

Resampling-Methoden, nichtparametrische Tests und asymptotische Näherungen

Werner Stahel

Seminar für Statistik, ETH Zürich

und

Diego Kuonen

Statoo Consulting, Bern

Januar 2010

1 Einleitung

1.1 Annahmen über Verteilungen

- a Die wichtigste Aufgabe der Statistik besteht darin, anzugeben, wie man interessierende Grössen aus Daten ermittelt und Genauigkeitsangaben zu den Resultaten erhält. Dazu gehen wir von Wahrscheinlichkeitsmodellen für die beobachteten Daten aus und bestimmen daraus mit Wahrscheinlichkeitsrechnung die Verteilung der geschätzten Grössen.
- b Stochastische Modelle setzen jeweils eine gewisse Verteilung der darin enthaltenen Zufallsvariablen voraus – sonst sind es keine wirklichen Wahrscheinlichkeitsmodelle! Sehr häufig spielt dabei die Normalverteilung eine entscheidende Rolle. Ein Beispiel liefert die einfache lineare Regression, bei der die Zufallsabweichungen als normalverteilt angenommen werden.
- c Bei der Verwendung der Modelle sollte man die **Voraussetzungen überprüfen**, denn die Eigenschaften der statistischen Methoden, die für eine adäquaten Interpretation nötig sind, hängen von diesen Voraussetzungen ab.

Die Abhängigkeit kann mehr oder weniger stark sein. Die **Robustheit** einer Methode wird durch Masse charakterisiert, die die Stärke dieser Abhängigkeit angeben sollen. Den robusten Methoden ist im WBL ein eigener Block gewidmet.

- d Am liebsten hätte man Methoden, für die **Abweichungen von den Voraussetzungen überhaupt keine Rolle** spielen. Beispielsweise soll in der einfachen linearen Regression eine Gerade geschätzt werden; man sucht eine Schätzmethode, deren Eigenschaften nicht von der Form der Verteilung der Zufallsabweichungen abhängen.

Exakt kann man dieses hohe Ziel nur in wenigen Problemstellungen erreichen. Und immer werden gewisse Voraussetzungen bleiben, beispielsweise die Unabhängigkeit der Zufallsabweichungen und die Linearität des Zusammenhangs in der einfachen Regression.

- e Aus der Einführung ist bekannt, dass die **Rangsummen-Tests** – der Vorzeichen-Rangsummen-Test für eine oder zwei verbundene Stichproben und der Wilcoxon-Mann-Whitney-Test für zwei unabhängige Stichproben – „funktionieren“, ohne dass eine bestimmte Verteilungsform für die Zufallsvariablen angenommen werden muss. Sie erreichen also das hoch gesteckte Ziel – bis zu welchem Grad, wird in Abschnitt 2.2.b diskutiert.

Eine Schätzung einer „interessierenden Grösse“ oder eines Parameters ist mit diesen Verfahren jeweils indirekt verknüpft. Mit dem Vorzeichen-Rangsummen-Test hängt die „Hodges-Lehmann-Schätzung“ zusammen: Man bestimmt den Wert μ , für den die Test-Statistik für die durch μ gegebene Nullhypothese minimiert. Die Genauigkeitsangabe erhält man durch das Vertrauensintervall, das ja aus allen Werten von μ besteht, für die die durch μ gegebene Nullhypothese nicht verworfen wird. Wie der Test selbst sind die daraus abgeleitete Schätzung und ihr Vertrauensintervall nicht von der Verteilungs-Annahme für die Beobachtungen abhängig.

- f Ein anderer Grundgedanke, wie man die Annahme bestimmter Verteilungen vermeiden kann, besteht darin, dass man diese **Verteilung selbst aus den Daten schätzt**. Bereits für den wohlbekannten **t-Test** verwendet man eine einfache Variante dieses Grundgedankens: Wenn man nichts über die Varianz der Daten annehmen will (was meistens der Fall ist), dann schätzt man sie eben aus den Daten. Methoden, die in diesem Sinne die ganze Verteilung als „Störparameter“ behandeln, werden wir im Kapitel über den „**Bootstrap**“ kennen lernen.

Die so genannten **Randomisierungstests** verwenden noch einen anderen Trick, um von Verteilungsannahmen unabhängig zu werden, wie in Kapitel 2.L.c erläutert wird.

- g Der Bootstrap liefert in vielen Fällen eine Methode, um für eine interessierende Grösse eine Genauigkeitsangabe zu erhalten – sogar, wenn die Grösse recht kompliziert zu ermitteln ist. Immer wieder trifft man auch eine andere Methode an, um in solchen Fällen eine Genauigkeit zu ermitteln. Man benützt den Zentralen Grenzwertsatz und bestimmt die so genannte **asymptotische Verteilung**. Wir befassen uns in diesem Block auch damit, um die wichtigsten allgemein anwendbaren Methoden der Statistik, die sozusagen als **Universalwerkzeuge** in den Werkzeugkasten der Statistik gehören, beisammen zu haben.

1.2 Einführendes Beispiel

- a Das folgende Beispiel soll in diesem Block ab und zu verwendet werden.
 ▷ **Beispiel Ausfallzeiten** des Air conditioning-Systems in Flugzeugen des Typs Boeing 720 (aus Davison and Hinkley (1997)). Es wurden folgende $n = 12$ Zeiten zwischen Ausfällen notiert (nach Länge sortiert):

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

Figur 1.2.a zeigt ein Histogramm der rohen und der logarithmierten Daten. Die mittlere Ausfallzeit beträgt 108.1. Wäre die längste Zeit (487) nicht gemessen worden, so erhielte man bloss 73.6. ◀

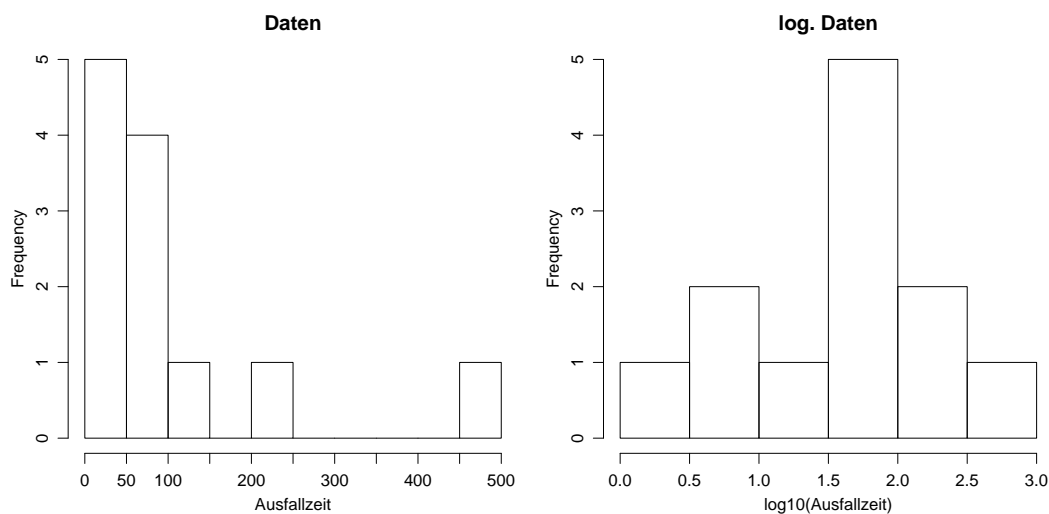


Abbildung 1.2.a: Histogramm der Air conditioning-Daten

- b Um die Genauigkeit des Mittelwertes (als Schätzung für den Erwartungswert) zu ermitteln, besteht das klassische Vorgehen darin, zunächst eine Verteilung für die Beobachtungen festzulegen. Für Ausfallzeiten eignet sich als einfachstes parametrisches Modell die **Exponential-Verteilung** Exp mit Dichte

$$f(y) = \frac{1}{\mu} e^{-y/\mu} \quad y > 0.$$

Passt dieses Modell? Um diese Frage zu beantworten, schätzen wir den Parameter und zeichnen entsprechende Dichte-Kurven ins Histogramm ein (Figur 1.2.b).

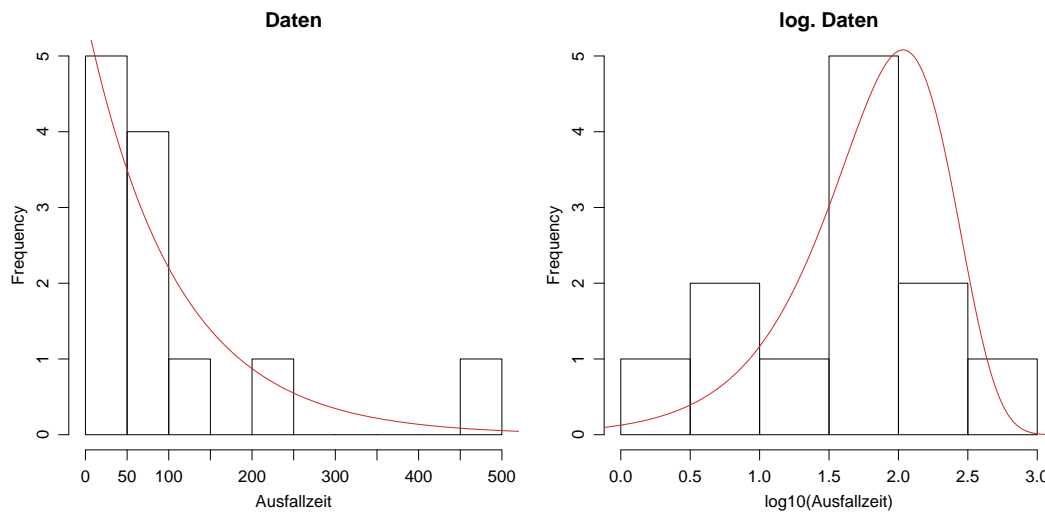


Abbildung 1.2.b: Histogramm der Air conditioning-Daten mit Dichten der angepassten Exponential-Verteilung

Die Modelle passen nicht schlecht. Aber der Datensatz ist zu klein, um einigermaßen sicher zu sein, dass die eine oder andere Verteilungsannahme stimmt.

Wir möchten die Genauigkeit des Mittelwertes angeben können, ohne die Verteilungsannahme machen zu müssen

Aus Gründen der Robustheit können wir daran interessiert sein, nicht den Mittelwert zur Charakterisierung der Länge der Ausfallzeiten zu verwenden, sondern ein gestutztes Mittel. Die Genauigkeit des gestutzten Mittels wäre auch dann schwierig zu ermitteln, wenn wir die Exponential-Verteilung voraussetzen würden. Einige Methoden dieses Blocks werden es erlauben, die Verteilung einer fast beliebig kompliziert aus den Daten ausgerechneten Größe zu schätzen.

2 Nichtparametrische Tests

2.1 „Nichtparametrisch“ heisst Vieles

- a Das Adjektiv „nichtparametrisch“ wurde in der Statistik auf viele Arten verwendet, um eine Methodik von der **parametrischen** Statistik abzugrenzen. Diese entspricht dem klassischen Vorgehen, wie es auch im Einführungsteil besprochen wurde:

Um eine Situation mit „Zufallswirkung“ zu erfassen, setzt man ein Modell an, das aus einem „strukturellen“ und einem „zufälligen“ Teil besteht. Der zufällige Teil wird durch Zufallsvariable beschrieben, deren Verteilung aus einer vorgegebenen **parametrischen Verteilungsfamilie** stammt, beispielsweise aus der Familie der Normalverteilungen. Ein Parameter (oder mehrere) wird durch den strukturellen Teil festgelegt. Der strukturelle Teil besteht beispielsweise aus einer Gruppierung mit einem „Lageparameter“ pro Gruppe oder aus einer Regressionsfunktion, die die Lageparameter der Zufallsvariablen festlegt und selber wieder neue Parameter enthält.

- b **Nichtparametrische Tests** verzichten darauf, für die Verteilung der Zufallsvariablen eine parameterische Verteilungsfamilie festzulegen. Ein weniger üblicher, aber klarerer Name dafür ist „**verteilungsfreie Tests**“.

Die im Einführungsteil besprochenen Rangtests gehören in diese Klasse. Sie benützen den naheliegendsten „Trick“, zu solchen Verfahren zu kommen: die Rangtransformation der Daten (oder abgeleiteter Grössen). Wir kommen auf diese Idee zurück (2.2).

- c „**Verteilungsfreie Tests**“ hat leider auch eine andere Bedeutung: Es sind Tests mit Teststatistiken, die für alle Verteilungen gleich sind. Es geht dabei gerade um Tests, die die Güte der Anpassung von Daten an eine bestimmte Verteilung prüfen (**goodness of fit tests**). Mehr dazu in 2.4.

Ein bekannter Anpassungstest ist der so genannte **Chiquadrat-Test**. (Bekanntlich gibt es viele „Chiquadrat-Tests“; man sollte also von Chiquadrat-Anpassungs-Test reden.)

- d Eine ganz andere Art, den Begriff nichtparametrisch zu verwenden, bezieht sich auf die Modellierung des **strukturellen Teils**. Statt einer parametrisierten Regressionsfunktion, wie wir sie in der Varianzanalyse, der gewöhnlichen linearen Regression, den verallgemeinerten linearen Modellen und der nichtlinearen Regression angetroffen haben, wird die Regressionsfunktion allgemeiner angesetzt – beispielsweise soll sie eine glatte, aber sonst beliebige Funktion sein. Wie man solche Vorstellungen konkretisiert, wird in der **nicht-parametrischen Regression** überlegt.

- e Methodisch eng mit der nichtparametrischen Regression verwandt ist das Problem der **Dichte-Schätzung**, das wieder den zufälligen Teil behandelt und versucht, eine Verteilung aus den Daten herauszulesen, die nicht aus einem parametrischen Modell stammt.

- f In diesem Kapitel beschäftigen wir uns mit nichtparametrischen Tests und daraus hergeleiteten Methoden. Von den Fragestellungen her werden wir eine ganze Reihe verschiedener Situationen antreffen: Varianzanalyse, Korrelation und Prüfung einer Verteilungsannahme.

2.2 Rang-Methoden

- a Die bekanntesten **Rangtests** haben wir bereits behandelt:
- Den Vorzeichen-Rangsummen-Test von Wilcoxon für zwei verbundene oder für eine einfache Stichprobe und
 - den Rangsummentest von Wilcoxon, Mann und Whitney (U-Test) für zwei unabhängige Stichproben.

Sie werden hier nicht nochmals beschrieben.

- b Die Rangsummentests vermeiden Voraussetzungen über die Verteilung. Das erreichen sie, indem sie von Daten, die einer Verteilung \mathcal{F} folgen, zu Rängen übergehen, die für jede Stichprobe vom Umfang n die Zahlen 1 bis n liefern.

Aber ganz ohne Voraussetzungen kommen sie doch nicht aus.

- Der **Vorzeichen-Rangsummentest** setzt die **Symmetrie** der Verteilung voraus. Wenn es um verbundene Stichproben geht und die Differenz zwischen den jeweiligen Werten getestet wird, ist die Symmetrie ihrer Verteilung unter der Nullhypothese meistens eine harmlose, äusserst plausible Annahme. Für den Fall einer einzelnen Stichprobe, aus dem auch die Schätzung eines „Lage-Parameters“ entsteht, ist die Symmetrie immer noch eine starke Annahme. Im Beispiel der Ausfallzeiten (1.2.a) ist sie offensichtlich verletzt – auch für die logarithmierten Werte.
 - Beim **U-Test** wird vorausgesetzt, dass die Beobachtungen unter der Nullhypothese alle die **gleiche Verteilung** haben. Es gibt Anwendungen, in denen damit gerechnet werden muss, dass die Streuungen in den beiden Gruppen unterschiedlich ist. In einem solchen Fall gibt es einen entsprechenden t-Test, der auch gleich die Annahme der Normalverteilung macht. Der U-Test ist dagegen in einem solchen Fall, genau genommen, nicht anwendbar. (Wie gross die Auswirkungen der verletzten Annahme sind, ist damit noch nicht gesagt.)
- c Die Verallgemeinerung des Zwei-Stichproben-Problems auf **mehrere Stichproben** führt zur **einfachen Varianzanalyse**. Genau wie beim U-Test kann man auch für dieses Problem die Daten durch ihre Ränge ersetzen und dann die Varianzanalyse durchführen – mit der nötigen Änderung der Verteilung der Teststatistik unter der Nullhypothese.
- d ▷ **Beispiel Nervenzellen.** (Aus Stahel (2007), Kap. 12.) Während der Entwicklung des Nervensystems und bei Regenerationsprozessen spielt das Auswachsen von Nervenzellfortsätzen (Neuriten) eine wichtige Rolle. In einem Versuch wurde die Wirkung verschiedener Stoffe auf diesen Prozess in Zellkulturen untersucht. Figur 2.2.d und Tabelle 2.2.d zeigen die Messwerte.

* Kleine Hirnstückchen (Explantate) wurden auf Gläschen in einer Kulturschale verteilt. Aus diesen Explantaten wandern Nervenzellen vorzugsweise in Richtung auf ein anderes Explantat aus. Zwischen benachbarten Explantaten wurden deshalb Felder abgegrenzt, in denen u.a. der längere Fortsatz jeder ausgewanderten Nervenzelle nach 4 Tagen in Kultur ausgemessen wurde. Für jedes Feld wurde als Mass für das Wachstum der Median gebildet (Quelle: S. Magyar, Doktorarbeit 1993, ETH Zürich).

Unterscheiden sich Felder, die aus verschiedenen behandelten Kulturen stammen? <

	Mediane																		
1	22.5	20.9	35.6	11.6	18.1	33.8	28.6	36.7	16.3	17.8									
2	8.6	43.0	19.3	25.5	26.1	24.4	32.0	16.5	17.5	19.2	8.6	27.5	14.9	23.1	25.8	15.9	27.2	6.8	
3	23.1	23.9	16.3	26.9	25.5	31.5	8.0	28.4	14.3	31.2	29.8	19.9	10.0	21.2	28.4	7.3	8.5	14.6	
4	16.3	30.9	21.2	9.4	45.3	27.6	29.9	15.2	13.2	15.1	26.7	25.2	24.5	19.2	6.4	21.0			
5	27.6	11.2	10.6	10.7	17.1	8.0	14.9	21.8	28.7	9.7	12.7	6.1	13.6	11.3	10.3	46.1	9.4		

Tabelle 2.2.d: Mediane von Neuritenlängen über „Felder“ für verschiedene Versuchsbedingungen

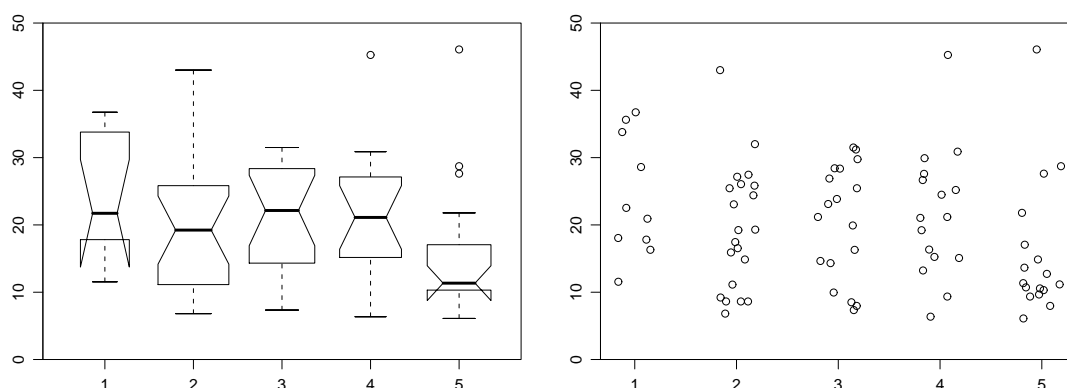


Abbildung 2.2.d: Gekerbte Kistendiagramme und Einzeldaten (rechts) mit horizontal verzerrten Werten für das Beispiel der Nervenzellen

- e Formulieren wir also das Test-Rezept für den Rangtest, der den Namen **Kruskal-Wallis-Test** trägt:

H_0 : $Y_{h,i} \sim \mathcal{F}$ (*i.i.d.*), wobei \mathcal{F} eine beliebige Verteilung ist – die gleiche für alle Beobachtungen in allen Stichproben.

H_A : $Y_{h,i} \sim \mathcal{F}_h$, wobei die \mathcal{F}_h bis auf eine Verschiebung die gleichen Verteilungen sind, $F_h(x) = F_1(x - \delta_h)$, wobei mindestens ein $\delta_h \neq 0$ ist.

U : Die Teststatistik wird wie folgt gebildet (vergleiche Beispiel):

- Bestimme den Rang $R_{h,i}$ bezüglich der „vereinigten Stichproben“ für jede Beobachtung $Y_{h,i}$,

$$R_{h,i} = \text{Rang}\langle Y_{h,i} \mid Y_{1,1}, \dots, Y_{1,n_1}, Y_{2,1}, \dots, Y_{g,n_g} \rangle .$$

- Bilde die Mittelwerte der Ränge für die einzelnen Gruppen, $\bar{R}_h = \sum_i R_{h,i} / n_h$. Ihr mit der Gruppengröße n_h gewichteter Mittelwert ist gleich dem Mittelwert aller Ränge, $(n + 1)/2$. Bilde als Testgröße die mit den n_h gewichtete Quadratsumme der Abweichungen $\bar{R}_h - (n + 1)/2$,

$$U = \sum_{h=1}^m n_h \left(\bar{R}_h - \frac{n+1}{2} \right)^2 .$$

$\mathcal{F}_0\langle U \rangle$: Die Verteilung von U hängt nicht von der Verteilung \mathcal{F} der $Y_{h,i}$ ab. Für kleine Stichproben kann man sie mit Hilfe der Kombinatorik berechnen, für grössere benützt man die standardisierte Teststatistik $T = \frac{12}{n(n+1)} U$, die sich auch mit den Rangsummen $S_h = \sum_i R_{h,i}$ schreiben lässt als

$$T = \frac{12}{n(n+1)} \sum_{h=1}^m \frac{S_h^2}{n_h} - 3(n+1).$$

Diese Grösse ist asymptotisch chiquadrat-verteilt mit $m - 1$ Freiheitsgraden, wenn m die Anzahl Gruppen ist. Das gilt allerdings nur, wenn F stetig ist und deshalb keine (wenige) „Bindungen“ auftreten. (Bindungen führen zu gleichen, „aufgeteilten“ Rängen. Solche stören nur, soweit sie für Beobachtungen aus verschiedenen Gruppen auftreten.) Wenn viele Bindungen auftreten, gibt es einen korrigierten Standardisierungsfaktor für T , siehe beispielsweise Hartung und Elpelt (1997, Kap. XI.1.1.B).

- f ▷ Im **Beispiel der Nervenzellen** erhält man von der R-Funktion `kruskal.test` die folgenden Angaben

```
> kruskal.test(medMaxL~treat, data=d.neurit)
Kruskal-Wallis rank sum test
data: medMaxL by treat
Kruskal-Wallis chi-squared = 7.2709, df = 4, p-value = 0.1222
```

Eine gewöhnliche einfache Varianzanalyse liefert für den F-Test einen p-Wert von 0.208. Die Daten zeigen 3 Ausreisser. Wenn man sie weglässt (was einem allzu liberalen Umgang mit den Daten gleichkommt), dann liefert die einfache Varianzanalyse einen p-Wert von 0.022, während der Rangtest bei 0.044 bleibt. ◁

- g Die naheliegende nächste Fragestellung ist die **Zweiweg-Varianzanalyse**. Zunächst betrachten wir einen Block-Versuch.

▷ Als **Beispiel** soll eine typische Fragestellung der **Sensorik** dienen: Um die Beliebtheit von 40 Produkten zu testen, wurden diese einem „Panel“ von 117 interessierten Personen vorgelegt. In diesem Falle ging es nicht um den sensorischen Geschmack, sondern um den ästhetischen Gesamteindruck von Feingebäck, der von jeder Testperson für jedes Produkt auf einer Skala von 1 bis 10 bewertet wurde. Wir werten hier die Daten für die 12 Produkte eines von vier Produkttypen aus. Figur 2.2.g zeigt, wie die Produkte abschnitten. Es stellt sich die Frage, ob sich die Produkte signifikant unterscheiden.

Die Testpersonen nützen die Skala nicht gleichmässig aus. Einige geben nur gute Beurteilungen ab, während andere stärker differenzieren. Es ist deshalb naheliegend, diesen Effekt zu eliminieren, indem man die Testperson als Block im Sinne eines Block-Versuchs auffasst. ◁

- h Analog zum Fall von zwei verbundenen Stichproben kann man von **mehreren verbundenen Stichproben** sprechen.

Neben dem Effekt des interessierenden Behandlungsfaktors wird ein Block-Effekt im Modell berücksichtigt. Das klassische Modell der Zweiweg-Varianzanalyse war

$$Y_{h,i} = \mu + \alpha_h + \beta_i + E_{h,i}.$$

Wenn sich die Behandlungen h stark unterscheiden, dann werden die Rangordnungen der $Y_{h,i}$ in jedem Block i die gleichen sein. Das ist die Idee, die hinter dem **Friedman-Test** steht. Er geht folgendermassen:

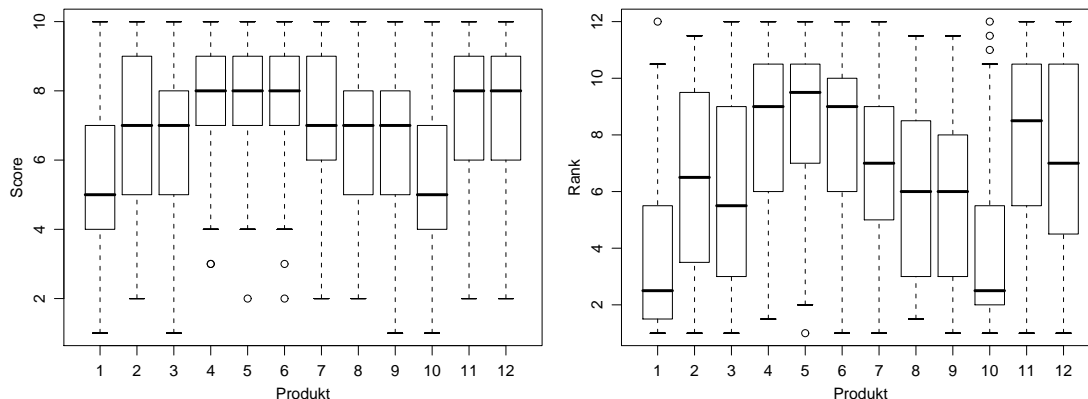


Abbildung 2.2.g: Daten des Beispiels aus der Sensorik. Die von den 117 Testpersonen abgegebenen Beurteilungen werden für jedes Produkt als Boxplot wiedergegeben (links). Rechts sind die Daten nach der Transformation auf Ränge dargestellt.

H_0 : Der Behandlungsfaktor hat keinen Effekt, $Y_{h,i} \sim \mathcal{F}_i$, unabhängig von h .

H_A : Wie oben für den klassischen Fall angegeben, mit einer beliebigen Verteilung für die $E_{h,i}$. Mindestens ein $\alpha_h \neq 0$.

U : Bilde die Ränge $R_{h,i}$ innerhalb jedes Blockes i (also anders als für den Kruskal-Wallis-Test). Dann sei $S_h = \sum_i R_{h,i}$ die Summe der Ränge für die Behandlung h . Die Summe der quadrierten Abweichungen der S_h von ihrem Erwartungswert $b(m+1)/2$ (bei m Behandlungen und b Blöcken) bilden die unstandardisierte Teststatistik

$$U = \sum_h (S_h - b(m+1)/2)^2 .$$

$\mathcal{F}_0\langle U \rangle$: Standardisiert man diese Testgröße nach

$$T = \frac{12}{bm(m+1)} U ,$$

dann kann man als genäherte Verteilung die Chiquadrat-Verteilung mit $g-1$ Freiheitsgraden verwenden. Für kleine Blockzahlen b kann man die exakte Verteilung berechnen oder Tabellen verwenden.

- i ▷ Im **Beispiel der Sensorik** ergibt der Test eindeutig signifikante Unterschiede zwischen den Produkten – was zu erwarten war und auch in der Abbildung zum Ausdruck kommt. Das Ergebnis der S-Funktion sieht so aus:

```
> friedman.test(as.matrix(t.dt))
Friedman rank sum test
data: as.matrix(t.dt)
Friedman chi-squared = 270.1325, df = 11, p-value < 2.2e-16
```

Der Test beruht ja darauf, dass die Daten zunächst innerhalb jeder Testperson auf Ränge transformiert werden. Die transformierten Daten sind, aufgeschlüsselt nach dem Produkt, in Abbildung 2.2.g rechts dargestellt. ◁

- j Leider gibt der Test keine Auskunft über die Frage, wie gross eine Differenz zwischen mittleren Rängen sein muss, damit der Unterschied noch als statistisch gesichert gelten kann. Eine solche Angabe erhält man, wenn man die rang-transformierten Daten mit gewöhnlicher Varianzanalyse untersucht. Das ist zwar eine fragwürdige Analyse, da die rang-transformierten Daten unter der Nullhypothese eher eine diskretisierte uniforme Verteilung als eine Normalverteilung zeigen sollten. Nun, Verteilungen mit kürzeren Schwänzen als die Normalverteilung sind für die Varianzanalyse ungefährlich. Ein Vergleich der P-Werte, die man mit dieser Art der Varianzanalyse erhält, mit denjenigen des Friedman-Tests zeigt (bei diesem Beispiel, auch mit Teildatensätzen) nur kleine Unterschiede.

In Figur 2.2.j sind die Mittelwerte der Ränge für die Produkte zusammen mit einem Balken dargestellt, der den kleinsten signifikanten Wert für eine Differenz zwischen zwei solchen Mittelwerten zeigt. Bei der Interpretation ist es wichtig, das Problem der multiplen Vergleiche zu beachten: Es würde auch unter der Nullhypothese der extremste Unterschied meistens signifikant erscheinen.

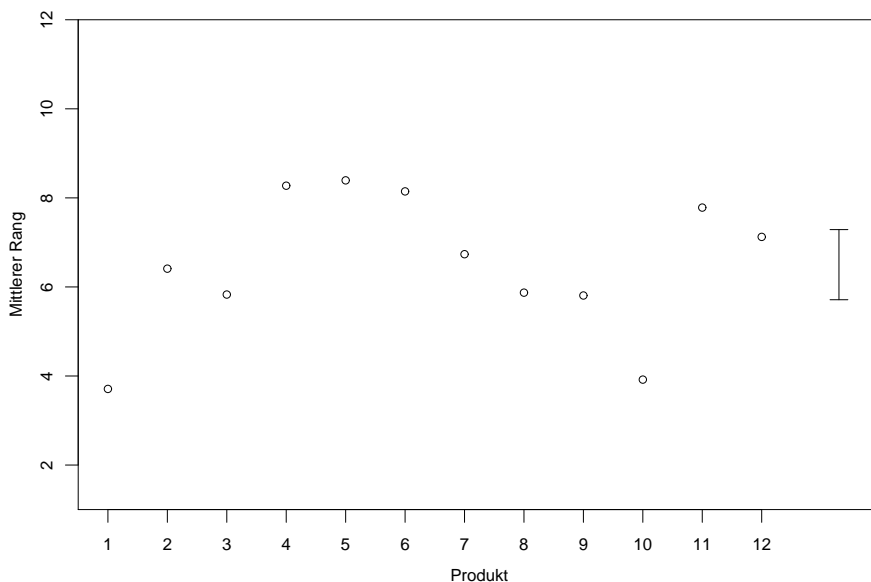


Abbildung 2.2.j: Mittelwerte der Ränge für die Produkte. Der Balken rechts zeigt den kleinsten signifikanten Wert für eine Differenz zwischen zwei Mittelwerten.

- k Für den Fall von zwei verbundenen Stichproben wurde der Vorzeichen-Rangsummen-Test als der geeignete Rangtest eingeführt. Was wird aus dem Friedman-Test, wenn $m = 2$ ist? Für jeden Block gibt es dann nur zwei Fälle: Entweder ist die Zielgrösse $Y_{h,i}$ für $h = 1$ grösser als für $h = 2$, was zu $R_{1,i} = 2$ und $R_{2,i} = 1$ führt, oder umgekehrt. Daraus lässt sich leicht folgern, dass der Friedman-Test in diesem Fall dem **Vorzeichentest** entspricht. Verglichen mit dem Vorzeichen-Rangsummen-Test ist dieser viel weniger mächtig und also ungeeignet.

Ein effizienter Test wurde von Doksum entwickelt. Er ist aber wenig bekannt. Siehe Hollander and Wolfe (1999, Ch. 7.11).

- l Nun wäre die Fortsetzung für **kompliziertere Varianzanalysen** fällig. Entsprechende Verfahren gibt es zwar, sie sind aber nicht gebräuchlich.

- m* **Rang-Regression.** In der multiplen linearen Regression kann die Quadratsumme der Residuen, die durch die Methode der Kleinsten Quadrate minimiert wird, durch eine andere Zielfunktion ersetzt werden. Dies wurde im Block über robuste Regression ausgeführt. Jaeckel führte eine Zielfunktion ein, die auf den Rängen der Residuen beruht. Die entsprechenden Schätzungen heissen **R-Schätzungen**. Ihre asymptotische Verteilung hängt nur von einem speziellen Aspekt der Fehlerverteilung ab (von der Dichte der Verteilung bei $E = 0$) und ist deshalb „teilweise verteilungsfrei“. Genauerer siehe Hettmansperger (1984).

2.3 Rangkorrelation

- a Ränge können auch zur Beschreibung der Beziehung zwischen zwei Variablen eingesetzt werden.

Die Rangkorrelation von **Spearman** ist definiert als die gewöhnliche („Pearson“-) Korrelation von rang-transformierten Daten. Man bildet also zunächst für jede Variable j die Ränge $R_i^{(j)}$ der Daten $X_i^{(j)}$ und berechnet dann deren gewöhnliche Korrelation.

Die gewöhnliche Korrelation ist dann gut interpretierbar, wenn die Daten bivariat normalverteilt sind (siehe Stahel (2007, Bild 3.2.i)). Wenn das der Fall ist, liefert die Rangkorrelation sehr ähnliche Werte. Andererseits ist die Rangkorrelation viel robuster als die gewöhnliche Korrelation. Transformiert man eine Variable (oder beide) monoton, dann ändert sie sich nicht. Sie bildet daher ein **Mass für die Stärke einer monotonen Beziehung** zwischen zwei Variablen.

Eine andere Definition eines Korrelationsmasses, das diese beiden Eigenschaften besitzt, stammt von **Kendall**.

- b* Es gibt auch Rang-Methoden für die multivariate Statistik. Da aber nicht eindeutig ist, wie im mehrdimensionalen Raum eine „Ordnung“ eingeführt werden soll, gibt es hier recht viele Vorschläge mit verschiedenen Vor- und Nachteilen. Die Forschung ist nicht konsolidiert.

2.4 Anpassungstests

- a Die Frage, ob die Daten mit der **Annahme einer bestimmten Verteilung** verträglich sind, wurde schon oft untersucht. Bisher haben wir meistens **QQ-Diagramme** verwendet und auf statistische Tests verzichtet. Die Rechtfertigung dieser „lockeren“ Haltung bestand darin, dass mit einem Test ja nicht nachgewiesen werden kann, dass eine bestimmte Verteilung wirklich den Daten zu Grunde liegt, sondern nur, gegebenenfalls, dass die Daten einer solchen Annahme widersprechen. Die Versuchung einer **missbräuchlichen Interpretation** eines Tests für die „Anpassung“ an eine bestimmte Verteilung ist gross: Wenn der Test keine signifikante Abweichung angibt, gilt die Voraussetzung als gegeben oder gar bewiesen. Trotzdem ist es an der Zeit, dass wir solche Verfahren genauer betrachten.

- b **Kolmogorov.** Die theoretische Verteilungsfunktion F charakterisiert die zu überprüfende Verteilung, also die Nullhypothese. Die empirische Verteilungsfunktion ist gegeben durch

$$\hat{F}_n(x) = \frac{\#\{i \mid X_i \leq x\}}{n}.$$

Wenn nun die empirische Verteilungsfunktion stark von der theoretischen abweicht, schliessen wir auf Verletzung der Nullhypothese. (Für das QQ-Diagramm haben wir die Quantilfunktionen, also die Umkehrfunktionen F^{-1} und \hat{F}_n^{-1} verglichen.) Kolmogorov führte deshalb als Teststatistik die Grösse

$$T = \max_x \langle |\hat{F}_n(x) - F(x)| \rangle$$

ein.

Es ist leicht einzusehen, dass die Verteilung dieser Grösse, unter der Annahme, dass die Beobachtungen X_i gemäss F verteilt sind, nicht von F abhängt (sofern F eine stetige Verteilung ist): Wenden wir auf die X_i irgendeine (differenzierbare) monotone Transformation g an. Wenn die Nullhypothese gilt, haben die transformierten Beobachtungen $Y_i = g(X_i)$ die kumulative Verteilungsfunktion $F^{(Y)}(y) = F(g^{-1}(y))$. Die Teststatistik T , berechnet für die Y_i und $F^{(Y)}$, ist gleich dem Wert von T für die X_i und F . Setzt man beispielsweise $g = F^{-1}$, so wird $F^{(Y)}$ gleich der Verteilungsfunktion für die uniforme Verteilung zwischen 0 und 1. Man kann die Verteilung der Teststatistik also für diese spezielle Verteilung berechnen und weiss, dass diese für alle Verteilungen F die gleiche sein muss.

Kolmogorov leitete 1933 die asymptotische Näherung an die Verteilung von T mit sehr eleganten und grundlegenden Methoden her. Die Bedeutung der mathematischen Methoden, die diesem Test zu Grunde liegen, ist wesentlich grösser als dessen praktische Bedeutung, wie wir gleich erläutern werden (2.4.e).

- c Üblicherweise will man nicht eine bestimmte Verteilung prüfen, sondern nur eine **Verteilungsform**, beispielsweise die Normalverteilung, mit beliebigen Werten für die Parameter. Diese müssen aus den Daten geschätzt werden, was für kleinere Stichproben natürlich Auswirkungen auf die Verteilung der Teststatistik hat (wie dies grundlegend beim t-Test diskutiert wurde, siehe Absatz 8.5.g in Stahel (2007)). Die Korrekturen hängen nun von der Verteilungsfamilie ab, siehe Kommentar 11.5.(28) in Hollander and Wolfe (1999).
- d **Chiquadrat-Test**. Siehe Abschnitt 10.2 in Stahel (2007).
- e **Welchen Test wählen?** Der Chiquadrat-Test ist dem Kolmogorov-Test generell vorzuziehen, da er – bei geeigneter Wahl der Klassen – gegen die üblicherweise wichtigen Alternativen von langschwänzigen Verteilungen bessere **Macht** zeigt. (Genauere Untersuchungen dazu muss ich allerdings noch suchen.)

2.5 Ausblick

- a Eine Durchsicht von Hollander and Wolfe (1999) zeigt weitere nichtparametrische Methoden (im Sinne von verteilungsfrei):
 - Test gegen unterschiedliche Streuungen und allgemeinere Alternativen für zwei unabhängige Stichproben, z.B. Vergleich zweier empirischer Verteilungsfunktionen (Kolmogorov-Smirnov-Test, verwandt mit dem Kolmogorov-Anpassungs-Test),
 - Varianzanalyse: Tests gegen speziellere Alternativen (monotone Effekte für einen geordneten Faktor u.a.), Methoden für Kontraste,
 - Methoden für Überlebenszeiten (vgl. Block Überlebenszeiten)

L Literatur

- a Ein neues, umfassendes Buch über nichtparametrische Statistik stammt von Hollander and Wolfe (1999). Jedes Verfahren ist zunächst rezeptartig beschrieben. Die Motivation, Eigenschaften, Theorie und andere Bemerkungen folgen nachher.
- b Ein älteres Buch in deutscher Sprache, das alle wichtigen Verfahren enthält, ist Büning und Trenkler (1994).
- c Hettmansperger (1984) diskutiert Rang-Methoden umfassend. Es gibt in diesem Gebiet auch neuere Forschungsergebnisse.

3 Randomisierungs- und Rangtests

3.1 Einführendes Beispiel

- a Verringert das „Impfen“ von potentiellen Hagelwolken mit Silberiodid die Hagelenergie? Das war die Fragestellung im „Grossversuch IV“, einem Feldexperiment, das 1978-1983 im Napfgebiet durchgeführt wurde.

[Wir stellen dieses Beispiel so ausführlich dar, dass es, zusammen mit dem Anhang, allenfalls als Material in einem Kurs über Statistik im Gymnasium verwendet werden kann.]

Um diese Frage wissenschaftlich sauber zu beantworten, verwendet man die folgende Grundidee: Man misst die Hagelenergie (indirekt, durch Radarbeobachtung) für n Wolken, wobei man eine zufällige Auswahl von $n/2$ Wolken „impft“, während man die andern ungestört lässt. (Genauerer siehe Federer et al., Pure and Applied Geophysics 117 (1978/79), 548-571).

- b Wir halten die folgenden Daten fest:

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

(G ist die „Indikatorvariable“ für das Impfen.)

Die Hoffnung der durch Hagel Betroffenen lautet: Die Y_i , für die Beobachtungen i mit $G_i = 1$ fallen tendenziell niedriger aus als die anderen.

- c Beobachtet wurde

$$\begin{array}{l} Y_i = y_i^* \\ G_i = g_i^* \end{array} \left| \begin{array}{cccccccc} 16672 & 25 & 855 & 0 & 152 & 0 & 46 & 1219 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \end{array} \right.$$

Die g_i^* repräsentieren die „Zufallsauswahl“ der zu impfenden Wolken. [Man kann der Frage, wie man eine **Zufallsauswahl** erzeugt, genauer nachgehen, vgl. Stahel (2007), Abschnitt 4.4 u.a.] In Wirklichkeit umfasste das Experiment 216 Wolken; davon wurden 94 geimpft.

- d Die Fragestellung ruft nach einem **statistischen Test**. [Im Anhang 3.A wird die Idee des statistischen Tests „ohne vorausgesetzte Kenntnisse in Statistik“ eingeführt.] Wir haben die wohlbekannte Situation des ungepaarten Zwei-Stichproben-Problems vor uns. Der Griff nach dem t-Test ist verführerisch – aber wenn man die Daten anschaut, nicht zu rechtfertigen, da eine Normalverteilung sehr unplausibel erscheint. Wir hätten gerne einen Test, der keine Annahmen über die Verteilung der Y_i voraussetzt.

3.2 Statistische Überlegung

- a Die **Nullhypothese** muss ein **Wahrscheinlichkeitsmodell** festlegen, das plausibel erscheint, wenn die „Hoffnung“ *nicht* richtig ist. Es gibt zwei prinzipiell verschiedene Möglichkeiten, ein Wahrscheinlichkeitsmodell für diese Situation aufzustellen. Die übliche, die dem t-Test zu Grunde liegt, besteht darin, die „Zielgrösse“ Y_i als zufällig zu betrachten und die Zugehörigkeit zu den Behandlungsgruppen als vorgegeben anzunehmen.

Die weniger übliche geht davon aus, dass für die Auswahl der zu impfenden Wolken (der zu behandelnden Einheiten) ein Zufallsmechanismus eingesetzt wurde. Die **Grundüberlegung** lautet: Wir beobachten bestimmte Werte y_i^* der Zielgrösse, wenn diejenigen Wolken geimpft werden, für die $g_i^* = 1$ ist. Wenn das Impfen keinen Einfluss auf die Hagelenergie hat, würden wir die genau gleichen Werte y_i^* erhalten, wenn die Wolken entsprechend den Werten

$$\underline{g}^{(1)} = [0, 1, 0, 0, 1, 1, 0, 1]$$

geimpft worden wären, oder entsprechend irgendeiner anderen Auswahl von zu impfenden Wolken. Diese Überlegung führt zum folgenden Wahrscheinlichkeitsmodell.

Wir betrachten die y_i^* als fest und die **Auswahl** $[G_1, \dots, G_n]$ **als zufällig**. Das ist sicher gerechtfertigt, wenn die Auswahl mit einem Zufallsmechanismus getroffen wurde, der (beispielsweise) jeder Auswahl von $n/2 = 4$ Elementen aus $n = 8$ Elementen gleiche Wahrscheinlichkeit gibt. In diesem Fall ist die Wahrscheinlichkeit für jede Auswahl $\binom{8}{4}^{-1} = \frac{1}{70}$. Damit ist die Nullhypothese festgelegt.

- b Nach dem allgemeinen Rezept (Stahel, 2007, Abschnitt 8.4) müssen wir als nächstes eine **Teststatistik** festlegen, die extreme Werte annimmt, wenn eine Alternative gilt. Falls das Impfen eine Wirkung zeigt, sind die Werte y_i^* , für die $g_i^* = 1$ ist, tendenziell kleiner als diejenigen mit $g_i^* = 0$. Die naheliegendste – wenn auch nicht die geeignetste – Grösse dieser Art ist die Differenz der Mittelwerte

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

- c Wie ist T unter der Nullhypothese verteilt? Wenn die y_1^*, \dots, y_n^* gegeben sind, gibt es (höchstens) $\binom{n}{n/2}$ mögliche Werte für T . Die Wahrscheinlichkeiten dieser Werte bestimmen sich nach der Regel „Anzahl günstige durch Anzahl mögliche Fälle“,

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

– meistens wird der Zähler =1 sein. Die Verteilung wird auch **Randomisierungs-Verteilung** genannt. Für das Beispiel ist sie in Figur 3.2.c dargestellt.

Der **Verwerfungsbereich** fasst die $\alpha = 5\%$ extremsten (z.B. die grössten) Werte zusammen. Das lässt sich allerdings meist nicht genau erreichen, da beispielsweise kein Vielfaches von $1/70$ gleich 0.05 ist. Also muss man genauer sagen: Der Verwerfungsbereich fasst die extremsten Werte zusammen, deren Wahrscheinlichkeit gesamthaft möglichst wenig kleiner als 5% ist. Im Beispiel ergibt sich als Verwerfungsbereich $\{t \mid t \geq 4643.25\}$ (für eine einseitige Fragestellung).

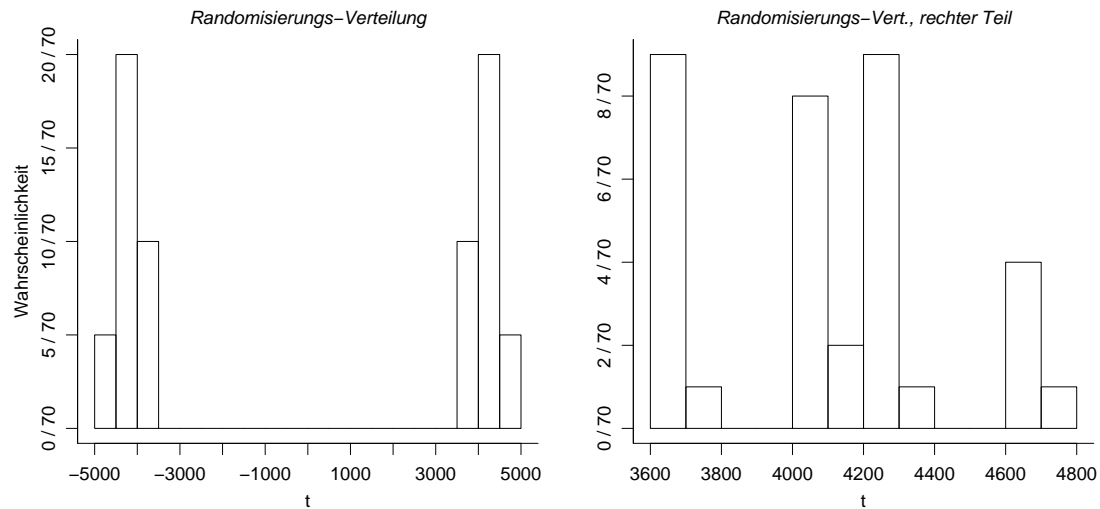


Abbildung 3.2.c: Randomisierungs-Verteilung im Beispiel. Links: ganze Verteilung, rechts: Verteilung der negativen Werte (Ausschnitt aus der Figur links)

- d Für die im Experiment getroffene Auswahl \underline{g}^* ergibt sich

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

also ein Effekt in die unerwartete Richtung! Da dies < 4643.25 ist, wird die Nullhypothese nicht verworfen; ein **Effekt der Impfung von Hagelwolken kann also nicht nachgewiesen werden**. (Wenn schon, könnte sich der Effekt in die unerwartete Richtung bei „umgekehrt einseitiger“ Fragestellung als signifikant erweisen, was aber auch nicht der Fall ist. Dies entspricht auch dem Ergebnis des gesamten Experiments.)

- e* In Wirklichkeit wurden 216 Wolken beobachtet, von denen 94 geimpft wurden. Um genau zu sein, wurden aus Gründen der statistischen Unabhängigkeit und der praktischen Durchführbarkeit jeweils für einen ganzen Tag bestimmt, ob alle Wolken an diesem Tag geimpft werden oder alle der Kontrollgruppe zugeordnet werden sollten. Die Teststatistik wurde dann für die mittlere Hagelenergie der Wolken eines Tages berechnet. Es wurden schliesslich von 76 „potentiellen Hageltagen“ 33 als Impftage ausgewählt.
- f Es ist auch für den stärksten Computer nicht möglich, für $\binom{76}{33} = 36 \cdot 10^{20}$ mögliche Auswahlen die Teststatistik zu berechnen und damit die Randomisierungs-Verteilung zu bestimmen. Man behilft sich (wie beim Bootstrap) mit **Simulation**, das heisst, man berechnet die Teststatistik für beispielsweise $r = 5000$ zufällig ausgewählte Zufallsauswahlen $\underline{g}^{(\ell)}$.

3.3 Tests für das Zwei-Stichproben-Problem

- a Ausgehend von diesem grundlegenden Beispiel wollen wir nun zeigen, dass das Prinzip des Randomisierungstest in wichtigen weiteren Situationen anwendbar ist.
- b Zunächst ist wichtig, festzustellen, dass **Randomisierungstests auch dann angewandt werden können, wenn die Durchführung des Versuchs keinen Randomisierungsschritt enthält**. Die einzigen Voraussetzungen, die dann gelten müssen, sind:
- Die Beobachtungen müssen *unter der Nullhypothese* gleich verteilt und
 - unabhängig sein.

Wenn diese Voraussetzungen gelten, **stimmt die gewählte Irrtumswahrscheinlichkeit α exakt**. Die Randomisierungstests bilden in diesem Sinne den „Goldstandard“ unter den statistischen Tests.

- c* Genauer gesagt genügt eine leicht schwächere Voraussetzung, genannt „Austauschbarkeit“ der Beobachtungen, an Stelle der beiden erwähnten.

Die Voraussetzungen beziehen sich auf die übliche Situation, in der die Beobachtungen, wie üblich, zufällig sind. Aus der Stichprobe $[Y_1, \dots, Y_n]$ kann man die geordnete Stichprobe $Y_{[1]}, \dots, Y_{[n]}$ oder die empirische kumulative Verteilungsfunktion \hat{F}_n (die auch beim Bootstrap verwendet wurde) bestimmen. Nun betrachten wir die Verteilung der Teststatistik, bedingt auf diese geordneten Beobachtungen oder eben auf \hat{F}_n – das ist die oben angegebene Randomisierungs-Verteilung der Teststatistik.

Nach Konstruktion beträgt nun die bedingte Wahrscheinlichkeit eines Fehlers erster Art, gegeben \hat{F}_n , (bis auf die oben erwähnten Unterschreitungen wegen Diskretheit der Randomisierungs-Verteilung) exakt α – für jede Bedingung \hat{F}_n , und deshalb auch ohne Bedingung.

- d Die **Teststatistik** kann **beliebig** gewählt werden. Bei den Überlegungen haben wir ja keine speziellen Eigenschaften der Teststatistik benutzt. Da die Daten offensichtlich nicht normalverteilt sind, ist die Differenz der Mittelwerte kaum die Grösse der Wahl. Sie ist ja „sehr unrobust“.

Wie wählt man eine optimale Teststatistik? Strikt lässt sich diese Frage nur beantworten, wenn man die Macht für die anvisierten Alternativen berechnet. Das lässt sich nicht tun, ohne wieder eine *bestimmte* Verteilung oder Verteilungsfamilie für die Y_i vorauszusetzen. Es ist dann nahe liegend, die Teststatistik des zugehörigen optimalen parametrischen Tests zu verwenden. (Dieser kann allgemein nach dem Prinzip des Likelihood-Ratio-Tests gefunden werden.)

Wenn die Teststatistik monoton transformiert wird, ändert sich am Test nichts, da die extremsten Randomisierungen (die \underline{g} , für die $T(\underline{g}, \underline{x})$ am extremsten ausfällt) die gleichen bleiben. Die Teststatistik hat die Aufgabe, für die Randomisierungen \underline{g} eine „Anordnung“ von der am wenigsten extremen zur „extremsten“ festzulegen. Die 5% extremsten Randomisierungen bilden dann den Verwerfungsbereich, und der P-Wert ist der Anteil der Randomisierungen, die in diesem Sinne extremer sind als die beobachtete.

- e Im Beispiel erscheint eine Logarithmus-Transformation der Daten angebracht (mit geeigneter Wahl des Transformationswertes von 0). Nachher kann die Mittelwertsdifferenz oder die Differenz für ein robustes Lagemass vernünftig erscheinen. Figur 3.3.e zeigt die Randomisierungs-Verteilung für die Mittelwerts-Differenz T' der logarithmierten Daten.

Macht die Änderung der Teststatistik einen Unterschied? In der rechten Hälfte der Figur 3.3.e werden die Ränge der $\binom{8}{4} = 70$ möglichen Randomisierungen $\underline{g}^{(\ell)}$ für diese neue Teststatistik mit den Rängen bezüglich der früher verwendeten Mittelwertsdifferenz der

unlogarithmierten Daten (Darstellung rechts) verglichen. Man sieht, dass diese Ränge sich deutlich unterscheiden; die beiden Teststatistiken führen also zu unterschiedlichen Anordnungen der Randomisierungen und damit zu verschiedenen Verwerfungsbereichen.

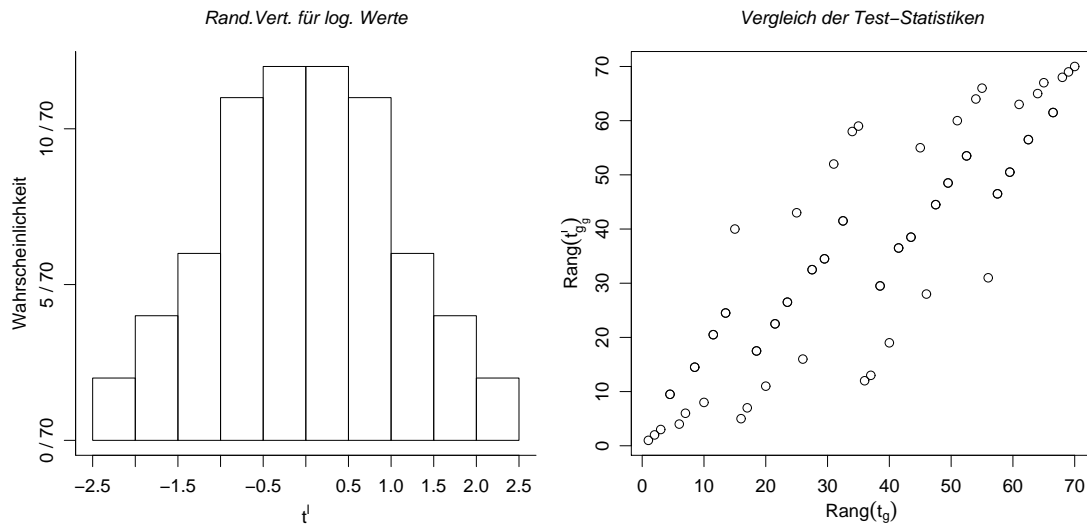


Abbildung 3.3.e: Randomisierungs-Verteilung für die Mittelwerts-Differenz von logarithmierten Daten im Beispiel (links) und Vergleich der Rangordnung der einzelnen Randomisierungen (rechts).

- f **Robustheit.** Wieso eine robuste Teststatistik verwenden, wenn der Test auch ohne diese „Vorsichtsmassnahme“ die Irrtumswahrscheinlichkeit genau einhält? Es kann Abweichungen geben, die eigentlich mit der Fragestellung nichts zu tun haben. Wenn zum Beispiel ein einzelner riesiger Wert in einer Gruppe auftaucht (ein „Ausreisser“), heisst das nicht, dass der „Idealwert“ (Lageparameter) für diese Gruppe grösser ist als für die andere. Man kann diesen Gesichtspunkt ebenfalls mit dem Begriff der Macht beschreiben: Der Test soll kleine Macht haben gegenüber uninteressanten Abweichungen von der Nullhypothese. Er soll grosse Macht haben gegen interessierenden Abweichungen, auch wenn diese durch uninteressante überlagert sein sollten.
- g Eine sehr bekannte Teststatistik ist diejenige des **Rangsummentests** von Wilcoxon, Mann und Whitney (auch U-Test genannt),

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

wobei R_i den Rang der Beobachtung y_i unter allen y_i -Werten bezeichnet. Da man nur mit Rängen arbeitet, ist diese Teststatistik recht robust. Das macht den Test, wie schon im Einführungsteil betont, zum Test der Wahl für das Zwei-Stichproben-Problem.

Die Berechnung der Verteilung dieser Teststatistik unter der Nullhypothese, die in den Einführungsvorlesungen normalerweise weggelassen wird, erfolgt nach dem Prinzip, das im vorhergehenden Abschnitt erläutert wurde. Da nur die Ränge benützt werden, die ja bei festgelegter Beobachtungszahl immer gleich sind (sofern keine „Bindungen“ auftreten), kann man die Verteilungen ein für alle Mal berechnen und die Grenzen des Verwerfungsbereiches in Tabellen festhalten; man braucht also nicht für jeden neuen Datensatz eine neue Berechnung der Randomisierungs-Verteilung, wie dies für andere Teststatistiken nötig ist.

h* Im Hagelversuch wurde eine wesentlich kompliziertere Konstruktion verwendet. Zunächst wurde mit den Daten eines Vorversuchs ein Regressionsmodell entwickelt, mit dem die zu erwartende Hagelenergie für jede Wolke vorhergesagt werden konnte auf Grund von erklärenden Grössen, die vor der allfälligen Impfung feststanden. Das diente dazu, die enorme natürliche Streuung der Hagelenergie ein wenig zu verringern und damit die Macht des Tests zu erhöhen. Die neue Zielgrösse war dann die Abweichung der gemessenen Hagelenergie von der Vorhersage. Da angenommen wurde, dass die Wirkung der Impfung für verschiedene Werte einer bestimmten Wettergrösse (der Wolkenbasistemperatur) verschieden gross ausfallen könnte, wurde mit diesen Abweichungen nochmals eine Regression gerechnet, mit der Wettergrösse und dem Impfungs-Indikator als erklärenden Grössen. Die Testgrösse für den Randomisierungstest war nun nicht nur der mittlere Unterschied zwischen den genannten Vorhersagefehlern für geimpfte und ungeimpfte Wolken, sondern auch der geschätzte Unterschied der Koeffizienten für die Wettervariable – also eine zweidimensionale Teststatistik, mit einer zweidimensionalen Randomisierungs-Verteilung. Schliesslich wurde der Verwerfungsbereich in diesem zweidimensionalen Raum durch ein Rechteck festgelegt, das zusammen 5% der Randomisierungsverteilung umfasste.

Das ist ausserordentlich kompliziert. Es soll zeigen, dass man beim Randomisierungstest die Teststatistik und den Verwerfungsbereich sehr genau auf die Situation – genauer auf die Alternativen, die möglichst gut sollen nachgewiesen werden können – abstimmen kann.

3.4 Eine Stichprobe oder zwei verbundene

- a **Beispiel Tranquilizer.** Die Wirkung eines Tranquilizers wurde an neun Patienten geprüft, indem vor und nach der Anwendung des Medikaments der „Hamilton depression scale factor IV“ gemessen wurde. Die Daten stammen aus Hollander and Wolfe (1999, Ex. 3.1) und sind in Tabelle 3.4.a wiedergegeben.

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

Tabelle 3.4.a: Daten im Beispiel Tranquilizer.

- b In einem solchen Problem der **verbundenen Stichproben** werden, wie bekannt, pro Beobachtungseinheit zwei Beobachtungen $Y_i^{(1)}$ und $Y_i^{(2)}$ gemacht, und die Frage lautet, ob die Verteilung der beiden Grössen gleich ist.

Der erste Schritt besteht jeweils darin, dass – allenfalls nach Transformation der Daten – die Differenzen $X_i = Y_i^{(2)} - Y_i^{(1)}$ gebildet werden und nun die Frage geprüft wird, ob die **Verteilung** der X_i **symmetrisch um 0** ist. Es wird vorausgesetzt, dass die X_i unabhängig sind.

Wenn Symmetrie gilt, ist für jedes X_i ein positives und ein negatives Vorzeichen gleich wahrscheinlich – unabhängig von seinem absoluten Betrag $|X_i|$. Die Beobachtungen mit $X_i = 0$ müssen aus der Stichprobe entfernt werden, da sie das Vorzeichen nicht festlegen. Sie geben ja auch keine Information über die Frage, welche Behandlung die bessere sei.

Die Vorzeichen übernehmen nun die Rolle der Gruppenzugehörigkeit G_i im Zwei-Stichproben-Problem, und die Beträge $|X_i|$ treten an die Stelle der Beobachtungen X_i . Die Wahrscheinlichkeit für jede Vorzeichen-Konstellation $\underline{g}^{(\ell)} = [g_1^{(\ell)}, \dots, g_n^{(\ell)}]$ (mit $g_i^{(\ell)} = +1$ oder $= -1$) ist gleich und damit $= 1/2^n$.

- c Wir brauchen also nur eine Teststatistik der Form $T\langle \underline{g}, \underline{z} \rangle$ festzulegen für Argumente $g_i = +1$ oder -1 und $z_i > 0$. Die Randomisierungs-Verteilung ist gegeben durch

$$P\langle T\langle \underline{G}, \underline{z} \rangle = t \rangle = \#\{\underline{g} \mid T\langle \underline{g}, \underline{z} \rangle = t\} / 2^n .$$

- d Einfache Festlegungen der Teststatistik führen zu bekannten Verfahren:

- $T\langle \underline{g}, \underline{z} \rangle = (1/n) \sum_i g_i z_i$ ist eine komplizierte Schreibweise für den Mittelwert $\text{ave}_i y_i$. Es entsteht der Randomisierungstest, der dem **t-Test** für gepaarte Stichproben am nächsten kommt.
- Die Anzahl der positiven Vorzeichen schreibt sich als $T\langle \underline{g}, \underline{z} \rangle = \#\{i : g_i = 1\}$ und liefert den **Vorzeichentest**.
- $T\langle \underline{g}, \underline{z} \rangle = \sum_{i: g_i=1} R_i$ wobei R_i den Rang von z_i unter allen $z_i = |y_i|$ bedeutet, führt zum **Vorzeichen-Rangsummen-Test** von Wilcoxon.

- e Im Beispiel des Tranquilizers erhält man für den Vorzeichen-Rangsummen-Test

```
> wilcox.test(d.tranquilizer[,1],d.tranquilizer[,2],paired=TRUE)
```

```
Wilcoxon signed rank test
```

```
data: d.tranquilizer[, 1] and d.tranquilizer[, 2]
```

```
V = 40, p-value = 0.03906
```

```
alternative hypothesis: true mu is not equal to 0
```

Der Tranquilizer scheint also einen knapp signifikanten Effekt zu haben. Dieser Schluss ist allerdings gewagt, da er auf einem „Vorher-Nachher-Vergleich“ beruht. Ein solcher könnte sich auch ohne wirksames Medikament ergeben. Für einen sauberen Nachweis müsste man mit einer Kontrollgruppe vergleichen oder einen Crossover-Versuch ansetzen.

3.5 Schätzungen und Vertrauensintervalle

- a Der Vorzeichen-Rangsummen-Test für eine einfache Stichprobe prüft, ob die Verteilung der Beobachtungen symmetrisch um 0 sei. Es ist einfach, den Test für einen beliebigen Wert μ eines Symmetriezentrums zu erweitern: Man zieht μ von allen Beobachtungen ab und testet auf Symmetrie um 0. Wir haben also für alle Hypothesen der Form „**Symmetriezentrum ist μ** “ einen Test. Daraus kann man nun eine Schätzung und ein Vertrauensintervall erhalten.
- b Für den Test wird eine Teststatistik gebildet, die misst, *wie gut Daten zur Nullhypothese*, hier also zum Symmetriezentrum μ , *passen*. Wir können nun die Nullhypothese, also μ , variieren und den Wert der Teststatistik als Funktion von μ auffassen. Dann kann man eine **Schätzung** festlegen: Man bestimmt dasjenige μ , das im Sinne der Teststatistik *am besten* zu den Daten passt.

* Mathematisch formuliert: Die Teststatistik hängt von der Nullhypothese ab, $T\langle \underline{g}, \underline{y}; \mu \rangle$. Wenn grosse Werte von T die Abweichung von $H_0 : \mu$ anzeigen, ist die Schätzung festgelegt als der Wert von μ , der die Teststatistik minimiert, $\hat{\mu} = \arg \min_{\mu} \langle T\langle \underline{g}, \underline{y}; \mu \rangle \rangle$.

- c Wenn man in diesem Prinzip die Teststatistik des Vorzeichen-Rangsummen-Tests (3.4.d) einsetzt, erhält man den so genannten **Hodges-Lehmann-Schätzer**. Man kann zeigen, dass sich die Schätzung aus den $n(n+1)/2$ Mittelwerten $(X_h + X_i)/2$ der Paare von Beobachtungen $[X_h, X_i]$ mit $h \leq i$, den so genannten **Walsh averages**, bestimmen lässt – als deren Median

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

Im Beispiel Tranquilizer ergeben sich 45 Walsh-Mittelwerte -0.1470, -0.1135, -0.0800, -0.0685, -0.0350, 0.0100, ..., 1.022 mit dem Median $\hat{\mu} = 0.46$.

- d* Die Herleitung ist nicht schwierig: Die Teststatistik des Wilcoxon-Tests ist ja gleich der Summe der Ränge R_i der positiven Beobachtungen, $\sum_{i:g_i=1} R_i = \sum_{i:X_i>0} R_i$. Es bezeichne $X_{[k]}$ den k -ten Wert der geordneten Stichprobe. Wir betrachten eine positive Beobachtung $X_{[k]} > 0$ und dazu die Walsh-Mittelwerte $Z_{hk} = (X_{[h]} + X_{[k]})/2$ mit den Beobachtungen, die in der geordneten Stichprobe vorher kommen, für die also $h < k$ ist. Das Vorzeichen dieser Z_{hk} ist negativ, wenn $|X_{[h]}| > |X_{[k]}|$ ist, und sonst positiv. Die Anzahl positiver Z_{hk} ist deshalb gleich der Anzahl Beobachtungen mit $|X_{[h]}| < |X_{[k]}|$. Zählt man eins dazu, erhält man den Rang $R_{[k]}$, der in die Teststatistik eingeht, $R_{[k]} = \#\{h \mid Z_{hk} > 0, h < k\} + 1$. Das kann man schreiben als $R_{[k]} = \#\{h \mid Z_{hk} > 0, h \leq k\}$, da $Z_{kk} = X_{[k]} > 0$ ist. Für negative Beobachtungen $X_{[k]} < 0$ werden alle Z_{hk} negativ. Wenn man jetzt über alle Beobachtungen aufsummiert, wird klar, dass die Testgrösse gleich der Anzahl positiver Walsh-Mittelwerte ist,

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

Testet man nun die Nullhypothese $\mu = \mu_0$, dann ist leicht einzusehen, dass die Teststatistik zur Anzahl Walsh-Mittelwerte wird, die $> \mu_0$ sind. Daraus wird klar, dass der Wert μ , der am besten zur Stichprobe passt, gleich dem Median der Walsh-Mittelwerte ist. Genauer: Der (zweiseitige) Wilcoxon-Test legt einen zweiseitigen Verwerfungsbereich für T fest, was allerdings jeweils so formuliert wird, dass das Minimum von T und $n(n+1)/2 - T$ mit der kritischen Grenze verglichen wird. Dieses Minimum wird dann möglichst klein, wenn T und $n(n+1)/2 - T$ gleich gross sind, und das trifft für den Median der Walsh-Mittelwerte zu.

- e Ein Test legt fest, wann Daten mit Parameterwerten verträglich sind. Damit kann man wie üblich **Vertrauensintervalle** konstruieren – sie fassen alle Parameterwerte zusammen, die gemäss Test mit den Daten verträglich sind.
- f Das Vertrauensintervall, das dem Vorzeichen-Rangsummen-Test von Wilcoxon entspricht, lässt sich ebenfalls aus den $n(n+1)/2$ Mittelwerten über die Paare ausrechnen: Man muss den kritischen Wert c für die Teststatistik des Wilcoxon-Tests aus einer Tabelle oder der asymptotischen Näherung bestimmen. Dann sind die Grenzen gleich dem c -ten und dem c' -ten Wert der geordneten Walsh-Mittelwerte, mit $c' = n(n+1)/2 + 1 - c$.

Im Beispiel Tranquilizer ist $c = 5$ (oder $c = 41$) und man erhält das Vertrauensintervall $[0.01, 0.786]$.

* Die kritischen Werte c wurden so bestimmt, dass die Wahrscheinlichkeit für $T \geq c$ höchstens gleich 5% ist. Weil die Verteilung von T diskret ist, wird diese Schranke „nie“ genau erreicht, und die Wahrscheinlichkeit des Fehlers erster Art ist $< 5\%$. Dem entsprechend hat das angegebene Vertrauensintervall eine höhere Überdeckungs-Wahrscheinlichkeit als 95%.

- g Schätzung und Vertrauensintervall erhält man mit der R-Funktion `wilcox.test`, indem man das Argument `conf.int = TRUE` setzt.

- h Das Prinzip der Schätzung und der Bestimmung des Vertrauensintervalls lässt sich auch für beliebige Teststatistiken an Stelle der Vorzeichen-Rangsumme anwenden. Dann müssen die Schätzung und die Vertrauensgrenzen als Nullstellen der Funktion

$$P\langle T\langle \underline{G}, \underline{z}^*; \mu \rangle > T\langle \underline{g}^*, \underline{z}^*; \mu \rangle \rangle - \beta$$

bestimmt werden; für $\beta = 0.5$ erhält man die Schätzung, für $\beta = 0.025$ und $= 0.975$ die Vertrauensgrenzen. Wenn die Wahrscheinlichkeit in dieser Formel mit der simulierten Randomisierungs-Verteilung berechnet werden muss, ist das recht aufwändig und wegen der Zufälligkeit der Simulation tückisch (immer die gleichen Zufallszahlen verwenden!).

- i **Mehrere unabhängige Stichproben.** Die Verallgemeinerung des Zwei-Stichproben-Problems auf **mehrere Gruppen** führt zur **einfachen Varianzanalyse**. Es ist nahe liegend, dass die Zugehörigkeit der Beobachtungen zu den Gruppen die Grundlage der Randomisierung bildet, und wieder jede mögliche Zuordnung gleich wahrscheinlich ist. Die Grundmenge der möglichen Zuordnungen wird jeweils so gewählt, dass die Anzahl der Beobachtungen in jeder Gruppe mit den Anzahlen in den vorliegenden Daten übereinstimmt.

Der bekannteste solche Test ist der **Kruskal-Wallis-Test** (2.2.e). Er beruht wie der U-Test auf den Rängen R_i der y_i unter allen Beobachtungen, und die Teststatistik ist diejenige einer einfachen Varianzanalyse, angewandt auf die Ränge.

- j **Mehrere verbundene Stichproben.** Die Verallgemeinerung von zwei verbundenen Stichproben auf mehrere führt zu dem, was in der Varianzanalyse ein einfacher Blockversuch genannt wird. Es liegen n Beobachtungseinheiten (Blöcke) i vor, für die jeweils die Werte $Y_i^{(j)}$ der Zielgröße unter m Bedingungen (m „verbundene Stichproben“) aufgenommen werden.

Randomisierungen müssen auf die Blöcke Rücksicht nehmen, da sich diese auch unter der Nullhypothese unterscheiden dürfen. Sie bestehen deshalb aus unabhängigen Permutationen der Werte der Zielgröße innerhalb jedes Blocks.

Der allgemein bekannte nichtparametrische Test für die Nullhypothese der Gleichheit der Zielgröße unter allen m Bedingungen ist der **Friedman-Test** (2.2.h). Man bildet die Ränge R_{ij} der m Werte $Y_i^{(j)}$ in jedem Block i und daraus die mittleren Ränge $\bar{R}_j = \text{ave}_i \langle R_{ij} \rangle$ der Bedingungen j . Die Teststatistik ist wie üblich die Quadratsumme der Abweichungen der R_j von ihrem erwarteten Wert $(m+1)/2$, mit einem geeigneten Faktor versehen, der zu einer genäherten Chiquadrat-Verteilung führt,

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

3.6 Korrelation und Regression

- a **Korrelation und einfache Regression.** Es werden zwei Variable X_i und Y_i beobachtet. (Die X_i können auch fest vorgegeben sein.) Um die Nullhypothese „kein Zusammenhang“ zu testen, wird die „Paarung“ als zufällig angesehen. Die Rolle der g_i spielt dabei eine Permutation g , also eine mögliche Anordnung der Zahlen $1, 2, \dots, n$, die zum „permutierten Datensatz $[X_i, Y_{g_i}]$ “ führt. Jede Permutation erhält die Wahrscheinlichkeit $1/n! = 1/(n(n-1)\dots 2 \cdot 1)$.

Als Teststatistik $T(\underline{X}, \underline{Y})$ kann man die gewöhnliche Korrelation, eine Rangkorrelation, eine robuste Schätzung des Regressions-Koeffizienten oder etwas nach eigenem Gusto wählen.

- b Das gleiche Vorgehen erlaubt es auch, in einer **multiplen Regression** die Frage zu klären, ob **überhaupt ein Zusammenhang** zwischen den erklärenden Variablen und der Zielgrösse besteht. Die Nullhypothese, dass die Koeffizienten aller erklärenden Variablen null seien, führt nämlich wieder dazu, dass alle Permutationen von \underline{Y} gleich wahrscheinlich sind.

* Setzt man die multiple Regression mit einer einzigen nominalen erklärenden Variablen als Modell für die einfache Varianzanalyse ein, dann erhält man die gleichen Tests wie oben (3.5.i).

- c Ebenso lässt sich für eine **Zeitreihe** feststellen, ob die Beobachtungen **unabhängig** seien. Als Testgrösse eignet sich beispielsweise die erste Autokorrelation.

- d Für den Test eines einzelnen (oder mehrerer) **Koeffizienten in der multiplen Regression** lässt sich leider kein strikt richtiges Randomisierungsmodell angeben.

- e* Die nahe liegende Art, wie man die Idee des Randomisierungstests hier anwenden kann, beruht auf der Tatsache, dass der geschätzte Koeffizient $\hat{\beta}_j$ in der multiplen Regression aus der einfachen Regression der „partiellen Residuen“ erhalten werden kann (siehe Regression 1, partial residual plot). Man bildet also die Residuen $R^{(Y|j)}$ der Regression der Zielgrösse auf alle erklärenden Variablen ohne $X^{(j)}$ und die Residuen $R^{(j|j)}$ der Regression von $X^{(j)}$ auf die gleichen erklärenden. Nun testet man die Steigung der einfachen Regression von $R^{(Y|j)}$ auf $R^{(j|j)}$ wie oben besprochen.

Ob dieser Test je genauer untersucht und angewandt wurde, ist dem Autor zur Zeit nicht bekannt.

- f* Ein Test, der auf Rang-Methoden beruht, stammt von Jaeckel, Hettmansperger und McKean und ist in Hollander and Wolfe (1999, Ch. 9.6) beschrieben.

- g* **Permutationen und andere Randomisierungen.** In den Tests für Regression und Korrelation legen alle Permutationen der Zahlen $1, 2, \dots, n$ die Randomisierungs-Verteilung fest. Bei zwei oder mehreren Gruppen waren es alle möglichen Auswahlen der Gruppenzugehörigkeit. Man kann auch hier die Permutationen verwenden; es sind viel mehr, aber entsprechend viele führen jeweils zur gleichen Gruppenzugehörigkeit und damit zum gleichen Wert der Teststatistik.

Die Permutationen sind aber nicht die allgemeinsten Randomisierungen. Im Hagelversuch wurde eigentlich keine Zufallsauswahl bestimmt, bei der fest lag, wie viele potentielle Hageltage welcher Gruppe zugehören sollten. Man wusste am Anfang nur ungefähr, wie viele solche Tage sich in der vorgesehenen Versuchsdauer von fünf Jahren zeigen würden. Es wurde deshalb für jeden Tag je mit Wahrscheinlichkeit $1/2$ Impfung oder „Kontrolle“ festgelegt, was jede Folge von Nullen und Einsen gleich wahrscheinlich machte. Es ergaben sich – nicht zur eitlen Freude der Forscher – zufälligerweise nur 33 von 76 Tagen als Impftage. Für die Randomisierungs-Verteilung wurden nur die Auswahlen von 33 aus 76 Tagen berücksichtigt, also nicht alle Zuordnungen, die gemäss Versuchsanlage möglich waren. Man kann von einem bedingten Test, gegeben die Anzahl Impftage, sprechen.

- h Es gibt einen allgemeinen Grundsatz, nach dem solche „Bedingungen“, die mit der Testfrage nichts zu tun haben, nach Möglichkeit für die Einschränkung des Ereignisraumes verwendet werden sollen. Es ergeben sich „**bedingte Tests**“, die im Allgemeinen präziser sind als diejenigen ohne Einschränkung.

Für die Randomisierungstests heisst das, dass die Menge der Randomisierungen, die für die Bestimmung der Randomisierungs-Verteilung benützt werden, eingeschränkt wird.

Eine eingeschränkte Menge von Randomisierungen betrachtet man auch beim Problem mehrerer Stichproben: Man lässt nur Permutationen der Zuordnung der Beobachtungen $Y_i^{(j)}$ zu den „Stichproben“ j innerhalb des gleichen „Blocks“ i zu – für jeden Block eine unabhängige Permutation der Zahlen $1, 2, \dots, m$.

3.A Anhang: Randomisierungstests im Unterricht

Dieser Anhang enthält zunächst einen ausführlicheren Text zur Einführung der Idee des statistischen Tests an Hand eines Randomisierungstests (mit Passagen, die mit dem vorhergehenden Text übereinstimmen). Dann folgen Überlegungen zur Verwendung dieses Stoffes im Unterricht auf der Gymnasialstufe.

3.A.1 Der statistische Test als Widerspruchsbeweis

- a Die Fragestellung kann allgemeiner formuliert werden als: „Hat die Behandlung der Untersuchungseinheiten i mit $G_i = 1$ einen Effekt auf die Zielgrösse Y_i ?“ Eine solche Frage wird mit einem **statistischen Test** beantwortet. Die Grundidee ist diejenige eines **Widerspruchsbeweises**: Wir nehmen an, dass *kein* Effekt vorhanden sei, und führen diese Annahme zu einem „statistischen Widerspruch“.
- b Wir stellen also ein **Wahrscheinlichkeitsmodell** auf, das plausibel erscheint, wenn kein Effekt vorhanden ist. Dieses Modell heisst **Nullhypothese**.

In unserem Beispiel gehen wir davon aus, dass für die Auswahl der zu impfenden Wolken (der zu behandelnden Einheiten) ein Zufallsmechanismus eingesetzt wurde. Die **Grundüberlegung** lautet also: Wir beobachten bestimmte Werte y_i^* der Zielgrösse, wenn die Wolken geimpft werden, für die $g_i^* = 1$ ist. Wenn das Impfen keinen Einfluss auf die Hagelenergie hat, würden wir die genau gleichen Werte y_i^* erhalten, wenn die Wolken entsprechend den Werten

$$\underline{g}^{(1)} = [0, 1, 0, 0, 1, 1, 0, 1]$$

geimpft worden wären, oder entsprechend irgendeiner anderen Auswahl von zu impfenden Wolken. Diese Überlegung führt zum folgenden Wahrscheinlichkeitsmodell.

Wir betrachten die y_i^* als fest und die Auswahl $[G_1, \dots, G_n]$ als zufällig. Das ist sicher gerechtfertigt, wenn die Auswahl mit einem Zufallsmechanismus getroffen wurde, der (beispielsweise) jeder Auswahl von $n/2 = 4$ Elementen aus $n = 8$ Elementen gleiche Wahrscheinlichkeit gibt. In diesem Fall ist die Wahrscheinlichkeit für jede Auswahl $\binom{8}{4}^{-1} = \frac{1}{70}$. Damit ist die Nullhypothese festgelegt.

- c **Idee des „statistischen Widerspruchs“**. Angenommen, wir fänden, dass alle geimpften Wolken kleinere Energien produzierten als alle ungeimpften. Das wäre der stärkste Beweis für einen Effekt, den wir erwarten können. Dennoch ist es kein strikter Beweis, da das auch auftreten kann, wenn das Impfen die Wolken nicht beeinflusst, nämlich genau für eine Zufallsauswahl, also mit Wahrscheinlichkeit $1/70$. Wenn wir überhaupt je einen Effekt als nachgewiesen bezeichnen wollen, können wir in einer solchen Situation, wo der Zufall mitspielt, keinen strikten Beweis verlangen. Eine sinnvolle **Entscheidungsregel** hat die folgende Form: Von allen möglichen Versuchsergebnissen scheiden wir eine Menge von extremen aus. Falls das beobachtete Ergebnis in dieser Menge liegt, erachten wir dies als Widerspruch zum Modell und „verwerfen“ damit die Nullhypothese. Die Menge heisst **Verwerfungsbereich**, ihr Komplement **Annahmebereich**.

d Wir müssen festlegen,

1. wie wir „Extremität“ messen wollen, und
2. wie viele Versuchsergebnisse wir als extrem bezeichnen wollen.

„Extremität“ soll heissen: Ein extrem starker Hinweis auf den erhofften (oder in andern Fällen einen befürchteten oder anderswie zu untersuchenden plausiblen) Effekt. Eine solche Grösse ist im Beispiel die Differenz zwischen den mittleren Energien für ungeimpfte und geimpfte Wolken,

$$T(\underline{G}) = \frac{1}{n/2} \sum_{i:G_i=0} y_i^* - \frac{1}{n/2} \sum_{i:G_i=1} y_i^* = \frac{2}{n} \sum_i y_i^* (1 - 2G_i).$$

Für die aktuelle Zufallsauswahl erhalten wir

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

T heisst die **Testgrösse** oder **Teststatistik**. Sie muss jeder Kombination von möglichen beobachteten Werten y_1, y_2, \dots, y_n und möglichen Auswahl g_1, g_2, \dots, g_n eine Zahl zuordnen.

- e Wenn die y_1^*, \dots, y_n^* gegeben sind, gibt es (höchstens) $\binom{n}{n/2}$ mögliche Werte für T . Wir fragen nun nach der **Verteilung** von T , d.h. nach den Wahrscheinlichkeiten, mit denen die möglichen Werte auftreten. Diese bestimmen sich nach der Regel „Anzahl günstige durch Anzahl mögliche Fälle“,

$$P\langle T(\underline{G}) = t \rangle = \#\{\underline{g} \mid T(\underline{g}) = t\} / \binom{n}{n/2}$$

– meistens wird der Zähler =1 sein. Die Verteilung für das Beispiel ist in Abbildung 3.2.c dargestellt.

Der Verwerfungsbereich fasst die extremsten (z.B. die grössten) Werte zusammen, deren Wahrscheinlichkeit gesamthaft 5% (oder eine andere durch Konvention festgelegte Prozentzahl) beträgt. Das lässt sich allerdings meist nicht genau erreichen, da beispielsweise kein Vielfaches von $1/70$ gleich 0.05 ist. Also muss man genauer sagen: Der Verwerfungsbereich fasst die extremsten Werte zusammen, deren Wahrscheinlichkeit gesamthaft möglichst wenig kleiner als 5% ist.

- f Im Beispiel ergibt sich als Verwerfungsbereich $\{t \mid t \geq 4643.25\}$ (für eine einseitige Fragestellung). Für die im Experiment getroffene Auswahl \underline{g}^* ergibt sich $T(\underline{g}^*) = -3629.25$, also ein Effekt in die unerwartete Richtung! Da dies < 4643.25 ist, wird die Nullhypothese nicht verworfen; ein **Effekt der Impfung von Hagelwolken kann also nicht nachgewiesen werden**. (Wenn schon, könnte sich der Effekt in die unerwartete Richtung bei „umgekehrt einseitiger“ Fragestellung als signifikant erweisen, was aber auch nicht der Fall ist. Dies entspricht auch dem Ergebnis des gesamten Experiments.)

Ein solches Ergebnis heisst nicht, dass das Impfen keinen Effekt hat. Wenn der Effekt klein ist gegenüber der zufälligen Streuung der Daten, wird ein wirklich vorhandener Effekt oft nicht zu einem statistisch „signifikanten“ Ergebnis führen, aber mit grösseren Stichproben lassen sich auch in diesem Sinne kleine Effekte immer besser statistisch nachweisen.

3.A.2 Simulation der Randomisierungs-Verteilung

- a Um die Randomisierungs-Verteilung zu bestimmen, muss man im allgemeinen die Werte der Test-Statistik für alle $\binom{n}{n/2}$ möglichen Auswahlen berechnen. Bereits bei $n = 20$ wird $\binom{n}{n/2} \approx 200'000$, was auch für Computer schon ein spürbarer Aufwand wird (vor allem bei komplizierteren Statistiken, s.u.) Für $n = 216$ ist an dieses Durchrechnen aller Möglichkeiten nicht mehr zu denken.

Die Frage lautet: Wie kann man mit einem solchen „unlösbaren“ Problem fertig werden?

- b Leute, die mit statistischen Methoden zu tun haben, kamen auf die Idee, mit Hilfe von Wahrscheinlichkeits-Überlegungen eine genäherte Lösung zu erhalten. Die grundlegende Interpretation von Wahrscheinlichkeiten ist ja die folgende: Wenn man sich für eine Grundgesamtheit interessiert, die man nicht vollständig untersuchen kann, zieht man eine Stichprobe. Wenn man sich für die Verteilung einer Grösse X in der Grundgesamtheit interessiert, dann bilden die X -Werte in der Stichprobe dafür eine Näherung; das Histogramm zeigt diese Verteilung anschaulich.

Diese Idee wenden wir nun auf unser genanntes Problem an. Die Grundgesamtheit bilden alle Randomisierungen (Auswahlen). Wir hätten gerne die Verteilung von T in dieser Grundgesamtheit bestimmt. Dazu ziehen wir eine Stichprobe von r^* Randomisierungen $[g_1^{(r)}, g_2^{(r)}, \dots, g_n^{(r)}]$ und berechnen für diese den Wert $t^{(r)} = T\langle g_1^{(r)}, g_2^{(r)}, \dots, g_n^{(r)} \rangle$. Die durch den Computer gezogenen einzelnen Randomisierungen werden auch „Replikate“ (Wiederholungen der Zufalls-Ziehung) genannt. Die gesuchte Verteilung der Teststatistik wird durch die empirische Verteilung dieser r^* Werte $t^{(r)}$ angenähert – je grösser die Anzahl, desto genauer die Näherung.

- c Wir haben also zur genäherten Lösung des „unlösbaren“ Problems den Zufall bewusst und genau kontrolliert eingesetzt. Es braucht ein klares Durchdenken, damit man die „technische Ebene“ der Wahrscheinlichkeit, die durch dieses Vorgehen verursacht wird, nicht verwechselt mit dem Zufall, der im ursprünglichen Problem steckt (im Beispiel die Zufallsauswahl, die zur aktuell benützten Behandlung der Wolken führte).

Das Vorgehen ist allgemein anwendbar, wenn man Probleme der Wahrscheinlichkeitsrechnung nicht rechnerisch lösen kann oder will. Es wird als stochastische **Simulation** bezeichnet. (Der Begriff Simulation wird auch in anderem Zusammenhang in der Mathematik verwendet, nämlich zur genäherten Lösung von Differentialgleichungen.)

3.A.3 Anmerkungen zur Simulation

- Man kann sich überlegen, wie gross r^* sein muss, damit die Entscheidung mit grosser Wahrscheinlichkeit richtig ist, d.h., dass die Test-Entscheidung (signifikant oder nicht) aufgrund der simulierten Verteilung mit der Entscheidung aufgrund der exakten Randomisierungs-Verteilung übereinstimmt. Dazu braucht man den Begriff des P-Wertes. Es ergibt sich ein nicht sehr schwieriges Problem der Binomial-Verteilung. Die Lösung ist unabhängig vom konkreten Kontext. Eine vernünftige Genauigkeit ergibt sich bei r^* -Werten in den Tausendern.
- Wenn man direkt auf die Simulation losgeht, ohne ein Wahrscheinlichkeitsmodell einzuführen, ergibt sich keine Schwierigkeit: Man knüpft direkt an die Grundüberlegung (s. oben) an und lässt den Mechanismus der Zufalls-Auswahl m mal laufen.

Dies kann zunächst zu einem intuitiven Verständnis von statistischen Überlegungen führen. Der Weg zu einem genauen Verständnis der Begriffe der klassischen Statistik von da aus müsste aber noch erkundet werden. Es ist denkbar, dass viele Anwender in (ferner?) Zukunft sich mit einem solchen Verständnis begnügen werden.

3.A.4 Seitenbemerkungen zur Einführung des statistischen Tests

Meine Art, wie ich den Begriff des Tests auffasse und vermittele, ist wie folgt:

- Die Schliessende Statistik untersucht den Zusammenhang zwischen Daten und Modellen. Als grundlegenden Begriff braucht man eine **Regel**, die angibt, welche Daten mit welchen Modellen **als verträglich gelten sollen**. Es gibt viele mögliche Regeln. Man muss sich durch **Konventionen** auf eine einigen. Ein Teil der Konvention, das Niveau (meistens 5%), ist willkürlich, zum anderen Teil, der Test-Statistik, kann man sich Nützlichkeits-Überlegungen (Macht) machen.
- Von einer solchen Regel leiten sich die Tests im üblichen Sinne und die Vertrauensintervalle her.
 - **Test:** Modell ist fixiert, der Annahmebereich bestimmt, welche Daten mit dem Modell verträglich sind.
 - **Vertrauensintervall** (oder allgemeiner -bereich): Daten sind fixiert, das Vertrauensintervall bestimmt, welche Parameterwerte, also welche Modelle innerhalb einer parametrischen Modellfamilie, mit den Daten verträglich sind.

Tests sind didaktisch einfacher zu erklären und sollten daher zuerst behandelt werden. Der Begriff des zufälligen Intervalls ist sehr schwierig. Ich habe deshalb Mühe mit Büchern, die den Begriff der „Intervallschätzung“ einführen, und dies erst noch vor dem Begriff Test. Ich führe das Vertrauensintervall, wie angegeben, mit dem Dualitäts-Prinzip ein, und erwähne erst nachher, dass es sich dabei um ein zufälliges Intervall handelt, das „mit Wahrscheinlichkeit 95% den wahren Wert des Parameters enthält“.

- Der statistische Test stellt also für das *Verständnis* den vielleicht wichtigsten Grundbegriff der Schliessenden Statistik dar. Für die *Praxis* sind Vertrauensintervalle nützlicher, soweit man mit parametrischen Modellfamilien arbeitet, da sie mehr Information geben.
- Der **P-Wert** ist ein standardisiertes Mass für die „Extremität“ der Daten bezüglich der Nullhypothese, also eigentlich eine „standardisierte Teststatistik“, die immer (für kontinuierliche Daten) unter der Nullhypothese uniform verteilt ist. Für die Praxis ist die Angabe des P-Wertes ebenfalls informativer als das „ja/nein“-Test-Ergebnis. Man könnte den P-Wert als Grundbegriff vor der Testregel einführen, aber das scheint mir technisch zu kompliziert zu sein.
- Vertrauensintervalle (oder -bereiche) können nach dem Dualitäts-Prinzip aus einem Randomisierungstest erhalten werden. Dabei scheint es zunächst, dass man die Grenzen des Intervalls durch Nullstellensuche einer komplizierten Funktion bestimmen

muss. Vereinfachungen für spezielle Teststatistiken sind möglich und teilweise Gegenstand der Forschung.

- Den formalen Begriff der Macht sollte man meiner Erfahrung nach erst mit grossem „Sicherheitsabstand“ nach der Einführung der Grundidee des Tests besprechen – also gegen Ende einer Einführungsvorlesung an der Hochschule. Im Gymnasium kann es angebracht sein, als Ausblick einmal eine Simulation eines Tests unter einer Alternative durchzuführen, um die Grundidee aufzuzeigen.

3.A.5 Randomisierungstests im (Zusatz-) Unterricht

Pluspunkte

- Man braucht den Begriff des parametrischen Modells nicht; es muss kein Verteilungstyp bekannt sein (ausser evtl. für Macht-Betrachtungen).
- Moderne Methode
- Sinnvolle Computer-Anwendung
- Herumspielen möglich, sogar Forschung.

Minuspunkte

- Man macht spezielle Überlegungen, die exakt nur auf wenige einfache Probleme der Statistik passen, oder anders gesagt:
- Man steigt durch die Hintertür in die Statistik ein.

Literaturverzeichnis

- Büning, H. und Trenkler, G. (1994). *Nichtparametrische statistische Methoden*, 2. Aufl., Walter de Gruyter, Berlin.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press. *includes 1 disk*
- Hartung, J. und Elpelt, B. (1997). *Multivariate Statistik: Lehr- und Handbuch der angewandten Statistik*, 6. Aufl., Oldenbourg, München.
- Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*, Wiley, N.Y.
- Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, Wiley Series in Probability and Statistics, 2nd edn, Wiley.
- Stahel, W. A. (2007). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 5. Aufl., Vieweg, Wiesbaden.