Name and Surname:

SDA Exam Stats Module

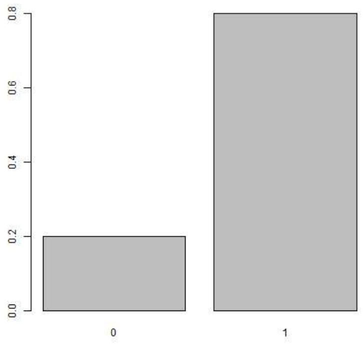
- **Clearly mark** the most appropriate answer to each question.
- You may leave unanswered questions.
- Each correct question **adds** 0.4 points to the final grade
- Each incorrect question **substracts** 0.1 point to the final grade

Questions:

1) From the following bar plot, we can say that the average of the data is at

a: 1; **b:** 0.8;

c: 0.5; **d:** 0



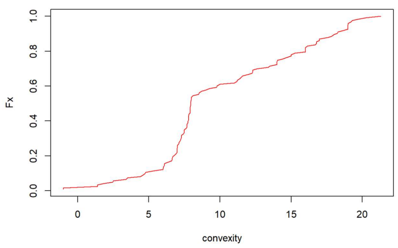
2) In the following cumulative frequency plot from the misophonic data, we can say that

a: most of the misophonics have jaws with convexity angles higher than 15;

b: most of the misophonics have jaws with convexity angles between 5 and 15;

c: most of the misophonics have jaws with convexity angles lower than 5 ;

d: we do not know which convexity angles are the most frequent



3) If the relative frequencies of a random experiment with outcomes $\{1, 2, 3, 4\}$ are:

$$f_1 = 0.15, \qquad f_2 = 0.60, \qquad f_3 = 0.05, \qquad f_4 = 0.2$$

Then the cumulative relative frequency for outcome 3 is

a: 0.8; **b:** 0.05;

c: 0.2; **d:** 0.75

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

$$3, \quad 3, \quad 10, \quad 2, \quad 6, \quad 5, \quad 4, \quad 9, \quad 9, \quad 10$$

The third quartile of the data is:

a: 0.75; **b:** 7.5;

c: 9; **d:** 6

5) A pie chart illustrates

a: the relative frequency

b: the absolute frequency

c: the absolute cumulative frequency

d: the relative cumulative frequency

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

In the leptin knock out experiment, researchers analyzed the weight of male and female mice in three conditions (control: normal mice, KOplus: leptin knock out mice and supplemented with leptin injection, leptinKO: leptin knock out mice). The contingency table for sex and conditions is

	$Female$	$Male$
$control$	24	16
$KOplus$	12	10
$leptinKO$	10	7

6) What is the relative frequency that a mouse is a knock out?

a: $17/79$; **b:** $17/22$;

c: $39/79$; **d:** $17/39$

7) What is the relative frequency that the mouse is a control?

a: $40/79$; **b:** $24/40$;

c: $40/100$; **d:** $24/79$

8) What is the estimated probability that the mouse has supplemented leptin injection if it is a knock out mouse?

a: $22/39$; **b:** $22/79$;

c: $22/40$; **d:** $39/79$;

9) What is the estimated probability that the mouse is female or control?

a: $62/79$; **b:** $24/79$;

c: $17/79$; **d:** $46/79$

10) When we applied the Bayes theorem on the PCRs for COVID19 detection, our main interest was to estimate

> **a:** the probability of having the infection if testing positive;
>
> **b:** the probability of testing positive if having the infection;
>
> **c:** the probability of having the infection;
>
> **d:** the probability of testing positive;

11) I have 15 minutes to take a taxi so I don't miss the train. If in the taxi stop where I am standing, one taxi arrives every 5 minutes on average, what is the probability that I miss the train?

a:

`poisson.pmf(x=0, lambda=1/3)=0.7` ;

b:

`1- poisson.pmf(x=0, lambda=3)=0.95` ;

c:

`poisson.pmf(x=0, lambda=3)=0.05` ;

d:

`1- poisson.pmf(x=0, lambda=1/3)=0.3`

12) Alzheimer's disease occurs in 1 out of 9 people over 65 years of age. What is the probability that, in a registry of 100 retired people over 65 years of age, we find at most 10 individuals with Alzheimer's disease?

a:

`binom.cdf(10, 100, 1/9)=0.43` ;

b:

`binom.pmf(10, 100, 1/9)=0.12` ;

c:

`1- binom.cdf(10, 100, 1/9)=0.56` ;

d:

`1- binom.pmf(10, 100, 1/9)=0.87` ;

13) A radioactive particle has an expected decay of 2 seconds. The probability that a particle decays in less than 10 seconds is better computed with

> **a:** the probability mass function of a Poissson model with parameter $\lambda = 2$;
>
> **b:** the probability distribution of a Poissson model with parameter $\lambda = 2$;
>
> **c:** the probability distribution of an exponential model with parameter $\lambda = 2$;
>
> **d:** the probability density of an exponential model with parameter $\lambda = 2$

14) In the lepting knockout experiment, control mice have a mean weight of $23gr$ and variance of $9gr^2$. If weight is a normal random variable, how do you calculate the weight that defines the top 95% of the weights?

a: `norm.ppf(0.95, 23, 3)`;

b: `norm.ppf(0.95, 23, 9)`;

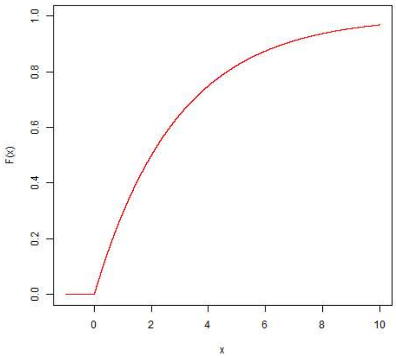
c: `norm.ppf(0.05, 23, 9)`;

d: `norm.ppf(0.05, 23, 3)`;

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

a: 0.25; **b:** 1;

c: 2; **d:** 3



16) What is the probability that a standard normal variable is between -2.57 and 2.57

a: 0.99; **b:** 0.75;

c: 0.5; **b:** 0.01

17) The probability for the number of tails when tossing 100 times a coin can be computed with the normal distribution using the central limit theorem. If the number of tails is a binomial variable with mean $100 * 1/2$ and variance $100 * 1/2 * 1/2$. What is the probability that we obtain between 45 and 55 tails

a:

`norm.cdf(50, 55, 5)-norm.cdf(50, 45, 5)`;

b:

`norm.cdf(50, 45, 5)-norm.cdf(50, 55, 5)`;

c:

`norm.cdf(45, 50, 5)-norm.cdf(55, 50, 5)`;

d:

`norm.cdf(55, 50, 5)-norm.cdf(45, 50, 5)`

18) The 95% margin of error for the mean is

a: The upper 95% quantile of $\bar{X}$

b: The distance at which μ falls from $\bar{X}$ with a probability of 95%;

c: The distance at which $\bar{X}$ falls from μ with a probability of 95%;

d: The upper 95% quantile of μ

19) The 95% confidence interval for the mean weight of a control mouse is $(22.25, 24.35)$. We can therefore claim that

a: we can be 95% confident that mean weight is not 22;

b: the mean weight is between $(22.25, 24.35)$ with probability of 95%;

c: we can be 95% confident that the mean weight is between $(22.25, 24.35)$;

d: we are 95% confident that the mean is $23.3gr$;

20) How would you calculate the confidence interval above

a: with a paired t-test; **b:** with one sample t-test;

c: with a proportion test; **d:** with a paired proportion test;

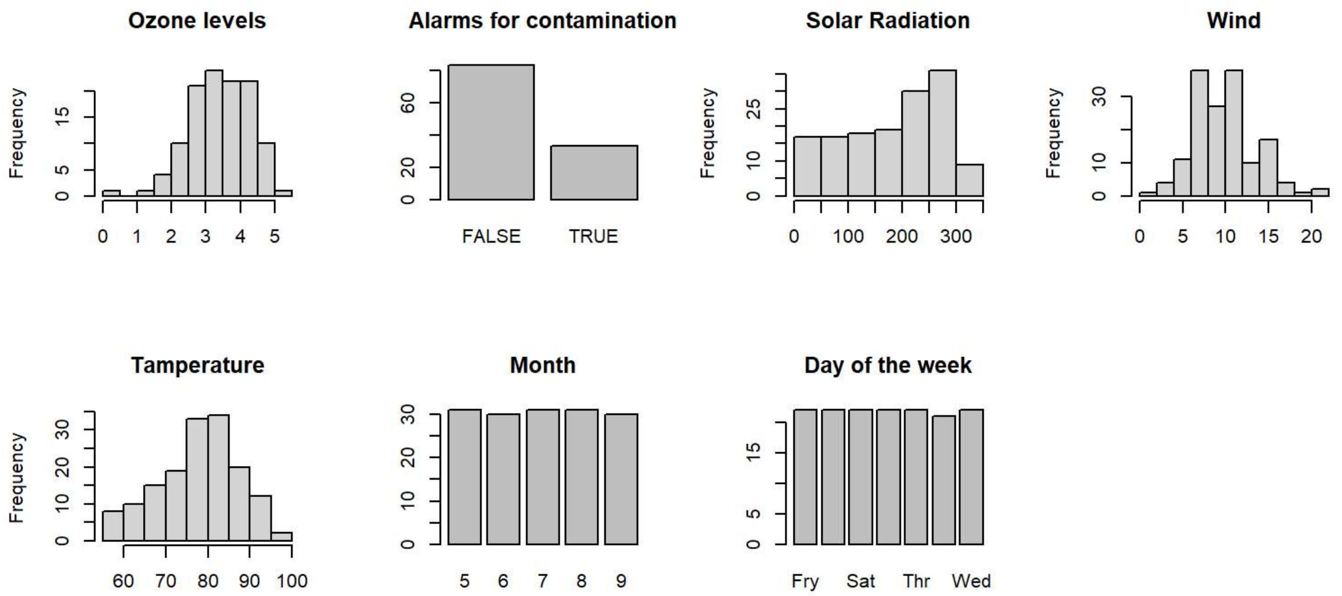
For questions **20-25)** consider the following:

You work at the meteorological office and have collected data relating to ozone levels in the atmosphere for each day from May to September (153 days). You want to determine which variables can predict ozone pollution levels to inform health agencies.

Here are the first six days of your 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 in histograms and barplots



21) If you want to know whether the mean ozone levels are different between spring (months: 5,6) and summer (months:7,8,9), what is the best hypothesis contrast?

a:

$$H_0: \mu_{5\cap 6} = \mu_{7\cap 8\cap 9}$$
$$H_1: \mu_{5\cap 6} \neq \mu_{7\cap 8\cap 9}$$

b:

$$H_0: \mu_{5\cap 6} < \mu_{7\cap 8\cap 9}$$
$$H_1: \mu_{5\cap 6} \geq \mu_{7\cap 8\cap 9} \neq 0$$

c:

$$H_0: \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9$$
H_1 : At least one $\mu_i \neq 0$

d:

$$H_0: p_{5\cap 6} = p_{7\cap 8\cap 9}$$
$$H_1: p_{5\cap 6} \neq p_{7\cap 8\cap 9}$$

22) You hire a statistician who runs a group t-test of Ozone pollution between spring and summer and shows you the following results (x:spring, y:summer)

```
##
##  Welch Two Sample t-test
##
## data:  spring and summer
## t = -4.204, df = 59.111, p-value = 9.001e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0510555 -0.3731867
## sample estimates:
## mean of x mean of y
##  2.921258  3.633379
```

We say that a result is significant when we reject the null hypothesis. In the previous test you, therefore, conclude that

a: The mean ozone levels in spring are significantly higher than the mean ozone levels in summer

b: The mean ozone levels in summer are significantly higher than the mean ozone levels in spring

c: The mean ozone levels in summer are not significantly higher than the mean ozone levels in spring

d: The mean ozone levels in spring are not significantly higher than the mean ozone levels in summer

23) You tell your statistician that you are interested on the relationship between ozone levels and Temperature, therefore she runs a regression model of Ozone on Temperature, adjusted by Solar Radiation, Wind and Month. These are the results

```
##
## Call:
## lm(formula = Ozone ~ Temp + Solar.R + Wind + Month, data = airquality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11552 -0.28963 -0.01603  0.29515  1.18693
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1841693  0.5589415  -0.329 0.742431
## Temp         0.0518836  0.0066581   7.793 4.74e-12 ***
## Solar.R      0.0023884  0.0005708   4.184 5.92e-05 ***
## Wind        -0.0613507  0.0157136  -3.904 0.000166 ***
## Month       -0.0370909  0.0368859  -1.006 0.316919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5085 on 106 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.6676, Adjusted R-squared:  0.6551
## F-statistic: 53.22 on 4 and 106 DF,  p-value: < 2.2e-16
```

Why do you think that the statistician ran this test?

a: To show you the interaction between Temperature with other variables

b: To show you that Ozone depends on other variables in addition to Temperature

c: To show you the statistical dependence between Ozone and Temperature adjusting by the effects of other variables

d: To show you that Wind is more important predictor of Ozone than Temperature.

24) We say that a result is significant when we reject the null hypothesis. Therefore, in the previous analysis you can conclude that

a: Ozone decreases with Temperature but not significantly

b: Ozone increases with Temperature but not significantly

c: Ozone significantly decreases with Temperature

d: Ozone significantly increases with Temperature

25) You are finally interested on testing whether the probability of sending an alarm of Ozone pollution depends on the day of the week. Therefore, the best way to test the statistical dependence between Alarms for for contamination and Day of the week is

a: a two sample t-test

b: an ANOVA

d: a Chi-squared test

d: a 2-way ANOVA