Banknotes Data

Is a banknote genuine or not?

The dataset

Goal: predict whether a banknote is genuine or not based on the following four characteristics obtained from wavelet transformed images of 1370 bills:

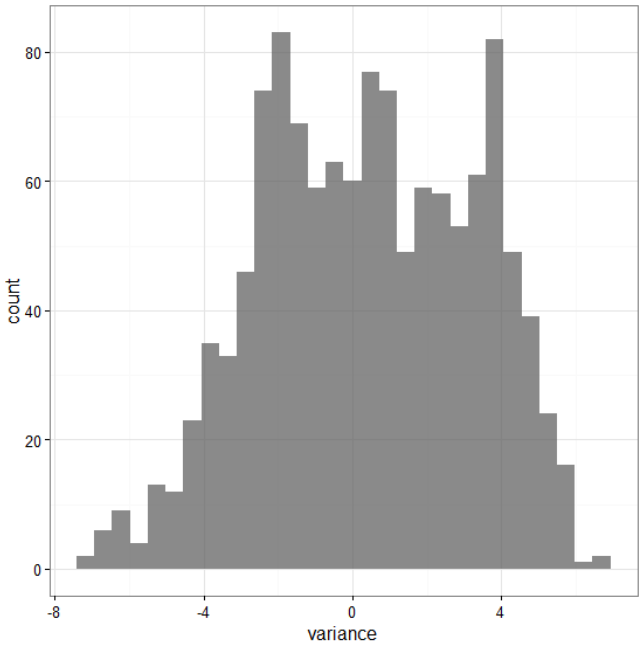
- Variance
- Skewness
- Kurtosis
- Entropy

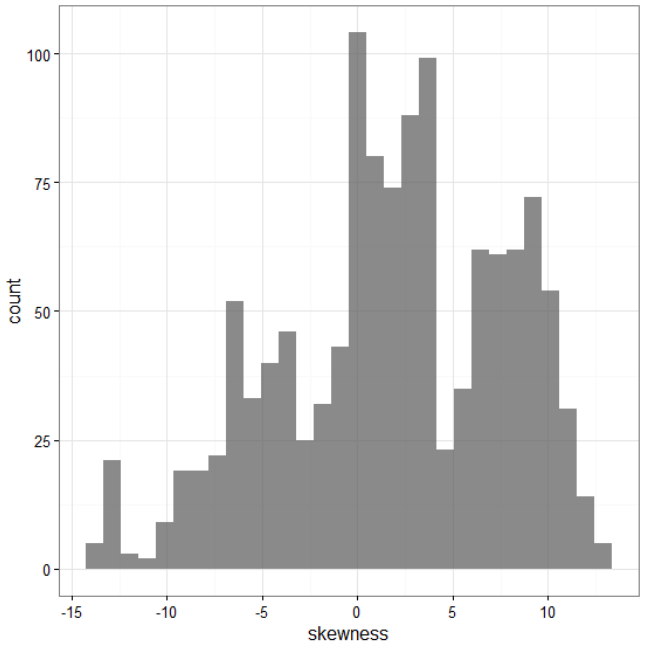
Can download from the UCI ML Repository

https://archive.ics.uci.edu/ml/datasets/banknote+authentication#

My set-up

My Goal: predict whether a banknote is genuine or not based on the following two characteristics obtained from wavelet transformed images of 1370 bills:





I set aside 10% of the data as a test set and will use the remaining data to train a logistic regression model, kNNs, LDA and QDA.

Notation

Random Variables:

- Say Y = 1 if a bill is genuine, 0 if fake.
- X_v= variance of wavelet transformed image
- X_s= skewness of wavelet transformed image

We have n=1235 observations of these variables in our training set.

Logistic Regression

- Assumes banknotes are “independent” and that the log odds is linear in the predictors:

$$\log\left(\frac{\pi}{1-\pi}\right) \text{ where } \pi = P(Y=1 \mid X_v = x_v, X_s = x_s)$$

Logistic regression model

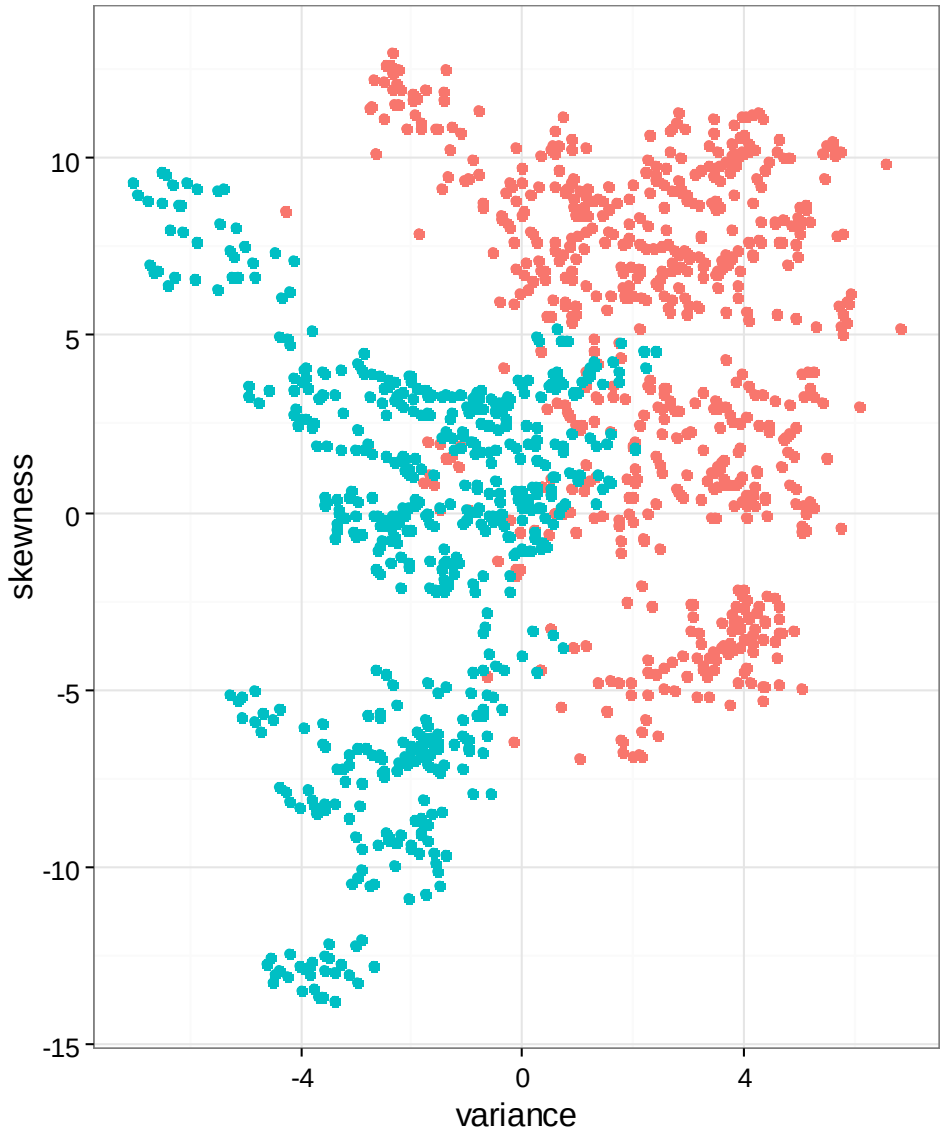
Using the training set of 1235 bills and the method of maximum likelihood, I found the coefficients of a logistic regression model

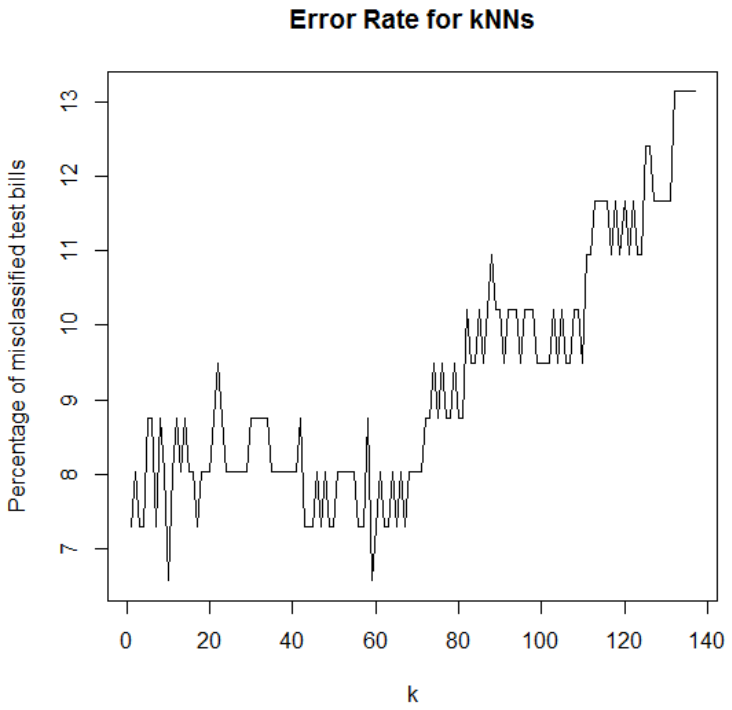
Proportion of test bills that were incorrectly classified: 16.79%

```
> model1<-glm(type~variance+skewness, data=ss,
family="binomial")
> model1

Call:  glm(formula = type ~ variance + skewness,
family = "binomial",
    data = ss)

Coefficients:
(Intercept)     variance     skewness
     0.6192      -1.1224      -0.2885
```

kNNs

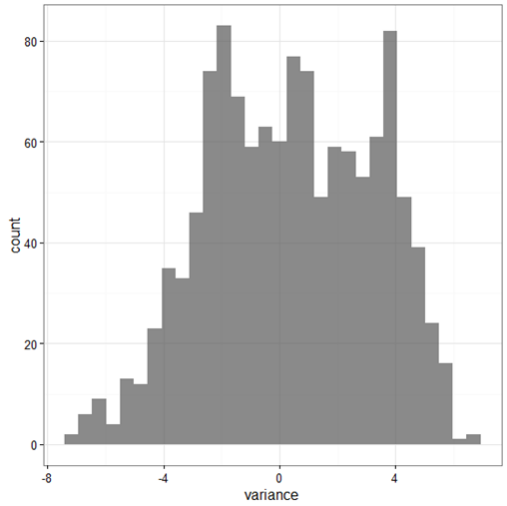
Using Euclidean distance, the minimum error rate was 6.57% for k=10



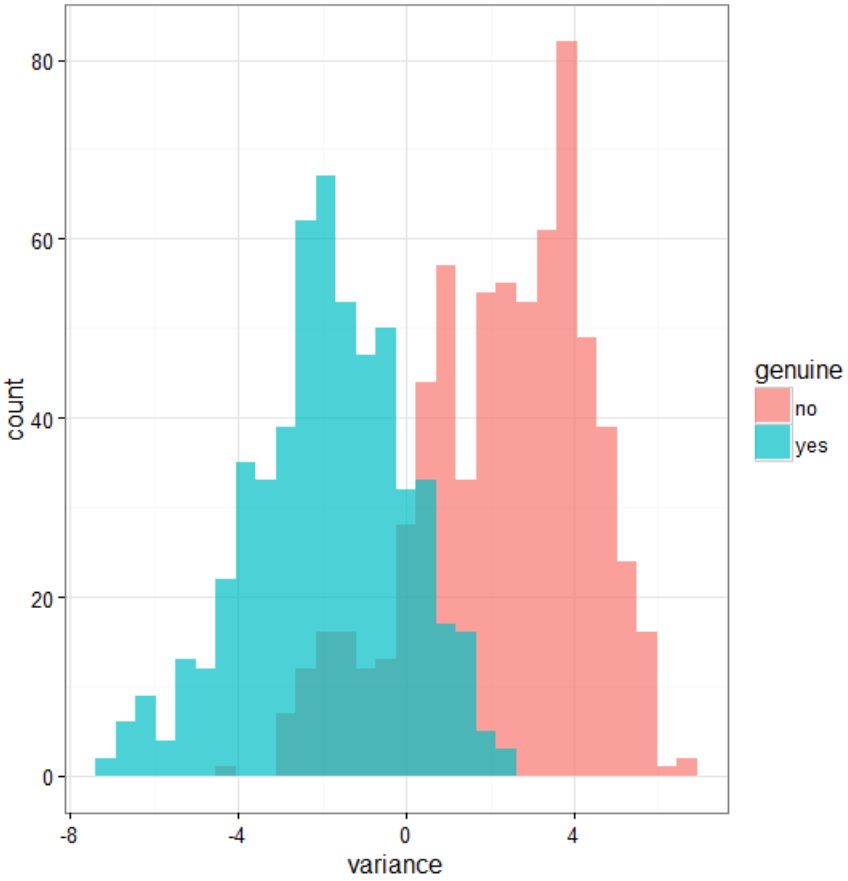


kNNs has a pretty good error rate – can we do any better?

Let's look at the dataset another way…



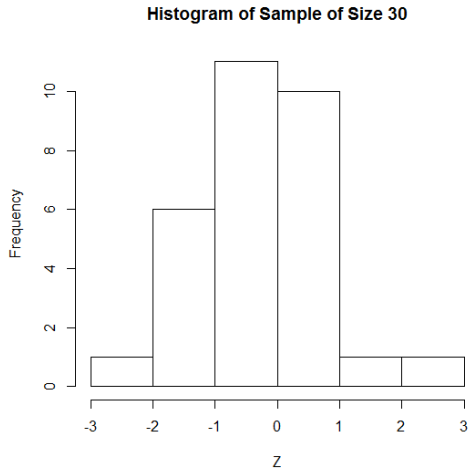
Split by bill type

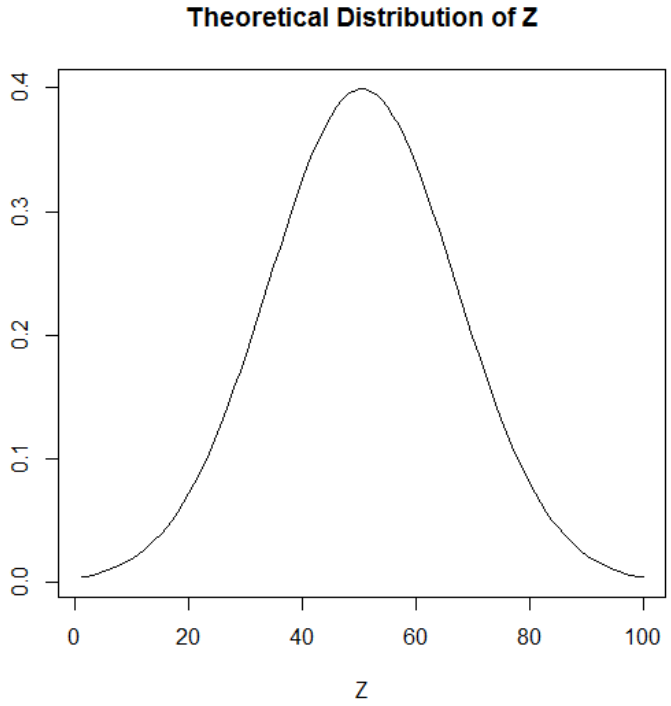


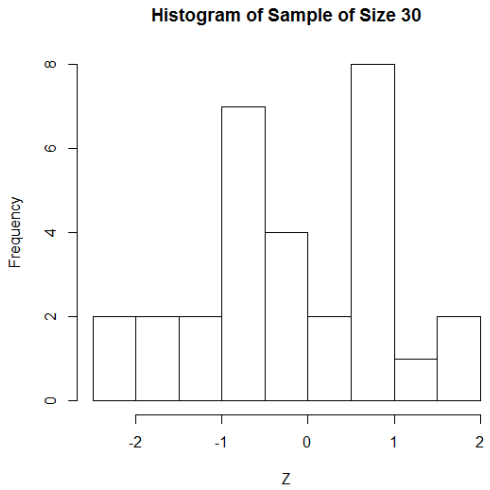
Do these shapes remind you of a famous distribution?

The Normal Distribution

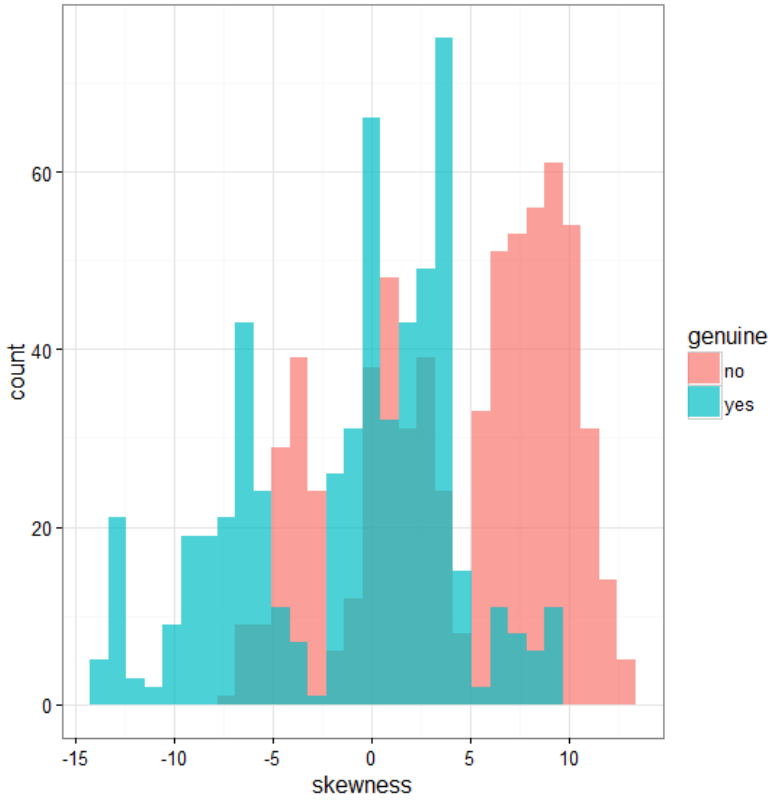
- Bell-shape
- can be described by mean μ and standard deviation σ
- Common: sums or means of enough iid RVs are always approximately normally distributed

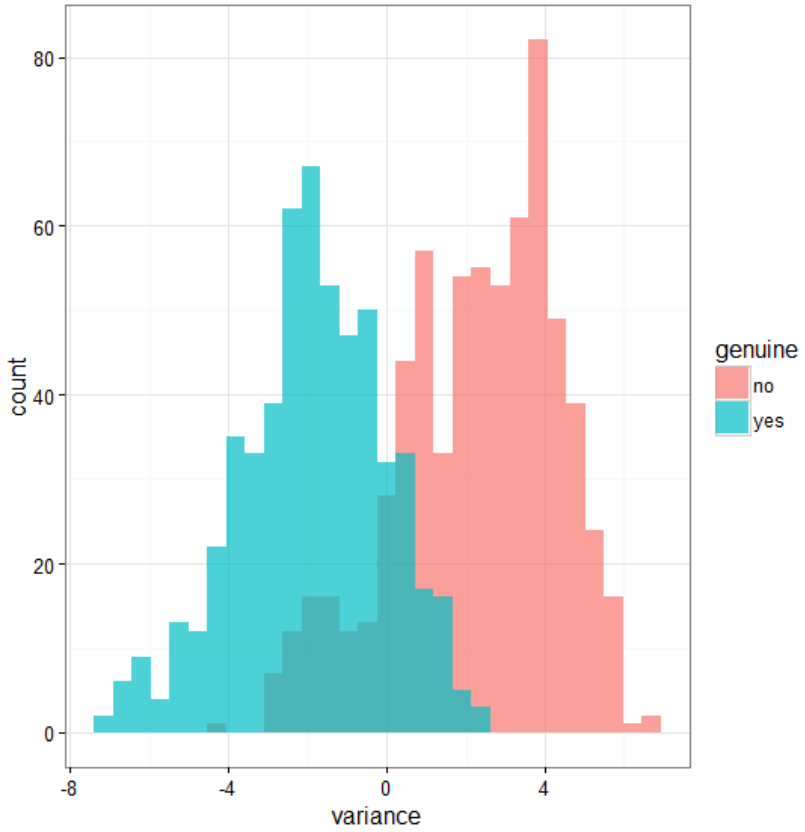




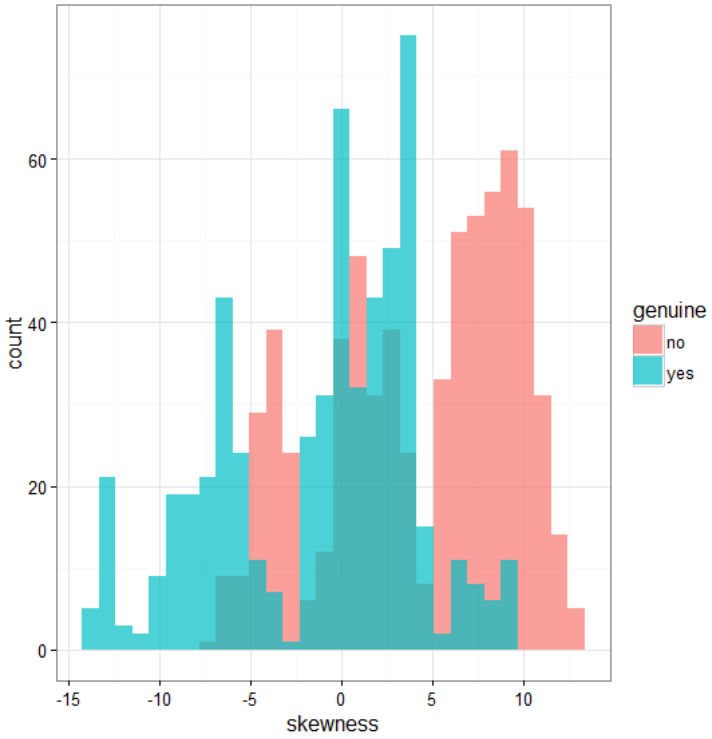


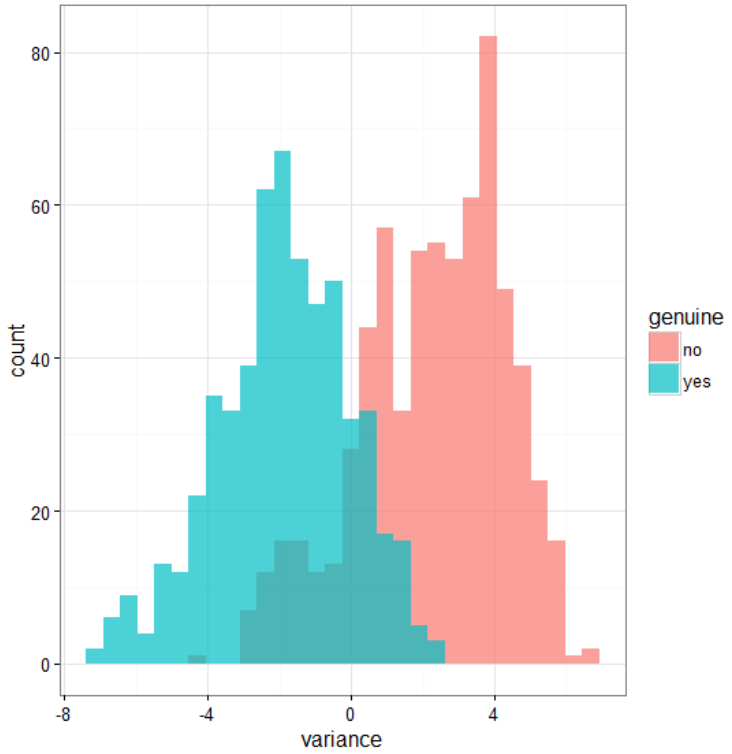
How would you describe these distributions in terms of mean and SD?





Given a bill with skewness = -2 and variance = -1.5, would you say it's real?





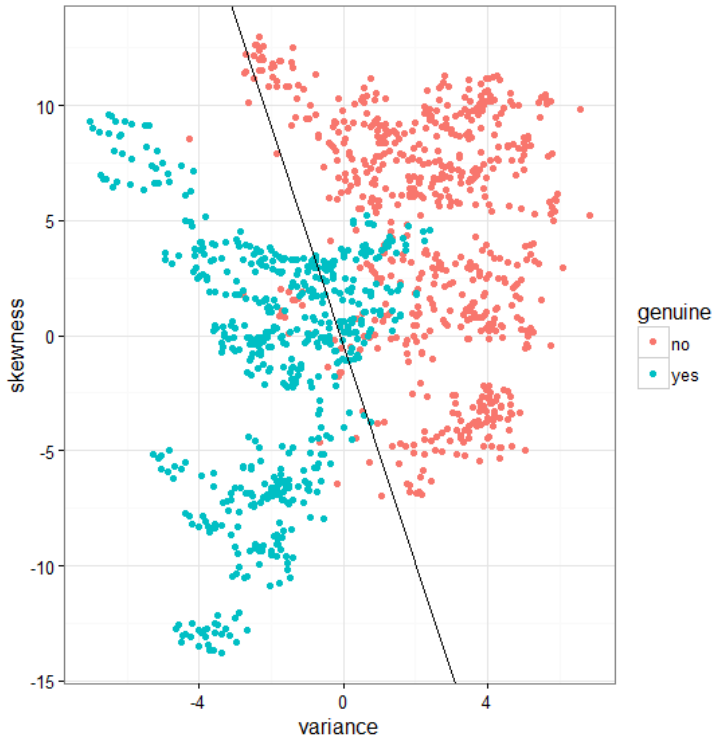
	Genuine		Fake	
	skewness	**variance**	**skewness**	**variance**
mean	-1.19	-1.89	4.31	2.28
SD	5.43	1.86	5.12	2.04

Linear Discriminant Analysis (LDA)

Idea: model the distributions of the predictor variables given the class of Y as normally distributed random variables with the same SD and then use Bayes Theorem to predict the class of Y given values of the predictors.

Results of LDA

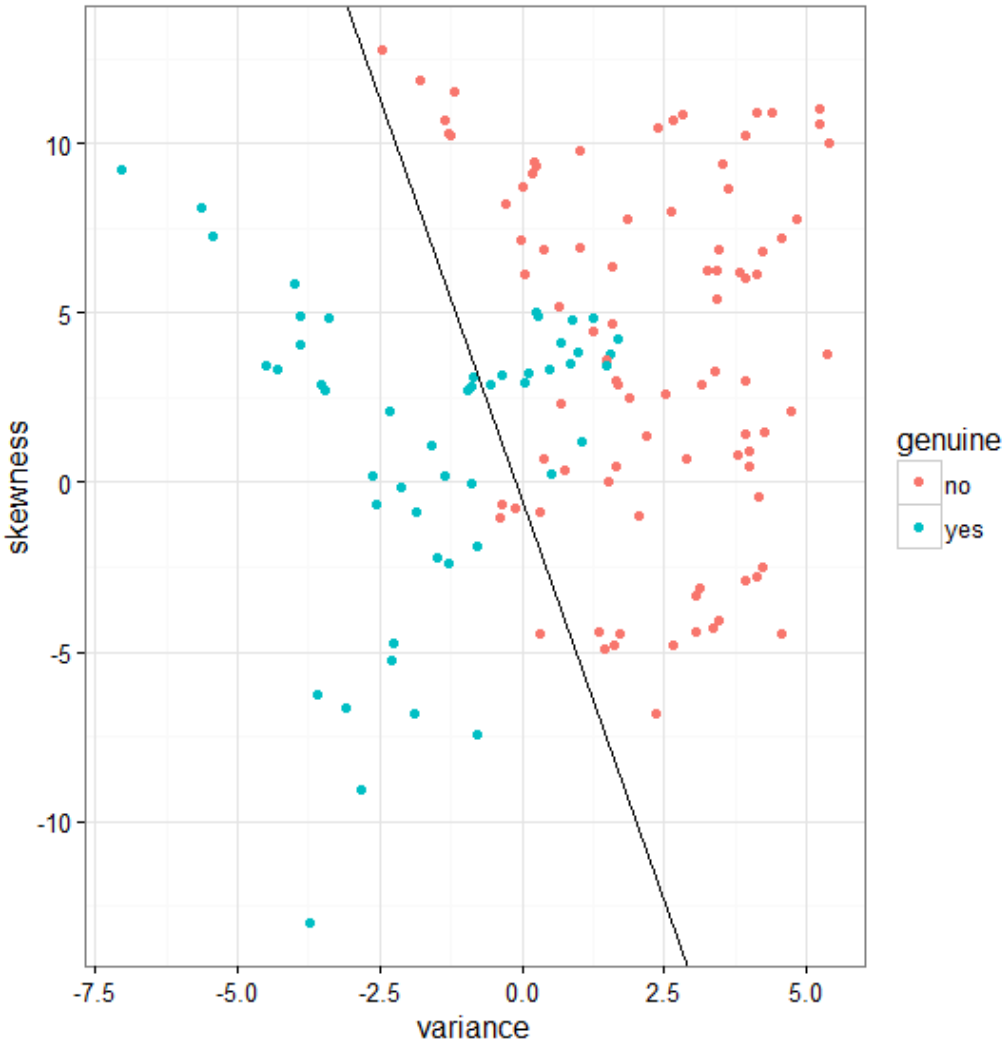
```
> model3 <- lda(formula = type ~ variance+skewness, data = ss)
> model3
Call:
lda(type ~ variance + skewness, data = ss)

Prior probabilities of groups:
        0         1
0.5465587 0.4534413

Group means:
   variance  skewness
0  2.280648  4.311509
1 -1.892714 -1.190899

Coefficients of linear discriminants:
                  LD1
variance -0.46505122
skewness -0.09733833
```



Why is this called ***Linear*** Discriminant Analysis?

Bills on one side of the black line are “real”, the others are “fake”

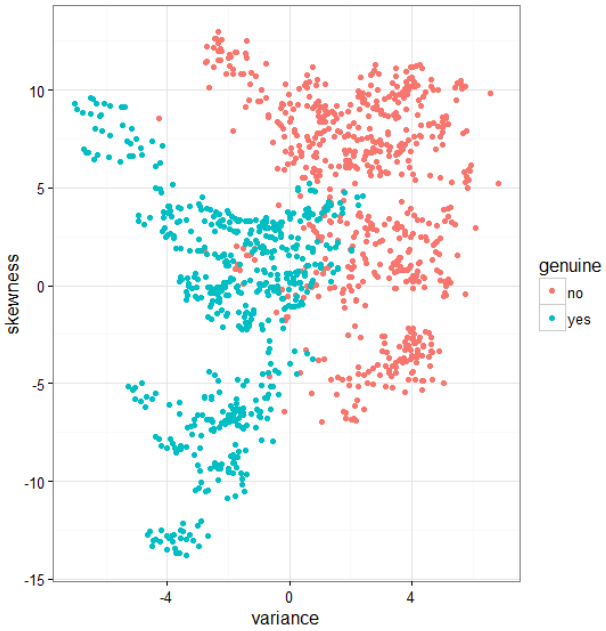
Error rate of 15.3%



Quadratic Discriminant Analysis (QDA)

What if we'd allowed a ***quadratic*** classification border instead of a linear one?

Error rate decreases to 14.6%...not much better in this case



Note: this corresponds to allowing unequal SD's in the normal distributions of the predictors.

So kNN with k=10 looks like the winner

We expect to misclassify 6.57% of bills.

But wait,

Is it equally bad to misclassify a genuine bill as it is to misclassify a fake bill?

Types of Misclassification:

False positive

True negative

Error Rates by Bill Type

	Overall Error Rate (%)	Misclassified Genuine Bills (%)	Misclassified Fake Bills(%)
Logistic Regression	16.6	17.2	16.0
kNNs, k=10	6.6	2.3	14.0
LDA	15.3	17.2	12.0
QDA	14.6	17.2	10.0

Error Rates by Bill Type

	Overall Error Rate (%)	Misclassified Genuine Bills (%)	Misclassified Fake Bills(%)
Logistic Regression	16.6	17.2	16.0
kNNs, k=10	6.6	2.3	14.0
LDA	15.3	17.2	12.0
QDA	14.6	17.2	10.0

QDA has the lowest error rate for fake bills!