GROUND RULES:

- Fill out the answer sheet below
- Upload a picture of the answer sheet only on LMS Week 15 final exam board
- This is a closed-book and closed-notes exam. You can not use external notes of any kind. You may use a calculator.
- This exam contains two parts:
  - Part 1. Multiple Choice. 25 questions, 1 point each (25 points total)
  - Part 2. Short Answer. 5 questions, 2 point each (10 points total)

  This exam is worth 35 points.
- You have 60 minutes to complete this exam.

ANSWER SHEET

Question	Answer	Short Answer
1		a
2		
3		
4		
5		
6		b
7		
8		
9		
10		
11		c
12		
13		
14		
15		
16		d
17		
18		
19		
20		
21		e
22		
23		
24		
25		

Use this table for questions **1-3**. This table represents the first 8 observations from a sample of 200 individuals, who reported their age, race, income, and job satisfaction score on a scale from 0 to 100.

`Age`	`Race`	`Income`	`Score`
21	W	less than \$10,000	29
33	B	\$20,000-23,000	32
41	B	more than \$100,000	84
26	A	\$30,000-40,000	78
22	O	\$10,000-20,000	87
19	A	\$40,000-50,000	42
34	W	\$50,000-60,000	21
26	W	less than \$10,000	91
⋮	⋮	⋮	⋮

1. Which of the following best describes the `Income` variable?

   a. quantitative, continuous

   b. quantitative, discrete

   c. categorical, dichotomous

   d. categorical, ordinal

   e. categorical

2. Which type of plot would be most useful for visualizing the relationship between `Age` and job satisfaction `Score`?

   a. side by side box plot

   b. scatter plot

   c. dot plot

   d. histogram

   e. single box plot

3. Below are some summary statistics from the `score` variable. Which of the following is <u>true</u>?

```
min Q1 median   Q3 max   mean       sd   n missing
 30 57   69.5   77  99 65.075 16.09361 200       0
```

   a. the minimum value of 30 would be identified as out outlier in a box plot

   b. there were more survey respondents who reported job satisfaction scores less than 57 than survey respondents reported job satisfaction scores greater than 77

   c. the standard deviation estimate is not possible because `score` is a whole number

   d. there is evidence that the distribution of `score` is right-skewed

   e. none of the above are true

4. The World Health Organization uses a normal distribution with mean = 125 and standard deviation = 15 to model the systolic blood pressure (SBP) of American males. John is an American male. His SBP is 110. What is his z-score?

   a. 2 b. 1 c. -1 d. -2

5. Which of the following statements **are** correct?

   a. A normal distribution is any distribution that typically occurs.

   b. In a normal distribution, the mean and median are equal.

   c. The graph of a normal distribution is bell-shaped.

   d. The graph of a normal distribution has one mode.

   e. In a normal distribution, the standard deviation is 1.

6. Which statement about the correlation r is false?

   a. The correlation r is a number between 0 and 1.

   b. The correlation r is the same between x and y as it is between y and x; i.e., reversing the roles of x and y in the calculation of r does not change its value.

   c. The correlation r is unitless.

   d. The correlation r describes linear relationships only.

7. Which of the following is true about the standard deviation of sample means ?

   a. It decreases as sample size n increases

   b. It increases as sample size n increases

   c. It stays the same as sample size n increases

   d. None of these

8. A school system gives every teacher a raise of $1,000. What will this do to the mean and standard deviation of the distribution of teacher salaries?

   a. The mean will increase by $1,000, and the standard deviation will increase by an amount more than $1,000.

   b. The mean will increase by $1,000, but the standard deviation will remain the same.

   c. The mean and standard deviation will both increase by $1,000.

   d. The mean will increase by $1,000, and the standard deviation will increase by an amount less than $1,000.

9. Based on a sample of students at our school, a 95% confidence interval for the true average number of instances that students commit academic misconduct is (1, 3). The margin of error of this interval is:

   a. 0.5 b. 1

   c. 2 d. cannot be determined without knowing the sample size

10. Which of the following is <u>true</u> regarding the Central Limit Theorem (CLT)?

    a. If your sample size is $n = 30$ exactly, then you are guaranteed to have an approximately normal sampling distribution of the sample mean.

    b. As the sample size n increases, the data distribution should become approximately normal.

    c. The Central Limit Theorem states that the sampling distribution of the sample mean should always have the same shape as the population distribution.

    d. none of the above

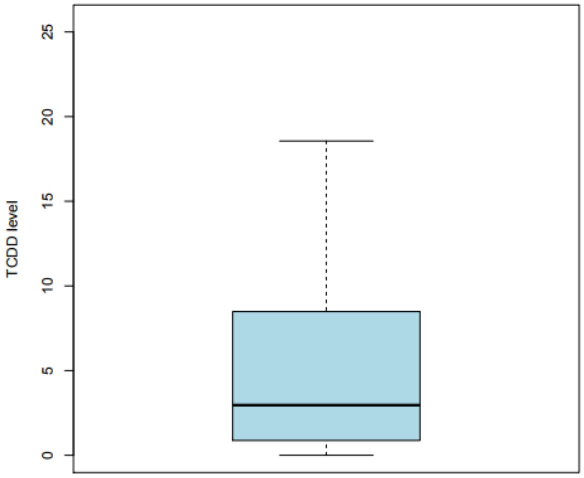
11. I recorded the time (in minutes) it took me to drive to work on 10 consecutive days:

    25 25 27 27 28 30 31 32 32 78

    The median is 29 and the mean is 33.5. If I removed the outlier "78" and recalculated the median and mean, what would happen?

    a. The median would change by more than the mean would.
    b. The median and mean would change by the same amount.
    c. The mean would change by more than the median would.
    d. Neither the median nor mean would change.

Use the information provided below to answer questions **12-14**. A recent study published in a leading journal in environmental science reported the levels of TCDD (a kind of dioxin) of a sample of 55 veterans who may have been exposed during Vietnam War. The boxplot of TCDD levels is given below.



12. The interquartile range (IQR) is closet to

    a. 15 b. 8 c. 3 d. 0

13. The median TCDD levels is closest to

    a. 15 b. 8 c. 3 d. 0

14. The shape of distribution of TCDD levels is best described as

    a. skewed left b. skewed right c. symmetric

15. Which of the following statements about z-scores is/are true?

    a. if a z-score is 2 that means that the observation is two times the value of the mean
    b. if a z-score is negative that means that the observation is less than mean
    c. larger z-scores are always better
    d. the z-score for an observation that is equal to the mean is 1
    e. none of the above are true

16. When constructing a confidence interval for a single proportion, which assumption should we check?

    a. $n > 30$ b. $n > 10$
    c. $n\hat{p} > 10$ d. $np_0 > 10$

17. When a statistic, like the median, is said to be resistant to outliers, this means that

    a. the statistic is not greatly influenced by the value of the outliers.
    b. the statistic itself is an outlier
    c. the statistic itself cannot be an outlier
    d. it is impossible for the data to have any outliers.
    e. the statistic is greatly influenced by the value of the outliers.

18. Based on a random sample of 120 rhesus monkeys, a 95% confidence interval for the proportion of rhesus monkeys that live in a captive breeding facility and were assigned to research studies is (0.67, 0.83). Which of the following is true?

    a. if we used a different confidence level, the interval would not be symmetric about the sample proportion
    b. 95 of the sampled monkeys were assigned to research studies
    c. the margin of error for the confidence interval is 0.16
    d. a larger sample size would yield a wider confidence interval
    e. none of the above are true

19. The World Bank reports that 1.7% of the US population lives on less than $2 per day. A policy maker claims that this number is misleading because of variation from state to state and rural to urban. To investigate this, she takes a random sample of 100 households in Atlanta to compare with the national average and finds that 2.1% of the Atlanta population live on less than $2/day. Select the null and alternative hypothesis to test whether Atlanta differs significantly from the national percentage.

    a. H_0: $\mu = 0.021$, H_a: $\mu \neq 0.021$
    b. H_0: $p = 1.7$, H_a: $p \neq 1.7$
    c. H_0: $p = 2.1$, H_a: $p \neq 2.1$
    d. H_0: $p = 0.017$, H_a: $p \neq 0.017$
    e. H_0: $\mu = 2$, H_a: $\mu \neq 2$

20. Complete the following sentence: When conducting a hypothesis test, we _____________ and then evaluate the test results to determine if there is enough evidence to _____________.

    a. Assume that the alternative hypothesis is true; reject the null hypothesis
    b. Assume the alternative hypothesis is false; reject the alternative hypothesis
    c. Assume that the null hypothesis is false; accept the null hypothesis
    d. Assume that the null hypothesis is true; reject the null hypothesis

21. The alumni association for our college has gathered a large dataset on graduating seniors. Some of the variables in the data set are gender, major, GPA, and starting salary. They are interested in looking at relationships among these variables. For which of the following pairs of variables would a two sample t-test be appropriate to examine whether there is a relationship between the two variables?

    a. GPA and salary
    b. gender and major
    c. GPA and gender
    d. major and salary
    e. none of the above

22. Ebay sellers wonder if the type of photo posted with an item affects the selling price of that item. One hundred and forty three MarioKart packages were analyzed, which were classified as having a "stock" photo or not. A 95% confidence interval for the average difference in selling price between those without and with "stock" photos ($\mu_{no} - \mu_{yes}$) is (-\$7.20, -\$1.14). Which of the following are correct interpretations of this interval?

    a. In general, the average selling price of the MarioKart packages is less than \$10.
    b. There is no evidence that photo type is associated selling price.
    c. We have evidence that packages with stock photos sell, on average, more than packages without stock photos.
    d. We have evidence that packages with stock photos sell, on average, less than packages without stock photos.
    e. More than one statement is correct.

23. A researcher conducts an experiment on human memory by randomizing 10 individuals to a treatment group and 10 individuals to a placebo group. After the intervention, she performs a two sample t-test on the memory score of the two groups. She obtains a p-value of .2. Which of the following is a reasonable interpretation of her results?

    a. There could be a treatment effect, but the sample size was too small to detect it.
    b. She should reject the null hypothesis.
    c. This proves that her experimental treatment has no effect on memory.
    d. There is evidence of a small effect on memory by her experimental treatment.
    e. There is a 20% chance that the two groups had the same memory scores.

24. An opinion poll asks a simple random sample of 1100 people whether they support reducing the number of legal immigrants the United States; 53% of the 1100 people sampled say "Yes." The number 53% is an example of

    a. a statistic
    b. a parameter
    c. the level of confidence
    d. the margin of error

25. Which of the following variables is categorical?

    (a) marital status of JBNU professors
    (b) starting salaries for statistics majors
    (c) the CD4 cell count (number of cells) for a group of HIV patients
    (d) the amount of fertilizer (in pounds) applied in an agricultural experiment

SHORT ANSWER The following scatter plot presents data collected in the 1960s for 21 countries on

- `Cigarette`: Annual Per Capita Cigarette Consumption
- `Coronary`: Deaths from Coronary Heart Disease per 100,000 persons of age 35-64

The least square fit output by `lm` function of R is also given along with residual plot

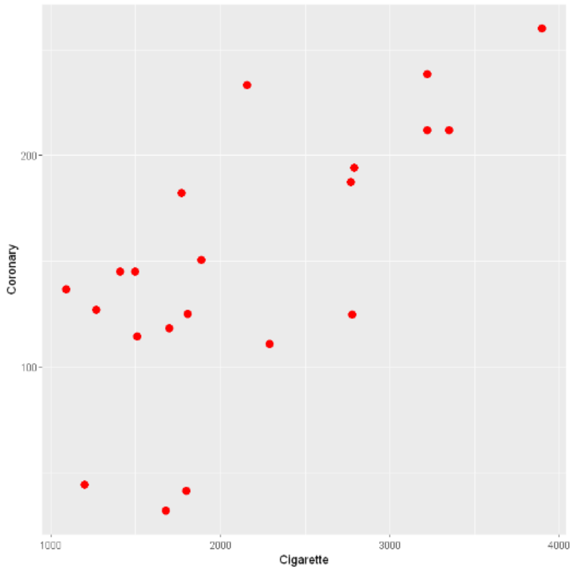


Figure 1: Scatter Plot

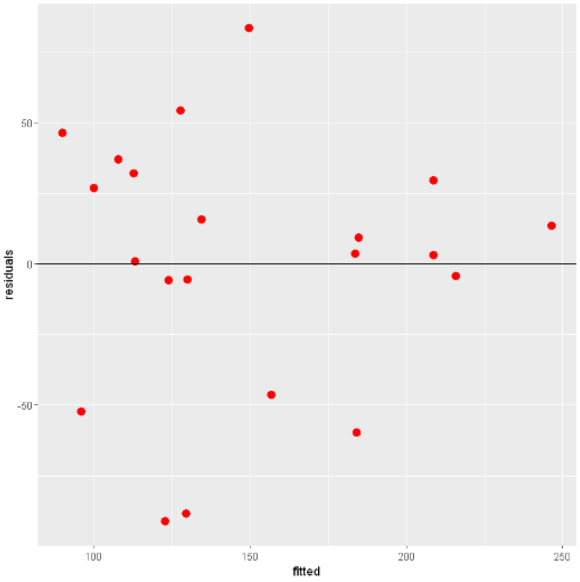
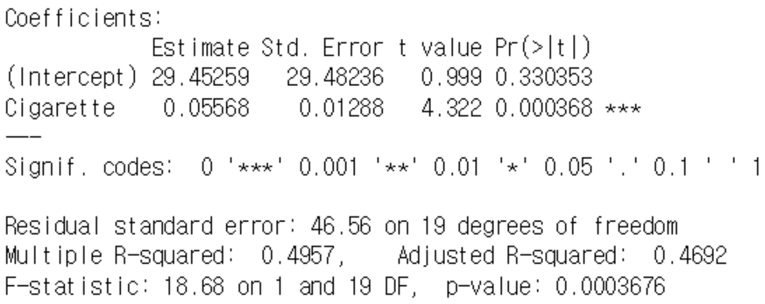


Figure 2: Residual vs Fitted value



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.45259   29.48236   0.999 0.330353
Cigarette    0.05568    0.01288   4.322 0.000368 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.56 on 19 degrees of freedom
Multiple R-squared:  0.4957,    Adjusted R-squared:  0.4692
F-statistic: 18.68 on 1 and 19 DF,  p-value: 0.0003676
```

a. Based on the scatterplot of Coronary versus Cigarette, does there appear to be a linear relationship between cigarette consumption and heart disease? If so, does the relationship appear to be negative or positive?

b. What patterns or problems, if any, do you see in the residuals versus fits plot? Would you feel reasonably comfortable in fitting a simple linear regression model to this data set?

c. Write the equation for the fitted model.

d. Give an interpretation of the fitted slope, β_1

e. From the output, we see the value of $R^2 \approx 0.496$ (or 49.6 as a percent). Interpret what this means