

# Chapter I

# 1. Introduction to Statistics

## 1.1. What is statistics

### Review

#### Exercise 1:

Probability or statistics

Determine whether each of the problems below is a probabilistic or a statistical problem. (You are **not** asked to solve them.)

Assume we have a population consisting of two subpopulations, A and B. A particular drug has a different chance of treatment success depending on the subpopulation, namely 70% for group A and 50% for group B.

Assume that subgroup A is 10% of the entire population and subgroup B is 90%. What is the chance of a successful treatment if we pick a random person from the entire population?

This is a

Probabilistic problem

Now, consider the scenario where we do not know the true composition of the population, which may be different from the previous setup. Among 1000 randomly chosen patients, we observe that the treatment was successful in 700 of them. What is a good estimate of the composition of the population?

This is a

Statistical problem

John Arbuthnot wrote a paper in 1710 entitled 'An Argument for Divine Providence', where he studied, based on the Christening records in London



## 1. Introduction to Statistics

---

for 1629-1710, the chances that a randomly chosen baby born is a girl or a boy. Is this a statistical problem, or a probabilistic problem?

Statistical problem

Next, you read Arbuthnot's paper, and went to a gyneacology facility, in which there are 1010 babies whom are expected to be born on the day you arrived, and you are interested in, what are the odds that 6 of those will be a boy, and the remaining will be a girl. Is this a statistical problem, or a probabilistic problem?

Probabilistic problem

A doctor realizes that there is an allergy medicine which is effective in treating seasonal allergies with probability at least 90%. From here, he claims:

Out of 100 patients admitted to clinic with seasonal allergies, this drug will cure 90 patients, on average.

At least 70 patients will be cured, with 99.99% chance.

Does he rely on statistics, or probability?

Probability

Now, a newly-hired scientist at a pharmacology company performs an experiment, and based his observations, deduces that, "I am 95% confident that if we repeat this experiment, then the drug will be effective on between 85% and 95% patients." Does he rely on statistics, or probability?

Statistics

### Exercise 2:

Assume that we observe three draws,  $X_1, X_2, X_3$  from a Bernoulli distribution with parameter  $p = 0.5$ . For example, imagine that in the model for the preferred head direction for kissing, either direction was actually equally likely and we observed three kissing couples.

What is the probability of observing at least two ones, i.e., what is

$$P\left(\sum_{i=1}^3 X_i \geq 2\right)$$

?



If in the model above, let us assume we decided to consider two or more right-turns as significant evidence for a predisposition of this direction for kissing. Now, 10 students go out and each observe three different couples kissing. How many of them would on average come to the conclusion that right-leaning is more common than left-leaning when kissing?

5

In a group of  $nn$  people indexed 1 through  $n$ , each pair  $(i,j)$  (there are  $(n 2)$  of them) are either friends, or not friends. To model this situation, we assign a random variable to each pair. Which one of the probability distributions below is the most appropriate model?

A Bernoulli Random Variable

With the setup as in the problem above, we say a group of four people is "interesting", if there are at most five pairs who are friends. Assume that each pair of people are friends, independent of every other pair, with probability 0.5. Let  $N$  be the number of pairs that are friends in this group.

Binomial,  $\frac{63}{64}$ 

Following the model above, if 128 different people each observe one randomly chosen groups of four people, how many times on average do these observations lead to the conclusion that the person's chosen group is interesting?

126

## Exercise 3:

Consider a probabilistic experiment where we roll a dice and toss a coin. We compute the probability that the fair dice gives 5 and the fair coin lands Heads. What assumptions are we implicitly using in this specific calculation:  $\Pr Pr (5, Heads) = \Pr Pr (5) \cdot \Pr Pr (Heads) = \frac{1}{6} \cdot \frac{1}{2}$ ? Choose all that apply, so that the chosen assumptions best capture the required concepts.

- Each dice roll is uniformly distributed within the set  $\{1,2,3,4,5,6\}$  and each coin toss is uniformly distributed in  $\{Heads,Tails\}$
- The dice roll and coin toss are independent.
- The random variables corresponding to outputs of each of these experiments are i.i.d.



## Population vs Samples

**A population data set** contains all members of a specified group (the entire list of possible data values). [Utilizes the count  $n$  in formulas.]

Example: The population may be "ALL people living in the US."

**A sample data set** contains a part, or a subset, of a population. The size of a sample is always less than the size of the population from which it is taken. [Utilizes the count  $n - 1$  in formulas.]

Example: The sample may be "SOME people living in the US."

We denote the sample average, or sample mean, of  $n$  random variables  $X_1, \dots, X_n$  by :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

### Exercise 4:

What happens with our assumption of the observations of the kiss orientation being i.i.d. Bernoulli if we assume that the preferred orientation changes with the time of day?

- a) The observations will always be dependent, so it is violated
- b) We will have to be more careful about how we collect observations
- c) No matter how we sample, we will still have i.i.d. observations

Answer b

