Stat 102

Introduction to Business Statistics

Class 2

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- The population/sample paradigm
- Standard error
- Central Limit Theorem
- Summary
- Next time

Last time

Main points:

- Deliverables
- Exams
- Expectations
- And we collected some data:

Figure 1: M&M packet

The big picture

Ultimately we want to ask the question:

Is the regression line flat?

If the line is flat then knowing X tells us nothing about Y (at least in a linear sense). So that this X is not useful in the regression prediction equation.
But asking a question in Statistics usually means running a hypothesis test:

$$H_0 : \beta_1 = 0 \quad v. \quad H_1 : \beta_1 \neq 0,$$

and this requires some concepts that we will discuss today.
The concepts are the

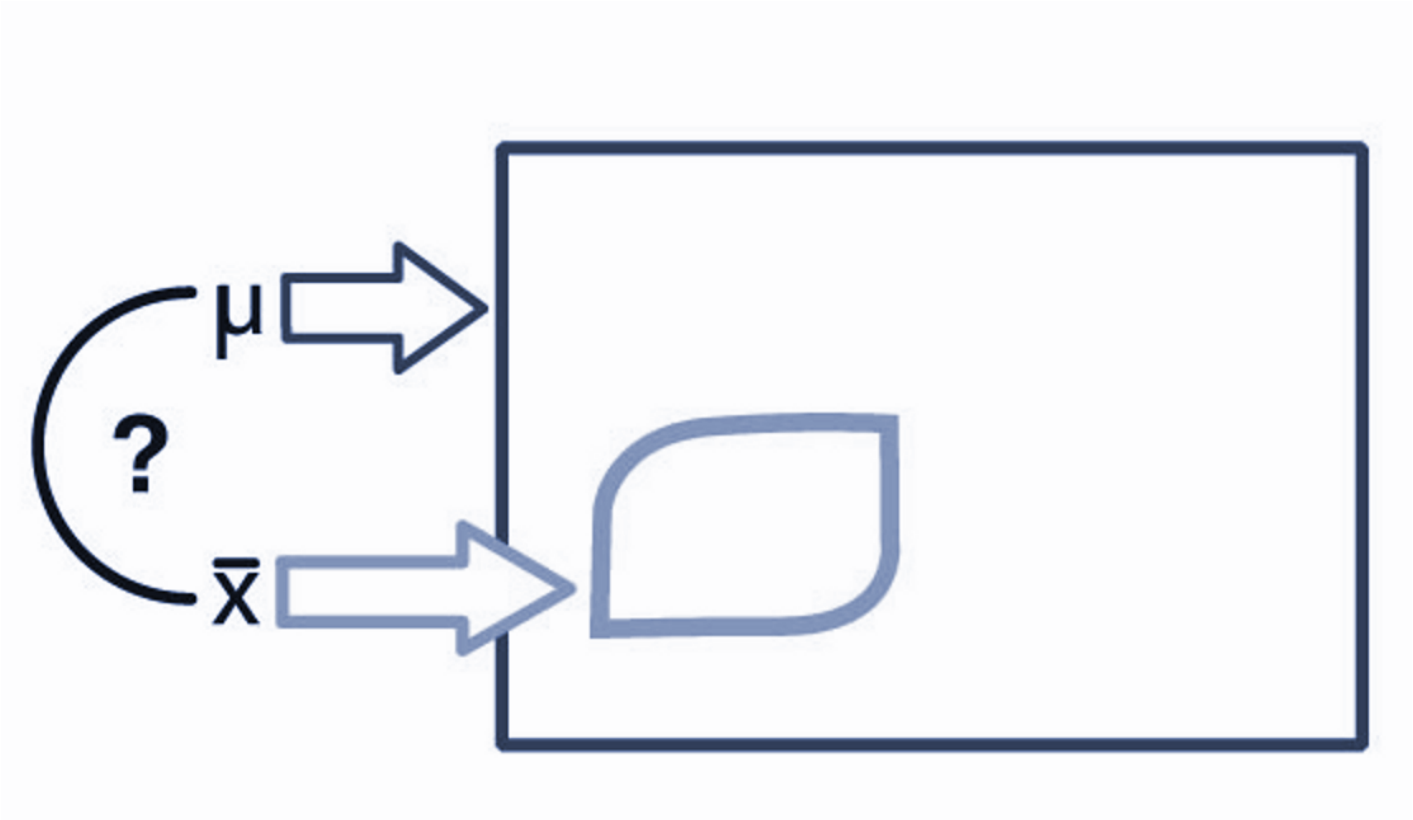
1. Standard error (p. 334-335)
2. Central Limit Theorem (p. 269-70)

which are both relevant in the regression context.

The population/sample paradigm

- Population parameters (μ) are the features we want to make inference about
- Sample statistics ($\bar{X}$) are what we have available from the data

Figure 2: The population/sample paradigm



Exploratory data analysis of the M&Ms

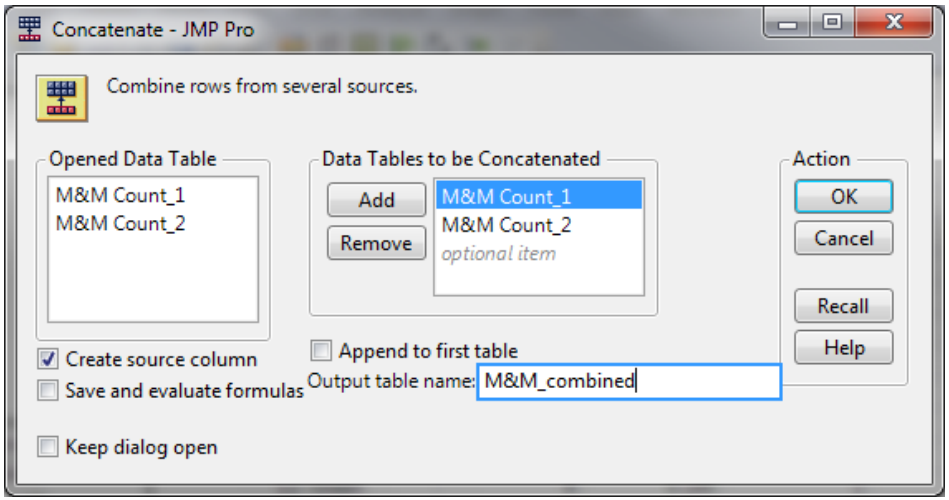
- Always, always plot the data
- Be on the look out for the shape of distributions and outliers

Two helpful students, one from each section, entered the data. Each made their own data table. The way to combine their JMP tables is with the command


Tables → Concatenate

Combining rows from sources

Figure 3: The Tables → Concatenate dialog



By checking the "Create source column" box, we will be able to compare the two sections at a later date.

Cleaning the data

- Deleted empty rows
- Removed "team 0" (that was my packet of M&Ms)
- Created a new column that would provide a *unique* team ID. This was because teams were numbered within sections, so that there were, for example, two team 1s.

The new column was created via the *Formula dialog* and the JMP function:

The Team variable had to be turned into a character for the concatenation. I also concatenated the underscore character "_" into the string for readability.

Cleaning the data, ctd.

Figure 4: The Column change numeric to character dialog

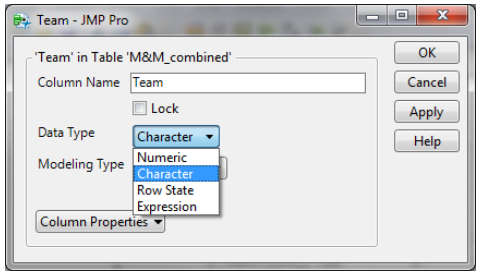
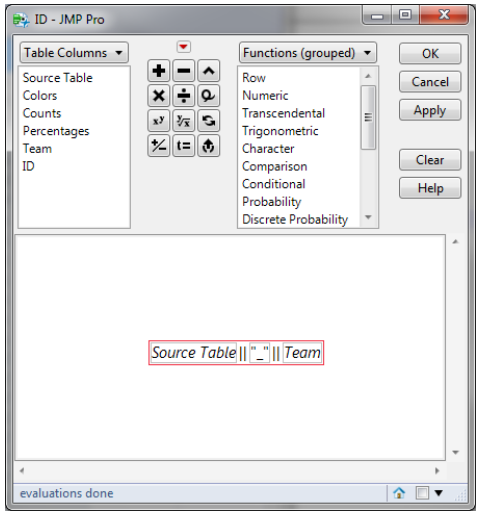
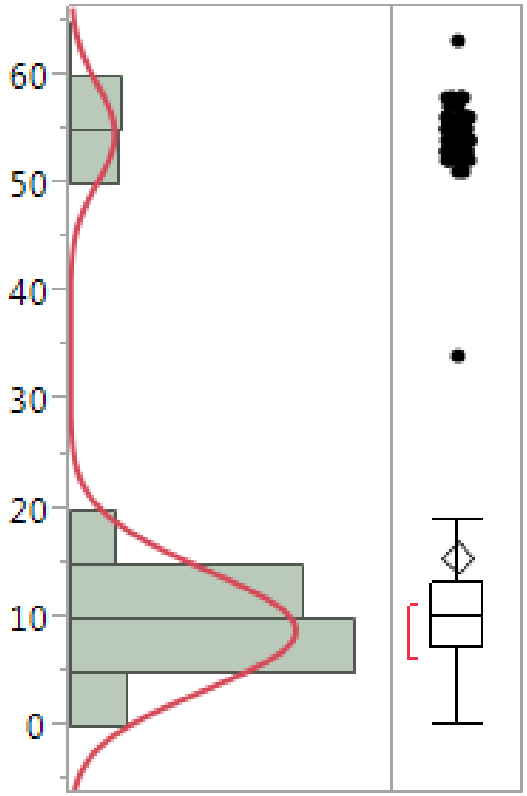


Figure 5: The Formula → Text, Concat dialog



The Count variable

Figure 6: Distribution of counts with smooth curve

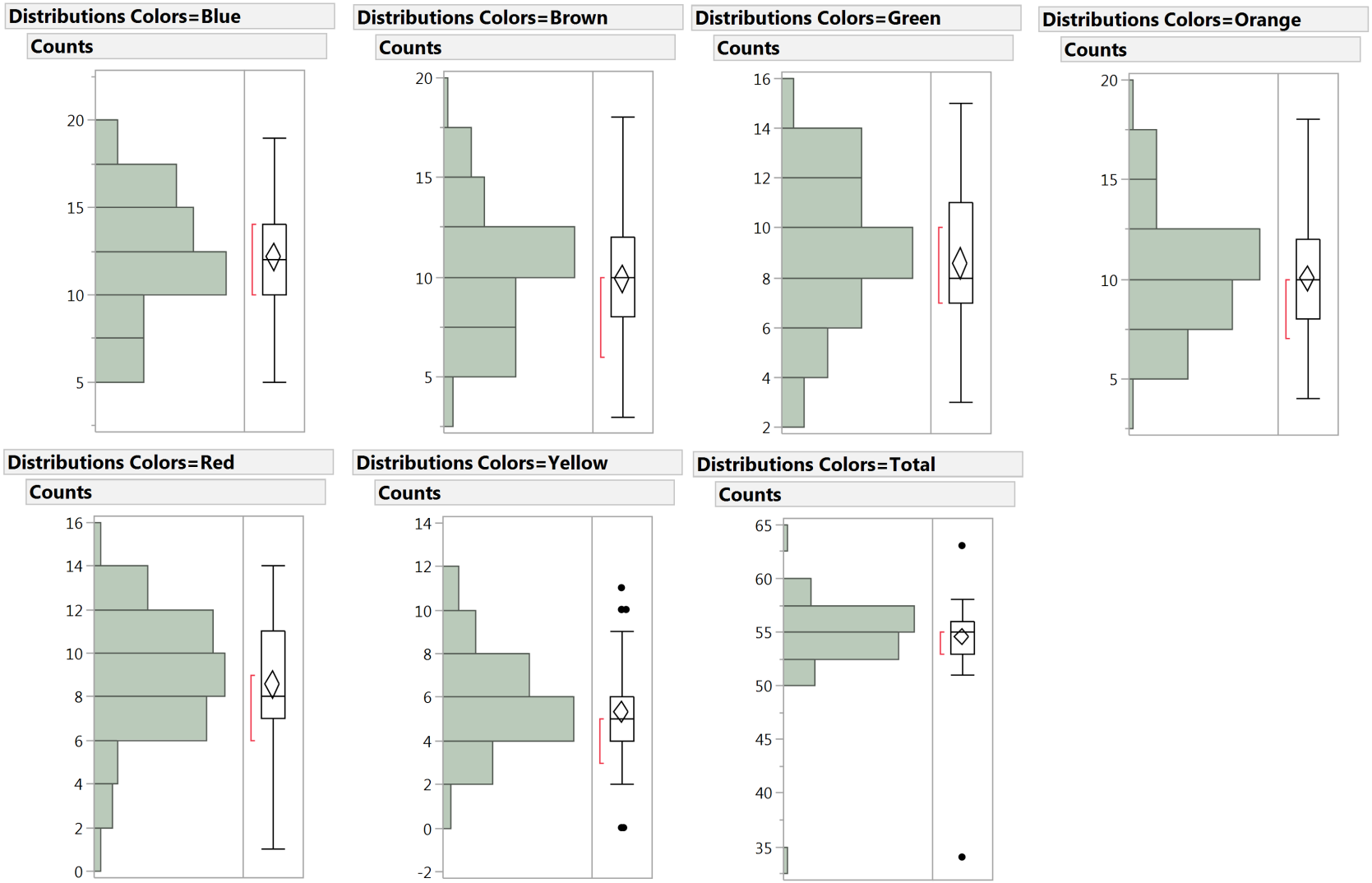


Two "bumps" usually tells a story. What is going on here?

Analysis "by" color

- We have inadvertently mixed up 2 groups: the count of each color and the total M&M count.
- To gain some insight we need to look at distributions *by* color.
- Whenever you see two bumps (a bi-modal distribution) think about whether there is a background variable that might explain group membership.

Figure 7: Results from the *Analyze*, *Distribution*, *By* dialog



Observations ...

- Wild outlier in the *Total* variable. The minimum value is only 34
- Also, an unusually high value for Total
- Outliers flagged for *Yellow* but that also is the least frequent color
- There were in fact some zero counts for yellow
- Somewhat *bell-shaped* distributions across the board
- My count for `blue` was only 3. That feels rather surprising, for me to get the lowest out of 80 groups

Try and explain outliers if you can. Don't just remove by default.
What explanations do we have?

Observations ...

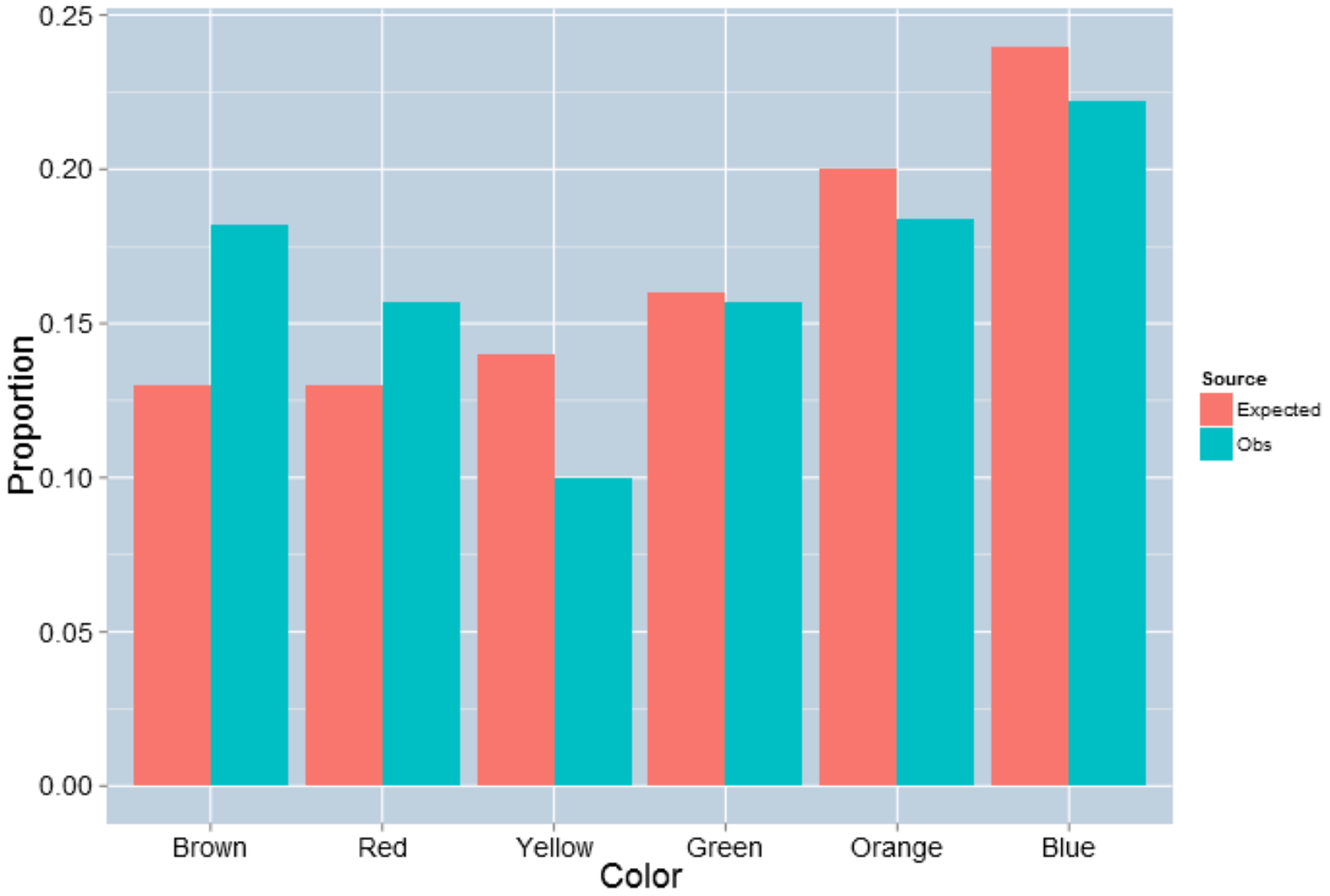
The average proportions and standard deviations (across the 80 groups):

Table 1: M&M average proportions and standard deviations

	Brown	Red	Yellow	Green	Orange	Blue
Expected	0.13	0.13	0.14	0.16	0.20	0.24
Average proportion	0.182	0.157	0.100	0.157	0.184	0.222
Standard deviation	0.056	0.049	0.039	0.051	0.052	0.061

It looks like more browns but fewer yellows.

Figure 8: Comparisons of observed and expected proportions for M&Ms (averaged over all student groups)



Focus on just the blues and a single student group (Team 2_24)

This is how the majority of Science works: you only have one experiment (one bag of M&Ms) to work with. Assume this group doesn't know the other groups exist, except for the sample to sample variation estimate, 0.061. Key question:

> *Does it look like their observed proportion, 0.182 is concordant with the expected/theoretical 0.24?*

Approach: count how many standard deviations 0.182 is away from 0.24. Is what they have observed unusual if the true proportion is 0.24?

$$Z = \frac{0.182 - 0.24}{0.061} = -0.951.$$

So this is an approximate one standard deviation event if the true proportion is 0.24. How **unusual/rare/surprising** an event is this?

Facts that would be very helpful for Team 2_24 to believe

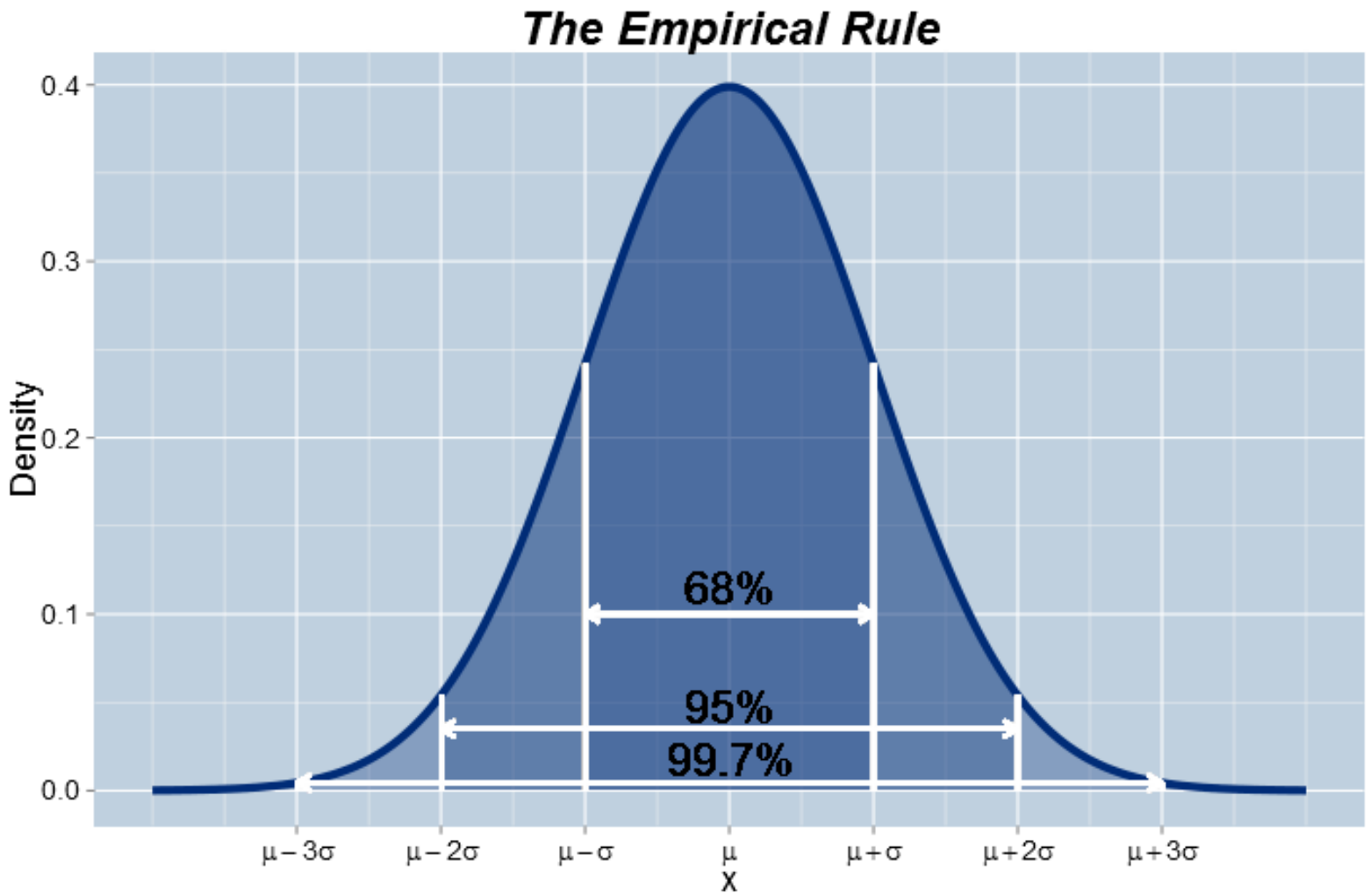
If

1. Sample proportions come from a Normal distribution, centered around the hypothesized value
2. The standard deviation of this Normal Distribution is about 0.061.

then we can use the Empirical Rule to judge how unusual a proportion that is 0.951 standard deviations below 0.24 really is.

The Empirical rule Illustrated

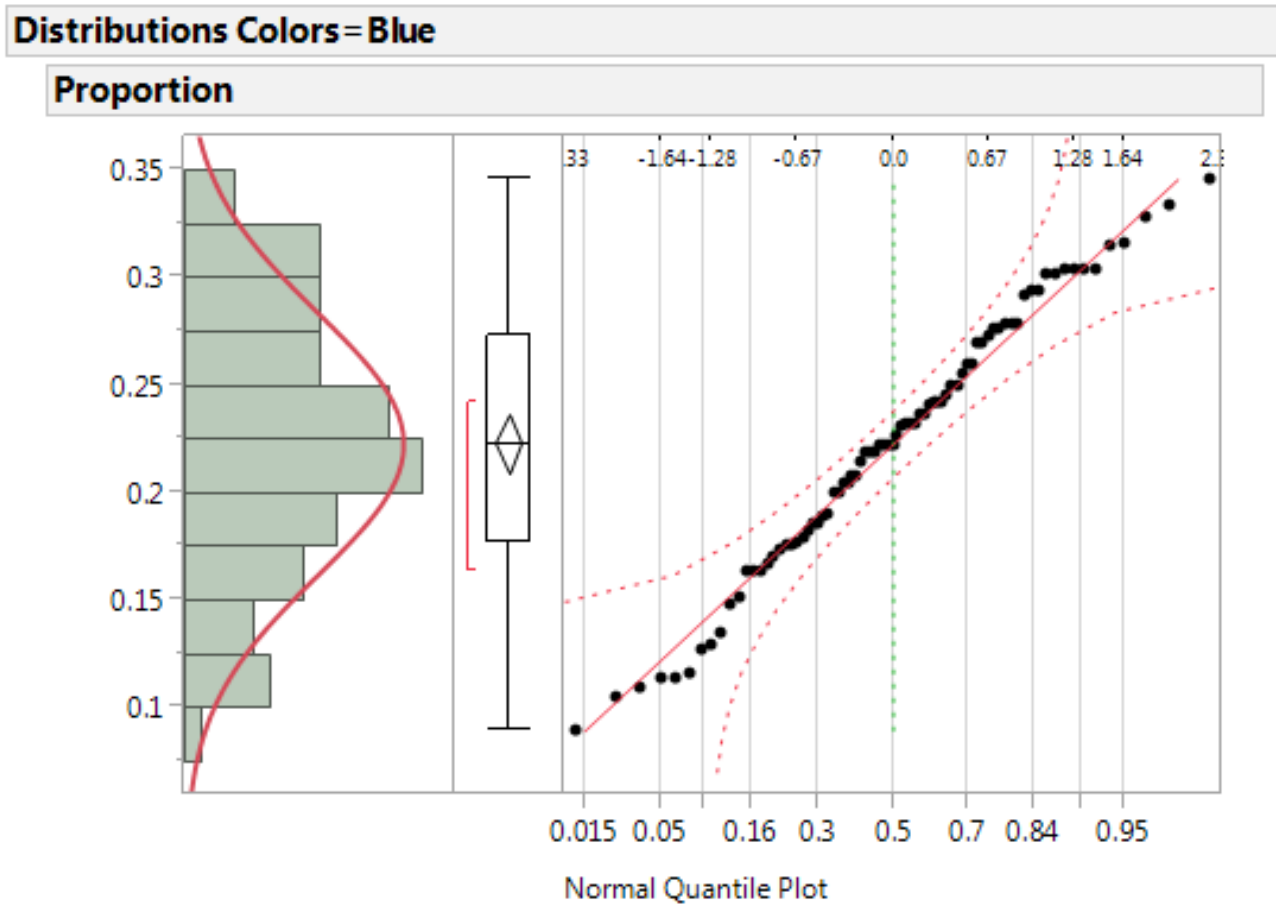
Figure 9: The Empirical Rule



The empirical distribution of the blue proportion from the 80 groups

The Normality assumption looks quite reasonable.

Figure 10: Distribution of just the blue proportions



The probability calculation

- Find
$$P(Z \leq -0.951 \text{ or } Z \geq 0.951)$$
- From the Empirical Rule this is approximately 34%. Using tables or software we get exactly 0.3416
- So this is not a rare event if the null (0.24) is true, so we have no evidence against the null
- This particular group has no basis to say that the proportion of blue M&Ms in the population is anything other than 0.24
- Aside: what defines a rare event? In this class, a probability of less than 0.05 is rare

What the group was relying on for their inference

1. The sample to sample variability of the proportions (0.061) so they could calculate their z-score
2. The reference Normal distribution, against which they compared their z-score

But:

- In real life, they will not know the sample variability of the proportions (0.061) because Science is usually done with a **single** bag of M&Ms
- They won't be able to look at the histogram of sample proportions to see whether or not it is Normally distributed (because they only have a single sample)

This is a conundrum: we don't have access or the ability to check the ideas we were relying on, because we only have a single sample.

The standard error of the mean

Solving the conundrum:

- The sample to sample variation of an estimate is called its *standard error*
- Because we had the luxury of many M&M experiments we could empirically estimate the standard error as 0.061
- In practice, this generally is not an option because we have a single sample

But, the fundamental insight of statistics, is that even with a **single** sample we are able to estimate sample-to-sample variability.
If we denote the standard deviation as σ and the sample size n, then

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

and we estimate the standard error as

$$se(\bar{X}) = \frac{s}{\sqrt{n}}.$$

Back to group 2_24

- They had a sample size of n = 55 and observed 10 blues
- Call a blue a 1 and any other color a 0
- Then the average of the 0s and 1s is 10/55 = 0.182 (the proportion of blues). The standard deviation of the 0s and 1s is s = 0.3892. This gives an estimated standard error of

$$\frac{0.3892}{\sqrt{55}} = 0.0525,$$

  which is pretty close to the 0.061 obtained from all of the samples
- Bottom line: you can use the standard error of the mean from a single sample to estimate the true sample-to-sample variation of the sample mean
- That's Science with a single bag of M&Ms

The Central Limit Theorem

Recall: the single group won't be able to look at the histogram of sample proportions to see whether or not they are Normally distributed (because they only have a single sample and you can't build a histogram from that one observation)
But, the **Central Limit Theorem** says that they don't have to.
The CLT states that with a sufficient sample size and an *i.i.d* sample, the sample mean must be **Normally** distributed:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

where $\bar{X}$ is the sample mean, μ is the expected value of X and σ its standard deviation

The Central Limit Theorem, ctd.

In summary:

- It doesn't matter what the distribution of the random variable X is, so long as the sample size is large, $\bar{X}$ is always going to have a Normal distribution
- Proportions are special cases of means. They are just the mean of a set of 0/1 data, so the CLT applies to the sample proportion too
- Sample proportions will be Normally distributed for sufficient sample size, n
- We want $n\,p$ or $n\,(1-p) \geq 10$ to believe that the Central **Limit** Theorem has *kicked in*

Summary: CLT and Standard Error

The two *nice to believe facts* that student Group 2_24 needed for their inference, that:

1. Sample proportions have a Normal distribution
2. The sample-to-sample variation of the proportions is about 0.061

are justified by

1. Invoking the Central Limit Theorem
2. Estimating the sample-to-sample variability with $\frac{s}{\sqrt{n}}$, which is calculated from a single sample

This means, that student group 2_24, with their single bag of M&Ms is able to do valid inference, without having to sample multiple bags of M&Ms.

Module summary

Topics covered today include:

- The Empirical Rule
- The standard error of the mean
- The Central Limit Theorem

Next time

- Confidence Intervals
- Hypothesis testing

Aside: the **surprise** paradigm

The eventual take home:

> *On observing a rare event, doubt the assumptions under which that event was defined to be rare.*
>
> *Statisticians don't believe in miracles.*

The paradigm:

- We have a belief
- Assuming this belief is true, we calculate the probability of an event of interest
- If we observe this event and the probability is small we have reason to reject this *belief*
- In this way we **learn**

The **surprise** paradigm illustrated

Lottery:

- Belief: a lottery is *fair* – each ticket is equally likely to win.
- Event of interest: a specific ticket wins.
- Observed event: the President wins the lottery.
- Learning: don't play this lottery.

Higgs boson:

- Belief: a Higgs type particle does **not** exist.
- Event of interest: certain decay products of collision in given proportions are observed.
- Observed event: decay products are seen in the right proportions, with a very small probability (5-σ) under the non-existence hypothesis.
- Learning: Overturn the initial belief of non-existence and conclude that the data supports the existence of a Higgs type particle.