1 Quiz 2 - Probability - Discrete Random Variables and Distributions

1.1 Part I: Probability

Exercise 1 (20 = 10 10 pts.)

(a) Suppose we roll two 6 -sided die. Consider the events: $A =$ 'odd on die 1 , $B =$ 'odd on die 2 ' $C =$ 'odd sum Are A, B, and C pairwise independent? Are they mutually independent?

(b) For families with n children, the events 'the family has children of both sexes' and 'there is at most one girl' are independent. What is n?

Hint:

1) Three events A, B, and C are pairwise independent if each pair is independent. They are mutually independent if they are pairwise independent and in addition

$$P(A \cap B \cap C) = P(A)P(B)P(C) \tag{1}$$

2) If you let X be the number of girls then we have

$$P(A) = P(1 \leq X \leq n-1) \tag{2}$$
$$P(B) = P(X \leq 1) \tag{3}$$
$$P(A \cap B) = P(X = 1) \tag{4}$$

Since $X \sim \text{binomial}(n, 1/2)$ compute the above probabilities.

```
**your answers here**
```

your answers here

Exercise 2 (15 pts.) **R simulation**. Suppose there is an experimental medical treatment for a cancer that if untreated is nearly always fatal within $12 - 15$ months. The doctors enroll 5000 patients in a study in which each patient is given the treatment and followed for 5 years. Let X be the length of time a random patient given the treatment survives. (If a patient is still alive at the end of the study, then $X = 5$ for this patient.)

(a) Compute the mean and standard deviation of the data in years

(b) Plot a frequency histogram of the data. Set the histogram so each bin has width 0.1 years.

(c) Using your answers in (a) and (b), approximate how many live after the mean and what percent are still alive after 5 years?

The following code loads the data. You must have the r code `problemdata.r` in the same directory with this notebook.

In [ ]:
```
# your code for (a) and (b)
source('problemdata.r')
years = getdata()
```

your answer for (c) here

your answer for (c) here

Exercise 3 (10 pts) Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both. What is the probability that a car has a CD player, given that it has AC ?

In [8]:
```
#your answer here
```

Definition 1 *ODDS: The odds in favor of a particular event are given by the ratio of the probability of the event divided by the probability of its complement*
The odds in favor of A are

$$\text{odds} = \frac{\text{P(A)}}{1 - \text{P(A)}} = \frac{\text{P(A)}}{\text{P}(\overline{\text{A}})} \tag{5}$$

Exercise 4 (5 pts) Calculate the probability of winning if the odds of winning are 5 to 1

In [ ]:
```
# your answer here
```

Definition 2

Bayes' Theorem-Rule

Bayes' theorem is a pillar of both probability and statistics. For two events A and B Bayes' theorem (also called Bayes' rule and Bayes' formula) says

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)} \tag{6}$$

Comments:

1. *Bayes' rule tells us how to 'invert' conditional probabilities, i.e. to find $P(B \mid A)$ from $P(A \mid B)$*

2. *In practice, $P(A)$ is often computed using the law of total probability.*

*Note that $P(B)$ is called the **prior probability**, $P(A \mid B)$ is called **posterior probabilities**, $P(B \mid A)$ the **likelihood probability**, and $P(A)$ is called evidence.*

Remark 1 Bayes' Rule: Hypothesis and Data *If we consider a disease to represent a specific hypothesis, and the symptoms to be some observed data then Bayes' rule becomes*

$$p(\text{hypothesis} \mid \text{ data}) = \frac{p(\text{ data} \mid \text{ hypothesis}) \times p(\text{ hypothesis})}{p(\text{ data})} \tag{7}$$

where the word "hypothesis" should be interpreted as, "hypothesis is true". Specifically, the probability that the proposed hypothesis is true given some data that were actually observed is the posterior probability

$$p(\text{ hypothesis} \mid \text{ data}) \tag{8}$$

whereas the probability of observing the data given that the hypothesis is true is the likelihood

$$p(\text{data} \mid \text{ hypothesis}) \tag{9}$$

Example 1 Doctor's perspective Using the results of a survey, a doctor can calculate the proportion of patients diagnosed with *smallpox* and *chickenpox_, and each patient's _symptoms* (eg _spots_). Using these data, we might find that the probability that a patient has spots given that the patient has smallpox is 90 or 0.9. We can write this

$$p(\text{ symptoms are spots} \mid \text{ disease is smallpox}) = 0.9 \tag{10}$$

So, this short-hand statement should be read as "the probability that the patient's symptoms are spots given that he has smallpox is 90% or 0.9 " Thus, the probability of spots is said to be conditional on the disease under consideration. For this reason, such probabilities are known as *conditional probabilities*.

Similarly, we might find that spots are observed in 80% of patients who have chickenpox, which is written as

$$p(\text{ spots} \mid \text{ chickenpox}) = 0.8 \tag{11}$$

Public health statistics inform us that the prevalence of **smallpox** in the general population is 0.001. This can be written as

$$p(\text{ smallpox}) = 0.001 \tag{12}$$

Thus, according to Baye's rule we have

$$p(\text{ smallpox} \mid \text{ spots}) = \frac{p(\text{ spots} \mid \text{ smallpox}) \times p(\text{ smallpox})}{p(\text{ spots})} \tag{13}$$

Thus,

$$\begin{aligned} p(\text{ smallpox} \mid \text{ spots}) &= 0.9 \times 0.001/0.081 \\ &= 0.011 \end{aligned} \tag{14}$$

which is the conditional probability that the patient has smallpox given that his symptoms are spots.

In [9]:
```
# likelihood =prob of spots given chickenpox
pSpotsGChickenpox <- 0.8;
# prior =prob of chickenpox
pChickenpox <- 0.1;
# marginal likelihood=prob of spots
pSpots <- 0.081;
# find posterior=prob of chickenpox given spots
pChickenpoxGSpots <- pSpotsGChickenpox * pChickenpox / pSpots;
#print
s <-sprintf('pChickenpoxGSpots = %.3f', pChickenpoxGSpots);
print(s)
# Output:pChickenpoxGSpots=0.988
```

```
[1] "pChickenpoxGSpots = 0.988"
```

Exercise 5 (10 pts) You might be interested in finding out a patient's probability of having liver disease if they are an alcoholic. "Being an alcoholic" is litmus test for liver disease.

Let A is the event "Patient has liver disease." Past data tells you that 10% of patients entering your clinic have liver disease. $P(A) = 0.10$.

Let B is the litmus test that "Patient is an alcoholic." Five percent of the clinic's patients are alcoholics that is $P(B) = 0.05$.

You also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your $B|A$: the probability that a patient is alcoholic, given that they have liver disease, is 7%.

Use Bayes theorem to find the probability of a patient to have liver decease.

In [ ]:
```
#your answer here
```

1.2 Part II: Discrete random variables and Distributions

1.2.1 Discrete random variables

Exercise 6 (10 = 5 5 pts) Let E be the experiment of flipping a coin twice.

(a) What is the sample space?

(b) Define the random variable X = the number of heads. What are the numbers that X takes?

1.2.2 Uniform (Discrete) Distribution

Exercise 7 (10 pts)

(a) What is the model of uniform distributions?

(b) what is used for?

(c) what are its parameters?

(d) What is its mean and variance?

your answers here

Exercise 8 (15 = 5 5 5 pts)

(a) Generate n=10 random numbers according to uniform distribution in the interval $[0, 100]$ and plot a histogram

(b) compute the mean and variance manually and computationally using the r functions of the above numbers. what do you observe? What do you observe if you increase n?

(c) describe mathematically (latex) the pdf and cdf of the uniform distribution

(d) Plot pdf and cdf for uniform distribution

In [ ]:
```
# your code here
```

1.3 Bernoulli Distributions

Exercise 9 (10 pts)

(a) What is the model of Bernoulli distributions?

(b) what is used for?

(c) what are its parameters?

(d) what is the theoretical mean and variance of the Bernoulli distribution?

```
**your answers here**
```

your answers here

Exercise 10 (20 = 5 5 5 5 pts)

(a) Generate n=10 random numbers according to Bernoulli distribution for $p = 0.7$ and plot a histogram

(b) compute the mean and variance manually and computationally using the r functions of the above numbers. what do you observe? What do you observe if you increase n?

(c) describe mathematically (latex) the pdf and cdf of the uniform distribution

(d) Plot pdf and cdf for uniform distribution

For Bernoulli we have the following R functions:

- Bernoulli Probability Density Function (dbern Function)
- Bernoulli Cumulative Distribution Function (pbern Function)
- Bernoulli Quantile Function (qbern Function)
- Generating Random Numbers (rbern Function)

In [7]:
```
# your answers here
install.packages("Rlab")                    # Install Rlab package
library("Rlab")                             # Load Rlab package

set.seed(98989)                             # Set seed for reproducib
N <- 10000
```

```
Warning message:
"package 'Rlab' is in use and will not be installed"
```

1.4 Binomial Distribution

Exercise 11 (15 - 3 3 3 3 3 pts)

(a) What is the model of binomial distributions?

(b) what is used for?

(c) what are its parameters?

(d) what is its relation with Bernoulli?

(f) What is the theoretical mean and variance of of binomial distributions?

```
**your answers here**
```

your answers here

Exercise 12 (5 pts) (e) What is the probability of 5 or more heads in 10 tosses of a fair coin?

In [ ]:
```
#your answer here
```

Exercise 13 In a learning quiz there are twelve multiple choice questions. Each question has five possible answers, and only one of them is correct. Find the probability of having 3 or less correct answers if a student attempts to answer every question at random.

In [ ]:
```
# you code here
```

1.5 Poisson Distribution

Exercise 14 (15 = 3 3 3 6 pts)

(a) What is the model of Poisson distributions?

(b) what is used for? Give an example

(c) what are its parameters?

(f) What is the theoretical mean and variance of Poisson distributions?

```
**your answers here**
```

your answers here

Exercise 15 (15 pts) Some vehicles pass through a junction on a busy road at an average rate of 300 per hour.

(a) Find out the probability that none passes in a given minute.

(b) What is the expected number of passing in two minutes?

(c) Find the probability that this expected number found above actually pass through in a given two-minute period.

In [ ]:
```
#your answers here
```

1.6 GEOMETRIC DISTRIBUTION

Conditions:

1. An experiment consists of repeating trials until first success.
2. Each trial has two possible outcomes;

(a) A success with probability p

(b) A failure with probability $q = 1 - p$.

3. Repeated trials are independent. X = number of trials to first success

X is a GEOMETRIC RANDOM VARIABLE.

PDF:

$$P(X = x) = q^{x-1}p; x = 1, 2, 3, \cdots \tag{15}$$

CDF:

$$\begin{aligned} P(X \leq x) &= P(X = 1) + P(X = 2) + \cdots + P(X = x) \\ &= p + qp + q^2p + \cdots + q^{x-1}p \\ &= p\left[1 - q^x\right]/(1 - q) \\ &= 1 - q^x \end{aligned} \tag{16}$$

Theorem 1

(a) The mean of a geometric random variable X is:

$$\mu = E(X) = \frac{1}{p} \tag{17}$$

(b) The variance of a geometric random variable X is:

$$\sigma^2 = \mathrm{Var}(X) = \frac{1 - p}{p^2} \tag{18}$$

Exercise 16 (15 pts)

(a) Products produced by a machine has a 3% defective rate. What is the probability that the first defective occurs in the fifth item inspected?

(b) What is the probability that the first defective occurs in the first five inspections?

(c) In a production line which has a 20% detective rate, what is the minimum number of inspections, that would be necessary so that the probability of observing a defective is more that 75%?

Type *Markdown* and LaTeX: α^2

In [ ]: