Statistics Worksheet – 6 Answers

1) D
2) A
3) A
4) C
5) A
6) A
7) C
8) B
9) B
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 before using advanced statistical analysis technique.

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.

Although histograms are better in displaying the distribution of data, you can use a box plot 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.

You can use histograms and box plots 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.

Many statistical applications allow the option of summarizing your data graphically. This can reveal unusual observations in your data that should be investigated before performing detailed statistical analysis.

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. It is recommended that you plot your data graphically before proceeding with further statistical analysis.

11) How to select metrics?

Selecting Metrics for Machine Learning

Fayrix Machine Learning expert shares performance metrics that are commonly used in Data Science for assessing performance of Machine Learning models

KEY STEPS TO SELECTING EVALUATION METRICS

First of all, metrics which we optimise tweaking a model and performance evaluation metrics in machine learning are not typically the same. Below, we discuss metrics used to

optimise Machine Learning models. For performance evaluation, initial business metrics can be used.

Understanding the task

Based on prerequisites, we need to understand what kind of problems we are trying to solve. Here is a list of some common problems in machine learning:

Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.

Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.

Ranking. The model will predict an order of items. For example, we have a student group and need to rank all the students depending on their height from the tallest to the shortest.

In our case, we are solving the problem of finding mathematical metrics which will also optimize the initial business problem.

CLASSIFICATION performance metrics

CONFUSION MATRIX

This matrix is used to evaluate the accuracy of a classifier and is presented in the table below.

False Positive (FP) moves a trusted email to junk in an anti-spam engine.

False Negative (FN) in medical screening can incorrectly show desease absense, when it is actually positive.

ACCURACY METRIC

This metric is the basis one. It indicates the number of correctly classified items compared to the total number of items.

Keep in mind that accuracy metric has some limitations: it doesn't work well with unbalanced classes that can have many items of the same class and few other classes.

RECALL/SENSITIVITY METRIC

Recall Metric shows how many True Positives the model has classified from the total number of positive values.

PRECISION METRIC

This metric represents the number of True Positives which are really positive compared to the total number of positively predicted values.

F1 SCORE

This metric is a combination of precision and recall metrics which serves as a comprise. The best F1 score equals 1, while the worst one is 0.

REGRESSION

performance metrics

MEAN ABSOLUTE ERROR (MAE)

This regression metric indicates the average sum of absolute difference between the actual and predicted value.

MEAN SQUARE ERROR (MSE)

Mean Squared Error (MSE) calculates the average sum of squared difference between the actual and predicted value for the entire data points. All related values are raised to the second power therefore all of negative values are not compensated by positives.

Moreover, due to the features of this metric, the impact of errors is higher. For example, if the error in our initial calculations is 1/2/3, MSE will equal 1/4/9 respectively. The less MSE is, the more accurate our predictions is. MSE = 0 is the optimal point in which our forecast is perfectly accurate.

MSE has some advantages over MAE:
1. MSE highlights large errors over small ones.
2. MSE is differentiable which helps find minimum and maximum values using mathematical methods more effectively.

ROOT MEAN SQUARE ERROR (RMSE)
RMSE is a square root of MSE. It is easy to interpret compared to MSE and it uses smaller absolute values which is helpful for computer calculations.

Fayrix Solutions
At Fayrix we are ready to provide our clients with effective technological competence approaches and build dedicated development team for your project in just 2 weeks.

Fault prediction
Fayrix Machine Learning solution forecasting machine failureand analyses equipment working conditions and predicts potential failures and downtime.

Recommender System by ML
Fayrix recommender system development services feature our proprietary recommender engine to create a personalized product offering and customer experience.

Preventive Maintenance by ML
Fayrix predictive maintenance solutions collect data from equipment sensors, analyses it, predicts faults and forecasts optimal schedule for maintenance.

Customer Churn Prediction
Fayrix customer churn prediction solution predicts customer churn rate to develop a customer retention plan and reduce customer attrition in a timely manner.

RANKING
performance metrics
BASIC METRIC
Best Predicted vs Human, BPH:
The most relevant item is taken from an algorithm-generated ranking and then compared to a human-generated ranking. This metric results in the binary vector that shows the difference in estimations of an algorithm and a human.
KENDALL'S TAU COEFFICIENT
Best Predicted vs Human, BPH:
Kendall's tau coefficient shows the correlation between the two lists of ranked items based on the number of concordant and discordant pairs in a pairwise: in each case we have two ranks (machine and human prediction). Firstly, the ranked items are turned into a pairwise comparison matrix with the correlation between the current rank and others. A concordant pair means an algorithm rank correlates with a human rank. Otherwise, this will be a discordant pair. Therefore, this coefficient is defined as following:

The values of τ varies from 0 to 1. The closer $|\tau|$ is to 1, the better ranking is. For instance, when τ-value is close -1, the ranking is just as accurate, however the order of its items

should be vice-a-versa. This is quite consistent with estimate indicators which assign the highest rank to the best values, whereas during manual human ranking the best ones receive the lowest ranks. τ-value=0 indicates the lack of any correlation between ranks.

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

Tests for statistical significance are used to estimate the probability that a relationship observed in the data occurred only by chance; the probability that the variables are really unrelated in the population. They can be used to filter out unpromising hypotheses.

Tests for statistical significance are used because they constitute a common yardstick that can be understood by a great many people, and they communicate essential information about a research project that can be compared to the findings of other projects.

However, they do not assure that the research has been carefully designed and executed. In fact, tests for statistical significance may be misleading, because they are precise numbers. But they have no relationship to the practical significance of the findings of the research.

Finally, one must always use measures of association along with tests for statistical significance. The latter estimate the probability that the relationship exists; while the former estimate the strength (and sometimes the direction) of the relationship. Each has its use, and they are best when used together.

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

The mean life of a calculator battery is 300 hours. Assume a normal distribution with a standard deviation of 4 hours. What is the probability that a randomly selected battery lasts for more than 305 hours?
Here's the procedure:

Convert your data value, mean and standard deviation to a z-score:
$z=x-\mu\sigma=305–3004=1.25$.
Look that number up in a table of z-scores or using some kind of technology. This should give you something in the neighborhood of 0.89435. This means that $P(x<305)=0.89435$. (Be sure you know what your table is giving you. Most, but not all, give the probability or area less than the z-score.)
Realize that that isn't what the question asked for. You want the probability that the battery lasts more than 305 hours, not less than. That makes your result
$P(x>305)=1-P(x<305)=0.10565$.

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

Median is the middle value in a rank-ordered sequence. Average is the sum of all observation values divided by the number of cases observed.

Medians are not affected by outliers, while averages can swing wildly due to extreme anomalies that are irrelevant to the norms.

The middle (median) remans the same middle value regardless of the size of the highest or the lowest case, which has great effects on the average.

Whenever Bill Gates enters a Starbucks, the average customer income skyrockets, but their median income remains about the same. If a homeless person left while Gates entered, the median would be constant.
In a statistically random population sample, the median remains very close to the mode (the single most frequently encountered value), so the median is a superior measure of the norm. The average can bounce all over the place, based on the outliers and the sample distribution.
If 10 kindergarten kids are visited in a room by one typical professional basketball player, the average height skyrockets, while the median height probably stays almost exactly the same.
The smaller the sample size and the more non-standard the population observation distribution, the larger the differences between the median and the average.
A universal sample of all possible observations will tend to produce a median and average values that are identical (or statistically the same) … pretty much.

15) What is the Likelihood?

Likelihood Function:

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. If Xo is the observed realization of vector X, an outcome of an experiment, then the function L(T | Xo) = P(Xo | T) is called a likelihood function. In other words, you have to substitute the observations instead of the random vector into the expression for probability of the random vector, and to consider the new expression as a function of parameters T. The likelihood function varies from outcome to outcome of the same experiment, for example, from sample to sample.

The likelihood function itself is not probability (nor density) because its argument is the parameter T of the distribution, not the random (vector) variable X itself. For example, the sum (or integral) of the likelihood function over all possible values of T should not be equal to 1.

Even if the set of all possible values of the vector T is discrete, the likelihood function still may be continuous (as far as the set of parameters T is continuous).

Suppose you have a sample of 50 balls - 10 white and 40 black. The balls have been drawn randomly from a large bag with black and white balls (population). The question of interest is the proportion of white balls in the population. A Binomial distribution Pb(X; N=50, p=T) is a reasonable statistical model for the number X of black balls in a sample of N=50 balls drawn from a population with proportion T of black balls. To obtain the likelihood function for your data you have to substitute observation X=10 into the formula for the binomial distribution, and to consider the expression as a function of T.