Die Analyse des Nichts
oder:
Zur Bedeutung des Zufalls bei empirischer Variablenselektion

Wilfried Schollenberger
WS Unternehmensberatung
und Controlling-Systeme GmbH

Friedrich-Weinbrenner-Straße 20

69126 Heidelberg

wisch@ws-unternehmensberatung.de

Zusammenfassung

Im Data Mining werden die relevanten Variablen häufig empirisch aus einer großen Menge potentieller Einflussfaktoren ausgewählt. Zunächst wird an einem konstruierten Beispiel demonstriert, wie leicht bei wiederholten Versuchen mit „Lern"- und „Test"-Stichproben Artefakte entstehen können.
Dabei wird gezeigt, wie Zufalls-Fehler systematisch in das scheinbare Ergebnis eingehen, und wie dieses Phänomen mit Hilfe der Inferenzstatistik kontrolliert werden kann. Wenn zu der jeweiligen Analyse eine Teststatistik (z.B. F-Test) existiert, ist eine Möglichkeit, die Signifikanzgrenzen für die p-Werte mit Hilfe der Bonferroni-Korrektur zu reduzieren.
Für die Bonferroni-Korrektur bei vielen Variablen wird die exakte Ausgabe kleiner p-Werte aus den Signifikanztests benötigt. Mit der Anweisung ODS OUTPUT gibt es die Möglichkeit, alle Prozedur-Ausgaben als exakte Werte zu verarbeiten.
Wenn Signifikanztests aufgrund verletzter Verteilungsannahmen oder der eingesetzten Verfahren nicht zur Verfügung stehen, ist es trotzdem sinnvoll, auf die Unabhängigkeit von Lern- und Test-Stichprobe zu verzichten, und stattdessen das erstellte Modell gegen viele kleinere Stichproben zu testen.

1 Anlass

Die Verfügbarkeit großer Datenmengen und die Möglichkeit, diese in kurzer Zeit mit anspruchsvollen Methoden auszuwerten, haben dazu geführt, dass diese Methoden nicht mehr nur zur Bestätigung und Präzisierung von gut begründeten Vermutungen, d.h. für Hypothesentests, sondern, unter dem Stichwort „Data Mining", auch zur Suche nach bis dahin nicht vermuteten Zusammenhängen eingesetzt werden.

Bei der Vorstellung solcher Projekte und insbesondere bei der Präsentation von Verfahren zur Variablenselektion ist mir aufgefallen, dass inferenzstatistische Aspekte oft vernachlässigt werden. Ich möchte deshalb zunächst an einem Experiment zeigen, wie Zufallsfehler zu reinen „Pseudo-Ergebnissen" führen können. Danach werde ich zwei Methoden vorstellen, um solche Fehler zu vermeiden.

2 Das Experiment

Für das Beispiel habe ich mit dem Zufallszahlengenerator von SAS eine Datei mit einer Million Daten-Sätzen, 100 unkorrelierten normalverteilten Variablen und einer diskreten Variablen erzeugt. Die diskrete Variable ist in rund 90% aller Fälle 0 und in 10 % aller Fälle 1. Nehmen wir an, es handle sich um teure Ausfälle, und eine Gruppierung, bei der sich die Ausfallwahrscheinlichkeit um einen Prozentpunkt unterscheidet, wäre schon ein Erfolg.

Natürlich bringt eine Analyse der Gesamtheit aller Beobachtungen „nichts". Deshalb ziehen wir zunächst eine geschichtete Stichprobe mit jeweils rund 2.500 Beobachtungen. Für diese Stichprobe berechnen wir die Korrelationskoeffizienten mit der „abhängigen" diskreten Variablen und sortieren das Ergebnis nach dem Betrag des Korrelationskoeffizienten. Korrelationskoeffizienten mit einem Betrag über 0.03 sind auf dem 5%-

Niveau signifikant. Daraus erzeugen wir eine Formel, mit der wir die Beobachtungen in zwei ungefähr gleichgroße Gruppen einteilen. In der „schlechteren“ Gruppe ist der Anteil der Ausfälle dann um ungefähr 1%-Punkt größer als in der „besseren“. Das Verfahren ist nicht besonders „gut“, aber es sollte mit Base-SAS nachvollziehbar sein.

Natürlich überprüfen wir das Ergebnis mit einer Teststichprobe und merken hoffentlich, dass es nicht bestätigt wird. Hatten wir Pech mit den Stichproben? Sollte man den Versuch nochmal mit anderen Stichproben wiederholen? Nehmen wir an, dieser Versuch würde nicht wiederholt, aber in 30 verschiedenen Unternehmen durchgeführt. Wahrscheinlich werden dann zwei oder drei Unternehmen durchaus „interessante“ Ergebnisse erhalten. Die folgende Tabelle zeigt die 10 „ähnlichsten“ Ergebnisse von Lern- und Teststichprobe aus 30 Versuchen.

Tabelle 1: Anteil der „Ausfälle“ in den durch das Modell bestimmten Gruppen

	Lernstichprobe			**Teststichprobe**		
Lauf-Nr	**gute Gruppe**	**schlechte Gruppe**	**Unter-schied**	**gute Gruppe**	**schlechte Gruppe**	**Unter-schied**
29	**9.54%**	**10.9%**	**1.38%**	**9.23%**	**11.1%**	**1.86%**
6	**9.52%**	**11.8%**	**2.26%**	**8.06%**	**9.76%**	**1.70%**
17	**8.71%**	**11.0%**	**2.25%**	**9.00%**	**10.6%**	**1.56%**
22	**8.44%**	**10.8%**	**2.33%**	**9.76%**	**11.4%**	**1.61%**
25	**8.59%**	**10.7%**	**2.14%**	**10.4%**	**11.7%**	**1.31%**
20	9.44%	10.8%	1.33%	9.65%	10.1%	0.48%
10	9.17%	10.5%	1.34%	9.46%	9.69%	0.23%
21	**9.01%**	**11.0%**	**1.99%**	**7.92%**	**11.2%**	**3.30%**
9	9.47%	11.5%	1.99%	9.32%	10.0%	0.68%
19	9.13%	10.4%	1.31%	9.73%	9.33%	(0.40%)

In 4 Fällen liegt der Unterschied der Ausfallquoten zwischen den beiden Gruppen in beiden Stichproben bei über 1,5 %-Punkten. In zwei weiteren Fällen liegt der Unterschied immerhin noch bei mehr als 1,3 %-Punkten[1].

3 Die Analyse des Zufallsfehlers

Mit der geschichteten Stichprobe wird hier in erster Linie ein Zufallsfehler produziert. Solche Stichproben sind sinnvoll, wenn die Beschaffung der Daten aufwendig ist oder Einzelfälle grafisch dargestellt werden. In fast allen anderen Fällen, ist es sinnvoller, die Gesamtheit zu analysieren[2].

Im nächsten Schritt werden dann mit der Variablenselektion die größten Stichprobenfehler gesucht. Bei unabhängigen 100 Variablen ist zu erwarten, dass 5 Variablen auf einem Signifikanz-Niveau von 5% korrelieren, und dann ist es nur noch eine Frage der Stichprobengröße, ob auch die Korrelationskoeffizienten groß genug sind, um einen scheinbar relevanten Effekt zu finden.

Erst mit der Teststichprobe findet dann ein echter „Test“ statt. Dabei wird aber nur getestet, ob eine, im Grunde zufällig erzeugte Berechnungsgröße mit der Zielvariablen korreliert. Und dieser Test liefert dann wieder in durchschnittlich 5% aller Fälle ein „signifikantes“ Ergebnis.

Nun mag eine 5%-Irrtumswahrscheinlichkeit bei gut begründeten Hypothesen und kleiner Fallzahl aufgrund schwer beschaffbarer Daten akzeptabel erscheinen. Bei zufälligen Tests ist sie es mit Sicherheit nicht.

[1] Dieser Wert wurde in der Lernstichprobe nur in einem Fall mit 0,9 %-Punkten unterschritten. Dagegen wurden dreimal Werte über 3 %-Punkte und weitere vier mal Werte über 2,5 %-Punkte erreicht – aber in der Teststichprobe nicht bestätigt.

[2] Z.B. habe ich in meinem Beitrag auf der KSFE 2001 gezeigt, wie man eine Diskriminanzanalyse durch apriori-Wahrscheinlichkeiten so parametrisieren kann, dass auch kleine Subgruppen erkannt werden.

4 Signifikanztests[3]

4.1 Bonferroni-Korrektur

Wenn man solche rein empirisch fundierten Analysen nicht prinzipiell ablehnen will[4], liefert die Inferenzstatistik zumindest das Instrumentarium, die Möglichkeit eines Irrtums zu abzuschätzen. Zuallererst sind hier die Signifikanztests zu nennen. „Wie groß ist Ihre Bereitschaft, ein Modell irrtümlich anzuwenden, wenn es in Wirklichkeit gar nicht zur Unterscheidung von „Guten“ und „Schlechten“ taugt?“ Mit der Antwort auf diese Frage, z.B. „in einem von zwanzig Fällen“, lässt sich zumindest ein Signifikanzniveau bestimmen, auf dem getestet wird.

Jetzt muss nur noch sichergestellt werden, dass dieses Niveau auch tatsächlich eingehalten wird. Die Prüfgröße und jede Einzelkomponente müssen also in der Lern- und in der Test-Stichprobe „signifikant“ mit der Ergebnis-Variablen korrelieren. Bei kleinen Effekten ist dabei die Größe der Stichproben von entscheidender Bedeutung. Zusätzlich muss der kritische P-Wert für die Lern-Stichprobe abgesenkt werden. Schließlich ist, wie oben beschrieben, bei 100 getesteten Korrelationen und einer tolerierten Irrtumswahrscheinlichkeit von 5% zu erwarten, dass 5 von 100 Variablen zufällig „signifikant“ korrelieren.

Die Bestimmung des kritischen P-Werts leistet die Bonferroni-Korrektur[5]. Ihr liegt folgende Überlegung zugrunde:

Wenn die Wahrscheinlichkeit, in einem Test unter der Null-Hypothese kein signifikantes Ergebnis zu erhalten, mit e(1) angegeben wird, dann ist die Wahrscheinlichkeit in n unabhängigen Tests, für die die Null-Hypothese zutrifft überhaupt kein signifikantes Ergebnis zu erhalten e(n) = e(1) ** n. Um eine Irrtumswahrscheinlichkeit p(n) = 1 - e(n) einzuhalten, muss also jeder Einzeltest auf dem Niveau von p(1) = 1 – ((1 - p(n)) ** (1/n)) getestet werden. Für das konkrete Beispiel bedeutet das bei einem Signifikanzniveau von 5%:

p(1) = (1 - ((1-0.05) ** (1/100)) = 0.00051

und bei einem Signifikanzniveau von 1%

p(1) = (1 - ((1-0.01) ** (1/100)) = 0.00010.

4.2 Automatisierte Verarbeitung mit SAS

Prinzipiell ist die automatisierte Verarbeitung der Ergebnisse von SAS-Prozeduren sehr einfach geworden: Während mit früheren SAS-Versionen noch umständlich das Ergebnis der Prozeduren in eine Text-Datei geschrieben und dann mit mühsam programmierten Data-Steps wieder eingelesen wurde, reicht es heute vollkommen aus, die Ergebnisse mit dem Output Delivery System, ODS, in SAS-Dateien zu schreiben. Das Problem besteht nur darin, die richtige „ods output“-Anweisung hinzuschreiben.

---

[3] Ein gewichtiger Einwand zu diesem Abschnitt: In vielen Fällen sind die Verteilungsannahmen für Signifikanztests nicht erfüllt. Das trifft auf dieses Experiment in extremer Weise zu. Dieser Umstand sollte aber nicht dazu führen, auf die Überlegungen aus der Inferenzstatistik vollständig zu verzichten und die Ergebnisse der Analysen unbesehen zu glauben. In Unkenntnis der Verteilung der Residuen ist die Annahme, der Stichprobenfehler sei kleiner als bei erfüllten Verteilungsannahmen nicht zulässig. Daraus folgt zwar, dass „signifikante Ergebnisse“, für sich genommen, in diesen Fällen nicht als ausreichende Bestätigung einer Hypothese angesehen werden dürfen. Umgekehrt bleibt aber der Vorbehalt, dass bei „nicht signifikanten Ergebnissen“ ein besonderes Misstrauen angebracht ist. Erst, wenn man in Kenntnis der Verteilungen zeigen könnte, dass der Stichprobenfehler im konkreten Fall kleiner sein muss, als unter den Bedingungen, die im konkreten Test angenommen werden, wäre das konservative Verwerfen „nicht signifikanter Ergebnisse“ in Frage zu stellen.
Wenn die Verteilungen, wie im konkreten Fall, bekannt sind, kann die Auswirkung der verletzten Verteilungsannahmen auch mit einer Simulation geprüft werden. Im konkreten Fall zeigt sich, dass der Signifikanztest sehr robust ist. (Das Testprogramm ist bei den Beispielprogrammen enthalten).

[4] Der „Risikomanager“ eines Finanzdienstleisters sagte neulich zu mir im Rahmen eines Basel II-Projekts: „Mir ist vollkommen egal, woran ich die Ausfallwahrscheinlichkeit eines Kunden erkennen kann, wenn ich sie erkennen kann.“

[5] Ausführliche Diskussion bei Westfall, P.H. and Young, S.S. (1993).

Wenn man eine SAS-Prozedur ausgeführt hat, lassen sich im Results-Fenster die einzelnen Elemente mit ihren Namen ausklappen.
Beispiel:

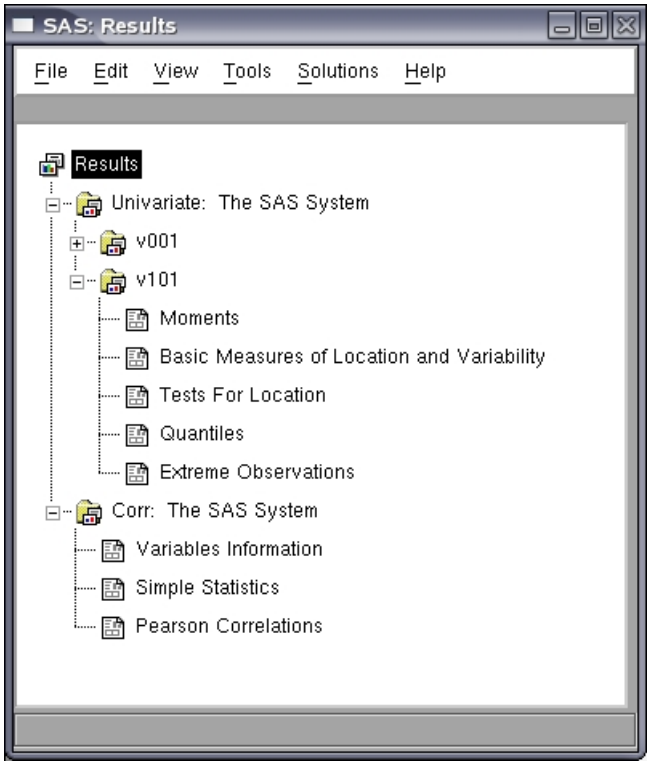


Abbildung 1: Das Results-Fenster in SAS

Aus jedem dieser Elemente lässt sich dann eine Datei erzeugen, indem man unmittelbar vor dem Prozedur-Aufruf mit einer „ods output"-Anweisung angibt, in welche Datei welches Element geschrieben werden soll.

Beispiele:

```
ODS output     'Moments' = work.moments
               'Tests for Location' = work.location;
Proc univariate data = wsksfe.daten;
Run;
```

oder:

```
ODS output     'Pearson Correlations' = work.corr ;
Proc corr data = wsksfe.daten;
     Var e1;
     With v001 - v100;
Run;
* und die Ausgabe: ;
Proc print data = work.corr;
     Where Pe1 LT 0.0001;
     Title "Signifikante Korrelationen bei Gesamt-P von 0.01";
Run;
```

4.3 Anwendung im Beispiel

Erwartungsgemäß korreliert keine der unabhängigen Variablen auf dem Signifikanzniveau von 0,01%. Allerdings lassen sich Effekte in der eben „gefundenen“ Größenordnung bei ausreichender Fallzahl sehr wohl identifizieren. Bei der Erstellung Beispieldaten wurde mit der folgenden Konstruktion eine solche Abhängigkeit erzeugt:

```
e1 = (ranuni(0)
      + 0.02 * (v101 )/3
      + 0.01 * (v102 )/3
      + 0.01 * (v103 )/3
     ) LT 0.1;
```

Die Variable V101 korreliert dann mit einem r von 0,02, die anderen beiden korrelieren mit einem r von 0,01. Das reicht, um in den beiden Gruppen einen Unterschied beim Anteil der Ausfälle von ungefähr einem Prozentpunkt (9,5% zu 10,5%) zu produzieren.

Bei einer Million Datensätzen, sind diese Korrelationen auch auf einem hohen Niveau signifikant ($p < 1*e^{-10}$). Die so erzeugte Formel liefert dann für die Grundgesamtheit das entsprechende Ergebnis. Beispiel:

Tabelle2: Korrelationen aus der Grundgesamtheit
Kritischer P-Wert für 1%-Niveau: 0.0000975

Obs	Variable	Korrelation	P-wert
1	v101	-0.02024	0,000000000
2	v102	-0.01168	0,000000000
3	v103	-0.01139	0,000000000
4	v045	0.00256	0,010553093
5	v012	0.00234	0,019465732

Daraus ergibt sich die Formel:

e1prob = -0.02024 * v101 + -0.01168 * v102 + -0.01139 * v103

und e1pred = e1prob > 0

mit dem Ergebnis von zwei Gruppen, in denen sich der Anteil der e1-Beobachtungen um etwas mehr rund 1%-Punkt unterscheidet.

Tabelle 3: Ergebnis

		e1		All
		0	**1**	
e1pred				
0	**N**	453.909	46.678	500.587
	PctN	90,68	**9,32**	100,00
1	**N**	446.437	52.976	499.413
	PctN	89,39	**10,61**	100,00
All	**N**	900.346	99.654	1000000
	PctN	90,03	9,97	100,00
e1prob	**Mean**	-,0002	0,0020	-,0000
	Std	0,0260	0,0259	0,0260
	Var	0,0007	0,0007	0,0007

Dieser Effekt kann in einer Stichprobe, wie oben beschrieben, nur noch zufällig entdeckt werden. Dabei sind die Ergebnisse natürlich zufällig. Ein noch relativ „gutes“ Ergebnis wäre das folgende, bei dem zumindest zwei der drei Prädiktoren auftauchen:

Tabelle 4.1: Korrelationen aus der Stichprobe
Kritischer P-Wert für 1%-Niveau: 0.0000975

Obs	Variable	Korrelation	P-Wert
1	v035	-0.03782	0,009125297
2	v068	0.03729	0,010147816
3	**v101**	-0.03052	0,035384296
4	**v102**	-0.03012	0,037873557
5	v086	0.02956	0,041562052
6	v005	0.02657	0,066985466
7	v077	0.02485	0,086770270
8	v059	0.02385	0,100131332
9	v078	0.02352	0,104929881
10	v017	0.02346	0,105890890

Es kann aber auch schlechter kommen:

Tabelle 4.2: Korrelationen aus der Stichprobe
zweites Beispiel

Obs	Variable	Korrelation	P-Wert
1	v047	0.04381	0,002822984
2	v003	-0.04348	0,003039038
3	v072	0.03695	0,011799971
4	v079	0.03528	0,016206294
5	v089	-0.03506	0,016894961
6	v083	-0.03480	0,017706439
7	v069	0.03434	0,019287246
8	v011	-0.03282	0,025295183
9	v071	-0.03130	0,032953242
10	v095	0.02971	0,042917979

Insgesamt zeigen diese Beispiele m.E. anschaulich, was theoretische Überlegungen ohnehin ergeben: Es kann nicht sinnvoll sein, ohne Not die Fallzahl und damit auch den Informationsgehalt eines Datenbestandes vor der Analyse zu reduzieren. Vielmehr muss die Qualität der Analyse auf anderen Wegen, z.B. über Signifikanztests, erreicht werden. Bei extremen Verteilungen muss das Analyse-Verfahren, z.B. die Diskriminanzanalyse, entsprechend parametrisiert werden.

5 Validierung

Ein wesentlicher Einwand gegen die Verwendung aller Daten in der Analyse besteht darin, dass es dann nicht mehr möglich ist, das Ergebnis mit einer unabhängigen Teststichprobe zu testen. Besondere Bedeutung erlangt dieses Argument, wenn die Verteilungsannahmen der Signifikanztests verletzt sind, was in diesem Beispiel sehr extrem der Fall ist. Und wenn für ein Verfahren, z.B. neuronale Netze, die Verteilung des Stichproben-Fehlers nicht bekannt ist, kann es auch gar keine Signifikanztests geben.

Trotzdem meine ich, dass es gute Gründe gibt, bei der Analyse alle bekannten Fälle[6] einzubeziehen:

1. Wenn sich z.B. bei einem neuronalen Netz zeigt, dass die Ergebnisse einer Teststichprobe den Erwartungen entsprechen, gibt es keine Möglichkeit, dieses Ereignis als Zufall zu erkennen. Und die Wahrscheinlichkeit einer „zufälligen Bestätigung“ ist nicht kleiner als in dem oben beschriebenen Experiment.
2. Wenn aber in der Teststichprobe die erwarteten Ergebnisse nicht erreicht werden, führt das in der Regel nicht zum Projektabbruch. Die nächsten Versuche, vielleicht mit neuen Lern- und Test-Stichproben, erhöhen dann die Wahrscheinlichkeit irgendwann doch eine „Bestätigung“ zu finden – auch dann, wenn dieses Modell in der Wirklichkeit nicht „besser“ als eines der vorhergegangen ist.
3. Die Analyse der Gesamtheit der Daten führt dazu, dass der Stichprobenfehler (nach dem Gesetz der großen Zahl) minimiert wird und bei jedem Versuch derselbe bleibt.
4. Die Bedeutung des Stichproben-Fehlers in einer Analyse lässt sich auch daran erkennen, dass das Analyse-Ergebnis für kleinere Stichproben aus dieser Gesamtheit gilt.

Das letzte Argument möchte ich noch etwas erläutern: Bei kleinerer Fallzahl ist zu erwarten, dass der Stichprobenfehler größer ist, als der in der zur Verfügung stehenden Gesamtheit. Wenn man also aus der Gesamtheit der zur Verfügung stehenden Daten viele kleinere Stichproben zieht, und in allen Fällen, der erwartete Effekt erkennbar ist, kann man daraus schließen, dass der Stichprobenfehler bei dieser kleineren Fallzahl kleiner ist, als der vorher gefundene Effekt des Modells. Der Verzicht auf die Unabhängigkeit von Lern- und Test-Stichprobe ermöglicht es dann, das Modell mit vielen kleineren Stichproben zu testen, und so das Risiko einer zufälligen „Bestätigung“ eines nicht validen Modells drastisch zu verringern.

Hinzu kommt in vielen Fällen ein „Nutzen-Argument“, welches oft vernachlässigt wird: Der Stichprobenfehler tritt ja nicht nur in den Analysen auf. Wenn z.B. für die Analyse ein großer Datenbestand zugrunde gelegt wurde, aber in der Praxis die darauf gestützten Entscheidungen relativ selten getroffen werden, kann es vorkommen, dass sogar ein valides Modell praktisch irrelevant wird[7]. Viele kleine Stichproben verdeutlichen diesen Effekt. Wenn in dem Beispielfall mit 1 Million Datensätzen monatlich nur 10.000 neue Verträge abgeschlossen werden, kann man mit dieser Stichprobengröße den erwarteten monatlichen Effekt, bzw. dessen Streuung schätzen. Dies wurde für die folgende Tabelle durchgeführt. Die in der Grundgesamtheit gefundene Formel wurde auf 20 Stichproben mit jeweils rund 10.000 Datensätzen angewandt.

Tabelle 5: Unterschied der „Ausfälle“ in den verschiedenen Stichproben

Lauf-Nr.	Quelle	"gute" Gruppe	"schlechte" Gruppe	Unterschied
1	Grundgesamtheit	9.32%	10.6%	1.28%
5	Stichprobe	8.31%	11.1%	2.77%
12	Stichprobe	8.43%	10.8%	2.40%
16	Stichprobe	9.12%	10.9%	1.77%
7	Stichprobe	9.01%	10.8%	1.77%
8	Stichprobe	9.54%	11.3%	1.73%
15	Stichprobe	8.95%	10.7%	1.72%
17	Stichprobe	9.39%	11.1%	1.67%
9	Stichprobe	9.00%	10.6%	1.63%
18	Stichprobe	8.62%	10.1%	1.50%
13	Stichprobe	9.37%	10.6%	1.23%
10	Stichprobe	9.91%	11.0%	1.07%

[6] Ausreißer sind ein vollkommen anderes Problem und werden in dieser Argumentation ignoriert.

[7] Das erleben z.B. Roulette-Spieler, die wissen, das „rot“ in (fast) der Hälfte aller Fälle fällt, und sie deshalb nichts verlieren können, wenn sie immer auf „rot“ setzen, und dabei ihren Einsatz solange verdoppeln, bis sie gewonnen haben. Statistisch gesehen reicht in 999 von 1000 Fällen ein Kapital von 2.046 Euro, um bei einem Einstand mit 2 Euro am Ende 2 Euro zu gewinnen. Aber nach der Wahrscheinlichkeitsrechnung sind über 6 von 100 Fälle zu erwarten, bei denen mindestens 4 mal hintereinander die falsche Farbe kommt.

Lauf-Nr.	Quelle	"gute" Gruppe	"schlechte" Gruppe	Unterschied
6	Stichprobe	10.1%	11.0%	0.92%
19	Stichprobe	9.51%	10.4%	0.85%
20	Stichprobe	9.60%	10.4%	0.85%
2	Stichprobe	9.67%	10.5%	0.84%
1	Stichprobe	9.75%	10.5%	0.78%
11	Stichprobe	9.82%	10.5%	0.72%
14	Stichprobe	9.62%	9.86%	0.24%
4	Stichprobe	9.52%	9.61%	0.09%
3	Stichprobe	9.93%	9.80%	(0.13%)

Man sieht, dass die Trennschärfe bei diesen kleineren Fallzahlen stark variiert, und in 4 Fällen unter 0,75%-Punkte fällt. Bei kleineren Stichprobengrößen wäre die Varianz entsprechend größer, und wenn z.B. eine kleine Sparkasse oder Volksbank monatlich nur 10 Großkredite vergibt, stellt sich die Frage, ob das ausgefeilte Rating-Verfahren einer Großbank wirklich anwendbar ist, um über die Abschätzung des Risikos hinaus die sinnvolle Unterlegung dieser Darlehen mit Eigenkapital zu bestimmen.

6 Zusammenfassung

Rein empirisch angelegte Analysen ohne theoretische Fundierung erfordern besonders große Sorgfalt, um zufällige Stichprobenfehler zu erkennen und ihre Fehlinterpretation zu vermeiden. Die Inferenzstatistik bietet dazu mit der Bonferroni-Korrektur ein geeignetes Instrument.

Der Versuch, ein Analyse-Ergebnis durch die Trennung in Lern- und Teststichprobe zu validieren, ist dagegen mit einer relativ hohen Fehler-Wahrscheinlichkeit behaftet. Besser ist es, die Gesamtheit aller verfügbaren Daten in die Analyse aufzunehmen, und das Ergebnis anschließend mit vielen kleineren Stichproben aus demselben Datenbestand zu validieren.

Literatur

[1] Schollenberger, W.: Die praktische Anwendung der Diskriminanzanalyse zur Gruppierung im Data Mining in: Proceedings der 5. Konferenz der SAS-Anwender in Forschung und Entwicklung, Universität Hohenheim, 2001

[2] Westfall, P.H., Young, S.S.: Resampling-based multiple testing, John Wiley & Sons, New York, 1993

Programme

Wenn Sie diese Experimente selbst durchführen wollen, finden Sie die zugehörigen SAS-Programme auf
http://www.ws-unternehmensberatung.de/KSFE2007/index.html

Anhang 1: Verteilung der Prüf-Stichproben

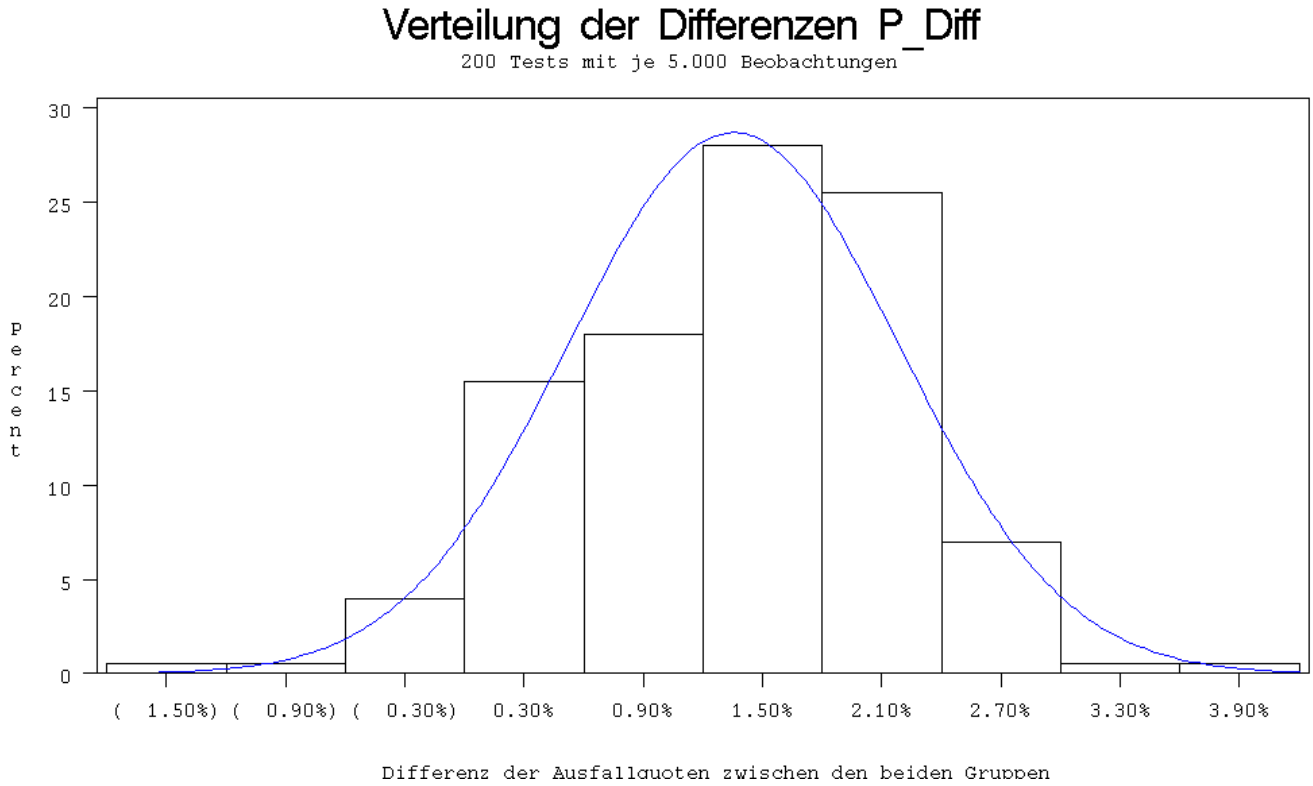
Im Anschluss an den Vortrag auf der KSFE ergab eine Diskussion, dass es sinnvoll ist, die Validierung im Abschnitt 5 mit einer Serie von Stichproben und der Verteilung der Stichprobenergebnisse durchzuführen. Z.B. könnte man die Daten zufällig in n Proben mit gleicher Fallzahl unterteilen, und die Verteilung der Differenzen zwischen der „guten“ und der „schlechten“ Gruppe prüfen.

Dies wurde hier mit 4 verschiedenen Probengrößen simuliert.

Anzahl der Proben	Größe der Proben	mittlere Differenz	maximale Differenz	minimale Differenz	Standard-abweichung
100	10.000	1,35%	3,00%	0,23%	0,57%
200	5.000	1,35%	4,01%	-1,44%	0,83%
500	2.000	1,35%	5,58%	-1,99%	1,32%
1.000	1.000	1,35%	9,30%	-5,80%	1,96%

Erwartungsgemäß steigt die Varianz mit sinkender Probengröße und die Extreme nehmen zu. Aber dir Mittelwert aus den Stichproben entspricht in allen Fällen der Gruppendifferenz in der Gesamtheit – es wurde hier ein neuer Datensatz mit anderen Werten verwendet als in den Beispielen des Vortrags.

Als Beispiel für die Verteilung wird hier das Ergebnis aus 200 Proben mit je 5.000 Beobachtungen wiedergegeben.

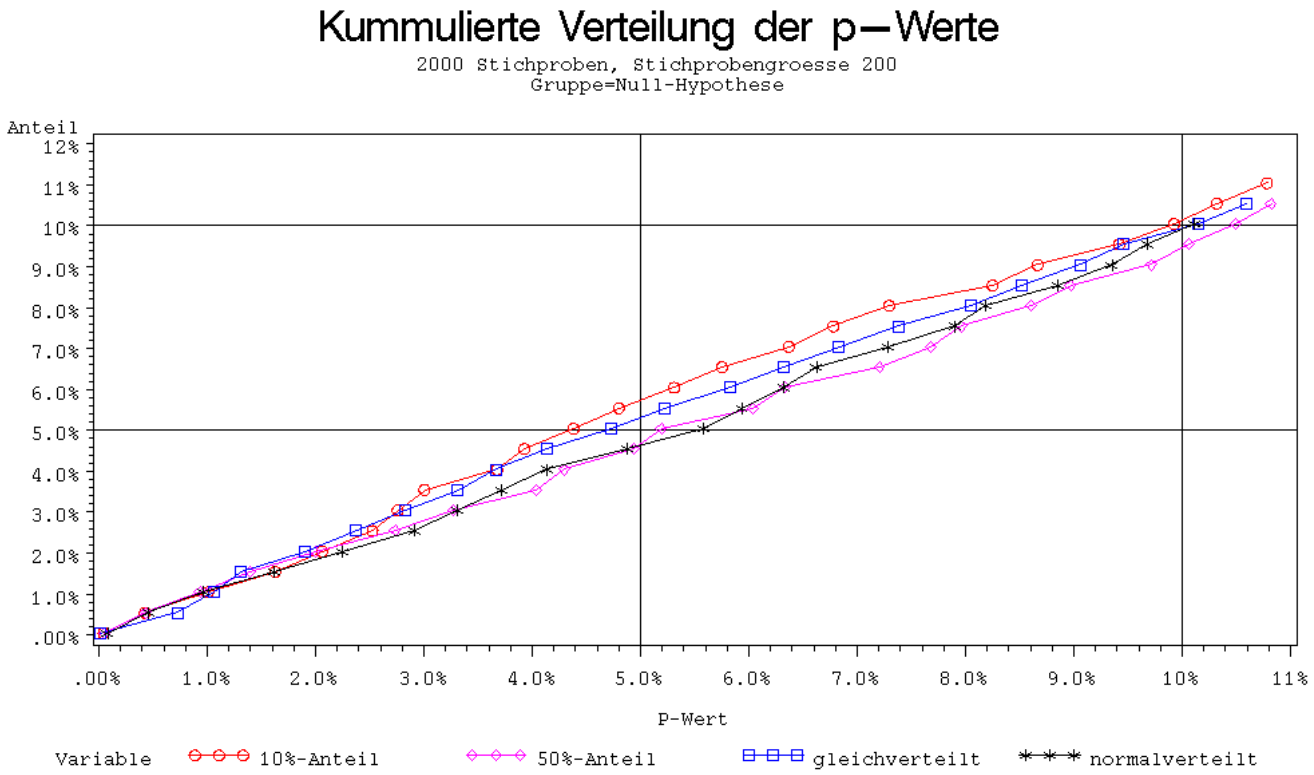


Anhang 2: Validierung des Signifikanztests

Der in dem Experiment verwendete t-Test[8] gilt als robust gegenüber nicht normalverteilten Residuen. Angeregt durch einen entsprechenden Vortag auf der KSFE lag es nahe, dies auch durch eine Simulation zu zeigen, und zu prüfen, wie gut die Power des Tests bei schwachen Korrelationen und kleineren Stichproben ist.

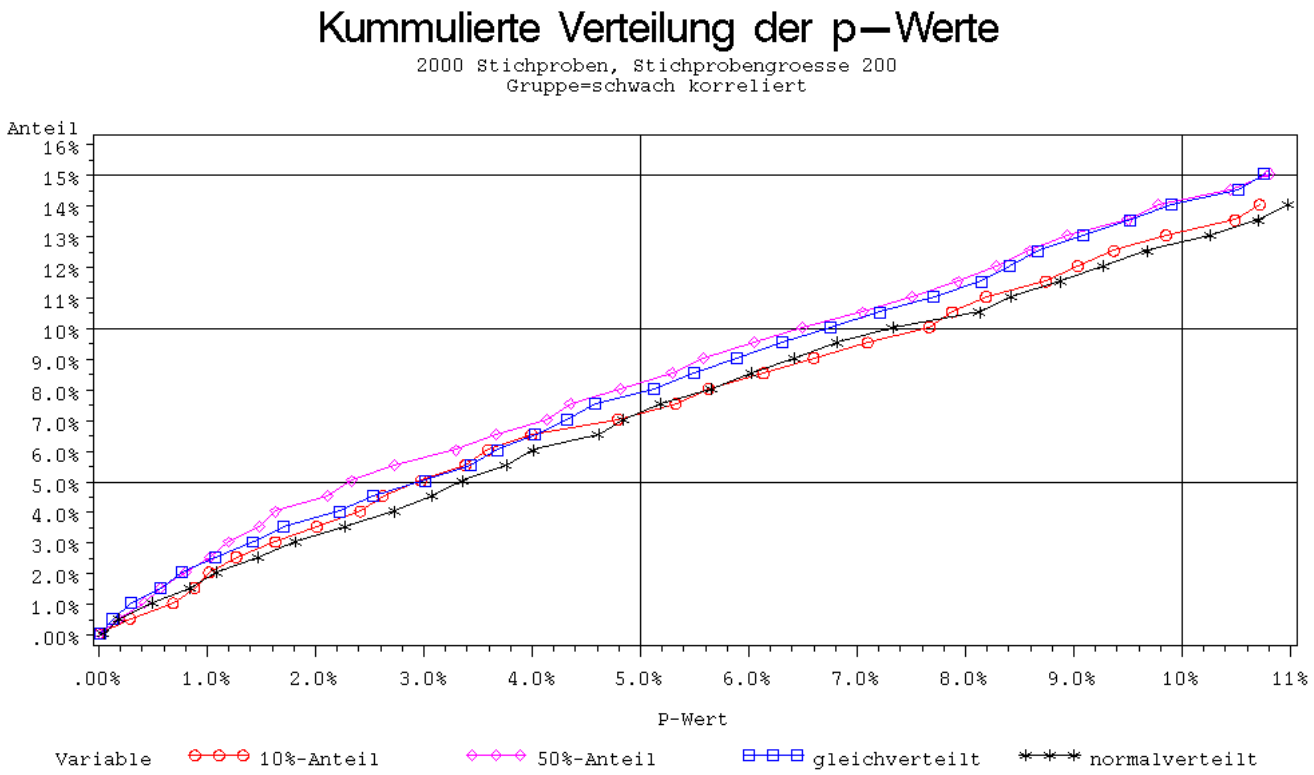
Für das Beispiel wurden 2000 Stichproben mit jeweils 200 Beobachtungen erzeugt. Die unabhängige Variable (Prädiktor) ist normalverteilt. Jeweils vier Variablen mit verschiedenen Verteilungen, normalverteilt, gleichverteilt, dichotom 50:50 und dichotom 10:90, sind mit der ersten unkorreliert und schwach korreliert (r um 0,03). Für jede Variable werden somit 2000 Korrelationen und die zugehörigen Signifikanztests gerechnet. Dann wird die kumulierte Verteilung der p-Werte ermittelt und kontrolliert.

Bei den unkorrelierten Variablen dürfen z.B. maximal 5 % aller p-Werte einen p-wert kleiner oder gleich 0.05 haben. Ein Beispiel für das Ergebnis einer solchen Prüfung ist im folgenden Plot dargestellt.



Bei den korrelierten Variablen soll natürlich der Anteil der kleinen p-Werte möglichst groß sein. Ein Beispiel ist im folgenden Plot dargestellt.

[8] $t = (n\text{-}2)^{1/2} \; ([(r^2)/(1\text{-}r^2)])^{1/2}$ ist mit n-2 Freiheitsgraden t-verteilt.



Wie man leicht erkennen kann, ist der t-Test bei dieser kleinen Stichprobengröße nicht geeignet, die schwachen Korrelationen zu erkennen. Bei größeren Stichproben wird das Ergebnis natürlich besser, und wenn man die Gesamtheit aus allen Stichproben analysiert, lassen sich die vier Korrelationen eindeutig erkennen, wie die folgende Tabelle zeigt:

Pearson Correlation Coefficients, N = 400000 Prob > |r| under H0: Rho=0	
	Normalverteilter Prädiktor
unkorreliert	
normalverteilt	0.00124 0.4338
gleichverteilt	-0.00185 0.2409
dichotom 50%	0.00256 0.1049
dichotom 10%	0.00149 0.3460
korreliert	
normalverteilt	0.03301 <.0001
gleichverteilt	0.03520 <.0001
dichotom 50%	-0.03596 <.0001
dichotom 10%	-0.03321 <.0001