Homework 5

April 25, 2020

1 Recitation Exercises

1.1 Exercise 7

a) Find the limit of the value shown in the text for Grubbs' test as m approached infinity?

$\lim_{m\to\infty} \frac{m-1}{\sqrt{m}} \sqrt{\frac{t_C^2}{m-2+t_C^2}} = \lim_{m\to\infty} \frac{m-1}{\sqrt{m(m-2+t_C^2)}} * t_C = 1 * t_C = t_C$

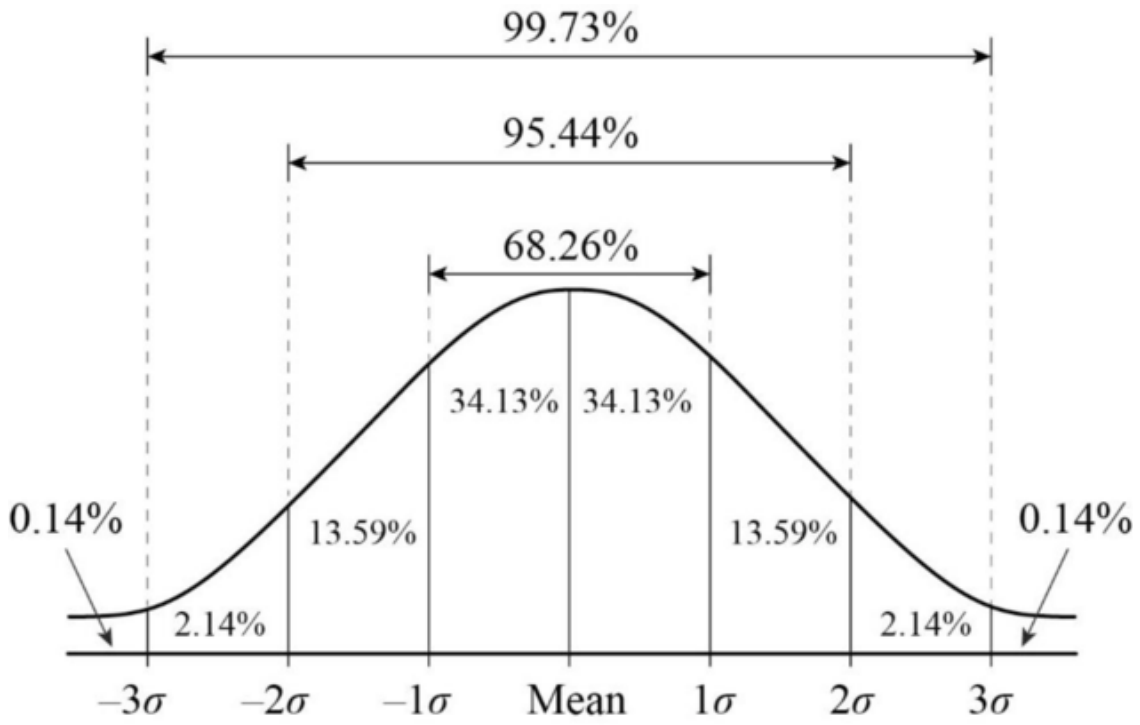
b) Describe the meaning of the result above.

I'm honestly unsure of the meaning. I can understand that the Grubbs test basically idenifies the maxiumum value from m as an outlier, but not really sure how limiting the value above proves this. From the algorithm, we can see that t_C is chosen so that the probability of the sample mean is greater than equal to t_C is equal to the significance level, which in this case is 0.05.

1.2 Exercise 8

Many statisical test for outliers were developed in an environment in which a few hundered observations was a large data set.

a) For a set of 1,000,000 values, how likely are we to have outliers according to the test that says a value is an outlier if it is more than 3 standard deviations from the average?



The image above is a normal bell-curve that shows up to three standard deviations from the mean. As shown above, 0.14% on each side of the bell-curve is considered as above three standard deviation. Because of this, you can say that 0.28% is more than 3 standard deviations. In our example with the 1,000,000 values, at least 280,000 would be outliers.

b) Does the approach that states an outlier is an object of unusually low probability need to be adjusted when dealing with large data sets? If so, how?

I feel like it depends on what the data set actually is. It's important to remember that outliers have a major influence of what is determined from the data. If the outliers are considered a negative influence, then you would try to minimize the amount of outliers present. If the outliers are the only thing that are considered interesting, then you would expect to see a little more.

1.3 Exercise 9

The probability density of a point x w/ respect to a multivariate normal distribution having a mean μ & a covariance matrix Σ is given by … if we log f(x), then the following results in:

log prob(x) = - log($(\sqrt{2\pi})^m * |\Sigma|^{1/2}$ - $\frac{1}{2} * (x - \mu) * \Sigma^{-1} * (x - \mu)^T$

If we use the sample mean and the covariance matrix as estimates of μ & Σ, then

log prob(x) = - log($(\sqrt{2\pi})^m * |S|^{1/2}$ - $\frac{1}{2} * (x - \overline{x}) * S^{-1} * (x - \overline{x})^T$

If you noticed that the left side represents a constant factor, then the right side wth the variables (sample mean) should be the closest to matching to the Mahalanobis distance.

1.4 Exercise 11

Considere the K-means scheme for outlier detecion described in Sec 9.5 & Fig. 9.10.

a) The pts. on the bottom of the compact cluster in Fig. 9.10 have a higher outlier score than those at the top of the compact cluster. Why?

The mean of the points (pt. D) is pulled somewhat upward from the center of the compact cluster.

b) Suppose that we choose the # of clusters to be larger (10). Would the proposed technique still be effective in finding the most extreme outlier at the top of the figure? Why or why not?

In this case, the point would just become a cluster by itself, so the proposed technique would never work.

c) Use of relative distance adjusts for differences in density. Provide an example of where the approach leads to a wrong conclusion.

Anything within the health department might not lead to the correct conclusion. If, for example, if someone's heart rate is above or below a certain range, then it would not be a good idea to ignore such an outlier.

1.5 Exercise 12

If the probability that a normal object is classified as an anomaly is 0.01 & the probability that an anomalous object is classified as an anomalous is 0.99, then what is the false alarm rate and detection rate if 99% of the objects are normal?

detection rate = # of anomalies detected/total # of anomalies = 99%.

false rate = # of false anomalies/# of objects classified as anomalies = (0.99m * 0.01)/(0.99m * 0.01 + 0.01m * 0.99) = 50%

1.6 Exercise 16

Consider a set of points that are uniformly distributed on the interval [0,1]. Is the statistical notion of an outlier as an infrequently observed value meaningful for this data?

If the set of points are uniformly distributed then each object should have at least the same probability. The only way that an outlier would be meaningful if it had a low probability, which in this case, it doesn't. So, no.