Exploratory Data Analysis

Psychology 3256

Introduction

- If you are going to find out anything about a data set you must first understand the data
- Basically getting a feel for you numbers
  - Easier to find mistakes
  - Easier to guess what actually happened
  - Easier to find odd values

Introduction

- One of the most important and overlooked part of statistics is Exploratory Data Analysis or EDA
- Developed by John Tukey
- Allows you to generate hypotheses as well as get a feel for you data
- Get an idea of how the experiment went without losing any richness in the data

Hey look, numbers!

x (the value)	f (frequency)
10	1
23	2
25	5
30	2
33	1
35	1

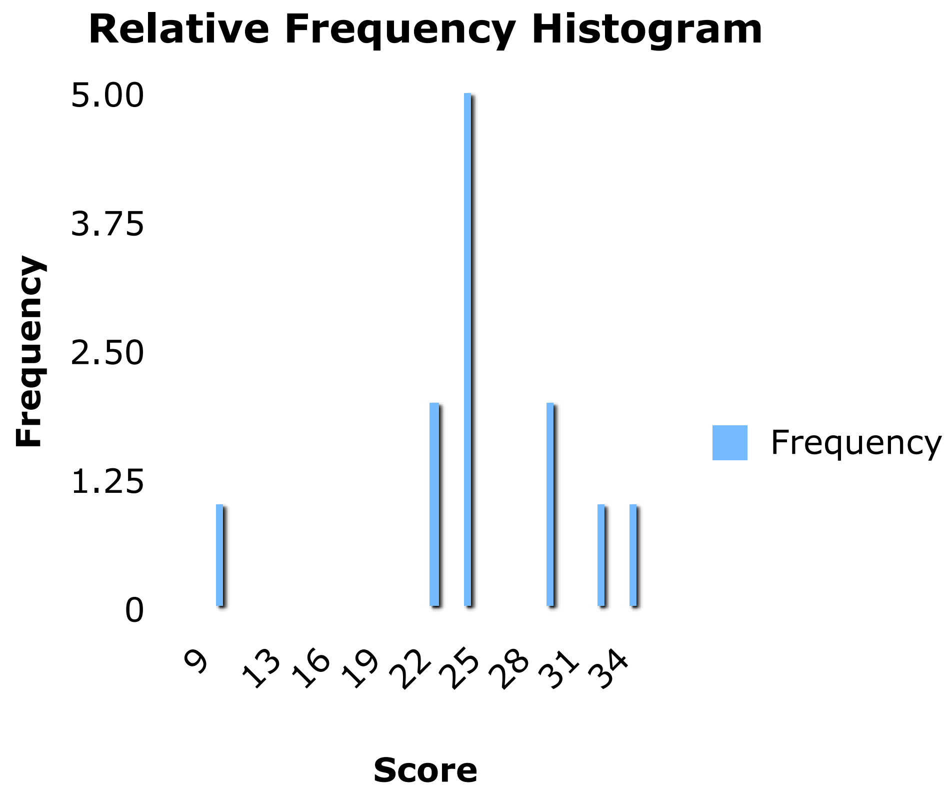
Frequency tables make stuff easy

$$\sum xf$$

- 10(1)+23(2)+25(5)+30(2)+33(1)+35(10
- = 309

Relative Frequency Histogram

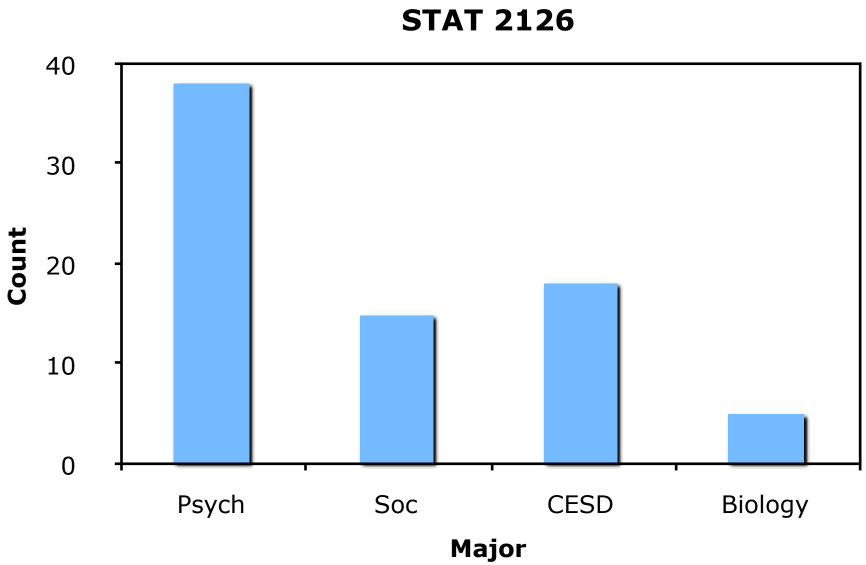
- You can use this to make a relative frequency histogram
- Lose no richness in the data
- Easy to reconstruct data set
- Allows you to spot oddities



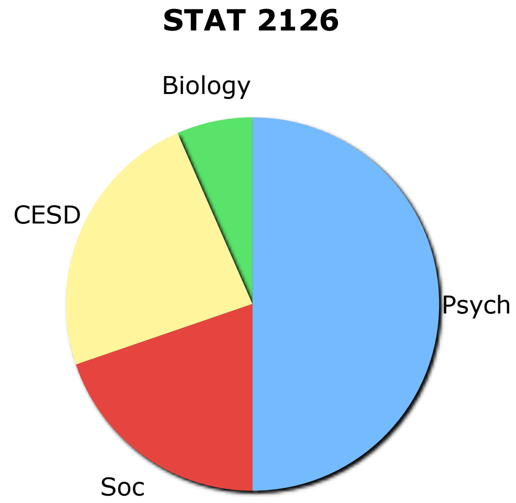
Categorical Data

- With categorical data you do not get a histogram, you get a bar graph
- You could do a pie chart too, though I hate them (but I love pie)
- Pretty much the same thing, but the x axis really does not have a scale so to speak
- So say we have a STAT 2126 class with 38 Psych majors, 15 Soc, 18 CESD majors and five Bio majors

Like this



STAT 2126
40
30
20
10
0
Count
Psych
Soc
CESD
Biology
Major

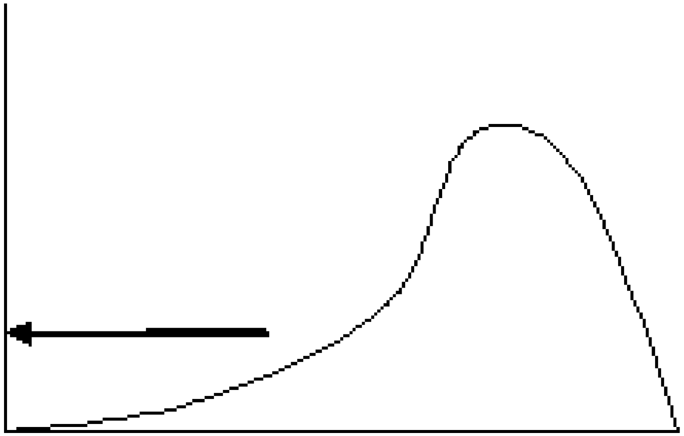


STAT 2126
Biology
CESD
Psych
Soc

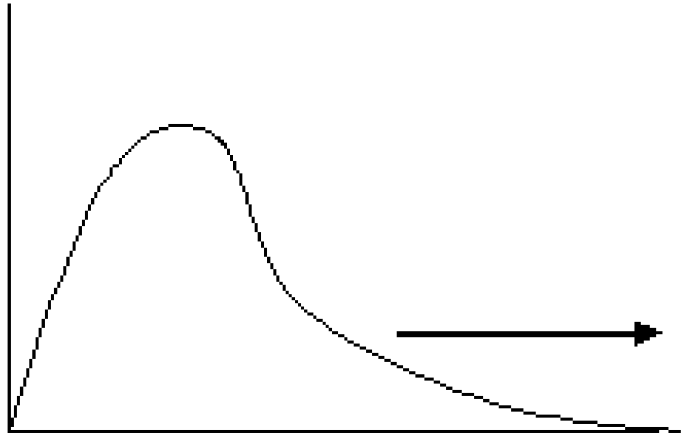
Quantitative Variables

- So with these of course we use a histogram
- We can see central tendency
- Spread
- shape

Skewness

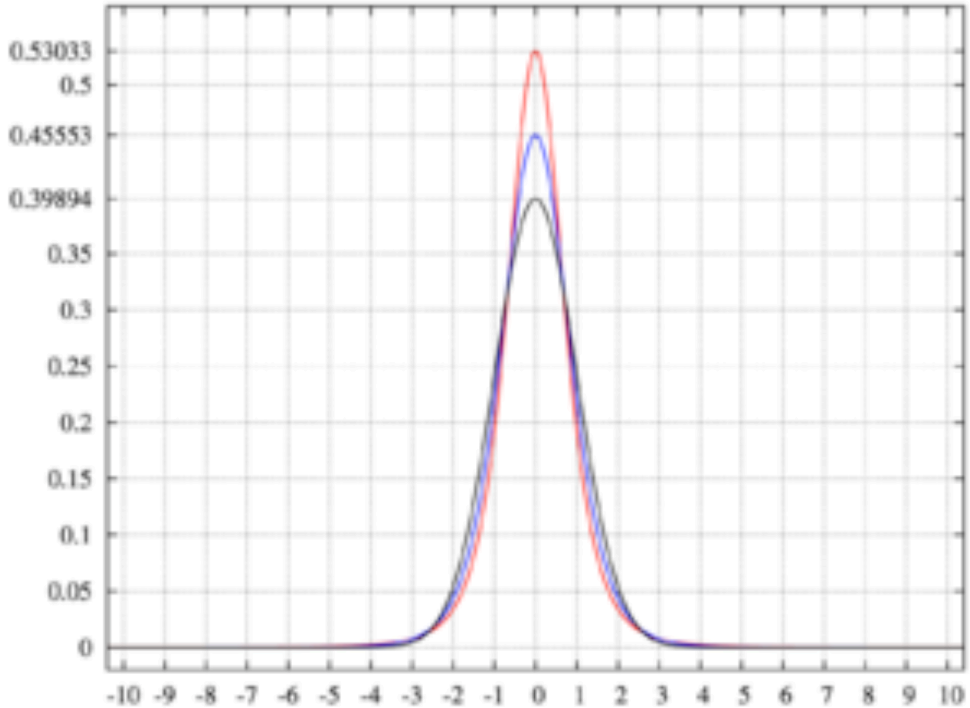


Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution



Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution

Kurtosis



- Leptokurtic means peaked
- Platykurtic means flat

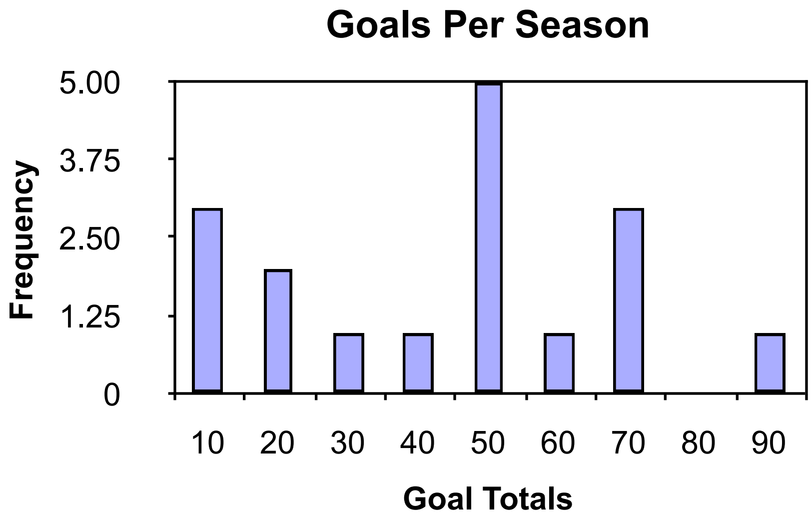
More on shape

- A distribution can be symmetrical or asymmetrical
- It may also be unimodal or bimodal
- It could be uniform

An example

- Number of goals scored per year by Mario Lemieux
- 43 48 54 70 85 45 19 44 69 17 69 50 35 6 28 1 7
- A histogram is a good start, but you probably need to group the values

Mario could sorta play



- Wait a second, what is with that 90?
- Labels are midpoints, limits are 5-14 … 85-94
- Real limits are 85.5 – 94.5

Careful

- You have to make sure the scale makes sense
- Especially the Y axis
- One of the problems with a histogram with grouped data like this is that you lose some of the richness of the data, which is OK with a big data set, perhaps not here though

Stem and Leaf Plot

0	1	6	7	
1	7	9		
2	8			
3	5			
4	3	4	5	8
5	0	4		
6	9	9		
7	0			
8	5			

- This one is an ordered stem and leaf
- You interpret this like a histogram
- Easy to sp ot outliers
- Preserves data
- Easy to get the middle or ***50th percentile*** which is 44 in this case

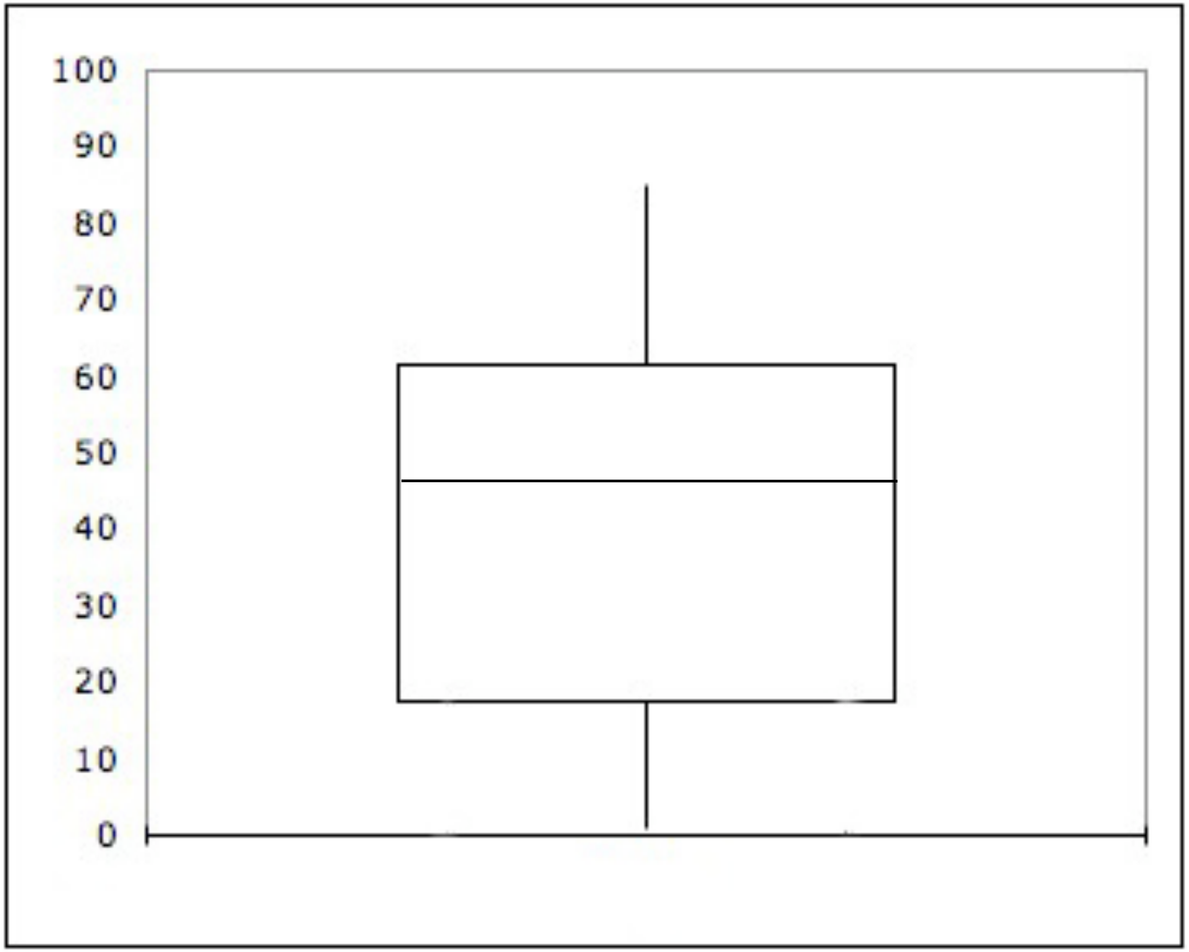
The Five Number Summary

- You can get other stuff from a stem and leaf as well
- Median
- First quartile (17.5 in our case)
- Third quartile (61.5 here)
- Quartiles are the 25th and 75th percentiles
- So halfway between the minimum and the median, and the median and the maximum

You said there were five numbers..

- Yeah so also there is the minimum 1
- And the maximum, 85
  – These two by the way, give you the range
- Now you take those five numbers and make what is called a box and whisker plot, or a boxplot
- Gives you an idea of the shape of the data

And here you go…



and it continues

- We talked about the central tendency of a distribution
- This is one of the three properties necessary to describe a distribution
- We can also talk about the shape
  - You know that kurtosis stuff and all of that

An Example

- Consider…
- 1 5 9 20 30
- 11 12 13 14 15
- Both have the same mean (13)
  - They both sum to 65, then divide 65 by 5, you get 13

The same, but different…

- 1 5 9 20 30
- 11 12 13 14 15
- So, they both have the same mean, and both are symmetrical
- How are they different?
- Well the one on the top is much more spread out

Spread

- Well how could we measure spreadoutedness?
- Well the range is a start
- 1 - 30 vs 11 - 15
- Seems pretty crude
- We could look at the IQD
- Still pretty crude

We need something better

- Something that is kind of like a mean really
- Like the average amount that the data are spread out
- Well why not do that?

Well here's why not

$$\sum \frac{(x-\bar{x})}{n} = \frac{(1-13)+(5-13)+(9-13)+(20-13)+(30-13)}{5}$$

$$= \frac{-12+(-8)+(-4)+7+17}{5}$$

$$= \frac{0}{5}$$

Hmm

- They will ALWAYS sum to zero
- Makes sense when you think about it
- If the mean is the balancing point, there should be as much mass on one side as the other
- So how do we get rid of negatives?
- Absolute value!

The Mean Absolute Deviation

$$\sum \frac{\left|(x-\overline{x})\right|}{n} = \frac{\left|(1-13)\right| + \left|(5-13)\right| + \left|(9-13)\right| + \left|(20-13)\right| + \left|(30-13)\right|}{5}$$

$$= \frac{12+8+4+7+17}{5}$$

$$= \frac{48}{5}$$

$$= 9.6$$

Cool!

- Well sometimes things you think are cool, well they aren’t
- Mullets for example…
- Anyway, for our purposes the MAD is just not that useful
- It is, in the type of stats we will do, a dead end
- Too bad, as it has intuitive appeal

There has to be another way

- Well of course there is or we would end now…
- OK, how else can we get rid of those nasty negatives?
- Square the deviations
- (you know, $-9^2 = 81$ for example)

We are getting closer…

$$\sum \frac{\left(x-\bar{x}\right)^2}{n} = \frac{\left(1-13\right)^2+\left(5-13\right)^2+\left(9-13\right)^2+\left(20-13\right)^2+\left(30-13\right)^2}{5}$$

$$= \frac{(-12)^2+(-8)^2+(-4)^2+7^2+17^2}{5}$$

$$= \frac{144+64+16+49+289}{5}$$

$$= 112.4$$

Hmmmm

- 112.4, seems like a mighty big number
- Well it is in squared units not in the original units
- What is the opposite of squaring something?
- Square root
- 10.6

There is a little problem here

- The formula I have shown you so far, has n on the bottom
- Yeah I know that just makes sense.
- In fact, it is supposed to be n-1
- We want something that will be an unbiased estimator of the same quantity in the population

Variance and standard deviation

- The population parameters, variance and the standard deviation have N on the bottom
- The sample statistics used to estimate them have n-1
- If they had n, they would underestimate the population parameters

Sample statistics

$$s^2 = \sum \frac{\left(x - \overline{x}\right)^2}{n-1}$$

$$s = \sqrt{\sum \frac{\left(x - \overline{x}\right)^2}{n-1}}$$

So in our case

$$s = \sqrt{\sum \frac{\left(x - \overline{x}\right)^2}{n - 1}} = \sqrt{\frac{\left(1 - 13\right)^2 + \left(5 - 13\right)^2 + \left(9 - 13\right)^2 + \left(20 - 13\right)^2 + \left(30 - 13\right)^2}{4}}$$

$$= \sqrt{\frac{(-12)^2 + (-8)^2 + (-4)^2 + 7^2 + 17^2}{4}}$$

$$= \sqrt{\frac{144 + 64 + 16 + 49 + 289}{4}}$$

$$= \sqrt{140.5}$$

$$= 11.85$$

For the Population

$$\sigma^2 = \sum \frac{(X - \mu)^2}{N}$$

$$\sigma = \sqrt{\sum \frac{(X - \mu)^2}{N}}$$

How are the variance and sd affected by extreme scores?

- 1 5 9 20 30
- s = 11.85
- OK let's throw in a new number, say 729
- 1 5 9 20 30 729
- Our new mean is 132.33
- Our new variance is 85555.067
- Our new standard deviation is 292.50
- Well the mean is affected by extreme scores, so of course so is the sd

How can we use this to our advantage?

- coefficient of variation
- Katz et al (1990)
- study mean 69.6 sd 10.6
- no study mean 46.6 sd 6.8
- one could conclude there is more variation with studying
- however the cvs are .152 and .146 respectively
- sd / mean

A couple of key points

- Remember, we want to learn about populations not samples
- we estimate population parameters with sample statistics
- we want unbiased estimators of parameters

Transformations

$$E(x + k) = \bar{x} + k$$

$$\mathrm{var}(x + k) = s_x^2$$

$$E(xk) = \bar{x}k$$

$$\mathrm{var}(xk) = s_x^2 k^2$$

40