**12.3 Data Exploration – Overview**

Chart, scatter chart

Description automatically generated**Scatterplots**

* Displays the relationship between **two quantitative** variables measured on the same individuals.
* Useful to determine if an **association** exists!
  + So is there a pattern where *some values of one variable* tend to occur with *some values of the other variable*.
  + Ex) Smaller carat diamonds tend to have lower prices, and as the carat increases prices tend to increase as well.
* Setup of axes
  + The explanatory (independent) variable goes on the X (horizontal) axis.
  + The response (dependent) variable goes on the Y (vertical) axis.
  + Ex) How large a diamond is impacts how much it costs 🡪 Carat = *X*; Price = *Y*.
* Interpreting a scatterplot (what we are looking for in a scatterplot)
  + **Form** (pattern of the dots)
    - Linear 🡪 Points follow a general linear trend; Straight line.
    - Curved 🡪 Points show some evidence of curvature; NOT a straight line.
    - Randomscatter 🡪 No pattern, points are just scattered about randomly kinda like a cloud of points.
  + **Direction** (of the association; only applies to linear relationships)
    - Positive 🡪 Upward trend.
    - Negative 🡪 Downward trend.
    - No Association 🡪There is no pattern or general trend (corresponds to random scatter).
  + **Strength** (how strong the association is; how well the data fits the pattern; only applying this to linear relationships)

Chart, scatter chart

Description automatically generated

Example



**Correlation**

* The **correlation (*r*)** is an index that expresses the direction and strength of the relationship.
  + It combines both of these aspects into a single number measure.
  + Often referred to as the correlation coefficient (or Pearson’s correlation).

Calendar

Description automatically generated

* Interpreting correlation
  + Sign = Direction
  + Absolute value |*r*| = Strength
* Properties of Correlation
  + Timeline

    Description automatically generatedScale goes from -1 to 1 →
  + Only applies to LINEAR relationships.
  + *r* has no units and is the same regardless of which variable is *X* or *Y*.
  + Does NOT imply a cause-and-effect relationship.

Ex) Ice cream sales and shark attacks have a strong positive correlation.

**Using your Calculator!**

Using TI-83/84 (and TI-30 XS MultiView / XIIS) to calculate correlation (and regression line).

Steps for the TI-83/84

1. Enter data: STAT → Edit →

Enter X data in L1

Enter Y data in L2

2. Calculate: STAT → CALC → LinReg(ax+b)

* 1. XList: L1.
  2. YList: L2.
  3. Rest leave blank.
  4. Calculate!

Steps for the TI-30 XIIS

1. 2nd 🡪 STAT 🡪 1-VAR (Enter)

2. DATA

X1 = # (scroll down)

FRQ = 1 (for ALL Xs, scroll down)

X2 = # (scroll down)

…

3. STATVAR (scroll across)

4. To exit this menu: 2nd 🡪 EXIT STAT 🡪 Y

b) FRQ: L2

c) CALC

Steps for the TI-30 XIIS

1. 2nd 🡪 STAT 🡪 2-VAR (Enter)

2. DATA

X1 = # (scroll down)

Y1 = # (scroll down)

… (repeat for all data points)

3. STATVAR (scroll across)

4. To exit this menu: 2nd 🡪 EXIT STAT 🡪 Y

Steps for the TI-30XS MultiView

1. Data 🡪

Enter X data in L1

Enter Y data in L2

2. 2nd 🡪 stat 🡪 2-Var Stats

a) *x*DATA: L1

b) *y*DATA: L2

c) CALC

*\*\*\* One time setup for TI-83/84: 2nd → Catalog → DiagnosticOn → Enter*

Graphical user interface

Description automatically generated

**Regression**

* Ultimately, we want to determine if we can use a straight line to model the relationship between two variables 🡪 If so, we can use that model to make predictions!
  + This process is called **Linear Regression**.

Chart, scatter chart

Description automatically generated

Step 1 🡪 Determine if there is a **significant correlation** (**linear relationship**).

1. Calculate correlation.
2. Compare it to the **Table of Critical Values or the Pearson Correlation Coefficient** to see if it is statistically significant.
   * Table

     Description automatically generated with medium confidenceMatch the sample size *n* and the Level of significance α (Probability our claims about the data are wrong) to the specific problem.
   * If |*r*| > Critical Value (CV) 🡪 *r* is statistically significant (unlikely to have occurred by chance).

Step 2 🡪 Once we have a significant correlation, we can find the **regression line**.

* Linear equation that fits our data best (aka ‘line of best fit’).
  + It is IMPORTANT to get the *X* and *Y* variables correct!
* Our calculator gives us our equation!

Demo ex)



Chart, scatter chart

Description automatically generated

Step 3 🡪 Make **predictions** using the regression line.

* We can think of our regression line, and specifically , as predicted values of *Y* for all *X* values *in the X range of our sample data*!
* Calculating these is simple:
  + Just plug in the new *X* value to our equation and this will give us the predicted *Y*.
  + Demo example) Predict *Y* for *X* = 7.

Full Example



1. Calculate the correlation for the dataset above and determine if it is statistically significant at a level of significance of .
2. If appropriate, determine the regression equation.
3. If a student spends 35 hours on homework, make a prediction for their grade on the test.
4. If a student spends 50 hours on homework, make a prediction for their grade on the test.