Introduction to Statistics

Topics

- Topics to be covered in the course:
  - Design of Experiments,
  - Exploratory Data Analysis and Descriptive Statistics,
  - Probability Theory,
  - Sampling Distributions and the Central Limit Theorem,
  - Estimation,
  - Statistical Inference,
  - Contingency tables,
  - Nonparametric Tests,
  - Power and sample size,
  - ANOVA,
  - Correlation and Regression

What is Statistics?

- **Statistics** can be defined as "a quantitative technology for empirical science; it is a logic and methodology for the measurement of uncertainty and for an examination of that uncertainty."

- The key word here is "uncertainty." Statistics become necessary when observations are variable.

The role of statistical analysis in science

- This course discusses statistical methods and their applications to engineering problems.
- We use **empirical evidence** to study data and make **informed decisions**.
- To study data, we measure a set of **characteristics**, which we refer to as variables.
- The objective of many scientific studies is to learn about the **variations** of a **specific characteristic**

The role of statistical analysis in science

- For engineering problems, it is important to find possible **relationships** among different **variables**.
- The variables that are the main focus of a study are as the **response (or target) variables**.
- In contrast, variables that explain or predict the variation in the response variable are called as **independent variables** or **predictors**
- **Statistical analysis** begins with a scientific problem usually presented in the form of a **hypothesis testing** or a **prediction problem**.

Goals of statistics

- **Estimate** the values of important **parameters**
- **Test hypotheses** about those parameters

Statistics is also about good scientific practice

- **Feline High-Rise Syndrome (FHRS)**
  - The injuries associated with a cat falling out of a window.

- "The diagnosis of high-rise syndrome is not difficult. Typically, the cat is found outdoors, several stories below, and a nearby window or patio door is open."

High falls show lower injury rates

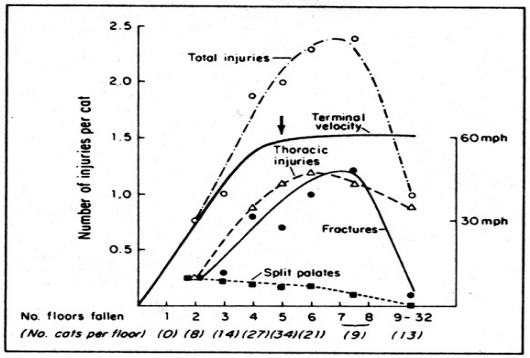


Figure 2—Relationship of injuries to distance fallen and velocity in 132 cats with high-rise syndrome: ↓ points to terminal velocity (——); total number of injuries/cat (○, —·—·—); number of thoracic injuries (pulmonary contusions +pneumothorax)/cat (△, – – –); number of fractures/cat (●, ——); number of split palates/cat (■, – – – – –).

Why?

- Cats have high surface-to-volume ratios
- Cats have excellent vestibular systems
- Cats reach terminal velocity quickly, relax, and therefore absorb impact better
- Cats land on their limbs and absorb shock through soft tissue

Jared Diamond, Nature 1988

Or not…

A *sample of convenience* is a collection of individuals that happen to be available at the time

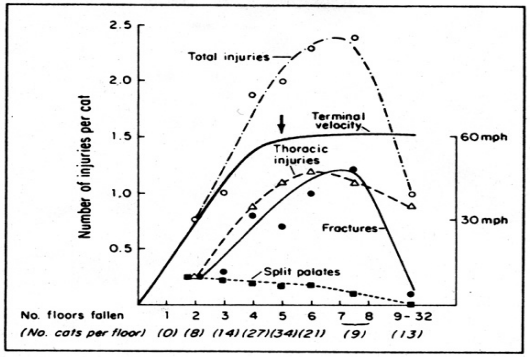


Figure 2—Relationship of injuries to distance fallen and velocity in 132 cats with high-rise syndrome: ↓ points to terminal velocity (——); total number of injuries/cat (○, —·—·—); number of thoracic injuries (pulmonary contusions +pneumothorax)/cat (△, – – –); number of fractures/cat (●, ——); number of split palates/cat (■, – – – – –).

Whitney and Mehloff, *Journal of the American Veterinary Medicine Association*, 1987

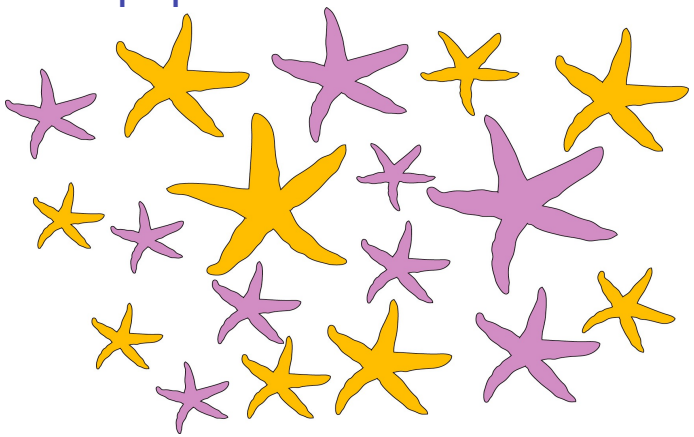
Sampling Principles

- **Samples** are typically selected randomly (i.e., with some probability) from the **population**.
- Unless stated otherwise, these **randomly selected** members of populations are treated as **independent**.
- The selected members (e.g., people, points) are called **sampling units**.
- The individual entities from which we collect information are called observation units, or simply **observations**.
- Samples must be **representative of the population**, and their environments should be comparable to the population.

Populations vs Samples

Populations <-> **P**arameters;
Samples <-> E**s**timates

A population of starfish



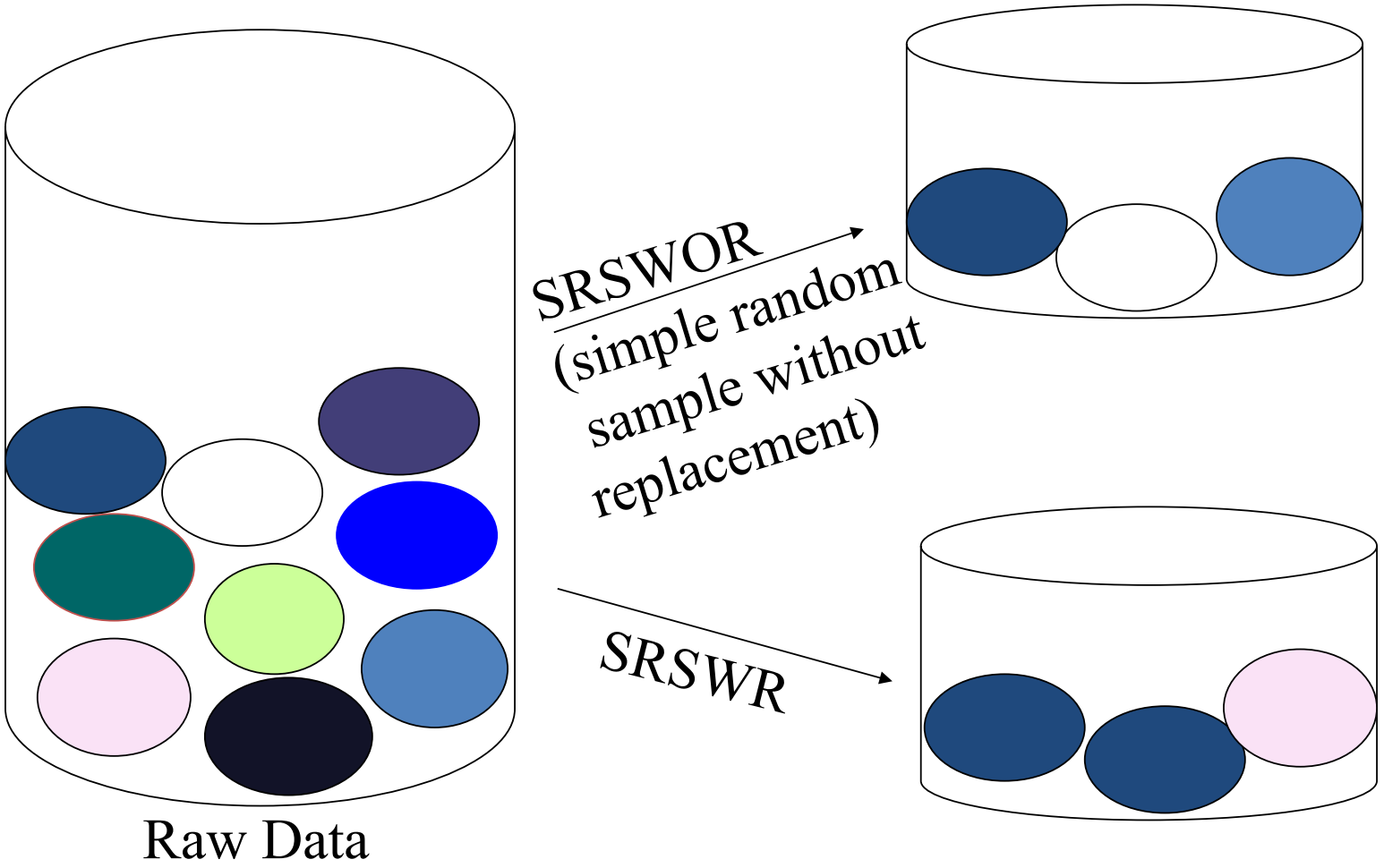
Sampling

Some of the most widely used sampling designs

- **Simple Random Sampling**: the chance of being selected is the same for any group of ***n*** members in the population

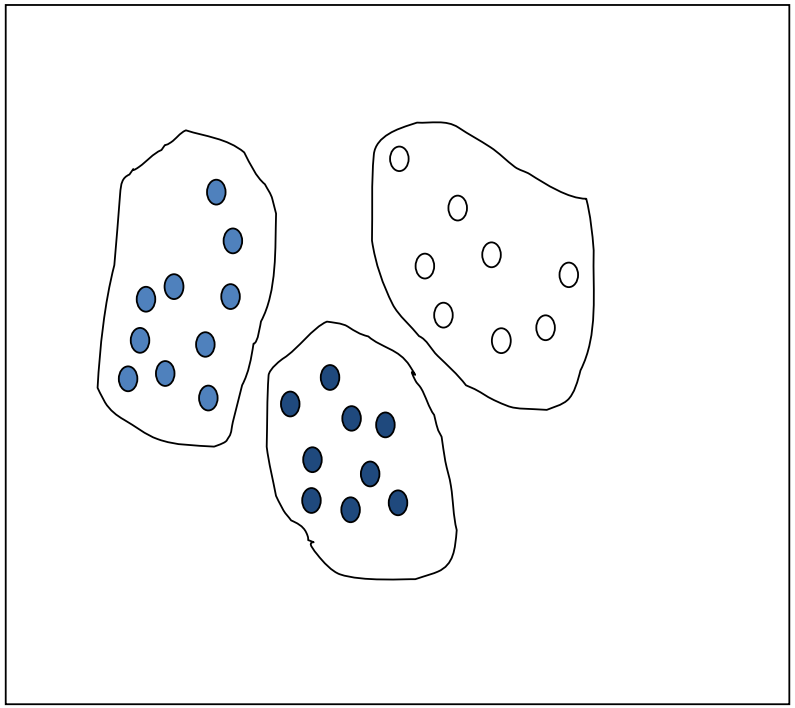
- **Cluster Sampling/Stratified Sampling**: The population is first partitioned into subpopulation, a.k.a. ***strata***, and sampling is performed seperately within each subpopulation.

Sampling: with or without Replacement

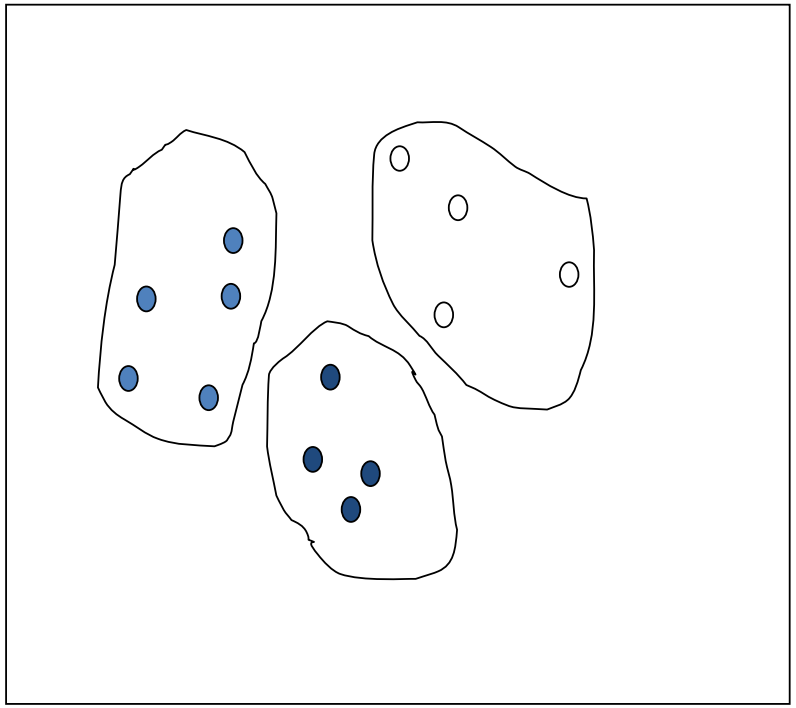


Sampling: Cluster or Stratified Sampling

Raw Data

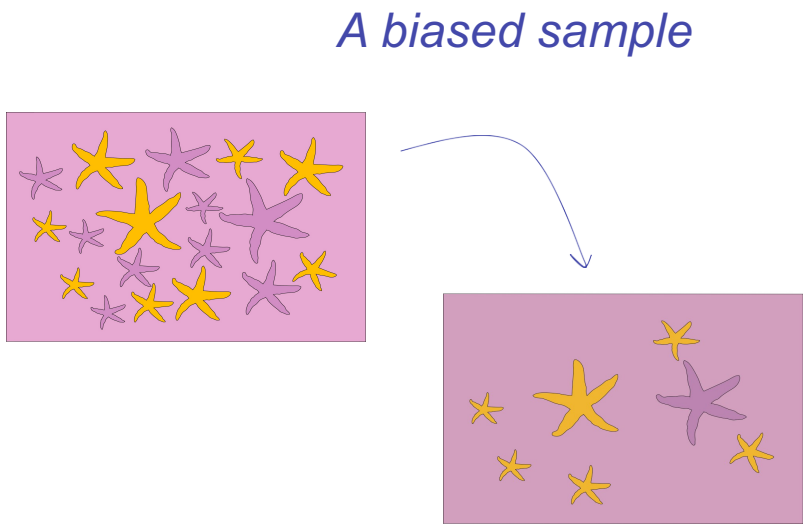


Cluster/Stratified Sample



Populations vs Samples

- ***Bias*** is a systematic discrepancy between estimates and the true population characteristic.



Sampling Bias

The 1936 US presidential election

Alf Landon Campaign Poster, 1936

Alf Landon
Republican

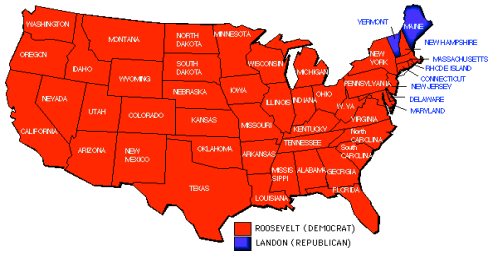
vs.

Franklin Roosevelt
Democrat

1936 *Literary Digest* Poll

- 2.4 million respondents
- Based on questionnaires mailed to 10 million people, chosen from telephone books and club lists
- Predicted Landon wins: Landen 57% over Roosevelt 43%

1936 election results



Roosevelt won with 62% of the vote

What went wrong?

- Subjects given the questionnaire were chosen from telephone books and clubs, biasing the respondents to be those with greater wealth

- Voting and party preference is correlated with personal wealth

Volunteer bias

- **Volunteers** for a study are **likely** to be **different**, on average, from the population

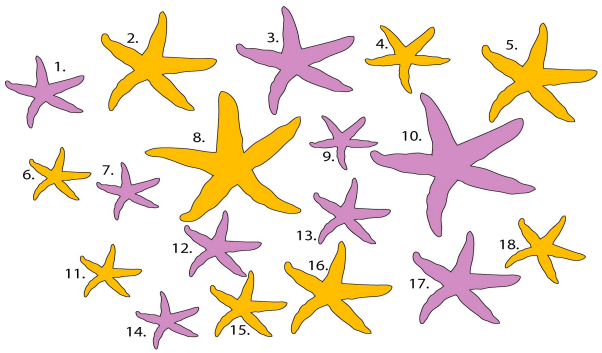
- For example:
  - Volunteers for sex studies are more likely to be open about sex
  - Volunteers for medical studies may be sicker than the general population
  - Volunteers for customer satisfaction surveys are likely to be very happy or upset about the service that they received

Properties of a good sample

- Independent selection of individuals
- Random selection of individuals
- Sufficiently large

One procedure for random sampling

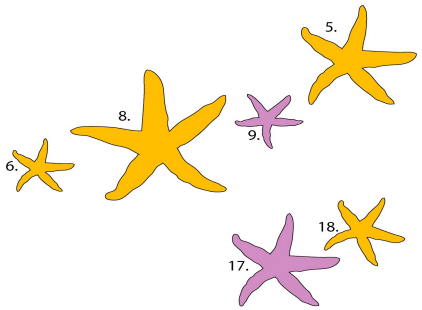
- In a random sample, each member of a population has an equal and independent chance of being selected.



Number each individual

18, 6, 8, 5, 9, 17

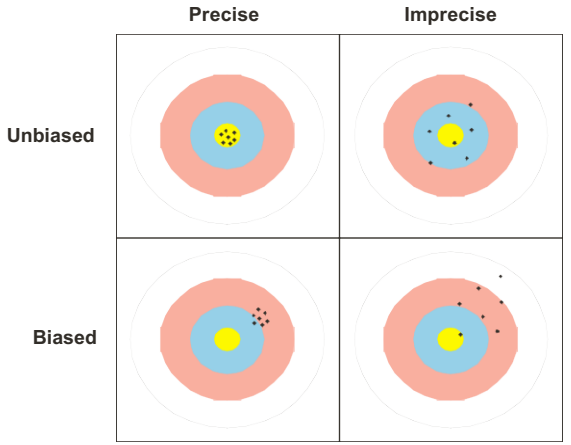
Choose random numbers



Sample those individuals with matching numbers

Sampling

- Population parameters are constants whereas estimates are random variables, changing from one random sample to the next from the same population.



Each point represents an estimate of a parameter.

Sampling error

- The difference between the estimate and average value of the estimate
- Larger samples on average will have smaller sampling error

Observational studies and experiments

- After obtaining the sample, the next step is **gathering** the **relevant information** from the selected members.
- In observational studies, researchers are **passive examiners**, trying to have the **least impact** on the **data collection** process.
- **Observational studies** are quite helpful in **detecting relationships** among characteristics.

Observational studies and experiments

- For relationships between characteristics, it is vital to **distinguish** between **association** and **causality**.
- The realationship is **casual** if one characteristic **influences the other one**.
- It is usually easier to **establish causality** by using **experiments**.
- In **experiments**, researchers attempt to **control the process** as much as possible.

Observational studies and experiments

- Observational Studies:
  - **Retrospective**: look into history
  - **Prospective**: observation over time
- Randomization and Replication
- Collecting Data
  - **Cross-Sectional**: at some fixed time
  - **Longitudinal**: follow the samples over time and repeatedly collect information and take measurements
  - **Time Series**: over a period of time. It is collected more frequently, but on smaller samples

Data exploration

- Towards statistical inference and decision making is to perform **data exploration**, which involves **visualizing and summarizing the data**.

- The **objective** of data visualization is to obtain a high level **understanding** of the sample and their observed (measured) **characteristics**.

- **Summary statistics**:

  To make the data **more manageable**, further reduce the amount of information in **meaningful ways** to **focus** on the **key aspects** of the data.

Data exploration

- Data exploration techniques to show **distribution** of a **variable**.
- Distributions tell us the **possible values** it can take, the **chance of observing** those values, and how often we expect to see them in a random sample.
- Data exploration can help **detect** previously **unknown patterns** and **relationships** that are worth further investigation.
- Can also **identify possible data issues**, such as unexpected or unusual measurements, known as outliers.

Statistical inference

- We collect data on a sample to **learn about the population**.
  - i.e., Mackowiak, et al. (1992) measure the normal body temperature for 148 people to learn about the normal body temperature for the entire population.
- In this case, we say we are **estimating** the **unknown population average**.
- However, the **characteristics** and **relationships** in the **whole population remain unknown**.
- Therefore, **there is always some uncertainty** associated with our estimations.

Statistical inference

- Mathematical tool to address **uncertainty** is **probability**.

- The **process of using data** to **draw conclusions** about whole population with a degree of uncertainty about our findings, is called **statistical inference**.

- The **knowledge acquired** from data through **statistical inference** allows us to **make decisions** w.r.t the scientific problem.

Computation

- Usually **computer programs** are used to perform most of statistical analysis and inference tasks.
- The computer programs commonly used for this purpose include SAS, STATA, SPSS, MATLAB, Python and R.
- R is free and arguably the **most common software** among **statisticians**.
- For the purpose of this course, we use **R programming language** for statistical analysis.