4/28/2022

ANOVA (multiple means) (13.1)
Tests of Independence/Chi-Square Tests (11.3)

ANOVA
Analysis of Variance
One-way ANOVA
It looks at the means (variances) of multiple sets of data (specifically 3 or more at a time)
If we have two sets of data to compare, we can do a t-test
But if we have more than 2 sets, we don't want to do pairwise comparisons unless we have to.
Allows us to compare all the sets to each other simultaneously.

The null hypothesis is that all the means are equal. The alternative hypothesis is that at least one mean is different from the others.

We can use a boxplot to compare as an intuition, or once the null is rejected, determine which set or sets is likely to be different from the others and would need to be tested further (in pairs).

(the data in the example is from Sheet 11 of example data sets file)

$$H_0: \mu_i = \mu_j, for\ all\ i, j$$
$$H_a: \mu_i \neq \mu_j, for\ at\ least\ one\ i \neq j$$

Practically:
H_0: the means are all the same
H_a: at least one mean is different

The statistic to be tested is called the F statistic. F distribution is skewed. Also depends on degrees of freedom related to the number of datapoints in each set of data (the number of means and the number of values). All of this is buried deep into the formulas built into Excel. The interpretation of the probabilities is the same. Assume that alpha 0.05 unless otherwise stated.

Go to Excel for Example (Sheet 2)

Anova: Single
Factor

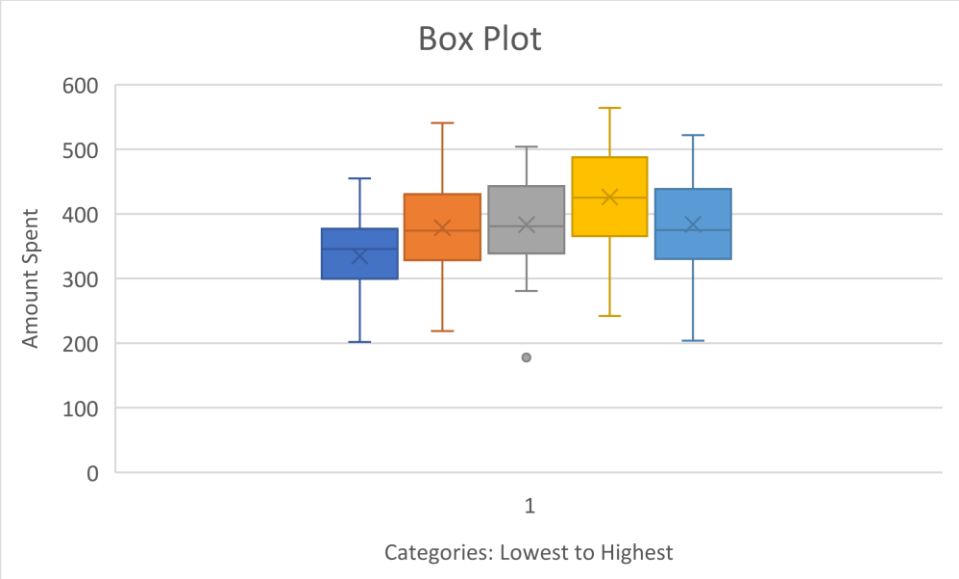
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
				3726.24333
Lowest	25	8373	334.92	3
Next-to-lowest	25	9467	378.68	7069.56
				5719.17333
Middle	25	9586	383.44	3
		1065		
Next-to-highest	25	7	426.28	7234.21

Highest	25	9597	383.88	4846.776667

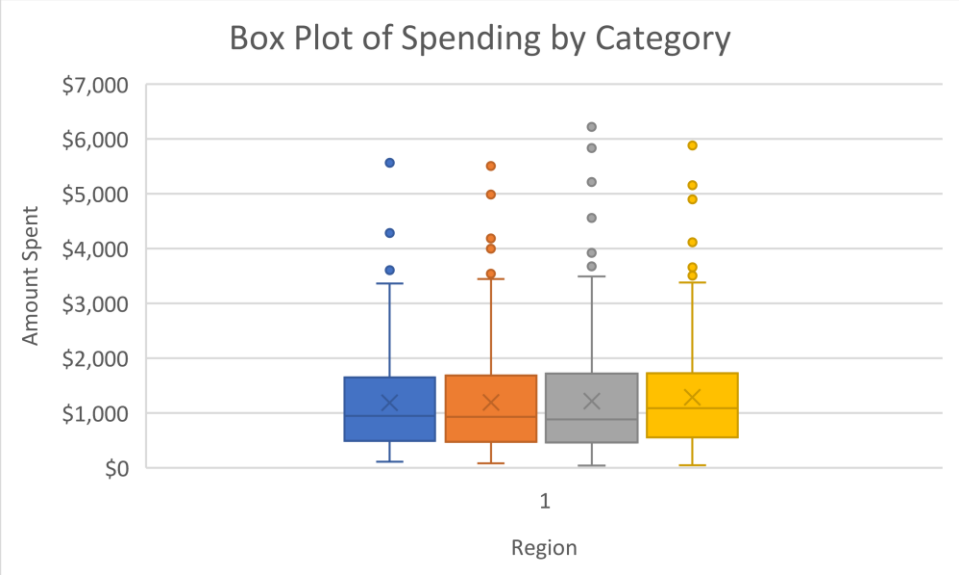
ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	104807.68	4	26201.92	4.581401874	0.001772104	2.447236511
Within Groups	686303.12	120	5719.192667			
Total	791110.8	124				



The p-value (highlighted) is less than our significance level of 0.05, and so we reject the null hypothesis. Our conclusion of the test is that at least one set of means is different from the others. At least one pair is different.
Based on the boxplot, the most likely culprit is Group 1 (Lowest) and Group 4 (Next-to-Highest).

Second Example on Sheet 4 (data taken from sheet 5 of the example datasets)



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
East	252	297952.193	1182.349972	798062.0388
MidWest	261	310828.779	1190.914862	861104.4921
South	253	307026.865	1213.544921	1095474.263
West	234	300960.026	1286.153957	947981.8606

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1602168.191	3	534056.0638	0.577458175	0.629900893	2.613839375
Within Groups	921140027.6	996	924839.3851			
Total	922742195.7	999				

Here, the p-value is very high, much bigger than 0.05 (0.6299>>0.05), so we fail to reject the null hypothesis. We conclude (in plain English) that the spending habits of the different regions are not different.

This agrees with our boxplot where the groups look very similar.

What you have to do:
You may need to reorganize the data (depending on how it is provided to you). It must be in different columns (or rows), each group in their own column.
State the null and alternative hypotheses.
Conduct the test.

Analyze the results by comparing the p-value to alpha, and stating the conclusion: reject or fail reject, and what this means in terms of the problem statement.

Only use the ANOVA when there are 3 or more groups. If you run it on two samples, if they are independent, you will get the same p-value, but you will get marked off for using the wrong test. If the samples are paired, it will be very incorrect.

Test of Independence
Usually conducted on data that is in a two-way table.
When we are doing just descriptive statistics, then for two variables to be independent we had to test whether $P(A) = P(A|B)$. This strict equality is only valid for descriptive statistics. We want to make inferences about the general population, and since we are working with a sample, with random noise, we want to look at the probability that this difference is just noise, or different enough to be a sign of real dependence.

(If you have raw data, use a pivot table to create a two-way table. And then so the formulas don't look insane, copy the data from the table as numbers.)

The null hypothesis should contain the equality statement. For independence, that is the equality statement.

H_0: the two variables are independent (not related to each other)
H_a: the two variables are dependent (they are related)

See Excel Sheet 1 for example.

Eye Color	Black	Brown	Blue	Green	Grey	Total
Female	20	30	10	15	10	85
Male	25	15	12	20	10	82
Total	45	45	22	35	20	167

Probabilities
In each category, if the data is independent, then $P(A \text{ and } B) = P(A) \times P(B)$

$$P(black \text{ and } Female) = P(black) \times P(Female) = \frac{45}{167} \times \frac{85}{167}$$

We need to convert this to a count, not a probability. So the way we do that is to multiply by the total number of observations in the table (sample size) to get the number of expected people.

$$E(black \text{ and } Female) = P(black \text{ and } Female) \times sample\ size = \frac{45}{167} \times \frac{85}{167} \times 167$$

We end up with: $\frac{is\ the\ number\ of\ people\ with\ black\ eyes \times the\ number\ of\ people\ who\ are\ female}{sample\ size\ (total)}$

For every entry in the table: multiply by the number at the bottom of the column times the number at the end of the row and then divide by the grand total.

Sometimes this test is referred to as Chi-Squared test χ^2. Has a distribution, it also skewed, but this requires one value for the degrees of freedom. Related to the size of the two-way table. Our example table is 5 by 2 (5×2). The rows are m and the columns are n, and degrees of freedom is $(m-1) \times (n-1)$, in our cases $(2-1) \times (5-1) = 1 \times 4 = 4$.

And the test statistic, it's calculated by comparing the observations with the expectations. Each entry is calculated as $\frac{(E-O)^2}{E}$, add them all up to get the test statistic, χ^2.

Excel will do all this for us, and then convert the test statistic to a probability. In our case, this is the probability that the differences between expectations and observations are due to chance alone.

Chi-Squared Test, p-value
0.171212

This p-value is greater than 0.05, and so we fail to reject the null. There is not enough evidence to conclude that eye color and gender are dependent.

Example.
On sheet 3 in Excel.
Chi-squared p-value
0.011320253
(Example from textbook)
Is the p-value less than our significance level? 0.01 < 0.05. So, can reject the null hypothesis. There is a difference between the ways CC students, 4-year college students and non-students volunteer. The two variables are dependent: there is a relationship.

The test for independence looks for relationships in categorical data (two categorical variables).
The ANOVA test looks for relationships in one categorical variable with one numerical variable.
Regression will be looking for a relationship between two numerical variables.

There are other uses for a chi-squared test. Skip "goodness of fit tests" in the homework. They depend much more on various distributions, etc., some of which we skipped, and so the set up for the table of expectations can change dramatically between problems.