# Distributions and Samples

## Clicker Question

A researcher studying the participation of students in class discussions sits in the classroom and discretely observes and makes notes about student participation. The research is

- engaged in participant observation
- engaged in naturalistic observation
- engaged in social observation
- generates reactivity bias

## Clicker Question

To determine how many vehicles travel a given road, a researcher installs a camera that takes a picture of traffic every 15 minutes. This researcher is using

- Continuous observation
- Time sampling
- Event sampling
- Situation sampling

## Clicker Question

To avoid reactivity bias in observation, a researcher should

A. not engage in participant observation
B. not employ video cameras
C. avoid antropormophic vocabulary
D. try to avoid being detected

## Clicker Question

The variable PARTY AFFILIATION (Libertarian, Green, Republican, Democrat, other) is

- A categorical or nominal variable
- An ordinal or rank variable
- An interval variable
- A ratio variable

## 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* values are distributed on a given variable

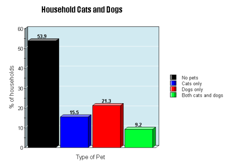
One of the best ways to present a distribution is to graph it

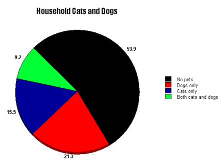
# Nominal variables and bar graphs

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



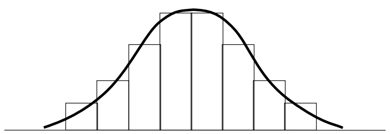




---

# 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!

---

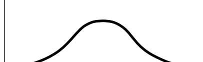
# Normal and non-normal distributions

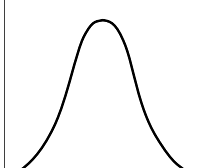
**Normal distributions**

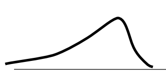
- Have a *single* peak
- Scores *equally* distributed around the peak
- Fewer scores *further* from the peak

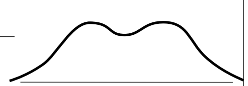
**Non-normal distributions**:

- ***Skewed***
- ***Bimodal***









# Clicker Question

The distribution below is

| < 70 | 70-74 | 75-79 | 80-84 | 85-89 | 90-94 | 95+ |
|---|---|---|---|---|---|---|
| 21 | 15 | 6 | 23 | 3 | 21 | 18 |

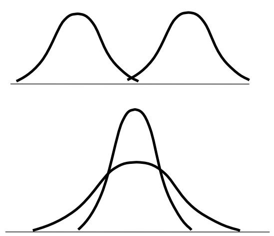
A. Normal since it has one peak
B. Normal since scores are equally distributed around the peak
C. Not normal since the scores are not equally distributed around the peak
D. Not normal since there are not fewer scores further from the peak

# 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
- Mean: the arithmetic average
  73 / 10 = 7.3
- Median: the score of which half are higher and half are lower = 7.5
- Mode: the most frequent score = 9

# 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 the distribution is not normal and there are extreme outliers 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 a better picture of the central tendency of the class

# Measures of variability

- How much do the scores vary?

  Range: the lowest value to the highest value

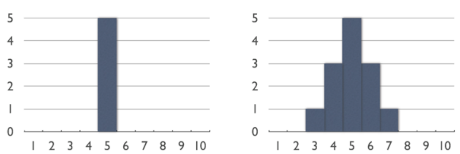
  Variance: $\frac{\sum (X\text{-mean})^2}{N}$

  Standard Deviation (SD): $\sqrt{\text{Variance}}$

- Intuitive interpretation (with normal distributions):
  - One standard deviation: the part of the range in which 68% of the scores fall
  - Two standard deviations: the part of the range in which 95% of the scores fall
  - Three standard deviations: the part of the range in which 99% of the scores fall.



# Same Mean, Different SD



Mean = 5.0 SD = 0

Mean = 5.0 SD = 1.04

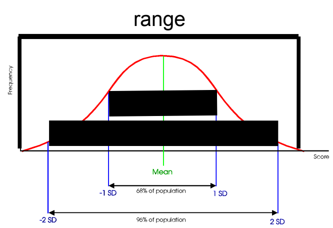
# Variance and Standard Deviation

- 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 | | $(X\text{-mean})^2$ |
| $\dfrac{\sum (X\text{-mean})^2}{N} = \dfrac{12}{9} = 1.33$ | | Variance |
| $\sqrt{1.33} = 1.15$ | | SD |

Range of 1 SD = 6 ± 1.15 = 4.85 to 7.15
Range of 2 SD = 6 ± 2.30 = 3.70 to 8.30

# Range and Standard Deviation




# Clicker Question

On the exam on which scores were distributed normally and the mean was 86 and the SD was 3.5,

A. 68% of the scores were between 82.5 and 89.5
B. 95% of the scores were between 82.5 and 89.5
C. 99% of the scores were between 79 and 93
D. 68% of the scores were between 79 and 93

## Populations

- The group 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
  - U.S. Census



## Does the sample reflect the population?

- Does the mean of the sample reflect the mean of the actual population?
  - Very unlikely that the mean of the same will exactly equal the mean of the population
  - Given the mean 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 this effect depends only on the size of the sample, not the size of 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



## From populations to samples

Start from the situation in which we know the distribution in the actual population: p (M) = .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 .001

What is the probability that we could get a sample of 7 males and 3 females?
It is about .117

What is the probability that we could get a sample of 5 males and 5 females?
It is about .246

## What happens as sample size gets larger?

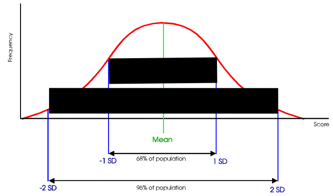
With larger sample sizes, the probability of a distribution in the sample closely approximating the distribution in the actual population increases

The important question is how much the mean of the samples will vary from the mean of the actual population

To determine this, we need to know the standard deviation (SD) of the sample.

# Standard deviation and mean

In ≈68% of samples, **the mean of the sample** will fall within 1 standard deviation of the **mean of the population**




In ≈ 95% of samples, the **mean of the sample** will fall within 2 standard deviations from the **mean of the population**

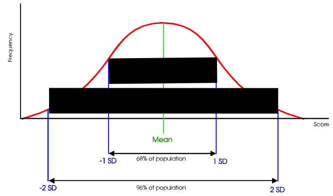
# Inferring Mean of Population from Mean of the Sample

- Just as we can determine from the mean and standard deviation of the actual population where the mean of the sample is likely to be
  - We can infer from the mean and standard deviation of the sample how far from the mean of the sample the mean of the actual population will likely be
    - 68% of the time it will be within one SD
    - 95% of the time it will be within two SD
    - 99% of the time it will be within three SD
- These percentages express our confidence that the mean will be in the range specified



# Standard deviation and mean

≈ 68% of the time, **the mean of the population** will fall within 1 standard deviation of the **mean of the sample**




≈ 95% of time, the **mean of the population** will fall within 2 standard deviations from 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 variability around the mean shrinks

Assume mean in the sample is 50%

| Sample size | +/- 2 SD (95%) | +/- 3 SD (99%) |
|---|---|---|
| 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 |

# Clicker Question

Why do most election polls study approx. 500 people even if the population is many million?

A. It gets hard to analyze data when too much is collected
B. It costs too much to survey more than about 500 people
C. With 500 people the SD is already small enough to make a good estimate of the actual population
D. With 500 people the SD is already large enough to make a good estimate of the actual population

# 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

| **<17** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** | **25** | **>25** |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 18 | 23 | 34 | 32 | 18 | 26 | 29 | 14 | 10 | 10 |
| | 2 | 1 | 3 | 1 | 2 | | 1 | | | |

# Estimating real distribution

| <17 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | >25 |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 18 | 23 | 34 | 32 | 18 | 26 | 29 | 14 | 10 | 10 |
| | 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 = 17.5-22.3 | 18.5-21.7 | |
| Range of 2 SD = 15.9-24.2 | 16.9-23.3 | |

Want to predict more accurately?

Use a larger sample size