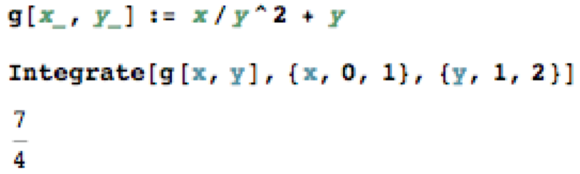
STT 351 HW5
Due at the close of class 10 - 8 - 08.

Define function x/y^2 + y for 0 < x < 1 and 1 < y < 2.

a. Integrate the above function over the indicated domain of x, y values.



```
g[x_, y_] := x / y^2 + y

Integrate[g[x, y], {x, 0, 1}, {y, 1, 2}]
```

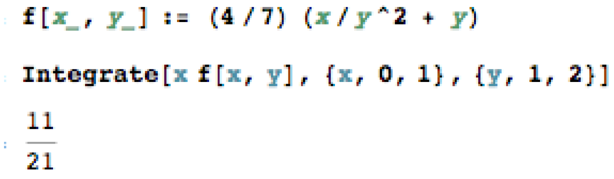
$$\frac{7}{4}$$

b. From the above, determine the probability density f(x, y) (or $f_{X,Y}(x,\ y)$) that is a constant multiple of the function given there.

f(x, y) := $\frac{4}{7}$(x/y^2 + y) (non-negative and integraltes to one)

Determine

c. E X = $\int\int x\, f(x,\ y)\, dx\, dy$

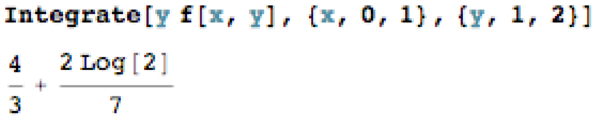


```
f[x_, y_] := (4 / 7) (x / y^2 + y)

Integrate[x f[x, y], {x, 0, 1}, {y, 1, 2}]
```

$$\frac{11}{21}$$

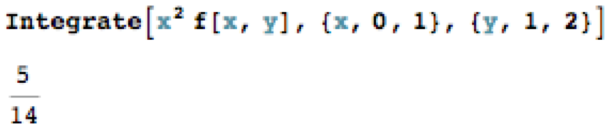
d. E Y



```
Integrate[y f[x, y], {x, 0, 1}, {y, 1, 2}]
```

$$\frac{4}{3}+\frac{2\,\text{Log}[2]}{7}$$

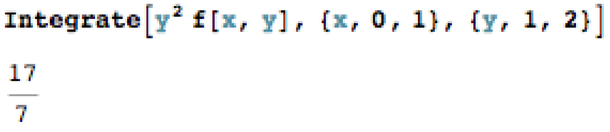
e. E X^2 = $\int\int x^2 f(x,\ y)\, dx\, dy$



```
Integrate[x^2 f[x, y], {x, 0, 1}, {y, 1, 2}]
```

$$\frac{5}{14}$$

f. E Y^2



```
Integrate[y^2 f[x, y], {x, 0, 1}, {y, 1, 2}]
```

$$\frac{17}{7}$$

g. E (XY) = $\int\int$xy $f(x,\ y)\, dx\, dy$

```
Integrate[x y f[x, y], {x, 0, 1}, {y, 1, 2}]
```

$$\frac{2}{21}\,(7+\text{Log}[4])$$

h. Var X

$5/14 - (11/21)^2$

i. $\sigma_X = \text{sd X}$

root of (h)

j. Var Y

$17/7 - (4/3 + 2\,(\text{Log}[2])\ /\ 7)^2$

k. $\sigma_Y = \text{sd Y}$

root of (j)

l. Covariance of X with Y defined by E(XY) - (E X)(E Y)

$\frac{2}{21}\,(7+\text{Log}[4])$ - (11/21) (4/3 + 2 (Log[2]) / 7)

m. Covariance of X with X

(7/4) - (11/21) (11/21) (just Var X)

n. Correlation between X, Y defined by $\dfrac{E\,XY - (E\,X)\,(E\,Y)}{\sigma_X\ \ \sigma_Y}$

$$\frac{\frac{2}{21}\,(7+\text{Log}[4]) - (11/21)\,(4/3 + 2\,(\text{Log}[2])\,/\,7)}{\sqrt{5/14-(11/21)^2}\ \sqrt{17/7-(4/3+2\,(\text{Log}[2])\,/\,7)^2}}$$

o. marginal density for X defined by $f_X(x) = \int_1^2 f(x,\ y)\,dy$

```
Integrate[f[x, y], {y, 1, 2}]
```

$$\frac{2\,(3+x)}{7}$$

for 0 < x < 1.

p. conditional density of y GIVEN x defined $f_{y\,|\,x}(y) = \dfrac{f(x,y)}{f_X(x)}$

```
f[x, y] / Integrate[f[x, y], {y, 1, 2}]
```

$$\frac{2\left(\frac{x}{y^2}+y\right)}{3+x}$$

q. conditional mean of y GIVEN x defined $E\big(Y \mid X = x\big) = \int y\, f_{y\mid x}(y)\, \mathbb{d}y$

$$\text{Integrate}\left[y\,\frac{2\left(\frac{x}{y^2}+y\right)}{3+x},\ \{y, 1, 2\}\right]$$

$$\frac{14 + x\,\text{Log}[64]}{9+3\,x}$$

r. using the definitions, prove that in general E Y = E (E(Y | X)), i.e.

$$\underbrace{\int y\, f_Y(y)\, \mathbb{d}y}_{\text{E Y}} = \int \big(\underbrace{\int y\, f_{y\mid x}(y)\ \mathbb{d}y}_{\text{E(Y | X = x)}}\big)\, f_X(x)\, \mathbb{d}\,\text{x}$$

Let's again calculate E Y directly from the joint density:

```
Integrate[y f[x, y], {x, 0, 1}, {y, 1, 2}]
```

$$\frac{4}{3}+\frac{2\,\text{Log}[2]}{7}$$

For comparison, calculate E Y from the marginal density for Y. In *Mathematica* it is neither necessary nor convenient to use the subscripted notation $f_Y(y)$for this density. We'll just have to pay attention when using f[y] to denote this density in *Mathematica*.

```
f[x_, y_] := 4 / 7 (x / y^2 + y)
```

```
f[y_] := Integrate[f[x, y], {x, 0, 1}]
```

```
f[y]
```

$$\frac{2}{7\,y^2}+\frac{4\,y}{7}$$

```
Integrate[y f[y], {y, 1, 2}]
```

$$\frac{4}{3}+\frac{2\,\text{Log}[2]}{7}$$

Once again, we obtain the same expectation E Y by either of the two methods. But there is even a third way to calculate E Y. It is to:

For each value X = x calculate the conditional expectation of Y for that value x. That produces a function of x denoted E(Y | X = x). This fuction of x, when evaluated at random variable X is denoted E(Y | X). The expectation of this r.v. is written E E(Y | X) and is equal to E Y. This may seem puzzling because we are, in the final step, calculating E Y (a y-

integral) as an x-integral

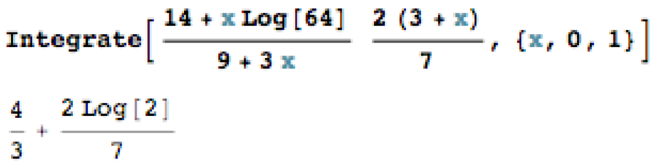
$$\mathbf{E\ Y = E\ E(Y \mid X) = \int dx\ f_X(x)\ [\ E(Y \mid X = x)\ ].}$$

You can see why it works by carefully looking at the following. I'll use some additional brackets **[** E[Y | X = x] **]** to keep track of that portion of the integrand.

$$E[Y \mid X\ ] = \int dx\ f_X(x)\ [\ E(Y \mid X = x)\ ] = \int dx\ f_X(x)\ [\ \int dy\ \ y\ f_{Y \mid X}(y \mid x)\ ]$$

$$= \int dx\ f_X(x)\ [\ \int dy\ \ y\ \frac{f_{X,Y}(x,y)}{f_{X,}(x)}\ ] = \int dx\ f_X(x)\ [\ \int dy\ \ y\ \frac{f_{X,Y}(x,y)}{f_{X,}(x)}\ ]\ = E\ Y$$

Here is the calculation of E Y done htis third way:



```
Integrate[(14 + x Log[64])/(9 + 3 x) (2 (3 + x))/7, {x, 0, 1}]

4/3 + (2 Log[2])/7
```

E E(Y | X) above.

The whole idea of E Y = E E(Y | X) is really very simple. Say I want to calculate the mean of Y = income next year taking into account X = income this year. I can calculate the mean of Y specific to each possible value of X = x then weigh these according to the probabilities of these different x.

Some motivational remarks.

Why study such a thing as E{Y | X) anyway?

Statistics is about expoiting simple relation ships of this kind. for example, there is very general "Decomposition of Variance" formula we can easily prove once we have the notios of conditional expectation and conditional variance in hand. It is:

$$\mathbf{Var\ Y = E\ Var(Y \mid X) + Var(\ E(Y \mid X)\ )}$$

I can briefly outline for you how this
"decomposition of variance Y in terms of X"
informs us how to design more efficient sampling methods.

Example: You have a budget of n = 100 samples from a population that is 34% Hispanic. You can of course ignore that fact and just sample 100 people at random from your combined population of Hispanics and nonHispanics.

If you are interested in overall population mean income $\mu_Y = \mathrm{E}\ \mathrm{Y}$, you need to know that the variance of your estimate is

$$\mathbf{Var}\ \overline{Y} = \frac{\mathbf{Var}\,Y}{100} = \frac{\sigma_Y{}^2}{100}.$$

Using the above "decomposition of variance" formula we can improve upon our sampling effort.

If you can manage it, instead sample 34 (at random) from the subpopulation of Hispanics and 66 from the subpopulation of non-Hispanics. If you do this, then your overall sample mean Y^* will have E Y^* = overall population mean and Var $Y^* = \frac{E(\mathrm{Var}(Y\,|\,X))}{100}$ which is only one part of $\frac{\mathrm{Var}\,Y}{100}$ as a consequence of the formula above.

So it is best to divide the sample of 100, sampling 34 at random from Hispanics and 66 at random from non-Hispanics, if possible, rather than to sample all 100 at random from the combined population!

The part of variance that dropped away when we chose to sample the "strata" of Hispanics vs nonHispanics separately is $\frac{\mathrm{Var}(E(Y\,|\,X))}{100}$. So sampling strata reduces variance to the extent that Hispanics have a different mean income than the nonHispanics (i.e. E(Y | X = x) varies with X = 1 for Hispanics and X = 0 for non-Hispanics).

These insights are possible precisely because we have the concepts of conditional mean and conditional variance and the decomposition of variance which follows.