Section 12.3 (Modified) Intro to Hypothesis Testing and Degrees of Freedom

What is Hypothesis Testing?
A hypothesis test involves making a conjecture about some part of our world, collecting data and performing mathematical calculations to determine if it's true.

There are many types of hypothesis tests depending on what you are trying to examine. In this class we will focus on chi-square (X^2) test of independence. It is used to determine whether two categorical variables are dependent.

When performing a hypothesis test, you must start with two hypotheses...

H_0 is the Null Hypothesis. For us that states that either the mean of the data is **equal** to the given mean OR that two variables are **independent (unrelated).**

H_1 is the Alternate Hypothesis. For us that states that either the mean of the data is **not equal** to the given mean OR that the two variables are **not independent (related).**

Example: Sean conjectures that the average price of a new laptop is €1300 (Euros). He randomly collects data by making a few phone calls and browsing the internet. His results are below

1395, 1290, 1400, 1490, 1100, 1535, 1370, 1480, 1270, 1430

Write the Null and Alternate Hypotheses for this situation. (Note: I am only given one type of data)

Answer:
H_0: The average price of a new laptop is €1300
H_1: The average price of a new laptop is not €1300.

Example: A survey was done in order to investigate whether the favorite genre of movie goers is dependent on gender. The results are shown in the table below

	Comedy	Action	Drama
Male	18	32	6
Female	35	11	14

Write a suitable Null and Alternate Hypotheses for this situation. (Note: I am given two types of data)

Answer:
H_0: Gender and type of movie are independent.
H_1: Gender and type of movie are not independent.

After performing some mathematical calculations, we will then come to a conclusion to either accept or reject the NULL Hypothesis. We will discuss that in the next section.

Most of the following is taken from http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-degrees-of-freedom-in-statistics

One of the calculations involved in the chi-square test of independence is called degrees of freedom. Degrees of freedom aren't easy to explain. They come up in many different contexts in statistics—some advanced and complicated. In mathematics, they're technically defined as the dimension of the domain of a random vector. Degrees of freedom are generally not something you *need* to understand to perform a statistical analysis—unless you're a research statistician, or someone studying statistical theory. They are easy to find and it is something you will need to compute on the IB Exam.

Here is an example that provides a basic gist of their meaning in statistics.

First, forget about statistics. Imagine you're a fun-loving person who loves to wear hats. You couldn't care less what a degree of freedom is. You believe that variety is the spice of life.

Unfortunately, you have constraints. You have only 7 hats. Yet you want to wear a different hat every day of the week.

On the first day, you can wear any of the 7 hats. On the second day, you can choose from the 6 remaining hats, on day 3 you can choose from 5 hats, and so on. When day 6 rolls around, you still have a choice between 2 hats that you haven't worn yet that week. But after you choose your hat for day 6, you have no choice for the hat that you wear on Day 7. You *must* wear the one remaining hat. You had 7-1 = 6 days of "hat" freedom—in which the hat you wore could vary!

That's kind of the idea behind degrees of freedom in statistics. Degrees of freedom are often broadly defined as the number of "observations" (pieces of information) in the data that are free to vary when estimating statistical parameters.

In the chi-square test of independence (in the next section) the degrees of freedom are the number of cells in the two-way table of the categorical variables that can vary, given the constraints of the row and column marginal totals.

If you experimented with different sized tables, eventually you'd find a general pattern. For a table with *r* rows and *c* columns, the number of cells that can vary is (# of rows – 1)(# of columns – 1) or (r – 1)(c – 1). And that's the formula for the degrees for freedom for the chi-square test of independence!

Example: Referring to the table on genre of movies and gender, what would be the degrees of freedom for the given data?

	Comedy	Action	Drama
Male	18	32	6
Female	35	11	14

Degrees of Freedom Formula: (# of rows – 1)(# of columns – 1)

# of rows of data: 2
# of columns of data: 3

(r – 1)(c – 1)
(2 – 1)(3 – 1)
(1)(2)

Answer:
2 Degrees of Freedom