Fruit snack analysis

Karl Broman

2016-06-24

# Introduction

My kids and I like [Welch's fruit snacks](http://www.welchsfruitsnacks.com/). We buy them from CostCo, in big boxes with 80 packages.

There are five flavors/colors. We're not entirely sure what the corresponding fruits are supposed to be. The kids proposed:

* blueberry (or grape?) - *purple*
* peach - *yellow*
* raspberry - *pink*
* strawberry - *red*
* orange - *orange*

The number of fruit snacks in a package is generally 12 or 13, but we've noticed that the five types are not equally frequent, and it seems like they are not completely random. For example, we prefer the dark purple ones, and they seem to be clustered: that you often get none or a number of them.

We wondered:

* How frequent are the five different flavors?
* Are the flavors in a package completely random?

To address these questions, we gathered some data: we counted the number of snacks of each flavor in each package in one box. (It turned out that there were *81* packages rather than 80 in our box.)

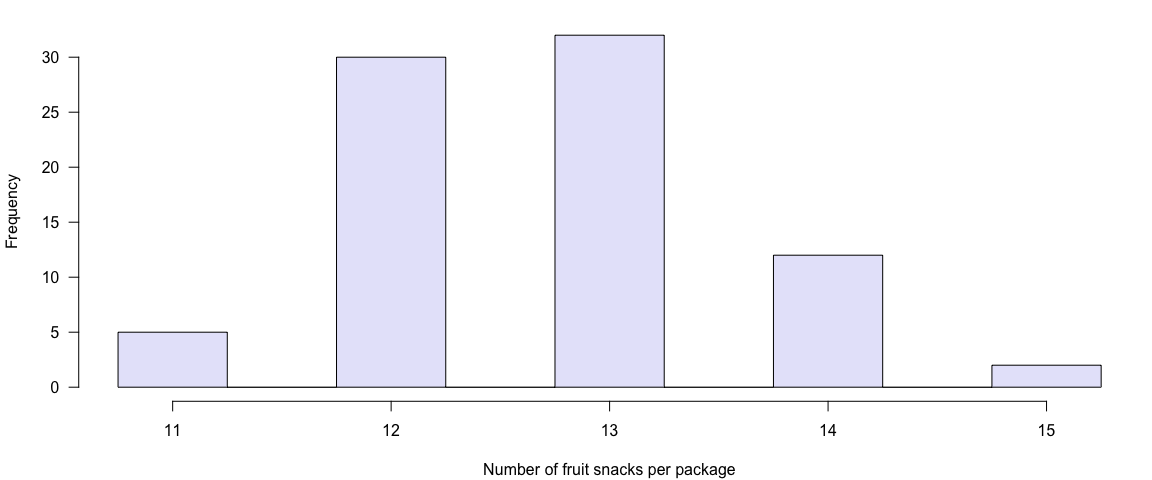
The data are available at GitHub. This document describes our results.

* [CSV file of the data](https://raw.githubusercontent.com/kbroman/FruitSnacks/master/Data/fruit_snacks.csv)
* [Photo gallery](https://github.com/kbroman/FruitSnacks/blob/master/PhotoGallery.md)
* [Scanned data](http://kbroman.org/FruitSnacks/assets/fruit_snacks_scanned.pdf)

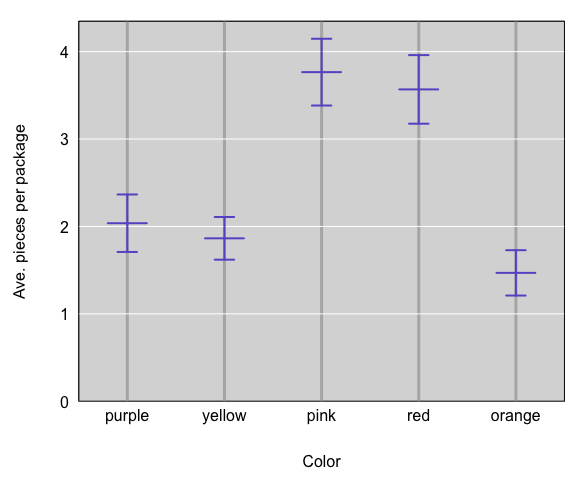
I don't give many methods details here; they're described in [a separate document](http://kbroman.org/FruitSnacks/assets/fruit_snacks_methods.html).

# Summaries

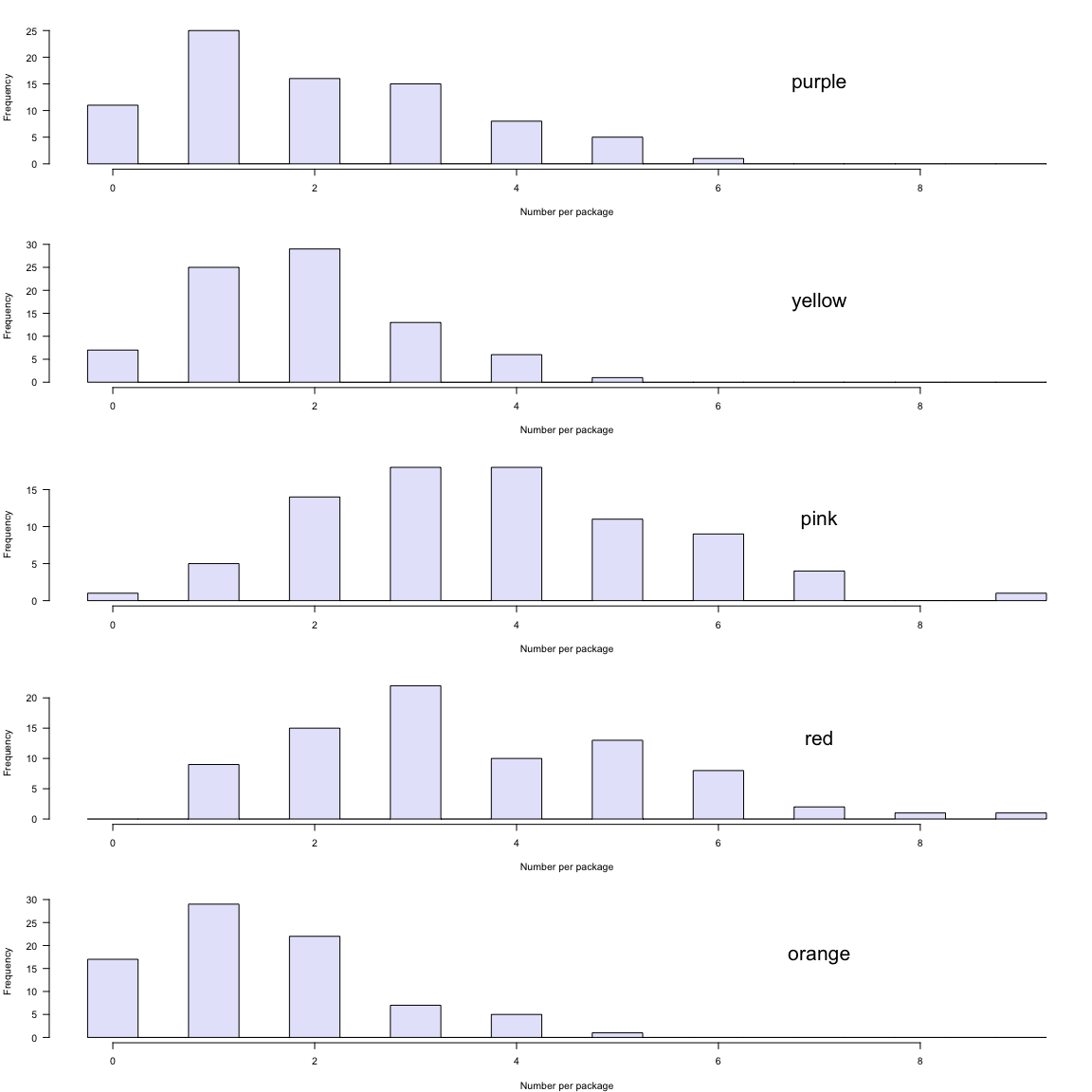
There were 81 packages containing a total of 1029 fruit snacks in 5 colors. (I'd call them "flavors" but we can't much distinguish among them. Maybe we just eat them too quickly) So there was an average of 12.7 fruit snacks per package, with a range of 11 – 15. Here's the distribution:



The different colors have quite different frequencies. Here's a plot of the average number of snacks of each color in a package, with 95% confidence intervals.



Here are histograms of the number of snacks per package for each color.



# Tests for differences

It seems pretty clear from the histograms above that pink and red are the most common. In the observed data, orange is least common. Can we infer a general rule here? In Welch's big vat of fruit snacks, are pink and red most common, purple and yellow in the middle, and orange least common?

We'll do some simple pairwise statistical tests, to check this. For example, if purple and pink are equally frequent in Welch's vat of fruit snacks, what would be the chance of seeing data as different as observed? I'll use pairwise, paired permutation tests.

The following are the p-values obtained from 10,000 permutations for each pair of colors.

## p-value  
## purple:yellow 0.461  
## purple:pink 0.000  
## purple:red 0.000  
## purple:orange 0.026  
## yellow:pink 0.000  
## yellow:red 0.000  
## yellow:orange 0.041  
## pink:red 0.562  
## pink:orange 0.000  
## red:orange 0.000

As seen in this table of p-values, there's very strong evidence that orange,purple,yellow < pink,red, and reasonably strong evidence that orange < purple,yellow. The observed differences in the frequency of the purple and yellow snacks can reasonably be ascribed to chance variation, as are the differences between pink and red.

# Is there clustering of colors?

I've had the impression that there is some clustering of colors. For example, there seemed to be a tendency to get either no purple snacks or many purple snacks.

If colors were randomly assigned to packages (but at color-specific frequencies), the number of snacks of a particular color, given the total number of snacks in a package, would follow a binomial distribution. The distribution of the counts of a particular color across packages would follow a mixture of binomial distributions. (A *mixture*, because the number of snacks in a package varies.)

It might be best to stratify based on the number of snacks in a package, but I'm going to just look overall. The simplest thing to do is to look at the varibility (as measured by the standard deviation SD) in the number of snacks of a particular color. If purple snacks are clustered, the SD should be higher than that expected under the binomial mixture model.

I'll compare the observed SDs of the counts of each color across packages to what would be expected if colors were assigned to packages completely at random, and I'll calculate a p-value from a randomization test: compare the observed SD to the distributed of estimated SDs you get when you randomly permute the snacks across packages.

I'll performed two-sided tests, with 10,000 permutations.

Here are the results: observed and expected SDs, and a p-value for the test.

## observed\_SD expected\_SD p-value  
## purple 1.49 1.32 0.081  
## yellow 1.10 1.27 0.095  
## pink 1.73 1.65 0.544  
## red 1.77 1.62 0.228  
## orange 1.17 1.14 0.747

The purple snacks are have higher-than-expected SD (indicating possible clustering), but the yellow snacks have lower-than-expected SD (indicating anti-clustering: that the snacks are more evenly distributed than would be expected under randomness). But in both cases, the observed difference could reasonably be ascribed to chance variation.

I was using counts there; it seems like maybe I should look at the *proportion* instead: for each color, the SD across packages of the proportion of snacks that are that color.

So I'll repeat the permutation tests, using the SD of the proportions as my test statistic, and again using 10,000 permutation replicates.

Here are the results: observed and expected SDs of the proportions, and a p-value for the test.

## observed\_SD\_prop expected\_SD\_prop p-value  
## purple 0.119 0.103 0.051  
## yellow 0.088 0.100 0.120  
## pink 0.134 0.128 0.552  
## red 0.132 0.126 0.572  
## orange 0.091 0.090 0.852

Using the SD of the proportions as the test statistic, the evidence for clustering in the purple snacks is a bit stronger, with P=0.051. I'll tentatively conclude that there *is* clustering. But maybe we should gather some more data.

[Source on GitHub](https://github.com/kbroman/FruitSnacks/blob/master/R/fruit_snacks_methods.Rmd)

[CC0](http://creativecommons.org/publicdomain/zero/1.0/)