Chapter 3

ST 511 - Descriptive Statistics

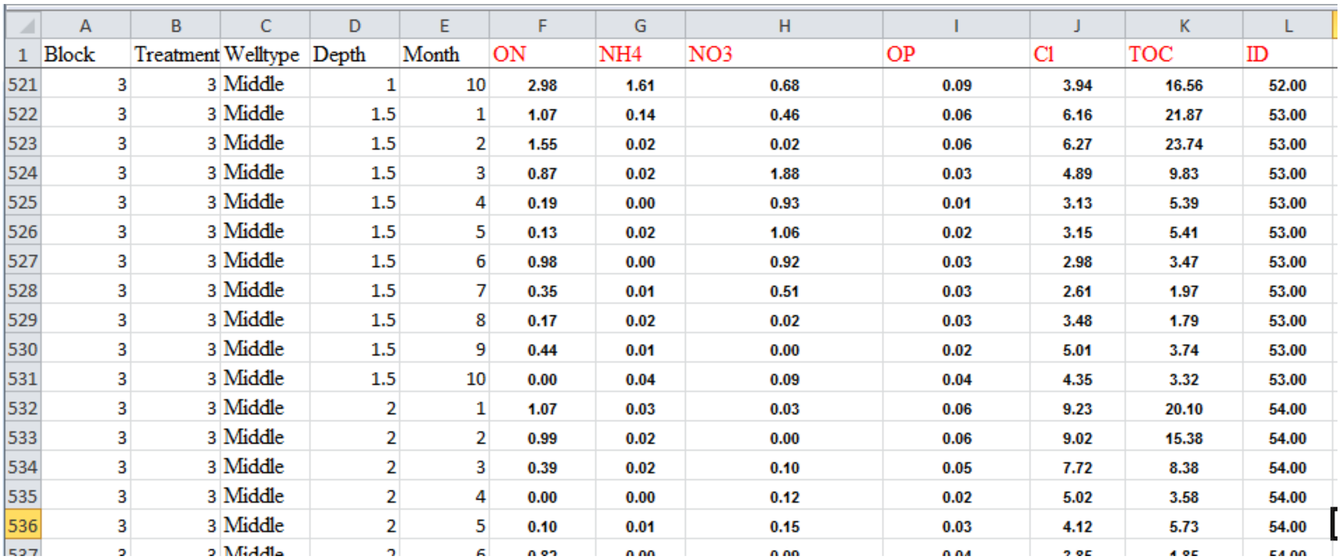
Readings: Chapter 3 (you can skip the guidelines for constructing class intervals, stem-and-leaf plots, grouped mean/median)

Recall: Process of a study involves

1. Identify the research objective
2. Collect the information needed to answer the questions
3. Organize and summarize the information.
4. Draw conclusions from the information.

We will now talk about step 3!

So you have data... now what??



	A	B	C	D	E	F	G	H	I	J	K	L
1	Block	Treatment	Welltype	Depth	Month	ON	NH4	NO3	OP	Cl	TOC	ID
521	3	3	Middle	1	10	2.98	1.61	0.68	0.09	3.94	16.56	52.00
522	3	3	Middle	1.5	1	1.07	0.14	0.46	0.06	6.16	21.87	53.00
523	3	3	Middle	1.5	2	1.55	0.02	0.02	0.06	6.27	23.74	53.00
524	3	3	Middle	1.5	3	0.87	0.02	1.88	0.03	4.89	9.83	53.00
525	3	3	Middle	1.5	4	0.19	0.00	0.93	0.01	3.13	5.39	53.00
526	3	3	Middle	1.5	5	0.13	0.02	1.06	0.02	3.15	5.41	53.00
527	3	3	Middle	1.5	6	0.98	0.00	0.92	0.03	2.98	3.47	53.00
528	3	3	Middle	1.5	7	0.35	0.01	0.51	0.03	2.61	1.97	53.00
529	3	3	Middle	1.5	8	0.17	0.02	0.02	0.03	3.48	1.79	53.00
530	3	3	Middle	1.5	9	0.44	0.01	0.00	0.02	5.01	3.74	53.00
531	3	3	Middle	1.5	10	0.00	0.04	0.09	0.04	4.35	3.32	53.00
532	3	3	Middle	2	1	1.07	0.03	0.03	0.06	9.23	20.10	54.00
533	3	3	Middle	2	2	0.99	0.02	0.00	0.06	9.02	15.38	54.00
534	3	3	Middle	2	3	0.39	0.02	0.10	0.05	7.72	8.38	54.00
535	3	3	Middle	2	4	0.00	0.00	0.12	0.02	5.02	3.58	54.00
536	3	3	Middle	2	5	0.10	0.01	0.15	0.03	4.12	5.73	54.00
537	3	3	Middle	2	6	0.82	0.00	0.00	0.04	3.85	4.85	54.00

Whether we are describing an observed population or using sampled data to draw an inference from the sample to the population, an insightful description of the data is an important step in drawing conclusions from it.

Good descriptive statistics enable us to make sense of the data by reducing a large set of measurements to a few summary measures that provide a good, rough picture of the original measurements.

Summary measure used for a variable depends on its ________________________.

Our goal will be to describe the variable's ________________________

i.e. the

Two major characteristics of the variable's distribution that we often describe are ________________________

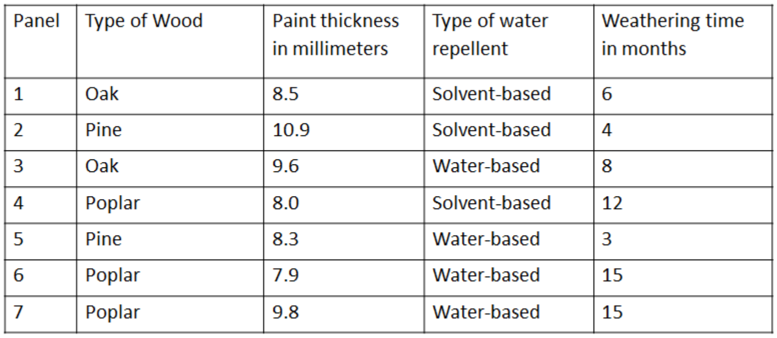
and ________________________

We will mostly deal with quantitative variables and our focus will be on their summary measures. However, we will briefly talk about graphs and statistics for categorical variables.

Categorical Variables

Numerical measure used for categorical variable:

For this simple study, we can find the sample proportion for each categorical variable:

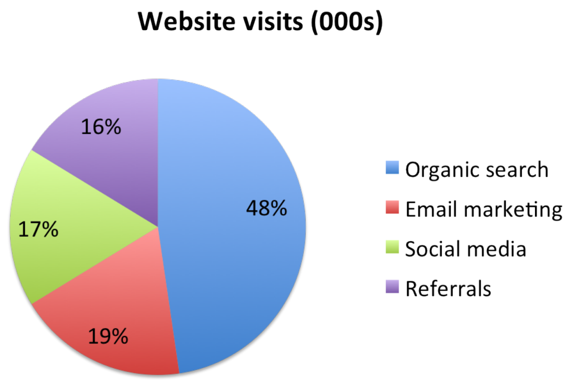


Panel	Type of Wood	Paint thickness in millimeters	Type of water repellent	Weathering time in months
1	Oak	8.5	Solvent-based	6
2	Pine	10.9	Solvent-based	4
3	Oak	9.6	Water-based	8
4	Poplar	8.0	Solvent-based	12
5	Pine	8.3	Water-based	3
6	Poplar	7.9	Water-based	15
7	Poplar	9.8	Water-based	15

The main plots used are ________________________ and ________________________.

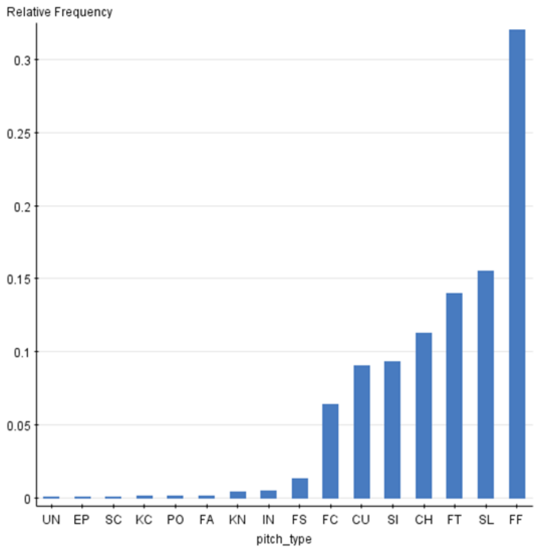
________________________ - use percents to display data.

- Order does not matter (although should order from highest percentage to lowest)



________________________ - Categories along the x-axis. Count, percent, or relative frequency (sample proportion) along the y-axis.

- Order does not matter (although should order from highest percentage to lowest).
- A gap should exist between the bars.



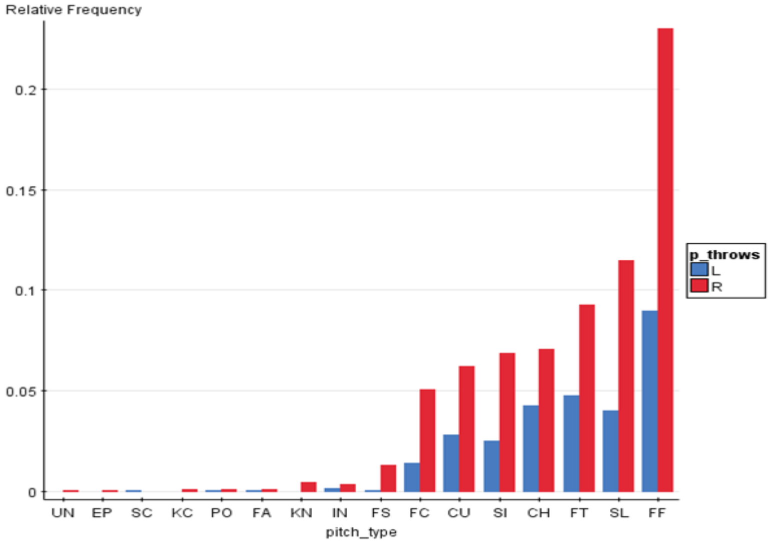
We might also have multiple categorical variables of interest. In which case, a good display of the data is

a ___________________________________.

Let's create one for the paint example from earlier.

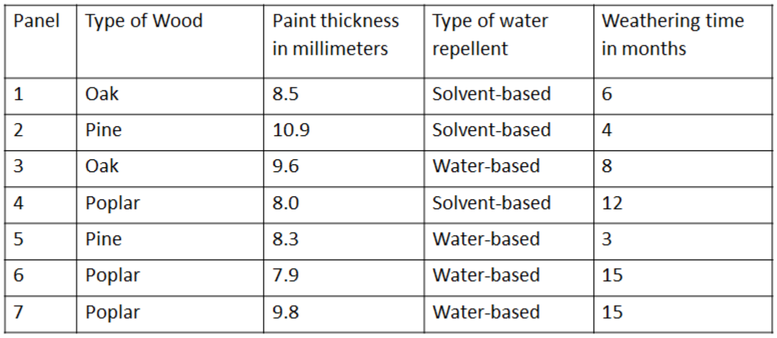
Book example on page 103 is nice.

Many other methods exist such as comparative bar charts:



Quantitative Variables

We will again consider the paint example:



Panel	Type of Wood	Paint thickness in millimeters	Type of water repellent	Weathering time in months
1	Oak	8.5	Solvent-based	6
2	Pine	10.9	Solvent-based	4
3	Oak	9.6	Water-based	8
4	Poplar	8.0	Solvent-based	12
5	Pine	8.3	Water-based	3
6	Poplar	7.9	Water-based	15
7	Poplar	9.8	Water-based	15

Numerical measures of location:

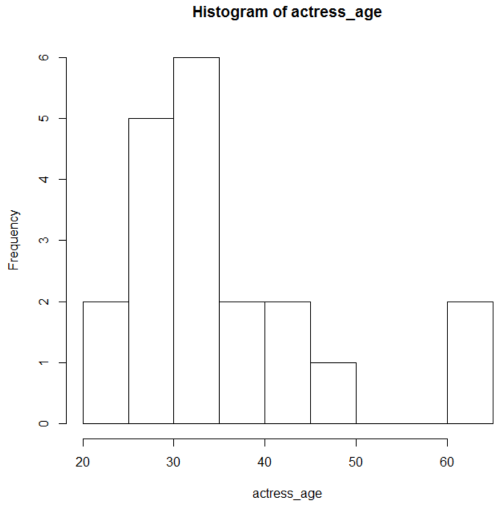
Numerical Measures of Spread

The main plots used are ________________________ and ________________________.

A ________________________ is obtained by splitting the range of the data into equal-sized bins. Then for each bin, we count the number of points that fall into each bin and that is the height of our bar (or use relative frequency - i.e. proportion in category).

- Typically, an observation equal to a boundary value is put in the higher interval.
- Bars should touch!
- Too many classes will spread the data out, thereby not revealing the pattern. Too few classes will lump the data.
- **** This is the most important graphical technique for displaying the distribution of a quantitative variable!

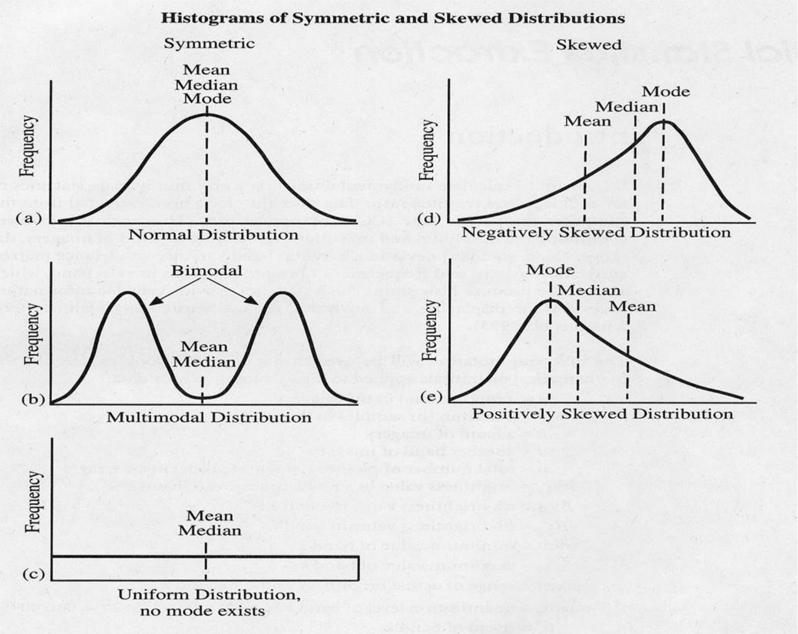
Ex. Ages of the winners of the best actress Academy Award in the recent 20 years (1994-2013) are: 36, 45, 49, 39, 34, 26, 25, 33, 35, 35, 28, 30, 29, 61, 32, 33, 45,29, 62 and 22



What are we looking for in a histogram?

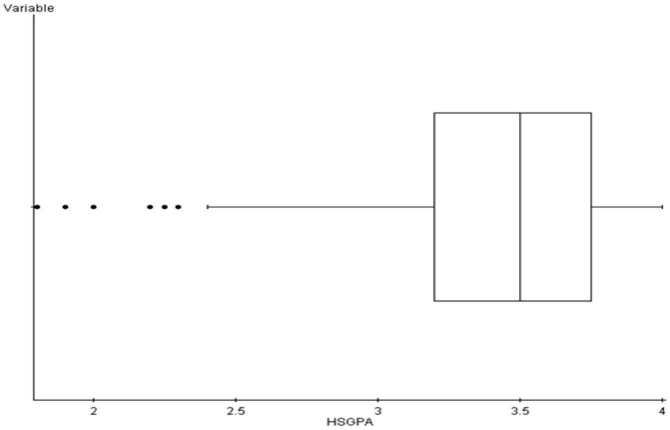
-
-
-

Relationship between mean and median for a histogram (note pictures use smooth curves, but same ideas hold):



A _______________________ displays the five number summary of the data.

Five number summary includes:



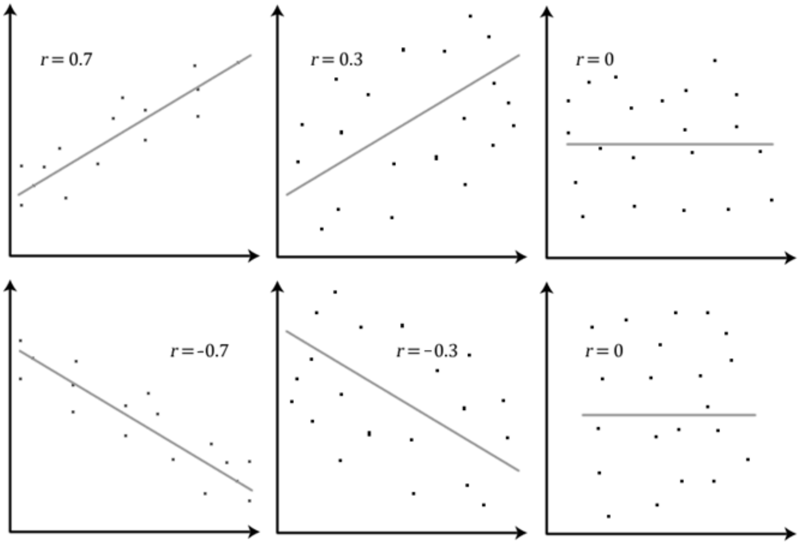
- Measure of center from a boxplot -

- Measures of spread from a boxplot -

- Can tell skewness but not modality!

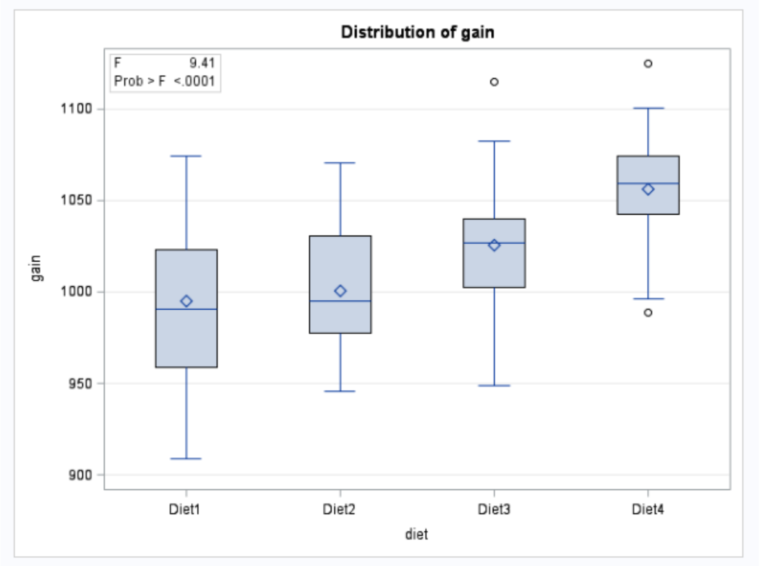
If we have two quantitative variables of interest, we often look at ____________________________________

and ____________________________________ to inspect the 'linear association' between the variables (call them x and y).



We will look at these more later in the course.

If we have a quantitative and a categorical variable, we often look at ____________________________________.



Review -
Numeric Summaries of Location: Mean/median/trimmed mean (quantitative), proportion (qualitative)
Numeric Summaries of Spread: Variance, SD, IQR, CV, Quartiles (quantitative)
Graphical Summaries for categorical: Bar Chart, Pie Graph
Graphical Summaries for quantitative: Histogram, Boxplot