Statistics Worksheet4

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

d) All of the mentioned

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

a) Discrete

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

a) pdf

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

c) mean

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

c) empirical mean

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

a) variance

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

c) 0 and 1

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

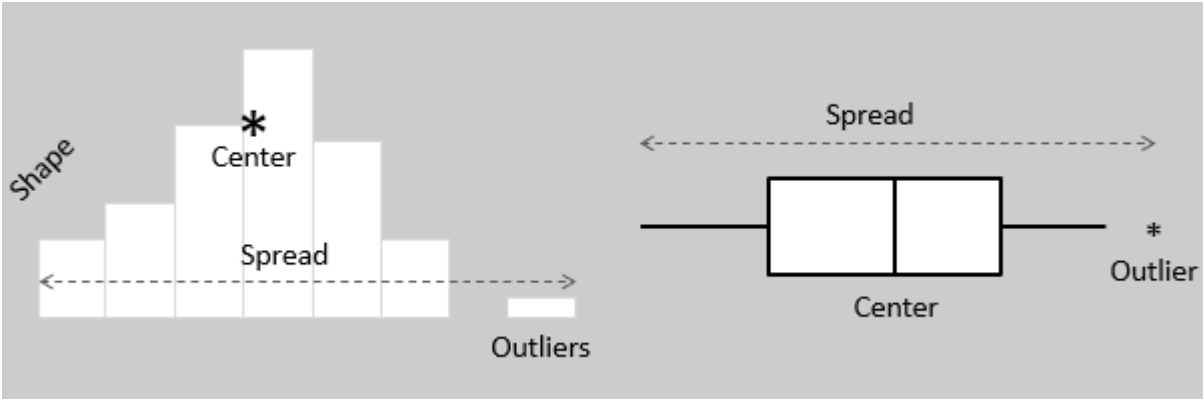
b) bootstrap

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

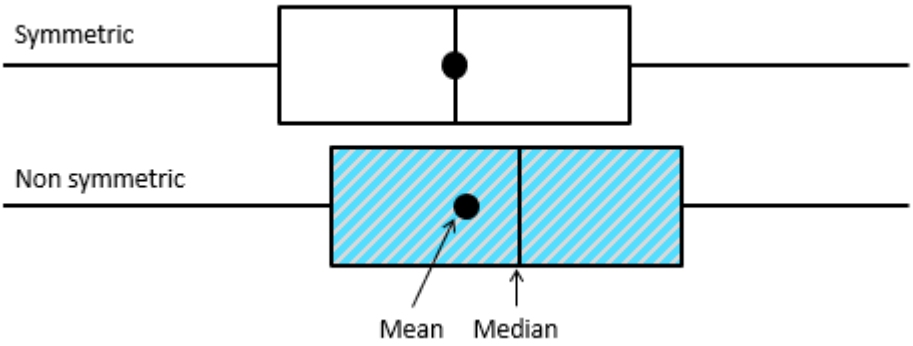
b) summarized

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

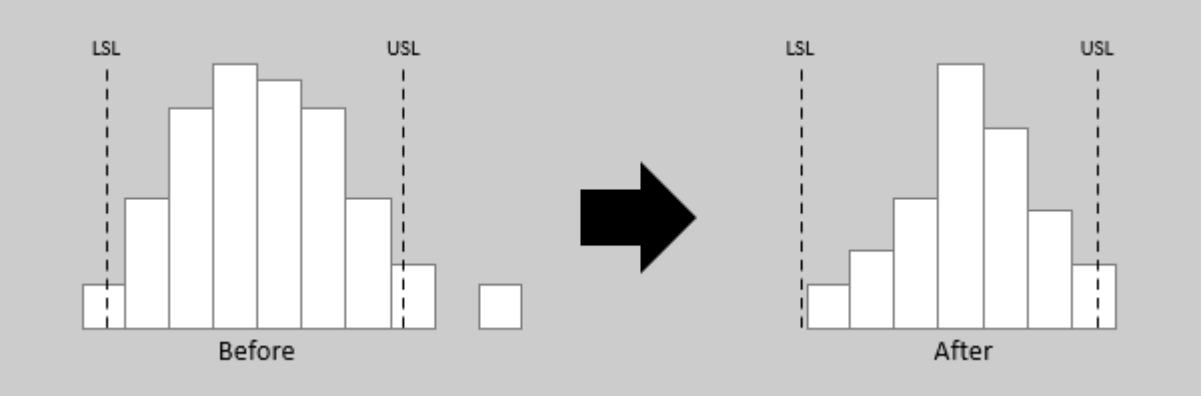
- Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability.
- Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



- 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.
- Although histograms are better in displaying the distribution of data, a box plot can be used to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.



- Histograms and box plots can be used to verify whether an improvement has been achieved by exploring the data before and after the improvement initiative. Both tools can be helpful to identify whether variability is within specification limits, whether the process is capable, and whether there is a shift in the process over time.



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 the probability of the result being observed, given that a certain statement (the null hypothesis) is true. 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, a p-value can be calculated and determine significance between two different groups of a dataset.

Setting up Your Experiment

1. **Define the hypotheses**
2. **Set the significance level to determine how unusual your data must be before it can be considered significant.**
3. **Decide to use a one-tailed or two-tailed test.**
4. **Determine sample size with a power analysis**

Calculating the Standard Deviation

1. Define the formula for standard deviation.
2. Average the samples in each group.
3. Subtract each sample from the average.
4. Square each of these numbers and add them together.
5. Divide by the total sample number minus 1
6. Take the square root.

Determining Significance

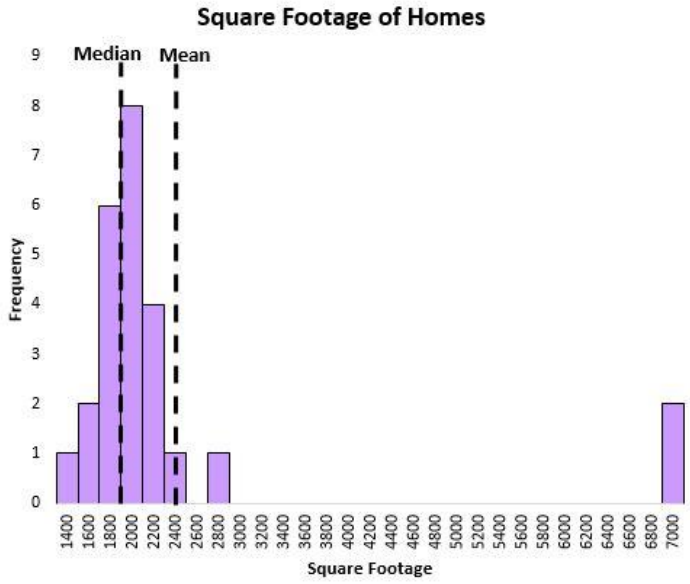
1. Calculate the variance between your 2 sample groups
2. Calculate the t-score of your data.
3. Determine the degrees of freedom of your sample
4. Use a t table to evaluate significance.

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

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

The median also does a better job of capturing the central location of a distribution when there are outliers present in the data. For example, consider the following chart that shows the square footage of houses on a certain street:



The mean is heavily influenced by a couple extremely large houses, while the median is not. Thus, the median does a better job of capturing the "typical" square footage of a house on this street compared to the mean.

When a distribution is skewed or outliers are present, the median does a better job of describing the center of the distribution than the mean.

15. What is the Likelihood?

- Likelihood is central to parametric statistical inference. The likelihood is a basis for the likelihood ratio test: a uniformly most powerful test for comparing two point hypotheses. It is also the basis for the maximum likelihood estimate**.** Likelihood refers to how well a sample provides support for particular values of a parameter in a model.

Suppose, flipping the coin 100 times and it only lands on heads 17 times. The likelihood that the coin is fair is quite low. If the coin was actually fair, the expectation is it to land on heads much more often**.**

when calculating the likelihood, it is trying to determine if the model parameter ($p = 0.5$) is actually correctly specified.

- Suppose a spinner is splited into thirds with three colors on it: red, green, and blue. Assuming that it's equally likely for the spinner to land on any of the three colors. It is rotated 100 times and it lands on red 2 times, green 90 times, and blue 8 times. It is considered that the likelihood that the spinner is actually equally likely to land on each color is very low.

when calculating the likelihood we're trying to determine if the model parameters ($P(red) = 1/3$, $P(green) = 1/3$, $P(blue) = 1/3$) are actually correctly specified.