Hw2 review

Validity of a test

- Valid:
  - p: The nominal p-value
  - P(p<= p0)=p0 for any constant p0 in (0,1)

- Suppose I have a valid test.

  t.test(rnorm(5), rnorm(5))

  What outcomes could I have?

  Could I see a p-value of 0.01?

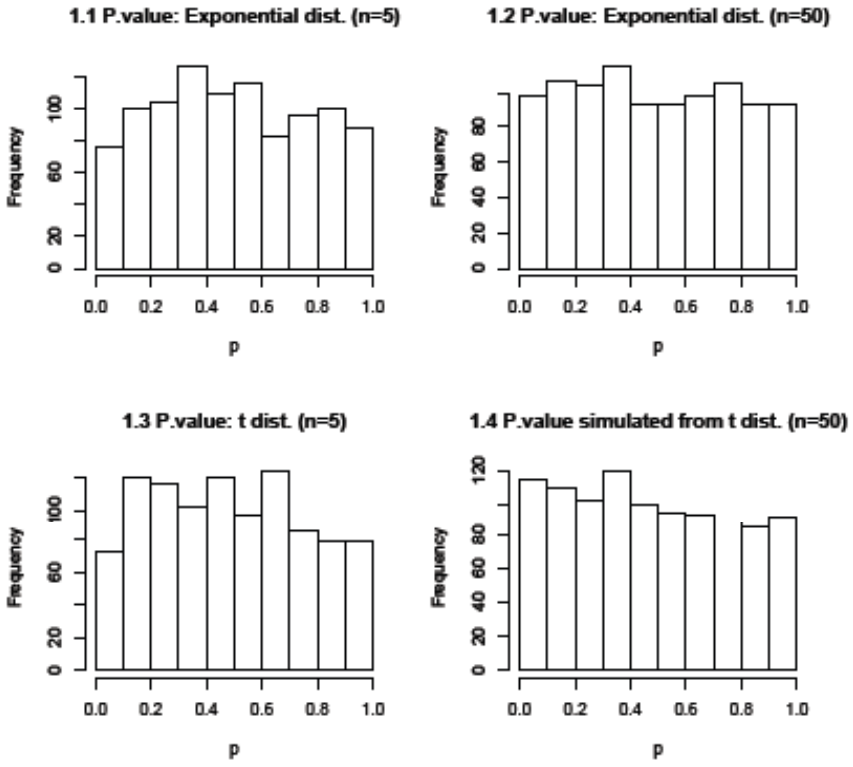
  Could I see a p-value of 0.8?

  Does it confirm that I have a valid test if I get a large p-value?

  Does it show that I have a bad, invalid test if I get a p-value that is less than 0.05?

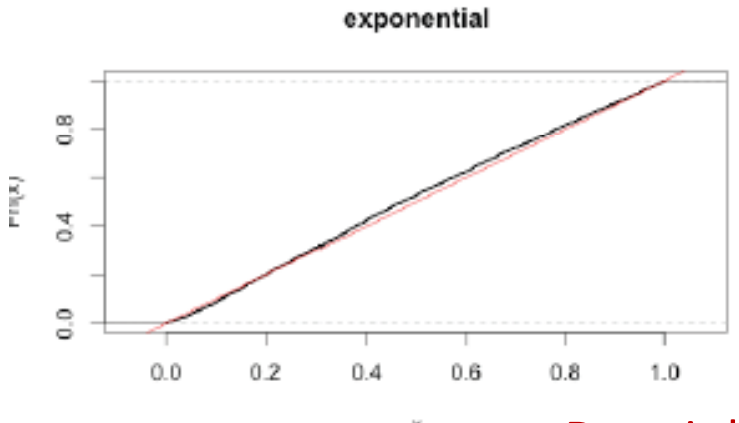
Validity of a test

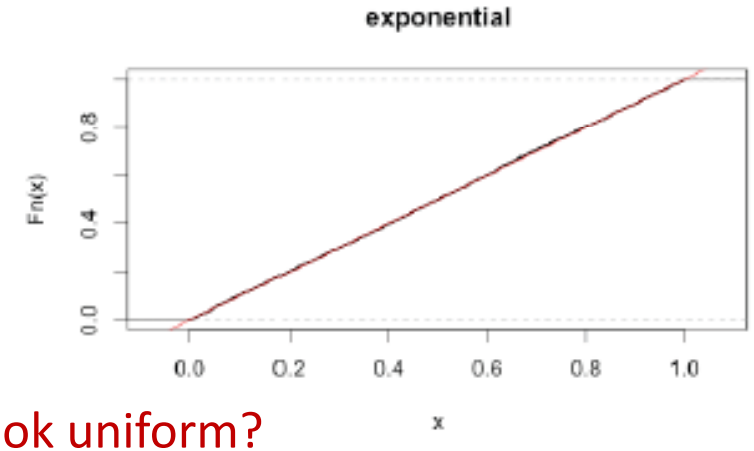
-- you have to do it more than once. Many times, in fact, to check the distribution.



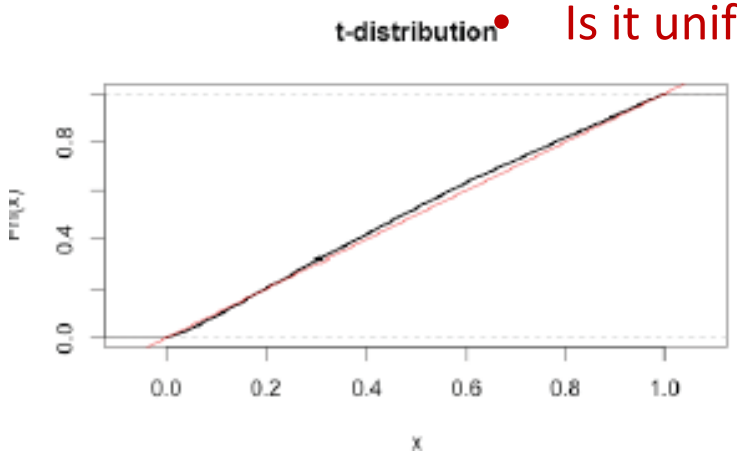
- Does it look uniform?
- Is it uniform?

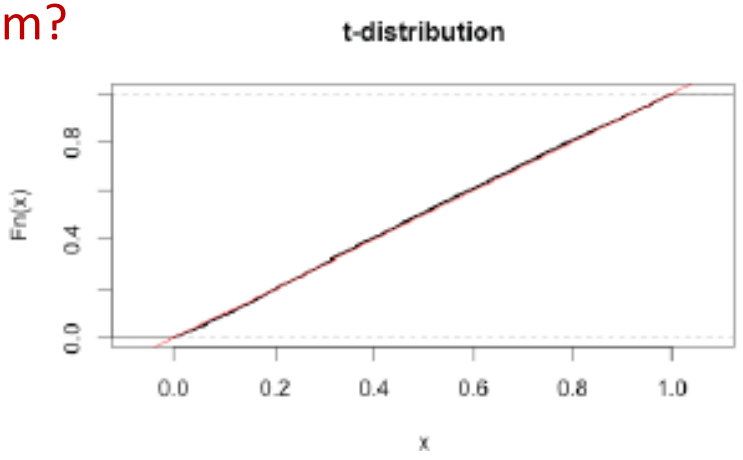
Validity of a test



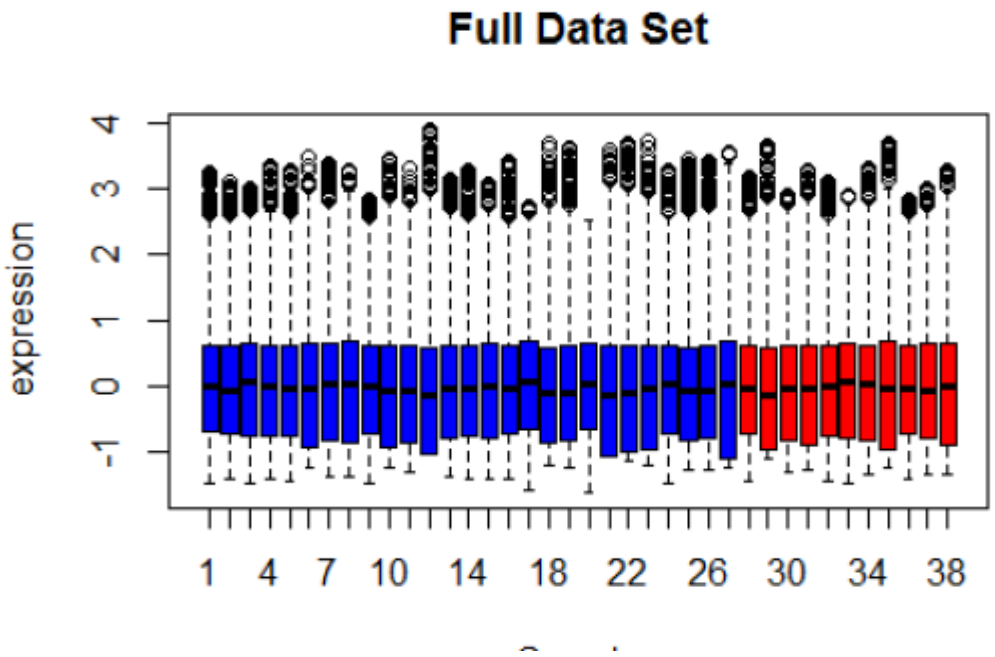


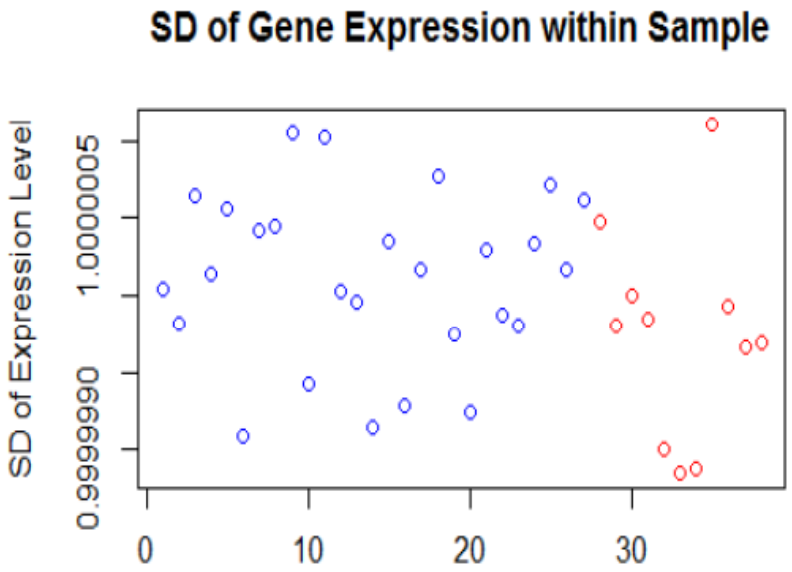
- Does it look uniform?
- Is it uniform?





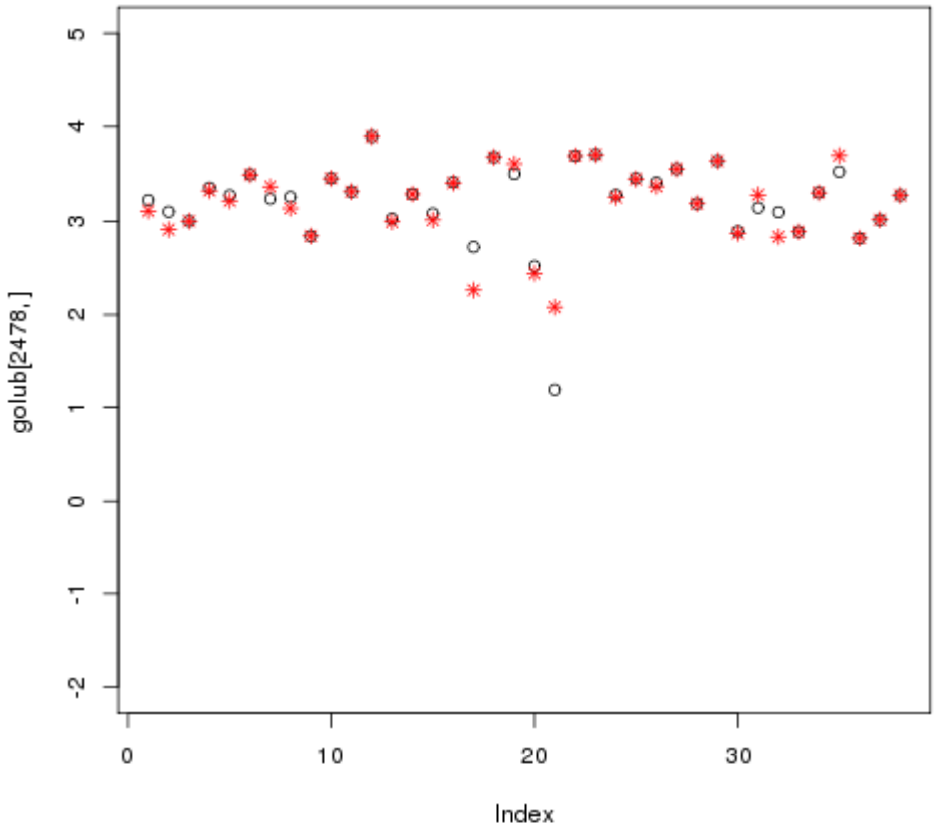
EDA





- Pay attention to pattern
- BUT don’t forget about the scale!

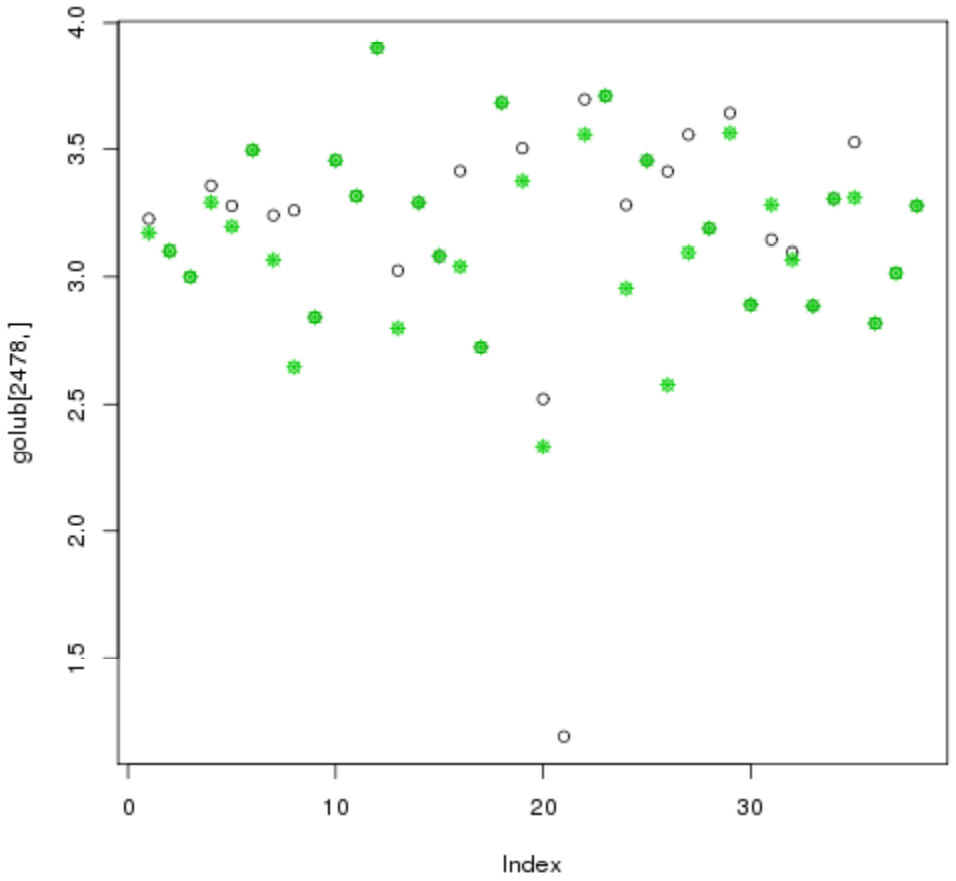
Gene with the highest expression




```
> plot(golub[2478,],ylim=c(-2,5))
>  points(golub[748,],col=2,pch=8)
> mean(golub[2478,])
[1] 3.203891
> mean(golub[748,])
[1] 3.201664
> golub.gnames[748,]
[1] "1707"
[2] "RPS14 gene (ribosomal protein S14) extracted
      from Human ribosomal protein S14 gene"
[3] "M13934_cds2_at"
> golub.gnames[2478,]
[1] "5729"
[2] "37 kD laminin receptor precursor/p40 ribosome
      associated protein gene"
[3] "U43901_rna1_s_at"
```

Gene with the highest expression

```
> apply(golub,2,which.max)
 [1]    9    7    7   10    9    7    4
    9    7    7    7    7   10    7    7
[16]   10    7    7  748   10  253    4
    4   10    7 1736   10    7    8    7
[31]    7    9    7    7  748    7    7
    7
> table(apply(golub,2,which.max))

   4    7    8    9   10  253  748 1736
   3   20    1    4    6    1    2    1
> plot(golub[2478,])
points(golub[7,],col=3,pch=8)

mean(golub[7,])
[1] 3.058682
> t.test(golub[7,],golub[748,])
        Welch Two Sample t-test
data:  golub[7, ] and golub[748, ]
t = -1.1231, df = 60.048, p-value = 0.2658
sample estimates:
mean of x mean of y
 3.058682  3.201664
```

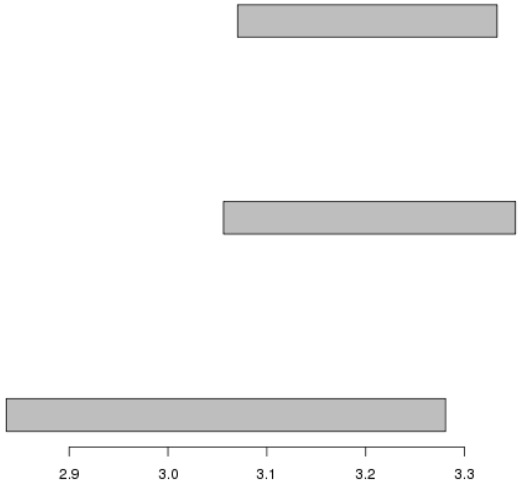


So who's right?

- Here's how statisticians differ from non-statistician programmers
  - Are these the same questions:
    - Which gene has the highest measured expression in these 38 samples?
    - Which gene has the highest expression in these 38 samples?
    - Which gene has the highest expression in these kind of cells, using the 38 samples as our evidence, assuming these are representative samples?

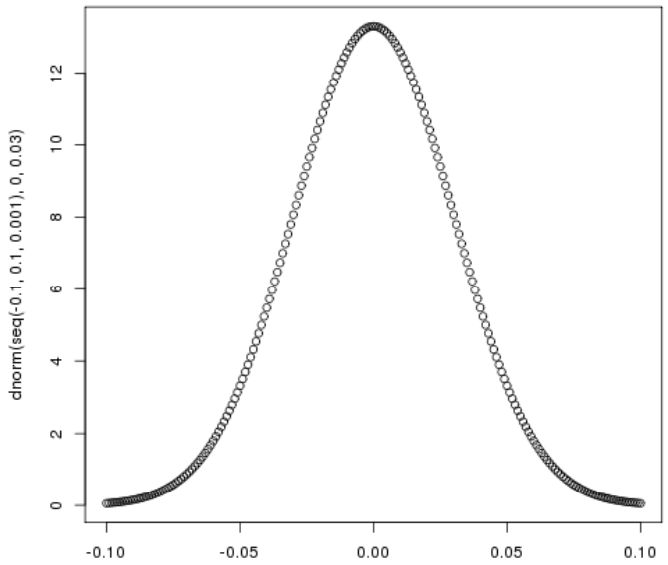
  Are some of these genes essentially tied?

```
> confint(lm(golub[7,]~1))
                2.5 %   97.5 %
(Intercept) 2.836638 3.280726
> confint(lm(golub[2478,]~1))
               2.5 %   97.5 %
(Intercept) 3.05621 3.351573
> confint(lm(golub[748,]~1))

                2.5 %   97.5 %
(Intercept) 3.070393 3.332936
```



Without getting into detailed modeling and testing, lets consider

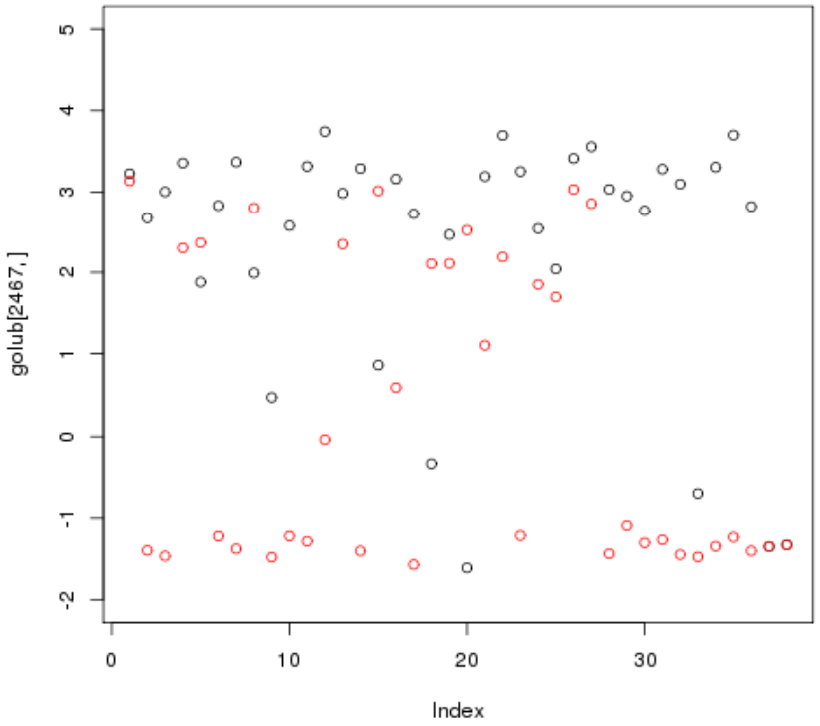
- Lowest SD ~ 0.19
- 0.19/sqrt(38) ~= 0.03

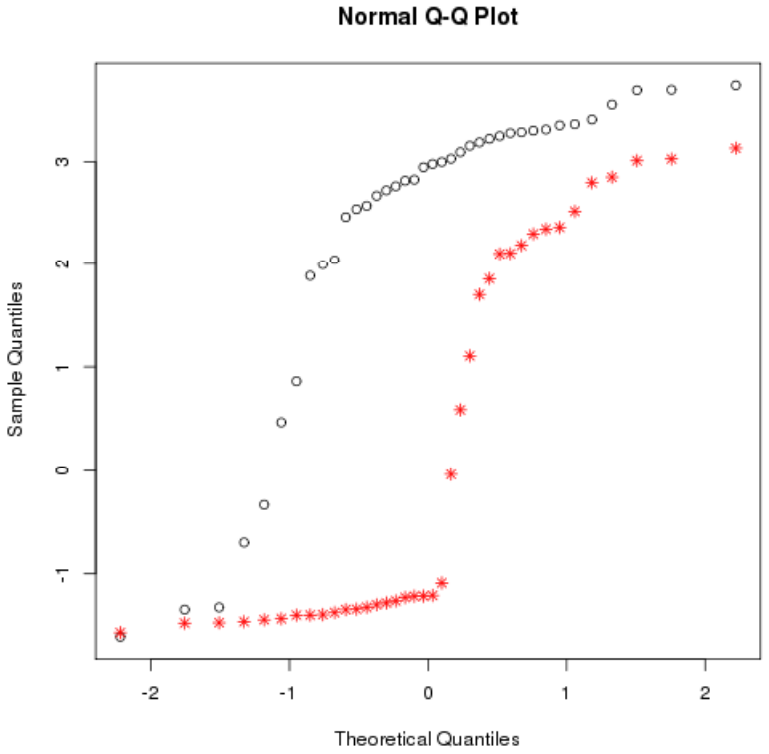


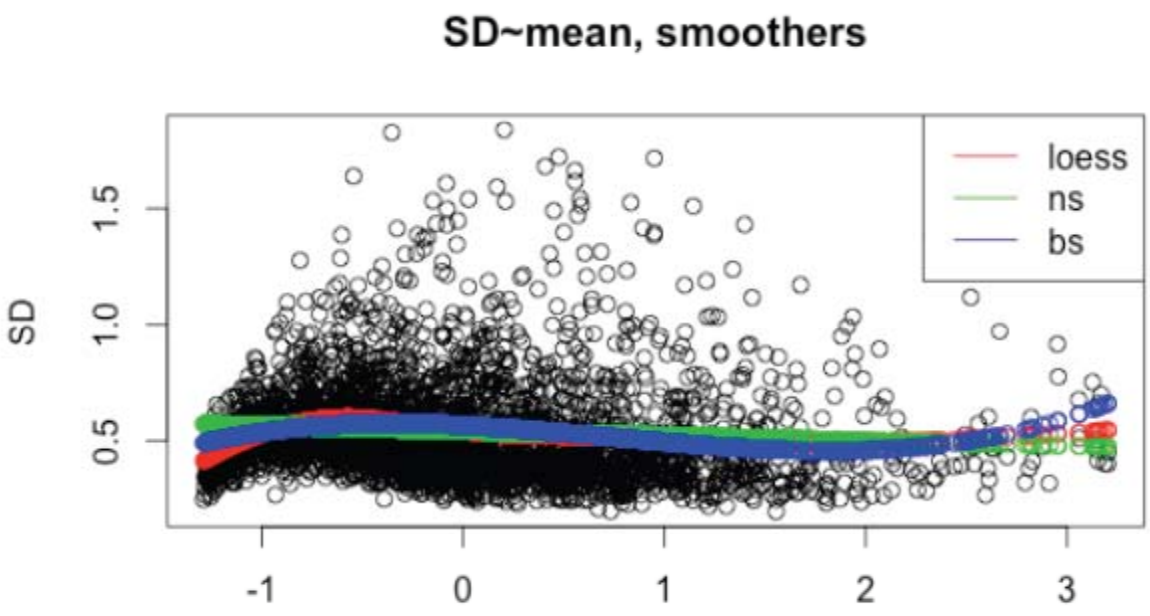
Consider a stable gene whose expression is measured in a random sample of 38 individuals. The sampling distribution of the "average expression from 38 samples" would have standard deviation of 0.03. (The standard error).

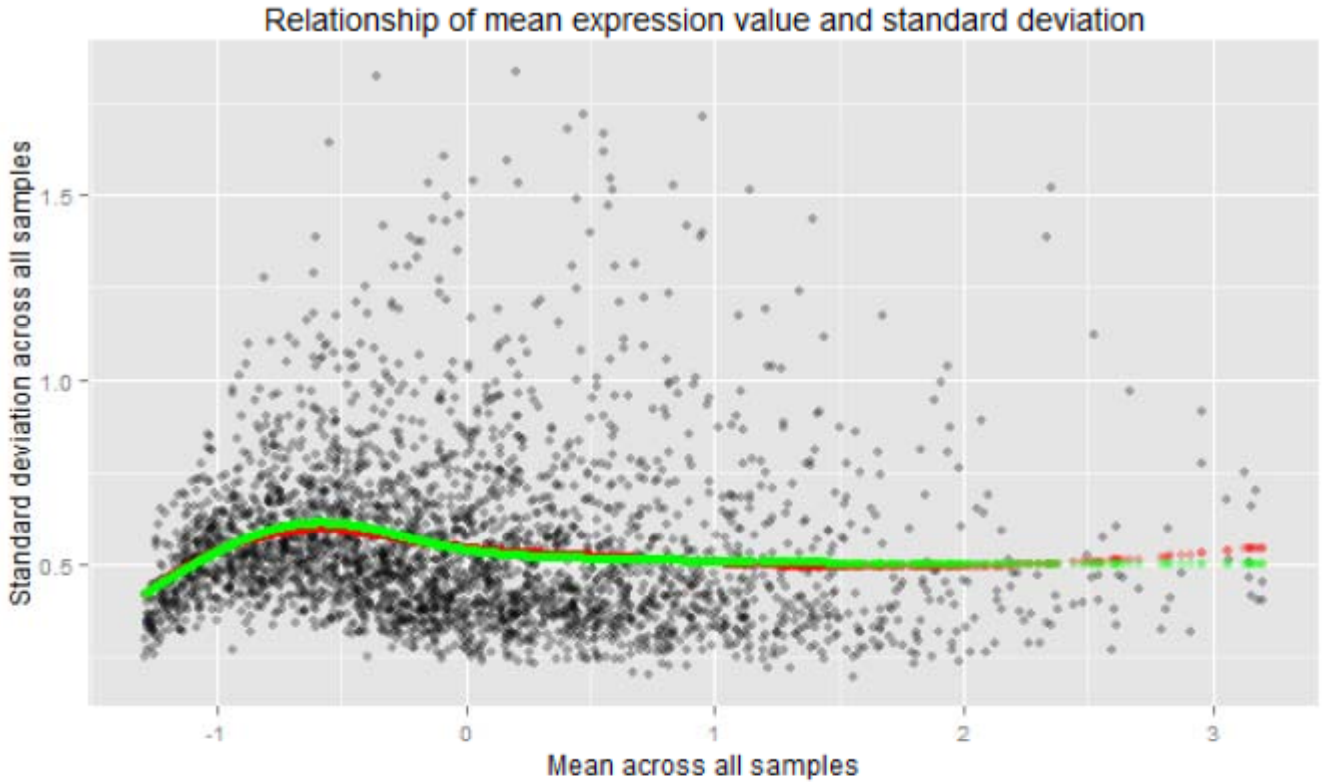
Gene that appears to vary most

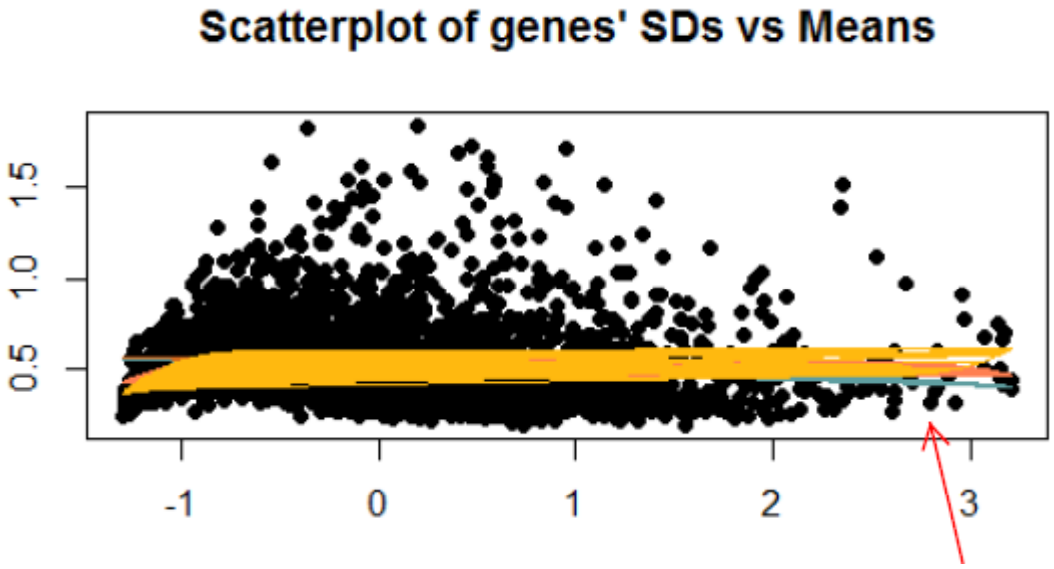
```
plot(golub[2467,],ylim=c(-2,5))
points(golub[2065,],col=2)
qqnorm(golub[2467,])
points(qqnorm(golub[2065,],plot.it=F),col=2,pch=8)
```

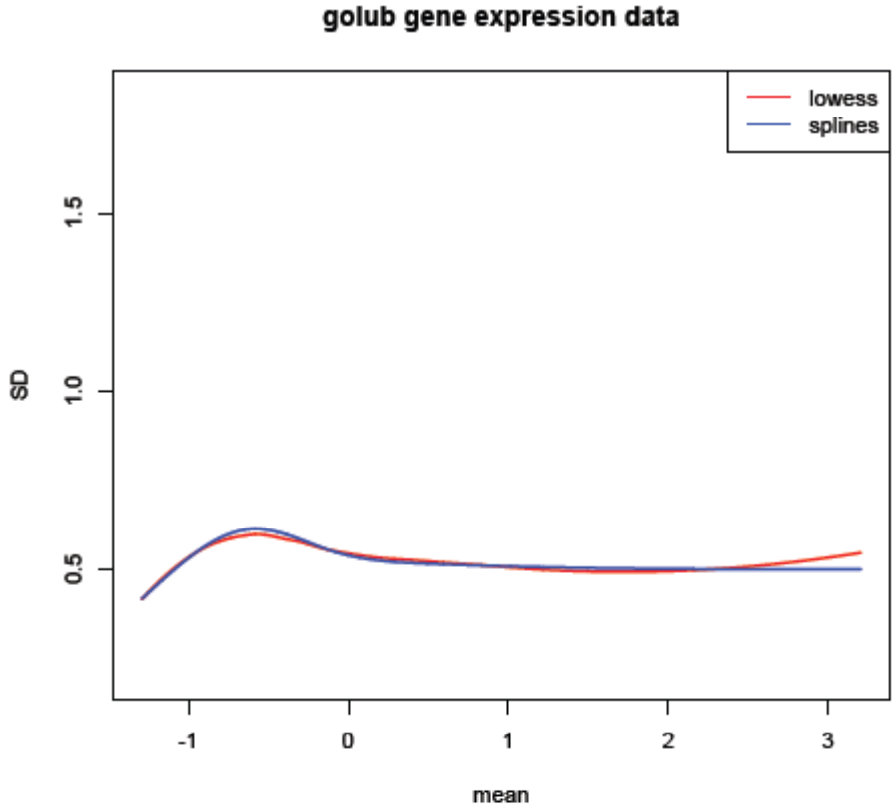




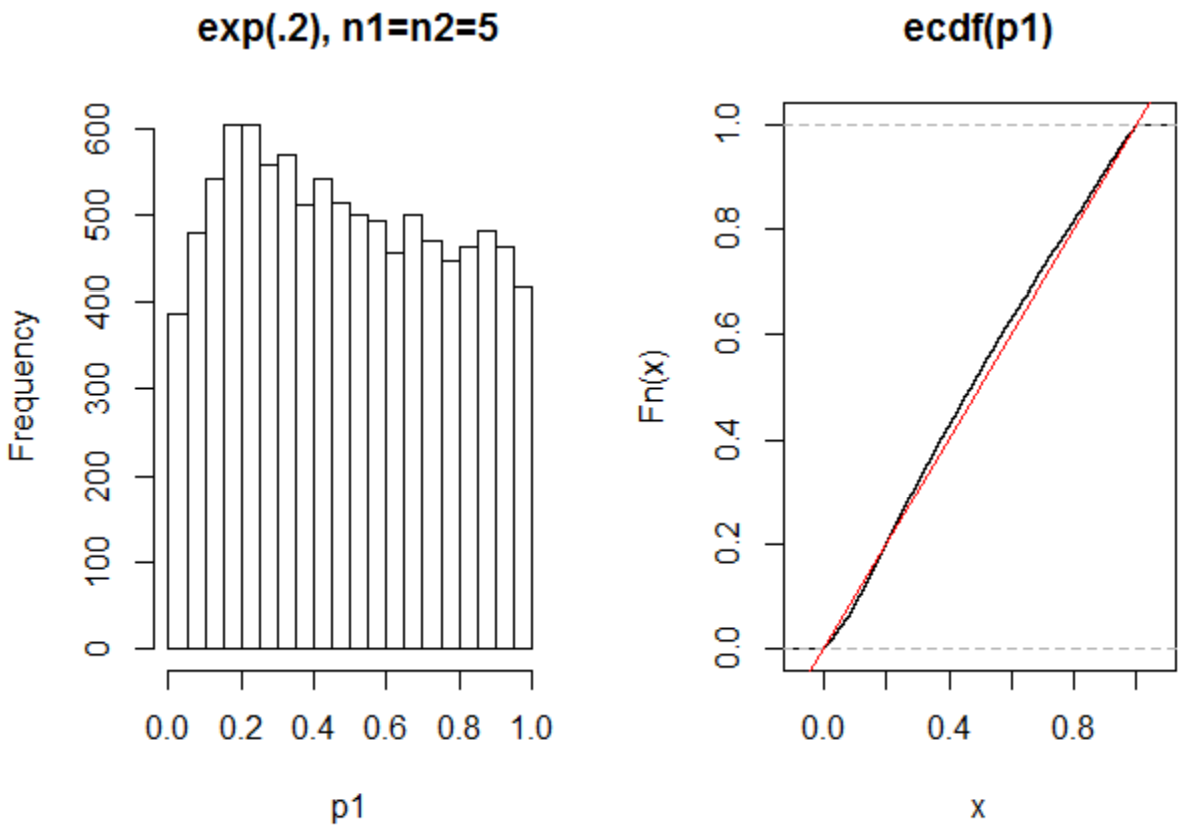

SD~mean, smoothers
SD
0.5
1.0
1.5
-1
0
1
2
3
loess
ns
bs


Relationship of mean expression value and standard deviation
Standard deviation across all samples
1.5
1.0
0.5
-1
0
1
2
3
Mean across all samples

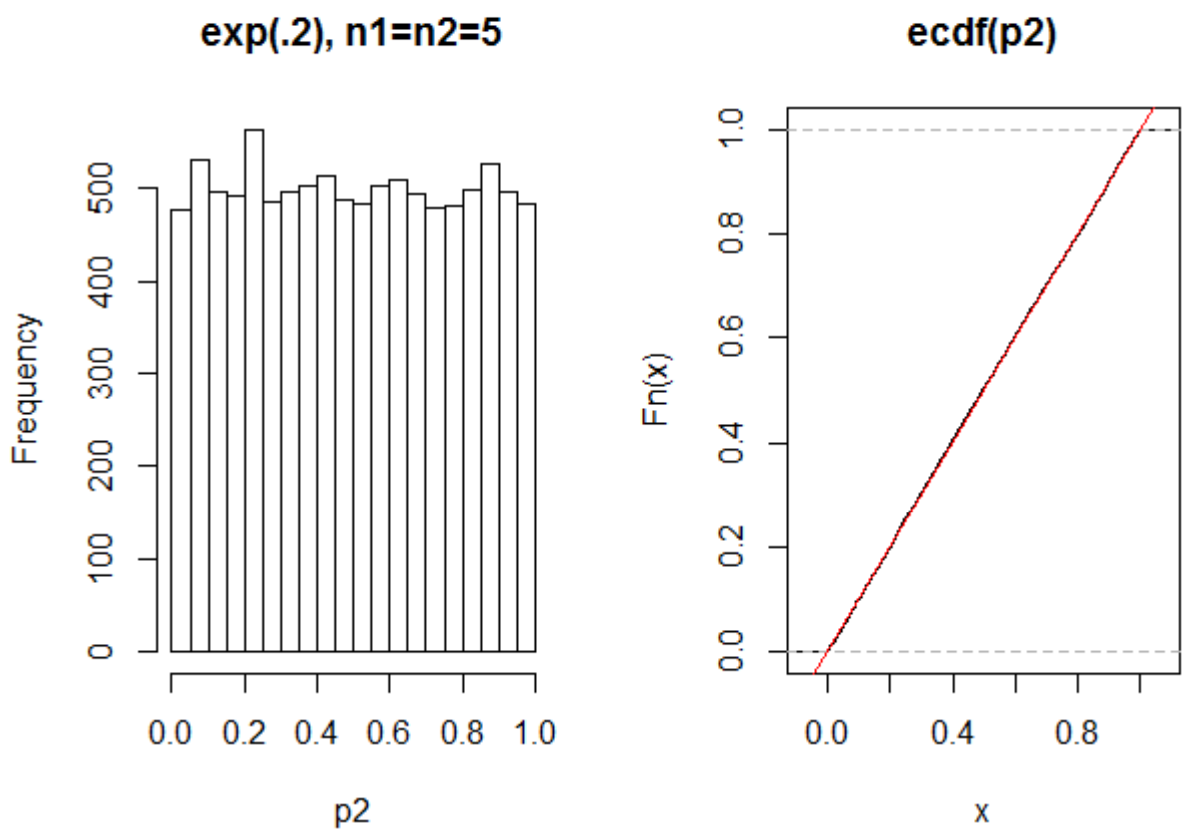

Scatterplot of genes' SDs vs Means
0.5
1.0
1.5
-1
0
1
2
3


golub gene expression data
lowess
splines
SD
0.5
1.0
1.5
-1
0
1
2
3
mean

```
set.seed(2014)
n1 =n2=5
X <- matrix(rexp(10000*(n1+n2),.2), ncol=n1+n2)
dim(X)
myt <-function(x) {
    t.test(x[1:n1], x[(n1+1):(n1+n2)], var.equal = TRUE)$p.value}
p1 <- apply(X,1,myt)
par(mfrow=c(1,2))
hist(p1,main = "exp(.2), n1=n2=5")
plot(ecdf(p1))
abline(0,1,col=2)
ks.test(p1,"punif")
        One-sample Kolmogorov-Smirnov test
data:  p1
D = 0.0333, p-value = 4.455e-10
alternative hypothesis: two-sided
```


exp(.2), n1=n2=5
Frequency
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0
p1
ecdf(p1)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.4
0.8
x

```
set.seed(2014)
n1 =n2=50
X <- matrix(rexp(10000*(n1+n2),.2), ncol=n1+n2)
dim(X)
myt <-function(x) {
    t.test(x[1:n1], x[(n1+1):(n1+n2)], var.equal = TRUE)$p.value}
p2 <- apply(X,1,myt)
hist(p2,main = "exp(.2), n1=n2=5")
plot(ecdf(p2))
abline(0,1,col=2)
ks.test(p2,"punif")

        One-sample Kolmogorov-Smirnov test

data:  p2
D = 0.0059, p-value = 0.8757
alternative hypothesis: two-sided
```


exp(.2), n1=n2=5
Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
p2
ecdf(p2)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.4
0.8
x

Kolmogorov–Smirnov test

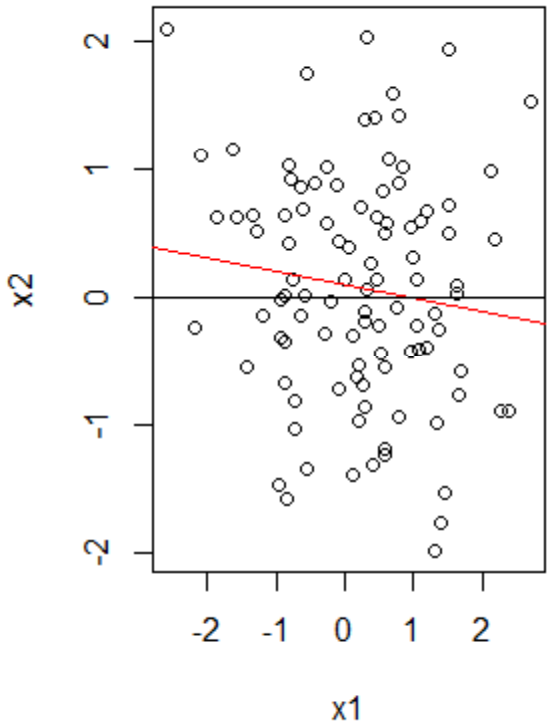
- Tests for the distribution of two random variables
  - Null: the two r.v. have the same marginal distribution
- Does not test whether the two r.v. are independent

```
>x1=rnorm(100)
>x2=rnorm(100)
> cor(x1,x2)
[1] -0.124407
> plot(x1,x2)
> ks.test(x1,x2)
        Two-sample Kolmogorov-Smirnov test
data:  x1 and x2
D = 0.14, p-value = 0.281 : The distribution of x1 and x2
   are not significantly different.
alternative hypothesis: two-sided
>summary(lm(x1~x2))
…
F-statistic: 1.541 on 1 and 98 : DF,  p-value: 0.2175: x1
   and x2 are not linearly dependent
```

The two different tests

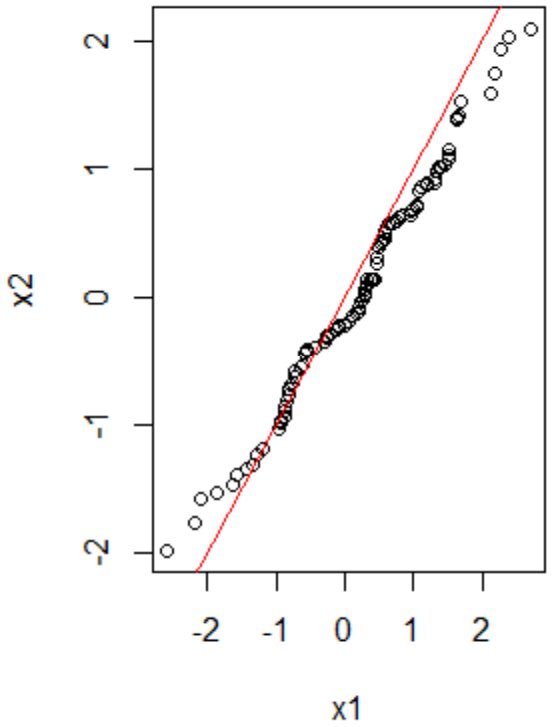
```
plot(x1,x2)
abline(lm(x2~x1),col=2)
abline(h=0)
```

- testing whether the straight line has slope 0 or not
- Visually: does it look random or correlated?



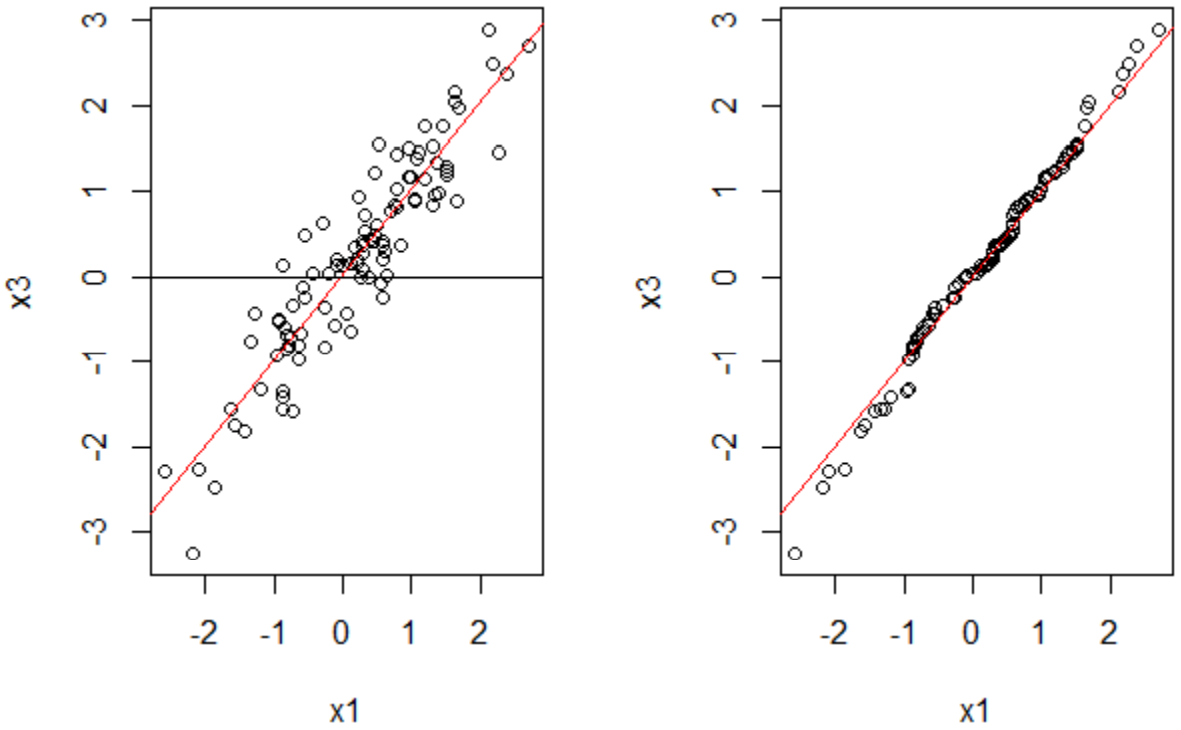

```
qqplot(x1,x2)
 abline(0,1,col=2)
```

- Test whether the distributions are the same
- Visually: whether the sorted data fall on identity line
-



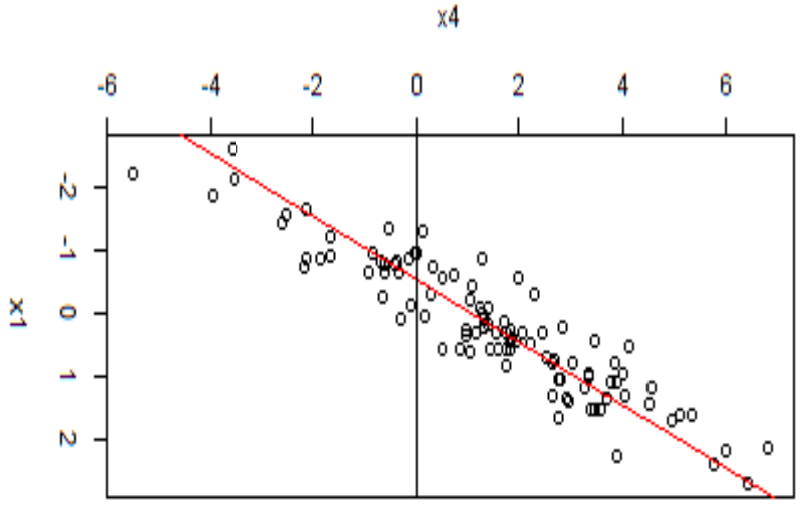
Closely correlated, and share the same distribution:

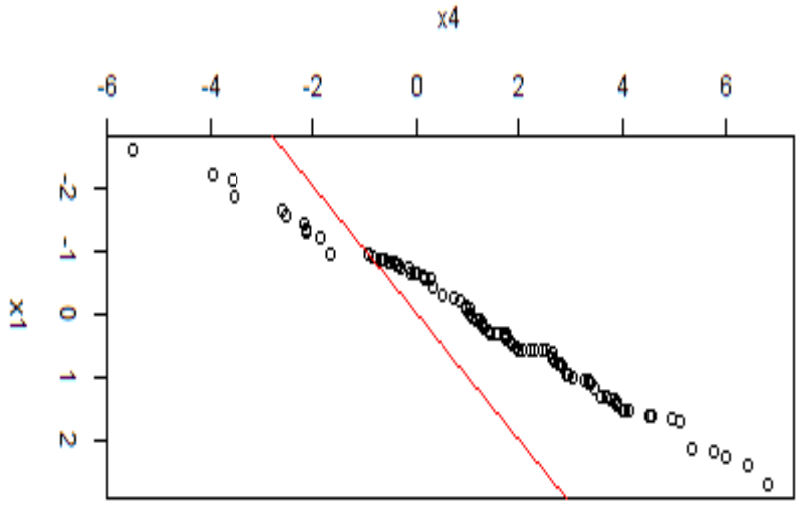
```
> x3=x1*.9+rnorm(100,0,sqrt(1-0.9^2))
> plot(x1,x3)
> qqplot(x1,x3)
> abline(0,1)
> ks.test(x1,x3)

        Two-sample Kolmogorov-Smirnov test

data:  x1 and x3
D = 0.06, p-value = 0.9938
alternative hypothesis: two-sided
>summary(lm(x3~x1))
….
F-statistic: 572.8 on 1 and 98 DF,  p-value: < 2.2e-16
>plot(x1,x2)
```

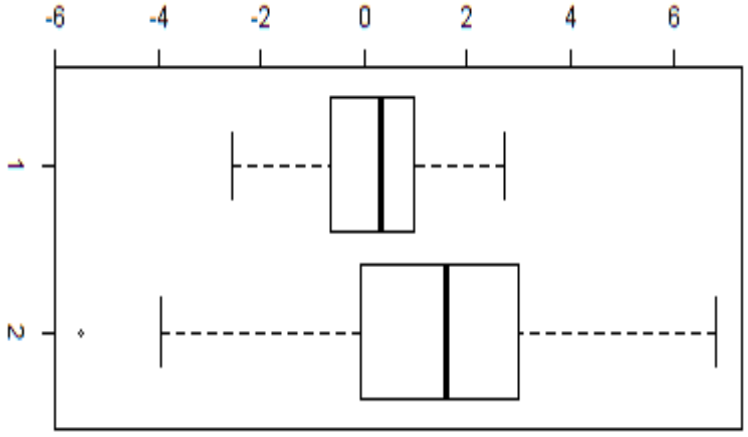

x3
x1
-3
-2
-1
0
1
2
3
-2
-1
0
1
2

Closely correlated, but do not have the same distribution:

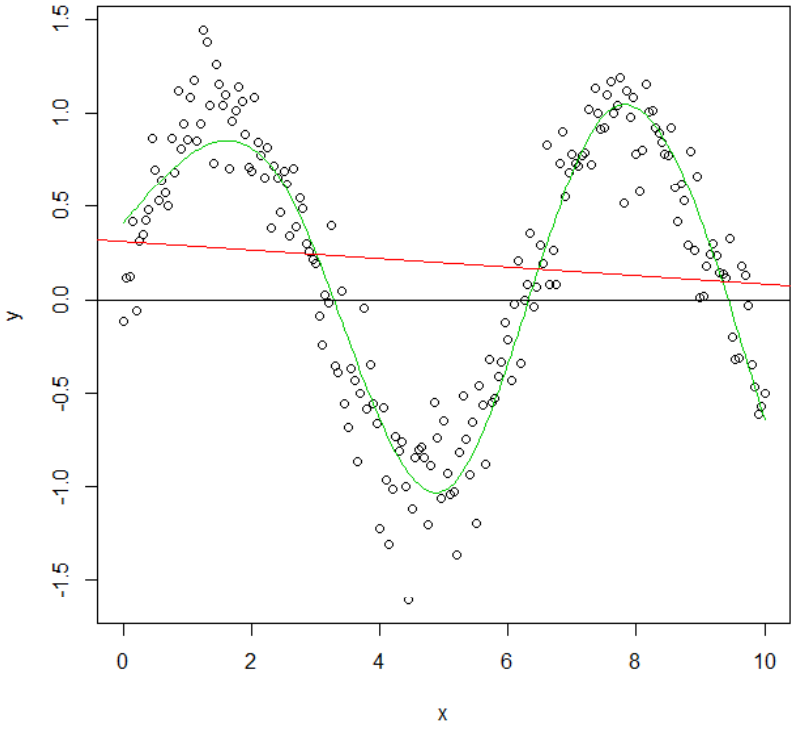
```
x4=x3*2+1
par(mfrow=c(1,3))
plot(x1,x4)
abline(lm(x4~x1),col=2)
abline(h=0)
summary(lm(x4~x1))
…
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.02294    0.09208   11.11   <2e-16 ***
x1           2.01758    0.08430   23.93   <2e-16 ***

qqplot(x1,x4)
abline(0,1,col=2)
boxplot(x1,x4)
ks.test(x1,x4)

        Two-sample Kolmogorov-Smirnov test

data:  x1 and x4
D = 0.44, p-value = 7.818e-09
alternative hypothesis: two-sided
```


x4
-6
-4
-2
0
2
4
6
x1
-2
-1
0
1
2


x4
-6
-4
-2
0
2
4
6
x1
-2
-1
0
1
2

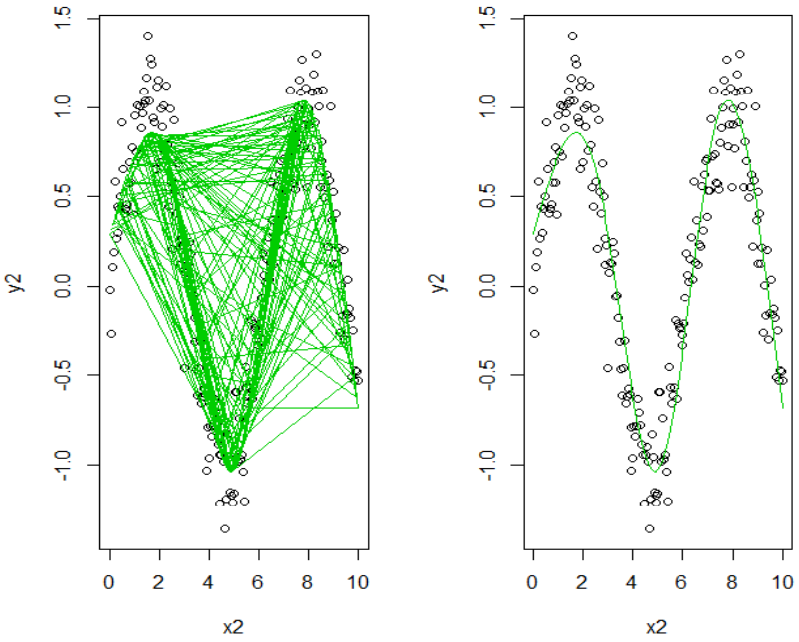

-6
-4
-2
0
2
4
6
1
2

Non-linear dependency




```
> set.seed(2014)
> x=seq(0,10,.05)
> y=sin(x)+rnorm(length(x),0,.2)
> plot(x,y)
summary(lm(y~x))
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.31126    0.09891   3.147   0.0019 **
x           -0.02239    0.01711  -1.308   0.1923
---
F-statistic: 1.712 on 1 and 199 DF,  p-value: 0.1923
> smooth1=lm(y~ns(x,4))
> lines(x,smooth1$fitted,col=3)
> summary(smooth1)
….
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.41024    0.06638   6.180 3.64e-09 ***
ns(x, 4)1   -2.86904    0.08347 -34.374  < 2e-16 ***
ns(x, 4)2    1.56181    0.08331  18.747  < 2e-16 ***
ns(x, 4)3    0.78939    0.16929   4.663 5.76e-06 ***
ns(x, 4)4   -1.63467    0.07856 -20.807  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
    ‘ ’ 1


Residual standard error: 0.2371 on 196 degrees of
    freedom
Multiple R-squared: 0.8892,     Adjusted R-squared:
    0.8869
F-statistic: 393.2 on 4 and 196 DF,  p-value: < 2.2e-16
```

Issue with making lines in R plots




```
par(mfrow=c(1,2))
> plot(x2,y2)
> lines(x2,smooth2$fitted,col=3)
> plot(x2,y2)
>lines(cbind(x2,smooth2$fitted)[order(x2),],col=3)
>
```