SAS – Macros zur Durchführung von Permutationstests

Matthias Frisch* und Erich Schumacher**
* Institut für Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
** Institut für Angewandte Mathematik und Statistik
Universität Hohenheim, 70593 Stuttgart
Email: frisch@uni-hohenheim.de, schumach@uni-hohenheim.de

Zusammenfassung

Permutationstests bieten verteilungsfreie Alternativen zur klassischen Varianzanalyse. Insbesondere im Feldversuchswesen bei der Auswertung von Boniturnoten testen sie Hypothesen, welche die zu untersuchende Fragestellung adäquater beschreiben als die einer auf Normalverteilungsannahme beruhenden Varianzanalyse mit fixen Effekten. Nach einer Einführung in die Theorie der Permutationstests werden SAS–Macros vorgestellt in denen Globaltests, paarweise Vergleiche und sequentielle Testverfahren zur Durchführung von Permutationstests implementiert sind. Der Aufruf der Macros sowie die Interpretation des Outputs werden beschrieben.

Einleitung

Bei der Auswertung von Feldversuchen bei denen Boniturnoten erfaßt werden, stellt sich oft das Problem, daß die durch eine Varianzanalyse getesteten Hypothesen die zu untersuchende Fragestellung nur schlecht beschreiben. So macht es offensichtlich bei den in Abbildung 1 gezeigten Boniturwerten keinen Sinn, die Nullhypothese

H_0: “Alle Stichproben stammen aus einer normalverteilten Grundgesamtheit“ gegen die Alternativhypothese

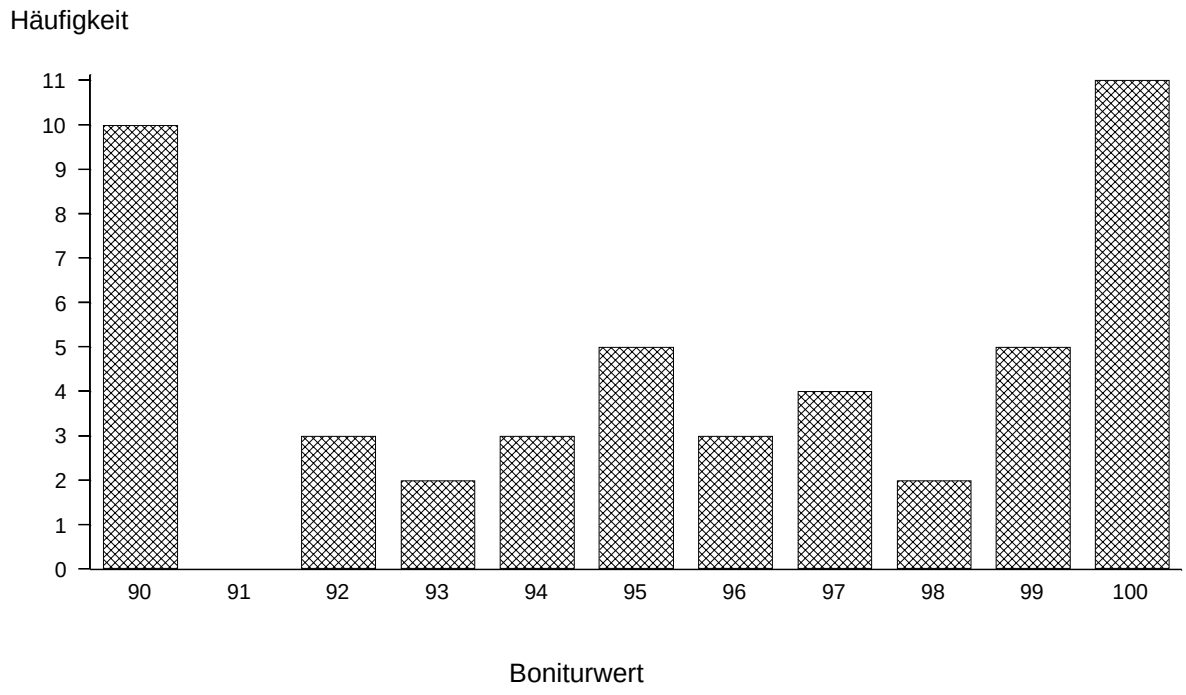
H_A: “Die Stichproben stammen aus mindestens zwei normalverteilten Grundgesamtheiten mit der selben Varianz und verschiedenen Erwartungwerten“ zu testen. Null- und Alternativhypothesen, welche die Frage nach Behandlungseffekten besser beschreiben, können z.B. lauten

H_0: “Alle Stichproben stammen aus einer Grundgesamtheit beliebiger Verteilung“ und

H_A: “Die Stichproben stammen aus mindestens zwei Grundgesamtheiten unterschiedlicher Verteilung.“

Solche Nullhypothesen lassen sich mit Hilfe von Permutationstests prüfen. Synonym werden auch die Begriffe Randomisations- oder Rerandomisationstests verwendet.

Abbildung 1: Häufigkeiten von Boniturnoten in einem Feldversuch.



Theoretische Grundlagen

Zur Entwicklung eines Tests muß ein Maß für die Ungleichmäßigkeit der Zuordnung der Beobachtungen zu den Prüfgliedern definiert werden. Bei balancierten einfaktoriellen Versuchen kann als Teststatistik die Summe der quadrierten Behandlungstotale verwendet werden (Edgington, 1995):

$$Z = \sum_{i=1}^{k} \left(\sum_{j=1}^{n} x_{ij} \right)^2 ,$$

mit *k* Prüfgliedern und *n* Beobachtungen je Prüfglied. Bei unbalancierten Versuchen wird eine Gewichtung mit dem Stichprobenumfang vorgenommen:

$$Z = \sum_{i=1}^{k} \left[\frac{1}{n_i} \left(\sum_{j=1}^{n_i} x_{ij} \right)^2 \right],$$

mit n_i = Anzahl der Beobachtungen des i-ten Versuchsglieds. Weitere Teststatistiken, die in den Makros umgesetzt wurden, z.B. für zweifaktorielle unbalancierte Versuche, beschreibt Edgington (1995).

Die Verteilung der Teststatistiken unter der Nullhypothese kann exakt bestimmt werden, indem sie für alle möglichen Zuordnungen der Beobachtungen zu den Prüfgliedern (Permutationen) berechnet wird. Eventuelle Randomisationsbeschränkungen in der Versuchsanlage werden hierbei berücksichtigt. Eine exakte Bestimmung der Verteilung ist jedoch oft nicht möglich da die Rechenkapazität hierzu nicht ausreicht, in diesem Fall wird die Verteilung durch Monte-Carlo Simulation geschätzt. Die Teststatistik wird für *nperm* rerandomisierte Anordnungen der Beobachtungen zu den Prüfgliedern berechnet. Je nach Größe des Versuches werden 1000 bis 10000 Rerandomisationen durchgeführt. Mit Hilfe der Verteilung der Prüfgröße unter der Nullhypothese berechnen die Macros Überschreitungswahrscheinlichkeiten für die beobachteten Realisierungen der Prüfgröße.

Analog zum multiplen t-Test können mit auch bei Randomisationstests paarweise Vergleiche zwischen zwei Prüfgliedern durchgeführt werden. Die vorgegebene

Irrtumswahrscheinlichkeit erster Art wird dann für jeden einzelnen Vergleich und nicht im multiplen Sinne eingehalten. Die unkorrigierten Überschreitungswahrscheinlichkeiten können jedoch für im Voraus geplante Vergleiche, z.B. mit einer Kontrollbehandlung, verwendet werden. Durch eine Adjustierung der Überschreitungswahrscheinlichkeiten nach Sidak kann eine Einhaltung der simultanen Irrtumswahrscheinlichkeit für den Fehler erster Art garantiert werden.

Ein Verfahren zum Vergleich von Prüfgliedmittelwerten analog dem Tukey-Test schlägt Miller (1981) vor. Es werden *n* zufällige Zuordnungen der Prüfglieder zu den Versuchseinheiten erzeugt. Aus diesen wird die Verteilung der Differenzen von Prüfgliedmittelwerten unter der Nullhypothese geschätzt. Anhand dieser geschätzten Verteilung wird die Überschreitungswahrscheinlichkeit für eine beobachtete Differenz zweier Prüfgliedmittelwerte berechnet. Diesem „Tukey-Type"-Test liegt die Rerandomisation des gesamten Versuches zugrunde, deshalb hält er die simultane Irrtumswahrscheinlichkeit erster Art nur im „schwachen Sinne" (d.h. unter der Bedingung, daß die globale Nullhypothese richtig ist; Miller, 1981) ein.

Auch sequentielle Tests wie der Newman-Keuls Test oder der Ryan Test lassen sich auf Permutationstests übertragen. Hierbei wird die H_0 zuerst für den gesamten Versuch mit *k* Prüfglieder geprüft, wird diese verworfen dann wird die H_0^* für jede Untergruppe mit *k*-1 Prüfgliedern getestet. Für alle Prüfgliedgruppen der Größe *k*-1 für die H_0^* verworfen wird, wird H_0^{**} für alle Untergruppen der Größe *k*-2 getestet. Dieses Verfahren wird solange fortgesetzt bis nur noch „homogene Gruppen" von Prüfgliedern gefunden werden, d.h. Gruppen von Prüfgliedern für die die Nullhypothese nicht verworfen werden kann, oder die Nullhypothese für alle paarweisen Vergleiche abgelehnt wurde. Zur Testentscheidung wird beim Newman-Keuls Typ eines sequentiellen Tests wird die Überschreitungswahrschenlichkeit einer Untergruppe mit α verglichen, die simultane Irrtumswahrscheinlichkeit erster Art zum Niveau α wird hier nicht eingehalten. Beim Ryan-Typ des sequentiellen Tests wird der berechnete *p*-Wert jeder Untergruppe mit $\alpha \frac{k_i}{k}$ verglichen. Dabei ist k_i Anzahl der Prüfglieder in der zu prüfenden Untergruppe. Der Ryan-Typ hält die simultane Irrtumswahrscheinlichkeit erster Art ein, hat jedoch im allgemeinen eine geringe Güte. So ist im eingangs vorgestellten Versuch (k = 12) bei einem Vergleich von zwei Prüfgliedern und α = 0,05 die zu unterschreitende Schranke für $\alpha^* = 0{,}05 \bullet 2 / 12 = 0{,}0083$. Da es jedoch maximal $\frac{1}{2}\frac{8!}{4!\cdot 4!} = 35$

verschiedene Permutationen der Beobachtungen gibt kann die beobachtete Überschreitungswahrscheinlichkeit nicht kleiner als p_{min} = 1/35 = 0,028 werden. Somit können auf dem multiplen Niveau α = 0,05 keine Signifikanzen gefunden werden.

Petrondas und Gabriel (1983) schlagen ein Verfahren zur Verbesserung der Güte bei sequentiellen Permutationstests vor. Hierbei werden alle Nullhypothesen die nach Newman-Keuls beibehalten werden bei ($\alpha < p_i$) beibehalten, alle Nullhypothesen die nach Ryan abgelehnt werden ($p_i < \alpha \frac{k_i}{k}$) werden abgelehnt. Ist ($\alpha \frac{k_i}{k} < p_i < \alpha$) dann wird H_0 dann beibehalten, wenn für alle Teilmengen T_j die im Komplement von T_i liegen gilt: $\alpha \frac{k_j}{k} < p_j$. Diese Modifikation garantiert die Einhaltung der simultanen Irrtumswahrscheinlichkeit erster Art zum Niveau α und eine deutlich bessere Güte als das Verfahren nach Newman-Keuls.

Die Grenzen für die Möglichkeit der exakten Berechnung der Verteilung der Prüfgröße unter der Nullhypothese werden schnell erreicht, so existieren im Beispielversuch (Globaler Test bei 12 Prüfgliedern, 4 Wiederholungen, vollständige Randomisation) $\frac{48!}{(4!)^{12}} = 3{,}39 \cdot 10^{44}$ mögliche Zuordnungen der Beobachtungen zu den Prüfgliedern. Die Prüfgröße läßt sich hier nicht mehr exakt berechnen. Wird die Verteilung mit Hilfe eine Monte-Carlo Simulation geschätzt, läßt sich Vertrauensintervall für die Überschreitungswahrscheinlichkeit über eine Normalapproximation bestimmen. Bei kleinen Versuche zeigen Permutationstests eine geringe Trennschärfe, so ist für die Durchführung von Paarweise Vergleiche in einer Blockanlage (5 Prüfglieder, 4 Blöcke) die kleinste zu erreichende Überschreitungswahrscheinlichkeit p_{min} = 1/2!^{4} = 0,17. Hier läßt sich durch Korrektur der Beobachtungen um einen Blockmittelwert (Median, Arithmetisches Mittel) und eine einfaktorielle Auswertung die Güte verbessern. Die kleinste zu erreichende Überschreitungswahrscheinlichkeit ist dann p_{min} = 1/ $\binom{8}{4}$ = 0,014. Voraussetzung für diese Vorgehensweise ist jedoch, daß die Blockeffekte des Modells additiv sind.

Macros

Die oben genannten Tests wurden für verschiedene Versuchsanlagen in Macros implementiert. Für einfaktorielle und zweifaktorielle vollständig randomisierte Versuchsanlagen (CRD), für einfaktorielle vollständig randomisierte Versuchsanlagen (RCBD) und für balancierte unvollständige Blockanlagen (BIBD) wurden Macros programmiert. Bei den Macros für Blockanlagen werden für Fehlstellen Ersatzwerte berechnet (auch hier ist dann jedoch die Additivität der Blockeffekte Voraussetzung). Die folgende Übersicht zeigt im einzelnen, welcher Test für welche Versuchsanlage zu umgesetzt wurde.

Die Macros lassen sich auch in anderen Bereichen als zur Auswertung von Boniturnoten in Feldversuchen anwenden. So läßt sich das Macro für ein CRD ganz allgemein auf Versuche anwenden bei denen die Voraussetzungen für eine zweifaktorielle balancierte Varianzanalyse nicht gegeben sind.

	CRD	**RCBD BIBD**	**BD nach Korrektur um Blockmedian**
Exakte Bestimmung der Verteilung der Prüfgröße			
Globaler Test	*	*	
Paarweise Vergleiche / mit Sidak Korrektur	*	*	
Newman-Keuls Test		*	
Petrondas-Gabriel Test		*	
Bestimmung der Verteilung der Prüfgröße durch Simulation			
Globaler Test	*	*	*
Paarweise Vergleiche / mit Sidak Korrektur	*	*	*
Mittelwertvergleich Tukey-Typ	*	*	*
Newman-Keuls Test	*	*	*
Petrondas-Gabriel Test	*	*	*

Das folgende Beispiel zeigt einen Aufruf des Marcros 'ribdperm.mac', es läßt sich zur Auswertung von RCBD, BIBD sowie von Blockanlagen mit Fehlstellen verwenden. Dargestellt ist die Auswertung einer vollständig randomisierten Blockanlage, bei der zwei Werte ausgefallen sind. Da die Auswertung von Fehlstellen additive Blockeffekte voraussetzt wird hier eine logit-Transformation der Daten vorgenommen. Nach dem Data-Step wird das Macro geladen (evtl. ist noch eine Pfadangabe der Quelldatei nötig) und aufgerufen. Beim Markroaufruf werden die durchzuführenden Tests sowie Angaben zum Dataset als Parameter übergeben. Für eine vollständige Übersicht der möglichen Optionen vgl. Schumacher und Frisch (1997).

```
DATA mehltau;
kennung = 'vers_1';
DO block = 1 to 5;
 DO pruefgl = 1 to 5;
  INPUT bonitur@@;
  logit=((bonitur+1)/ log(100+1-bonitur));
  OUTPUT;
 END;
END;
CARDS;
 8  7  3  .    0.1
12  8  4  0.2  0.1
11  9  .  0    0.3
10  5  1  1    0
15  6  2  1    0.2
RUN;

%inc 'ribdperm.mac';       /* Einfügen der Makrodatei       */
%ribdperm (                /* Aufruf des Makros             */
g,                         /* Globaltest                    */
pw_c,                      /* Blockkorrigierter Sidak Test  */
hg_pg,                     /* Petrodas Gabriel Test         */
daten   = mehltau,         /* Name des Datasets             */
versuch = kennung,         /* Bezeichung für den Versuch    */
block   = block,           /* Variable für Block            */
beh     = pruefgl,         /* Variable für Behandlung       */
wert    = logit,           /* Zielgröße                     */
nperm   = 10000,           /* Anzahl der Simulationen       */
alpha   = 0.05             /* Fehler erster Art             */
);
```

In jedem Fall wird vom Macro der folgende Output erzeugt:

```
Bezeichnung des Versuches, Anzahl der Simulationsschritte und
multiples Alpha

                     VERS       NPERM     ALPHA
                     vers_1     10000      0.05
```

Für den globalen Test wird die Nullhypothese verworfen, aufgrund der hohen Anzahl von Monte-Carlo Simulationen wird das Konfidenzinterfall für die Überschreitungswahrscheinlichkeit klein.

```
Globaler Test (Simulation) H0:"Keine Behandlungsunterschiede"

                              ERGEBNIS
                    Pr>Perm          0

        0.99 Konfidenzintervall für Pr
                    ERGEBNIS
                    0 0.0005297
```

Der Petrondas-Gabriel Test findet keine Prüfgliedunterschiede zum Niveau α=0.05.

```
Homogene Gruppen, Petrondas/Gabriel (Simulation)

ERGEBNIS
1        2        3        4        5        Pr>Perm  Pr*>Perm
                  a        a        a        0.047    0.078
         b                          b        0.066
         c                 c                0.065
         d        d                          0.060
e                                   e        0.063
f                          f                 0.062
g                 g                          0.066
h        h                                   0.060
```

Beim t-Test-Typ werden signifikante Prüfgliedunterschiede gefunden,

```
Paarweise Vergleiche (Simulation, Korrektur um Blockmedian)
p-Werte nicht adjustiert

                ERGEBNIS
         Pr>Perm      1     2     3     4     5
                1           0.044 0.009 0.010 0.007
                2                 0.008 0.008 0.005
                3                       0.277 0.012
                4                             0.312
                5
```

auch nach der Sidak-Adjustierung wird noch ein Prüfgliedunterschied zum Niveau α=0.05 gefunden.

```
Paarweise Vergleiche (Simulation, Korrektur um Blockmedian)
p-Werte nach Sidak adjustiert

                ERGEBNIS
         Pr>Perm      1     2     3     4     5
                1           0.360 0.086 0.092 0.071
                2                 0.076 0.073 0.047
                3                       0.961 0.116
                4                             0.976
                5
```

Literatur

Edgington, E. S., 1995: Randomisation tests. Marcel Dekker, New York.

Good, P., 1994: Permutation Tests. Springer Verlag, New York.

Miller, R.G., 1981: Simultaneous Statistical Inference. Springer Verlag, New York.

Petrondas, D.A., and K.R. Gabriel, 1983: Multiple Comparisons by randomisation Tests. Journal of the American Statistical Association, *78*, 949-957.

Schumacher, E. und M. Frisch, 1995: Permutationstests zur Analyse von Boniturwerten in einfachen Versuchsanlagen. Zeitschrift für Agrarinformatik *3*, 107-113.

Schumacher, E. und M. Frisch, 1997: Ein SAS-Macro zur Durchführung von Permutationstests in vollständigen und unvollständigen Blockanlagen. Zeitschrift für Agrarinformatik *5:* 125-130.