Homework_13

April 28, 2025

Probability for Data Science

UC Berkeley, Spring 2025

Michael Xiao and Ani Adhikari

1 Homework 13 (Due Monday, April 28th at 5 PM)

[41]:

```
import warnings
warnings.filterwarnings('ignore')

from prob140 import *
from datascience import *
import numpy as np
from scipy import stats

import matplotlib.pyplot as plt
%matplotlib inline
import matplotlib
matplotlib.style.use('fivethirtyeight')
```

1.0.1 How to Do Your Homework

The point of homework is for you to try your hand at using what you've learned in class. The steps to follow:

- Go to lecture and sections, and also go over the relevant text sections before starting on the homework. This will remind you what was covered in class, and the text will typically contain examples not covered in lecture. The weekly Study Guide will list what you should read.
- Work on some of the practice problems before starting on the homework.
- Attempt the homework problems by yourself with the text, section work, and practice materials all at hand. Sometimes the week's lab will help as well. The two steps above will help this step go faster and be more fruitful.
- At this point, seek help if you need it. Don't ask how to do the problem — ask how to get started, or for a nudge to get you past where you are stuck. Always say what you have already tried. That helps us help you more effectively.

- For a good measure of your understanding, keep track of the fraction of the homework you can do by yourself or with minimal help. It's a better measure than your homework score, and only you can measure it.

1.0.2 Rules for Homework

- Every answer should contain a calculation or reasoning. For example, a calculation such as $(1/3)(0.8) + (2/3)(0.7)$ or `sum([(1/3)*0.8, (2/3)*0.7])`is fine without further explanation or simplification. If we want you to simplify, we'll ask you to. But just $\binom{5}{2}$ by itself is not fine; write "we want any 2 out of the 5 frogs and they can appear in any order" or whatever reasoning you used. Reasoning can be brief and abbreviated, e.g. "product rule" or "not mutually exclusive."
- You may consult others (see "How to Do Your Homework" above) but you must write up your own answers using your own words, notation, and sequence of steps.
- We'll be using Gradescope. You must submit the homework according to the instructions at the end of homework set.

1.1 We will not grade assignments which do not have pages correctly selected for each question.

1.2 1. Overlapping Counts

Consider a sequence of i.i.d. Bernoulli (p) trials. Consider the three variables X, Y, and V defined by:

- X is the number of successes in trials 1 through 100
- Y is the number of successes in trials 51 through 100
- V is the number of successes in trials 51 through 150

a) For each of X, Y, and V, say what the distribution is and provide the parameters.

b) Fix k in the range $0, 1, \ldots, 100$ and find the conditional distribution of Y given $X = k$. Recognize this as a famous one and provide the parameters.

c) Find the least squares predictor of Y based on X and say whether it is a linear function of X. (If it is, then the best linear predictor is in fact the best among all predictors.) Find $Var(Y \mid X)$.

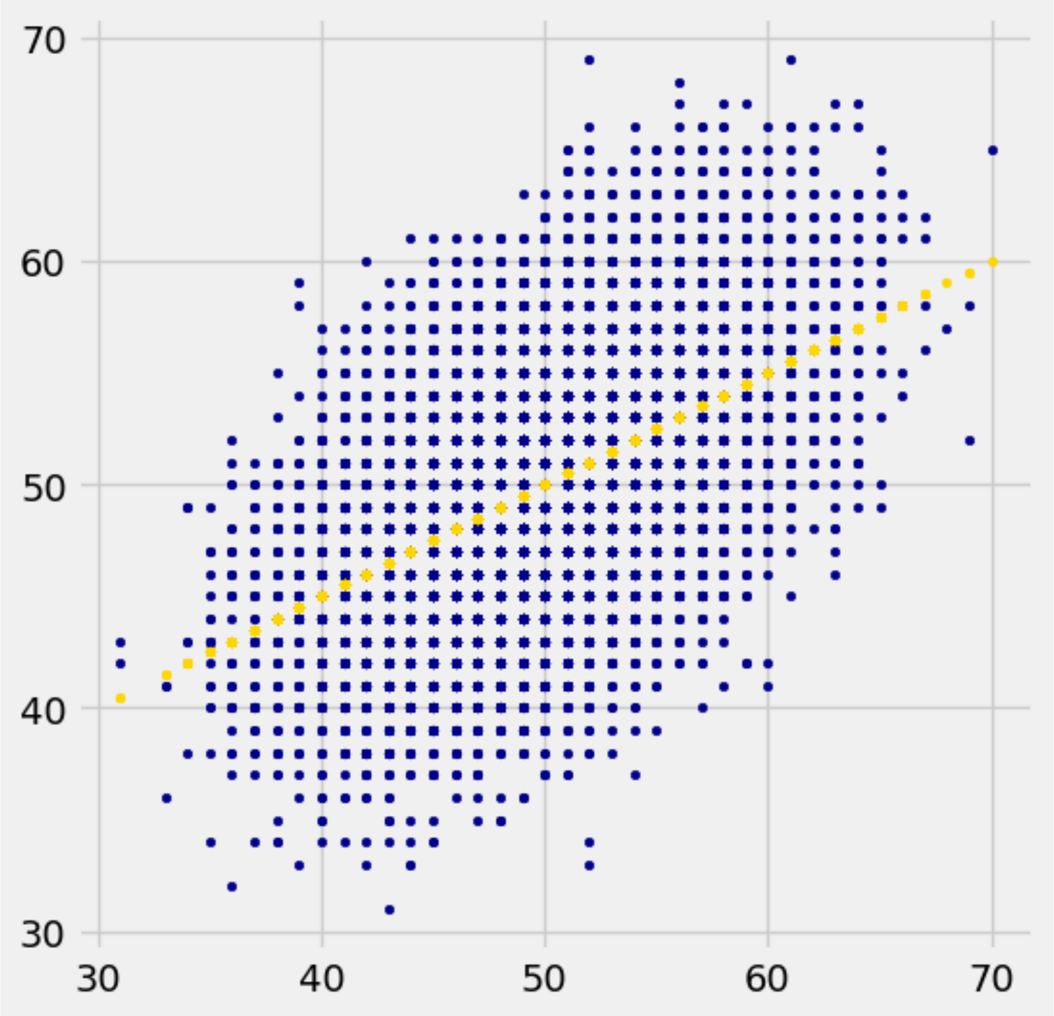
d) Find $E(V \mid X)$, $Var(V \mid X)$, and the correlation $r(X, V)$.

e) Simulate 20,000 (X, V) pairs for $p = 0.5$ and draw the scatter plot of the observed points. Plot $E(V \mid X)$ as a function of X on the same plot. Use the cell below. The arrays `x` and `v` should contain the observed values of X and V. The array `exp_V_given_x` should contain $E(V \mid X = x)$ for each x in `x`, using the formula you derived in **d**.

[50]:

```
# Simulation for e

p = 0.5
n = 20000

trials = stats.binom.rvs(1, p, size=(n, 150))

x = np.sum(trials[:, :100], axis=1)
```

```
v = np.sum(trials[:, 50:150], axis=1)

exp_V_given_x = 0.5 * x + 50 * p

# Don't change the lines below
plt.figure(figsize=(6, 6))
plt.axes().set_aspect('equal')
plt.xticks(np.arange(30, 71, 10))
plt.yticks(np.arange(30, 71, 10))
plt.scatter(x, v, color='darkblue', s=10)
plt.scatter(x, exp_V_given_x, color='gold', s=10);
```



Complete the cell below so that the x_su is x in standard units, v_su is v in standard units, and the last line evaluates to the observed correlation between x and v. Check that the output is consistent with your calculation in **d**.

[51]:

```
# Part e continued

x_mean = np.mean(x)
v_mean = np.mean(v)
x_std = np.std(x)
v_std = np.std(v)

x_su = (x - x_mean) / x_std
v_su = (v - v_mean) / v_std

obs_corr = np.mean(x_su * v_su)
obs_corr
```

[51]:

```
0.50063818204889554
```

1.3 2. Midterms 1 and 2

In a large class, the distribution of scores in Midterms 1 and 2 is approximately bivariate normal. Define the following notation.

- For $i = 1, 2$, μ_i is the class average on Midterm i, and σ_i^2 is the class variance on Midterm i
- The correlation between Midterm 1 and 2 scores in the class is ρ

a) For approximately what proportion of the students was the Midterm 2 score (in standard units) higher than the Midterm 1 score (in standard units)?

b) For approximately what proportion of the students did the two midterm scores (each in standard units) differ by more than 1?

1.4 3. Normals and Coins

Let X be standard normal. Construct a random variable Y as follows:

- Toss a fair coin.
- If the coin lands heads, let $Y = X$.
- If the coin lands tails, let $Y = -X$.

a) Find the cdf of Y and hence identify the distribution of Y.

b) Find $E(XY)$ by conditioning on the result of the toss.

c) Are X and Y uncorrelated?

d) Are X and Y independent?

e) Is the joint distribution of X and Y bivariate normal?

1.5 4. Slices of a Normal Cake

This problem needs only the material of a few weeks ago, but it's here because the ideas and visualization will be helpful in the next exercise. A former 140 staff member told me how he used this method in his interview at a major quant firm and surprised the interviewer. (Yes, he got the job.) Simple, insightful solutions tend to beat tons of calculus even when the calculus is done correctly.

Let X and Y be independent standard normal random variables.

a) Find $P(X > 0, Y > 0)$.

Yes, it's easy. But get a piece of paper and draw the event on the plane anyway. Imagine the joint density surface over the plane, and try to imagine the relevant volume under the joint density surface as a quadrant-shaped slice of a bell-shaped cake. **Then use the same approach for the next two parts.**

b) Find $P(X > 0, Y > X)$.

c) Find $P(X > 0, Y > \sqrt{3}X)$.

1.6 5. Heights of Mothers and Daughters

The heights of a population of mother-daughter pairs have a bivariate normal distribution with correlation 0.5.

a) Of the mothers on the 90th percentile of mothers' heights, what proportion have daughters who are taller than the 90th percentile of daughters' heights?

b) In what proportion of mother-daughter pairs are both women taller than average? (This means the mothers are taller than the average mother and the daughters are taller than the average daughter.)

[Hint: Express standard bivariate normal variables in terms of two independent standard normal variables, and then apply the "slices of a normal cake" method.]

1.7 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

1.7.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using applications such as CamScanner. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image in CamScanner or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.
- If you used LaTeX to do the written portions, you do not need to do any scanning; you can just download the whole notebook as a PDF via LaTeX.

1.7.2 Code Portion

- Save your notebook using `File > Save and Checkpoint`.
- Generate a PDF file using `File > Download As > PDF via LaTeX`. This might take a few seconds and will automatically download a PDF version of this notebook.
  - If you have issues, please post a follow-up on the general Homework 13 Ed thread.

1.7.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 13 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

If you are having difficulties scanning, uploading, or submitting your work, please read the Ed Thread on this topic and post a follow-up on the general Homework 13 Ed thread.

[ ]:

```
```

1a. $X \sim \text{Binomial}(100, p)$

$Y \sim \text{Binomial}(50, p)$

$V \sim \text{Binomial}(100, p)$

1b. $Y \mid X = k$

second 51-100 trials

$X = Y + Z$

$\therefore Z \sim \text{Binomial}(50, p)$

first 1-50 trials

$k = Y + Z$

$Y = k - Z$

idea: Y and Z are independent, BUT NOT independent when conditioned on $X = k$

$$P(Y=y \mid X=k) = \frac{P(Y=y,\ X=k)}{P(X=k)}$$

$$= \frac{P(Y=y,\ Z=k-y)}{P(X=k)}$$

$$= \frac{\binom{50}{y} p^{y} (1-p)^{50-y} \cdot \binom{50}{k-y} p^{k-y} (1-p)^{50-(k-y)}}{\binom{100}{k} p^{k} (1-p)^{100-k}}$$

$$= \frac{\binom{50}{y}\binom{50}{k-y} p^{y+k-y} (1-p)^{50-y+50-k+y}}{\binom{100}{k} p^{k} (1-p)^{100-k}}$$

$$= \frac{\binom{50}{y}\binom{50}{k-y} p^{k} (1-p)^{100-k}}{\binom{100}{k} p^{k} (1-p)^{100-k}}$$

$$= \frac{\binom{50}{y}\binom{50}{k-y}}{\binom{100}{k}}$$

$\therefore$ Hypergeometric$(100, 50, k)$

1c. $E[Y \mid X = k] = k \cdot \frac{50}{100}$

$= \frac{k}{2}$

$\hat{Y} = E[Y \mid X]$

$= \frac{x}{2}$ linear function of X

$$\mathrm{Var}(Y \mid X = k) = 50 \cdot \frac{k}{100} \cdot \frac{100-k}{100} \cdot \frac{50}{99}$$

$$= \frac{2500k(100-k)}{990000}$$

$$= \frac{k(100-k)}{396}$$

$$\mathrm{Var}(Y \mid X) = \frac{x(100-X)}{396}$$

1d. $X = Y + Z$

$V = Y + W$

$\therefore W \sim \text{Binomial}(50, p)$

$\hookrightarrow$ second 50 trials (101 - 150)

W and X are independent!

$$E[V \mid X] = E[Y + W \mid X]$$

$$= E[Y \mid X] + E[W \mid X]$$

$$= \frac{x}{2} + 50p$$

$$\mathrm{Var}(V \mid X) = \mathrm{Var}(Y + W \mid X)$$

$$= \mathrm{Var}(Y \mid X) + \mathrm{Var}(W \mid X)$$

$$= \frac{x(100-x)}{396} + 50p(1-p)$$

$$r(X, V) = \frac{\mathrm{Cov}(X, V)}{SD(X)\,SD(V)}$$

$$= \frac{\mathrm{Cov}(Y + Z, Y + W)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(V)}}$$

$$\operatorname{cov}(Y+Z, Y+W) = \operatorname{Cov}(Y,Y) + \operatorname{cov}\underline{(Y,W)} + \operatorname{Cov}(\underline{Z,Y}) + \operatorname{Cov}(\underline{Z,W})$$

independent
= 0

$$= \operatorname{Var}(Y) + 0$$

$$= 50p(1-p)$$

$$\operatorname{Var}(X) = 100p(1-p)$$

$$\operatorname{Var}(V) = 100p(1-p)$$

$$\therefore\ r(X,V) = \frac{50p(1-p)}{\sqrt{100p(1-p)\cdot 100p(1-p)}}$$

$$= \frac{50p(1-p)}{100p(1-p)}$$

$$= \frac{1}{2}$$

2a. $X \sim$ Midterm 1 scores

$\sim$ Normal $(0, 1)$

$Y \sim$ Midterm 2 scores

$\sim$ Normal $(0, 1)$

$E[X] = E[Y] = 0$

$\text{Var}(X) = \text{Var}(Y) = 1$

$X, Y \sim$ Bivariate Normal

$\text{Corr}(X, Y) = \rho$

$\sim$ proportion of students where $M2 > M1$

$\sim Y - X$

$$E[A] = E[Y - X]$$
$$= E[Y] - E[X]$$
$$= 0$$

$$\text{Var}(A) = \text{Var}(Y - X)$$
$$= \text{Var}(Y) + \text{Var}(X) - 2\,\text{Cov}(Y, X)$$
$$= 1 + 1 - 2\rho$$
$$= 2 + 2\rho$$

$\therefore A \sim$ Normal $(0, 2(1 + \rho))$

$P(A > 0) = 0.5$

$\hookrightarrow$ normal distribution, symmetric around mean, and mean $= 0$

26. $P(|A| > 1) = P(A > 1) + P(A < -1)$

$$Z = \frac{A}{\sqrt{2(1+\rho)}}$$

$\sim \text{Normal}(0,1)$

$$P(|A| > 1) = P\left(Z\sqrt{2(1+\rho)} > 1\right) + P\left(Z\sqrt{2(1+\rho)} < -1\right)$$

$$= \underbrace{P\left(Z > \frac{1}{\sqrt{2(1+\rho)}}\right)} + \underbrace{P\left(Z < -\frac{1}{\sqrt{2(1+\rho)}}\right)}$$

symmetric

∴ Normal distribution

$$= 2P\left(Z > \frac{1}{\sqrt{2(1+\rho)}}\right)$$

$$= 2\left(1 - \Phi\left(\frac{1}{\sqrt{2(1+\rho)}}\right.\right)$$

3a. $X \sim \text{Normal}(0,1)$ $\begin{cases} E[x] = 0 \\ Var(X) = 1 \end{cases}$

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(Y \leq y \mid \text{heads}) P(\text{heads}) + P(Y \leq y \mid \text{tails}) P(\text{tails}) \\
&= P(Y \leq y \mid Y = X)(0.5) + P(Y \leq y \mid Y = -X)(0.5) \\
&= 0.5 P(X \leq y) + 0.5 P(-X \leq y)
\end{aligned}$$

$$\begin{aligned}
P(-X \leq y) &= P(X \geq -y) \\
&= 1 - P(X \leq -y) \\
&= 1 - (1 - P(X \leq y)) \\
&= P(X \leq y)
\end{aligned}$$

$$\begin{aligned}
&= 0.5 P(X \leq y) + 0.5 P(X \leq y) \\
&= P(X \leq y) \\
&= \underline{\Phi}(y)
\end{aligned}$$

$\therefore Y \sim \text{Normal}(0,1)$

3b.
$$\begin{aligned}
E[XY] &= E[XY \mid \text{heads}) P(\text{heads}) + E[XY \mid \text{tails}] P(\text{tails}) \\
&= E[XY \mid Y = X)(0.5) + E[XY \mid Y = -X](0.5) \\
&= E[XX](0.5) - E[XX](0.5) \\
&= 0.5 E[X^2] - 0.5 E[X^2]
\end{aligned}$$

$$\begin{aligned}
Var(X) &= E[X^2] - (E[X])^2 \\
1 &= E[X^2] - 0^2 \\
E[X^2] &= 1
\end{aligned}$$

$$\begin{aligned}
&= (1)(0.5) - (1)(0.5) \\
&= 0.5 - 0.5 \\
&= 0
\end{aligned}$$

3c.
$$\begin{aligned}
cov(X,Y) &= E[XY] - E[x]E[Y] \\
&= 0 - 0(0) \\
&= 0
\end{aligned}$$

$\therefore X, Y$ are uncorrelated

3d. X, Y are NOT independent. We know that X determines the outcome of Y. For example, if $X = 1$, Y can either be 1 or -1 depending on how the fair coin lands. If the coin lands heads $X = Y = 1$, but if the coin lands tails $X = 1, Y = -1$.

3e. $Z = aX + bY$

$= X + Y \qquad a, b = 1$

heads : $Z = X + X$

$= 2X$

tails : $Z = X - X$

$= 0$

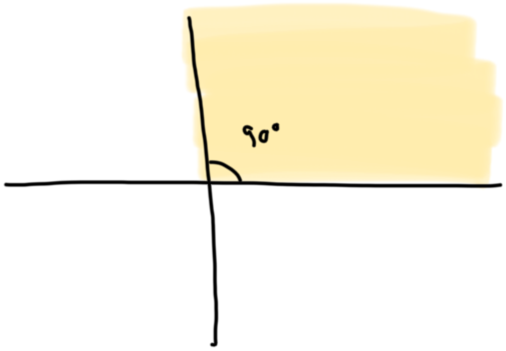
$$Z = \begin{cases} 2x & P(\text{heads}) = \frac{1}{2} \\ 0 & P(\text{tails}) = \frac{1}{2} \end{cases}$$

$\therefore$ X, Y are NOT Bivariate Normal. The RVs need to be both uncorrelated and independent (textbook 24.2). X and Y are uncorrelated, but we found that they are NOT independent. Also, the linear combination is NOT normal either.

4a. $X \sim \text{Normal}(0,1)$

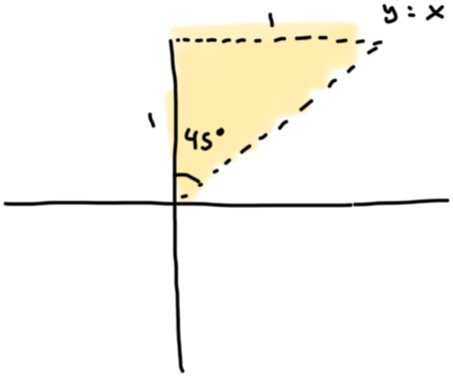
$Y \sim \text{Normal}(0,1)$

$$P(X > 0, Y > 0) = P(X > 0)\,P(Y > 0)$$
$$= \frac{1}{2} \cdot \frac{1}{2}$$
$$= \frac{1}{4}$$



4b. $P(X > 0, Y > X) = \frac{45°}{360°}$

$= \frac{1}{8}$



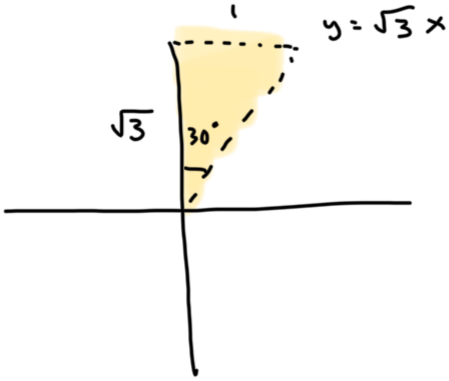
$\tan\theta = 1$

$\theta = \tan^{-1}(1)$

$= 45°$

4c. $P(X > 0, Y > \sqrt{3} X) = \frac{30°}{360°}$

$= \frac{1}{12}$



$\tan\theta = \frac{1}{\sqrt{3}}$

$\theta = \tan^{-1}\left(\frac{1}{\sqrt{3}}\right)$

$= 30°$

5a. $X \sim$ mothers' heights

$\sim \text{Normal}(0,1)$

$Y \sim$ daughters' heights

$\sim \text{Normal}(0,1)$

$X, Y \sim$ bivariate normal

$$\therefore Y = \rho X + \sqrt{1-\rho^2}\, Z \quad , \rho = 0.5$$
$$= 0.5X + \sqrt{1-(0.5)^2}\, Z$$
$$= 0.5X + \sqrt{0.75}\, Z$$

90th percentile ≈ 1.282

$$P(Y > 1.282 \mid X = 1.282) = P(0.5X + \sqrt{0.75}\, Z > 1.282 \mid X = 1.282)$$
$$= P(0.5(1.282) + \sqrt{0.75}\, Z > 1.282)$$
$$= P(0.64 + \sqrt{0.75}\, Z > 1.282)$$
$$= P(\sqrt{0.75}\, Z > 1.282 - 0.64)$$
$$= P(\sqrt{0.75}\, Z > 0.642)$$
$$= P\left(Z > \frac{0.642}{\sqrt{0.75}}\right)$$
$$= P(Z > 0.741)$$
$$= 1 - \Phi(0.741)$$

5b. $$Y = \rho X + \sqrt{1-\rho^2}\, Z > 0$$
$$\sqrt{1-\rho^2}\, Z > -\rho X$$
$$Z > \underbrace{\frac{-\rho}{\sqrt{1-\rho^2}} X}_{\text{slope}}$$

$$\therefore \tan\theta = \frac{-\rho}{\sqrt{1-\rho^2}}$$
$$\theta = \tan^{-1}\left(\frac{-\rho}{\sqrt{1-\rho^2}}\right)$$

$$= \frac{\pi}{2} + \tan^{-1}\left(-\frac{\rho}{\sqrt{1-\rho^2}}\right)$$

$$= \frac{\pi}{2} - \tan^{-1}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$$

$$= \frac{1}{2\pi}\left(\frac{\pi}{2} - \tan^{-1}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)\right)$$

$$P(X > 0, Y > 0) = \frac{1}{4} + \tan^{-1}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$$

$$\rho = 0.5 : \frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\left(\frac{0.5}{\sqrt{1-0.5^2}}\right)$$

$$= \frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\left(\frac{0.5}{\sqrt{0.75}}\right)$$

$$= \frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\left(\frac{\frac{1}{2}}{\sqrt{\frac{3}{4}}}\right)$$

$$= \frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\left(\frac{1}{2}\cdot\frac{2}{\sqrt{3}}\right)$$

$$= \frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\left(\frac{1}{\sqrt{3}}\right)$$

$$= \frac{1}{4} + \frac{1}{2\pi}\cdot\frac{\pi}{6}$$

$$= \frac{1}{4} + \frac{1}{12}$$

$$= \frac{3}{12} + \frac{1}{12}$$

$$= \frac{1}{3}$$