

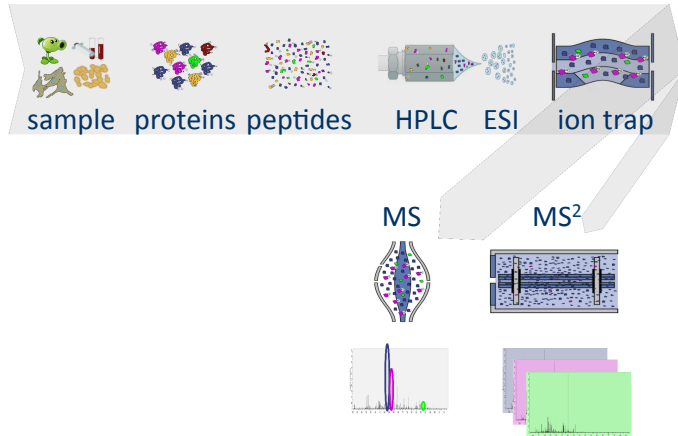
Statistical Methods for Quantitative MS-Based Proteomics:

1. Identification

Lieven Clement

Statistics and Genomics Seminar, UC Berkeley, California

Challenges in Label Free MS-based Quantitative proteomics



Quantification Identification

Challenges in Label Free MS-based Quantitative proteomics

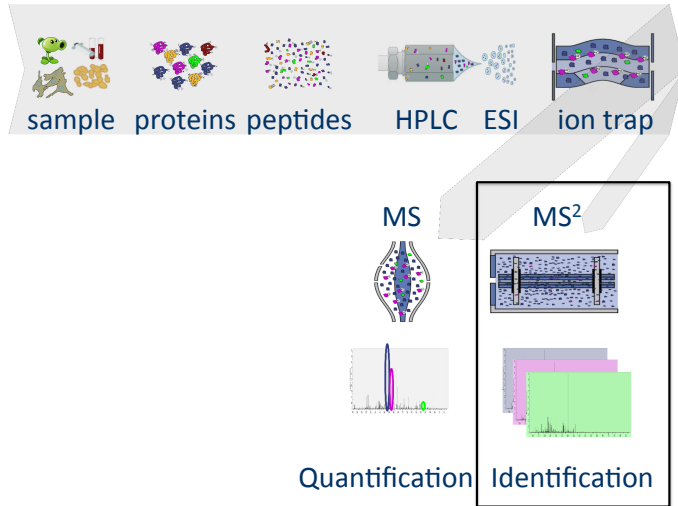




Table of Outcomes

	Called Bad	Called Correct	
Bad hit	TN	FP	m_0
Correct hit	FN	TP	m_1
Total	NR	R	m

- TN: number of true negatives
- FP: number of false positives
- FN: number of false negatives
- TP: number of true positives
- NR: number of non-rejections, R: number of rejections

Table of Outcomes

	Called Bad	Called Correct	
Bad hit	TN	FP	m_0
Correct hit	FN	TP	m_1
Total	NR	R	m
	Random Variables		Constant

Table of Outcomes

		Called Bad	Called Correct	
Unobservable	Bad hit	TN	FP	m_0
	Correct hit	FN	TP	m_1
Observable	Total	NR	R	m

Table of Outcomes

		Called Bad	Called Correct	
Unobservable	Bad hit	TN	FP	m_0
	Correct hit	FN	TP	m_1
Observable	Total	NR	R	m

$FDP = \frac{FP}{FP+TP}$. But is unknown! (FDP: false discovery proportion)

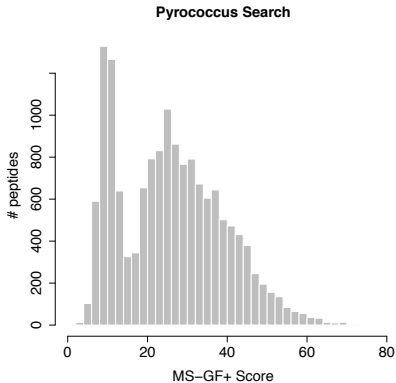
Table of Outcomes

		Called Bad	Called Correct	
Unobservable	Bad hit	TN	FP	m_0
	Correct hit	FN	TP	m_1
Observable	Total	NR	R	m

$FDP = \frac{FP}{FP+TP}$. But is unknown! (FDP: false discovery proportion)

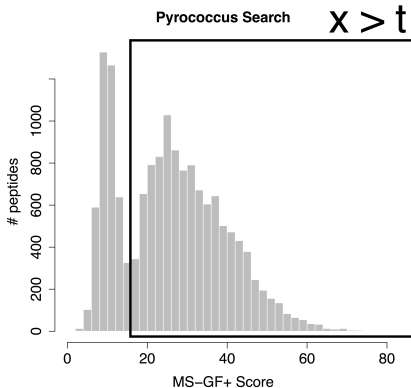
$FDR = E \left[\frac{FP}{FP+TP} \right]$. (FDR: false discovery rate)

Search engines return score that discriminates good from bad matches

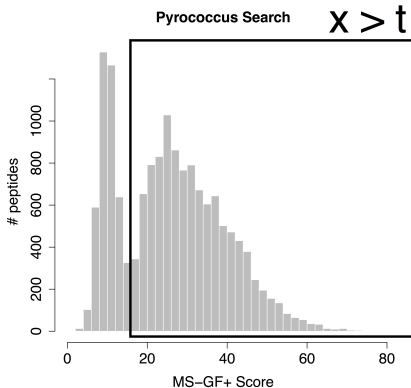


Search engines return score that discriminates good from bad matches

Score threshold t ?



Search engines return score that discriminates good from bad matches

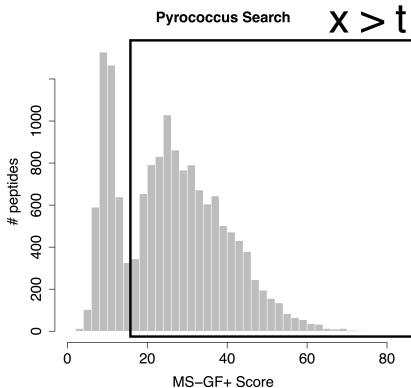


Score threshold t ?

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

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

Search engines return score that discriminates good from bad matches



Score threshold t ?

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

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

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

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

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

Link to Benjamini Hochberg FDR

- Bayesian FDR

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

- 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

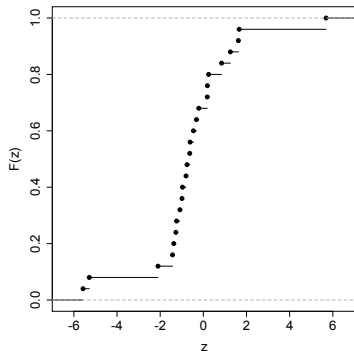
$$\text{FDR}(t) = \frac{\pi_0 [1 - F_0(t)]}{1 - F(t)}$$

- BH for one sided test

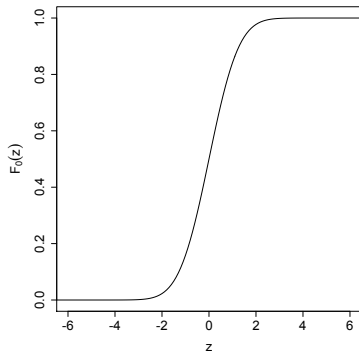
$$\text{FDR}(t) = \frac{mp(t)}{\#t_i \geq t} = \frac{1 - F_0(t)}{\frac{\#t_i \geq t}{m}} = \frac{1 - F_0(t)}{1 - F(t)}$$

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

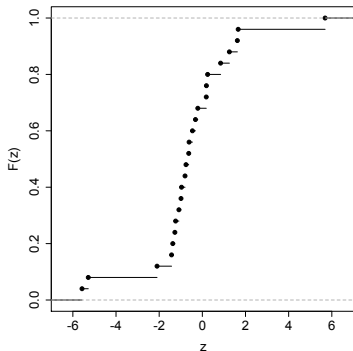
- Benjamini Hochberg 1995:
 $F(t)$ using the ECDF



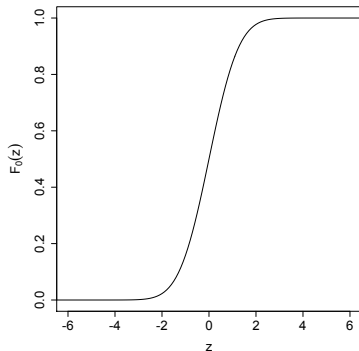
$F_0(t)$ theoretical CDF



- Benjamini Hochberg 1995:
 $F(t)$ using the ECDF

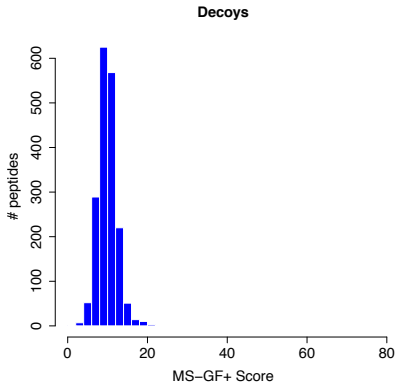


$F_0(t)$ theoretical CDF



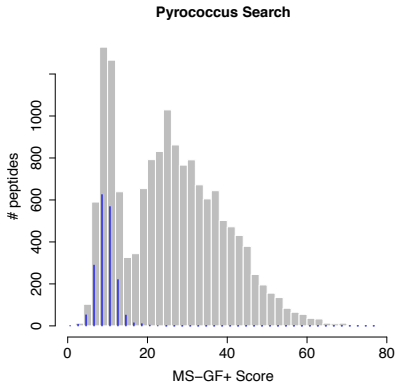
- How to define $F_0(t)$ in proteomics?

Target-Decoy approach to establish null distribution



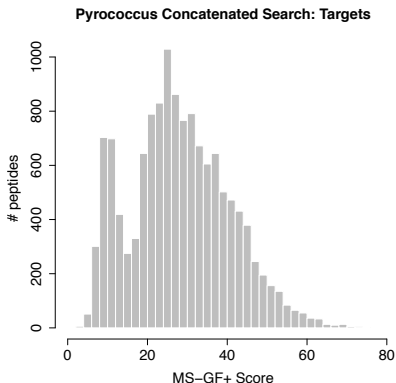
- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice

Target-Decoy approach to establish null distribution



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

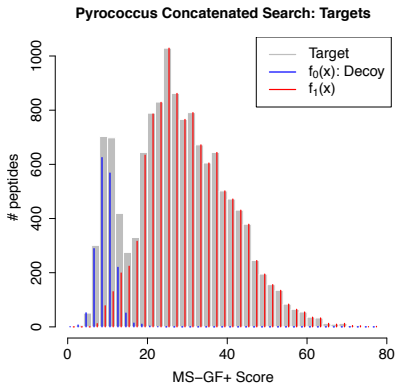
Target-Decoy approach to establish null distribution



- 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)

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

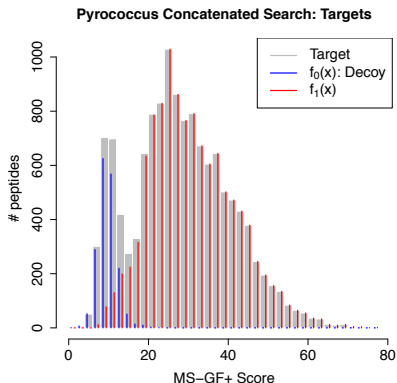
Target-Decoy approach to establish null distribution



- 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



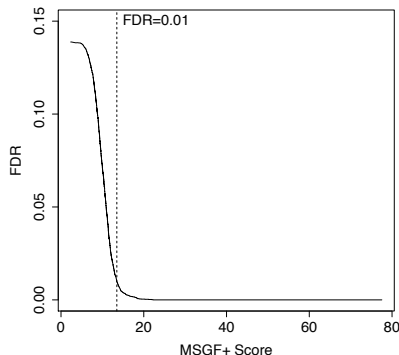
- Score cutoff?
- Competitive Target - decoy:

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

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

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

Target-Decoy approach to establish null distribution



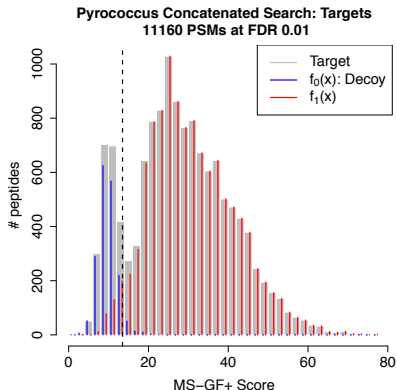
- Score cutoff?
- Competitive Target - decoy:

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

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

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

Target-Decoy approach to establish null distribution



- Score cutoff?
- Competitive Target - decoy:

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

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

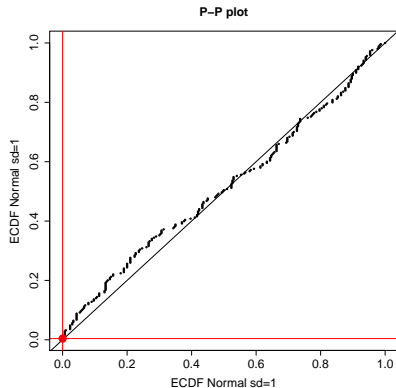
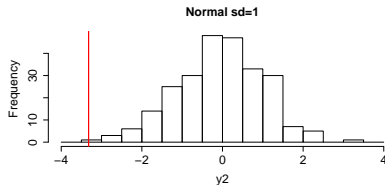
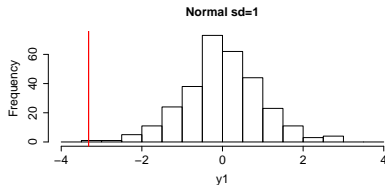
$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{1 - \bar{F}_0(x)}{1 - \bar{F}(x)}$$

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.

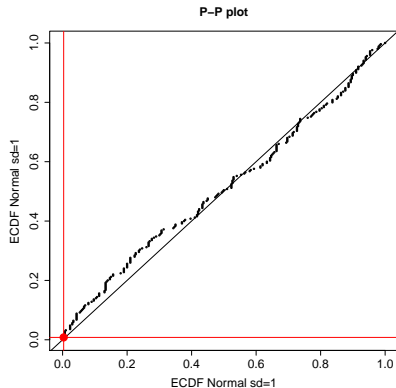
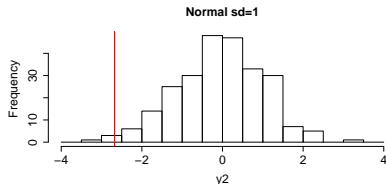
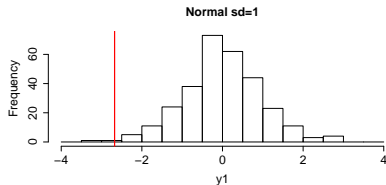
PP-plot

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



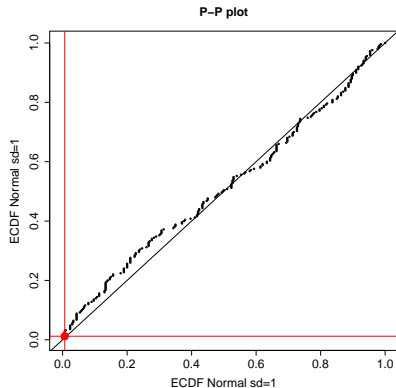
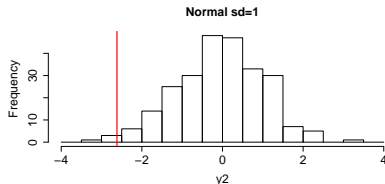
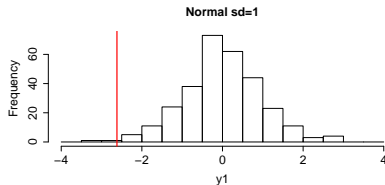
PP-plot

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



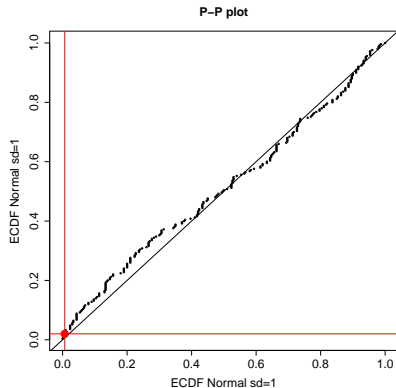
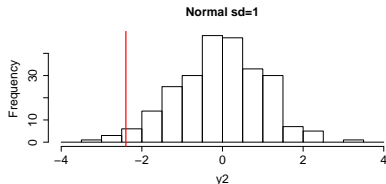
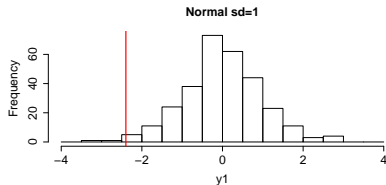
PP-plot

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



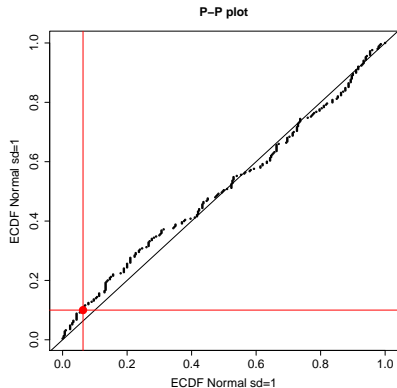
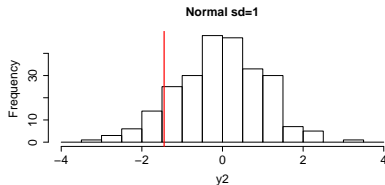
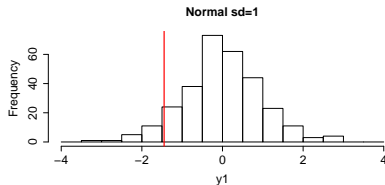
PP-plot

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



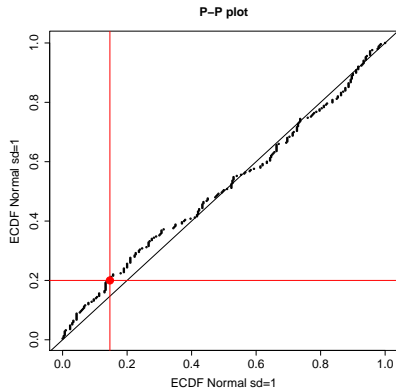
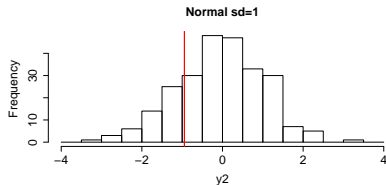
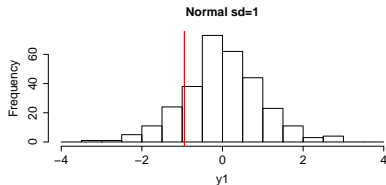
PP-plot

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



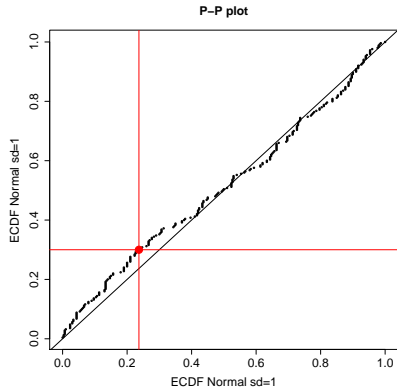
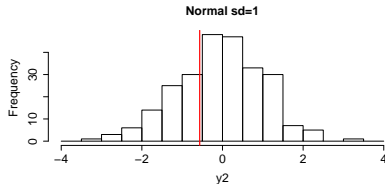
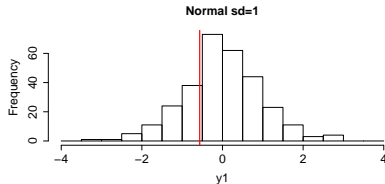
PP-plot

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



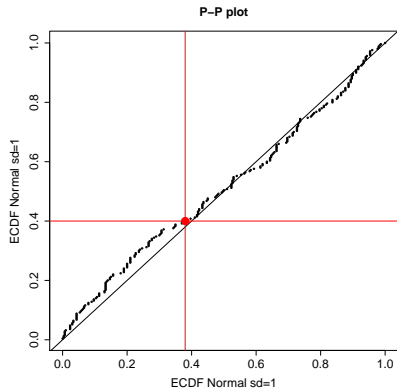
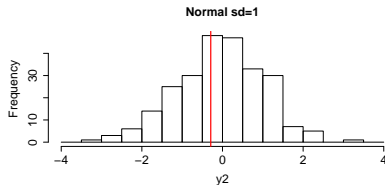
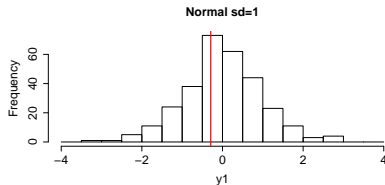
PP-plot

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



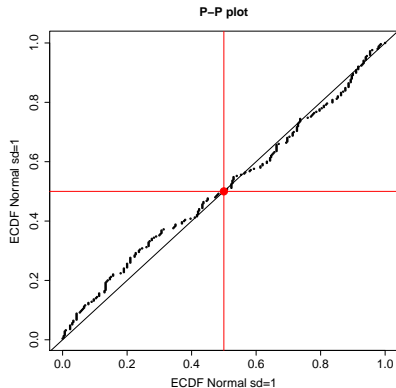
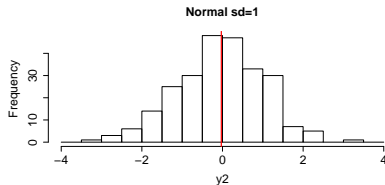
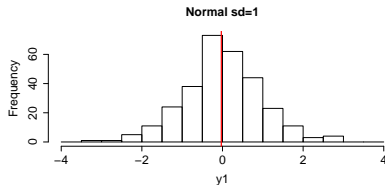
PP-plot

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



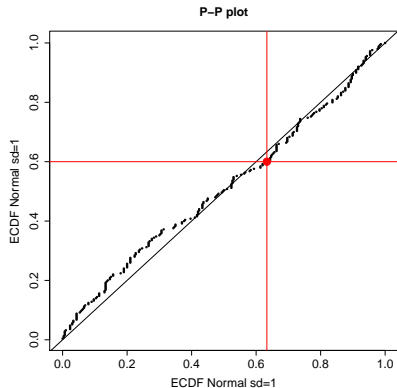
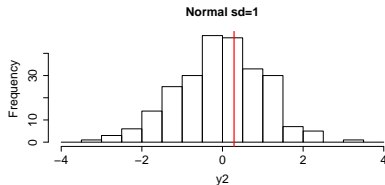
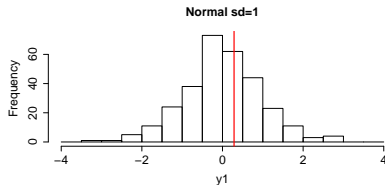
PP-plot

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



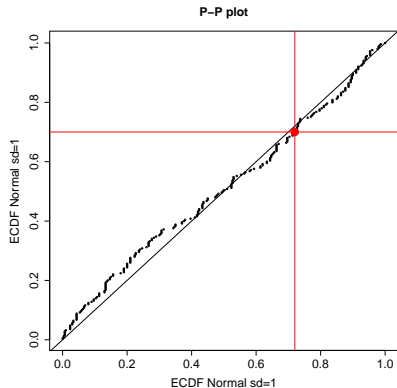
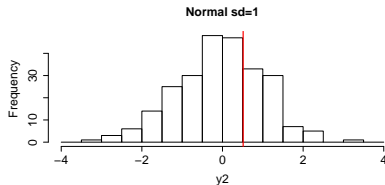
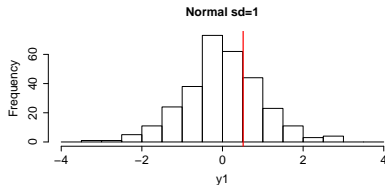
PP-plot

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



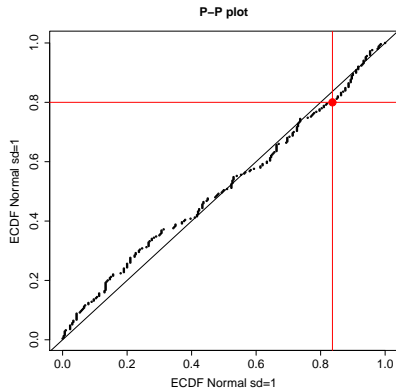
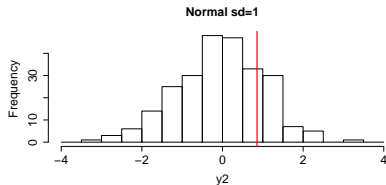
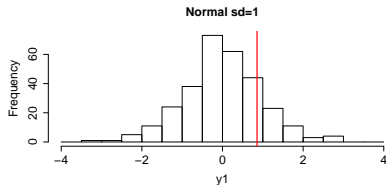
PP-plot

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



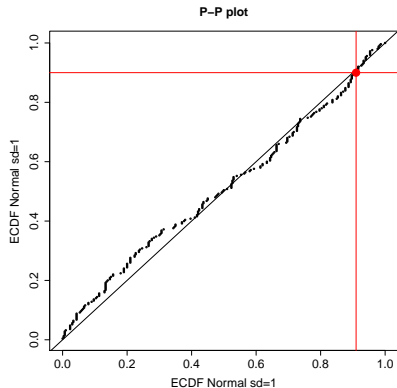
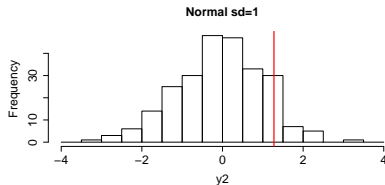
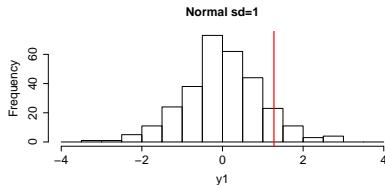
PP-plot

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



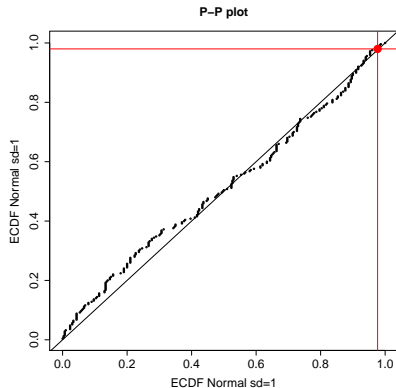
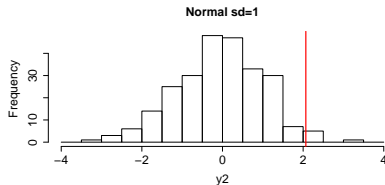
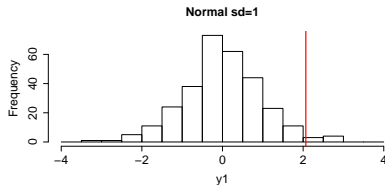
PP-plot

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



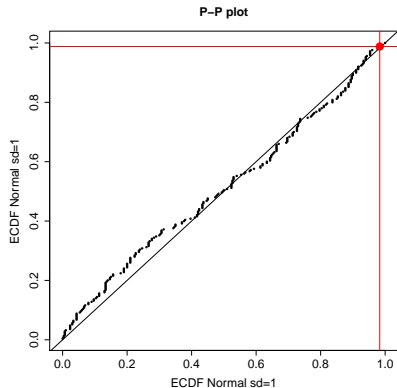
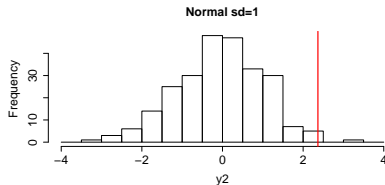
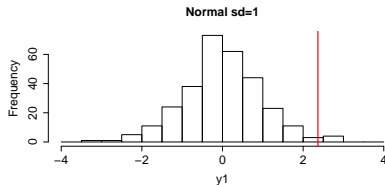
PP-plot

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



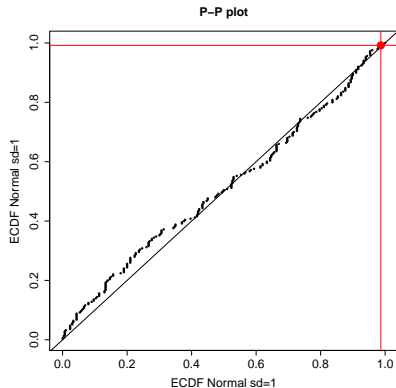
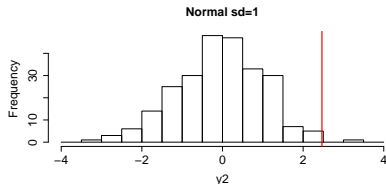
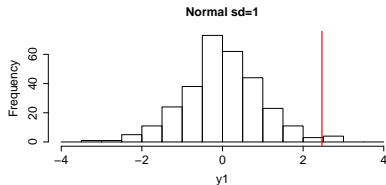
PP-plot

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



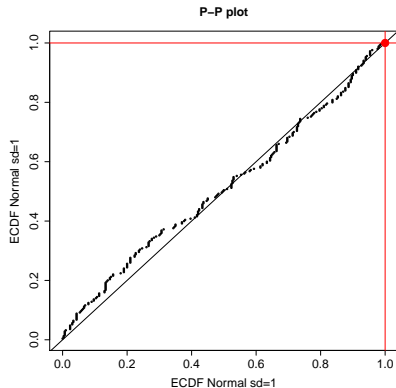
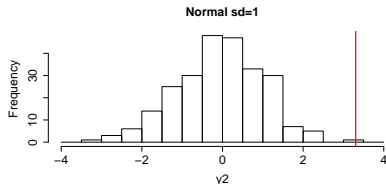
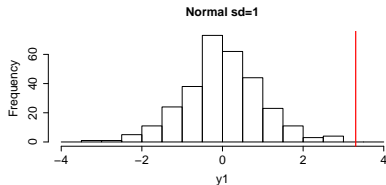
PP-plot

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

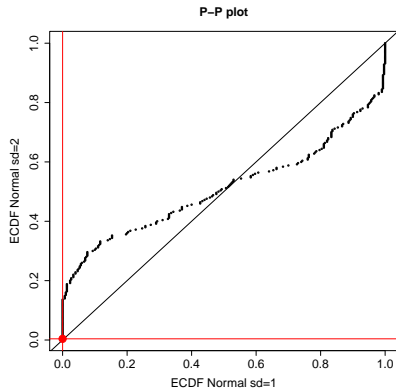
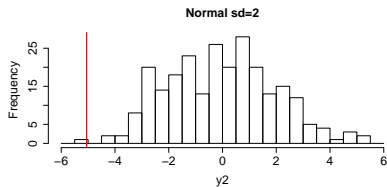
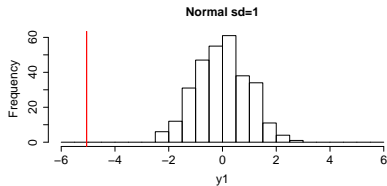


PP-plot

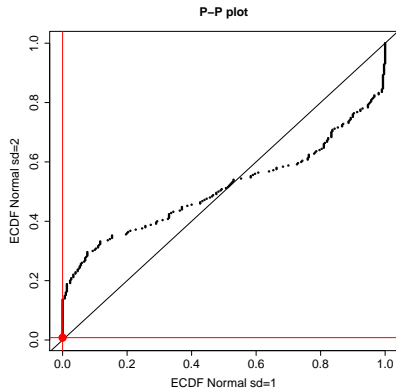
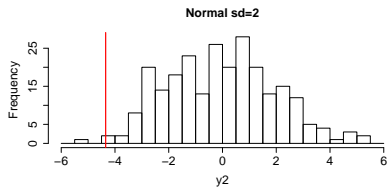
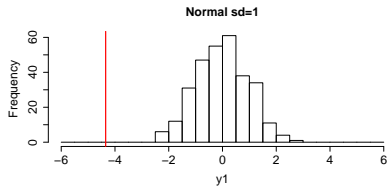
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



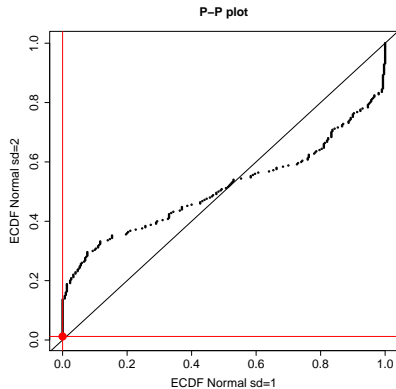
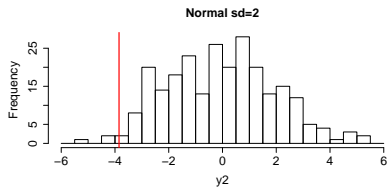
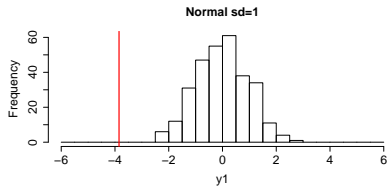
PP-plot



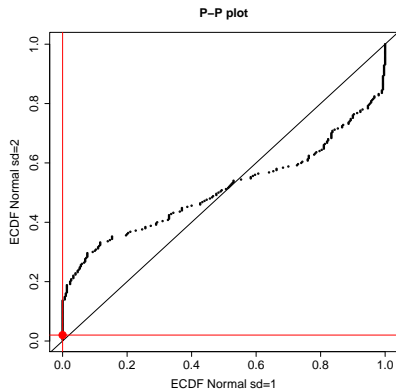
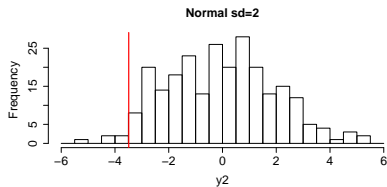
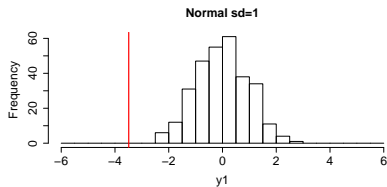
PP-plot



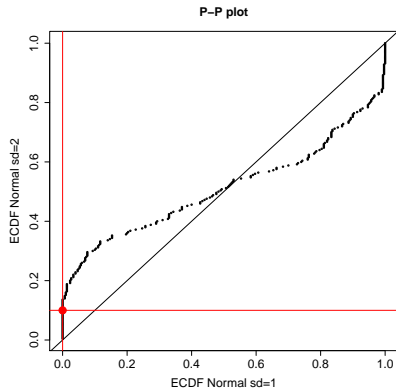
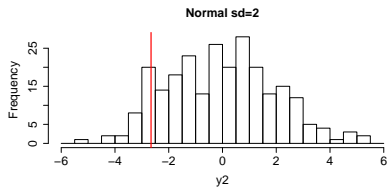
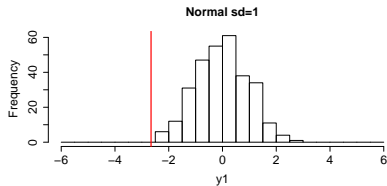
PP-plot



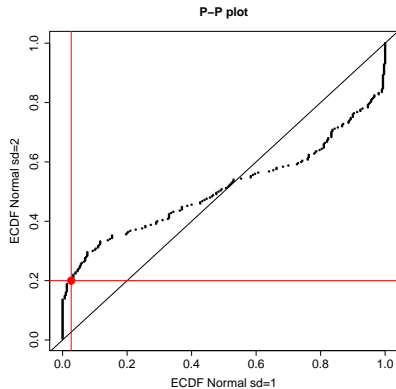
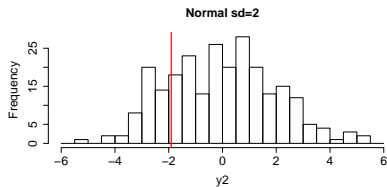
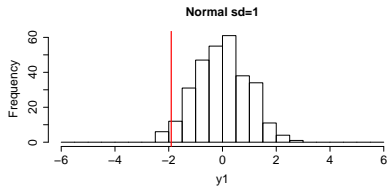
PP-plot



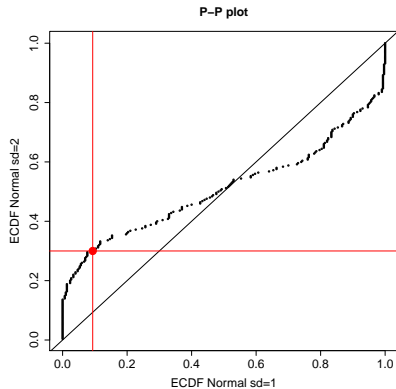
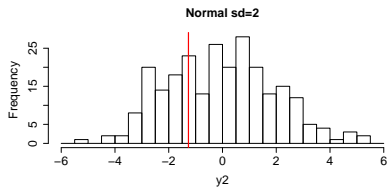
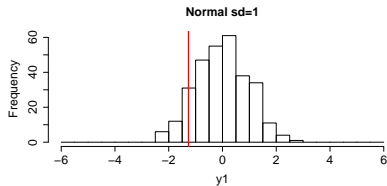
PP-plot



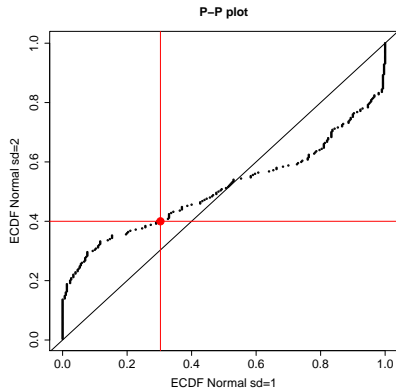
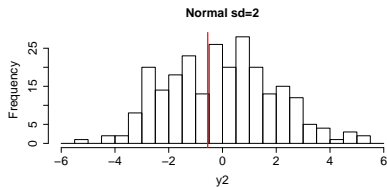
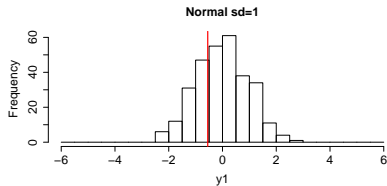
PP-plot



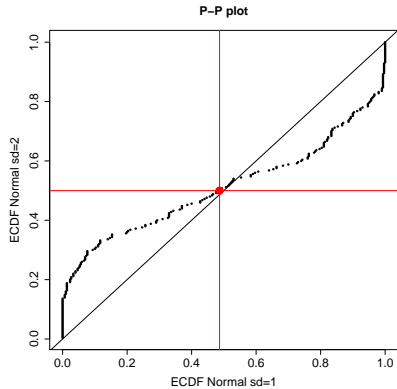
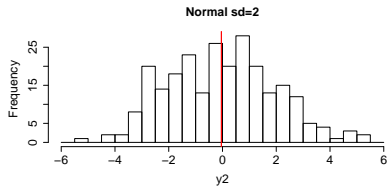
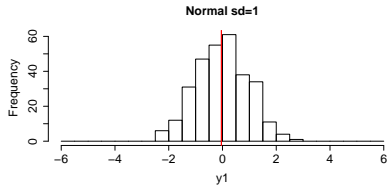
PP-plot



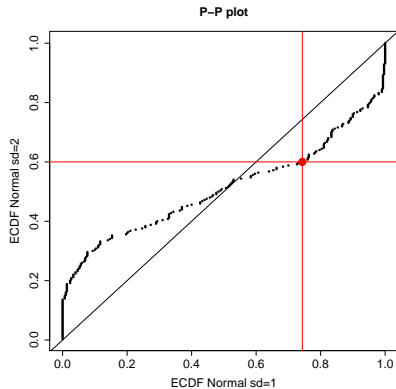
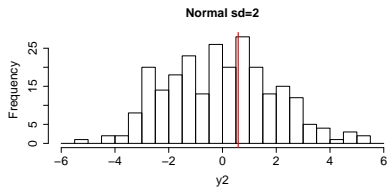
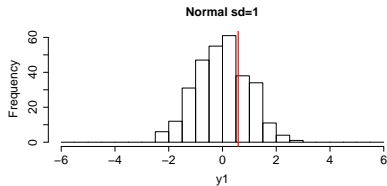
PP-plot



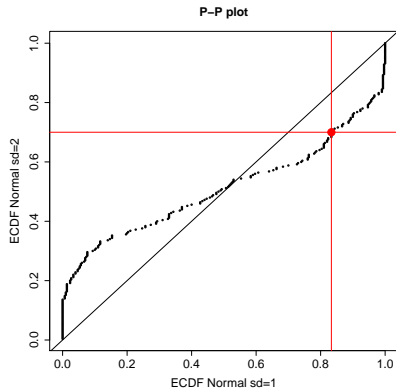
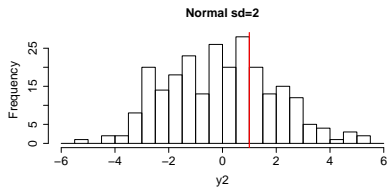
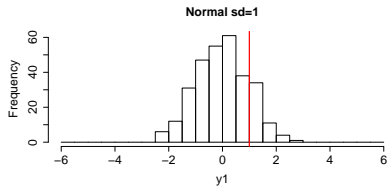
PP-plot



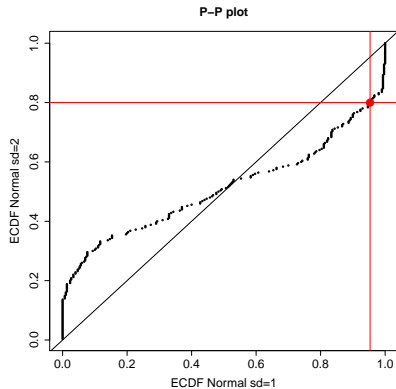
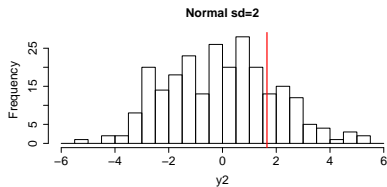
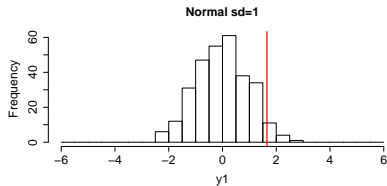
PP-plot



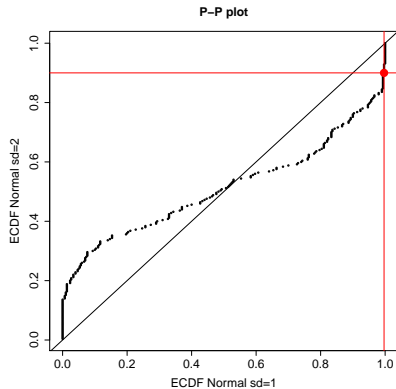
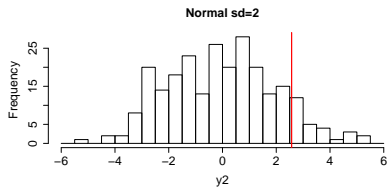
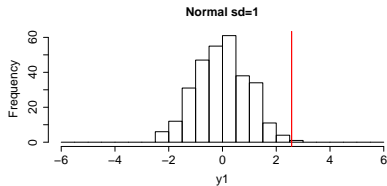
PP-plot



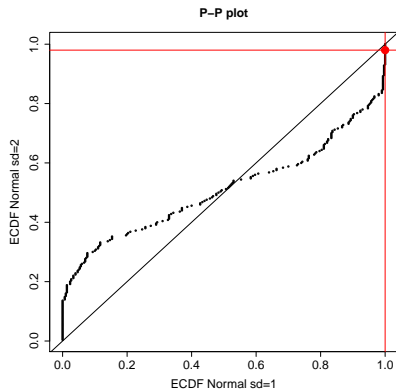
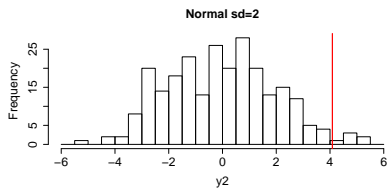
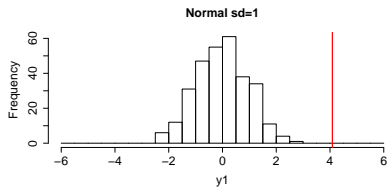
PP-plot



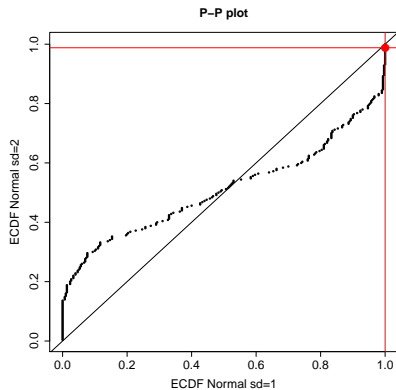
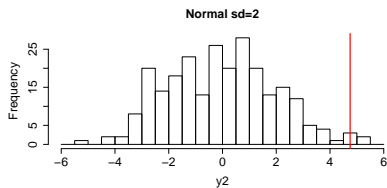
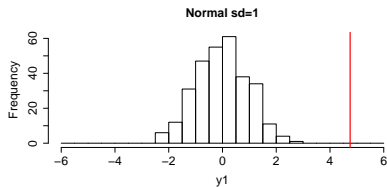
PP-plot



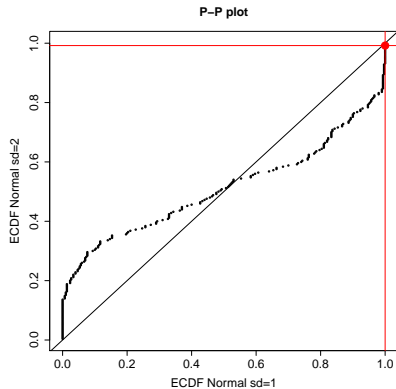
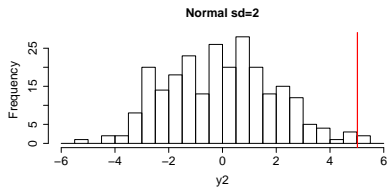
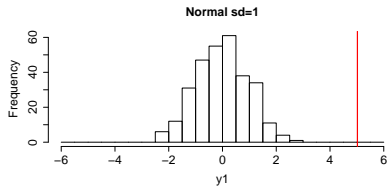
PP-plot



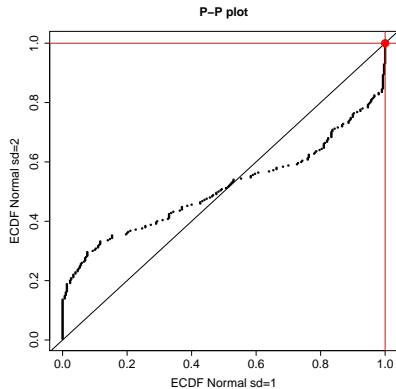
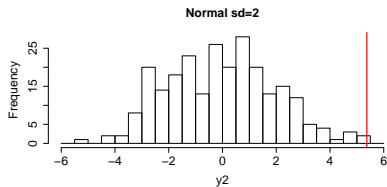
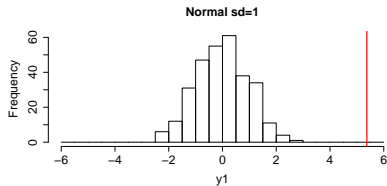
PP-plot



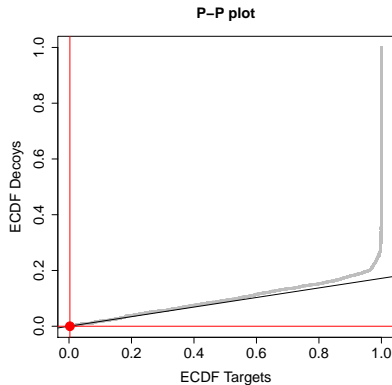
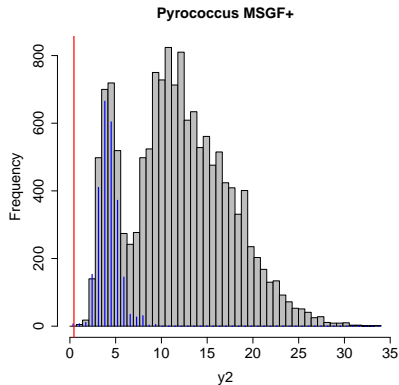
PP-plot



PP-plot



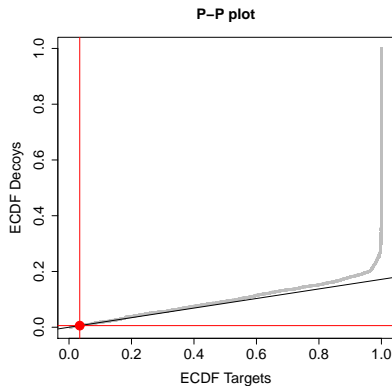
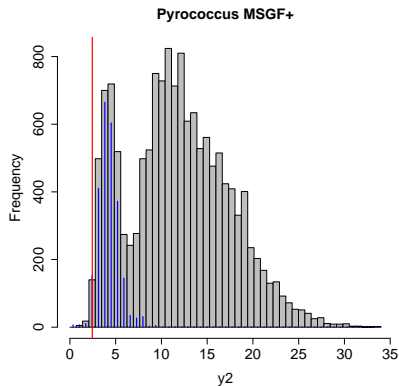
PP-plot: pyrococcus



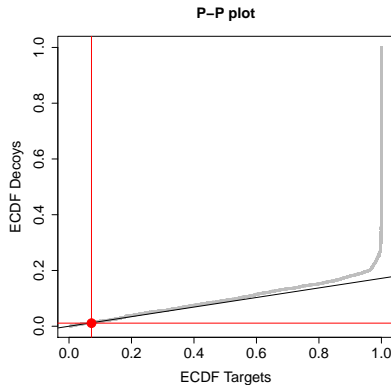
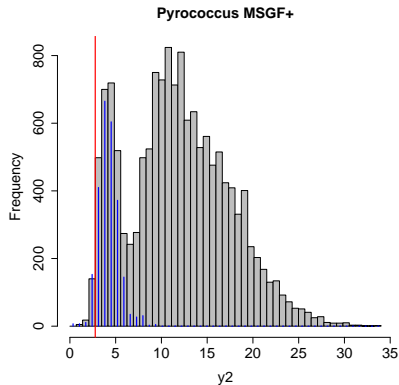
PP-plot: pyrococcus

What about π_0 ?

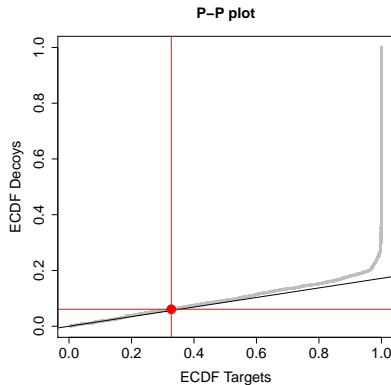
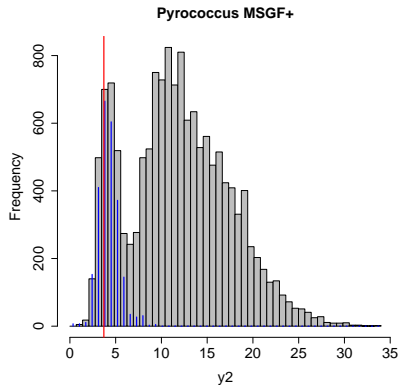
PP-plot: pyrococcus



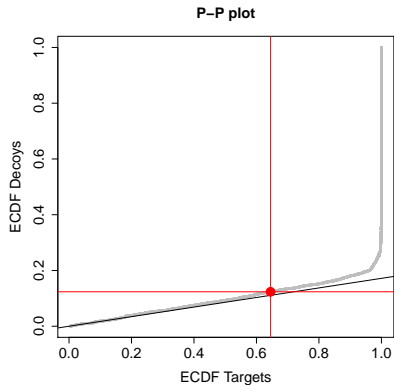
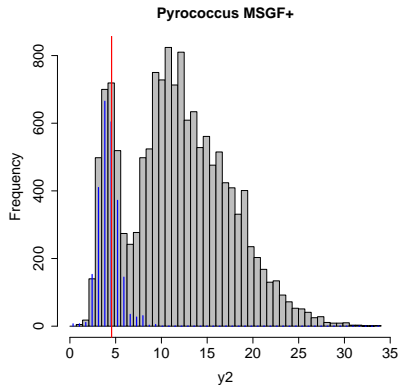
PP-plot: pyrococcus



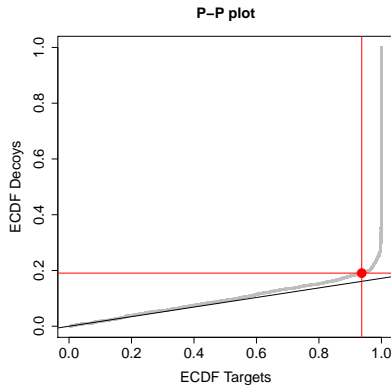
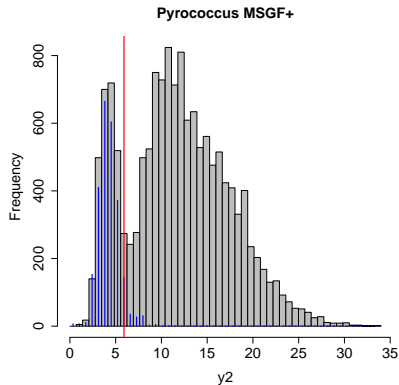
PP-plot: pyrococcus



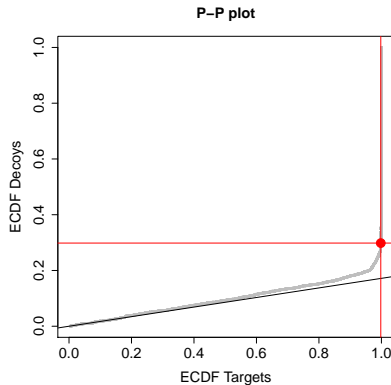
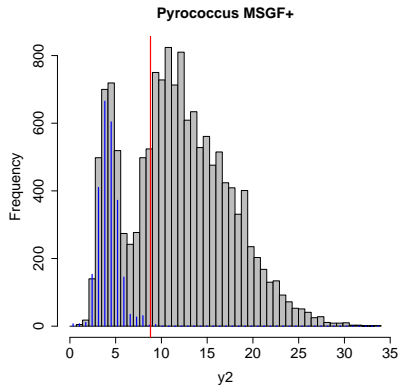
PP-plot: pyrococcus



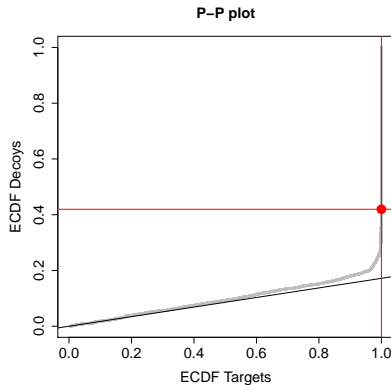
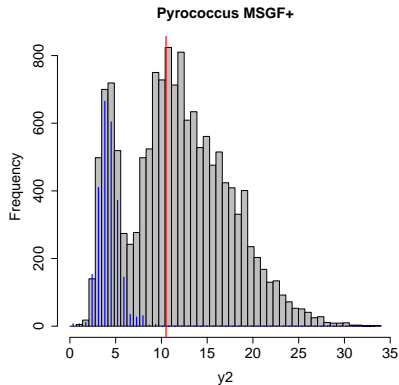
PP-plot: pyrococcus



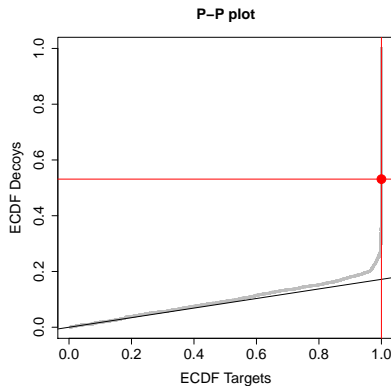
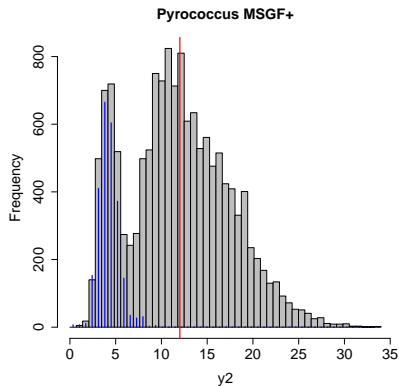
PP-plot: pyrococcus



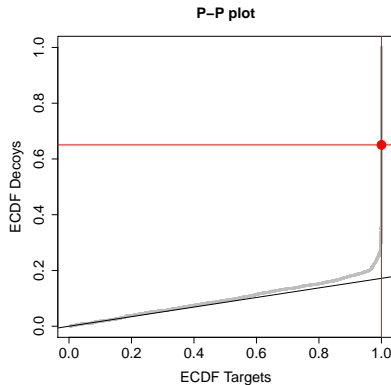
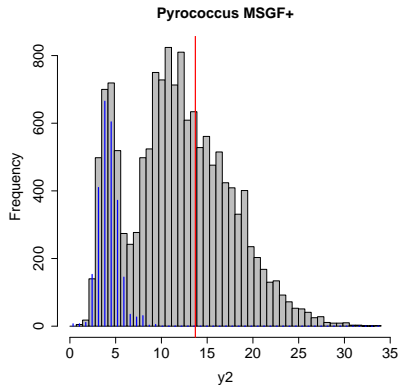
PP-plot: pyrococcus



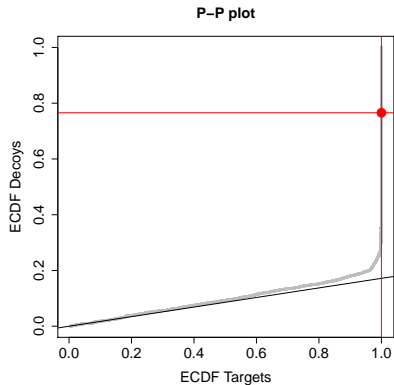
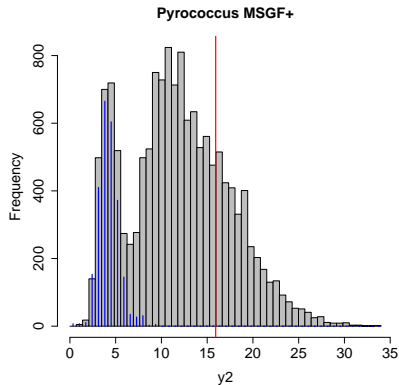
PP-plot: pyrococcus



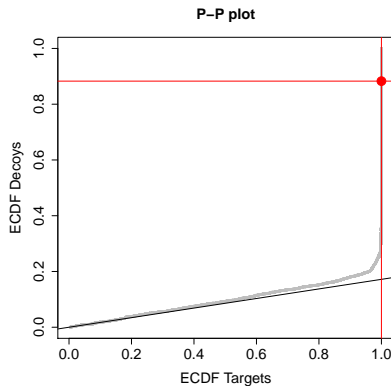
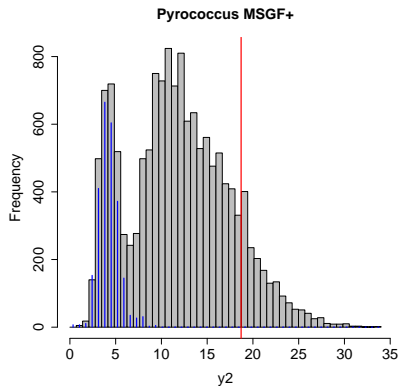
PP-plot: pyrococcus



PP-plot: pyrococcus



PP-plot: pyrococcus



PP-plot: pyrococcus

