Lecture 14 Population, Sample and Random Sample

BIO210 Biostatistics

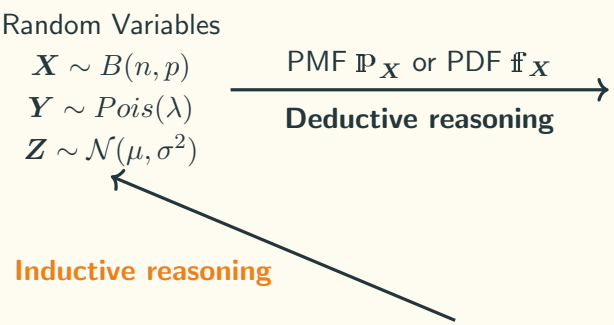
Xi Chen

Fall, 2023

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

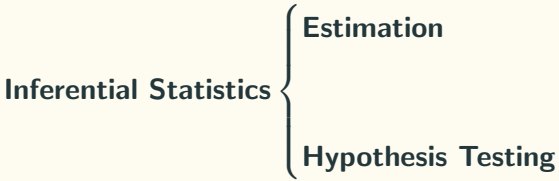
Theoretical Probability Distribution



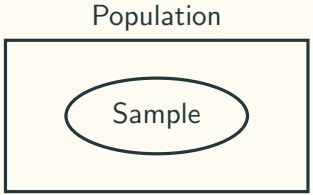
- How many heads do we expect to get after 100 flips?
- How many trains do we expect to see in an hour?
- What proportion of people are there with height > 175 cm?

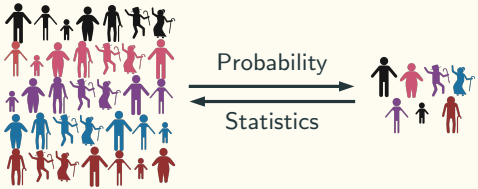
- Flips the coin for a reasonable number of times
- Count the # of train for a reasonable number of hours
- Check the height of a reasonable number of people

Inferential Statistics



Population vs. Sample





Population: all elements, individuals, cells, items, objects *etc.* of your interest. You can relate this to sample space. We often use random variables ($\boldsymbol{X}$) to describe the population.

Population distribution: $\mathbb{P}_{\boldsymbol{X}}(x)$ or $\mathbb{f}_{\boldsymbol{X}}(x)$

Sample: A subset of population to study.

Sampling: The process of generating a sample.

Population vs. Sample by AI

Prompt: `/imagine statistical concepts of population and sample, schematic view, simple and minimalism`

Population vs. Sample

	Population	Sample
mean	$\mu = \mathbb{E}[\boldsymbol{X}]$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
variance	$\sigma^2 = \mathbb{V}\mathrm{ar}(\boldsymbol{X})$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
standard deviation	$\sigma = \sqrt{\mathbb{V}\mathrm{ar}(\boldsymbol{X})}$	$s = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}}$
	Population parameters	**Sample statistics**

Examples of population and sample

Population	Sample
Advertisements for University jobs in China	The top 50 search results for advertisements for University jobs in China on Mar 1, 2021
Undergraduate students in China	Undergraduate students in Shenzhen
Undergraduate students in Shenzhen	Undergraduate students in SUSTech
All countries of the world	Countries with published data available on birth rates and GDP since 2000

Sample and sampling

- Not every sample or sampling process is appropriate

- **A good sample**: a representative sample, a micro-version of the entire population.

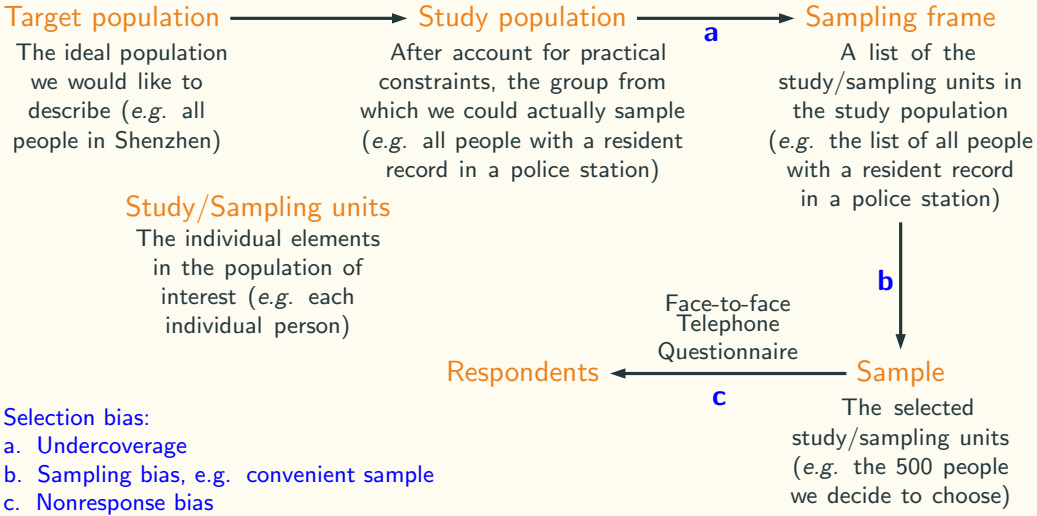
Random sampling

- **Random sample**: a random sample is a selection of some members of the population such that each member is independently chosen and has a known nonzero probability of being selected.

- **A simple random sample**: a simple random sample is a random sample in which each group member has the same probability of being selected.

- Most often, we rely on computer programs to generate (pseudo-)random number for us.

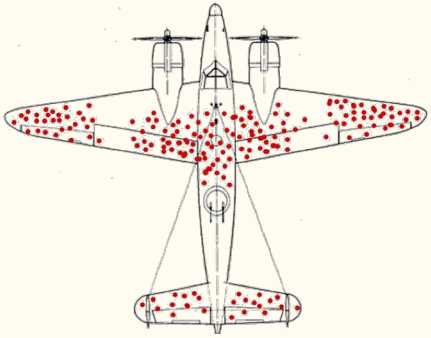
Random sampling procedures

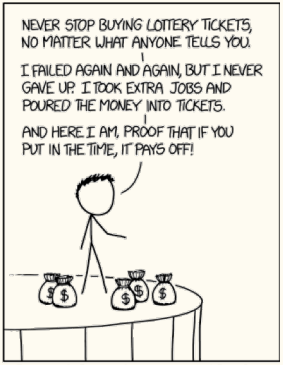


Selection bias:

a. Undercoverage
b. Sampling bias, e.g. convenient sample
c. Nonresponse bias

Survivorship bias





xkcd.com

Nature 332, 586–587 (1988)

Animal behaviour

Why cats have nine lives

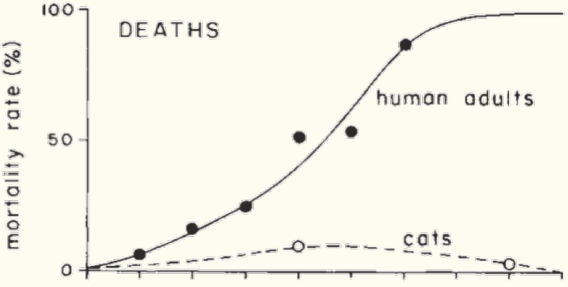
Jared M. Diamond

THE famous adage that cats have nine lives stems in part from their ability to survive falls lethal to most people. This phenomenon has not received the scientific attention that it deserves. Filling this lacuna, a new study by W.O. Whitney and C.J. Mehlhoff (*J. Am. Vet. Med. Assoc.* **191**, 1399–1403; 1987) applies principles of anatomy, physics and evolutionary biology to falling cats.

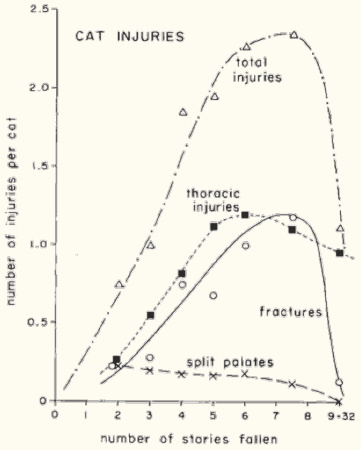
The authors were veterinarians at an animal hospital in New York City, where skyscrapers, open windows and paved ground combined to generate a database of 132 cats injured by falls of 2 or more stories, with a maximum of 32 stories and a mean of 5.5 ± 0.3 (s.e.m.) (1 storey = 15 feet). Most victims landed on concrete after a free-fall. Omitting 17 cats that were euthanized by owners unable to afford treatment, 90 per cent of the cats (104 of 115) survived, whereas 11 died (mainly because of thoracic injuries and shock). The most remarkable feature of the results (see figure) is that incidence both of injuries and of mortality peaked for falls of around seven stories and decreased for falls from greater heights. For instance, the cat that free-fell 32 stories onto concrete was released after 2 days of observation in the hospital, having suffered nothing worse than a chipped tooth and mild pneumothorax.

Falling adult humans differ from falling cats in their much higher mortality rate, monotonic mortality/height relation, different causes of death, and different sub-lethal injuries (Warner, K.G. & Demling,

Survivorship bias



Nature 332, 586–587 (1988)



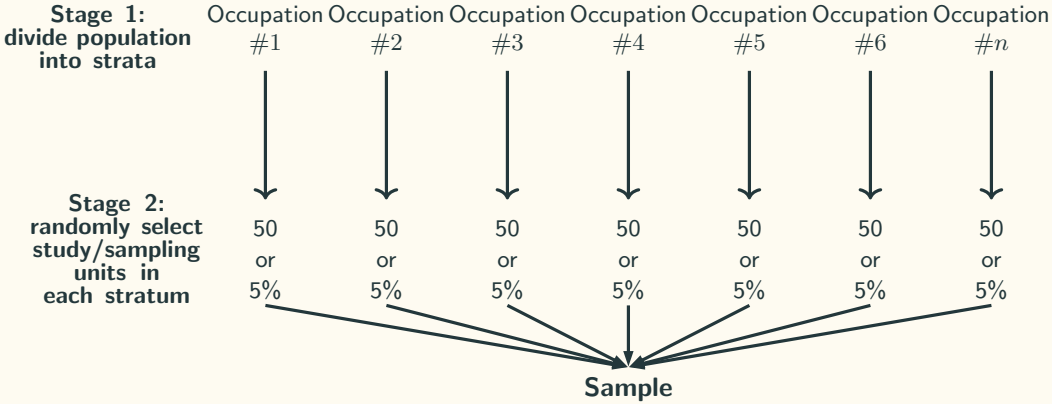
Random multi-stage clustering sampling

A Random multi-stage cluster sampling:

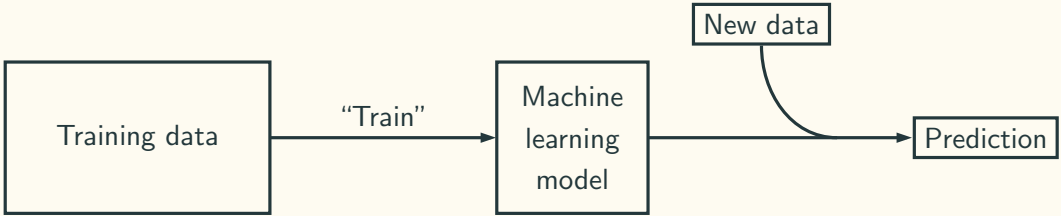
Stage 1: divide population into clusters	Police Station $\#1$	Police Station $\#2$	Police Station $\#3$	Police Station $\#4$	Police Station $\#5$	Police Station $\#6$	Police Station $\#n$
Stage 2: randomly select clusters			10 police station				
Stage 3: choose all study/sampling units in the selecterd clusters			**Sample**				

Stratified Sampling

A stratified sampling:



Stratified K-fold cross validation in supervised machine learning:

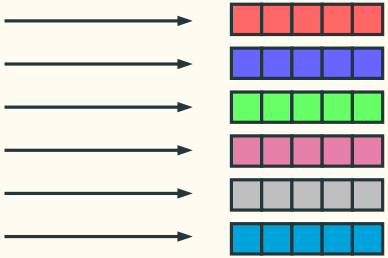


Stratified K-fold Cross Validation

Cellular compsition of a mouse spleen

Cell types	Abundance
B cells	57.9%
Dendritic cells	2.0%
Macrophages	1.7%
Nature Killer Cells	4.4%
Neutrophils	2.7%
T cells	31.3%

Split each cell types into 5 pieces



Randomised Clinical Trials

- **A randomised clinical trial** is a type of research design used for comparing different treatments, in which patients are assigned to a specific treatment by some random mechanism. The process of assigning treatments to patients is called **randomisation**.

- **Example**: Aminoglycosides to treat gram-negative organisms in patients. For several decades, studies have been performed to compare the efficacy and safety of different aminoglycosides. The earliest studies were nonrandomized studies. No random mechanism was used to assign treatments to patients. The more effective antibiotic might actually perform worse because this antibiotic is prescribed more often for the sickest patients.

Randomised Clinical Trials

- **Block randomisation** is defined as follows in clinical trials comparing two treatments (treatments A and B). A block size of 2n is determined in advance, where for every 2n patients entering the study, n patients are randomly assigned to treatment A and the remaining n patients are assigned to treatment B.
- **Blindness**: A clinical trial is called **double blind** if neither the physician nor the patient knows what treatment he or she is getting. A clinical trial is called **single blind** if the patient is blinded as to treatment assignment but the physician is not. A clinical trial is **unblinded** if both the physician and patient are aware of the treatment assignment.

Sample size

Bigger is always better.

Bigger sample size does not compensate for bad sampling.