REDLINE DATA

Fallstricke der Statistik!

Jens Medenwaldt
Redline DATA GmbH

Was wir uns anschauen wollen

- Wer sind die eigentlich?
  - Der Trick mit der Stichprobe!
- Was ist die Mitte von was?
  - Die Crux mit dem Mittelwert
- Was ist eigentlich wieviel von was?
  - Jonglieren mit Prozenten!
- Und wer fehlt hier?
  - Missingwerte und ihr Tücken
- Was hat sich wie verändert?
  - Das Kreuz mit den Kreuztabellen

Der Trick mit der Stichprobe

Redline Data

statistisch gesehen ist die Tierart
mit dem geringsten Risiko
von einem Auto angefahren zu werden…

…die Fische!

Stichprobe

- „Mehr als die Hälfte (63,5%) der HSV-Fans hat einen Hochschulabschluss“
- Quelle: SHZ (Schleswig-Holstein Zeitung)
- Stichprobe: 45.000 Fußballfans
- und zwar in Profilen von XING-Mitgliedern

Teilkategorie Produktqualität Kundenbefragung Antifalten-Creme 2014			
Unternehmen	**Punkte***	**Rang**	**Qualitätsurteil**
Sebamed	85,0	1	sehr gut
Garnier	83,5	2	
Helena Rubinstein	81,2	3	
Christian Dior	81,2	4	
Estée Lauder	80,9	5	
Dr. Hauschka	80,4	6	
L'Oréal	80,3	7	
Olaz	79,9	8	gut
Nivea	79,5	9	
Biotherm	79,3	10	
Chanel	79,1	11	
Weleda	78,1	12	
Vichy	77,6	13	
Avène	77,0	14	
Shiseido	76,8	15	
Aok	76,5	16	
Balea	76,1	17	
Louis Widmer	75,3	18	
Florena	75,3	19	
Eucerin	74,8	20	
Clinique	74,7	21	
Lancôme	72,4	22	
Lancaster	71,3	23	
Diadermine	70,4	24	
Alterra	68,7	25	befriedigend
Rival de Loop	67,3	26	
Sonstige**	78,5		

„Berücksichtigt wurden alle Marken, zu denen sich jeweils mindestens 80 Kunden geäußert haben“

Stichproben

- Stichprobe genau beschreiben
  - Auswahl und Entstehung
  - z.B. alle entlassenen Klienten im Jahr
- Stichprobe nicht wechseln
  - und wenn, dann wieder beschreiben warum gewechselt wurde
- Stichprobe mit Bedacht auswählen
  - Begonnene Behandlungen und Auswertung auf Entlassungsform?

Sinnvolle Stichproben

- Behandlungsverläufe = alle Entlassungen eines Jahres
- Soziodemografische Daten = alle Zugänge eines Jahres
- Katamnesen = immer Entlassungsjahrgänge (auch für Soziodemografische Daten)

Teilstichproben

- weiblich vs. männlich
- alt vs. jung
- Migrationshintergrund vs. kein Migrationshintergrund
- „Lange“ Suchtkarriere vs. kurze Suchtkarriere

Die Crux mit dem Mittelwert

Redline Data

Wer auf der Jagd ist
und mit dem ersten Schuss
50cm links daneben schießt
und mit dem zweiten Schuss
50 cm rechts daneben schießt
der hat im Mittelwert genau getroffen…

…aber trotzdem nichts zu essen!

Mittelwerte

- A: Meine Klienten sind im Durchschnitt 40 Jahre alt
- B: Auch meine Klienten haben ein Durchschnittsalter von 40 Jahren
- A: Prima, dann haben wir ja ein ganz ähnliches Klientel
- C: Komisch, meine Klienten sind eigentlich alle Anfang 20, aber mein Mittelwert sagt 40 Jahre?

Mittelwert

- A:
- Mittelwert 40,0

- B:
- Mittelwert 40,0

- C:
- Mittelwert 40,0

Mittelwert

- A hat 6 Klienten
- 18, 20, 22, 58, 60 und 62 Jahre

- B hat ebenfalls 6 Klienten
- 38, 39, 40, 40, 41, 42 Jahre

- C hat ebenfalls 6 Klienten
- 20, 23, 24, 25, 25 Jahre

Mittelwert

- A:
- Mittelwert 40,0
- Standardabweichung 21,98

- B:
- Mittelwert 40,0
- Standardabweichung 1,41

- C:
- Mittelwert 40,0
- Standardabweichung 39,21 (???)

Mittelwert

- Schauen wir die eingegebenen Werte zu C noch einmal genau an:
- 120, 23, 24, 25, 25 Jahre

Mittelwerte

		A	B	C
N	Gültig	6	6	6
Mittelwert		40,0	40,0	40,0
Median		40,0	40,0	24,5
Modus		18*	40	23*
Standardabweichung		21,98	1,41	39,20

*Mehrere Modi, der niedrigste Wert wurde verwendet

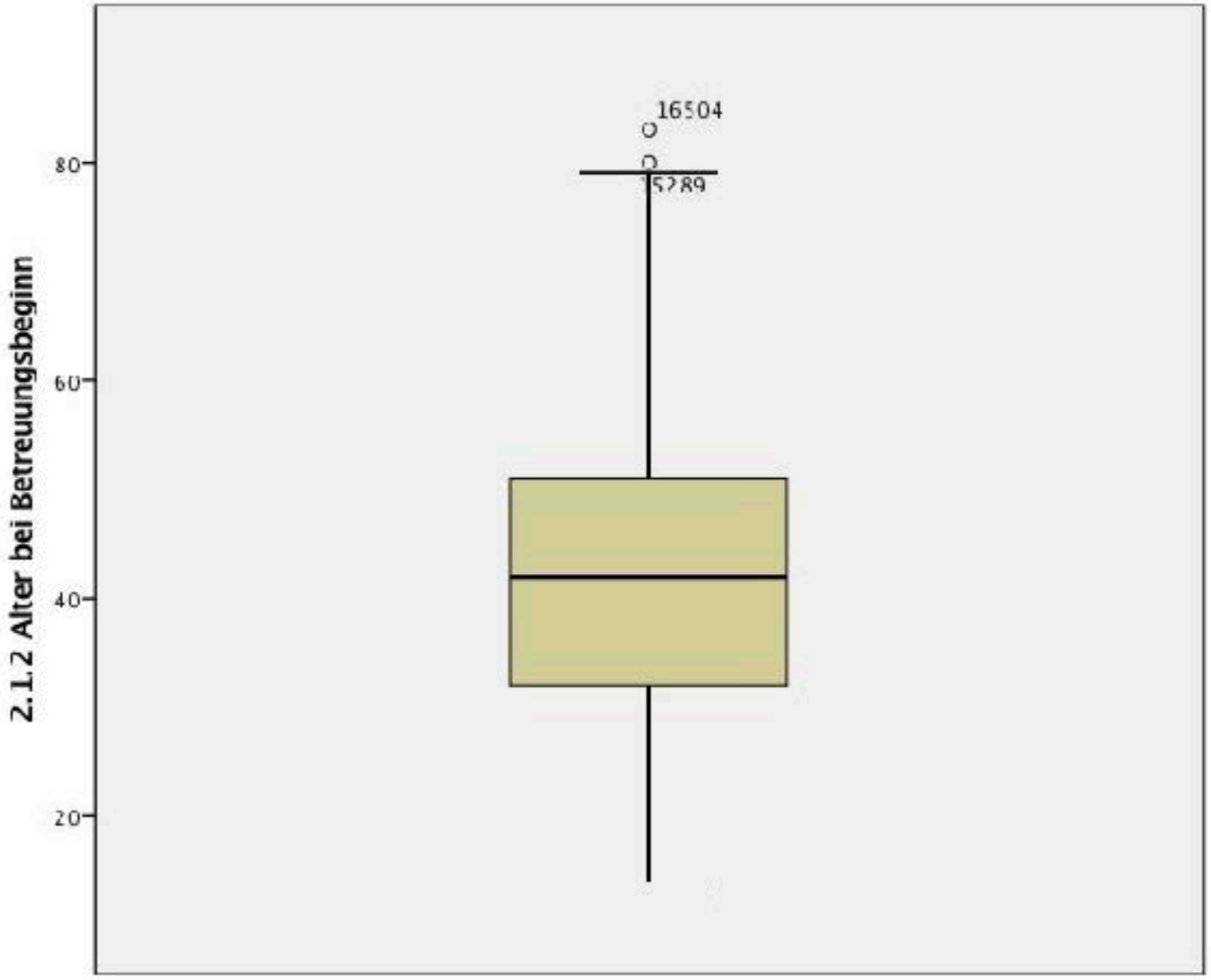
Der Median ist deutlich robuster gegen „Ausreißer“

Statistiken
2.1.2 Alter bei Betreuungsbeginn

N Gültig 19097
Fehlend 0

Mittelwert 41,50
Median 42,00
Modus 48
Std.abweichung 12,176
Minimum 14
Maximum 83

Mittelwerte



Was ist zu beachten

- Mittelwerte nie allein angeben
- mindestens mit Standardabweichung
- besser noch Median, Min und Max
- Für Mittelwertvergleiche gibt es spezielle statische Verfahren
- Bei der Berechnung auf „künstliche“ Missings achten (99= unbekannt)

Jonglieren mit Prozenten

Bezugsgrösse!

- Der Anteil der weiblichen Angestellten konnte im vergangenen Jahr um 100% gesteigert werden.

- Frau Weber hat eine Kollegin bekommen

Prozente addieren

- Eine dänische Untersuchung besagt:

- 25 % der Jungen fühlen sich in der KiTa unwohl
- 10% der Mädchen fühlen sich in der KiTa unwohl

- Damit fühlen sich mehr als 1/3 der Kinder (35%) in der KiTa unwohl !!!

Prozente addieren

- In der KiTa sind 80 Jungen und 100 Mädchen
- 20 Jungen (25%) fühlen sich unwohl
- 10 Mädchen (10%) fühlen sich unwohl
- damit fühlen sich 30 Kinder unwohl
- 30 von 180 sind 16,7%

Mit Prozenten Politik machen

- Nach der Armutsstudie des Paritätischen liegt die Armutsquote bei 15,5%
- definiert als die Menschen, die pro Monat weniger als 60 % des Durchschnittseinkommens zur Verfügung haben.
- Dieser Prozentsatz bleibt gleich, auch wenn sich das reale Einkommen aller Bundesbürger verdoppeln würde.

Relative Steigerung

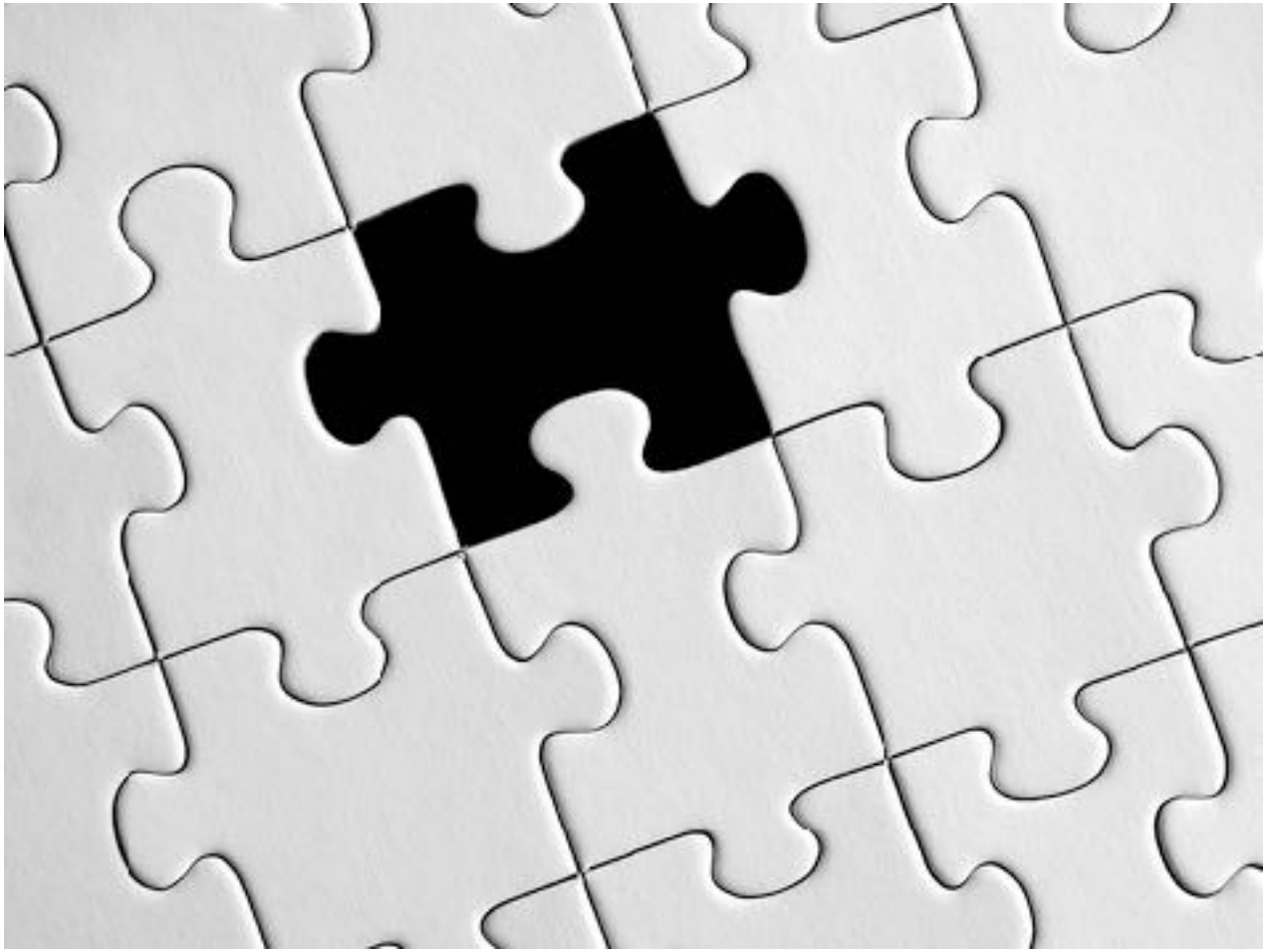
- Wien im Januar 2015
- die Durchnittstemperatur betrug 3,5° Celsius
- das langjährige Mittel beträgt 0,8° Celsius
- eine Steigerung um 2,7° Celsius = 337%
  (2,7/0,8*100)
- Dramatisch !!!

- Rechnen wir das Ganze mal um in ° Fahrenheit
- 3,5° C = 38,5° F - 0,8° C = 33,4° F
- Steigerung um 5,1° F = 15,3% (5,1/33,4*100

Was ist zu beachten

- Bei Prozentwerten immer die Basis angeben
- Prozente addieren oder Subtrahieren nur bei gleicher Basis

Was fehlt denn da?



Missingwerte

REDLINE DATA

2.3.11 Problematische Schulden

	Häufigkeit	Prozent	Häufigkeit	Gültige Prozente
Keine Angabe	1262	15,3		
Keine Schulden	4638	56,1	4638	66,1
Bis 10.000 Euro	1232	14,9	1232	17,6
Bis 25.000 Euro	541	6,5	541	7,7
Bis 50.000 Euro	334	4,0	334	4,8
Über 50.000 Euro	267	3,2	267	3,8
Gesamt	8274	100,0	7012	100,0

Missingwerte

- 56,1% haben keine Schulden
- 28,6% haben Schulden
- 15,3% machen keine Angabe

- Sind alle ohne Angabe schuldenfrei, so wären 71,4% ohne Schulden
- Haben alle ohne Angabe Schulden, so wären 43,9% mit Schulden

Schulden vs. Schuldenfrei

- Schuldenfrei sind
- 56,1% oder 66,1% oder 71,4%

- Mit Schulden sind
- 28,4% oder 33,9% oder 43,9%

Was ist zu beachten

- Missingwerte möglichst vermeiden
- Ggf. Missingwerte herausrechnen (aber angeben)

Strecken in Irland

	km	Zeit	km	Zeit	km	Zeit	km	Zeit	km	Zeit
	Dublin		Cork		Larne		Rosslare		Belfast	
Athlone	127	02:15	219	04:15	261	05:15	201	04:00	220	04:30
Bangor	187	03:30	438	08:00	60	01:15	351	07:15	25	00:45
Belfast	166	03:15	417	08:00	34	01:00	330	07:15		
Cavan	114	01:45	302	06:00	166	03:00	267	05:30	138	02:45
Cork	254	05:00			451	08:15	208	04:15	417	08:00
Derry	232	04:45	428	08:00	114	01:45	397	07:45	131	02:15
Donegal	233	04:45	402	08:00	228	04:30	391	07:45	174	03:00
Dublin			254	05:00	204	04:15	153	03:00	166	02:45
Ennis	230	04:45	137	02:30	372	07:30	245	05:00	325	07:15
Enniskillen	184	03:30	360	07:30	172	03:00	312	07:00	136	02:30
Galway	212	04:00	209	04:00	330	07:15	274	06:00	300	06:00
Killarney	304	06:00	87	01:30	470	08:45	275	06:00	425	08:15
Larne	204	04:00	451	08:15			364	07:30	34	00:45
Letterkenny	237	04:45	449	08:15	146	03:00	391	08:00	155	02:45
Limerick	193	03:30	105	01:45	357	07:30	211	04:15	320	07:00
Mullingar	87	01:30	245	05:00	211	04:15	198	04:00	175	03:00
Navan	48	01:00	275	05:30	171	03:15	203	04:00	134	02:30
Portlaoise	85	01:30	175	03:00	285	06:00	135	02:15	244	04:45
Roscommon	156	02:45	251	05:00	249	05:30	241	05:00	222	04:30
Rosslare	153	02:45	208	04:00	364	07:30			330	07:15
Sligo	214	04:00	336	07:15	240	05:30	327	07:15	200	04:00
Tralee	298	06:00	119	01:45	459	08:45	291	06:00	420	08:15
Waterford	163	03:15	126	01:45	367	07:30	79	01:30	325	07:15

Kreuztabellen

REDLINE DATA

Behandlungsbeginn	Erwerbssituation Katamnesezeitpunkt				
	Keine Angabe	Erwerbstätig	Arbeitslos	Nicht erwerbstätig	Gesamt
Keine Angabe	0	32	23	18	73
Erwerbstätig	30	1064	[illegible]	196	1480
Arbeitslos	13	367	542	163	1085
Nicht erwerbstätig	1	74	[illegible]	382	496
Gesamt	44	1537	794	759	3134

Sind es 50,0% oder 68,3% oder 17,3%

- 50,0% der zu Beginn Arbeitslosen sind auch zur Katamnese arbeitslos (Basis 1085)
- 68,3% der zur Katamnese Arbeitslosen waren auch zu Beginn arbeitslos (Basis 794)
- 17,3% aller Klienten waren zu Beginn und zur Katamense arbeitslos (Basis 3134)

Behandlungsbeginn	Erwerbssituation Katamnesezeitpunkt			
	Erwerbstätig	Arbeitslos	Nicht erwerbstätig	Gesamt
Erwerbstätig	1064		196	1450
Arbeitslos	367	542	163	1072
Nicht erwerbstätig	74		382	495
Gesamt	1505	771	741	3017

Sind es 50,6% oder 70,3% oder 18%

- 50,6% der zu Beginn definitiv Arbeitslosen sind auch zur Katamnese arbeitslos (Basis 1072)
- 70,3% der zur Katamnese definitiv Arbeitslosen waren auch zu Beginn definitiv arbeitslos (Basis 771)
- 18,0% aller Klienten mit Angaben waren zu Beginn und zur Katamense arbeitslos (Basis 3017)

Zeilenprozent

Behandlungsbeginn	Erwerbssituation Katamnese: Erwerbstätig	Arbeitslos	Nicht erwerbstätig	Gesamt
Erwerbstätig	73,4 %	13,1 %	13,5 %	100,0 %
Arbeitslos	34,2 %	50,6 %	15,2 %	100,0 %
Nicht erwerbstätig	14,9 %	7,9 %	77,2 %	100,0 %
Gesamt	49,9 %	25,6 %	24,6 %	100,0 %

Spaltenprozent

Behandlungsbeginn	Erwerbssituation Katamnese: Erwerbstätig	Arbeitslos	Nicht erwerbstätig	Gesamt
Erwerbstätig	70,7 %	24,6 %	26,5 %	48,1 %
Arbeitslos	24,4 %	70,3 %	22,0 %	35,5 %
Nicht erwerbstätig	4,9 %	5,1 %	51,6 %	16,4 %
Gesamt	100,0 %	100,0 %	100,0 %	100,0 %

Tabellenprozent

Behandlungsbeginn	Erwerbssituation Katamnese: Erwerbstätig	Arbeitslos	Nicht erwerbstätig	Gesamt
Erwerbstätig	35,3 %	6,3 %	6,5 %	48,1 %
Arbeitslos	12,2 %	18,0 %	5,4 %	35,5 %
Nicht erwerbstätig	2,5 %	1,3 %	12,7 %	16,4 %
Gesamt	49,9 %	25,6 %	24,6 %	100,0 %

Was ist zu beachten

- In Kreuztabellen zunächst mit absoluten Zahlen arbeiten
- Genau überlegen welche Prozentwerte angegeben werden
- Bei Verläufen Vorzugsweise Tabellenprozente
- Missingwerte genau beachten und ggf. herausrechnen

Und zum Schluss: VORSICHT

- Korrelation ist nicht immer gleich ein kausaler Zusammenhang

In Dörfern mit vielen Storchennestern werden mehr Kinder geboren, als in Dörfern mit wenig Storchennestern.

Also bringen die Störche Kindersegen!

Mögliche tatsächliche Erklärung:
Dörfer mit vielen Nestern haben schlichtweg mehr Häuser in denen Menschen leben