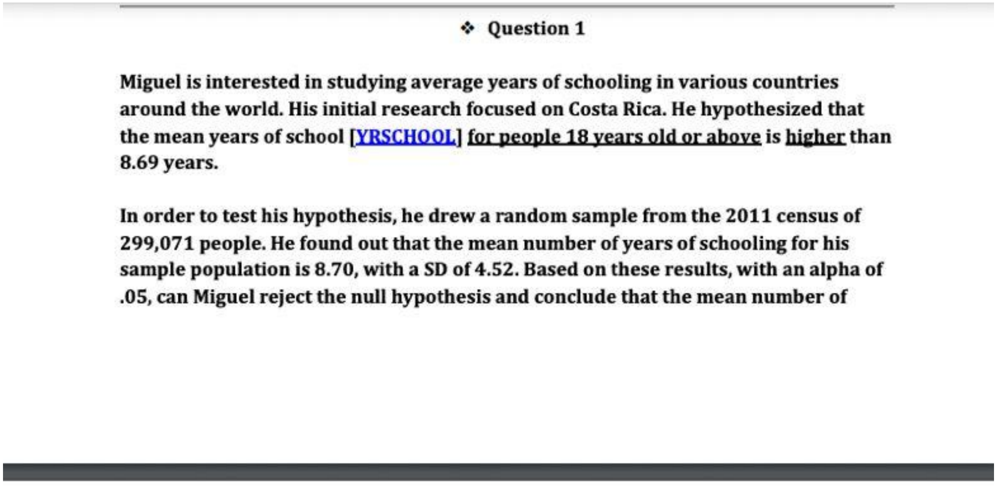
Exercise 18.5 (Casella and Berger 8.14) For a random sample $X_1, \cdots, X_n$ of $Bernoulli(p)$ variables, it is desired to test $H_0 : p = 0.49$ versus $H_1 : p = 0.51$. Use the Central Limit Theorem to determine, approximately, the sample size needed so that the two probabilities of error are both about 0.01. Use a test function that rejects H_0 if $\sum_{i=1}^n X_i$ is large.

Proof. By CLT, $Z = \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}}$ is approximately $N(0, 1)$. For a test that rejects H_0 when $\sum_{i=1}^n X_i > c$, we need to find c and n to satisfy

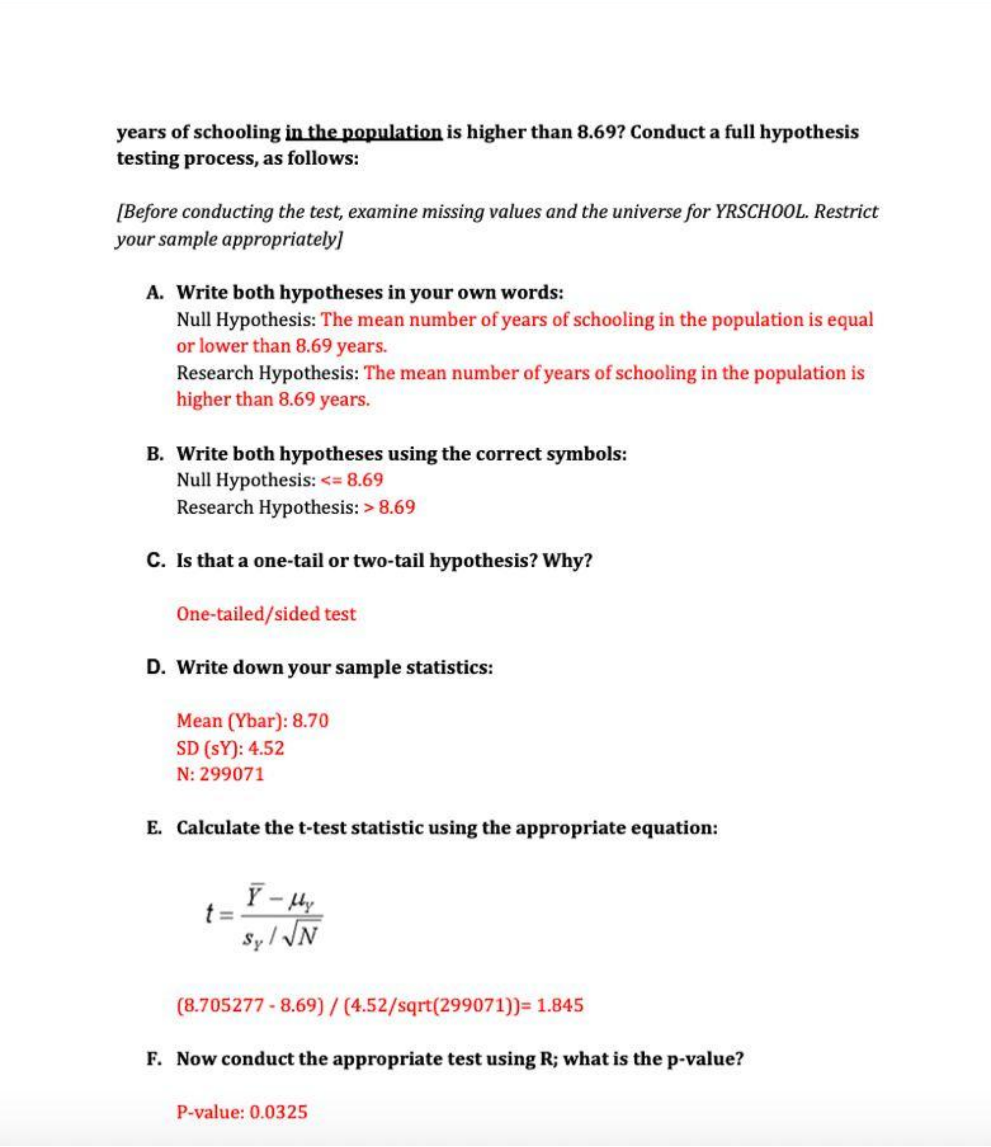
$$\begin{aligned} P(Z > \frac{c - 0.49n}{\sqrt{0.49 \cdot 0.51n}}) = 0.01 \\ P(Z > \frac{c - 0.51n}{\sqrt{0.49 \cdot 0.51n}}) = 0.99 \end{aligned} \tag{18.17}$$



❖ **Question 1**

Miguel is interested in studying average years of schooling in various countries around the world. His initial research focused on Costa Rica. He hypothesized that the mean years of school [YRSCHOOL] for people 18 years old or above is higher than 8.69 years.

In order to test his hypothesis, he drew a random sample from the 2011 census of 299,071 people. He found out that the mean number of years of schooling for his sample population is 8.70, with a SD of 4.52. Based on these results, with an alpha of .05, can Miguel reject the null hypothesis and conclude that the mean number of



years of schooling in the population is higher than 8.69? Conduct a full hypothesis testing process, as follows:

[Before conducting the test, examine missing values and the universe for YRSCHOOL. Restrict your sample appropriately]

A. Write both hypotheses in your own words:

Null Hypothesis: The mean number of years of schooling in the population is equal or lower than 8.69 years.

Research Hypothesis: The mean number of years of schooling in the population is higher than 8.69 years.

B. Write both hypotheses using the correct symbols:

Null Hypothesis: <= 8.69

Research Hypothesis: > 8.69

C. Is that a one-tail or two-tail hypothesis? Why?

One-tailed/sided test

D. Write down your sample statistics:

Mean (Ybar): 8.70
SD (sY): 4.52
N: 299071

E. Calculate the t-test statistic using the appropriate equation:

$$t = \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{N}}$$

(8.705277 - 8.69) / (4.52/sqrt(299071))= 1.845

F. Now conduct the appropriate test using R; what is the p-value?

P-value: 0.0325

G. **What is the relationship between the t-test statistic and p-value stated above? Explain.**

The p-value is calculated based on the t-test statistic. In this case, we use the t-test statistic (1.845), and search in the z-table for the matching probability, which is 0.0325 (i.e. the p-value)

H. **What is the *meaning* of the p-value? Explain in your own words.**

If the mean number of years of schooling in the (hypothetical) population (i.e. all Costa Ricans age 18 and older) is equal or lower than 8.69 years (in other words: if the null is true), then the probability of obtaining a test statistic as or more extreme than the one calculated is 0.0325.

I. **What conclusion can Carlos draw from these results?**

Since the p-value is less than 0.05, we reject the null hypothesis at the 0.05 level. We have enough evidence to conclude that the mean number of years of schooling in the population (i.e. people in Costa Rica) is higher than 8.69 years.

J. **What *is* a Type I error, and what is its probability in our case?**

A Type 1 error occurs if the null hypothesis is rejected when, in fact, the null is true. Therefore, the probability of a Type I error is equal to alpha, which in this case is 0.05.

Example 2

A group of 5 patients treated with medicine. A is of weight 42,39,38,60 &41 kgs. Second group of 7 patients from the same hospital treated with medicine B is of weight 38, 42, 56, 64, 68, 69, & 62 kgs. Find whether there is any difference between medicines?

Solution

Ho:., μ_1=μ_2 (i.e) there is no significant difference between the medicines A and B as regards on increase in weight.

H_1 $\mu_{1\neq}\mu_2$ (i.e) there is a significant difference between the medicines A and B

Level of significance = 5%

Before we go to test the means first we have to test their variability using F-test.

F-test

Ho:., $\sigma_1^2=\sigma_2^2$

H1:., $\sigma_1^2\neq\sigma_2^2$

$$S_1^2 = \frac{\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n1}}{n1-1} = 82.5$$

$$S_2^2 = \frac{\sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n2}}{n2-1} = 154.33$$

$$\therefore F = \frac{S_2^2}{S_1^2} \sim F_{(n_2-1, n_{1-}1)}\, d.f \text{ if } S_2^2 > S_1^2$$

$$F_{cal} = \frac{154.33}{32.5} = 1.8707$$

F_{tab}(6,4) d.f=6.16

$\Rightarrow F_{cal}<F_{tab}$

We accept the null hypothesis H_0.(i.e) the variances are equal.

Test statistic

$$t = \frac{\left|(\bar{x}_1 - \bar{x}_2)\ \right|}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{(n1+n2-2)}d.f$$

Where

$$S^2 = \frac{\left[\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n1}\right] + \left[\sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n2}\right]}{n_1+n_2-2} = \frac{330+926}{10} = 125.6$$

$$t = \frac{|44-57|}{\sqrt{125.6\left(\frac{1}{7}+\frac{1}{75}\right)}} = 1.98$$

Table value

$t_{tab[(5+7-2)=10]}$d.f at 5% l.o.s = 2.228

Inference:

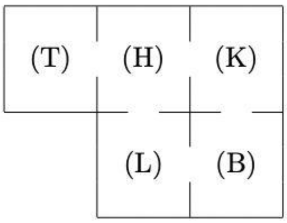
$t_{cal} < t_{tab}$

We accept the null hypothesis H_0

We conclude that the medicines A and also B do not differ significantly.

-4

Exercise 44. There is a mouse in the house! The mouse moves in such a way that in each room, it chooses one of the adjacent rooms (each with the same probability) and runs there (the movements occur at times $n = 1, 2, \ldots$). This is what our flat looks like:



The rooms are the hall (H), kitchen (K), toilet (T), bedroom (B) and the living room (L). We can set two traps - one is mischievously installed in the bedroom and the other one in the kitchen since we really should not have a mouse there. As soon as the mouse enters a room with a trap, it is caught and it will never ever run into another room. Denote by X_n the position of the mouse at time n. Classify the states of the Markov chain $X = (X_n, n \in \mathbb{N}_0)$ and compute the matrix of absorption probabilities $\boldsymbol{U}$.

Answer to Exercise 44. The states T, H, L are 2-periodic, transient and K and B are both absorbing. The absorption probabilities are given by

$$\hat{\boldsymbol{U}} = \begin{pmatrix} u_{TK} & u_{TB} \\ u_{HK} & u_{HB} \\ u_{LK} & u_{LB} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

Answer to Exercise 45. The states T, H, L are 2-periodic, transient and the states K, B and G are absorbing. The absorption probabilities are given by

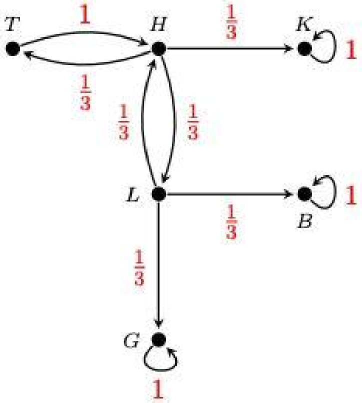
$$\hat{\boldsymbol{U}} = \begin{pmatrix} u_{TK} & u_{TB} & u_{TG} \\ u_{HK} & u_{HB} & u_{HG} \\ u_{LK} & u_{LB} & u_{LG} \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{3}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{pmatrix}.$$

In particular, the probability that the mouse will escape the house before being caught in a trap when it starts at the toilet is $u_{TG} = 1/5$.

Exercise 45. Modify the previous Exercise. Suppose now that we open the door from our living room (L) to the garden (G). Once the mouse leaves the flat and enters the garden, it will never come inside again. If the mouse starts in the toilet, what is the probability that it will escape the flat before it is caught in a trap?

Solution to Exercise 45. Building the formal model (i.e. defining X_n and proving that it is a homogeneous discrete time Markov chain) is simple and could be done in a similar way as for any random walk with absorbing states (see e.g. Exercise 22). We will focus on the task at hand: finding the probabilities u_{ij}. First, we shall find the transition probabilities $\boldsymbol{P}$ and the transition diagram describing X. We have that

$$\boldsymbol{P} = \begin{array}{c} \\ K \\ B \\ G \\ T \\ H \\ L \end{array} \begin{array}{c} \begin{array}{cccccc} K & B & G & T & H & L \end{array} \\ \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \end{array}\right) \end{array}$$



Now we have two options. Either to compute the whole matrix $\hat{\boldsymbol{U}}$ or just to compute u_{TG}.
Computing $\hat{\boldsymbol{U}}$: This can be done as in the previous exercises using the formula (2.4.2). We have

$$\hat{\boldsymbol{U}} = (\boldsymbol{I}_3 - \boldsymbol{R})^{-1}\boldsymbol{Q} = \begin{pmatrix} 1 & -1 & 0 \\ -\frac{1}{3} & 1 & -\frac{1}{3} \\ 0 & -\frac{1}{3} & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{3}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{pmatrix}.$$

Of course, one can make many mistakes in computing the inverse of a matrix 3 × 3. You should always check that all the rows of your final matrix $\hat{\boldsymbol{U}}$ are probability distributions (i.e. that they sum up to 1).
Computing only u_{TG}: We can also appeal to the formula (2.4.1) and write down only those equations which interest us. We are interested in u_{LG} so let us rewrite (2.4.2) for $i = L$ and $j = G$. We obtain

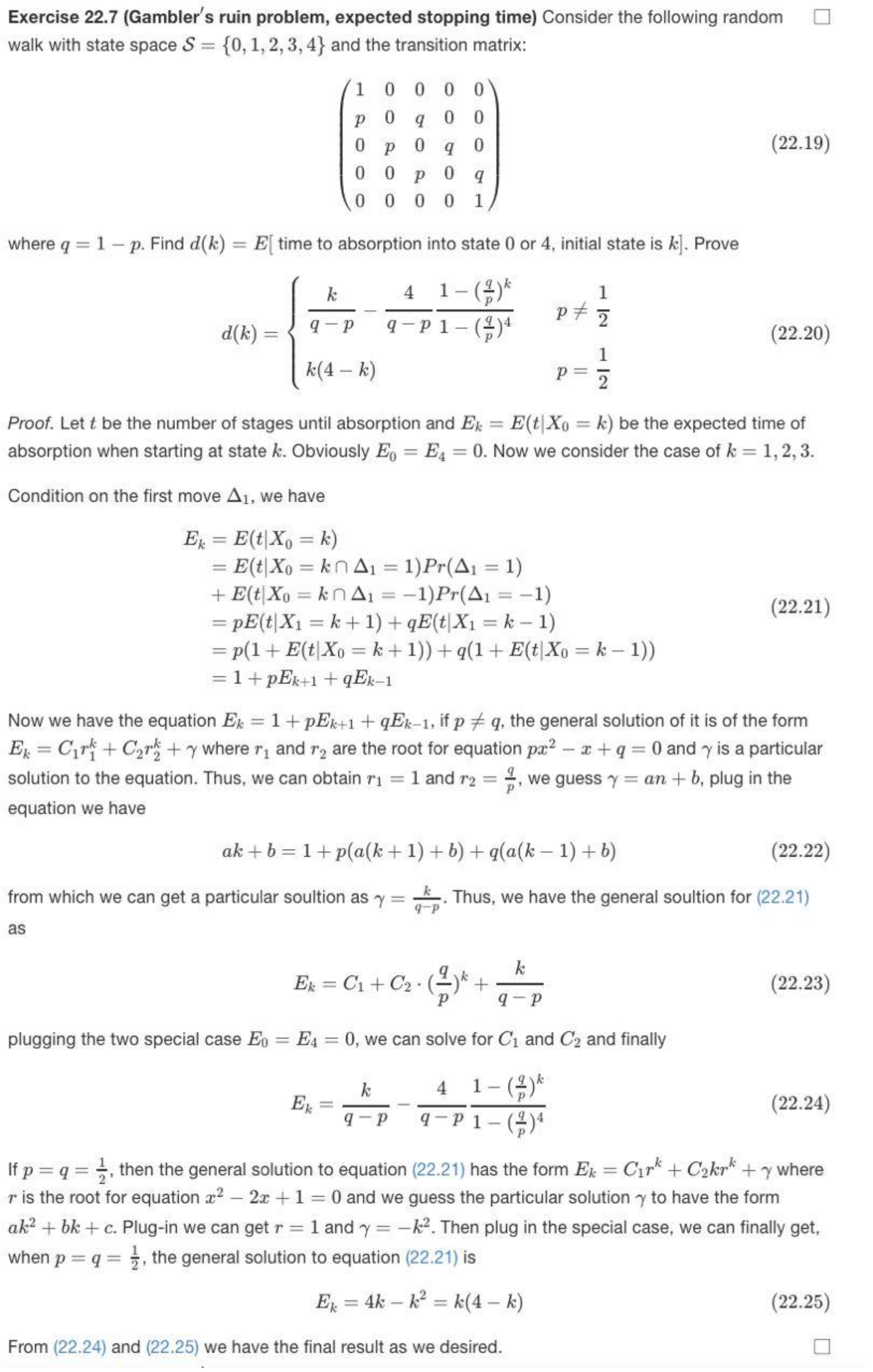
$$u_{TG} = p_{TG} + p_{TT}u_{TG} + p_{TH}u_{HG} + p_{TL}u_{LG} = u_{HG}$$

Hence, we also need an equation for u_{HG}. We again apply formula (2.4.2) and look into $\boldsymbol{P}$ to obtain

$$u_{HG} = p_{HG} + p_{HT}u_{TG} + p_{HH}u_{HG} + p_{HL}u_{LG} = \frac{1}{3}u_{TG} + \frac{1}{3}u_{LG}$$

Exercise 48. Consider the mouse moving in our flat from Exercise 44. Assume that there are no traps and the doors to the garden are closed. We wish to analyse the behavioural patterns of our mouse. The mouse moves in the same way as before - in each room, it chooses (uniformly randomly) one of the adjacent rooms and moves there. What is the long-run proportion of time the mouse spends in each of the rooms?

Answer to Exercise 48. The long-run proportion of time the mouse spends in the rooms T, H, K, B, L, is $1/10, 3/10, 2/10, 2/10, 2/10$, respectively.



Exercise 22.7 (Gambler′s ruin problem, expected stopping time) Consider the following random walk with state space $\mathcal{S} = \{0, 1, 2, 3, 4\}$ and the transition matrix: □

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ p & 0 & q & 0 & 0 \\ 0 & p & 0 & q & 0 \\ 0 & 0 & p & 0 & q \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{22.19}$$

where $q = 1 - p$. Find $d(k) = E[$ time to absorption into state 0 or 4, initial state is $k]$. Prove

$$d(k) = \begin{cases} \dfrac{k}{q-p} - \dfrac{4}{q-p}\dfrac{1-(\frac{q}{p})^k}{1-(\frac{q}{p})^4} & p \neq \dfrac{1}{2} \\ k(4-k) & p = \dfrac{1}{2} \end{cases} \tag{22.20}$$

Proof. Let t be the number of stages until absorption and $E_k = E(t|X_0 = k)$ be the expected time of absorption when starting at state k. Obviously $E_0 = E_4 = 0$. Now we consider the case of $k = 1, 2, 3$.

Condition on the first move Δ_1, we have

$$\begin{aligned} E_k &= E(t|X_0 = k) \\ &= E(t|X_0 = k \cap \Delta_1 = 1)Pr(\Delta_1 = 1) \\ &+ E(t|X_0 = k \cap \Delta_1 = -1)Pr(\Delta_1 = -1) \\ &= pE(t|X_1 = k+1) + qE(t|X_1 = k-1) \\ &= p(1 + E(t|X_0 = k+1)) + q(1 + E(t|X_0 = k-1)) \\ &= 1 + pE_{k+1} + qE_{k-1} \end{aligned} \tag{22.21}$$

Now we have the equation $E_k = 1 + pE_{k+1} + qE_{k-1}$, if $p \neq q$, the general solution of it is of the form $E_k = C_1 r_1^k + C_2 r_2^k + \gamma$ where r_1 and r_2 are the root for equation $px^2 - x + q = 0$ and γ is a particular solution to the equation. Thus, we can obtain $r_1 = 1$ and $r_2 = \frac{q}{p}$, we guess $\gamma = an + b$, plug in the equation we have

$$ak + b = 1 + p(a(k+1) + b) + q(a(k-1) + b) \tag{22.22}$$

from which we can get a particular soultion as $\gamma = \frac{k}{q-p}$. Thus, we have the general soultion for (22.21) as

$$E_k = C_1 + C_2 \cdot (\frac{q}{p})^k + \frac{k}{q-p} \tag{22.23}$$

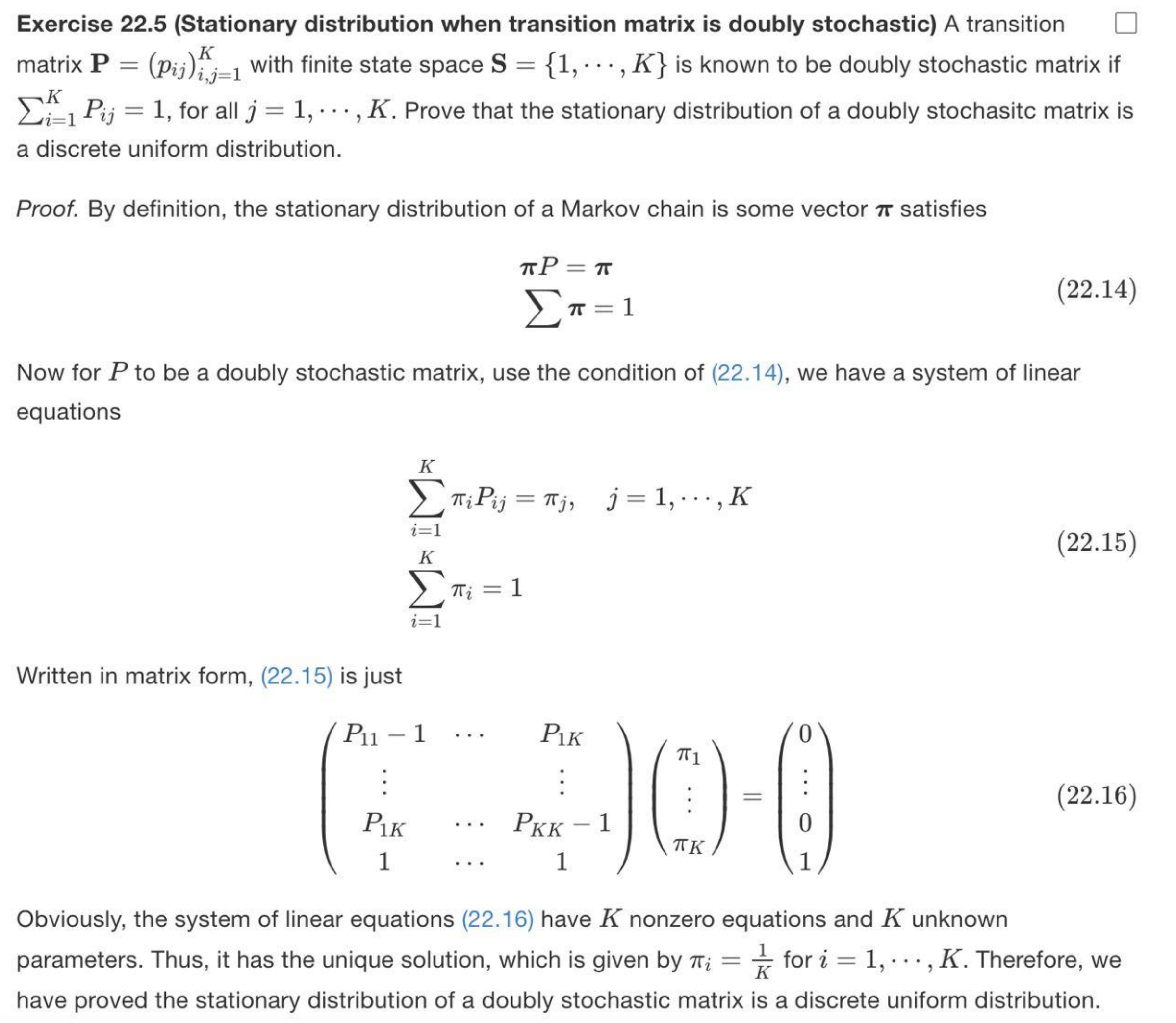
plugging the two special case $E_0 = E_4 = 0$, we can solve for C_1 and C_2 and finally

$$E_k = \frac{k}{q-p} - \frac{4}{q-p}\frac{1-(\frac{q}{p})^k}{1-(\frac{q}{p})^4} \tag{22.24}$$

If $p = q = \frac{1}{2}$, then the general solution to equation (22.21) has the form $E_k = C_1 r^k + C_2 k r^k + \gamma$ where r is the root for equation $x^2 - 2x + 1 = 0$ and we guess the particular solution γ to have the form $ak^2 + bk + c$. Plug-in we can get $r = 1$ and $\gamma = -k^2$. Then plug in the special case, we can finally get, when $p = q = \frac{1}{2}$, the general solution to equation (22.21) is

$$E_k = 4k - k^2 = k(4-k) \tag{22.25}$$

From (22.24) and (22.25) we have the final result as we desired. □



Exercise 22.5 (Stationary distribution when transition matrix is doubly stochastic) A transition matrix $\mathbf{P} = (p_{ij})_{i,j=1}^{K}$ with finite state space $\mathbf{S} = \{1, \cdots, K\}$ is known to be doubly stochastic matrix if $\sum_{i=1}^{K} P_{ij} = 1$, for all $j = 1, \cdots, K$. Prove that the stationary distribution of a doubly stochasitc matrix is a discrete uniform distribution. □

Proof. By definition, the stationary distribution of a Markov chain is some vector $\boldsymbol{\pi}$ satisfies

$$\begin{aligned} \boldsymbol{\pi} P &= \boldsymbol{\pi} \\ \sum \boldsymbol{\pi} &= 1 \end{aligned} \tag{22.14}$$

Now for P to be a doubly stochastic matrix, use the condition of (22.14), we have a system of linear equations

$$\begin{aligned} &\sum_{i=1}^{K} \pi_i P_{ij} = \pi_j, \quad j = 1, \cdots, K \\ &\sum_{i=1}^{K} \pi_i = 1 \end{aligned} \tag{22.15}$$

Written in matrix form, (22.15) is just

$$\begin{pmatrix} P_{11} - 1 & \cdots & P_{1K} \\ \vdots & & \vdots \\ P_{1K} & \cdots & P_{KK} - 1 \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_K \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{22.16}$$

Obviously, the system of linear equations (22.16) have K nonzero equations and K unknown parameters. Thus, it has the unique solution, which is given by $\pi_i = \frac{1}{K}$ for $i = 1, \cdots, K$. Therefore, we have proved the stationary distribution of a doubly stochastic matrix is a discrete uniform distribution.

1. Assume that we have the Hidden Markov Model (HMM) depicted in the figure above. If each of the states can take on k different values and a total of m different observations are possible (across all states), how many parameters are required to fully define this HMM? Justify your answer.

★ **SOLUTION:** There are a total of three probability distributions that define the HMM, the initial probability distribution, the transition probability distribution, and the emission probability distribution. There are a total of k states, so k parameters are required to define the initial probability distribution (we'll ignore all of the -1s for this problem to make things cleaner). For the transition distribution, we can transition from any one of k states to any of the k states (including staying in the same state), so k^2 parameters are required. Then, we need a total of km parameters for the emission probability distribution, since each of the k states can emit each of the m observations.
Thus, the total number of parameters required are $k + k^2 + km$. Note that the number of parameters does not depend on the length of the HMMs.

3. Using the forward algorithm, compute the probability that we observe the sequence $O_1 = 0$, $O_2 = 1$, and $O_3 = 0$. Show your work (i.e., show each of your alphas).

★ **SOLUTION:** The values of the different alphas and the probability of the sequence are as follows (notice that your answers may be slightly different if you kept a different number of decimal places):
$\alpha_1^A = 0.8 \times 0.99 = 0.792$
$\alpha_1^B = 0.1 \times 0.001 = 0.001$
$\alpha_2^A = 0.2(0.792(0.99) + 0.001(0.01)) = 0.156818$
$\alpha_2^B = 0.9(0.792(0.01) + 0.001(0.99)) = 0.008019$
$\alpha_3^A = 0.8(0.156818(0.99) + 0.008019(0.01)) = 0.124264$
$\alpha_3^B = 0.1(0.156818(0.01) + 0.008019(0.99)) = 0.000950699$
$P(\{O_T\}_{t=1}^T) = 0.1252147$

State	$P(S_1)$
A	0.99
B	0.01

(a) Initial probs.

S_1	S_2	$P(S_2\|S_1)$
A	A	0.99
A	B	0.01
B	A	0.01
B	B	0.99

(b) Transition probs.

S	O	$P(O\|S)$
A	0	0.8
A	1	0.2
B	0	0.1
B	1	0.9

(c) Emission probs.

4. Using the backward algorithm, compute the probability that we observe the aforementioned sequence ($O_1 = 0$, $O_2 = 1$, and $O_3 = 0$). Again, show your work (i.e., show each of your betas).

★ **SOLUTION:** The values of the different betas and the probability of the sequence are as follows (again, your answers may vary slightly due to rounding):
$\beta_3^A = 1$
$\beta_3^B = 1$
$\beta_2^A = 0.99(0.8)(1) + 0.01(0.1)(1) = 0.793$
$\beta_2^B = 0.01(0.08)(1) + 0.99(0.01)(1) = 0.107$
$\beta_1^A = 0.99(0.02)(0.793) + 0.01(0.9)(0.107) = 0.157977$
$\beta_1^B = 0.01(0.2)(0.793) + 0.99(0.9)(0.107) = 0.096923$
$P(\{O_T\}_{t=1}^{T}) = 0.792(0.157977) + 0.001(0.096923) = 0.1252147$

7. Use the Viterbi algorithm to compute (and report) the most likely sequence of states. Show your work (i.e., show each of your Vs).

★ **SOLUTION:** The Viterbi algorithm predicts that the most likely sequence of states is A, A, A. The relevant computations are:
$V_1^A = 0.99 \times 0.8 = 0.792$
$V_1^B = 0.01 \times 0.1 = 0.001$
$V_2^A = 0.2(0.792)(0.99) = 0.156816$
$V_2^B = 0.9(0.792)(0.01) = 0.007128$
$V_3^A = 0.8(0.156816)(0.99) = 0.1241983$
$V_3^B = 0.1(0.007128)(0.99) = 0.000705672$