M&M's® Simulation

James Mason

Friday, April 03, 2015

This R Markdown code explores the issues with a misspecified ANOVA model for investigating the homogeneity of the distribution of colors within packages of m&m's® candies[1].

```
colors=factor(c('red', 'yellow', 'orange', 'green', 'brown', 'blue'))
n.candies=20             # Number of m&m's per box
n.boxes=990              # Number of boxes sampled in each experiment,
                         # should be a multiple of the number of colors
n.replications=50000     # Number of experiments

library(foreach)
set.seed(1337)           # For replicability; remove this line for a random seed
```

The data-generating model is that boxes of m&m's® are assumed to be filled with a *fixed* number of candies, drawn at random from an infinite vat containing unknown proportions of each color. The research question is whether the colors in this vat are in *equal proportions*, or whether those *proportions differ*.

To answer this question, a researcher proposes to draw a sample of boxes of m&m's®, and count the numbers of each color in each box. ANOVA would then be conducted to see if the counts differ between the colors, using the color for group membership and the count as the outcome variable. Thus, if there are c different colors of m&m's®, and N boxes were sampled, $c \times N$ data points would be entered into ANOVA, c for each box. This model is misspecified because the count of colors are *not independent* within each box: for example if the box is mostly full of blue candies, there will be less room for each of the other colors.

For the statistical test implied by the ANOVA model, hypothesis test is as follows:

- H_0: The colors in the vat are in equal proportion.
- H_A: At least one color in the vat has a different proportion than the others.

In a correctly-specified model, if the null hypothesis is true, with a significance threshold of α, we would expect to obtain $p < \alpha$ (and thus falsely reject the null hypothesis) in exactly α proportion of experiments. (For example if our significance threshold is $p < 0.05$, then we would expect to falsely reject the null 20% of the time). This is, in fact, the *definition* of the p-value, and thus the *sampling distribution of p-values* should have a cumulative distribution function (CDF) which is:

$$g(p) = 1 \quad \text{if } 0 < p < 1$$
$$g(p) = 0 \quad \text{otherwise}$$

We investigate these models using statistical simulation, a computationally-intensive[2] process, with 50000 replications per model.

[1] "m&m's" is a trademark of Mars Chocolate North America, LLC

In this first simulation, we simulate a *correctly* specified ANOVA model where, for each color, we draw an independent sample from the standard normal distribution $N(0, 1)$.

```
p.values = times (n.replications) %dopar% {
  # For each color, draw N standard normal deviates
  counts=data.frame(color=rep(colors, each=n.boxes),
                    count=rnorm(n.boxes*length(colors)))

  # To run ANOVA in R, we first fit ANOVA as a linear model using lm(),
  # then use the anova() function to compute the required F-tests.
  model=lm(count~color, data=counts)
  results=anova(model)

  # Next, we extract and return the p-value; these are collected by times() into a vector
  p.value=results['color','Pr(>F)']
  p.value
}

# Set smaller borders for our plots
par(mar=c(4, 4, 0.5, 2)) # (bottom, left, top, right)

# Histogram of the p-values
hist(p.values, xlim=c(0, 1),
     breaks=min(20, n.replications/10),
     main=NA, freq=FALSE)

# Overlay a kernel density plot
lines(density(p.values, from=0, to=1))

# Overlay the density=1.0 criterion line
abline(h=1.0, lty=4)
```

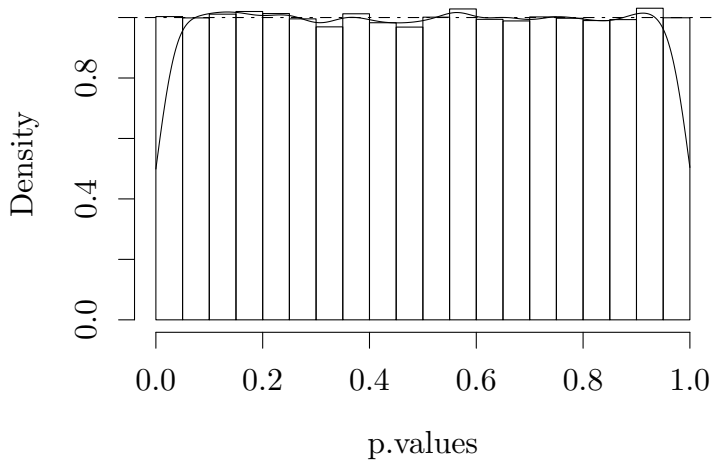


Figure 1: Empiricial Distribution of p-values under simple ANOVA

With a large number of replications, no obvious bias is visible.

In this second simulation, we simulate experiments under the null hypothesis: *many* experiments are simulated in which we sample a large number of boxes of m&m's®, drawn from a vat with equal proportions of the colors (defined above), and run the misspecified ANOVA described above. We then investigate the distribution of p-values from these experiments, to see how this distribution deviates from the expected $Uniform(0, 1)$ distribution.

```
p.values = times (n.replications) %dopar% {
  # Draw a sample from a vat with equal proportions of each color
  sample=t(replicate(n.boxes, sample(colors, n.candies, replace=TRUE)))

  # We count each color in each box and combine these counts into a single dataset.
  counts=data.frame(foreach(color=colors, .combine=rbind) %do%
                      data.frame(color=color,
                                 count=apply(sample, 1, function(row) sum(row == color))))
  model=lm(count~color, data=counts)
  results=anova(model)
  p.value=results['color','Pr(>F)']
  p.value
}

# Histogram of the p-values, density plot, and criterion line
par(mar=c(4, 4, 0.5, 2)) # (bottom, left, top, right)
hist(p.values, xlim=c(0, 1),
     breaks=min(20, n.replications/10),
     main=NA, freq=FALSE)
lines(density(p.values, from=0, to=1))
abline(h=1.0, lty=4)
```

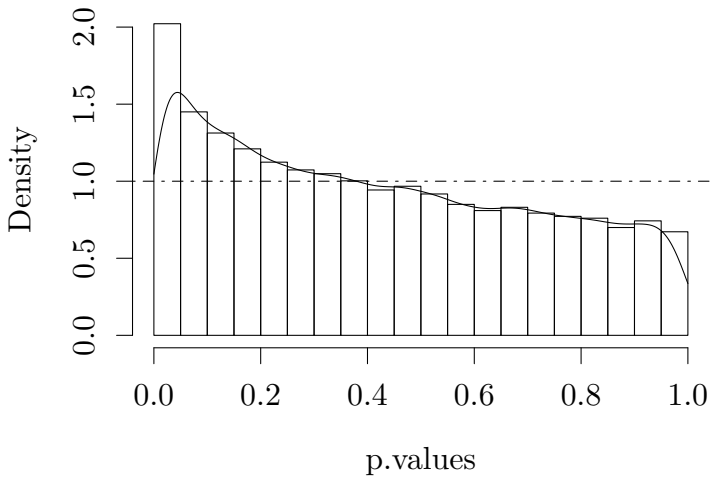


Figure 2: Empiricial Distribution of p-values under misspecified ANOVA

Even with a large number of replications, bias is clearly visible: there is too much density near 0. This indicates that they null hypothesis would be rejected too often.

In this third simulation, we attempt to use ANOVA correctly in the above situation. Again we simulate the same null hypothesis: *many* experiments are simulated in which we sample a large number of boxes of m&m's®, drawn from a vat with equal proportions of the colors (defined above). This time, we divide the sample into different portions, of size $\frac{n.boxes}{n.colors}$, and count the colors in separate fractions of the sample. Although this reduces our overall sample size, it will mean that the counts of each color are independent of each other.

```
p.values = times (n.replications) %dopar% {
  # Draw a sample from a vat with equal proportions of each color
  sample=t(replicate(n.boxes, sample(colors, n.candies, replace=TRUE)))

  # Split the sample into equal parts for each color.
  split = split(data.frame(sample), colors)

  # We count each color in it's separate portion, and recombine the
  # results into a single dataset.
  counts=data.frame(foreach(color=colors, .combine=rbind) %do%
                    data.frame(color=color,
                               count=apply(split[[color]], 1,
                                           function(row) sum(row == color))))
  model=lm(count~color, data=counts)
  results=anova(model)
  p.value=results['color','Pr(>F)']
  p.value
}

# Histogram of the p-values, density plot, and criterion line
par(mar=c(4, 4, 0.5, 2)) # (bottom, left, top, right)
hist(p.values, xlim=c(0, 1),
     breaks=min(20, n.replications/10),
     main=NA, freq=FALSE)
lines(density(p.values, from=0, to=1))
abline(h=1.0, lty=4)
```

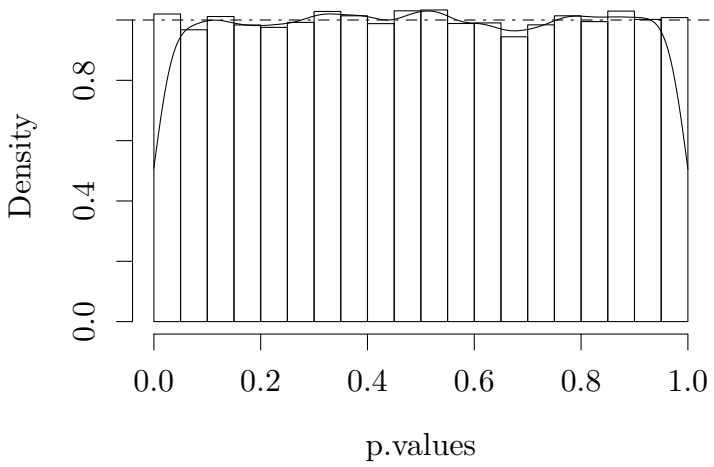


Figure 3: Empiricial Distribution of p-values under corrected ANOVA

With a large number of replications, no obvious bias is visible.

In this fourth simulation, alter the data-generating model so that the vat contains unequal proportions of the different colors of m&m's®. We investigate the power of the corrected ANOVA to detect a situation where the null hypothesis is, in fact, false.

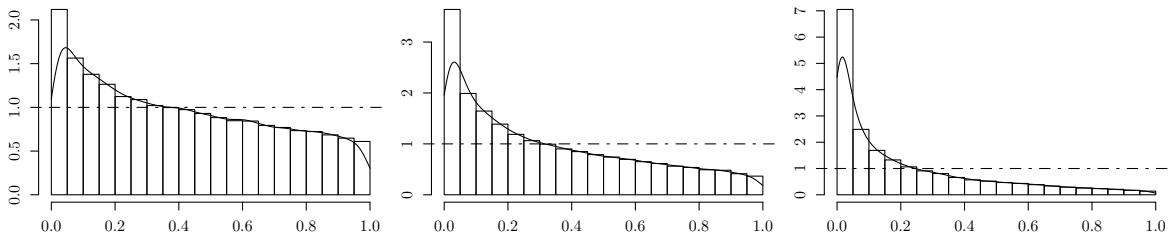
As this is a power study, we these experiments use a smaller sample size; instead of 990 boxes, we try smaller samples: 60, 120, and 240.

```
for (n.boxes in sample.sizes) {
  p.values = times (n.replications) %dopar% {
    # Unequal proportions in the vat:
    weights=rep(1, length(colors))
    names(weights)=colors
    weights['red']=1.2

    # As before, draw a sample from a the vat, and split it
    sample=t(replicate(n.boxes, sample(colors, n.candies, replace=TRUE, prob=weights)))
    split = split(data.frame(sample), colors)

    # Now assemble the counts, as before.
    counts=data.frame(foreach(color=colors, .combine=rbind) %do%
                        data.frame(color=color,
                                   count=apply(split[[color]], 1,
                                               function(row) sum(row == color))))
    model=lm(count~color, data=counts)
    results=anova(model)
    p.value=results['color','Pr(>F)']
    p.value
    }

  # Histogram of the p-values, density plot, and criterion line
  par(mar=c(2, 2, 0.5, 0.5), cex=0.5) # (bottom, left, top, right)
  hist(p.values, xlim=c(0, 1),
       breaks=min(20, n.replications/10),
       xlab=NA, ylab=NA, main=NA, freq=FALSE)
  lines(density(p.values, from=0, to=1))
  abline(h=1.0, lty=4)
  }
```



(a) 60 boxes (b) 120 boxes (c) 240 boxes

Figure 4: Empiricial Distributions of p-values under unequal proportions

At reasonable sample sizes (about 40 *per color*), this test appears to reliably detect that the proportion of one color is different.

[2]Computation took 169 minutes using 32 cores.