Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

Exercise 2

Social Data Science

1 The Monty Hall Problem

Imagine you participate in a game show, in which you have to choose one out of three doors. Behind one door, there is a car, whereas behind both other doors there is goat. However, which door conceals which is unknown to you, the player. Your aim is to select the door behind which the car is, and you make your initial guess by approaching the door of your choice without opening it. At this point, regardless of which door you selected, the game show host chooses and opens one of the remaining two doors. If you chose the door with the car, the host selects one of the two remaining doors at random (with equal probability) and opens that door. If you chose a door with a goat, the host selects and opens the other door with a goat. You are given the option of standing where you are and switching to the other closed door. Apply Bayes Theorem to argue whether or not switching to the other door increases your chances of winning!
Hint: Consider the conditional probability that the car is behind your chosen door, given that the game show host reveals a door with a goat behind it

W.l.o.g. we selct door 1, and the moderator then opens door 3

- C_i denotes the event that the car is behind door i
- M_j denotes the event that the moderator opens door j

↳ we want to know prob. $P(C_2 | M_3)$

We know that $P(C_i) = \frac{1}{3}$,
and further:

- $P(M_3 | C_1) = \frac{1}{2}$
- $P(M_3 | C_2) = 1$

- $P(M_3 | C_3) = 0$

With Bayes Theorem, we have that

$$P(C_2 | M_3) = \frac{P(M_3 | C_2) \cdot P(C_2)}{P(M_3)}$$

Law of total probability

$$= \frac{P(M_3 | C_2) \cdot P(C_2)}{\underbrace{P(M_3 | C_1) \cdot P(C_1)}_{P(M_3, C_1)} + \underbrace{P(M_3 | C_2) \cdot P(C_2)}_{P(M_3, C_2)} + \underbrace{P(M_3 | C_3) \cdot P(C_3)}_{P(M_3, C_3)}} = P(M_3)$$

$$= \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{2}{3}$$

↳ Because $P(C_1 | M_3) + P(C_2 | M_3) = 1$, it is now more likely that the car is behind door 2 than behind door 1.

2 Descriptive Statistics

Assume that we have counted the calendar weeks in Aachen in it has snowed at some point in time for the last 20 years, with the counts per year given as in the following tables.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Snow Weeks	5	4	5	5	3	4	3	3	3	4

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Snow Weeks	3	4	2	2	3	1	2	3	2	1

a) Compute the mean, median, and variance of the number of snow weeks in the last 20 years!

b) Sketch the probability mass function (PMF) and the cumulative distribution function (CDF) of the probability distribution of snow weeks per year.

Value	1	2	3	4	5
#	2	4	7	4	3

$n = 20$

a) mean: $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{20} (2 \cdot 1 + 4 \cdot 2 + 7 \cdot 3 + 4 \cdot 4 + 3 \cdot 5) = 3.1$

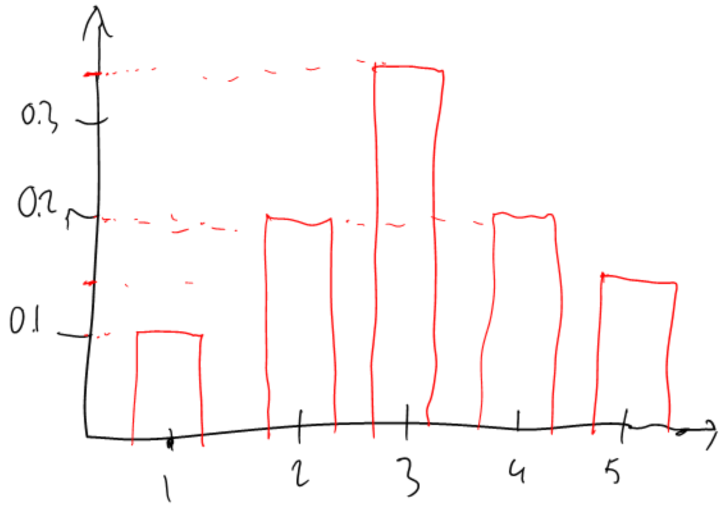
median $= 3$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{20} \cdot \Big[2 \cdot (1-3.1)^2 + 4 \cdot (2-3.1)^2 + 7 \cdot (3-3.1)^2 + 4 \cdot (4-3.1)^2 + 3 \cdot (5-3.1)^2 \Big]$$

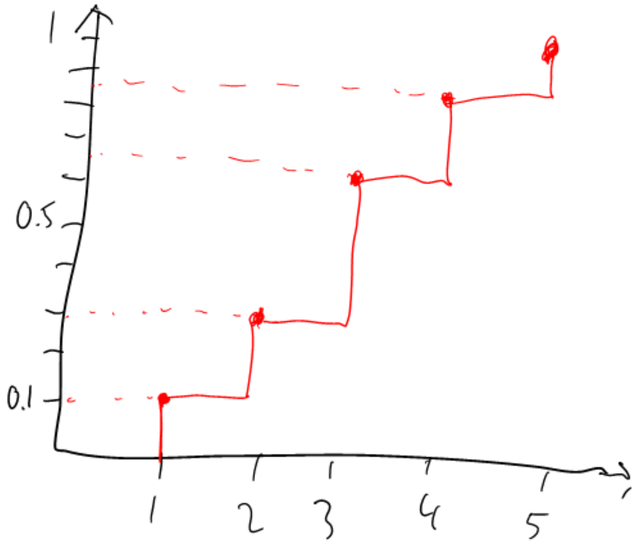
$$= 1.409$$

Value	1	2	3	4	5
#	2	4	7	4	3
PMF	$\frac{2}{20}=0.1$	$\frac{4}{20}=0.2$	0.35	0.2	0.15
CDF $=\sum PMF$	0.1	0.3	0.65	0.85	1

PMF:



CDF:

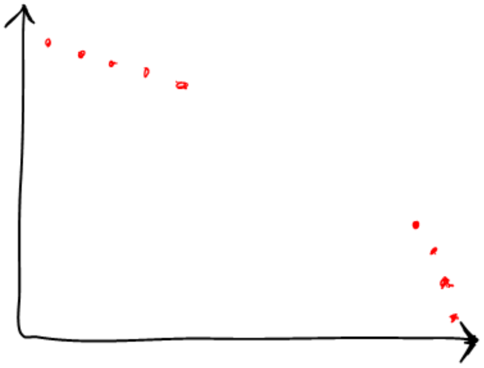


3 Correlation

a) Sketch a scatterplot of two variables $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$, where the Pearson correlation coefficient between X and Y would be low, but Spearman's correlation coefficient would be high. Explain your answer!

b) Suppose you have two numerical vectors of empirical observations, X and Y, both consisting of positive real numbers strictly greater than 0. The Spearman correlation between X and Y is .5. Does the Spearman correlation between $\log(X)$ and $\log(Y)$ increase, decrease, or stay the same? Explain your answer.

a)



- no linear relationship any more, but still a "monotonic" correlation

b) The Spearman correlation would stay the same, because the log function is strictly monotonic and thus would not change the rankings of X and Y.