Biostatistik, WS 2013/2014

Exkurs: Faktorielle Varianzanalyse und F-Test, Diskriminanzanalyse

Matthias Birkner

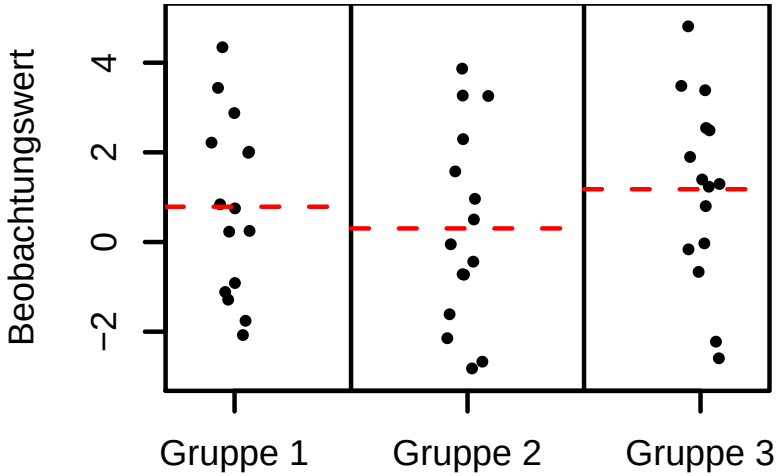
http://www.mathematik.uni-mainz.de/~birkner/Biostatistik1314/

31.1.2014

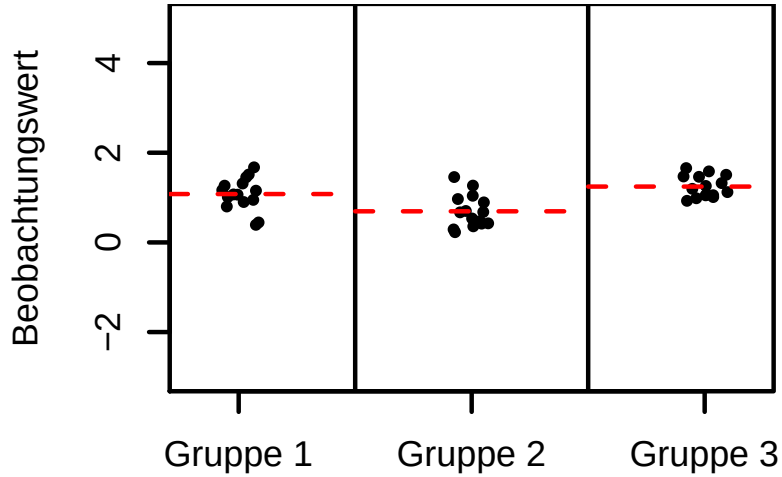
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Grundidee der Varianzanalyse

Wir beobachten unterschiedliche Gruppenmittelwerte:



Variabilität innerhalb der Gruppen groß



Variabilität innerhalb der Gruppen klein

Sind die beobachteten Unterschiede der Gruppenmittelwerte ernst zu nehmen — oder könnte das alles Zufall sein?

Das hängt vom Verhältnis der Variabilität der Gruppenmittelwerte und der Variabilität der Beobachtungen innerhalb der Gruppen ab: die Varianzanalyse gibt eine (quantitative) Antwort.

Beispiel: Blutgerinnungszeiten

Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

Gruppe	Beobachtung							
1	62	60	63	59				
2	63	67	71	64	65	66		
3	68	66	71	67	68	68		
4	56	62	60	61	63	64	63	59

Globalmittelwert $\overline{x}_{\cdot\cdot} = 64$,
Gruppenmittelwerte $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

Bemerkung: Der Globalmittelwert ist in diesem Beispiel auch der Mittelwert der Gruppenmittelwerte. Das muss aber nicht immer so sein!

Beispiel

Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

Gr.	$\overline{x}_{i\cdot}$	Beobachtung							
1	61	62	60	63	59				
		$(62-61)^2$	$(60-61)^2$	$(63-61)^2$	$(59-61)^2$				
2	66	63	67	71	64	65	66		
		$(63-66)^2$	$(67-66)^2$	$(71-66)^2$	$(64-66)^2$	$(65-66)^2$	$(66-66)^2$		
3	68	68	66	71	67	68	68		
		$(68-68)^2$	$(66-68)^2$	$(71-68)^2$	$(67-68)^2$	$(68-68)^2$	$(68-68)^2$		
4	61	56	62	60	61	63	64	63	59
		$(56-61)^2$	$(62-61)^2$	$(60-61)^2$	$(61-61)^2$	$(63-61)^2$	$(64-61)^2$	$(63-61)^2$	$(59-61)^2$

Globalmittelwert $\overline{x}_{\cdot\cdot} = 64$,
Gruppenmittelwerte $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

Die roten Werte (ohne die Quadrate) heißen Residuen: die „Restvariabilität“ der Beobachtungen, die das Modell nicht erklärt.

Quadratsumme innerhalb der Gruppen:
$ss_{\text{innerh}} = 112$, 20 Freiheitsgrade
Quadratsumme zwischen den Gruppen:
$ss_{\text{zw}} = 4 \cdot (61-64)^2 + 6 \cdot (66-64)^2 + 6 \cdot (68-64)^2 + 8 \cdot (61-64)^2 = 228$,
3 Freiheitsgrade

$$F = \frac{ss_{\text{zw}}/3}{ss_{\text{innerh}}/20} = \frac{76}{5{,}6} = 13{,}57$$

Beispiel: Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

ANOVA-Tafel („ANalysis Of VAriance“)

	Freiheits-grade (DF)	Quadrat-summe (SS)	mittlere Quadrat-summe (SS/DF)	F-Wert
Gruppe	3	228	76	13,57
Residuen	20	112	5,6	

Unter der Hypothese H_0 „die Gruppenmittelwerte sind gleich“
(und einer Normalverteilungsannahme an die Beobachtungen)
ist F Fisher-verteilt mit 3 und 20 Freiheitsgraden,
$p = \mathrm{Fisher}_{3,20}([13{,}57, \infty)) \leq 5 \cdot 10^{-5}$.

Wir lehnen demnach H_0 ab.

Sir Ronald Aylmer Fisher,
1890–1962

F-Test

$n = n_1 + n_2 + \cdots + n_I$ Beobachtungen in I Gruppen,
$X_{ij} = j$-te Beobachtung in der i-ten Gruppe, $j = 1, \ldots, n_i$.
Modellannahme: $X_{ij} = \mu_i + \varepsilon_{ij}$,
mit unabhängigen, normalverteilten ε_{ij}, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$
(μ_i ist der „wahre“ Mittelwert innerhalb der i-ten Gruppe.)
$\overline{X}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} X_{ij}$ (empirisches) „Globalmittel“
$\overline{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ (empirischer) Mittelwert der i-ten Gruppe

$$SS_{\text{innerh}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$$

Quadratsumme innerhalb d. Gruppen,
$n - I$ Freiheitsgrade

$$SS_{\text{zw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

Quadratsumme zwischen d. Gruppen,
$I - 1$ Freiheitsgrade

$$F = \frac{SS_{\text{zw}}/(I-1)}{SS_{\text{innerh}}/(n-I)}$$

F-Test

$X_{ij} = j$-te Beobachtung in der i-ten Gruppe, $j = 1, \ldots, n_i$,
Modellannahme: $X_{ij} = \mu_i + \varepsilon_{ij}$. $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

$$SS_{\text{innerh}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$$

Quadratsumme innerhalb d. Gruppen, $n - I$ Freiheitsgrade

$$SS_{\text{zw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

Quadratsumme zwischen d. Gruppen, $I - 1$ Freiheitsgrade

$$F = \frac{SS_{\text{zw}}/(I-1)}{SS_{\text{innerh}}/(n-I)}$$

Unter der Hypothese $H_0 : \mu_1 = \cdots = \mu_I$ („alle μ_i sind gleich“)
ist F Fisher-verteilt mit $I - 1$ und $n - I$ Freiheitsgraden
(unabhängig vom tatsächlichen gemeinsamen Wert der μ_i).

F-Test: Wir lehnen H_0 zum Signifikanzniveau α ab, wenn
$F \geq q_\alpha$, wobei q_α das $(1-\alpha)$-Quantil der Fisher-Verteilung mit
$I - 1$ und $n - I$ Freiheitsgraden ist.

Berechnung der Signifikanz mit R

Wie muss man q wählen, damit $\mathbb{P}(F \leq q) = 0.95$ für Fisher(6,63)-verteiltes F?

```
> qf(0.95,df1=6,df2=63)
[1] 2.246408
```

p-Wert-Berechnung: Wie wahrscheinlich ist es, dass eine Fisher(3,20)-verteilte Zufallsgröße einen Wert ≥ 13.57 annimmt?

```
> pf(13.57, df1=3, df2=20, lower.tail=FALSE)
[1] 4.66169e-05
```

Tabelle der 95%-Quantile der F-Verteilung

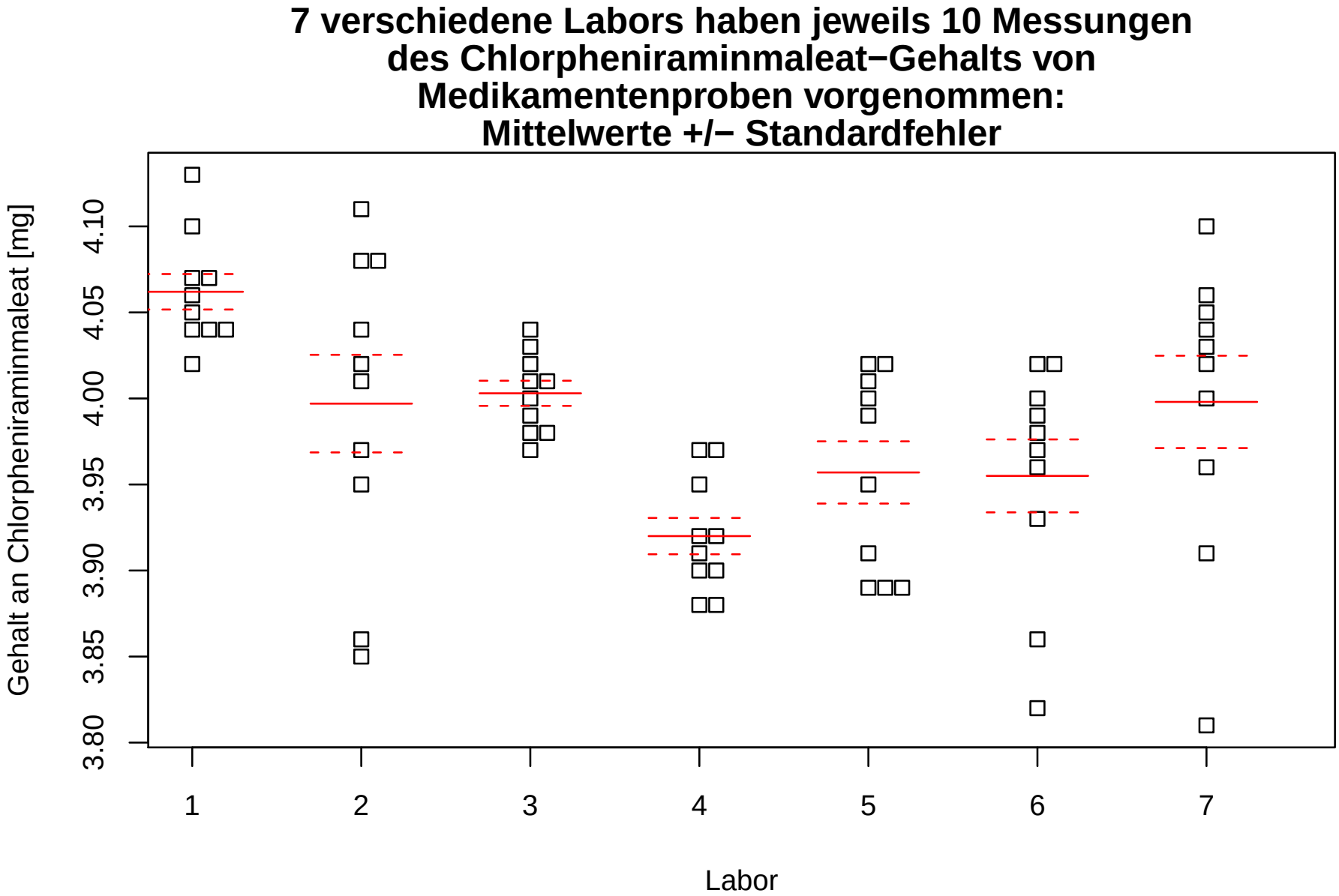
Die folgende Tabelle zeigt (auf 2 Nachkommastellen gerundet) das 95%-Quantil der Fisher-Verteilung mit k_1 und k_2 Freiheitsgraden (k_1 Zähler- und k_2 Nennerfreiheitsgrade)

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	11
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.7
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.6
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.1
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.72
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.57
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.51
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.41
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.34
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.31

Varianzanalyse komplett in R

Die Text-Datei gerinnung.txt enthält eine Spalte “bgz” mit den Blutgerinnungszeiten und eine Spalte “beh” mit der Behandlung (A,B,C,D).

```
> rat<-read.table("gerinnung.txt",header=TRUE)
> rat.aov <- aov(bgz~beh,data=rat)
> summary(rat.aov)
            Df Sum Sq Mean Sq F value    Pr(>F)
beh          3    228    76.0  13.571 4.658e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```



Daten aus R.D. Kirchhoefer, Semiautomated method for the analysis of chlorpheniramine maleate tablets: collaborative study, *J. Assoc. Off. Anal. Chem.* 62(6):1197-1201 (1979),
zitiert nach John A. Rice, *Mathematical statistics and data analysis*, 2nd ed., Wadsworth, 1995

Beachte: Die Labore sind sind mit Zahlen nummeriert. Damit R das nicht als numerische Werte sondern als Nummern der Labore auffasst, müssen wir die Variable “Labor” in einen sog. Factor umwandeln:

```
> chlor <- read.table("chlorpheniraminmaleat.txt")
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ Gehalt: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4
 $ Labor : int  1 1 1 1 1 1 1 1 1 1 ...
> chlor$Labor <- as.factor(chlor$Labor)
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ Gehalt: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4
 $ Labor : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1
```

Nun können wir die Varianzanalyse durchführen:

```
> chlor.aov <- aov(Gehalt~Labor,data=chlor)
> summary(chlor.aov)
             Df  Sum Sq  Mean Sq F value    Pr(>F)
Labor         6 0.12474 0.020789  5.6601 9.453e-05 ***
Residuals    63 0.23140 0.003673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Die Varianzanalyse zeigte, dass es signifikante Unterschiede zwischen den Laboren gibt.
Aber welche Labore unterscheiden sich signifikant?
p-Werte aus paarweisen Vergleichen mittels t-Tests:

	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
Lab1	0.05357	0.00025	0.00000	0.00017	0.00055	0.04657
Lab2		0.84173	0.02654	0.25251	0.25224	0.97985
Lab3			0.00001	0.03633	0.05532	0.86076
Lab4				0.09808	0.16280	0.01944
Lab5					0.94358	0.22336
Lab6						0.22543

Wir haben 21 paarweise Vergleiche; auf dem 5%-Niveau zeigen einige davon Signifikanz an.
Problem des Multiplen Testens: Wenn die Nullhypothese (“alles nur Zufallsschwankungen”) stimmt, verwirft man im Schnitt bei 5% der Tests die Nullhypothese zu Unrecht. Testet man mehr als 20 mal und gelten jeweils die Nullhypothesen, wird man im Schnitt mehr als eine Nullhypothese zu Unrecht verwerfen.
Daher sollte man bei multiplen Tests mit korrigierten p-Werten arbeiten.

Eine ganz allgemeine Korrektur für multiples Testen ist die **Bonferroni-Methode**: Multipliziere jeden p-Wert mit der Anzahl n der durchgeführten Tests.

Beispiel: Paarweise Vergleiche (mittels t-Test) für die Blutgerinnungszeiten bei vier verschiedenen Behandlungen, zunächst ohne Korrektur für multiples Testen:

```
        B       C       D
A 0.00941 0.00078 1.00000
B         0.17383 0.00663
C                 0.00006
```

Nun mit Bonferroni-Korrektur (alle Werte mit $\binom{4}{2} = 6$ multiplizieren):

```
        B       C       D
A 0.05646 0.00468 6.00000
B         1.04298 0.03978
C                 0.00036
```

Nach Bonferroni-Korrektur führen folgende Paare von Behandlungen zu jeweils signifikant unterschiedlichen Ergebnissen: A/C, B/D sowie C/D. (Der Bonferroni-korrigierte p-Wert von 6.0 für den Vergleich der Behandlungen A und D ist natürlich nicht als echter p-Wert zu interpretieren.)

Die Bonferroni-Methode ist sehr *konservativ*, d.h. um auf der sicheren Seite zu sein, lässt man sich lieber die eine oder andere Signifikanz entgehen.
Eine Verbesserung der Bonferroni-Methode ist die **Bonferroni-Holm-Methode**: Ist k die Anzahl der Tests, so multipliziere den kleinsten p-Wert mit k, den zweitkleinsten mit $k-1$, den drittkleinsten mit $k-2$ usw.
In R gibt des den Befehl `p.adjust`, der p-Werte für multiples Testen korrigiert und dabei defaultmäßig Bonferroni-Holm verwendet:

```
> pv <- c(0.00941, 0.00078, 1.00000, 0.17383,
+         0.00663, 0.00006)
> p.adjust(pv)
[1] 0.02823 0.00390 1.00000 0.34766 0.02652 0.00036
> p.adjust(pv, method="bonferroni")
[1] 0.05646 0.00468 1.00000 1.00000 0.03978 0.00036
```

Für paarweise t-Tests gibt es ebenfalls eine R-Funktion, die per default die Bonferroni-Holm-Korrektur verwendet:

```
> pairwise.t.test(rat$bgz,rat$beh,pool.sd=FALSE)

Pairwise comparisons using t tests with non-pooled SD

data:  rat$bgz and rat$beh

  A       B       C
B 0.02823 -       -
C 0.00391 0.34766 -
D 1.00000 0.02654 0.00035

P value adjustment method: holm
```

photo (c) Thermos

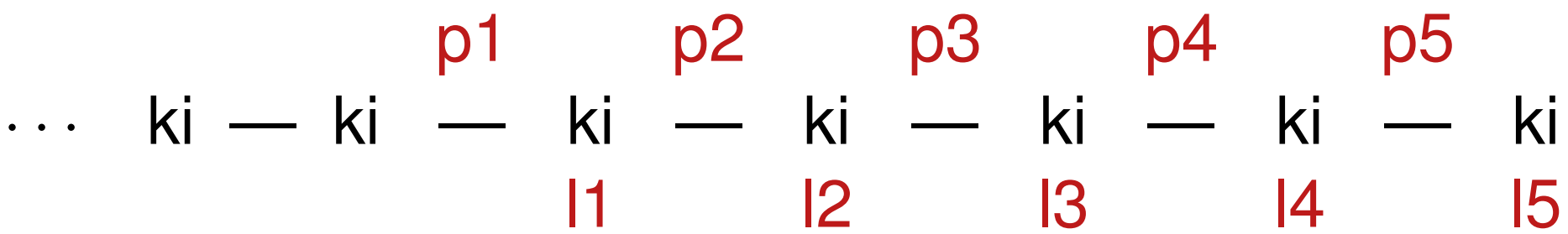
ein Kleinspecht (*Picoides minor*)

Man kann die Geschlechter optisch unterscheiden.

Frage: Geht es auch akustisch?

Ruf des Kleinspechts:

Längen der letzten fünf Pausen und Laute



Frage:

Kann man aus den Längen der Pausen und der Laute

(p1, p2, p3, p4, p5, l1, l2, l3, l4, l5)

das Geschlecht bestimmen?

Daten: 62 Rufe von Kleinspechten

18 Rufe von Männchen

44 Rufe von Weibchen

Daten von Dr. Kerstin Höntsch, Frankfurt
(siehe `http:www.kleinspecht.de`)

aufbereitet von Dr. Brooks Ferebee, Frankfurt

Die Daten in computergerechter Form:

	G	p1	p2	p3	p4	p5	l1	l2	l3	l4	l5
1	1	0.1719	0.1581	0.1726	0.1785	0.1697	0.0740	0.0703	0.0674	0.0725	0.0660
2	1	0.1052	0.1175	0.0986	0.1008	0.1052	0.0957	0.1023	0.0950	0.0957	0.0943
3	1	0.1473	0.1407	0.1393	0.1407	0.1465	0.0754	0.0776	0.0769	0.0725	0.0653
4	1	0.1378	0.1400	0.1552	0.1828	0.1393	0.0718	0.0667	0.0645	0.0754	0.0747
5	1	0.1473	0.1371	0.1284	0.1509	0.1371	0.0740	0.0696	0.0725	0.0718	0.0718
6	1	0.1175	0.1451	0.1393	0.1407	0.1661	0.0740	0.0711	0.0754	0.0689	0.0565
7	1	0.1385	0.1262	0.1487	0.1407	0.1603	0.0653	0.0696	0.0747	0.0776	0.0725
8	1	0.1197	0.1146	0.1204	0.1182	0.1161	0.0783	0.0805	0.0783	0.0878	0.0696
9	1	0.1393	0.1269	0.1458	0.1429	0.1291	0.0761	0.0761	0.0769	0.0856	0.0725
10	1	0.1197	0.1204	0.1124	0.1146	0.1240	0.0754	0.0769	0.0848	0.0798	0.0645
11	1	0.1625	0.1589	0.1385	0.1502	0.1690	0.0638	0.0689	0.0696	0.0645	0.0529
12	1	0.1298	0.1465	0.1349	0.1400	0.1756	0.0812	0.0747	0.0747	0.0689	0.0602
13	1	0.1204	0.1226	0.1306	0.1465	0.1581	0.0761	0.0754	0.0674	0.0631	0.0689
14	1	0.1110	0.1081	0.1233	0.1248	0.1385	0.0732	0.0747	0.0732	0.0660	0.0587
15	1	0.1139	0.1313	0.1371	0.1589	0.1777	0.0689	0.0674	0.0682	0.0682	0.0711
16	1	0.1335	0.1168	0.1248	0.1313	0.1306	0.0718	0.0703	0.0689	0.0682	0.0667
17	1	0.1407	0.1407	0.1284	0.1400	0.1516	0.0725	0.0696	0.0740	0.0667	0.0696
18	1	0.1204	0.1182	0.1204	0.1269	0.1538	0.0805	0.0718	0.0769	0.0696	0.0645
19	2	0.1044	0.1204	0.1298	0.1393	0.1153	0.1110	0.1211	0.1342	0.0972	0.1037
20	2	0.1436	0.1342	0.1248	0.1581	0.1966	0.1451	0.1400	0.1335	0.1371	0.1240
21	2	0.0907	0.0943	0.0936	0.0936	0.1168	0.0921	0.0812	0.0798	0.0761	0.0674
22	2	0.0921	0.0979	0.1015	0.1015	0.1385	0.0827	0.0827	0.0754	0.0696	0.0653
23	2	0.1052	0.1168	0.1161	0.1306	0.1545	0.0776	0.0732	0.0725	0.0711	0.0609
24	2	0.0928	0.0936	0.0943	0.1066	0.1197	0.0819	0.0863	0.0812	0.0819	0.0805
25	2	0.1516	0.1494	0.1603	0.2140	0.1915	0.1414	0.1429	0.1306	0.1385	0.1044

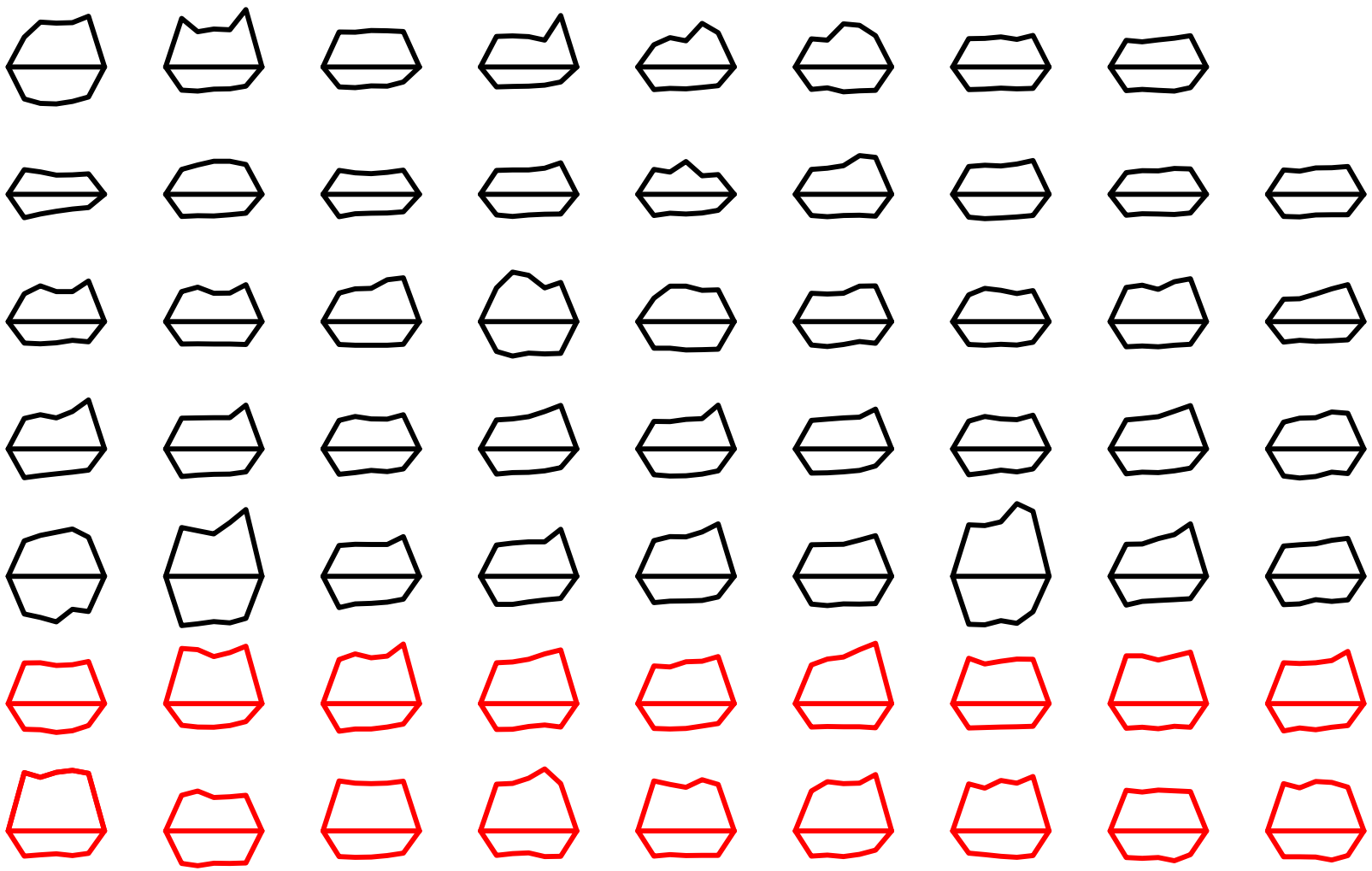
...

Gesucht:

eine dem
menschlichen Gehirn gerechte
Darstellung des Vektors

(p1, p2,p3, p4,p5, l1, l2, l3, l4, l5)

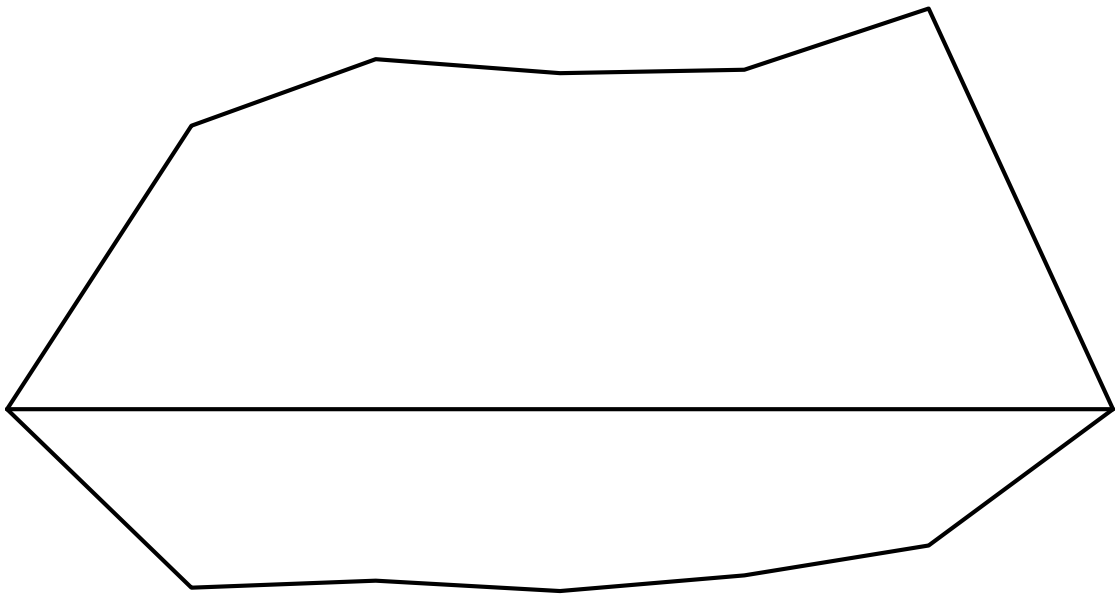
Alle 62 Rufe: rot=Männchen, schwarz=Weibchen



Mit dem Auge kann man Unterschiede erkennen:

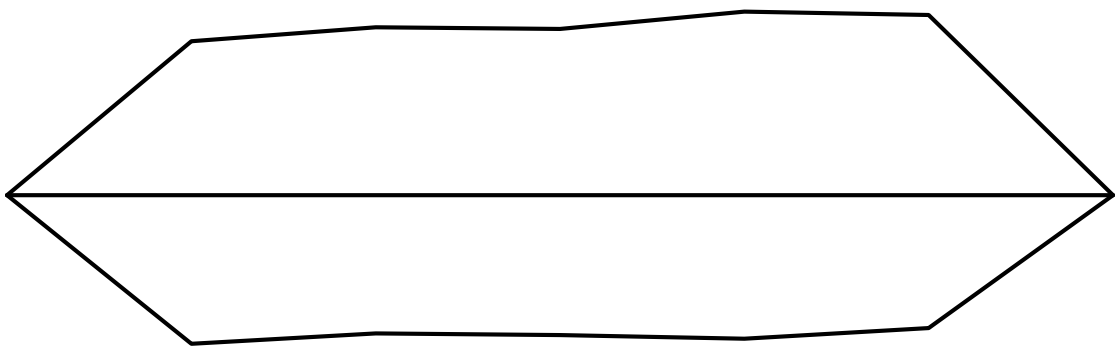
Männchen oder Weibchen?

Typisch Männchen



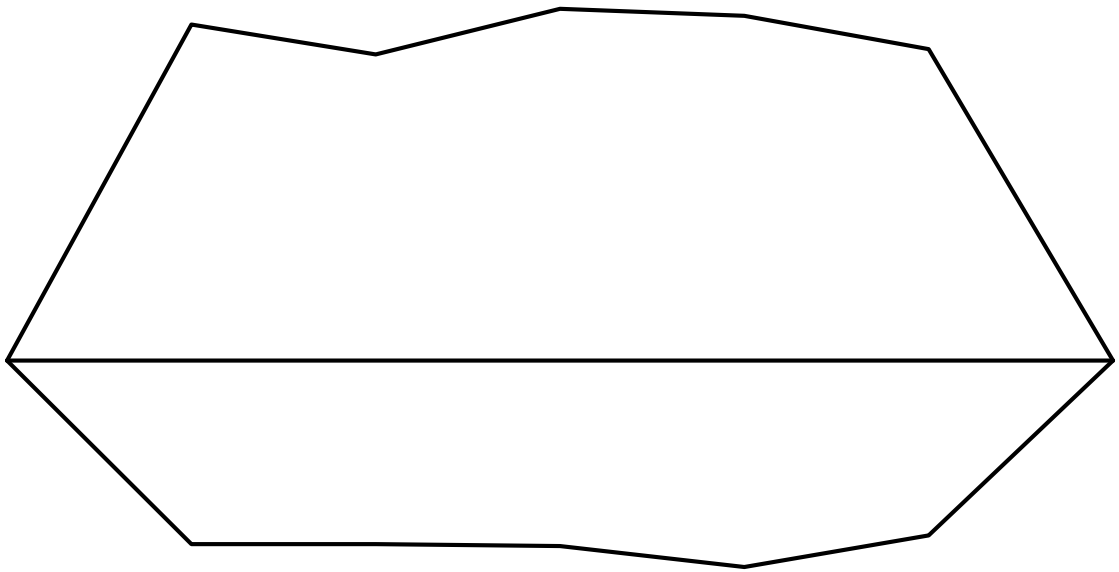
Männchen oder Weibchen?

Typisch Weibchen



Männchen oder Weibchen?

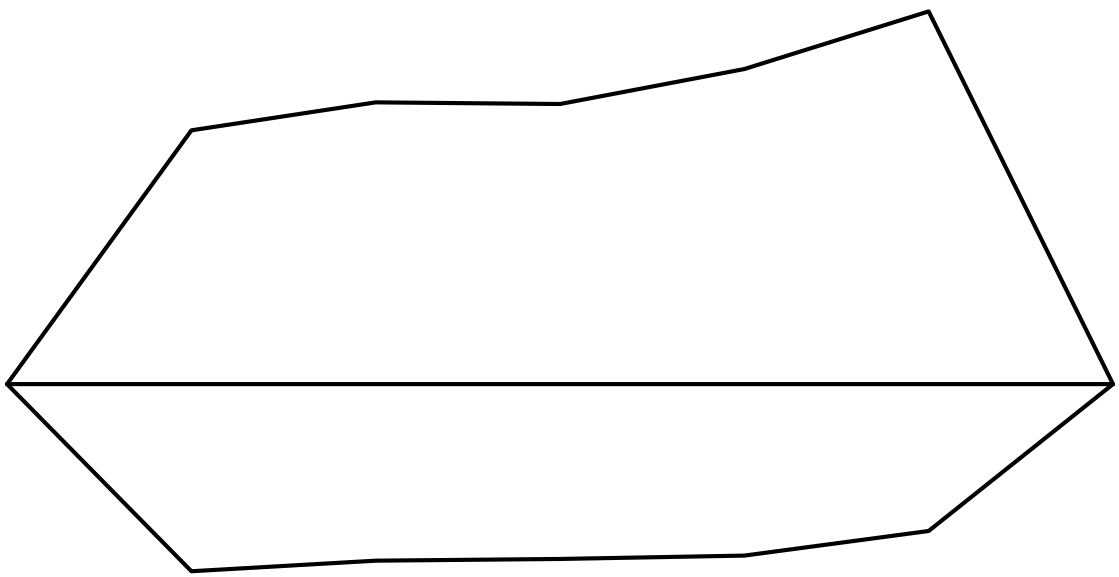
Männchen



Manchmal ist es schwierig:

Männchen oder Weibchen?

Weibchen (untypisch)



Das Auge
(das Gehirn)
sieht Unterschiede.

Schafft es
der Computer
(mit Hilfe der Mathematik)
auch?

Die 10 Zahlen

(p1, p2,p3, p4,p5, l1, l2, l3, l4, l5)

fassen wir als die Koordinaten eines Punktes im 10-dimensionalen Raum $\mathbb{R}^{10}$ auf.

Jeder Ruf entspricht einem Zufallspunkt im $\mathbb{R}^{10}$:

Männchenrufe aus einer Population mit Dichte f_m

Weibchenrufe aus einer Population mit Dichte f_w

Gesucht: Eine Regel, die jeden neuen Punkt

x = (p1, p2,p3, p4,p5, l1, l2, l3, l4, l5)

einer der beiden Populationen zuweist.

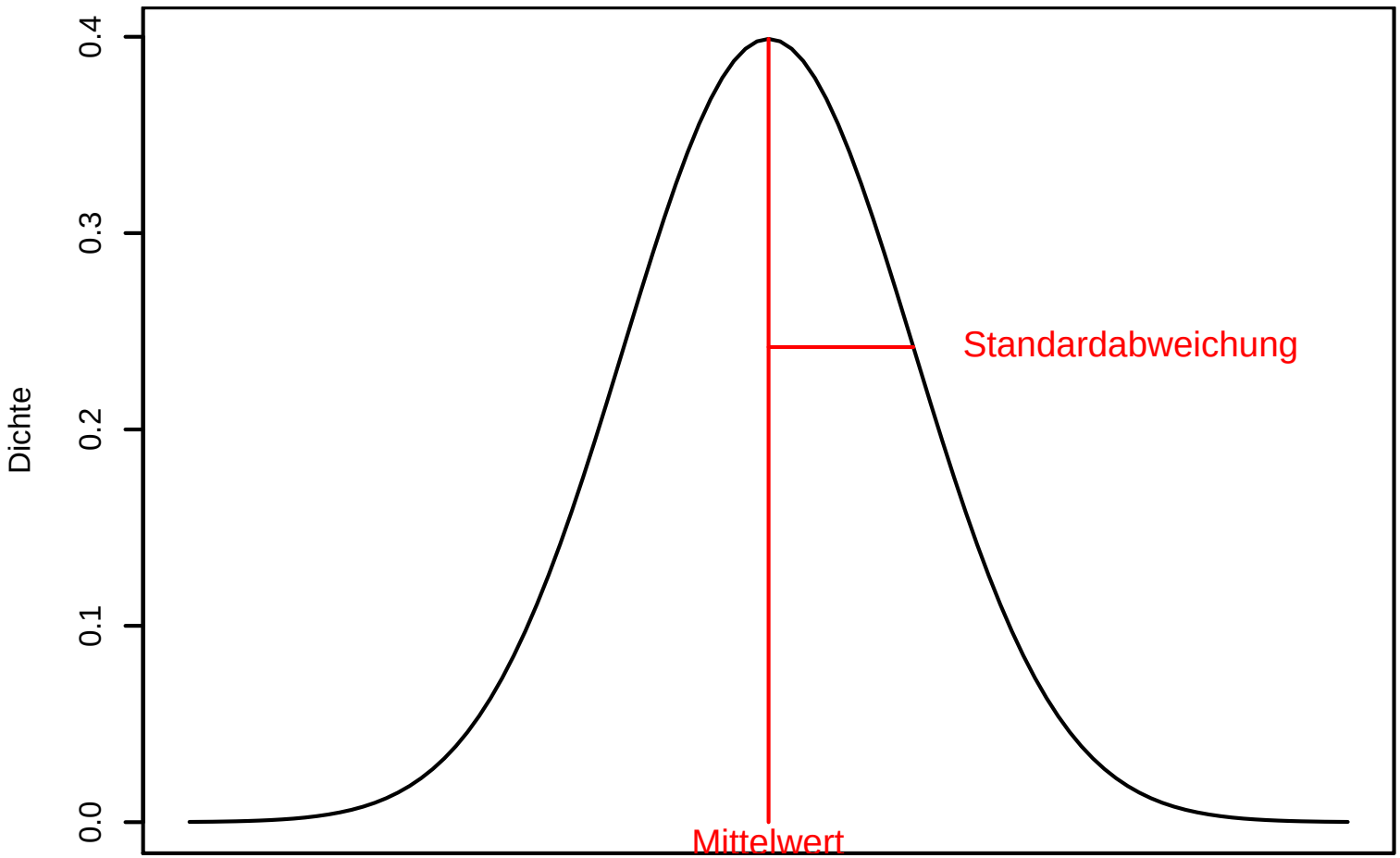
Verfahren

1. Schätze f_m und f_w
2. Ordne x der Population mit dem größeren f-Wert zu.

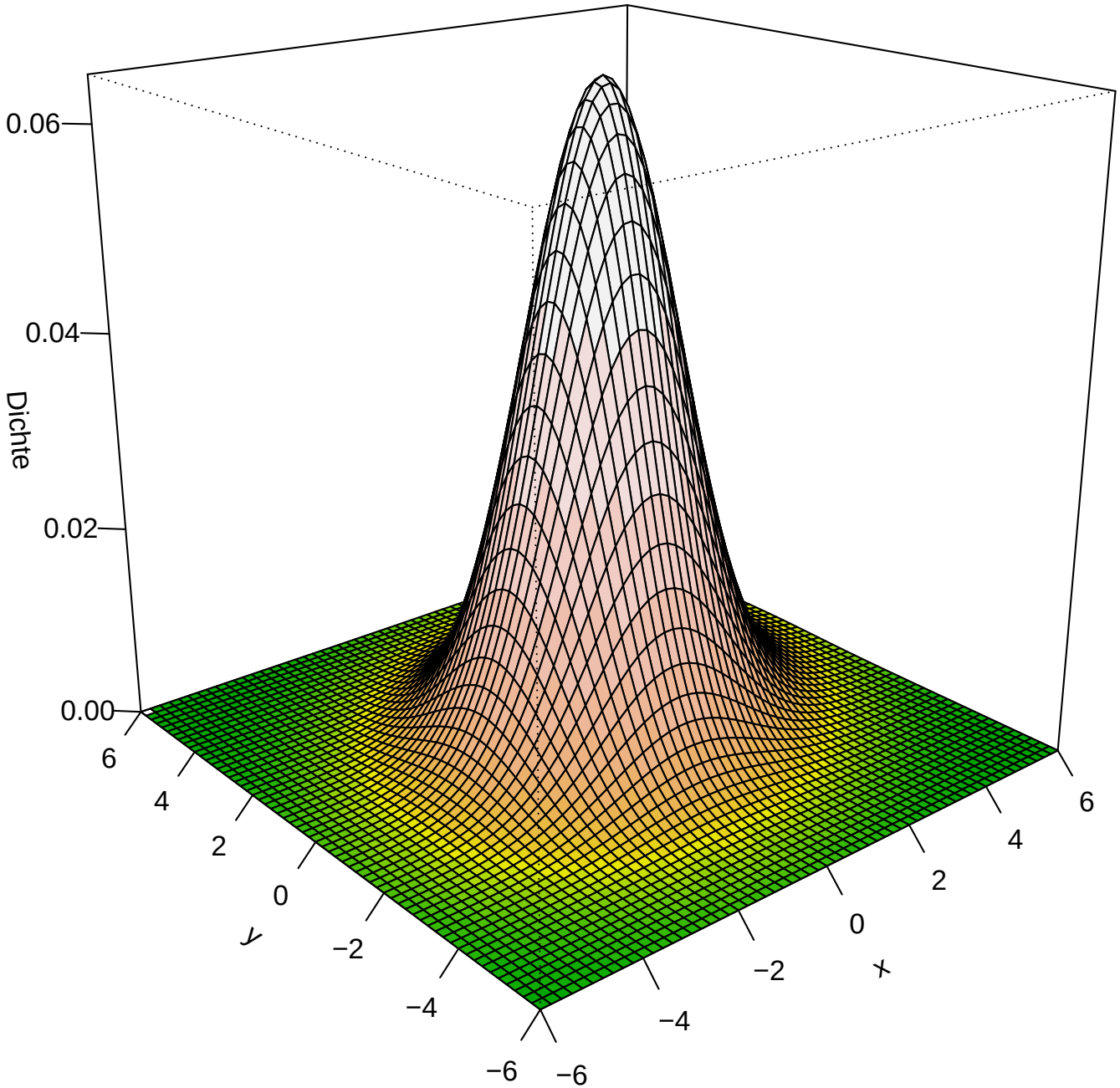
Wir benutzen für f_m und f_w mehrdimensionale Normalverteilungen.

Vorteil: Leicht anzupassen. Wir müssen nur Mittelwert(svektor) und Varianz (mehrdimensional: die Kovarianzmatrix) schätzen.

Erinnerung: Eindimensionale Normalverteilung

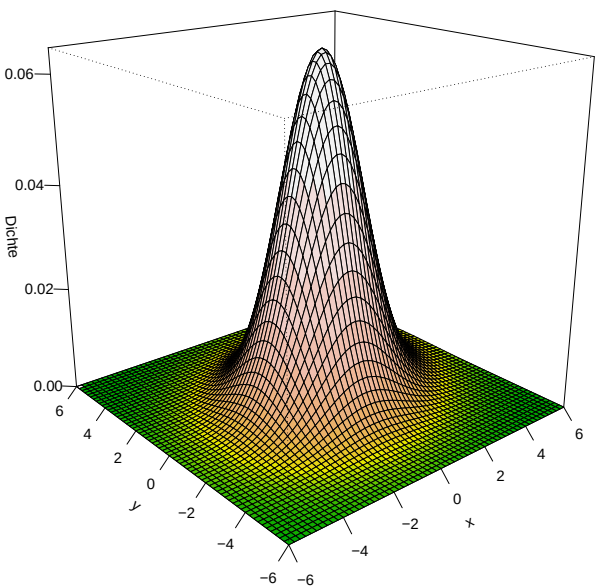


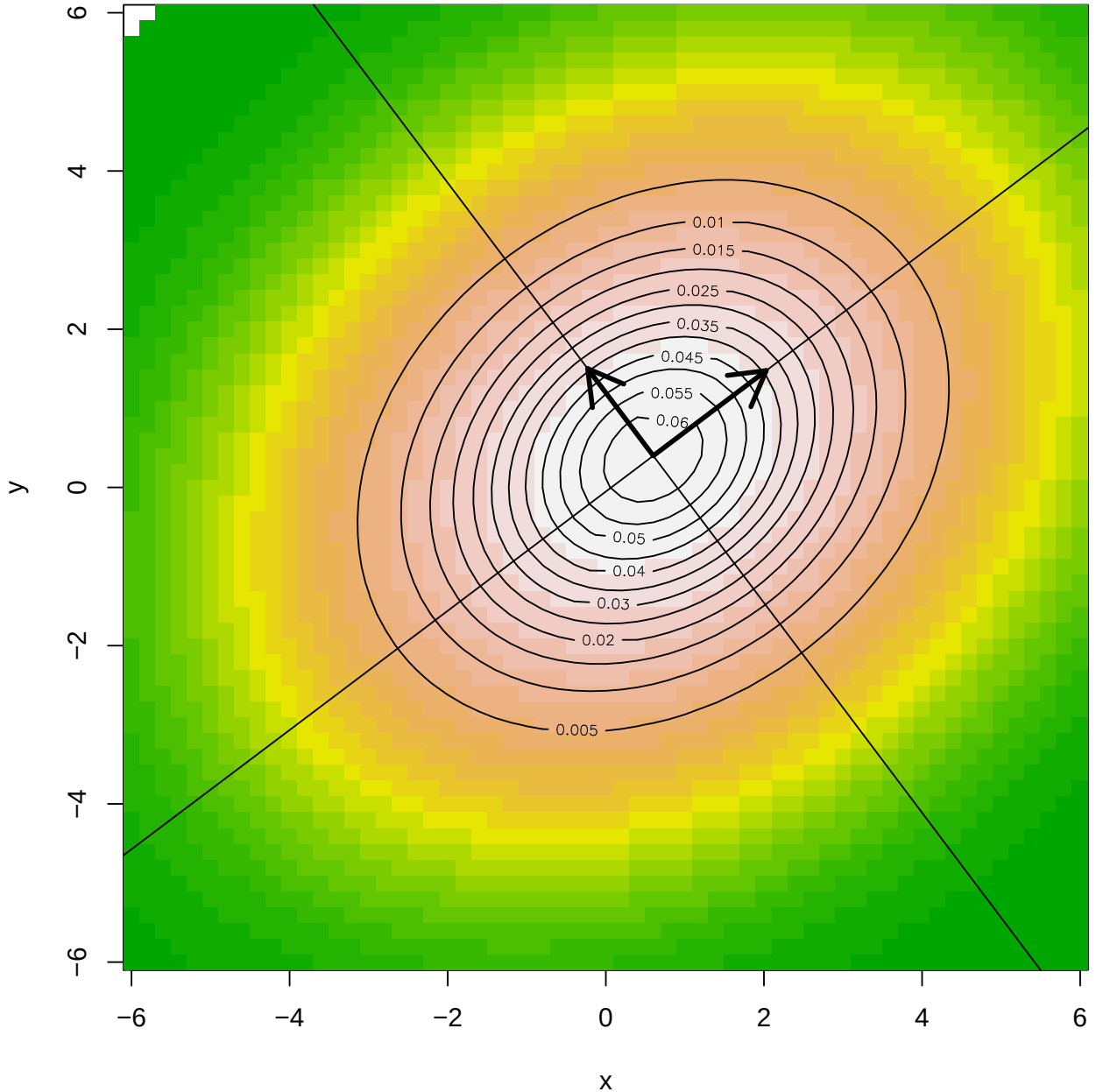
Dichte einer zweidimensionalen Normalverteilung



Zur Beschreibung einer mehrdimensionalen Normalverteilung benötigt man

- Einen Mittelwertvektor μ
- Ein Achsenkreuz (die „Hauptachsen“)
- Standardabweichungen in den Achsenrichtungen


0.06
0.04
0.02
0.00
Dichte
6
4
2
0
−2
−4
−6
y
x

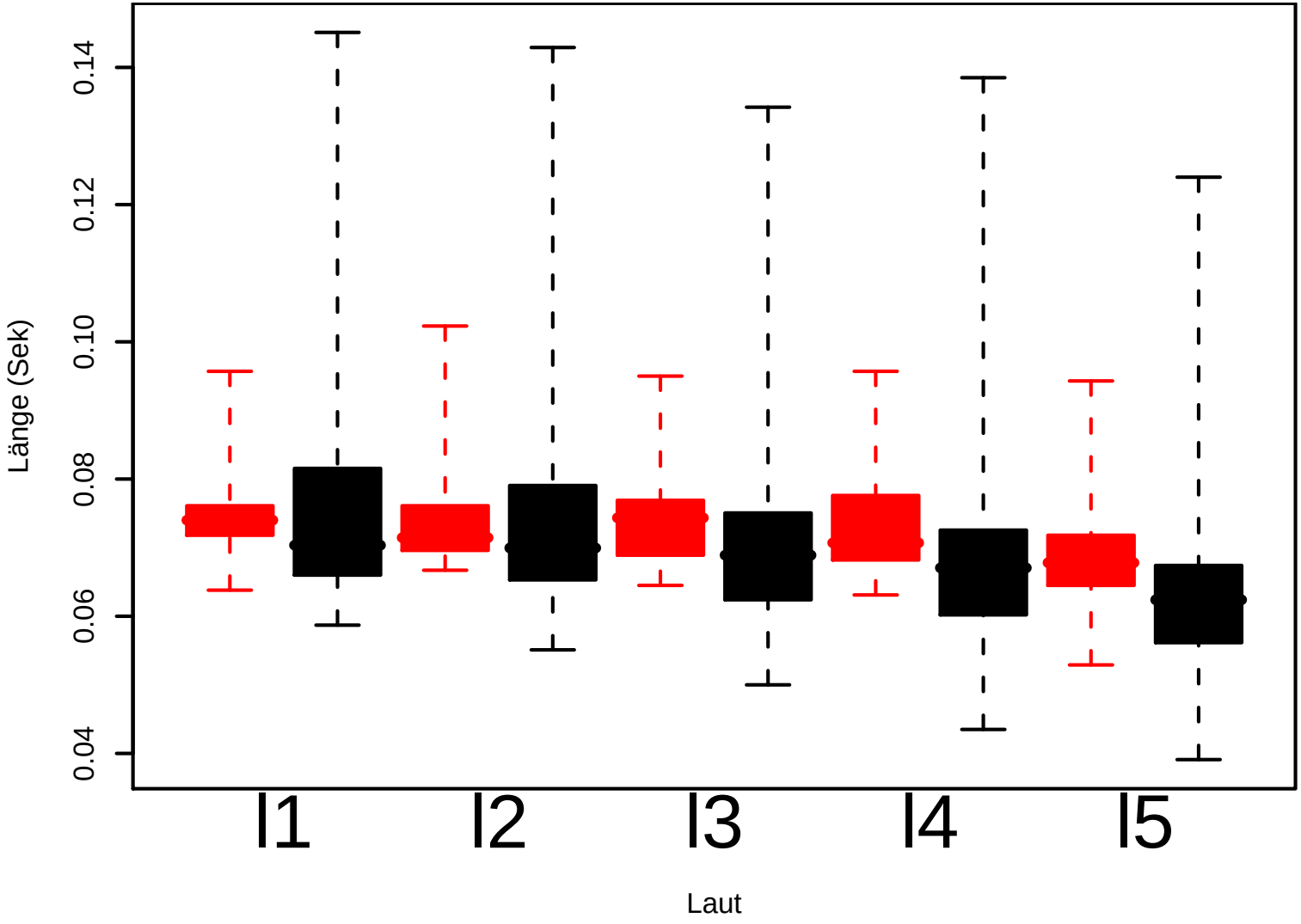

0.01
0.015
0.025
0.035
0.045
0.055
0.06
0.05
0.04
0.03
0.02
0.005
6
4
2
0
−2
−4
−6
y
x

In unserem Problem gibt es 10 Dimensionen.

Wir beginnen eindimensional.

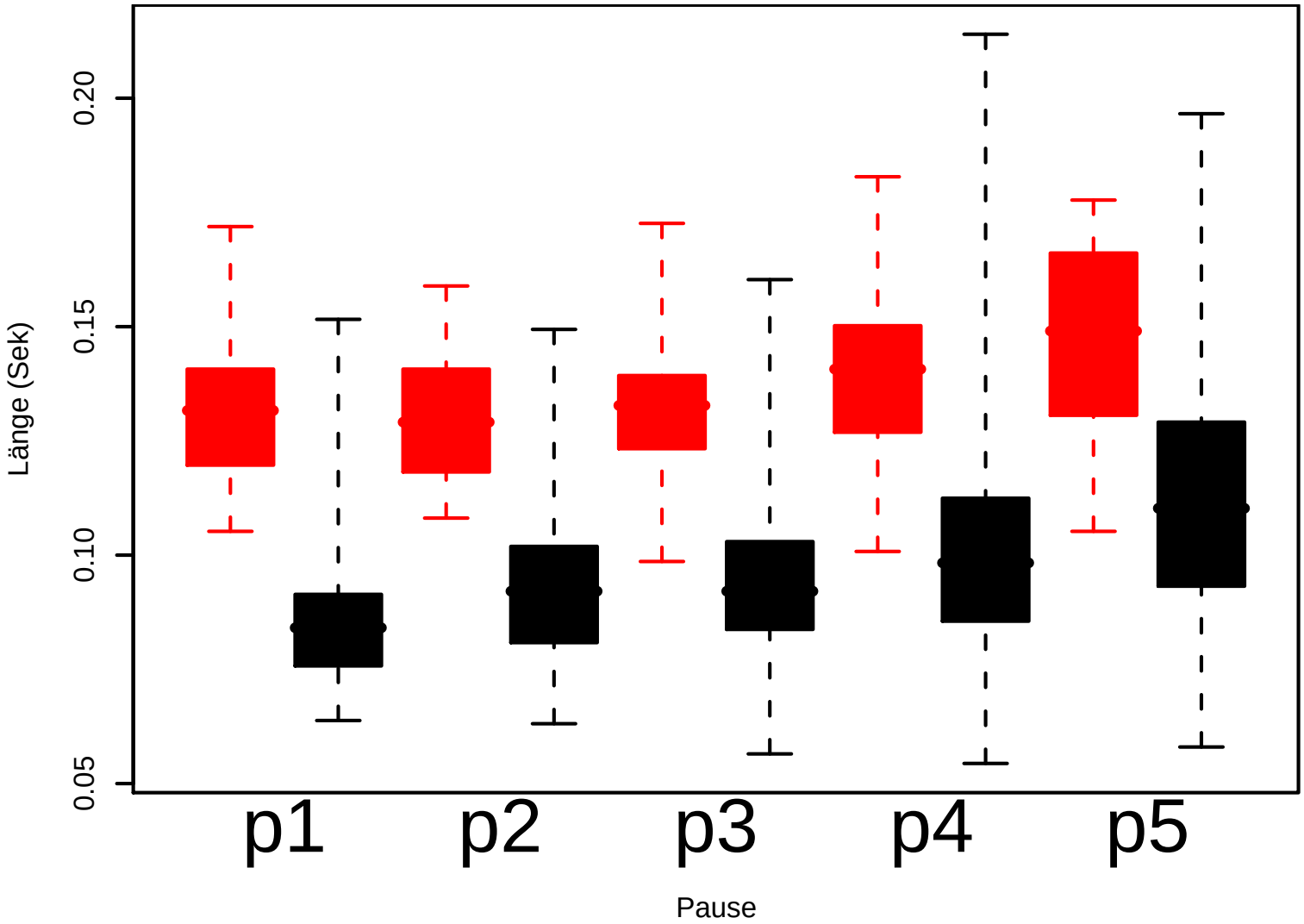
Frage: Welche eine der 10 Variablen sollen wir wählen?

Länge der Laute bei Männchen (rot) und Weibchen (schwarz)



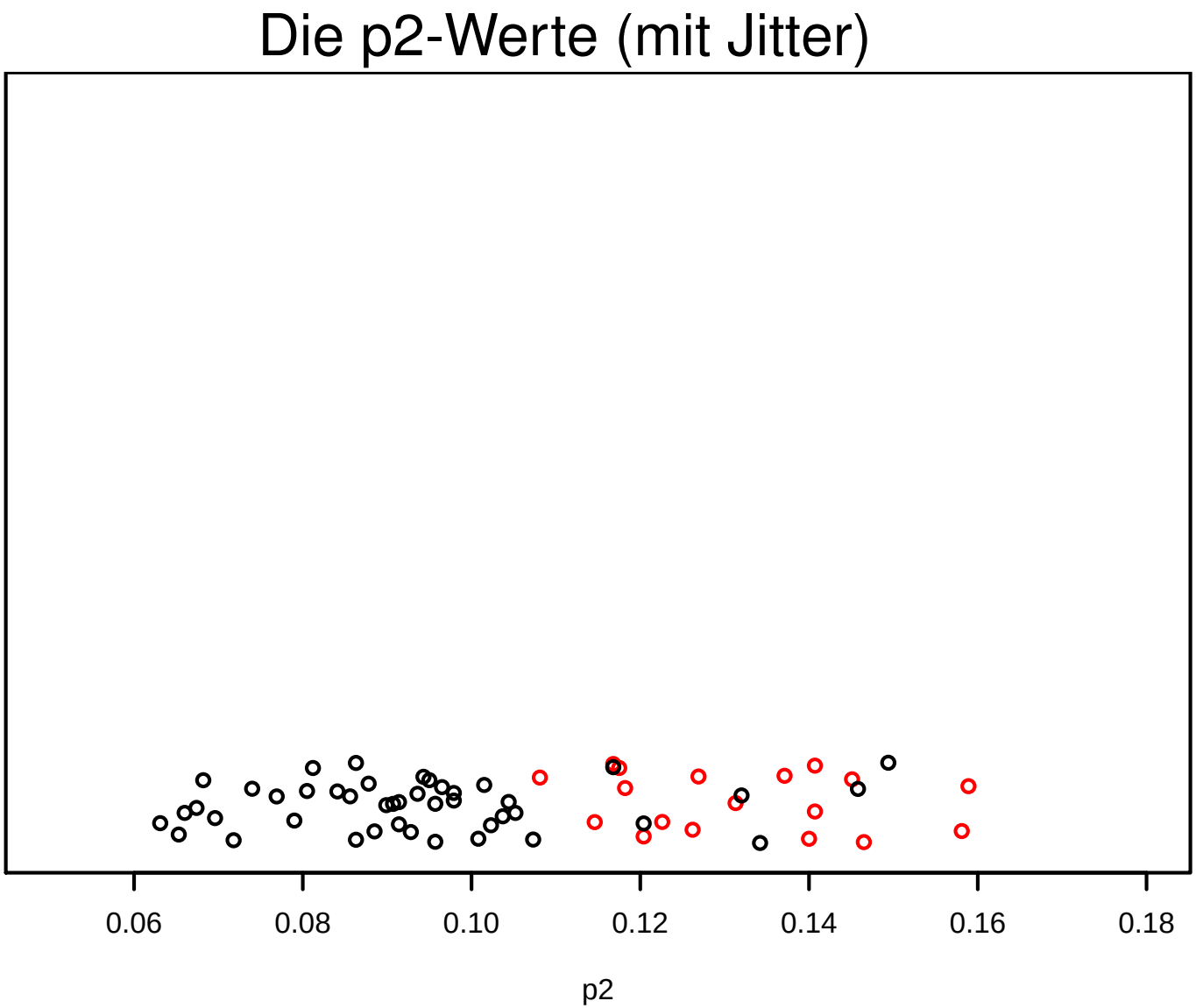
Keine gute Trennung der Geschlechter

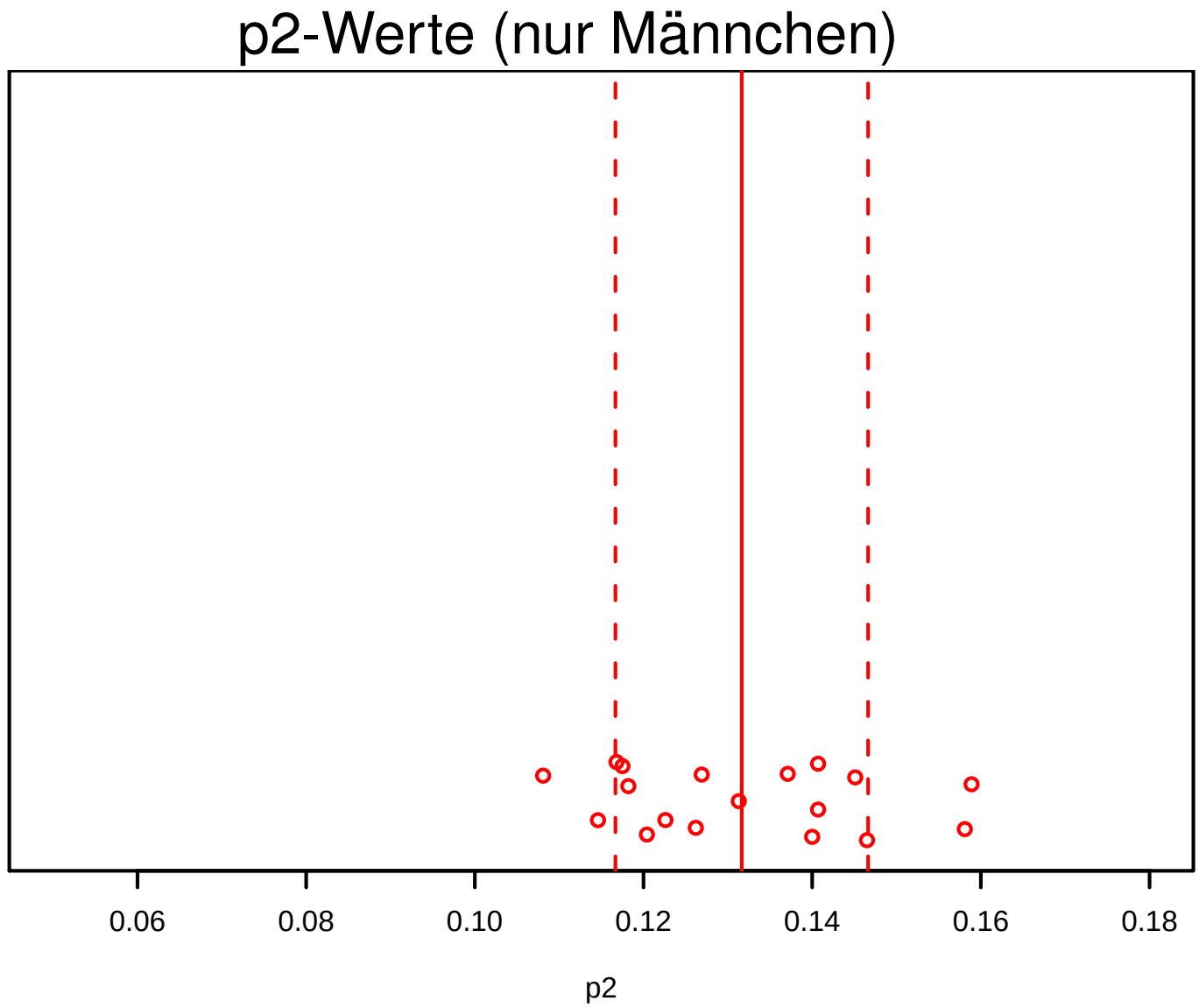
Länge der Pausen bei Männchen (rot) und Weibchen (schwarz)



Bei den Männchen sind die Pausen typischerweise länger

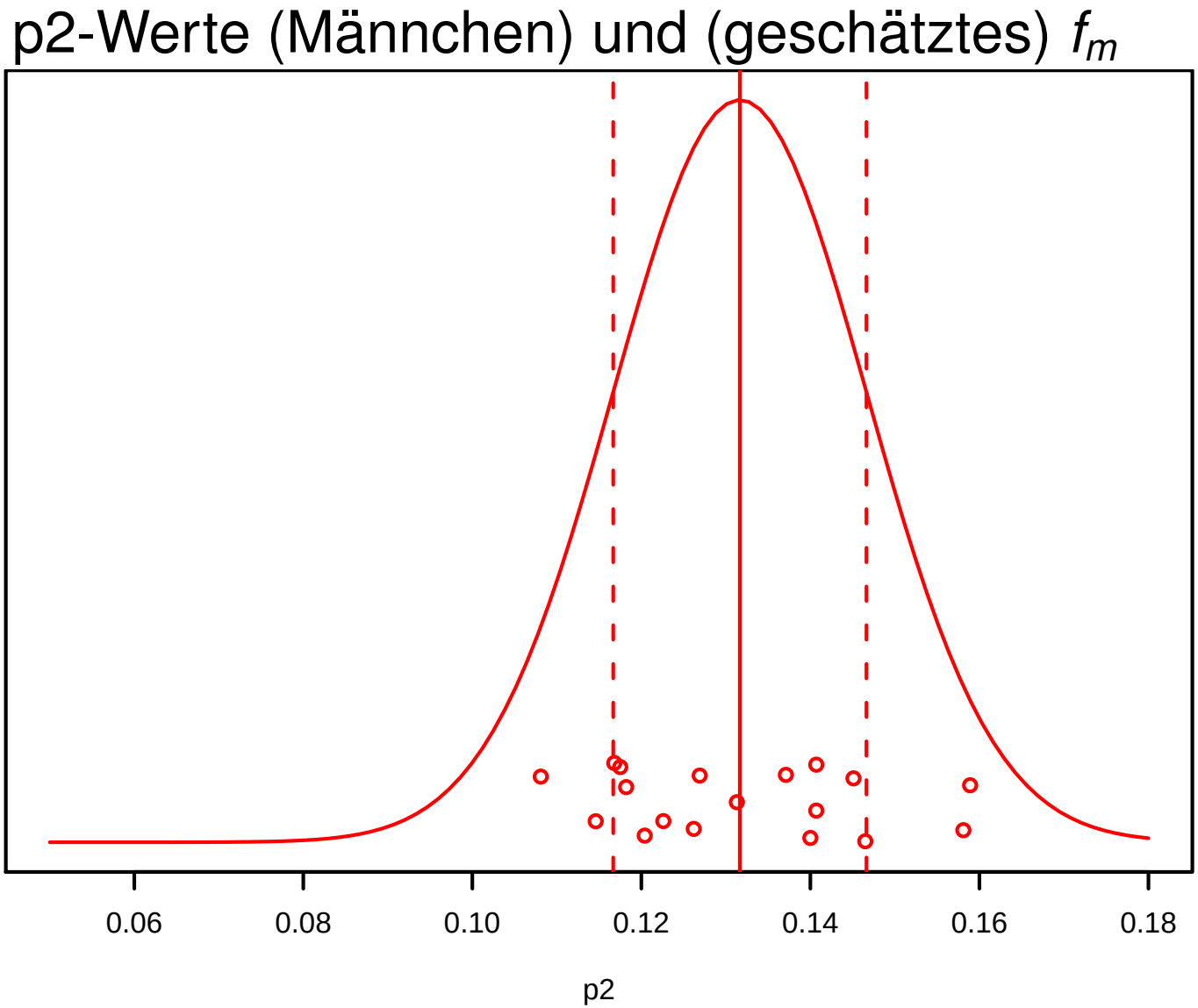
Wie gut läßt sich das Geschlecht anhand von p2, der Länge der zweiten Pause, bestimmen?

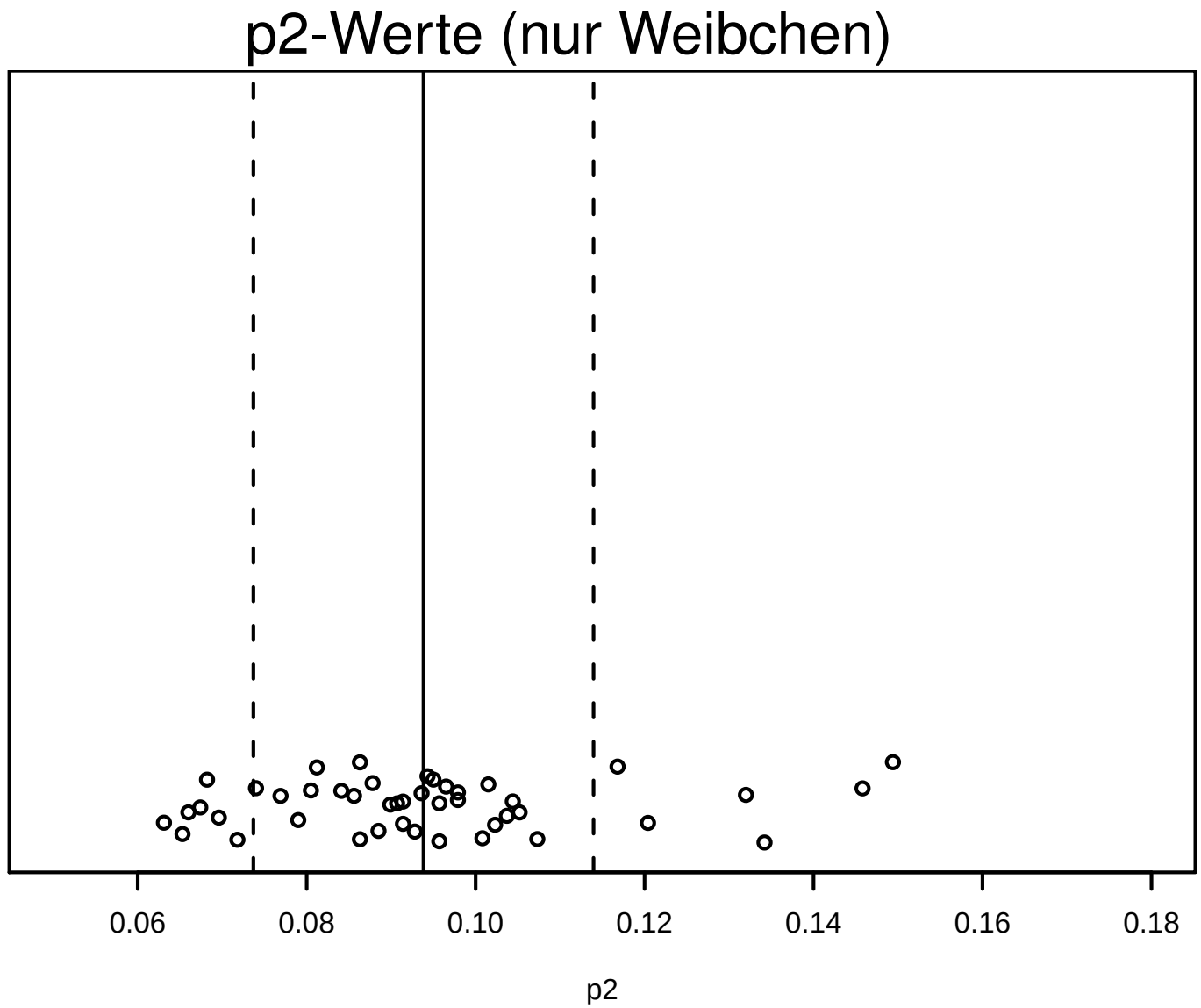

Die p2-Werte (mit Jitter)
0.06
0.08
0.10
0.12
0.14
0.16
0.18
p2



Mittelwert $\mu_m = 0{,}1316$, Standardabweichung $\sigma_m = 0{,}0150$

Wir approximieren f_m durch die Normalverteilung mit Mittelwert μ_m und Standardabweichung σ_m

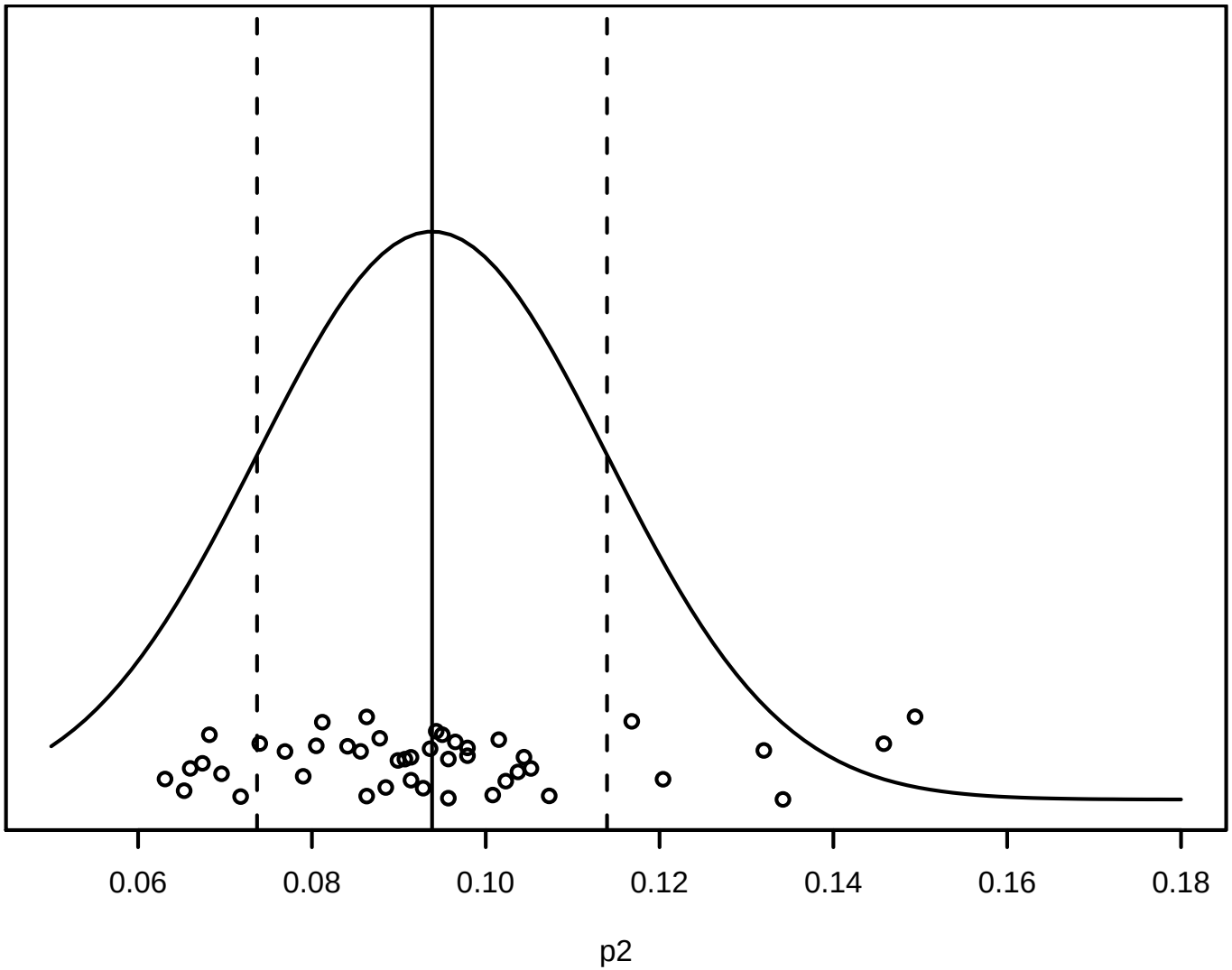




Mittelwert $\mu_w = 0{,}0938$, Standardabweichung $\sigma_m = 0{,}0201$

Wir approximieren f_w durch die Normalverteilung mit Mittelwert μ_w und Standardabweichung σ_w

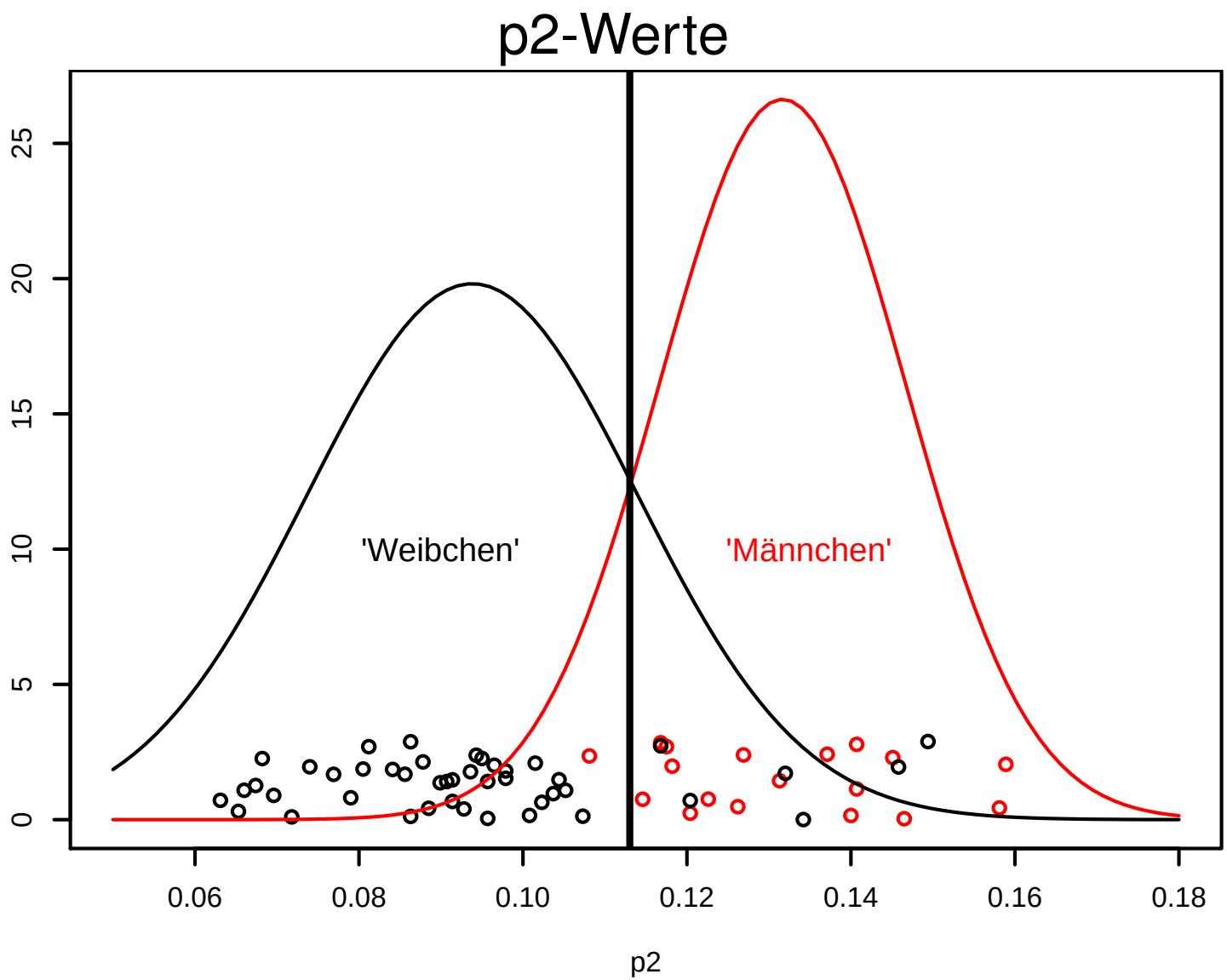
p2-Werte (Weibchen) und (geschätztes) f_w



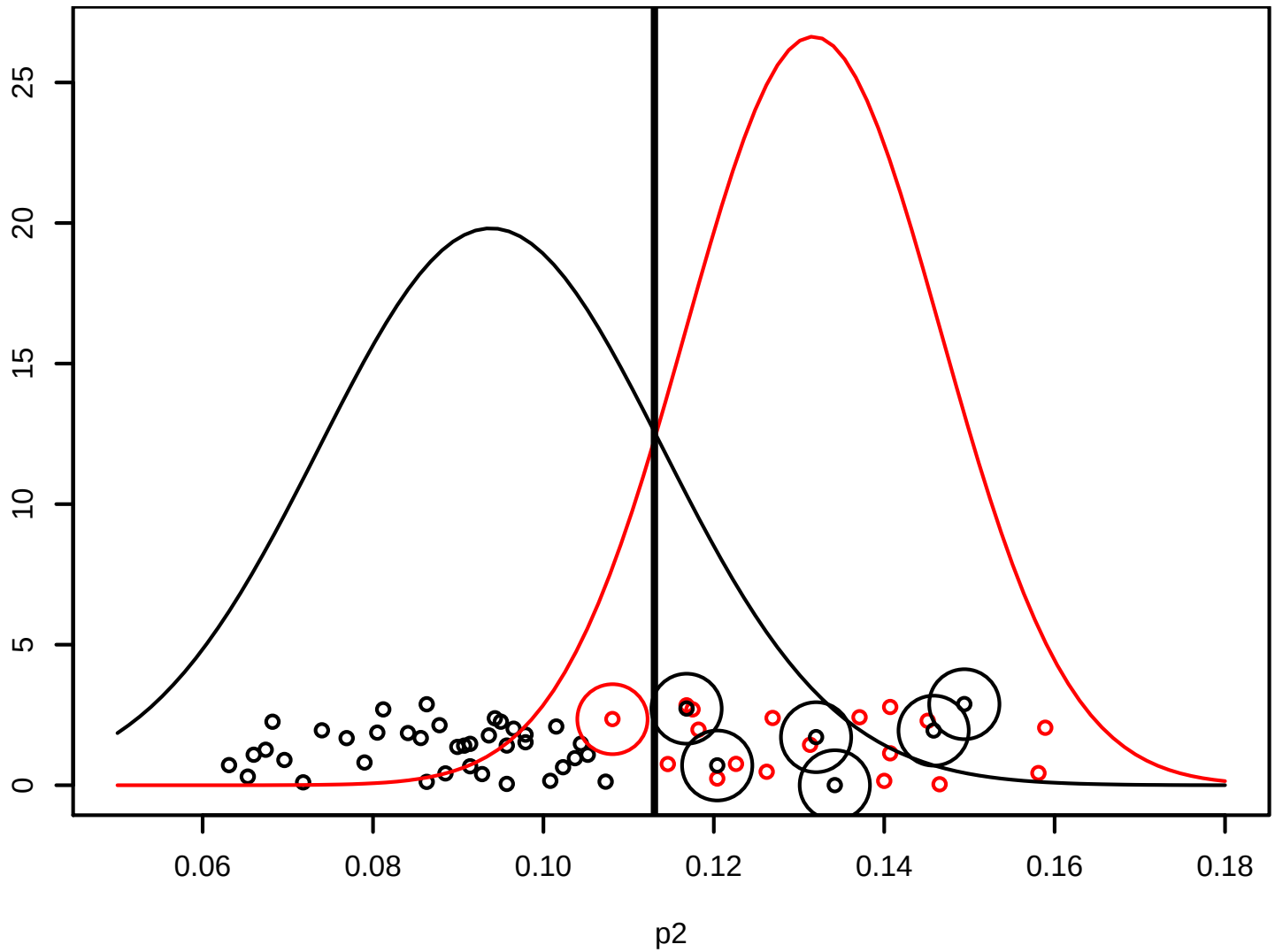
Klassifikationsregel:

f_m größer $\longrightarrow$ „Männchen“

f_w größer $\longrightarrow$ „Weibchen“


p2-Werte
'Weibchen'
'Männchen'
0
5
10
15
20
25
0.06
0.08
0.10
0.12
0.14
0.16
0.18
p2

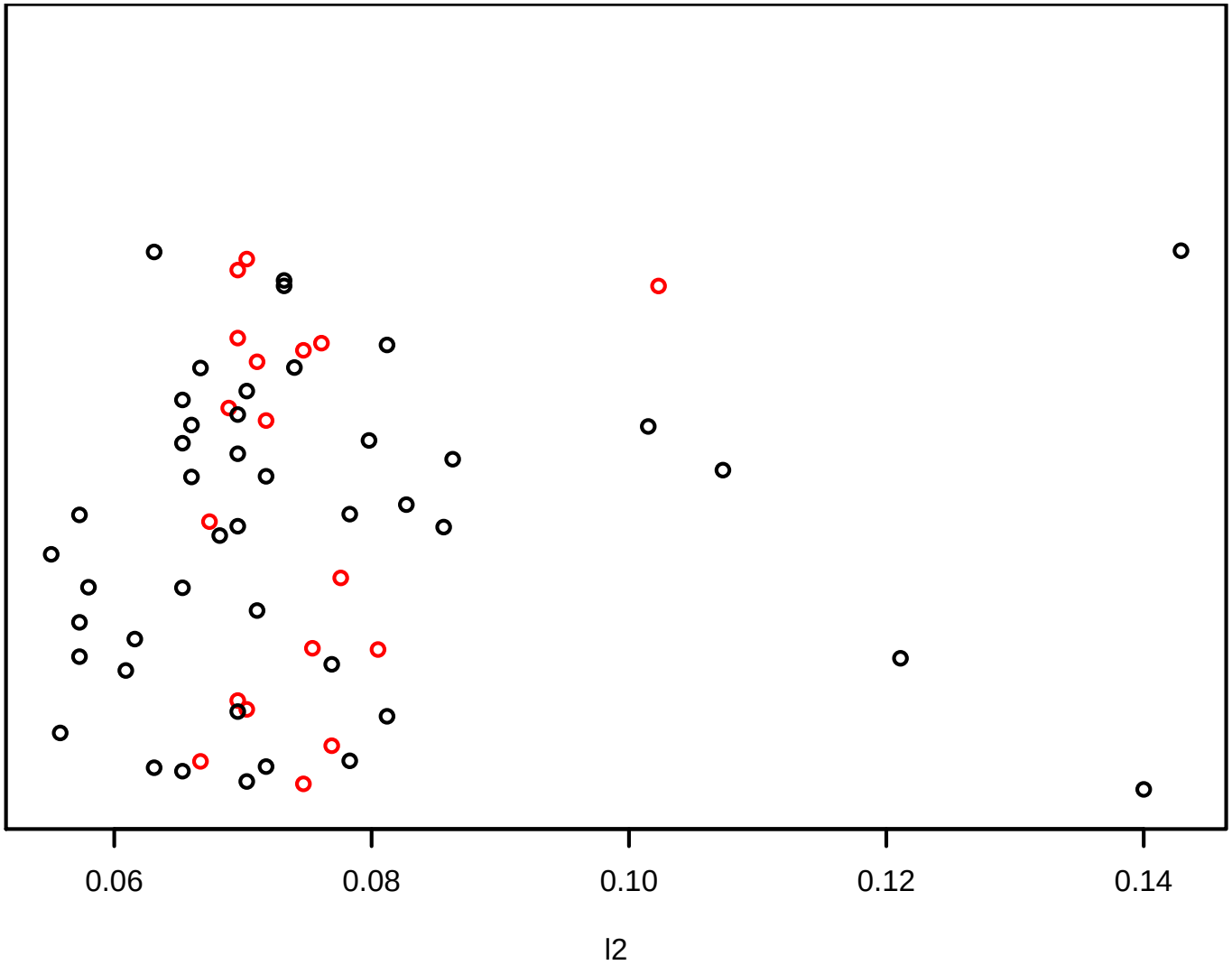
Falsch klassifiziert:
1 Männchen 6 Weibchen



Zur Verbesserung der Klassifikation nehmen wir mehr Information hinzu, z.B. eine weitere Variable.

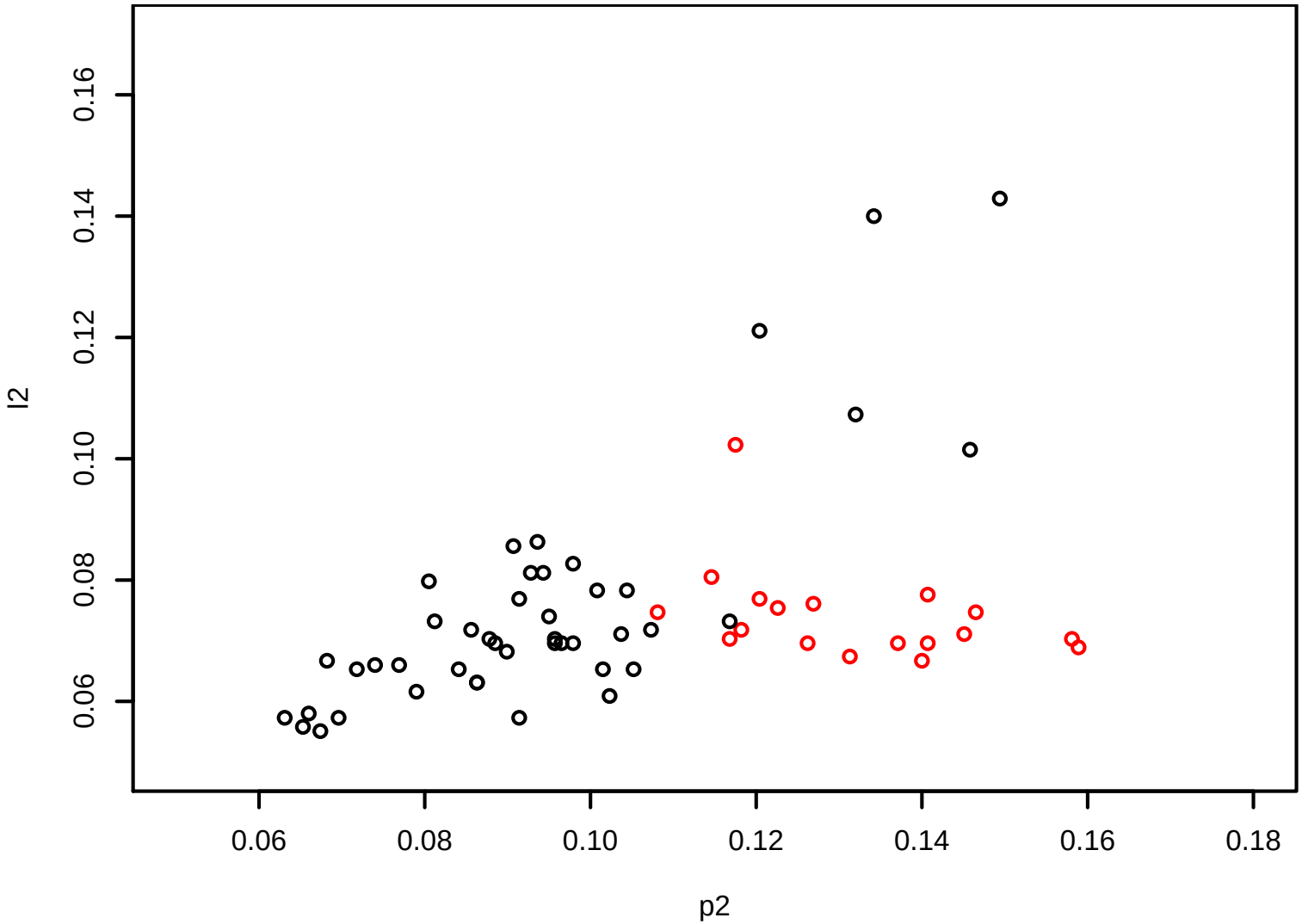
Wir betrachten:

Erste Variable = p2
Zweite Variable = l2



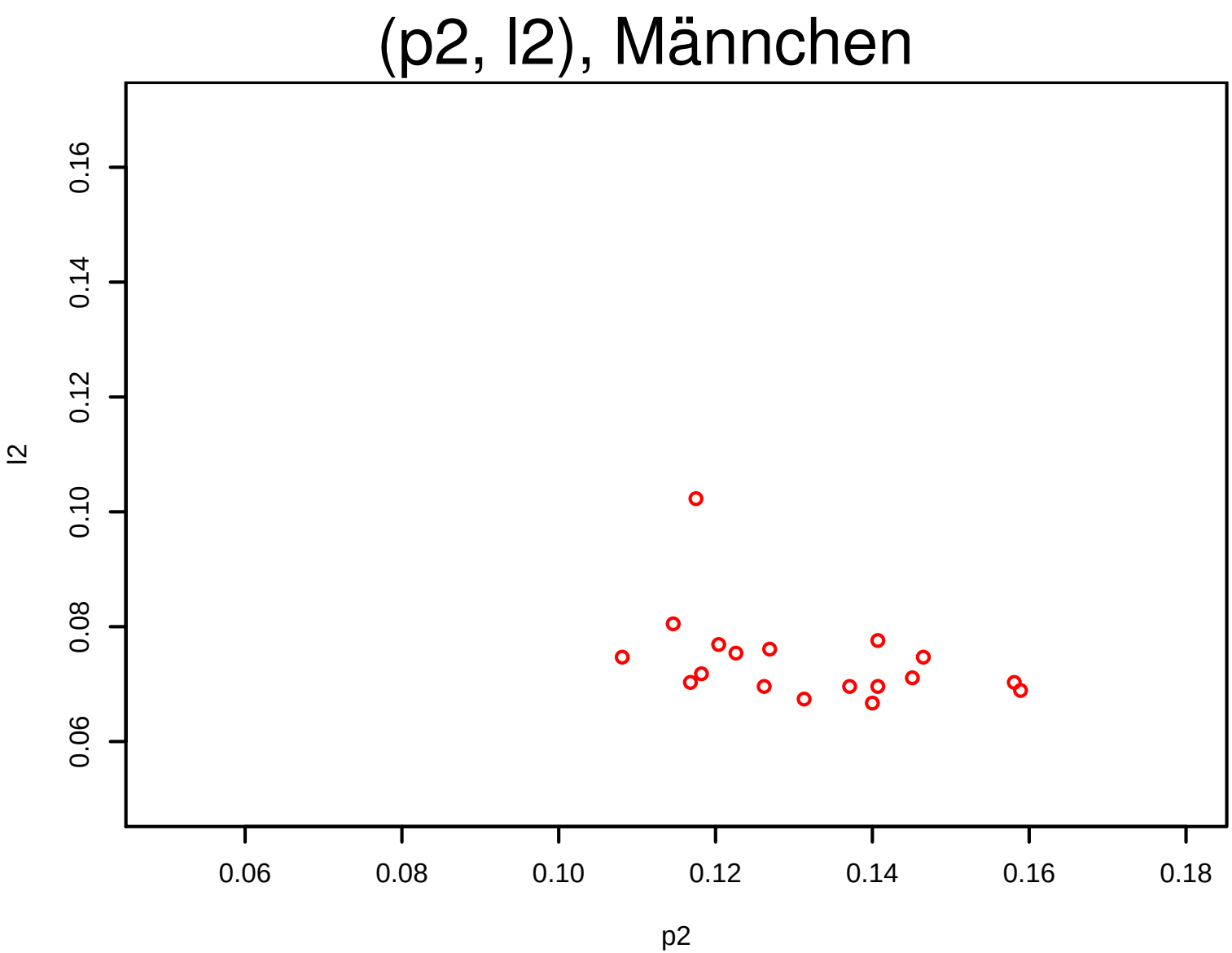
Beobachtung: l2 allein trennt die Geschlechter sehr schlecht.

Aber: l2 zusammen mit p2 gibt zusätzliche Information:



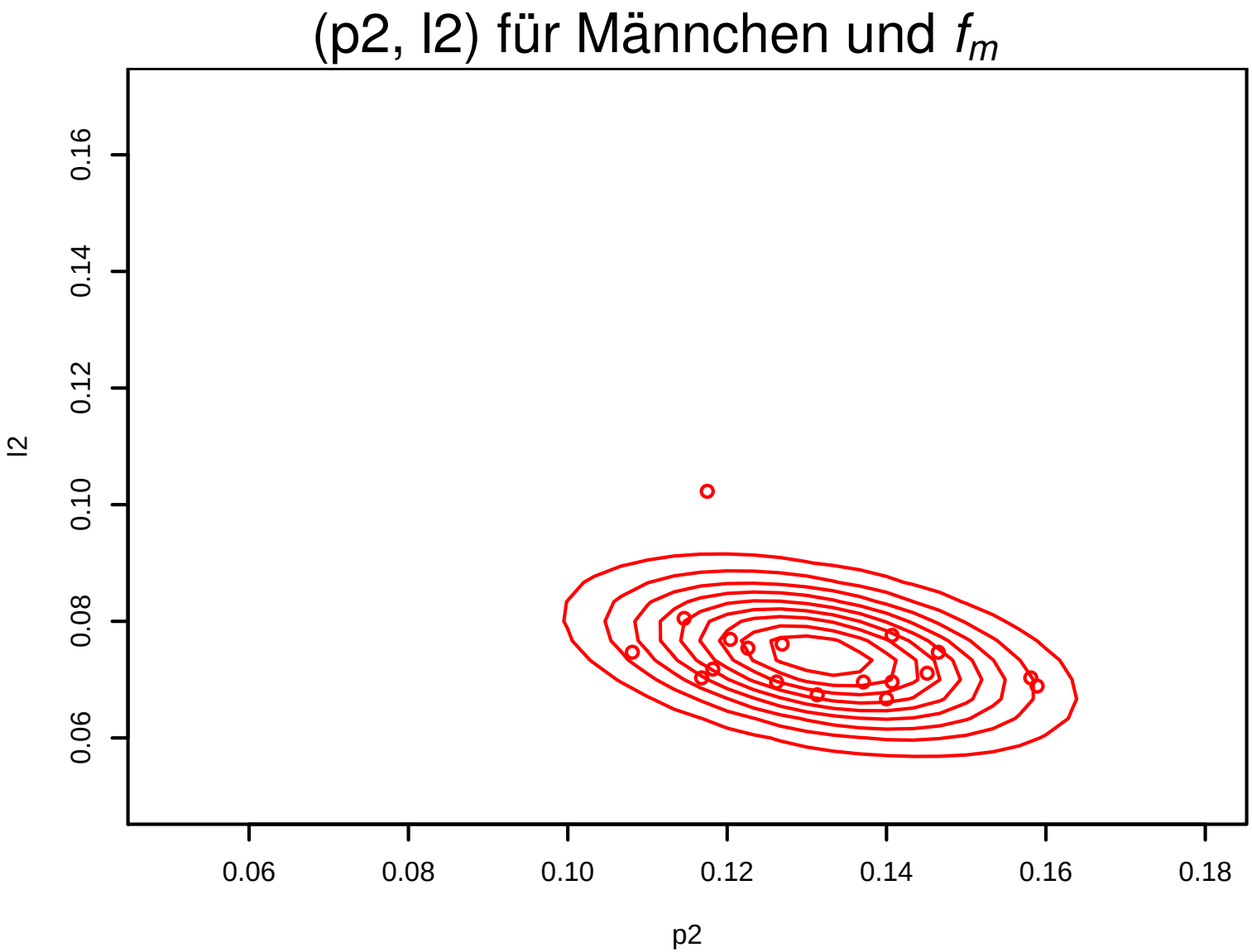
Beispielsweise zeigt die Hinzunahme von l2, dass die 5 Punkte oben rechts besser zu den Weibchen passen.

Wir approximieren
die Verteilungen
von (p2, l2) bei Männchen und bei Weibchen
durch
zweidimensionale
Normalverteilungen.

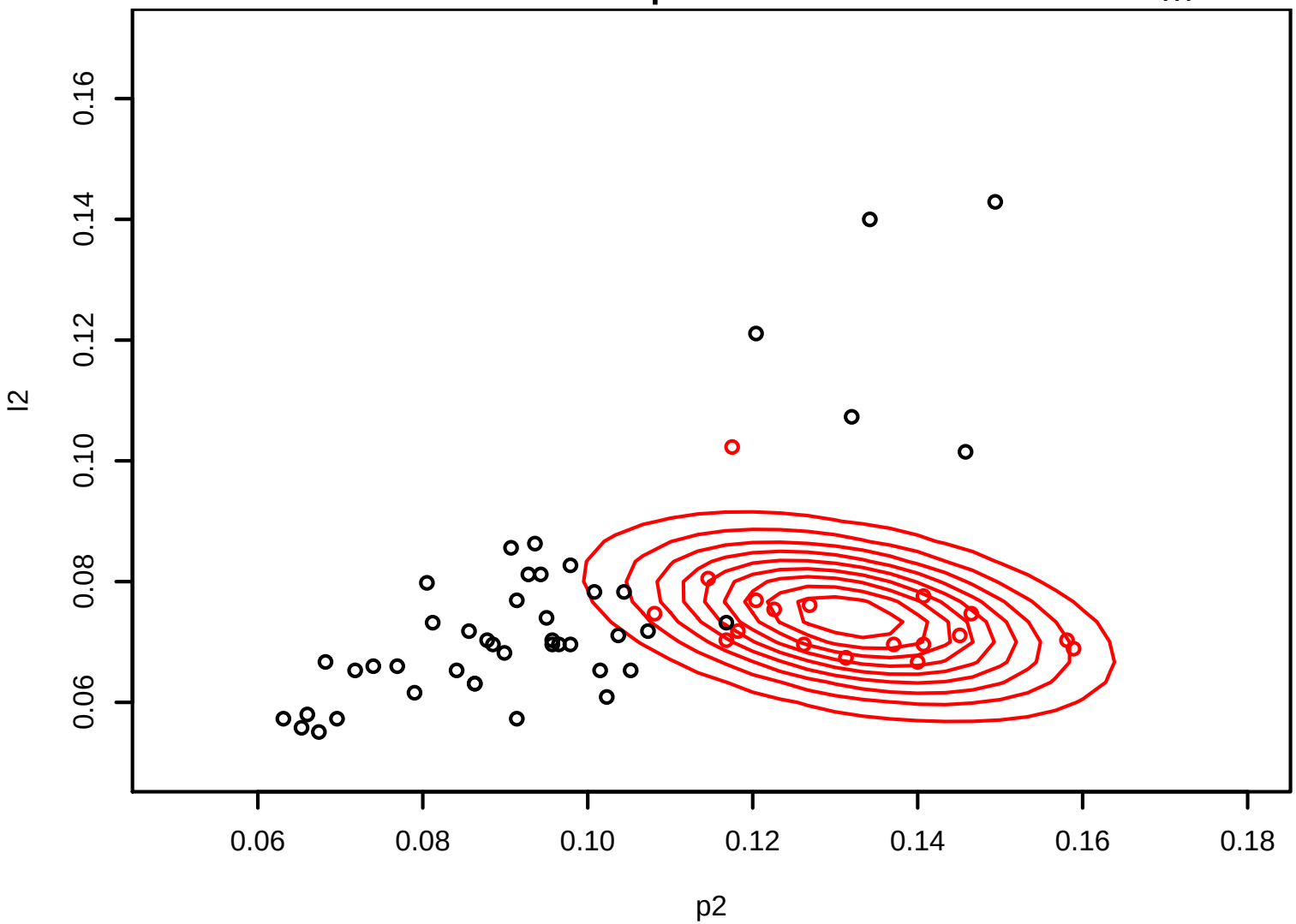

(p2, l2), Männchen
l2
0.06
0.08
0.10
0.12
0.14
0.16
p2
0.06
0.08
0.10
0.12
0.14
0.16
0.18

Wie im eindimensionalen Fall schätzen wir
den (zweidimensionalen) Mittelwert
und die (zweidimensionale) Varianz
(d.h. die sog. Kovarianzmatrix)

und approximieren f_m durch eine
zweidimensionale Normalverteilung
mit dem geschätzten Mittelwert
und der geschätzten Varianz.

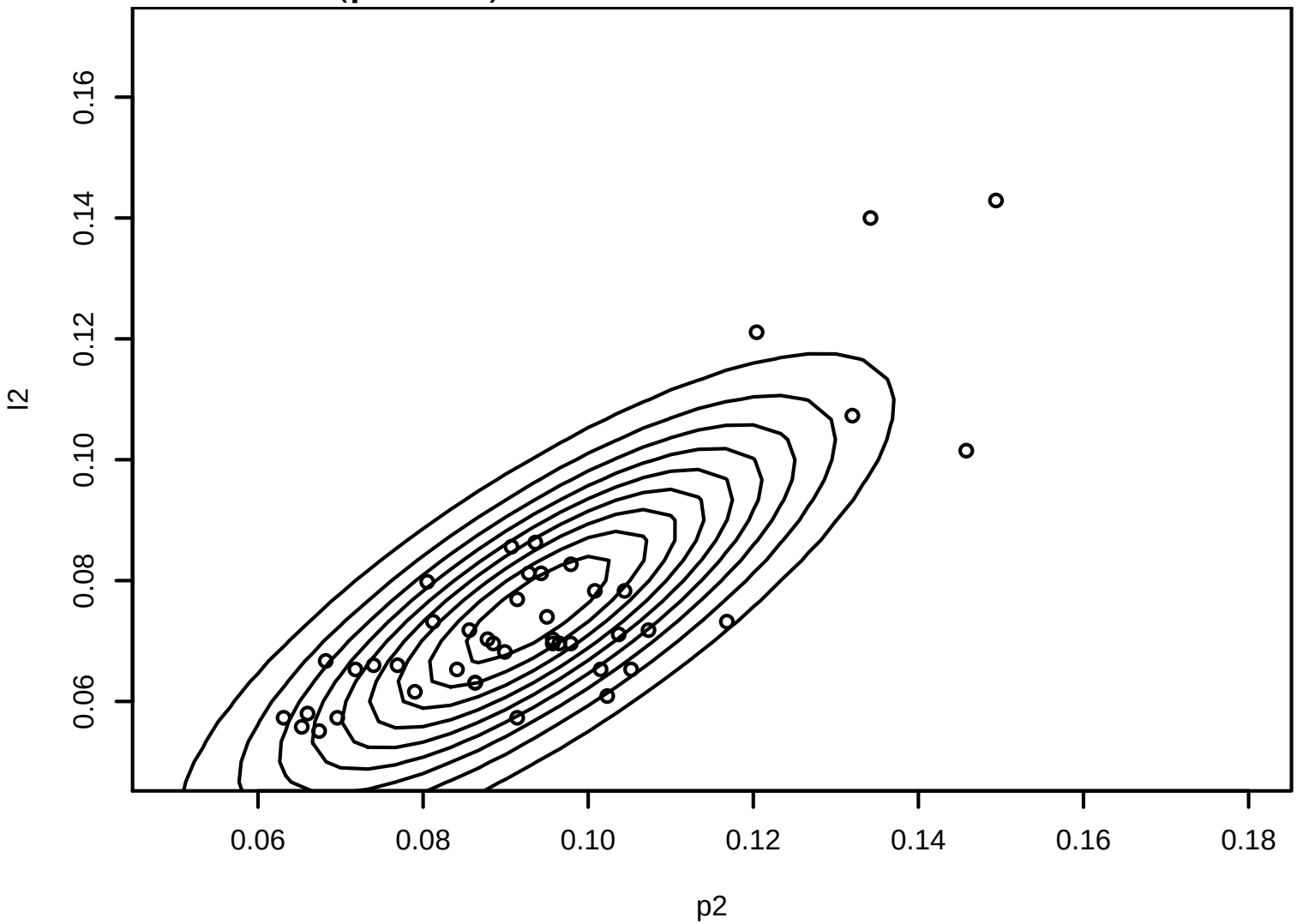

(p2, l2) für Männchen und f_m
l2
0.06
0.08
0.10
0.12
0.14
0.16
p2
0.06
0.08
0.10
0.12
0.14
0.16
0.18

Viele der Weibchen passen schlecht zu f_m:

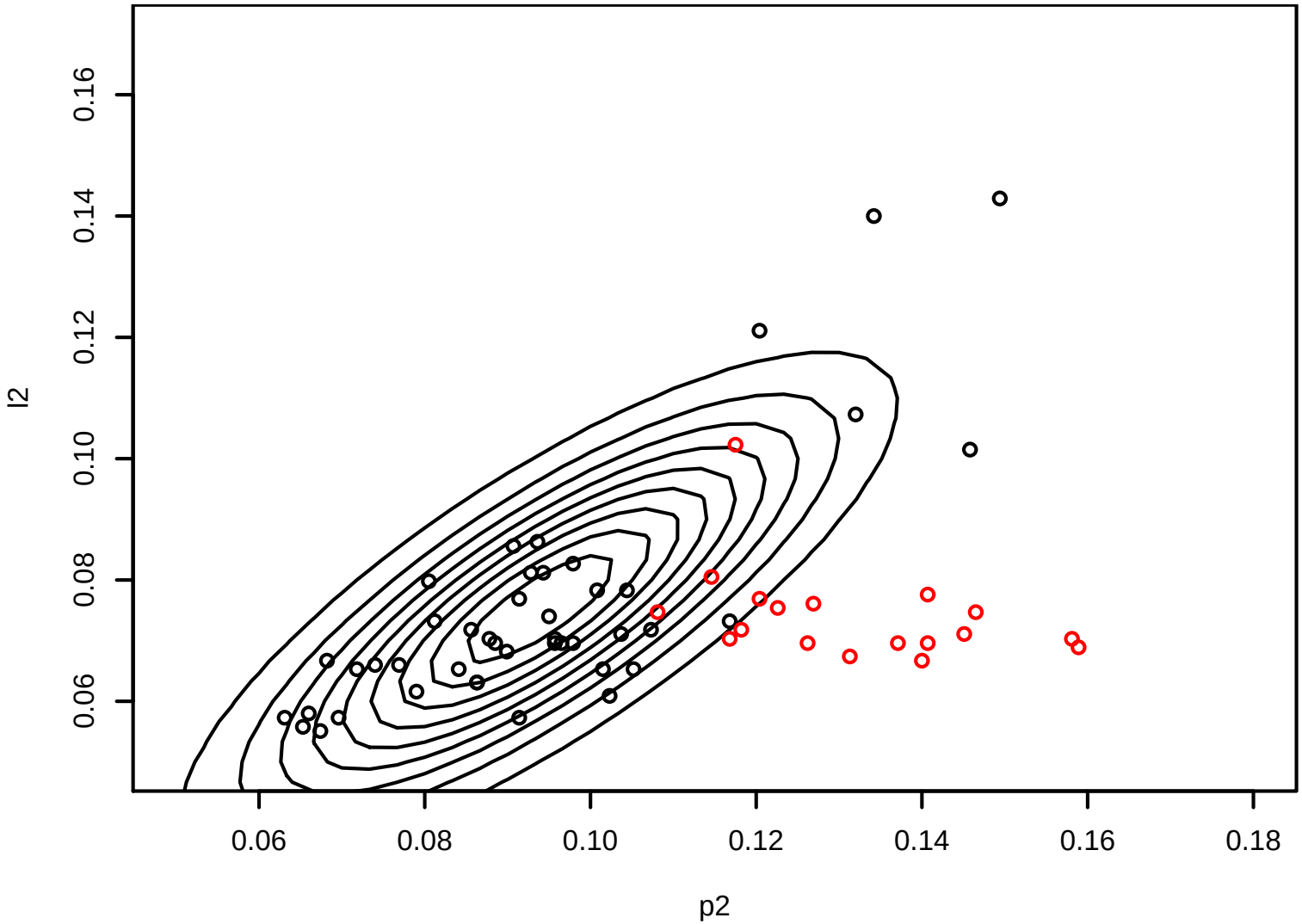


Analog für die Weibchen:

(p2, l2) für Weibchen und f_w



Viele der Männchen passen schlecht zu f_W:



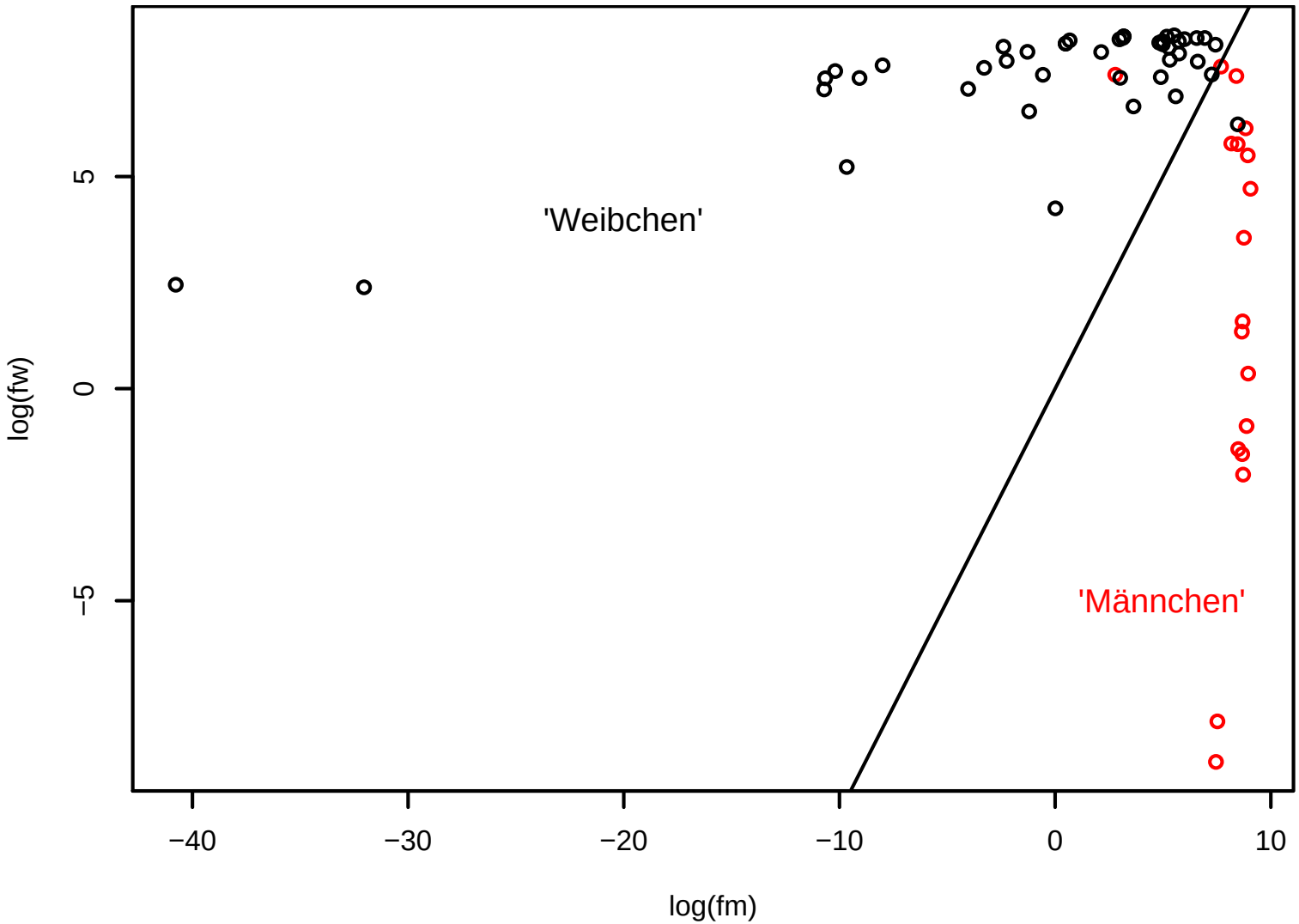
Klassifikation:

Für jeden Punkt berechnen wir $f_m(x)$ und $f_w(x)$.

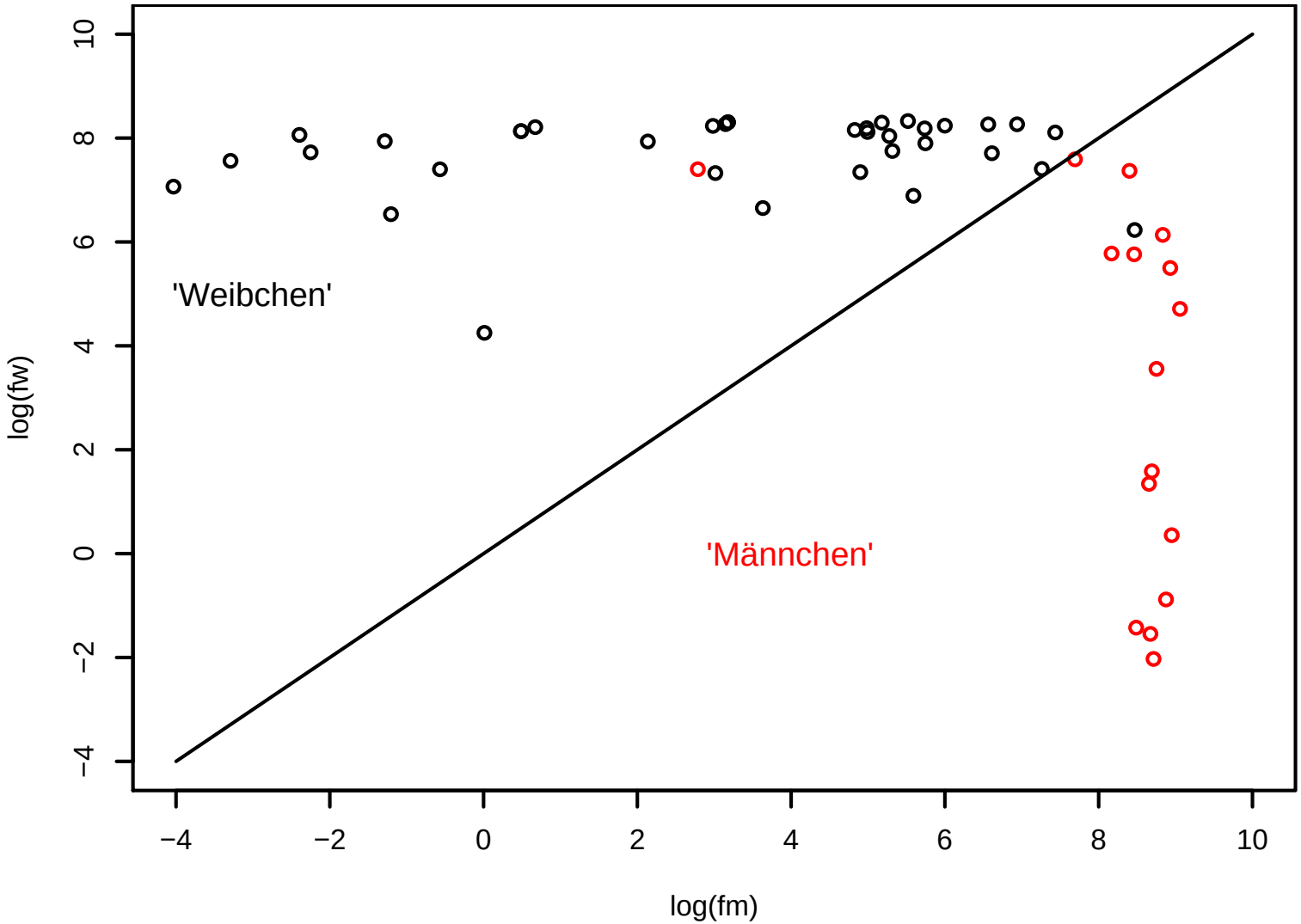
$f_m(x)$ größer $\longrightarrow$ „Männchen“

$f_w(x)$ größer $\longrightarrow$ „Weibchen“

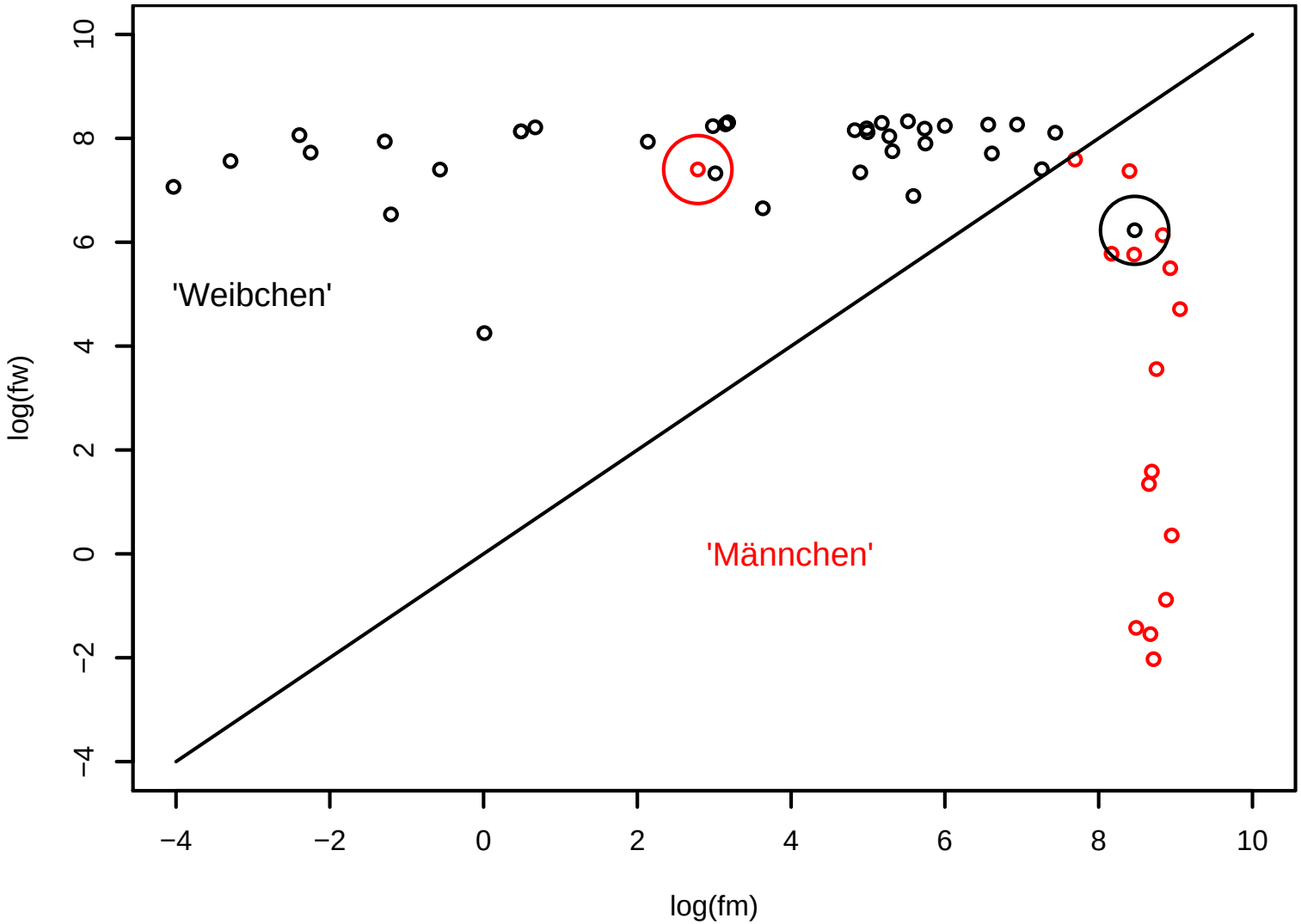
$\log(f_w)$ gegen $\log(f_m)$ und Diagonale:



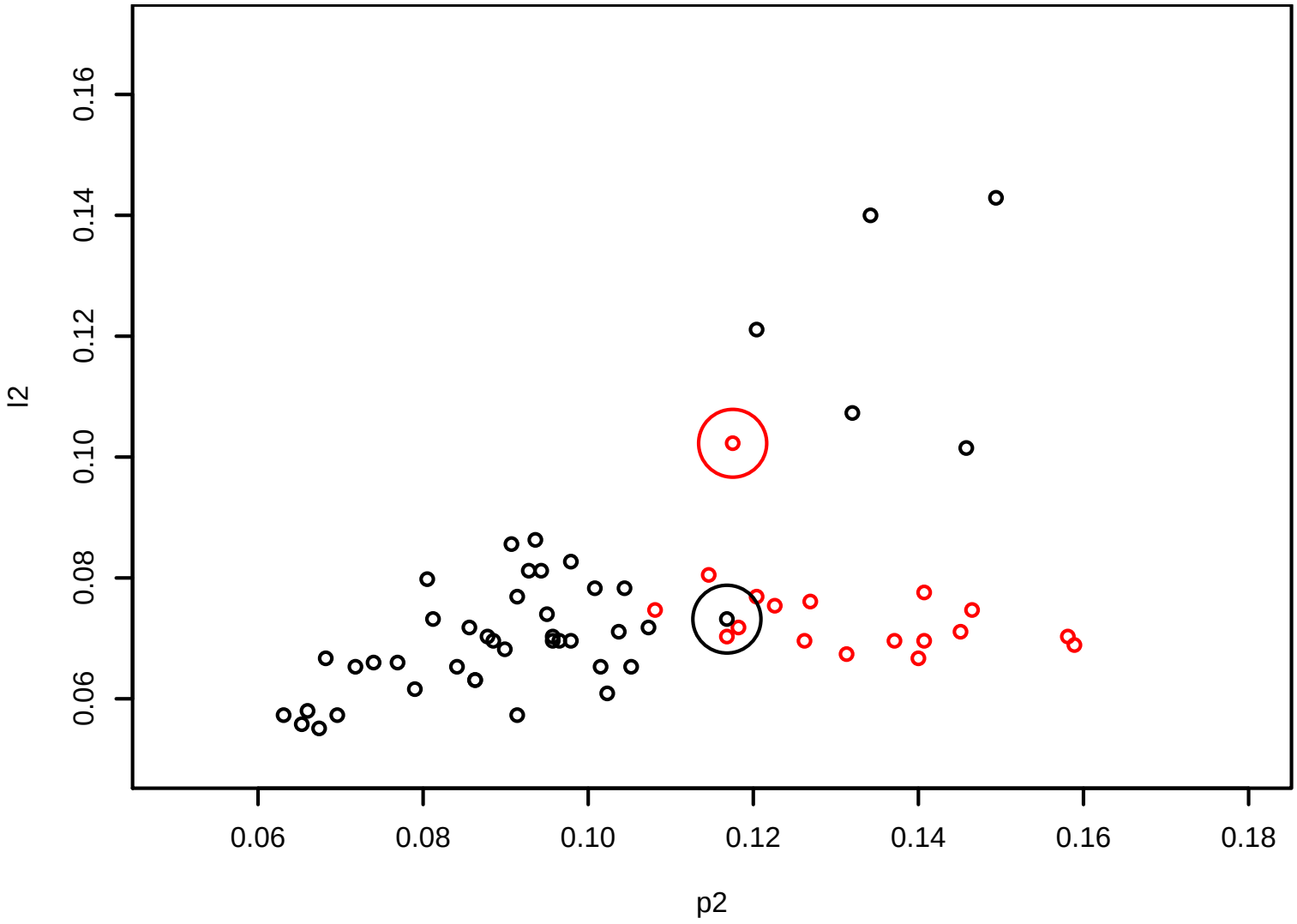
$\log(f_w)$ gegen $\log(f_m)$ und Diagonale, Ausschnittvergrößerung:



Falsch klassifiziert: 1 Männchen, 1 Weibchen (und eigentlich 2 „unentschieden“)



Welche Fälle wurden falsch zugeordnet?

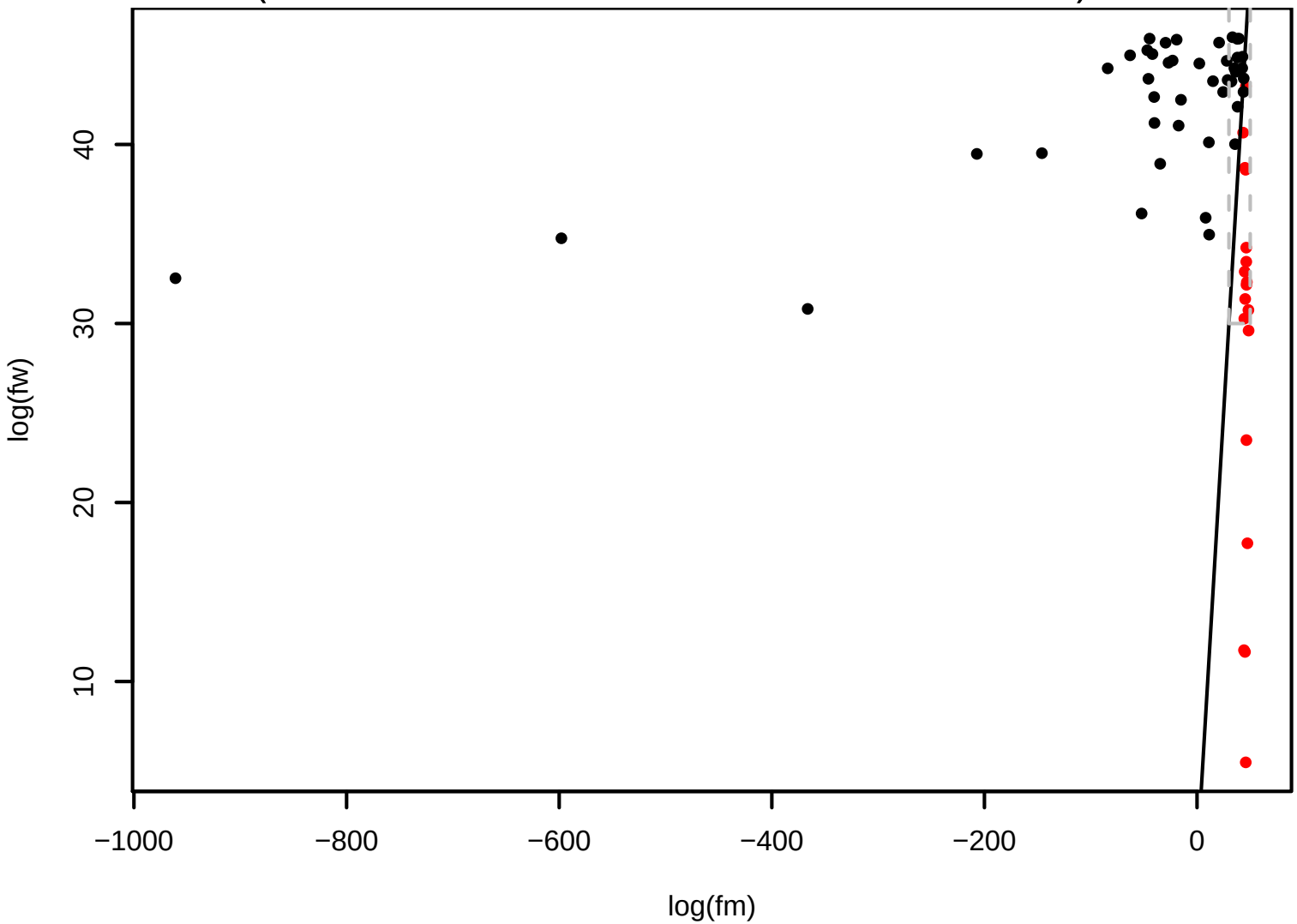


Wenn man nur p2 und l2 kennt, ist es sehr verständlich, dass diese Fälle falsch klassifiziert werden.

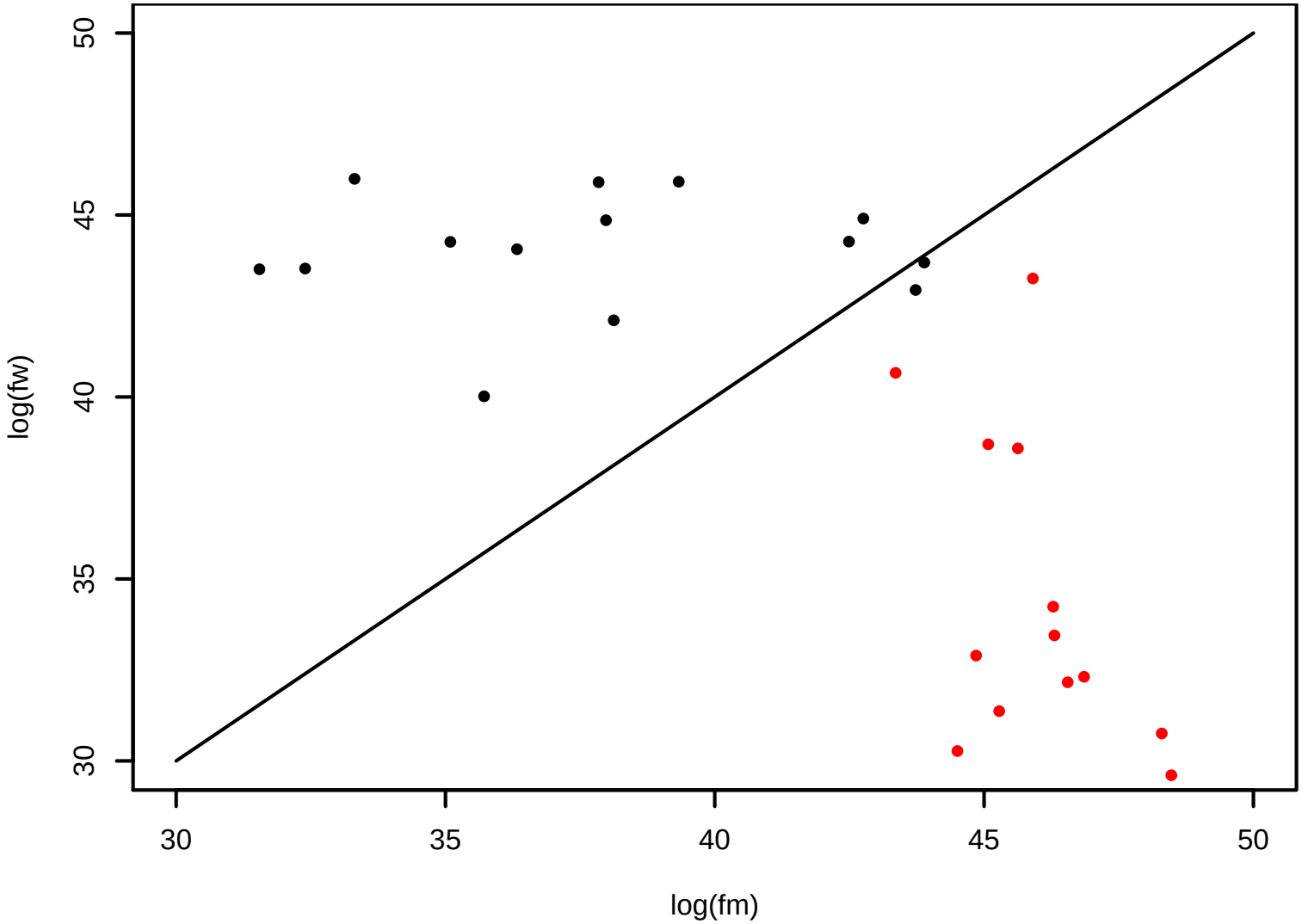
Wir verfahren genause mit allen Variablen (p1, p2, p3, p4, p5, l1, l2, l3, l4, l5) gemeinsam — mathematisch analog, allerdings geometrisch sehr schwierig darzustellen.

Ergebnis:

$\log(f_w)$ gegen $\log(f_m)$ und Diagonale (basierend auf allen 10 Variablen):

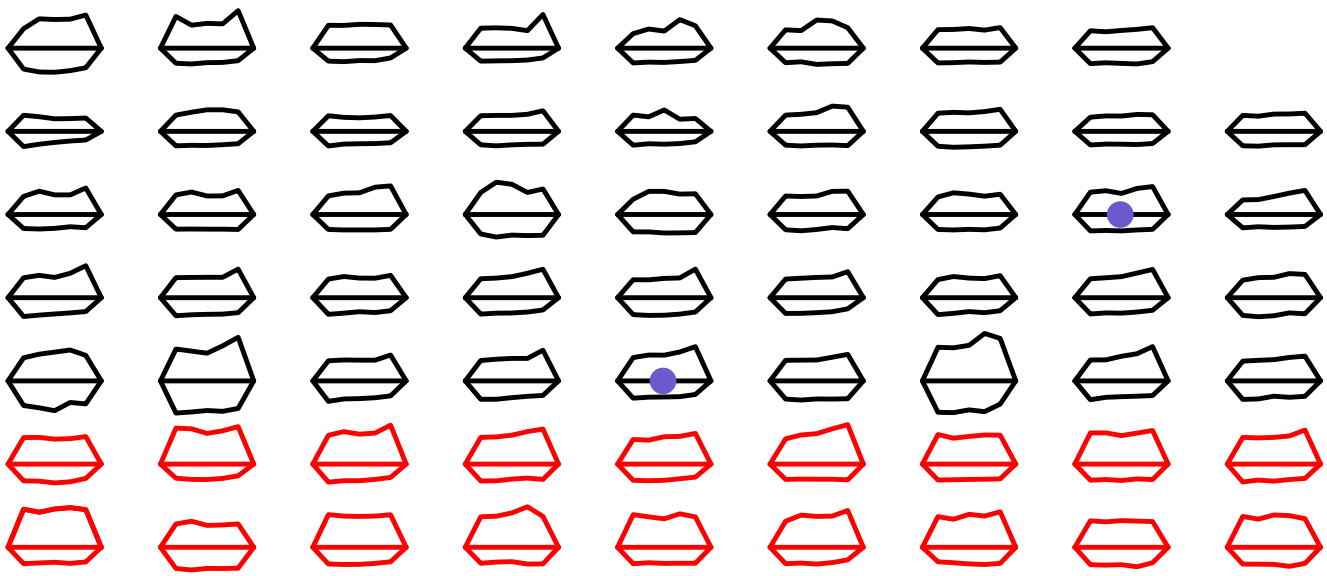


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale (basierend auf allen 10 Variablen, Ausschnittvergrößerung):



Die zwei mit (p2,l2) falsch klassifizierten Fälle wurden nun richtig klassifiziert

Falsch klassifiziert



Die beiden falsch klassifierten Rufe: sie sehen ziemlich „männlich“ aus.