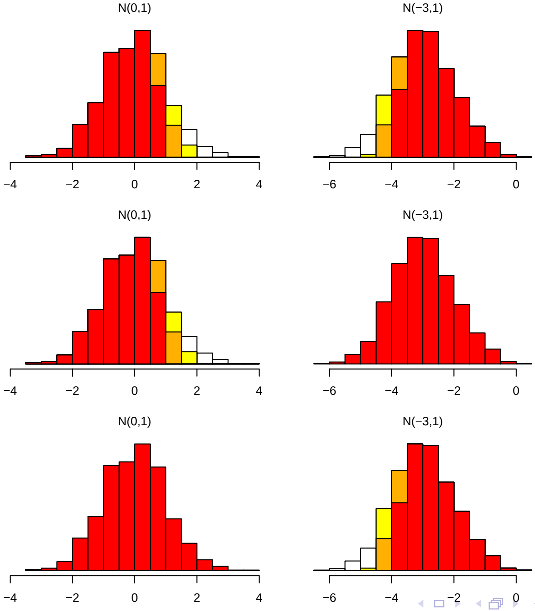
Statistical Applications of Over-fitting due to Trimmings

Pedro C. Alvarez
(joint work with E. del Barrio, J.A. Cuesta-Albertos and C. Matrán)

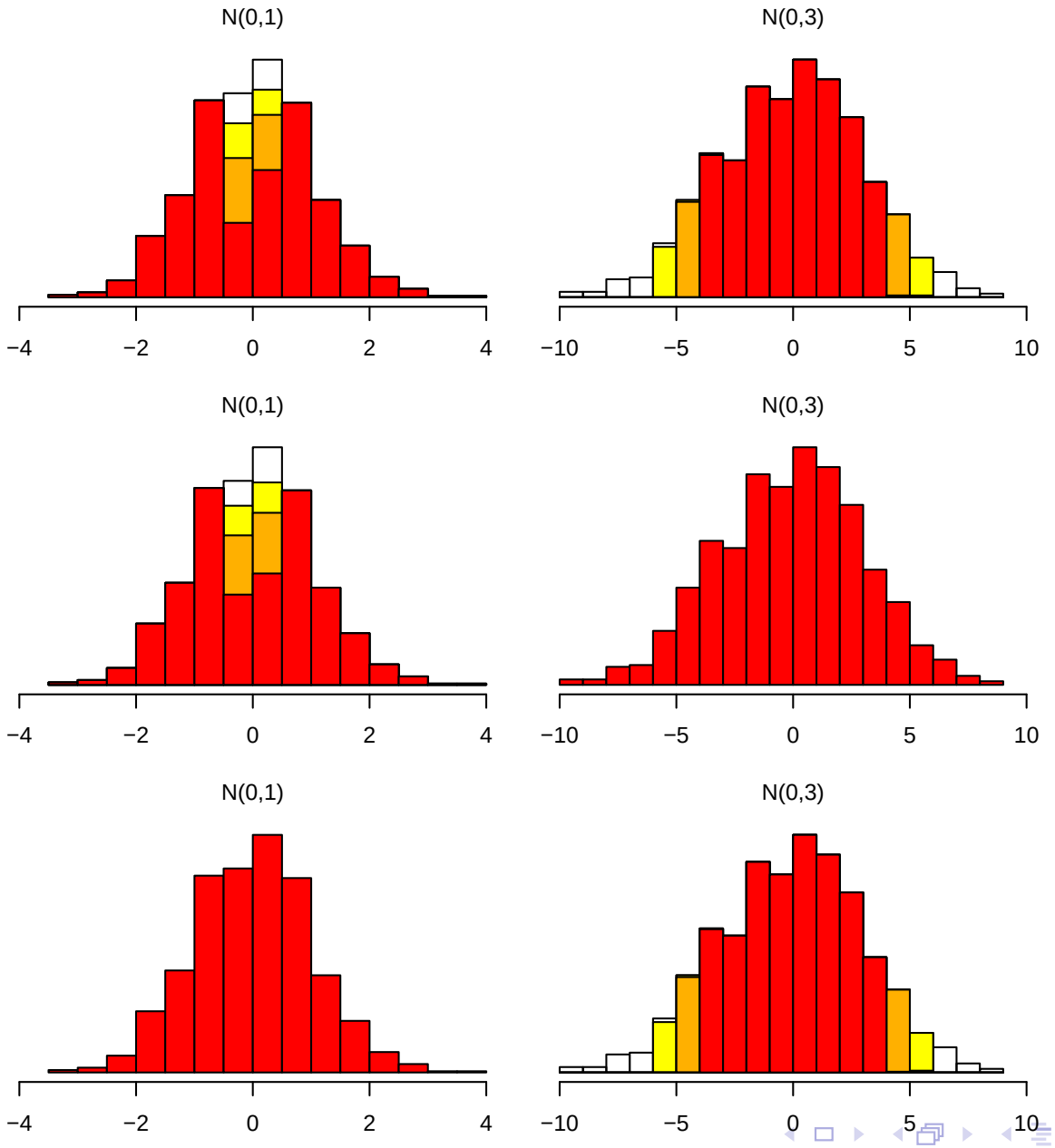
5èmes Rencontre de Statistiques Mathématiques BoSanTouVal09,
Parc du Teich (Bordeaux, France), du 3 au 5 juin 2009

Outline

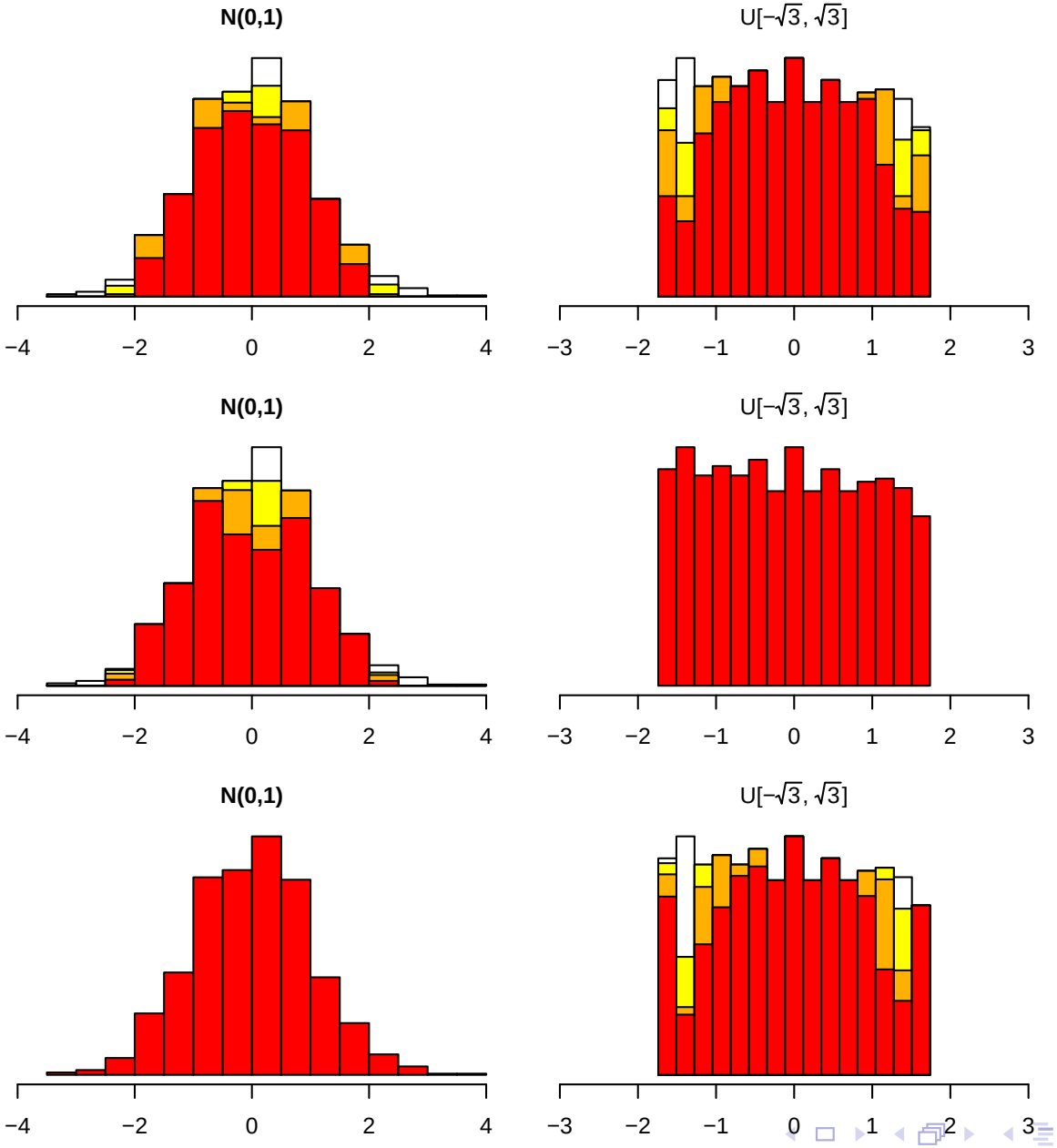
How impartial trimmings work?: N(0,1) vs N(-3,1)



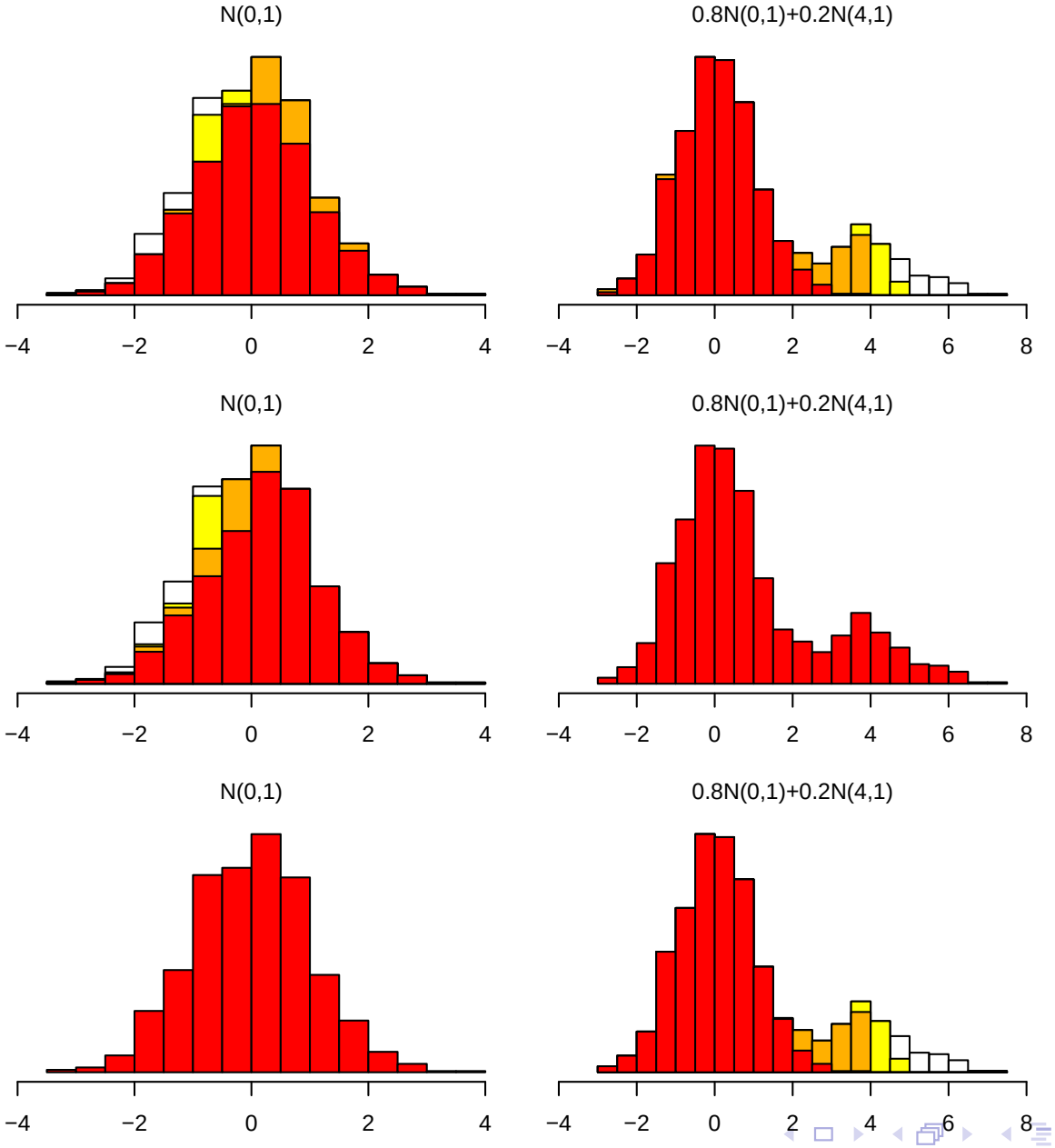
How impartial trimmings work?: N(0,1) vs N(0,3)



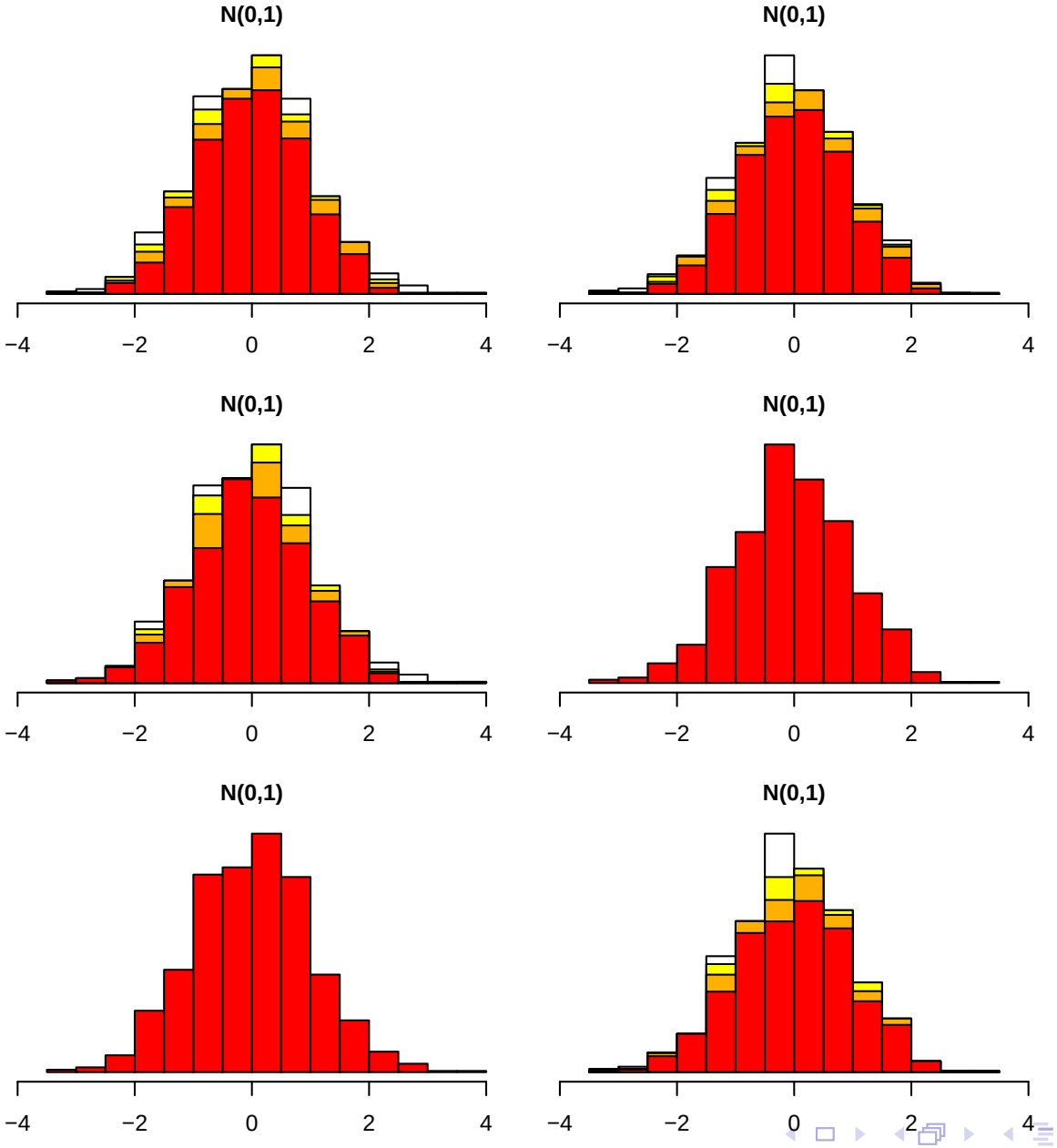
How impartial trimmings work?: N(0,1) vs U($-\sqrt{3},\sqrt{3}$)



How impartial .. work?: N(0,1) vs 0.8*N(0,1)+0.2*N(4,1)

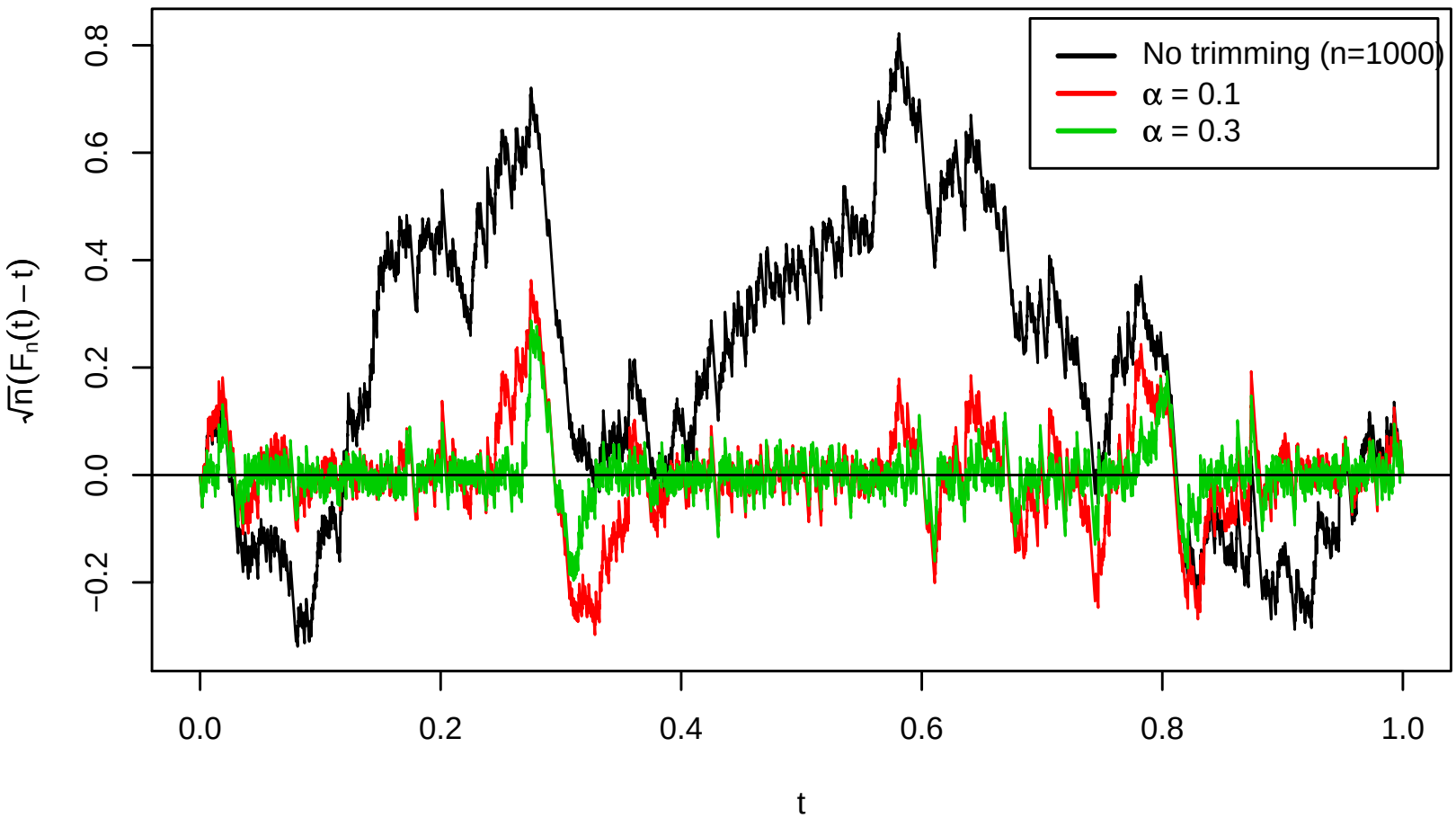


How impartial trimmings work?: N(0,1) vs N(0,1)



Overfitting.

Example(one sample problem): 1000 observations from U[0,1] vs U[0,1]



Overfitting.

Let $P \in \mathcal{P}_2(\mathbb{R})$, $P \ll \ell$, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ i.i.d $\sim P$. P_n and Q_n the respective empirical distributions,

	1 sample	
No trimming	$n\mathcal{W}_2^2(P_n, P) = O_P(1)$	
Trimming	$n\mathcal{W}_2^2(P_{n,\alpha}, P) = o_P(1)$	

Overfitting.

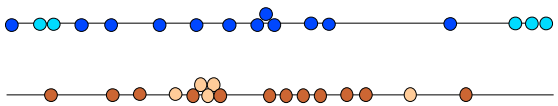
Let $P \in \mathcal{P}_2(\mathbb{R})$, $P \ll \ell$, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ i.i.d $\sim P$. P_n and Q_n the respective empirical distributions,

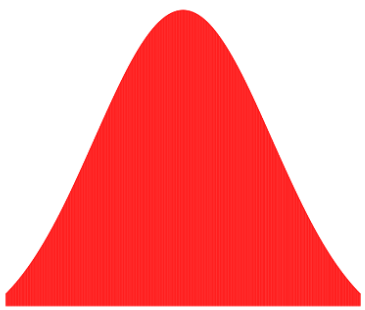
	1 sample	2 samples
No trimming	$n\mathcal{W}_2^2(P_n, P) = O_P(1)$	$n\mathcal{W}_2^2(P_n, Q_n) = O_P(1)$
Trimming	$n\mathcal{W}_2^2(P_{n,\alpha}, P) = o_P(1)$	$n\mathcal{W}_2^2(P_{n,\alpha}, Q_{n,\alpha}) = o_P(1)$

What does it mean? how we can take advantage of it?

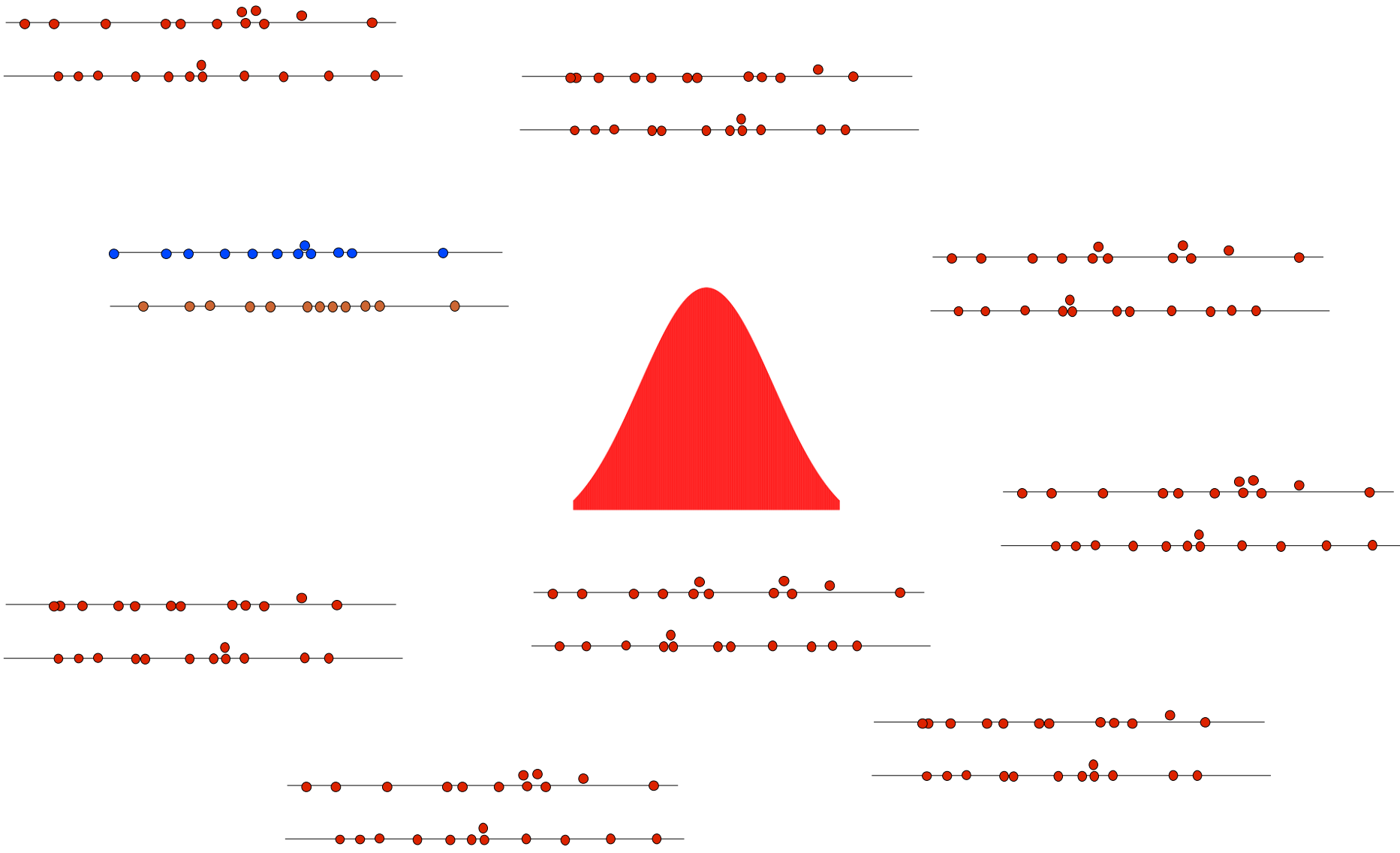
$$\mathcal{W}_2(P_n, Q_n) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$$

Overfitting.

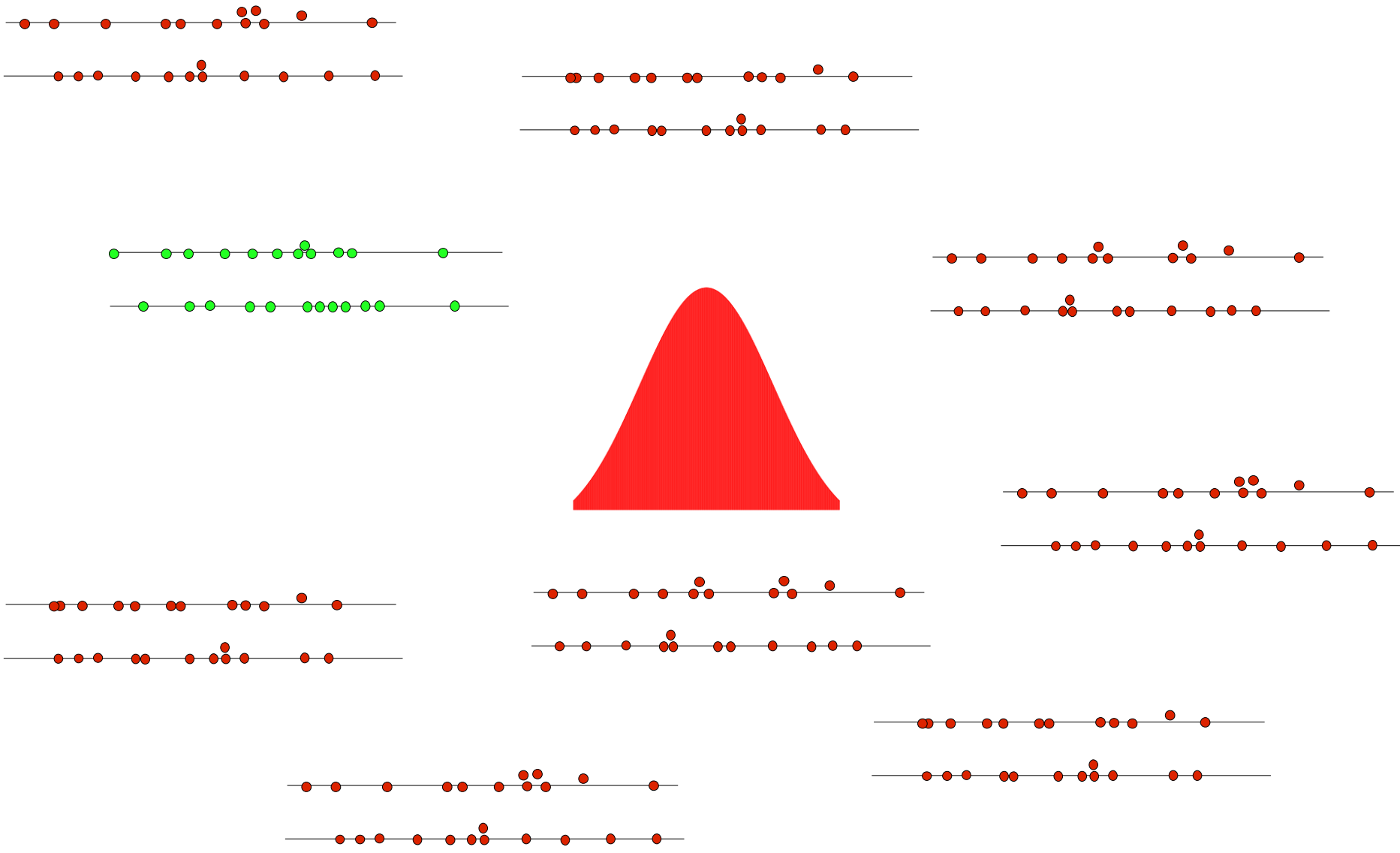




Overfitting.



Overfitting.



Overfitting: numerical example.

1. Generate two random samples from a $N(0,1)$ of size n, then trim them (α) and compute the $\mathcal{W}_2$ distance between the trimmed samples (of size $[n(1-\alpha)]$): $\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$.
2. Generate $R = 1000$ pairs of random samples from a $N(0,1)$ of size $m = [n(1-\alpha)]$ and compute $\mathcal{W}_2(P_m^i, Q_m^i)$, $i = 1, \ldots, 1000$.
3. Calculate the frequency of "$\mathcal{W}_2(P_m^i, Q_m^i) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$" (p-value).

(1): Pair of samples n.1 (n=100),

	trimming size							
	0%	1%	2%	3%	4%	5%	10%	15%
(1)	0.199	0.447	0.738	0.914	0.985	1	1	1

Overfitting: numerical example.

1. Generate two random samples from a $N(0,1)$ of size n, then trim them (α) and compute the $\mathcal{W}_2$ distance between the trimmed samples (of size $[n(1-\alpha)]$): $\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$.
2. Generate $R = 1000$ pairs of random samples from a $N(0,1)$ of size $m = [n(1-\alpha)]$ and compute $\mathcal{W}_2(P_m^i, Q_m^i)$, $i = 1, \ldots, 1000$.
3. Calculate the frequency of "$\mathcal{W}_2(P_m^i, Q_m^i) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$" (p-value).

(1): Pair of samples n.1 (n=100),
(2): Pair of samples n.2 (n=100),

	trimming size							
	0%	1%	2%	3%	4%	5%	10%	15%
(1)	0.199	0.447	0.738	0.914	0.985	1	1	1
(2)	0.020	0.030	0.062	0.109	0.235	0.383	0.978	1

Overfitting: numerical example.

1. Generate two random samples from a $N(0,1)$ of size n, then trim them (α) and compute the $\mathcal{W}_2$ distance between the trimmed samples (of size $[n(1-\alpha)]$): $\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$.
2. Generate $R = 1000$ pairs of random samples from a $N(0,1)$ of size $m = [n(1-\alpha)]$ and compute $\mathcal{W}_2(P_m^i, Q_m^i)$, $i = 1, \ldots, 1000$.
3. Calculate the frequency of "$\mathcal{W}_2(P_m^i, Q_m^i) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$" (p-value).

(1): Pair of samples n.1 (n=100),
(2): Pair of samples n.2 (n=100),
(3): Pair of samples n.3 (n=1000).

	trimming size							
	0%	1%	2%	3%	4%	5%	10%	15%
(1)	0.199	0.447	0.738	0.914	0.985	1	1	1
(2)	0.020	0.030	0.062	0.109	0.235	0.383	0.978	1
(3)	0.364	0.918	1	1	1	1	1	1

Overfitting: numerical example.

1. Generate two random samples from a $N(0,1)$ of size n, then trim them (α) and compute the $\mathcal{W}_2$ distance between the trimmed samples (of size $[n(1-\alpha)]$): $\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$.
2. Generate $R = 1000$ pairs of random samples from a $N(0,1)$ of size $m = [n(1-\alpha)]$ and compute $\mathcal{W}_2(P_m^i, Q_m^i)$, $i = 1, \ldots, 1000$.
3. Calculate the frequency of "$\mathcal{W}_2(P_m^i, Q_m^i) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$" (p-value).
4. Repeat 1-3 100 times and compute the median p-value:

	trimming size							
n	0%	1%	2%	3%	4%	5%	10%	15%
100	0.433	0.703	0.850	0.932	0.976	0.993	1	1
300	0.513	0.880	0.977	0.997	1	1	1	1
1000	0.499	0.986	1	1	1	1	1	1

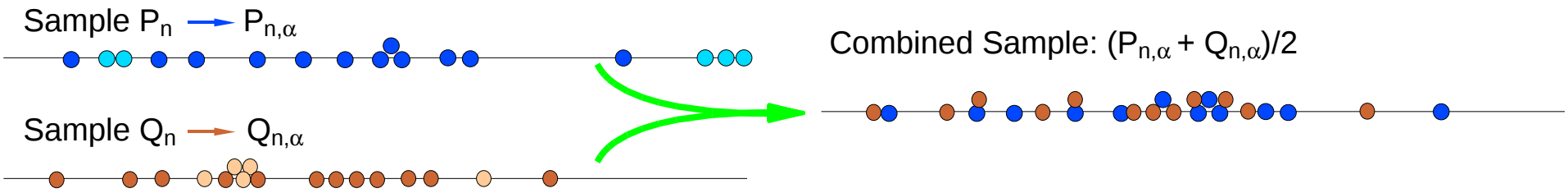
Overfitting: numerical example.

1. Generate two random samples from a $N(0,1)$ of size n, then trim them (α) and compute the $\mathcal{W}_2$ distance between the trimmed samples (of size $[n(1-\alpha)]$): $\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$.
2. Generate $R = 1000$ pairs of random samples from a $N(0,1)$ of size $m = [n(1-\alpha)]$ and compute $\mathcal{W}_2(P_m^i, Q_m^i)$, $i = 1, \ldots, 1000$.
3. Calculate the frequency of "$\mathcal{W}_2(P_m^i, Q_m^i) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$" (p-value).
4. Repeat 1-3 100 times and compute the median p-value:

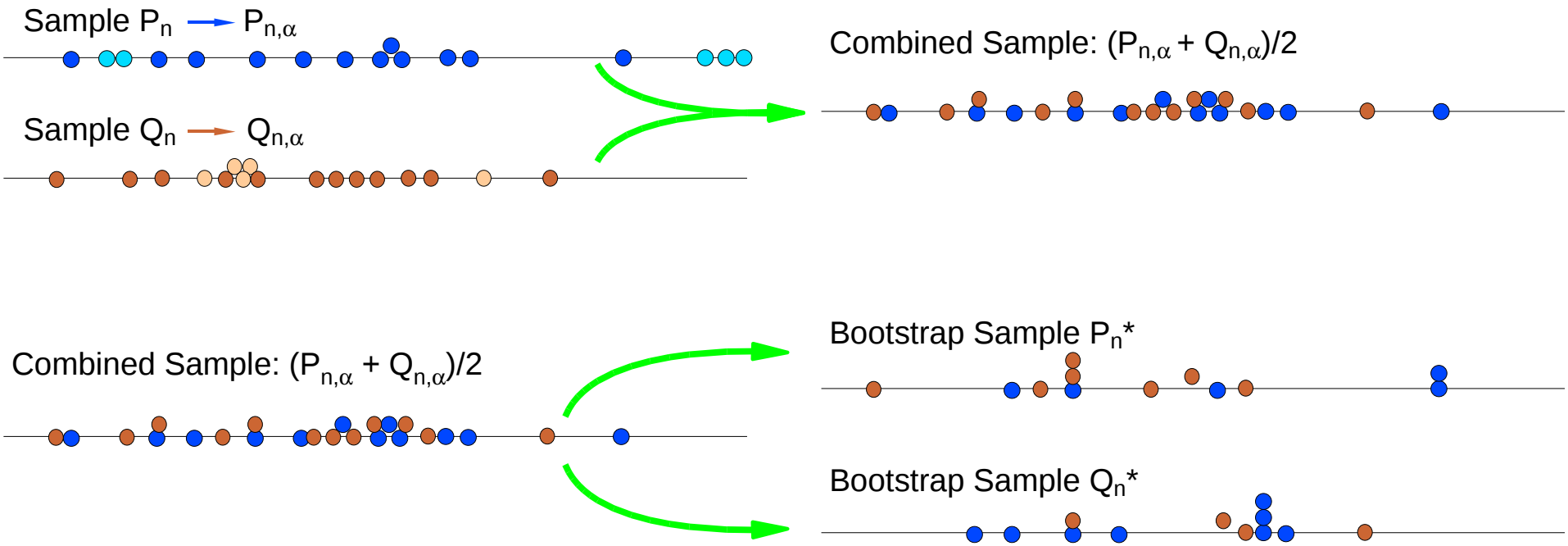
	trimming size							
n	0%	1%	2%	3%	4%	5%	10%	15%
100	0.433	0.703	0.850	0.932	0.976	0.993	1	1
300	0.513	0.880	0.977	0.997	1	1	1	1
1000	0.499	0.986	1	1	1	1	1	1

But, in practice we don't know the true model...

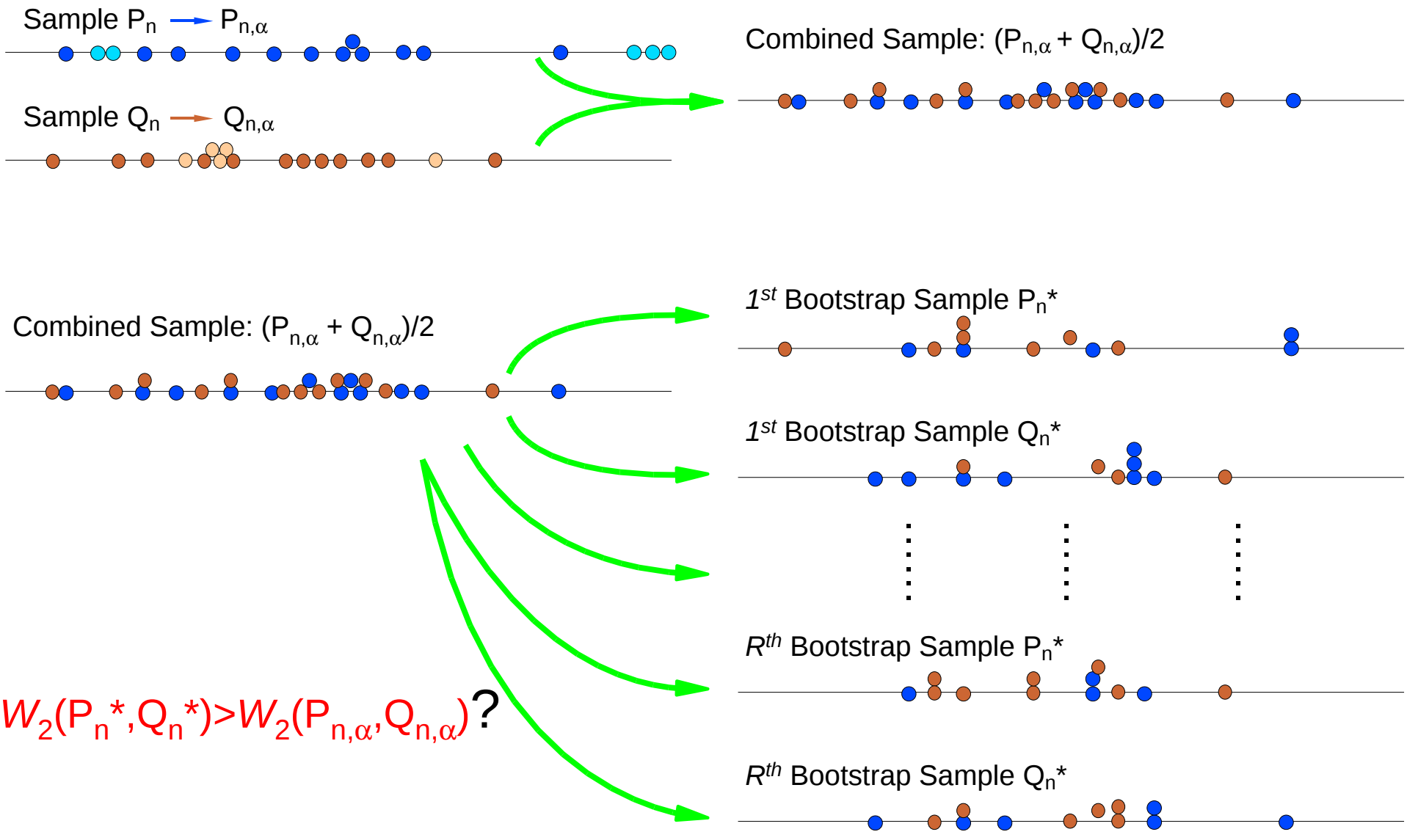
Overfitting and Bootstrap.



Overfitting and Bootstrap.



Overfitting and Bootstrap.



Bootstrap: revisiting the numerical example.

1. Compute the **bootstrap p-value** as the frequency of times:

$$\text{“}\mathcal{W}_2(P_m^*, Q_m^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\text{”}.$$

Bootstrap: revisiting the numerical example.

1. Compute the **bootstrap p-value** as the frequency of times:

$$\text{“}\mathcal{W}_2(P_m^*, Q_m^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\text{”}.$$

2. Repeat the previous process for 100 pairs of N(0,1) samples of size n and compute the median p-value, then

	trimming size							
n	0%	1%	2%	3%	4%	5%	10%	15%
100	0.580	0.755	0.846	0.914	0.964	0.982	1	1
300	0.490	0.820	0.955	0.994	1	1	1	1
1000	0.577	0.973	1	1	1	1	1	1

Bootstrap: revisiting the numerical example.

1. Compute the **bootstrap p-value** as the frequency of times:

$$\text{“}\mathcal{W}_2(P_m^*, Q_m^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\text{”}.$$

2. Repeat the previous process for 100 pairs of N(0,1) samples of size n and compute the median p-value, then

	trimming size							
n	0%	1%	2%	3%	4%	5%	10%	15%
100	0.580	0.755	0.846	0.914	0.964	0.982	1	1
300	0.490	0.820	0.955	0.994	1	1	1	1
1000	0.577	0.973	1	1	1	1	1	1

Are these results similar to those previous to the bootstrap procedure?

Bootstrap: revisiting the numerical example.

1. Compute the **bootstrap p-value** as the frequency of times:

$$\text{“}\mathcal{W}_2(P_m^*, Q_m^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\text{”}.$$

2. Repeat the previous process for 100 pairs of N(0,1) samples of size n and compute the median p-value, then

	trimming size							
n	0%	1%	2%	3%	4%	5%	10%	15%
100	0.580	0.755	0.846	0.914	0.964	0.982	1	1
300	0.490	0.820	0.955	0.994	1	1	1	1
1000	0.577	0.973	1	1	1	1	1	1

	trimming size							
n	0%	1%	2%	3%	4%	5%	10%	15%
100	0.433	0.703	0.850	0.932	0.976	0.993	1	1
300	0.513	0.880	0.977	0.997	1	1	1	1
1000	0.499	0.986	1	1	1	1	1	1

Bootstrap.

Theorem

Let $\alpha > 0$ and $P, Q \in \mathcal{P}_{2+\delta}(\mathbb{R})$, for some $\delta > 0$. Let us suppose that P and Q have density functions f and g, with support in an interval (possibly non-bounded), with continuous derivatives,

(a) *If $\mathcal{W}_2(\mathcal{R}_{\alpha'}(P), \mathcal{R}_{\alpha'}(Q)) = 0$ for some $\alpha' \in (0, \alpha)$, then*

$$\mathbb{P}\left(\mathcal{W}_2(P_n^*, Q_n^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\right) \to 1.$$

(b) *If $\mathcal{W}_2(\mathcal{R}_{\alpha}(P), \mathcal{R}_{\alpha}(Q)) > 0$, then*

$$\mathbb{P}\left(\mathcal{W}_2(P_n^*, Q_n^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\right) \to 0.$$

Bootstrap.

Theorem

Let $\alpha > 0$ and $P, Q \in \mathcal{P}_{2+\delta}(\mathbb{R})$, for some $\delta > 0$. Let us suppose that P and Q have density functions f and g, with support in an interval (possibly non-bounded), with continuous derivatives,

(a) *If $\mathcal{W}_2(\mathcal{R}_{\alpha'}(P), \mathcal{R}_{\alpha'}(Q)) = 0$ for some $\alpha' \in (0, \alpha)$, then*

$$\mathbb{P}\left(\mathcal{W}_2(P_n^*, Q_n^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\right) \to 1.$$

(b) *If $\mathcal{W}_2(\mathcal{R}_{\alpha}(P), \mathcal{R}_{\alpha}(Q)) > 0$, then*

$$\mathbb{P}\left(\mathcal{W}_2(P_n^*, Q_n^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\right) \to 0.$$

- If $\mathcal{W}_2(\mathcal{R}_{\alpha}(P), \mathcal{R}_{\alpha}(Q)) > 0$:

$$n\mathcal{W}_2^2(P_n^*, Q_n^*) = O_P(1), \quad \text{and}$$

$$\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha}) \to_{a.s} \mathcal{W}_2(\mathcal{R}_{\alpha}(P), \mathcal{R}_{\alpha}(Q)), \text{ then } n\mathcal{W}_2^2(P_{n,\alpha}, Q_{n,\alpha}) \to \infty.$$

Assessing the similarity of two distributions

- Given two distributions, P and Q, we say that they are similar at level α if $\mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0$.

Assessing the similarity of two distributions

- Given two distributions, P and Q, we say that they are similar at level α if $\mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0$.

  or equivalently, if there exists μ (a "common part") such that

$$\begin{cases} P &= (1-\alpha)\mu + \alpha P' \\ Q &= (1-\alpha)\mu + \alpha Q' \end{cases}$$

Assessing the similarity of two distributions

- Given two distributions, P and Q, we say that they are similar at level α if $\mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0$.

  or equivalently, if there exists μ (a "common part") such that

$$\begin{cases} P &= (1-\alpha)\mu + \alpha P' \\ Q &= (1-\alpha)\mu + \alpha Q' \end{cases}$$

- We define the level of similarity between P and Q as

$$\mathcal{S}(P, Q) = \min_\alpha \{\alpha : \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0\} \qquad (= d_{TV}(P, Q)).$$

Assessing the similarity of two distributions

- Given two distributions, P and Q, we say that they are similar at level α if $\mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0$.

  or equivalently, if there exists μ (a "common part") such that

$$\begin{cases} P &= (1-\alpha)\mu + \alpha P' \\ Q &= (1-\alpha)\mu + \alpha Q' \end{cases}$$

- We define the level of similarity between P and Q as

$$\mathcal{S}(P, Q) = \min_\alpha \{\alpha : \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0\} \qquad (= d_{TV}(P, Q)).$$

- Using the previous theorem, and given $\alpha \in (0, 1)$, we can assess whether

$$\mathcal{S}(P, Q) < \alpha \Leftrightarrow \exists \alpha' \in (0, \alpha) \text{ such that } \mathcal{W}_2(\mathcal{R}_{\alpha'}(P), \mathcal{R}_{\alpha'}(Q)) = 0,$$

$$\text{or } \mathcal{S}(P, Q) > \alpha \Leftrightarrow \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) > 0.$$

Comparison of two distributions

We have generated 5 random samples: A $\sim$ N(0,1), B $\sim$ N(0,1), C $\sim$ N(1,1), D $\sim$ N(1,2) and E $\sim$ N(2,1) for two sample sizes ($n = 30, 100$).

	0%	1%	5%	10%	20%	30%
A vs B	0.669	0.706	0.855	0.968	0.996	1.000
A vs C	0.002	0.000	0.004	0.011	0.146	0.897
A vs D	0.002	0.005	0.013	0.032	0.164	0.683
A vs E	0.000	0.000	0.000	0.000	0.000	0.000
C vs D	0.006	0.016	0.018	0.033	0.163	0.433

bootstrap p-values when $n = 30$.

Comparison of two distributions

We have generated 5 random samples: A $\sim$ N(0,1), B $\sim$ N(0,1), C $\sim$ N(1,1), D $\sim$ N(1,2) and E $\sim$ N(2,1) for two sample sizes ($n = 30, 100$).

	0%	1%	5%	10%	20%	30%
A vs B	0.669	0.706	0.855	0.968	0.996	1.000
A vs C	0.002	0.000	0.004	0.011	0.146	0.897
A vs D	0.002	0.005	0.013	0.032	0.164	0.683
A vs E	0.000	0.000	0.000	0.000	0.000	0.000
C vs D	0.006	0.016	0.018	0.033	0.163	0.433

bootstrap p-values when $n = 30$.

	0%	1%	5%	10%	20%	30%
A vs B	0.964	0.974	0.998	1.000	1.000	1.000
A vs C	0.000	0.000	0.000	0.000	0.004	0.371
A vs D	0.000	0.000	0.000	0.000	0.091	1.000
A vs E	0.000	0.000	0.000	0.000	0.000	0.000
C vs D	0.000	0.000	0.000	0.001	0.134	0.905

bootstrap p-values when $n = 100$.

Comparison of two distributions

We have generated 5 random samples: A $\sim$ N(0,1), B $\sim$ N(0,1), C $\sim$ N(1,1), D $\sim$ N(1,2) and E $\sim$ N(2,1) for two sample sizes ($n = 30, 100$).

	0%	1%	5%	10%	20%	30%
A vs B	0.669	0.706	0.855	0.968	0.996	1.000
A vs C	0.002	0.000	0.004	0.011	0.146	0.897
A vs D	0.002	0.005	0.013	0.032	0.164	0.683
A vs E	0.000	0.000	0.000	0.000	0.000	0.000
C vs D	0.006	0.016	0.018	0.033	0.163	0.433

bootstrap p-values when $n = 30$.

	0%	1%	5%	10%	20%	30%	$\mathcal{S}(P, Q)$
A vs B	0.964	0.974	0.998	1.000	1.000	1.000	0%
A vs C	0.000	0.000	0.000	0.000	0.004	0.371	38%
A vs D	0.000	0.000	0.000	0.000	0.091	1.000	39%
A vs E	0.000	0.000	0.000	0.000	0.000	0.000	68%
C vs D	0.000	0.000	0.000	0.001	0.134	0.905	32%

bootstrap p-values when $n = 100$.

Other examples

Generate 100 replicates of the bootstrap p-values:

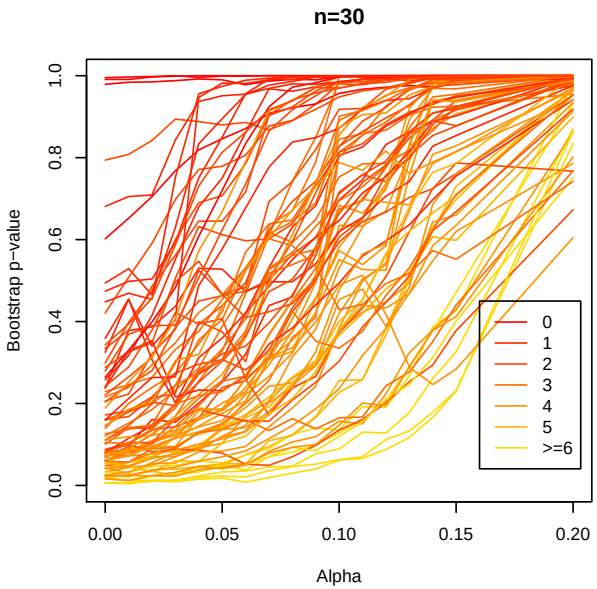
Case 1. $n = 100$. X's sample taken from $.90N(0,1) + .10N(5,1)$
Y's sample taken from $.90N(0,1) + .10N(-5,1)$
$\mathcal{S}(P,Q) = .1000$

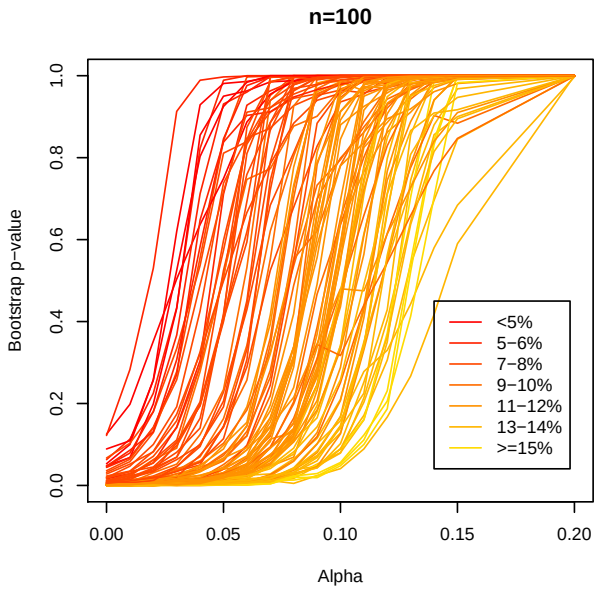
Case 2. $n = 100$. X's sample taken from $N(0,1)$
Y's sample taken from $.80N(0,1) + .20N(0,3)$
$\mathcal{S}(P,Q) = .0969$

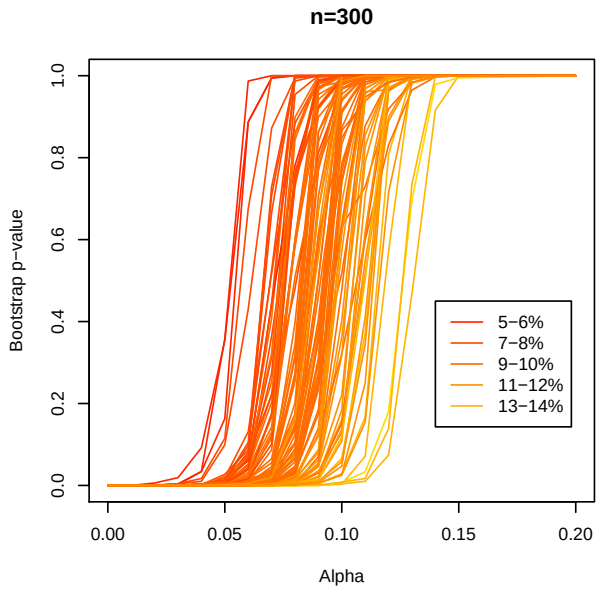
Medians of bootstrap p-values:

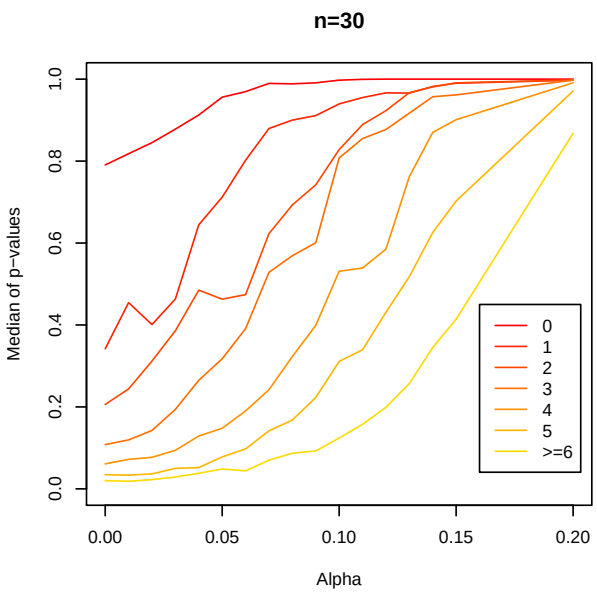
α	0	5	6	7	8	9	10	11	12	13	14	15
Case 1	0	.01	.03	.06	.12	.26	.47	.70	.87	.92	.97	.99
Case 2	.04	.46	.63	.80	.91	.98	.99	1	1	1	1	1

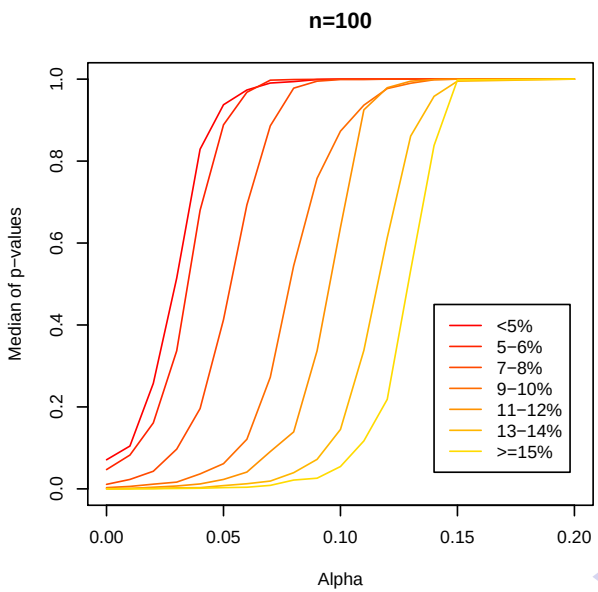
N(0,1) vs .9N(0,1)+.1N(5,1): 100 replicates

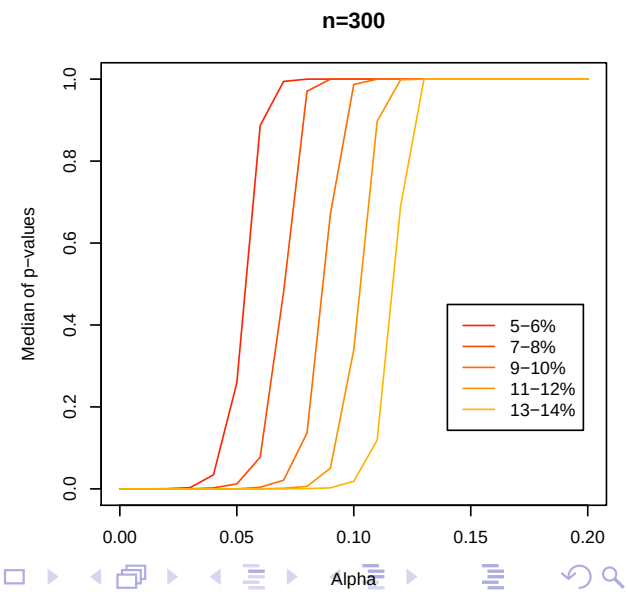












Conclusion

Test $H_0 : \mathcal{W}_2(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0 \quad (\mathcal{S}(P, Q) \leq \alpha)$

Rejecting H_0 for small values of the bootstrap p-value.

This procedure is asymptotically error free.

Example: markers of *selectividad* exam

Marks of the access-to-university exam in the university district of Valladolid: 1550 exams of the same subject distributed between 10 markers.

Marker	1	2	3	4	5	6	7	8	9	10
N^o of exams	155	152	155	156	156	156	156	154	156	154

Do they mark in a homogeneous way? Do they use a common pattern to mark?

Example: markers of *selectividad* exam

Marks of the access-to-university exam in the university district of Valladolid: 1550 exams of the same subject distributed between 10 markers.

Marker	1	2	3	4	5	6	7	8	9	10
N^o of exams	155	152	155	156	156	156	156	154	156	154

Do they mark in a homogeneous way? Do they use a common pattern to mark?

What subset of them marks more similarly?

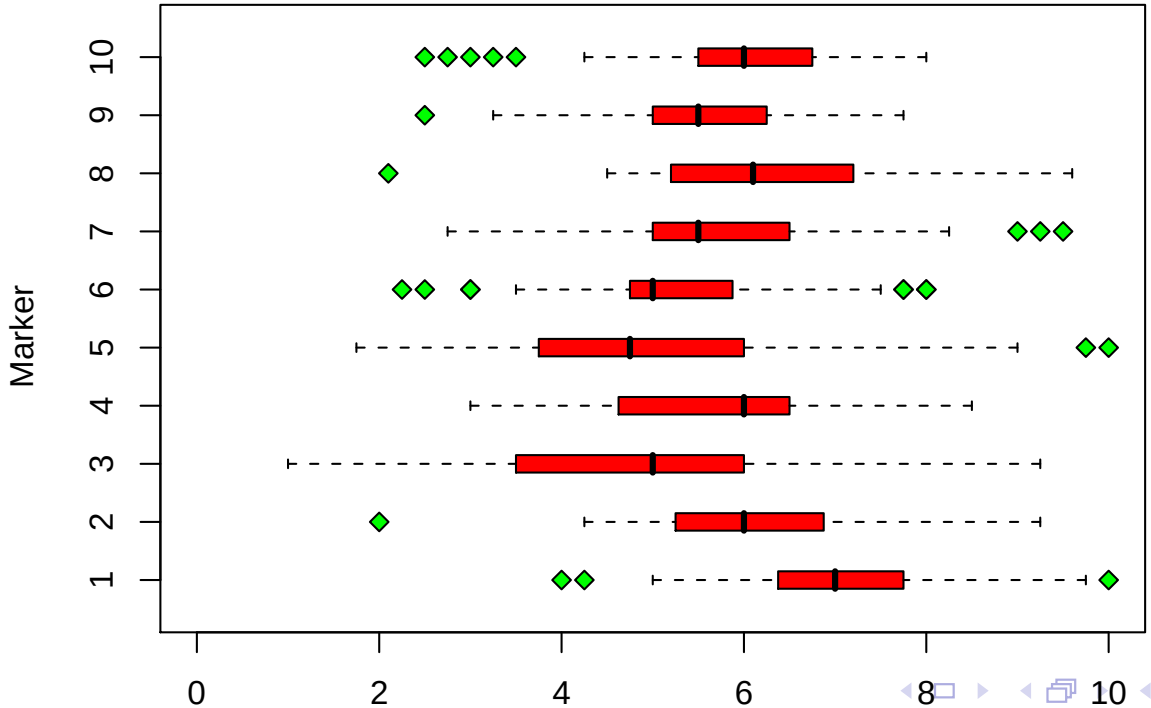
Example: markers of *selectividad* exam

Marks of the access-to-university exam in the university district of Valladolid: 1550 exams of the same subject distributed between 10 markers.

Marker	1	2	3	4	5	6	7	8	9	10
N^o of exams	155	152	155	156	156	156	156	154	156	154

Do they mark in a homogeneous way? Do they use a common pattern to mark?

What subset of them marks more similarly?



Looking for the "common pattern" of several distributions

- We know how to assess the similarity of two distributions, but now, we have k distributions ...**then?**

Looking for the "common pattern" of several distributions

- We know how to assess the similarity of two distributions, but now, we have k distributions ...**then?**
- **Design a sequential procedure**:

Looking for the "common pattern" of several distributions

- We know how to assess the similarity of two distributions, but now, we have k distributions ...**then?**
- **Design a sequential procedure**:
  1. Start considering all distributions in the group of the similar ones, (and perhaps, fix the level of similarity, α).

Looking for the "common pattern" of several distributions

- We know how to assess the similarity of two distributions, but now, we have k distributions ...**then?**
- **Design a sequential procedure**:
  1. Start considering all distributions in the group of the similar ones, (and perhaps, fix the level of similarity, α).
  2. Compare each distribution in the group with the pool of all distributions in the group, except, the one you are considering.

Looking for the "common pattern" of several distributions

- We know how to assess the similarity of two distributions, but now, we have k distributions ...**then?**
- **Design a sequential procedure**:
  1. Start considering all distributions in the group of the similar ones, (and perhaps, fix the level of similarity, α).
  2. Compare each distribution in the group with the pool of all distributions in the group, except, the one you are considering.
  3. Take the less similar distribution, if you have enough evidence of dissimilarity, this distribution leaves the group.

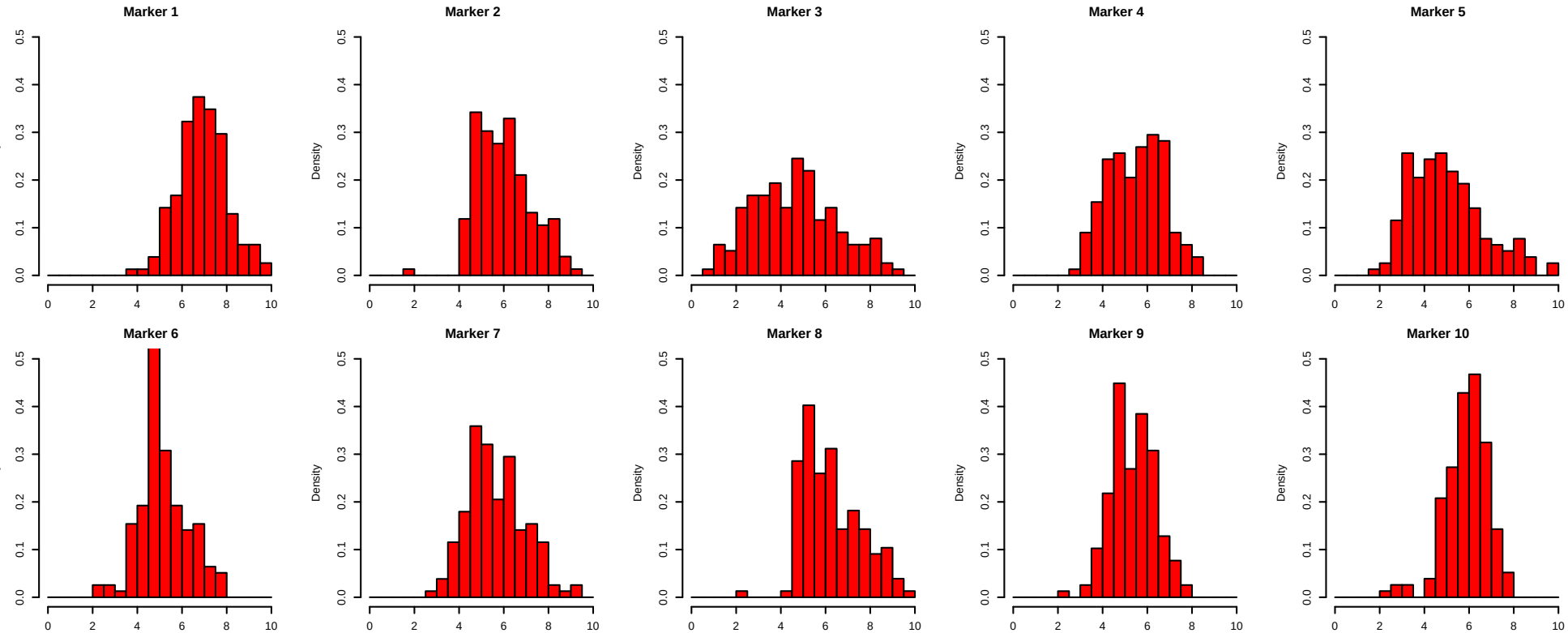
Looking for the "common pattern" of several distributions

- We know how to assess the similarity of two distributions, but now, we have k distributions ...**then?**
- **Design a sequential procedure**:
  1. Start considering all distributions in the group of the similar ones, (and perhaps, fix the level of similarity, α).
  2. Compare each distribution in the group with the pool of all distributions in the group, except, the one you are considering.
  3. Take the less similar distribution, if you have enough evidence of dissimilarity, this distribution leaves the group.
  4. Iterate steps 2-3 until none distribution leaves the group.

Looking for the "common pattern" of several distributions

- We know how to assess the similarity of two distributions, but now, we have k distributions ...**then?**
- **Design a sequential procedure**:
  1. Start considering all distributions in the group of the similar ones, (and perhaps, fix the level of similarity, α).
  2. Compare each distribution in the group with the pool of all distributions in the group, except, the one you are considering.
  3. Take the less similar distribution, if you have enough evidence of dissimilarity, this distribution leaves the group.
  4. Iterate steps 2-3 until none distribution leaves the group.
  5. Consider to recover the distributions out of the group.

Example: markers of *selectividad* exam



Example: sequential process

Step 0: All markers are in the group. Compute bootstrap p-values.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%	0.000	0.000	0.000	0.079	0.000	0.000	0.406	0.000	0.000	0.000
5%	0.000	0.134	0.000	0.497	0.000	0.000	0.997	0.000	0.001	0.000
10%	0.000	0.996	0.000	0.951	0.001	0.003	1.000	0.450	0.058	0.003
20%	0.000	1.000	0.036	1.000	0.572	0.642	1.000	1.000	1.000	0.591

Example: sequential process

Step 1: Marker 1 is the most *different* and goes out.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%	0.000	0.000	0.000	0.079	0.000	0.000	0.406	0.000	0.000	0.000
5%	0.000	0.134	0.000	0.497	0.000	0.000	0.997	0.000	0.001	0.000
10%	0.000	0.996	0.000	0.951	0.001	0.003	1.000	0.450	0.058	0.003
20%	0.000	1.000	0.036	1.000	0.572	0.642	1.000	1.000	1.000	0.591

Example: sequential process

Step 2: Recompute bootstrap p-values.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.000	0.000	0.083	0.000	0.000	0.254	0.000	0.000	0.000
5%		0.001	0.000	0.466	0.000	0.000	0.996	0.000	0.007	0.000
10%		0.740	0.000	0.949	0.006	0.031	1.000	0.036	0.399	0.000
20%		1.000	0.124	1.000	0.766	0.996	1.000	1.000	1.000	0.367

Example: sequential process

Step 2: Next marker that goes out is num. 3.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.000	0.000	0.083	0.000	0.000	0.254	0.000	0.000	0.000
5%		0.001	0.000	0.466	0.000	0.000	0.996	0.000	0.007	0.000
10%		0.740	0.000	0.949	0.006	0.031	1.000	0.036	0.399	0.000
20%		1.000	0.124	1.000	0.766	0.996	1.000	1.000	1.000	0.367

Example: sequential process

Step 3: Recompute bootstrap p-values.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.000		0.173	0.000	0.000	0.920	0.000	0.000	0.000
5%		0.089		0.518	0.000	0.000	0.999	0.000	0.055	0.000
10%		0.992		0.966	0.000	0.037	1.000	0.299	0.931	0.010
20%		1.000		1.000	0.146	0.993	1.000	1.000	1.000	0.652

Example: sequential process

Step 3: Marker 5 goes out.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.000		0.173	0.000	0.000	0.920	0.000	0.000	0.000
5%		0.089		0.518	0.000	0.000	0.999	0.000	0.055	0.000
10%		0.992		0.966	0.000	0.037	1.000	0.299	0.931	0.010
20%		1.000		1.000	0.146	0.993	1.000	1.000	1.000	0.652

Example: sequential process

Step 4: Recompute bootstrap p-values (1, 3 and 5 are out).

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.006		0.058		0.000	0.712	0.000	0.000	0.000
5%		0.619		0.349		0.000	0.998	0.005	0.056	0.001
10%		1.000		0.888		0.001	1.000	0.709	0.975	0.038
20%		1.000		1.000		0.769	1.000	1.000	1.000	0.879

Example: sequential process

Step 4: Marker 6 goes out (1, 3 and 5 are out).

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.006		0.058		0.000	0.712	0.000	0.000	0.000
5%		0.619		0.349		0.000	0.998	0.005	0.056	0.001
10%		1.000		0.888		0.001	1.000	0.709	0.975	0.038
20%		1.000		1.000		0.769	1.000	1.000	1.000	0.879

Example: sequential process

Step 5: Recompute bootstrap p-values of the markers in.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.067		0.013			0.322	0.001	0.000	0.000
5%		0.992		0.196			0.914	0.020	0.004	0.005
10%		1.000		0.825			1.000	0.865	0.338	0.120
20%		1.000		1.000			1.000	1.000	1.000	0.979

Example: sequential process

Step 5: Marker 10 goes out ($W_2(5\%)$ is 0.33 points and $W_2(10\%)$=0.25).

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.067		0.013			0.322	0.001	0.000	0.000
5%		0.992		0.196			0.914	0.020	0.004	0.005
10%		1.000		0.825			1.000	0.865	0.338	0.120
20%		1.000		1.000			1.000	1.000	1.000	0.979

Example: sequential process

Step 6: Recompute bootstrap p-values of the markers in.

	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.027		0.017			0.771	0.000	0.000	
5%		0.857		0.165			0.999	0.017	0.004	
10%		1.000		0.684			1.000	0.819	0.352	
20%		1.000		1.000			1.000	1.000	1.000	

Example: sequential process

Step 6: None of the markers out enters into the group.

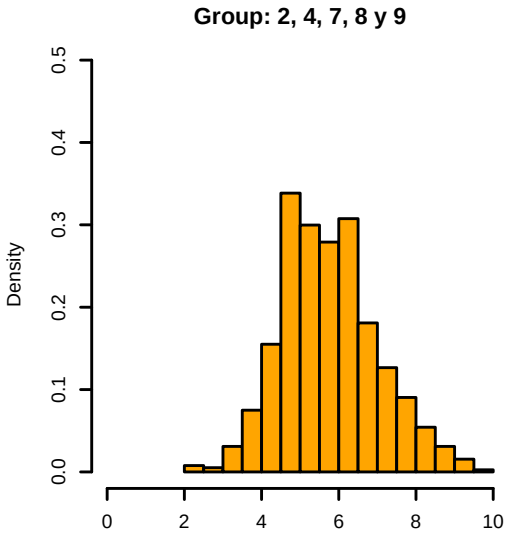
	Marker									
	1	2	3	4	5	6	7	8	9	10
1%	0.000	0.027	0.000	0.017	0.000	0.000	0.771	0.000	0.000	
5%	0.000	0.857	0.000	0.165	0.000	0.000	0.999	0.017	0.004	
10%	0.000	1.000	0.000	0.684	0.000	0.000	1.000	0.819	0.352	
20%	0.000	1.000	0.008	1.000	0.001	0.000	1.000	1.000	1.000	

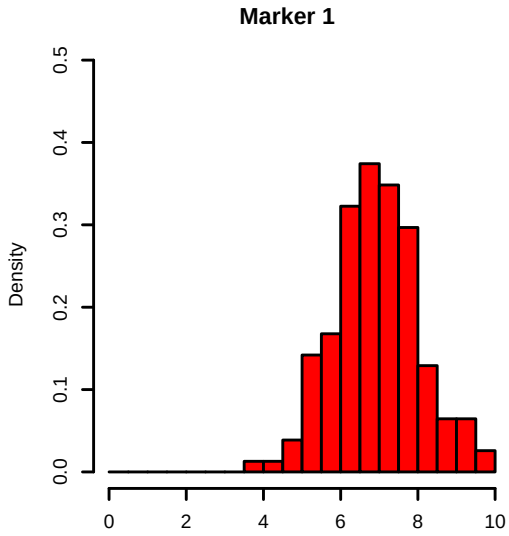
Example: sequential process

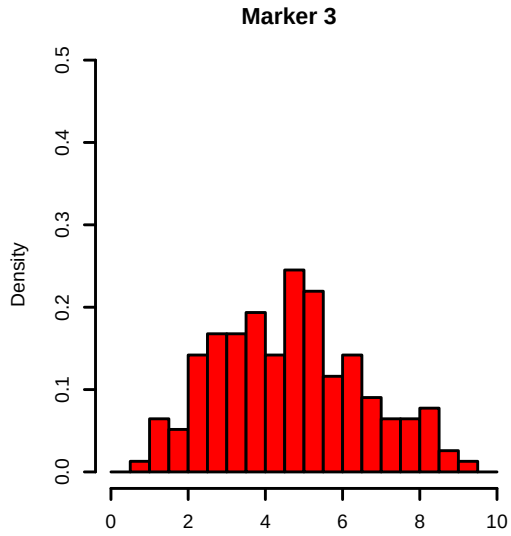
Step 6 (End): The group of markers that mark most similarly is 2, 4, 7, 8 and 9.

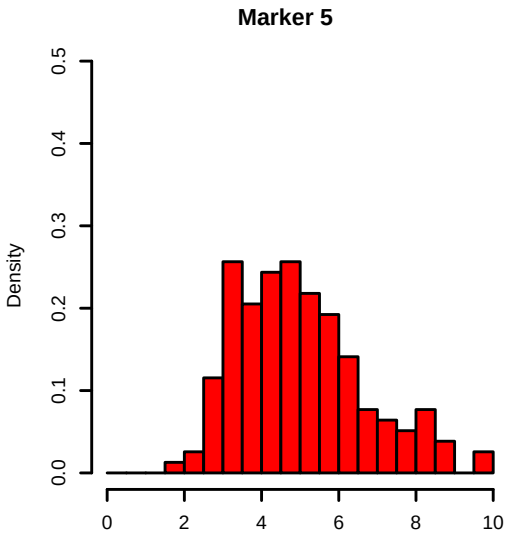
	Marker									
	1	2	3	4	5	6	7	8	9	10
1%		0.027		0.017			0.771	0.000	0.000	
5%		0.857		0.165			0.999	0.017	0.004	
10%		1.000		0.684			1.000	0.819	0.352	
20%		1.000		1.000			1.000	1.000	1.000	

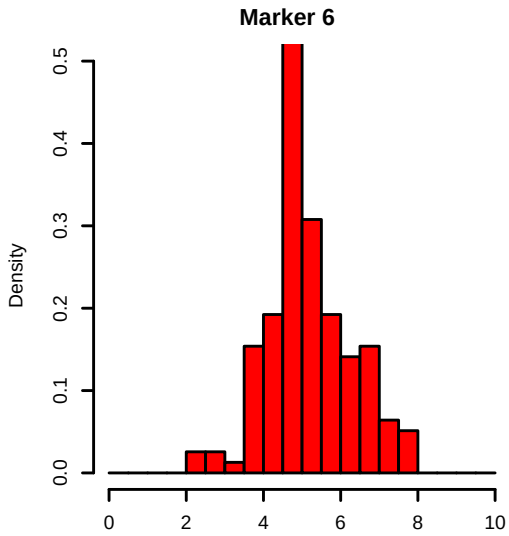
Example: markers of *selectividad* exam

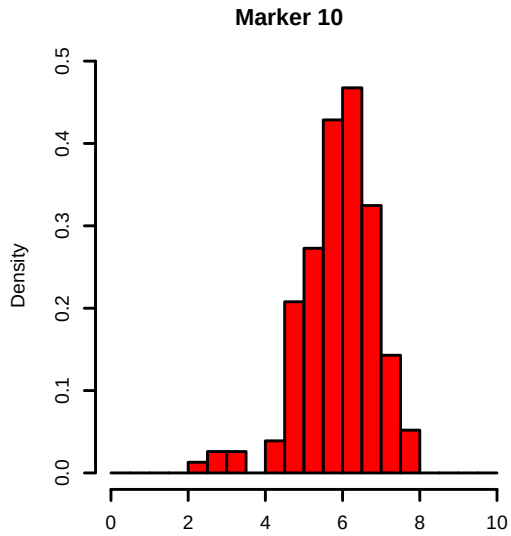




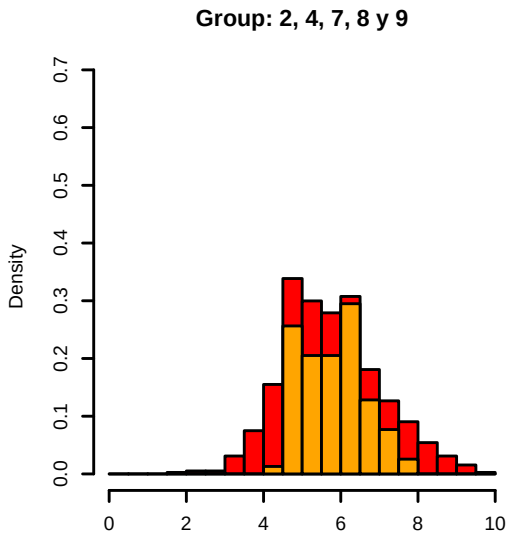


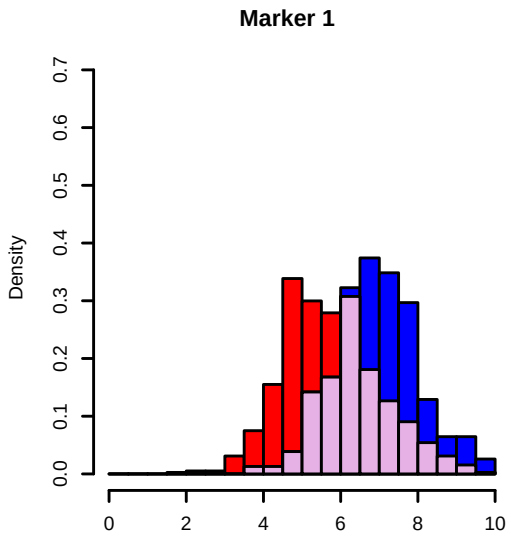


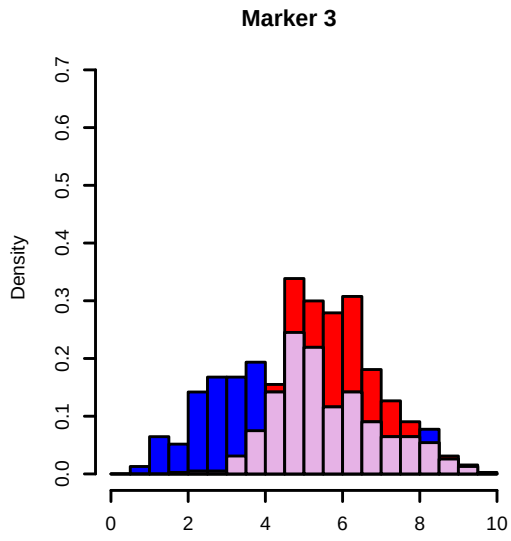


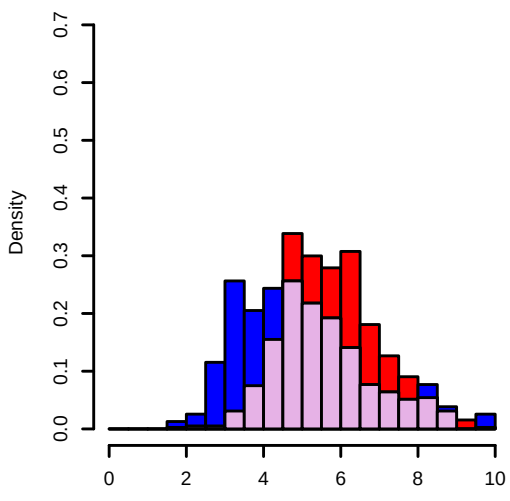


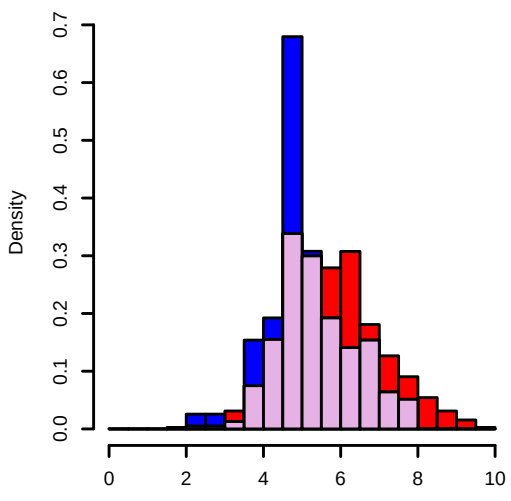
Example: markers of *selectividad* exam

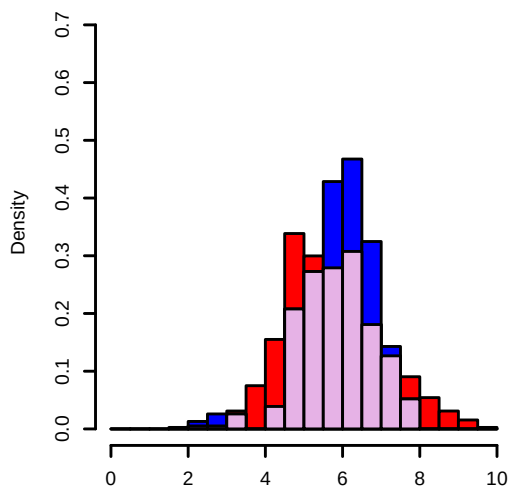




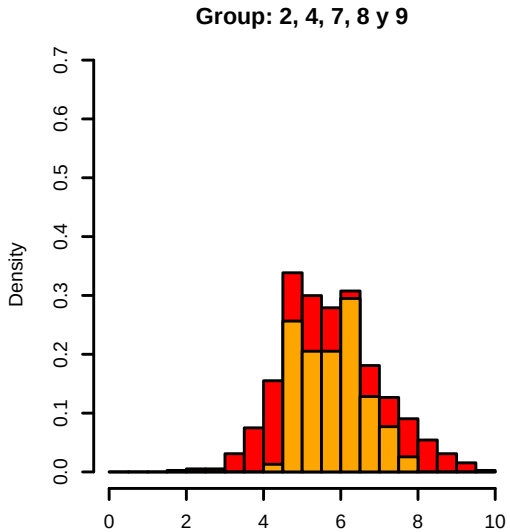


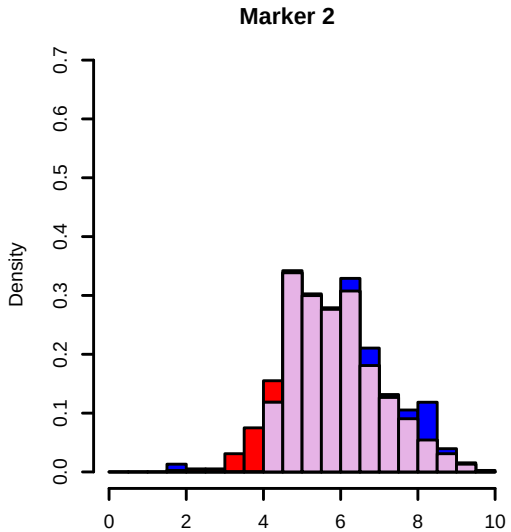


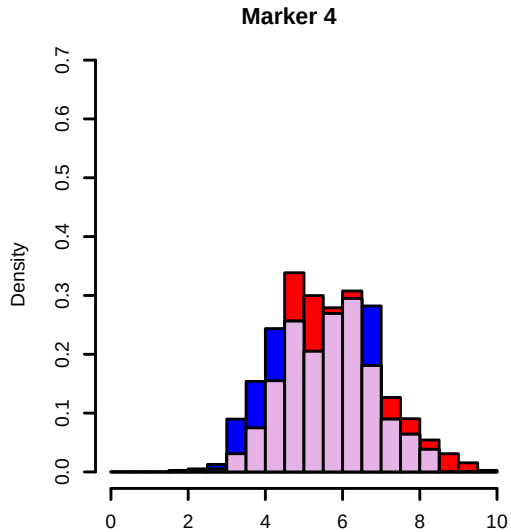


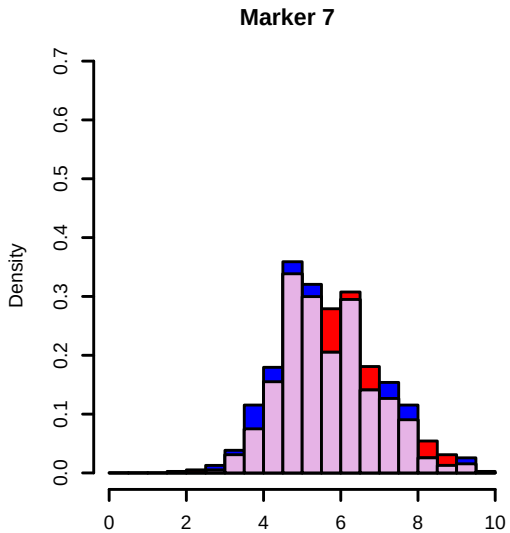


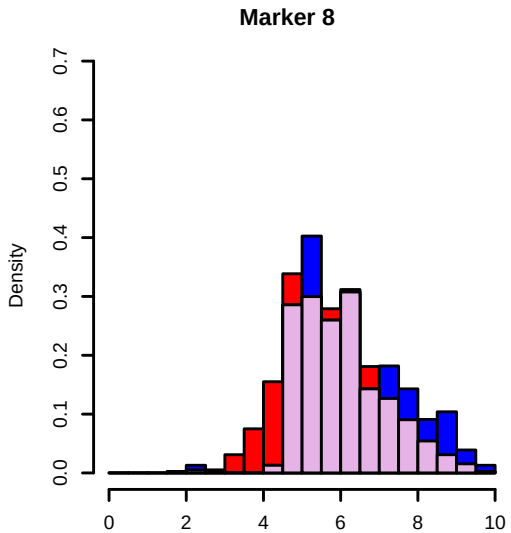
Example: markers of *selectividad* exam

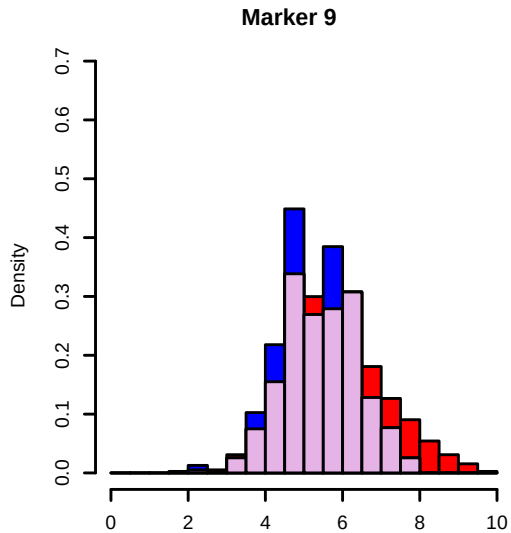




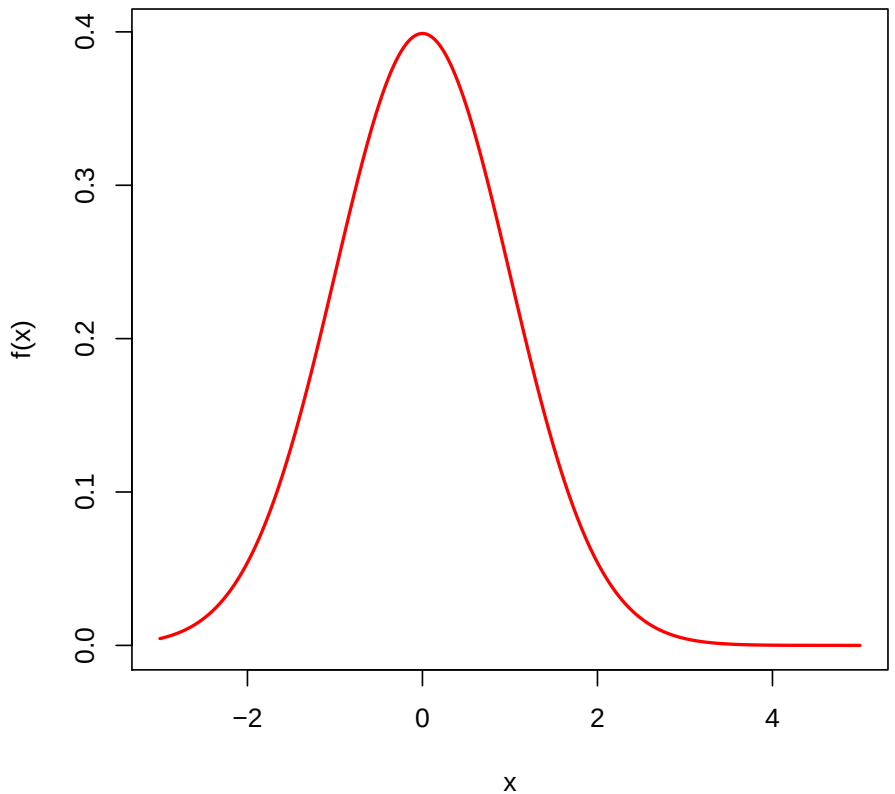




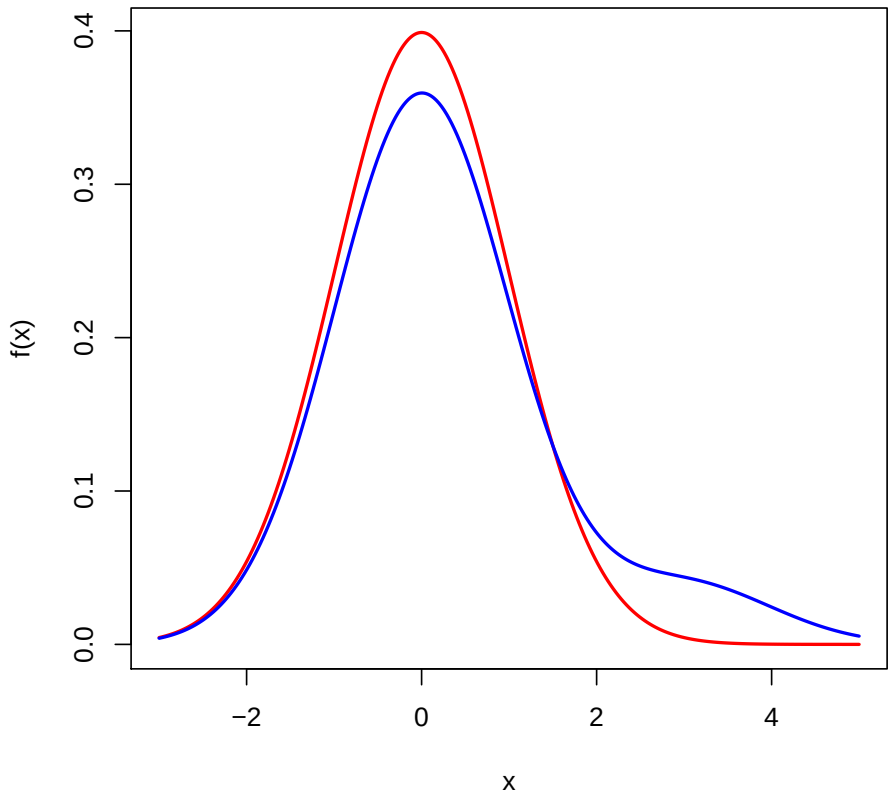




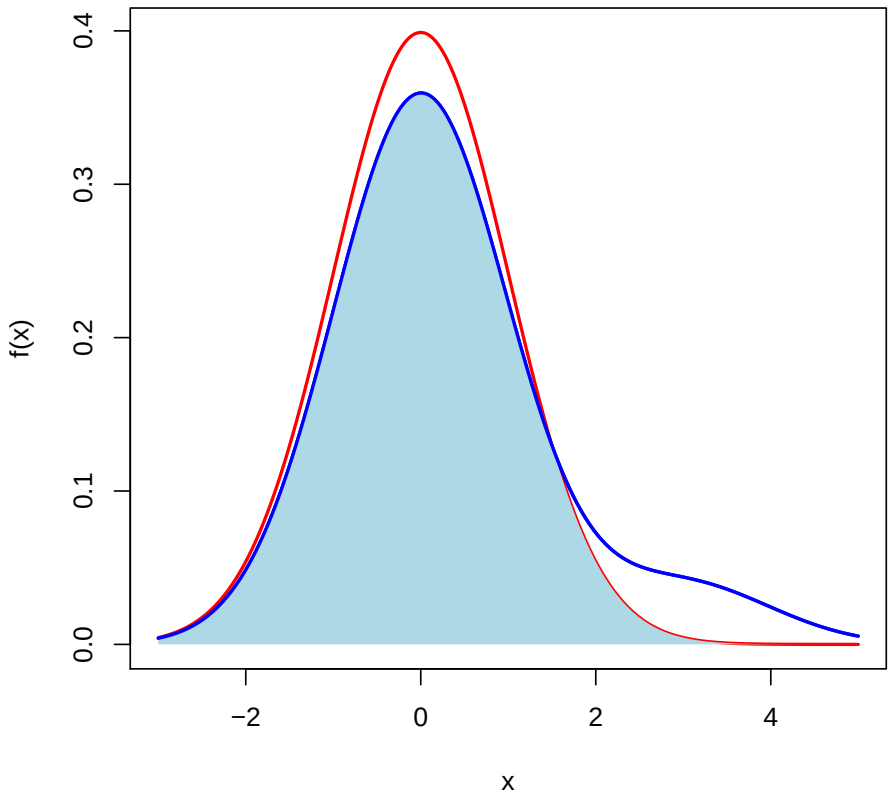
The 'Core' of several distributions



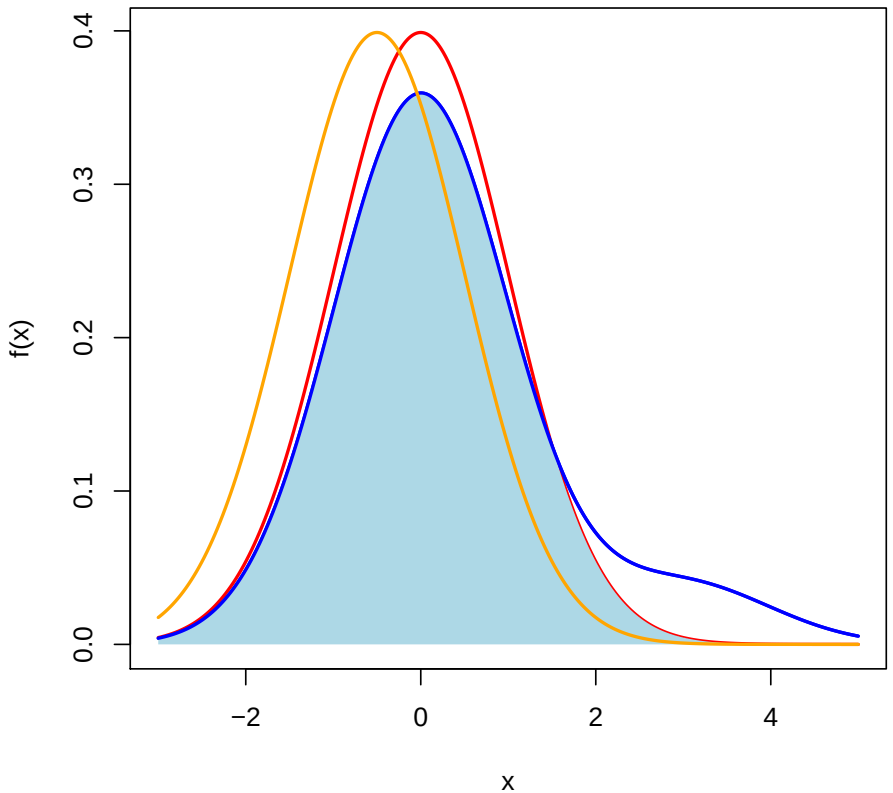
The 'Core' of several distributions



The 'Core' of several distributions



The 'Core' of several distributions



The 'Core' of several distributions