Universität Bremen

ipp bremen
Institut für Public Health und Pflegeforschung
Universität Bremen | Fachbereich 11

Abteilung Sozialepidemiologie

Statistisches Testen

Statistik in der Epidemiologie II
Wintersemester 2021/22

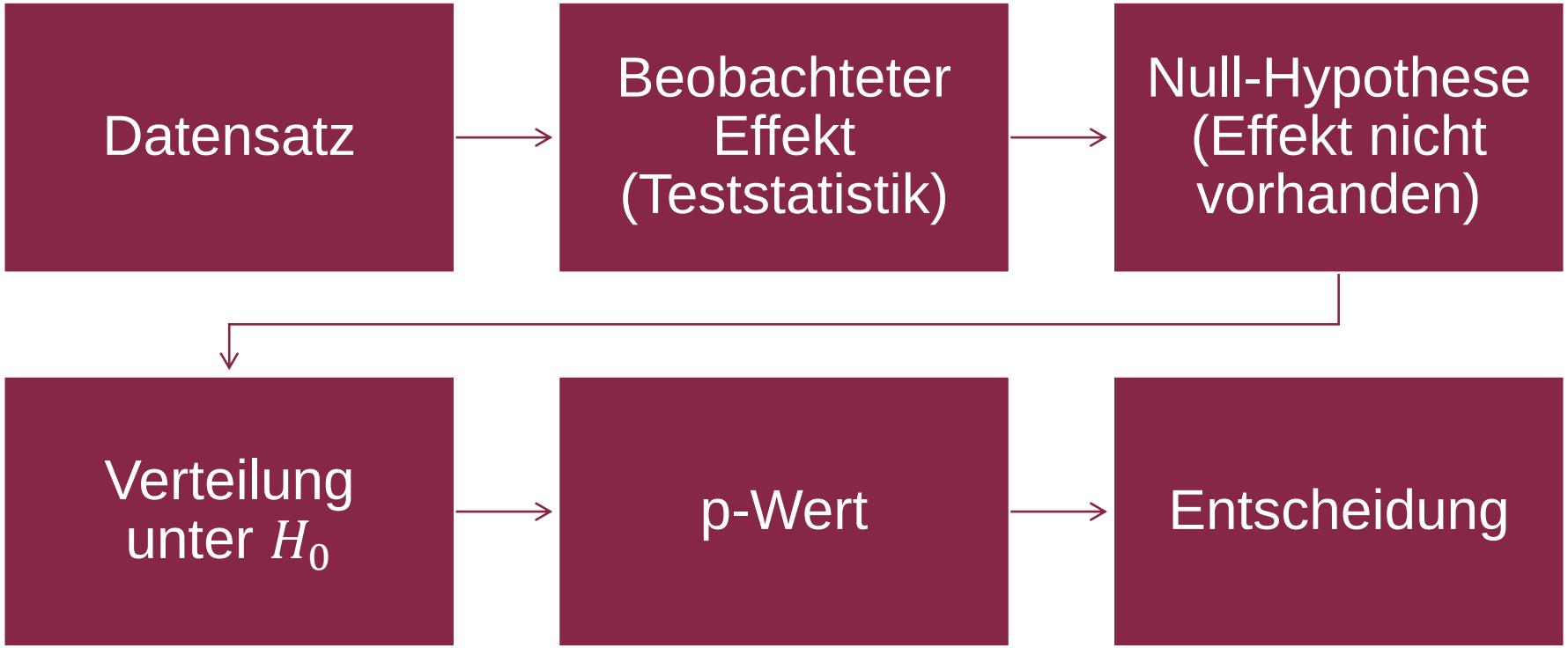
17.01.2022

Klaus Telkmann

Lernziele

- Ein etwas anderer Blick auf statistisches Testen
- Wie konstruiere ich einen geeigneten Test? Insbesondere: Wie modelliere ich die Verteilung unter der Nullhypothese?
- Computer-gestützte Tests (Monte Carlo Simulationen, Permutationstests)

Klassische Vorgehensweise



Motivation

- Angenommen wir haben einen Datensatz und wollen einen angeblichen Effekt untersuchen
- Dann müssen wir zuerst eine Teststatistik (Prüfgröße) bestimmen, die den angeblichen Effekt quantifiziert
- Diese Teststatistik kann z.B. die absolute Differenz der Mittelwerte zweier Gruppen sein (z.B. Differenz des Netto-Einkommens zwischen Frauen und Männern)
- Oder das Odds Ratio (bzw. relatives Risiko) einer Exposition auf eine Zielgröße (z.B. OR für Raucher auf die Zielerkrankung Krebs)

Motivation

- Ein statistischer Test versucht die Frage zu beantworten: „Ist mein beobachteter Effekt real oder nur durch Zufall entstanden?“

- Dazu formulieren wir zwei Hypothesen:
  - H_0: Der Effekt ist nicht vorhanden und wir haben ihn nur zufällig beobachtet
  - H_A: Der Effekt ist real

Motivation

- Insbesondere ist H_0 ein Modell der Welt, in dem der Effekt nicht existiert und nur zufällig zu beobachten ist

- Im Gegensatz dazu ist H_A ein Modell der Welt, in dem der Effekt real ist

Motivation

- Im Idealfall wollen wir die Wahrscheinlichkeit, den Effekt zu beobachten, unter beiden Hypothesen berechnen, also $\mathbb{P}(E \mid H_0)$ und $\mathbb{P}(E \mid H_A)$

- Im Allgemeinen ist H_A aber zu vage formuliert, als dass man sie modellieren könnte

- Daher berechnen wir $\mathbb{P}(E \mid H_0)$, die Wahrscheinlichkeit, einen solchen Effekt zu beobachten unter der Bedingung, dass H_0 wahr ist

Motivation

- Beachte, dass $\mathbb{P}(E \mid H_0)$ gerade der p-Wert ist

- Falls der p-Wert klein ist, schließen wir daraus, dass es unwahrscheinlich ist, diesen Effekt beobachtet zu haben wenn die Nullhypothese wahr ist. Wir gehen dann davon aus, dass der Effekt tatsächlich vorhanden ist

Klassische Herangehensweise

- Die meisten klassischen Tests nutzen die asymptotische Verteilung der Teststatistik unter der Nullhypothese

- Die Tests basieren also auf einer Approximation, die eine effiziente Berechnung des p-Wertes zulässt

Problem 1: Würfeln

- Wir betreiben ein Casino und vermuten, ein Gast hat seinen eigenen gezinkten Würfel für ein Würfelspiel reingeschmuggelt

- Wir müssen unsere Vermutung natürlich beweisen. Daher werfen wir den Würfel 60 mal und schauen uns das Ergebnis an

Wert	1	2	3	4	5	6
Häufigkeit	8	9	19	6	8	10

- Was ist die Wahrscheinlichkeit, dieses Ergebnis durch Zufall erzielt zu haben?

Quelle: http://allendowney.blogspot.com/2011/05/there-is-only-one-test.html

Problem 1: Würfeln

- Nullhypothese: Der Würfel ist nicht gezinkt, d.h.

$$\mathbb{P}(W = 1) = \cdots = \mathbb{P}(W = 6) = \frac{1}{6}$$

- Unter der Nullhypothese erwarten wir die folgenden Häufigkeiten

Wert	1	2	3	4	5	6
Häufigkeit	8	9	19	6	8	10
erwartet	10	10	10	10	10	10

Problem 1: Würfeln

- Wie stark weichen die beobachteten Häufigkeiten von den erwarteten ab?

Wert	1	2	3	4	5	6
Häufigkeit	8	9	19	6	8	10
erwartet	10	10	10	10	10	10
Differenz	-2	-1	+9	-4	-2	0

Chi-Quadrat Test

- Klassisch beurteilt man die Wahrscheinlichkeit, dieses Ergebnis beobachtet zu haben, falls die Nullhypothese wahr ist, mit einem χ^2 –Test

- Dabei bildet man als Teststatistik die Summe der quadrierten relativen Differenzen (SRD) $\frac{(beobachtet-erwartet)^2}{erwartet}$ und summiert diese auf

Wert	1	2	3	4	5	6
Häufigkeit	8	9	19	6	8	10
erwartet	10	10	10	10	10	10
Differenz	-2	-1	+9	-4	-2	0
SRD	0.4	0.1	8.1	1.6	0.4	0

Chi-Quadrat Test

- Es ergibt sich als Teststatistik $T = \sum SRD_i = 10.6$
- Je größer die Teststatisik, desto größer die Abweichung von einem erwarteten Ergebnis eines fairen Würfels
- Asymptotisch (also für $n \to \infty$) ist diese Teststatistik χ^2 –verteilt mit $k - 1$ Freiheitsgraden (k bezeichnet dabei die Anzahl der verschiedenen Ausprägungen, hier 6)
- Also $T \sim \chi^2(5)$

Chi-Quadrat Test

Diesen Wert kann man nun mit dem theoretischen Quantil der χ^2 –Verteilung abgleichen und damit ergibt sich ein p-Wert von 0.0599

Alternativ schaut man in einer Tabelle nach

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Monte Carlo Simulation

- Man kann das Problem auch etwas experimenteller angehen. Bei einer Monte-Carlo Simulation werden wiederholt Zufallsstichproben aus einer gegebenen Verteilung gezogen

- Im Prinzip führen wir das 60-malige Würfeln mit einem **fairen** Würfel wiederholt aus und notieren uns das Ergebnis (also die Teststatistik der Summe der quadrierten relativen Abweichungen)

- Beachte, dass dies die Nullhypothese simuliert, da jeder Durchgang mit einem fairen Würfel erzeugt wurde

Monte Carlo Simulation

- Wir schauen dann, wie häufig wir eine höhere Teststatistik (als 10.6) beobachtet haben. Dies sind ja gerade die „noch extremeren Ereignisse“ obwohl wir hier mit einem fairen Würfel (also unter der Nullhypothese) gewürfelt haben.

- Dieser prozentuale Anteil ist der p-Wert. Als Erinnerung: Der p-Wert ist die Wahrscheinlichkeit ein solches oder noch extremeres Ereignis unter Annahme der Nullhypothese beobachtet zu haben

- Wir haben hier also die Verteilung unter der Nullhypothese modelliert und unsere beobachtete Größe damit abgeglichen

Monte Carlo Simulation

Wert	1	2	3	4	5	6	T^*
Test 1	8	9	11	11	11	10	0.8
Test 2	15	4	7	8	10	16	11
⋮							
Test 1000	13	7	10	8	14	8	4.2

Beispiel für 1000 maliges Wiederholen des 60-maligen Werfens mit einem fairen Würfel. Berechne für jeden Durchgang die Summe der quadrierten relativen Differenzen T^*. Wie oft haben wir hier ein „scheinbar unfaireres“ Ergebnis (also größere Teststatistik) als 10.6 erhalten?

R liefert dafür einen p-Wert von ca. 0.052

Problem 2: Zwei-Stichproben-Test

- Wir betrachten 13 Personen, denen entweder das Medikament A oder B (Placebo) verabreicht wurde. Das Outcome könnte hier z.B. ein Score für einen Schmerzwert sein.

Gruppe											Mittelwert	n
A	23	17	8	12							15	**4**
B	13	12	4	9	4	7	2	11	12		8.22	**9**
Diff											6.78	**13**

- Unterscheiden sich die Mittelwerte der beiden Gruppen signifikant?

Problem 2: Zwei-Stichproben-Test

- Klassisches Vorgehen: Zwei-Stichproben t-Test,

- Abgleichen einer Teststatistik basierend auf der Differenz der Mittelwerte der beiden Gruppen mit dem theoretischen Quantil der Student‘schen t-Verteilung (nach William Sealy Gosset 1908, der die Arbeit unter dem Pseudonym „Student“ veröffentlichte, da sein Arbeitgeber, die Guinness-Brauerei ihm das Publizieren untersagte)

t-Test

- Statistisch gesehen: Wir betrachten zwei Stichproben $X_1, \dots X_n$ und $Y_1, \dots, Y_m$.

- Dann berechnen wir die Differenz der Mittelwerte $\bar{X} - \bar{Y}$ und das gewichtete Mittel der Stichprobenvarianzen

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

t-Test

- Unter der Annahme, dass beide Mittelwerte asymptotisch normalverteilt sind und unter der Nullhypothese, dass die beiden Mittelwerte gleich sind, ist die Teststatistik

$$\mathrm{T} = \sqrt{\frac{nm}{n+m}}\frac{\bar{X}-\bar{Y}}{S}$$

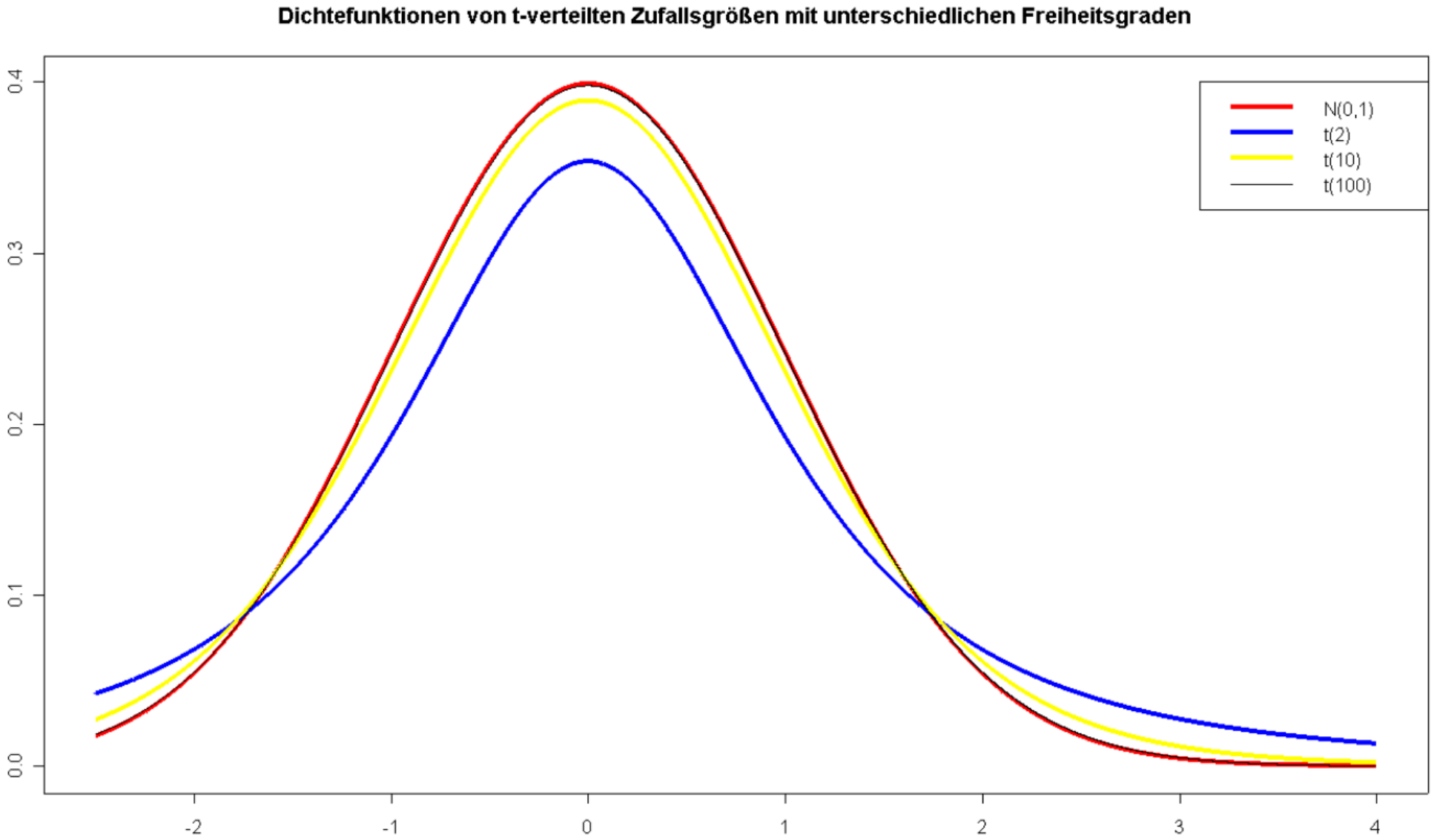
t-verteilt mit $m + n - 2$ Freiheitsgraden

- Wir gleichen den Wert T nun mit den Quantilen der t-Verteilung ab. Falls $|T| > t_{n+m-2}\left(1 - \frac{\alpha}{2}\right)$ so wird die Nullhypothese zum Signifikanzniveau α abgelehnt

t-Test

- In unserem Beispiel ist $\mathrm{T} = 1.93$ und $n + m - 2 = 4 + 9 - 2 = 11$
- Wir gleichen also mit einer t –Verteilung mit 11 Freiheitsgraden ab
- Der Welch t-Test liefert einen p-Wert von 0.1243 und wir würden die Nullhypothese (kein Unterschied zwischen den Mittelwerten) **nicht** ablehnen
- Allerdings setzt dieser Test hinreichend große Gruppengrößen oder Normalverteilung der Werte voraus (beides nicht gegeben)

t-Verteilung



Permutationstest

- Da man mit heutigen Computern den Effizienz-Faktor vernachlässigen kann, können wir auf die „Abkürzung“ über die Approximation einer bekannten Verteilung verzichten und die exakte Verteilung unter der Nullhypothese berechnen.

- Eine nicht-parametrische Möglichkeit dazu bieten sogenannte Permutationstests.

Permutationstest

- Gleiche Fragestellung wie eben: Wirkt Medikament A besser als Medikament B?

Gruppe											Mittelwert	n
A	23	17	8	12							15	**4**
B	13	12	4	9	4	7	2	11	12		8.22	**9**
Diff											6.78	**13**

- Teststatistik: $\mathrm{T} = \bar{X} - \bar{Y} = 6.78$ (Differenz der Mittelwerte)

Permutationstest

- Idee: Falls kein Unterschied zwischen den Mittelwerten der beiden Gruppen besteht, sollte es egal sein, welches Medikament jede Person erhalten hat

- Wir können nun alle Beobachtungen poolen und jeweils zwei Stichproben mit zufällig gezogenen Werten den Gruppen „A“ und „B“ zuordnen und die Differenz der Mittelwerte bilden. Die Gruppengrößen sollen dabei aber gleich bleiben (4 und 9 im Beispiel)

Permutationstest

- Originaler Datensatz

Gruppe											Mittelwert
A	23	17	8	12							15
B	13	12	4	9	4	7	2	11	12		8.22
Diff											6.78

- Beispiel für eine Permutation

Gruppe											Mittelwert
A*	12	7	2	12							8.25
B*	23	12	8	13	17	4	9	4	11		11.22
Diff*											-2.98

Permutationstest

- Wir wiederholen diese Prozedur nun für jede mögliche Aufteilung der Werte in zwei Gruppen der Größen 4 und 9 und schreiben die Differenzen der Mittelwerte in eine Liste

- Diese Differenzen bilden die exakte Verteilung. Wenn ein Unterschied zwischen den Gruppen besteht, sollte unsere Teststatistik am Rand der Verteilung liegen (wie bei einem t-Test)

Permutationstest

- Wir sortieren also die aus den Permutationen gebildeten Mittelwerte und gleichen mit der originalen Teststatistik ab

- Falls die Teststatistik außerhalb der mittleren $(1-\alpha)\cdot 100\%$ der Werte liegt, so können wir die Nullhypothese zum Signifikanzniveau $1-\alpha$ ablehnen (wie beim t-Test)

- Bzw. $p = \frac{1}{\#Permutationen}\sum 1(|T| \leq |T_i^*|)$

  (Was ist der prozentuale Anteil, so dass genauso oder noch extremere absolute Differenzen als die beobachtete unter der Nullhypothese erzeugt wurden?)

Permutationstest

Wir erhalten also eine Liste mit allen möglichen absoluten Differenzen zwischen zwei zufällig gezogenen Gruppen

Im Allgemeinen ist die Liste sehr lang (in diesem Beispiel hätten wir 715 Einträge)

1.4
1.3
5
3.5
2.7
1.3
⋮
3.6

Wir sortieren diese Liste nun in aufsteigender Reihenfolge.

Uns interessiert, an welcher Stelle unsere beobachtete Differenz auftaucht bzw. wieviele Werte noch größer sind.

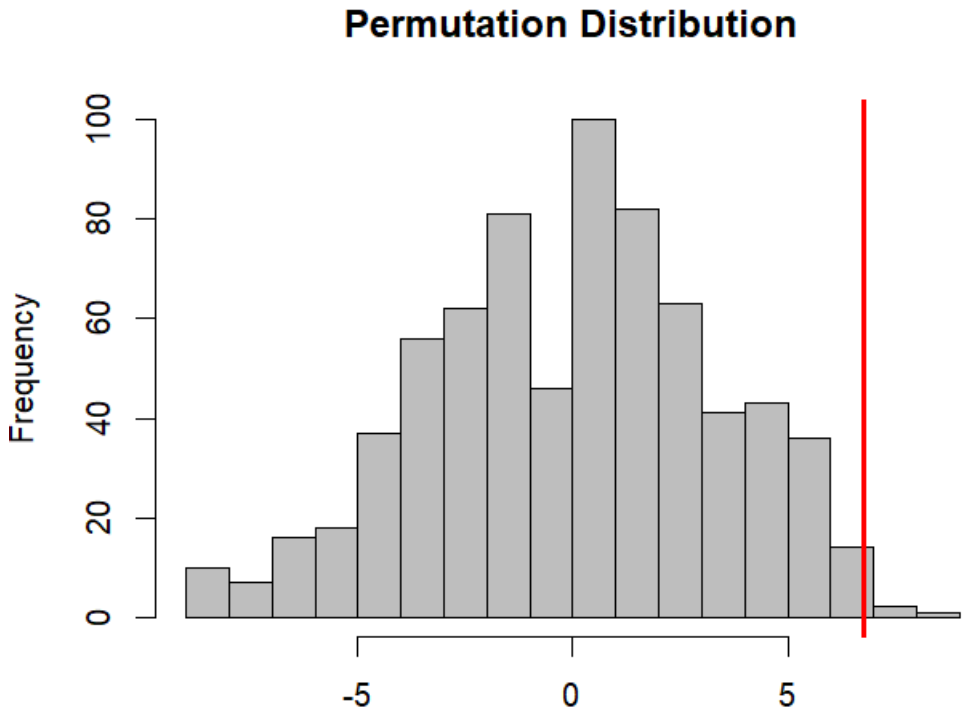
34 Werte sind größer oder gleich 6.78, also

$p = \frac{34}{715} \approx 0.047$

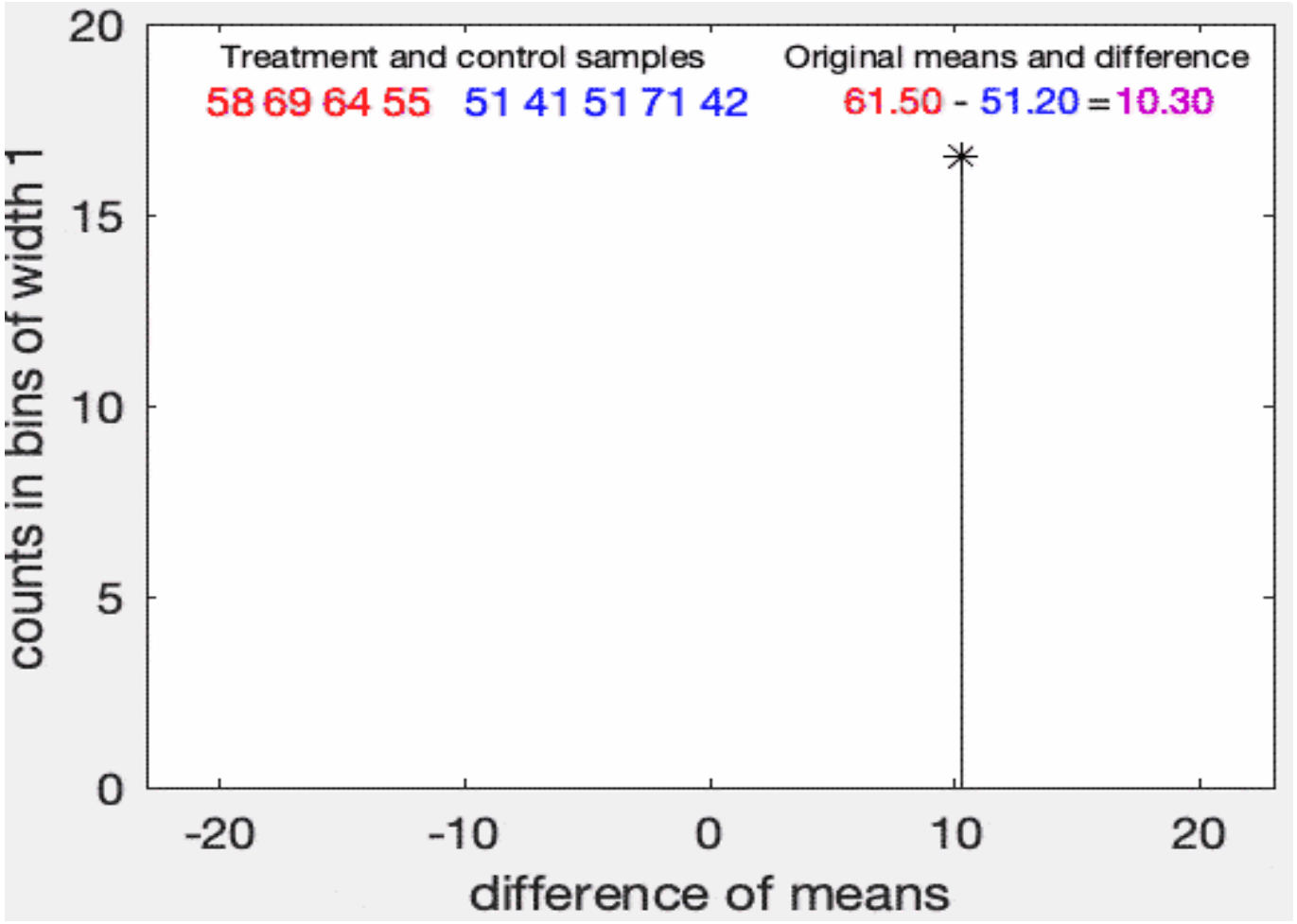
0.08
0.08
0.09
⋮
6.78
⋮
8.57
8.58

Permutationstest

Es ergibt sich die folgende Verteilung der Differenzen der Mittelwerte in den permutierten Gruppen (also unter der Nullhypothese). Hier als Histogramm dargestellt. In rot eingezeichnet ist die „echte“ Differenz der Mittelwerte.



Permutationstest



https://commons.wikimedia.org/wiki/File:Permutation_test_example_animation.gif#/media/File:Permutation_test_example_animation.gif

Permutationstest

- Man nennt diesen Test auch „Fisher-Pitman (Randomization) Test“
- In R z.B. implementiert im Package „EnvStats“ in der Funktion „twoSamplePermutationTestLocation“ (mit der Option „exact=TRUE“)
- In SAS: https://blogs.sas.com/content/iml/2014/11/21/resampling-in-sas.html oder http://www.utstat.toronto.edu/~brunner/oldclass/305s14/lectures/305s14PermutationTestSAS.pdf
- Lässt sich aber auch in wenigen Zeilen selbst programmieren

Permutationstest

- Wieviele mögliche Permutationen gibt es für diesen Test denn eigentlich?

- Die Anzahl der Permutationen lässt sich mit dem Binomialkoeffizienten bestimmen: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

- Dies ist die Anzahl der Möglichkeiten, k aus n Elementen zu ziehen ohne Zurücklegen und ohne Berücksichtigung der Reihenfolge. Hier ist n die Anzahl aller Elemente und k die Anzahl der Elemente einer (beliebigen) Gruppe.

Permutationstest

- Die Anzahl der möglichen Permutationen explodiert sehr schnell
- Beispiel: Angenommen wir haben 100 Beobachtungen aufgeteilt in zwei gleich große Gruppen. Wie lange bräuchte ein Supercomputer, der die Differenz für 1 Billion Permutationen pro Sekunde berechnen kann, für die exakte Verteilung?
- $\binom{100}{50} \approx 10^{29}$ Permutationen. Also $\frac{10^{29}}{10^{12}} = 10^{17}$ Sekunden. Das entspricht ungefähr 3.000.000.000 Jahren.

Permutationstest

- Es ist also für die allermeisten Probleme quasi unmöglich alle Permutationen zu berechnen

- Das ist aber zum Glück auch gar nicht notwendig: Wir können den Test auch für eine relativ kleine Zahl (z.B. 10.000) an zufälligen Permutationen durchführen und erhalten immer noch zuverlässige Ergebnisse

Permutationstest

- Von einem Permutationstest spricht man streng genommen nur dann, wenn man **alle** möglichen Permutationen berechnet hat
- Wenn dies nicht möglich ist und man eine große Zahl an zufälligen Permutationen betrachtet, so spricht man von einem „Random Permutation Test“ oder auch „Monte Carlo Permutation Test“
- In der Literatur werden beide aber oft Permutationstest genannt
- Ein randomisierter Test ist übrigens etwas völlig anderes und damit nicht zu verwechseln

Problem 3: Verbundene Stichproben

- Angenommen wir betrachten eine Kohorte und messen für jede Person zu zwei unterschiedlichen Zeitpunkten eine Zielgröße

t_0	2	4	7	3	5	6	1	7	2
t_1	5	8	6	8	6	5	6	7	6
Δt	+3	+4	-1	+5	+1	-1	+5	0	+4

- Unterscheiden sich die Messwerte zwischen der ersten und zweiten Messreihe signifikant voneinander?

Problem 3: Verbundene Stichproben

- Auch hier kann man klassisch wieder einen t-Test für verbundene Stichproben durchführen. Dabei greifen wieder die selben (hier nicht erfüllten) Annahmen

- Wir können aber auch hier mit einem Permutationstest die Verteilung unter der Nullhypothese bestimmen

Problem 3: Verbundene Stichproben

- Vorüberlegung: Im Prinzip interessieren uns nur die einzelnen Differenzen

t_0	2	4	7	3	5	6	1	7	2
t_1	5	8	6	8	6	5	6	7	6
Δt	+3	+4	-1	+5	+1	-1	+5	0	+4

- Die Summe der Differenzen (hier 20) ist unsere Prüfgröße

- Wie können wir nun sinnvoll permutieren?

Problem 3: Verbundene Stichproben

- Unter der Nullhypothese (Differenzen sind im Mittel 0) sollte es egal sein, ob wir die Zielgröße für eine Person zum Zeitpunkt t_0 oder t_1 messen
- Wenn wir nun für eine Person diese beiden Messwerte vertauschen (permutieren) wie ändert sich dann die Differenz?
- Es ändert sich offensichtlich nur das Vorzeichen

Problem 3: Verbundene Stichproben

t_0	2	4	7	3	5	6	1	7	2
t_1	5	8	6	8	6	5	6	7	6
Δt	+3	+4	-1	+5	+1	-1	+5	0	+4

Beispiel: Wir vertauschen die Messwerte für die Personen 1,4,6 und 8

t_0	5	4	7	8	5	5	1	7	2
t_1	2	8	6	3	6	6	6	7	6
Δt	-3	+4	-1	-5	+1	-1	+5	0	+4

Die Differenzen sind bis auf das Vorzeichen gleich geblieben. Die Summe der Differenzen ist nach dem Permutieren aber nur noch 4 (im Vergleich zu 20 bei den Original-Daten)

Problem 3: Verbundene Stichproben

- Das Verfahren ist nun wie eben: Wir bilden alle möglichen Permutationen, schreiben die Summe der Differenzen in eine Liste und gleichen unsere ursprüngliche Differenz damit ab

- Beachte, dass es hier viel weniger mögliche Permutationen gibt. Wieviele, und warum?

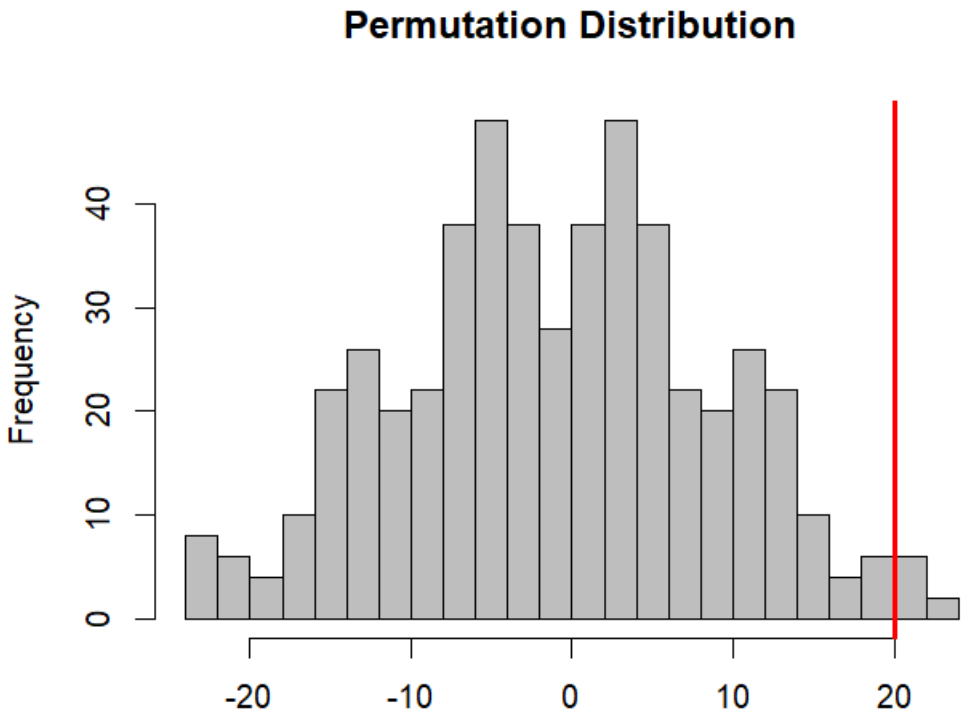
- Es gibt $2^{\#Personen}$ Permutationen, da wir für jede Person jeweils nur das Vorzeichen der Differenz ändern können. Im Beispiel also $2^9 = 512$ Permutationen

Problem 3: Verbundene Stichproben

Für einen einseitigen Test (Sind die Werte zum Zeitpunkt t_1 größer?) schauen wir nur wieviele der Differenz-Summen gleich oder größer als unsere Teststatistik sind.

Der Test liefert einen p-Wert von $\frac{14}{512} \approx 0.027$

(Für einen zweiseitigen Test würden wir nur die absoluten Summen der Differenzen betrachten)



Problem 3: Verbundene Stichproben

- In R z.B. implementiert im Package „EnvStats“ in der Funktion „twoSamplePermutationTestLocation“ (mit den Optionen „exact=TRUE“ und „paired=TRUE“)

- Auch hier gilt: Falls die Anzahl an Permutationen zu groß ist, lässt sich ein random permutation test mit einer hinreichend großen Anzahl an Permutationen (z.B. 10.000) durchführen

Problem 4: Unabhängigkeit

- Wir wollen untersuchen, ob zwei kategorielle Variablen unabhängig voneinander sind

- Hängen Exposition und Erkrankung zusammen?

	Krank	Nicht krank	
Exponiert	6	4	10
Nicht exponiert	3	12	15
	9	16	25

Problem 4: Unabhängigkeit

- Was bedeutet Unabhängigkeit im statistischen Sinne?
- Zwei Ereignisse A und B sind unabhängig, wenn

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

- D.h. die Wahrscheinlichkeit, dass A und B gemeinsam auftreten entspricht dem Produkt der Wahrscheinlichkeiten von A und B

Problem 4: Unabhängigkeit

- In unserem Beispiel:
  - $\mathbb{P}(exponiert\ und\ krank) = \frac{6}{25} = 0.24$
  - $\mathbb{P}(exponiert) \cdot \mathbb{P}(krank) = \frac{9}{25} \cdot \frac{10}{25} = 0.144$
- Bei Unabhängigkeit erwarten wir also von $0.144 \cdot 25 = 3.6$ Personen krank und exponiert zu sein

	Krank	Nicht krank	
Exponiert	6	4	10
Nicht exponiert	3	12	15
	9	16	25

Problem 4: Unabhängigkeit

- Ähnlich wie im ersten Beispiel (Würfeln) können wir also für jede Kombination die erwarteten Häufigkeiten berechnen durch

$$\mathbb{P}(A) \cdot \mathbb{P}(B) \cdot n$$

wobei n die Anzahl Beobachtungen ist und A exponiert oder nicht exponiert und B krank oder nicht krank sein kann

- Wir bilden nun wieder die Summe der quadrierten relativen Differenzen der beobachteten und erwarteten Häufigkeiten

Problem 4: Unabhängigkeit

Beobachtet	Krank	Nicht krank	
Exponiert	6	4	10
Nicht exponiert	3	12	15
	9	16	25

Erwartet	Krank	Nicht krank	
Exponiert	3.6	6.4	10
Nicht exponiert	5.4	9.6	15
	9	16	25

$$T = \sum \frac{(beobachet_i - erwartet_i)^2}{erwartet_i} \approx 4.167$$

Chi-Quadrat Unabhängigkeitstest

- Klassisch würde man hier wieder einen χ^2-Unabhängigkeitstest verwenden

- Asymptotisch folgt T einer χ^2-Verteilung mit einem Freiheitsgrad (die Anzahl der Freiheitsgrade ist i.A. [#Zeilen-1]·[#Spalten-1]). Man gleicht dann wieder mit den theoretischen Quantilen dieser Verteilung ab.

- Hier ergibt sich ein p-Wert von 0.04123

Chi-Quadrat Unabhängigkeitstest

- Eine Voraussetzung, um den χ^2 –Unabhängigkeitstest anzuwenden, ist, dass wir für jede Kombination von Ereignissen mindestens 5 erwarten.

- Das ist hier nicht der Fall (für exponiert und krank erwarten wir nur 3.6). Daher ist der p-Wert womöglich verzerrt

- Eine Alternative bietet der exakte Test nach Fisher, der auch für kleine Stichproben das geforderte Signifikanzniveau einhält

Permutationstest

- Wir können hier einen Permutationstest anwenden, um die Verteilung unter der Nullhypothese zu bestimmen

- Vorüberlegung: Die Nullhypothese besagt, dass Expositionsstatus und Krankheitsstatus unabhängig voneinander sind. Unter dieser Annahme sollte es also egal sein, welchen Krankheitsstatus wir jeder beliebigen Person zuordnen.

Permutationstest

- Wir teilen die 9 beobachteten Krankheitsfälle und 16 Gesunden nun zufällig auf die 10 Exponierten und 15 Nicht-Exponierten auf. Beachte, dass die Randsummen stets gleich bleiben!

Beobachtet	Krank	Nicht krank	
Exponiert	6	4	10
Nicht exponiert	3	12	15
	9	16	25

$T = 4.167$

Permutiert	Krank	Nicht krank	
Exponiert	4	6	10
Nicht exponiert	5	10	15
	9	16	25

$T^* = 0.116$

Permutationstest

- Wir wiederholen das Ganze nun für alle möglichen Kombinationen und schreiben die Teststatistiken T^* in eine (ziemlich lange) Liste

- Wie oft wurde nun eine Teststatistik T^* durch Permutationen erzeugt, die gleich oder größer als unsere beobachtete Statistik T ist? Anders ausgedrückt: Wieviele Ereignisse weichen genauso stark oder noch stärker von der Unabhängigkeit ab, auch wenn sie unter der Annahme der Unabhängigkeit (sprich Nullhypothese) erzeugt wurden?

- Der prozentuale Anteil entspricht wieder dem p-Wert (hier 0.0872)

Permutationstest

- Diese Art von Test lässt sich problemlos auch auf größere Kontingenztafeln übertragen
- Im Falle einer Vierfeldertafel entspricht dieses Vorgehen dem exakten Test nach Fisher
- Für große Stichproben und unter den gegebenen Annahmen liefert der χ^2 –Test aber sehr gute Resultate

Implementierungen

- In SAS mit proc freq: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/procstat/procstat_freq_details106.htm#procstat.freq.freqexmc

- In R mit chisq_test im package EnvStats

Was gibt es noch so?

- Das waren nur ein paar simple Beispiele. Ein Permutationstest lässt sich aber prinzipiell für jede Teststatistik konstruieren ohne Annahmen über ihre Verteilung

- Bootstrapping ist eine weitere Resampling Methode. Damit lassen sich auch Konfidenzintervalle berechnen. Ein Unterschied zum Permutationstest besteht darin, dass hierbei mit Zurücklegen aus der Stichprobe gezogen wird.