$ is the number of uniform $(0,1)$ random variables we must add until their sum exceeds $x$, and $m(x)$ is its expected value.

  We will now derive an equation for $m(x)$ by conditioning on $U_1$.

  This gives, from a previous equation,

$$m(x) = \int_0^1 E[N(x)|U_1 = y]dy.$$

# Example (Cont'd)

- We have

$$E[N(x)|U_1 = y] = \begin{cases} 1, & \text{if } y > x \\ 1 + m(x-y), & \text{if } y \leq x \end{cases}$$

The preceding formula is:

- Obviously true when $y > x$;
- True when $y \leq x$, since, if the first uniform value is $y$, then the remaining number of uniform random variables needed is the same as if we were just starting and were going to add uniform random variables until their sum exceeded $x - y$.

Now we get

$$\begin{aligned} m(x) &= 1 + \int_0^x m(x-y)dy \\ &= 1 + \int_0^x m(u)du \quad (u = x - y). \end{aligned}$$

Differentiating the preceding equation yields

$m'(x) = m(x) \Leftrightarrow \frac{m'(x)}{m(x)} = 1 \Rightarrow \log[m(x)] = x + c \Rightarrow m(x) = ke^x.$

Since $m(0) = 1$, it follows that $k = 1$. So $m(x) = e^x$.

# Computing Probabilities by Conditioning

- Let $E$ denote an arbitrary event.
- Define the indicator random variable $X$ by

$$X = \begin{cases} 1, & \text{if } E \text{ occurs} \\ 0, & \text{if } E \text{ does not occur} \end{cases}$$

- It follows from the definition of $X$ that

$$\begin{aligned} E[X] &= P(E); \\ E[X|Y=y] &= P(E|Y=y), \text{ for any random variable } Y. \end{aligned}$$

- Therefore, we obtain

$$P(E) = \begin{cases} \sum_y P(E|Y=y)P(Y=y), & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} P(E|Y=y)f_Y(y)dy, & \text{if } Y \text{ is continuous} \end{cases}$$

# Computing Probabilities by Conditioning (Cont'd)

- For $Y$ discrete, we got

$$P(E) = \sum_y P(E|Y = y)P(Y = y).$$

- Suppose $Y$ is a discrete random variable taking on one of the values $y_1, \ldots, y_n$.
- Define the events $F_i$, $i = 1, \ldots, n$, by $F_i = \{Y = y_i\}$.
- Then $F_1, \ldots, F_n$ are mutually exclusive events whose union is the sample space.
- Thus, the equation reduces to the familiar equation

$$P(E) = \sum_{i=1}^{n} P(E|F_i)P(F_i).$$

# Example: The Best-Prize Problem

- Suppose we are presented with $n$ distinct prizes in sequence.

  After being presented with a prize, we must immediately decide whether to accept it or to reject it and consider the next prize.

  The only information we are given when deciding whether to accept a prize is the relative rank of that prize compared to ones already seen.

  > E.g., when the fifth prize is presented, we learn how it compares with the four prizes we've already seen.

  Suppose that once a prize is rejected, it is lost.

  Our objective is to maximize the probability of obtaining the best prize.

  Assuming that all $n!$ orderings of the prizes are equally likely, how well can we do?

# Example: The Best-Prize Problem (Cont'd)

- Fix a value $k$, $0 \leq k < n$.

  Consider the strategy that rejects the first $k$ prizes and then accepts the first one that is better than all of those first $k$.

  Let $P_k(\text{best})$ be the probability that the best prize is selected under this strategy.

  To compute it, we condition on $X$, the position of the best prize.

  This gives

$$\begin{aligned} P_k(\text{best}) &= \sum_{i=1}^{n} P_k(\text{best}|X=i)P(X=i) \\ &= \frac{1}{n}\sum_{i=1}^{n} P_k(\text{best}|X=i). \end{aligned}$$

  - If the overall best prize is among the first $k$, then no prize is ever selected under the strategy considered, i.e.,

$$P_k(\text{best}|X=i) = 0, \text{ if } i \leq k.$$

# Example: The Best-Prize Problem (Cont'd)

- Fix a value $i > k$. If the best prize is in position $i$, then the best prize will be selected if the best of the first $i-1$ prizes is among the first $k$. Then none of the prizes in positions $k+1, k+2, \ldots, i-1$ would be selected.
  But, conditional on the best prize being in position $i$, all possible orderings of the other prizes remain equally likely. So each of the first $i-1$ prizes is equally likely to be the best of that batch.
  Hence, we have

$$\begin{aligned} & P_k(\text{best}|X=i) \\ &= P\{\text{best of first } i-1 \text{ is among the first } k|X=i\} \\ &= \tfrac{k}{i-1}, \quad \text{if } i > k. \end{aligned}$$

Now we get

$$\begin{aligned} P_k(\text{best}) &= \tfrac{k}{n}\sum_{i=k+1}^{n}\tfrac{1}{i-1} \approx \tfrac{k}{n}\int_{k+1}^{n}\tfrac{1}{x-1}dx \\ &= \tfrac{k}{n}\log\left(\tfrac{n-1}{k}\right) \approx \tfrac{k}{n}\log\left(\tfrac{n}{k}\right). \end{aligned}$$

# Example: The Best-Prize Problem (Cont'd)

- Consider the function

$$g(x) = \frac{x}{n} \log\left(\frac{n}{x}\right).$$

Then

$$g'(x) = \frac{1}{n} \log\left(\frac{n}{x}\right) - \frac{1}{n}.$$

$$g'(x) = 0 \;\Rightarrow\; \log\left(\frac{n}{x}\right) = 1 \;\Rightarrow\; x = \frac{n}{e}.$$

But we saw that $P_k(\text{best}) \approx g(k)$.

Thus, the best strategy of the type considered is to:

- Let the first $\frac{n}{e}$ prizes go by;
- Accept the first one to appear that is better than all of those.

The probability that this strategy selects the best prize is approximately $g(\frac{n}{e}) = \frac{1}{e} \approx 0.36788$.

# Example: The Best-Prize Problem (Cont'd)

- Even without detailed calculations, we can see that the probability of obtaining the best prize can be made reasonably large.
- Consider the strategy of:
  - Letting half of the prizes go by;
  - Selecting the first one to appear that is better than all of those.
- The probability that a prize is actually selected is the probability that the overall best is among the second half; This is $\frac{1}{2}$.
- Given that a prize is selected, at the time of selection that prize would have been the best of more than $\frac{n}{2}$ prizes to have appeared.

  So it would have probability of at least $\frac{1}{2}$ of being the overall best.
- Hence, the strategy above has a probability greater than $\frac{1}{4}$ of obtaining the best prize.

# Example

- Let $U$ be a uniform random variable on $(0, 1)$.

  Suppose that the conditional distribution of $X$, given that $U = p$, is binomial with parameters $n$ and $p$.

  Find the probability mass function of $X$.

  Conditioning on the value of $U$ gives

$$\begin{aligned} P\{X = i\} &= \int_0^1 P\{X = i|U = p\} f_U(p) dp \\ &= \int_0^1 P\{X = i|U = p\} dp \\ &= \frac{n!}{i!(n-i)!} \int_0^1 p^i (1-p)^{n-i} dp. \end{aligned}$$

# Example (Cont'd)

- It can be shown that

$$\int_0^1 p^i(1-p)^{n-i}dp = \frac{i!(n-i)!}{(n+1)!}.$$

Hence, we obtain

$$P\{X=i\} = \frac{n!}{i!(n-i)!}\frac{i!(n-i)!}{(n+1)!} = \frac{1}{n+1}, \quad i = 0,\ldots,n.$$

If a coin whose probability of coming up heads is uniformly distributed over $(0,1)$ is flipped $n$ times, then the number of heads occurring is equally likely to be any of the values $0,\ldots,n$.

# Example (Cont'd)

- Because the preceding conditional distribution has such a nice form, it is worth trying to find another argument to enhance our intuition as to why such a result is true.
- To do so, let $U, U_1, \ldots, U_n$ be $n+1$ independent uniform $(0,1)$ random variables.
- Let $X$ denote the number of the random variables $U_1, \ldots, U_n$ that are smaller than $U$.
  - Since all the random variables $U, U_1, \ldots, U_n$ have the same distribution, it follows that $U$ is equally likely to be the smallest, or second smallest, or largest of them.
    So $X$ is equally likely to be any of the values $0, 1, \ldots, n$.
  - On the other hand, given that $U = p$, the number of the $U_i$ that are less than $U$ is a binomial random variable with parameters $n$ and $p$.

This establishes our previous result.

# Example

- Suppose that $X$ and $Y$ are independent continuous random variables having densities $f_X$ and $f_Y$, respectively.

  Compute $P\{X < Y\}$.

  Conditioning on the value of $Y$ yields

$$\begin{aligned} P\{X < Y\} &= \int_{-\infty}^{\infty} P\{X < Y | Y = y\} f_Y(y) dy \\ &= \int_{-\infty}^{\infty} P\{X < y | Y = y\} f_Y(y) dy \\ &= \int_{-\infty}^{\infty} P\{X < y\} f_Y(y) dy \text{ (independence)} \\ &= \int_{-\infty}^{\infty} F_X(y) f_Y(y) dy. \end{aligned}$$

  Here

$$F_X(y) = \int_{-\infty}^{y} f_X(x) dx.$$

# Example

- Suppose that $X$ and $Y$ are independent continuous random variables. Find the distribution of $X + Y$.

  By conditioning on the value of $Y$, we obtain

$$\begin{aligned} P\{X+Y<a\} &= \int_{-\infty}^{\infty} P\{X+Y<a|Y=y\}f_Y(y)dy \\ &= \int_{-\infty}^{\infty} P\{X+y<a|Y=y\}fY(y)dy \\ &= \int_{-\infty}^{\infty} P\{X<a-y\}f_Y(y)dy \\ &= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy. \end{aligned}$$

# Conditional Variance

- Just as we have defined the conditional expectation of $X$ given the value of $Y$, we can also define the **conditional variance of $X$ given that $Y = y$**:

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

- $\text{Var}(X|Y)$ is equal to the (conditional) expected square of the difference between $X$ and its (conditional) mean when the value of $Y$ is given.
- In other words, $\text{Var}(X|Y)$ is exactly analogous to the usual definition of variance, but now all expectations are conditional on the fact that $Y$ is known.

# Conditional and Unconditional Variance

## Proposition (The Conditional Variance Formula)

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]).$$

- By the same reasoning that yields $\text{Var}(X) = E[X^2] - (E[X])^2$, we have

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

So

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

Also,

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

Finally adding, we get

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

# Example

- Suppose that by any time $t$ the number of people that have arrived at a train depot is a Poisson random variable with mean $\lambda t$.
  Suppose the initial train arrives at the depot at a time (independent of passenger arrivals) uniformly distributed over $(0, T)$.
  (a) Find the mean of the number of passengers who enter the train;
  (b) What is the variance of the number of passengers who enter the train?

  Let $N(t)$ be the number of arrivals by time $t$, $t \geq 0$;

  Let $Y$ be the time at which the train arrives.

  The random variable of interest is then $N(Y)$.

(a) Conditioning on $Y$ gives

$$\begin{aligned} E[N(Y)|Y=t] &= E[N(t)|Y=t] \\ &= E[N(t)] \text{ (independence of } Y \text{ and } N(t)) \\ &= \lambda t \text{ } (N(t) \text{ is Poisson with mean } \lambda t). \end{aligned}$$

Hence, $E[N(Y)|Y] = \lambda Y$.

# Example (Cont'd)

- So taking expectations gives $E[N(Y)] = \lambda E[Y] = \frac{\lambda T}{2}$.

(b) To obtain $\mathrm{Var}(N(Y))$, we use the conditional variance formula:

$$\begin{aligned} \mathrm{Var}(N(Y)|Y=t) &= \mathrm{Var}(N(t)|Y=t) \\ &= \mathrm{Var}(N(t)) \text{ (by independence)} \\ &= \lambda t. \end{aligned}$$

Thus,

$$\mathrm{Var}(N(Y)|Y) = \lambda Y, \quad E[N(Y)|Y] = \lambda Y.$$

Hence, from the conditional variance formula,

$$\begin{aligned} \mathrm{Var}(N(Y)) &= E[\mathrm{Var}(N(Y)|Y)] + \mathrm{Var}(E[N(Y)|Y]) \\ &= E[\lambda Y] + \mathrm{Var}(\lambda Y) \\ &= \lambda \frac{T}{2} + \lambda^2 \frac{T^2}{12}. \quad (\text{since } \mathrm{Var}(Y) = \frac{T^2}{12}) \end{aligned}$$

# Variance of Sum of Random Number of Random Variables

- Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables.
- Let $N$ be a nonnegative integer-valued random variable that is independent of the sequence $X_i$, $i \geq 1$.
- Note that, given $N$,
  - $\sum_{i=1}^{N} X_i$ is the sum of a fixed number of independent random variables;
  - So its expectation and variance are just the sums of the individual means and variances, respectively.

Thus, we get:

$$E\left[\sum_{i=1}^{N} X_i | N\right] = NE[X]; \quad \text{Var}\left(\sum_{i=1}^{N} X_i | N\right) = N\text{Var}(X).$$

From the conditional variance formula,

$$\text{Var}\left(\sum_{i=1}^{N} X_i\right) = E[N]\text{Var}(X) + (E[X])^2\text{Var}(N).$$

Subsection 6

## Conditional Expectation and Prediction

# Prediction

- Sometimes a situation arises in which the value of a random variable $X$ is observed and then, on the basis of the observed value, an attempt is made to predict the value of a second random variable $Y$.
- Let $g(X)$ denote the predictor:

$$\text{If } X \text{ is observed to equal } x, \text{ then}$$
$$g(x) \text{ is our prediction for the value of } Y.$$

- We would like to choose $g$ so that $g(X)$ tends to be close to $Y$.
- One possible criterion for closeness is:

$$\text{Choose } g \text{ so as to minimize } E[(Y - g(X))^2].$$

- We show that:

$$\text{Under this criterion, the best possible predictor is } g(X) = E[Y|X].$$

# Best Predictor

## Proposition

$$E[(Y - g(X))^2] \geq E[(Y - E[Y|X])^2].$$

- We have

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

Given $X$, $E[Y|X] - g(X)$, being a function of $X$, can be treated as a constant:

$$\begin{aligned} &E[(Y - E[Y|X])(E[Y|X] - g(X))|X] \\ &= (E[Y|X] - g(X))E[Y - E[Y|X]|X] \\ &= (E[Y|X] - g(X))(E[Y|X] - E[Y|X]) = 0. \end{aligned}$$

# Best Predictor (Cont'd)

- Hence, from these equations we obtain

$$E[(Y - g(X))^2|X] \geq E[(Y - E[Y|X])^2|X].$$

Taking expectations, we get

$$E[E[(Y - g(X))^2|X]] \geq E[E[(Y - E[Y|X])^2|X]].$$

We conclude that

$$E[(Y - g(X))^2] \geq E[(Y - E[Y|X])^2].$$

# Best Predictor (Alternative Argument)

- A more intuitive, although less rigorous, argument verifying the proposition is as follows.
- We can verify that $E[(Y - c)^2]$ is minimized at $c = E[Y]$.
- Thus, in the absence of data, if we want to predict the value of $Y$, the best possible prediction, in the sense of minimizing the mean square error, is to predict that $Y$ will equal its mean.
- However, if the value of the random variable $X$ is observed to be $x$, then the prediction problem remains exactly as in the previous (no-data) case, with the exception that all probabilities and expectations are now conditional on the event that $X = x$.
- Hence, the best prediction in this situation is to predict that $Y$ will equal its conditional expected value given that $X = x$.

# Example

- Suppose that the son of a man of height $x$ (in inches) attains a height that is normally distributed with mean $x+1$ and variance 4. What is the best prediction of the height at full growth of the son of a man who is 6 feet tall?
  - Let $X$ represent the height of the man;
  - Let $Y$ represent the height of his son;
  - Let $e$ is a normal random variable, independent of $X$, having mean 0 and variance 4.

  The model can be written as

$$Y = X + 1 + e.$$

  The best prediction is equal to

$$\begin{aligned} E[Y|X=72] &= E[X+1+e|X=72] \\ &= 73 + E[e|X=72] \\ &= 73 + E(e) \text{ (by independence)} \\ &= 73. \end{aligned}$$

# Example

- Suppose that:
  - A signal value $s$ is sent from location A;
  - The value received at B is normally distributed with parameters $(s, 1)$.

Suppose $S$, the value of the signal sent at A, is normally distributed with parameters $(\mu, \sigma^2)$.

Given that $R$, the value received at B, is equal to $r$, what is the best estimate of the signal sent?

We start by computing the conditional density of $S$ given $R$.

$$\begin{aligned} f_{S|R}(s|r) &= \frac{f_{S,R}(s,r)}{f_R(r)} = \frac{f_S(s) f_{R|S}(r|s)}{f_R(r)} \\ &= K e^{-(s-\mu)^2/2\sigma^2} e^{-(r-s)^2/2}, \end{aligned}$$

Here $K$ does not depend on $s$.

# Example (Cont'd)

- We calculate the negative of the exponent:

$$\begin{aligned}
\frac{(s-\mu)^2}{2\sigma^2}+\frac{(r-s)^2}{2} &= \frac{s^2}{2\sigma^2}-\frac{s\mu}{\sigma^2}+\frac{\mu^2}{2\sigma^2}+\frac{r^2}{2}-rs+\frac{s^2}{2} \\
&= s^2\left(\frac{1}{2\sigma^2}+\frac{1}{2}\right)-\left(\frac{\mu}{\sigma^2}+r\right)s+C_1 \\
&= s^2\left(\frac{1+\sigma^2}{2\sigma^2}\right)-\left(\frac{\mu+r\sigma^2}{\sigma^2}\right)s+C_1 \\
&= s^2\left(\frac{1+\sigma^2}{2\sigma^2}\right)-2\left(\frac{\mu+r\sigma^2}{1+\sigma^2}\right)\left(\frac{1+\sigma^2}{2\sigma^2}\right)s+C_1 \\
&= \frac{1+\sigma^2}{2\sigma^2}\left[s^2-2\left(\frac{\mu+r\sigma^2}{1+\sigma^2}\right)s\right]+C_1 \\
&= \frac{1+\sigma^2}{2\sigma^2}\left[s^2-2\frac{\mu+r\sigma^2}{1+\sigma^2}s+\left(\frac{\mu+r\sigma^2}{1+\sigma^2}\right)^2-\left(\frac{\mu+r\sigma^2}{1+\sigma^2}\right)^2\right]+C_1 \\
&= \frac{1+\sigma^2}{2\sigma^2}\left(s-\frac{\mu+r\sigma^2}{1+\sigma^2}\right)^2-\frac{1+\sigma^2}{2\sigma^2}\left(\frac{1+r\sigma^2}{1+\sigma^2}\right)^2+C_1 \\
&= \frac{1+\sigma^2}{2\sigma^2}\left(s-\frac{(\mu+r\sigma^2)}{1+\sigma^2}\right)^2+C_2.
\end{aligned}$$

Here $C_1$ and $C_2$ do not depend on $s$.

# Example (Cont'd)

- Hence,

$$f_{S|R}(s|r) = C \exp\left\{ \frac{-\left[s - \frac{(\mu + r\sigma^2)}{1+\sigma^2}\right]^2}{2\left(\frac{\sigma^2}{1+\sigma^2}\right)} \right\}.$$

  Here $C$ does not depend on $s$.

- Thus, the conditional distribution of $S$, given that $r$ is received, is normal with mean and variance now given by

$$E[S|R = r] = \frac{\mu + r\sigma^2}{1+\sigma^2}, \quad \text{Var}(S|R = r) = \frac{\sigma^2}{1+\sigma^2}.$$

  By the proposition, given that the value received is $r$, the best estimate, in the sense of minimizing the mean square error, for the signal sent is

$$E[S|R = r] = \frac{1}{1+\sigma^2}\mu + \frac{\sigma^2}{1+\sigma^2} r.$$

# Example

- In digital signal processing, raw continuous analog data $X$ must be quantized, or discretized, in order to obtain a digital representation.

  In order to quantize the raw data $X$:
  - An increasing set of numbers $a_i$, $i = 0, \pm 1, \pm 2, \ldots$, such that $\lim_{i \to +\infty} a_i = \infty$ and $\lim_{i \to -\infty} a_i = -\infty$ is fixed;
  - The raw data are quantized according to the interval $(a_i, a_{i+1}]$ in which $X$ lies.

  Let $y_i$ be the discretized value when $X \in (a_i, a_{i+1}]$.

  Let $Y$ denote the observed discretized value:

$$Y = y_i, \text{ if } a_i < X \leq a_{i+1}.$$

  The distribution of $Y$ is given by

$$P\{Y = y_i\} = F_X(a_{i+1}) - F_X(a_i).$$

# Example (Cont'd)

- Suppose now that we want to choose the values $y_i$, $i = 0, \pm 1, \pm 2, \ldots$ so as to minimize

$$E[(X - Y)^2],$$

the expected mean square difference between the raw data and their quantized version.

(a) Find the optimal values $y_i$, $i = 0, \pm 1, \ldots$.
For the optimal quantizer $Y$, show that:

(b) $E[Y] = E[X]$, so the mean square error quantizer preserves the input mean;

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

# Example (Cont'd)

(a) For any quantizer $Y$, upon conditioning on the value of $Y$, we obtain

$$E[(X-Y)^2] = \sum_i E[(X-y_i)^2 | a_i < X \leq a_{i+1}] P\{a_i < X \leq a_{i+1}\}.$$

Now, if we let $I = i$ if $a_i < X \leq a_{i+1}$, then

$$E[(X-y_i)^2 | a_i < X \leq a_{i+1}] = E[(X-y_i)^2 | I = i].$$

By the proposition, this quantity is minimized when

$$\begin{aligned} y_i &= E[X|I=i] = E[X|a_i < X \leq a_{i+1}] \\ &= \int_{a_i}^{a_{i+1}} \frac{x f_X(x) dx}{F_X(a_{i+1}) - F_X(a_i)}. \end{aligned}$$

Since the optimal quantizer is given by $Y = E[X|I]$, we get:

(b) $E[Y] = E[X]$;

(c) $\mathrm{Var}(X) = E[\mathrm{Var}(X|I)] + \mathrm{Var}(E[X|I]) =$
$E[E[(X-Y)^2|I]] + \mathrm{Var}(Y) = E[(X-Y)^2] + \mathrm{Var}(Y)$.

# The Best Linear Predictor

- Sometimes, the joint probability distribution of $X$ and $Y$ is not known.
- If it is known, it may be that the calculation of $E[Y|X = x]$ is difficult.
- If, however, the means and variances of $X$ and $Y$ and the correlation of $X$ and $Y$ are known, then we can at least determine the best linear predictor of $Y$ with respect to $X$.
- To obtain the best linear predictor of $Y$ with respect to $X$, we need to choose $a$ and $b$ so as to minimize

$$E[(Y - (a + bX))^2].$$

- We have

$$\begin{aligned} &E[(Y - (a + bX))^2] \\ &= E[Y^2 - 2aY - 2bXY + a^2 + 2abX + b^2X^2] \\ &= E[Y^2] - 2aE[Y] - 2bE[XY] + a^2 + 2abE[X] + b^2E[X^2]. \end{aligned}$$

# The Best Linear Predictor (Cont'd)

- We found

$$E[(Y-(a+bX))^2] = E[Y^2]-2aE[Y]-2bE[XY]+a^2+2abE[X]+b^2E[X^2].$$

Taking partial derivatives, we obtain

$$\begin{aligned}\frac{\partial}{\partial a}E[(Y-a-bX)^2] &= -2E[Y]+2a+2bE[X];\\ \frac{\partial}{\partial b}E[(Y-a-bX)^2] &= -2E[XY]+2aE[X]+2bE[X^2].\end{aligned}$$

- Setting these to 0 and solving for $a$ and $b$ yields the solutions

$$\begin{aligned}b &= \frac{E[XY]-E[X]E[Y]}{E[X^2]-(E[X])^2} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x^2} = \rho\frac{\sigma_y}{\sigma_x};\\ a &= E[Y]-bE[X] = E[Y]-\frac{\rho\sigma_y E[X]}{\sigma_x}.\end{aligned}$$

Here $\rho = \mathrm{Cor}(X,Y)$, $\sigma_y^2 = \mathrm{Var}(Y)$ and $\sigma_x^2 = \mathrm{Var}(X)$.

- Thus, the best (in the sense of mean square error) linear predictor $Y$ with respect to $X$ is $\mu_y + \frac{\rho\sigma_y}{\sigma_x}(X-\mu_x)$, where $\mu_y = E[Y]$, $\mu_x = E[X]$.

# The Best Linear Predictor (Cont'd)

- The mean square error of this predictor is given by

$$
\begin{aligned}
&E[(Y-\mu_y-\rho\tfrac{\sigma_y}{\sigma_x}(X-\mu_x))^2]\\
&=E[(Y-\mu_y)^2]+\rho^2\tfrac{\sigma_y^2}{\sigma_x^2}E[(X-\mu_x)^2]-2\rho\tfrac{\sigma_y}{\sigma_x}E[(Y-\mu_y)(X-\mu_x)]\\
&=\sigma_y^2+\rho^2\sigma_y^2-2\rho^2\sigma_y^2\\
&=\sigma_y^2(1-\rho^2).
\end{aligned}
$$

- If $\rho$ is near $+1$ or $-1$, then the mean square error of the best linear predictor is near zero.

Subsection 7

## Moment Generating Functions

# Moment Generating Functions

- The **moment generating function** $M(t)$ of the random variable $X$ is defined for all real values of $t$ by

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \begin{cases} \sum_x e^{tx} p(x), & \text{if } X \text{ is discrete with mass function } p(x) \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx, & \text{if } X \text{ is continuous with density } f(x) \end{cases} \end{aligned}$$

- We call $M(t)$ the moment generating function because all of the moments of $X$ can be obtained by:
  - Successively differentiating $M(t)$;
  - Evaluating the result at $t = 0$.

# Interchanging Differentiation and Expectation

- For the first moment (i.e., the expectation), We have

$$M'(t) = \frac{d}{dt}E[e^{tX}] = E\left[\frac{d}{dt}(e^{tX})\right] = E[Xe^{tX}].$$

Here we have assumed that the interchange of the differentiation and expectation operators is legitimate, i.e., that

  - In the discrete case,

$$\frac{d}{dt}\left[\sum_x e^{tx}p(x)\right] = \sum_x \frac{d}{dt}[e^{tx}p(x)];$$

  - In the continuous case,

$$\frac{d}{dt}\left[\int e^{tx}f(x)dx\right] = \int \frac{d}{dt}[e^{tx}f(x)]dx.$$

- This assumption is valid for all of the distributions considered here.

# Computing Moments

- $M(t) = E[e^{tX}] \Rightarrow \frac{dM}{dt} = \frac{d}{dt}E[e^{tX}] = E\left[\frac{d}{dt}e^{tX}\right] = E[Xe^{tX}]$.
  Evaluating at $t = 0$, we obtain

$$M'(0) = E[X].$$

- Similarly,

$$\begin{aligned} M''(t) &= \frac{d}{dt}M'(t) = \frac{d}{dt}E[Xe^{tX}] \\ &= E\left[\frac{d}{dt}(Xe^{tX})\right] = E[X^2e^{tX}]. \end{aligned}$$

  Thus, $M''(0) = E[X^2]$.
- In general, the $n$th derivative of $M(t)$ is given by

$$M^{(n)}(t) = E[X^n e^{tX}], \quad n \geq 1.$$

  So, we get

$$M^{(n)}(0) = E[X^n], \quad n \geq 1.$$

# Binomial Distribution With Parameters $n$ and $p$

- If $X$ is a binomial random variable with parameters $n$ and $p$, then

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\ &= (pe^t + 1 - p)^n. \end{aligned}$$

  The last equality follows from the binomial theorem.

- Differentiation yields

$$M'(t) = n(pe^t + 1 - p)^{n-1} pe^t.$$

  Thus,

$$E[X] = M'(0) = np.$$

# Binomial Distribution With Parameters $n$ and $p$ (Cont'd)

- Differentiating a second time yields

$$\begin{aligned} M''(t) &= (n(pe^t + 1 - p)^{n-1} pe^t)' \\ &= n(n-1)(pe^t + 1 - p)^{n-2}(pe^t)^2 \\ &\quad + n(pe^t + 1 - p)^{n-1} pe^t. \end{aligned}$$

  So

$$E[X^2] = M''(0) = n(n-1)p^2 + np.$$

- The variance of $X$ is given by

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= np(1-p). \end{aligned}$$

# Poisson Distribution With Mean $\lambda$

- If $X$ is a Poisson random variable with parameter $\lambda$, then

$$\begin{aligned} M(t) &= E[e^{tX}] = \sum_{n=0}^{\infty} \frac{e^{tn} e^{-\lambda} \lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda e^t} = \exp\{\lambda(e^t - 1)\}. \end{aligned}$$

- Differentiation yields

$$\begin{aligned} M'(t) &= \lambda e^t \exp\{\lambda(e^t - 1)\} \\ M''(t) &= (\lambda e^t)^2 \exp\{\lambda(e^t - 1)\} + \lambda e^t \exp\{\lambda(e^t - 1)\}. \end{aligned}$$

- Thus,

$$\begin{aligned} E[X] &= M'(0) = \lambda; \\ E[X^2] &= M''(0) = \lambda^2 + \lambda; \\ \text{Var}(X) &= E[X^2] - (E[X])^2 = \lambda. \end{aligned}$$

# Exponential Distribution With Parameter $\lambda$

- We have

$$\begin{aligned} M(t) &= E[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx \\ &= \lambda\int_0^\infty e^{-(\lambda-t)x}dx = \frac{\lambda}{\lambda-t}, \quad \text{for } t<\lambda. \end{aligned}$$

- We note from this derivation that, for the exponential distribution, $M(t)$ is defined only for values of $t$ less than $\lambda$.
- Differentiation of $M(t)$ yields

$$M'(t) = \frac{\lambda}{(\lambda-t)^2}, \quad M''(t) = \frac{2\lambda}{(\lambda-t)^3}.$$

- Hence,

$$\begin{aligned} E[X] &= M'(0) = \frac{1}{\lambda}; \\ E[X^2] &= M''(0) = \frac{2}{\lambda^2}; \\ \text{Var}(X) &= E[X^2] - (E[X])^2 = \frac{1}{\lambda^2}. \end{aligned}$$

# Normal Distribution

- We first compute the moment generating function of a unit normal random variable $Z$, with parameters 0 and 1.

$$\begin{aligned} M_Z(t) &= E[e^{tZ}] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x^2-2tx)}{2}\right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-t)^2}{2} + \frac{t^2}{2}\right\} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2}. \end{aligned}$$

- Hence, the moment generating function of the unit normal random variable $Z$ is given by $M_Z(t) = e^{t^2/2}$.

# Normal Distribution (Cont'd)

- Now recall that $X = \mu + \sigma Z$ will have a normal distribution with parameters $\mu$ and $\sigma^2$ whenever $Z$ is a unit normal random variable.
- Hence, the moment generating function of $X$ is given by

$$\begin{aligned} M_X(t) &= E[e^{tX}] = E[e^{t(\mu+\sigma Z)}] \\ &= E[e^{t\mu}e^{t\sigma Z}] = e^{t\mu}E[e^{t\sigma Z}] \\ &= e^{t\mu}M_Z(t\sigma) = e^{t\mu}e^{(t\sigma)^2/2} \\ &= \exp\{\frac{\sigma^2 t^2}{2} + \mu t\}. \end{aligned}$$

- By differentiating, we obtain

$$\begin{aligned} M'_X(t) &= (\mu + t\sigma^2)\exp\{\frac{\sigma^2 t^2}{2} + \mu t\}; \\ M''_X(t) &= (\mu + t\sigma^2)^2\exp\{\frac{\sigma^2 t^2}{2} + \mu t\} + \sigma^2\exp\{\frac{\sigma^2 t^2}{2} + \mu t\}. \end{aligned}$$

- Thus,

$$E[X] = M'(0) = \mu; \quad E[X^2] = M''(0) = \mu^2 + \sigma^2;$$
$$\mathrm{Var}(X) = E[X^2] - E([X])^2 = \sigma^2.$$

# Generating Function of Sum of Independent Variables

- The moment generating function of the sum of independent random variables equals the product of the individual moment generating functions.
- Suppose that $X$ and $Y$ are independent and have moment generating functions $M_X(t)$ and $M_Y(t)$, respectively.
- Then $M_{X+Y}(t)$, the moment generating function of $X+Y$, is given by

$$\begin{aligned} M_{X+Y}(t) &= E[e^{t(X+Y)}] \\ &= E[e^{tX}e^{tY}] \\ &= E[e^{tX}]E[e^{tY}] \\ &= M_X(t)M_Y(t). \end{aligned}$$

The next-to-last equality follows from a previous proposition, since $X$ and $Y$ are independent.

# Moment Generating Functions and Distributions

- The moment generating function uniquely determines the distribution.

  That is, if $M_X(t)$ exists and is finite in some region about $t = 0$, then the distribution of $X$ is uniquely determined.
- For instance, we know that the binomial random variable with parameters $n$ and $p$ has moment generating function

$$(pe^t + 1 - p)^n.$$

Suppose

$$M_X(t) = \left(\frac{1}{2}\right)^{10} (e^t + 1)^{10} = \left(\frac{1}{2}e^t + 1 - \frac{1}{2}\right)^{10}.$$

It follows that $X$ is a binomial random variable with parameters 10 and $\frac{1}{2}$.

# Example

- Suppose that the moment generating function of a random variable $X$ is given by

$$M(t) = e^{3(e^t - 1)}.$$

What is $P\{X = 0\}$?

We know that $M(t) = e^{3(e^t-1)}$ is the moment generating function of a Poisson random variable with mean 3.

Hence, by the one-to-one correspondence between moment generating functions and distribution functions, it follows that $X$ must be a Poisson random variable with mean 3, i.e., $f_X(x) = e^{-3}\frac{3^x}{x!}$.

Thus, $P\{X = 0\} = e^{-3}$.

# Sums of Independent Binomial Random Variables

- If $X$ and $Y$ are independent binomial random variables with parameters $(n, p)$ and $(m, p)$, respectively, what is the distribution of $X + Y$?

  The moment generating function of $X + Y$ is given by

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= (pe^t + 1 - p)^n(pe^t + 1 - p)^m \\ &= (pe^t + 1 - p)^{m+n}. \end{aligned}$$

  However, $(pe^t + 1 - p)^{m+n}$ is the moment generating function of a binomial random variable having parameters $m + n$ and $p$.

  Thus, this must be the distribution of $X + Y$.

# Sums of Independent Poisson Random Variables

- Calculate the distribution of $X + Y$ when $X$ and $Y$ are independent Poisson random variables with means respectively $\lambda_1$ and $\lambda_2$.

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= \exp\{\lambda_1(e^t - 1)\}\exp\{\lambda_2(e^t - 1)\} \\ &= \exp\{(\lambda_1 + \lambda_2)(e^t - 1)\}. \end{aligned}$$

Hence, $X + Y$ is Poisson distributed with mean $\lambda_1 + \lambda_2$.

# Sums of Independent Normal Random Variables

- Show that if $X$ and $Y$ are independent normal random variables with respective parameters $(\mu_1, \sigma_1^2)$ and $(\mu_2, \sigma_2^2)$, then $X + Y$ is normal with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) M_Y(t) \\ &= \exp\left\{\frac{\sigma_1^2 t^2}{2} + \mu_1 t\right\} \exp\left\{\frac{\sigma_2^2 t^2}{2} + \mu_2 t\right\} \\ &= \exp\left\{\frac{(\sigma_1^2+\sigma_2^2) t^2}{2} + (\mu_1 + \mu_2) t\right\}. \end{aligned}$$

This is the moment generating function of a normal random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

The desired result follows because the moment generating function uniquely determines the distribution.

# Sum of a Random Number of Random Variables

- Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables.
- Let $N$ be a nonnegative, integer-valued random variable that is independent of the sequence $X_i$, $i \geq 1$.
- We compute the moment generating function of $Y = \sum_{i=1}^{N} X_i$.
- We first condition on $N$:

$$\begin{aligned} E[\exp\{t\sum_1^N X_i\}|N=n] &= E\left[\exp\{t\sum_1^n X_i\}|N=n\right] \\ &= E\left[\exp\{t\sum_1^n X_i\}\right] = [M_X(t)]^n. \end{aligned}$$

  Here $M_X(t) = E[e^{tX_i}]$.
- Hence, $E[e^{tY}|N] = (M_X(t))^N$.
  Thus, $M_Y(t) = E[(M_X(t))^N]$.

# Sum of a Random Number of Random Variables (Cont'd)

- We found $M_Y(t) = E[(M_X(t))^N]$.
- The moments of $Y$ can now be obtained upon differentiation, as follows:

$$M'_Y(t) = E[N(M_X(t))^{N-1} M'_X(t)].$$

- So

$$\begin{aligned} E[Y] &= M'_Y(0) \\ &= E[N(M_X(0))^{N-1} M'_X(0)] \\ &= E[NE[X]] \\ &= E[N]E[X]. \end{aligned}$$

# Sum of a Random Number of Random Variables (Cont'd)

- We have $M'_Y(t) = E[N(M_X(t))^{N-1}M'_X(t)]$.
- Hence

$$M''_Y(t) = E[N(N-1)(M_X(t))^{N-2}(M'_X(t))^2 + N(M_X(t))^{N-1}M''_X(t)].$$

- So

$$\begin{aligned} E[Y^2] &= M''_Y(0) \\ &= E[N(N-1)(E[X])^2 + NE[X^2]] \\ &= (E[X])^2(E[N^2] - E[N]) + E[N]E[X^2] \\ &= E[N](E[X^2] - (E[X])^2) + (E[X])^2E[N^2] \\ &= E[N]\mathrm{Var}(X) + (E[X])^2E[N^2]. \end{aligned}$$

- Hence, we have

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

# Example

- Let $Y$ denote a uniform random variable on $(0, 1)$.
- Suppose that, conditional on $Y = p$, the random variable $X$ has a binomial distribution with parameters $n$ and $p$.
- We showed that $X$ is equally likely to take on any of the values $0, 1, \ldots, n$.
- Now, we establish this result by using moment generating functions.
- To compute the moment generating function of $X$, we start by conditioning on the value of $Y$.
- Using the formula for the binomial moment generating function gives

$$E[e^{tX} | Y = p] = (pe^t + 1 - p)^n.$$

# Example

- $Y$ is uniform on $(0, 1)$.
- Taking expectations, we obtain

$$\begin{aligned} E[e^{tX}] &= \int_0^1 (pe^t + 1 - p)^n dp \\ &= \frac{1}{e^t - 1} \int_1^{e^t} y^n dy \text{ (by the substitution } y = pe^t + 1 - p) \\ &= \frac{1}{n+1} \frac{e^{t(n+1)} - 1}{e^t - 1} \\ &= \frac{1}{n+1}(1 + e^t + e^{2t} + \cdots + e^{nt}). \end{aligned}$$

- This is the moment generating function of a random variable that is equally likely to be any of the values $0, 1, \ldots, n$.
- The desired result follows from the fact that the moment generating function of a random variable uniquely determines its distribution.

# Joint Moment Generating Functions

- For any $n$ random variables $X_1, \ldots, X_n$, the joint moment generating function, $M(t_1, \ldots, t_n)$, is defined, for all real values of $t_1, \ldots, t_n$, by

$$M(t_1, \ldots, t_n) = E[e^{t_1 X_1 + \cdots + t_n X_n}].$$

- The individual moment generating functions can be obtained from $M(t_1, \ldots, t_n)$ by letting all but one of the $t_j$'s be 0.
- That is,

$$M_{X_i}(t) = E[e^{tX_i}] = M(0, \ldots, 0, t, 0, \ldots, 0),$$

where the $t$ is in the $i$th place.

# Uniqueness and Independence

- It can be proven that the joint moment generating function $M(t_1, \ldots, t_n)$ uniquely determines the joint distribution of $X_1, \ldots, X_n$.
- This result can then be used to prove that the $n$ random variables $X_1, \ldots, X_n$ are independent if and only if

$$M(t_1, \ldots, t_n) = M_{X_1}(t_1) \cdots M_{X_n}(t_n).$$

- Suppose, first, that the $n$ random variables are independent.

  Then we have:

$$\begin{aligned} M(t_1, \ldots, t_n) &= E[e^{t_1 X_1 + \cdots + t_n X_n}] \\ &= E[e^{t_1 X_1} \cdots e^{t_n X_n}] \\ &= E[e^{t_1 X_1}] \cdots E[e^{t_n X_n}] \text{ (by independence)} \\ &= M_{X_1}(t_1) \cdots M_{X_n}(t_n). \end{aligned}$$

# Uniqueness and Independence (Cont'd)

- In the other direction, suppose the equation is satisfied.

  Then the joint moment generating function $M(t_1, \ldots, t_n)$ is the same as the joint moment generating function of $n$ independent random variables, the $i$th of which has the same distribution as $X_i$.

  But the joint moment generating function uniquely determines the joint distribution.

  Thus, this must be the joint distribution.

  Hence, the random variables are independent.

# Example

- Let $X$ and $Y$ be independent normal random variables, each with mean $\mu$ and variance $\sigma^2$.
- We showed that $X+Y$ and $X-Y$ are independent.
- We now establish this result by computing their joint moment generating function:

$$\begin{aligned} E[e^{t(X+Y)+s(X-Y)}] &= E[e^{(t+s)X+(t-s)Y}] \\ &= E[e^{(t+s)X}]E[e^{(t-s)Y}] \\ &= e^{\mu(t+s)+\sigma^2(t+s)^2/2}e^{\mu(t-s)+\sigma^2(t-s)^2/2} \\ &= e^{2\mu t+\sigma^2 t^2}e^{\sigma^2 s^2}. \end{aligned}$$

This is the joint moment generating function of the sum of two independent normal random variables:

- One with with mean $2\mu$ and variance $2\sigma^2$;
- One with mean 0 and variance $2\sigma^2$.

The joint moment generating function uniquely determines the joint distribution. So $X+Y$ and $X-Y$ are independent normal.

# Example

- Suppose that the number of events that occur is a Poisson random variable with mean $\lambda$ and that each event is independently counted with probability $p$.

  Show that the number of counted events and the number of uncounted events are independent Poisson random variables with respective means $\lambda p$ and $\lambda(1-p)$.

  - Let $X$ denote the total number of events.
  - Let $X_c$ denote the number of them that are counted.

  We start by conditioning on $X$ to obtain

$$\begin{aligned} E[e^{sX_c+t(X-X_c)}|X=n] &= e^{tn}E[e^{(s-t)X_c}|X=n] \\ &= e^{tn}(pe^{s-t}+1-p)^n \\ &= (p^{es}+(1-p)e^t)^n. \end{aligned}$$

  The last equation follows because, conditional on $X=n$, $X_c$ is a binomial random variable with parameters $n$ and $p$.

# Example

- Hence,
$$E[e^{sX_c+t(X-X_c)}|X] = (pe^s + (1-p)e^t)^X.$$
- Taking expectations of both sides of this equation yields
$$E[e^{sX_c+t(X-X_c)}] = E[(pe^s + (1-p)e^t)^X].$$
- Now, since $X$ is Poisson with mean $\lambda$, $E[e^{tX}] = e^{\lambda(e^t-1)}$.
- Therefore, for any $a > 0$, by letting $a = e^t$, we get $E[a^X] = e^{\lambda(a-1)}$.
- Thus,
$$\begin{aligned} E[e^{sX_c+t(X-X_c)}] &= e^{\lambda(pe^s+(1-p)e^t-1)} \\ &= e^{\lambda pe^s-\lambda p+\lambda e^t-\lambda pe^t-\lambda+\lambda p} \\ &= e^{\lambda p(e^s-1)}e^{\lambda(1-p)(e^t-1)}. \end{aligned}$$

This is the joint moment generating function of independent Poisson random variables with respective means $\lambda p$ and $\lambda(1-p)$.

Subsection 8

## Additional Properties of Normal Random Variables

# The Multivariate Normal Distribution

- Let $Z_1, \ldots, Z_n$ be $n$ independent unit normal random variables.
- If, for some constants $a_{ij}$, $1 \leq i \leq m$, $1 \leq j \leq n$, and $\mu_i$, $1 \leq i \leq m$,

$$\begin{aligned} X_1 &= a_{11}Z_1 + \cdots + a_{1n}Z_n + \mu_1 \\ X_2 &= a_{21}Z_1 + \cdots + a_{2n}Z_n + \mu_2 \\ &\vdots \\ X_m &= a_{m1}Z_1 + \cdots + a_{mn}Z_n + \mu_m, \end{aligned}$$

  then the random variables $X_1, \ldots, X_m$ are said to have a **multivariate normal distribution**.
- From the fact that the sum of independent normal random variables is itself a normal random variable, it follows that each $X_i$ is a normal random variable with mean and variance given, respectively, by

$$E[X_i] = \mu_i, \quad \text{Var}(X_i) = \sum_{j=1}^{n} a_{ij}^2.$$

# Joint Moment Generating Function of Normals

- Consider

$$M(t_1, \ldots, t_m) = E[\exp\{t_1 X_1 + \cdots + t_m X_m\}]$$

the joint moment generating function of $X_1, \ldots, X_m$.

- Since $\sum_{i=1}^m t_i X_i$ is itself a linear combination of the independent normal random variables $Z_1, \ldots, Z_n$, it is also normally distributed.
- Its mean and variance are

$$\begin{aligned} E[\textstyle\sum_{i=1}^m t_i X_i] &= \textstyle\sum_{i=1}^m t_i \mu_i; \\ \mathrm{Var}(\textstyle\sum_{i=1}^m t_i X_i) &= \mathrm{Cov}(\textstyle\sum_{i=1}^m t_i X_i, \sum_{j=1}^m t_j X_j) \\ &= \textstyle\sum_{i=1}^m \sum_{j=1}^m t_i t_j \mathrm{Cov}(X_i, X_j). \end{aligned}$$

# Joint Moment Generating Function of Normals (Cont'd)

- If $Y$ is a normal random variable with mean $\mu$ and variance $\sigma^2$, then

$$E[e^Y] = M_Y(t)|_{t=1} = e^{\mu+\sigma^2/2}.$$

- Thus,

$$M(t_1, \ldots, t_m) = \exp\left\{\sum_{i=1}^{m} t_i \mu_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} t_i t_j \mathrm{Cov}(X_i, X_j)\right\}.$$

- This shows that the joint distribution of $X_1, \ldots, X_m$ is completely determined from a knowledge of the values of

$$E[X_i] \quad \text{and} \quad \mathrm{Cov}(X_i, X_j), \quad i, j = 1, \ldots, m.$$

# Example

- Find $P(X < Y)$ for bivariate normal random variables $X$ and $Y$ having parameters

$$\mu_x = E[X],\ \mu_y = E[Y],\ \sigma_x^2 = \text{Var}(X),\ \sigma_y^2 = \text{Var}(Y),\ \rho = \text{Corr}(X, Y).$$

- $X - Y$ is normal with mean

$$\begin{aligned} E[X - Y] &= \mu_x - \mu_y; \\ \text{Var}(X - Y) &= \text{Var}(X) + \text{Var}(-Y) + 2\text{Cov}(X, -Y) \\ &= \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y. \end{aligned}$$

- Thus,

$$\begin{aligned} P\{X < Y\} &= P\{X - Y < 0\} \\ &= P\left\{ \frac{X-Y-(\mu_x-\mu_y)}{\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}} < \frac{-(\mu_x-\mu_y)}{\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}} \right\} \\ &= \Phi\left( \frac{\mu_y-\mu_x}{\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}} \right). \end{aligned}$$

# Example

- Suppose that:
  - The conditional distribution of $X$, given that $\Theta = \theta$, is normal with mean $\theta$ and variance 1;
  - $\Theta$ itself is a normal random variable with mean $\mu$ and variance $\sigma^2$.

  Find the conditional distribution of $\Theta$ given that $X = x$.
- We show that $X, \Theta$ has a bivariate normal distribution.
- The joint density function of $X, \Theta$ can be written as

$$f_{X,\Theta}(x, \theta) = f_{X|\Theta}(x|\theta) f_{\Theta}(\theta),$$

  where $f_{X|\Theta}(x|\theta)$ is a normal density with mean $\theta$ and variance 1.
- Let $Z$ be a standard normal random variable that is independent of $\Theta$.

  Then the conditional distribution of $Z + \Theta$, given that $\Theta = \theta$, is also normal with mean $\theta$ and variance 1.

  Thus, the joint density of $Z + \Theta$, $\Theta$ is the same as that of $X, \Theta$.

# Example (Cont'd)

- The joint density of $Z+\Theta$ and $\Theta$ is clearly bivariate normal (both are linear combinations of the independent normal $Z$ and $\Theta$).
- Hence, $X, \Theta$ has a bivariate normal distribution.
- Now,

$$\begin{aligned} E[X] &= E[Z+\Theta] = \mu; \\ \text{Var}(X) &= \text{Var}(Z+\Theta) = 1+\sigma^2; \\ \rho &= \text{Corr}(X,\Theta) = \text{Corr}(Z+\Theta,\Theta) \\ &= \frac{\text{Cov}(Z+\Theta,\Theta)}{\sqrt{\text{Var}(Z+\Theta)\text{Var}(\Theta)}} = \frac{\sigma}{\sqrt{1+\sigma^2}}. \end{aligned}$$

- The conditional distribution of $\Theta$, given $X=x$, is normal with

$$\begin{aligned} E[\Theta|X=x] &= E[\Theta] + \rho\sqrt{\frac{\text{Var}(\Theta)}{\text{Var}(X)}}(x-E[X]) \\ &= \mu + \frac{\sigma^2}{1+\sigma^2}(x-\mu); \\ \text{Var}(\Theta|X=x) &= \text{Var}(\Theta)(1-\rho^2) = \frac{\sigma^2}{1+\sigma^2}. \end{aligned}$$

# Joint Distribution of Sample Mean and Sample Variance

- Let $X_1, \ldots, X_n$ be independent normal random variables, each with mean $\mu$ and variance $\sigma^2$.
- Let $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ denote their sample mean.
- As a sum of independent normal random variables, $\overline{X}$ is also a normal random variable.

  As we have seen, $\overline{X}$ has expected value $\mu$ and variance $\frac{\sigma^2}{n}$.
- Recall that $\mathrm{Cov}(\overline{X}, X_i - \overline{X}) = 0$, $i = 1, \ldots, n$.
- Note $\overline{X}$, $X_1 - \overline{X}$, $X_2 - \overline{X}$, $\ldots$, $X_n - \overline{X}$ are all linear combinations of the independent standard normals $\frac{X_i - \mu}{\sigma}$, $i = 1, \ldots, n$.

  Thus, $\overline{X}$, $X_i - \overline{X}$, $i = 1, \ldots, n$ has a joint distribution that is multivariate normal.

# Sample Mean and Sample Variance (Cont'd)

- Let $Y$ be a normal random variable, with mean $\mu$ and variance $\frac{\sigma^2}{n}$, that is independent of the $X_i$, $i = 1, \ldots, n$.

  Then $Y, X_i - \overline{X}$, $i = 1, \ldots, n$ also has a multivariate normal distribution with the same expected values and covariances as the random variables $\overline{X}, X_i - \overline{X}$, $i = 1, \ldots, n$.
- But a multivariate normal distribution is determined completely by its expected values and covariances.

  Thus, $Y, X_i - \overline{X}$, $i = 1, \ldots, n$ and $\overline{X}, X_i - \overline{X}$, $i = 1, \ldots, n$ have the same joint distribution.
- This shows that $\overline{X}$ is independent of the sequence of deviations $X_i - \overline{X}$, $i = 1, \ldots, n$.
- $\overline{X}$ being independent of $X_i - \overline{X}$, $i = 1, \ldots, n$, it is also independent of the sample variance $S^2 \equiv \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$.

# Sample Mean and Sample Variance (Cont'd)

- Since we already know that $\overline{X}$ is normal with mean $\mu$ and variance $\frac{\sigma^2}{n}$, it remains only to determine the distribution of $S^2$.
- Recall, from a previous example, the algebraic identity

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i-\overline{X})^2 \\ &= \sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2.\end{aligned}$$

- Upon dividing by $\sigma^2$, we obtain

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2.$$

- Now, $\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2$ is the sum of the squares of $n$ independent standard normal random variables.
- This is called a **chi-squared random variable with $n$ degrees of freedom**.

# Moment Generating Function of Chi-Squared

- Compute the moment generating function of a chi-squared random variable with n degrees of freedom.
- Represent such a random variable as

$$Z_1^2 + \cdots + Z_n^2,$$

where $Z_1, \ldots, Z_n$ are independent standard normal random variables.

- Let M(t) be its moment generating function.
- Then, $M(t) = (E[e^{tZ^2}])^n$, where $Z$ is standard normal.
- Now,

$$\begin{aligned} E[e^{tZ^2}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx \quad (\sigma^2 = (1-2t)^{-1}) \\ &= \sigma = (1-2t)^{-1/2}. \end{aligned}$$

We used the fact that the normal density with mean 0 and variance $\sigma^2$ integrates to 1. Thus, $M(t) = (1-2t)^{-n/2}$.

# Sample Mean and Sample Variance (Cont'd)

- A chi-squared with $n$ degrees of freedom has moment generating function is $(1-2t)^{-n/2}$.
- But $\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$ is the square of a standard normal random variable.

  Hence, it is a chi-squared random variable with 1 degree of freedom.

  So it has moment generating function $(1-2t)^{-1/2}$.
- But, as we saw, the two random variables on the left side of the displayed equation above are independent.

  Hence, as the moment generating function of the sum of independent random variables is equal to the product of their individual moment generating functions:

$$E[e^{t(n-1)S^2/\sigma^2}](1-2t)^{-1/2} = (1-2t)^{-n/2};$$

$$E[e^{t(n-1)S^2/\sigma^2}] = (1-2t)^{-(n-1)/2}.$$

# Sample Mean and Sample Variance (Cont'd)

- But $(1-2t)^{-(n-1)/2}$ is the moment generating function of a chi-squared random variable with $n-1$ degrees of freedom.

  A moment generating function determines the distribution uniquely.

  We conclude that the distribution of $\frac{(n-1)S^2}{\sigma^2}$ must be a chi-squared with $n-1$ degrees of freedom.

## Proposition

If $X_1, \ldots, X_n$ are independent and identically distributed normal random variables with mean $\mu$ and variance $\sigma^2$, then:

- The sample mean $\overline{X}$ and the sample variance $S^2$ are independent;
- $\overline{X}$ is a normal random variable with mean $\mu$ and variance $\frac{\sigma^2}{n}$;
- $\frac{(n-1)S^2}{\sigma^2}$ is a chi-squared random variable with $n-1$ degrees of freedom.