WORKSHEET

STATISTICS WORKSHEET- 6

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following can be considered as random variable?

a) The outcome from the roll of a die

b) The outcome of flip of a coin

c) The outcome of exam

d) All of the mentioned

Ans: d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?

a) Discrete

b) Non Discrete

c) Continuous

d) All of the mentioned

Ans: a) Discrete

3. Which of the following function is associated with a continuous random variable?

a) pdf

b) pmv

c) pmf

d) all of the mentioned

ans: **a) pdf**

4. The expected value or _____ of a random variable is the center of its distribution.

a) mode

b) median

c) mean

d) bayesian inference

ans : c) mean

5. Which of the following of a random variable is not a measure of spread?

a) variance

b) standard deviation

c) empirical mean

d) all of the mentioned

ans: a) variance

6. The _____ of the Chi-squared distribution is twice the degrees of freedom.

a) variance

b) standard deviation

c) mode

d) none of the mentioned

ans: a) variance

7. The beta distribution is the default prior for parameters between _____

a) 0 and 10

b) 1 and 2

c) 0 and 1

d) None of the mentioned

ans: c) 0 and 1

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for

difficult statistics?

a) baggyer

b) bootstrap

c) jacknife

d) none of the mentioned

ans: b) bootstrap

WORKSHEET

9. Data that summarize all observations in a category are called _____ data.

a) frequency

b) summarized

c) raw

d) none of the mentioned

ans: b) summarized

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

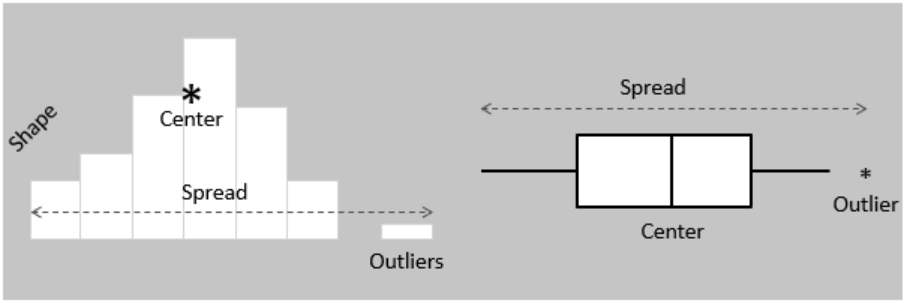
10. What is the difference between a boxplot and histogram?

Ans: Histograms and box plots are very similar in that they both help to visualize and describe numeric data.

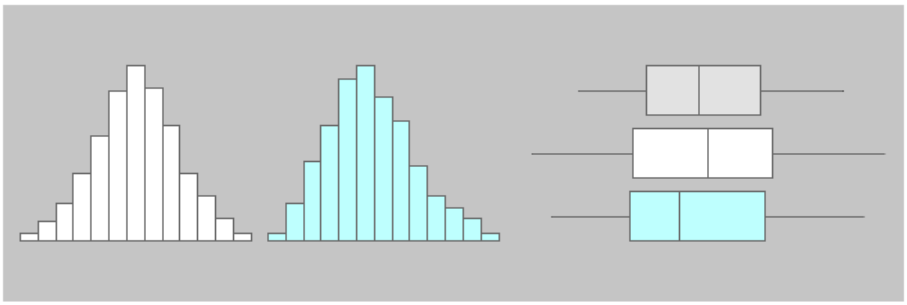
Although histograms are better in determining the underlying distribution of the data, **box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space**.

Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.



11. How to select metrics?

Ans: To select metrics : **prioritize objectives, examine which metric consistently predicts their achievement, and identify which activities influence predictors, in that order**

. And continuously re-evaluate this process to keep up with the times.

1. Define your primary objective

Before you even begin to sift through the various metrics and statistics available to you, it is essential that your company's governing objectives have been clearly established. As a B2B finance company, a primary objective could be to increase market share by 3% before the end of the year.

While an overarching goal such as this may seem somewhat abstract, if marketing metrics aren't considered with this objective in mind, you're bound to pick increasingly irrelevant ones over time.

2. Choose your metric(s) - determine cause and effect

Once a clear, overarching objective has been established, most marketing companies look to major metrics to determine their success—factors such as the generation of sales and leads.

But these metrics aren't the only indicator of a company's success. Less easily quantifiable factors such as customer satisfaction and brand loyalty also play a significant role in the ability to achieve overall marketing objectives, especially in the long term.

Examining the relationship between these metrics can allow marketers and others to develop a cause-and-effect theory to determine what drives the end results. You may be tempted to aim your campaign at a huge but diverse audience, so as to widen the top of your sales funnel. But as you start to develop a clearer view on the effect of you actions on the achievement of objectives, it may turn out that spending your budget on potential customers further down the funnel would yield more sales.

3. Create relevant activities

Digital technology has made it easier than ever to track the engagement of various types of marketing materials, be they a video, article, or even a podcast. Let's look at how a marketing agency could help its clients to improve performance.

Once a marketing agency has determined that engaging content is what drives sales and leads for their clients, the agency must determine which types of content reliably generates that engagement.

By taking advantage of the analytical tools provided by various online platforms, a marketing agency could easily find that for one client, video content is the chief driver of engagement, while for another,

lengthy, informative blogs are the most effective type of content. Within these types of content, further factors could also impact engagement (such as the length of a video or the number of list items contained in a blog post).

The marketing agency can then use this information to generate more effective and engaging work for their clients. Videos may be kept within a certain length, and articles may adhere to a particular format. These items are completely within control of the marketing company, and as a result, can remain consistent to drive engagement and sales.

4. Evaluate periodically

Of course, the ever-changing nature of marketing (and the business world as a whole), ensures that the measures you use to link activities with your primary goals must be constantly re-evaluated. The metrics and statistics that drive value for your clients can change over time, especially as new technologies emerge and target demographics shift.

Regularly evaluating your methods and adapting when necessary may cause you to throw away some of your work. But this is by no means a waste. Adjusting course on a regular basis, whether it is for objectives, metrics or activities, will ensure you remain competitive in the years to come. You'll take a couple of small losses now, so you don't have to unexpectedly take a huge loss in the future. Develop a habit of making small improvements in the present, so you'll become a huge success in the future.

12. How do you assess the statistical significance of an insight?

Ans: Statistical significance is often calculated with **statistical hypothesis testing**, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance.

To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd calculate the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha — in other words, the result is statistically significant.

Hypothesis testing is guided by statistical analysis. Statistical significance is calculated using a p-value, which tells you the probability of your result being observed, given that a certain statement (the null hypothesis) is true.[1] If this p-value is less than the significance level set (usually 0.05), the experimenter can assume that the null hypothesis is false and accept the alternative hypothesis. Using a simple t-test, you can calculate a p-value and determine significance between two different groups of a dataset.

1.Define your hypotheses. The first step in assessing statistical significance is defining the question you want to answer and stating your hypothesis. The hypothesis is a statement about your experimental data and the differences that may be occurring in the population. For any experiment, there is both a null and an alternative hypothesis.[2] Generally, you will be comparing two groups to see if they are the same or different.

- The null hypothesis (H_0) generally states that there is no difference between your two data sets. For example: Students who read the material before class do not get better final grades.

- The alternative hypothesis (H_a) is the opposite of the null hypothesis and is the statement you are trying to support with your experimental data. For example: Students who read the material before class do get better final grades.

2. Set the significance level to determine how unusual your data must be before it can be considered significant. The significance level (also called alpha) is the threshold that you set to determine significance. If your p-value is less than or equal to the set significance level, the data is considered statistically significant.[3]

- As a general rule, the significance level (or alpha) is commonly set to 0.05, meaning that the probability of observing the differences seen in your data by chance is just 5%.
- A higher confidence level (and, thus, a lower p-value) means the results are more significant.
- If you want higher confidence in your data, set the p-value lower to 0.01. Lower p-values are generally used in manufacturing when detecting flaws in products. It is very important to have high confidence that every part will work exactly as it is supposed to.
- For most hypothesis-driven experiments, a significance level of 0.05 is acceptable.

3. Decide to use a one-tailed or two-tailed test. One of the assumptions a t-test makes is that your data is distributed normally. A normal distribution of data forms a bell curve with the majority of the samples falling in the middle. The t-test is a mathematical test to see if your data falls outside of the normal distribution, either above or below, in the "tails" of the curve.[4]

- A one-tailed test is more powerful than a two-tailed test, as it examines the potential of a relationship in a single direction (such as above the control group), while a two-tailed test examines the potential of a relationship in both directions (such as either above or below the control group).[5]
- If you are not sure if your data will be above or below the control group, use a two-tailed test. This allows you to test for significance in either direction.
- If you know which direction you are expecting your data to trend towards, use a one-tailed test. In the given example, you expect the student's grades to improve; therefore, you will use a one-tailed test.

4.Determine sample size with a power analysis. The power of a test is the probability of observing the expected result, given a specific sample size. The common threshold for power (or β) is 80%. A power analysis can be a bit tricky without some preliminary data, as you need some information about your expected means between each group and their standard deviations. Use a power analysis calculator online to determine the optimal sample size for your data.[6]

- Researchers usually do a small pilot study to inform their power analysis and determine the sample size needed for a larger, comprehensive study.

- If you do not have the means to do a complex pilot study, make some estimations about possible means based on reading the literature and studies that other individuals may have performed. This will give you a good place to start for sample size.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

Ans: Exponential distributions do not have a log-normal distribution or a Gaussian distribution.

In fact, any type of data that is categorical will not have these distributions as well

. Example: **Duration of a phone car, time until the next earthquake**, etc.

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.

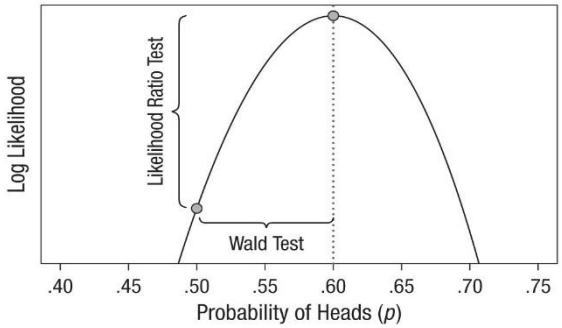
Example : allocation of wealth among individuals.

14. Give an example where the median is a better measure than the mean.

Ans: **Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed**. The median indicates that half of all incomes fall below 27581, and half are above it. For these data, the mean overestimates where most household incomes fall.

15. What is the Likelihood?

Ans: The likelihood is **the probability that a particular outcome is observed when the true value of the parameter is , equivalent to the probability mass on** ; it is not a probability density over the parameter .



The likelihood, , should not be confused with , which is the posterior probability of given the data .

Likelihood is a strange concept in that it is not a probability but is proportional to a probability. The likelihood of a hypothesis (H) given some data (D) is the probability of obtaining D given that H is true multiplied by an arbitrary positive constant K: $L(H) = K \times P(D|H)$.

The likelihood function, parameterized by a (possibly multivariate) parameter , is usually defined differently for discrete and continuous probability distributions .