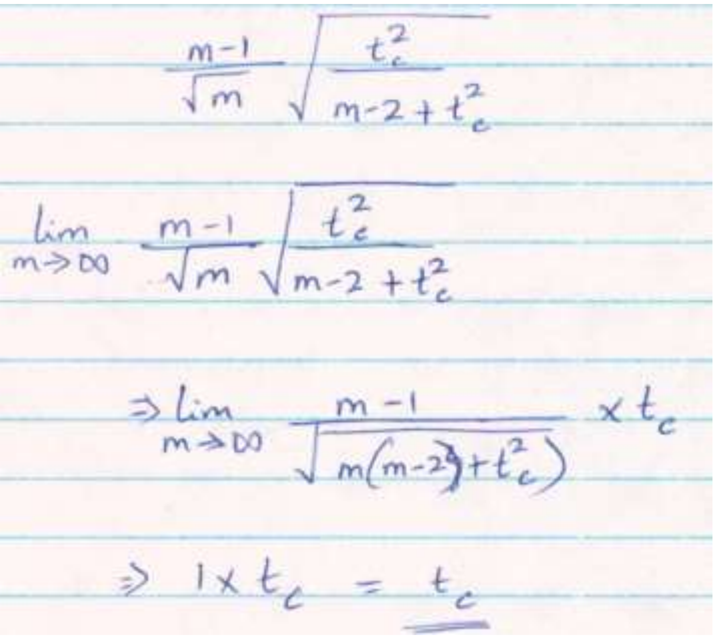
Problem 7:

Part a) Formula:

(m-1/sqrt(m))*sqrt(tc^2/m-2+tc^2)

lim m-> infinite m-1/sqrt(m(m-2)+tc^2)*tc

= 1*tc = tc



$$\frac{m-1}{\sqrt{m}}\sqrt{\frac{t_c^2}{m-2+t_c^2}}$$

$$\lim_{m\to\infty}\frac{m-1}{\sqrt{m}}\sqrt{\frac{t_c^2}{m-2+t_c^2}}$$

$$\Rightarrow \lim_{m\to\infty}\frac{m-1}{\sqrt{m(m-2)+t_c^2)}}\times t_c$$

$$\Rightarrow 1\times t_c = \underline{\underline{t_c}}$$

Also, the value of tc continue to increase with m.

For m = 10^20, tc = 93 for significance value = 0.05

Part b):

The distribution of ‘g’ becomes a ‘t’ distribution as ‘m’ increase. This relationship we observe in the part A where we found that the value of tc continues to increase with m.

Problem 8:

Part a)

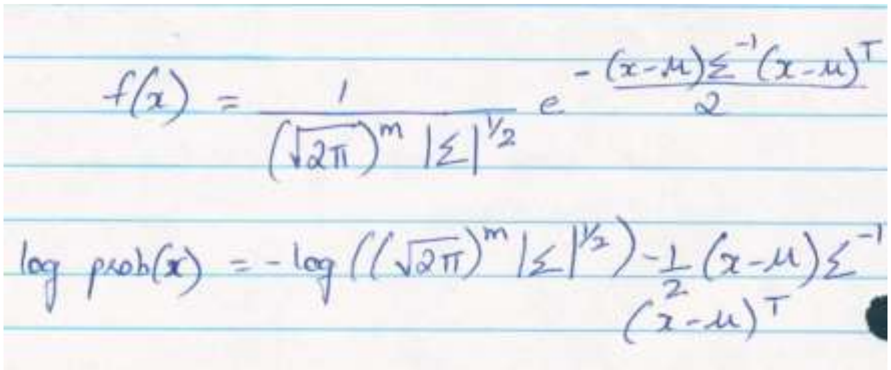
The probability is either 0.00135 for a single-sided deviation of 3 standard deviations or 0.0027 for a double-sided deviation. Thus, the number of anomalous objects will be either 1350 or 2700.

Hence, that even a small probability of an outlier yields a large number of outliers for a large data set. The probability is unaffected by the number of objects.

Part b)

These are thousands of outliers in a million objects. We may choose to accept these objects as outliers or prefer to increase the threshold so that fewer outlier result.

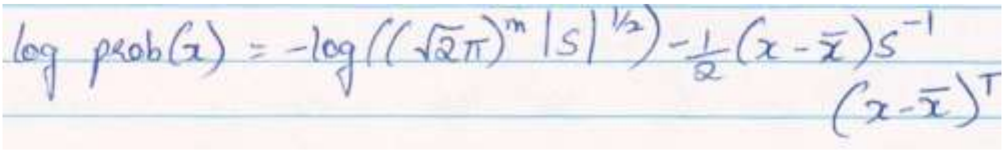
Problem 9:



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

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

We use sample mean and covariance as estimates of μ and Σ



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

The constant and the constant factor do not affect the ordering of this quantity only their magnitude. Thus, if we want to base a distance on this quantity, we can keep only the variable part, which is Mahalanobis Distance.

Problem 11:

Part a)

The mean of the points is pulled somewhat upward from the center of the compact cluster by point D.

Part b)

No, this point would become a cluster.

Part c)

If absolute distances are important. Examples consider heart rate monitors for patients. If heart rate above or below a specified range of values, then this has a physical meaning. It would be incorrect not to identify any patient outside that range as abnormal even though these may be a group of patients that are relatively similar to each other and all have abnormal heart rates.

Problem 12:

Deduction rate = number of anomalies detected / total no of anomalies

False alarm rate = no of false anomalies / no of objects classified anomalies

Given: deduction rate= 99%

False alarm rate = 0.99m*0.01/ (0.99m*0.01+0.01m*0.99) = **50%**

Problem 16:

I do **not** think, because the statistical definition for an outlier is relying on the idea that an object with a relatively low probability is suspect. Now we know it is not necessarily important that having low probability is an outlier, it could be an anomaly. With uniform distribution, no such distinction can be made.