Homework_07

March 10, 2025

Probability for Data Science

UC Berkeley, Spring 2025

Michael Xiao and Ani Adhikari

1 Homework 7 (Due March 10th at 5 PM)

[13]:
```
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 Instructions

Your homeworks will generally have two components: a written portion and a portion that also involves code. Written work should be completed on paper, and coding questions should be done in the notebook. Start the work for the written portions of each section on a new page. You are welcome to LaTeX your answers to the written portions, but staff will not be able to assist you with LaTeX related issues.

It is your responsibility to ensure that both components of the homework are submitted completely and properly to Gradescope. **Make sure to assign each page of your pdf to the correct question. Refer to the bottom of the notebook for submission instructions.**

1.0.2 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.

1.0.3 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 selected for each question.

1.2 1. Correlation

The *correlation coefficient* between random variables X and Y is defined as

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

It is called the correlation, for short. The definition explains why X and Y are called *uncorrelated* if $Cov(X,Y) = 0$.

a) Let X^* be X in standard units and let Y^* be Y in standard units. Check that

$$r(X,Y) = E(X^*Y^*)$$

This is the random variable version of the Data 8 definition of the correlation between two data variables: convert each variable to standard units; multiply each pair; take the mean of the products.

b) Use the fact that $(X^* + Y^*)^2$ and $(X^* - Y^*)^2$ are non-negative random variables to show that $-1 \le r(X,Y) \le 1$.

[First find the numerical values of $E(X^*)$ and $E\left({X^*}^2\right)$. Then find $E\left((X^* + Y^*)^2\right)$.]

c) Show that if $Y = aX + b$ where $a \neq 0$, then $r(X,Y)$ is 1 or -1 depending on whether the sign of a is positive or negative.

d) Consider a sequence of i.i.d. Bernoulli (p) trials. For any positive integer k let X_k be the number of successes in trials 1 through k. **Use bilinearity** to find $Cov(X_n, X_{n+m})$ and hence find $r(X_n, X_{n+m})$.

e) Fix n and find the limit of $r(X_n, X_{n+m})$ as $m \to \infty$. Explain why the limit is consistent with intuition.

1.3 2. Collecting Distinct Values

In Homework 4 you found the expectation of each of the random variables below. **Go back and see how you did that, and then use the same ideas** to find the variance of each one.

For one part you will need the fact that the SD of a geometric (p) random variable is $\frac{\sqrt{q}}{p}$ where $q = 1 - p$. We haven't proved that as the algebra takes a bit of work. We will prove it later in the course by conditioning.

(a) A die is rolled n times. Find the variance of number of faces that *do not* appear.

(b) Use your answer to (a) to find the variance of the number of distinct faces that *do* appear in n rolls of a die.

(c) Find the variance of the number of times you have to roll a die till you have seen all of the faces.

1.4 3. The "Sample Variance"

Let $X_1, X_2, \ldots, X_n$ be i.i.d. (discrete or continuous), each with mean μ and SD σ. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean.

(a) Find $E(\bar{X})$ and $SD(\bar{X})$. It's fine to just quote answers derived in class or in the textbook.

(b) For each i, find $Cov(X_i, \bar{X})$. [Plug in the definition of $\bar{X}$ and use bilinearity.]

(c) For each i in the range 1 through n, define the *ith deviation in the sample* as $D_i = X_i - \bar{X}$. Find $E(D_i)$ and $Var(D_i)$. [Write the variance as $Cov(D_i, D_i)$, plug in the definition of D_i, and use bilinearity.]

(d) Define the random variable $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} D_i^2$$

Find $E(\hat{\sigma}^2)$.

For this random variable, the notation $\hat{\sigma}^2$ is pretty standard in statistics. Just think of $\hat{\sigma}^2$ as a symbol; it doesn't help to start thinking about the random variable that is its square root.

(e) Use Part **d** to construct a random variable denoted S^2 that is an unbiased estimator of σ^2. This random variable S^2 is called the *sample variance* and is frequently used in inference.

1.5 4. Poisson-Binomial Distribution

For this exercise, please refer to the theory in Section 14.1 and the code in Section 14.2.

In Lab 3B you saw that a *Poisson-binomial* random variable is a sum of independent indicators that are not necessarily identically distributed:

$X = I_1 + I_2 + \cdots + I_n$ where I_j has the Bernoulli (p_j) distribution and $I_1, I_2, \ldots, I_n$ are independent.

(a) What is the probability generating function of a Bernoulli (p) random variable? Provide a formula and then use the code cell below to define a function `indicator_pgf` that takes p as its argument and returns the probability generating function of a Bernoulli (p) random variable as a `NumPy` polynomial. Use as many lines as you need. The last line of the cell is there for you to check that your function is working.

[14]:
```
# Answer to a

def indicator_pgf(p):
    return np.poly1d([p, 1-p])

print(indicator_pgf(0.4))
```

```
0.4 x + 0.6
```

(b) For $j = 1, 2, \ldots, 20$, let $p_j = 1/(j+1)$. Let $I_1, I_2, \ldots, I_{20}$ be independent indicators such that I_j has the Bernoulli (p_j) distribution, and let $X = I_1 + I_2 + \cdots + I_{20}$. Complete the code cell below so that `pgf_X` is the probability generating function of X as a `NumPy` polynomial. Use as many lines as you need. The last two lines are there for you to check that your polynomial has the correct degree and that it is indeed a probability generating function.

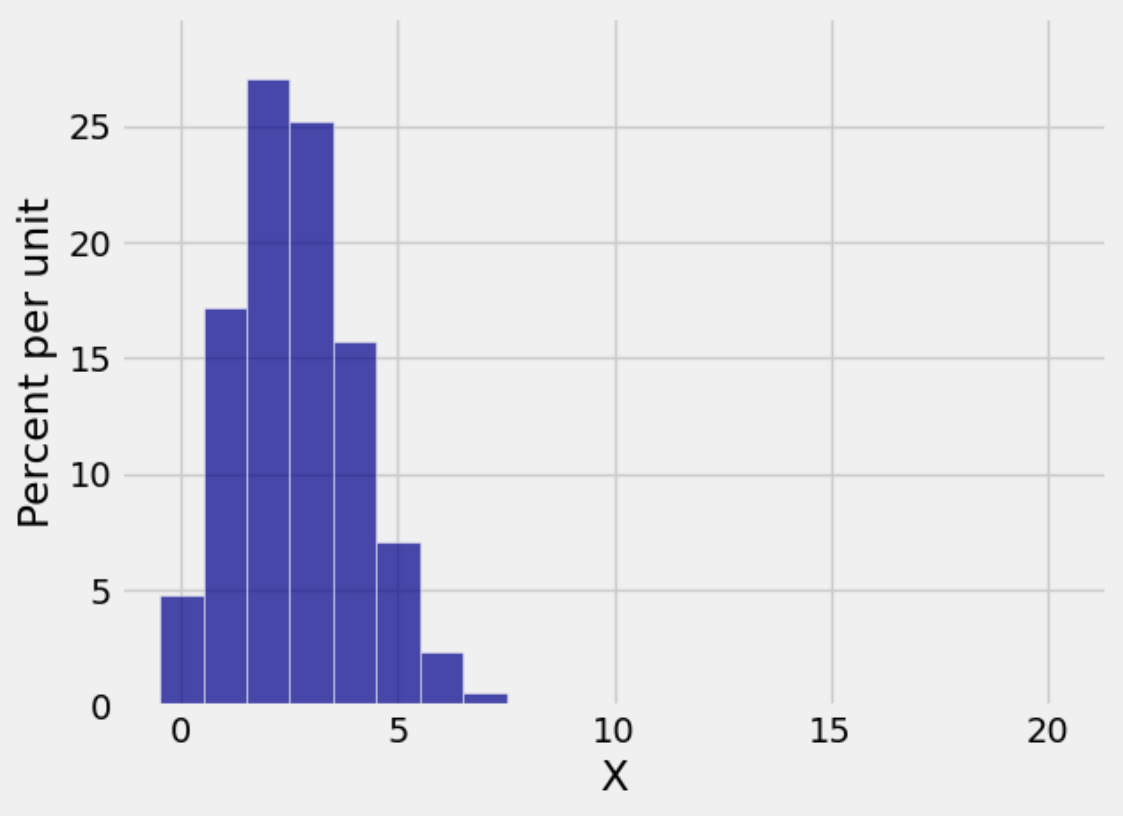
[15]:
```
# Answer to b

pgf_X = np.poly1d([1])

for j in np.arange(1, 21):
    p_j = 1 / (j + 1)
    pgf_X *= indicator_pgf(p_j)

print(pgf_X)
sum(pgf_X.c) # sum of coefficients
```

```
           20             19             18             17             16
1.957e-20 x  + 4.11e-18 x  + 4.035e-16 x  + 2.46e-14 x  + 1.044e-12 x
              15             14             13             12
 + 3.273e-11 x  + 7.863e-10 x  + 1.48e-08 x  + 2.214e-07 x
              11             10             9             8             7
 + 2.654e-06 x  + 2.559e-05 x  + 0.0001985 x + 0.001234 x + 0.006094 x
            6          5          4          3          2
 + 0.02362 x + 0.07046 x + 0.1573 x + 0.2519 x + 0.2702 x + 0.1713 x + 0.04762
```

[15]: 1.0000000000000002

(c) Complete the cell below to plot the probability histogram of X. Do not add any more lines.

[16]:
```
# Answer to c

vals_X = np.arange(21)
probs_X = np.flipud(pgf_X)
dist_X = Table().with_columns("X", vals_X, "Probability", probs_X)
Plot(dist_X)
```



(d) Complete the cell below to find the expectation, variance, and SD of X using `p_array`. Do not add any more lines. Then run the cell below that to check your answers.

[17]:
```
# Answer to d

p_array = 1/np.arange(2, 22)
ev_X = np.sum(p_array)
var_X = np.sum(p_array * (1 - p_array))
sd_X = np.sqrt(var_X)
ev_X, var_X, sd_X
```

[17]:
```
(2.6453587047627294, 2.046927887153561, 1.4307088757513042)
```

[18]:
```
dist_X.ev(), dist_X.var(), dist_X.sd()
```

[18]:
```
(2.6453587047627281, 2.0469278871535597, 1.4307088757513038)
```

(e) Explain why the distribution of X cannot be Poisson. Then show that the distribution of X is not binomial either, as follows. If X were binomial, what would n have to be? Use that and your answer to Part (d) to see what p would have to be. Use the code cell below to find the variance of that binomial distribution, and compare with your answer to Part (d).

[19]:
```
# Answer to e

n = 20
p = ev_X / n
binomial_variance = n * p * (1 - p)
binomial_variance
```

[19]:
```
2.2954625709195318
```

1.6 5. Testing Hypotheses in the Gauss Model

The Gauss model for measurement error says that repeated measurements $X_1, X_2, \ldots, X_n$ of the same quantity have the structure

$$X_i = \mu + \epsilon_i, \qquad 1 \le i \le n$$

where μ is an unknown constant called the *true value* and $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are random error terms assumed to be i.i.d. with mean 0 and variance σ^2.

From a practical perspective, the true value μ comes from the quantity being measured (for example, the true weight of an object). The error terms come from the measuring process (for example, from the balance being used for weighing). Thus σ is sometimes known because of extensive experience with the measuring process (for example, having used the same balance to weigh many different objects).

So assume that the Gauss model holds with $\sigma = 1$, and let $n = 100$. Suppose a data scientist wants to test the following hypotheses:

- Null hypothesis H_0: $\mu = 20$
- Alternative hypothesis H_A: $\mu \neq 20$

Suppose the data scientist wants to use the average measurement $\bar{X}$ as the test statistic and reject the null hypothesis if $|\bar{X} - 20| > 0.175$.

(a) Rewrite the decision rule by filling in the blanks with numbers: $|\bar{X} - 20| > 0.175 \iff \bar{X} <$ _____ or $\bar{X} >$ _____

(b) Level: Find the approximate distribution of the test statistic $\bar{X}$ under H_0, and use this distribution to find the approximate probability that the test rejects the null hypothesis if the null hypothesis is true. This probability is called the *level* of the test. In Data 8 we called it the cutoff for the p-value.

Please write out your answer in math notation. Then use the code cell below for scratch work. Remember that `stats.norm.cdf(x, mean, SD)` evaluates to the cdf of the normal (mean, SD^2) distribution at the point x. The necessary modules have been imported at the top of this notebook.

(c) Power: Suppose that in fact $\mu = 20.5$ though the data scientist doesn't know this and is still performing the same test as above. Find the approximate distribution of the test statistic $\bar{X}$ under the condition $\mu = 20.5$, and use this distribution to find the approximate probability that the test rejects the null hypothesis if $\mu = 20.5$. This probability is called the *power of the test against the fixed alternative* $\mu = 20.5$.

Please write out your answer in math notation. Then use the code cell below for scratch work.

[20]:
```
# Scratch work for b and c
phi_1_75 = stats.norm.cdf(1.75, 0, 1)
level = 2 * (1 - phi_1_75)

phi_3_25 = stats.norm.cdf(3.25, 0, 1)
power = phi_3_25
```

```
level, power
```

[20]: (0.080118313727634227, 0.99942297495760923)

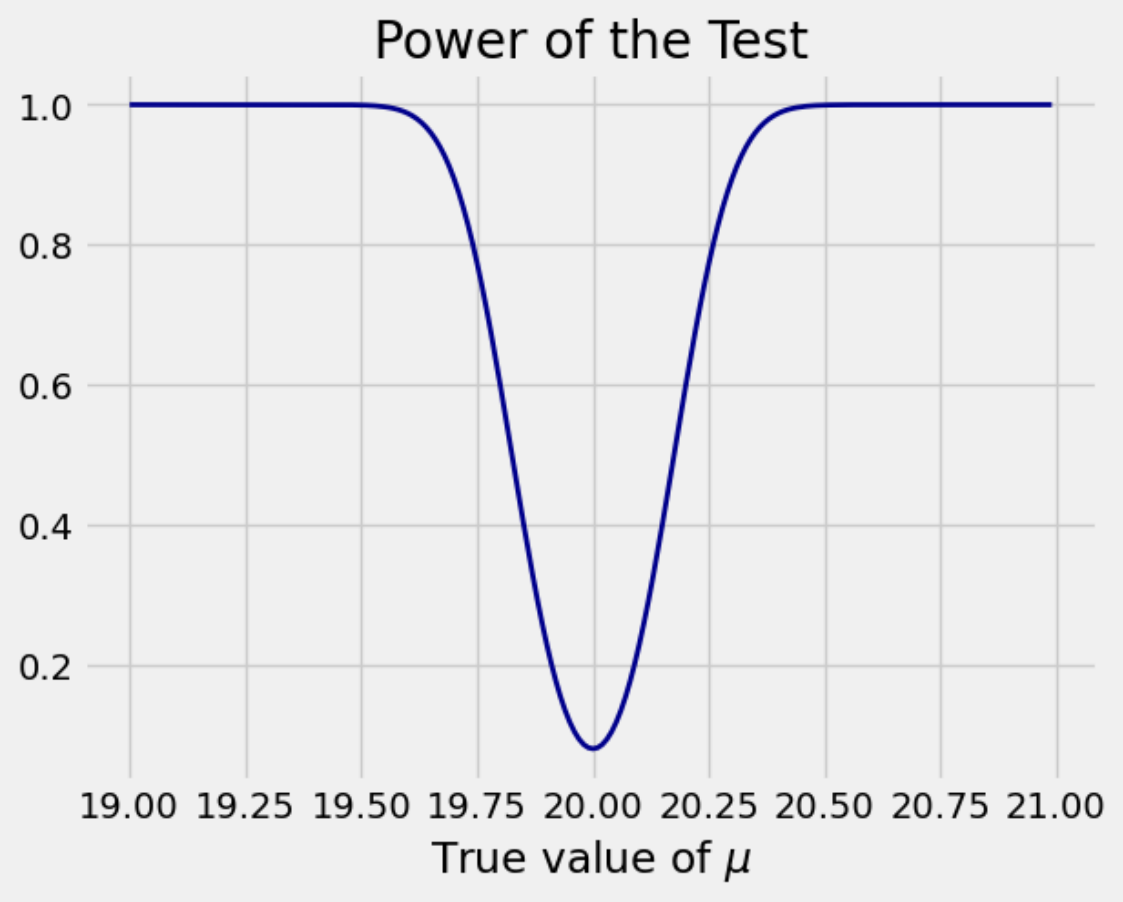
(d) Complete the code cell below to plot the graph of the power of the test under the fixed alternative $\mu = \mu_A$ for μ_A in the range `true_mu` below. Do not add any more lines.

Computational note: First study the code below and compare with the output of the cell.

[21]:
```
mu_list = [10, 15, 20]  # It's also fine for this to be an array.

# array of P(X_i < 12)
# for X_i normal with mean = ith element of mu_list
# and SD = 8
stats.norm.cdf(12, mu_list, 8)
```

[21]: array([0.59870633, 0.35383023, 0.15865525])

[22]:
```
# Answer to d

true_mu = np.arange(19, 21, 0.01)
power = stats.norm.cdf((19.825 - true_mu) / 0.1) + 1 - stats.norm.cdf((20.175 -␣
 ↪true_mu) / 0.1)

plt.plot(true_mu, power, color='darkblue', lw=2)
plt.xlabel(r'True value of $\mu$')
plt.title('Power of the Test');
```



(e) Interpret the graph. What is the test likely to do if the true value of μ is far from 20, and what does the power converge to (be careful!) when the true value gets close to 20?

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 Notebook`.
- Generate a PDF file using `File > Save and Export Notebook As > PDF`. 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 7 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 7 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 7 Ed thread.

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

1a. $$r(X,Y) = \frac{Cov(X,Y)}{SD(X)\,SD(Y)}$$

$$= \frac{E[(X - E[X]),(Y - E[Y])]}{SD(X)\,SD(Y)}$$

$$X^* = \frac{X - E[X]}{SD(X)} \qquad Y^* = \frac{Y - E[Y]}{SD(Y)}$$

$$r(X,Y) = E[X^*, Y^*]$$

$$= E\left[\frac{X - E[X]}{SD(X)}, \frac{Y - E[Y]}{SD(Y)}\right]$$

$$= \frac{E[(X - E[X]),(Y - E[Y])]}{SD(X)\,SD(Y)}$$

$$= \frac{Cov(X,Y)}{SD(X)\,SD(Y)}$$

$$\therefore r(X,Y) = E[X^*, Y^*]$$

1b. $$E[(X^* + Y^*)] \geq 0$$

$$E[(X^* - Y^*)] \geq 0$$

$E[X^*] = 0$ mean of any RV in standard units is 0

$$E[X^{*2}] = Var[X^*]$$

$$= \sqrt{SD(X^*)}$$

$$= \sqrt{1}$$

$$= 1$$

SD of any RV in standard units is 1

$$E\left[(X^* + Y^*)^2\right] = E\left[(X^* + Y^*)(X^* + Y^*)\right]$$

$$= E\left[X^{*2} + 2X^*Y^* + Y^{*2}\right]$$

$$= E\left[X^{*2}\right] + 2E\left[X^*Y^*\right] + E\left[Y^{*2}\right]$$

$$= 1 + 2E\left[X^*Y^*\right] + 1$$

$$= 2 + 2E\left[X^*Y^*\right]$$

$$= 2 + 2r(X, Y) \geq 0$$

$$2r(X, Y) \geq -2$$

$$r(X, Y) \geq -1$$

$$E\left[(X^* - Y^*)^2\right] = E\left[(X^* - Y^*)(X^* - Y^*)\right]$$

$$= E\left[X^{*2} - 2X^*Y^* + Y^{*2}\right]$$

$$= E\left[X^{*2}\right] - 2E\left[X^*Y^*\right] + E\left[Y^{*2}\right]$$

$$= 1 - 2E\left[X^*Y^*\right] + 1$$

$$= 2 - 2E\left[X^*Y^*\right]$$

$$= 2 - 2r(X^*, Y^*) \geq 0$$

$$2r(X^*, Y^*) \leq 2$$

$$r(X^*, Y^*) \leq 1$$

$$\therefore \quad -1 \leq r(X^*, Y^*) \leq 1$$

1c. $Y = aX + b$

$E[Y] = E[aX + b]$

$E[Y] = aE[X] + b$

$Y - E[Y] = aX + b - (aE[X] + b)$

$Y - E[Y] = aX + b - aE[X] - b$

$Y - E[Y] = aX - aE[X]$

$Y - E[Y] = a[X - E[X]]$

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

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[(X - E[X])(a(X - E[X]))] \\ &= a\,E[(X - E[X])^2] \\ &= a\,Var(X) \end{aligned}$$

$$SD(Y) = \sqrt{Var(Y)}$$

$$\begin{aligned} Var(Y) &= Var(aX + b) \\ &= a^2 Var(X) \qquad \text{textbook 12.1.3} \end{aligned}$$

$$\begin{aligned} SD(Y) &= \sqrt{a^2 Var(X)} \\ &= |a|\, SD(X) \end{aligned}$$

$$\therefore r(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

$$= \frac{a\,Var(X)}{SD(X)\left(|a|\,SD(X)\right)}$$

$$= \frac{a\,SD(X)^2}{|a|\,SD(X)^2}$$

$$= \frac{a}{|a|}$$

$$r(X,Y) = \begin{cases} 1 & a > 0 \\ -1 & a < 0 \end{cases}$$

1d. $Cov(X_n, X_{n+m}) = Cov(X_n, X_n + X_{n+1-n+m})$

$X_n \sim Binomial(n,p)$
$X_{n+m} \sim Binomial(n+m,p)$ $\Big\rangle$ dependent

$$\therefore Cov(X_n, X_{n+m}) = Cov(X_n, X_n + X_{(n+1)-(n+m)})$$

$$= Cov(X_n, X_n) + \underbrace{Cov(X_n, X_{(n+1)-(n+m)})}_{\text{independent} \therefore 0}$$

$$= Var(X_n) + 0$$

$$= np(1-p)$$

$$r(X_n, X_{n+m}) = \frac{Cov(X_n, X_{n+m})}{SD(X_n)\,SD(X_{n+m})}$$

$$SD(X_n) = \sqrt{Var(X_n)}$$
$$= \sqrt{np(1-p)}$$

$$SD(X_{n+m}) = \sqrt{Var(X_{n+m})}$$
$$= \sqrt{(n+m)p(1-p)}$$

$$\therefore r(X_n, X_{n+m}) = \frac{np(1-p)}{(\sqrt{np(1-p)})(\sqrt{(n+m)p(1-p)})}$$
$$= \frac{np(1-p)}{\sqrt{n(n+m)}\ p(1-p)}$$
$$= \frac{n}{\sqrt{n(n+m)}} \cdot \frac{\sqrt{n}}{\sqrt{n}}$$
$$= \frac{\sqrt{n}}{\sqrt{n+m}}$$

i.e. $\lim_{m \to \infty} r(X_n, X_{n+m}) = \frac{\sqrt{n}}{\sqrt{n+m}}$ ($\sqrt{n+m} \to \infty$)
$$= 0$$

$\therefore$ As $m \to \infty$, the total number of trials $(n+m)$ increases significantly due to the additional m trials, which are independent from the first n trials. So, the correlation between X_n and X_{n+m} approaches 0.

2a. $X \sim$ # faces that do NOT appear

$$I_i = \begin{cases} 1 & i^{th} \text{ face doesn't appear on } n^{th} \text{ roll} \\ 0 & \text{otherwise} \end{cases}$$

$$E[I_i] = P(I_i = 1)$$
$$= 1 - \frac{1}{6}$$
$$= \left(\frac{5}{6}\right)^n$$

$$\therefore E[X] = \sum_{i=1}^{6} I_i$$
$$= \sum_{i=1}^{6} \left(\frac{5}{6}\right)^n$$
$$= 6\left(\frac{5}{6}\right)^n$$

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

$$(E[X])^2 = \left(6\left(\frac{5}{6}\right)^n\right)^2$$
$$= 36\left(\frac{5}{6}\right)^{2n}$$

$$E[X^2] = E\left[\left(\sum_{i=1}^{6} I_i\right)\left(\sum_{j=1}^{6} I_j\right)\right]$$
$$= E\left[\sum_{i=1}^{6}\sum_{j=1}^{6} I_i I_j\right]$$
$$= \sum_{i=1}^{6}\sum_{j=1}^{6} E[I_i I_j]$$
$$= \begin{cases} 6\left(\frac{5}{6}\right)^n & i = j \end{cases}$$

$$\left(30\left(\frac{4}{6}\right)^n \quad i \neq j\right.$$

case 1: $i = j$

$$E[I_i I_j] = E[I_i]$$
$$= \sum_{i=1}^{6} E[I_i]$$
$$= 6\left(\frac{5}{6}\right)^n$$

case 2: $i \neq j$

$$E[I_i I_j] = P(I_i = 1 \text{ and } P_j = 1)$$
$$= 1 - \frac{2}{6}$$
$$= \left(\frac{4}{6}\right)^n$$

missing faces $= 36 - 6$
$= 30$

$$\therefore 30\left(\frac{4}{6}\right)^n$$

$$\therefore E[X^2] = 6\left(\frac{5}{6}\right)^n + 30\left(\frac{4}{6}\right)^n$$

$$\therefore \text{Var}(X) = E[X^2] - (E[X])^2$$
$$= 6\left(\frac{5}{6}\right)^n + 30\left(\frac{4}{6}\right)^n - 36\left(\frac{5}{6}\right)^{2n}$$

2b. $Y \sim$ # distinct faces that do appear in n rolls of a die

$$Y = 6 - X$$
$$\text{Var}(Y) = \text{Var}(6 - X)$$
$$\text{Var}(Y) = \text{Var}(X)$$
$$= 6\left(\frac{5}{6}\right)^n + 30\left(\frac{4}{60}\right)^n - 36\left(\frac{5}{6}\right)^{2n}$$

$\therefore$ 6 doesn't affect spread of the distribution

2c. $X \sim$ # rolls till all of the faces have been seen

$X_1 \sim$ # rolls to get 1st unique face
$X_2 \sim$ # rolls to get 2nd unique face
$\vdots$
$X_6 \sim$ # rolls to get last unique face

$X_i \sim \text{Geometric}(p)$

$P(X_1) = 1$
$P(X_2) = \frac{5}{6}$
$P(X_3) = \frac{4}{6}$
$P(X_4) = \frac{3}{6}$
$P(X_5) = \frac{2}{6}$
$P(X_6) = \frac{1}{6}$

$X_1, X_2, \ldots, X_6$ are independent
$\therefore$ # rolls till you see a face you haven't before only depends on the # faces left to be seen
↳ outcome of a roll doesn't affect probability dist. of next roll

$\therefore$ idea: $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$

$$X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$$
$$\text{Var}(X) = \text{Var}(X_1 + X_2 + X_3 + X_4 + X_5 + X_6)$$
$$= \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \text{Var}(X_4) + \text{Var}(X_5) + \text{Var}(X_6)$$

$$SD(X) = \frac{\sqrt{q}}{p}$$

$$\therefore \text{Var}(X) = \left(\frac{\sqrt{q}}{p}\right)^2$$

$$= \frac{q}{p^2}$$

$$\text{Var}(X_1) = \frac{1 - 1}{1^2}$$

$$= 0$$

$$\text{Var}(X_2) = \frac{1 - \frac{5}{6}}{\left(\frac{5}{6}\right)^2}$$

$$= \frac{1}{6} \times \frac{36}{25}$$

$$= \frac{6}{25}$$

$$\text{Var}(X_3) = \frac{1 - \frac{4}{6}}{\left(\frac{4}{6}\right)^2}$$

$$= \frac{2}{6} \times \frac{36}{16}$$

$$= \frac{6}{8}$$

$$= \frac{3}{4}$$

$$\text{Var}(X_4) = \frac{1 - \frac{3}{6}}{\left(\frac{3}{6}\right)^2}$$

$$= \frac{3}{6} \times \frac{36}{9}$$

$$= \frac{6}{3}$$

$$= 2$$

$$\text{Var}(X_5) = \frac{1 - \frac{2}{6}}{\left(\frac{2}{6}\right)^2}$$

$$= \frac{4}{6} \times \frac{36}{4}$$

$$= 6$$

$$\text{Var}(X_6) = \frac{1 - \frac{1}{6}}{\left(\frac{1}{6}\right)^2}$$

$$= \frac{5}{6} \times \frac{36}{1}$$

$$= 30$$

$$\therefore \text{Var}(X) = 0 + \frac{6}{25} + \frac{3}{4} + 2 + 6 + 30$$

$$= \frac{24}{100} + \frac{75}{100} + \frac{3800}{100}$$

$$= \frac{3899}{100}$$

$$= 38.99$$

3a. $E[S_n] = n\mu$ textbook 13.3.3

$\therefore S_n$ = sample sum

$$\therefore E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

$$= \frac{1}{n}(n\mu)$$

$$= \mu$$

$Var(S_n) = n\sigma^2$ textbook 13.3.3

$\therefore SD(\bar{X}) = \sqrt{Var(\bar{X})}$

$Var(aX + b) = a^2 Var(X)$ textbook 12.1.3

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2} \times n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

$$\therefore SD(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}$$

$$= \frac{\sigma}{\sqrt{n}}$$

3b. $Cov(X_i, \bar{X}) = E[X_i \bar{X}] - E[X_i]E[\bar{X}]$

$E[X_i] = \mu$

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n}(n\mu)$$ textbook 13.3.3

$$= \mu$$

$$E[X_i \bar{X}] = E\left[X_i \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)\right]$$

$$= \frac{1}{n}\sum_{j=1}^{n} E[X_i X_j]$$

$$= \begin{cases} \sigma^2 + \mu^2 & i = j \\ \mu^2 & i \neq j \end{cases} \quad \text{(work below)}$$

case 1: $i = j$: $E[X_i X_i] = E[X_i^2]$

$$= Var(X_i) + (E[X_i])^2$$

$$= (SD(X_i))^2 + \mu^2$$

$$= \sigma^2 + \mu^2$$

case 2: $i \neq j$: $E[X_i X_j] = E[X_i]\,E[X_j]$

$$= \mu \times \mu$$

$$= \mu^2$$

$$E[X_i \bar{X}] = \frac{1}{n}\left[E[X_i X_i] + \sum_{j \neq i} E[X_i X_j]\right]$$

$$= \frac{1}{n}\left[\sigma^2 + \mu^2 + (n-1)\mu^2\right]$$

$n-1 \rightarrow$ ensures that j NEVER equals i

$$= \frac{1}{n}\left[\sigma^2 + \mu^2 + n\mu^2 - \mu^2\right]$$

$$= \frac{\sigma^2}{n} + \mu^2$$

$$\therefore Cov(X_i, \bar{X}) = E[X_i \bar{X}] - E[X_i]\,E[\bar{X}]$$

$$= \frac{\sigma^2}{n} + \mu^2 - \mu \times \mu$$

$$= \frac{\sigma^2}{n}$$

3c. $E[D_i] = E[x_i - \bar{x}]$

$= E[x_i] - E[\bar{x}]$

$= \mu - \mu$

$= 0$

$Var(D_i) = Cov(D_i, D_i)$

$= Cov(x_i - \bar{x}, x_i - \bar{x})$

$= Cov(x_i, x_i) - 2Cov(x_i, \bar{x}) + Cov(\bar{x}, \bar{x})$

$= Var(x_i) - 2Cov(x_i, \bar{x}) + Var(\bar{x})$

$= \sigma^2 - 2 \times \frac{\sigma^2}{n} + \frac{\sigma^2}{n}$

$= \sigma^2 - \frac{\sigma^2}{n}$

3d. $E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n} D_i^2\right]$

$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{E[D_i^2]}$

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

$= E[D_i^2] - 0^2$

$Var(D_i) = E[D_i^2]$

$= \frac{1}{n}\sum_{i=1}^{n} \left(\sigma^2 - \frac{\sigma^2}{n}\right)$

$= \frac{1}{n} \times n \left(\sigma^2 - \frac{\sigma^2}{n}\right)$

$= \sigma^2 - \frac{\sigma^2}{n}$

3e. $E[\hat{\sigma}^2] \rightarrow$ right now biased for σ^2

$$E[\hat{\sigma}^2] = \frac{\sigma^2 n - \sigma^2}{n}$$

$$= \sigma^2 \frac{(n-1)}{n} \times \frac{n}{(n-1)}$$

$$= \sigma^2$$

$\therefore$ unbiased

idea: $\frac{\sigma^2 n - \sigma^2}{n}$

$\sigma^2 \frac{(n-1)}{n} \times \frac{n}{n-1} = \sigma^2$

$$S_n = \hat{\sigma}^2 \rightarrow \text{bias}$$

$$\therefore E[S_n] = E\left[\frac{n}{n-1} \cdot \hat{\sigma}^2\right]$$

$$= \frac{n}{n-1} E[\hat{\sigma}^2]$$

$$= \frac{n}{n-1}\left(\sigma^2 - \frac{\sigma^2}{n}\right)$$

$$= \frac{n}{n-1} \times \sigma^2\left(\frac{n-1}{n}\right)$$

$$= \sigma^2$$

4a. $I = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}$

$$G_I(s) = \sum_{k=1}^{n} P(I=k)\, s^k$$

$$= E[S^I]$$
$$= P(I=0)\, s^0 + P(I=1)\, s^1$$
$$= (1-p) \times 1 + p \times s$$
$$= (1-p) + ps$$

4e. The distribution of X cannot be Poisson because the probabilities vary for each p_j (ie. $j=3 \rightarrow p_3 = \frac{1}{3+1} = \frac{1}{4}$ vs $j=10 \rightarrow p_{10} = \frac{1}{10+1} = \frac{1}{11}$). This does NOT satisfy the condition of having a constant probability. Also, the number of trials is relatively small for using Poisson approximation. Similarly, the distribution of X is also not binomial because it doesn't have a constant probability for each trial, failing to satisfy the iid condition.

The code calculates the variance of a binomial distribution when ALL the probabilities are fixed. Comparing the output with the actual variance in part d shows that the distribution of X is not truly binomial, as the values do NOT match.

5a. $|\bar{x} - 20| > 0.175$

$\therefore \bar{x} > 20 + 0.175$ or $\bar{x} < 20 - 0.175$

$\bar{x} > 20.175$ $\bar{x} < 19.825$

5b. $\bar{X} \sim N(E[\bar{X}], Var(\bar{X}))$

$\sim N(\mu, \frac{\sigma^2}{n})$

solved in 3a.

$\sim N(20, \frac{1^2}{100})$

$\sim N(20, \frac{1}{100})$

$P(\bar{X} < 19.825) = \Phi\left(\frac{19.825 - 20}{\sqrt{1/100}}\right)$

$= \Phi\left(\frac{-0.175}{0.1}\right)$

$= \Phi(-1.75)$

$P(\bar{X} > 20.175) = 1 - \Phi\left(\frac{20.175 - 20}{\sqrt{1/100}}\right)$

$= 1 - \Phi\left(\frac{0.175}{0.1}\right)$

$= 1 - \Phi(1.75)$

level : $\Phi(-1.75) + 1 - \Phi(1.75)$

$\therefore$ symmetry $\rightarrow$ normal distribution

$\therefore \Phi(-1.75) = 1 - \Phi(1.75)$

$= 2(1 - \Phi(1.75))$

5c. $\bar{X} \sim N(20.5, \frac{1}{100})$

$$P(\bar{X} < 19.825) = \Phi\left(\frac{19.825 - 20.5}{0.1}\right)$$
$$= \Phi\left(\frac{-0.675}{0.1}\right)$$
$$= \Phi(-6.75)$$

$$P(\bar{X} > 20.175) = 1 - \Phi\left(\frac{20.175 - 20.5}{0.1}\right)$$
$$= 1 - \Phi\left(\frac{-0.325}{0.1}\right)$$
$$= 1 - \Phi(-3.25)$$

$$\text{power} = \underbrace{\Phi(-6.75)}_{\text{area under curve will be VERY small } \therefore \text{ assume it's 0}} + 1 - \underbrace{\Phi(-3.25)}_{\text{symmetry} \rightarrow \text{normal distribution}}$$

$$\therefore \Phi(-3.25) = 1 - \Phi(3.25)$$

$$\approx 0 + 1 - (1 - \Phi(3.25))$$
$$\approx \Phi(3.25)$$

5e. The graph shows how the power of the test changes as the true value of μ varies between 19 and 21. When μ is far from 20, the power increases, approaching 1, meaning that the test becomes more likely to correctly reject the null hypothesis $H_0 : \mu = 20$ when there is a true difference. As μ approaches 20, the power decreases, converging to the test's level (≈ 0.0801) when $\mu = 20$. This represents the probability of incorrectly rejecting H_0. The test is effective at detecting large deviations from $\mu = 20$, but not so effective for smaller deviations, where the power remains relatively low.