STAT 305: Lecture 4

Amin Shirazi

2019-09-02

STAT 305: Lecture 4

Chapter 3

Elementary Descriptive Statistics

Course page:
ashirazist.github.io/stat305.github.io

Section 3.1

Elementary Graphical and Tabular Treatment

of

Quantitative Data

Summarizing Univariate Data

Introduction: Creative Writing Workshops

Two methods of teaching a creative writing workshop are being studied for their effectiveness of improving writing skills. First, two groups of creative writing students who were randomly assigned to one of two different 3-hour workshops. At the end of the workshop, the students were given a standard creative writing test and their score on the test was recorded.

Exam Scores for Two Groups of Students Following Different Courses

Group 1	Group 2
74 79 77 81	65 77 78 74
68 79 81 76	76 73 71 71
81 80 80 78	86 81 76 89
88 83 79 91	79 78 77 76
79 75 74 73	72 76 75 79

Intro

Exam Scores for Two Groups of Students Following Different Courses

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

We may have several questions we are interested in answering using this data. For instance,

- Which group did better on average?
- Which group has the most consistent scores?
- Were there any unusually low or high scores in either group?
- If we ignore unusual scores, which group is better?
- Which group had the most scores over 80?
- ...

However, none of these are immediately clear looking at the raw recorded data.

Summarizing The Purpose of Summaries

Intro

Certain questions can and should be asked across many types of experiments.

Purpose

But seeing data in this kind of *flat* format presents lots of problems for a person trying to understand the relationship between the two groups.

Summaries are tools (mainly mathematical or graphical) which help researchers quickly understand the data they have collected.

The purpose of a summary is to faithfully present aspects of the data in such a way that

- we are capable of answering the types of core questions about the data asked on the previous page,
- we are able to identify more complicated aspects of the data that we may want to investigate further.

Key Idea: Good summaries should be quickly interpreted, provide clear insight into the data, and be widely applicable.

Summarizing Simple Graphical Summaries

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

Simple graphical summaries aim to provide a better view of the entire set of data. The best graphs are able to make important points clearly and give valuable insights with closer study. Producing good graphs is an art.

Two common graphical summaries

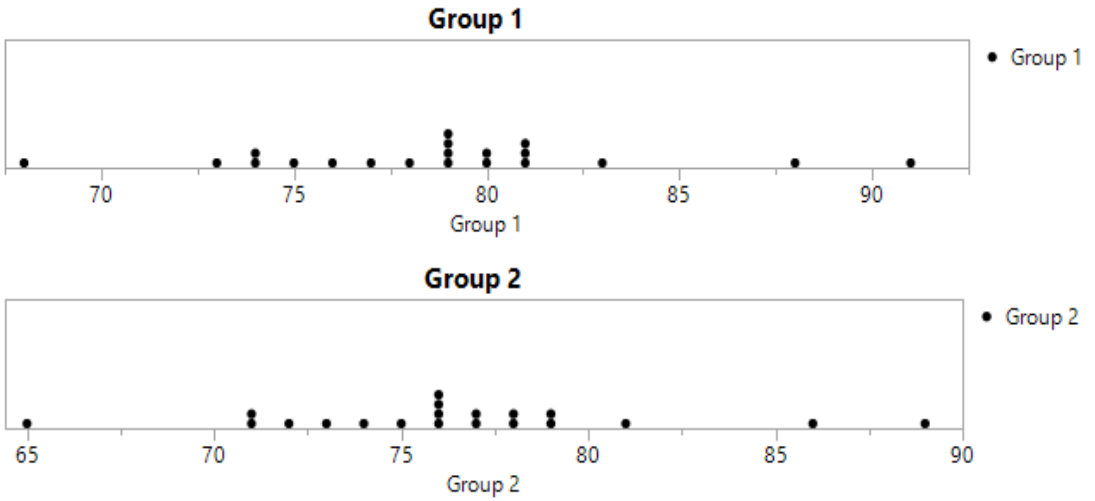
- Dot Diagrams
- Stem and Leaf Diagrams

  Carries much the same visual information as a dot diagram while preserving the original values exactly

Summarizing Simple Graphical Summaries

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

Dot Diagrams



Summarizing Simple Graphical Summaries

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

Stem and Leaf Diagrams

Stem and Leaf

Stem	Leaf	Count
9	1	1
8	8	1
8		
8		
8	3	1
8	00111	5
7	89999	5
7	67	2
7	445	3
7	3	1
7		
6	8	1

6|8 represents 68

Stem and Leaf

Stem	Leaf	Count
8	9	1
8	6	1
8		
8		
8	1	1
7	8899	4
7	666677	6
7	45	2
7	23	2
7	11	2
6		
6		
6	5	1

6|5 represents 65

Summarizing Frequency Tables

Dot diagrams and stem-and-leaf plots are useful devices when analyazing a data set, but not commonly used in presentations and reports. In such more formal contexts, **frequency tables** and **histograms** are more often used.

A frequency table is made by

- First breaking an interval containing all the data into an apropriate number of smaller intervals of **equal length.**
- Then tally marks can be recorded to indicate the number of data points falling into each interval.
- Finally, add frequency, relative frequency and cumlative relative frequency can be added.

Summarizing Frequency Tables

- **Class**: A grouping of the observations
- **Frequency**: The number of observations in a class
- **Relative Frequency**: The proportion of the observations in the class
- **Cumulative Relative Frequency**: The proportion of observations in the current class or a previous class.

Table 3.2
Frequency Table for Laid Gear Thrust Face Runouts

Runout (.0001 in.)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
5–8	III	3	.079	.079
9–12	𝍸 𝍸 𝍸 III	18	.474	.553
13–16	𝍸 𝍸 II	12	.316	.868
17–20	IIII	4	.105	.974
21–24		0	0	.974
25–28	I	1	.026	1.000
		38	1.000	

Summarizing Histograms

After making a frequency table, it is common to use the organization provided by the table to create a **histogram**.

A **histogram** is essentially a graphical representation of a frequency table.

Tips for useful frequency tables

1. Use equal class intervals
2. When the goal is to compare multiple groups, use uniform scales on each graph (i.e., keep lengths consistent)
3. Show the entire vertical axis (especially for relative frequency histograms)

Summarizing Histograms

Purpose

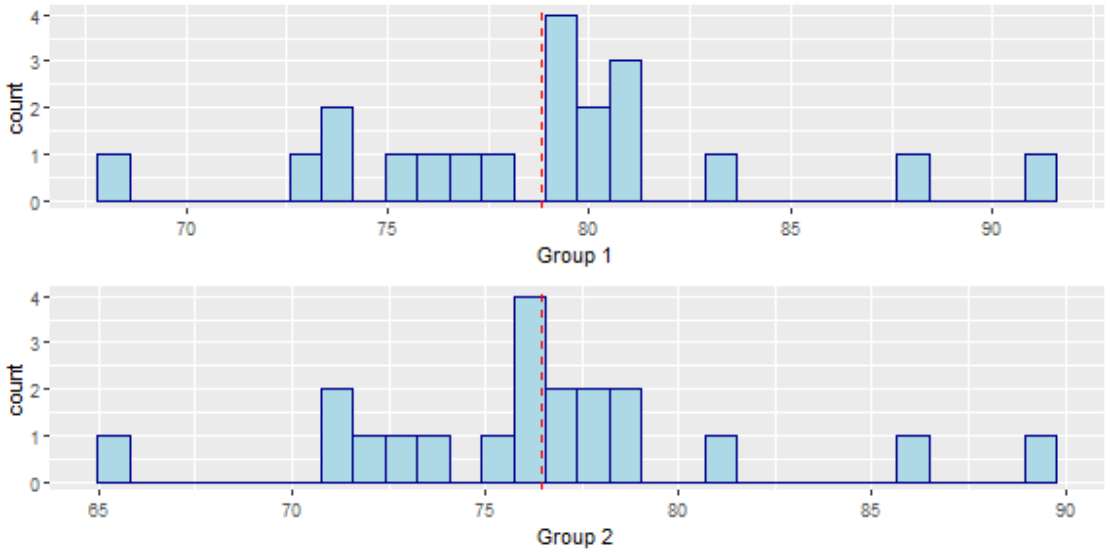
Simple Graphs

Freq Tables

Histograms

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

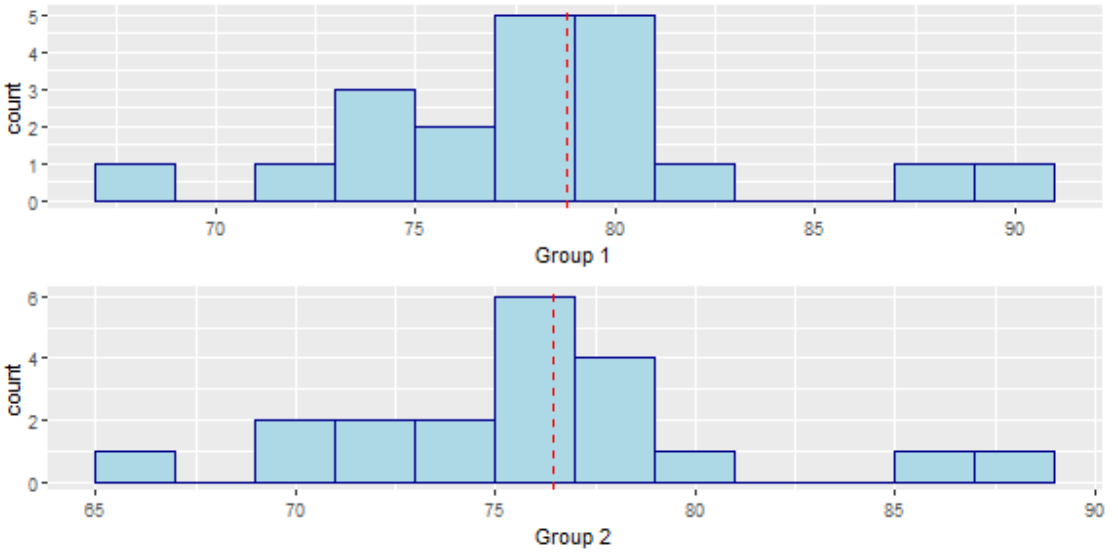
Unit interval



Summarizing Histograms

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

Interval of length two



Summarizing Summaries of Location and Central Tendency

Motivated by asking what is *normal/common/expected* for this data. There are three main types used:

Mean: A "fair" center value. The symbol used differs depending on whether we are dealing with a sample or population:

		Mean
Population		$\mu = \sum_1^N x_i$
Sample		$\bar{x} = \sum_1^n x_i$

Mode: The most commonly occurring data value in set.

Quantiles: The number that divides our data values so that the proportion, p, of the data values are below the number and the proportion $1 - p$ are above the number.

Median: The value dividing the data values in half (the middle of the values). The median is just the 50th quantile.

Summarizing Summaries of Location and Central Tendency

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

Calculating Mean Think of it as an equal division of the total

- each value in the data is an "x_i" (i is a **subscript**)
- Group 1: $x_1 = 74, x_2 = 79, \ldots, x_{20} = 73$
- The sum: $x_1 + x_2 + x_3 + \ldots + x_{20}$
- divides : $(x_1 + x_2 + x_3 + \ldots + x_{20})/20$
- Or using summation notation: $\frac{1}{20} \sum_{i=1}^{20} x_i$

The Quantile Function

Two useful pieces of notation:

floor: $\lfloor x \rfloor$ is the largest integer smaller than or equal to x

ceiling: $\lceil x \rceil$ is the smallest integer larger than or equal to x

Examples

- $\lfloor 55.2 \rfloor = 55$
- $\lceil 55.2 \rceil = 56$
- $\lfloor 19 \rfloor = 19$
- $\lceil 19 \rceil = 19$
- $\lceil -3.2 \rceil = -3$
- $\lfloor -3.2 \rfloor = -4$

The Quantile Function

For a data set consisting of n values that when ordered are $x_1 \leq x_2 \leq \ldots \leq x_n$ and $0 \leq p \leq 1$. We define the **quantile function** $Q(p)$ as:

$$Q(p) = \begin{cases} x_i & \lfloor n \cdot p + .5 \rfloor = n \cdot p + \\ x_i + (np - i + .5)\,(x_{i+1} - x_i) & \lfloor n \cdot p + .5 \rfloor \neq n \cdot p + \end{cases}$$

(note: this is the definition used in the book - it's just written using *floor* and *ceiling* instead of in words)

Example: Find the median, first quartile, 17th quantile and 65th quantile for the following set of data values:

58, 76, 66, 61, 50, 77, 67, 64, 41, 61

Summarizing Summaries of Variablity (or "Spread")

Motivated by asking what kind of *variability* is seen in our data or *how spread out* our observed values are.

- **Range**: the total distance the data values are spread across
- **Interquartile Range (IQR)**: the distance *the middle of data values* are spread across.
- **Variance** and **standard deviation**: measures for average distance from the center. Calculation differs depending on whether we have a population or sample:

	Variance	Standard Deviat
Population	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i -$
Sample	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i$

Summarizing Boxplots

Group 1				Group 2			
74	79	77	81	65	77	78	74
68	79	81	76	76	73	71	71
81	80	80	78	86	81	76	89
88	83	79	91	79	78	77	76
79	75	74	73	72	76	75	79

A boxplot can be used to summarize the values of a single quantitative variable. It does this by making use of both many of the statistics we have discussed up to this point. It depicts both

- spread (with IQR, range, etc.)

and

- location statistics (min, median, max, etc.)

Quantile Plots:

Scatterplots using quatiles and their corresponding values

For each x_i in the data set, we plot $\left(\frac{i-.5}{n}, x_i\right)$ - meaning we are plotting $(p, Q(p))$. We connect the points with a straight line, which follows the values of $Q(p)$ exactly.

Consider the sample: 13, 15, 18, 19, 21, 34, 35, 35, 36, 39. The following table which helps create the plot:

	1	2	3	4	5	6	7	8	9	10
p										
$Q(p)$										

Quantile-Quantile Plots:

QQ plots are created by plotting the values of $Q(p)$ for a data set against values of $Q(p)$ coming from some other source.

- Empirical QQ plots: the other source are quantiles from another actual data set.
- Theoretical QQ plots: the other source are quantiles from a theoretical set - we know the quantiles without having any data.

Example

- Set 1: 36, 15, 35, 34, 18, 13, 19, 21, 39, 35
- Set 2: 37, 39, 79, 31, 69, 71, 43, 27, 73, 71

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
p										
Set 1 $Q(p)$										
Set 2 $Q(p)$										

Quantile-Quantile Plots:

The resulting plot shows some kind of linear pattern - this means that the quantiles increase at the same rate, even if the sizes of the values themselves are very different.