ECON2843 Elements of Statistics

Part 1 Descriptive Statics, Summary Measures, and Data Visualization

Roadmap for Course

First half of course - learn some fundamental statistical concepts and tools:

- Descriptive statistics for data
- Summary measures
- Data visualization
- Probability
- Probability distributions
- Sampling and sampling distributions
- etc..

Roadmap for Course

Second half of course - use the tools and techniques from first half to tackle more complex statistical questions:

- Estimation
- Test a hypothesis
- Compare two distributions: χ-square test; analysis of variance
- Correlation and regression
- Applications
- etc..

I have a Question...

- I ask myself at night, "Am I smarter than the average person?"
- How can I find out?

I have a Question...

- Step 1: Establish a relevant variable of interest.
  - Perhaps IQ might be appropriate.
- Step2: Compare my IQ to a benchmark.
  - How do we form this benchmark?
  - We need to think back to the original question.

I have a Question...

- A good benchmark might be the average IQ in the population.
- Ideal setting is to survey everyone on the planet and calculate the average IQ.
- In practice, this is impossible.
- What to do?

I have a Question...

- Use a (representative) sample from the population.
- Compare my IQ to the average IQ of the sample.
- Then, answer my original question.

The Study of Statistics

To summaries we

- Established a question of interest, or *hypothesis*, that can be tested.
- Determine some relevant *variables*.
- Identified our *population* of interest.
- Gathered some data by taking a *sample* from the population.
- Analyze the data we gathered.
- Form a *causal inference* or conclusion regarding the original hypothesis.

The Study of Statistics

- Statistics is essentially the study of **data**.
- More specifically, **statistical inference** refers to the problem of determining the behavior of a large population by studying a small sample of data from that population.

Terminology and Notation

- Population: Every observation of interest available in the physical world.

  For example:

  - Every single person in Norman, Oklahoma.
  - Every single student who has ever taken ECON2843.
- Usually use N to denote the total number of observations in the population.

Terminology and Notation

- Populations have **parameters**: A descriptive measure of a population that is usually unobservable and unknown. The parameters are characteristic of a population.
- Parameters are typically denoted by Greek letters:
  - Population average/mean: μ
  - Population variance: σ^2

Terminology and Notation

- **Sample**: A selection of observations drawn randomly from the population of interest.

  For example:

  - A random sample of 50 OU students from the entire university.
  - A random sample of 10 ECON2843 students.
- Usually use n to denote the total number of observations in the sample.

Terminology and Notation

- Sample have **statistics**: A descriptive measure of a sample that can be observed (calculated) and is known.
- Sample statistics are used to make inferences about population parameters and are typically denoted by Roman letters:
  - Sample average/mean: $\bar{X}$
  - Sample variance: s^2

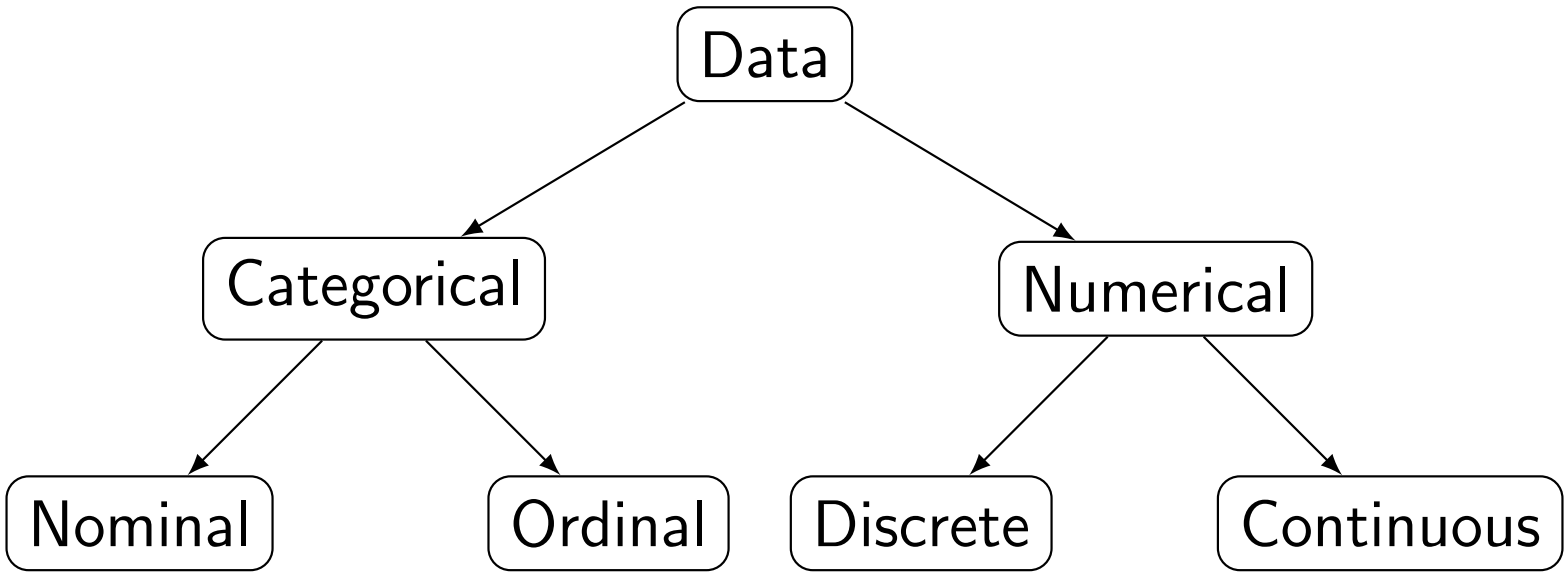
Terminology and Notation

- **Variable**: Any quantity or characteristic that can be measured or recorded for *each observation* in a population or sample.

  For example:

  - A person's IQ.
  - A person's eye color.
  - A stock's price.
- The value of a variable is likely to be different from observation to observation, hence the name variable.

Types of Data



Categorical Data

- Data where the values fall into categories.
- Nominal data: The categories have no ordering or relationship.

  Examples: Marital status, eye color, job, etc.
- Ordinal data: The categories have a distinct ordering.

  Examples: Ranking teacher performance as "poor/fair/good", survey answer "strongly disagree/disagree/agree/strongly agree", etc.

Descriptive Tools for Categorical Data

Numerically:

- Frequency of each category.
- **Mode**: The most frequently occurring observation.

Graphically:

- Bar charts.
- Pie charts.

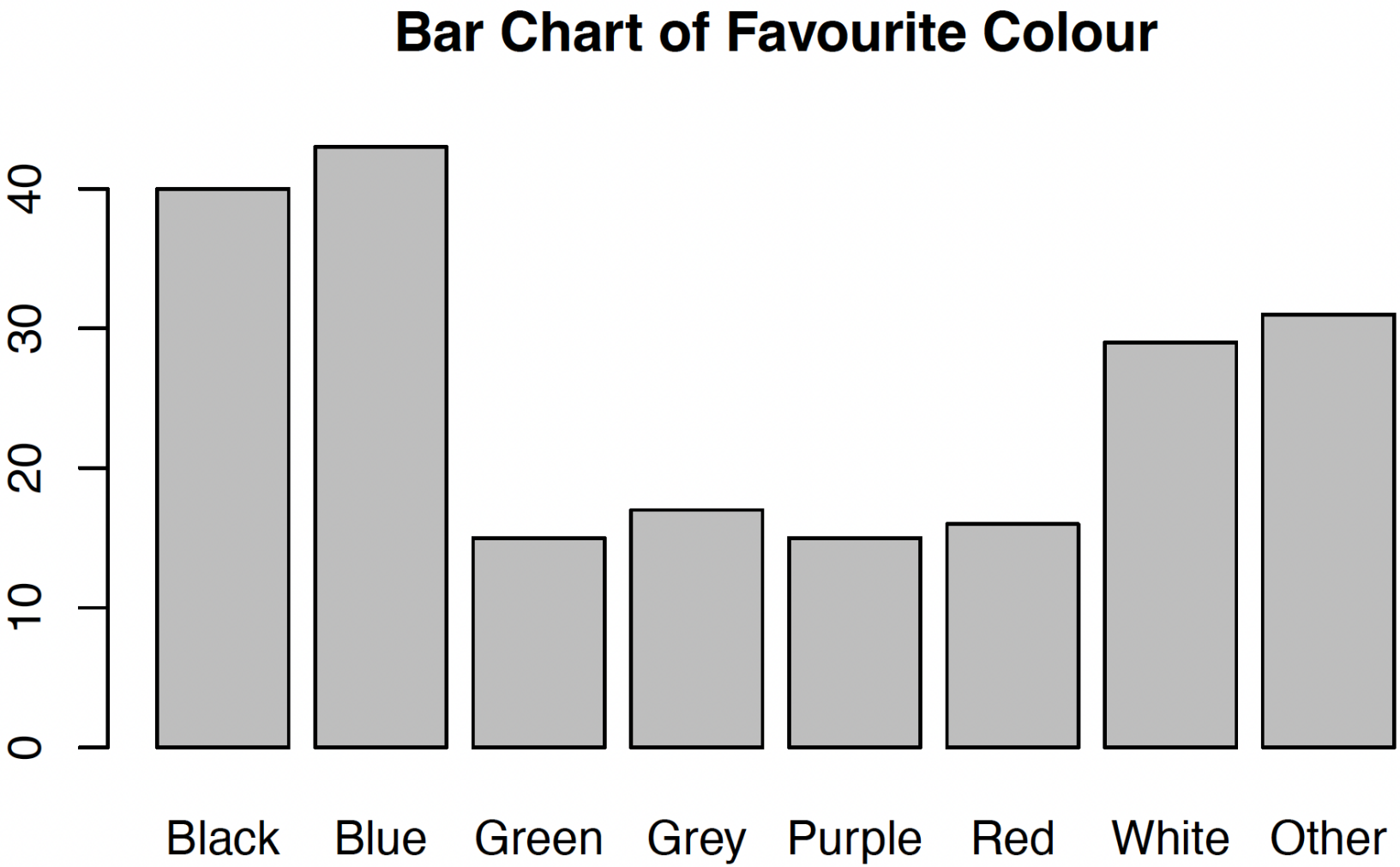
Presentation of Nominal Data: Example

- Experiment done to determine the favorite colors of students at a university.
- We surveyed 206 students and asked the question:What is your favorite color?

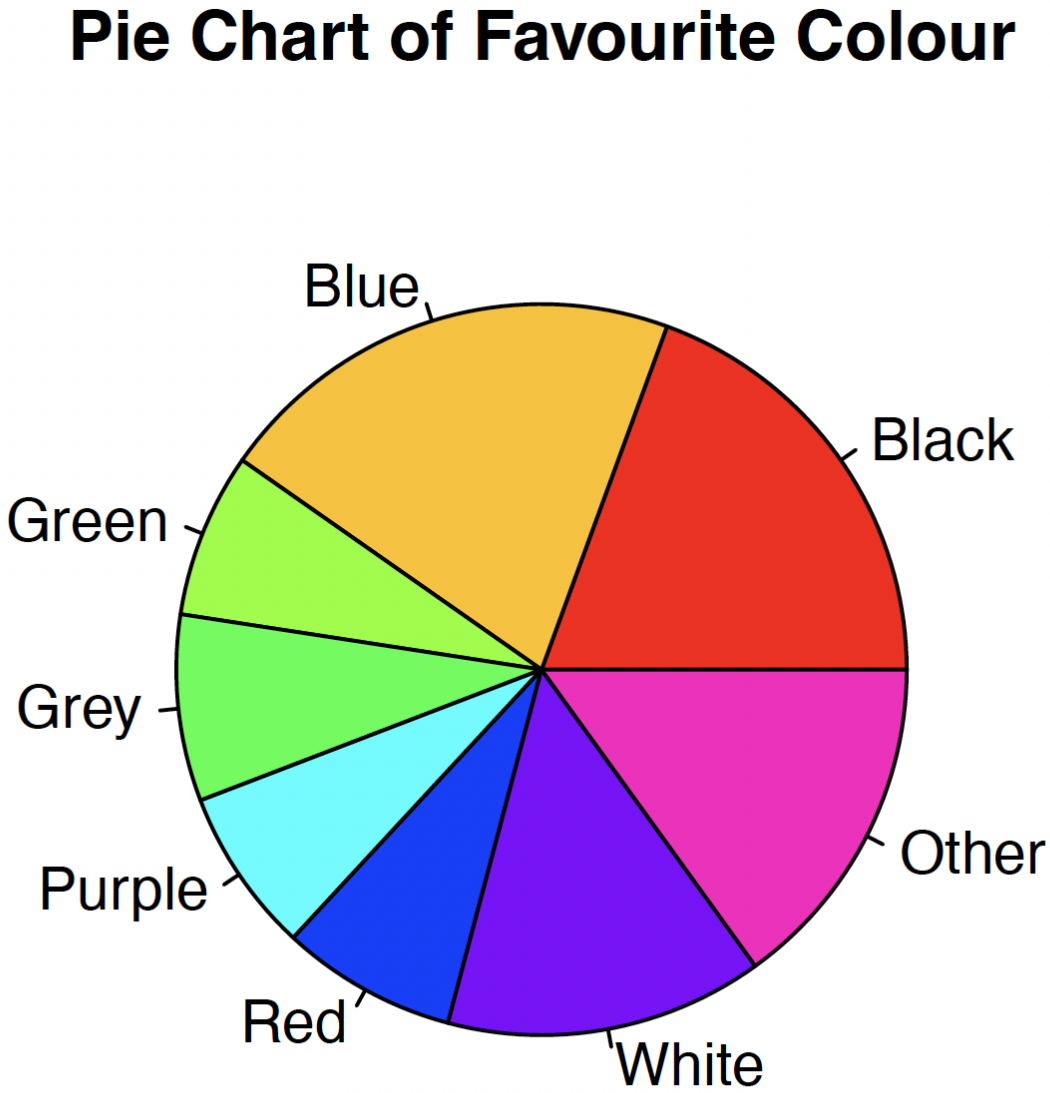
Color	Black	Blue	Green	Grey	Purple	Red	White	Other
Count	40	43	15	17	15	16	29	31

- The mode is *blue*, most frequently occurring observation.

Presentation of Nominal Data: Example



Presentation of Nominal Data: Example



Presentation of Ordinal Data

- Can use exactly the same tools we used for nominal data, e.g., bar charts, pie charts.
- But, the most important thing is that we preserve the order of the categories.
- But, the most important thing is that we preserve the order of the categories. For example, the bars in a bar chart should be in increasing or decreasing ordinal value.

Time Series vs Cross-Sectional Data

- **Time Series Data**:
  - Data points collected over time.
  - Each observation corresponds to a different time point.
  - Example: Daily stock prices, monthly unemployment rates, etc.
- **Cross-Sectional Data**:
  - Data collected at a single point in time.
  - Each observation represents a different individual or entity.
  - Example: GDP of various countries in a given year, employee counts across companies on a specific day, etc.

Numerical Data

- Data where the values can be *measured*.
- **Continuous data**: Anything that can be measured in infinitely small increments. Example: Weight, Height, etc.
- **Discrete data**: Anything that can be measured in fixed increments.
  - The number of cars you own, number of heads in three coin flips, etc.

To Describe Numerical Data

Numerically:

- Mean, median, mode.
- Quantile.
- Range, variance, coefficient of variance.
- Covariance, correlation.

Graphically:

- Histograms.
- Boxplots.

Measures of Central Tendency

- A **measure of central tendency** measures the location of the middle or center of the distribution of your data.
- Common measures include the arithmetic mean, the median and the mode.

Measures of Central Tendency

- Suppose we have two tutorial sessions, A and B.
- We would like to establish which tutorial session performed better in a recent quiz:

A	5	6	5	7	8	7	8	8
B	9	5	6	7	7	6	5	

Mean

- The **arithmetic mean** is the average of all the observations.
- Population mean:

$$\mu = \frac{1}{N}(X_1 + \cdots + X_N) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- Sample mean:

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Mean

So for the two tutorial sessions:

A	5	6	5	7	8	7	8	8
B	9	5	6	7	7	6	5	

Tutorial A: The mean of students' grades is 6.75.
Tutorial B: The mean of students' grades is 6.43.

Median

The median is the middle observation. Rank the observations in ascending order; median is the middle observation if n is odd, or the average of the middle two observations if n is even.

A	5	6	5	7	8	7	8	8
B	9	5	6	7	7	6	5	

Tutorial A: The median is ?
Tutorial B: The median is ?

Mode

The mode is the most frequently occurring observation.

A	5	6	5	7	8	7	8	8
B	9	5	6	7	7	6	5	

Tutorial A: The mode is ?
Tutorial B: The mode is ?

Mean vs Median

The mean is the most commonly used measure.

But, the median is more robust to extreme observations.

B	9	5	6	7	7	6	5

Mean $= 6.43$, median $= 6$.

B	90	5	6	7	7	6	5

Mean $= 18$, median $= 6$.

Measures of Variability

Let's say we receive the final grades for the semester for students in the two tutorial sessions:

A	75	80	70	77	73	75	90	60
B	75	100	50	85	65	98	52	

Measures of Variability

If we calculate the mean of each tutorial, we find that they are both equal to 75.

									$\bar{X}$
A	75	80	70	77	73	75	90	60	75
B	75	100	50	85	65	98	52		75

Measures of Variability

- However, it is clear that these two tutorials are not the same.
- Is there another characteristic of the distributions of marks that we can measure and compare?
- Variability!
  - Which tutor is more consistent in their teaching methods?
- How can we quantify the difference in variability or "spread" in the marks?

Range

- The range of a data set is defined to be:

  Range = Largest Value − Smallest Value

- Session A: The range is $90 - 60 = 30$.
- Session B: The range is $100 - 50 = 50$.
- Range is simple to understand and calculate, but can be affected by extreme observations or "outliers".

Measures of Variability

- The range is useful, but it is calculated using only two numbers.
- What about the other observations in our data set?
- A better idea for measuring the variability might be to look at the distance of each observation from a central measure...

Measures of Variability

- Distances from the mean, i.e., $(X_i - \bar{X})$.

	X_i								$\bar{X}$
A	75	80	70	77	73	75	90	60	75
B	75	100	50	85	65	98	52		75

	$X_i - \bar{X}$								$\sum$
A	0	5	-5	2	-2	0	15	-15	0
B	0	25	-25	10	-10	23	-23		0

Measures of Variability

- Squared distances from the mean, i.e., $(X_i - \bar{X})^2$.

$(X_i - \bar{X})^2$									$\sum$
A	0	25	25	4	4	0	225	225	508
B	0	625	625	100	100	529	529		2508

Measures of Variability

- Not a good comparison ☹, as tutorial B is smaller than tutorial A (i.e., we must scale by class size).

	$(X_i - \bar{X})^2$								$\frac{\sum}{n-1}$
A	0	25	25	4	4	0	225	225	72.57
B	0	625	625	100	100	529	529		418

Variance

- What we just calculated is known as the **variance**,which measures the spread or variability of a given distribution of data.
- Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

- Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Variance

To be continued...