

**Nichtparametrische Statistik:
Resampling-Methoden anhand
eines Experimentes zur Hagelabwehr
und anderen Beispielen**

Werner Stahel, Seminar für Statistik, ETH Zürich

FORMI-Kurs für Gymnasiallehrer

1. September 2004, St. Gallen

1 Einleitung

1.1 Das Grundschemata der parametrischen Statistik

- a Wahrscheinlichkeits-Theorie: **Modell**.
Typischerweise parametrische Familie, z.B. Normalverteilung $\mathcal{N}(\mu, \sigma^2)$.

b **Statistik: Brücke zwischen Modell und Daten.**

Drei Grundfragen der Schliessenden Statistik

[1.] **Welcher Wert ist für den (jeden) Parameter am plausibelsten?**

→ **Schätzung**

[2.] **Ist ein bestimmter Wert plausibel?**

→ **Test.**

[3.] **Welche Werte sind insgesamt plausibel?**

→ **Vertrauens- oder Konfidenzintervall**

1.2 Beispiele

^a Hagelabwehr: Grossversuch IV

Frage: Vermindert „Impfung“ von Gewitterwolken mit AgI die Schäden ?

Methode: Raketen mit AgI, russische Vorschrift

Zielgrösse: Schäden ungeeignet → Ersatzgrösse (Radar-Reflektiv.)
Beobachtungseinheit: Wolke

Versuchsplanung: Vergleich von „behandelten“ und „Kontrolle“

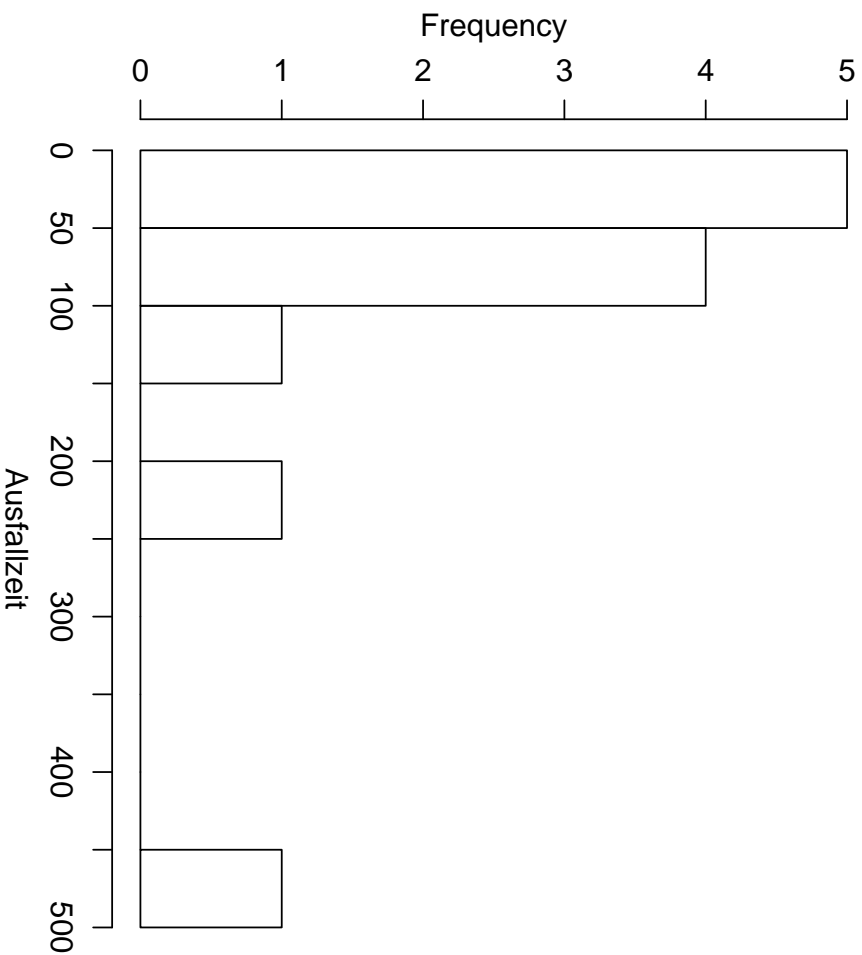
→ Zufällige Zuteilung v. potent. Hagel-Tagen zu den Gruppen
Grosse Streuung → Gewitter von 5 Jahren

b **Ausfallzeiten** des Air conditioning-Systems in Boeing 720

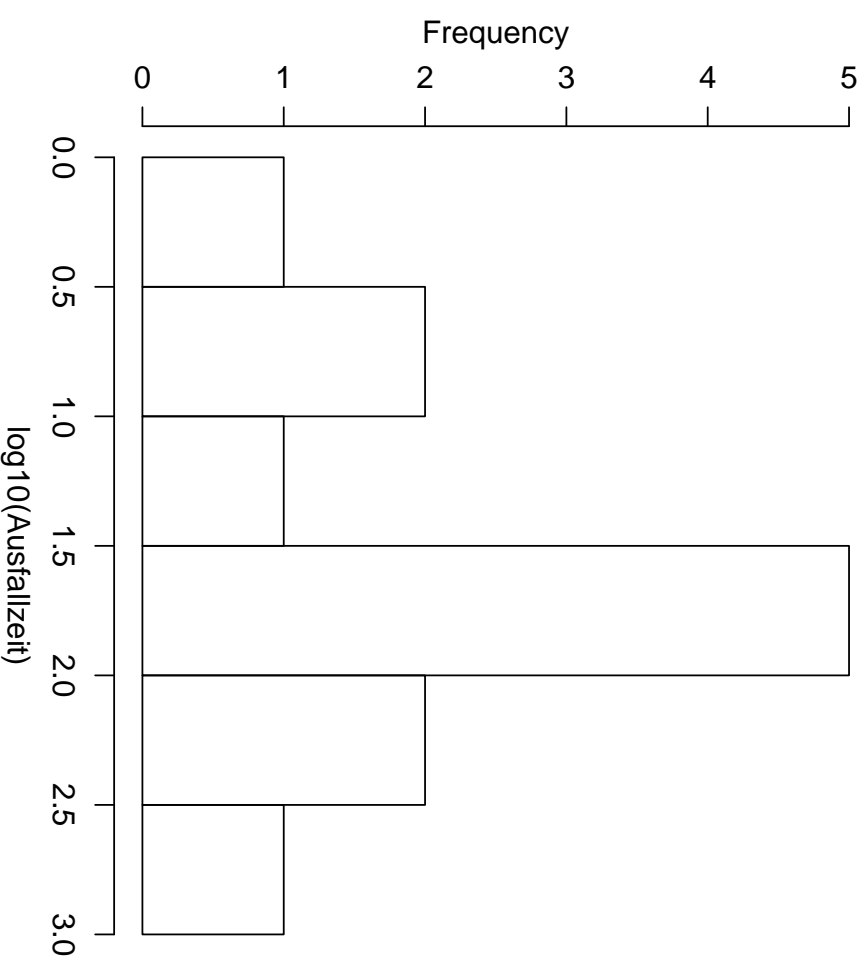
$n = 12$ Zeiten zwischen Ausfällen (sortiert):

3 5 7 18 43 85 91 98 100 130 230 487

Daten



log. Daten



- c Einfachstes parametrisches Modell für Ausfallzeiten:

Exponential-Verteilung &xp mit Dichte

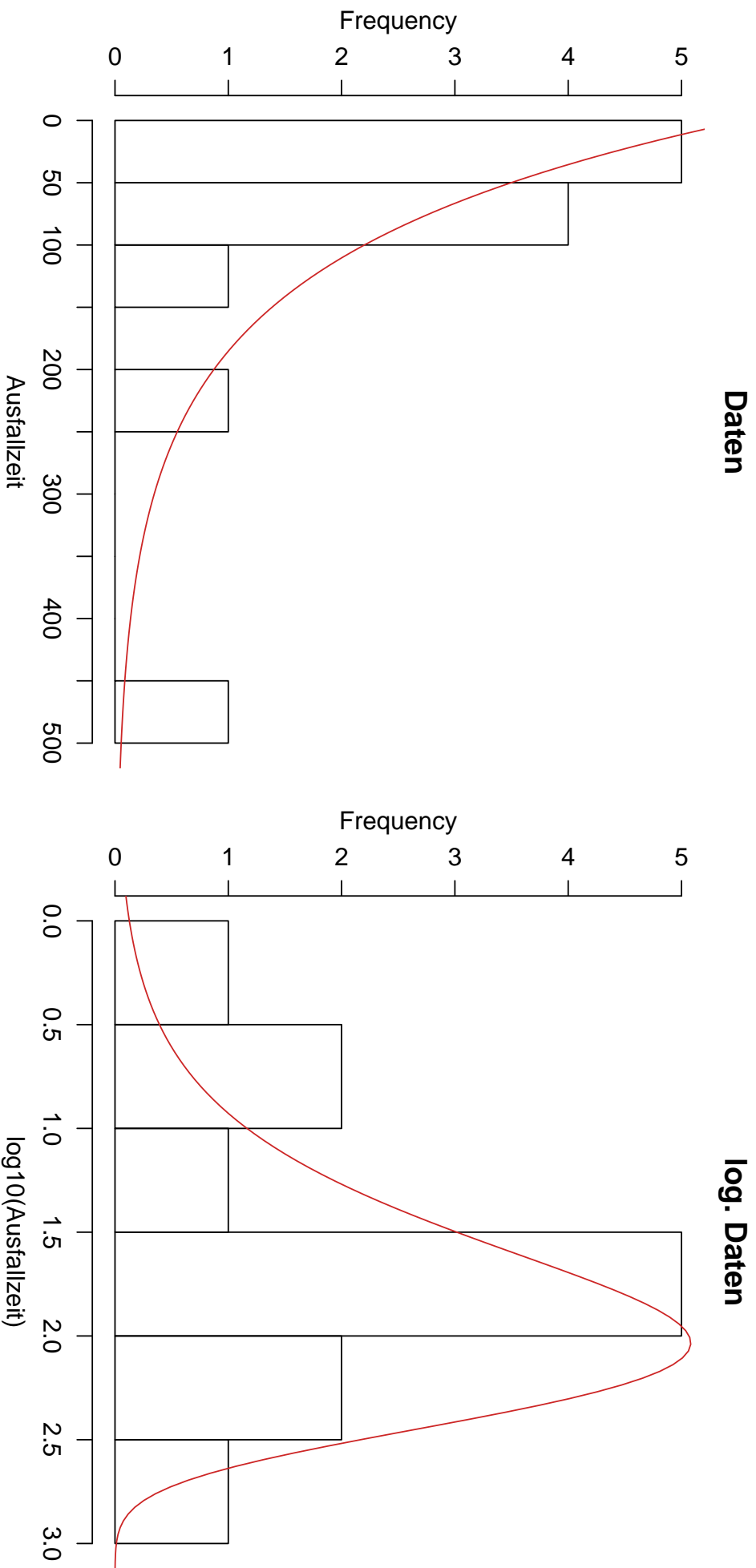
$$f\langle y \rangle = \frac{1}{\mu} e^{-y/\mu} \quad y > 0$$

Oder allgemeiner: **Gamma-Verteilung** mit Dichte

$$f\langle y \rangle = \frac{1}{\Gamma\langle \kappa \rangle} \left(\frac{\kappa}{\mu} \right)^{\kappa} e^{-\kappa y / \mu} \quad y > 0, \quad \mu > 0, \quad \kappa > 0$$

Passen diese Modelle?

→ Parameter schätzen, Kurven einzeichnen.



Histogramm der Air conditioning-Daten mit Dichten der angepassten Exponential- und Gamma-Verteilung

d Modelle passen nicht schlecht. Aber der Datensatz ist klein.

Besser: **Keine Verteilung voraussetzen!**

Frage: Mittelwert? $\bar{x} = 108.1$.

Etwas spannender: 20% gestutztes Mittel?

= Lasse die 20% kleinsten & 20% grössten Daten weg,
bilde Mittel der übrigen!

Für $n = 12$ je 2 Beob. weglassen.

→ $(7 + 18 + 43 + 85 + 91 + 98 + 100 + 130) / 8 = 71.5$

„Eine Zahl ohne Genauigkeitsangabe ist wertlos!“

→ Schliessende Statistik, Vertrauensintervall.

Dafür braucht man Wahrscheinlichkeitsmodelle!

1.3 Parametrische & nichtparametr. Statistik

- a **Wahrscheinlichkeitsmodell** wird gebraucht, um zu beschreiben, was „auch noch hätte herauskommen können, und mit welchen Chancen“.

Besser: W.modell besteht, bevor wir die Daten sehen, und beschreibt unsere Vorstellung, was für Resultate wir mit welcher „Plausibilität“ erwarten.

b Parametrische Verteilungsfamilien

Im Beispiel: Exponential- (oder Gamma-) Verteilung, Parameter μ (oder $[\mu, \kappa]$).

Am bekanntesten: Normal- und Binomial-Verteilung.

Fragestellung meist mit Bezug auf die Parameter formuliert:

Schätzung von μ ; Vertrauensintervall; Test für Nullhyp. $\mu = \mu_0$.

c Nichtparametrische Statistik

Wir wollen die Annahme einer parametr. Familie vermeiden.

Es braucht trotzdem Annahmen!

Es bleibt: X_i sind **unabhängig und gleich verteilt**.

$$X_i \sim \mathcal{G}, \quad \text{unabhängig}$$

→ Frage so formulieren, dass sie für jedes \mathcal{G} Sinn macht.

Beispiel: Median ist für alle Verteilungen definiert.

Ebenso Erwartungswert, Varianz, andere Quantile etc.

= „**Funktional**“.

- d **Das Wort „nichtparametrisch“** wird auch anders verwendet:
Nichtparametrische Regression:
Regressionfunktion nicht über Parameter festgelegt („glatt“)
setzt meist Normalverteilung der Zufallsfehler voraus!
- e **Grundidee des Resampling:**
Die Daten selber verwenden, um ihre Verteilung \mathcal{G} zu schätzen.

1.4 Überblick

- Simulation (Gewöhnung an Notation und Jargon)
- Bootstrap
- Randomisierungstests, inkl. bekannte nichtparametr. Tests
- Ausblick auf andere Resampling-Verfahren

2 Simulation

2.1 Zufallszahlen

- a Zurück zum parametr. Wahrscheinlichkeitsmodell, Bsp. Exponential-Vert.

Die Wahrscheinlichkeit liefert komplexe Modelle:

Zur Beschreibung einer (Zufalls-) Zahl brauchen wir

eine ganze Funktion (W.-Funktion, -Dichte oder kumul. Vt.fn.)!

Anderer anschauliche Vorstellung: Modell legt Möglichkeit fest, ihm entsprechende Zufallszahlen zu ziehen.

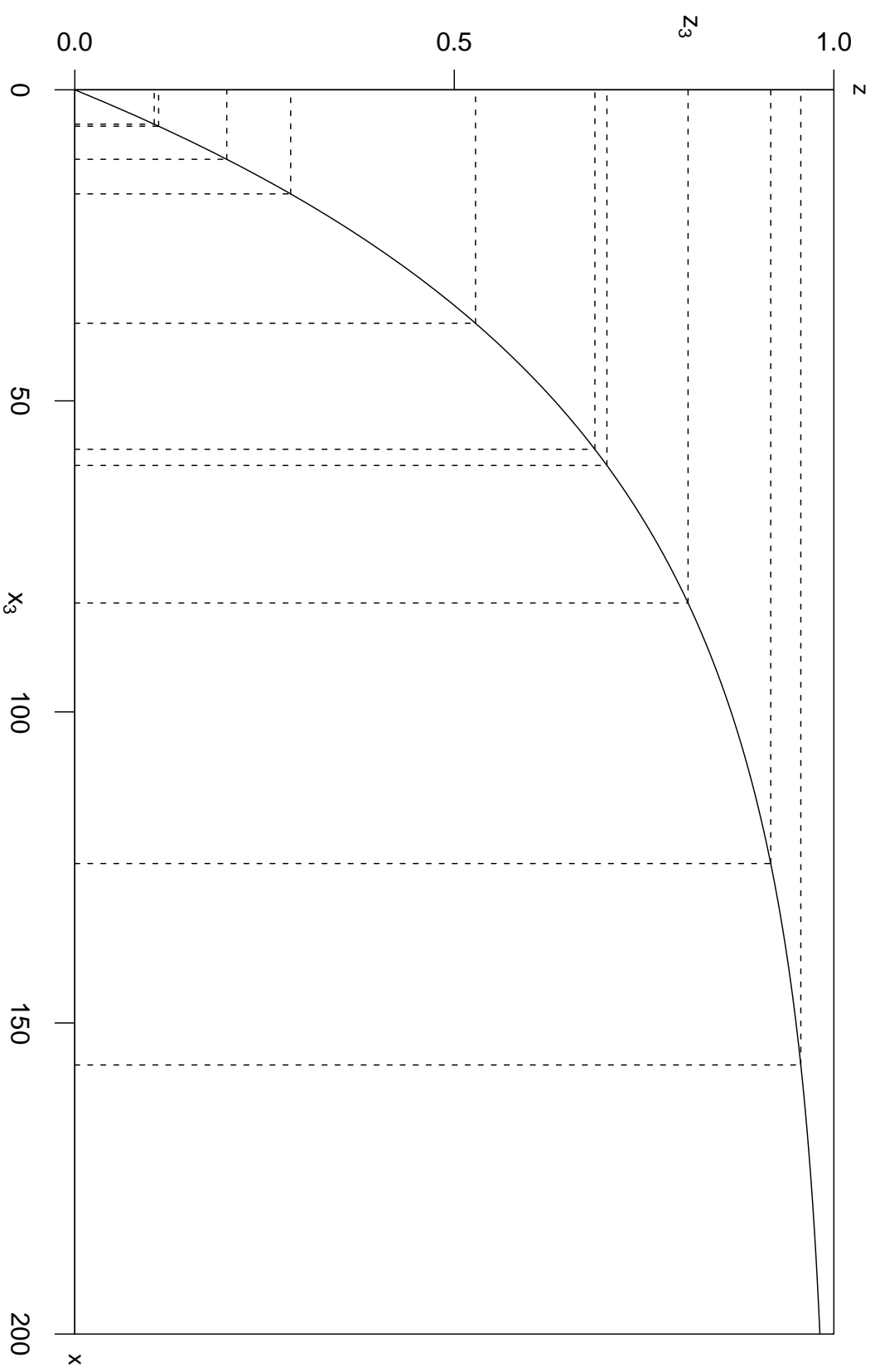
→ **Anschauung für experimentell denkende Leute.**

b Zufallszahlen-Generatoren erzeugen Zahlen z_i , die der **uniformen Verteilung** entsprechen.

Aus ihnen erhält man Zufallszahlen mit beliebiger, geg. Vt. \mathcal{F}
Kumulative Verteilungsfunktion F , inverse F^{-1}

$$x_i = F^{-1}(z_i)$$

sind Zufallszahlen entsprechend der Vt. \mathcal{F} .



Simulation von stetigen Zufallsvariablen:

10 Zufallszahlen z_i werden aus uniform verteilten z_i berechnet.

c **Aufgabe**

Die Dichte der Exponential-Verteilung ist gegeben durch

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma}, \quad x > 0$$

Wie muss man eponential verteilte Zufallszahlen aus uniform verteilten berechnen?

2.2 Verteilung einer Schätzung

a Modell für eine **Stichprobe**:

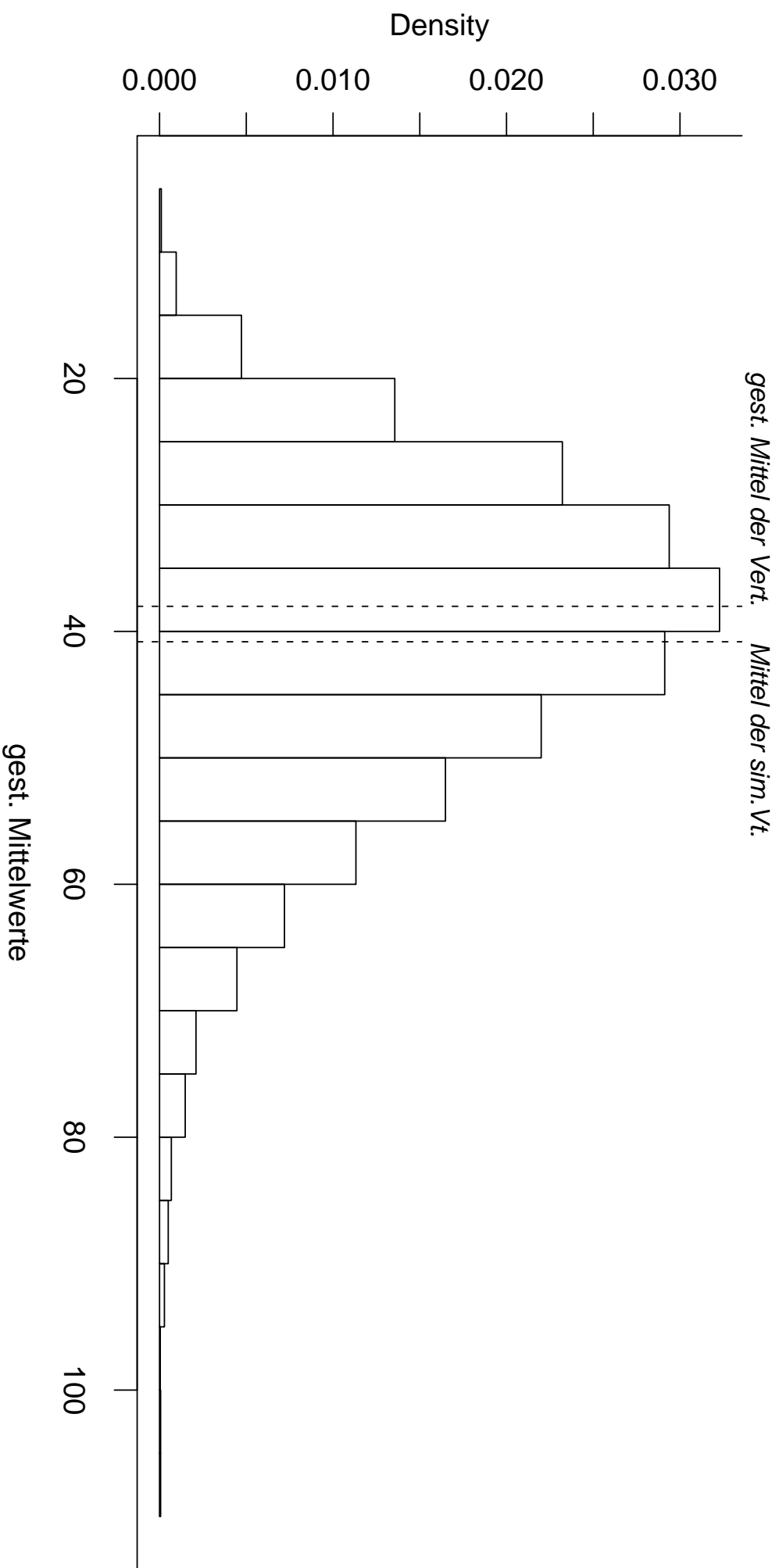
- n unabhängige, identisch verteilte Zufallsvariable X_i .
noch abstrakter!
- n Zufallszahlen zusammenfassen.

Aus je n Zufallszahlen den Wert $T\langle x_1, x_2, \dots, x_n \rangle$ berechnen.

nrep Mal \longrightarrow „**Stichprobe**“ der **Verteilung von T**

Verteilung einer Schätzung simulieren. „Ersetzt“ die W.rechnung!

Beispiel: 20% gestutztes Mittel einer Stichprobe
von 12 exponential-verteilten Beobachtungen



Simulierte Verteilung des 20% gestutzten Mittels
von 12 exponential-verteilten Beobachtungen mit $\sigma = 50$

b **Das ist der Grund-Baustein der Statistik!**

Gegeben ist die Vt. der Beobachtungen: Stichprobe $X_i \sim \mathcal{G}$, unabhängig.

Gesucht ist die **Verteilung** \mathcal{L} **einer Funktion** $T\langle X_1, \dots, X_n \rangle$ der Beobachtungen.

T typischerweise Schätzung eines Parameters oder Test-Statistik.

Die Vt. hängt von T und \mathcal{G} ab,

$$X_i \sim \mathcal{G}, \quad \text{unabhängig} \quad \implies \quad T\langle X_1, \dots, X_n \rangle \sim \mathcal{L}\langle T, \mathcal{G} \rangle$$

Beispiel $T = \text{gest. Mittel}$, $\mathcal{G} = \text{Exp}\langle \sigma \rangle$, $\sigma = 50$.

c **Schwierigkeit:**

Zusätzliche Wirkung des Zufalls → Verwirrung!

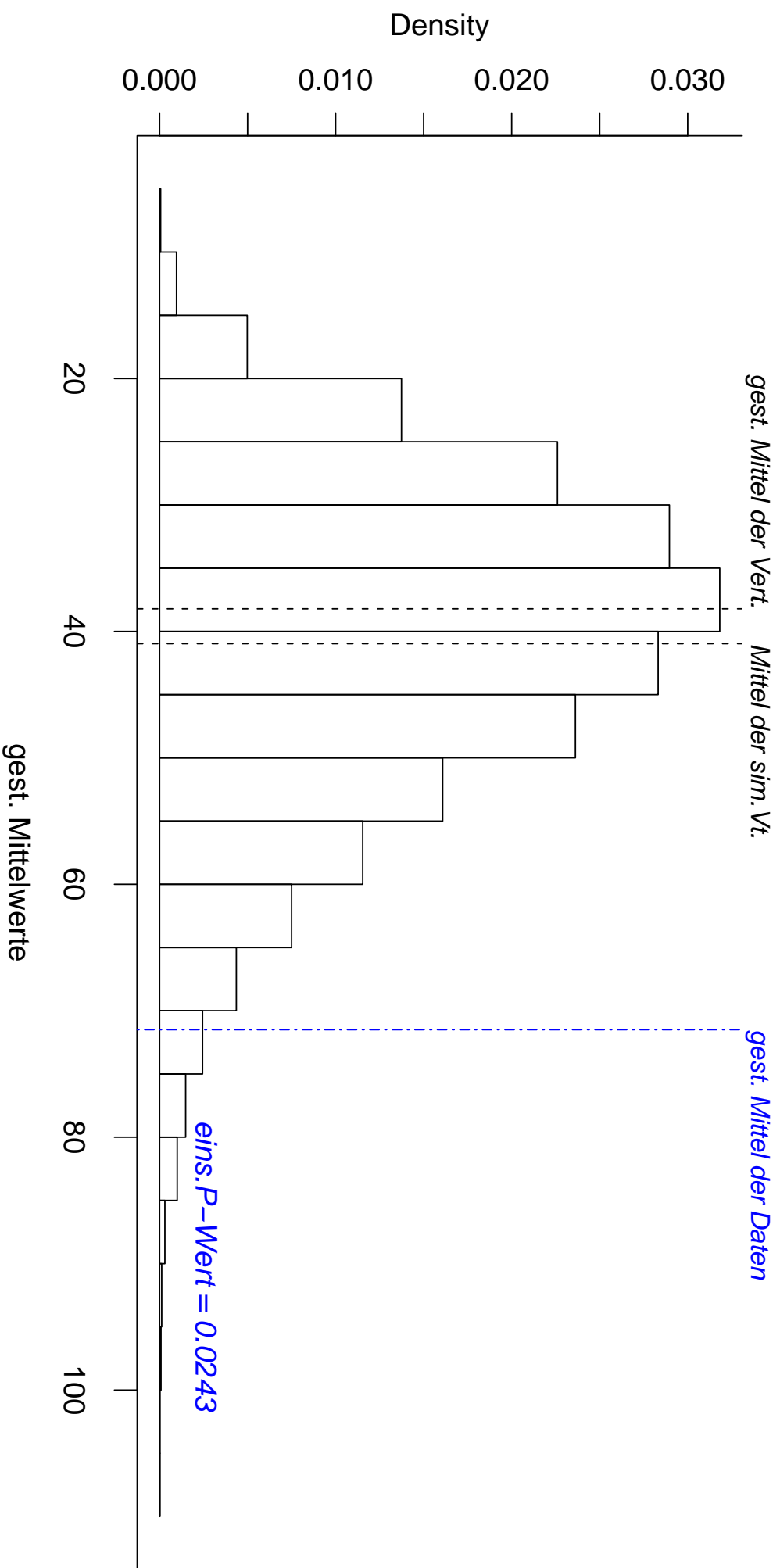
- **Essenzieller Zufall:** Modell für die Daten
Unsicherheiten durch Modell & Stichprobenumfang. bestimmt.
- **Technischer Zufall: Simulation**
Unsicherheiten durch beschränkte Zahl n_{rep} bestimmt
→ Genauigkeit kann mit mehr Computer-Aufwand
beliebig erhöht werden.

d **Simulierter Test:**

Simulation der V_t . des 20% gest. Mittels

unter der Nullhypothese $\sigma = 50$

gest. Mittel(Daten) = 71.5 \longrightarrow Ablehnung.



Simul. Verteilung des 20% gestutzten Mittels mit beob. Wert und P-Wert

e **Aufgabe**

Der P-Wert für einen Test wurde aus

$nrep=1000$ Simulations-Replikaten als

$p=0.02$ berechnet.

(a) Wie genau ist das?

(b) Muss man $nrep$ erhöhen?

→ Bestimmung der Anzahl nötiger Replikate
hängt nicht vom spezifischen Test-Problem ab!

f Theoretische Untersuchungen

Eigenschaften von T ?

- Verteilung des geschätzten gestutzten Mittels ist schief!
- T sollte das 20% gestutzte Mittel der V_t G schätzen.
Nicht erwartungstreu.

Eigenschaften eines Tests (Test-Statistik und Verw.bereich):

- Wahres Niveau = W . des Fehlers erster Art.
Simulation unter Nullhyp. \longrightarrow Häufigk. der Verwerfung.
- Macht = W . des Verwerfens unter (best.) Alternative.

Simulation wird verwendet, wenn neue statistische Methoden begründet werden sollen.

g Anwendung für grafische Methoden der Datenanalyse

Quantil-Quantil-Diagramme:

Vergleich von empirischer Verteilung (Daten) mit Modell-Verteilung

Kumulative Verteilungsfunktion, empirisch und theoretisch

Umkehrfunktion = Quantilfunktion, empirisch und theoretisch

Empirische Quantile sollten \approx theoretische sein

Diagramm: Empirische Quantile vs. theoretische = QQ-Diagramm

Oft benützt für Residuen-Analyse in der Regression.

Ab wann sind Abweichungen ernst zu nehmen?

Simuliere 19 QQ-Plots entsprechend dem Modell!

Ist QQ-Plot der Daten auffällig gegenüber diesen 19 simulierten?

Merkmale

Simulation

- Simulation kann Wahrscheinlichkeitsrechnung ersetzen.
 - **Spielderscher Umgang mit Wahrscheinlichkeit.**
- Schwierigkeit: Zusätzliche Wirkung des Zufalls

Quad Arrow Verwirrung!

- 1. Anwendung in der Statistik: „Monte Carlo“-Untersuchungen der **Eigenschaften von statistischen Verfahren.**
- 2. Anwendung: „Kalibrierung“ für **grafische Datenanalyse.**

3 Bootstrap

3.1 Die grundlegende Idee

- a Die grundlegende Aufgabe war, für ein best. T (gest. Mittel) die Verteilung $\mathcal{L}(T, G)$ zu bestimmen.

Anschaulich formuliert:

Ich habe für die vorliegenden Daten $T = t$ erhalten.

Wie unsicher ist dieser Wert?

→ Verteilung von T – unter welcher Verteilung G der X_i ?

- Letztes Kapitel: \mathcal{G} aus parametrischer Familie $\mathcal{G} = \mathcal{F}_\theta$. –
Welches θ ?
Naheliegend: θ aus den Daten schätzen $= \hat{\theta}$, $\mathcal{G} = \mathcal{F}_{\hat{\theta}}$.
→ Simulation unter geschätzter parametrischer Vert.
Neuer Name: „Parametrischer Bootstrap“.
- „Nichtparametrischer Bootstrap:“
 \mathcal{G} = empirische Verteilung der Daten
Inhalt dieses Kapitels.

b Schliessende Statistik ohne Modell?

- $X_i \sim \mathcal{G}$, unabhängig (= Zufalls-Stichprobe)
- Ohne zusätzliche Struktur keine sinnvollen Fragen; diese brauchen **Alternativen**.

c Stichprobe benützen, um ihre Verteilung zu schätzen, tönt nach Münchhausen.

Eigenschaften untersuchen, wie für jedes andere stat. Verfahren
→ Verfahren bewähren sich für viele (nicht alle!) Probleme.

3.2 Nichtparametrischer Bootstrap konkret

- a Empirische Verteilung, bezeichnet als \hat{G} , ist definiert durch
- $$P\langle X^* = x_i \rangle = 1/n \quad \text{für jedes } x_i \text{ aus der beob. Stichprobe}$$
- sonst = 0.

In seltenen Fällen kann man $\mathcal{L}\langle T; \hat{G} \rangle$ analytisch bestimmen.

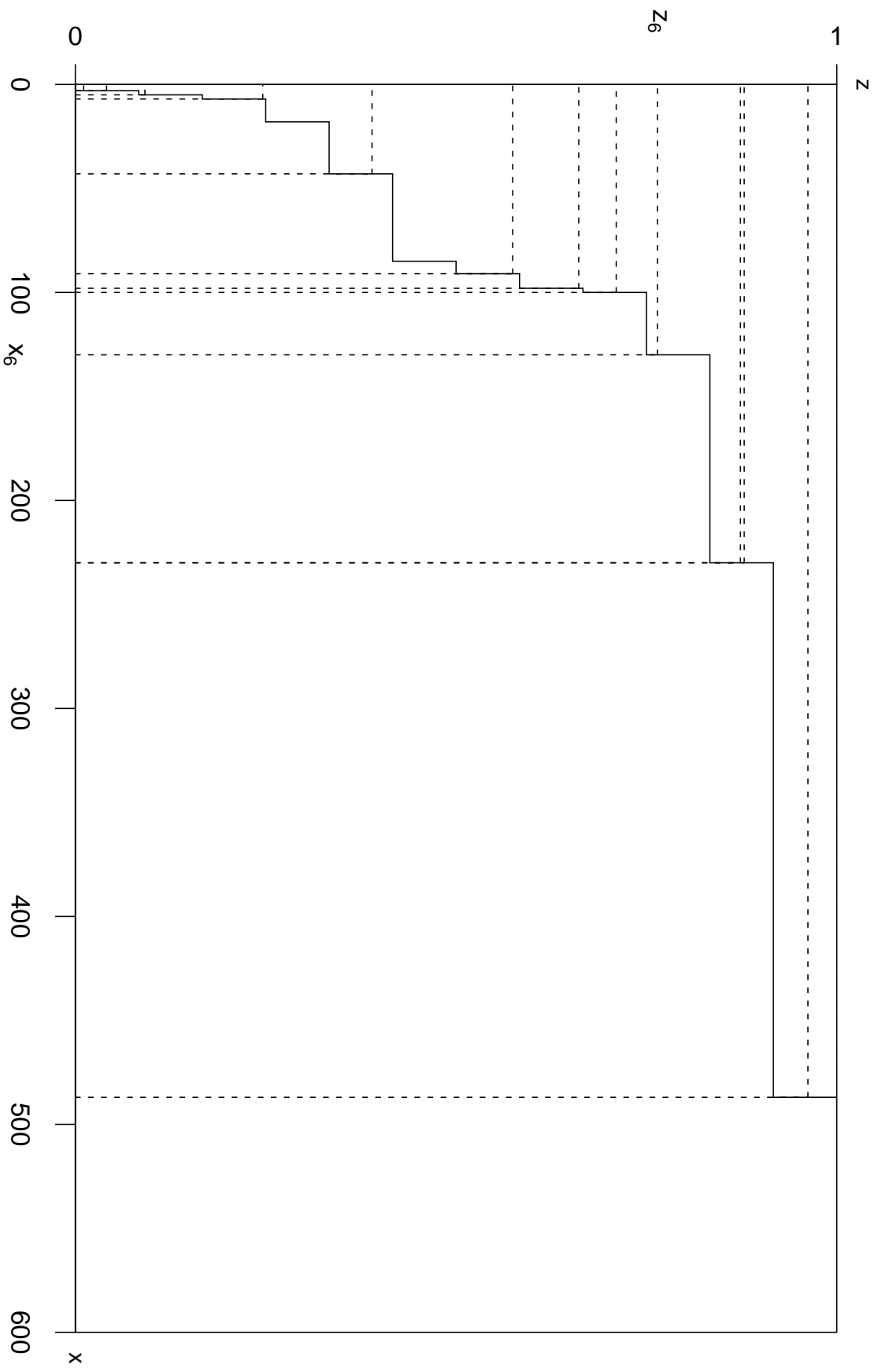
Meistens gehört zum nicht-parametr. Bootstrap die Simulation

- b **Simulation gemäss der empirischen Verteilung**

\hat{G} ist eine diskrete Verteilung, charakterisiert durch empirische Verteilungsfunktion $\hat{G} =$ Treppenfunktion.

Simulation von Zufalls-Stichproben $[X_1^*, \dots, X_n^*]$?

So, wie allgemein für diskrete Verteilungen.



Simulation von Zufallszahlen gemäss \hat{G}

- c Führt zu: Ziehen von n Werten aus den geg. n Werten x_i mit Zurücklegen.

Resultat:

- Bootstrap-Stichprobe enthält nur Werte der beobachteten Stichprobe.
- Einige Werte kommen nicht vor, einige 1, 2, 3 ... Mal

- d **Aufgabe:**

Wie gross sind die Wahrscheinlichkeiten dafür?

W., dass der grösste Wert der Stichprobe im bootstrap-sample

- nicht
- 1, 2, 3 ... Mal vorkommt?

- e Name „**Resampling**“: Wiederverwertung der Stichprobe.
Wenn analysiert. Lösung möglich ist, ist der Begriff irreführend!
(Bootstrap verwendet zwar die Daten wieder,
aber „sample-t“ nicht.)

f Weiter geht's wie vorher:

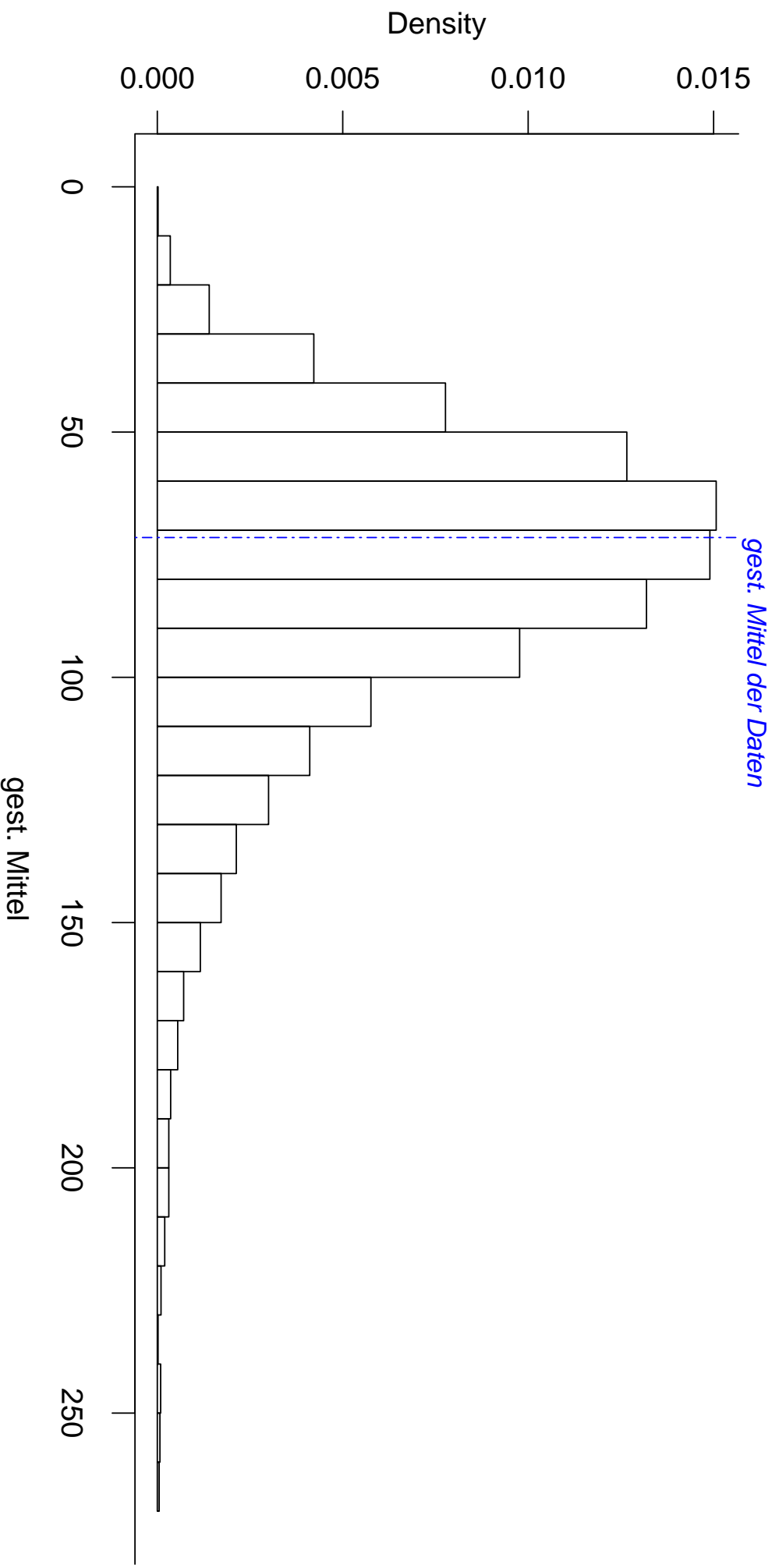
- Erzeuge `nrep` Stichproben $[X_1^{*r}, \dots, X_n^{*r}]$, $r = 1, \dots, nrep$,
- Rechne die simulierten Werte von T aus:

$$t^{(r)} = T(X_1^{*r}, \dots, X_n^{*r})$$

→ (simulierte) **Bootstrap-Verteilung** von T .

`nrep` meistens zwischen 500 und 10 000.

g Bootstrap-Vt. für das Beispiel



Bootstrap-Verteilung des gestutzten Mittels für die Air condition-Daten

3.3 Bootstrap-Tests und -Vertrauensintervalle

- a Grundgedanke des statistischen Tests:
 - Vorgegebenes Modell \mathcal{F}_0 mit Daten vergleichen:
 - Diskrepanz akzeptabel oder zu gross?
 - Vert. unter Modell bestimmt Annahmebereich für Daten.

Nichtparametrischer Bootstrap: Modell „gleich“ Daten.

→ Prinzipielle Schwierigkeit:

Wie kommen wir zu einem Modell, das der Nullhyp. entspricht, ohne eine parametrische Familie zu verwenden?

Schon besprochen:

Um eine sinnvolle statistische Unsicherheit zu bestimmen, müssen wir alternative Modelle zu \mathcal{G} einbeziehen.

b Vertrauensintervalle

In vielen klassischen Anwendungen haben Vertrauensintervalle die Form

$$\hat{\theta} \pm q \operatorname{se}^{(\theta)}, \quad \operatorname{se}^{(\theta)} = \text{Standardfehler von } \theta$$

q Quantil einer t-Verteilung oder Normalverteilung, also $q \approx 2$.
Standardfehler von $\theta =$ Standardabweichung der V_t von θ .

Die Verteilung haben wir bestimmt (Bootstrap-Verteilung).

→ Verwende St.dev. der Bootstrap-Verteilung als $\operatorname{se}^{(\theta)}$.

„Bootstrap normal confidence interval“

Problem gelöst !

???

- c Diese Form stimmt (approx.), wenn gilt:
Wenn sich θ verändert, verschiebt sich die Verteilung von T ,
ändert aber ihre Form und Streuung nicht (stark).
Normalerweise kommt das davon, dass auch die Vert. der X_i
sich nur verschiebt.

Stimmt nicht z.B. f. Binomial-Vt. $X_i \sim \mathcal{B}(n, p)$, Korrelation, ...

Eine Variante gilt für Exponential-Vert.:

Vert. wird skaliert statt verschoben.

Log-Transformation \longrightarrow Verschiebung

Allgemeiner: Nach geeigneter **Transformation der Test-Statistik** kann die Verteilung die „**Verschiebungseigenschaft**“ (approx.) erhalten.

- d Falls Bootstrap-Vt. normal ist (prüfen mit QQ-Plot), dann ist $\hat{\theta} \pm 2 \cdot \text{Bootstrap-se}^{(\theta)}$ meistens gut. (Varianz stabil?)

e Grundlegendes Verschiebungs-Beispiel

- Beobachtungen X_i lassen sich schreiben als

$$X_i = \theta + Z_i, \quad Z_i \sim \mathcal{G}_0$$

→ „Lokations-Familie“

- T ist eine „translations-äquivalente“ Funktion

$$T\langle \theta + Z_1, Z_2, \dots, Z_n \rangle = \theta + T\langle Z_1, Z_2, \dots, Z_n \rangle$$

- T schätzt θ im Sinne von

$$\mathcal{E}\langle T\langle Z_1, Z_2, \dots, Z_n \rangle \rangle = 0$$

Dann ...

kommt man ans Ziel – wenn man die Fallstricke beachtet!

f Falls Bootstrap-Vt. **schief** – was tun?

Wir haben die ganze Verteilung, können **Quantile** bestimmen

→ $\hat{q}_{0.025}^{(\theta)}$ und $\hat{q}_{0.975}^{(\theta)}$ liefern Grenzen des Vertrauensintervalls.

„**Bootstrap percentile confidence interval**“

Wirklich??? – Das wären Grenzen des **Annahmebereichs**,

wenn Nullhypothese $\theta = \hat{\theta}$ geprüft werden müsste! (!)

Verschiebungs-Prinzip: Annahmebereich

$$\left[\theta_0 - (\hat{\theta} - \hat{q}_{0.025}^{(\theta)}) , \theta_0 + (\hat{q}_{0.975}^{(\theta)} - \hat{\theta}) \right]$$

Vertrauensintervall-Grenzen definiert durch:

- untere Grenze: θ_0 so, dass $\hat{\theta} =$ obere Grenze des Annahmebereichs.
- obere Grenze: θ_1 so, dass $\hat{\theta} =$ untere Grenze des A.

Vertrauensintervall:

$$\left[\hat{\theta} - (\hat{q}_{0.975}^{(\theta)} - \hat{\theta}) \quad , \quad \hat{\theta} + (\hat{\theta} - \hat{q}_{0.025}^{(\theta)}) \right]$$

→ Abw. des oberen Quantils von $\hat{\theta}$ nach unten abtragen
und umgekehrt!

„Bootstrap standard confidence interval“

g Hinweis:

Wenn die Bootstrap-Verteilung schief ist,
hängt ihre Streuung oft von θ ab.

→ Transformation von θ (und T),
Vertrauensintervall bestimmen,
zurücktransformieren.

h Beispiel Exponential-Verteilung

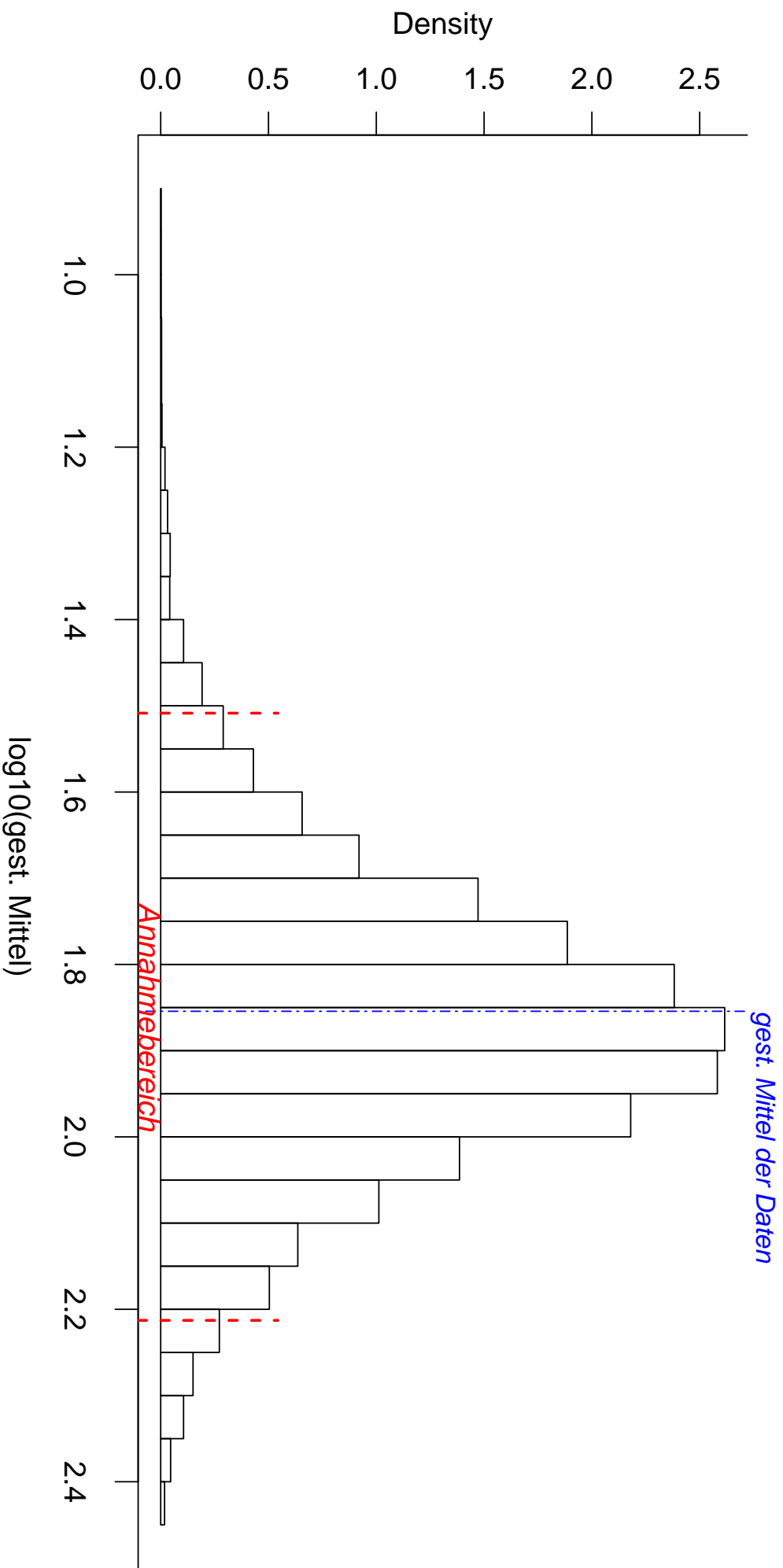
Bootstrap-Verteilung schief.

Streuung proportional zu σ (Plausibilitätsbetrachtung!)

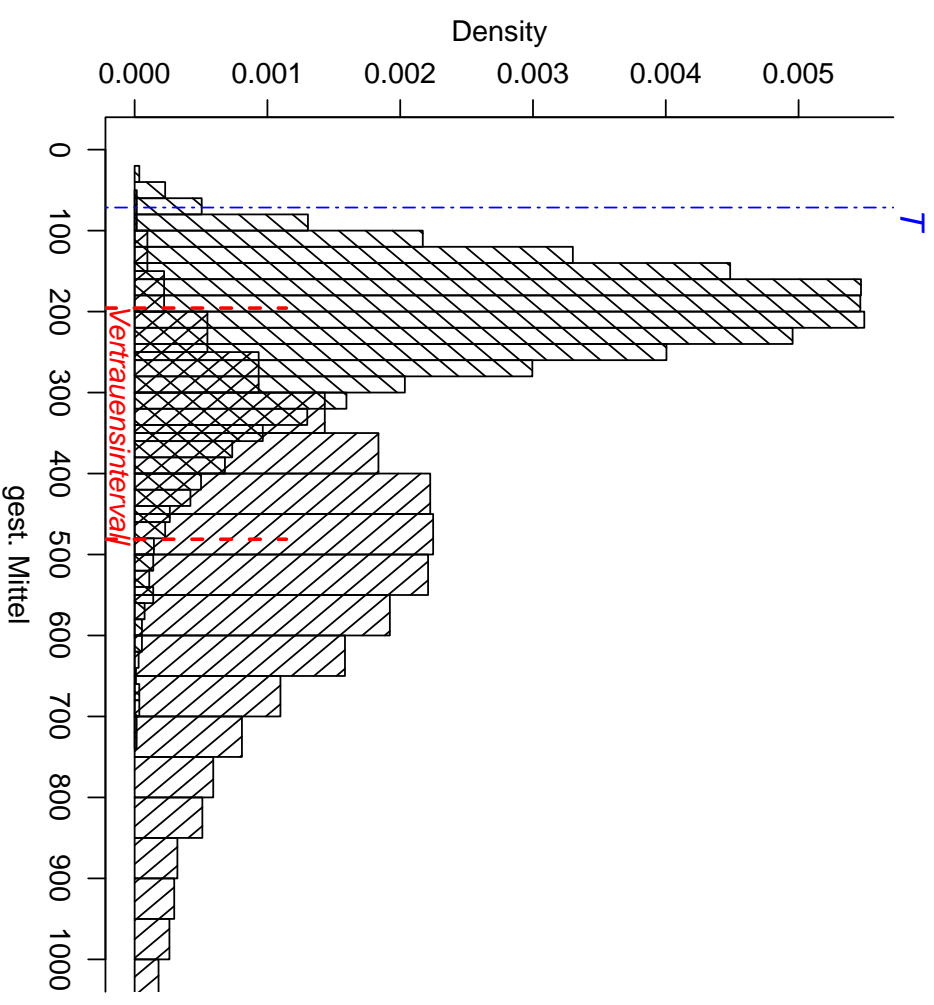
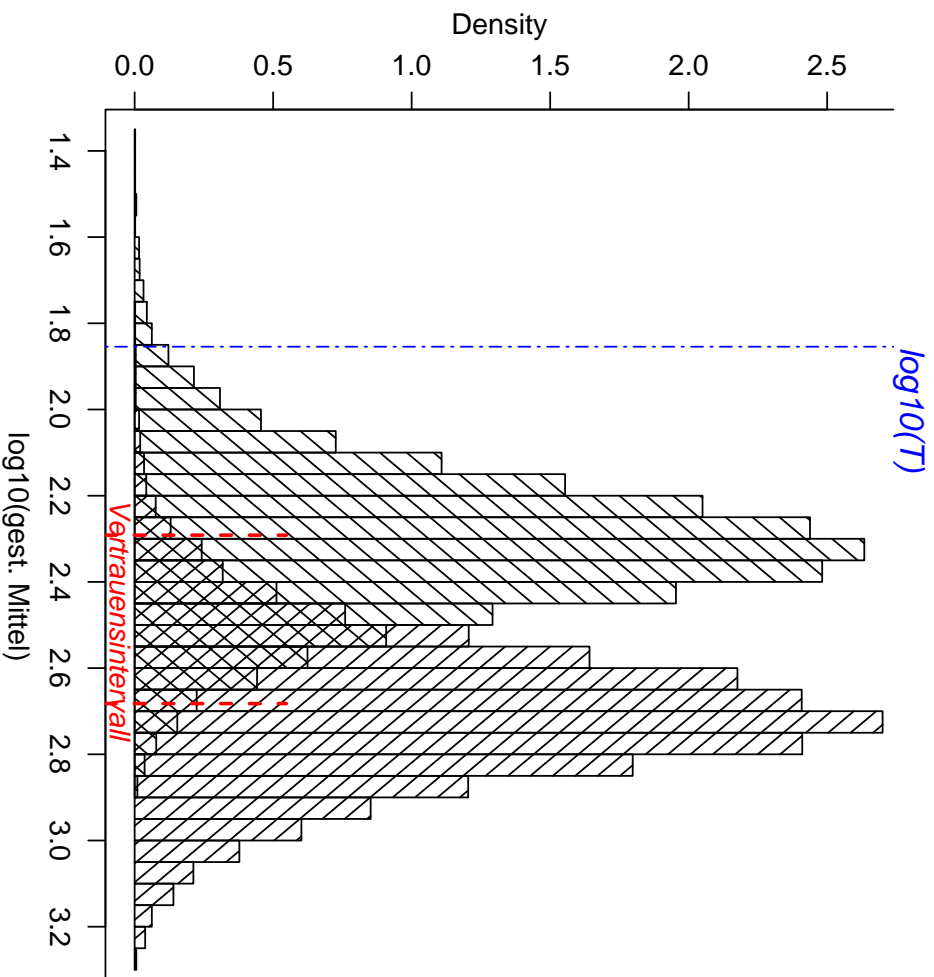
→ Log-Transformation: macht Exponential-Vt. zu Gumbel=Vt.

Skalen-Parameter σ wird zu Lokations-Par. $\theta = -\log\langle\sigma\rangle$

→ Verschiebungs-Eigenschaft



Bootstrap-Verteilung von $\log\langle T \rangle$ mit „Annahmebereich“



Bootstrap-Verteilung, verschoben auf die Grenzen des Vertrauensintervalls
für $\log\langle T \rangle$ und zurücktransformiert

- i Diese Idee, **allgemeiner formuliert**:
 T sei eine Schätzung von $\theta = T(G)$.

Idealerweise hängt die Verteilung von $T - \theta$ nicht von θ ab.

Gilt z.B. für Schätzung des Parameters einer Lokations-Familie.

Für die Binomial-Verteilung $X \sim \mathcal{B}(n, \pi)$ ist

$$T = X/n \text{ Schätzung von } \pi.$$

Vert. von $T - \pi$ hat Erwartungswert 0 und Varianz $\pi(1 - \pi)/n$.

also: **Varianz hängt von Parameter ab.**

Annahmebereich f. untere Grenze des Vertrauensint. schmaler

als für obere Grenze, falls $\pi < 0.5$.

→ nicht aus der Breite für $\pi = T$ zu schätzen!

Verbesserung: Suche Transformation $g\langle\theta\rangle$, so dass

$\text{var}\langle g\langle T\rangle\rangle$ nicht von θ abhängt.

= „**varianzstabilisierende Transf.**“

(Für Binomial-Verteilung: Arc-Sin-Transformation $g\langle\pi\rangle = \arcsin\langle\sqrt{\pi}\rangle$)

Bestimme Vertrauensintervall für $g\langle\pi\rangle =$ Bootstrap-Vi. aus $g\langle T\rangle$

→ Rücktransformation.

j Stör-Parameter

Bei den meisten Fragestellungen gibt es Stör-Parameter.

$X_i \sim \mathcal{N}(\mu, \sigma^2)$, μ von Interesse, σ Stör-Parameter.

→ Student's t-Test statt z-Test, „Studentisieren“.

Das lohnt sich auch für die Bootstrap-Version.

Idee:

$$\tilde{T} = \frac{T - \theta}{\widehat{se}} \quad \text{mit} \quad se = \hat{\sigma} / \sqrt{n}$$

hat eine Verteilung, die nicht von den Parametern abhängt –

falls $X_i \sim$ Lokations-Skalen-Familie, z.B. $\mathcal{N}(\mu, \sigma^2)$.

Die Verteilung von \tilde{T} eignet sich besser zum Bootstrappen
als die Verteilung von T .

k Konkret:

- Berechne Bootstrap-Vt. von $[T, se]$ und daraus \tilde{T} .
- Bilde Bootstrap-Annahmebereich für \tilde{T}
Quantile der B-Vt. $\longrightarrow [\tilde{t}_0, \tilde{t}_1]$
- \longrightarrow Vertrauensintervall für θ :
$$[\hat{\theta} - \hat{se} \cdot \tilde{t}_1 , \hat{\theta} - \hat{se} \cdot \tilde{t}_0]$$

Merkpunkte

Bootstrap

- Der Bootstrap liefert ohne Annahmen über die Form der Vert. der Beobachtungen und für beliebig komplizierte Funktionen T der Beob.
 - Verteilung von T
 - insbesondere Erwartungswert, Varianz und Quantile
 - Vertrauensintervall(Bedingungen an die „Glattheit“ von T werden gebraucht!)
- Es lohnt sich, Parameter θ & Schätzung T so zu transformieren, dass die Verteilung von $T - \theta$ möglichst wenig von θ abhängt, oder eine standardisierte Grösse \tilde{T} für das Bootstrappen zu benutzen, deren Verteilung möglichst wenig von θ abhängt.

4 Randomisierungs-Tests

4.1 Einführendes Beispiel

- a Hagel-Experiment: („Grossversuch IV“ im Napfgebiet 1978-1983)
 Verringert das „Impfen“ von potenziellen Hagelwolken
 mit Silberiodid die Hagelenergie?

Zielgrösse: Hagelenergie, gemessen für n Wolken

Zwei Gruppen: ca. $n/2$ „geimpft“, Rest „Kontrolle“.

$$Y_i : \text{Hagelenergie der Wolke } i$$

$$G_i = \begin{cases} 1 & \text{falls Wolke } i \text{ geimpft,} \\ 0 & \text{sonst.} \end{cases}$$

Hoffnung: Y_i mit $G_i = 1$ fallen tendenziell niedriger aus.

b Beobachtet:

$Y_i = y_i^*$		16672	25	855	0	152	0	46	1219
$G_i = g_i^*$		1	1	0	0	0	1	1	0

g_i^* : Zufallsauswahl der zu impfenden Wolken.

(In Wirklichkeit 216 Wolken; davon wurden 94 geimpft.)

Statistischer Test! H_0 : Keine Wirkung.

(\longrightarrow Widerspruchsbeweis!)

Ungeparter Zwei-Stichproben-Problem. \longrightarrow t-Test ?

Keine Annahmen über die Verteilung der Y_i !!

4.2 Statistische Überlegung

a **Nullhypothese** = Wahrscheinlichkeitsmodell.

Üblich: Verteilung für Y_i ; $G_i = g_i^*$ fest vorgegeben.

Randomisierungstests: G_i zufällig; $Y_i = y_i^*$ als fest betrachtet
(Analyse „bedingt auf die y_i^* “.)

Falls das Impfen keinen Einfluss auf die Hagelenergie hat,
würden wir die genau gleichen Werte y_i^* erhalten,
wenn die Wolken entspr. $\vec{g}^{(1)} = [0, 1, 0, 0, 1, 1, 0, 1]$ oder
entspr. irgendeiner anderen Auswahl geimpft worden wären.

Zufallsauswahl:

Jede Auswahl von $n/2 = 4$ Elementen aus $n = 8$ hat gleiche Wahrscheinlichkeit

$$p = \binom{8}{4}^{-1} = \frac{1}{70}$$

Damit ist die **Nullhypothese** festgelegt.

- b **Teststatistik:** Soll extreme Werte annehmen, wenn Alternative gilt.

Alternative: y_i^* mit $g_i^* = 1$ sind tendenziell kleiner.

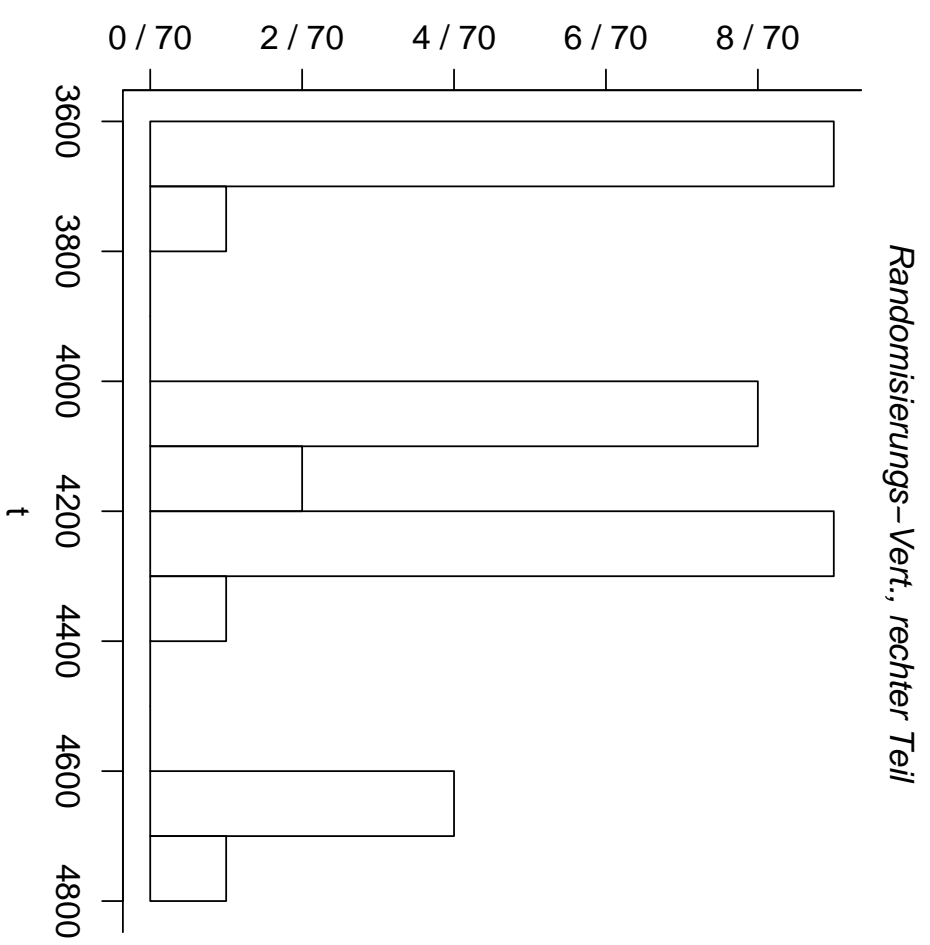
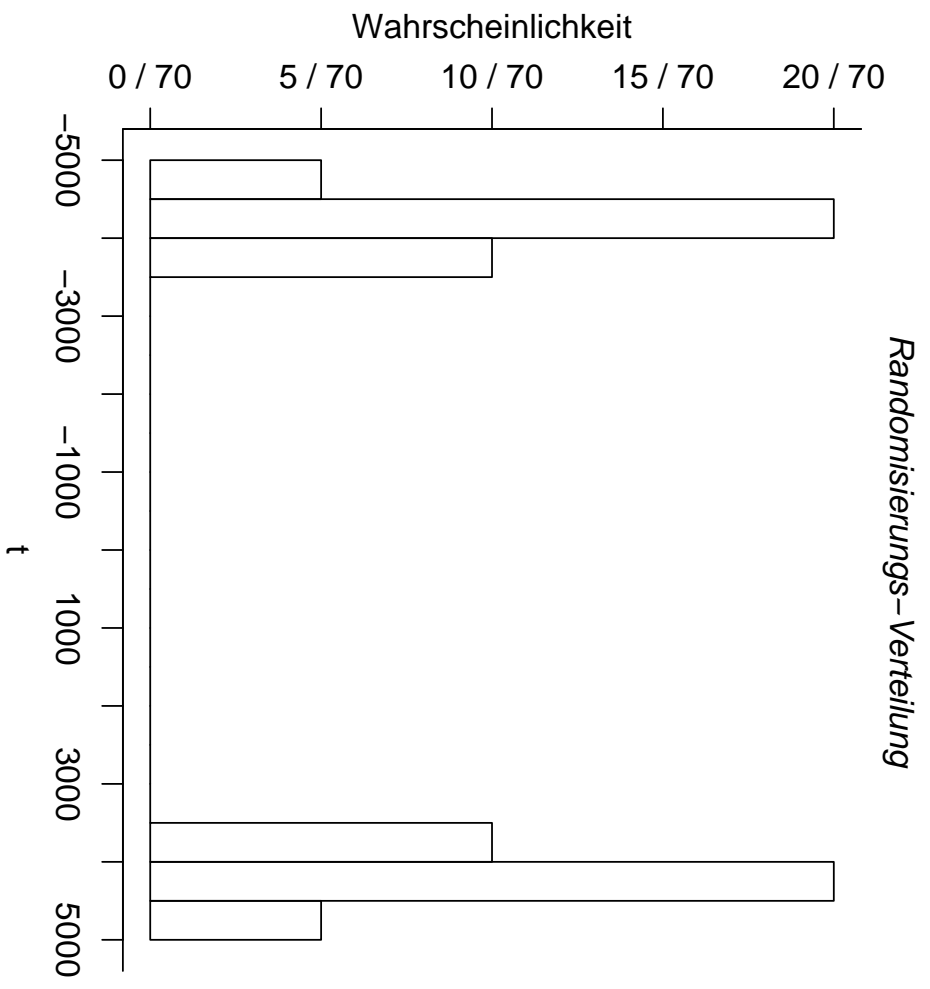
$$T\langle \vec{g}, \vec{y}^* \rangle = \frac{1}{n/2} \sum_{i:g_i=0} y_i^* - \frac{1}{n/2} \sum_{i:g_i=1} y_i^* = \frac{2}{n} \sum_i y_i^* (1 - 2g_i).$$

- c Wie ist T unter der Nullhypothese verteilt?

y_1^*, \dots, y_n^* gegeben $\longrightarrow \leq \binom{n}{n/2}$ mögliche Werte für T .

$$P\langle T\langle \vec{G}, \vec{y}^* \rangle = t \rangle = \frac{\#\{\vec{g} \mid T\langle \vec{g}, \vec{y}^* \rangle = t\}}{\binom{n}{n/2}}$$

„Randomisierungs-Verteilung“



- d **Verwerfungsbereich**: $\alpha = 5\%$ extremste Werte
(so genau als möglich).

Beispiel: $\{t \mid t \geq 4643.25\}$ (einseitig).

- e Experiment:

$$T(\vec{g}^*, \vec{y}^*) = \frac{1}{4}(855 + 0 + 152 + 1219) \\ - \frac{1}{4}(16672 + 25 + 0 + 46) = -3629.25$$

Effekt in die unerwartete Richtung!

Nullhypothese nicht verworfen; Effekt nicht nachgewiesen.

(Auch nicht in umgekehrter Richtung.)

* Voraussetzung des Tests: **Unabhängigkeit**

→ Randomisierung über 76 „potentielle Hageltage“

Davon 33 als Impftage ausgewählt. Anzahl Impftage zufällig.

→ Analyse bedingt auf Anzahl Hageltage mit Impfung.
Eingeschränkte Randomisierung.

g $\binom{76}{33} = 36 \cdot 10^{20}$ mögliche Auswahlen

→ Simulation der Randomisierungs-Verteilung.

4.3 Tests für das Zwei-Stichproben-Problem

- a Randomisierungstests sind auch dann anwendbar, wenn die Durchführung des Versuchs keinen Randomisierungsschritt enthält.

Voraussetzungen, die dann gelten müssen:

- Die Beobachtungen müssen unter H_0 gleich verteilt und
- unabhängig sein.

Dann stimmt die gewählte Irrtumswahrscheinlichkeit α exakt.

Die Randomisierungstests bilden in diesem Sinne den

„Goldstandard“ unter den statistischen Tests.

(* Schwächere Voraussetzung: „Austauschbarkeit“.)

b Wenn **Beobachtungen zufällig**:

Stichprobe $[Y_1, \dots, Y_n]$ \longrightarrow geordnete St. $Y_{[1]}, \dots, Y_{[n]}$
oder empirische Verteilungsfunktion \hat{F}_n (s. Bootstrap)

Vert. der Teststatistik, bedingt auf \hat{F}_n , = Randomisierungs-Vt.

Bedingte W. eines Fehlers erster Art, gegeben \hat{F}_n , = α
— für jede Bedingung \hat{F}_n , und deshalb auch ohne Bedingung.

c **Beliebige Teststatistik.**

Differenz der Mittelwerte unrobust.

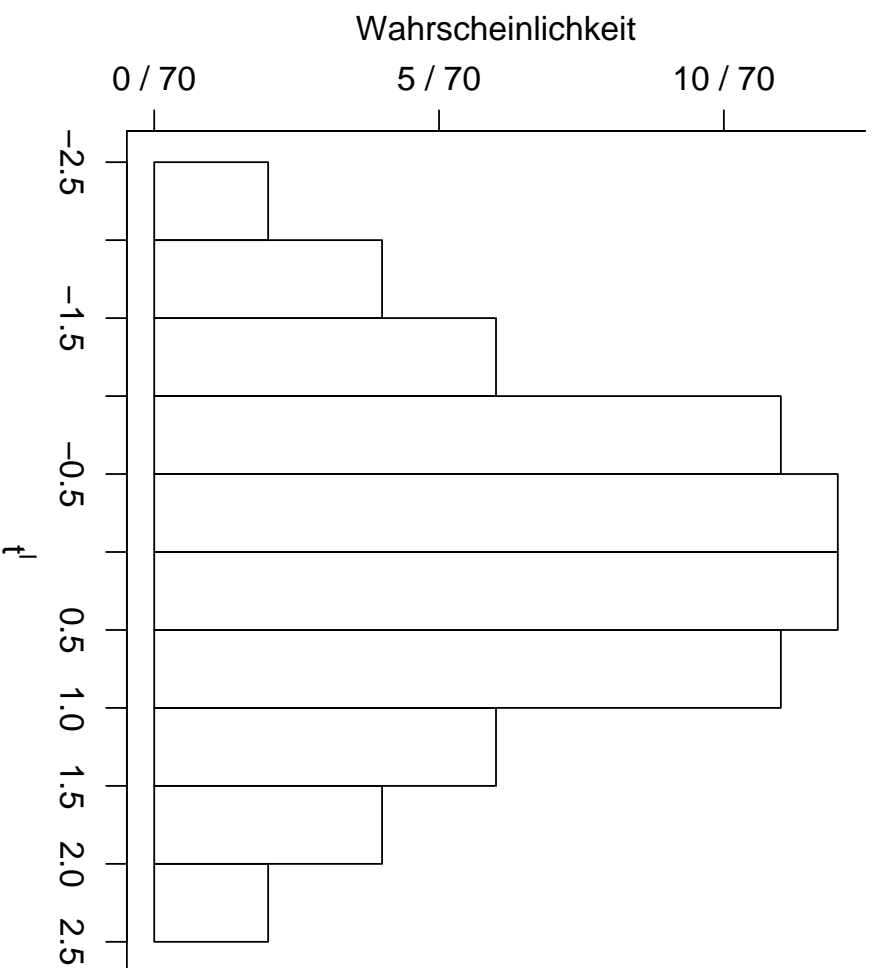
Optimale Teststatistik? → Macht für die Alternative(n) opt.!

Braucht **bestimmte** Verteilung(s-Familie)

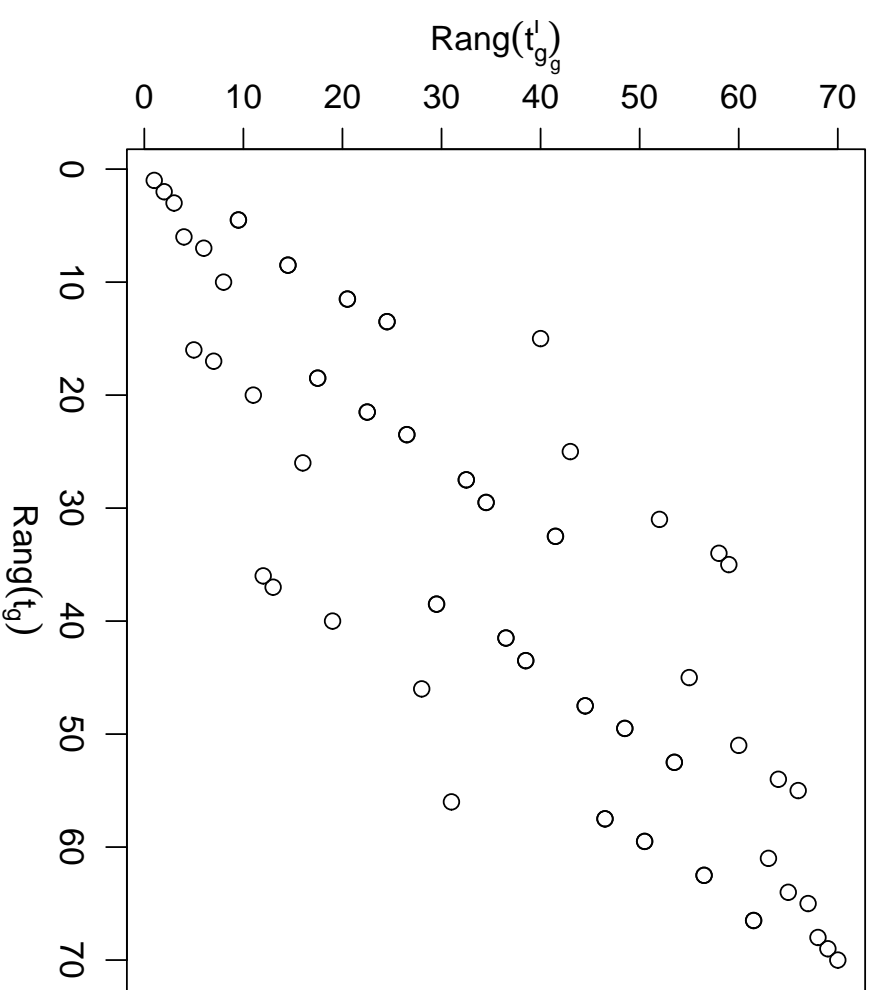
→ optimale Teststatistik (Likelihood-Ratio-Test)

d Beispiel: Logarithmus-Transformation, dann
Mittelwertsdifferenz (robustifiziert).

Rand. Vert. für log. Werte



Vergleich der Test-Statistiken



Randomisierungs-Verteilung für die Mittelwerts-Differenz von logarithmierten Daten im Beispiel (links) und Vergleich der Rangordnung der einzelnen Randomisierungen (rechts).

e **Robustheit.** Wieso eine robuste Teststatistik verwenden, wenn der Test auch ohne diese „Vorsichtsmassnahme“ die Irrtumswahrscheinlichkeit genau einhält?

f **Rangsummentest** von Wilcoxon, Mann und Whitney (U-Test),

$$T(\vec{g}, \vec{y}) = \sum_{g_i=1} R_i = \sum_i g_i R_i ,$$

Recht robust \longrightarrow Test der Wahl für das 2-Stichpr.-Problem
Verteilung der Teststatistik unter H_0 wie gehabt.

g^* Hagel-Experiment: Komplizierte Teststatistik, zweidimensional
 \longrightarrow zweidim. Verwerfungsbereich.

4.4 Eine Stichprobe oder zwei verbundene

a Beispiel Tranquilizer.

Zielgrösse: „Hamilton depression scale factor IV“.
 9 Patienten, vor und nach Anwendung des Tranquilizers.

vorher ($X_i^{(1)}$)	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
nachher ($X_i^{(2)}$)	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29
Abnahme ($-Y_i$)	0.952	-0.147	1.022	0.43	0.62	0.59	0.49	-0.08	0.01

b **Verbundene Stichproben.**

Differenzen $Y_i = X_i^{(2)} - X_i^{(1)}$ **symmetrisch um 0** verteilt?

H_0 : Für jedes Y_i ist + und - Vorzeichen gleich wahrscheinlich.

$G_i =$ Vorzeichen, $|Y_i| =$ „ Y_i “ im Zwei-Stichproben-Problem.

Für jede Vorzeichen-Konstellation $\vec{g}^{(\ell)} = [g_1^{(\ell)}, \dots, g_n^{(\ell)}]$

ist $W. = 1/2^n$.

c Teststatistik $T(\vec{g}, \vec{z})$ festlegen,

$$g_i = +1 \text{ oder } = -1, z_i > 0.$$

Rand.-Vert. $P(T(\vec{G}, \vec{z}) = t) = \#\{\vec{g} \mid T(\vec{g}, \vec{z}) = t\} / 2^n$

- $T(\vec{g}, \vec{z}) = (1/n) \sum_i g_i z_i = \text{ave}_i \langle y_i \rangle$
entspricht dem t-Test für gepaarte Stichproben.
- $T(\vec{g}, \vec{z}) = \#\{i : g_i = 1\}$: Vorzeichentest.
- $T(\vec{g}, \vec{z}) = \sum_{i:g_i=1} R_i$, R_i : Rang von z_i :
Vorzeichen-Rangsummen-Test von Wilcoxon.

e Beispiel:

```
> wilcox.test(d.tranquillizer[,1], d.tranquillizer[,2],  
             paired=TRUE)  
      Wilcoxon signed rank test  
data:  d.tranquillizer[, 1] and d.tranquillizer[, 2]  
V = 40, p-value = 0.03906  
alternative hypothesis: true mu is not equal to 0  
knapp signifikant.
```

Achtung: „Vorher-Nachher-Vergleich“!

Richtig: Vergleich mit Kontrollgruppe oder Crossover-Versuch.

4.5 Schätzungen und Vertrauensintervalle

a **Modell:** Testfrage war: Ist Verteilung symmetrisch um 0?

Allgemeineres Modell: Verteilung symmetrisch um μ

$\Leftrightarrow Y_i - \mu$ symmetrisch um 0.

Test: Teststatistik $T\langle \vec{g}, \vec{y} - \mu \vec{1} \rangle$.

Grosse Werte = Abweichung von $H_0 : \mu$.

b Daraus ergibt sich eine **Schätzung:**

$$\hat{\mu} = \arg \min_{\mu} \langle T \langle \vec{g}, \vec{y} - \mu \vec{1} \rangle \rangle$$

- c Vorzeichen-Rangsummen-Test \longrightarrow Hodges-Lehmann-Schätzer.

Betrachte Walsh averages $(X_h + X_i)/2$.

$$\hat{\mu} = \text{med}_{h \leq i} \langle (X_h + X_i)/2 \rangle .$$

Beispiel Tranquilizier: 45 Walsh-Mittelwerte

-0.1470, -0.1135, -0.0800, -0.0685, -0.0350, 0.0100, ..., 1.022

Median $\hat{\mu} = 0.46$

d* Herleitung: $X_{[k]}$ k -t-kleinsten Wert.

$$X_{[k]} > 0, Z_{hk} = (X_{[h]} + X_{[k]})/2, h < k$$

$$Z_{hk} < 0, \text{ wenn } |X_{[h]}| > |X_{[k]}|.$$

$$\#\{Z_{hk} < 0\} = \#\{h \mid |X_{[h]}| < |X_{[k]})\} = R_{[k]} - 1$$

$$R_{[k]} = \#\{h \mid Z_{hk} > 0, h \leq k\}.$$

$$X_{[k]} < 0 \implies Z_{hk} < 0, \text{ wenn } h < k.$$

$$T(\vec{g}, \vec{z}) = \sum_{i: g_i=1} R_i = \#\{[h, k] \mid Z_{hk} > 0, h \leq k\}$$

Nullhypothese $\mu = \mu_0$:

$$T(\vec{g}, \vec{z}) = \sum_{i: g_i=1} R_i = \#\{[h, k] \mid Z_{hk} > \mu_0, h \leq k\}$$

Test am wenigsten signifikant, wenn dies $= \frac{n(n+1)}{2}$ ist

$$\longrightarrow \hat{\mu} = \text{median}\langle Z_{hk} \mid h \leq k \rangle.$$

f **Vertrauensintervall** für Vorzeichen-Rangsummen-Test:

Grenzen des Ann.bereichs von T : c und $c' = n(n+1)/2 + 1 - c$
 Vertrauensgrenzen = c -ter und c' -ter Walsh-Mittelwert.

Beispiel Tranquilizer: $c = 6$, $c' = 40$, Vertrauensintervall $[0.01, 0.786]$.

h Allgemeine Teststatistik $T(\vec{G}, \vec{z}^*; \mu)$: Betrachte

$$Q(\beta) = P\langle T(\vec{G}, \vec{z}^*; \mu) > T(\vec{g}^*, \vec{z}^*; \mu) \rangle - \beta$$

Schätzung = Nullstelle für $\beta = 0.5$.

Vertrauensgrenzen = Nullstellen für $\beta = 0.025$ und $\beta = 0.975$.

Lösbar!

4.6 Mehrere Stichproben

a Einfache Varianzanalyse.

Randomisierung = Zugehörigkeit der Beob. zu den Gruppen.
Anzahl der Beobachtungen in jeder Gruppe fest.

Ränge R_i der y_i unter allen Beobachtungen.

„Gruppensummen“ $\bar{R}_h = \text{ave}_{g_i=h} R_i$. $E\langle \bar{R}_h \rangle = (n+1)/2$.

Abweichungsquadrate $(\bar{R}_h - (n+1)/2)^2$, gew. Mittel

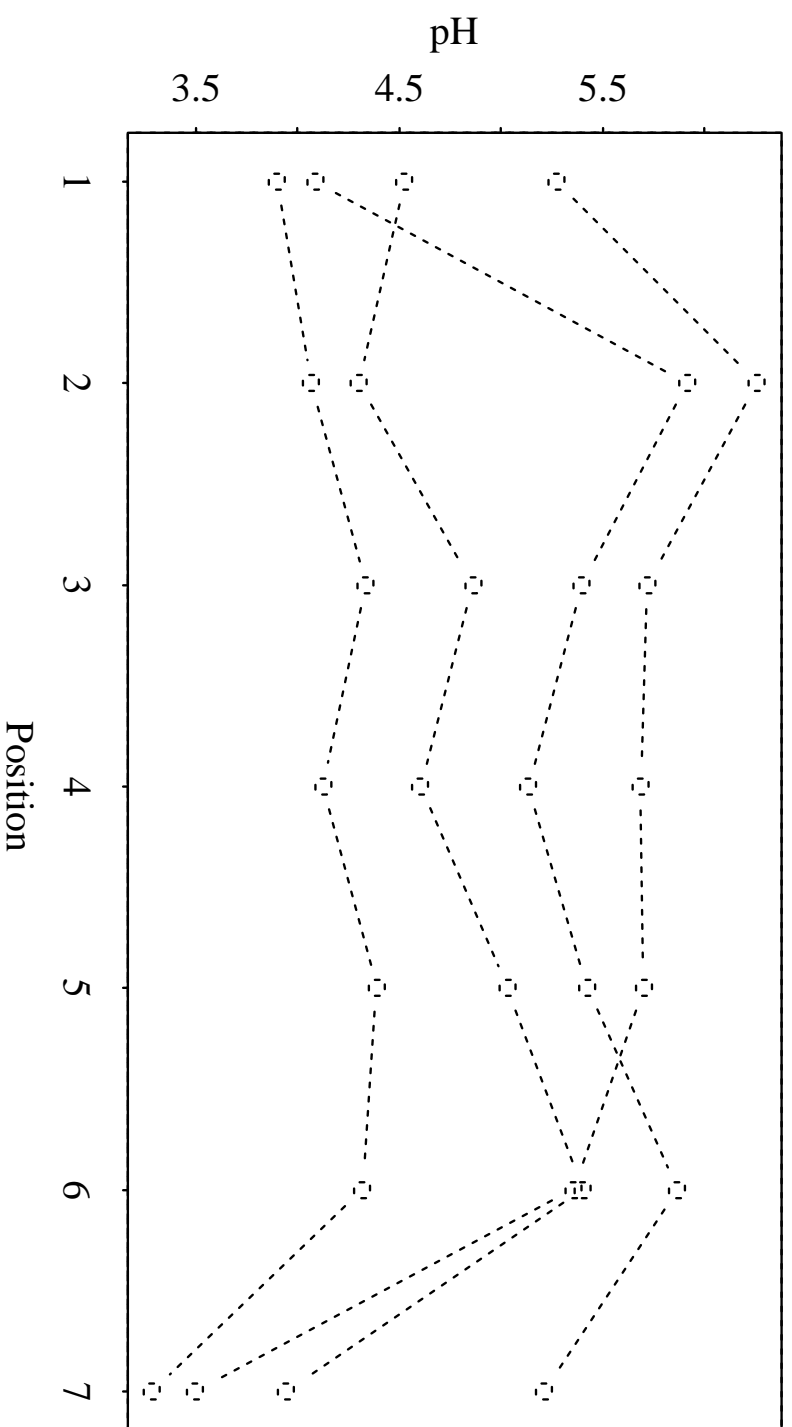
$$T\langle \vec{g}, \vec{y} \rangle = \frac{12}{n(n+1)} \sum_h m_h \left(\bar{R}_h - \frac{n+1}{2} \right)^2$$

Kruskal-Wallis-Test. 2 Stichproben \longrightarrow U-Test.

- b **Mehrere verbundene Stichproben** = einfacher Blockversuch.
 m Blöcke, m Bedingungen.
 Randomisierungen?

- c **Beispiel saure Böden**

Block	Position						
	1	2	3	4	5	6	7
1	4.09	5.91	5.40	5.13	5.43	5.87	5.21
2	3.90	4.07	4.34	4.13	4.39	4.32	3.29
3	5.27	6.26	5.72	5.69	5.70	5.36	3.50
4	4.53	4.30	4.86	4.61	5.03	5.40	3.95



Beispiel saure Böden

Friedman-Test. R_{ij} = Rang der Beobachtung j im Block i .

$\tilde{R}_j = \text{ave}_i \langle R_{ij} \rangle$ mittlerer Rang der „Stichprobe“ j .

$$T = \frac{12n}{m(m+1)} \sum_{j=1}^m (\tilde{R}_j - (m+1)/2)^2.$$

d

Block	Position						
	1	2	3	4	5	6	7
1	1	7	4	2	5	6	3
2	2	3	6	4	7	5	1
3	2	7	6	4	5	3	1
4	3	2	5	4	6	7	1
Summe	8	19	21	14	23	21	6
Mittel	2	4.75	5.25	3.5	5.75	5.25	1.5

```
> friedman.test(t.dt)
      Friedman rank sum test
data:  t.dt
Friedman chi-squared = 14.8, df = 6,
p-value = 0.02199
```

e Varianzanalyse

```
> summary(aov(pH~trans+pos, data=t.d))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trans	1	0.12	0.12	0.17	0.68
pos	1	0.18	0.18	0.27	0.61
Residuals	25	16.57	0.66		

4.7 Korrelation und Regression

a **Korrelation und einfache Regression.**

X_i, Y_i (X_i zufällig oder fest)

Nullhypothese: „kein Zusammenhang“

Randomisierung = „Paarung“ = Permutation von \vec{Y} .

W. jeder Permutation = $1/n! = 1/(n(n-1)\dots 2 \cdot 1)$.

Teststatistik:

- gewöhnliche Korrelation,
- Rangkorrelation,
- robuste Schätzung des Regressions-Koeffizienten, ...

- b **Multiple Regression:**
Permutation von \vec{Y} für Test der Hypothese, dass überhaupt kein Zusammenhang zwischen den Eingangs-Variablen und der Zielgrösse besteht.
- c **Zeitreihen:** Beobachtungen unabhängig?
Randomisierung: Permutation.
Testgrösse: z.B. erste Autokorrelation.
- d **Multiple Regression: Einzelner Koeffizient (oder mehrere)**
→ kein strikt richtiges Randomisierungsmodell.
- e* → Partielle Korrelation wie einfache Korrelation testen.
- f* Ein Test, der auf Rang-Methoden beruht
→ Jaeckel, Hettmansperger und McKeen.

g^* **Permutationen und andere Randomisierungen.**

Regression und Korrelation: Permutationen.

Bei zwei oder mehreren Gruppen: Auswahlen.

Permutationen: viel mehr;

viele führen zur gleichen Gruppenzugehörigkeit

→ gleiche Randomisierungs-Verteilung.

Hagelversuch: Anzahl potentielle Hageltage zufällig,

Anteil geimpfter zufällig.

Randomisierungs-Verteilung: Auswahlen von 33 aus 76 Tagen

→ **bedingter Test**, geg. die Anzahlen Impf- und Kontrolltage.

Merkpunkte

Randomisierungs-Tests

- **Randomisierungstests halten das Niveau exakt ein,** ohne Voraussetzungen an die Verteilung. (Unabhängigkeit von Beobachtungen vorausgesetzt.)
- Die **Teststatistik kann beliebig kompliziert sein.** Wahl mit (informellen) Überlegungen zur Macht. Robuste Teststatistik (z.B. aus Rängen) wählen!
- Es können auch Vertrauensintervalle konstruiert werden.