Name and Surname:

SDA Exam Stats Module

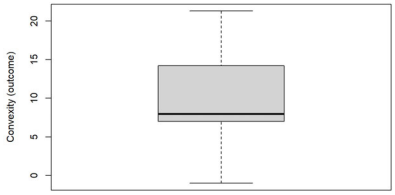
Part 1 (5 Points):

- **Clearly mark** the most appropriate answer to each question.
- You may leave unanswered questions.
- Each correct question **adds** 0.25 points to the final grade
- Each incorrect question **substracts** 0.1 point to the final grade

Questions:

1) In the the box plot below, the 1st quartile and 2nd quartlie of the data are:

a: $(-1.00, 21.30);$ **b:** $(-1.00, 7.02);$ **c:** $(7.02, 7.96);$ **d:** $(7.02, 14.22)$



2) The main disadvantage of a histogram is that:

a: It depends on the size of the bin; **b:** It cannot be used for categorical variables; **c:** It cannot be used when the bin size is small; **d:** It is used for relative frequencies only;

3) If the relative cumulative frequencies of a random experiment with outcomes $\{1, 2, 3, 4\}$ are:
$F(1) = 0.15, \quad F(2) = 0.60, \quad F(3) = 0.85, \quad F(4) = 1.$

Then the relative frequency for outcome 3 is

a: $0.15;$ **b:** $0.85;$ **c:** $0.45;$ **d:** 0.25

4) In a sample of size 10 of a random experiment we obtained the following data:

$8, \quad 3, \quad 3, \quad 7, \quad 3, \quad 6, \quad 5, \quad 10, \quad 3, \quad 8.$

The first quartile of the data is:

a: $3.5;$ **b:** $4;$ **c:** $5;$ **d:** 3

5) Imagine we collect data for two events that are not mutually exclusive, for example, the sex and nationality of passengers on a flight. If we want to make only one pie chart for the data, which of these statements is true

a: We can only make a pie chart of nationality because it has more than two possible outcomes; **b:** We can make one pie chart for the joint events of sex and nationality; **c:** We can make one pie chart for the events of sex or nationality; **d:** We can only choose whether we make one pie chart for sex or one pie chart for nationality

For questions **6-9)** consider the following:

We collect the age and category of 100 athletes in a competition

	$junior$	$senior$
$1st$	14	12
$2nd$	21	18
$3rd$	22	13

6) What is the estimated probability that an athlete is in the 2nd category and senior

a: $18/100$; **b:** $18/43$; **c:** 18; **d:** $18/39$

7) What is the estimated probability that the athlete is not in the third category and is a senior?

a: $35/100$; **b:** $30/100$; **c:** $22/100$; **d:** $13/100$

8) What is the estimated probability that the athlete is in the third category if the athlete is junior?

a: 22; **b:** $22/100$; **c:** $22/57$; **d:** $22/35$;

9) What is the estimated probability that the athlete is junior and in the 1st category if the athlete is not in the 3rd category?

a: $14/35$; **b:** $14/65$; **c:** $14/100$; **d:** $14/26$

10) A diagnostic test has a probability of $8/9$ for detecting a disease if patients are ill and a probability of $3/9$ for detecting the disease if patients are healthy. If the probability of being ill is $1/9$. What is the probability that a patient is ill if a test detects the disease?

a: $\frac{8/9}{8/9+3/9} * 1/9$; **b:** $\frac{3/9}{8/9+3/9} * 1/9$; **c:** $\frac{3/9*8/9}{8/9*1/9+3/9*8/9}$; **d:** $\frac{8/9*1/9}{8/9*1/9+3/9*8/9}$;

11) During WWII in London, the expected number of bombs that hit an area of $3km^2$ was 0.92. The probability that, in one day, one area received two bombs was

a: `1-ppois(x=2, lambda=0.92)` ; **b:** `ppois(x=2, lambda=0.92)` ;
c: `1-dpois(x=2, lambda=0.92)` ; **d:** `dpois(x=2, lambda=0.92)`

12) Opinion polls for the USA 2022 election give a probability of 0.55 that a voter favors the republican party. If we conduct our own poll and ask 100 random people on the street, How would you compute the probability that in our poll democrats win the election?

a: `pbinom(x=49, n=100, p=0.55)=0.13` ; **b:** `1-pbinom(x=49, n=100, p=0.55)=0.86` ;
c: `pbinom(x=51, n=100, p=0.45)=0.90` ; **d:** `1-pbinom(x=51, n=100, p=0.45)=0.095`

13) In an exam a student chooses at random one of the four answers that he does not know. If he doesn't know 10 questions what is the probability that at least 5 questions are correct?

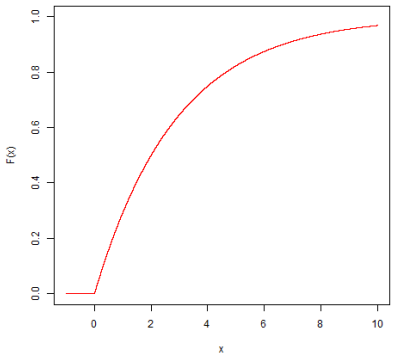
a: `dbinom(x=5, n=10, p=0.25)~ 0.05` ; **b:** `pbinom(x=5, n=10, p=0.75)~ 0.07` ;
c: `dbinom(x=5, n=10, p=0.75)~ 0.05` ; **d:** `1-pbinom(x=5, n=10, p=0.25)~ 0.02`

14) The probability that a passenger has to wait less than 20 minutes until the next bus arrives at her stop is better described by

a: A poisson model on the number of buses per 20 minutes;
b: An exponential distribution at 20 minutes with a given expectation of buses per minute; **c:** A binomial model that counts the number of buses per 20 minutes **d:** A normal distribution with an average of buses per 20 minutes and a given standard deviation;

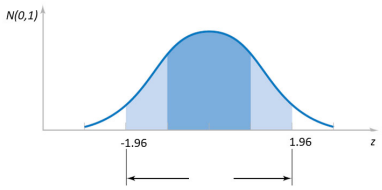
15) From the exponential probability distribution in the figure below, what is the most likely value of the median

a: 2; **b:** 3; **c:** 4; **d:** 5



16) What is the probability that a standard normal variable is between -1.96 and 1.96

a: 0.90; **b:** 0.925; **c:** 0.95; **b:** 0.975



17) A magnetic resonance imaging of the brain's hippocampus has 100 pixels. We expect 90% of the pixels to be white (brain tissue). According to the central limit theorem, what is the probability that the scanning of a patient has at most 85% of white pixels?

a: `pnorm(0.9, 0.85, sqrt(0.85*0.15)/10)` ; **b:** `dnorm(0.85, 0.9, sqrt(0.9*0.1)/10)` ;
c: `pnorm(0.85, 0.9, sqrt(0.9*0.1)/10)` ; **d:** `dnorm(0.9, 0.85, sqrt(0.85*0.15)/10)`

18) For a standard normal variable, the number $z_{0.025}$ in the definition of the margin of error $m = z_{0.025}\frac{\sigma}{\sqrt{n}}$ refers to

a: The first quartile; **b:** The number at which the distribution has accumulated 0.975 of probability;
c: The number at which the distribution has accumulated 0.025 probability; **d:** The third quartile;

19) Why is the statistic $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ used instead of $S_n^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$ to estimate the variance of a random variable?

a: because its variance is 0; **b:** because it is a consistent estimator of σ^2;
c: because it is an unbiased estimator of σ^2; **d:** because it is the averaged square distance to the sample mean ($\bar{X}$);

20) What is the variance of the sample mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X_i$?

a:σ; **b:**$\frac{\sigma}{\sqrt{n}}$; **c:**σ^2; **d:**$\frac{\sigma^2}{n}$;

R functions of probability models

Model	**X**	$f(x)$	$F(x) = P(X \leq x)$
Uniform	real number	`dunif(x, a, b)`	`punif(x, a, b)`
Binomial	# of A events in n repetitions of Bernoulli trials	`dbimon(x,n,p)`	`pbimon(x,n,p)`
Negative Binomial for events	# of B events in Bernoulli repetitions before r As are observed	`dnbinom(x,r,p)`	`pnbinom(x,r,p)`
Hypergeometric	# A events in a sample n from population N with K As	`dhyper(x, K, N-K, n)`	`phyper(x, K, N-K, n)`
Poisson	# of events A in an interval	`dpois(x, lambda)`	`ppois(x, lambda)`
Exponential	Interval between two events A	`dexp(x, lambda)`	`pexp(x, lambda)`
Normal	measurement with symmetric errors whose most likely value is the average	`dnomr(x, mu, sigma)`	`pnomr(x, mu, sigma)`

Answers Part 1:

1 c, 2 a, 3 d, 4 d, 5 b, 6 a, 7 b, 8 c, 9 b, 10 d, 11 d, 12 a, 13 d, 14 b, 15 a, 16 c, 17 c, 18 b, 19 c, 20 d,

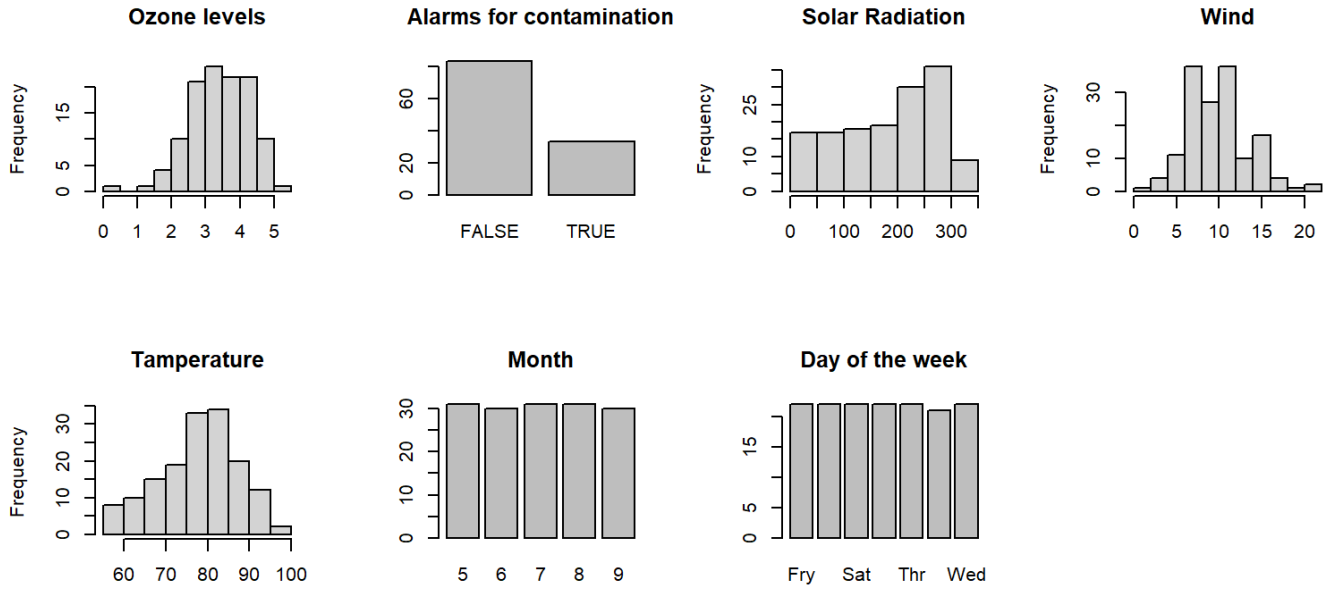
Part 2 (5 Points):

The meteorological office has given you data relating to the levels of ozone in the atmosphere for each day from May to September (153 days). They are interested to know the variables that may predict the levels of ozone pollution to inform health agencies.

Here are the first six days of the data

```
##      Ozone Solar.R Wind Temp Month Day Day.week Alarm
## 1 3.713572     190  7.4   67     5   1      Wed FALSE
## 2 3.583519     118  8.0   72     5   2      Thr FALSE
## 3 2.484907     149 12.6   74     5   3      Fry FALSE
## 4 2.890372     313 11.5   62     5   4      Sat FALSE
## 5       NA      NA 14.3   56     5   5      Sun    NA
## 6 3.332205      NA 14.9   66     5   6      Mon FALSE
```

Here is the illustration of the data



They want to know

a. Whether ozone levels are different between spring (months: 5,6) and summer (months:7,8,9).
b. Whether ozone levels exclusively depend on the temperature.
c. Whether the alarms of contamination depend on the day of the week.

Question: Write an analysis protocol, in which you specify for **each** of the research questions mentioned above clearly stating:

1. the statistical hypothesis
2. the test statistic
3. the significance criteria
4. the plot to illustrate the results
5. possible challenges