Errors in the Statistical Analysis of
Polarotactic tabanids find striped patterns with brightness and/or polarization modulation least attractive: an advantage of zebra stripes.

C. J. Schwarz
Department of Statistics and Actuarial Science, Simon Fraser University
cschwarz@stat.sfu.ca
December 27, 2013

Contents

1 Introduction 2
2 Experiment 1 3
3 Experiment 2 8
4 Experiment 3 9
5 Experiment 4 13
6 Experiment 5 13
7 Experiment 6 16
8 Summary 18
9 Appendix: Why are the authors’ \( \chi^2 \) tests valid but their ANOVA tests not valid? 19
Abstract

Egri et al. (2012) investigated how the attractiveness of substrates to tabanids (horseflies) varied when painted with different numbers, orientation, and density of stripes. The basic experimental design was a randomized complete block design where the number of flies attracted was counted at regular intervals. The authors analyzed this data using a complete-randomized design ANOVA which is inappropriate as their analyses method did not match the design of the studies. The authors also used a simple $\chi^2$ test which, fortunately for the authors, is appropriate under a conditioning argument. However, their $\chi^2$ analysis cannot readily provide estimates of the effect sizes for the various factors. The appropriate analyses for these experiments is a generalized linear model assuming a Poisson distribution for the number of counts. We demonstrate the proper analyses of these experiments.

1 Introduction

Egri et al. (2012) investigated the influence of density and types of stripes on the attractiveness of surfaces to horseflies (tabanids) as a model of why zebra have stripes. This experiment attracted much media attention such as reported in the Economist.

The experimental protocol is explained in their paper. Briefly, plates or models with different striping patterns were exposed to tabanids (horseflies) outdoors and the number of flies attracted to the various patterns was compared. The counts were measured at regular intervals (e.g. weekly) and a separate count was obtained each week. After the count for the interval was determined, the plates or model positions were cleaned and re-randomized.

The authors performed two types of analyses.

In one type of analysis, they compared the mean number of flies trapped in each intervals using a single-factor completely randomized design (CRD) ANOVA. These analyses are incorrect for several reasons. The data were NOT collected using a CRD design; rather the design is a variant of randomized complete block (RCB) design where the sampling intervals are the blocks. Consequently the authors should have used a variant of a RCB ANOVA. Second, the number of flies trapped in each interval on each pattern are typically smallish counts and would more closely be described by a Poisson distribution rather than a normal distribution. Third, it is quite clear that variance of the responses are not equal for all treatment groups and appears to be proportional to the mean response which violates a key assumption of standard normal-distribution ANOVA that the variance is the same in all treatment groups. Lastly, the standard ANOVA on the counts assumes that the treatment effects are constant DIFFERENCES, i.e. the difference in mean counts between two treatments is the same regardless of the underlying means. This additive effect on the counts is unlikely to be true – it is more likely from a biological perspective that the mean counts differ by a MULTIPLICATIVE constant, e.g. treatment A attracts twice as many flies as treatment B regardless of the underlying mean counts. This also appears to be supported by the data at hand.

---


In the second type of analysis, the authors often compared the total counts (over the multiple sampling intervals) using a Pearson $\chi^2$ test of the hypothesis of equal proportions. Under the null hypothesis of no difference in attractiveness, it can be shown (see Appendix A) that by conditioning on the totals in each sampling interval, these $\chi^2$ tests are appropriate. However, this method cannot be readily used to derive the estimated multiplicative effect and its standard error because of this conditioning process. Furthermore, this method cannot be readily extended for the more complex experimental design found in two of the experiments in the paper.

For these reasons, the proper analysis should use a generalized linear model (GLIMs) using a Poisson-distribution with a logarithmic link function.

In the remainder of this paper, I will illustrate the proper analyses of this data. Complete R code is available at: [http://www.stat.sfu.ca/~cschwarz/CourseNotes/Reanalysis](http://www.stat.sfu.ca/~cschwarz/CourseNotes/Reanalysis). The authors are fortunate that the correct analysis does not lead to substantial differences in their conclusions – this is more of matter of good luck vs. good management.

I have ignored completely the issue of pseudo-replication (Hurlbert, 1984) in these experiments. In all of the experiments, a single replicate of the same substrate is used through out the experiment rather than separate experimental units constructed for each sampling interval. One could argue that the results of this paper are only applicable to those particular sets of substrates used in the experiment and not generalizable to the different patterns in general. For example, perhaps there was something peculiar to a particular substrate used in the experiment that, while seeming identical to the other substrate, is particularly appealing to tabanids.

2 Experiment 1

Experiment 1 investigated the influence of the number of white orthogonal stripes on the attractiveness of the surfaces to tabanids. Three plastic trays were painted with differing number of stripes (0, 2, or 6). They were covered with vegetable oil and placed on the ground. Approximately every two days, the number of flies in each tray were counted, the trays cleaned, and the position of the trays re-randomized. The raw data is available in Table S1 of the authors paper. It is presented in my Table 1 and a profile plot is shown in my Figure 1. The authors conclude:

"There was a significant difference in the number of trapped tabanids between the three test surfaces (one-way ANOVA test: $F_{2;24} = 19.95; P << 0.001$)" (Legend to Table-S1)

"The white framed black tray without orthogonal white stripes (supplementary material Fig.S1A) captured a significantly higher number ($m = 247$) of tabanids than the white-framed black trays with 2 (Fig. 3A; supplementary material Fig.S1B) and 6 (supplementary material Fig.S1C) orthogonal white stripes ($m = 185$ and 33, respectively) ($\chi^2$ test: $\chi^2 = 108.60$, d.f.=2, $P << 0.001$)" (Results)

3 This was reproduced in my analyses
4 I was unable to reproduce this result in my analyses.
Table 1: Raw data from Experiment 1.

<table>
<thead>
<tr>
<th>Week</th>
<th>stripe0</th>
<th>stripe2</th>
<th>stripe6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26-27 August</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>28-29 August</td>
<td>34</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>30-31 August</td>
<td>32</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>1-2 September</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>3-4 September</td>
<td>31</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>5-6 September</td>
<td>28</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>7-8 September</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>9-10 September</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>11-12 September</td>
<td>16</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 1: Profile plot of the raw data from Experiment 1.
This first experiment illustrates quite nicely why the authors’ analyses are not appropriate. First, the experimental design is NOT a completely randomized design (CRD). In a CRD, it would be possible that two replicates of same treatment could occur in a particular sampling interval. However, within each sampling interval in this design, all three traps were used, i.e. each sampling interval is a block. Second, notice that the total flies captured in each interval is generally declining over time. Consequently, the difference in the number of flies captured between $N = 2$ and $N = 6$ stripes is not a fixed arithmetic amount, but also declines as the total number of flies declines. However, the ratio of the number of flies attracted to 2 strips is generally about $5 \times$ larger than with 6 stripes, suggesting a multiplicative effect. The variability in counts with $N = 0$ stripes is also substantially larger than the variability with $N = 6$ stripes.

The proper analysis of such an experiment is a generalized linear model (GLIM) using a Poisson distribution to describe the response variable with a logarithmic link function. Using a standard statistical shorthand:

$Y_{ij} \sim \text{Poisson}(\mu_{ij})$

$\theta_{ij} = \log(\mu_{ij})$

$\theta_{ij} = \text{Date} + \text{Stripes}$

where $Y_{ij}$ is the number of tabanids counted in interval $i$ and treatment $j$ ($j = 0, 2,$ or $6$ stripes); $\mu_{ij}$ is the mean number of tabanids that would be captured; and the terms Date and Stripes refer to the effects of the different dates and number of stripes. The Date term captures the change in the the number of flies per interval while the term Stripes captures the effect of the different number of stripes. The last line indicates a model where the $\log(\mu)$ varies additively across dates and number of stripes, i.e. a standard randomized block model. Because the means are transformed to the logarithmic scale, the effect of Dates and Stripes acts multiplicatively, e.g. the mean count under $N = 6$ stripes will be modeled as a fixed fraction of the mean count under $N = 2$ stripes regardless of date.

The model is easily fit using R:

```r
result1.glm <- glm(tabanids ~ Week + Stripes,
                   family=poisson(link='log'), data=fly1)
anova(result1.glm, test="Chi")
```

giving

```
Analysis of Deviance Table
Model: poisson, link: log
Response: tabanids
```

\footnote{Given that no count are zero, one could get an approximate similar analysis using a normal-theory ANOVA on the logarithm of the counts. See R code for details.}

©2013 Carl James Schwarz
2 EXPERIMENT 1

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>26</td>
<td>274.365</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td>8</td>
<td>63.922</td>
<td>18</td>
<td>210.443</td>
</tr>
<tr>
<td>Stripes</td>
<td>2</td>
<td>193.554</td>
<td>16</td>
<td>16.889</td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

There is strong evidence that the mean counts are not equal across the three treatments \( (p < .0001) \)

Furthermore, a Tukey-adjusted multiple comparison can be constructed to drill down to investigate where the means could differ among the pairs of treatments. Note that because the model operates at the logarithmic level, the estimates of the pairwise differences are the differences on the logarithmic scale and a back transformation yields estimates of the RATIO of the mean counts on the anti-logarithmic scale.

```r
library(lsmeans)
result1.glm.lsmeans <- lsmeans(result1.glm, cld ~ Stripes, adjust='tukey')
result1.glm.lsmeans[[2]]$ratio <- exp(result1.glm.lsmeans[[2]]$estimate)
result1.glm.lsmeans[[2]]$se.ratio <- result1.glm.lsmeans[[2]]$ratio * result1.glm.lsmeans[[2]]$SE
```

giving

<table>
<thead>
<tr>
<th>estimate</th>
<th>SE</th>
<th>df</th>
<th>z.ratio</th>
<th>p.value</th>
<th>ratio</th>
<th>se.ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 stripes - 2 stripes</td>
<td>0.2890325</td>
<td>0.09723161</td>
<td>NA</td>
<td>2.97262</td>
<td>0.00832</td>
<td>1.335135</td>
</tr>
<tr>
<td>0 stripes - 6 stripes</td>
<td>2.0128808</td>
<td>0.18534184</td>
<td>NA</td>
<td>10.86037</td>
<td>0.00000</td>
<td>7.484848</td>
</tr>
<tr>
<td>2 stripes - 6 stripes</td>
<td>1.7238483</td>
<td>0.18896671</td>
<td>NA</td>
<td>9.12250</td>
<td>0.00000</td>
<td>5.606061</td>
</tr>
</tbody>
</table>

p values are adjusted using the tukey method for 3 means

Because the (Tukey-adjusted) \( p \)-values are all small, this indicates there is evidence that the mean differs among all pairs of treatments. Furthermore, the estimated ratio of mean counts at \( N = 0 \) to \( N = 2 \) stripes is 1.33 (SE 0.12); of \( N = 0 \) to \( N = 6 \) stripes is 7.48 (SE 1.39); and \( N = 2 \) to \( N = 6 \) stripes is 5.61 (SE 1.06). The can be interpreted as implying the \( N = 0 \) stripe pattern is about 1.33 \( \times \) more attractive to tabanids than the \( N = 2 \) striped pattern. This information was not available from the analysis by the authors.

The above analysis did not take into account the different area of the black stripes. The Poisson generalized linear model is also easily adjusted for the different area of the stripes while the chi-square analysis

---

6 The Analysis of Deviance Table is analogous to the ANOVA table in normal-theory ANOVA. The results from this and other models fit in this paper were also examined for over-dispersion and lack-of-fit. No evidence of lack-of-fit was found for any of the GLIM models.

7 The SE of the anti-logarithmic transformation is obtained by an application of the delta method.
performed by the authors cannot be so easily adjusted. The mean number of flies trapped by each type of tray is the product of the area $A_s$ and the underlying density per unit area of attracted flies $d_s$, i.e. $\mu_s = A_s \times d_s$ or $\log(\mu_s) = \log(A_s) + \log(d_s)$. Consequently, the adjustment for the area of the black stripes also enters on the logarithmic scale as an offset:

```r
glm(tabanids ~ Week + Stripes, family=poisson(link='log'),
    offset=log(fly1$area.black), data=fly1)
anova(result1.glm.ad, test="Chi")
```

result1.glm.ad <- glm(tabanids ~ Week + Stripes, family=poisson(link='log'),
    offset=log(fly1$area.black), data=fly1)
anova(result1.glm.ad, test="Chi")

```

Analysis of Deviance Table

Model: poisson, link: log
Response: tabanids
Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>26</td>
<td>217.707</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td>8</td>
<td>63.922</td>
<td>18</td>
<td>153.785</td>
</tr>
<tr>
<td>Stripes</td>
<td>2</td>
<td>136.895</td>
<td>16</td>
<td>16.889</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is strong evidence that the mean counts are not equal across the three treatments even after adjusting for the area of the stripes ($p < 0.0001$).

A Tukey-adjusted multiple comparison can again be constructed as shown earlier:

```
estimate       SE  df   z.ratio  p.value ratio  se.ratio
0 stripes - 2 stripes 0.200129 0.09723161 NA  2.05827 0.09873 1.221560 0.1187743
0 stripes - 6 stripes 1.733357 0.18534184 NA  9.35222 0.00000 5.659621 1.0489645
2 stripes - 6 stripes 1.533228 0.18896671 NA  8.11375 0.00000 4.633108 0.8755032
```

p values are adjusted using the tukey method for 3 means

However, now after adjusting for the area of black, there is no evidence that the attractiveness to flies differs between the $N = 0$ and $N = 2$ stripes treatment ($p = .097$; estimated ratio is $1.22$ (SE .12)). There still is strong evidence that the attractiveness to flies differs in the other pairwise comparisons.
In summary, the ANOVA analysis reported in Table S1 is inappropriate; I was unable to reproduce the results of the $\chi^2$ test; and the authors missed the followup comparisons of no evidence in the attractiveness to flies using $N = 0$ and $N = 2$ stripes after adjusting for the area of the black stripes.

3 Experiment 2

In experiment 2, the effect of stripe width on the attractiveness to tabanids was investigated. There were five treatments (Solid White, 12, 6, 2 stripes and Solid Black patterns). The experimental protocol was similar to that of experiment 1, i.e. counts were made at approximately weekly intervals and the order of the traps was rerandomized each week. The data is presented in Table S2 and has a similar structure to that of Experiment 1.

The authors concluded:

“There was a significant difference in the number of tabanids trapped between the five test surfaces (one-way ANOVA: $F_{4,60} = 5.93; P << 0.001$)” (Legend to Table S3)

“In experiment 2, the white and black trays filled with salad oil were, respectively, the least and the most attractive to tabanid flies, with 3 versus 145 trapped individuals (supplementary material Tables S3, S4). The smaller the number $N$ of black/white stripes, the more tabanids were trapped; the trays with $N = 12, 6$ and 2 stripes captured 24, 66 and 138 tabanids, respectively ($\chi^2$ test: $\chi^2 = 88.62, \text{d.f.} = 4, P << 0.001$; supplementary material Tables S3, S4)$.

(Results)

Once again the use of a single-factor CRD ANOVA is not appropriate because of the reasons listed earlier. Furthermore, inspection of the table shows some weeks (e.g. 6-10 July) with 0 flies captured under any treatment condition. These weeks are non-informative about differences in capture efficiency and should be deleted prior to analysis (which the authors did not do).

A generalized linear model can once again be fit to this data (after removing weeks with 0 flies captured in all traps) in much the same way as in Experiment 1: giving

<table>
<thead>
<tr>
<th>Analysis of Deviance Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model: poisson, link: log</td>
</tr>
<tr>
<td>Response: tabanids</td>
</tr>
<tr>
<td>Terms added sequentially (first to last)</td>
</tr>
</tbody>
</table>

8Verified
9I was unable to reproduce this result in my analyses.
Again, very strong evidence that the attractiveness of the treatments to flies is not equal.

As before, a Tukey-like multiple comparison can be done and can be summarized using the traditional joined-letter displays:\textsuperscript{10}

\begin{verbatim}
.group Stripes
1 A White
2 B 12 stripes
3 C 6 stripes
4 D 2 stripes
5 D Black
\end{verbatim}

This indicates that the attractiveness could NOT be distinguished between the $N = 2$ stripes and Black treatments, but there was evidence of differences among all other treatment means.

In summary, the ANOVA analysis reported in Table S3 is inappropriate; I was unable to reproduce the results of the chi-square test; and the authors missed the followup comparisons of no evidence in the mean number of flies attracted using $N = 2$ stripes the and Black treatment.

4 Experiment 3

While Experiments 1 and 2 used the total number of flies attracted to each treatment substrate in each sampling interval, it could not distinguish if the flies were attracted to the white or black portion of pattern. In Experiment 3, glue was placed on the substrate with $N = 2$, $N = 4$, and $N = 8$ stripes patterns, and the number of flies stuck on the white or black portion of the pattern was counted. The area of white or black was equal and identical for all numbers of stripes. The raw data is presented in Table S5 and reproduced in Table 2.

The authors’ concluded:

\textsuperscript{10}There is no evidence of a difference in attractiveness to flies among treatment groups joined by the same letter. The actual letter is not relevant.
Table 2: Raw data from Experiment 3.

<table>
<thead>
<tr>
<th>Date</th>
<th>Black.8</th>
<th>White.8</th>
<th>Black.4</th>
<th>White.4</th>
<th>Black.2</th>
<th>White.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 17 August</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>2 18 August</td>
<td>1</td>
<td>1</td>
<td>33</td>
<td>12</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td>3 24 August</td>
<td>4</td>
<td>3</td>
<td>78</td>
<td>14</td>
<td>70</td>
<td>11</td>
</tr>
<tr>
<td>4 26 August</td>
<td>1</td>
<td>1</td>
<td>28</td>
<td>16</td>
<td>46</td>
<td>9</td>
</tr>
<tr>
<td>5 28 August</td>
<td>5</td>
<td>0</td>
<td>29</td>
<td>7</td>
<td>29</td>
<td>5</td>
</tr>
<tr>
<td>6 1 September</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>3</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>7 4 September</td>
<td>1</td>
<td>0</td>
<td>26</td>
<td>8</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>8 8 September</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>9 13 September</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>

“In experiment 3, the horizontal sticky 2-striped test surface attracted most tabanids ($N = 321$), the 4-striped surface was less attractive ($N = 290$) and the 8-striped surface was the least attractive ($N = 19$), resulting in significant differences in the number of trapped tabanids between the different groups ($\chi^2$ test: $\chi^2 = 262.87$, d.f.=2, $P << 0.001$; supplementary material Table S5[11])” (Results)

“... Under all conditions, we found significantly higher numbers of tabanids trapped on the black stripes compared with the white stripes ($\chi^2$ test: 2 stripes, $chi^2 = 136.08$, d.f.=1, $P < 0.001$; 4 stripes, $\chi^2 = 88.28$, d.f.=1, $P < 0.001$; 8 stripes, $\chi^2 = 5.0$, d.f.=1, $P < 0.025$[12])” (Results)

“There was a significant difference in the number of trapped tabanids between the three test surfaces (one-way ANOVA: $F_{2;51} = 6.23 P < 0.004$[13])” (Legend of Table S5)

Once again the single-factor CRD ANOVA is not appropriate for this data for similar reasons as in Experiments 1 and 2.

The analysis of this design is more complex because each pattern was measured twice – on the white and black surfaces. This is an example of split-plot design. There are two factors in this experiment – the number of stripes (2, 4 or 8) and the color of the background (black or white). There are two sizes of experimental unit - the entire substrate, and then the portion of the board covered in white or black. The experimental design to investigate the effect of the number of stripes is similar to previous experiment – a Poisson RCB generalized linear model. The experimental design to investigate the effect of color is similar to a paired design within each number of stripes. The split-plot design combines these two designs.

A generalized linear MIXED model (GLIMM) must be used for this combined design because of the

---

11 I was unable to reproduce these results
12 Verified except I was unable to match results for $N = 8$ stripes.
13 Verified
two sizes of experimental units. The model is:

\[
Y_{ijk} \sim \text{Poisson}(\mu_{ijk})
\]

\[
\theta_{ijk} = \log(\mu_{ijk})
\]

\[
\theta_{ijk} = \text{Date}(R) + \text{Stripes} + \text{Dates:Stripes}(R) + \text{Color} + \text{Stripes:Color}
\]

where \(Y_{ijk}\) is the count in interval \(i\), treatment \(j\), and color \(k\); the terms \text{Stripes}, \text{Color} represent the main effect of the two factors; the term \text{Stripes:Color} represents the interaction between the effects; the terms \text{Date} and \text{Dates:Stripes} represent the random effects of the sampling intervals and the experimental unit within each date. This model is analogous to those used in split-plot designs using normal-theory ANOVAs.

These models can be fit in R using:

```r
library(lme4)
result3.glmer <- glmer( tabanids ~ Color + Stripes + Color:Stripes + 
                          (1|Date/Stripes),
                          data=fly3,
                          family=poisson(link=log))
```

The results of the hypothesis tests about the individual main effects and their interaction can be summarized:

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>Chi</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>1</td>
<td>46.652602</td>
<td>8.475435e-12</td>
</tr>
<tr>
<td>Stripes</td>
<td>2</td>
<td>22.788619</td>
<td>1.267735e-10</td>
</tr>
<tr>
<td>Stripes:Color</td>
<td>2</td>
<td>1.410173</td>
<td>2.441011e-01</td>
</tr>
</tbody>
</table>

There is strong evidence of effects of stripes and color on the attractiveness to tabanids, but no evidence of an interaction between the effects of the two factors. This is illustrated in the interaction plot shown in Figure 2.

A pairwise multiple comparison (see R code), showed there is no evidence of a difference in attractiveness between \(N = 2\) and \(N = 4\) stripes, but there is evidence that both patterns are more attractive to flies than the \(N = 8\) stripes pattern. The black background is consistently more attractive than the white background under all numbers of stripes. The estimated ratio of attractiveness is:

<table>
<thead>
<tr>
<th>estimate</th>
<th>SE</th>
<th>df</th>
<th>z.ratio</th>
<th>p.value</th>
<th>ratio</th>
<th>se.ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black - White</td>
<td>1.275243</td>
<td>0.1867046</td>
<td>NA 6.83027</td>
<td>0</td>
<td>3.579572</td>
<td>0.6683228</td>
</tr>
</tbody>
</table>

p values are adjusted using the tukey method for 2 means

i.e. the black area appears to be approximately 3.6 (SE 0.67) times more attractive than the white area.
Figure 2: Estimated interaction plot from Experiment 3 from a split-plot analysis.
6 EXPERIMENT 5

In summary, the ANOVA analysis reported in Table S5 is inappropriate; and I was unable to reproduce some of the results of the chi-square tests. A split-plot analysis showed no evidence of an interaction between the effects color and strip number. It also showed no evidence of a difference in attractiveness between $N = 2$ and $N = 4$ stripes (the authors’ concluded that these differed), and estimated a common color effect (the authors’ estimated different effects for each number of stripes). Their methods could not cope with this more complex design.

5 Experiment 4

In Experiment 4, four different models of zebra with different strip patterns was used to investigate differences in attractiveness to tabanids. The results from this experiment were published by Horvath (2012) and the raw data can be found in this other paper.

This experimental design is similar to Experiments 1 and 2, i.e. a randomized complete block design. Again the authors used a single-factor CRD ANOVA which is inappropriate. Refer to the R code for details on the appropriate analysis. Once again, intervals where the counts are 0 for all models within a sampling are non-informative and should be removed prior to analysis.

6 Experiment 5

Experiment 5 was again a two-factor design investigating the impact of the pattern (5 levels) and color (black or white) on the attractiveness of tabanids. The experimental design is complex because two of the patterns (all white or all grey) did not have a contrast between two colors on the same test surface. The data are presented in Table S6 and summarized in Table 3.

<table>
<thead>
<tr>
<th>Period</th>
<th>Date</th>
<th>White</th>
<th>Grey</th>
<th>Z13_B</th>
<th>Z13_W</th>
<th>Z5_B</th>
<th>Z5_W</th>
<th>Z2_B</th>
<th>Z2_W</th>
<th>Z1_B</th>
<th>Z1_W</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>16-21 July</td>
<td>32</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>3</td>
<td>89</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>22-29 July</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>4</td>
<td>23</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>30 July-5 August</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>30</td>
<td>9</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>6-17 August</td>
<td>30</td>
<td>14</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>92</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>18-26 August</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>23</td>
</tr>
</tbody>
</table>

It is not clear exactly how the authors’ analyzed this experiment, but it appears that they did two separate analyses.

First, was a simple comparison of the effect of pattern. The authors first added the counts from the black
and white colors within each week. Then, based on Table S7, the authors appears to do separate single-factor CRD ANOVA for each pair of patterns using the counts as data. For example, when comparing patterns Z13 vs G, they used the 5 (weeks) × 2 (treatments) = 10 data values and did a two-sample t-test or a chi-square test.

The authors concluded:

“According to the one-way ANOVA test, there were significant differences between Z5/Z13 and Z2, Z2 and Z1 (supplementary material Table S7)\(^{14}\). (Results)

The reduced data (i.e. after summing over the black and white parts of the pattern) again as a similar design as from Experiments 1, and 2 and so a single-factor CRD ANOVA is not appropriate. It is also not appropriate to do separate ANOVAs (or chi-square tests) for all possible pairs of treatments as this does not properly control the overall false positive rate.

As in earlier experiments, a Poisson RCB ANOVA is the appropriate analysis (see R code) and the results can be summarized in the joined-line plots:

\[
\begin{array}{ccc}
\text{group} & \text{Pattern} \\
6 & A & Z5 \\
4 & A & Z13 \\
1 & A & Grey \\
2 & B & White \\
5 & B & Z2 \\
3 & C & Z1 \\
\end{array}
\]

This is a much more compact and appropriate presentation of the results.

The second part of the analysis looked at the relative attractiveness of the black and white portions of the Z* patterns. The authors’ appear to do a series of single-factor CRD ANOVAs on the black and white counts for each individual patterns. This analysis is not appropriate because of the design is more akin to a paired comparison within each pattern (induced by the 5 sampling periods) than a completely randomized design. The analysis is also inefficient because separate analyses are done for each Z* patterns. The authors concluded:

“According to the \(\chi^2\)-test, on surfaces Z2 and Z1, the black stripes trapped significantly more tabanids than the white stripes (supplementary material Table S8)\(^{15}\). Thus, only when the white and black stripes are wide enough is there a difference in the number of tabanids caught by the black and white stripes.”

\(^{14}\)Verified
\(^{15}\)Verified except for results for Z5
Here the authors concluded that failing to detect an effect means no effect was present. However, looking at the raw data, it is clear that failure to detect an effect of color in Z13 and Z5 patterns is simply an artifact of a very small counts.

This portion of the experiment is again better analyzed as a split-plot design as in Experiment 3 (see R code for details). The overall test of the main effects and interactions show strong evidence of main effect, but no evidence of interaction. The interaction plot is shown in Figure 3.

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>Chi</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>1</td>
<td>7.038646</td>
<td>7.976930e-03</td>
</tr>
<tr>
<td>Pattern</td>
<td>2</td>
<td>22.788619</td>
<td>1.267735e-10</td>
</tr>
<tr>
<td>Color:Pattern</td>
<td>3</td>
<td>2.164912</td>
<td>8.987032e-02</td>
</tr>
</tbody>
</table>

The two curves again appear to be parallel and the difference between the curves estimates the relative attractiveness of the two colors:

<table>
<thead>
<tr>
<th>estimate</th>
<th>SE</th>
<th>df</th>
<th>z.ratio</th>
<th>p.value</th>
<th>ratio</th>
<th>se.ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>B - W</td>
<td>0.5562664</td>
<td>0.209671</td>
<td>NA</td>
<td>2.65304</td>
<td>0.00799</td>
<td>1.744148</td>
</tr>
</tbody>
</table>

p values are adjusted using the tukey method for 2 means.

which indicates that black surfaces are about 1.7 \times more attractive than white surfaces.
In summary, the authors again used an inappropriate single-factor CRD ANOVA when the data were collected under a randomized block design and/or a split-plot design. They also made the error that not detecting an effect is not the same as no effect existing – in this case, failure to detect an effect is an artifact of a low sample size.

7 Experiment 6

The authors investigated the effect of two factors, pattern and mounting, on the attractiveness to tabanids. The experiment had 3 patterns (labelled Z9+, Z17+, and Z17-). Each pattern was placed either horizontally or vertically. The experiment was repeated over several sampling intervals as was done in the previous experiments.

The raw data is available in Table S9 or Table 4.

<table>
<thead>
<tr>
<th>Date</th>
<th>H_09p</th>
<th>H_17p</th>
<th>H_17m</th>
<th>V_09p</th>
<th>V_17p</th>
<th>V_17m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 27 June</td>
<td>16</td>
<td>6</td>
<td>20</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>2 30 June</td>
<td>7</td>
<td>3</td>
<td>15</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3 4 July</td>
<td>6</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>4 8 July</td>
<td>6</td>
<td>1</td>
<td>21</td>
<td>3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>5 10 July</td>
<td>10</td>
<td>3</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>6 12 July</td>
<td>9</td>
<td>4</td>
<td>17</td>
<td>6</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>7 14 July</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8 16 July</td>
<td>8</td>
<td>5</td>
<td>22</td>
<td>5</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>9 18 July</td>
<td>10</td>
<td>6</td>
<td>25</td>
<td>7</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>10 20 July</td>
<td>14</td>
<td>8</td>
<td>23</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>11 28 July</td>
<td>34</td>
<td>24</td>
<td>45</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>12 5 August</td>
<td>18</td>
<td>14</td>
<td>25</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>13 7 August</td>
<td>16</td>
<td>9</td>
<td>22</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>14 9 August</td>
<td>13</td>
<td>8</td>
<td>19</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>15 11 August</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>16 13 August</td>
<td>8</td>
<td>6</td>
<td>14</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>17 15 August</td>
<td>20</td>
<td>10</td>
<td>33</td>
<td>10</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>18 17 August</td>
<td>26</td>
<td>16</td>
<td>41</td>
<td>7</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>19 22 August</td>
<td>28</td>
<td>14</td>
<td>45</td>
<td>6</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>20 24 August</td>
<td>29</td>
<td>16</td>
<td>48</td>
<td>9</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>21 26 August</td>
<td>30</td>
<td>17</td>
<td>56</td>
<td>12</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>22 28 August</td>
<td>21</td>
<td>15</td>
<td>32</td>
<td>3</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>23 30 August</td>
<td>17</td>
<td>11</td>
<td>23</td>
<td>8</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>24 1 September</td>
<td>12</td>
<td>7</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
As in experiment 5, the authors did a series of pairwise single-factor CRD ANOVAs or chi-square tests comparing the various patterns (their Table S10). The ANOVAs are not appropriate, and the authors approach does not control the false positive rate.

In this case, the experimental unit for the vertical and horizontal placement are separate panels in each week and are independent substrates. This is now an example of randomized block design with a two-factor experiment done in each block rather than a split-plot design. The appropriate model is

\[ Y_{ijk} \sim \text{Poisson}(\mu_{ijk}) \]
\[ \theta_{ijk} = \log(\mu_{ijk}) \]
\[ \theta_{ijk} = \text{Date} + \text{Stripe} + \text{Position} + \text{Stripe}:\text{Position} \]

where the term Date represents the blocking effect; and the remaining terms represent the main effects and interaction of the two factors. This model can be fit in R:

```r
result6.glm <- glm(tabanids ~ Date + Position + Stripes + Position:Stripes, data=fly6, family=poisson(link="log"))
anova(result6.glm, test='Chi')
```

giving:

<table>
<thead>
<tr>
<th>Terms added sequentially (first to last)</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>143</td>
<td>1275.39</td>
<td>120</td>
<td>833.26</td>
<td>&lt;2e-16 ***</td>
</tr>
<tr>
<td>Date</td>
<td>23</td>
<td>442.13</td>
<td>120</td>
<td>398.91</td>
<td>&lt;2e-16 ***</td>
</tr>
<tr>
<td>Position</td>
<td>1</td>
<td>434.35</td>
<td>119</td>
<td>126.34</td>
<td>&lt;2e-16 ***</td>
</tr>
<tr>
<td>Stripes</td>
<td>2</td>
<td>272.57</td>
<td>117</td>
<td>126.16</td>
<td>0.9166</td>
</tr>
<tr>
<td>--- Signif. codes:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 ‘<em><strong>’ 0.001 ‘</strong>’ 0.01 ‘</em>’ 0.05 ‘.’ 0.1 ‘ ’ 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Not unexpectedly, there is strong evidence of main effects, but no evidence of an interaction between the effects of stripes and color. The interaction plot is shown in Figure 4. Estimates of effect sizes can be obtained in the usual way (see R code for details).
There were several errors made in the analysis of this data by the authors.

- All of the ANOVAs conducted by the authors are inappropriate because they used a completely randomized design ANOVA when, in fact, the data were collected under a randomized complete block or split-plot design.
- The $\chi^2$ tests may be appropriate under a conditioning argument, but I was unable to reproduce some of the results.

Fortunately, the results are striking enough that even a poor analysis generally leads to the correct conclusions about the relative attractiveness of various patterns to tabanids. However, the incorrect analyses should be corrected so that future experimenters do not make the same errors.

The authors published the complete data on the journal website so it was possible to reproduce exactly how they proceeded in most cases. However, there were a number of results that could not be reproduced – it would also be helpful to post the computer code that performed the analyses to see how these results were obtained.
9 Appendix: Why are the authors’ $\chi^2$ tests valid but their ANOVA tests not valid?

While the authors’ use of a single-factor completely randomized design (CRD) ANOVA is NOT appropriate for the way the experiment was collected, their use of a $\chi^2$ test to test for equality of attractiveness may be appropriate.

Consider, for example, the data from Table[1] A model for this data is

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$\mu_{ij} = \mu \times \lambda_i \times f_j$$

where $\mu_{ij}$ is the mean count of flies attracted in sampling interval $i$ and stripe treatment $j$; $\mu$ is the overall mean per unit area of the number of flies that are attracted; $\lambda_i$ is a factor that adjusts the means upwards and downwards (e.g. some sampling intervals have higher means and some intervals have lower means); and $f_j$ is the factor for stripe condition $j$.

Under this model, the distribution of flies between the three conditions conditional on the total number of flies observed in interval $i$ ($Y_i$), is multinomial with probabilities given by the ratio of the respective means.

$$Y_{ij} | Y_i \sim \text{Multinomial}(\frac{\mu_{ij}}{\mu_i} = \frac{f_j}{f})$$

Under the null hypothesis of no difference in attractiveness of the treatments, the $f_j$ would be all equal, and the multinomial distribution would have equal probability in each treatment group.

Because each sampling interval has the same conditional probabilities, the sum over the sampling intervals also has the same multinomial distribution which then leads to the stated $\chi^2$ test.

However, it is not clear how to do the multiple-comparison tests after the omnibus test using this approach, nor is it obvious how to compute the standard errors of the ratios of the means to estimate effect sizes. Also, if over dispersion was present in the data, this approach would not work.

So while valid, the $\chi^2$ tests do not readily extract all of the information from the data. The generalized linear model approach is preferred.