**STATISTICS WORSHEET – 4 SOLUTIONS:**

**Q1:**

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution

**Q2:**

The **sample** is the specific group of individuals that you will collect data from. The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).

There are two types of sampling methods:

* [**Probability sampling**](https://www.scribbr.com/methodology/sampling-methods/#probability-sampling) involves random selection, allowing you to make strong statistical inferences about the whole group.
* [**Non-probability sampling**](https://www.scribbr.com/methodology/sampling-methods/#non-probability-sampling) involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

**Q3:**

|  |  |  |
| --- | --- | --- |
| **Basis for comparison** | **Type I error** | **Type II error** |
| **Definition** | Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true. | Type II error is the error that occurs when the null hypothesis is accepted when it is not true. |
| **Also termed** | Type I error is equivalent to false positive. | Type II error is equivalent to a false negative. |
| **Meaning** | It is a false rejection of a true hypothesis. | It is the false acceptance of an incorrect hypothesis. |
| **Symbol** | Type I error is denoted by α. | Type II error is denoted by β. |
| **Probability** | The probability of type I error is equal to the level of significance. | The probability of type II error is equal to one minus the power of the test. |
| **Reduced** | It can be reduced by decreasing the level of significance. | It can be reduced by increasing the level of significance. |
| **Cause** | It is caused by luck or chance. | It is caused by a smaller sample size or a less powerful test. |
| **What is it?** | Type I error is similar to a false hit. | Type II error is similar to a miss. |
| **Hypothesis** | Type I error is associated with rejecting the null hypothesis. | Type II error is associated with rejecting the alternative hypothesis. |
| **When does it happen?** | It happens when the acceptance levels are set too lenient. | It happens when the acceptance levels are set too stringent. |

**Q4:**

**Normal distribution**, also known as the Gaussian **distribution**, is a probability **distribution** that is symmetric about the **mean**, showing that data near the **mean** are more frequent in occurrence than data far from the **mean**. In graph form, **normal distribution** will appear as a bell **curve**.

**Q5:**

**Covariance** and **Correlation** are two mathematical concepts which are commonly used in the field of probability and statistics. Both concepts describe the relationship between two variables.

**Covariance –**

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

**Correlation –**

1. It show whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

**Q6:**

* **Univariate** statistics summarize only **one**[**variable**](https://www.scribbr.com/methodology/types-of-variables/) at a time. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
* **Bivariate** statistics compare**two variables**.
* **Multivariate** statistics compare **more than two variables**.

**Q7:**

A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty. This technique is used within specific boundaries that depend on one or more input variables.

The **sensitivity** is calculated by dividing the percentage change in output by the percentage change in input.

**Q8:**

Hypothesis testing is an act in statistics whereby an analyst [tests](https://www.investopedia.com/terms/w/wilcoxon-test.asp) an assumption regarding a population parameter. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

All analysts use a random population sample to test two different hypotheses: the [null hypothesis](https://www.investopedia.com/terms/n/null_hypothesis.asp) and the alternative hypothesis.

The null hypothesis would be represented as Ho: P = 0.5. The alternative hypothesis would be denoted as "H1" and be identical to the null hypothesis.

For a two tailed test, the null hypothesis (H0) should be rejected when the test value is in either of two critical regions on either side of the distribution of the test value and vice versa for alternate hypothesis.

**Q9:**

**Quantitative data** is information about quantities, and therefore numbers, **examples** are length, mass, temperature, and time whereas **qualitative data** is descriptive, and regards phenomenon which can be observed but not measured, such as language.

**Q10:**

The **Range** is the difference between the lowest and highest values. Example: In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9. So the **range** is 9 − 3 = 6.

We can find the interquartile range or IQR in four simple steps:

1. Order the data from least to greatest
2. Find the median
3. Calculate the median of both the lower and upper half of the data
4. The IQR is the difference between the upper and lower medians

**Q11:**

The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean. The mean, in turn, refers to the average of all data points in the data set or sequence and will be found at the highest point on the bell curve.

**Q12:**

**One of the methods to Calculate the Outlier is by using the Interquartile Range**

1. Take your IQR and multiply it by 1.5 and 3. We'll use these values to obtain the inner and outer fences. ...
2. **Calculate** the inner and outer lower fences. Take the Q1 value and subtract the two values from step 1. ...
3. **Calculate** the inner and outer upper fences.

**Q13:**

The **p**-**value**, or probability **value**, **tells you** how likely it is that your data could have occurred under the null hypothesis.The **p**-**value** is a proportion: if your **p**-**value** is 0.05, that means that 5% of the time **you** would see a test statistic at least as extreme as the one **you** found if the null hypothesis was true.

**Q14:**

The binomial distribution formula is:

**b(x; n, P) = nCx \* Px \* (1 – P)n – x**

Where:  
b = binomial probability  
x = total number of “successes” (pass or fail, heads or tails etc.)  
P = probability of a success on an individual trial  
n = number of trials

**Q15:**

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

There are many industries that can use the ANOVA test to identify issues or variances between samples. The ANOVA is a good statistical technique for testing. Businesses that might consider the use of the ANOVA include manufacturing, healthcare, service, food, and more.