# Chapter 2: Statistical Foundations

Nischal Regmi

Everest Engineering College

2022

# Variable Types

1. **Numeric variables**: Values can be numbers
   - 1.1 **Continuous variable**: Values are real numbers (or, complex numbers sometimes)
   - 1.2 **Discrete variables**: Values are countable (usually, integers)
2. **Categorical variable**: Values can be one among a fixed number of possible values, conceptually representing a category
   - 2.1 **Nominal variable**: There is no order (greater-equal-less relation) between the categories
   - 2.2 **Ordinal variable**: There is an order between the categories

**Classwork**: Give real-life examples for each type of variables

# Understanding Variable Distributions

# Mean, Median, Mode

Consider a random variable $X \in \mathbf{R}$ with $N$ observed values $\{x_1, x_2, \ldots, x_N\}$.

- **(Sample) Mean**

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- **Median**
  - The median is the middle value when the measurements are arranged from lowest to highest. (What if $N$ is odd?)
- **Mode**
  - *Mode* is the highest value of the distribution function. Intuitively, the mode is the value that appears most often.

# Distribution Function

The distribution function for a random variable $X$ is defined as

$$F(x) = P(X \leq x) \qquad x \in \mathbb{R}$$

where $P$ denotes the 'probability of'.
A distribution function has the following properties

1. $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$
2. $F(X)$ is non-decreasing, i.e. if $x_1 \leq x_2$ then $F(x_1) \leq F(x_2)$
3. $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$

**Note**: Some authors call the function $F(x)$ as the 'cumulative distribution function'.

# Distribution - The Continuous Case

- For a continuous random variable $X$, the probability that its value is between $a$ and $b$ is

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

where function $f(x)$ is a non-negative function called the **probability density function**. In addition, note that $P(X = x) = 0$.

- The density function is related with the distribution function in the following way. $P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$

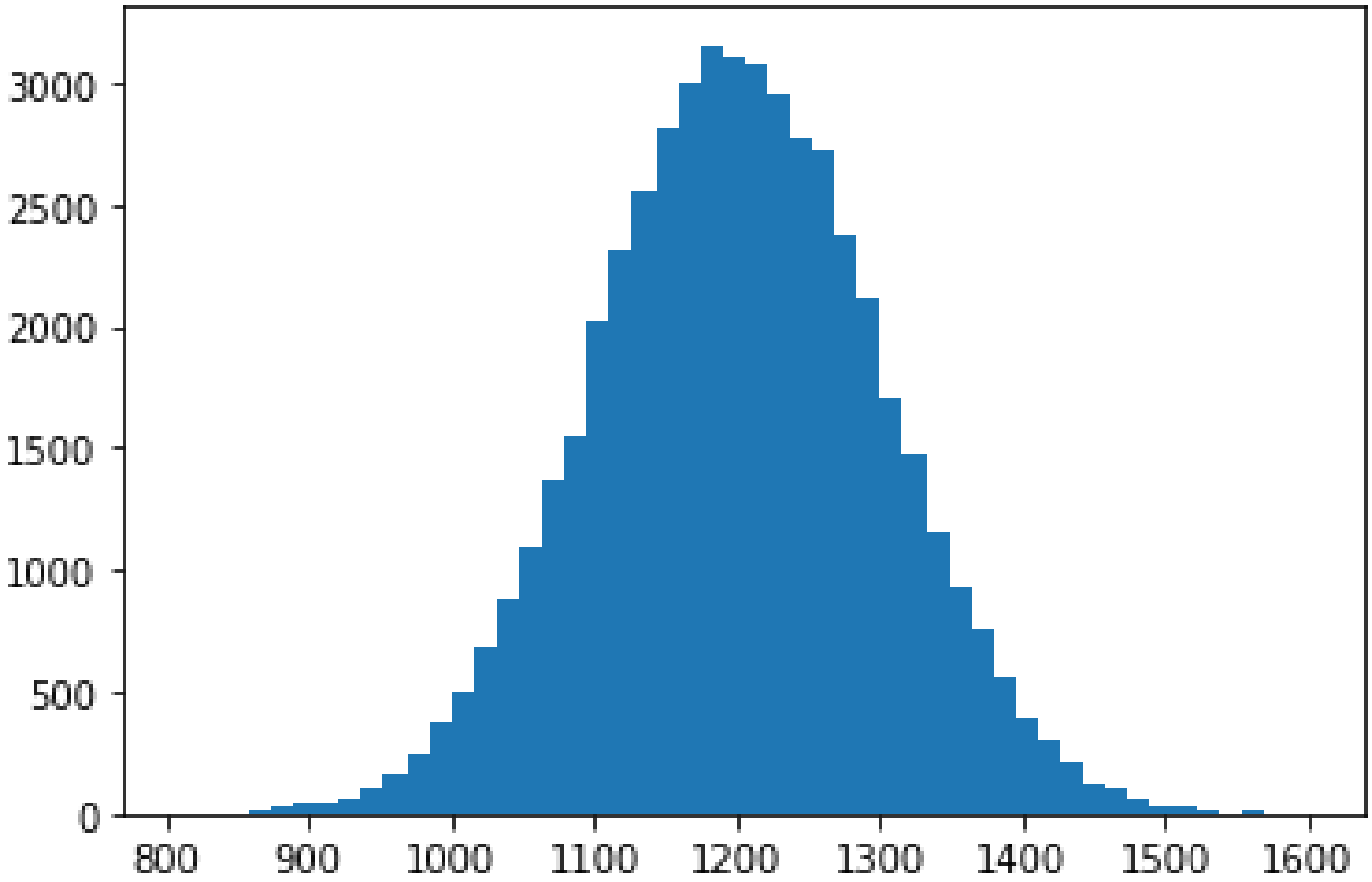



Figure: Empirical distribution for a normal random variable with mean 1200 and standard deviation 100. Sample size is 50,000

**Classwork**: Real-life examples please?

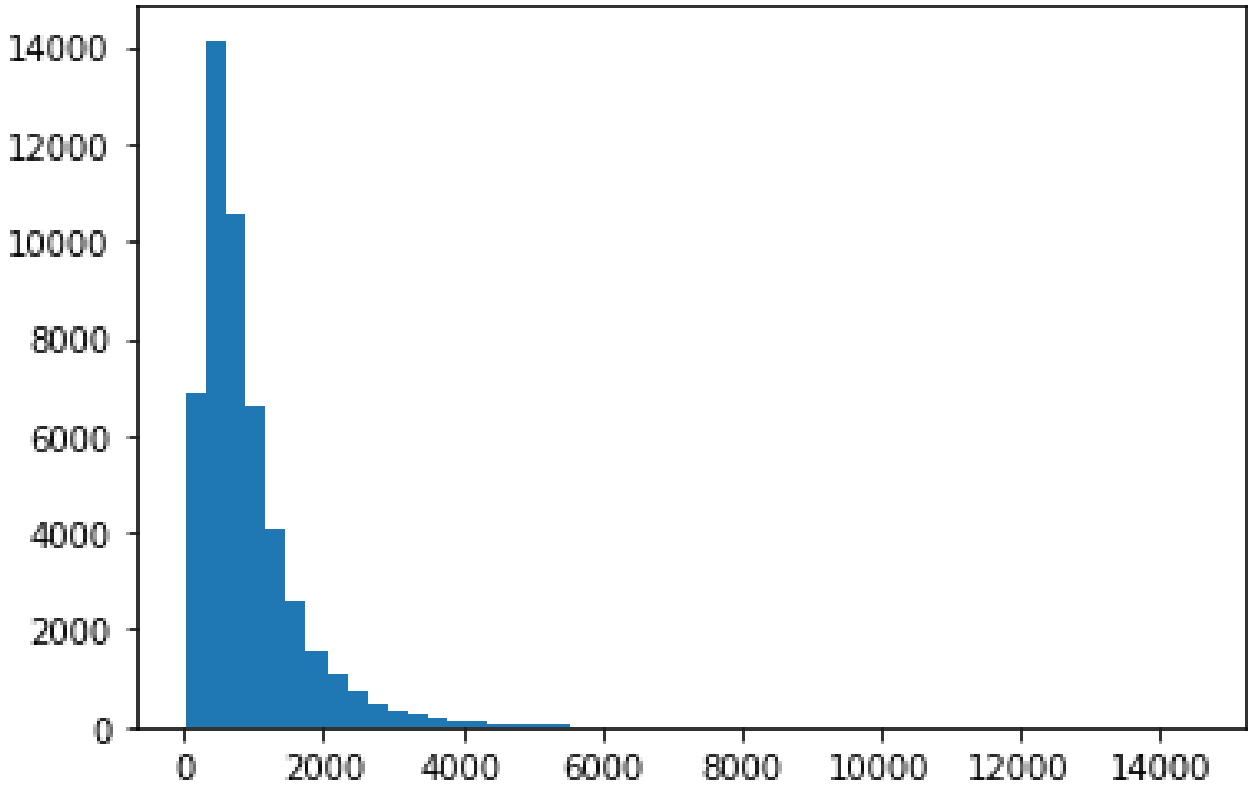



Figure: Empirical distribution for a lognormal distribution corresponding to a sample with mean 1200 and median 700. Sample size is 50,000. The highest value in the sample is 14,485

**Note:**

- $X$ has a *lognormal distribution* if $Y = \ln X$ has a normal distribution.

**Classwork**: Real-life examples please?

# Skewness

Did you understand skewness from the preceding discussions?

- A distribution is *skewed* if its mean differs significantly from its median.
- If the left tail of the distribution is longer, we call it a negatively skewed (or left skewed) distribution. In this case

$$\text{mean} < \text{median} < \text{mode}$$

- If the right tail of the distribution is longer, we all it a positively skewed (or right skewed) distribution. In this case

$$\text{mode} < \text{median} < \text{mean}$$

- If both the left and right tails are equal, the distribution is symmetric. In this case

$$\text{mean} = \text{median} = \text{mode}$$

**Note**: I am skipping the formal definitions of various skewness measures. You can refer to standard statistics textbooks if needed for the practicals.

# Distribution - the Discrete Case

- A discrete random variable can have countable number of values. e.g. $X =$ the total number of vehicles crossing Pulchowk bridge between 6 to 7 AM in a day.
- **Probability Mass Function**: Suppose a discrete random number $X = \{x_1, x_2, x_3, \ldots\}$. The the probability mass function is defined as

$$p(x_i) = P(X = x_i) \qquad i = 1, 2, 3, \ldots$$

We must have

$$\sum_{i=1}^{\infty} P(x_i) = 1$$

- Suppose that $I = [a, b]$ is an interval and $X$ is a discrete random variable. Then

$$P(X \in I) = \sum_{a \leq x_i \leq b} p(x_i)$$

- The distribution function for a discrete random variable $X$ is given by

$$F(x) = \sum_{x_i \leq X} p(x_i) \qquad -\infty < x < \infty$$

- Concepts of skewness etc. for the discrete as well as the categorical variables are similar to the continuous case. See the following slide.

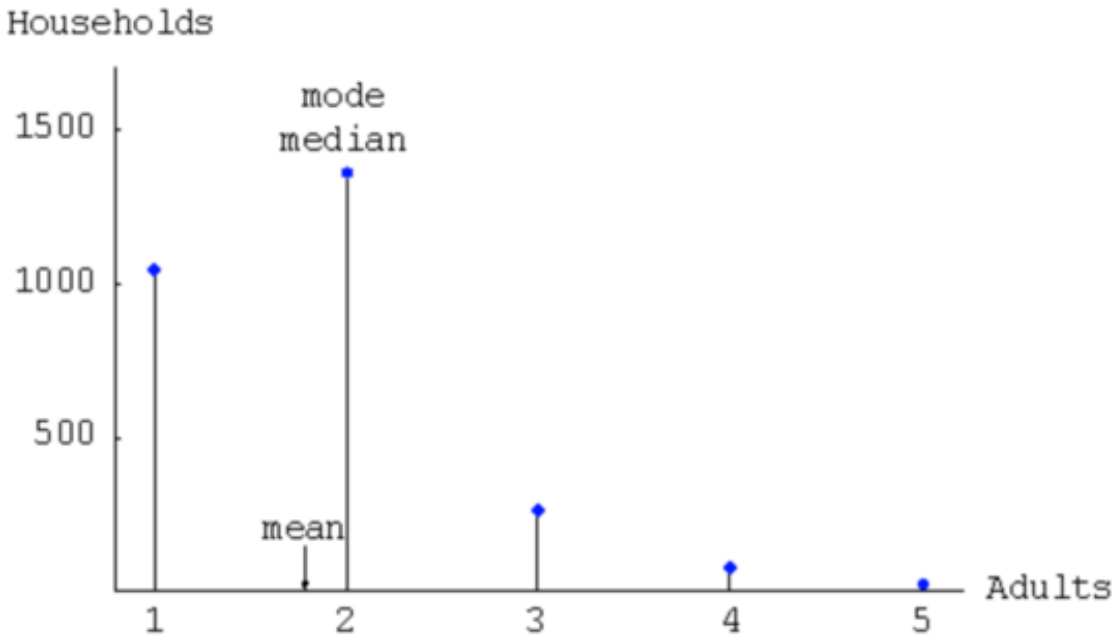



Figure: Distribution of adult members in US households (Figure source: Wikipedia). It is an example of a right-skewed discrete distribution

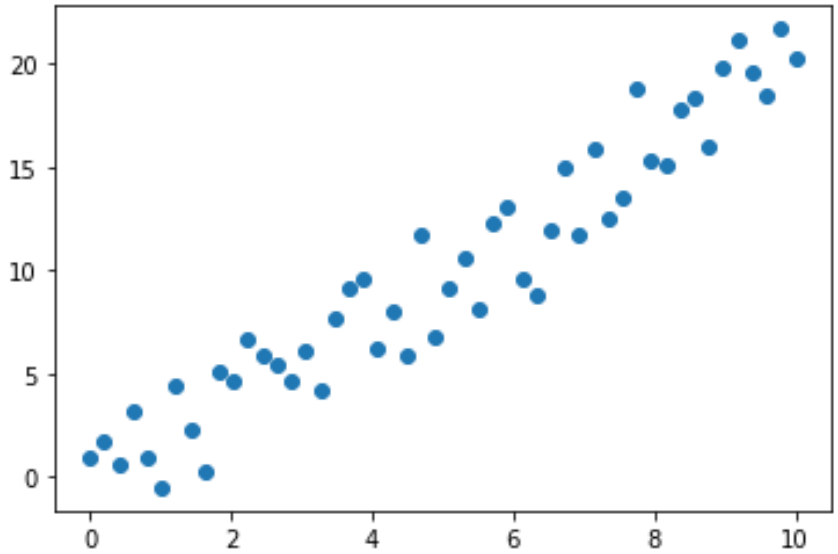
# Inferring Preliminary Relationships between Variables

# Relationship between Numeric variables

- The basic measure that defines the strength of two numeric variables is the correlation coefficient.
- There are different types of correlation coefficients. The most usual is the Karl Pearson's correlation coefficient, defined for a sample as following.

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

- It should be noted that Pearson's correlation coefficient is a linear measure. Two variables may be non-linearly associated, but have a low value of Pearson correlation.



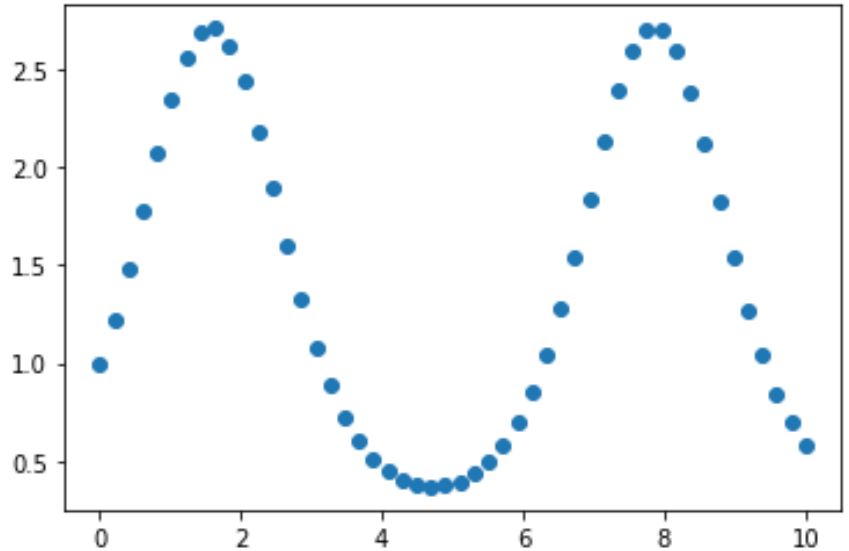




Figure: The Karl Pearson correlation correlation between $x$ and $y$ in the left figure is 0.96, which is high, as well as obvious from the figure. However, although there is a precise relationship ($y = \sin x$) in the right figure, correlation is -0.09, which is extremely low. Thus, Pearson correlation coefficient only identifies a linear relationship.

# Relationship between Categorical Variables

- For categorical variables in general, a good strategy of observing correlation between is *cross tabulation*. Below example shows a cross-tab between 'food preferences' and gender.

| | Female | Male | Other |
|---|---|---|---|
| Likes buff momo | 8 | 75 | 1 |
| Likes chicken momo | 63 | 3 | 0 |
| Likes veg momo | 23 | 19 | 3 |

- If single entry in each of the rows of the cross-tab is dominantly higher than others in the same rows, it is likely that the variables corresponding to rows and column are correlated.

- In above example, for the first row, the 2nd entry is dominantly higher than the remaining two.
- Similarly, for the second row, the first entry is dominantly higher than the other two entries.
- However, the first and second entries in the third row are not so different.
- It is plausible from the above table that food preference and gender are correlated.
- A mere observation of the cross-tab may not provide satisfactory answers all of the time. We need to resort to the statistical procedures, mainly, the chi-squared test.
- For ordinal variables, one could use the Spearman's Rank correlation and Kendall's Tau – (please read about them yourselves)

# Additional Topics to Ponder Upon

# Statistical Significance

- Conceptually, statistical significance indicates the reliability of inferences or estimates that we make from data.
  - e.g. We survey 100 persons comprising equally males and females, and ask if they like buff momo. How can we assess that male and female have different preferences regarding buff momo?
- Statistical significance can be calculated only by assuming that a certain hypothesis is true.
  - In the previous example, we could begin with the hypothesis that both gender have same food preference, and test whether the data we have is consistent with our hypothesis or not.
- In practice, parametric indicators like t-value/p-value, and non-parametric techniques like bootstrapping, both are used for assessing statistical significance.
- As a rule of thumb, inferences/estimates become more and more statistically reliable with the increase in sample size.

# Statistical Significance – Example

Suppose you survey 5 male students how much they spend per day on refreshments. You found that, on an average, males spend Rs. 730. Suppose the sample standard deviation for your sample is 200. Also assume that from a previous study, it is almost certain that female students spend Rs. 810 per day on an average. We ask the question: "*Is the per day expense of male students different from the female?*"

*The Big Point*: The reliability can only be assessed assuming a null hypothesis. So, let us assume the null hypothesis that mean expenses of male and female are equal. We have

$$\mu_{\text{male}} = 730$$
$$\mu_{\text{female}} = 810$$
$$n = 5 \qquad \text{(sample size)}$$
$$s = 200$$
$$\text{H0: } \mu_{\text{male}} = \mu_{\text{female}}$$

We now calculate the $t$-statistics

$$t = \frac{\mu_{\text{male}} - \mu_{\text{female}}}{s/\sqrt{n}} = -0.89$$

As per the rule of thumb, t-value should be larger than 1.8 (without considering the sign). Since our t-value is -0.89, we cannot conclude that if male and female expenses are different.

Suppose that our sample size $n$ is now 100, and all the other data are same. Then, we would get $t$-value equal to $-4$. In this case we would reject the null hypothesis and conclude that the male and female expenses are different.

**Note**: In the t-test, there is an implicit assumption that sample means are normally distributed.

# Stability of Algorithms

- If a model estimated using an algorithm data differs significantly with slight change in the data, the algorithm is said to be *unstable* for the data we have.
- Stability is a result of both the data and algorithm. In other words, a same algorithm may produce stable output for a dataset while unstable output for another dataset.
- Algorithms for inferring decision trees from data are examples of unstable algorithms.
- *Stability analysis* is thus an important step in machine learning applications and should not be ignored.
- In machine learning, there is a broad class of techniques called *regularization* that can be used to deal with unstable models/data.

# Imbalanced Data

- Data is imbalanced if the sample does not sufficiently represent each of categories in the population.
- For example, in demographic surveys, it is usual that the sample consisted of relatively low number of representatives for minority ethnic groups.
- Another example, if you collect image dataset for classifying vehicles by randomly taking pictures in a street in Kathmandu, it is very likely that your sample will be imbalanced. The number of bikes would be very large whereas the number of heavy-load trucks/lorries would be very few.
- Data imbalance make the inferred results unreliable (why?)
- Adjusting the effects of data imbalanced is difficult and controversial topic. There are several books dedicated confined solely to the question of data imbalance!