# MATH 201: Lecture 3a Handout

## Section 2.1 Examining numerical data

Name:_____ Date:_____

### Learning goals for today

By the end of this lecture, you should be able to:

- Describe the shape, center, spread, and unusual features of a numerical distribution
- Choose an appropriate graph for numerical data
- Explain how outliers and skewness affect numerical summaries
- Decide when to use robust statistics vs non-robust

We will start by using a dataset containing a a random sample of 50 emails, `email50`.

### Scatterplots & Relationships

Before analyzing a single variable, we often look at how two numerical variables interact. A **Scatterplot** provides a case-by-case view of data for two numerical variables.

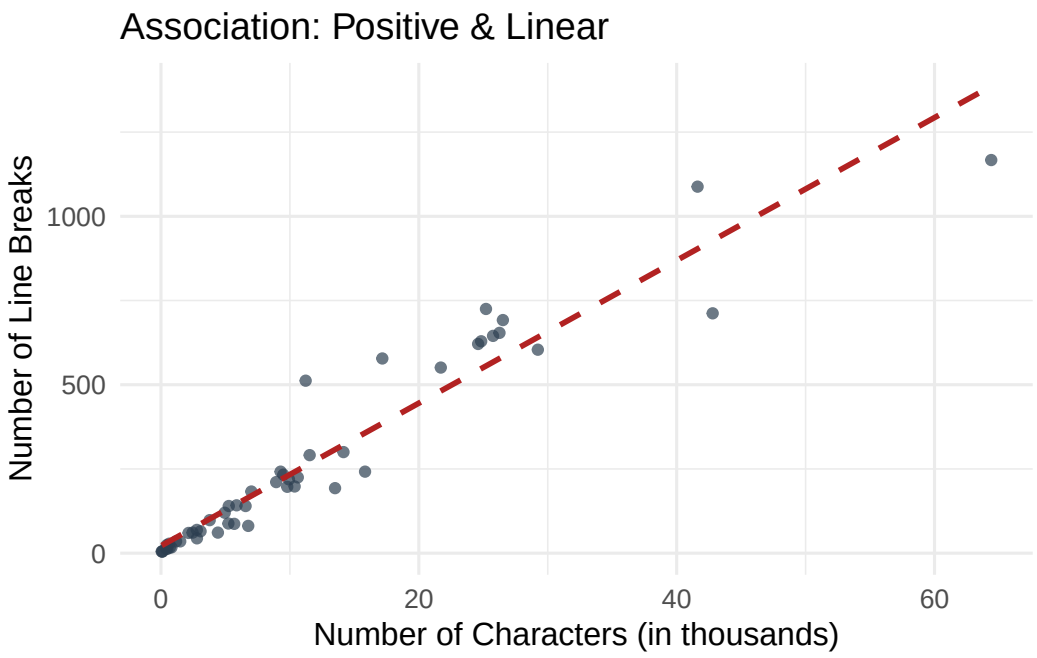



Figure 1: Scatterplot of Line Breaks vs. Character Count in the email50 dataset.

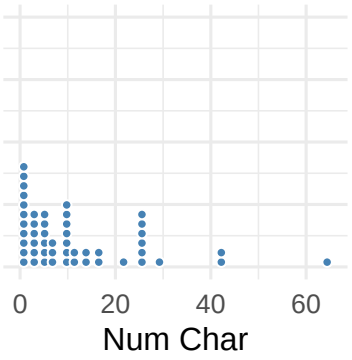
### Key Concepts

- **Association:**
  - **Positive:** As $x$ increases, $y$ tends to increase.
  - **Negative:** As $x$ increases, $y$ tends to decrease.
- **Linearity:** Does the trend follow a straight line (linear) or a curve (nonlinear)?

> ⚠ Key Limitation: Correlation $\neq$ Causation
>
> An association in a scatterplot does not imply that changing $x$ causes a change in $y$.

## Visualizing a single variable

Looking at the **dot plot**, where each dot is a datapoint, of number of characters in an email:




- Where do most values seem to fall?
- Are there values that seem unusual/extreme relative to the rest of the data?
- Would you say there are multiple peaks to the data? If so, how many?

We will learn some vocabulary to help us describe data. There are a lot of software and AI tools available to help us compute summary statistics, so in this class we will focus on understanding them.

### Center

- **Mean:** The arithmetic average = sum your data and divide by the number of data points: $\bar{x} = (x_1 + x_2 + ... + x_n)/n$
- **Median:** The midpoint of your data.

The mean number of characters per email is 11.6 and the median is 6.89. Why do you think these differ?

### Spread (Variability)

- **Standard Deviation (SD):** "Average" distance from the mean.
- **IQR:** Range of the middle 50% of data.

For number of characters per email the standard deviation and IQR are 13.13 and 12.88, respectively. In the plot we see the maximum number of characters in an email is around 64. If it were instead 600, how would the standard deviation and IQR change, if at all?

### Shape (Skewness & Modality)

- **Right Skewed:** Tail extends to the right (positive). Typical of income, house prices. $(Mean > Median)$.
- **Left Skewed:** Tail extends to the left (negative). Typical of easy test scores. $(Mean < Median)$.
- **Symmetric:** Data mirrors around the center.
- **Modality:** Unimodal (1 peak), Bimodal (2 peaks), Multimodal (many peaks).

How would you describe the shape of the distirbution of number of characters per email?

## Plot options for a single numeric variable

| Feature | Dot Plot | Histogram | Box Plot |
|---|---|---|---|
| **Best Use** | Small datasets ($n < 50$) | Large datasets ($n > 50$) | Comparing groups & spotting **outliers** |
| **Visual** | Shows **exact** values (dots). | Groups values into **bins** (bars). | Shows **quartiles** and whiskers. |
| **Concept** | Preserves individual data points. | Shows **data density** (Shape). | Shows **spread** (IQR) & **Median**. |

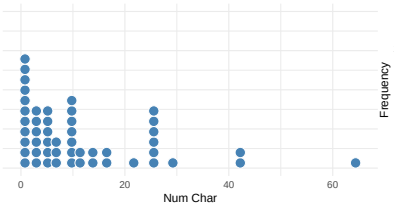



Figure 2: Dot Plot

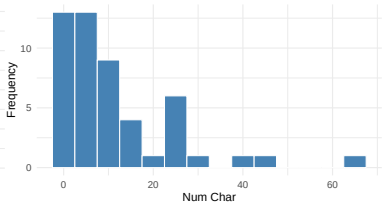



Figure 3: Histogram

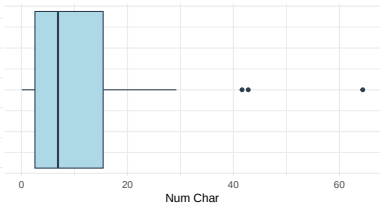



Figure 4: Box Plot

An **outlier** is an observation that appears extreme relative to the rest of the data. What is an outlier in the number of characters data?

**Example: Exam scores**

Let's consider another example. Imagine we had exam scores for a test that turned out to be really easy.

0, 78, 82, 85, 88, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99

Create a dot plot for this data.

What is the median exam score? Approximately what do you think the mean exam score is?

Do you think the standard deviation or IQR for exam scores will be larger? Why?

Sketch a box plot for this data.