Love the Way You Chi: *An M&M Masterpiece*

1. What is the distribution of colors in a M&M bag?
2. How does that differ from what we expect?
3. Is that difference “statistically significant” or due to chance?

Count your M&M's. If there were an *equal* amount of each color the **expected** number of M&Ms for each color would be:

brown	yellow	red	blue	orange	green

Your **actual** color count distribution is:

brown	yellow	red	blue	orange	green

Mars Company is Not Afraid to release the proportion of colors that they *supposedly* produce[1]:

Milk Chocolate				Peanut			
Brown:	13%	Blue:	24%	Brown:	12%	Blue:	23%
Yellow:	14%	Orange:	20%	Yellow:	15%	Orange:	23%
Red:	13%	Green:	16%	Red:	12%	Green:	15%

Your expected count distribution if Mars is correct:

brown	yellow	red	blue	orange	green

Is this difference small enough to attribute it to **chance sample error** or must our expected distributions be wrong? Let's quantitatively find an answer.

I Need a Statistician

[1] us.mms.com/us/about/products/milkchocolate and us.mms.com/us/about/products/peanut

Stats God

We're going to be using **Chi-square analysis** to test **goodness-of-fit** *between our actual and expected distributions*.

Any good stats test has a specific statement their trying to prove or disprove (a **null hypothesis**).

Example: There is no difference in the expected class average of 100% on Ms. Gerry's tests and the actual average.

What is a possible null hypothesis?

χ^2 test-statistic:

a value that quantifies the difference between an observed and expected distributions. Formally,

$$\chi^2 = \sum_{\text{all categories}} \frac{\text{observed value} - \text{expected value}}{\text{expected value}}$$

Calculate the χ^2 statistic for both expected distributions.

Show all work. You may a table useful to keep track of (O-E)/E for each color before you sum them.

The Real Slim Chance

Higher χ^2 correspond to bigger departure from expected values.

But what χ^2 values are small enough to be ignored and labeled as chance sampling errors from the expected?

Let's measure the probability of getting the deviation from expected that we did (read: probability of getting your χ^2).

Checkout this almighty table:

http://sites.stat.psu.edu/~mga/401/tables/Chi-square-table.pdf

df on the left column refers to a test's **degrees of freedom**: the number of categories (that you summed) minus one. What is your *df*?

Now, you can look up the probability of getting the χ^2 value(s) that you did (AKA look-up the **p-value**).

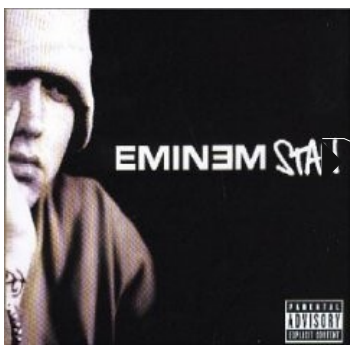
What are your p-values?

So what do these p-values even mean?

The p-value is the probability that you get χ^2 value atleast as high as the one you calculated, randomly sampling from the expected, underlying distribution.

A low p-value (less than 0.05) means that it is pretty unlikely that you got your sample of M&Ms from your hypothesis's expected distribution. If such is the case, you'd reject your null hypothesis.

Should you reject your null hypothesis?



KICKING AROUND CHI SCWAIR

Messi thinks starting lineups in La Liga generally have an average height for each position:

Goalkeeper: 1.82, Defender: 1.87, Midfield: 1.85, Striker: 1.87

Madrid has a starting height line-up as follows:

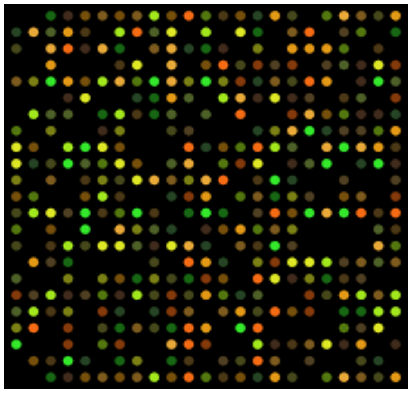
Goalkeeper: 1.84
Defender: 1.91 1.83 1.72 1.84
Midfield: 1.89 1.82 1.8 1.76
Striker: 1.87 1.8

Does Real Madrid disprove Messi's theory?

COMPUTATIONAL DIAGNOSIS

Genes have quantifiable expression levels that vary depending on how much protein is transcribed from that gene.

Kick-ass technologies like *microarrays* can quickly profile a cell and determine these expression values.



A microarray scan result. Color and intensity of each dot represent the expression of a single gene

One issue with symptom-based diagnosis when you go to the doctor is that it's often subjective. *It's estimated about 10-20% of cancer-related patients are misdiagnosed*.

Just one reason why we have blood tests to gather things like expression values.

COMPUTATIONAL DIAGNOSIS CONT'D

To get our feet wet, let's try out chi-square to crunch some expression values.

Known model breast cancer patient has the following expression values for cancer related genes:

Gene FGFR2	Gene TOX3	Gene ESR1	Gene CASC16
0.63	0.51	0.81	0.71

A patient comes in with a breast lump, and the profile:

Gene FGFR2	Gene TOX3	Gene ESR1	Gene CASC16
0.60	0.45	0.75	0.92

Cancer or not? Show work.

Is there ever such a thing as a “model” cancer patient? What might happen if I include a gene that isn't related to cancer at all in the profile?

LOOKING AHEAD

A lot of innovation in biology in the past few years boils down to smart handling of large amounts of biological data.

Along those lines, lined up in the next few labs: new modeling techniques, automated Punnett square generators, predicting disease in children with Bayes rule, and applying AI to find gene pathways underlying disease...

Stay tuned until next time.