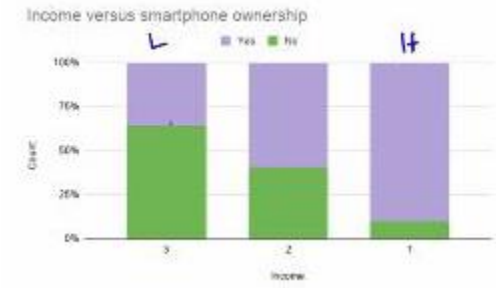
1. Contingency tables can be used to understand the association between two categorical variables (bivariate).
2. Order of the values of variables are not relevant in the case of a *nominal* variable. But, in the case of an *ordinal* variable, it's a good idea to keep the order intact.
3. In contingency table, when you populate relative frequencies on each cell using row totals, cells are said to have row-relative frequencies. Similarly, when you populate relative frequencies on each cell using column totals, cells are said to have column-relative frequencies.
4. If the row(column) frequencies are same for all rows(columns), then the two variables we're studying are not **associated** with each other.
5. If the row(column) frequencies are different for some rows(columns), then the two variables we're studying are **associated** with each other.
6. Association can be clearly visualized with a 100% stacked bar chart, where row(column) relative frequencies for each variable are stacked, and proportioned to 100%.
7. Following is an example that plots 100% stacked bar chart of row-relative frequencies for Income and Smartphone ownership.



It's evident from the above graph that in high income groups there're more smartphone owners, than in low income groups. Hence, we conclude that smartphone ownership is associated with Income.

Raw data for the above graph is given in the below row-relative frequency table.

	Own a smartphone		
Income level	**No**	**Yes**	**Row total**
High	2/20	18/20	**20**
Medium	27/66	39/66	**66**
Low	9/14	5/14	**14**
Column total	**38/100**	**62/100**	**100**

Below table represents the row-relative frequencies for above raw data. This table has been used to plot the 100% stacked-bar chart given above.

	Own a smartphone		
Income level	No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

8. Use scatter plots to study association between two numerical variables.

9. Scatter plots are graphs that display pairs of values (explanatory, response) as points on a 2-dimensional plane. Explanatory variable is typically plotted along X-axis and response variable along Y-axis.
10. Any patterns like upward/downward trend, linear/curved nature in the scatter plot, will reveal an association between the variables. Scatter plot will also reveal variability of the data (identification of clustering of data) and outliers if any.
11. Covariance and correlation are two measures used to quantify the *linear* association between two numerical variables.
12. If x_i and y_i denote the i-th observation of variable x and y respectively, and (x_i, y_i) be the i-th paired observation of a population (sample) dataset having N(n) observations, then covariance between the variables x and y is given by

- *Population covariance:* $Cov(x,y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N}$
- *Sample covariance:* $Cov(x,y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

13. When large (small) values of x tend to be associated with large(small) values of y, the signs of the deviations (x - x_i) and (y - y_i) tend to be same.
14. When large (small) values of x tend to be associated with small (large) values of y, the signs of the deviations (x - x_i) and (y - y_i) tend to be different.
15. In the case of a positive association (upward trend), *covariance* is positive. In the case of negative association (downward trend), *covariance* is negative.
16. *Covariance* is difficult to interpret because it has units.
17. A more easily interpreted measure of linear association between two numerical variables is *correlation*, which is derived from *covariance*.
18. *Correlation* between two variables x and y is defined as the covariance between the variables divided by the product of the std.deviations of both.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{s_x s_y}$$

19. *Correlation* is a unitless measure between -1 and +1. The std.deviation of the variables cancel out the units of the numerator values.
20. *Correlation* is affected by outliers.
21. The linear association between two variables can also be described using the equation of a line of the form y = mx + c.

22. Best line fit between a set of pairs of (x, y) will have it slope calculated using

$$m = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

where n is the number of pairs, $\bar{x}$ and $\bar{y}$ are mean value of x-values and y-values respectively.

(Reference: https://www.varsitytutors.com/hotmath/hotmath_help/topics/line-of-best-fit)

23. Square of the correlation coefficient is referred to as R^2 and is the *goodness to fit* measure, and takes values between 0 and 1. R^2 is a positive value irrespective of whether the correlation is negative (typically downward slope) or positive (typically upward slope).
24. R^2 helps capture the proportion of variability in a data set that. It is very low, if R^2 is closer to 0. And it is pretty high, if the R^2=1.
25. If you've to measure the association between a categorical (assuming only two levels; also called dichotomous) and numerical variable, we use a concept called bi-serial correlation measure. To start with, code the levels of the categorical variable. Thus, code 'Male' as 0 and 'Female' as 1 in a gender-wise dataset. Now, scatter-plot the variables. If the cluster for 'Male' is visually different from that for 'Female', correlation is low.
26. In order to quantify the correlation of a categorical and numerical variable, point bi-serial correlation coefficient can be used. Following steps are used to calculate it.

- Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).
- The following steps are used for calculating the Point Bi-serial correlation between these two variables:

Step 1 Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1.

Step 2 Calculate the mean values of two groups: Let $\bar{Y}_0$ and $\bar{Y}_1$ be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.

Step 3 Let p_0 and p_1 be the proportion of observations in a group with $Y = 0$ and $Y = 1$, respectively, and s_X be the standard deviation of the random variable X.

The correlation coefficient

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{s_X}\right)\sqrt{p_0 p_1}$$

27. The PBS coefficient is between 0 and 1. If the value is closer to 0, the variables don't have any association. Closer to 1, the variables are tightly associated.
28. How data manipulation affects statistical measures?

	Adding constant (+C)	Multiplying constant (*C)	Outliers
Mean	+C	* C	Affected
Median	+C	* C	Unaffected
Mode	+C	* C	Unaffected
Range	Unaffected	* C	Affected
IQR	+ C	* C	Unaffected
Variance	Unaffected	* C^2	Affected
Standard deviation	Unaffected	* C	Affected
Covariance	Unaffected	* C^2	Affected
Correlation Coefficient	Unaffected	Unaffected	Unaffected