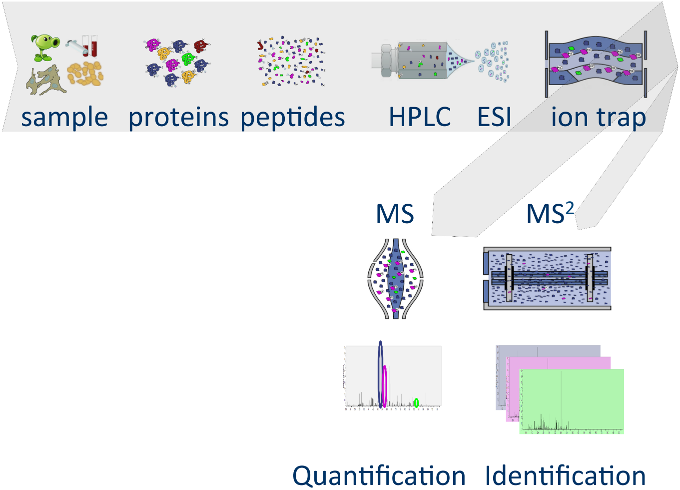
Statistical Methods for Quantitative MS-Based Proteomics:
1. Identification

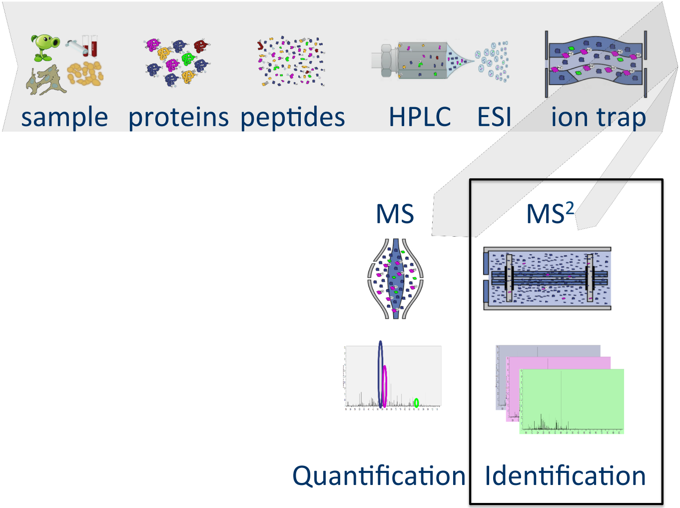
Lieven Clement

Statistics and Genomics Seminar, UCBerkeley, California

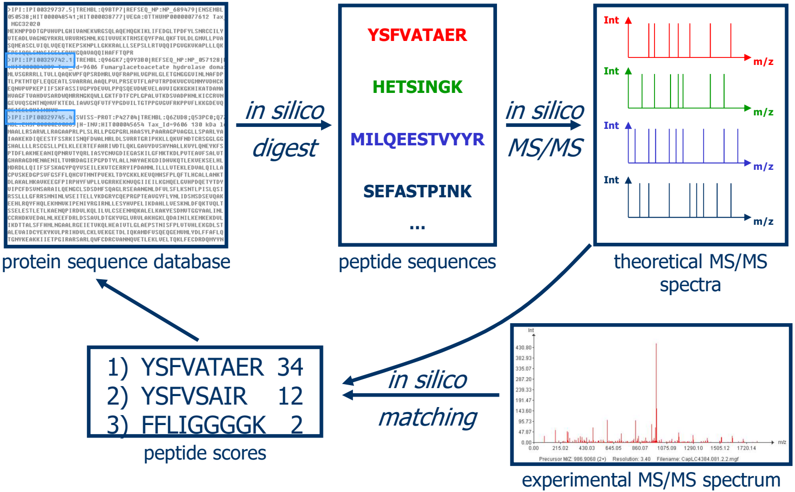
Challenges in Label Free MS-based Quatitative proteomics



Challenges in Label Free MS-based Quatitative proteomics

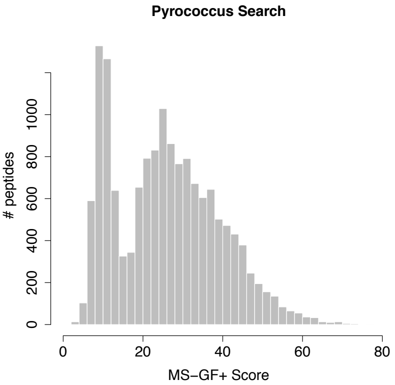


Identification

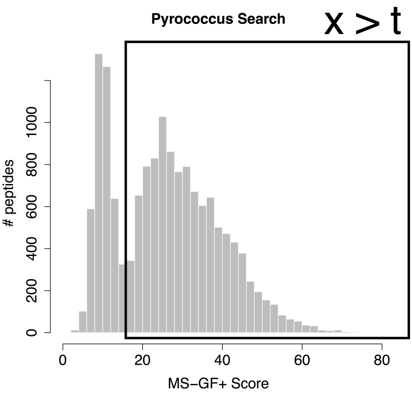


(slide courtesy to Lennart Martens)

Search engines return score that discriminates good from bad matches

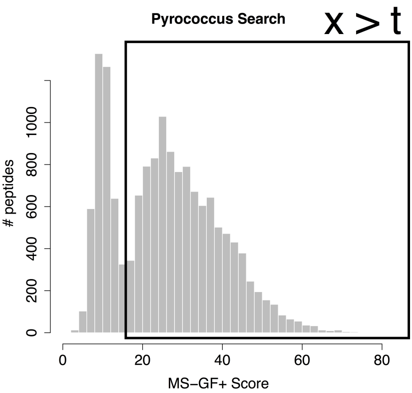


Search engines return score that discriminates good from bad matches



Score threshold t?

Search engines return score that discriminates good from bad matches

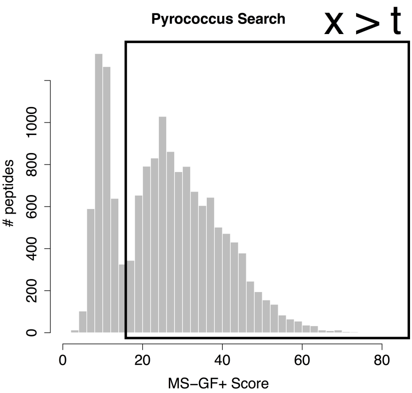


Score threshold t?

$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x)$

$\text{FDR}(t) = Pr\left[FP|x \geq t\right]$

Search engines return score that discriminates good from bad matches



Score threshold t?

$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x)$$

$$\mathrm{FDR}(t) = Pr\left[FP | x \geq t\right]$$

$$\int_{-\infty}^{x=t} f(x) = \pi_0 \int_{-\infty}^{x=t} f_0(x) + (1-\pi_0) \int_{-\infty}^{x=t} f_1(x)$$

$$F(x) = \pi_0 F_0(t) + (1-\pi_0) F_1(t)$$

$$\mathrm{FDR}(t) = \frac{\pi_0[1-F_0(t)]}{1-F(t)}$$

Link to Benjamini Hochberg FDR

- Bayesian FDR

$$\text{FDR}(t) = \frac{\pi_0\left[1 - F_0(t)\right]}{1 - F(t)}$$

- BH for one sided test

$$FDR(t) = \frac{mp(t)}{\#t_i \geq t}$$

with

- m the number of tests
- $p(t)$ the p-value corresponding to a test statistic with value t

Link to Benjamini Hochberg FDR

- Bayesian FDR

$$\text{FDR}(t) = \frac{\pi_0 \left[1 - F_0(t)\right]}{1 - F(t)}$$

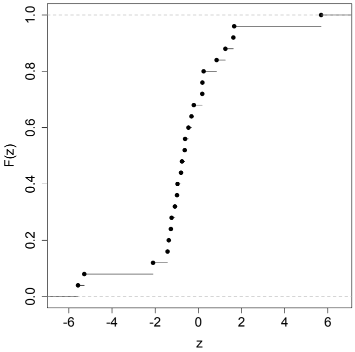
- BH for one sided test

$$FDR(t) = \frac{mp(t)}{\#t_i \geq t} = \frac{1 - F_0(t)}{\frac{\#t_i \geq t}{m}} = \frac{1 - F_0(t)}{1 - F(t)}$$

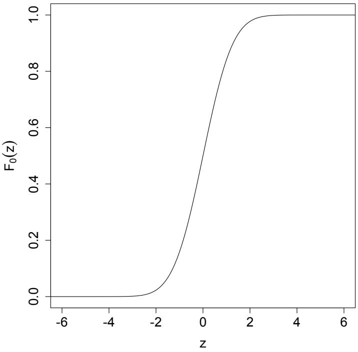
  - Use theoretical distribution for $p(t) = 1 - F_0(t)$
  - Conservative estimate $\pi_0 = 1$
  - ECDF: $1 - F(t) = \frac{\#t_i \geq t}{m}$

- Benjamini Hochberg 1995:

$F(t)$ using the ECDF

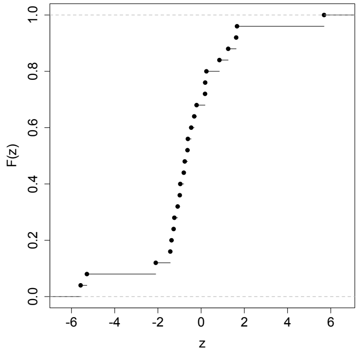


$F_0(t) theoreticalCDF$

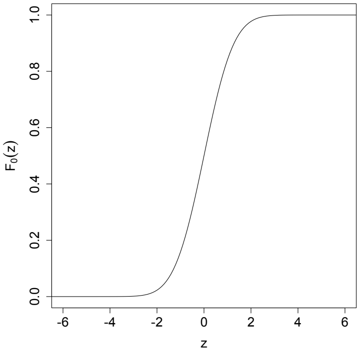


- Benjamini Hochberg 1995:

$F(t)$ using the ECDF

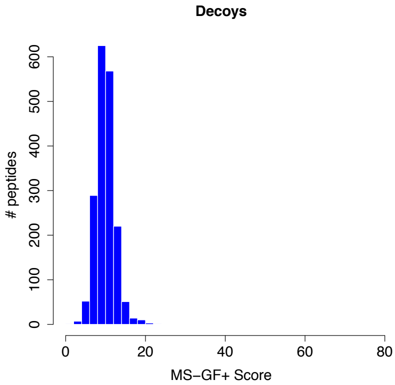


$F_0(t) theoretical CDF$



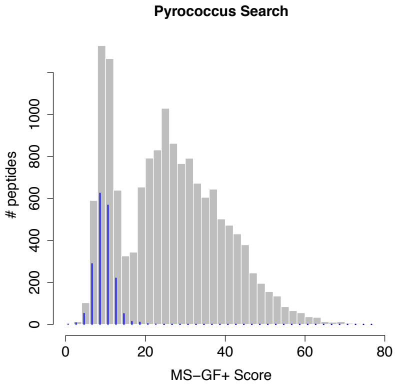
- How to define $F_0(t)$ in proteomics?

Target-Decoy approach to establish null distribution



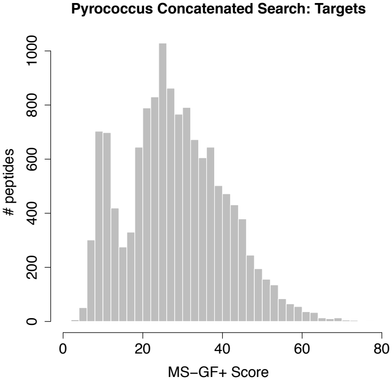
- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice

Target-Decoy approach to establish null distribution



- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice
- Concatenated search

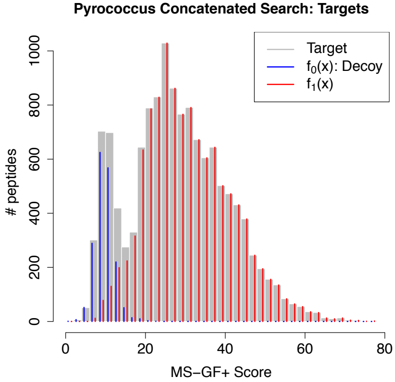
Target-Decoy approach to establish null distribution



- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice
- Concatenated search
- Assumption that bad hits have an equal probability to map on forward (target) and reverse database (decoy)

$$\hat{\pi}_0 = \frac{\#decoys}{\#targets}$$

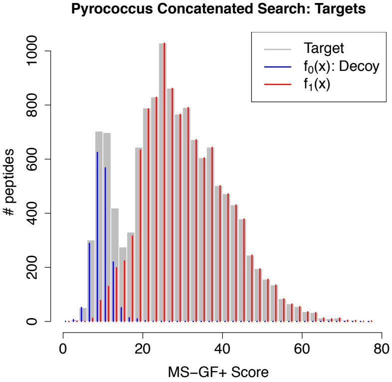
Target-Decoy approach to establish null distribution



- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice
- Concatenated search
- Assumption that bad hits have an equal probability to map on forward (target) and reverse database (decoy)

$$\hat{\pi}_0 = \frac{\#decoys}{\#targets}$$

Target-Decoy approach to establish null distribution



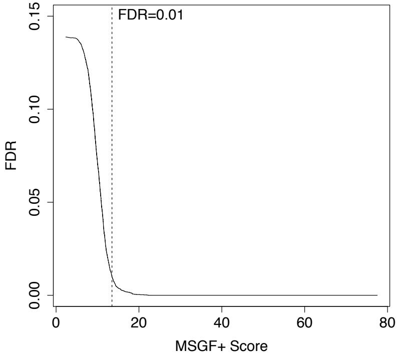
- Score cuttoff?
- Competitive Target - decoy:

$$\widehat{\text{FDR}}(x) = \frac{\#decoys|X \geq x}{\#targets|X \geq x}$$

$$\widehat{\text{FDR}}(x) = \frac{\#decoys}{\#targets}\frac{\frac{\#decoys|X\geq x}{\#decoys}}{\frac{\#targets|X\geq x}{\#targets}}$$

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{1-\bar{F}_0(x)}{1-\bar{F}(x)}$$

Target-Decoy approach to establish null distribution



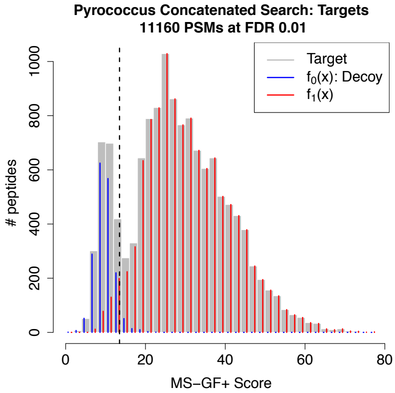
- Score cuttoff?
- Competitive Target - decoy:

$$\widehat{\text{FDR}}(x) = \frac{\#decoys|X \geq x}{\#targets|X \geq x}$$

$$\widehat{\text{FDR}}(x) = \frac{\#decoys}{\#targets}\frac{\frac{\#decoys|X\geq x}{\#decoys}}{\frac{\#targets|X\geq x}{\#targets}}$$

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{1-\bar{F}_0(x)}{1-\bar{F}(x)}$$

Target-Decoy approach to establish null distribution



- Score cuttoff?
- Competitive Target - decoy:

$$\widehat{\text{FDR}}(x) = \frac{\#decoys|X \geq x}{\#targets|X \geq x}$$

$$\widehat{\text{FDR}}(x) = \frac{\#decoys}{\#targets}\frac{\frac{\#decoys|X\geq x}{\#decoys}}{\frac{\#targets|X\geq x}{\#targets}}$$

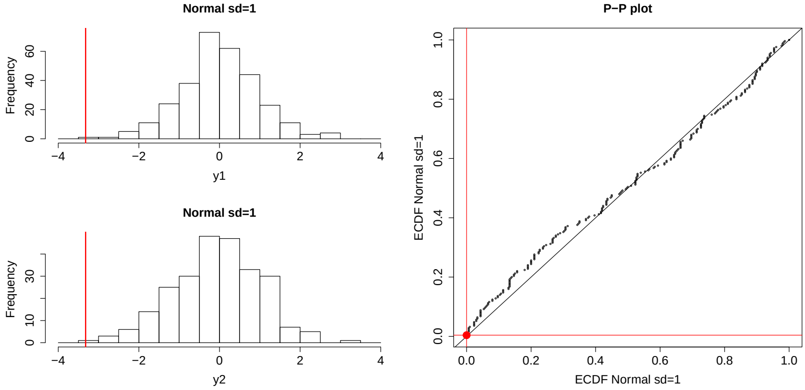
$$\widehat{\text{FDR}}(x) = \hat{\pi}_0\frac{1-\bar{F}_0(x)}{1-\bar{F}(x)}$$

We have to evaluate that

- The decoys are good simulations of the targets: compare $\bar{F}_0(x)$ with $\bar{F}(x)$
- $\hat{\pi}_0 = \frac{\#decoys}{\#targets}$ are a good estimator for π_0.
- We will use Probability-Probability-plots for this purpose.
- They plot the ECDFs from two samples in function of each other.

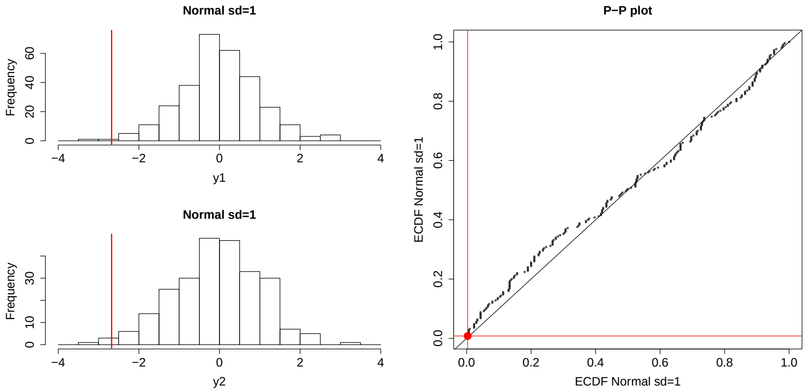
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



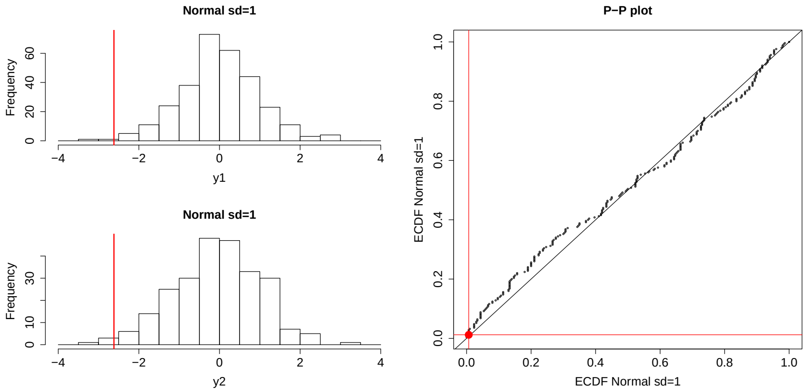
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



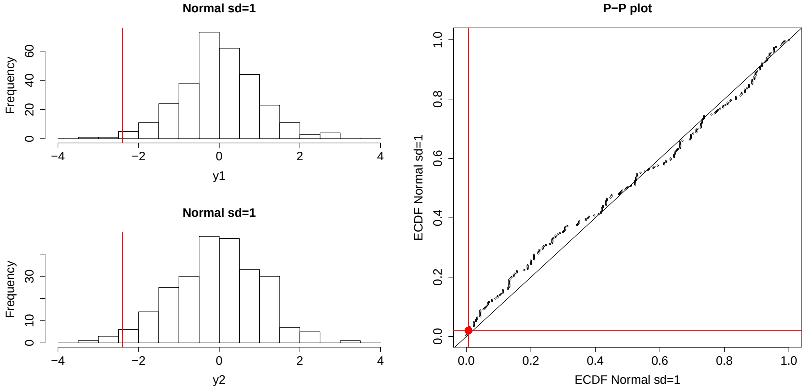
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



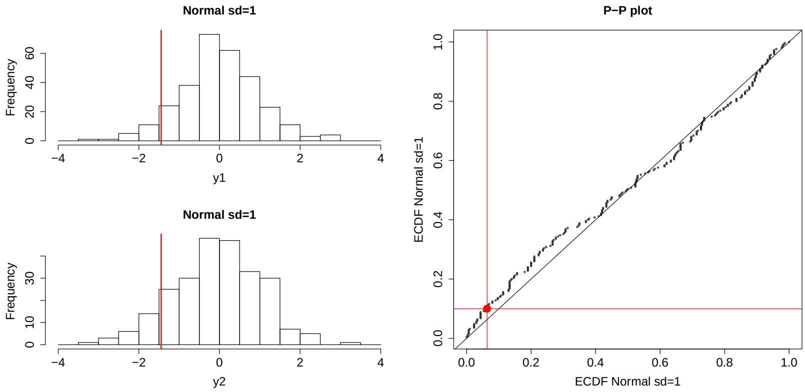
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



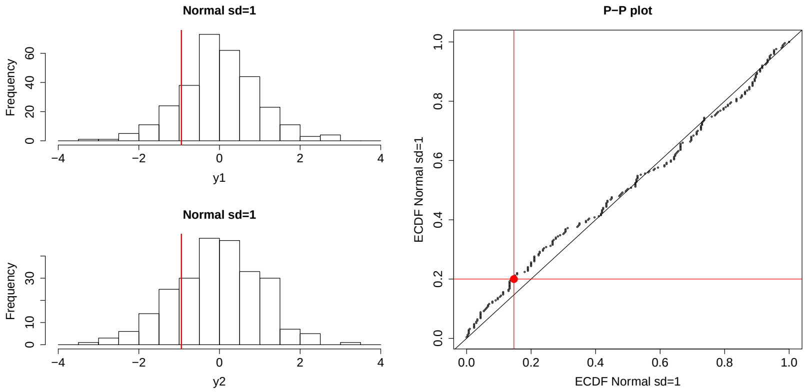
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



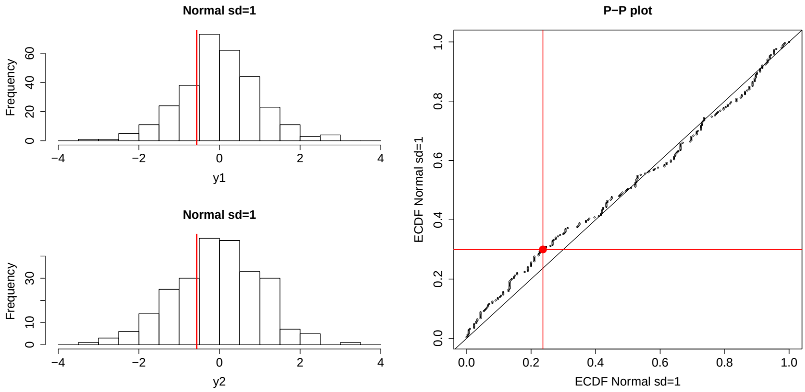
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



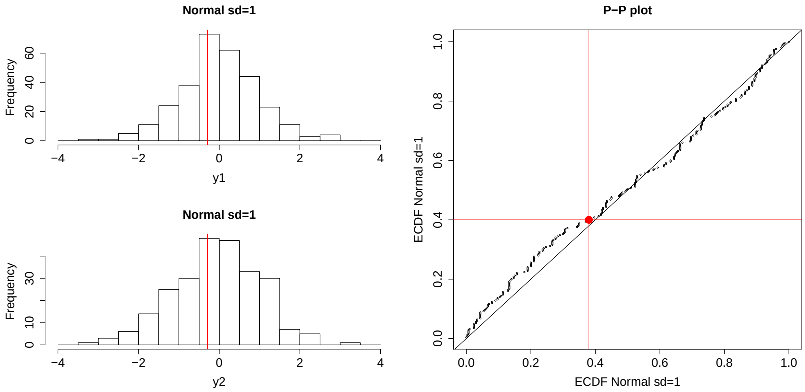
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



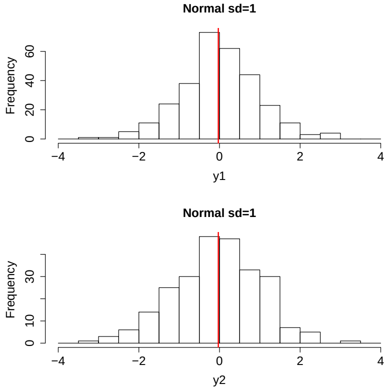
PP-plot

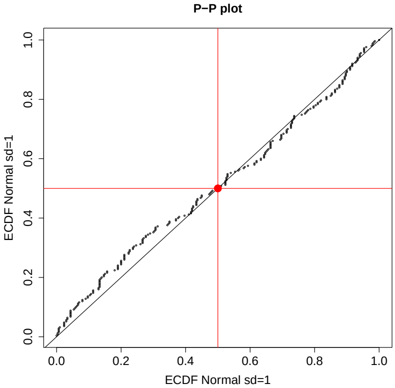
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



PP-plot

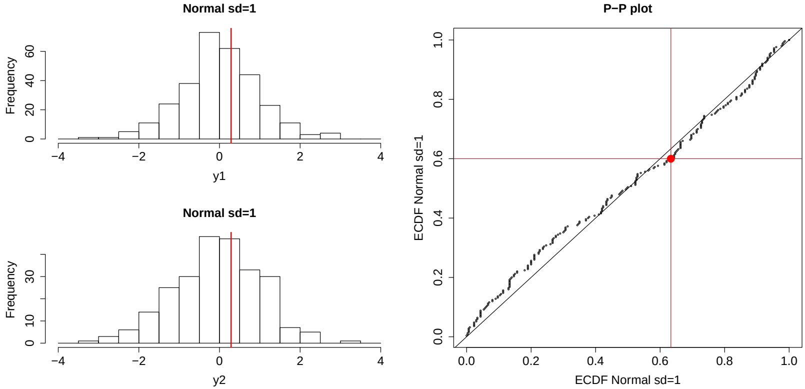
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.





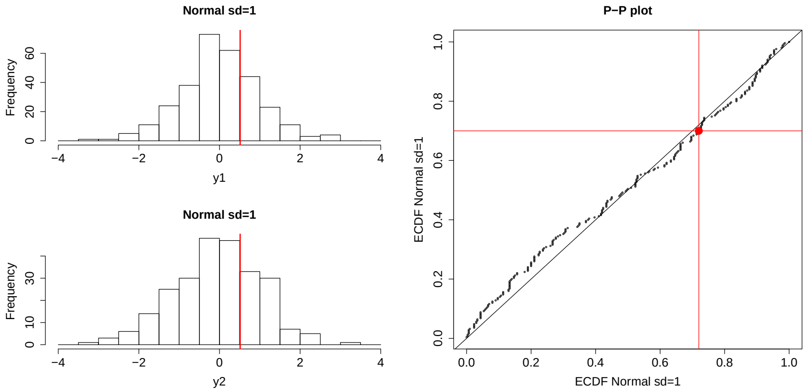
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



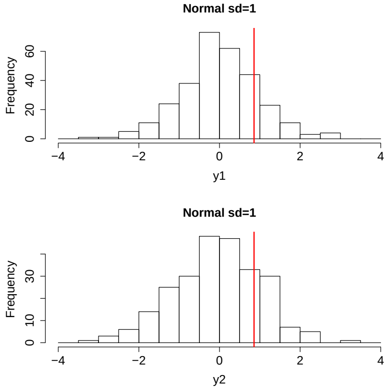
PP-plot

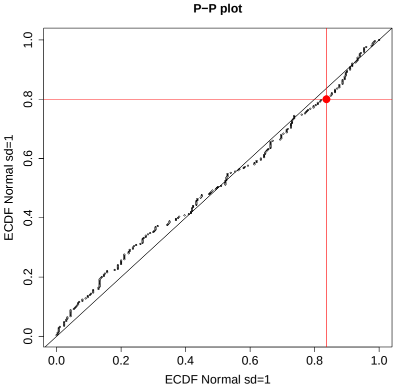
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



PP-plot

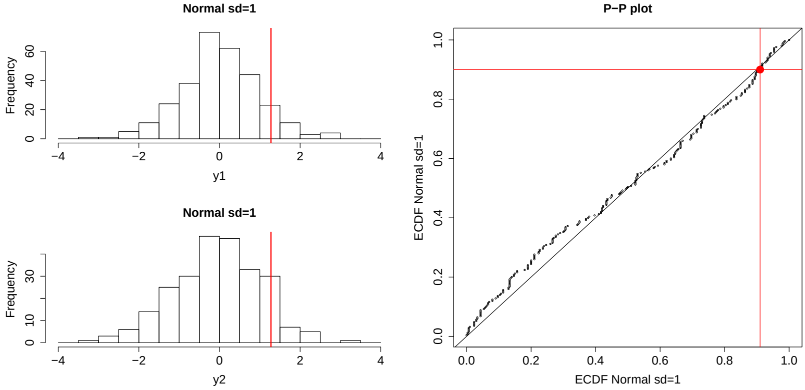
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.





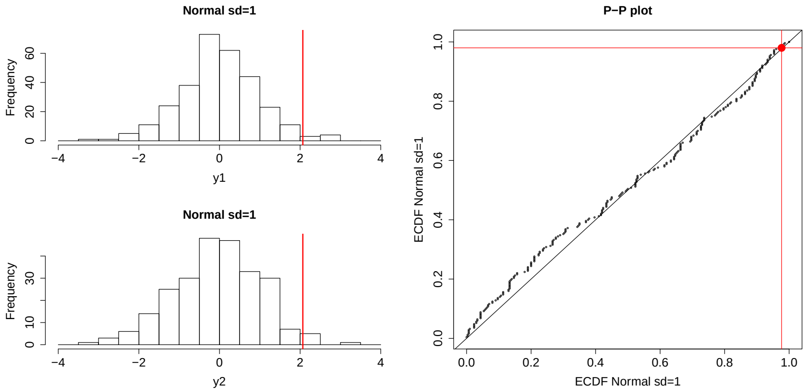
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



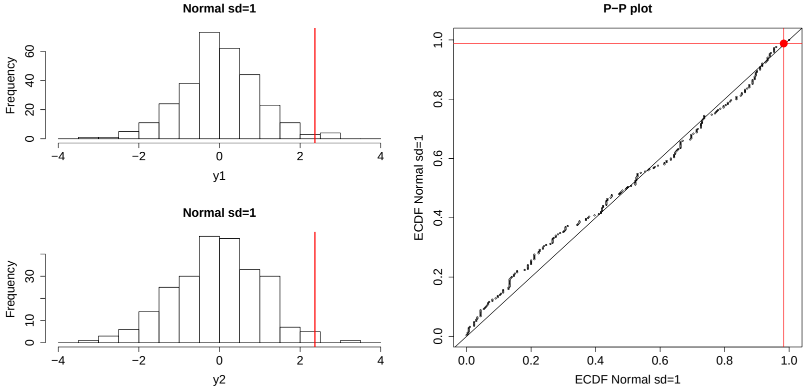
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



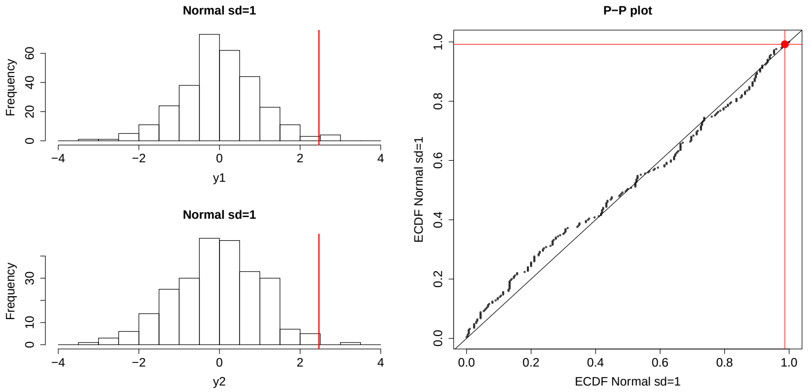
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



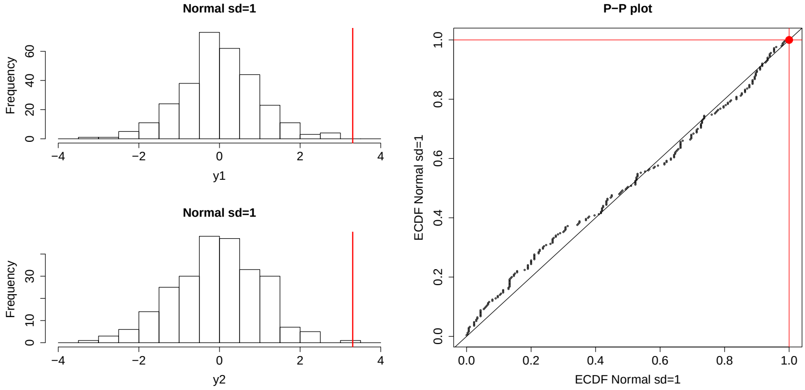
PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.

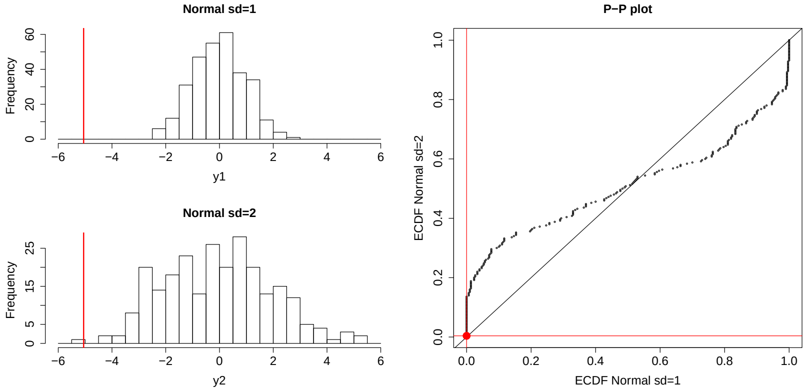


PP-plot

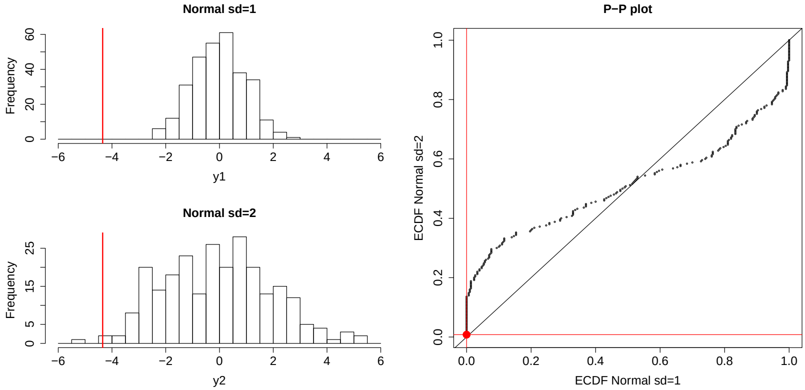
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



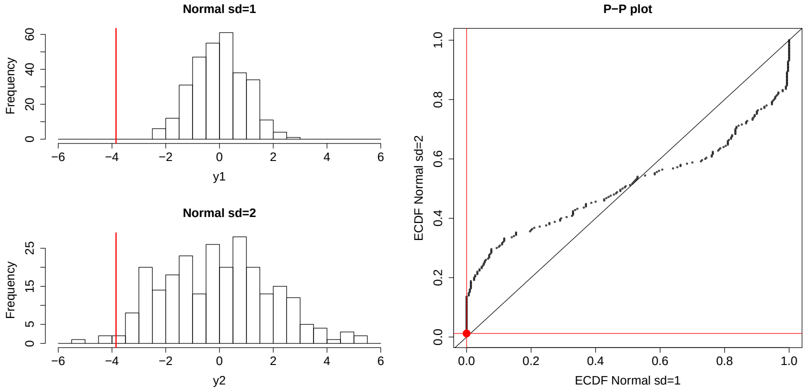
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

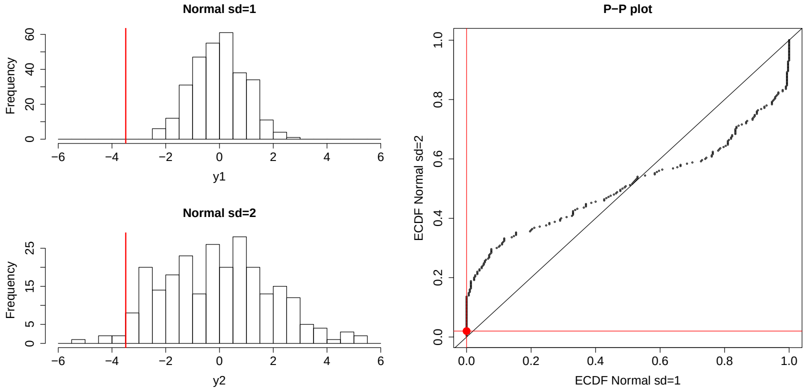
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

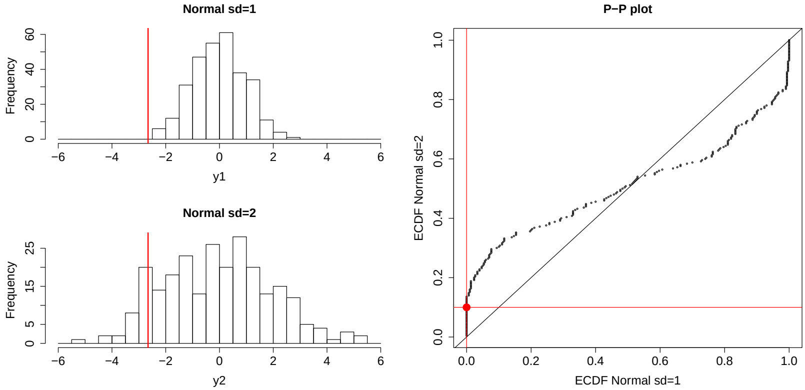
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

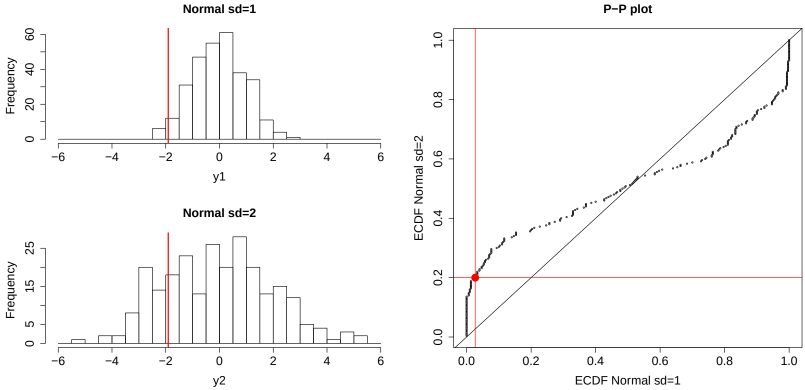
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

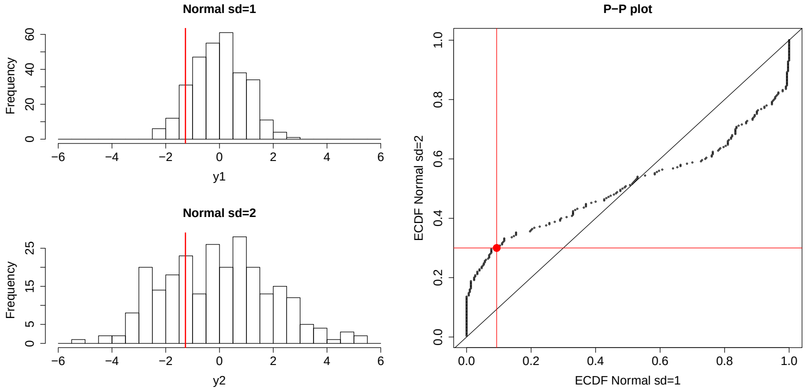
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

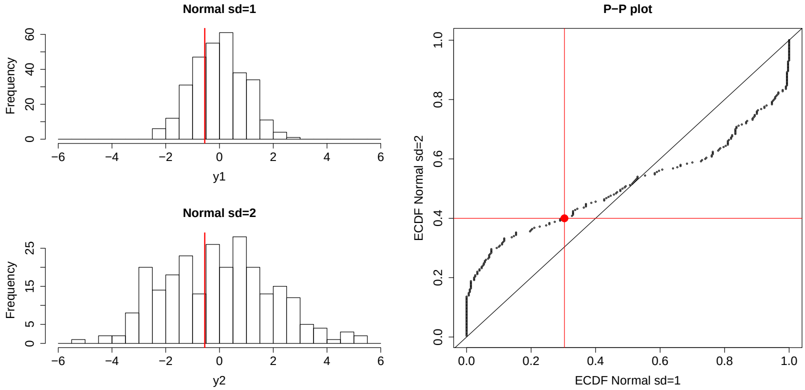
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

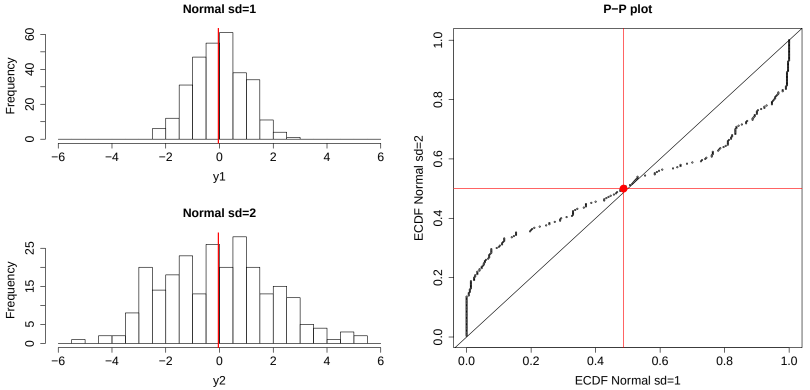
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

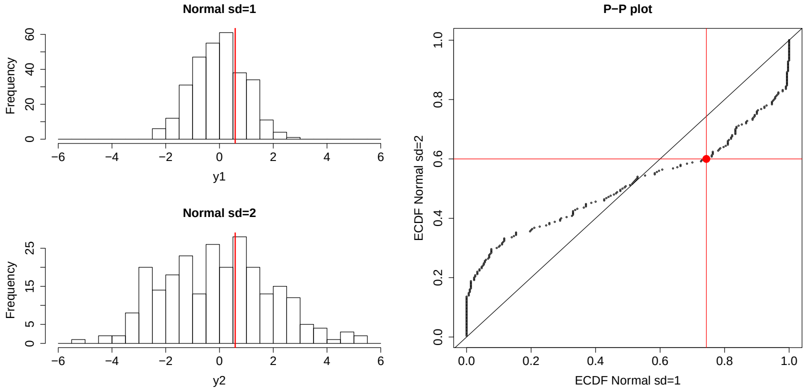
PP-plot


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

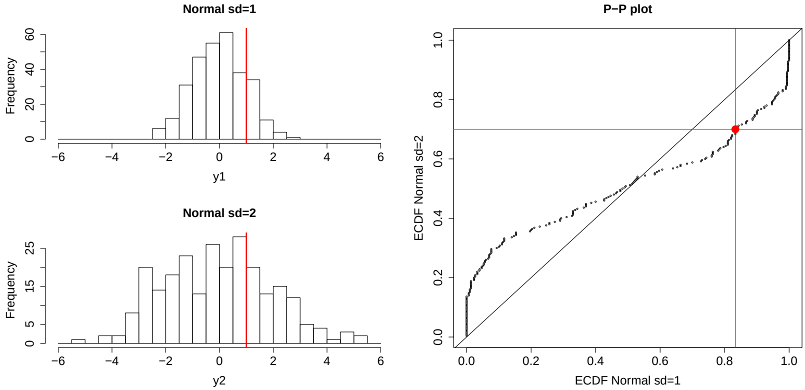
PP-plot


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

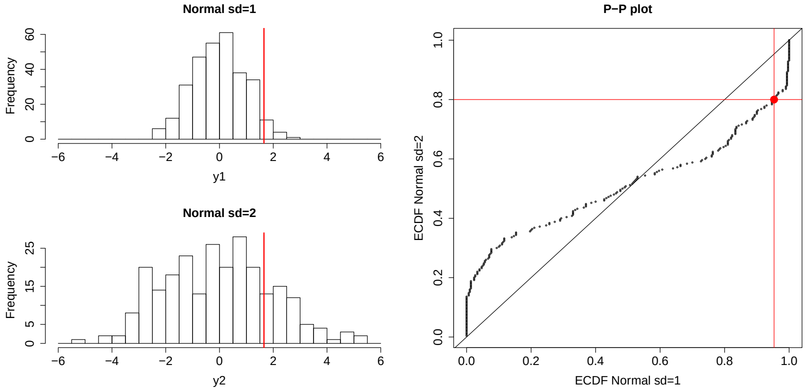
PP-plot


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

PP-plot

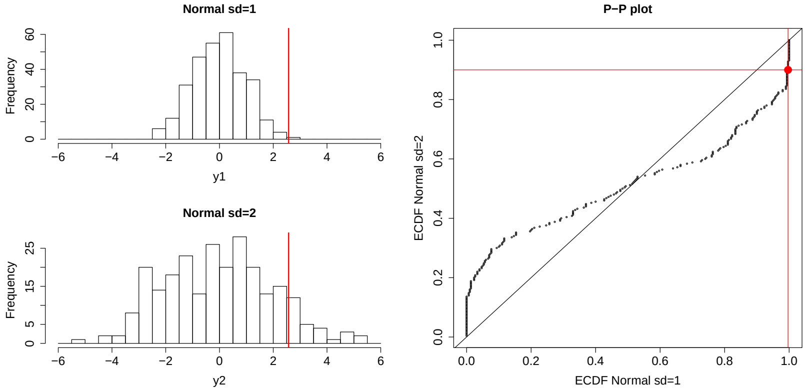

Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

PP-plot

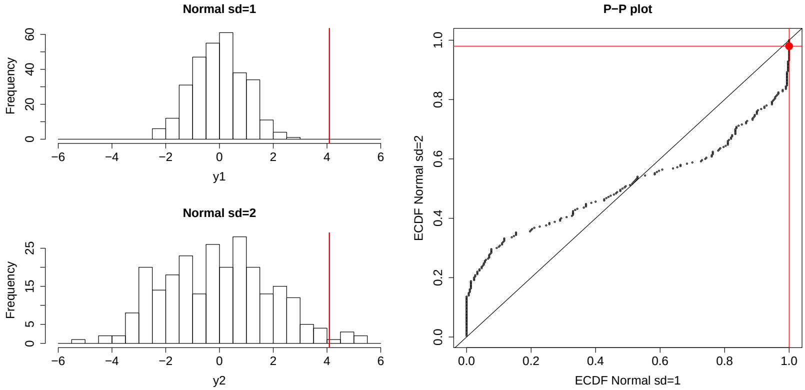


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

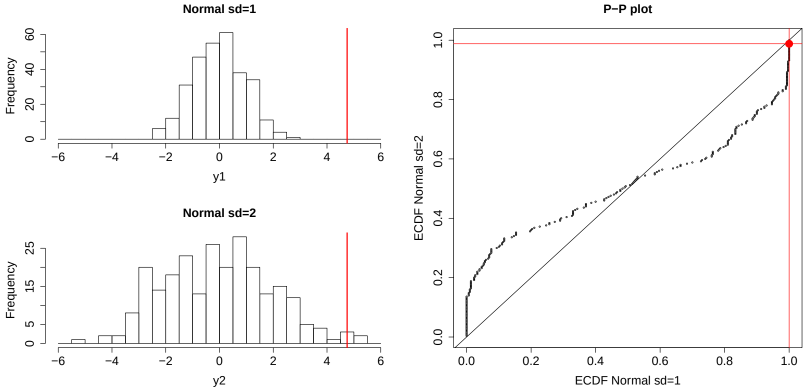
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

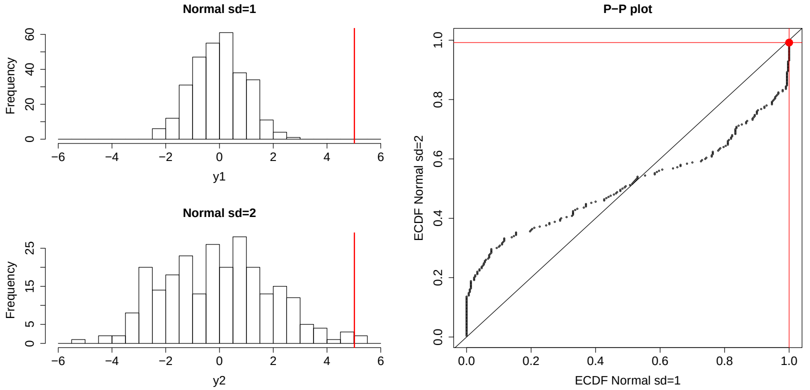
PP-plot


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

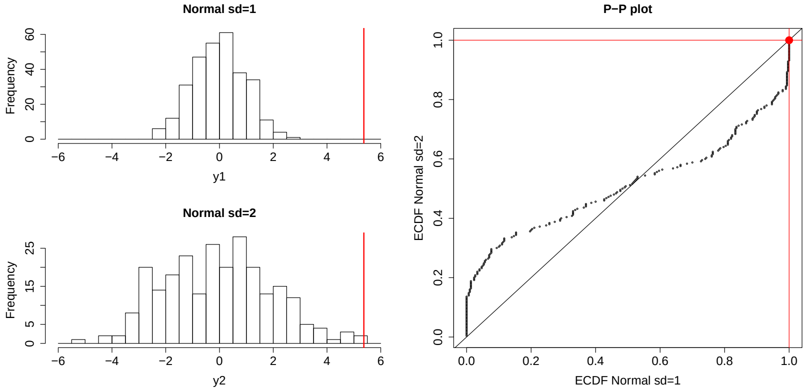
PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

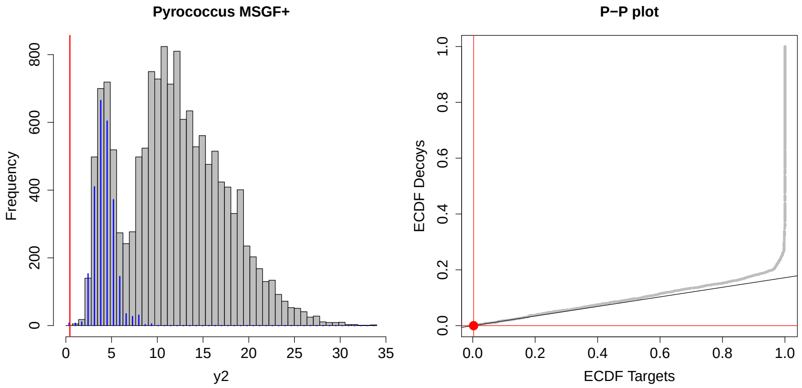
PP-plot


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

PP-plot


Normal sd=1
Frequency
60
40
20
0
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
25
15
5
0
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

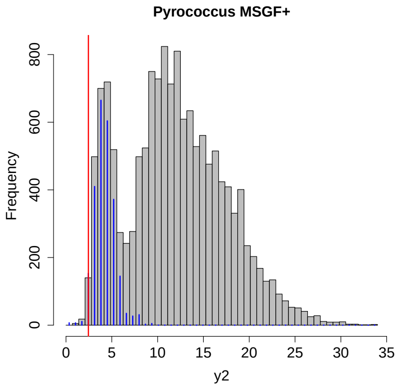
PP-plot: pyrococcus



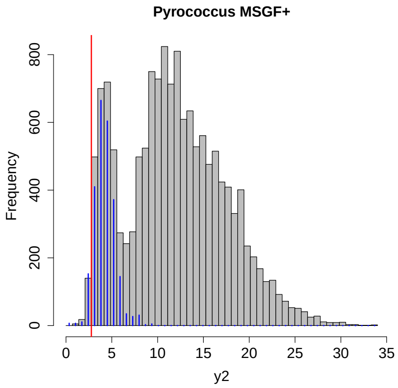
PP-plot: pyrococcus

What about π_0?

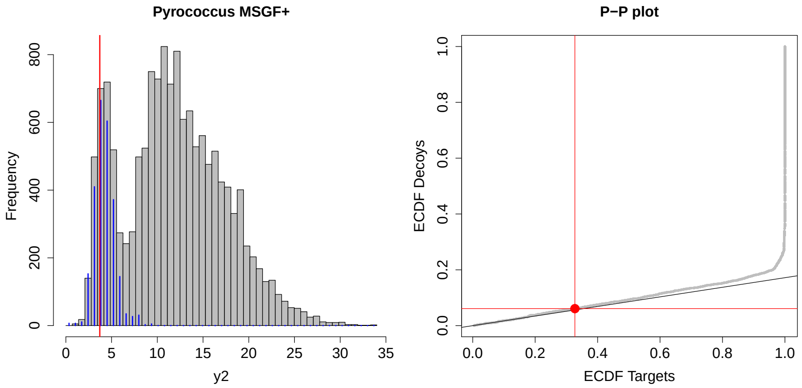
PP-plot: pyrococcus



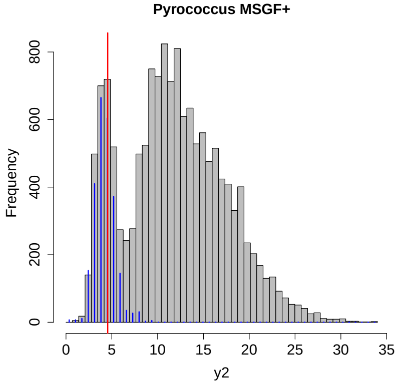
PP-plot: pyrococcus

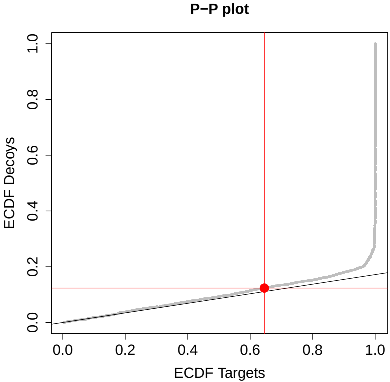


PP-plot: pyrococcus

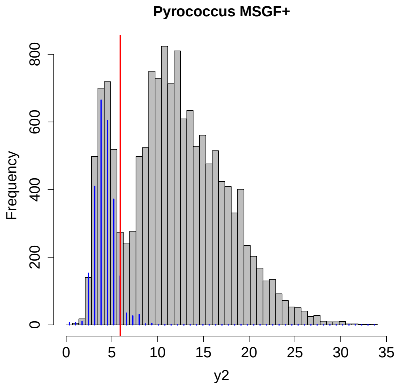


PP-plot: pyrococcus

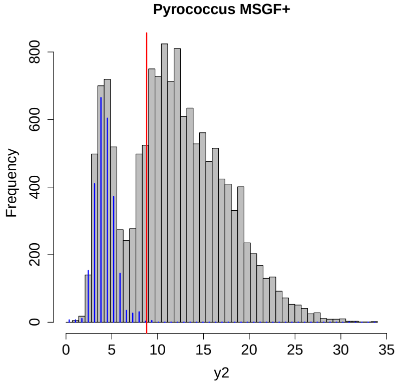


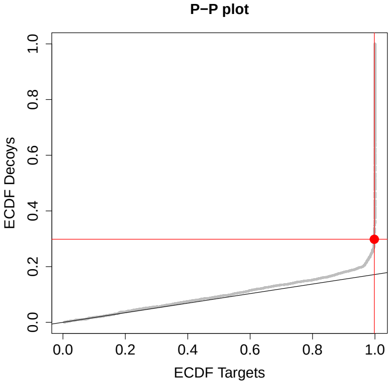


PP-plot: pyrococcus

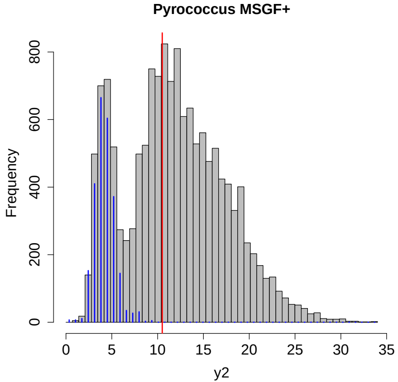


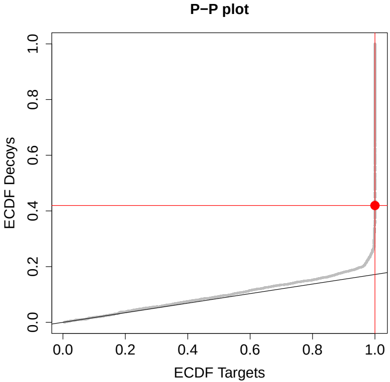
PP-plot: pyrococcus



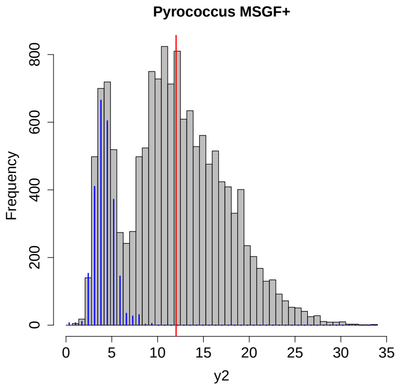


PP-plot: pyrococcus

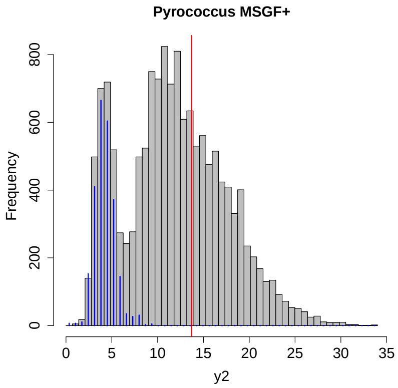




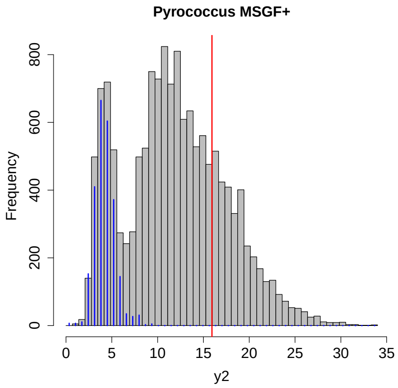
PP-plot: pyrococcus



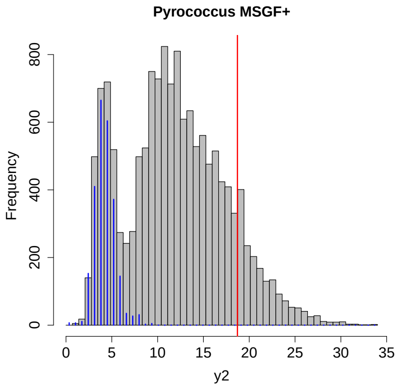
PP-plot: pyrococcus

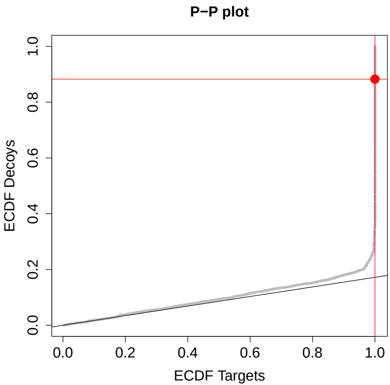


PP-plot: pyrococcus



PP-plot: pyrococcus





PP-plot: pyrococcus