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 : c) Continuous

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

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?

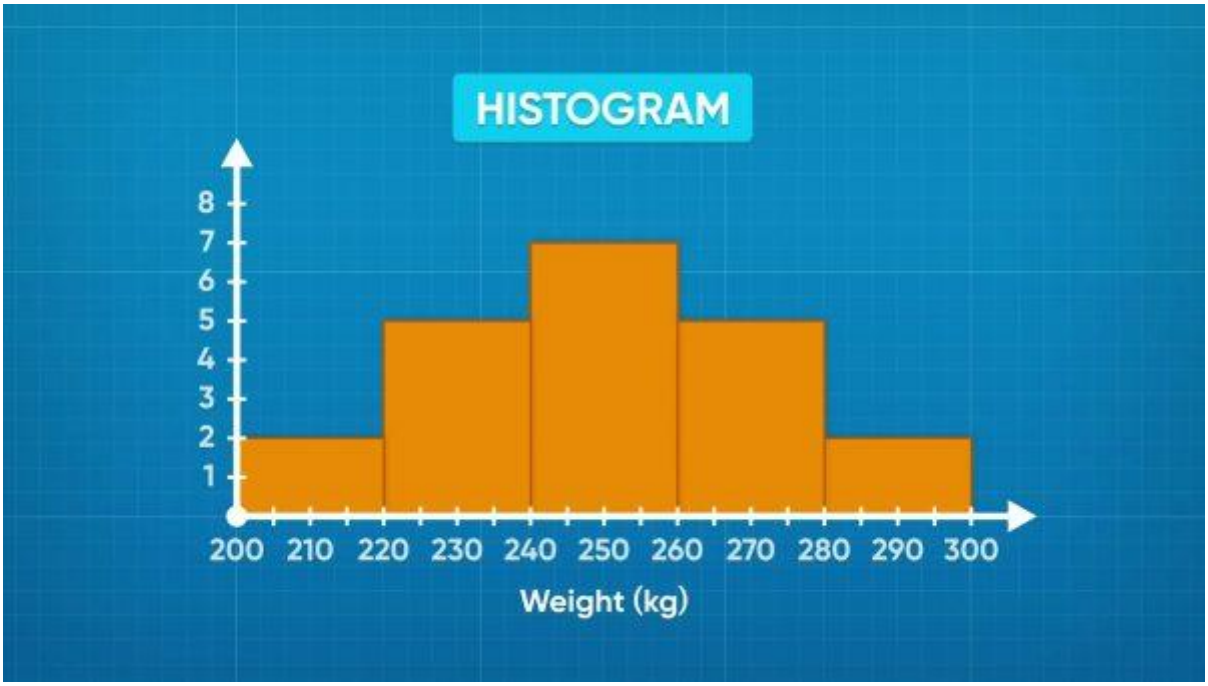
Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data. The 'whiskers' of a box plot show the least and greatest values in the data set.

WHAT ARE HISTOGRAMS AND BOX PLOTS?

Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data. The ‘whiskers’ of a box plot show the least and greatest values in the data set.

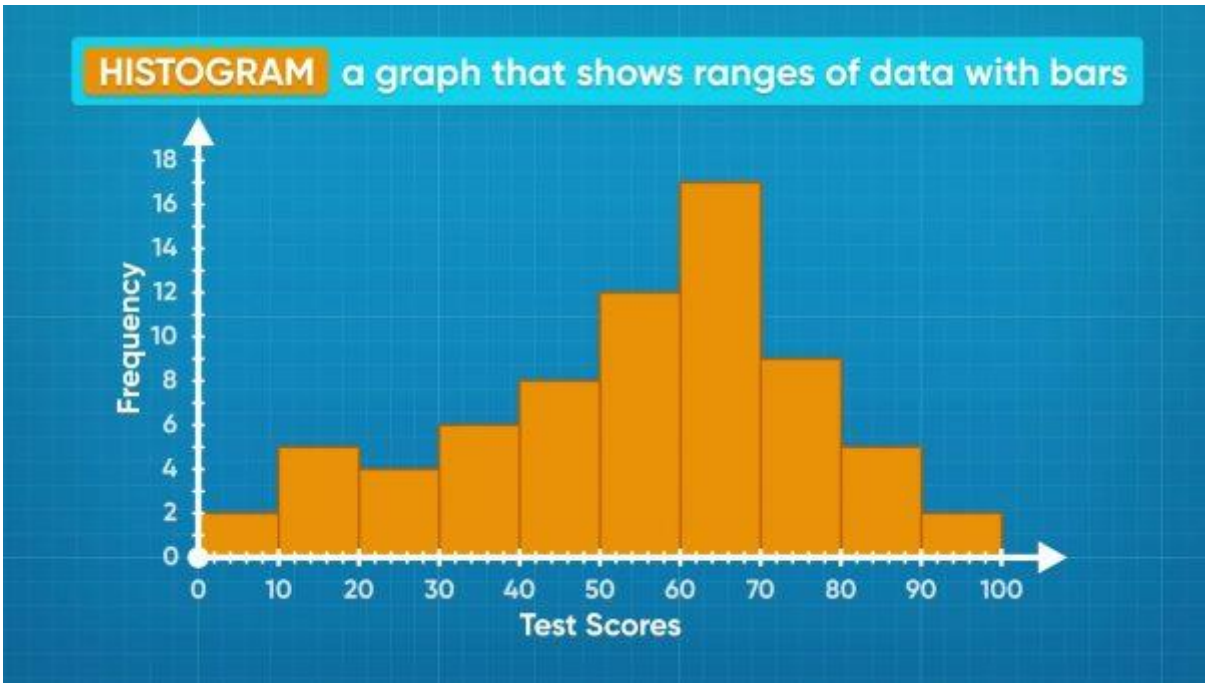
To better understand histograms & box plots…

What are histograms?



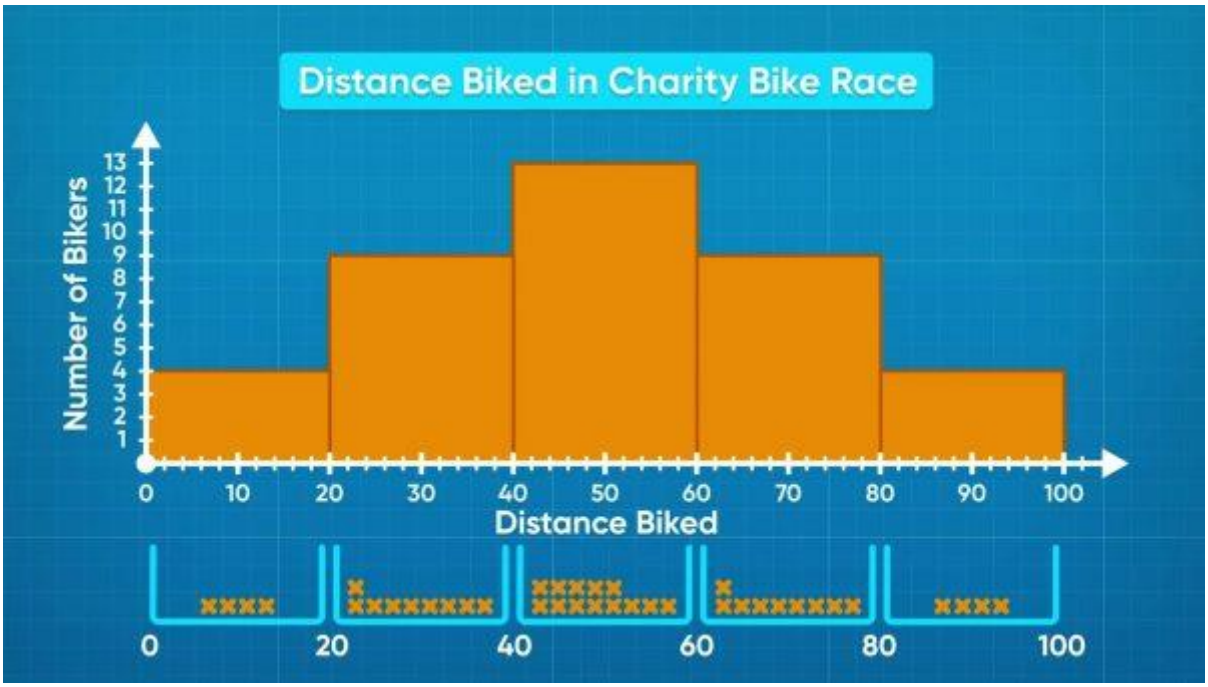
If a data set has a lot of different measurements, displaying it using line graphs does not always help to interpret the information. Instead, you can display the data using a type of bar graph called a histogram. In a histogram, bars represent ranges instead of individual values. These bars are called bins, and they are presented continuously with no spaces between them. Each bar represents a range of data points, and the height of the bar tells us how many data points are in that range. Now you try: **Find a picture of a histogram online or in your textbook. Identify the bins and the height of each bar.**

What can we learn from reading a histogram?



On a test, the students on your class got the following scores: 58, 55, 58, 55, 59, 56, 54, 62, 66, 62, 62, 61, 68, 70, 66, 70, 69, 66, 63, 70, 66, 66, 62, 61, 76, 77, 71, 75, 71, 79, 79, 78, 85, 88, 85, 81, 85, 86, 95, 99, 91, 95. You can separate the data into ranges. Here, it makes sense to choose ranges 50-59, 60-69, 70-79, 80-89, and 90-99. You want to have enough bins to make several distinct ranges, but not so many that they are hard to interpret. If you look at the histogram represented by the data, you can see that the most common grades are in the 60-69% range. The least common grades are in the 90-100% range. Those are called outliers. Now you try: **Find a picture of a histogram online or in your textbook. Which bin contained the least number of values? Where there any outliers?**

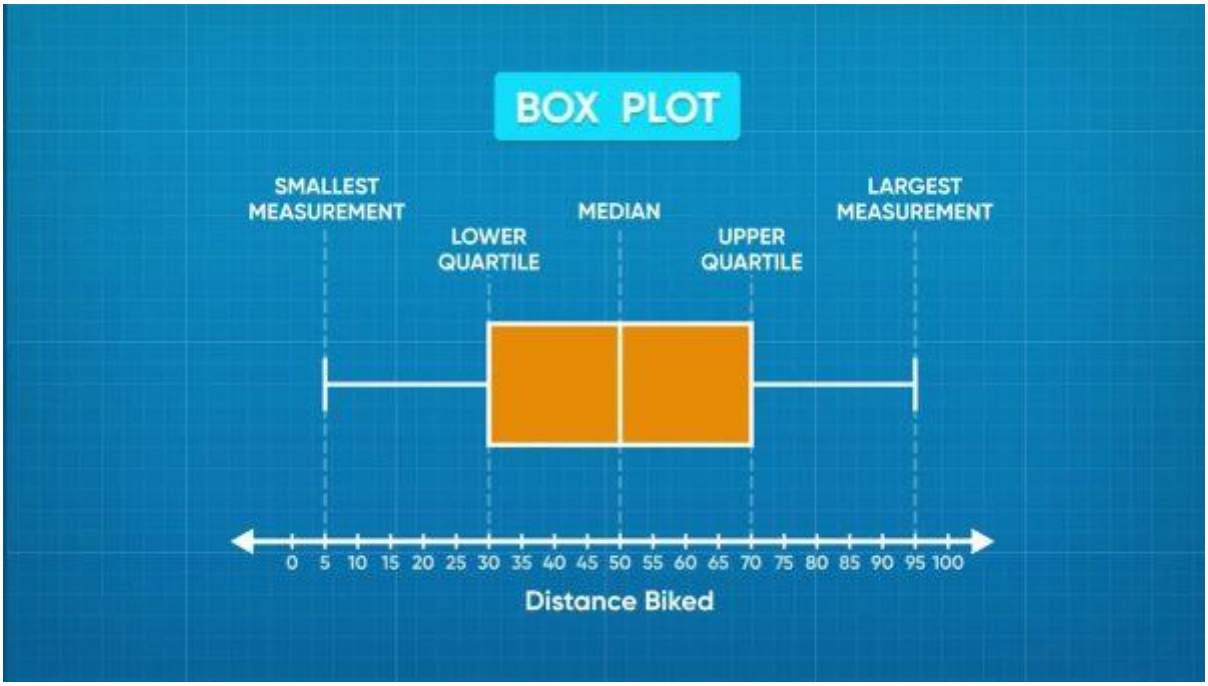
How can we organize our data for a histogram?



When you have a lot of data, you first have to decide how many bins you would like to use, and what the range of each bin should be. In a bike race, the distances in kilometers that cyclists rode are: 5, 8.25, 15.5, 18, 20, 22.5, 28, 28, 29.5, 30, 30, 36.5, 38, 42.5, 45.75, 46, 47, 48, 48, 50, 50, 52, 55, 58, 58, 59, 63.25, 65.5, 67, 70, 70, 72, 75, 75, 76, 83, 87.75, 94.5, 95. You can organize the data into 5 bins, the first one 0-19, then 20-39, 40-59, 60-79, and 80-99. You can then draw the axes for the graph and label the bin sizes at the bottom along the x-axis, and label it “Distance Biked.” The y-axis can be labeled “Number of Cyclists.” Now as you read the data, you can make a tick to count each time a point contributes to a bin. The 0-19 bin has a frequency of 4, 20-39 has 9, 40-59 has 13, 60-79 has 9, and 80-99 has 4. You can see that the most common distance biked is 40-59 kilometers! If you choose bin size 5, you would have a lot more bars and they would be harder to interpret. If you choose bin size 40, you would only have 3 bars, which is not clear either. It is important to choose a bin size that helps you make sense of the data. Now you try: **A data set contains the following data: 12,**

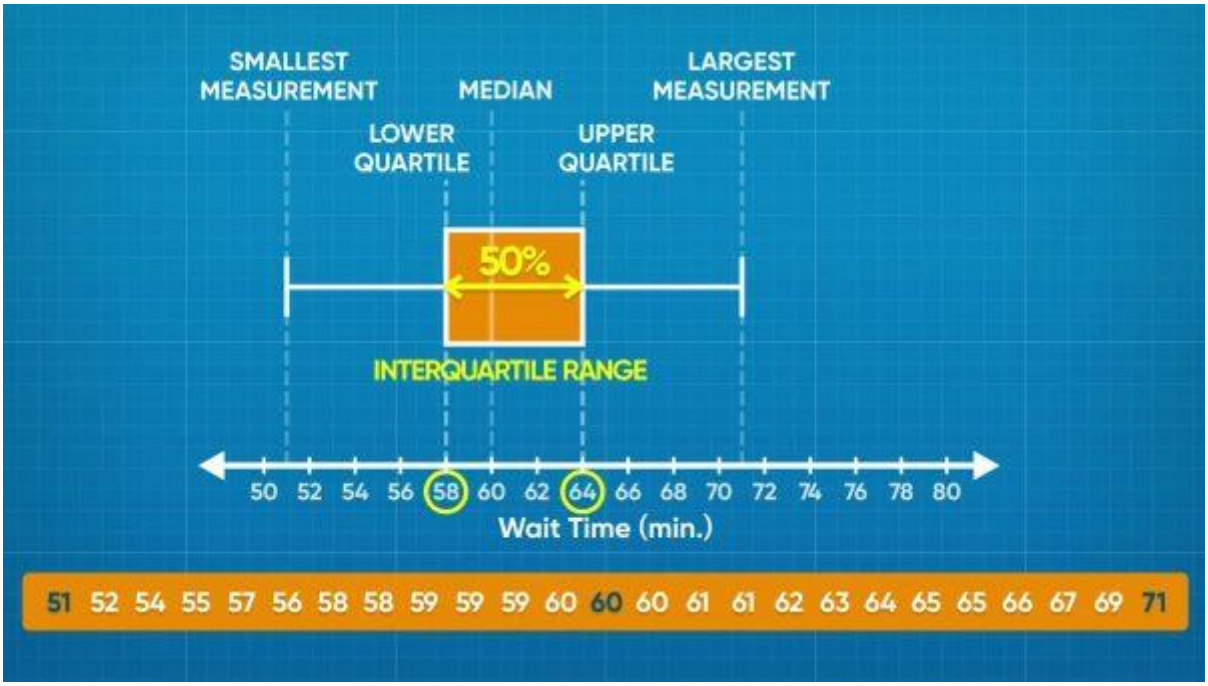
18, 19, 5, 17, 10, 3, 2, 24, 1, 22. What size bins would you choose for your histogram?

What are box plots?



Histograms are great for showing what data ranges are most and least common, but they do not tell details like the range or the median. You can use box plots to present these values. They have 5 vertical lines. The lines farthest on the left and right tell the least and greatest values of the data set. The line in the middle is the median. The other two lines are called the lower quartile and upper quartile. The lower quartile line is on the left of the median, and it tells us that one-quarter of the data points are less than or equal to the lower quartile. The upper quartile is on the right of the median and tells us that one-quarter of the data points are greatest than or equal to the upper quartile. Now you try: Find an image of a box plot online or in your textbook. Identify the 5 key values in the box plot.

How do we make a box plot?



The wait times for a rollercoaster, in minutes, are: 51, 54, 55, 56, 57, 57, 58, 58, 58, 59, 59, 59, 59, 59, 60, 61, 61, 61, 61, 62, 62, 64, 64, 66, 67, 69, 70, 71, 71. To start making the box plot, locate the least and greatest values. Luckily, this data set is already sorted from least to greatest, so you can see that these values are 51 and 71. Place and label these values on the plot. Next, find the median. The number in the middle is 60. That means half of the people waited greater than or equal to 60 minutes, and half the people waited less than or equal to 60 minutes. To find the lower quartile, we find the time that is the middle data point in the range 51 to 60, and to find the upper quartile we find the middle data point in the range 60 to 71. The lower quartile is 58 and the upper quartile is 64. If 25% of people waited 58 minutes or less, and 25% of people waited 64 minutes or more, that means that at least half of the people waited between 58 and 64 minutes. This is called the interquartile range. Now you try: **A set of data has a range of 3 to 85. The median is 72. The lower and upper quartiles are 35 and 78. Draw a box plot using this information.**

11. How to select metrics?

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

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.

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.

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.

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.

Many statistical properties are just valid for Gaussian, other might just need different treatment.
E.g. the standard deviation for Laplace or Bernoulli does directly describe their exponential behaviour, where Gaussians are L2 Distributions, which means their exponential part is dependent from the variance and the squared distance between mean and data.

Many distributions do not even provide a mean value. From my perspective they explicitly describe a specific behaviour of the data values rather than just the deviation from the mean estimate.

Specifically Poissons for example are directly more dense around the satisfaction of their specific expectation value. It does not provide any separate variance from the expectation value. Thus it is tending to have higher masses at lim x->0.

LDA and QDA seem to be like PQ equations of estimating any underlying distribution, similar to Gaussians in other regimes, like Bayesian statistics.

There are many ways to fit data into a specific model by transforming it, the important part is that those transformations are invertible if necessary and at minimal information loss, as well as the model does recover the problem well enough.

Any distribution of money or value will be non--Gaussian. For example: **distributions of income; distributions of house prices; distributions of bets placed on a sporting event**. These distributions cannot have negative values and will usually have extended right hand tails.

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

For skewed distribution, the median is better. For example, **house prices are often skewed to the right, meaning some of them are abnormally high**. The small number of high prices are gonna have a large impact on the mean. Thus the median is better at showing the price of a 'typical' house.

.

15. What is the Likelihood?

Likelihood function is a fundamental concept in statistical inference. It **indicates how likely a particular population is to produce an observed sample**. Let P(X; T) be the distribution of a random vector X, where T is the vector of parameters of the distribution

The likelihood function (often simply called the likelihood) is **the joint probability of the observed data viewed as a function of the parameters of the chosen statistical model**.