Beta Distribution (Sec. 7.5).
We say a random variable $X$ has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ if it has density

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where

$B(\alpha, \beta)$ is the beta function defined as

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, dx$$

We have seen that

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

(see Feb. 13 lecture).

Also from the Feb. 13 lecture we have that if $X_1, \ldots, X_n$ are independent with

$$X_i \sim \text{Gamma}(\alpha_i, \lambda), \quad i = 1, \ldots, n$$
then
\[
\frac{X_i}{X_1 + \ldots + X_n} \sim \text{Beta} \left( r_i, \frac{n}{\sum_{j=1}^{n} r_j} \right)
\]
This is so since if we let
\[
Y_1 = \sum_{j \neq i} X_j, \quad Y_2 = X_i
\]
Then
\[
Y_1 \sim \text{Gamma} \left( \sum_{j \neq i} r_j, \lambda \right),
\]
\[
Y_2 \sim \text{Gamma} \left( r_i, \lambda \right)
\]
and \( Y_1 \) and \( Y_2 \) are independent. Then
\[
\frac{Y_2}{Y_1 + Y_2} \sim \text{Beta} \left( r_i, \frac{\sum_{j \neq i} r_j + r_i}{\lambda} \right)
\]
So we can view the Beta distribution as modelling the proportion of the total sum obtained from independent Gamma random variables (with the same 2nd parameter \( \lambda \)).

**Special Case**

If \( \alpha = 1 \) and \( \beta = 1 \) then \( X \) has a Uniform \((0,1)\) distribution.
Moments

For a fixed positive integer $k$, we have

$$E[X^k] = \int x^k \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} dx$$

$$= \frac{B(\alpha+\beta, \beta)}{B(\alpha, \beta)} \int_0^1 x^{\alpha+\beta-1}(1-x)^{\beta-1} dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+\beta)}$$

$$= \frac{(\alpha+\beta-1) \ldots \alpha}{(\alpha+\beta+\beta-1) \ldots (\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+\beta)}$$

$K=1$ : $E[X] = \frac{\alpha}{\alpha+\beta}$

$K=2$ : $E[X^2] = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$

Then

$$Var(X) = E[X^2] - E[X]^2$$

$$= \frac{(\alpha+1)\alpha(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta+1)(\alpha+\beta)}$$
\[
= \frac{(\alpha^2 + \alpha)(\alpha + \beta) - \alpha^2(\alpha + \beta + 1)}{(\alpha + \beta + 1)(\alpha + \beta)^2}
\]
\[
= \frac{\alpha \beta}{(\alpha + \beta + 1)(\alpha + \beta)^2} = \text{Var}(X)
\]

\[E(X) = \frac{1}{1+1} = \frac{1}{2}\]
\[\text{Var}(X) = \frac{(1)(1)}{(1+1+1)(1+1)^2} = \frac{1}{12}\]

**Conditional Expectation**

Let \(X\) be a random variable and \(Y\) a random vector.

**Conditional Distributions**

If \(X\) and \(Y\) are both discrete then the conditional pmf of \(X\) given \(Y = y\) gives the conditional probabilities
\[
P(X = x \mid Y = y), \text{ denoted by } p_{x \mid y}(x \mid y)
\]
We have

\[ p_{x|y}(x|y) = \frac{p(x = x, y = y)}{p(y = y)} \]

\[ = \frac{\varphi_{x,y}(x,y)}{p_y(y)} \]

where \( \varphi_{x,y} \) and \( p_y \) are the joint pmf of \( X \) and \( Y \), and the marginal pmf of \( Y \), respectively.

Similarly, if \( X \) and \( Y \) are both continuous with joint pdf \( f_{x,y}(x,y) \) and marginal pdf \( f_y(y) \) for \( Y \), then the conditional pdf of \( X \) given \( Y = y \) is

\[ f_{x|y}(x|y) = \frac{f_{x,y}(x,y)}{f_y(y)} \]

Mar. 3.

**Conditional Means and Variances**

The conditional mean of \( X \) given \( Y = y \) denoted \( \mathbb{E}(X|Y=y) \) is the mean of
the conditional distribution of $X$ given $Y = y$

$$E(X | Y = y) = \begin{cases} \sum_x x \cdot p_{x|y}(x|y) & \text{(discrete case)} \\ \int x \cdot f_{x|y}(x|y) \, dx & \text{(continuous case)} \end{cases}$$

Similarly, the conditional variance of $X$ given $Y = y$, denoted $\text{Var}(X | Y = y)$, is the variance of the conditional distribution of $X$ given $Y = y$. Both $E(X | Y = y)$ and $\text{Var}(X | Y = y)$ are functions of $y$. Since $Y$ is a random vector we can also view the conditional mean and variance as functions of the random vector $Y$. We write this as $E[X | Y]$ and $\text{Var}(X | Y)$. These quantities are random variables!

The law of total expectation says that if we average all the conditional expectations $E[X | Y = y]$ over all $y$ with respect to the distribution of $Y$ we get back $E[X]$.
Theorem  Law of Total Expectation  

\[ E[ E[X \mid Y] ] = E[X] \]

w.r.t. dist'n of \( Y \)  
w.r.t. dist'n of \( X \)

Proof (discrete case)

\[ E[ E[X \mid Y] ] = \sum_y E[X \mid Y = y] \cdot p_Y(y) \]

\[ = \sum_y \sum_x x \cdot p_{X \mid Y}(x \mid y) \cdot p_Y(y) \]

\[ = \sum_y \sum_x \frac{p_{X,Y}(x,y)}{p_Y(y)} \cdot p_Y(y) \]

\[ = \sum_x \sum_y p_{X,Y}(x,y) \]

\[ = \sum_x p_X(x) = E[X] \] .

Example  Mean of a Geometric (\( p \)) distribution

\( X \) is the trial where the first success occurred.

Let \( Y = \begin{cases} 1 & \text{if first trial is success} \\ 0 & \text{if first trial is failure} \end{cases} \)
Then by the law of total expectation,
\[ E(X) = E(X | Y=1)P(Y=1) + E(X | Y=0)P(Y=0) \]
\[ = (1)p + E(X | Y=0)(1-p) \]

\[ E(X | Y=0) = 1 + E(X) \]
(given first trial is failure, then \( X = 1 + X_1 \), where \( X_1 \) is the number of trials, starting with trial 2, until the first success. But \( X_1 \) and \( X \) both have the same distribution).

So \( E(X) = p + (1 + E(X)) (1-p) \)

\[ E(X) (1-(1-p)) = p + 1-p \]

\[ \Rightarrow E(X) = \frac{1}{p} \]

**Example** Branching Process

Let \( Y_{nj}, n=1,2,\ldots \)
\[ j=1,2,\ldots \]
be i.i.d. random variables on \( \{0,1,2,\ldots\} \) with common pmf \( f_Y \).

Let \( X_0 = 1 \)

Recursively, let
\[X_n = Y_{n-1,1} + \ldots + Y_{n-1,X_{n-1}}\]

\(Y_{n-1,j}\) is the "family size" of individual \(j\) in generation \(n-1\).

\(X_{n-1}\) is the number of individuals in generation \(n-1\).

\(X_n\) is the number of individuals in generation \(n\).

We wish to compute \(E[X_n]\).

Condition on \(X_{n-1}\). If \(X_{n-1} = k\) then \(X_n = Y_{n-1,1} + \ldots + Y_{n-1,k}\) and so

\[E[X_n | X_{n-1} = k] = k \mu\]

where \(\mu\) is the mean family size

\[\mu = \sum_{y=0}^{\infty} y f_Y(y)\]

and \(E[X_n | X_{n-1}] = \mu X_{n-1}\)

So

\[E[X_n] = E[E[X_n | X_{n-1}]]\]

\[= E[E[X_n | X_{n-1}]]\]

\[= E[E[X_n | X_{n-1}]]\]

\[= \mu E[X_n]\]

\[= \mu^n E[X_0]\]

\[= \mu^n\]
We have seen that we can compute \( E(X) \) as \( E(E[X|Y]) = E(X) \).

Can we similarly compute \( \text{Var}(X) \) by conditioning?

In particular, can we average all the conditional variances \( \text{Var}(X|Y=y) \) over \( y \) with respect to the distribution of \( Y \)?

Note that \( \text{Var}(X|Y=y) = E(X^2|Y=y) - E(X|Y=y)^2 \)

so as a random variable

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

Then

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

w.r.t. dist'n of \( Y \)

by law of total expectation

So \( E[\text{Var}(X|Y)] \) does not in fact give us \( \text{Var}(X) \).
But now consider
\[
\text{Var}(E(X|Y)) = E(E(X|Y)^2) - E(E(X|Y))^2
\]
\[
= E(E(X|Y)^2) - E(X)^2
\]
by law of total expectation.

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

Example Branching Process

\[X_0 = 1\]
\[X_n = \text{population size of generation } n.\]
Let \(\sigma^2\) be the family size variance
and \(\mu = \text{mean family size}\)
Let us compute \(\text{Var}(X_n)\).
First, given \(X_{n-1} = k\) we have
\[E(X_n | X_{n-1} = k) = \mu k\]
\[\text{Var}(X_n | X_{n-1} = k) = k \sigma^2\]
variance of the sum of \(k\) independent family sizes.

So \(E(X_n | X_{n-1}) = \mu X_{n-1}\), \(\text{Var}(X_n | X_{n-1}) = \sigma^2 X_{n-1}\)
$\sum_0 \text{Var}(X_n) = E[\sigma^2 X_{n-1}] + \text{Var}(\mu X_{n-1})$

$= \sigma^2 \mu^{n-1} + \mu^2 \text{Var}(X_{n-1})$

$= \sigma^2 \mu^{n-1} + \mu^2 (\sigma^2 \mu^{n-2} + \mu^2 \text{Var}(X_{n-2}))$

$= \sigma^2 (\mu^{n-1} + \mu^n) + \mu^4 \text{Var}(X_{n-2})$

$= \sigma^2 (\mu^{n-1} + \mu^n) + \mu^4 (\sigma^2 \mu^{n-3} + \mu^2 \text{Var}(X_{n-3}))$

$= \sigma^2 (\mu^{n-1} + \mu^n + \mu^{n+1}) + \mu^6 \text{Var}(X_{n-3})$

$: \vdots$

$= \sigma^2 (\mu^{n-1} + \mu^n + \cdots + \mu^{2n-2}) + \mu^{2n} \text{Var}(X_0)$

$= \sigma^2 \mu^{n-1} (1 + \mu + \cdots + \mu^{n-1})$

$= \begin{cases} 
\sigma^2 \mu^{n-1} \frac{1 - \mu^n}{1 - \mu} & \mu \neq 1 \\
\pi \sigma^2 & \mu = 1
\end{cases}$

Note if $\mu < 1$ then $\text{Var}(X) \to 0$ as $n \to \infty$

**Example**

Suppose we have $n$ coins, one of which is fair. The others are biased. Each coin is flipped once. What is the probability that the number of heads is even?
Condition on outcome of fair coin:
Let \( A = \text{"# of heads is even"} \). 
\[
P(A) = E[I_A] \quad \text{where}
\]
\[
I_A = \begin{cases} 
1 & \text{if } A \text{ occurs} \\
0 & \text{if } A \text{ does not occur}
\end{cases}
\]
Let \( Y = \begin{cases} 
1 & \text{if fair coin is heads} \\
0 & \text{if fair coin is tails}
\end{cases} \)

Then
\[
E[I_A] = E[I_A \mid Y = 1] \left( \frac{1}{2} \right) + E[I_A \mid Y = 0] \left( \frac{1}{2} \right)
\]
\[
= \frac{1}{2} \left( E[I_A \mid Y = 1] + E[I_A \mid Y = 0] \right)
\]
\[
= \frac{1}{2} \left( P(\text{# of heads among biased coins is odd}) + P(\text{# of heads among biased coins is even}) \right)
\]
\[
= \frac{1}{2}
\]