Biostatistik, WS 2015/2016

Exkurs: Ein Beispiel zur Diskriminanzanalyse

Matthias Birkner

http://www.staff.uni-mainz.de/birkner/Biostatistik1516/

29.1.2016

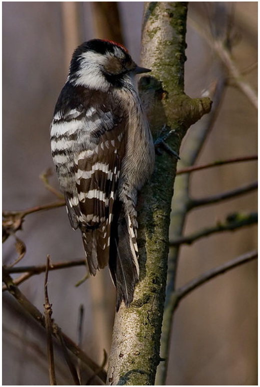


photo (c) Thermos

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

				p1		p2		p3		p4		p5		
…	ki	—	ki	—	ki	—	ki	—	ki	—	ki	—	ki	
					l1		l2		l3		l4		l5	

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

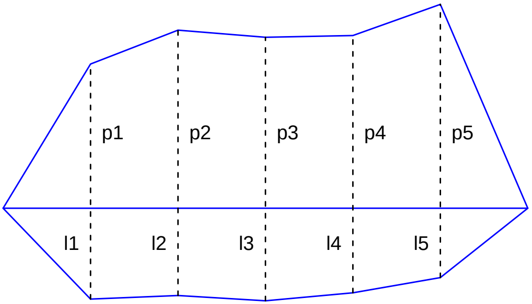
...

Gesucht:

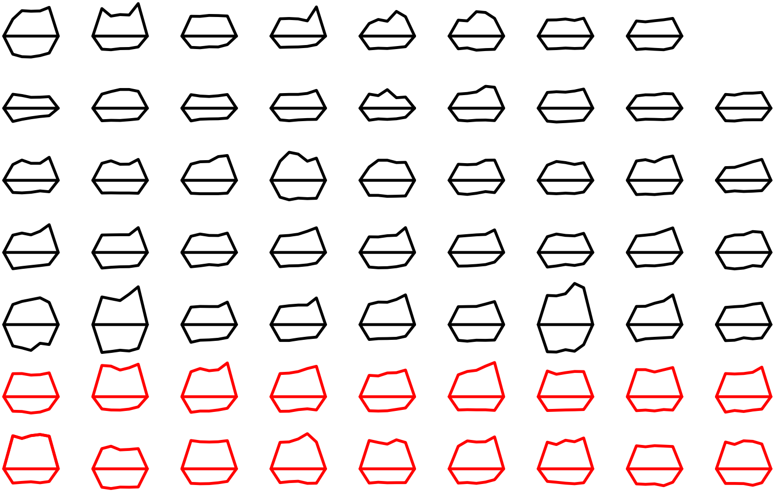
eine dem menschlichen Gehirn gerechte Darstellung des Vektors

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

Vorschlag:



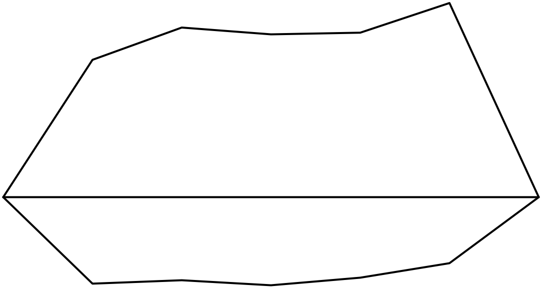
Alle 62 Rufe: rot=Männchen, schwarz=Weibchen



Mit dem Auge kann man Unterschiede erkennen:

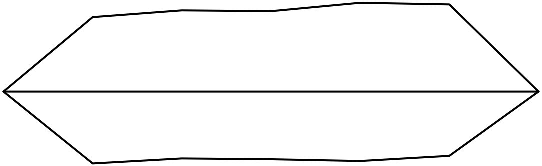
Männchen oder Weibchen?

(Männchen)



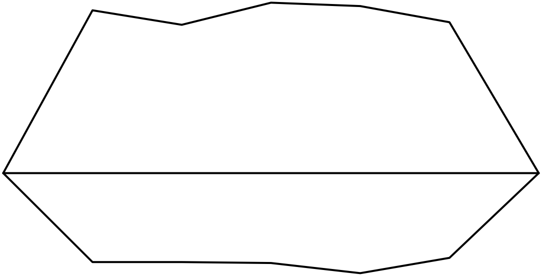
Männchen oder Weibchen?

(Weibchen)



Männchen oder Weibchen?

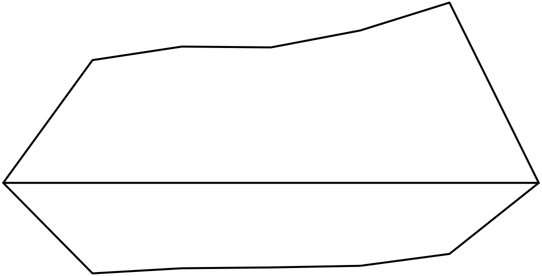
(Männchen)



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.

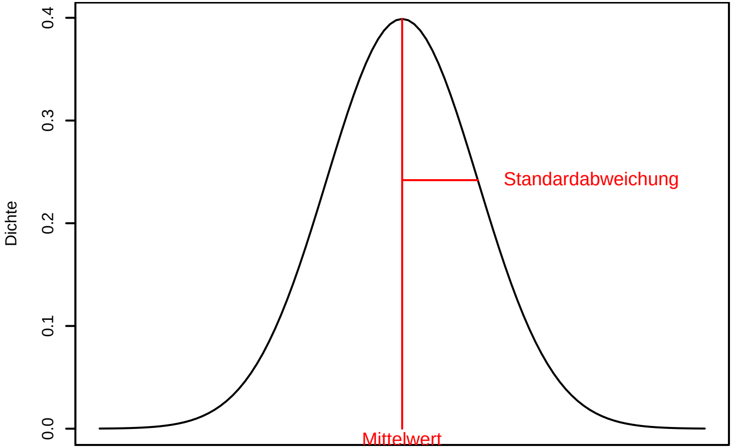
Verfahren

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

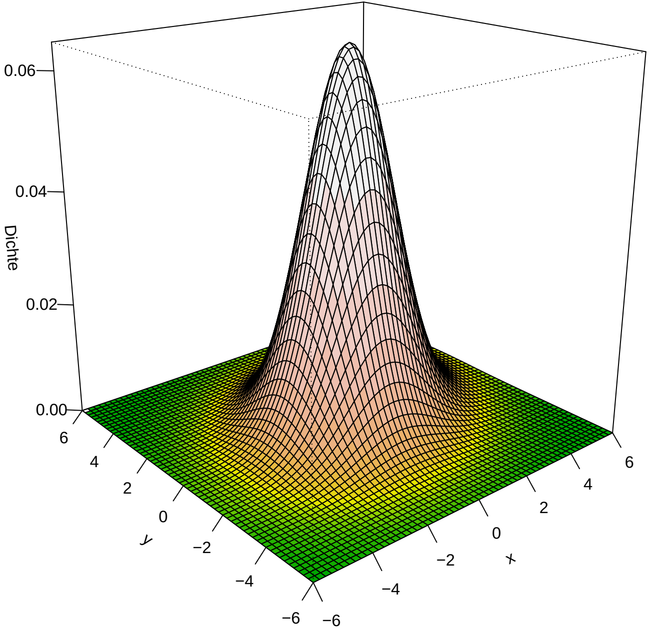
Wir benutzen für f_m und f_w mehrdimensionale Normalverteilungen.

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

Erinnerung: Eindimensionale Normalverteilung

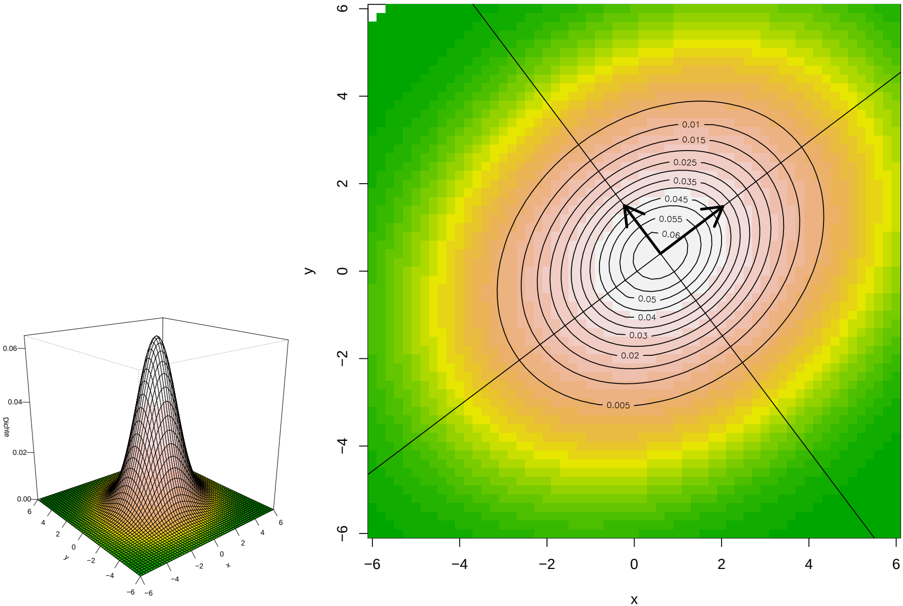


Dichte einer zweidimensionalen Normalverteilung



Zur Beschreibung einer mehrdimensionalen Normalverteilung benötigt man

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

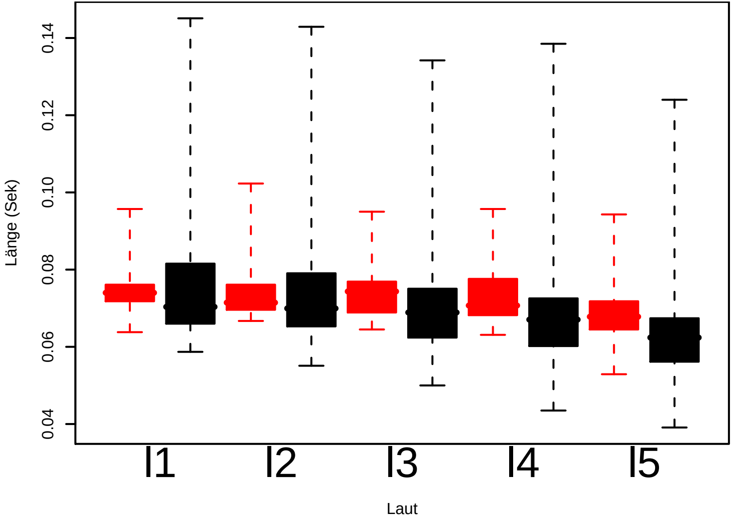

Dichte
x
y
0.01
0.015
0.025
0.035
0.045
0.055
0.06
0.05
0.04
0.03
0.02
0.005

In unserem Problem gibt es 10 Dimensionen.

Wir beginnen eindimensional.

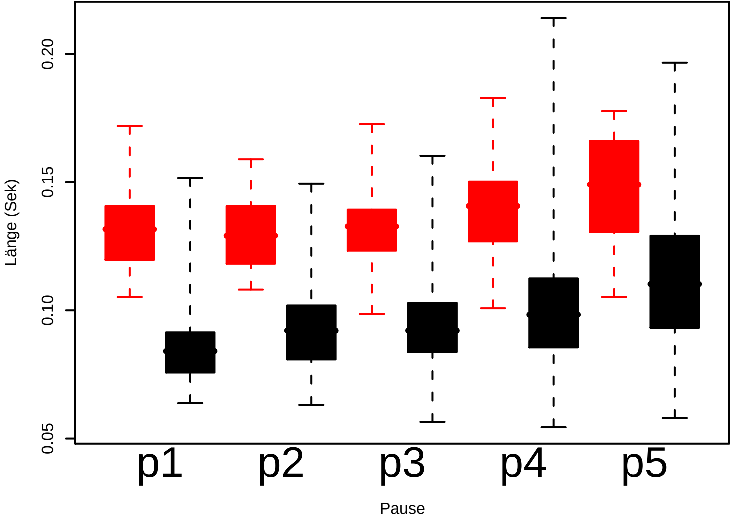
Frage: Welche eine der 10 Variablen sollen wir wählen?

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



Keine gute Trennung der Geschlechter

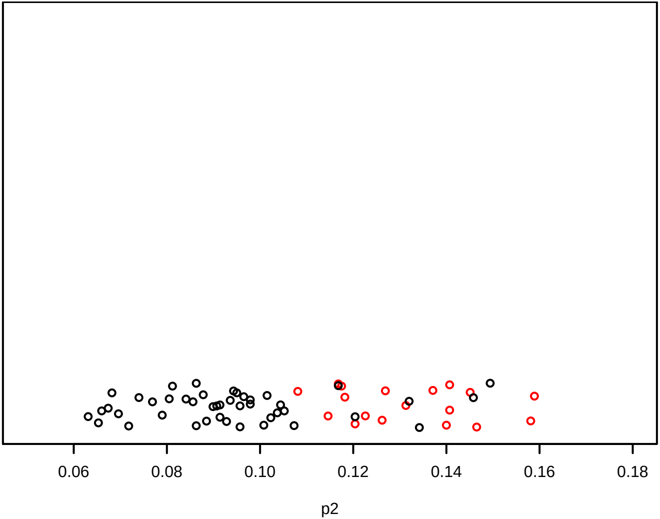
Länge der Pausen bei Männchen (rot) und Weibchen (schwarz)

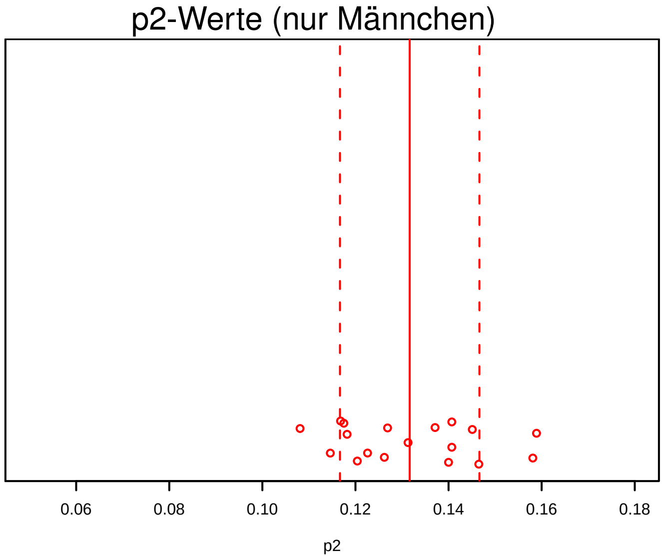


Bei den Männchen sind die Pausen typischerweise länger

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

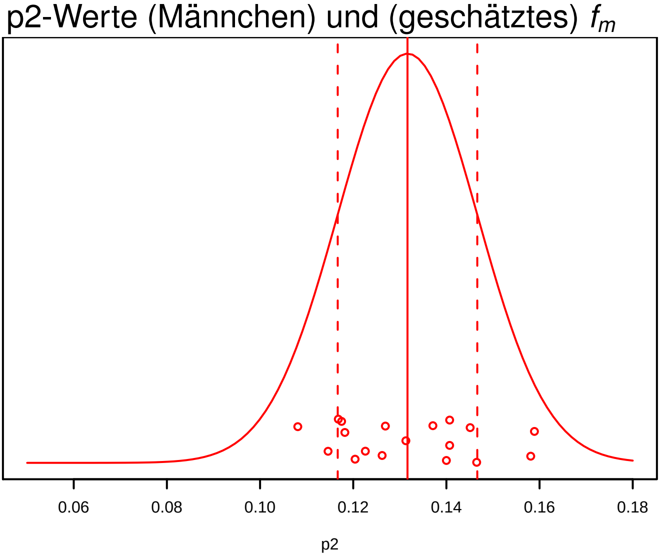
Die p2-Werte (mit Jitter)

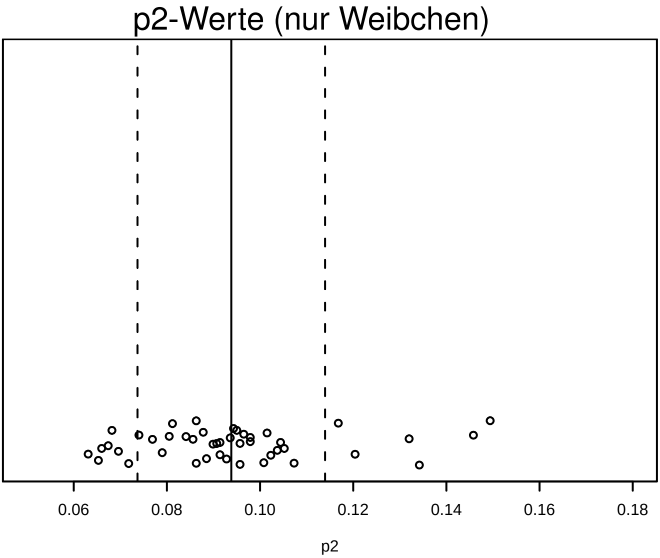




Mittelwert $\mu_m = 0{,}1316$, Standardabweichung $\sigma_m = 0{,}0150$

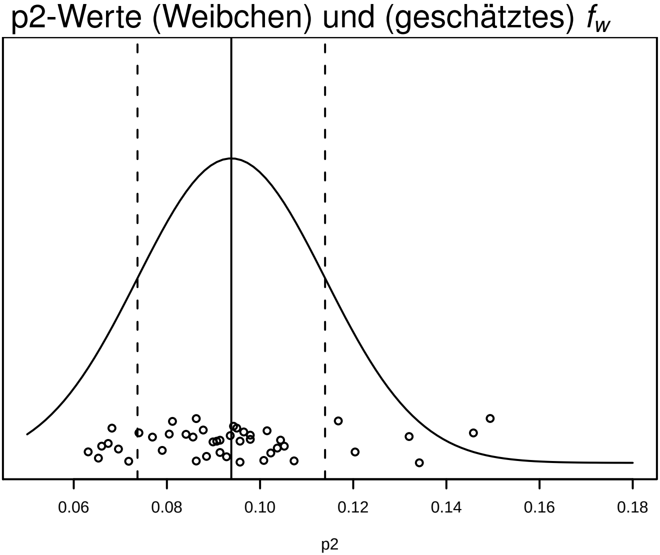
Wir approximieren f_m durch die Normalverteilung mit Mittelwert μ_m und Standardabweichung σ_m





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

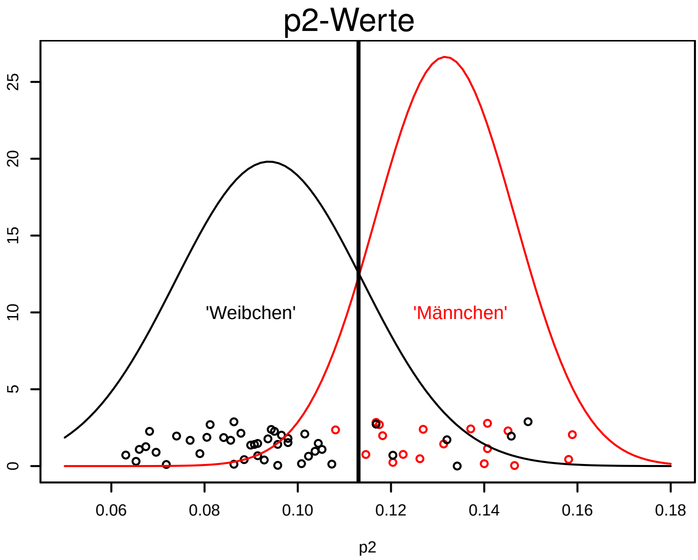
Wir approximieren f_w durch die Normalverteilung mit Mittelwert μ_w und Standardabweichung σ_w



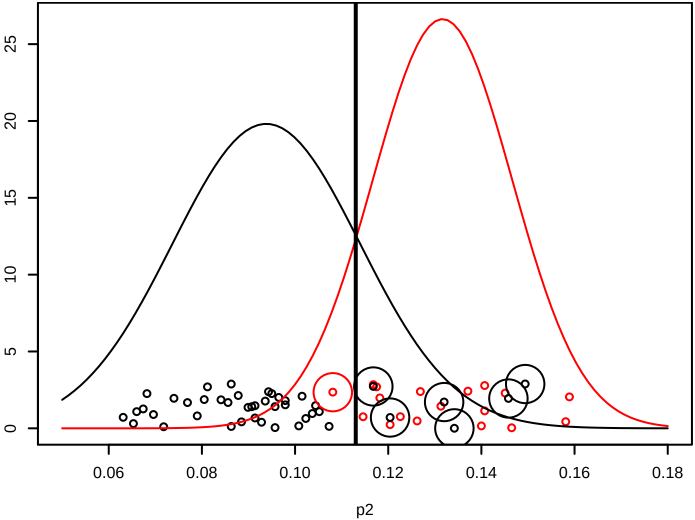
Klassifikationsregel:

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

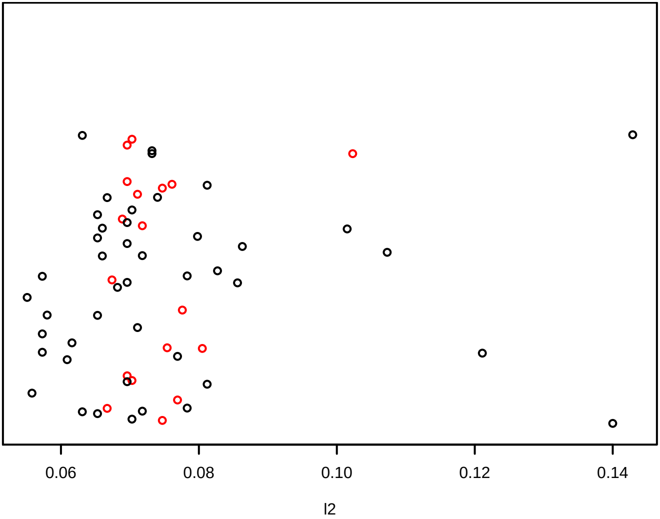
Falsch klassifiziert:
1 Männchen 6 Weibchen



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

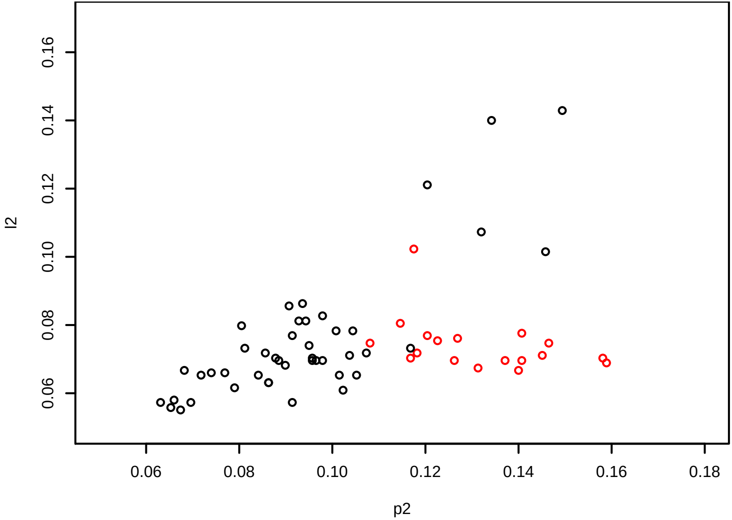
Wir betrachten:

Erste Variable = p2
Zweite Variable = l2



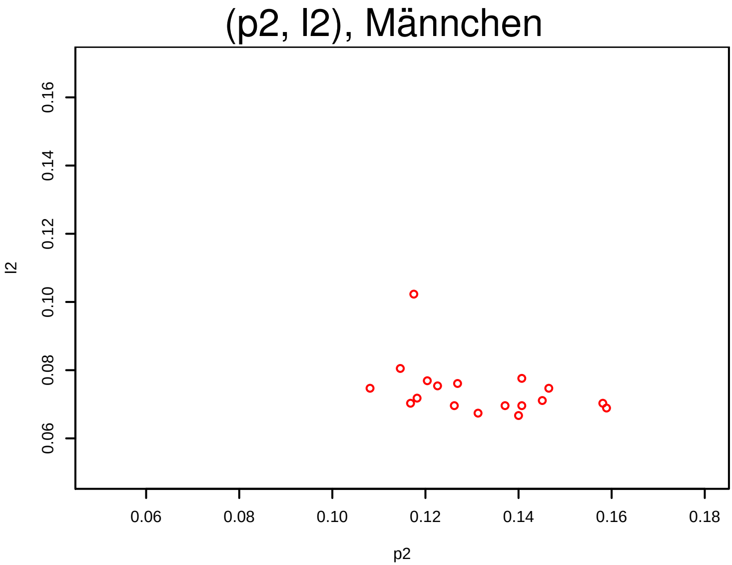
Beobachtung: l2 allein trennt die Geschlechter sehr schlecht.

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.

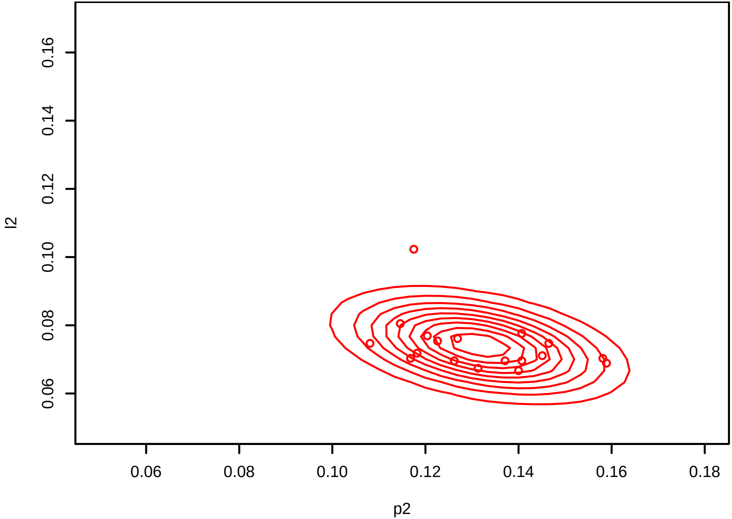
Wir approximieren
die Verteilungen
von (p2, l2) bei Männchen und bei Weibchen
durch
zweidimensionale
Normalverteilungen.


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

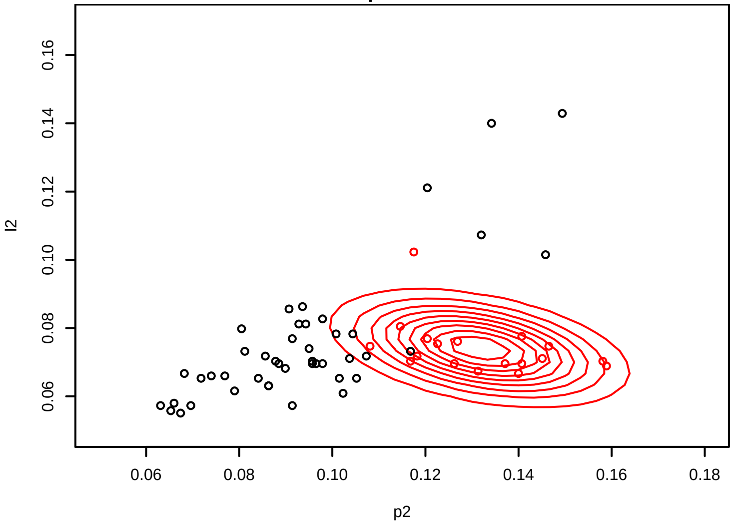
Wie im eindimensionalen Fall schätzen wir
den (zweidimensionalen) Mittelwert
und die (zweidimensionale) Varianz
(d.h. die sog. Kovarianzmatrix)

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

(p2, l2) für Männchen und f_m

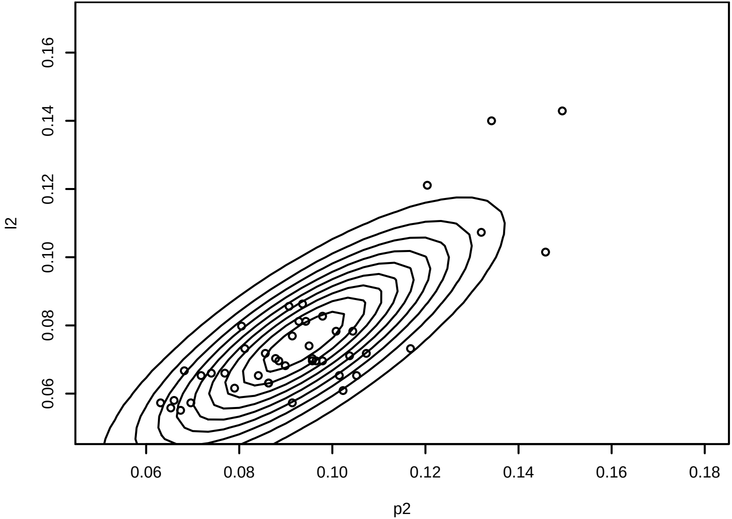


Viele der Weibchen passen schlecht zu f_m:

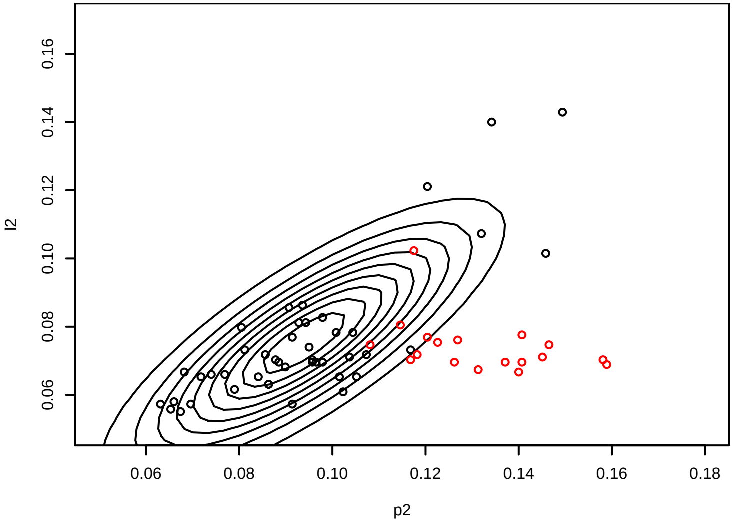


Analog für die Weibchen:

(p2, l2) für Weibchen und f_w



Viele der Männchen passen schlecht zu f_w:



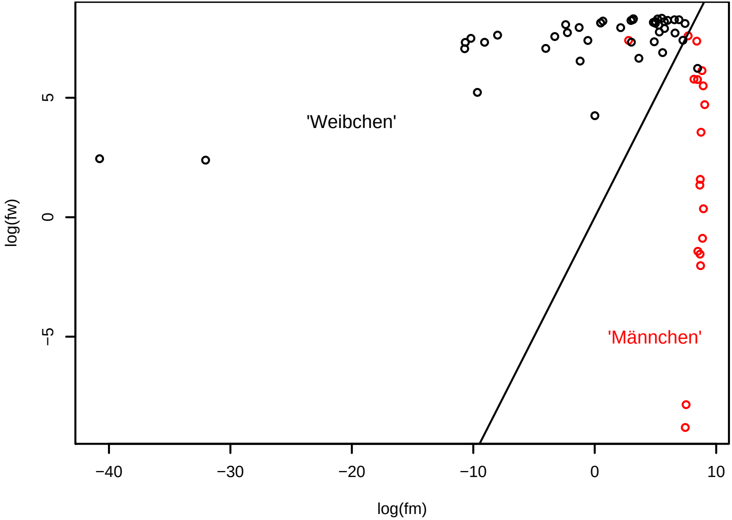
Klassifikation:

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

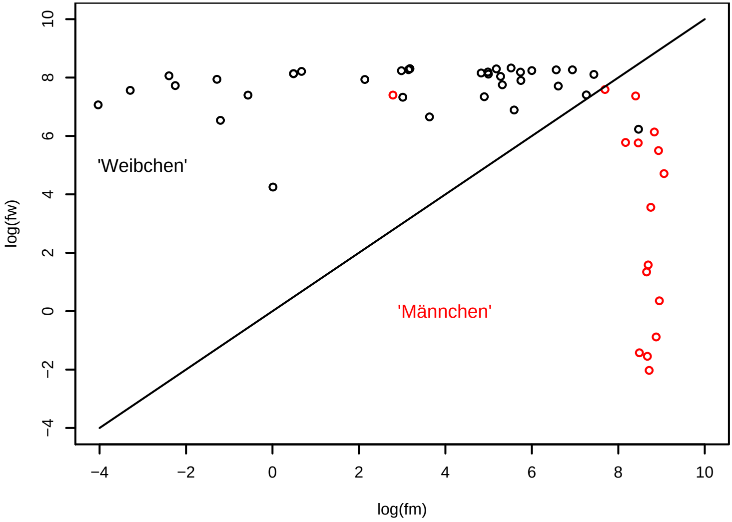
$f_m(x)$ größer $\longrightarrow$ „Männchen“

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

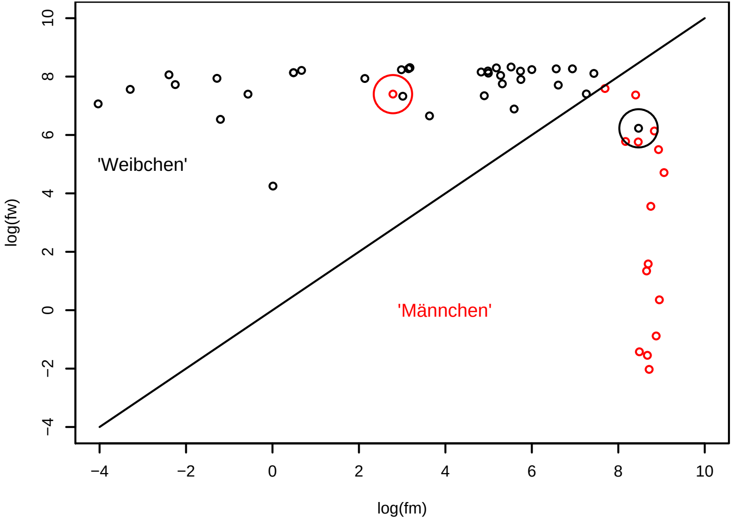
$\log(f_w)$ gegen $\log(f_m)$ und Diagonale:



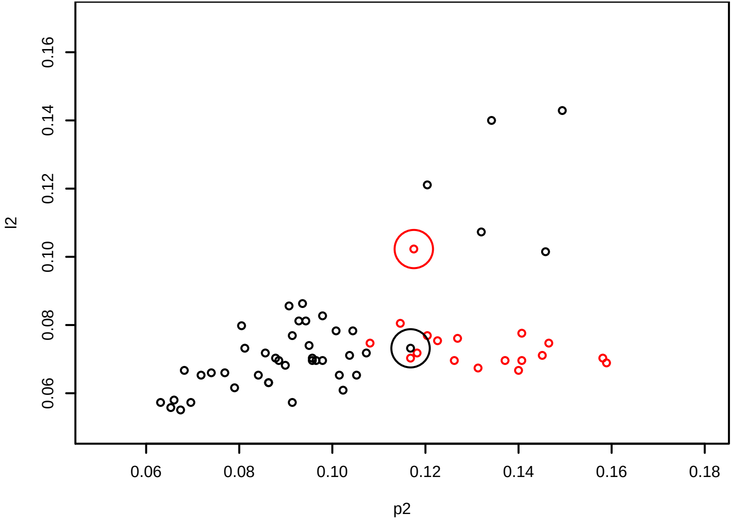
$\log(f_w)$ gegen $\log(f_m)$ und Diagonale, Ausschnittvergrößerung:



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



Welche Fälle wurden falsch zugeordnet?

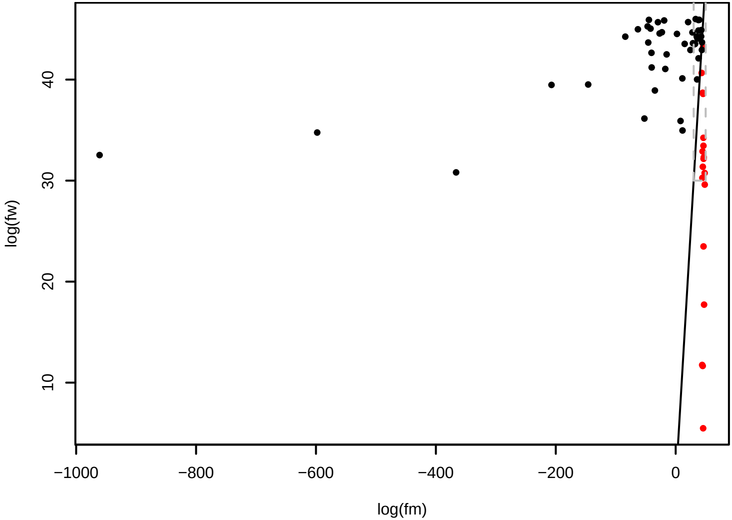


Wenn man nur p2 und l2 kennt, ist es sehr verständlich, dass diese Fälle falsch klassifiziert werden.

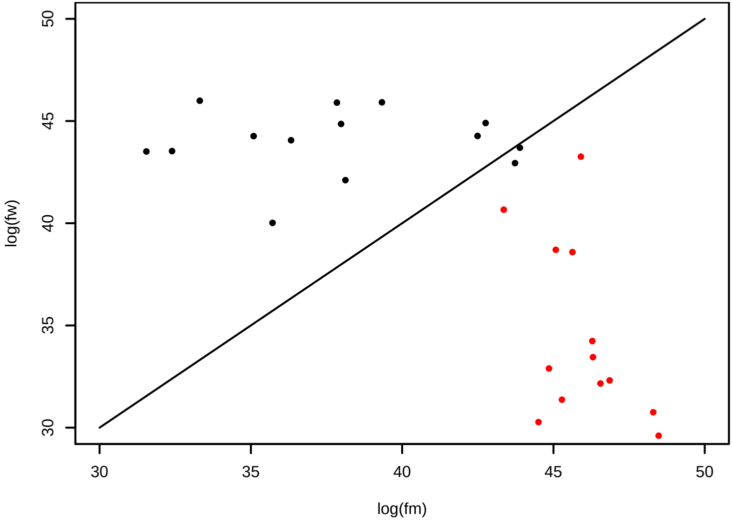
Wir verfahren genauso mit allen Variablen (p1, p2, p3, p4, p5, l1, l2, l3, l4, l5) gemeinsam — mathematisch analog, allerdings geometrisch sehr schwierig darzustellen.

Ergebnis:

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

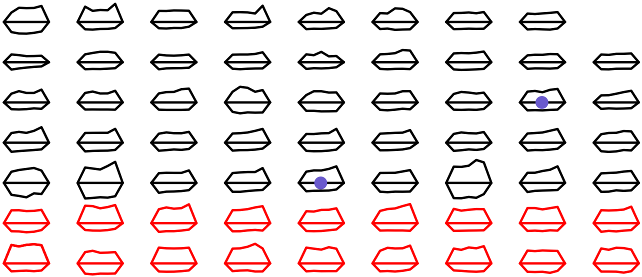


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale (basierend auf allen 10 Variablen, Ausschnittvergrößerung):



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

Falsch klassifiziert



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