Distributions and Samples

http://philosophy.ucsd.edu/faculty/wuthrich/

12 Scientific Reasoning

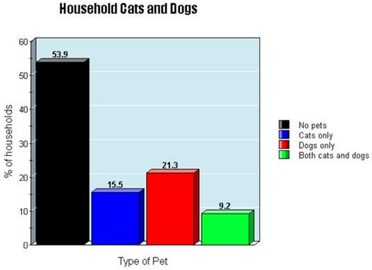
Acknowledgements: Bill Bechtel

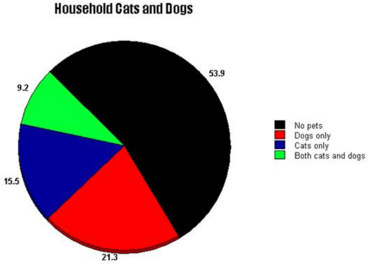
Distributions of values

- Since the values of a variable vary, they will be distributed.
- A major part of understanding a domain of objects is to describe how they are distributed on a given variable.
- One of the best ways to present a distribution is to graph it.

Nominal variables and bar graphs

Example: profile of pet ownership in San Diego County

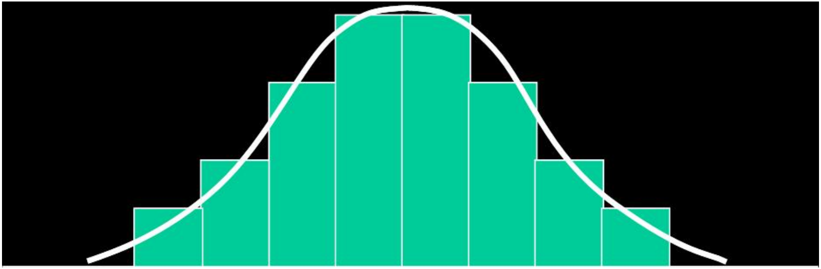




- Value of graphs: provide an intuitive appreciation of the data
- Bar graphs and pie charts work well with nominal and ordinal variables
- Bar graphs idea for relative comparison of size of groups (not so much for each one's share of the total); pie charts ideal to show each group's share of total (but harder to compare directly)

Score variables and histograms

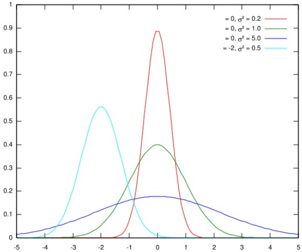
- Since score variables are continuous, histograms rather than bar graphs are used.
- This is done by creating bins and tabulating the number of items in each bin:



- The size of bins can create radically different pictures of the distribution!

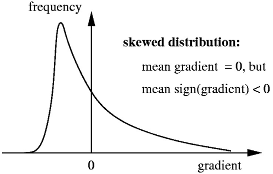
Normal and non-normal distributions

Normal distributions (More details on `Wikipedia`):

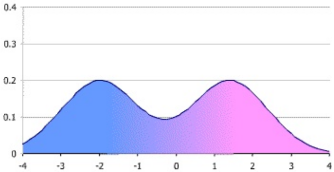


- have a single peak
- scores equally distributed around the peak
- fewer scores further from the peak

Non-normal distributions:



Skewed distribution

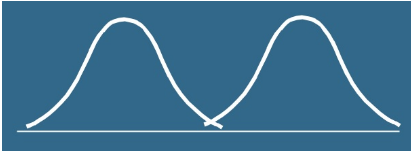


Bimodal distribution

Describing distributions: Two principal measures

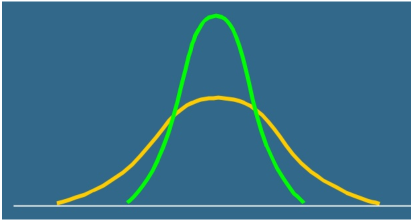
Central tendency:

- Two comparable distributions differing in central tendency



Variability:

- Two distributions with same central tendency but differing in variability



Three measures of central tendency

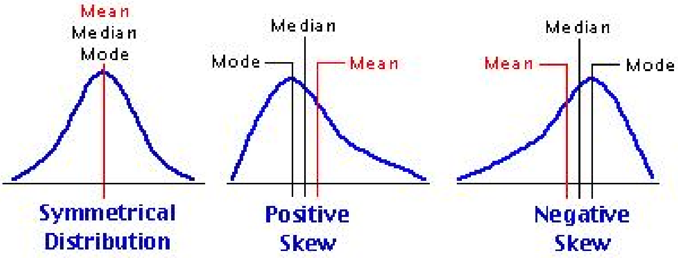
Consider this distribution of values: 2, 6, 9, 7, 9, 9, 10, 8, 6, 7

1. Mean: the arithmetic average, i.e. the sum of the scores divided by their number ($73 : 10 = 7.3$)
2. Median: the score of which half of the scores are higher and half are lower ($= 7.5$) (Note: for an even number of scores, the median is the mean of the middle two scores)
3. Mode: the most frequent score ($= 9$) (Note: in general not unique, since different scores may have same frequency)

Which measure to use?

- If the distribution is normal, all three measures of central tendency give the same result.
  - The mean is the easiest to calculate and the most frequently reported.
- If there are extreme outliners in one direction, the mean may be distorted:
  - exam scores: 21, 72, 76, 79, 82, 84, 87, 88, 90, 91, 95
  - mean: 78.6
  - median: 84
- In such a case, the median gives the better picture of the central tendency of the class.

The central tendencies for skewed distributions



Measures of variability

How much do the scores in a class vary?

Definition (Range)

The range is given by subtracting the smallest score from the largest score in the class. It thus provides a measure of statistical dispersion.

Definition (Variance)

The variance of a distribution of scores is given by

$$\frac{\sum (X - mean)^2}{N},$$

where the sum is over all the N scores X in the class. It gives a measure of how far the scores lie from the mean.

Definition (Standard deviation)

The standard deviation is given by the square root of the variance.

Intuitive interpretation of the standard deviation:

- one standard deviation: the part of the range in which 68% of the scores fall
- two standard deviation: the part of the range in which 95% of the scores fall
- three standard deviation: the part of the range in which 99% of the scores fall

Variance

Consider a distribution:

4	5	5	6	6	6	7	7	8	mean = 6
-2	-1	-1	0	0	0	1	1	2	$X-$ mean
4	1	1	0	0	0	1	1	4	$(X-\text{mean})^2$

Variance:

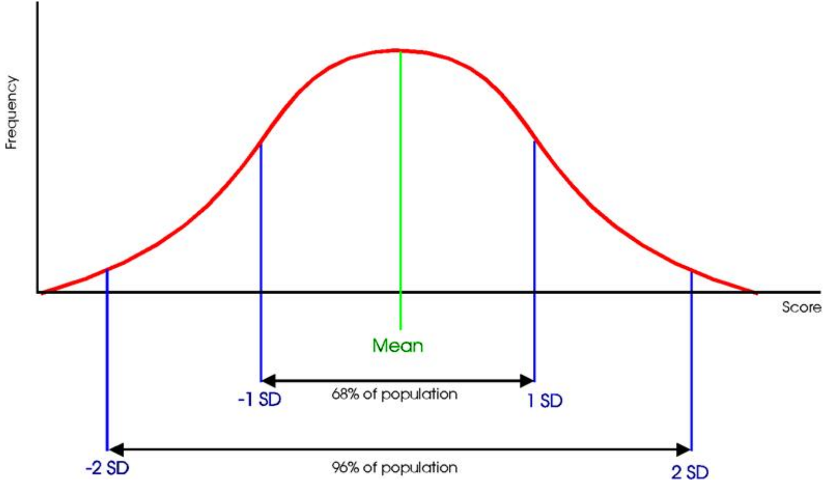
$$\frac{\sum(X-\text{mean})^2}{N} = \frac{12}{9} = 1.33$$

Standard deviation (SD):

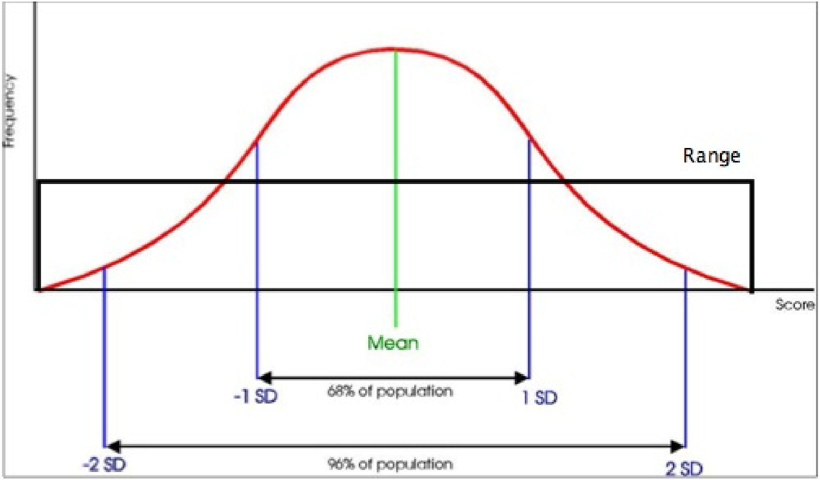
$$\sqrt{1.33} = 1.15$$

1 SD $= 6 \pm 1.15 = 4.85$ to 7.17
2 SD $= 6 \pm 2.30 = 3.70$ to 8.30

Range and standard deviation



Range and standard deviation



Populations

- The phenomena about which we seek to draw conclusions in a study are known as the population.
- Sometimes one can study each member of the population of interest.
- But if the population is large:
  - It may be impossible to study the whole population.
  - There may be no need to study the whole population.

Samples

- A sample is a subset of the population chosen for study.
- From studying the distribution of a variable in a sample one makes an estimate of the distribution in the actual population.
- Sometimes the estimate from a sample may be more accurate than trying to study the population itself.
  - Example: US Census

Does the sample reflect the population?

Question

Does the mean of the sample reflect the mean of the actual population?

- very unlikely that the mean of the sample will exactly equal the mean of the population

$\Rightarrow$ Given the means of a sample, what is the range within which the mean of the actual population lies?

- Bottom line: with larger samples this range becomes smaller and smaller
- And the effect depends only on the size of the sample, not the population sampled!

Is the sample biased?

- If information about the sample is to be informative about the actual population, the sample must be representative.
- Randomization: attempt to insure that the sample is representative by avoiding bias in selecting the sample
- Risk: inadvertently developing a misrepresentative sample
    - E.g., using telephone numbers in the phonebook to sample electorate

Two famous examples of sampling bias

(1) The *Literary Digest* and the 1936 US presidential election

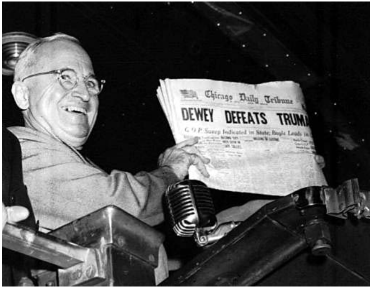
- *Literary Digest* magazine collected over two million postal surveys to predict that Republican candidate Alf Landon would beat the incumbent Franklin Roosevelt by large margin
- outcome was exact opposite (Landon only carried Maine and Vermont for a total of 8 electoral votes, compared to Roosevelt's 523)
- sample was collected from their readership, supplemented by records of registered automobile owners and telephone users

⇒ over-representation of the wealthy, who tended to vote Republican

- Footnote: in contrast, George Gallup's organization successfully predicted outcome from polling only 50 thousand

Two famous examples of sampling bias

(2) The "World's Greatest Newspaper" and the 1948 US presidential election



- frontline of *Chicago Tribune* on 3 November 1948 (election night) read DEWEY DEFEATS TRUMAN
- Next morning, a grinning President-Elect Harry Truman was photographed (right).
- reason for mistake: wrongly trusted their phone survey, when sample of telephone users was not representative of general population
- (and Gallup poll on which the prediction was partly based was outdated by two weeks)

Distribution on nominal variables

Take the special case of a variables with two values (exhaustive and exclusive)

- heads/tails
- true/false
- born in January/not-born in January
- male/female

where the value for each item is independent of that for other items, and consider the likely distributions.

Birth order

Consider these to be orders of births of babies in a hospital. Which of the following is more/most likely?

1. MFMFFMFMFF
2. MMMMMMMMMM
3. FFFFFMMMMM

Each pattern is equally likely! ($0.5^{10} \approx 0.1\%$)

A very different question

Consider these to be totals of births of babies in a hospital on a given day. Which of these outcomes is more/most likely?

1. 5 males / 5 females
2. 7 males / 3 females
3. 10 males / 0 females

From populations to samples

Start from the situation in which we know the distribution in the actual population: $p(M) = 0.5$

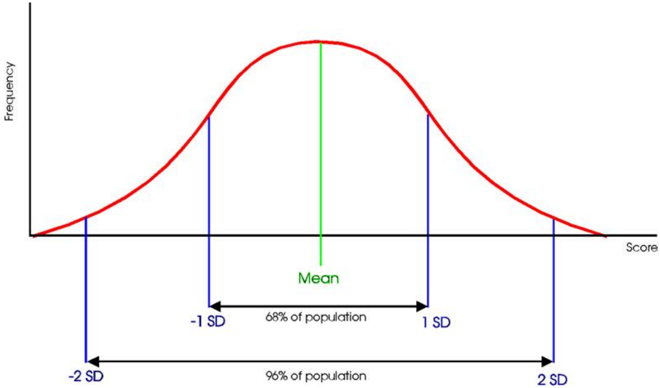
We draw a sample of a given size, say 10.

- Is it possible that we could get a sample of all males? Yes, the probability is about 0.001.
- What is the probability that we could get a sample of 7 males and 3 females? It is about 0.117.
- What is the probability that we could get a sample of 5 males and 5 females? It is about 0.246.

What happens as sample size gets larger?

- With larger sample sizes, the probability of a distribution in the sample closely approximating the distribution in the actual population increases.
- The important question is how much the mean of the samples will vary from the mean of the actual population.
- To determine this, the standard deviation measure is very useful.

Standard deviation and mean



- In $\approx$ 68% of samples, the mean of the population will fall within 1 standard deviation of the mean of the sample.
- In $\approx$ 95% of samples, the mean of the population will fall within 2 standard deviation of the mean of the sample.

SD and larger sample size

- As sample size grows, the SD of the sample shrinks.
- So with larger samples, the range of 2 standard deviations shrinks.
- Assume mean in the sample is 0.50.

Sample size	Percentage range of 2 SD	Percentage range of 3 SD
10	34.5–65.5	29.5–70.5
20	39–61	35.6–64.4
50	43–57	40.9–59.1
100	45–55	43.5–56.5
500	47.8–52.2	47.1–52.9
1000	48.4–51.6	48–52

Generalize to score variables

- Score variables: interval and ratio variables
- With score variables, it is the scores that are distributed (not the items in a given category).
- Example: age of person eating at the Food Court
- Draw a sample to make inference of average age of person eating at the Food Court

Age	<17	17	18	19	20	21	22	23	24	25	>25
Pop.	(6)	18	23	34	32	18	26	29	14	10	(10)
Sam.		2	1	3	1	2		1			

Estimating real distribution

Age	<17	17	18	19	20	21	22	23	24	25	>25
Pop.	(6)	18	23	34	32	18	26	29	14	10	(10)
Sam.		2	1	3	1	2		1			

- mean of the actual population: 20.63
- mean of the sample: 19.4
- SD of the sample: 1.9
- range of 1 SD: 17.5–21.3
- range of 2 SD: 15.6–23.2

Want to predict more accurately? Use a larger sample size

Estimating real distribution

Age	<17	17	18	19	20	21	22	23	24	25	>25
Pop.	(6)	18	23	34	32	18	26	29	14	10	(10)
Sam.		2	1	3	1	2		1			
		1	2	4	6	3	2	2			

- mean of the actual population: 20.63
- mean of the sample: 19.4 ⇒ 20.1
- SD of the sample: 1.9 ⇒ 1.6
- range of 1 SD: 17.5–21.3 ⇒ 18.5–21.7
- range of 2 SD: 15.6–23.2 ⇒ 16.9–23.3

Want to predict more accurately? Use a larger sample size

Summary and review

- four types of variables: nominal, ordinal, interval, ratio (score variables)
- values of variables are distributed and it is an important goal to characterize this distribution
- graphs:
  - bar graphs for nominal variables
  - histograms for score variables
- normal vs. non-normal distributions (skewed, bimodal etc.)
- two principal measures of distributions:
  1. central tendency: mean, median, mode
  2. variability: range, variance, standard deviation
     - 1 SD includes approx. 68% of scores
     - 2 SD includes approx. 95% of scores
     - 3 SD includes approx. 99% of scores

- population and samples
  - From studying the distribution in sample, estimate the distribution in the actual population.
  - mean of actual population will
    - fall within one SD of mean of sample 68% of time
    - fall within two SD of mean of sample 95% of time
    - fall within three SD of mean of sample 99% of time
  - larger sample yields smaller SD and hence more precise estimate
  - hence, to improve the precision of an estimate, use a larger sample