Informelle statistische Inferenz

Manfred Borovcnik

Universität Klagenfurt

ÖMG, Fortbildungstag für Lehrkräfte 2021, Wien, 9. 4. 2021

• Informelle und „Informal“ Inference

1. Ein elementarer Zugang zum Signifikanz-Test – Rangtests, Re-Randomisierung und der *p*-Wert
2. Informelle Inferenz – eine Analogie zur Medizin
3. Informelle Wege zur statistischen Inferenz – Beispiele
4. „Informal Inference“ – Eine vereinfachte Inferenz
5. Resümee – Vereinfachung oder Reduktion

1. Ein elementarer Zugang zum Signifikanz-Test – Rangtests, Re-Randomisierung und der *p*-Wert

Aufgabe:

Empirischer Nachweis der Effizienz eines blutdrucksenkenden Medikaments durch eine placebo-kontrollierte, randomisierte, doppel-blinde klinische Studie

Zielvariable:

Intra-individuelle Differenz des Blutdrucks $\Delta = sys_{Basis} - sys_{4.Woche}$ [mm Hg]

Große Werte entsprechen einer starken Wirkung.

Hypothesen:

Nullhypothese (H_0): Verum = Placebo

Alternativhypothese: Verum ist besser als Placebo

Wenn Verum besser ist, dann sind große Werte unter der Behandlung im Vergleich zu Placebo zu erwarten.

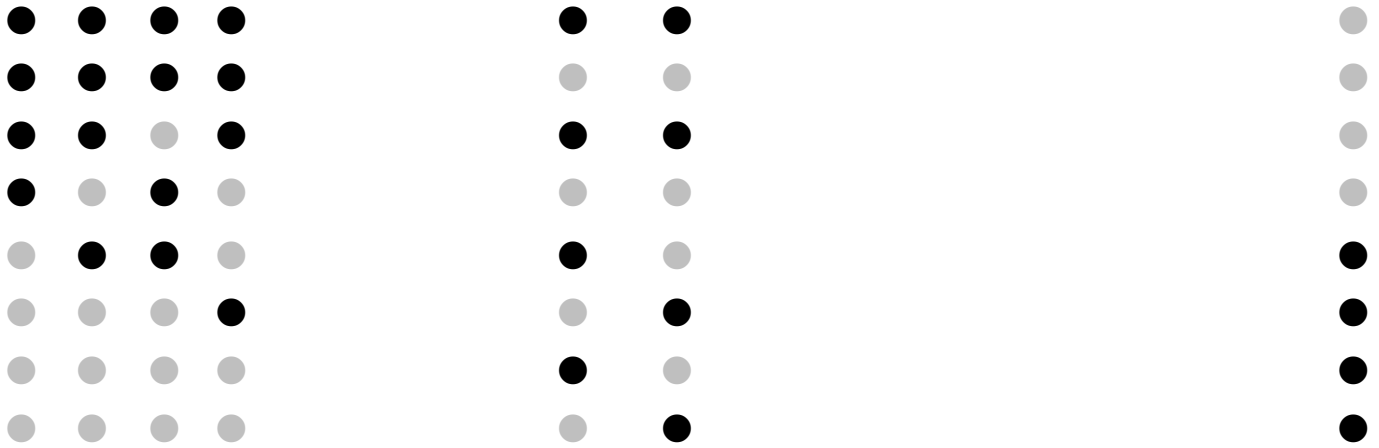
Mann-Whitney-Test für unabhängige Stichproben

1.1 Umordnung und Ränge

	Originaldaten	Geordnet	Rang	Rangsumme
Placebo	2,5 0,9 1,8 3,6	0,9 1,8 2,5 3,6	1 2 3 4	$\Sigma = 10$
Verum	3,7 5,2 4,8 6,1	3,7 4,8 5,2 6,1	5 6 7 8	$\Sigma = 26$

Unter H_0 sind alle Zuordnungen gleich wahrscheinlich!

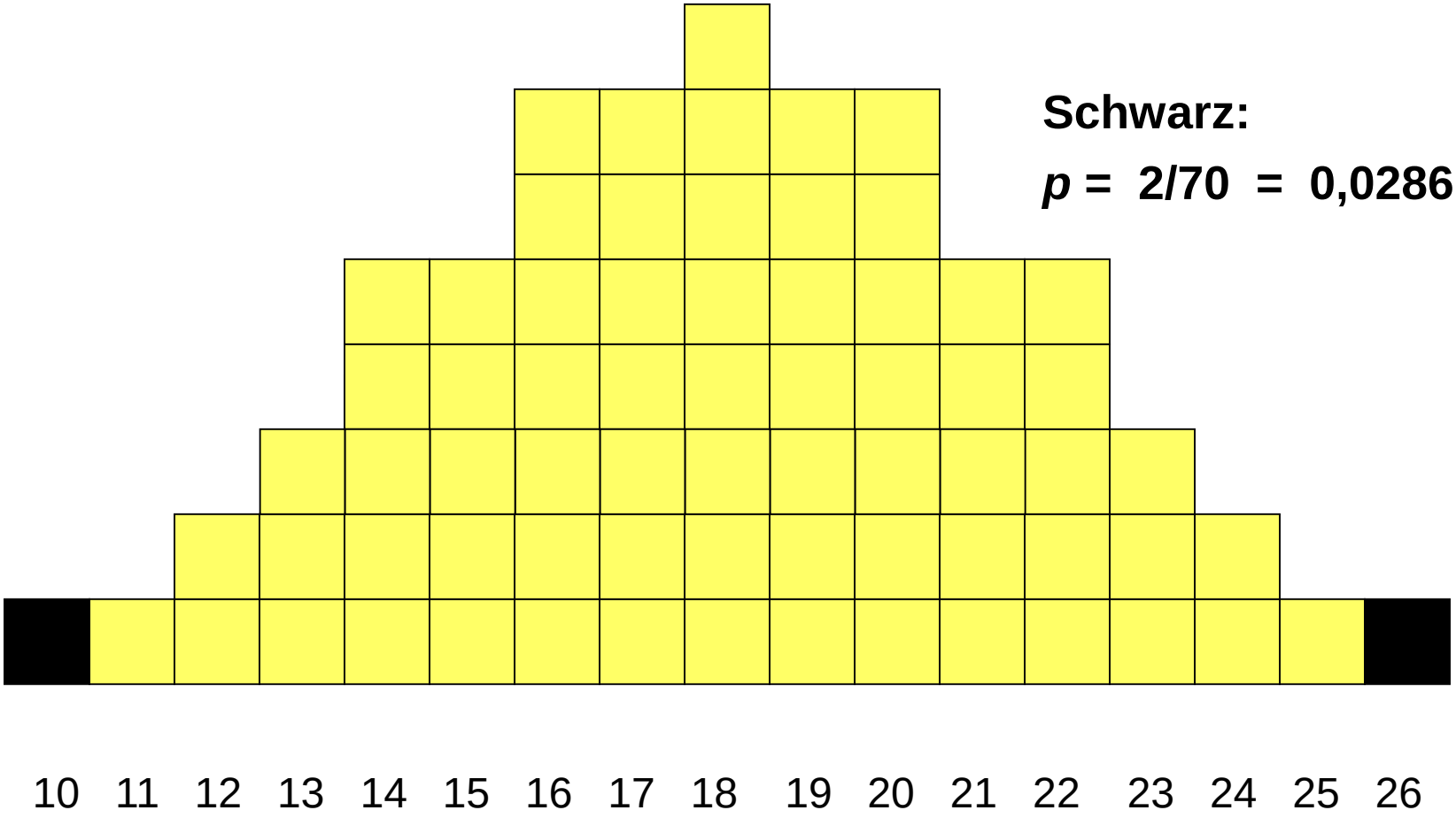
Anzahl der Umordnungen: $\binom{8}{4}$ = ‚8 über 4‘ = 70



Rangsumme **10 11 12 13 … 16 18 … 26**

Verteilung der Rangsumme (n = 70 Umordnungen)

Szenario-Annahme: Es gibt keinen Unterschied zwischen Verum & Placebo (Nullhypothese)



1.2 Der *p*-Wert: erste Bedenken

p = Wahrscheinlichkeit für ein beobachtetes Resultat, wenn H_0 zutrifft.

Wenn *p* kleiner als 5% ist, wird die Nullhypothese abgelehnt;

p ist die Wahrscheinlichkeit für eine falsch-positive Aussage,

d.h., der Test ergibt ein signifikantes Ergebnis, wenn das Droge/Medikament nicht wirksam ist:

p (Test signifikant | Droge ist nicht wirksam).

Wir haben etwas beobachtet, was weniger Wahrscheinlichkeit als 5% hat, **wenn** H_0 zutrifft (Droge nicht wirksam).

Aber, wir sind **nur** an dieser Zahl interessiert :

P (Droge ist wirksam | Test signifikant) ??

Formale Methoden und wissenschaftliche Prinzipien

Eine Entscheidung über eine klinische Studie basiert auf statistischen Methoden.

Ärzte sind keine Experten in Statistik und sie müssen es auch nicht sein.
Dennoch sollten sie die Prinzipien wissenschaftlicher Methoden kennen.

No test based upon a theory of probability
can by itself provide any valuable evidence of
the truth or falsehood of a hypothesis.

Neyman J., Pearson E. (1933): On the problem of most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. A, 231*, 289-337.

Ein Dialog dazu zwischen einem Mediziner und einem Statistiker

M: Du hast dich so bemüht, mir den statistischen Test zu erklären – aber was bedeutet es, wenn mein Test ein signifikantes Resultat ergibt? Kann ich dann behaupten, dass die Droge wirksam ist?

ST: Nein – Du kannst nur berechnen, wie wahrscheinlich solch ein Resultat ist, **WENN** die Droge tatsächlich nicht wirksam ist.

M: Die Ethikkommission hat die Studie über die Wirksamkeit dieser Droge bewilligt. Ich habe dich gefragt, ob man das durch einen statistischen Test beweisen kann. Weil das Resultat signifikant ist, dachte ich, dass die Wahrscheinlichkeit, dass die Droge wirksam ist, 95% beträgt, weil der *p*-Wert 5% ist.

ST: Du hast mich etwas gefragt, auf das der *p*-Wert keine Antwort gibt. Die Fehler-Wahrscheinlichkeit für deine Aussage ist höher – **dennoch, ich kann das nicht berechnen**.

M: Du magst ja Recht haben, aber ich habe das getan, wie es alle tun – warum sollte das falsch sein? **Das Resultat** des statistischen Tests **ist signifikant und wird publiziert**: Die Droge ist wirksam ($p < 0.05$).

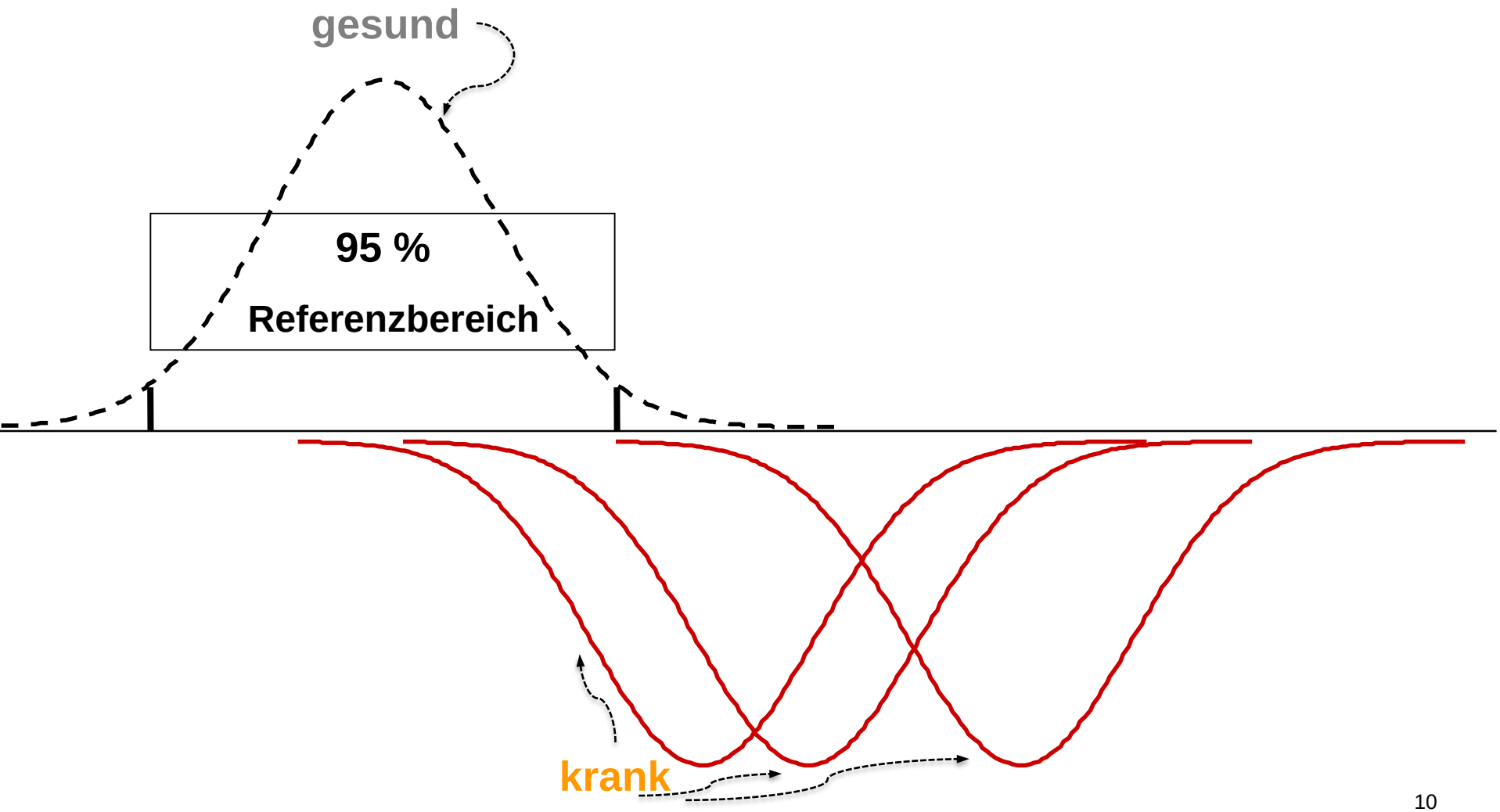
2. Informelle Inferenz – eine Analogie zur Situation in der Medizin

Wir untersuchen die Situation in der Medizin, wo es immer eine Entscheidung gibt, die dann zu diversen Fehlern führen kann, wie auch immer die Entscheidung ausfällt.

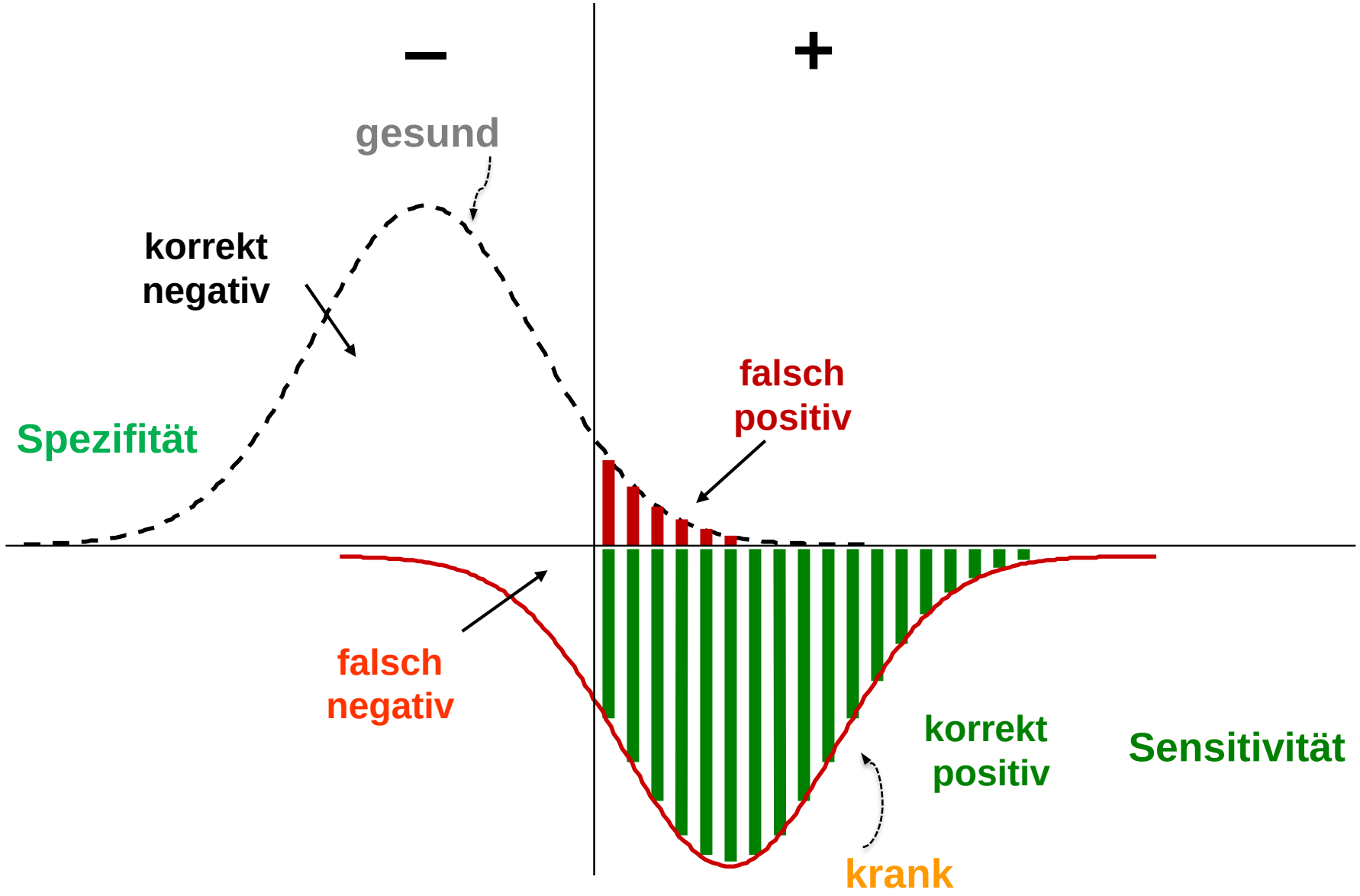
Ein diagnostischer Test kann mit einem statistischen Test verglichen werden:

- Dies dient dazu, statistische Tests besser zu verstehen.
- Dies kann auch dazu dienen, die medizinischen Entscheidung besser zu verstehen und zu untersuchen.

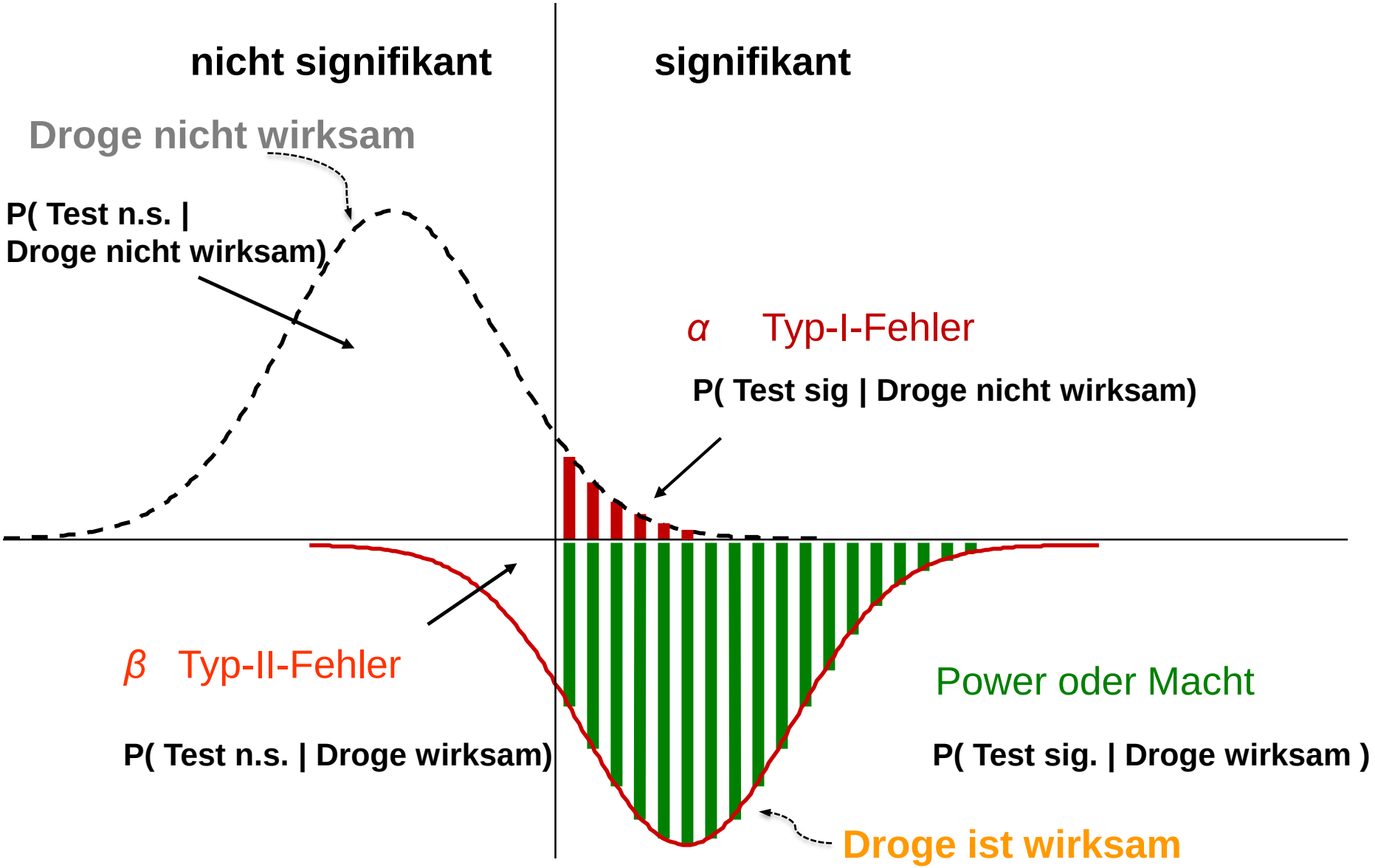
2.1 Trennung der Verteilung einer Variablen zwischen gesunden und kranken Menschen



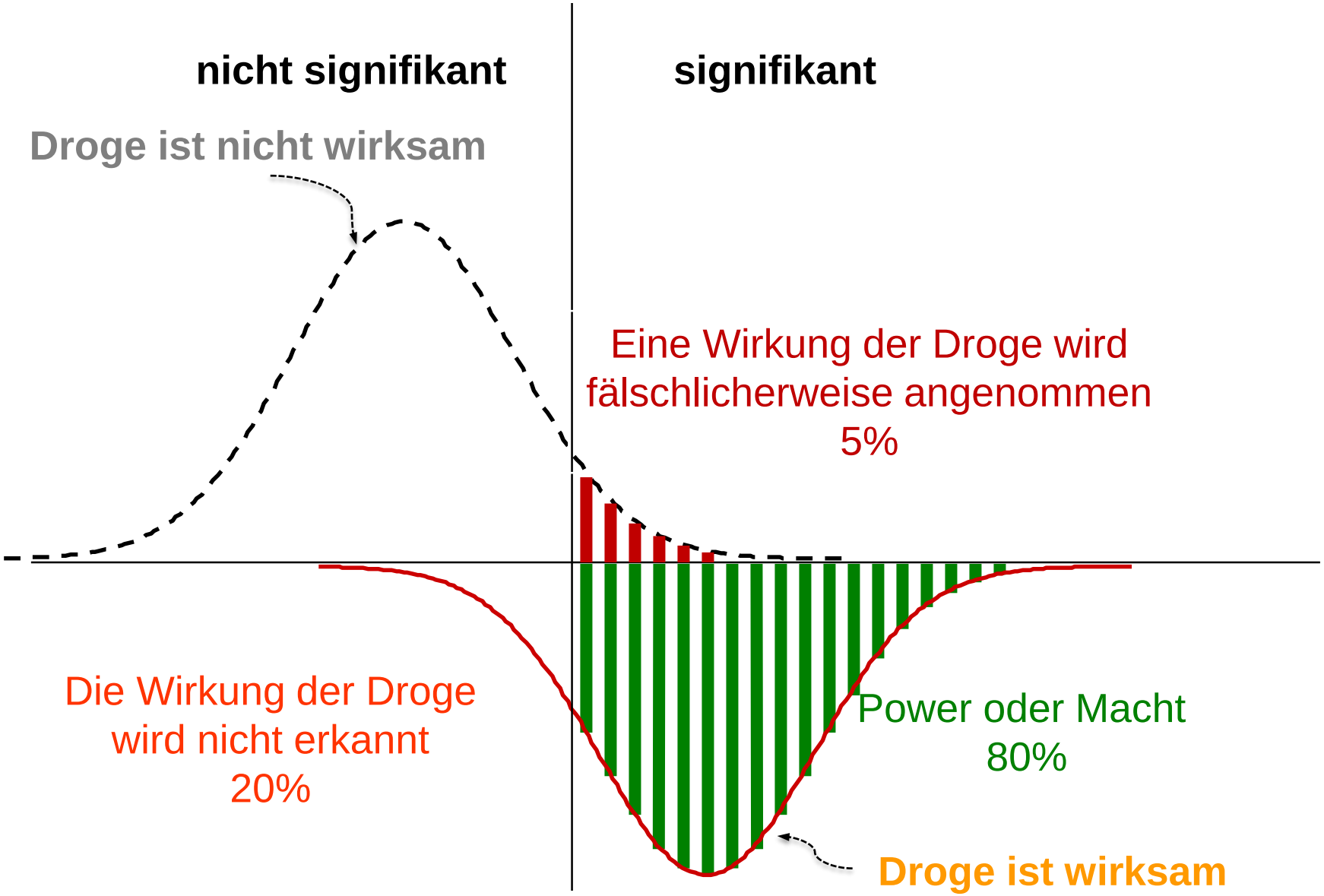
Trennung der Gruppen: Diagnostischer Test



Trennung der Gruppen: Statistischer Test



Klinische Versuche von Drogen als statistischer Test



2.2 Medizinischer Test als Entscheidung

- 2 Stichproben: Placebo (P) , Verum (V)
- Hypothesen: H_0: P = V (Nullhypothese)
  H_A: P ≠ V (Alternativhypothese)
- Entscheidung über H_0 oder H_A

Wirklichkeit

Testentscheidung		H_0	H_A
		Droge nicht wirksam	**Droge wirksam**
Droge nicht wirksam	H_0	korrekt $1-\alpha$	Falsch-negative Entscheidung β Konsumentenrisiko
Droge wirksam	H_A	Falsch-pos. Entscheidung α Produzentenrisiko	korrekt $1-\beta$ Power / Macht

Die Anordnung von H_0 und H_A und der Entscheidungen ist umgekehrt zu zuvor!

Die Analogie zwischen diagnostischem und statistischen Test

p (Test + | **krank**)

Sensitivität

p (Test sig. | **Droge wirksam**)

Power oder Macht

Fehlt: wir haben keinerlei Information über:

p (**krank** | Test +)

Positiver Prädiktiver Wert

p (**Droge wirksam** | Test sig.)

Abhängig von der Prävalenz

Abhängig von der Qualität der Forschungshypothesen

Das fehlende Bindeglied: Prävalenz und die Bayesformel

Mammographie in radiologischer Klinik und im Screening

	Ca	No Ca	
+	**80** Sensitivität ↑	**4** Falsch pos. ↑	84
–	**20** Falsch neg. ↑	**96** Spezifität ↑	116
	100	**100**	200

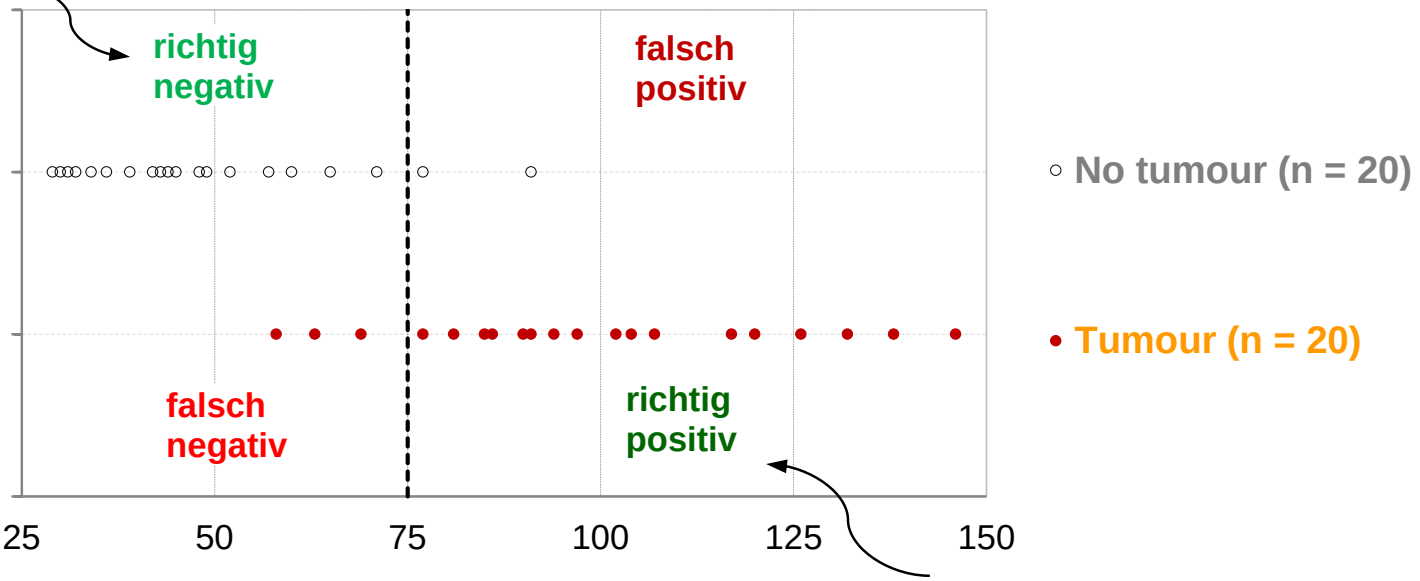
	Ca	No Ca	
+	**640**	**3 968**	4 608
–	**160**	**95 232**	95 392
	800	**99 200**	100 000

Prävalenz	**Klinik** **50 %**	**Screening** **0.8 %**	
Sensitivität ↑	80/100 = 80.0%	80.0%	P (+| Ca)
Spezifität ↑	96/100 = 96.0%	96.0%	P (– | No Ca)
Pos.präd.Wert (PPV) →	80/84 = 95.2%	13.9%	P (Ca| +)
Neg.präd.Wert (NPV) →	96/116 = 82.8%	99.8%	P (No Ca| –)

2.3 Trennpunkte, um zwischen den Gruppen von Gesunden und Kranken zu unterscheiden

Immunchem. faecal occult blood test FOBT ng/ml

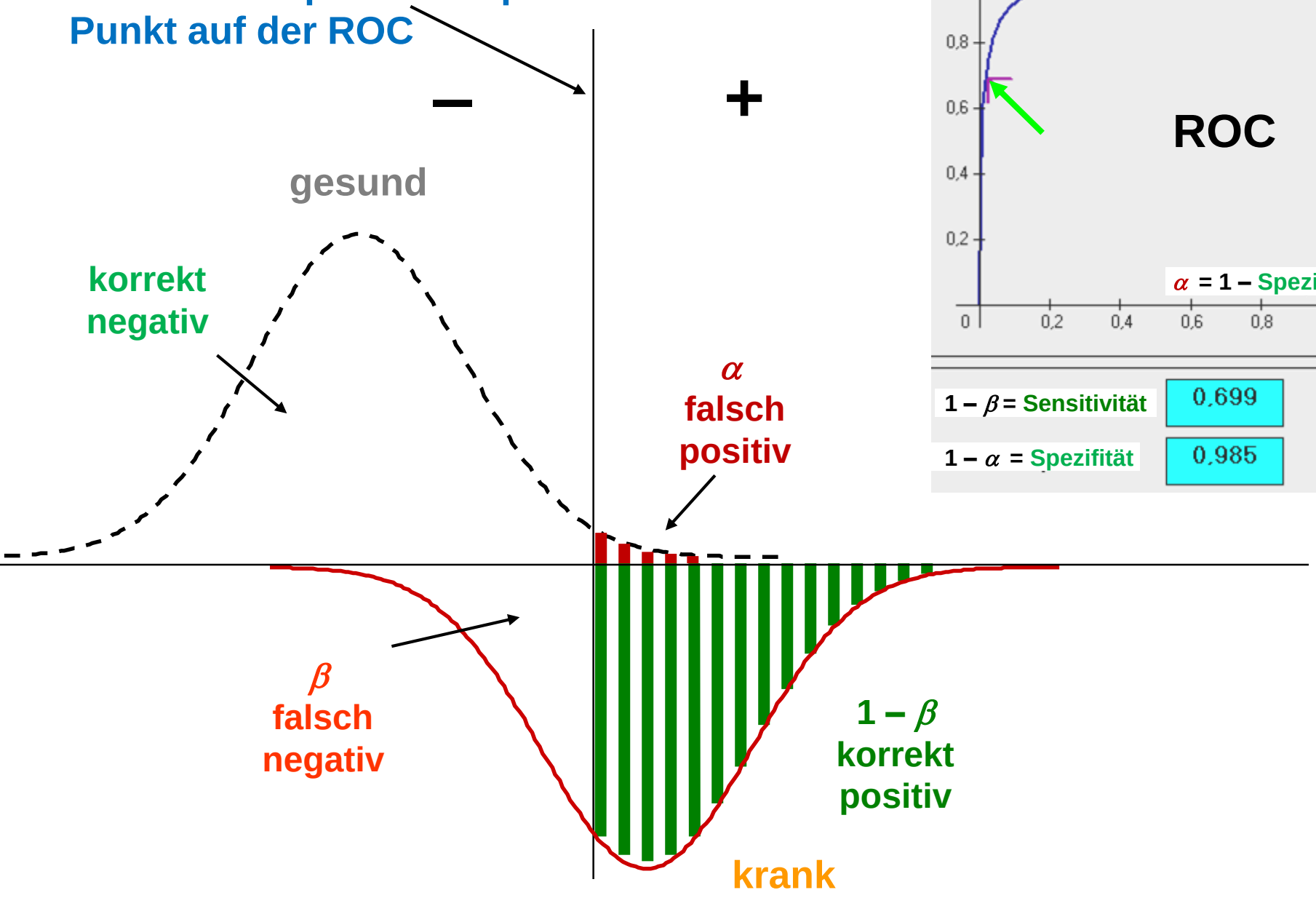
Spezifität: 18/20 (90%)



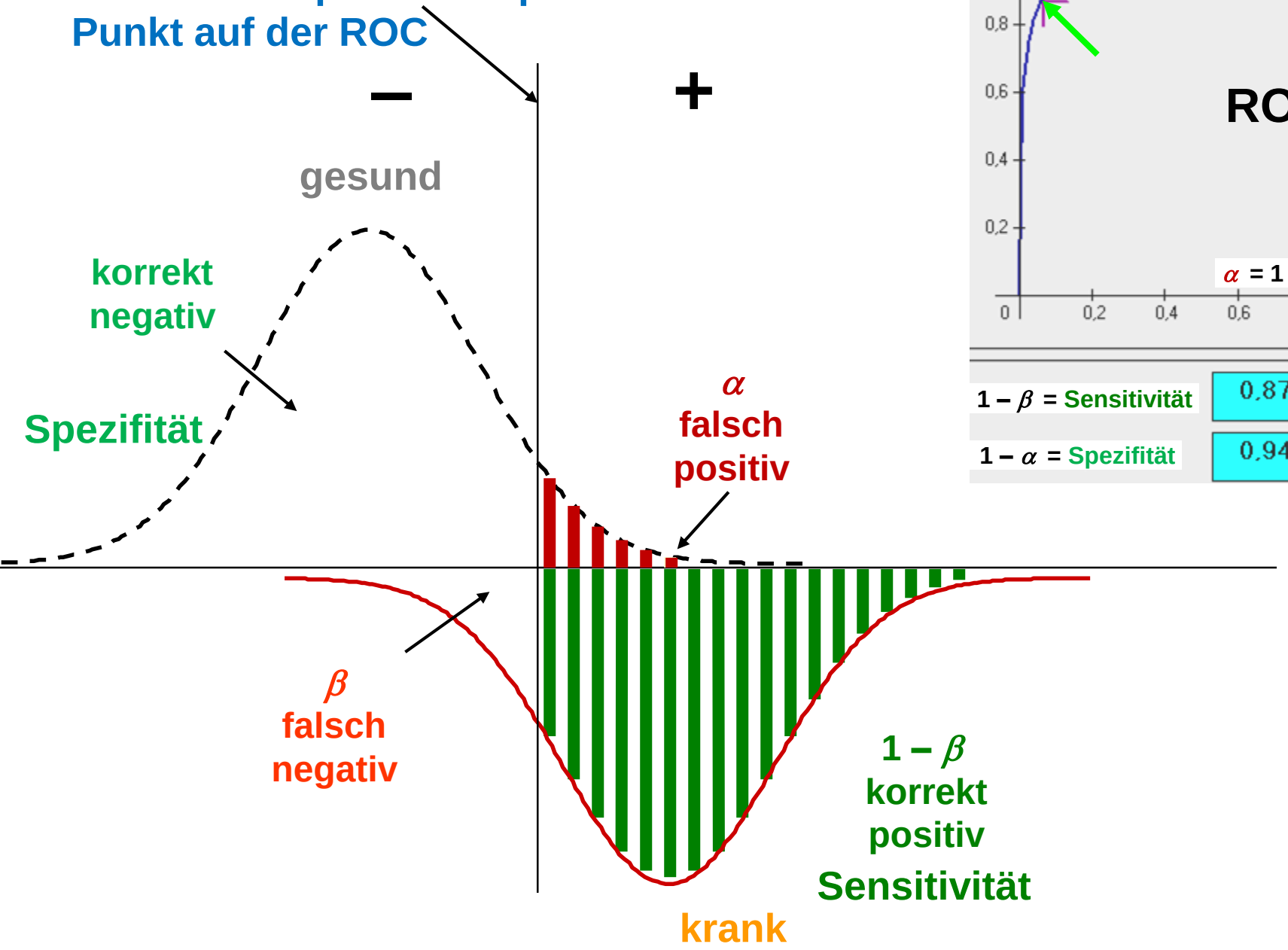
Bei **diesem** Trennpunkt	**Spezifität**	**Sensitivität**
	18 / 20 (90%)	**17 / 20 (85%)**

Welchen Trennpunkt sollte man für die Diagnose verwenden?

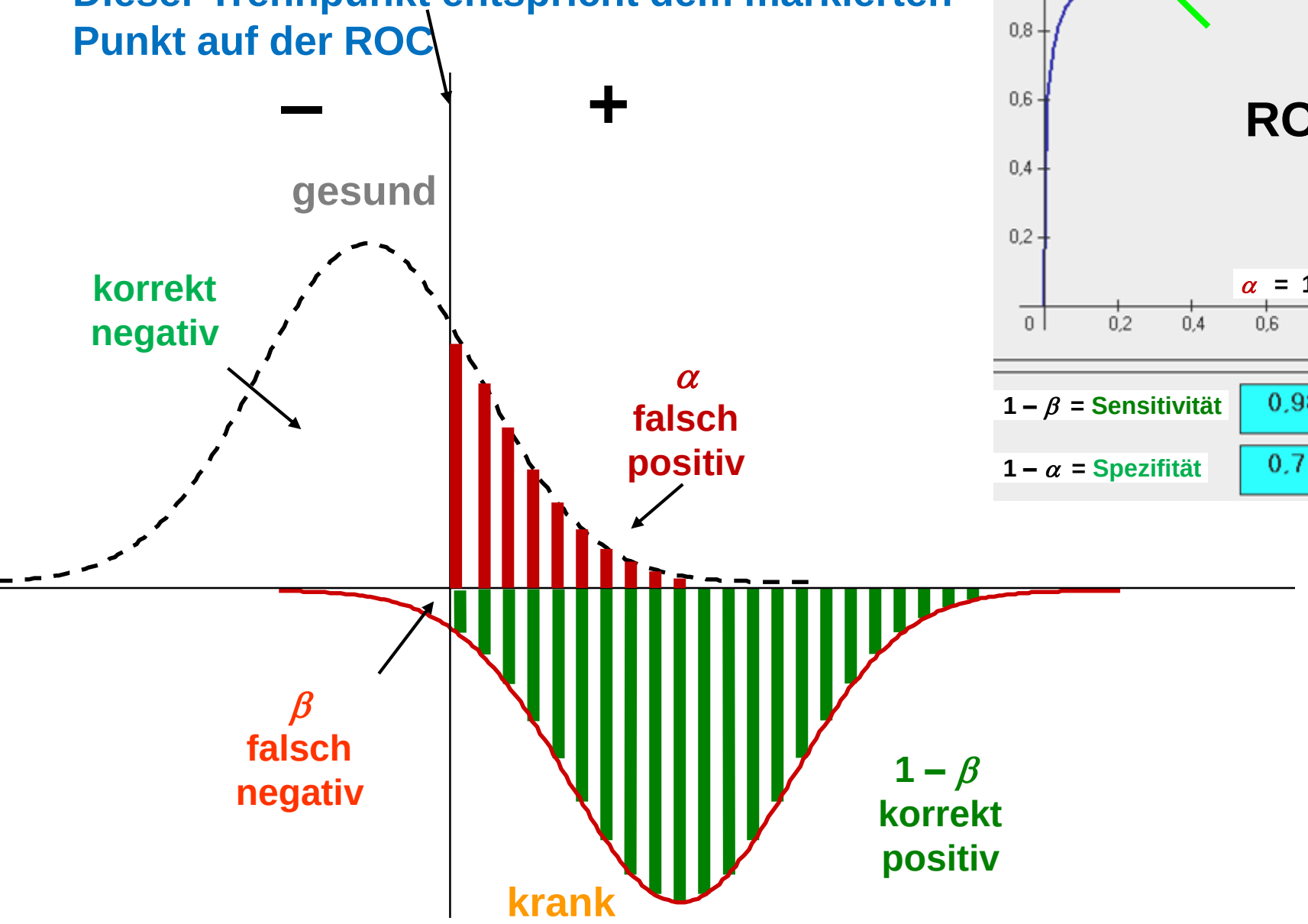
Dieser Trennpunkt entspricht dem markierten Punkt auf der ROC



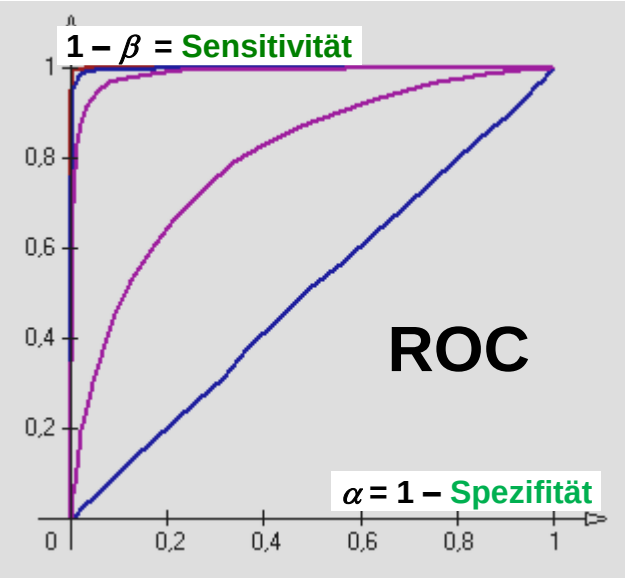
Dieser Trennpunkt entspricht dem markierten Punkt auf der ROC



Dieser Trennpunkt entspricht dem markierten Punkt auf der ROC



Verschiedene Krankheiten haben verschiedene Verteilungen und ROCs



gesund

Je steiler die ROC, umso besser.
Eine diagnostische Prozedur sollte einem Punkt weit oben links entsprechen.

krank

2.4 Einige Schlüsse aus der Analogie zur Medizin

Der *p*-Wert ist nur schwer sinnvoll zu interpretieren.

Diagnose von Krankheiten ist ein Entscheidungsproblem, in welchem man die Verteilungen unter dem Szenario von gesund & krank vergleicht.

Es sind immer zwei divergierende Fehler im Spiel:

- Diagnose der Krankheit, wenn die Person gesund ist.
- Nicht-Erkennen der Krankheit, obwohl die Person diese tatsächlich hat.

Verschiedene Trennpunkte zur Trennung von gesund und krank bedingen unterschiedliche Größen dieser Fehler. Es gibt Krankheiten, welche leicht zu diagnostizieren sind.

Es gibt einen 3. Fehler: Ob die Entscheidung gut ist, hängt nicht nur von den Trennpunkten ab, sondern auch von der Prävalenz der Krankheit.

In vielen Fällen erhalten wir nur schlecht interpretierbare Kenngrößen für die Qualität der Entscheidungen.

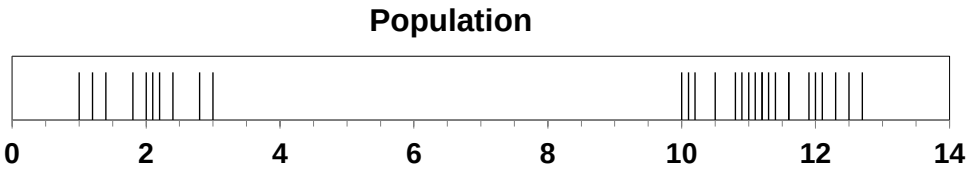
3. Informelle Wege zur statistischen Inferenz

Beispiele

- Zwei verschiedene Methoden zur Messung des Mittelwerts
- Unbekannte Wahrscheinlichkeiten messen – Weg zum Gesetz der Großen Zahlen
- Kleine Risiken, dass einzelne Daten mehr als 2 SD vom Mittelwert abweichen
- Stichprobenverteilung des Mittelwerts
- Single-choice-Prüfungen – Erfolg unter verschiedenen Szenarios
- Lady tasting tea – Einführung in den Signifikanztest
- Trennen von guter und schlechter Qualität – Folge einer Ablehnungszahl
- Statistische Prozesskontrolle als Exploration von Szenarios – Informelle Tests
- Überdeckungswahrscheinlichkeit von Konfidenzintervallen

Explorationen, um Schlüsselbegriffe kennenzulernen; Meta-Wissen jenseits der Mathematik. Reduzieren der Komplexität, wobei man den Weg zur allgemeinen Situation offenhält.

3.1 Zwei Methoden den Mittelwert zu schätzen



Die Werte der Population sind durch einen Strich dargestellt.
Zwei homogene Schichten werden sichtbar.
Wenn eine solche Schichtung bekannt ist, ist es ratsam, das bei der Rekrutierung der Stichprobe zu berücksichtigen.

- Zufällig: 6 Elemente aus allen, Vernachlässigen der Schichten
- Zufällig 2 aus Schicht 1 (kleine Werte) und 4 aus Schicht 2.

Für beide Methoden:

- Mittel der simulierten Daten ≈ Mittel der Population (unverzerrt).
- Schichtenmethode (Methode 2) liefert viel präzisere Ergebnisse.

Die Verbesserung der Schätzung durch die Schichtung im Vergleich zur einfachen Zufallsstichprobe ist stabil in der Wiederholung.

Sampling from the population

Population

Nr.	Strata	Data		Rd. Nr.	Rank
1	1	1,0	0,9	0,2362	18
2	1	1,2	0,9	0,3845	12
3	1	1,4	0,9	0,1215	26
4	1	1,8	0,9	0,2328	19
5	1	2,0	0,9	0,1242	25
6	1	2,1	0,9	0,6044	10
7	1	2,2	0,9	0,1980	22
8	1	2,4	0,9	0,8156	3
9	1	2,8	0,9	0,2492	16
10	1	3,0	0,9	0,2969	15
11	2	10,0	0,9	0,0758	27
12	2	10,1	0,9	0,0144	30
13	2	10,2	0,9	0,6751	7
14	2	10,5	0,9	0,2228	20
15	2	10,8	0,9	0,3073	14
16	2	10,9	0,9	0,7538	4
17	2	11,0	0,9	0,3248	13
18	2	11,2	0,9	0,2489	17
19	2	11,4	0,9	0,2027	21
20	2	11,6	0,9	0,0145	29
21	2	11,1	0,9	0,7024	6
22	2	11,2	0,9	0,4284	11
23	2	11,3	0,9	0,9400	2
24	2	11,6	0,9	0,9574	1
25	2	11,9	0,9	0,1869	23
26	2	12,0	0,9	0,6579	9
27	2	12,1	0,9	0,7246	5
28	2	12,3	0,9	0,1400	24
29	2	12,5	0,9	0,0260	28
30	2	12,7	0,9	0,6746	8

1. Sample neglecting strata structure of population

Size: 6

Sample

Rank	Nr.	Data
1	24	11,6
2	23	11,3
3	8	2,4
4	16	10,9
5	27	12,1
6	21	11,1

Mean	9,90
St. Dev.	3,698

Parameter of the population

Mean	8,21
St. Dev.	4,457

Task

To estimate the mean of the population by the mean of a sample

The method is judged by the quality of the estimate in repeated samples.

1000 samples are generated

To investigate the properties of the estimation.

Sampling by drawing without replacement

Sampling within strata

Population

Nr.	Strata	Data	Rd. Nr.	Rank
1	1	1,0	0,6516	2
2	1	1,2	0,5796	3
3	1	1,4	0,2428	7
4	1	1,8	0,3945	6
5	1	2,0	0,5015	4
6	1	2,1	0,1815	8
7	1	2,2	0,4864	5
8	1	2,4	0,1764	9
9	1	2,8	0,0709	10
10	1	3,0	0,8578	1
11	2	10,0	0,5762	11
12	2	10,1	0,5037	12
13	2	10,2	0,8726	3
14	2	10,5	0,6001	10
15	2	10,8	0,7414	7
16	2	10,9	0,4921	13
17	2	11,0	0,2603	17
18	2	11,2	0,9282	1
19	2	11,4	0,8277	5
20	2	11,6	0,2755	16
21	2	11,1	0,2816	15
22	2	11,2	0,3335	14
23	2	11,3	0,7034	9
24	2	11,6	0,0229	19
25	2	11,9	0,7978	6
26	2	12,0	0,0122	20
27	2	12,1	0,1997	18
28	2	12,3	0,8321	4
29	2	12,5	0,7215	8
30	2	12,7	0,9118	2

1. Sample within strata of population

Size	Stratum 1	Stratum 2	**Sample**
6	2	4	
	Rank	Nr.	Data
Stratum 1	1	10	3,0
	2	1	1,0
Stratum 2	1	18	11,2
	2	30	12,7
	3	13	10,2
	4	28	12,3
Mean			8,40
St. Dev.			5,073

Parameter of the population

Mean	8,21
St. Dev.	4,457

Task

To estimate the mean of the population by the mean of a sample

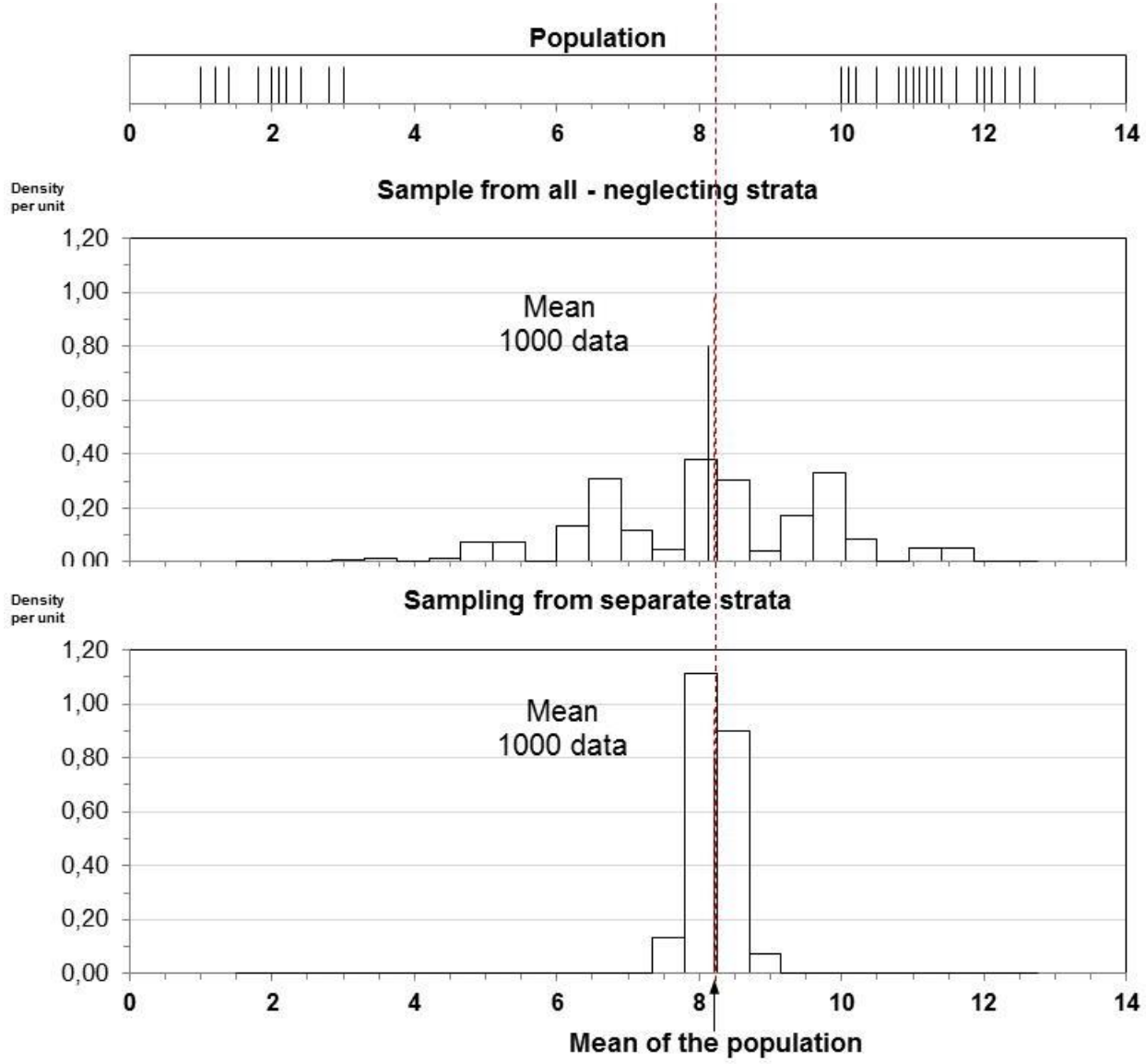
The method is judged by the quality of the estimate in repeated samples.

1000 samples are generated

To investigate the properties of the estimation.

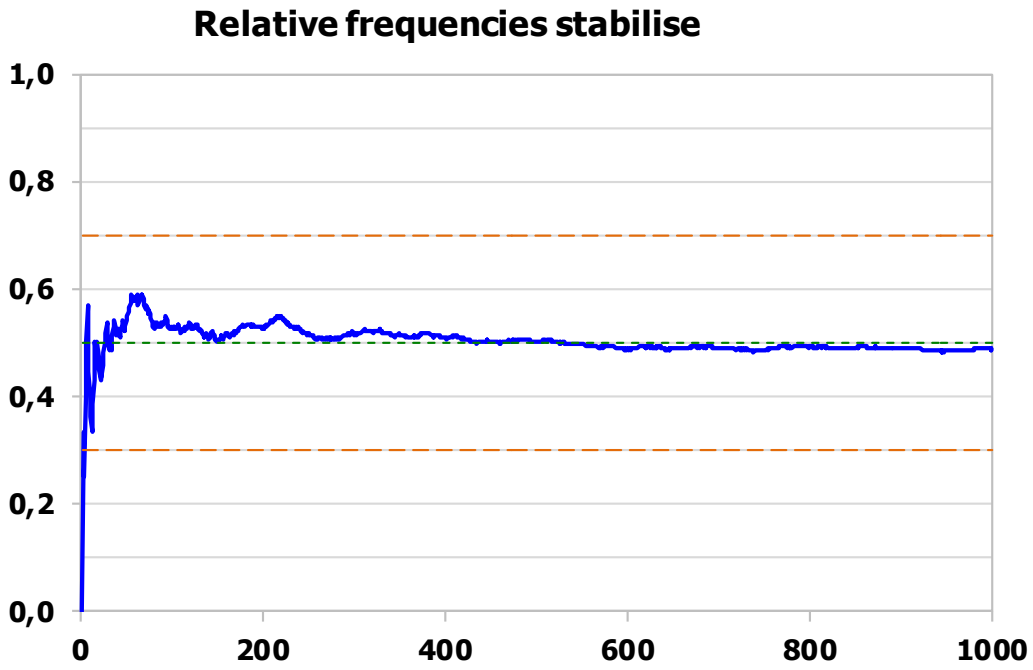
Sampling by drawing without replacement

- **Beide Methoden treffen den Parameter im Durchschnitt.**
- **Die Schichtenmethode liefert viel präzisere Ergebnisse.**



3.2 Messen unbekannter Wahrscheinlichkeiten – Weg zu Gesetz der Großen Zahlen

Coin tossing: Investigate the development of the relative frequency of heads



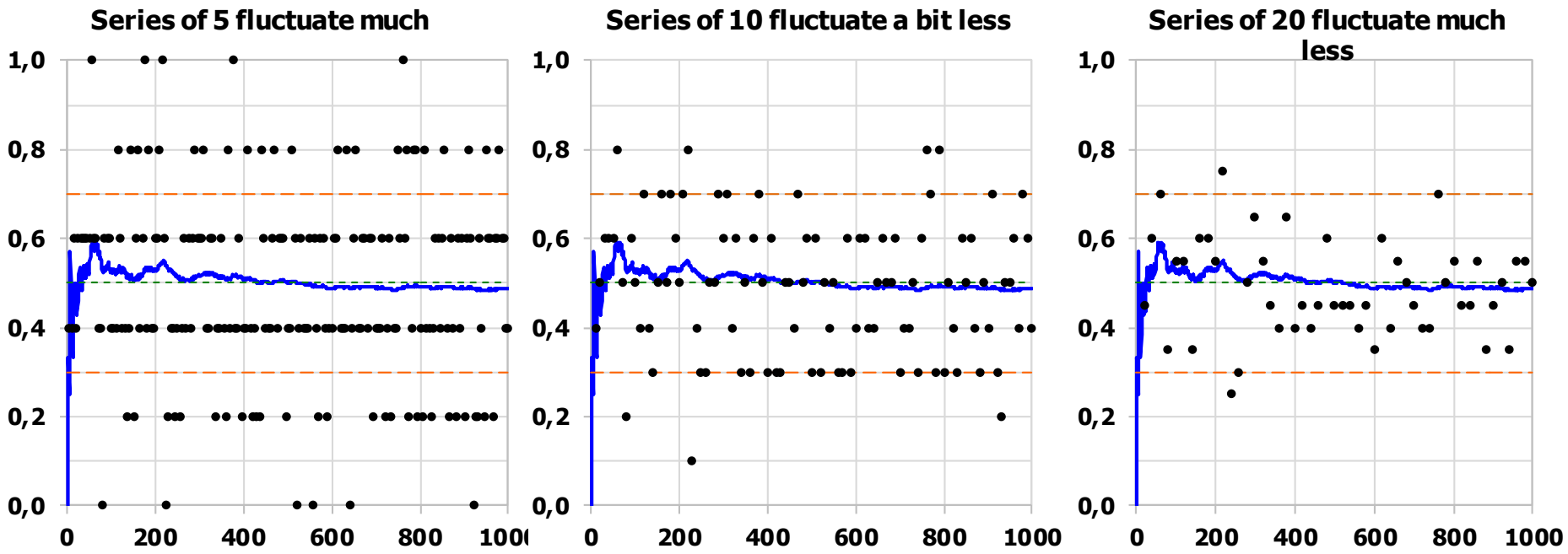
n	Measure of p
987	0,488
988	0,489
989	0,488
990	0,489
991	0,488
992	0,489
993	0,488
994	0,489
995	0,488
996	0,488
997	0,488
998	0,488
999	0,487
1000	0,488

The current series can now more fluctuate because of the weight of close to 1000 values.
The limiting curve suggests a great precision of less than 0.5 percentage points of fluctuation.
Yet, a new experiment (F9!) shows another curve with another "limiting point" within +/- 4 % points.

The law of large numbers: relative frequencies "converge" to the unknown probability

- The convergence hides: the current results are still completely prone to randomness
- What about measuring the unknown probability by short series
  and investigate the precision of such a measurement

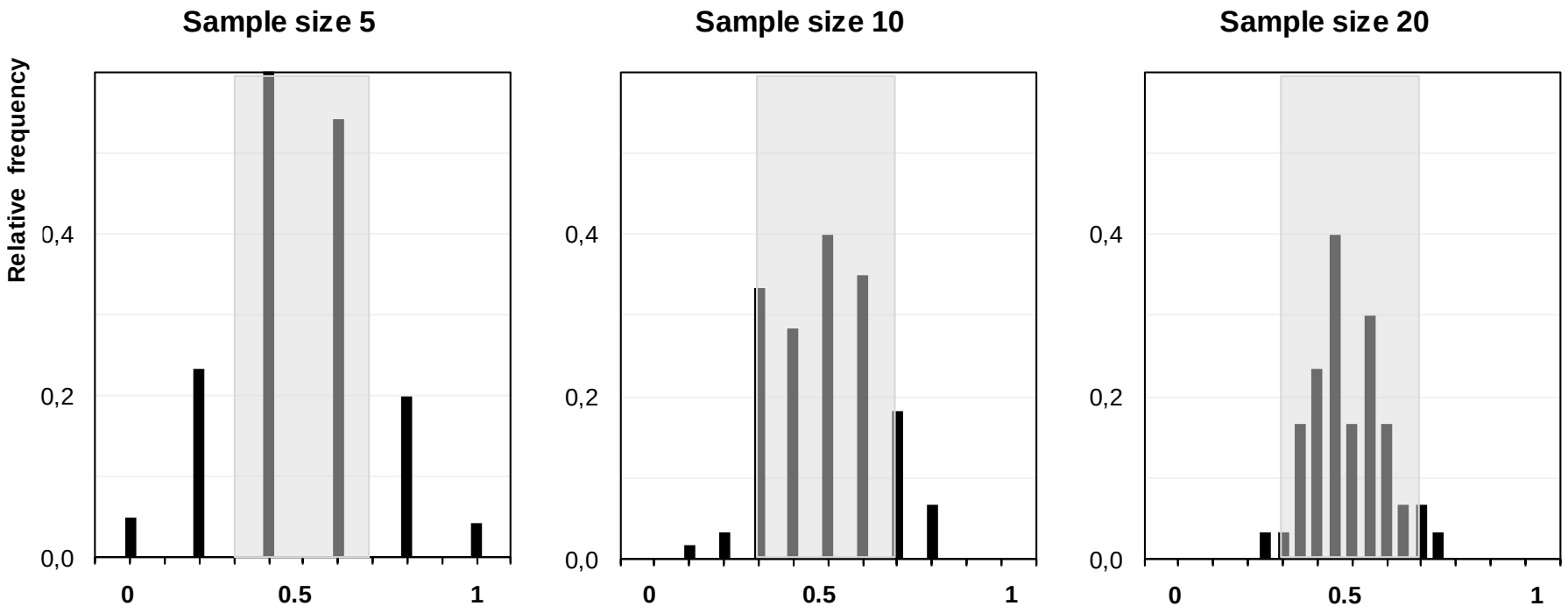
Messen einer unbekannten Wahrscheinlichkeit – Untersuchung der Präzision



We see: despite the convergence, the current series of 5 (10, 20) still vary!
This variation decreases if the series of measurement values is longer.

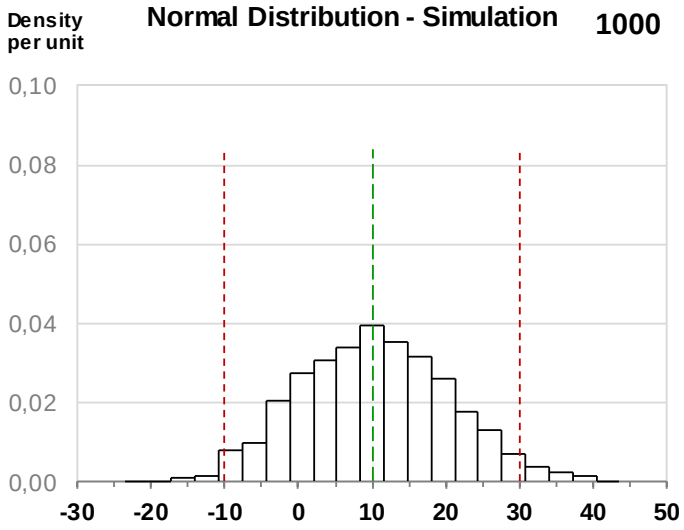
Longer series provide more precise measurements of the unknown probability.
The risk of larger errors decreases with the length of the series. Less values exceed the boundaries.

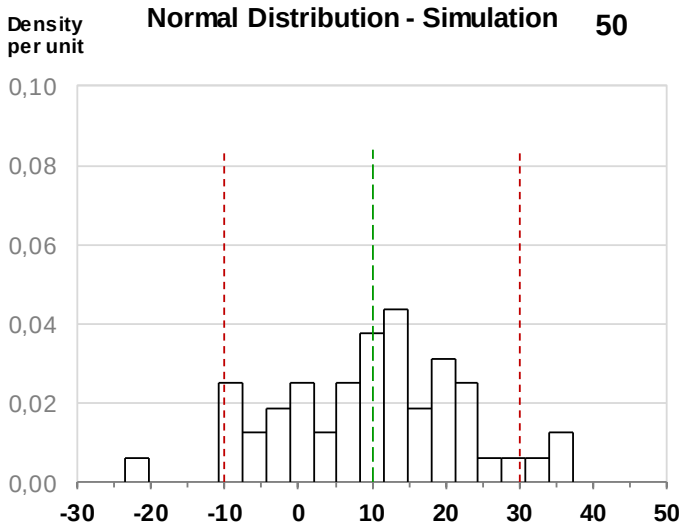
Untersuchung der Verteilung der Messfehler



We see the impact of the length of the series. Longer series provide a more reliable measuring "device".

3.3 Kleine Risiken, das *ein* Datum weiter als 2 SD vom Mittel abweicht





First:

See the great variability in the shape of a normal sample!

Second:

Sample size	Percentage of data within $\mu - 2\sigma$ and $\mu + 2\sigma$:
1000	95,3%
50	88,0%

In conclusion:

about 95% of the data lie within 2 standard deviations from the mean; regardless of the values of the parameters.

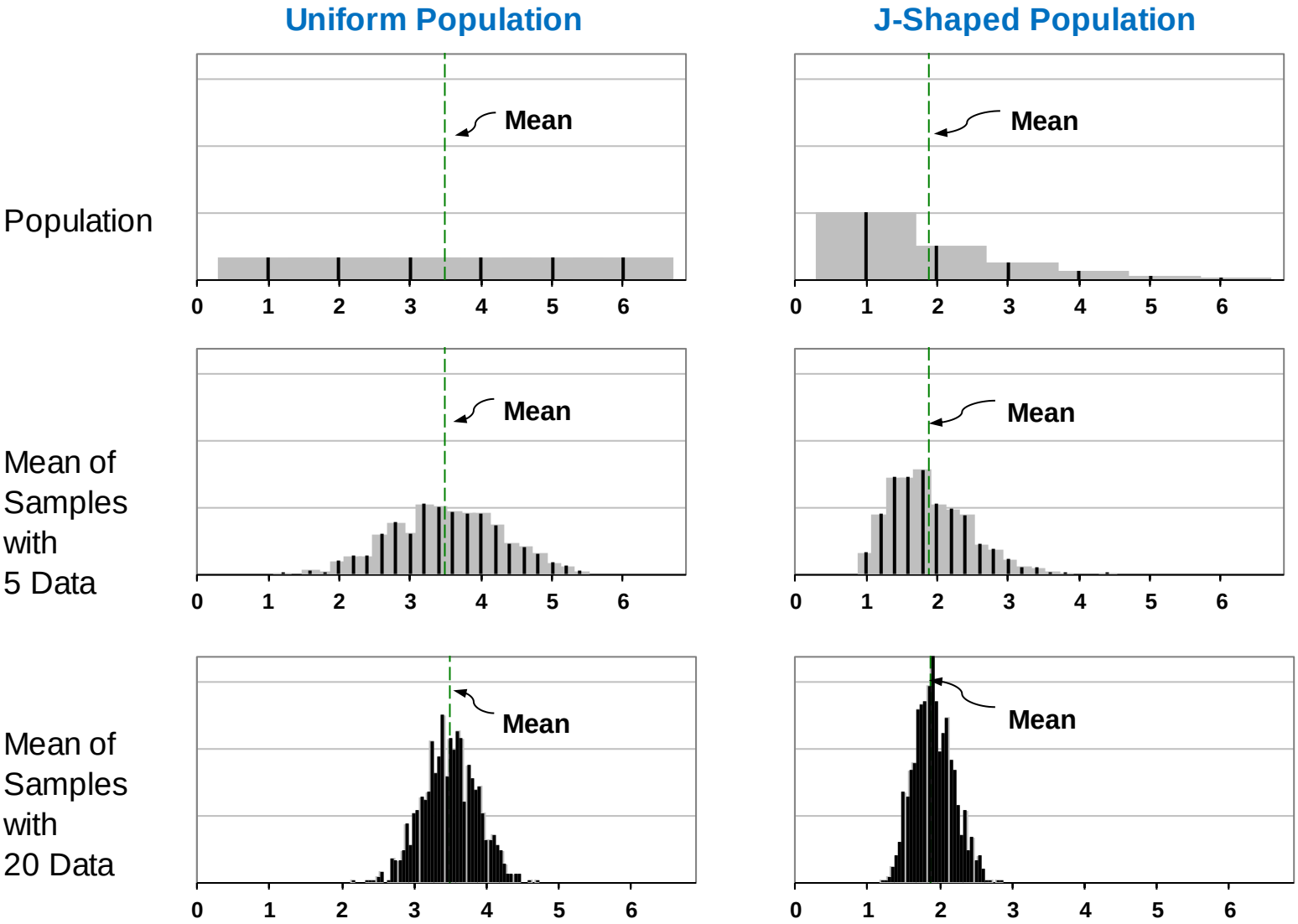
If sigma is small, then *one* piece of data can indicate the location of the population mean!

There is a small risk that the distance is larger than two standard deviations.

Third:

For mean values, the normal distribution applies approximately!

3.4 Die Stichprobenverteilung des Mittels ist artifiziell



The larger the sample, the smaller the variability of the distribution of the mean.

..., the closer the mean of an arbitrary random sample will be to the mean of the population.

… the better the normal shape of the distribution of the mean of samples.

3.5 Single-choice-Prüfungen – Erfolg unter verschiedenen Szenarios

Binomial distribution to model the number of correctly solved items

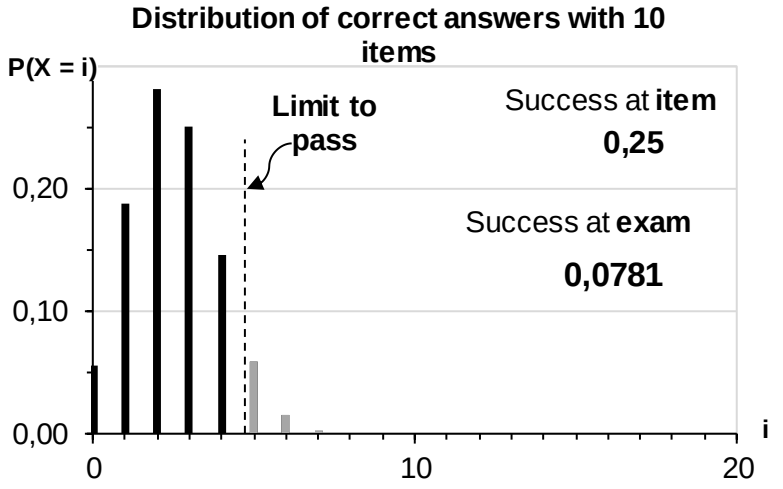
No. Items	Success
n	p
10	0,25

n Number of Items (questions)

p Probability of success (answer the question correctly)

X_i i-th item, how is it solved; 1 correct; 0 wrong

$P(X_i = 1) = p$ $P(X_i = 0) = 1-p$



Assumptions

Independence of items; i.e., the probability to answer items correctly, is independent of answering the other items; e. g.,

$P(X_i = 1 \text{ and } X_j = 1) = p * p$

We model answering the items as if we drew balls out of an urn

N Number of balls in the urn

p Proportion of marked (= 1 = correct) in the urn

n we draw n balls with replacement

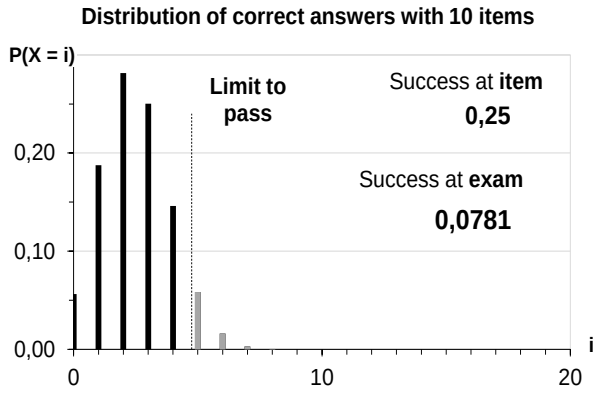
X_i = 1 or 0; depending on ball we draw.

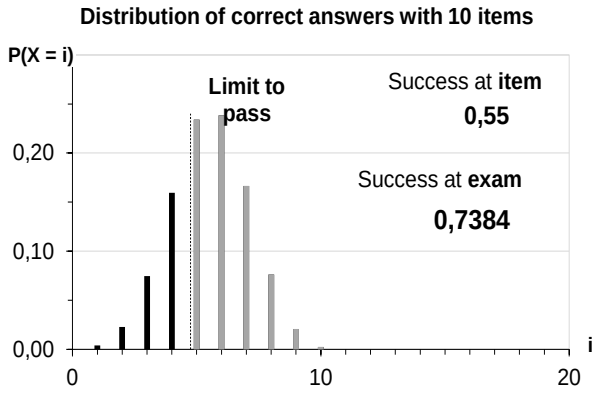
X ΣX_i = number of marked balls = number of correct answers.

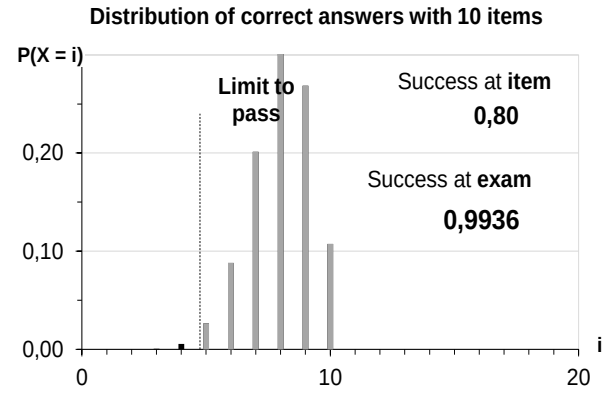
X ~ B(n, p) , i. e., is binomially distributed

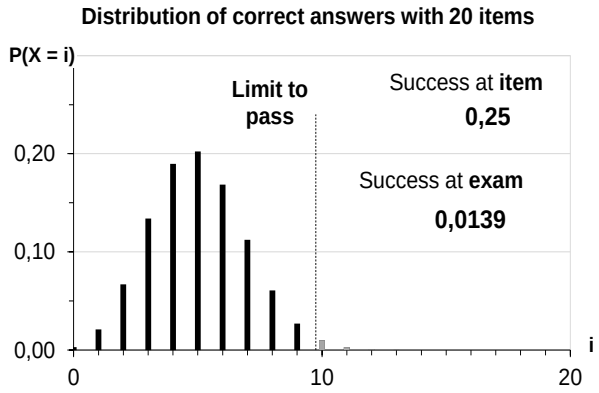
Plausibility of assumptions

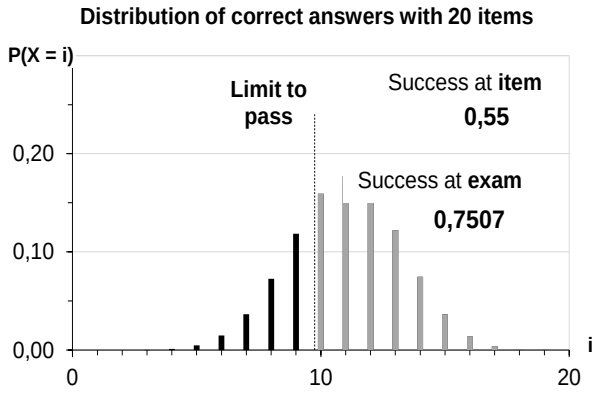
It makes sense to model the number of correctly solved items under the assumption of guessing.
For a learner who has learned relatively much, and should be modelled by a solving capacity p,
the assumptions of the binomial distribution are not really met.
Neither the same solution probability for each item, nor the independence of solving different items.
Yet, as a scenario, the binomial distribution might give insight how to tune the parameters in a single-choice exam.

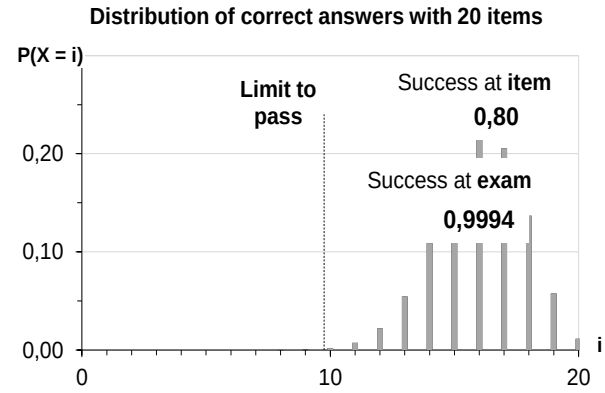

Distribution of correct answers with 10 items
P(X = i)
Limit to pass
Success at item
0,25
Success at exam
0,0781
0,20
0,10
0,00
0
10
20
i

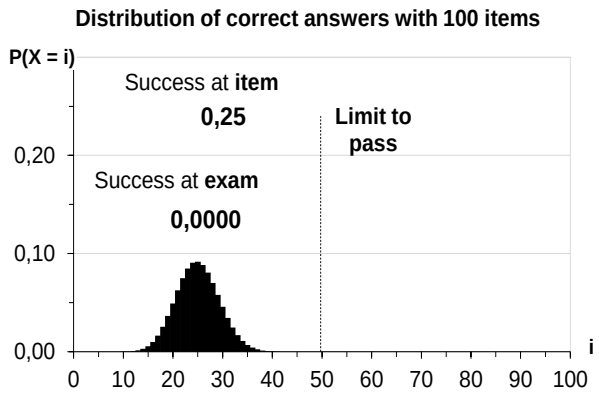

Distribution of correct answers with 10 items
P(X = i)
Limit to pass
Success at item
0,55
Success at exam
0,7384
0,20
0,10
0,00
0
10
20
i

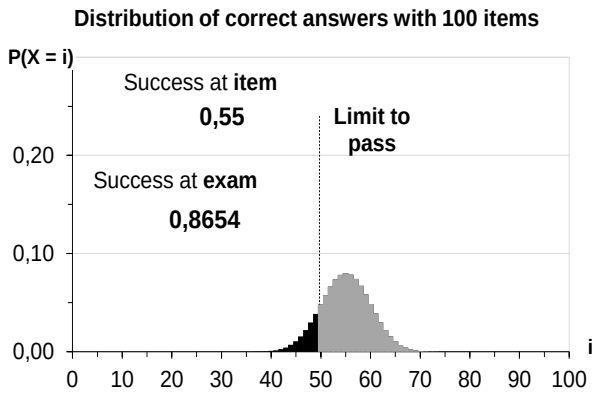

Distribution of correct answers with 10 items
P(X = i)
Limit to pass
Success at item
0,80
Success at exam
0,9936
0,20
0,10
0,00
0
10
20
i

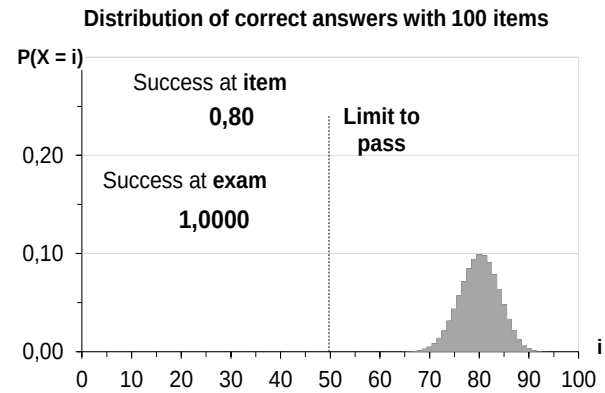

Distribution of correct answers with 20 items
P(X = i)
Limit to pass
Success at item
0,25
Success at exam
0,0139
0,20
0,10
0,00
0
10
20
i


Distribution of correct answers with 20 items
P(X = i)
Limit to pass
Success at item
0,55
Success at exam
0,7507
0,20
0,10
0,00
0
10
20
i


Distribution of correct answers with 20 items
P(X = i)
Limit to pass
Success at item
0,80
Success at exam
0,9994
0,20
0,10
0,00
0
10
20
i


Distribution of correct answers with 100 items
P(X = i)
Success at item
0,25
Limit to pass
Success at exam
0,0000
0,20
0,10
0,00
0
10
20
30
40
50
60
70
80
90
100
i


Distribution of correct answers with 100 items
P(X = i)
Success at item
0,55
Limit to pass
Success at exam
0,8654
0,20
0,10
0,00
0
10
20
30
40
50
60
70
80
90
100
i


Distribution of correct answers with 100 items
P(X = i)
Success at item
0,80
Limit to pass
Success at exam
1,0000
0,20
0,10
0,00
0
10
20
30
40
50
60
70
80
90
100
i

3.6 Lady tasting tea – Einführung in den Signifikanztest

Eine Dame behauptet, dass sie durch Schmecken erkennen kann, ob Tee oder Milch zuerst in die Tasse gefüllt wurde.

Wie können wir ein Experiment organisieren, um dies zu prüfen?

Experiment mit 8 Tassen, T und M in zufälliger Reihenfolgen (nicht bekannt, wie viele in welcher Reihenfolge) Nullhypothese: Einfach raten. Ws. für Erfolg (Erraten der Reihenfolge T & M): 1/2 wie beim Münzwerfen.

Wir untersuchen das Experiment

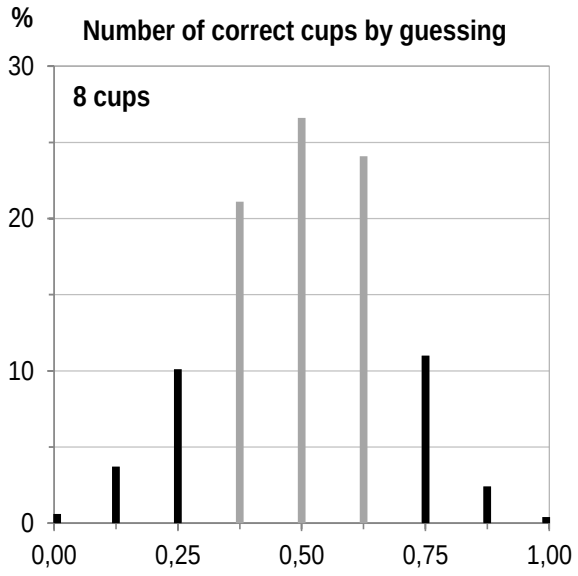
- Wir nutzen die Binomialverteilung (n=8, p=0.5).
- Oder wir simulieren Werfen einer Münze 8 Mal; und wiederholen dieses Experiment 1000 Mal und schätzen die Ws. aus den relativen Häufigkeiten dieses Szenarios.

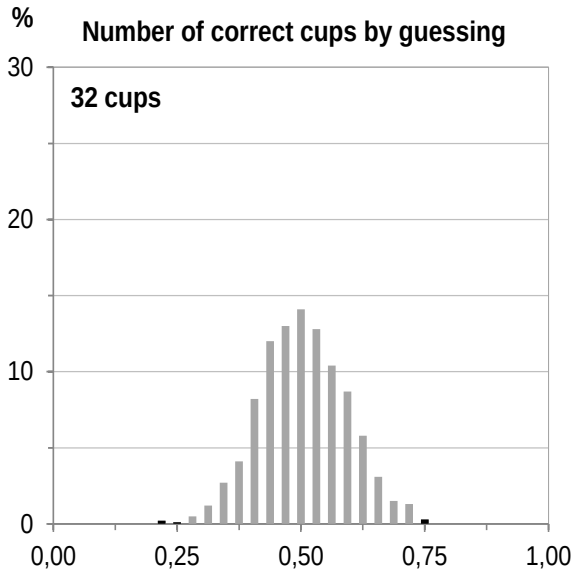
Wenn wir 6 von 8 Tassen korrekt klassifiziert (75% Erfolg) als Leistung ansehen, können wir nun sehen, dass dies ziemlich leicht ist; die Ws. ist 14.5%.

Bei 32 Tassen ist eine Erfolgsrate von 75% und mehr sehr selten: 0.4%.

Tea-tasting (8 cups)

No. correct	Proport. correct	abs. frequ.	rel. frequ.
0	0,006	6	0,6
1	0,125	37	3,7
2	0,250	101	10,1
3	0,375	211	21,1
4	0,500	266	26,6
5	0,625	241	24,1
6	0,750	110	11,0
7	0,875	24	2,4
8	0,994	4	0,4
Total		1000	100





R. A. Fisher arrangierte das Experiment etwas anders.
Er ließ in 4 Tassen zuerst Tee dann Milch eingießen und dann 4 mit Milch zuerst. Er ordnete die Tassen zufällig an, etwa: TMTT MMTM.

Seine Überlegung beruhte auf einem Re-randomisierungsargument: Wenn jede Permutation gleich ws. ist – dies entspricht einfach Raten – dann ist eine von 70 Permutationen korrekt, was bedeutet, es hat eine Ws. von 1/70 = 1.4%; und 6 oder mehr korrekt entsprechen 4 Mal 1.4 = 5.7% (es kann nur 8 oder 6 etc. richtige in diesem Arrangement geben).

Fisher begründete den Signifikanztest durch eine Re-randomisierung.

3.7 Separieren guter & schlechter Qualität – Folgen einer Ablehnungszahl

Good quality: Process under control
Represented by 4% defectives

Bad quality: Process out of control
Represented by 10% defectives

We simulate rather than apply the binomial distribution

Null hypothesis 100 B(100, 0.04) **reject if ≥ 8**

N. of defects	Absolute frequency	Relative frequency		
0	68	1,36%	1,36	0,00
1	348	6,96%	6,96	0,00
2	732	14,64%	14,64	0,00
3	991	19,82%	19,82	0,00
4	1015	20,30%	20,30	0,00
5	814	16,28%	16,28	0,00
6	510	10,20%	10,20	0,00
7	296	5,92%	5,92	0,00
8	128	2,56%	2,56	2,56
9	63	1,26%	1,26	1,26
10	25	0,50%	0,50	0,50
11	9	0,18%	0,18	0,18
20	0	0,00%	0,00	0,00
21	0	0,00%	0,00	0,00
	5000	100,00%		

Alternative hypothesis 100 B(100, 0.1) **not reject if ≤ 7**

N. of defects	Absolute frequency	Relative Frequency		
0	0	0,00%	0,00%	0,00
1	1	0,02%	-0,02%	-0,02
2	5	0,10%	-0,10%	-0,10
3	30	0,60%	-0,60%	-0,60
4	79	1,58%	-1,58%	-1,58
5	175	3,50%	-3,50%	-3,50
6	311	6,22%	-6,22%	-6,22
7	436	8,72%	-8,72%	-8,72
8	597	11,94%	-11,94%	-11,94
9	678	13,56%	-13,56%	-13,56
10	639	12,78%	-12,78%	-12,78
11	539	10,78%	-10,78%	-10,78
20	5	0,10%	-0,10%	-0,10
21	1	0,02%	-0,02%	-0,02
	5000		-20,74	

Fehler beim Separieren der Hypothesen guter & schlechter Qualität

Small sample n = 100

Two types of errors: reject the null even if it applies: not reject the null even if the alternative hypothesis applies

Rejection number currently **8**

Type I error (α) **4,52%**

Type II error (β) **20,74%**

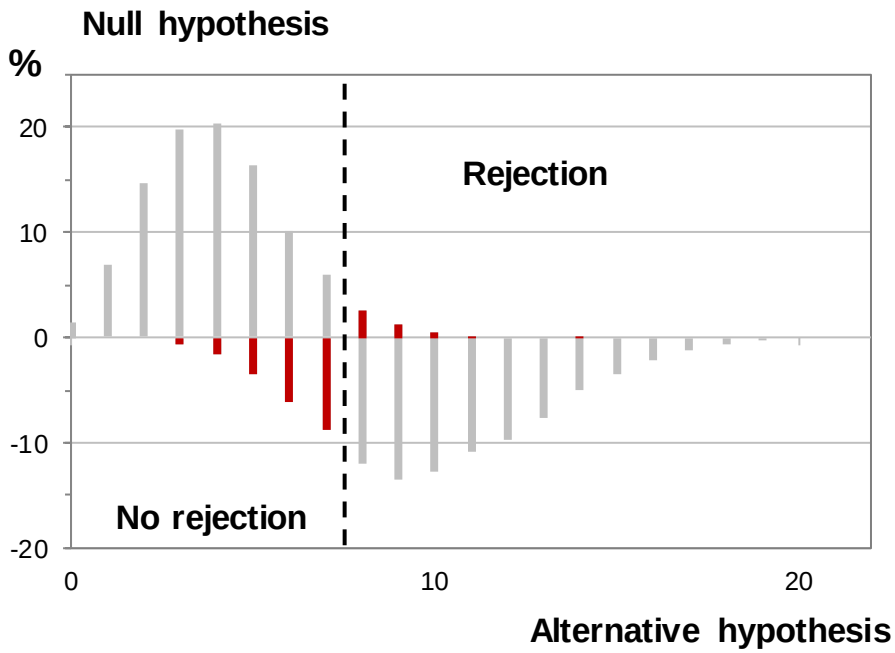
Large sample n = 400

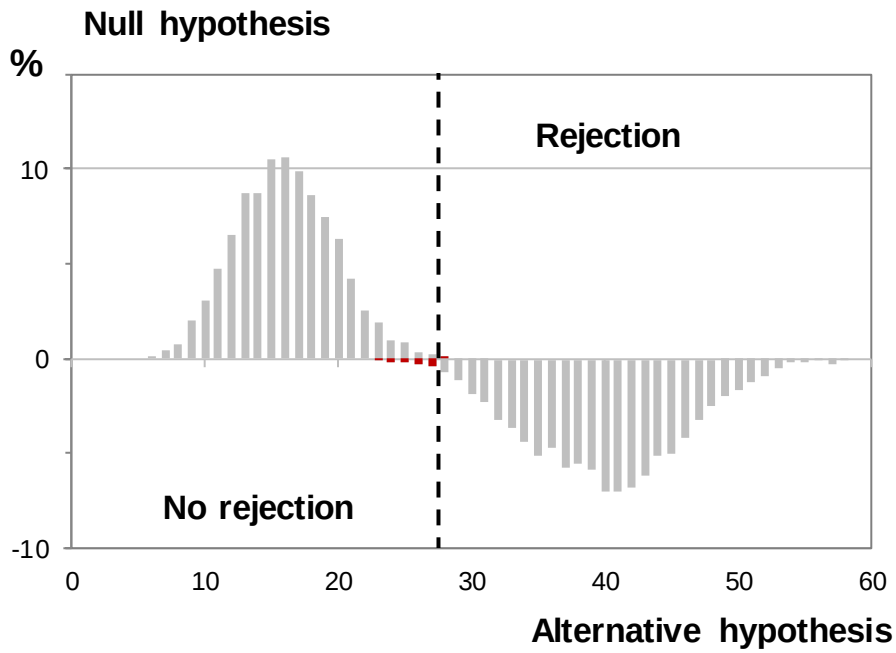
Rejection number current **28**

Type I error (α) **0,18%**

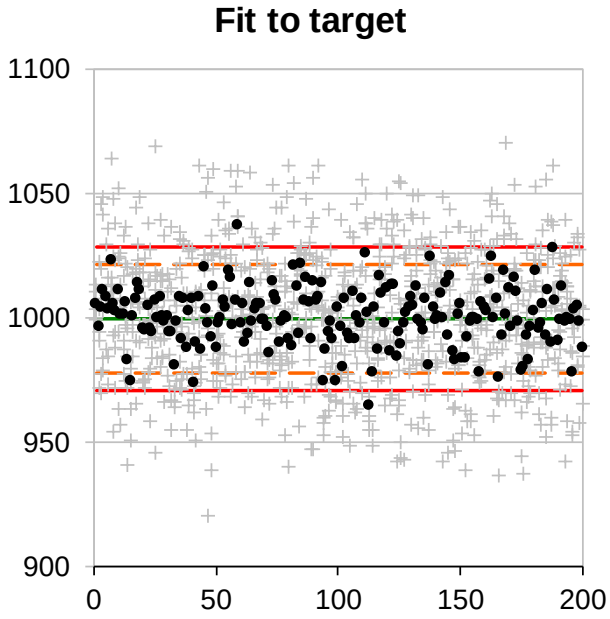
Type II error (β) **1,16%**

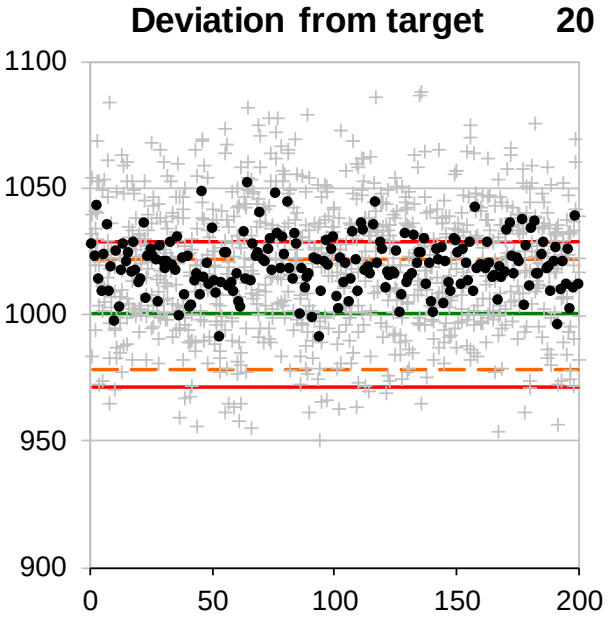
The rejection number may be shifted to see that the **errors are developing in the opposite direction.**
The choice of the rejection number balances the diverging interests.

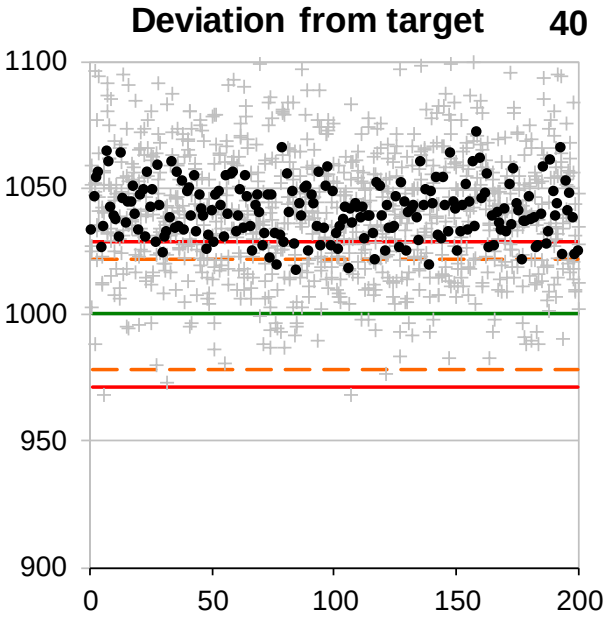




3.8 Statistische Prozesskontrolle: Exploration von Szenarios – Informelle Tests







	Scenario 1	Scenario 2	Scenario 3
Deviation from target	0	20	40
SD of process	25	25	25
No of Samples	200	200	200
Within limits W, no W alarm	187	124	6
Within limits C, no C alarm	198	166	32
W alarms rate	0,065	0,380	0,970
C alarm rate	0,010	0,170	0,840

In this scenario, alarms are false alarms. ..., alarms are correct. ..., alarms are correct.

wrong
correct

False alarm rate = alpha error

False rate of missing = beta error

False rate of missing = beta error

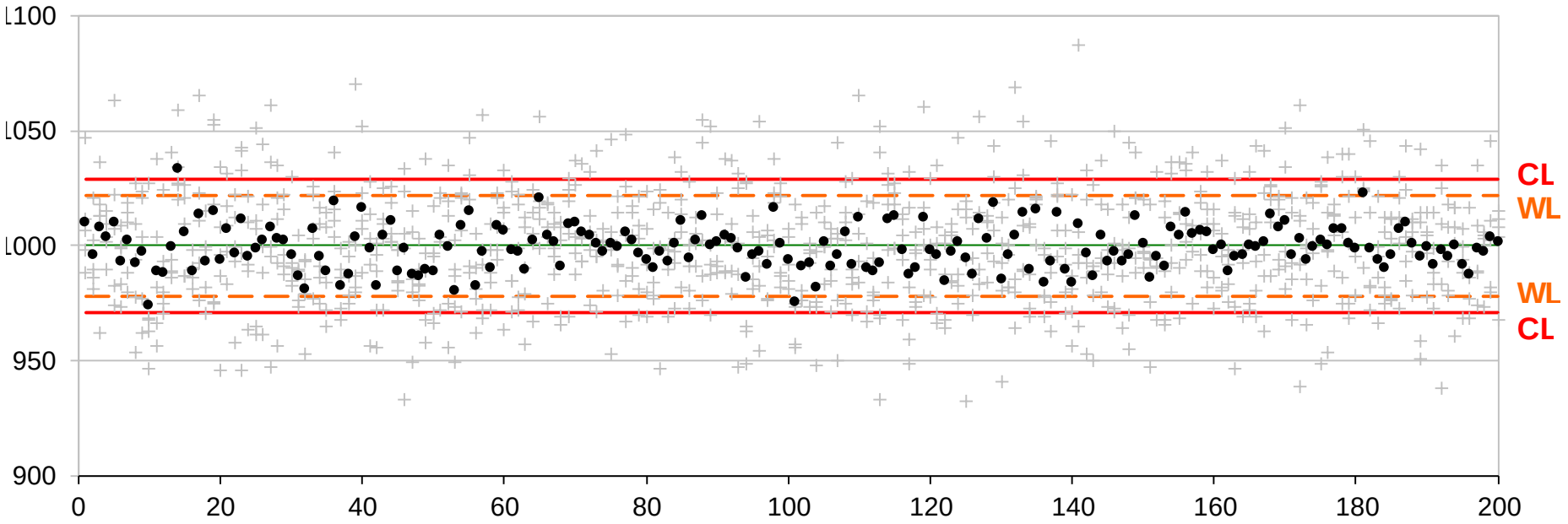
Produktion unter regulären Bedingungen – Alles ist unter Kontrolle

At each control time, inspection involves 5 single items; 200 inspections are simulated (a month).
It is checked how the prescribed control (CL) and warn limits (WL) behave.

Chart for mean values of 5 items	Control Limits (CL)	Warning limits (WL)
Number of samples	200	200
Samples within the limits, no alarm	199	196
Proportion of samples within the limits, no alarm	0,995	0,980
Sample rate outside limits, alarm	0,005	0,020

wrong
correct

Control chart for mean values of samples of 5 items



Produktion unter Abweichungen von regulären Bedingungen – Abweichung im Mittel von 20

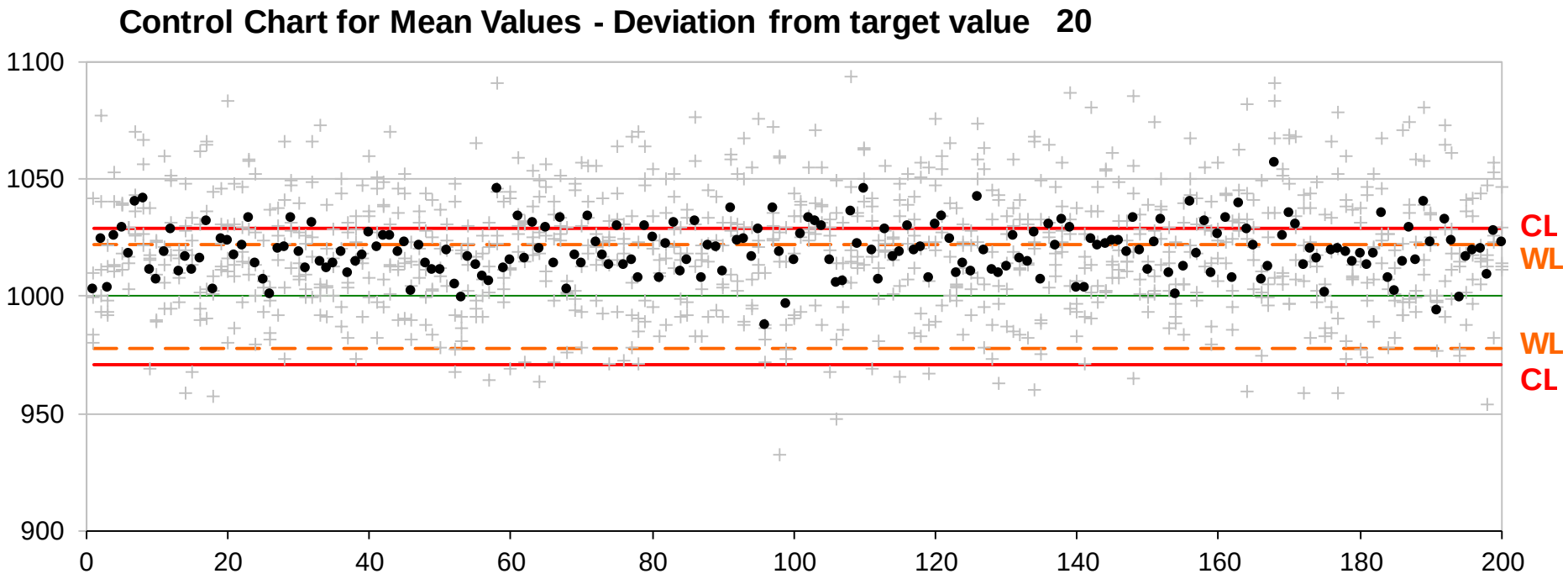
It is checked how the prescribed control (CL) and warn limits (WL) behave.

Chart for mean values of 5 items

	Control Limits (CL)	Warning limits (WL)
Mean values within		
Number of samples	200	200
Samples within limits	154	122
Sample rate within limits, no alarm	0,770	0,610
Sample rate outside limits, alarm	0,230	0,390

wrong

correct



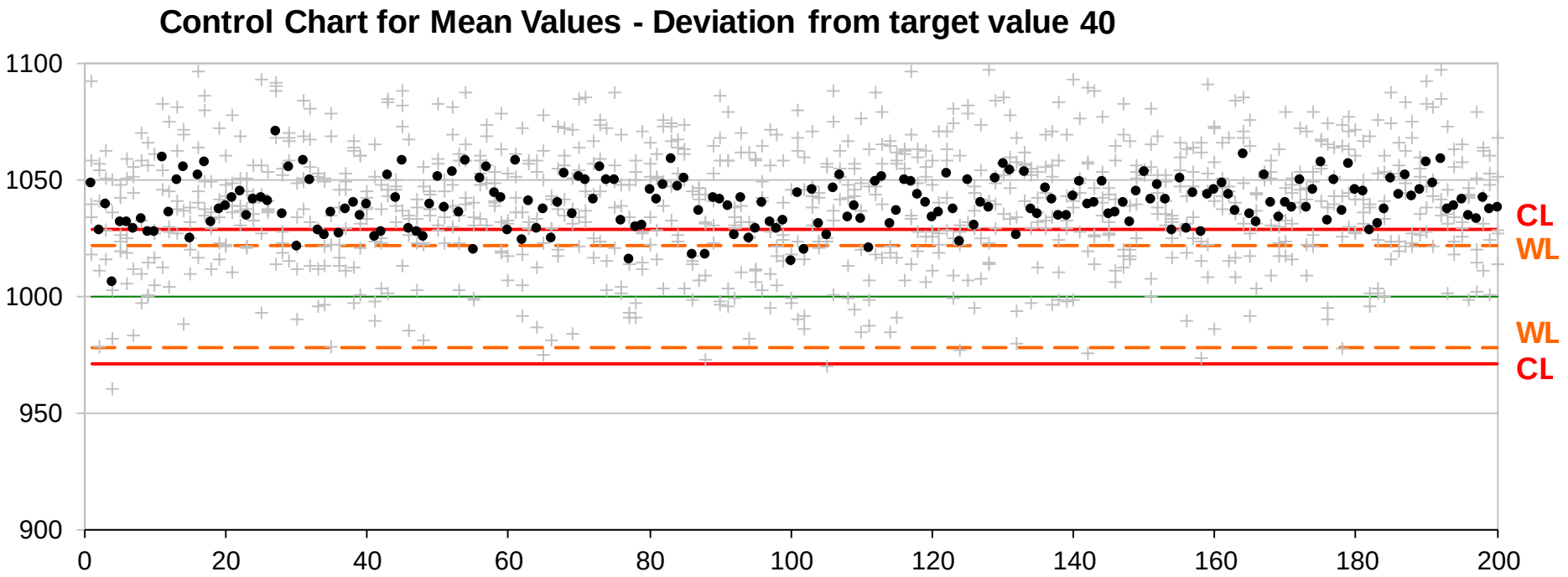
Produktion unter Abweichungen von regulären Bedingungen – Abweichung im Mittel von 40

It is checked how the prescribed control (CL) and warn limits (WL) behave.

Chart for mean values of 5 items

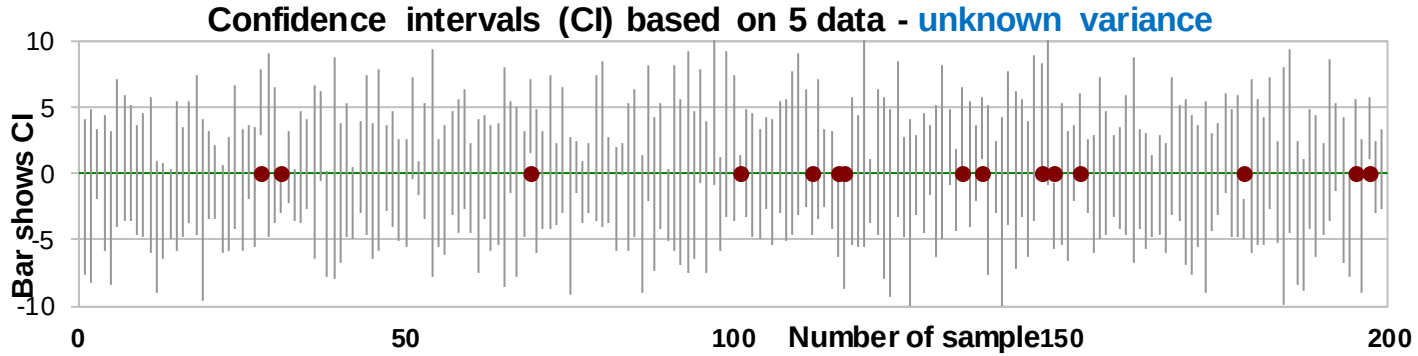
Mean values within	Control Limits (CL)	Warning limits (WL)
Number of samples	200	200
Samples within limits	30	9
Sample rate within limits, no alarm	0,150	0,045
Sample rate outside limits, alarm	0,850	0,955

wrong
correct



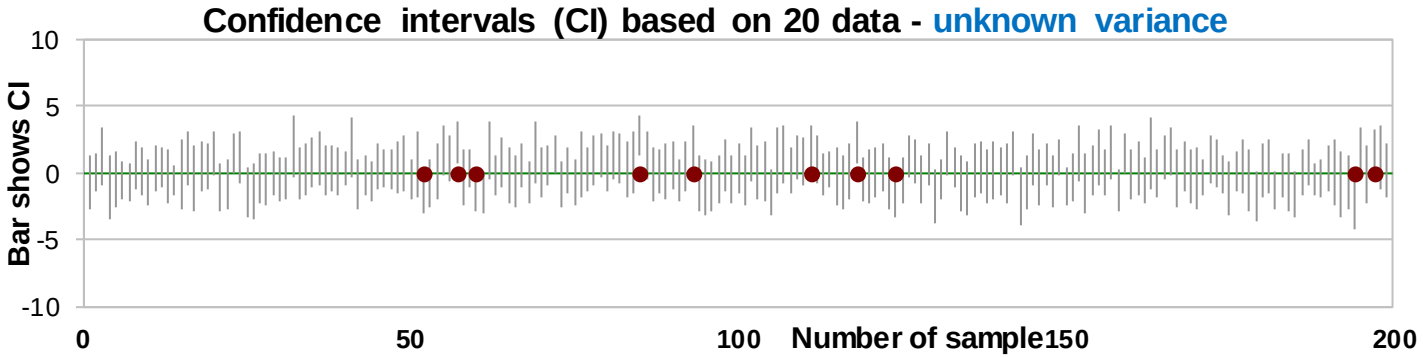
3.9 Überdeckungswahrscheinlichkeit von Konfidenzintervallen

The confidence level (coefficient) relates to the coverage of the unknown parameter if samples are repeated.
Model for the population (single measurement) $X \sim N(0, \sigma^2 = 16)$; the variance is unknown.
Repeatedly samples are drawn; i. with 5 data; ii. with 20 data;
for each sample, a confidence interval is calculated for the expected value (which is "known" in the laboratory).
Bars show the repeated confidence intervals. If μ (0 in our scenario) lies outside, this will be marked by a red dot on the axis.



% of covering intervals:

ALL	SHORT
92,5%	80,6%
MIDDLE	**LONG**
97,0%	100,0%



ALL	SHORT
95,0%	86,6%
MIDDLE	**LONG**
100,0%	98,5%

The present sample may have an SD above/below average with no clear interpretation of the confidence level.

4. „Informal Inference“ – Eine vereinfachte Inferenz

Inferenz basiert einzig auf vorhandenen Daten – keine theoretische Verteilung wird unterstellt. Wenn Hypothesen verwendet werden, sind sie “natürlich”.

- **Schätzung** ↓ Bootstrap-Stichproben, um den Fehler zu schätzen

  Stichproben aus den Daten mit Zurücklegen.

  Anstatt Stichproben aus der wahren Verteilungsfunktion *F* zu nehmen, nimmt man sie aus der Schätzung von *F*, die aus der ersten Stichprobe resultiert.

  Bootstrap liefert approximative Intervalle und kann und muss durch andere Methoden korrigiert werden (BCa, ABC).

- **Hypothesentesten** ↓ Randomisierungstests

  Re-randomisierung der Zuordnung zu Gruppen werden verglichen.
  Permutationen der Daten oder Stichproben aus den Daten ohne Zurücklegen.

  Liefert exakte nonparametrische Tests in einzelnen Fällen.

4.1 Bootstrap-Interval & klassisches Konfidenzintervall

Für das Mittel einer Population
Gegeben: eine Stichprobe vom Umfang n mit Mittel und SD für eine Variable
Wie präzise ist das Mittel aus der Stichprobe als Messung für die Population?
Time = Zeitaufwand für ein Seminar.

Raw data

Nr	Times
1	12
2	2
3	6
4	2
5	19
6	5
7	34
8	4
9	1
10	4

n	Mean
10	8,90

1. Bootstrap

Nr	Times
10	4
1	12
8	4
9	1
2	2
9	1
7	34
5	19
1	12
8	4

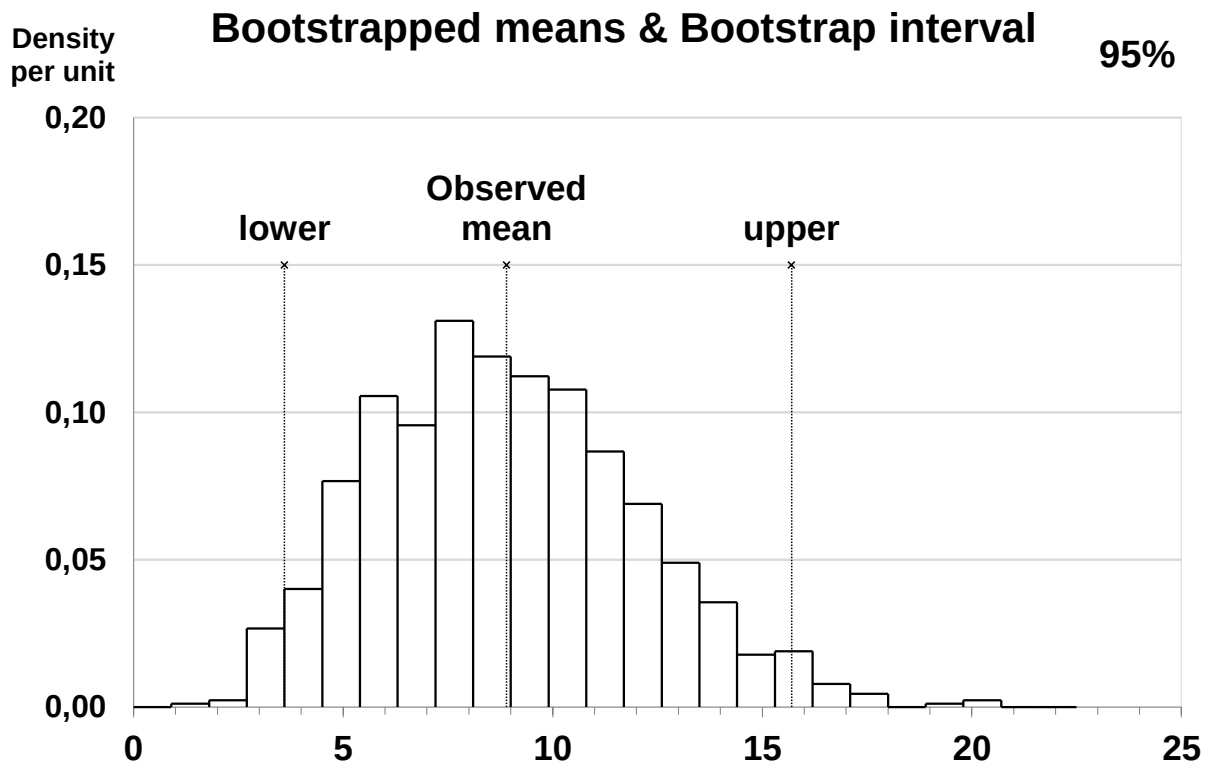
Mean
9,30

Statt wieder aus der Population zu ziehen, was ja unmöglich ist, nehmen wir die Stichprobe aus der ersten Stichprobe (mit Zurücklegen).

Der erste Bootstrap liefert eine neue Messung des Mittels der Population.

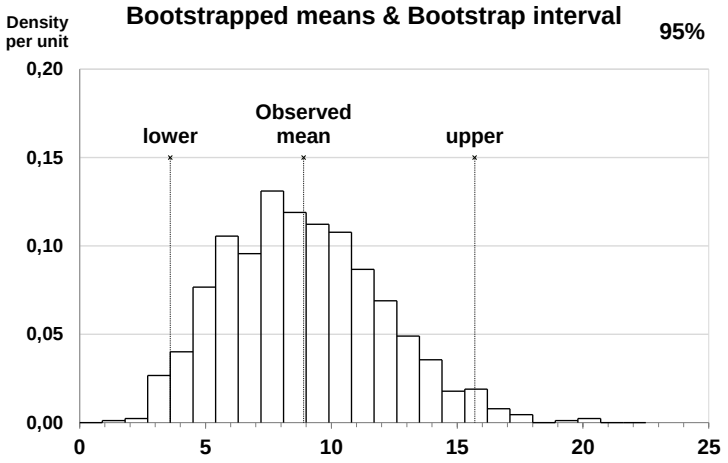
Wir wiederholen den Bootstrap und erhalten 1000 (oder mehr) artifizielle Messungen.
Wir analysieren diese mit den üblichen Methoden.

Bootstrap-Verteilung für das Mittel



Das ergibt ein Bild der Präzision der Messung des Mittels basierend auf der ersten Stichprobe. Wir vergleichen die numerischen Resultate mit dem klassischen Konfidenzintervall.

Bootstrap-Verteilung für das Mittel



95% Bootstrap interval	3,60	15,70

contains ex. 95% of Bootstrapped means

The Bootstrap interval reflects the precision of measurements of the mean of the population

95% Confidence interval	2,46	15,34

contains population mean in 95% of "repeated" samples

Wir sehen eine gute Übereinstimmung beider Methoden.

Natürlich ist die Interpretation ganz unterschiedlich.

4.2 Bootstrap-Intervall für andere Parameter

Der Vorteil des Bootstraps ist, dass die Strategie immer dieselbe bleibt, egal, um welchen Parameter es sich handelt. Wir zeigen das für die Korrelation. Gegeben: eine Stichprobe vom Umfang n Paare von Daten. Wie präzise ist die Korrelation der Stichprobe als Messung für die ganze Population?

Raw data

Nr	Ca	M
1	105	1247
2	17	1668
3	5	1466
4	14	1800
5	18	1609
6	10	1558
7	15	1807
8	78	1299
9	10	1637
10	84	1359
11	73	1392
12	12	1755
13	78	1307

n	r
13	-0,851

1. Bootstrap

Nr	Ca	M
13	78	1307
10	84	1359
11	73	1392
8	78	1299
12	12	1755
7	15	1807
13	78	1307
9	10	1637
3	5	1466
11	73	1392
10	84	1359
6	10	1558
3	5	1466

r
-0,782

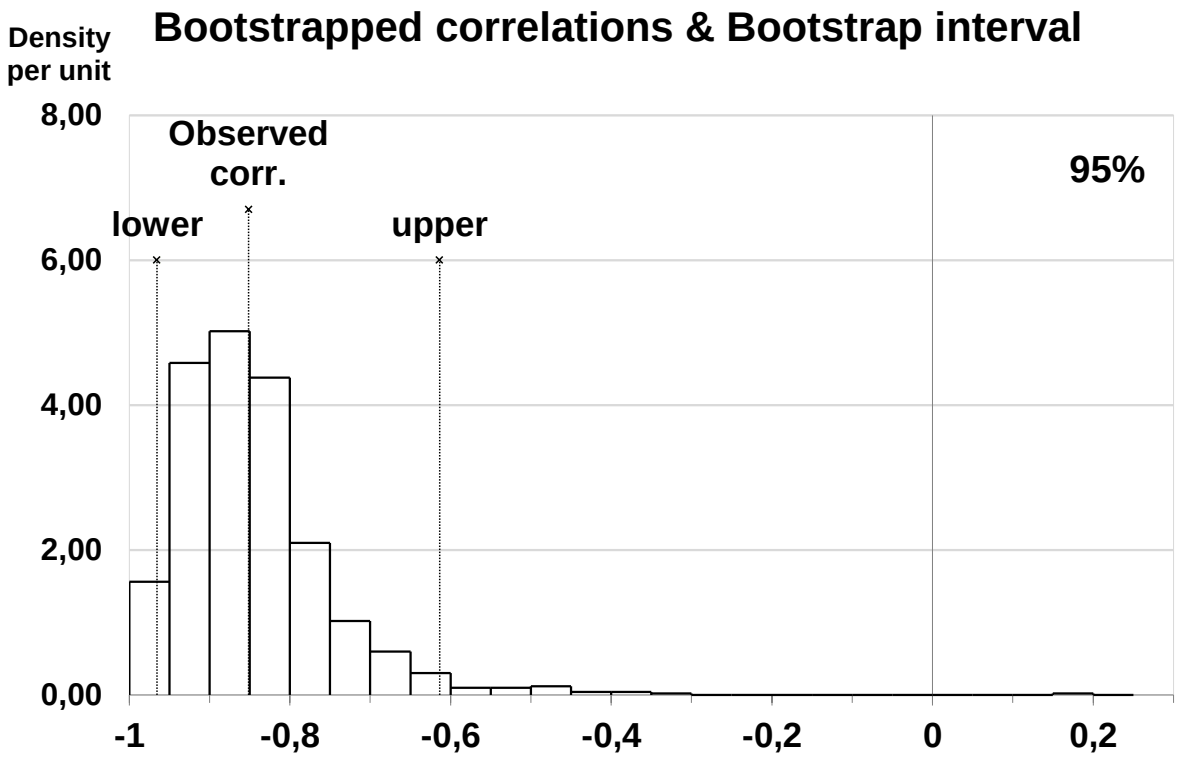
Statt wieder von der Population zu sampeln, nehmen wir die Stichprobe aus der ersten Stichprobe (mit Zurücklegen).

Der erste Bootstrap liefert eine neue Messung der Korrelation der Population.

Beachte, dass Paare von Daten resampelt wurden.

Wir wiederholen den Bootstrap und erhalten 1000 (oder mehr) artifizielle Messungen. Wir analysieren diese mit den üblichen Methoden.

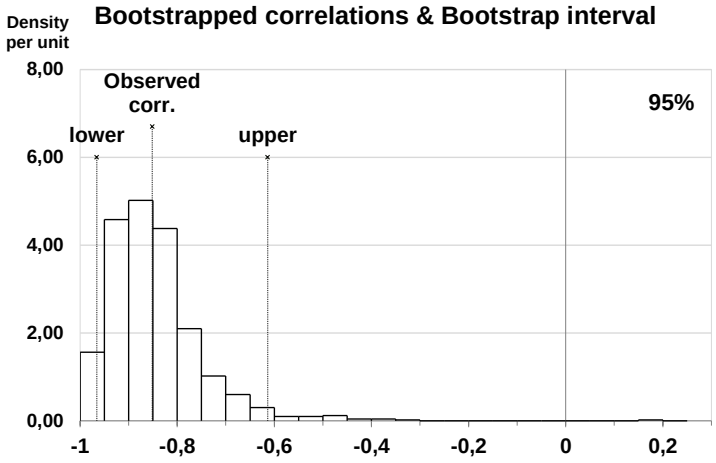
Bootstrap-Verteilung für die Korrelation



Das ergibt ein Bild der Präzision der Messung der Korrelation basierend auf der ersten Stichprobe.

Wir vergleichen die numerischen Resultate mit dem klassischen Konfidenz-intervall.

Bootstrap-Verteilung für die Korrelation



95% Bootstrap interval	-0,97	-0,61

contains ex. 95% of Bootstrapped correlations

The Bootstrap interval reflects the precision of measurements of the corr. of the population

95% Confidence interval	-0,95	-0,57

contains population corr. in 95% of "repeated" samples

Wieder sehen wir eine gute Übereinstimmung beider Methoden. Doch, die Interpretation unterscheidet sich sehr.

Die Methode des Untersuchens der wiederholten Messungen der Korrelation durch artifizielle Messungen, die durch Ziehen aus der ersten Stichprobe entstehen, ist sehr intuitiv.

Was sonst kann man tun, um Hypothesen über die Population zu vermeiden?

4.3 Re-Randomisierungstest für die Differenz von Mitteln

Ist die Behandlung wirksam (im Hinblick auf eine Zielvariable) ? Treatment group TG bekommt VERUM – control group CG bekommt Placebo. Re-Randomisierung als Alternative zum Zwei-Stichproben-Test.

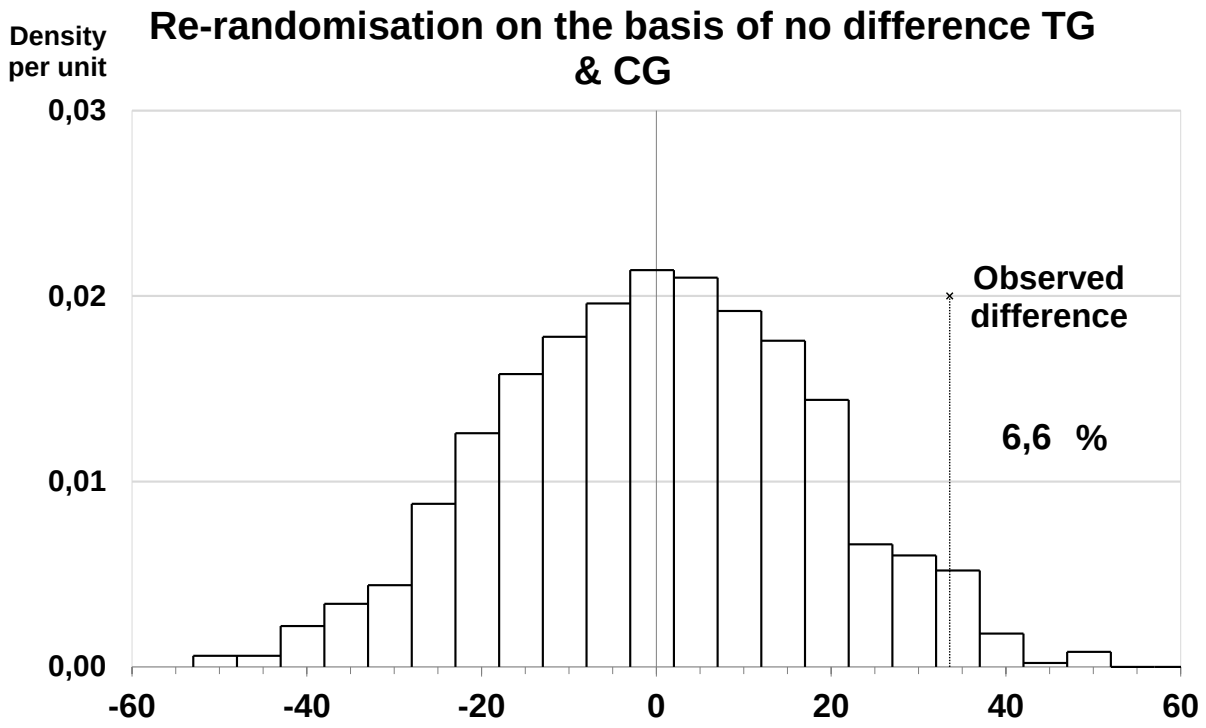
Raw data			1. Rerandomisation			
			Random		Nr	E
1	69,0	Treatment TG	0,48	new TG	6	40,0
2	24,0		0,74		5	77,5
3	63,0		0,17		11	-7,5
4	87,5		0,39		7	9,0
5	77,5		0,26		10	77,5
6	40,0		0,36		8	12,0
7	9,0	Control CG	0,78	new CG	4	87,5
8	12,0		0,36		9	36,0
9	36,0		0,99		1	69,0
10	77,5		0,98		2	24,0
11	-7,5		0,16		12	32,5
12	32,5		0,81		3	63,0
TG	60,17		Mean		TG	34,75
CG	26,58				CG	52,00
Diff	33,58		Effect		Diff	-17,25

Unter der Nullhypothesis KEINE DIFF ist es intuitiv, dass jede Neu-Zuordnung von Personen zu den Behandlungen KEINEN EFFEKT haben sollte.

Daher permutieren wir die Zahlen der Personen, sodass die nächste Behandlungsgruppe aus den Personen Nr. 6, 5, 11, 7, 10, und 8 besteht.

Die erste Re-Attribution liefert eine neue Messung der Differenz der Mittel (zur Messung des Effekts der Behandlung).

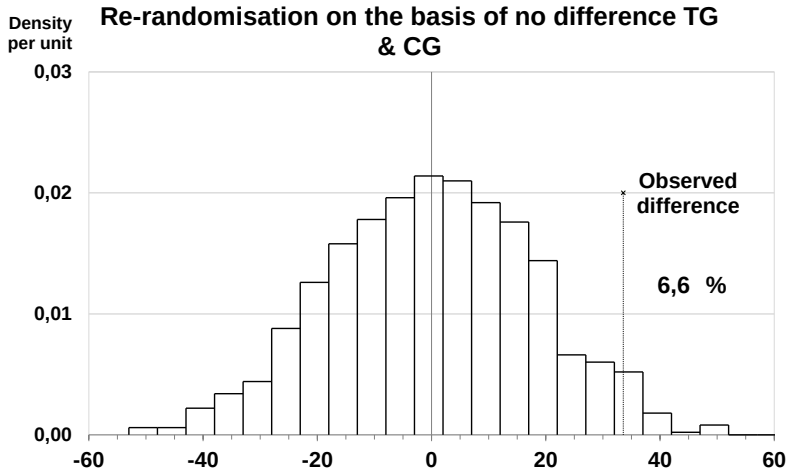
Re-Randomisierung der Differenz der Mittelwerte unter H_0



Die Verteilung für die Mittelwertdifferenz für die wiederholte Re-attribution (durch Zufall) ist im Bild enthalten. Sie liefert artifizielle Ergebnisse basierend auf der Hypothese KEINE DIFFERENZ, d.h., der Nullhypothese.

Wir können das Ergebnis der ersten Stichprobe in diese Verteilung einordnen, um zu unserer Einschätzung zu kommen.

Re-Randomisierungsverteilung unter der Nullhypothese



Is the observed difference (of the means) only due to randomness?

"what if there is no difference between TG & CG?"

lower 2.5%	upper 2.5%	observed	p value %
-34,25	**35,08**	**33,58**	**6,6**
			two-sided

t test	**Welch test**	
equal	unequal	variances
2,16	2,16	Test statistic
5,6%	**5,9%**	**p value**

Wieder ergibt sich eine gute Übereinstimmung mit beiden Methoden.

Die Re-Randomisierung ist überzeugend für die Bedingungen unter der Nullhypothese.

Aber es gibt keine Möglichkeit, eine Alternativhypothese über die Differenz im Mittel durch eine Re-Randomisierung zu repräsentieren, weil wir dazu ja keine Daten haben.

4.4 Re-Randomisierungstest für die Korrelation

Eine Stichprobe für 2 Variable ergibt einen Korrelationskoeffizienten von -0,85. Können wir behaupten, dass der beobachtete Wert signifikant von Null verschieden ist?

Raw data

Nr	Ca	M
1	105	1247
2	17	1668
3	5	1466
4	14	1800
5	18	1609
6	10	1558
7	15	1807
8	78	1299
9	10	1637
10	84	1359
11	73	1392
12	12	1755
13	78	1307

n	r
13	-0,851

1. Re-sampling based on Corr = 0

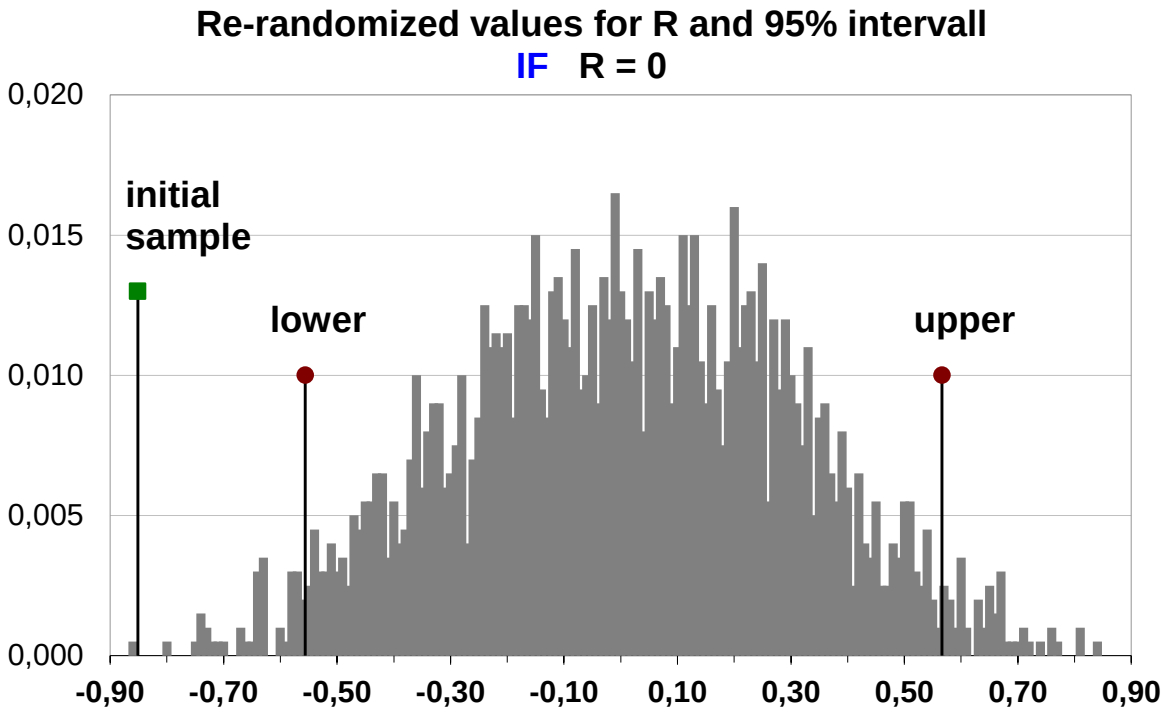
Rand	Rank	Ca	M
0,689	4	105	1800
0,489	8	17	1299
0,919	2	5	1668
0,223	13	14	1307
0,541	5	18	1609
0,303	12	10	1755
0,377	11	15	1392
0,433	9	78	1637
0,425	10	10	1359
0,504	7	84	1807
0,966	1	73	1247
0,880	3	12	1466
0,522	6	78	1558

r
0,384

Da die Annahme ist, dass KEINE Korrelation (Corr = 0) besteht,
können wir die Werte der zweiten Variablen zur ersten durch eine Permutation neu zuordnen.

Schon die zweite Neu-Zuordnung unter Corr = 0 ergibt eine neue Messung der Korrelation, welche sehr weit weg ist von der ersten Stichprobe ist, was andeutet, dass die Korrelation signifikant ist.

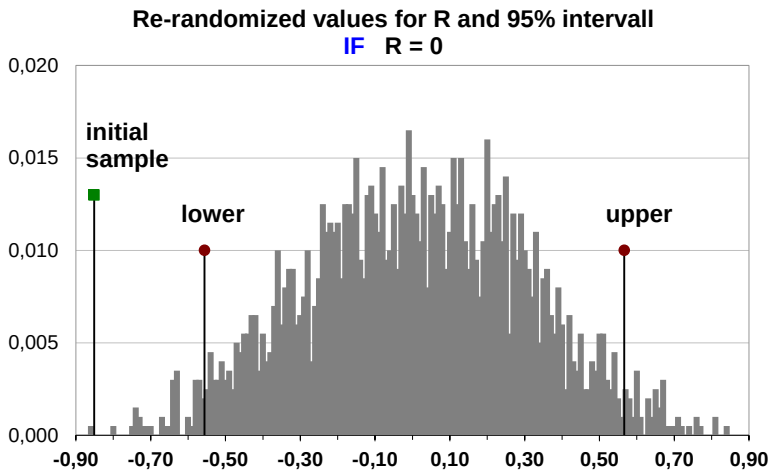
Re-Randomisierung der Korrelationen unter H_0



Die Verteilung der wiederholten Re-randomierung ist im Bild. Sie zeigt die artifiziellen Ergebnisse, die auf der Hypothese von KEINE KORRELATION, d.h., auf der Nullhypothese, basieren.

Wir können das Ergebnis der ursprünglichen Stichprobe in diese Verteilung einordnen, um unsere Einschätzung zu treffen.

Re-Randomisierungs-Verteilung unter der Nullhypothese



Distribution of re-randomised R's IF Corr = 0		
95% Re-random interval	-0,56	0,57
95% Confid. interval	-0,95	-0,57

Die beobachtete Korrelation liegt weit außerhalb des resampelten Intervalls.
Das klassische Konfidenzintervall ist weit weg von der Null.
Mit beiden Methoden: die Korrelation ist signifikant von Null verschieden!

Die Re-Randomisierung ist überzeugend unter den Bedingungen der Null-hypothese.

Aber es gibt keine Möglichkeit, die Situation unter irgendeiner Alterna-tivhypothese über den Wert der Korrelation durch Rerandomisierung zu repräsentieren.

5. Resümee – Vereinfachung oder Reduktion

"Informal inference" geht viel weiter als informelles Explorieren stochastischer Modelle durch Simulation; es zielt darauf ab, traditionelle statistische Inferenz zu ersetzen.

Eine Überlegung zum Ansatz der "Informal Inference"

- Konzepte verstehen $\neq$ Lösen von Aufgaben.
- Wie soll man mit Wahrscheinlichkeit umgehen? Normalverteilung weglassen? Und andere Verteilungen (etwa zur Risikoanalyse)? Wie soll man andere Zugänge und Interpretationen (Bayes) damit vereinbaren, wenn alles durch Simulation erledigt wird.
- Modellieren wird in der Simulation absorbiert. Daten werden zu Fakten, während Modelle einem hypothetischen Denken entsprechen.
- Re-Randomisierung erlaubt nicht, β-Fehler anzusprechen.
- Bootstrap ist intuitiv; aber Korrekturen für die Verzerrtheit der Schätzungen sind komplex. Bootstrap versagt bei kleinen Wsn.
- Wie soll man das Curriculum zu Inferenz / Bayes-Methoden fortsetzen?

Der Zugang engt die Sicht auf Modellieren mit Wahrscheinlichkeiten ein.

Didaktische Fragen, die durch “Informal Inference” auftauchen:

- Die Statistiker nützen immer komplexere Modelle, aber wir haben es noch nicht geschafft, die einfacheren unter ihnen zu unterrichten.
- Wie will man Experten herausfordern, wenn man nur diesen Seitenstrang statistischer Inferenz kennenlernt?
- Soll Statistik für die Sekundarstufe wirklich etwas sein, das nichts mit Statistik an der Universität und in den vielfältigen Anwendungen, die in alle Bereiche des öffentlichen und privaten Lebens hereinspielen, zu tun hat?
- Wollen wir die nächste Generation so ausbilden, dass sie weder die Expertise wertschätzen noch unsachgemäße Anwendungen kritisieren kann?

Informelle Inferenz

- Für exploratorische Wege zum Erforschen der Begriffe in statistischer Inferenz findet man auch Anregungen in Batanero & Borovcnik (2016), oder in meinen statistischen Applets.

ResearchGate: **https://www.researchgate.net/profile/Manfred_Borovcnik**
Feedback bitte an: **manfred.borovcnik@aau.at**