Statistical Methods for Quantitative MS-Based Proteomics: 1. Identification & False discovery rate

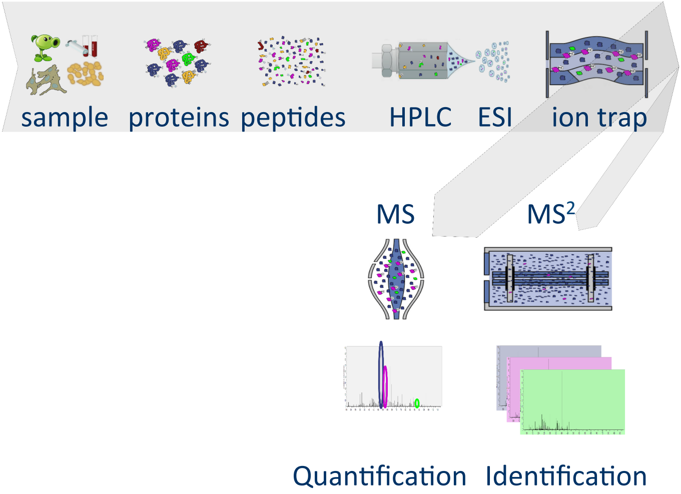
Lieven Clement

Statistics and Genomics Seminar, UCBerkeley, California

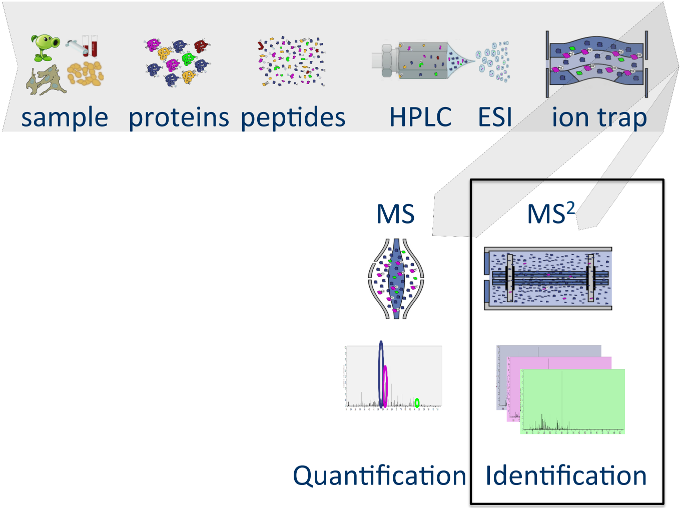
statOmics, Krijgslaan 281 (S9), Gent, Belgium

lieven.clement@ugent.be

Challenges in Label Free MS-based Quantitative Proteomics



Challenges in Label Free MS-based Quantitative Proteomics



sample proteins peptides HPLC ESI ion trap

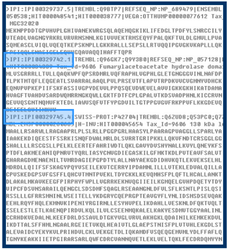
MS

MS^2

Quantification

Identification

Identification



```
>IPI:IPI00329737.5|TREMBL:Q9BTP7|REFSEQ_NP:NP_689479|ENSEMBL
050538;HIT000048541;HIT000038777|VEGA:OTTHUMP00000077612 Tax
 MGC32020
MEKNPPDDTGPVHVPLGHIVANEKWRGSQLAQEMQGKIKLIFEDGLTPDFYLSNRCCILY
VTEADLVAGNGYRKRLVRVRNSNNLKGIVVVEKTRMSEQYFPALQKFTVLDLGMVLLPVA
SQMEASCLVIQLVQEQTKEPSKNPLLGKKRALLLSEPSLLRTVQQIPGVGKVKAPLLLQK
FPSIQQLSNASIGELEQVVGQAVAQQIHAFFTQPR
>IPI:IPI00329742.1|TREMBL:Q96GK7;Q9Y3B0|REFSEQ_NP:NP_057128|
;HIT000004009 Tax_Id=9606 Fumarylacetoacetate hydrolase doma
MLVSGRRRLLTVLLQAQKWPFQPSRDMRLVQFRAPHLVGPHLGLETGNGGGVINLNAFDP
TLPKTMTQFLEQGEATLSVARRALAAQLPVLPRSEVTFLAPVTRPDKVVCVGMNYVDHCK
EQNVPVPKEPIIFSKFASSIVGPYDEVVLPPQSQEVDWEVELAVVIGKKGKHIKATDAMA
HVAGFTVAHDVSARDWQMRRNGKQWLLGKTFDTFCPLGPALVTKDSVADPHNLKICCRVN
GEVVQSGNTNQMVFKTEDLIAWVSQFVTFYPGDVILTGTPPGVGVFRKPPVFLKKGDEVQ
QEIEELQVIINKVV
>IPI:IPI00329745.4|SWISS-PROT:P42704|TREMBL:Q6ZUD8;Q53PC0;Q7
MBL:ENSP00000260665|H-INV:HIT000045654 Tax_Id=9606 130 kDa le
MAALLRSARWLLRAGAAPRLPLSLRLLPGGPGRLHAASYLPAARAGPVAGGLLSPARLYA
IAAKEKDIQEESTFSSRKISNQFDWALMRLDLSVRRTGRIPKKLLQKVFNDTCRSGGLGG
SHALLLLRSCGSLLPELKLEERTEFAHRIWDTLQKLGAVYDVSHYNALLKVYLQNEYKFS
PTDFLAKMEEANIQPNRVTYQRLIASYCNVGDIEGASKILGFMKTKDLPVTEAVFSALVT
GHARAGDMENAENILTVMRDAGIEPGPDTYLALLNAYAEKGDIDHVKQTLEKVEKSELHL
MDRDLLQIIFSFSKAGYPQYVSEILEKVTCERRYIPDAMNLILLLVTEKLEDVALQILLA
CPVSKEDGPSVFGSFFLQHCVTMNTPVEKLTDYCKKLKEVQMHSFPLQFTLHCALLANKT
DLAKALMKAVKEEGFPIRPHYFWPLLVGRRKEKNVQGIIEILKGMQELGVHPDQETYTDY
VIPCFDSVNSARAILQENGCLSDSDMFSQAGLRSEAANGNLDFVLSFLKSNTLPISLQSI
RSSLLLGFRRSMNINLWSEITELLYKDGRYCQEPRGPTEAVGYFLYNLIDSMSDSEVQAK
EEHLRQYFHQLEKMNVKIPENIYRGIRNLLESYHVPELIKDAHLLVESKNLDFQKTVQLT
SSELESTLETLKAENQPIRDVLKQLILVLCSEENMQKALELKAKYESDMVTGGYAALINL
CCRHDKVEDALNLKEEFDRLDSSAVLDTGKYVGLVRVLAKHGKLQDAINILKEMKEKDVL
IKDTTALSFFHMLNGAALRGEIETVKQLHEAIVTLGLAEPSTNISFPLVTVHLEKGDLST
ALEVAIDCYEKYKVLPRIHDVLCKLVEKGETDLIQKAMDFVSQEQGEMVMLYDLFFAFLQ
TGNYKEAKKIIETPGIRARSARLQWFCDRCVANNQVETLEKLVELTQKLFECDRDQMYYN
```

protein sequence database

in silico digest

YSFVATAER

HETSINGK

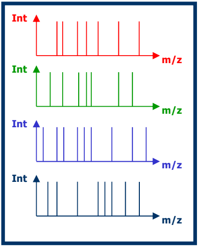
MILQEESTVYYR

SEFASTPINK

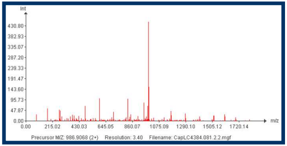
...

peptide sequences

in silico MS/MS



theoretical MS/MS spectra



experimental MS/MS spectrum

in silico matching

1)	YSFVATAER	34
2)	YSFVSAIR	12
3)	FFLIGGGGK	2

peptide scores

(slide courtesy to Lennart Martens)

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

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

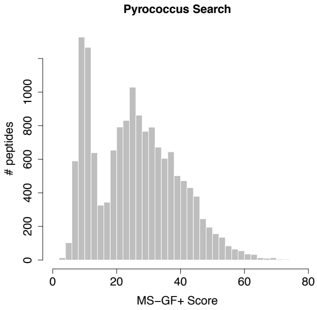
$FDP = \frac{FP}{FP+TP}$. But is unkown! (FDP: false discovery proportion)

Table of Outcomes

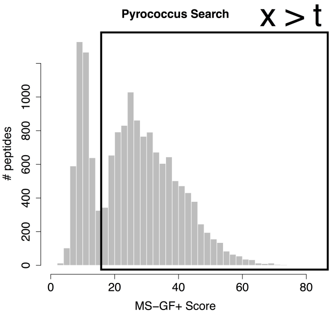
		Called Bad	Called Correct	
Unobservable	Bad hit	TN	FP	m_0
	Correct hit	FN	TP	m_1
Observable	Total	NR	R	m

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

Search engines return score that discriminates good from bad matches

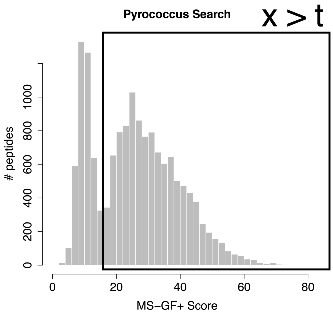


Search engines return score that discriminates good from bad matches



Score threshold t?

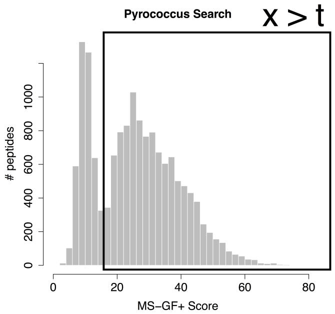
Search engines return score that discriminates good from bad matches



Score threshold t?

$$f(x) = \color{blue}{\pi_0 f_0(x)} + \color{red}{(1-\pi_0) f_1(x)}$$

Search engines return score that discriminates good from bad matches

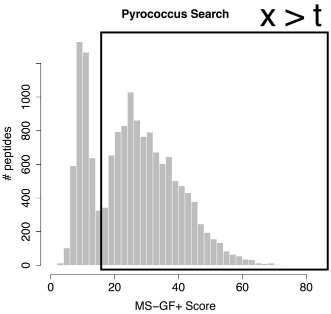


Score threshold t?

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

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

Search engines return score that discriminates good from bad matches



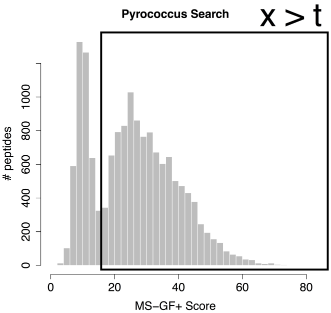
Score threshold t?

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

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

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

Search engines return score that discriminates good from bad matches



Score threshold t?

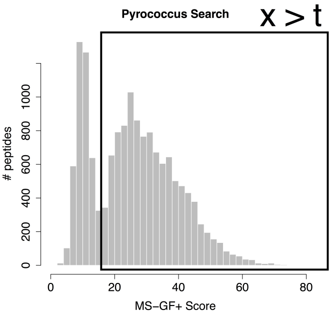
$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x)$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

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

$$\text{FDR}(t) = \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

Search engines return score that discriminates good from bad matches



Score threshold t?

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

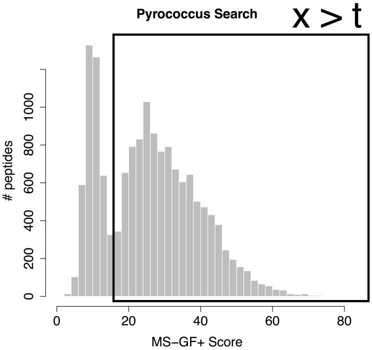
$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

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

$$\text{FDR}(t) = \frac{\pi_0 P_0[x\geq t]}{P[x\geq t]}$$

$$P_.[x \geq t] = \int_t^\infty f_.(x)$$

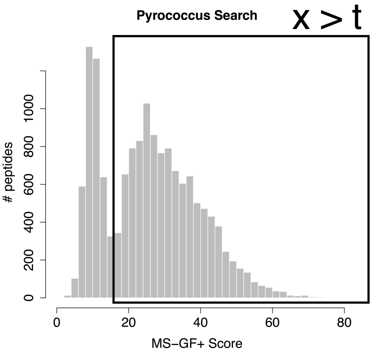
How to estimate FDR?



$$P_{.}[x \geq t] = \int_{t}^{\infty} f_{.}(x)$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right] = \frac{m\pi_0 P[x \geq t|FP]}{mP[x \geq t]}$$

How to estimate FDR?



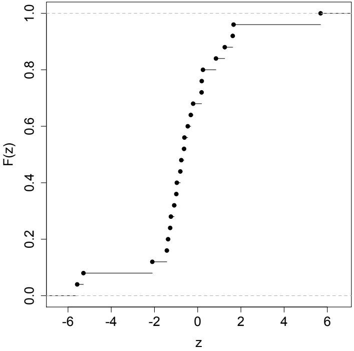
$$P_.[x \geq t] = \int_t^{\infty} f_.(x)$$

$$P_.[x \geq t] \approx \frac{\#x \geq t}{m}$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP + TP}\right] = \frac{m\pi_0 P[x \geq t|FP]}{mP[x \geq t]}$$

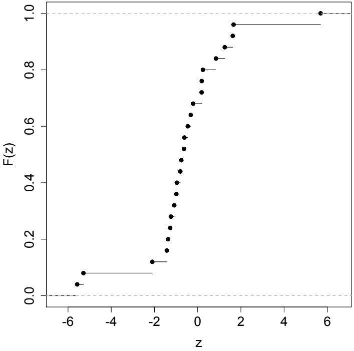
$$\widehat{FDR}(t) = \frac{m\pi_0 P[x \geq t|FP]}{\#x \geq t}$$

- $F(t) = \int\limits_{-\infty}^{t} f(x)$ using the Empirical cumulative distribution function (ECDF): $\bar{F}(t)$



$$\rightarrow \widehat{FDR}(t) = \frac{m\pi_0 P[x \geq t|FP]}{\#x \geq t} = \frac{\pi_0\,[1 - F(t)]}{1 - \bar{F}(t)}$$

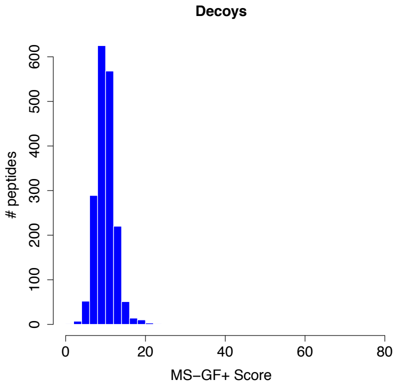
- $F(t) = \int_{-\infty}^{t} f(x)$ using the Empirical cumulative distribution function (ECDF): $\bar{F}(t)$



$$\rightarrow \widehat{FDR}(t) = \frac{m\pi_0 P[x \geq t|FP]}{\#x \geq t} = \frac{\pi_0\left[1 - F(t)\right]}{1 - \bar{F}(t)}$$

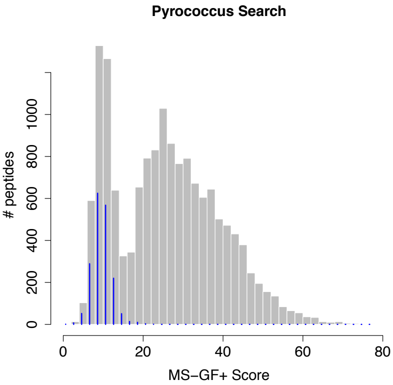
- How to characterize $F_0(t)$ and π_0 in proteomics?

Target-Decoy approach to establish null distribution



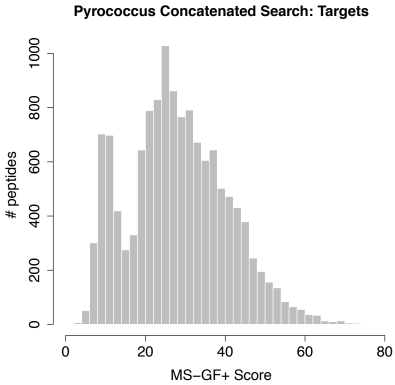
- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice

Target-Decoy approach to establish null distribution



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

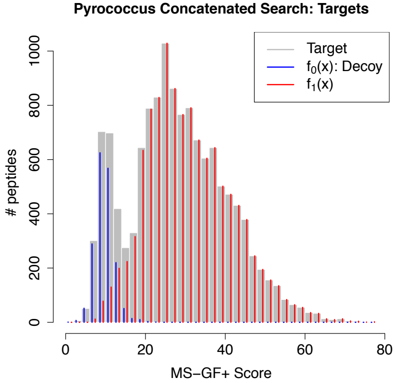
Target-Decoy approach to establish null distribution



- 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)

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

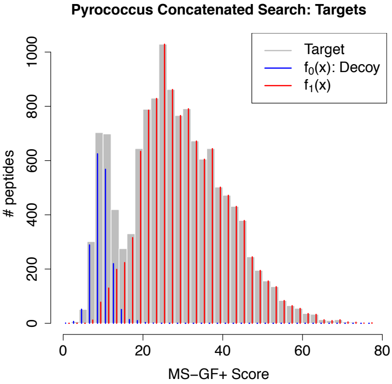
Target-Decoy approach to establish null distribution



- 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



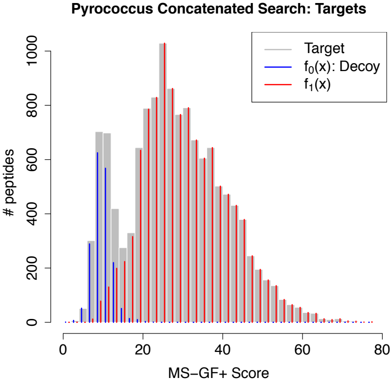
- Score cuttoff?

$$\text{FDR}(x) = E\left[\frac{FP}{FP + TP}\right]$$

- Competitive Target - decoy:

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

Target-Decoy approach to establish null distribution



- Score cuttoff?

$$\text{FDR}(x) = E\left[\frac{FP}{FP+TP}\right]$$

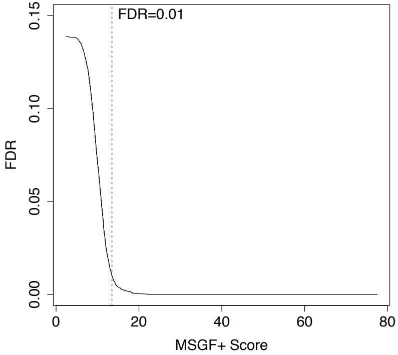
- Competitive Target - decoy:

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

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

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

Target-Decoy approach to establish null distribution



- Score cuttoff?

$$\text{FDR}(x) = E\left[\frac{FP}{FP + TP}\right]$$

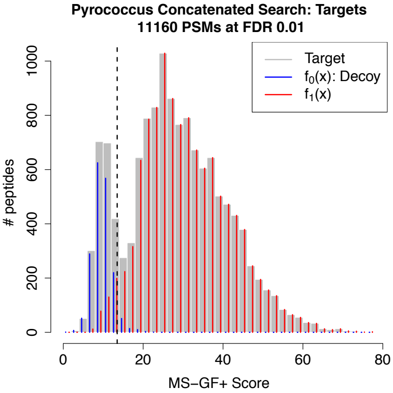
- Competitive Target - decoy:

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

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

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

Target-Decoy approach to establish null distribution



- Score cuttoff?

$$\mathrm{FDR}(x) = E\left[\frac{FP}{FP+TP}\right]$$

- Competitive Target - decoy:

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

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

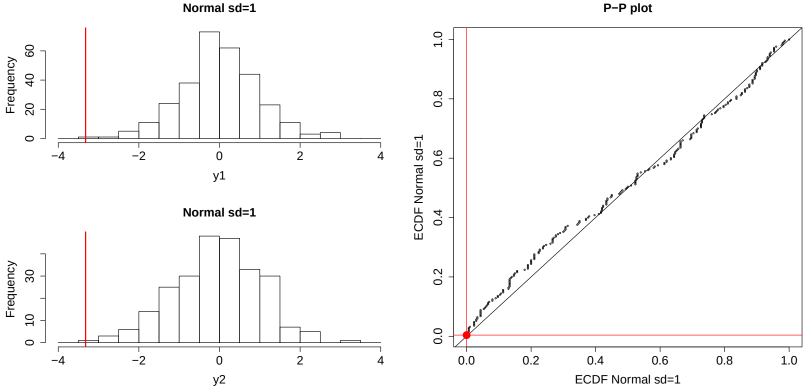
$$\widehat{\mathrm{FDR}}(x) = \hat{\pi}_0\frac{1-\bar{F}_0(x)}{1-\bar{F}(x)}$$

We have to evaluate that

- The decoys are good simulations of the targets: compare $\bar{F}_0(x)$ with $\bar{F}(x)$
- $\hat{\pi}_0 = \frac{\#decoys}{\#targets}$ is 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.

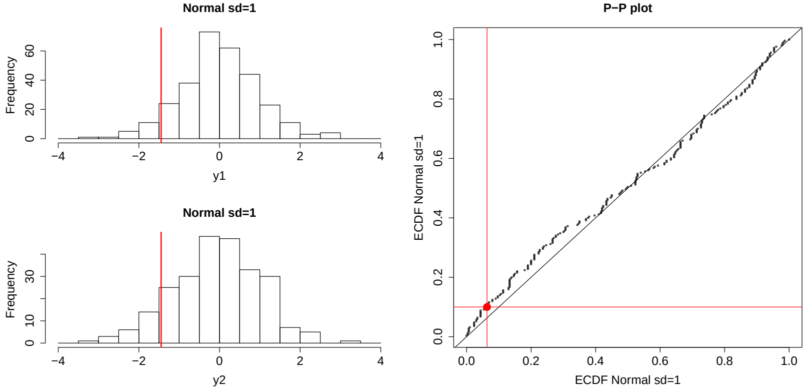
PP-plot

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



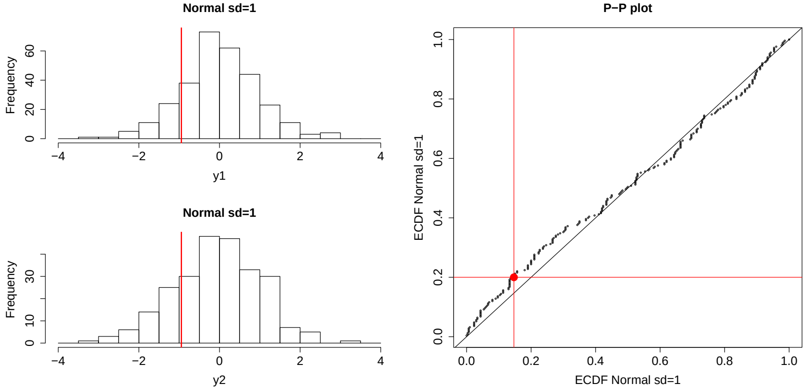
PP-plot

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



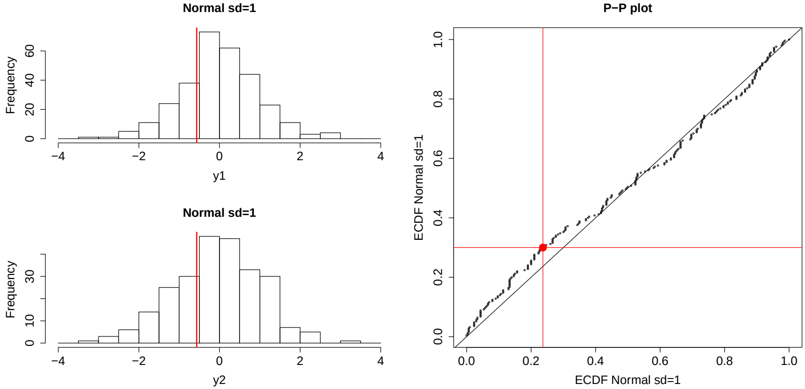
PP-plot

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



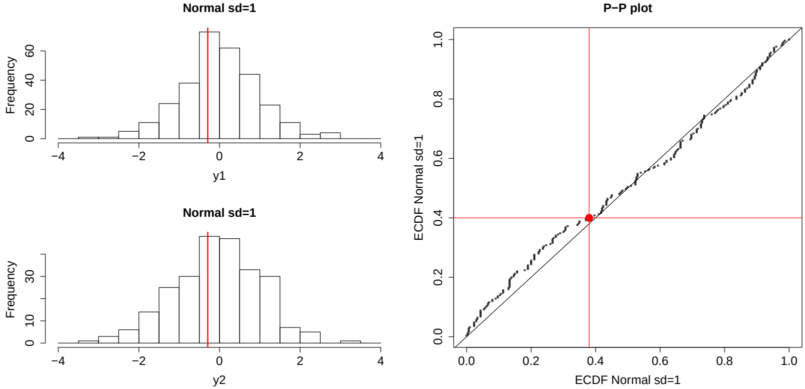
PP-plot

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



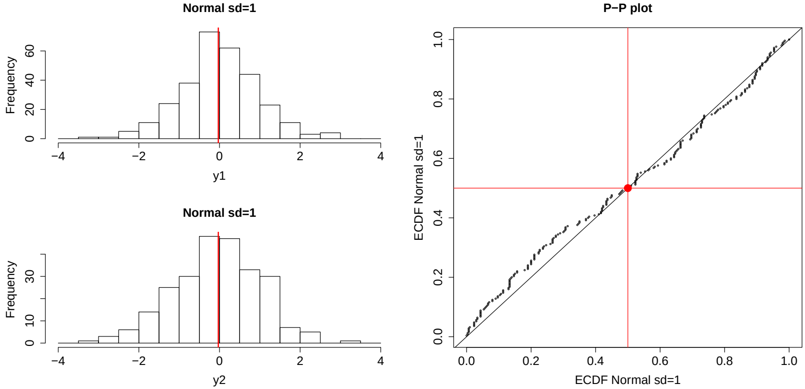
PP-plot

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



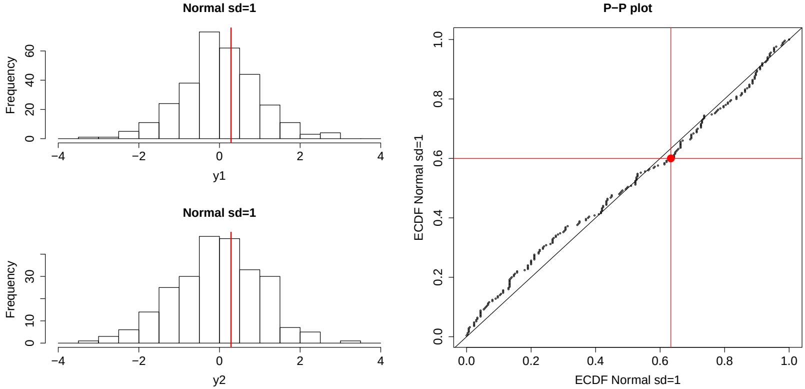
PP-plot

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



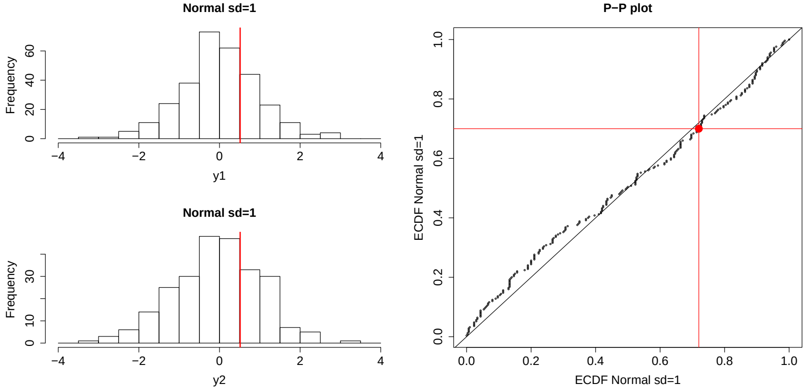
PP-plot

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



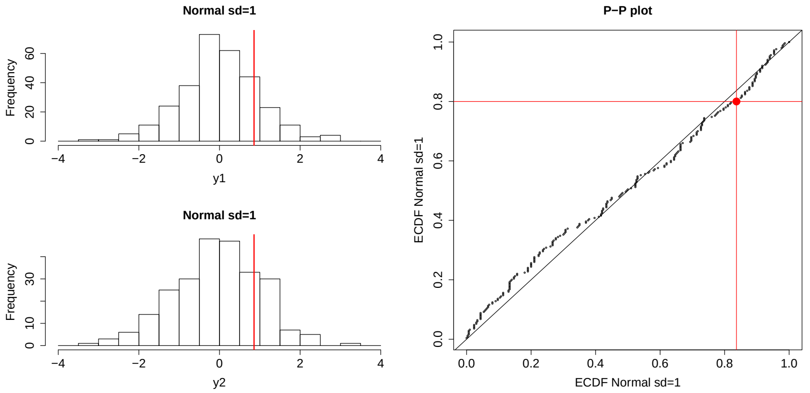
PP-plot

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



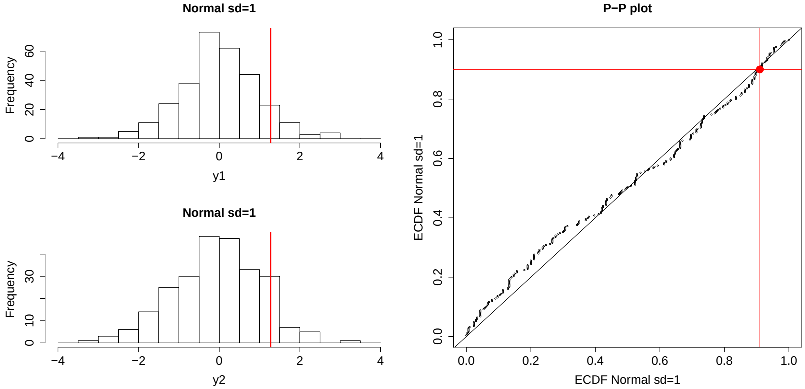
PP-plot

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



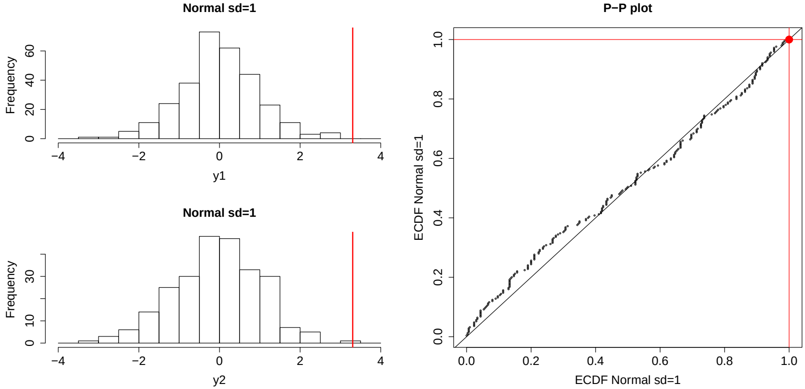
PP-plot

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

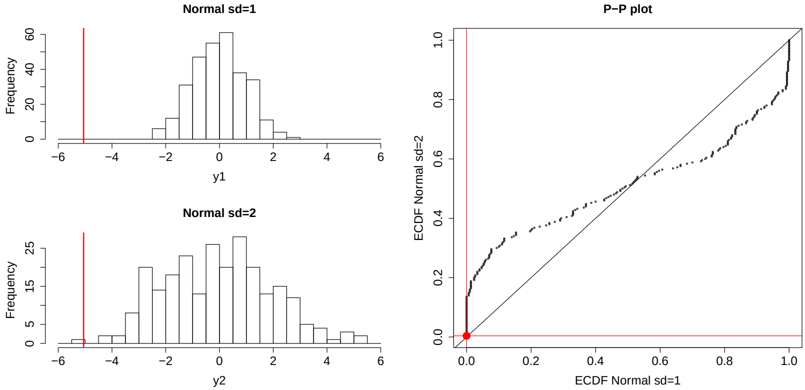


PP-plot

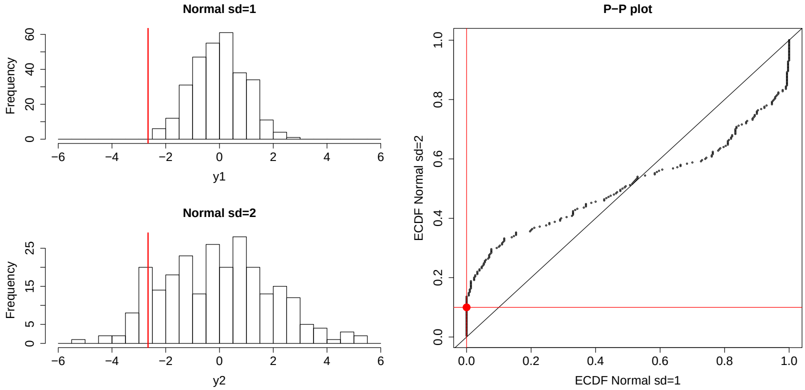
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



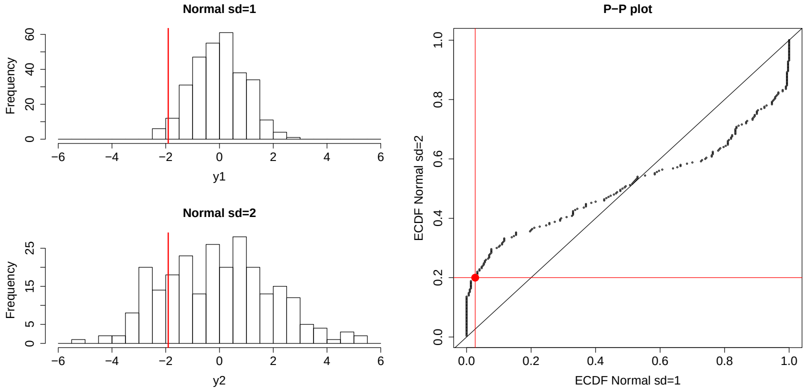
PP-plot


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

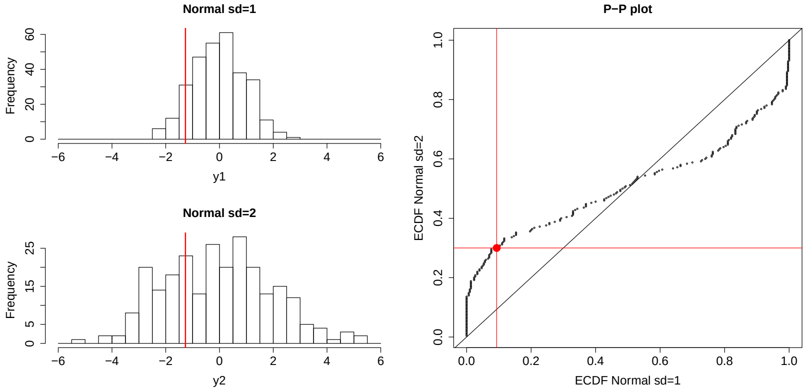
PP-plot


Normal sd=1
Frequency
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
0.0
0.2
0.4
0.6
0.8
1.0

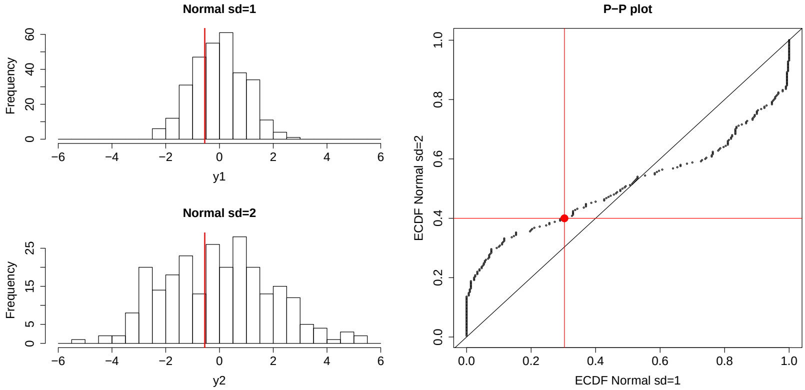
PP-plot


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

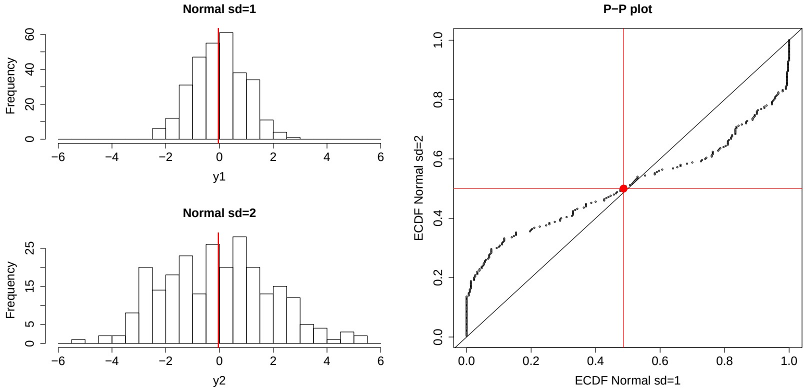
PP-plot


Normal sd=1
Frequency
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
0.0
0.2
0.4
0.6
0.8
1.0

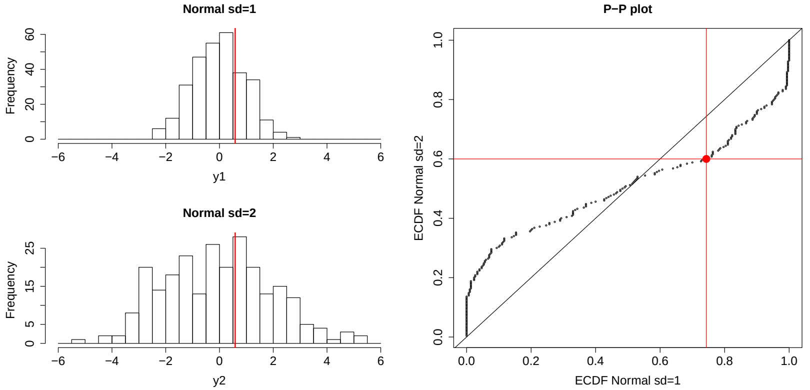
PP-plot


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

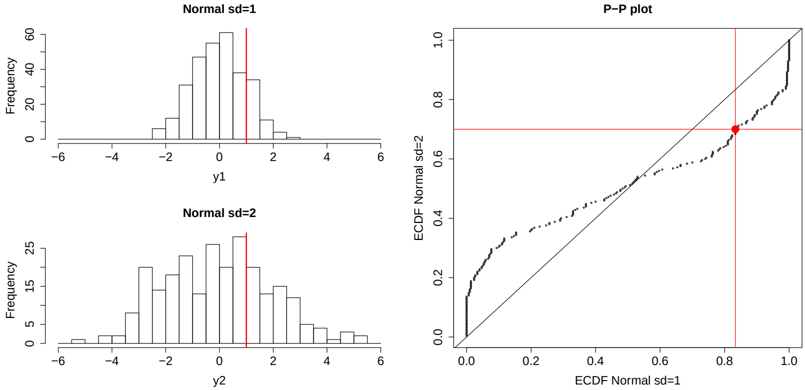
PP-plot


Normal sd=1
Frequency
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
0.0
0.2
0.4
0.6
0.8
1.0

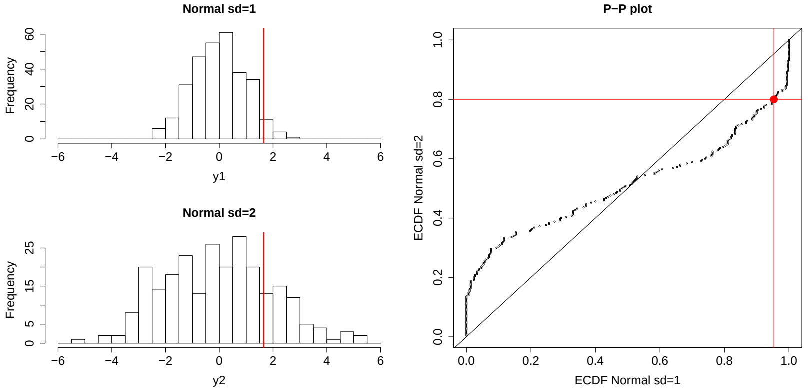
PP-plot


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
0.0
0.2
0.4
0.6
0.8
1.0

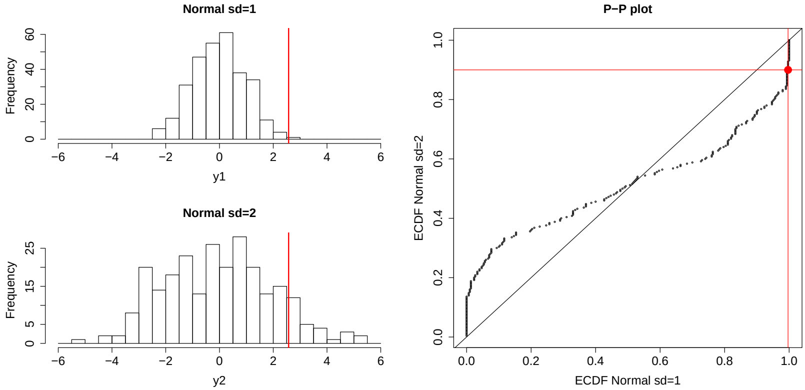
PP-plot


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
0.0
0.2
0.4
0.6
0.8
1.0

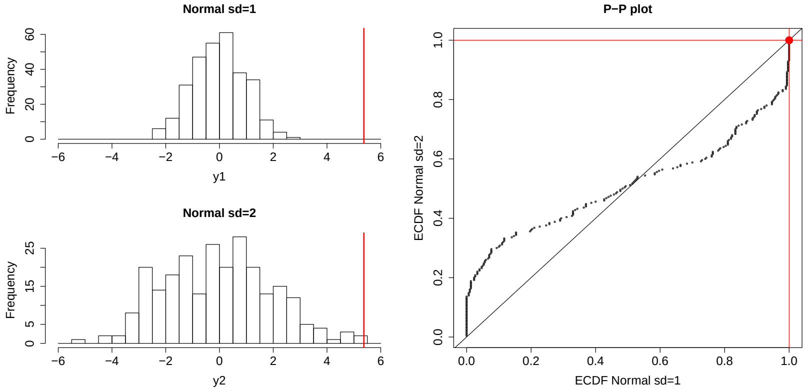
PP-plot


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
0.0
0.2
0.4
0.6
0.8
1.0

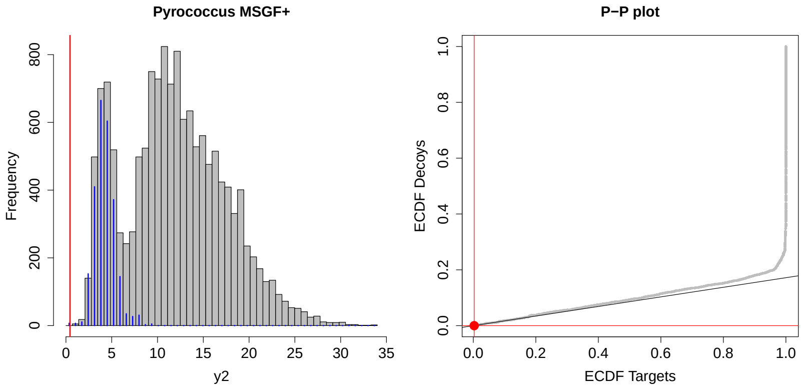
PP-plot


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

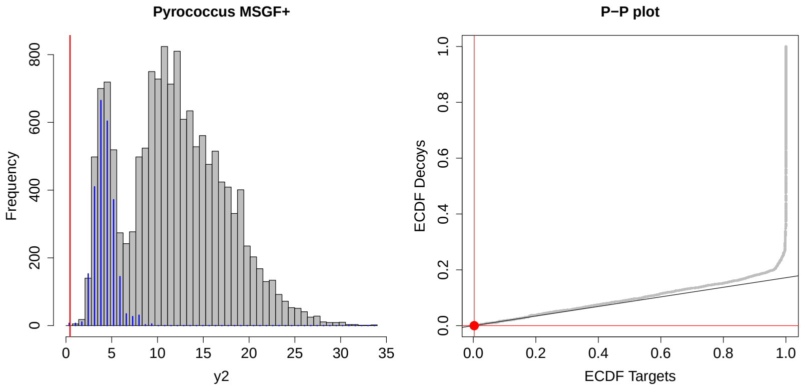
PP-plot


Normal sd=1
Frequency
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
0.0
0.2
0.4
0.6
0.8
1.0

PP-plot: pyrococcus

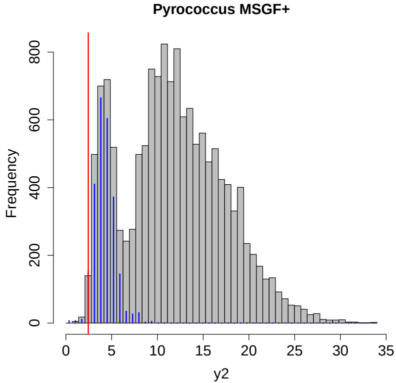


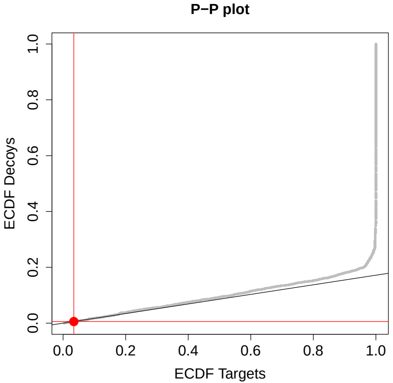
PP-plot: pyrococcus



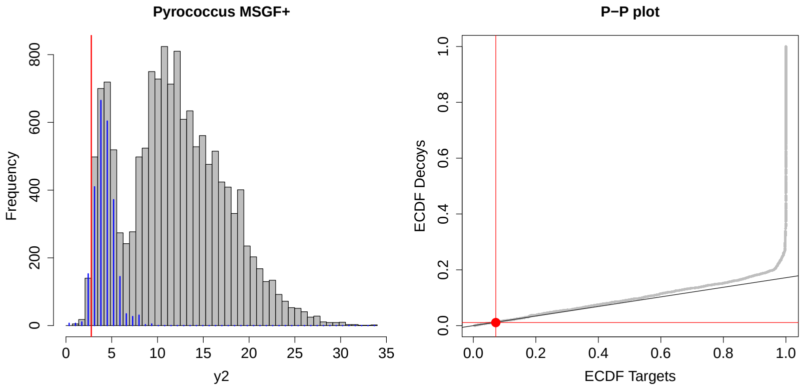
What about $\hat{\pi}_0$?

PP-plot: pyrococcus

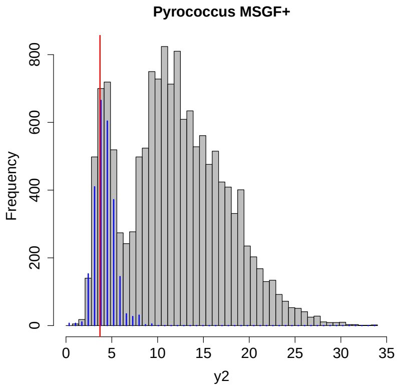


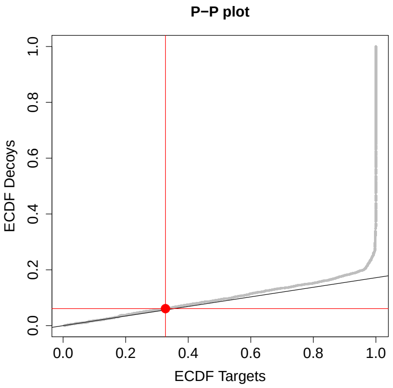


PP-plot: pyrococcus

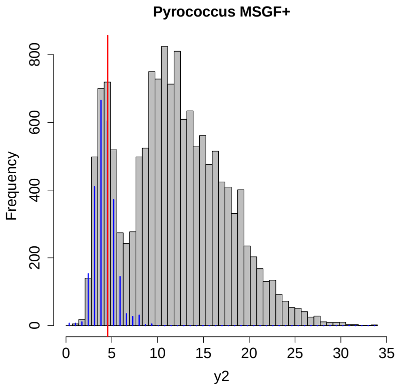


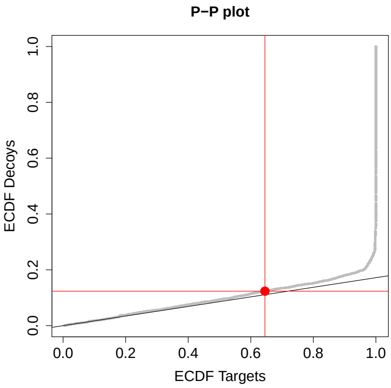
PP-plot: pyrococcus



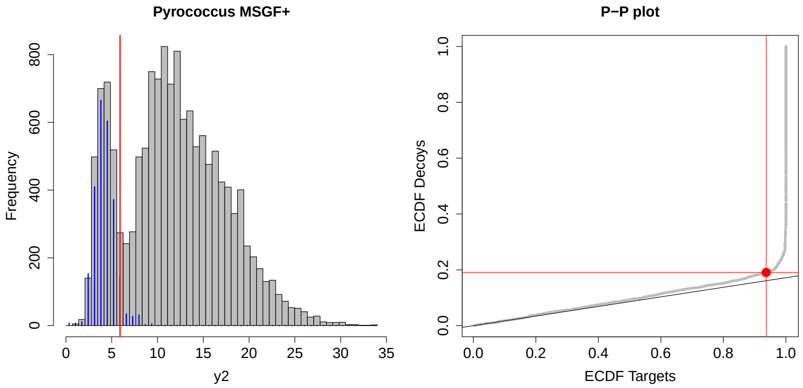


PP-plot: pyrococcus

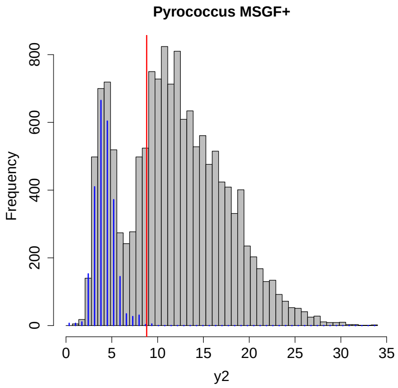


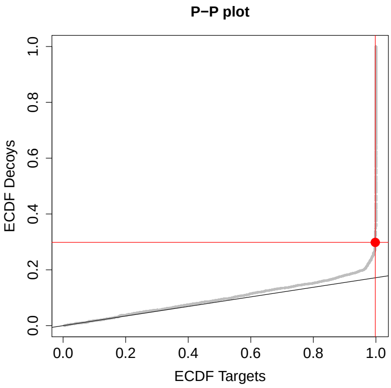


PP-plot: pyrococcus

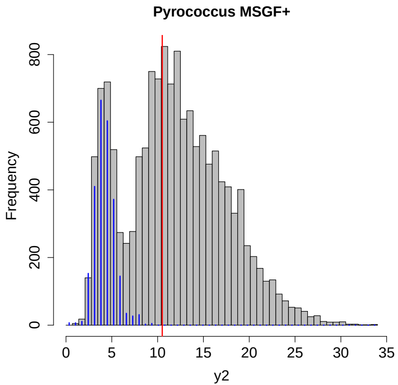


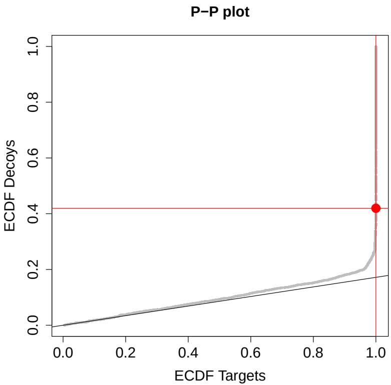
PP-plot: pyrococcus



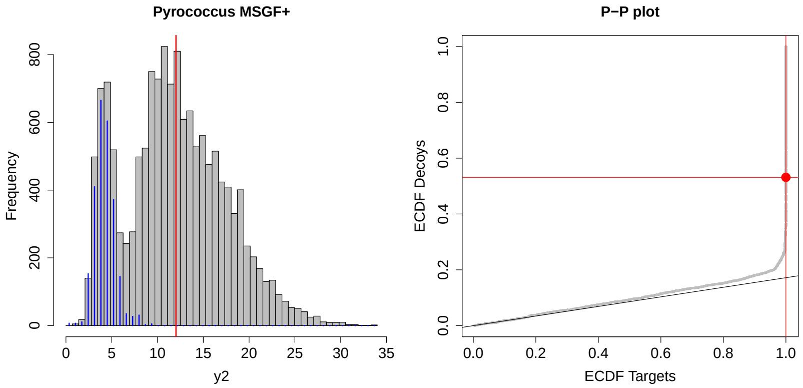


PP-plot: pyrococcus

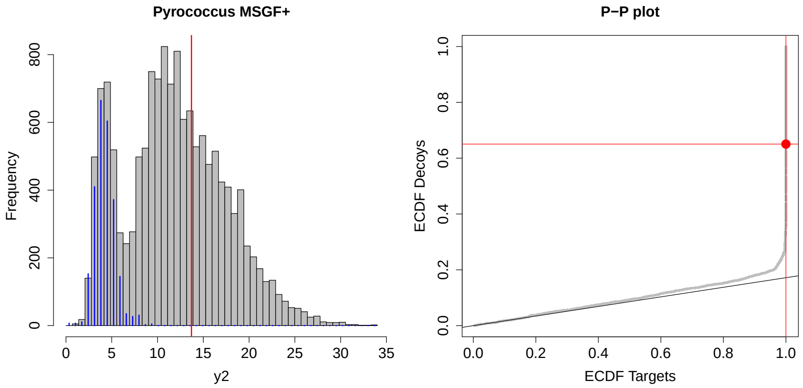




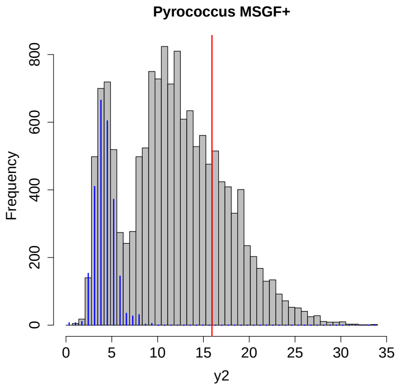
PP-plot: pyrococcus

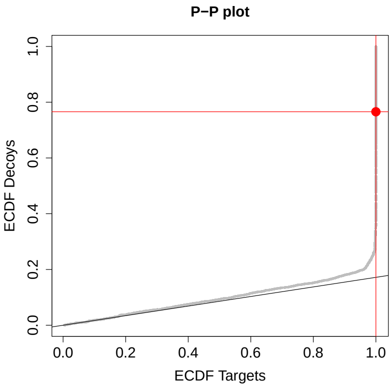


PP-plot: pyrococcus

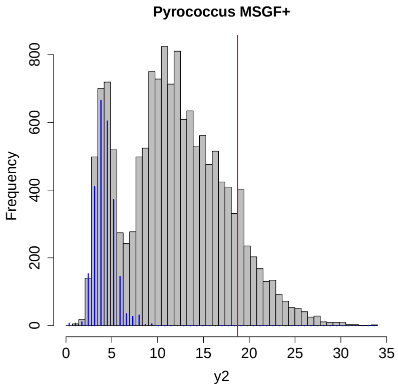


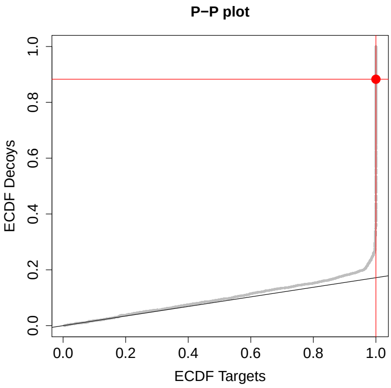
PP-plot: pyrococcus



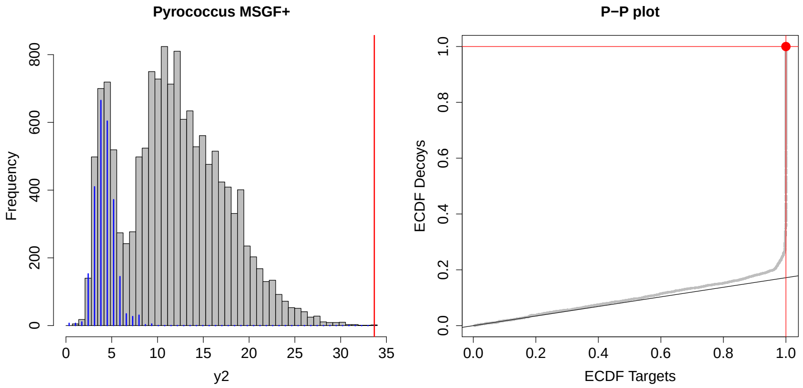


PP-plot: pyrococcus





PP-plot: pyrococcus


Pyrococcus MSGF+
Frequency
y2
P−P plot
ECDF Decoys
ECDF Targets