PAMANTASAN NG LUNGSOD NG MAYNILA
(University of the City of Manila)

Department of Tourism and Hospitality Management

Name: Maria Elizabeth M. Mercado **November 13, 2024**
Block/Section: 3-2 **Prof. Ivan Cassidy F. Villena**

MIDTERM EXAMINATION

APPLIED STATISTICS FOR BUSINESS AND ECONOMICS

1. Explain the difference between descriptive and inferential statistics. Provide an example of each (10 pts.)

According to the discussion of Professor Ivan Cassidy F. Villena, I learned that the definition of Descriptive statistics is a concerned with the collection, description, analysis of a set of data with drawing conclusions or inferences about a large set while the definition of Inferential Statistics is a centers on making predictions or inferences about a larger set of data using the information gathered from a subset of a larger set. Basically, The Differences between Descriptive Statistics and Inferential Statistics are the **analysis of a set of data with drawing conclusions or inferences about a Large set** which means it talked about the **overall number of collections** of things in our life. while, the Inferential Statistics is **a Predictions or inferences about a larger set of data or a subset of a larger set which** means this is the predictions of things in our life that don't have a right or assurance that the things that will do in our life. The Example of Descriptive Statistics is **the Average of a Student on her/his major subject for the whole first semester.** Next example is Inferential Statistics is **The Student predicting that her/his Grade in Applied Statistics For Business and Economics for the First Semester is Higher than her/his Blockmates because She/He Active on Class and Do her/his Assignments on time, so that She/He Predicting that her/his grade for the First semester is High.**

2. Identify and describe the three main measures of central tendency. When might each measure be most appropriate to use? (10 pts.)

After the discussion, I learned the Definition of **Measures of Central Tendency** and also the Three main measures of Central tendency. The Central Tendency is a single value that is used to identify the "Center" of the data or the typical value and It is precise yet simple. In addition to that, The Central tendency is a most representative value of the data. The Three main measures of Central tendency are **Mean, Median,** and **Mode.** The Mean is the most frequently used measure of central tendency and sum of the observations divided by the total number of observations. Also, The Advantages of Mean is Takes into account all observations and can be used for further statistical calculations and mathematical manipulation and mean always exists and unique. The Notations of Mean is μ it is used to denote population mean and another notation symbol is $\bar{x}$ It is used to denote sample mean. Next Central Tendency is Median, This is the Central Value of a

Distribution and The Value that divides the Distribution into two equal parts. In Addition to that, The Advantages of Median is computed even for grouped data with open ended class intervals and exact middle value of the distribution. Also, This is not affected by extreme values. As I learned in Discussion, The First step in finding the median, denoted by **Md** is to arrange the observations in an array. Lastly, The Central Tendency is Mode. The Mode is that value of a variable that occurs most frequently in a distribution and It is also referred to as the nominal average. As I also learned in the discussion, It is Mode by counting the frequency of each observed value and finding the observed value with the highest frequency of occurrence. The advantages of Mode are easily identified through ocular inspection and Extreme values do not easily affect the mode. Also, Value is always one of the observed values in the data set and it can be obtained both for quantitative and qualitative types of data. In addition to that, I learned that Mode is the highest or most frequent in distribution of things or even in calculating the highest number. In my own perspective, **The most frequently used measure of Central Tendency is Arithmetic Mean** because it can be computed in two ways such as for ungrouped data and for grouped data and also it can be used for Population mean and Sample mean in computing.

3. Given the dataset: [10, 10, 10, 11, 11, 11, 11, 13, 13, 14, 14, 16, 17, 18, 19, 19, 20, 25, 25,30], calculate the following: (10 pts.)

Mean

Formula: $\bar{x} = \sum$ x1, x2, x3, x4……/n

[**10, 10, 10, 11, 11, 11, 11, 13, 13, 14, 14, 16, 17, 18, 19, 19, 20, 25, 25,30**]

$\bar{x} = \frac{10 + 10 + 10 + 11 + 11 + 11 + 11 + 13 + 13 + 14 + 14 + 16 + 17 + 18 + 19 + 19 + 20 + 25 + 25 +}{20}$

$\bar{x}$= **317** divided by **20**

$\bar{x}$=**15.85**

Median

[10, 10, 10, 11, 11, 11, 11, 13, 13, **14, 14,** 16, 17, 18, 19, 19, 20, 25, 25,30]

$\bar{x}$= 14 (middle number)

Mode

[10, 10, 10, **11, 11, 11, 11,** 13, 13, 14, 14, 16, 17, 18, 19, 19, 20, 25, 25,30]

m= 11 (most numbers)

Variance

Formula: $\sigma 2 = \frac{\sum (xi - \bar{x})2}{N}$

μ=15.85 (population mean)

10, 10, 10, 11, 11, 11, 11, 13, 13, 14, 14, 16, 17, 18, 19, 19, 20, 25, 25,30

$$\sigma 2 = \frac{\sum(10-15.85) + (10-15.85)^2 + (10-15.85)^2 + (11-15.85)^2 + (11-15.85)^2 + (11-15.85)^2 + (11-15.85)^2 + (13-15.85)^2 + (13-15.85)^2 + (14-15.85)^2 + (14-15.85)^2 + (16-15.85)^2 + (17-15.85)^2 + (18-15.85)^2 + (19-15.85)^2 + (19-15.85)^2 + (20-15.85)^2 + (25-15.85)^2 + (25-15.85)^2 + (30-15.85)^2}{20}$$

$$\sigma 2 = \frac{\sum (-5.85)^2 + (-5.85)^2 + (-5.85)^2 + (-4.85)^2 + (-4.85)^2 + (-4.85)^2 + (-4.85)^2 + (-2.85)^2 + (-2.85)^2 + (-1.85)^2 + (-1.85)^2 + (0.15)^2 + (1.15)^2 + (2.15) + (3.15)^2 + (3.15)^2 + (4.15)^2 + (9.15)^2 + (9.15)^2 + (14.15)^2}{20\text{VariabiVa}}$$

$$\sigma 2 = \frac{34.22 + 34.22 + 34.22 + 23.52 + 23.52 + 23.52 + 23.52 + 8.12 + 8.12 + 3.42 + 3.42 + 0.02 + 1.13 + 4.62 + 9.92 + 9.92 + 17.22 + 83.72 + 83.72 + 200.22}{20}$$

$\sigma 2$= **630.48** divided by **20**

$\sigma 2$= **31.52**

Standard Deviation

Formula: $\sqrt{\sigma 2}$

$\sigma = \sum (X - \mu) 2 \, n = \sqrt{}$ **31.52**

σ = **5.61** Standard Deviation

4. Why is it important to understand the variability in a dataset? Briefly explain how the range and standard deviation provide different perspectives on data spread (10 pts.)

In my own perspective, It's important to understand the variability in a dataset because it characterizes between data points and the center of a distribution. In addition to that, Dataset is very helpful since it can compute measures of variability and central tendency in descriptive statistics that may be used to summarize the data. Furthermore, When it comes to Range it can give information about data. The range provides the most convenient method for calculating the dataset's dispersion. In the case of five newborns, for instance, the heaviest or maximum weight is twenty pounds, and the smallest or minimum weight is ten pounds. Therefore, the weight range of babies that are the heaviest to the least is 20-10, or 10 pounds. Based on this, we can infer that babies weigh between 10 and 20 pounds. Furthermore, by taking the square root of the average squared variances, the standard deviation gives the data dispersion. It is often referred to as the root mean square of the deviations from the mean, and it is very helpful in determining how representative the mean is. It is also commonly used to assess dispersion and is the most straightforward to handle algebraically and spread in the most basic way possible, including the ranges in which each item occurred.The ranges and standard deviations that give us different perspectives on the data spread are also important to understand because we can use them to assess how data points diverge from the mean and to understand variability. The next step is to compare data because standard deviation can be beneficial when comparing different datasets. Finding outliers is the next step, which aids in spotting extreme data. The next step is Quality Management and Control, which is used to evaluate the reliability of data.

5. Suppose you are given the test scores of two classes, and each class has the same mean score but different standard deviations. What does this tell you about the score distribution in each class? (10 pts.)

In my own perspective, If I will give the test scores to the two classes and each class has the same mean score but different standard deviations, I will tell about the score distribution in each class by measuring how spread out and dispersed the scores around the mean. Additionally, a lower standard deviation indicates the scores are significantly clustered around the mean, although a higher standard deviation indicates that the outcomes are more widely distributed. The two classes' scores will be assigned according to how well they perform and present it to the class. Two classes will serve as our basis for progress. The first is the class with a high standard deviation, whose scores deviate considerably from the mean. There is greater variation in student performance, with scores that are both significantly higher and lower than the mean. The second class has a lower standard deviation, meaning that they perform similarly with a smaller variance and their scores are closer to the mean. Furthermore, even if the class has the same mean, the class with the lowest standard deviation will receive low points for their class performances, while the class with the highest standard deviation will have a greater variability in performance.

Practical Exam: (50 pts)

In part, please refer to the examdata.csv and Jupyter Noterbook link (Google Colab). You will be tasked to clean the data and compute for the summary measures. Please save your output showing the codes in PDF format. Merge the separate PDF files before submitting.

(scroll down so that you will see my answer po, thank you)

```
#Data Structures
#Python packages for data processing
import numpy as np #numerical python
import pandas as pd #pandas


 df = pd.read_csv ('examdata (1).csv')

#check the data set
df.head (30) #print the first 5 rows
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 30,\n  \"fields\": [\n
{\n      \"column\": \"overtime_hours\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 3.394112549695428,\n
\"min\": 1.0,\n        \"max\": 15.0,\n        \"num_unique_values\":
13,\n        \"samples\": [\n          11.0,\n          15.0,\n
5.0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"monthly_wage\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 607.4279102526045,\n        \"min\":
1101.0,\n        \"max\": 2935.0,\n        \"num_unique_values\": 26,\
n        \"samples\": [\n          1400.0,\n          1455.0,\n
2345.0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe"}
```


```
#check the name of columns
df.columns
```

```
Index(['overtime_hours', 'monthly_wage'], dtype='object')
```

```
df.dtypes
```

```
overtime_hours    float64
monthly_wage      float64
dtype: object
```

```
df.shape #(rows, columns)
```

```
(30, 2)
```

```
#Count the missing values
df.isnull ().sum()
```

```
overtime_hours    4
monthly_wage      4
dtype: int64
```

```
#replace the missing values with mean
#replace the missing values with mean

df['monthly_wage'].fillna(df['monthly_wage'].mean(), inplace=True)
df['overtime_hours'].fillna(df['overtime_hours'].mean(), inplace=True)
```

```
<ipython-input-44-2e24cb6715f4>:4: FutureWarning: A value is trying to
be set on a copy of a DataFrame or Series through chained assignment
using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never
work because the intermediate object on which we are setting values
always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try
using 'df.method({col: value}, inplace=True)' or df[col] =
df[col].method(value) instead, to perform the operation inplace on the
original object.


  df['monthly_wage'].fillna(df['monthly_wage'].mean(), inplace=True)
<ipython-input-44-2e24cb6715f4>:5: FutureWarning: A value is trying to
be set on a copy of a DataFrame or Series through chained assignment
using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never
work because the intermediate object on which we are setting values
always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try
using 'df.method({col: value}, inplace=True)' or df[col] =
df[col].method(value) instead, to perform the operation inplace on the
original object.


  df['overtime_hours'].fillna(df['overtime_hours'].mean(),
inplace=True)
```

```
#Check for the missing values of monthly wage

df.isnull().sum()
```

```
overtime_hours    0
monthly_wage      0
dtype: int64
```

```
#Create another column patient classification
classification = {1: 'overtime_wage', 0: 'monthly_wage'}
df['overtime_wage'] = df['monthly_wage'].map(classification)

df.head()
```

```
{"repr_error":"Out of range float values are not JSON compliant:
nan","type":"dataframe","variable_name":"df"}
```

```
df.dtypes
```

```
overtime_hours    float64
monthly_wage      float64
```

```
overtime_wage      object
dtype: object
```

```
#Compute for Summary Measures

df.describe()
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 8,\n  \"fields\": [\n
{\n      \"column\": \"overtime_hours\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 9.42795078452886,\n
\"min\": 1.0,\n        \"max\": 30.0,\n        \"num_unique_values\":
7,\n        \"samples\": [\n          30.0,\n          6.0,\n
7.0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"monthly_wage\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 954.5984522449825,\n        \"min\":
30.0,\n        \"max\": 2935.0,\n        \"num_unique_values\": 7,\n
\"samples\": [\n          30.0,\n          1937.1153846153845,\n
2363.25\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe"}
```