Module 1 – Section 6

Goodness of Fit Test for One Categorical Variable

Outline

- Review Multinomial Random Variables
- Goodness of Fit
- Expected Values
- Test Statistic and P-value
- Chi-square Distribution
- Example

Multinomial Random Variables

- Random event with J outcomes
- Probability of each Outcome = p_j
- $\sum_{j=1}^{J} p_j = 1$

Multinomial Random Variables

- Y_j = number of observations in j^{th} outcome in n independent and identical trials of random event.
- Independent – outcome on one trial does not affect outcomes on other trials.
- Identical – same probabilities for outcomes.

Goodness of Fit

- Values of p_j are unknown.
- Assume values of p_j for $j = 1, \dots, J$
- Are the observations Y_j consistent with these assumed values of p_j?

Null and Alternative Hypotheses

- $H_0: p_1 = p_{1_0}, p_2 = p_{2_0}, \ldots, p_J = p_{J_0}$
- H_A: At least one $p_j \neq p_{j_0}$ for $j = 1, \ldots, J$

Expected Values

- If the null hypothesis is correct:

$$E(Y_j) = np_{j_0} \text{ for all } j = 1, \ldots, J$$

Test Statistic

- Compare observed values Y_j to expected values $E(Y_j)$.

$$X^2 = \sum_{j=1}^{J} \frac{(Y_j - E(Y_j))^2}{E(Y_j)}$$

Distribution of Test Statistic

- As long as $E(Y_j) \geq 5$ for each j, X^2 will have an approximate chi-square distribution (χ^2) with $J - 1$ degrees of freedom.
- Denote this distribution as χ^2_{J-1}

P-value

- Large values of test statistic X^2 indicate significant differences between observed values Y_j and expected values $E(Y_j)$.
- p-value = $P\left(\chi^2_{J-1} > X^2\right)$

Ex. M&Ms

- In June 2008, I purchased one large bag of M&Ms.
- Company's website provided information on proportion of each color produced.
- Is color distribution of a randomly selected purchased bag consistent with company information?

Ex. Distribution of Colors of M&Ms (Milk Chocolate)*

- Model Probabilities
  - Blue – 24%
  - Orange – 20%
  - Yellow – 14%
  - Red – 13%
  - Green – 16%
  - Brown – 13 %

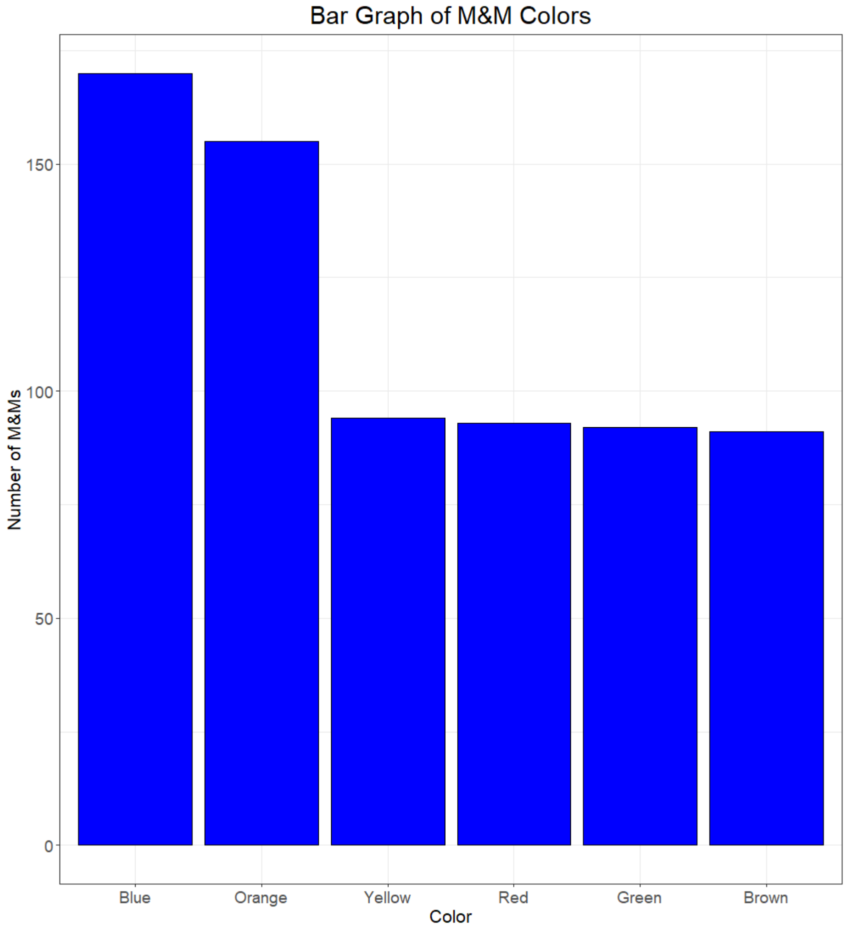
*From company's website – June 2008

Ex. M&Ms Data

Color
Blue
Blue
⋮
⋮
Brown
Brown

Ex. M&Ms Summary Data

Color	Count	Proportion
Blue	170	0.2446
Orange	155	0.2230
Yellow	94	0.1353
Red	93	0.1338
Green	92	0.1324
Brown	91	0.1309
Total	695	1.0000



Ex. M&Ms

- Null Hypothesis

  $$H_0\colon p_{\text{blue}} = 0.24, p_{\text{orange}} = 0.20, p_{\text{yellow}} = 0.14,$$
  $$p_{\text{red}} = 0.13, p_{\text{green}} = 0.16, p_{\text{brown}} = 0.13$$

- Alternative Hypothesis

  H_a: At least one of the probabilities in the null hypothesis is incorrect.

Ex. M&Ms

Color	Count	Model (p_j)	Expected Value (np_j)	Contribution to X^2 $\left(\frac{(Y_j - E(Y_j))^2}{E(Y_j)}\right)$
Blue	170	0.24	$695 * 0.24 = 166.80$	$(170 - 166.80)^2/166.80$
Orange	155	0.20	$695 * 0.20 = 139.00$	$(155 - 139.00)^2/139.00$
Yellow	94	0.14	$695 * 0.14 = 97.30$	$(94 - 97.30)^2/97.30$
Red	93	0.13	$695 * 0.13 = 90.35$	$(93 - 90.35)^2/90.35$
Green	92	0.16	$695 * 0.16 = 111.20$	$(92 - 111.20)^2/111.20$
Brown	91	0.13	$695 * 0.13 = 90.35$	$(91 - 90.35)^2/90.35$
Total	695	1.00	695	5.5604

Ex. M&Ms

- Test Statistic: $X^2 = 5.5604$
- p-value = $P\left(\chi_5^2 > 5.5604\right) = 0.3514$
- We do not have evidence of lack of model fit. The color data in our randomly selected bag are consistent with the model from the website.