Hinweise zur statistischen Beratung

Prof. Dr. Mario Hasler
Lehrfach Variationsstatistik

Christian-Albrechts-Universität zu Kiel
hasler@email.uni-kiel.de

1 Allgemein

Der größte zeitliche Anteil statistischer Beratung entfällt i.d.R auf das Kommunizieren zwischen Statistiker und Versuchsansteller. Dies ist unumgänglich, schon allein deshalb, weil beide Seiten unterschiedliche Sichtweisen haben und unterschiedliche Fachbegriffe verwenden. Darüber hinaus ist aus Sicht des Statistikers leider oft viel Zeit nötig für allgemeine Sachverhalte, über die bereits vor der Beratung Klarheit herrschen sollte, vom Versuchsansteller häufig gestellte Fragen und häufig begangene Fehler bzw. Missverständnisse. Mögliche Fragen sind:

- Wie lauten die Versuchsfragen? ($\rightarrow$ 2)
- Wie war der Versuchsaufbau? ($\rightarrow$ 3)
- Wie müssen die Daten aufbereitet werden? ($\rightarrow$ 4)
- Wie werden Testvoraussetzungen überprüft? ($\rightarrow$ 5)
- Welche Software ist zu verwenden? ($\rightarrow$ 6)
- Wie stellt man Testergebnisse dar? ($\rightarrow$ 7)

Einige Erläuterungen hierzu sind im folgenden gegeben und sollen einerseits das statistische Verständnis verbessern und andererseits der Vereinfachung der statistischen Beratung dienen. Statistische Grundkenntnisse können und sollen hiermit nicht vermittelt werden, sondern werden vorausgesetzt.

2 Versuchsfragen

Am Anfang jeder statistischer Auswertung stehen eine oder mehrere praktische Versuchsfragen. Diese sind vom Versuchsansteller vor dem Versuch, also bevor überhaupt Daten vorliegen, präzise zu formulieren. Die statistische Auswertung/Beratung soll und kann nicht die Frage klären, was mit gegebenen Daten gezeigt werden kann! Aus den Versuchsfragen, dem Versuchsaufbau und der Verteilung der Daten resultieren dann die anzuwendenden statistischen Tests bzw. Verfahren.

Wichtig ist zudem – und hier liegt ein typischer Fehler der Versuchsauswertung – die Versuchsfragen sauber voneinander zu trennen! Wenn z.B. in einem Versuch der Einfluss der Sorte und der Düngung auf den Ertrag geklärt werden soll, dann resultieren daraus zwei voneinander unabhängige Versuchsfragen, nämlich: „(Wie) Unterscheiden sich die Erträge abhängig von der Sorte?“ und „(Wie) Unterscheiden sich die Erträge abhängig von der Düngung?“. Bei eventuell folgenden Mittelwertsvergleichen der unterschiedlichen Sorten muss nur für die Anzahl der Sortenvergleiche multiplizitätsadjustiert werden. Bei eventuell folgenden Mittelwertsvergleichen der unterschiedlichen Dünger muss nur für die Anzahl der Düngervergleiche multiplizitätsadjustiert werden. Die Sortenvergleiche sind unabhängig von den Düngervergleichen. Es sei denn, die Versuchsfrage lautet: „(Wie) Unterscheiden sich die Erträge abhängig von der Kombination aus Sorte und Düngung?“.

3 Versuchsaufbau

Beim Erläutern des Versuchsaufbaus ist es notwendig, den Versuch in Gänze zu erklären. Wichtig ist dabei weniger der fachliche Hintergrund, sondern dass alle Einflussgrößen, Kovariablen und Messgrößen benannt werden. Ziel ist es, falls möglich, ein statistisches Modell/Konzept für den Gesamtversuch zu entwickeln. Aus diesem können später Aussagen zu Teilversuchen abgeleitet werden, nicht umgekehrt! Auch wenn beispielsweise die statistische Analyse eines Versuches für zwei verschiedene Standorte separat durchgeführt werden soll, erlaubt ein gesamtheitliches statistisches Modell beider Standorte (falls möglich) i.d.R. präzisere Effektschätzungen. Außerdem wird – unwissentlich oder nicht – durch die vorschnelle Aufsplittung des Gesamtversuches oft eine nötige Multiplizitätsadjustierungen umgangen und Wechselwirkungseffekte ignoriert.

Ein Punkt, der sehr oft zu Missverständnissen führt ist: Blöcke sind formell keine Wiederholungen, auch wenn sie oft als solche fungieren. Daten

zweier unterschiedlicher Blöcke haben erwartungsgemäß – also nicht einfach aufgrund der Reststreuung – unterschiedliche Werte. Der Blockfaktor ist eine Einflussgröße. Diese kann gegebenenfalls als fixer oder zufälliger Effekt berücksichtigt werden.

4 Datenaufbereitung

Es empfiehlt sich grundsätzlich, also unabhängig von der zu verwendenden Statistiksoftware, die Daten in einem sog. Flat-File-Format aufzulisten. Darunter versteht man eine Wertetabelle (vorzugsweise im xls- oder txt-Format) mit bestimmten Vorgaben für einen optimalen Überblick und zur fehlerfreien Softwarebearbeitung. Folgende Vorgaben sind zu beachten:

- Zeilen entsprechen Messobjekten, Spalten entsprechen Einflussgrößen, Kovariablen, Messgrößen
- alle Spalten mit (möglichst kurzen) Namen
- keine Verwendung von Sonderzeichen (mathematische Symbole, Umlaute, Leerzeichen o.ä.)
- möglichst sinnvolle Reihenfolge der Zeilen und Spalten (Spalten z.B. analog zum Versuchsaufbau)
- keine Leerzeilen oder -spalten
- Angabe fehlender Werte entweder durch Punkt (SAS) oder Auslassen (R), nicht durch den Wert 0!
- einheitliche Verwendung von entweder Komma oder Punkt als Dezimaltrennzeichen
- keine zusätzlichen Erläuterungen, Kommentare, Graphiken o.ä. innerhalb des Datenfiles
- keine Zwischengrößen wie Mittelwerte oder Varianzen innerhalb der Wertetabelle

Tabelle 1 dient als allgemeines, idealisiertes Beispiel. Weitere typische Spalten sind „Termin“, „Aufwuchs“ o.ä.. Achtung: Blöcke sind formell keine Wiederholungen (siehe 3)! Deswegen kann gegebenenfalls auch „Wiederholung“ als separate Spalte aufgeführt werden. Nicht alle Spalten müssen gegeben sein. Andererseits können auch Spalten aufgeführt werden, die später nicht statistisch ausgewertet werden sollen. Die Wertetabelle sollte möglichst die kompletten Daten des Gesamtversuchs als ein Tabellenblatt beinhalten. Dies er-

Mo (Parzelle)	Bf1 (Standort)	Bf2 (Block)	Eg1 (Sorte)	Eg2 (Duenger)	Kv (Bodenfeuchte)	Mg1 (Ertrag)	Mg2 (Staerkegehalt)
01	Hsch	B1	Weizen	org	4.7	20.3	69.2
02	Hsch	B2	Weizen	org	3.9	23.5	65.3
03	Hsch	B3	Weizen	org	...	...	...
04	Hsch	B1	Weizen	min	3.6	21.5	61.9
05	Hsch	B2	Weizen	min	3.5	20.9	93.3
06	Hsch	B3	Weizen	min	...	...	...
07	Hsch	B1	Gerste	org	4.1	19.3	55.4
08	Hsch	B2	Gerste	org	4.9	18.6	57.3
09	Hsch	B3	Gerste	org	...	...	...
10	Hsch	B1	Gerste	min	3.6	20.5	62.1
11	Hsch	B2	Gerste	min	4.3	17.9	61.1
12	Hsch	B3	Gerste	min	...	...	...
13	Lh	B1	Weizen	org	5.7	26.6	70.3
14	Lh	B2	Weizen	org	4.3	28.7	75.7
15	Lh	B3	Weizen	org	...	...	...
16	Lh	B1	Weizen	min	4.5		
17	Lh	B2	Weizen	min	4.5	29.8	72.1
18	Lh	B3	Weizen	min	...	...	...
19	Lh	B1	Gerste	org	5.1	29.3	59.8
20	Lh	B2	Gerste	org	4.9	28.6	62.1
21	Lh	B3	Gerste	org	...	...	...
22	Lh	B1	Gerste	min	3.9	27.4	58.7
23	Lh	B2	Gerste	min	4.6	27.9	60.1
24	Lh	B3	Gerste	min	...	...	...
...	...	...	...	...	...	...	...

Tabelle 1: Beispiel eines Flat-File-Formats; Mo=Messobjekt, Bf=Blockfaktor, Eg=Einflussgröße, Kv=Kovariable, Mg=Messgröße

laubt eine ganzheitliche Übersicht und gibt zudem erste Aufschlüsse darüber, ob bzw. welche Versuchsfragen anhand der Daten beantwortet werden können (und welche nicht).

5 Testvoraussetzungen

Jedes statistische Verfahren hat Voraussetzungen bzw. Annahmen an die Daten, welche erfüllt sein müssen. D.h., jedes Testergebnis gilt nur konditioniert auf diese Annahmen. Sind bestimmte Voraussetzungen nicht erfüllt, bzw. sind bestimmte Annahmen nicht gerechtfertigt, ist das Testergebnis fragwürdig, denn es kann zu Effektüber- oder -unterschätzungen kommen. Beispielsweise

sind Zähldaten bekanntermaßen nicht normalverteilt. Die Auswertung solcher Daten mittels eines t-Tests unterstellt den Daten eine nicht vorhandene Normalverteilung.

Prinzipiell resultiert daraus ein praktisches Dilemma. Ein Testverfahren passt umso besser zu gegebenen Daten, je mehr Informationen über die Daten verwertet werden können. D.h., je mehr Annahmen/Voraussetzungen ein Test hat, desto besser seine Güte. Jedoch nur, falls diese auch erfüllt sind. In praxi fällt es eher schwer, Daten strenge Eigenschaften zuzuweisen bzw. die Erfüllung gewisser Voraussetzungen zu unterstellen. Im Zweifelsfall ist es daher angebrachter, robustere Testverfahren mit weniger Annahmen zu verwenden. Diese liefern sicherere Testergebnisse, haben aber i.d.R. eine geringere Güte.

Hieraus ergibt sich die Notwendigkeit, die Testvoraussetzungen an die Daten zu überprüfen. An dieser Stelle sei explizit hervorgehoben, dass es entgegen vielen Lehrbüchern und veralteten Lehrmeinungen absolut nicht notwendig ist, Testvoraussetzungen durch Vortests für die Daten zu prüfen (Easterling and Anderson, 1978; Moser and Stevens, 1992; Schucany and Ng, 2006; Rasch et al., 2011; Rochon et al., 2012; Kozak and Piepho, 2018)! Typische Vortests sind: Kolmogorow-Smirnow-Test und χ^2-Test (Normalverteilung), F-Test und Levene-Test (Varianzhomogenität). Diese sind umgekehrt auch nicht „verboten“, sondern genau wie andere Vorgehensweisen als Indiz zu betrachten. Jedoch darf die Wahl des (Haupt-) Testverfahrens nicht strikt von einem Vortest abhängen. Beispielsweise könnte man sich abhängig vom Ergebnis eines Varianztests entweder für einen t-Test oder einen Welch-t-Test entscheiden. Dies führt aber dann u.U. zu Verletzungen des Fehlers 1. Art (α) im anschließenden Haupttest. Ein Grund dafür ist, dass die Testrichtung aller Vortests nicht korrekt ist. Kein statistischer Test, also auch kein besagter Vortest, kann die zugrunde liegende Nullhypothese zeigen, lediglich die Alternativhypothese durch Ablehnen der Nullhypothese. Kolmogorow-Smirnow-Test und χ^2-Test können also lediglich zeigen, dass die Daten signifikant nichtnormalverteilt sind, F-Test und Levene-Test, dass die Daten signifikant varianzheterogen sind. Dies ist aber nicht das Ziel. Gezeigt werden soll Normalverteilung und Varianzhomogenität. Genau das sind aber die zugrunde liegenden Nullhypothesen. Liefert nun beispielsweise ein F-Test zu $\alpha = 0.05$ einen p-Wert von 0.07, sind die Daten also nicht signifikant varianzheterogen, können eben aber auch nicht als signifikant varianzhomogen betrachtet werden. “*Absence of evidence is not evidence of absence*” (Altman and Bland, 1995). Ein solcher Testausgang ließe streng genommen einfach keine klare Aussage zu.

Eine korrekte Datenüberprüfung auf Testvoraussetzungen erfordert daher ein diffizileres Vorgehen. Am Anfang haben stets **sachlogische Überlegungen** zu stehen. Gibt es triftige Gründe, warum die Daten nicht normalverteilt oder varianzhomogen sein sollten? Zähldaten sind bekanntermaßen nicht normalverteilt, weil diskret verteilt. Dosisfindungsstudien erzeugen oft varianzheterogene Daten, weil oft (nicht immer) mit steigenden Messwerten auch eine steigende Streuung einhergeht. Dies gilt unabhängig von Ergebnissen etwaiger Vortests. Gelegentlich können auch aus **Vorversuchen** Verteilungsannahmen abgeleitet werden. Des weiteren ist es ratsam, sich die vorliegenden Daten stets „anzuschauen", sprich **graphisch zu analysieren**. Barplots sind hierfür absolut ungeeignet, da sie lediglich Mittelwert und Standardfehler darstellen. Die Angabe dieser beiden Kenngrößen ist aber nur dann sinnvoll, wenn man von symmetrisch verteilten Daten ausgehen kann. Bei einfacheren Testproblemen sind Boxplots ein besseres Mittel der Wahl. Sie zeigen in gewissem Sinne die kompletten Daten in Form wichtiger Kenngrößen wie Minimum, 1. Quartil, Median, 3. Quartil, Maximum sowie Ausreißer. Dies erlaubt im Gegensatz zu Barplots eine Einschätzung der Verteilung der Daten. Die beste Alternative stellt die Residuenanalyse dar. Residuen sind, grob gesagt, die Messwerte bereinigt um sämtliche Effekte der Einflussgrößen. Sprich, Residuen sind Messwerte abzüglich der laut statistischem Modell berechneten Werte. Residuenplots stellen in unterschiedlicher Weise die Residuen der Daten bzw. des Testproblems dar. Anhand solcher Plots kann man nicht nur beurteilen, ob die Residuen – und damit die Daten – den Testvoraussetzungen genügen, sondern auch, ob das verwendete statistische Modell korrekt ist. Die Überprüfung der Testvoraussetzungen stützt sich also gewissermaßen auf mehrere Indizien, nicht auf „harte Beweise". In diesem Sinne – als Indiz also – sind natürlich auch Vortests erlaubt, aber nicht in dem strengen Maß, wie bisher in der Literatur beschrieben.

Darüber hinaus gibt es die Möglichkeit der Datentransformation. Die Idee dabei ist, die Daten geschickt umzuwandeln und statt der Originaldaten die transformierten Daten statistisch auszuwerten. Z.B. könnte man varianzheterogene Messwerte gegebenenfalls log-transformieren, um homogene Varianzen (der logarithmierten Werte) zu erzwingen. Diese Vorgehensweise wird leider zu oft verwendet, ohne sich über deren Konsequenzen im Klaren zu sein. Jede Datentransformation verzerrt die Daten. Dies ist gerade der Sinn und Zweck. Mit der Verzerrung der Daten gehen aber auch Verzerrungen der Mittelwertsdifferenzen einher. Signifikante Mittelwertsdifferenzen der transformierten Daten bedeuten nicht zwangsläufig auch signifikante Mittelwertsdifferenzen der originalen Daten (und umgekehrt). Auch statistische Modelle werden durch Datentransformation beeinflusst. Beispielsweise unterstellt ein

„normales“ statistisches Modell der originalen Messwerte, dass sämtliche Effekte der Einflussgrößen additiv wirken. Das selbe additive Modell, angewendet auf die log-transformierten Werte, impliziert aber multiplikative Effekte auf den originalen Werten. Die beste Vorgehensweise ist es daher immer, wenn möglich nicht die Daten dem Verfahren, sondern das Verfahren den Daten anzupassen.

6 Software

Excel ist keine Statistiksoftware, auch wenn sich damit einige statistische Anwendungen, und zudem noch ziemlich simpel, ausführen lassen. Generell gilt: Je „entgegenkommender“ eine Statistiksoftware scheint, desto mehr Fehler können vom Anwender gemacht werden. Gerade „Klickersoftware“ macht es dem Anwender zu leicht, mit den Daten herumzuspielen und verschiedene Testverfahren so lang auszuprobieren, bis ein gewünschter Output erscheint. Besonders der statistische Laie ist meist versucht, Entscheidungsverantwortung von sich zu weisen und agiert oft nach dem Motto: „Wenn eine Software ein bestimmtes Testverfahren auf gegebene Daten anwenden kann, dann wird dies schon seine Richtigkeit haben“. Keine Software überprüft, ob die Voraussetzungen des beabsichtigten Testverfahrens auch erfüllt sind. Und keine Software kann entscheiden, welches Testverfahren anzuwenden ist. Dies hängt ab von den Versuchsfragen, dem Versuchsaufbau und den Eigenschaften der Daten. Die Entscheidungsverantwortung bleibt beim Anwender, egal, was die Software „erlaubt“!

Unabhängig von Voraussetzungen verleitet manche Software dazu, veraltete oder unpassende Methoden anzuwenden. Drei Beispiele seien im folgenden gegeben: i) Nur weil, eine Software möglicherweise einen Fisher's LSD-Test anbietet, bedeutet das nicht, dass dieser auch angebracht ist. Bei eben diesem z.B. ist seit langer Zeit bekannt, dass er den Gesamtfehler 1. Art (α) nicht im strengen Sinn einhält. ii) Barplots sind in mehrerlei Hinsicht ungeeignet. Ein Grund, aus dem Barplots verwendet werden, ist schlichtweg der, dass sie simpel sind und Excel sie anbietet. iii) Der Tukey-Test hat sich in vielen Bereichen zu einer Art Standardtest entwickelt. Viel zu selten werden hierbei dessen Sinnhaftigkeit hinterfragt und Alternativen verwendet. Auch hier ist oftmals lediglich die Methodenbeschränktheit der Sofware der Anwendungsgrund. Statistische Verfahren oder Methoden sollten also nicht angewendet werden, weil die Software es ermöglicht, sondern nach Maßgabe dessen, was sinnvoll ist. Bietet eine Software ein bestimmtes Testverfahren oder eine Methode nicht an, könnte es angebracht sein, eine andere Software

zu verwenden.

In gewissem Sinne, d.h. abhängig von der Situation und abgesehen von Ausnahmen, sind (meiner Meinung nach) SAS und R die beiden führenden Statistikprogramme. Besonders vor dem Hintergrund ihrer universellen Anwendbarkeit; beide besitzen ein sehr großes Anwendungsspektrum. Für spezielle Testprobleme kann es aber auch durchaus sinnvoll sein, auf andere Software auszuweichen. Die Vorteile von SAS liegen eindeutig bei den gemischten Modellen, die Vorteile von R bei multiplen Vergleichen. Viele Anwender, besonders aus der Industrie, scheuen sich vor der Benutzung von R, da es nicht validiert ist, sondern eine Open-Source-Software. Dies ist gleichzeitig auch ein Vorteil, da R dadurch moderner ist in dem Sinne, dass neueste statistische Verfahren hier eher Einzug halten. Ganz abgesehen davon, dass es nichts kostet.

7 Darstellung der Testergebnisse

Wie die Daten selbst, lassen sich auch deren Testergebnisse sowohl in Form von Zahlen als auch graphisch darstellen. Graphiken haben dabei den Vorteil der Kompaktheit und Eingänglichkeit und werden deshalb oft bevorzugt. Allerdings verleiten sie auch zu mehreren Fehlern im Rahmen der statistischen Auswertung und deren Interpretation.

Meist aus Platzmangel ist der Versuchsansteller dazu geneigt, sowohl die Daten als auch die Testergebnisse in eine Graphik zu integrieren. Vor diesem Hintergrund ist die Buchstabenvergabe leider ein weit verbreitetes Mittel der Signifikanzdarstellung. Die wenigsten Statistiksoftwares stellen Signifikanzen anhand von Buchstaben dar bzw. zumindest nicht automatisch. Dies hat folgende Gründe: Vor dem Hintergrund von Mittelwertsvergleichen beziehen sich Signifikanzen nicht auf Mittelwerte, sondern (i.d.R.) auf Mittelwertsdifferenzen. Streng genommen werden nämlich nicht Mittelwerte getestet, sondern eben deren Differenzen. Am sinnvollsten und eindeutigsten ist es daher, nicht die Mittelwerte, sondern die Vergleiche/Mittelwertsdifferenzen mit Kennzeichnungen (Buchstaben, Sternchen o.ä.) zu versehen. Wenn überhaupt, denn selbst dies ist unnötig, da ein zugehöriger p-Wert viel mehr Aussagekraft besitzt und zudem nicht an ein festes Signifikanzniveau α gebunden ist. Z.B. deutet ein Mittelwertsvergleich mit p-Wert 0.07 offensichtlich schon auf einen gewissen Effekt hin, der bei $\alpha = 0.05$ zwar nicht signifikant ist, bei $\alpha = 0.10$ aber schon und eventuell als „Trend“ oder „Tendenz“ interpretiert werden könnte. D.h., die Buchstabenvergabe unterschlägt Informationen, indem sie Testergebnisse gewissermaßen auf Ja-Nein-Aussagen reduziert. Darüber hin-

aus implizieren gleiche Buchstaben fälschlicherweise gleiche Mittelwerte. Für die Aussagen über Gleichheit sind aber sog. Äquivalenztests nötig. Signifikanztests können lediglich signifikante Unterschiede aufzeigen. „*Absence of evidence is not evidence of absence*“ (Altman and Bland, 1995), siehe 5.

Die meiste Aussagekraft besitzen jedoch Konfidenzintervalle, genauer gesagt: (simultane) Konfidenzintervalle der Mittelwertsdifferenzen. Diese lassen sich gleichzeitig für Testentscheidungen und für Parameterschätzungen verwenden und sie erlauben die Darstellung sowohl in Form von Zahlen als auch als Graphik. Auch für Mittelwerte selbst lassen sich Konfidenzintervalle angeben. Dies ist aber i.d.R. wenig zielführend. Manche Anwender benutzen Konfidenzintervalle für Mittelwerte dafür, Signifikanzentscheidungen über Mittelwertsdifferenzen zu gewinnen nach der Regel: „Wenn sich die Intervalle der Mittelwerte überschneiden, ist die Mittelwertsdifferenz nicht signifikant“. Analoge Regeln existieren auch für Barplots oder Boxplots. Sie sind schlichtweg Unsinn, auch wenn sie in Ausnahmefällen zu korrekten Ergebnissen führen können. Für eine Signifikanzaussage über Differenzen benötigt man Konfidenzintervalle für Differenzen!

Außerdem, und wiederum meist aus Platzmangel, ist der Versuchsansteller dazu geneigt, die Ergebnisse sämtlicher Versuchsfragen in einer Graphik zu vereinen. Dies verleitet dazu, auch alle Versuchsfragen miteinander zu vermischen (siehe 2). Es ist also geraten, falls Graphiken verwendet werden, diese zu trennen nach Daten und Testergebnissen einerseits und nach den verschiedenen Versuchsfragen andererseits. Dies ist platzintensiver, sorgt aber für mehr Übersicht und Klarheit.

8 Zusammenfassung

- Die Versuchsfragen sind vom Versuchsansteller vor dem Versuch sauber zu definieren und voneinander zu trennen.
- Der Gesamtversuch muss beschrieben werden, also der vollstandige Versuchsaufbau. Teilfragestellungen können gegebenenfalls später daraus abgeleitet werden.
- Blöcke sind formell keine Wiederholungen.
- Die Daten sollten am besten im Flat-File-Format vorliegen.
- Barplots als Darstellung der Daten sind nicht zwingend notwendig. Boxplots besitzen mehr Aussagekraft.

- Vortests zur Überprüfung von Testvoraussetzungen sind nicht zwingend notwendig. Residuenanalysen sind die bessere Alternative.
- Wenn möglich, ist Datentransformation zu vermeiden und stattdessen ein zu den Daten passendes Verfahren oder statistisches Modell zu wählen.
- Die zu verwendende Software hängt ab von den benötigten Testverfahren und Methoden, nicht umgekehrt.
- Buchstaben als Darstellung der Testergebnisse sind nicht zwingend notwendig. Die meiste Aussagekraft besitzen Konfidenzintervalle, falls existent.
- Graphiken sollten möglichst nach Daten und Testergebnissen einerseits und nach den verschiedenen Versuchsfragen andererseits getrennt werden.

Literatur

D. G. Altman and J. M. Bland. Statistics notes - Absence of evidence is not evidence of absence. *British Medical Journal*, 311(7003):485–485, 1995.

R. G. Easterling and H. E. Anderson. The effect of preliminary normality goodness of fit tests on subsequent inference. *Journal of Statistical Computation and Simulation*, 8(1):1–11, 1978.

M. Kozak and H. P. Piepho. What's normal anyway? residual plots are more telling than significance tests when checking anova assumptions. *Journal of Agronomy and Crop Science*, 204(1):86–98, 2018. doi: 10.1111/jac.12220.

B. K. Moser and G. R. Stevens. Homogeneity of variance in the 2-sample means test. *American Statistician*, 46(1):19–21, 1992.

D. Rasch, K. D. Kubinger, and K. Moder. The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers*, 52(1):219–231, 2011.

J. Rochon, M. Gondan, and M. Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12:81, 2012.

W. R. Schucany and H. K. T. Ng. Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. *Communications in Statistics-theory and Methods*, 35(12):2275–2286, 2006.