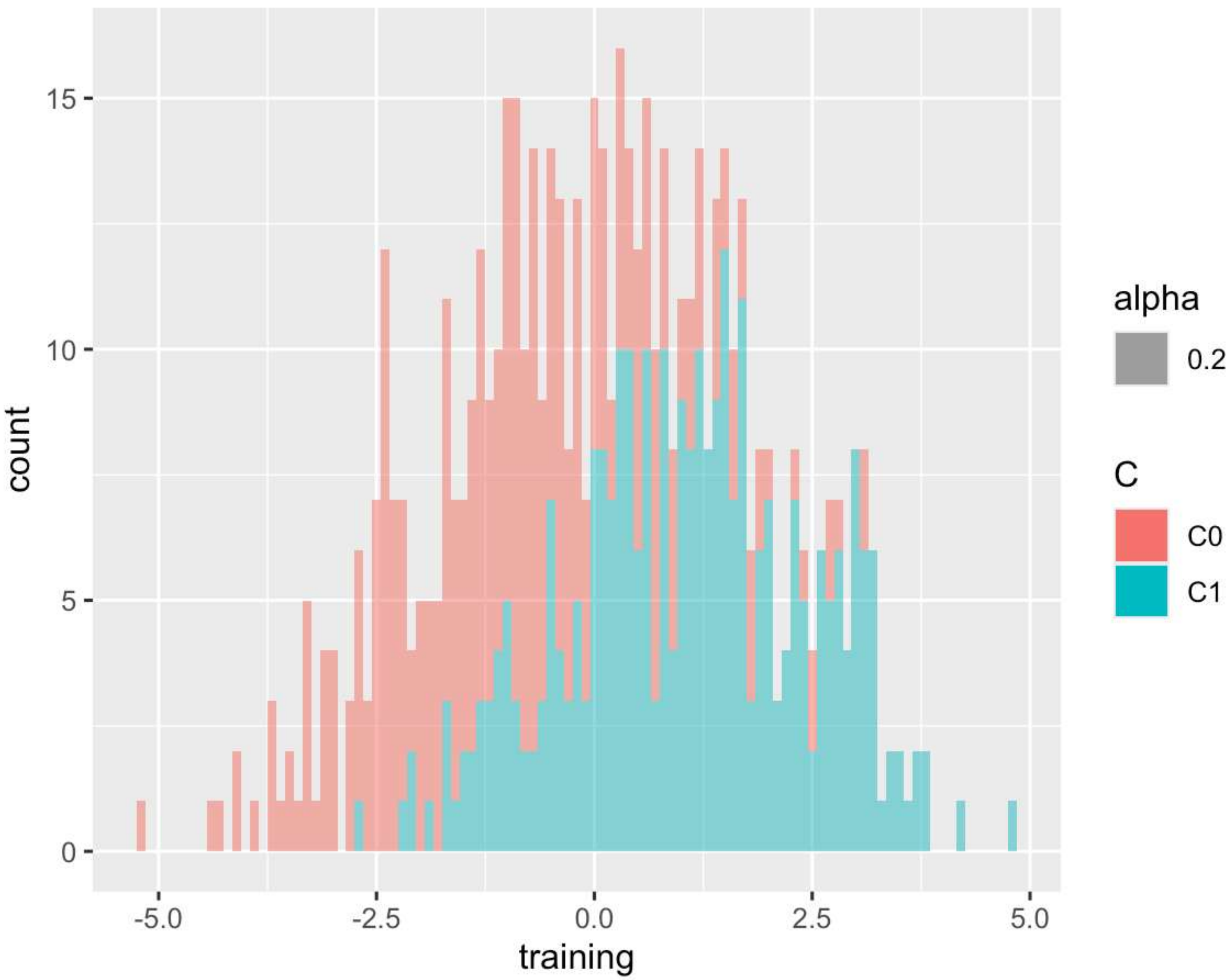
Midterm Examination Solution

Directions: There are unlimited uploads to Canvas for this exam, so upload frequently and often. You may not upload after the deadline.

1.) (5 points) Please upload the training data from the files: **trainingData0-1.csv (https://northeastern.instructure.com/courses/193161/files/30363191?wrap=1) (https://northeastern.instructure.com/courses/193161/files/30363191/download?download_frd=1)** and **trainingData1-1.csv (https://northeastern.instructure.com/courses/193161/files/30363205?wrap=1) (https://northeastern.instructure.com/courses/193161/files/30363205/download?download_frd=1)** . From these files produce a single plot showing two histograms with the same binwidth. Label the vertical axis "count" and the horizontal axis "training". Provide a legend for "C0" and "C1" and choose a binwidth that shows the curve shape. What family of distributions do these histograms resemble?



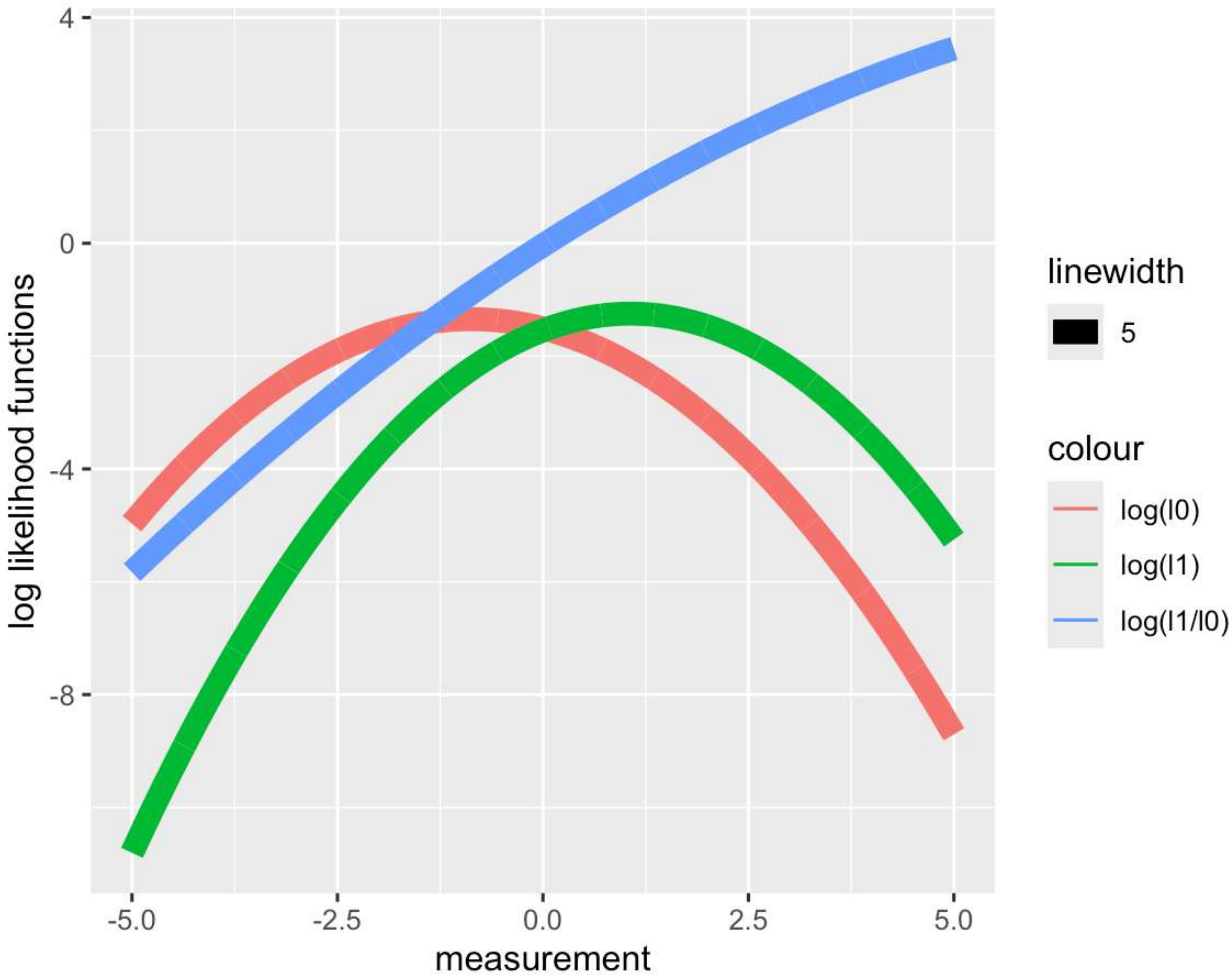
These histograms appear to have a Gaussian shape.

2.) (20 points) Let's assume that you chose "normal distributions" in 1.). From Wikipedia, find the parameters for the normal distribution and the methods to estimate them. Using the training data from class 0, estimate the parameters using these formulae. Repeat the procedure for class 1. Provide your formulae and estimates here. Also, produce three curve on a single plot, versus the measurement axis from -5 to 5: the logarithms of the likelihood function for class 0, the logarithm of the likelihood function for class 1, and the logarithm of the likelihood ratio (with class 1 on top). Label the vertical axis "log likelihood functions", the horizontal axis "measurement", and provide a legend with the labels: log(l0), log(l1), and log(l1/l0).

The two parameters for the normal distribution are the mean and the standard deviation. They are estimated by the sample mean and the sample variance:

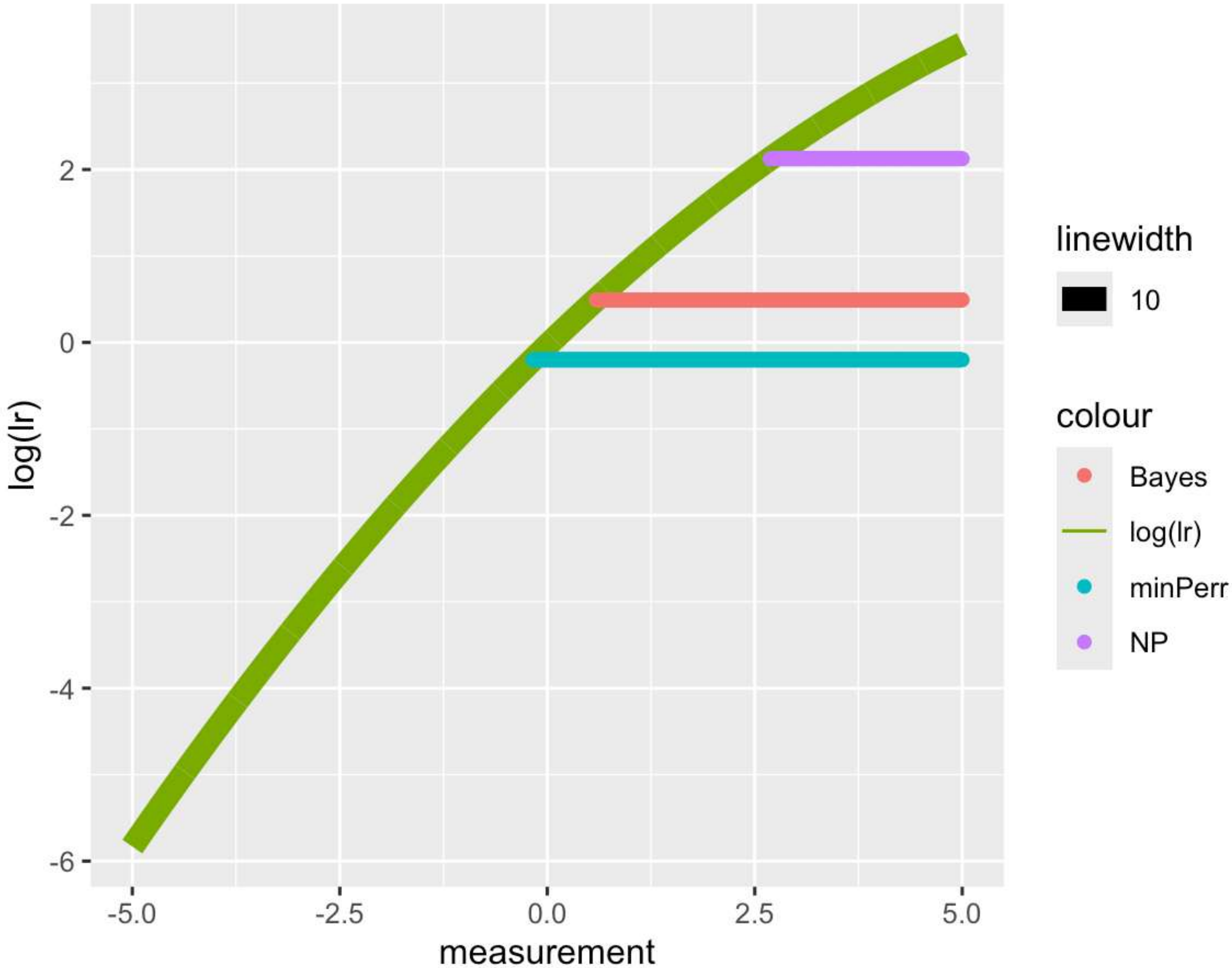
$$\hat{\mu} = \overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Here, the square root of the sample variance provides the sample standard deviation. For class 0, we have estm0=-.87, estsigma0=1.53. For class 1, we have estm1=1.07 and estsigma1=1.38. Using these estimates as the parameters for the likelihood functions we obtain the following plot of their logarithms:



3.) (15 points) Please set the thresholds for the single-measurement tests: Bayes classifier, minimum probability of error classifier, and Neyman Pearson classifier. To do this, you will need the log(l1/l0) from above, and prior probabilities (p0=0.45, p1=0.55), costs (dc0=1, dc1=2) and maximum false alarm probability (alpha=0.01). Produce a single plot showing four curves: the logarithm of the likelihood ratio from 2., and three horizontal lines at heights given by the logarithm of the Bayes threshold, the logarithm of the minimum probability of error threshold, and the NP threshold. Draw these horizontal lines only over the decision region for class 1 on the measurement axis. This will clearly show which measurement values would make a decision for class 1. Provide a legend for the plot with the labels: "Bayes", "minPerr", " NP" and "log(lr)".

Here are all the requested information in this plot:



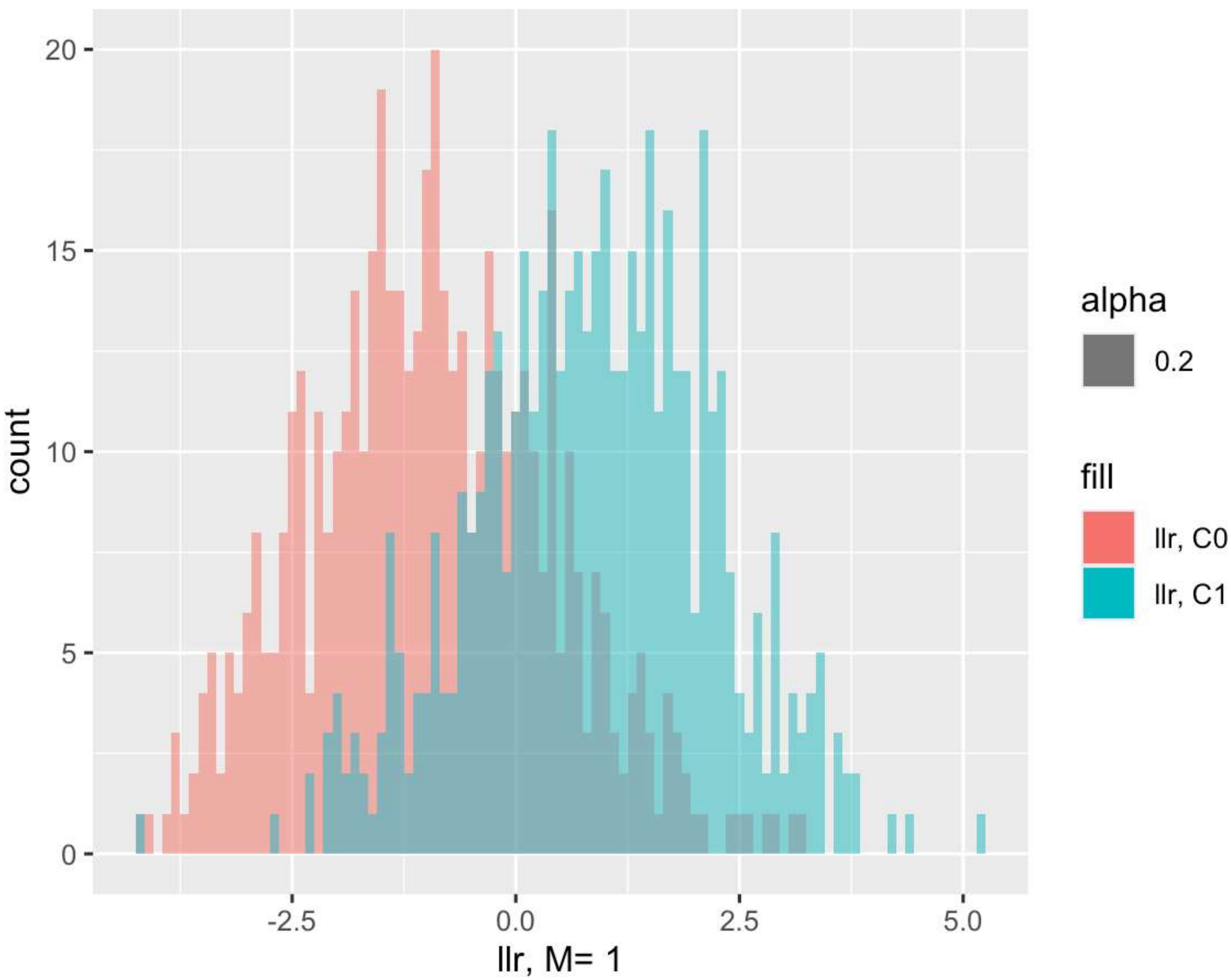
4.) (15 points) Find the minimum risk for the Bayes classifier, the error rate for the minimum error rate test, and the power for the NP classifier. For a single measurement M=1, are you satisfied with these metrics? Why or why not?

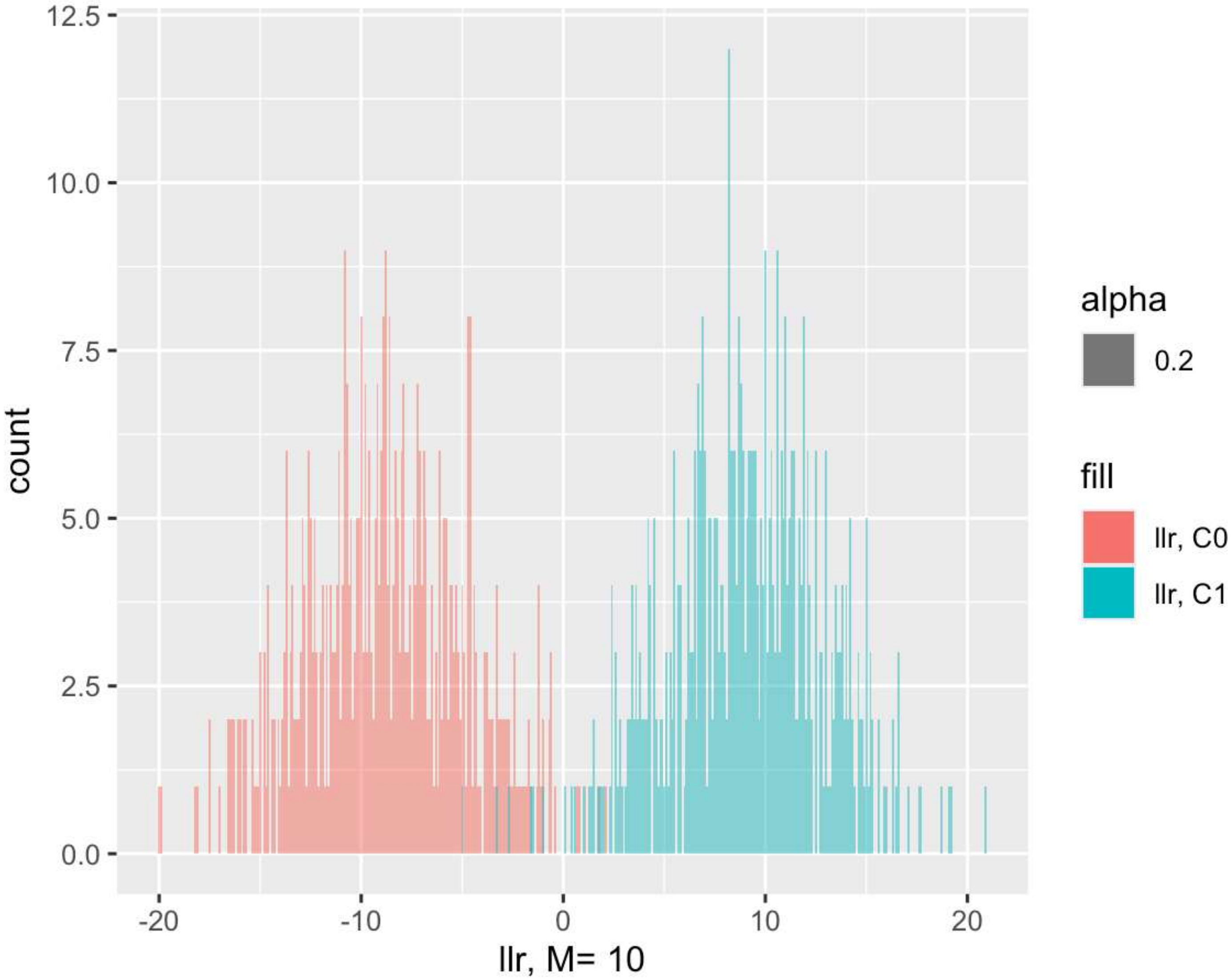
M	Bayes risk	minimum error rate	power
1	0.365	0.242	0.064

I am not happy with this metrics, and seek better perfomance with more measurements.

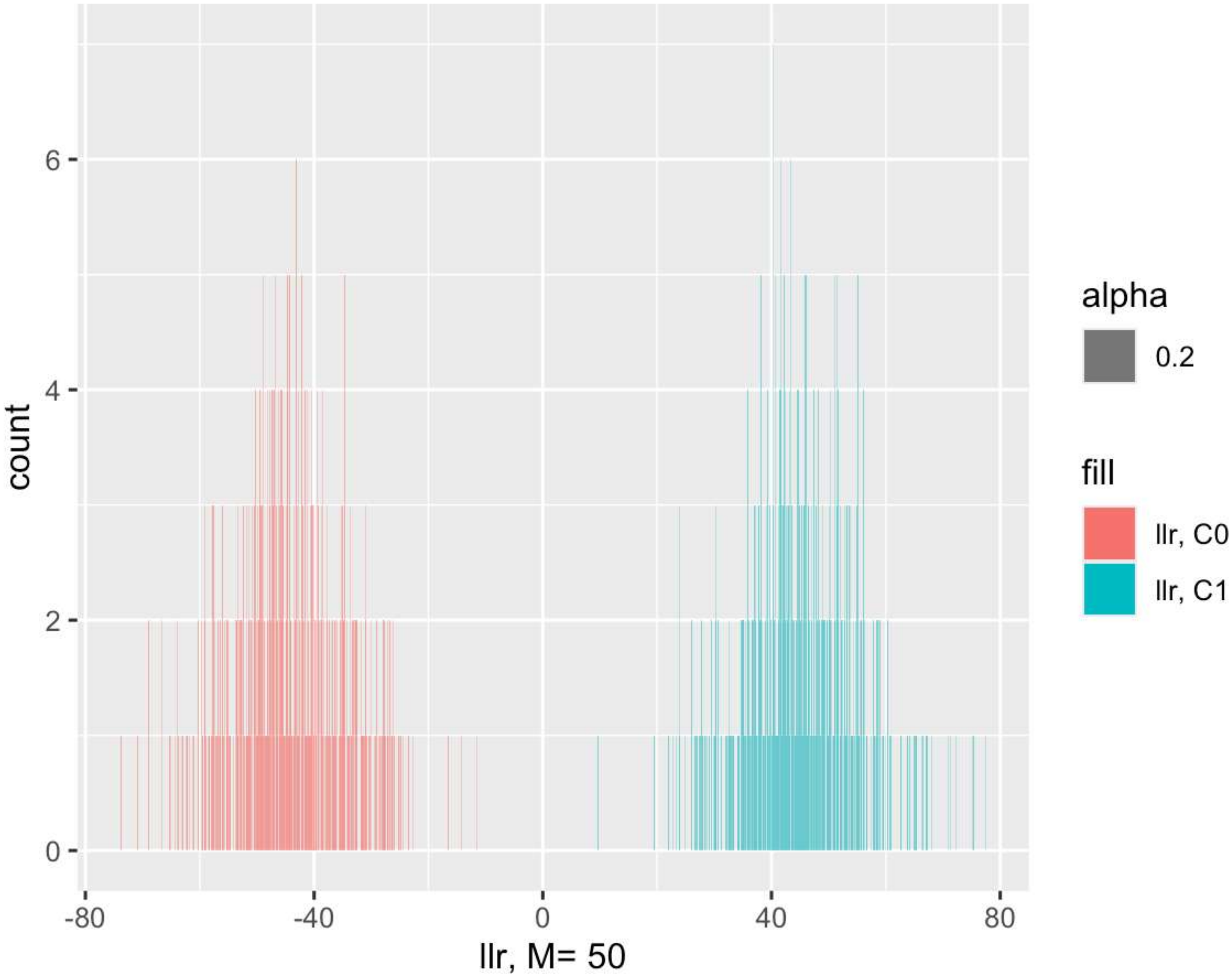
5.). (15 points) When there are M, independent measurements from the same class, we can also form optimal classifiers with almost no additional work. This is because the log likelihood ratio for M samples is the sum of the single-sample llr's, evaluated once at each measurement. Also, the Central Limit Theorem allows us to approximate this sum as Gaussian for large M. Write a script which performs 500 trials of the following: generate M independent sample from class 0, M independent sample from class 1, form a sum of the llr for each class, and plot 2 histograms of these results on the same graph. Use the estimates of the parameters obtained for each class above. Label the horizontal axis "llr, M=" with your value of M, the vertical axis with "count", and provide a legend with the labels: "C0" and "C1". Produce 4 executions of this script, and four graphs, for. M=1,10,50,100.

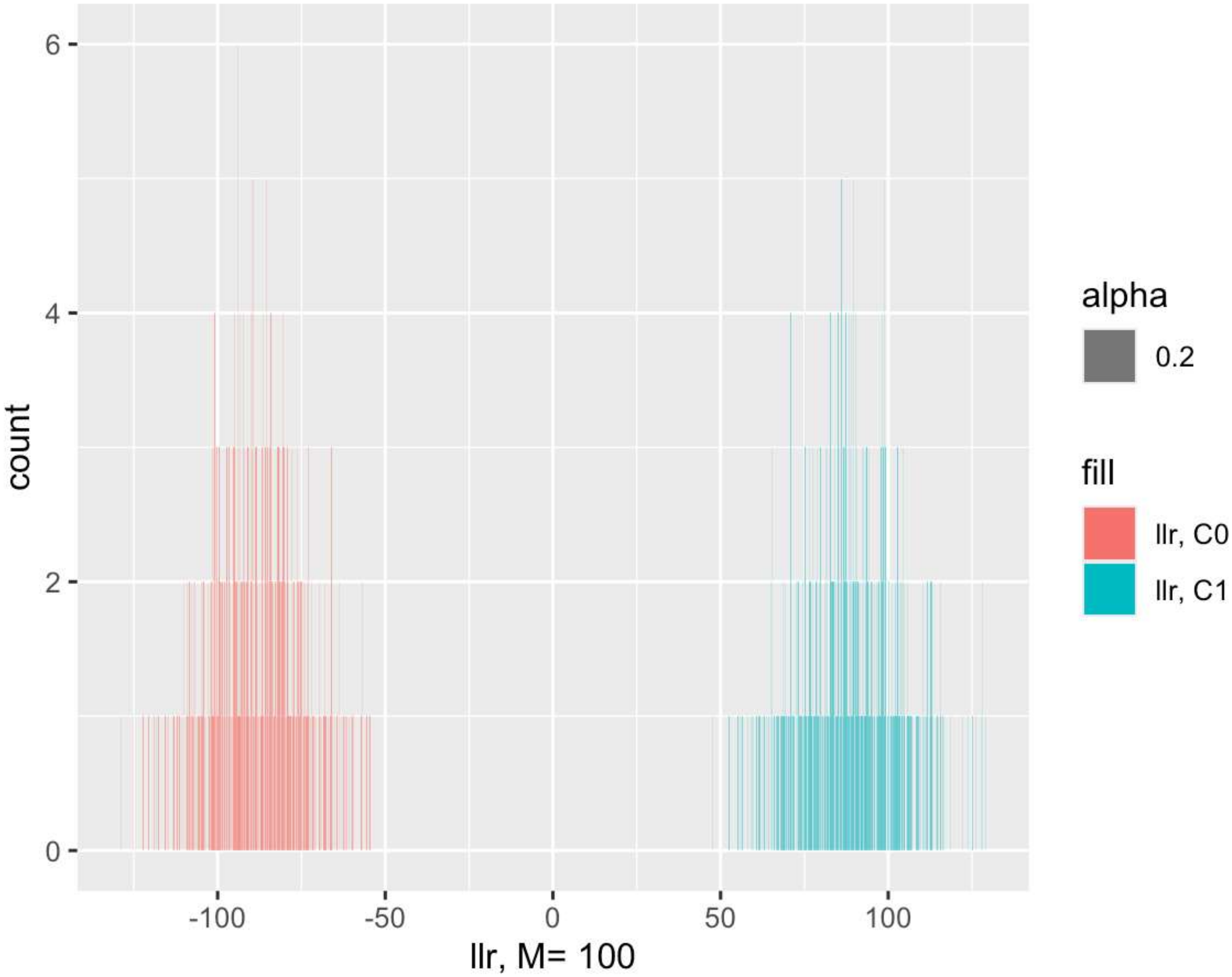
Here are the four plots:





12.5
10.0
7.5
5.0
2.5
0.0
count
-20
-10
0
10
20
llr, M= 10
alpha
0.2
fill
llr, C0
llr, C1


count
6
4
2
0
-80
-40
0
40
80
llr, M= 50
alpha
0.2
fill
llr, C0
llr, C1



6.). (10 points) What is the smallest value of M for which the histograms look Gaussian? Provide your reasoning. Also, provide the mean and variance of the llr for M=1, class 0, and the mean and variance for the llr for M=1, class 1.

The histograms look Gaussian for M>=10. For M=1, the class-0 llr mean and variance are

m0 = -0.99

var0 = 2.36

For M=1, the class-1 llr mean and variance are

m1 =0.82

var1 = 1.35

7.) (20 points) Assume here that the llr look Gaussian for M>9 and provide a table of performances as follows: let the ordered columns of the table be M, minimum risk, minimum error rate, and power. Complete the table entries. for M=1, 10, 50, 100. For M=1, use your results from 4.). For the rest, use the Gaussian approximation.

Here is the table:

M	Bayes risk	minimum error rate	power
1	0.365	0.242	0.064
10	0.025	0.017	0.966
50	2e-6	1e-6	1 - 1e-10
100	4e-11	3e-11	1