

Introduction to Descriptive Statistics and Probability for Data Science

- Sandeep Chaurasia

Content

- Introduction
- Measure of Central Tendency (Mean, Mode, Median)
- Measures of Variability (Range, IQR, Variance, Standard Deviation)
- Probability (Bernoulli Trials, Normal Distribution)
- Central Limit Theorem
- Z scores

Descriptive Statistics

- Statistics has become the universal language of the sciences, and data analysis can lead to powerful results.
 - Has there been a significant change in the mean sawtimber volume in the red pine stands?
 - Has there been an increase in the number of invasive species found in the Great Lakes?
 - What proportion of white tail deer in New Hampshire have weights below the limit considered healthy?
 - Did fertilizer A, B, or C have an effect on the corn yield?

Statistics is the science of collecting, organizing, summarizing, analyzing, and interpreting information.

Cont.

- Good statistics come from good samples, and are used to draw conclusions or answer questions about a population

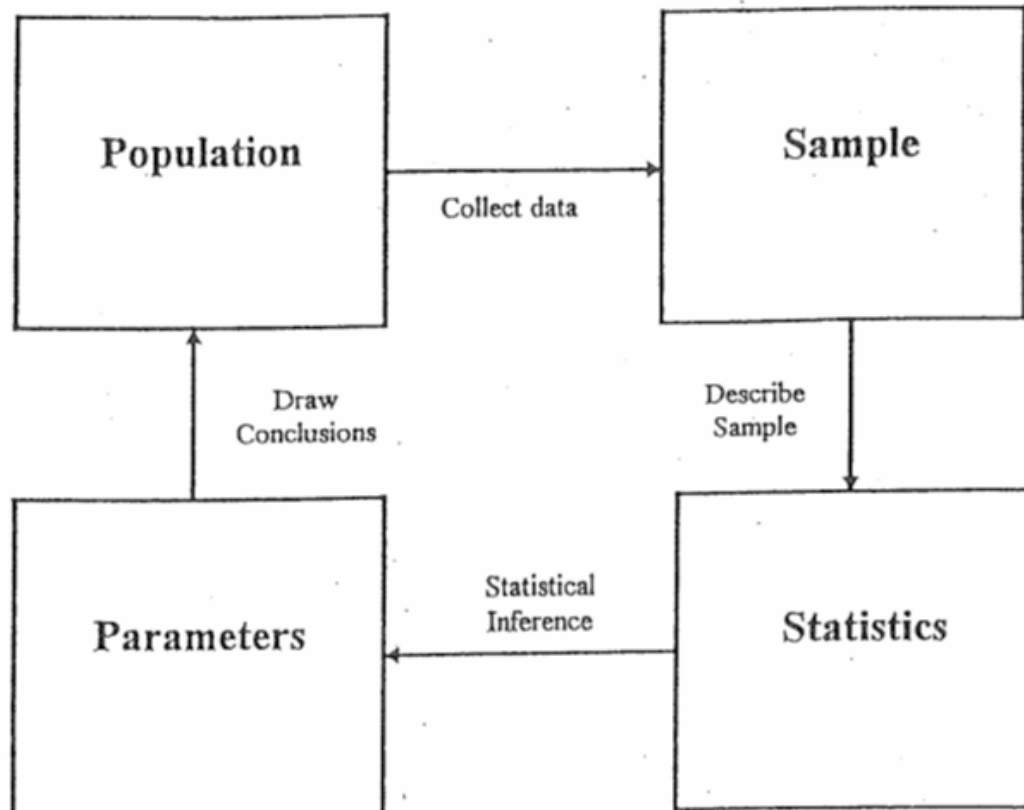


Fig. Using sample statistics to estimate population parameters.

Descriptive Statistics

- A population is the group to be studied, and population data is a collection of all elements in the population. For example:
 - All the fish in Long Lake.
 - All the lakes in the Adirondack Park.
 - All the grizzly bears in Yellowstone National Park.
- A sample is a subset of data drawn from the population of interest. For example:
 - 100 fish randomly sampled from Long Lake.
 - 25 lakes randomly selected from the Adirondack Park.
 - 60 grizzly bears with a home range in Yellowstone National Park.

Descriptive Statistics

- Populations are characterized by descriptive measures called parameters. Inferences about parameters are based on sample statistics. For example, the population mean (μ) is estimated by the sample mean (\bar{x}). The population variance (σ^2) is estimated by the sample variance (s^2).
- Variables are the characteristics we are interested in. For example:
 - The length of fish in Long Lake.
 - The pH of lakes in the Adirondack Park.
 - The weight of grizzly bears in Yellowstone National Park.

Variables

- Variables are divided into two major groups: qualitative and quantitative. Qualitative variables have values that are attributes or categories. Mathematical operations cannot be applied to qualitative variables.
 - Examples of qualitative variables are gender, race, and petal color.
 - Examples of quantitative variables are age, height, and length.
- Quantitative variables can be broken down further into two more categories: discrete and continuous variables

Cont.

Is the variable qualitative or quantitative?

Species

Weight

Diameter

Zip Code

- Descriptive measures of samples are called statistics and are typically written using Roman letters. The sample mean is \bar{x}
- The sample variance is s^2 and the sample standard deviation is s . Sample statistics are used to estimate unknown population parameters.

Measures of Center

Mean

The arithmetic mean of a variable, often called the average, is computed by adding up all the values and dividing by the total number of values.

The sample mean is usually the best, unbiased estimate of the population mean. However, the mean is influenced by extreme values (outliers) and may not be the best measure of center with strongly skewed data. The following equations compute the population mean and sample mean.

$$\mu = \frac{\sum x_i}{N}$$

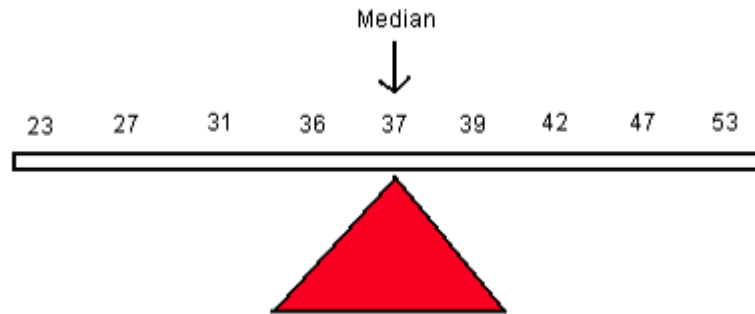
$$\bar{x} = \frac{\sum x_i}{n}$$

where x_i is an element in the data set, N is the number of elements in the population, and n is the number of elements in the sample data set.

- Find the mean for the following sample data set: 6.4, 5.2, 7.9, 3.4

Median

- The median of a variable is the middle value of the data set when the data are sorted in order from least to greatest. It splits the data into two equal halves with 50% of the data below the median and 50% above the median.

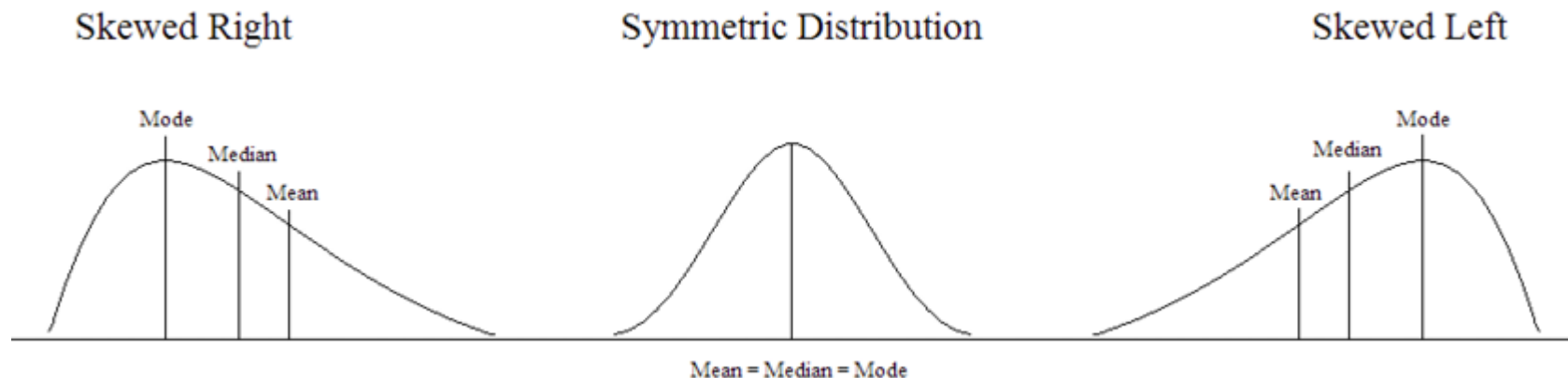


#1 : 23, 27, 29, 31, 35, 39, 40, 42, 44, 47, 51

#2 : 23, 27, 29, 31, 35, 39, 40, 42, 44, 47

Mode

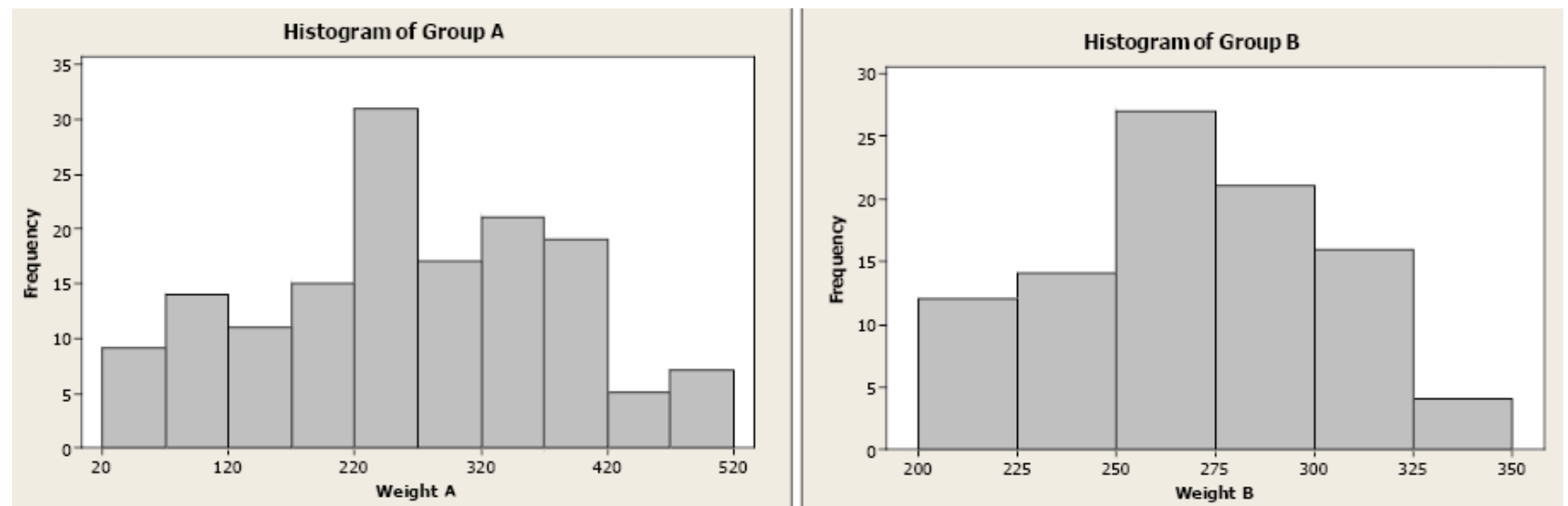
- The mode is the most frequently occurring value and is commonly used with qualitative data as the values are categorical. Categorical data cannot be added, subtracted, multiplied or divided, so the mean and median cannot be computed.
- Understanding the relationship between the mean and median is important. It gives us insight into the distribution of the variable. For example, if the distribution is skewed right (positively skewed), the mean will increase to account for the few larger observations that pull the distribution to the right.



Measures of Dispersion

- Measures of center look at the average or middle values of a data set. Measures of dispersion look at the spread or variation of the data. Variation refers to the amount that the values vary among themselves. Values in a data set that are relatively close to each other have lower measures of variation. Values that are spread farther apart have higher measures of variation.

Examine the two histograms below. Both groups have the same mean weight, but the values of Group A are more spread out compared to the values in Group B. Both groups have an average weight of 267 lb. but the weights of Group A are more variable.



Range

- The range of a variable is the largest value minus the smallest value. It is the simplest measure and uses only these two values in a quantitative data set.
- Find the range for the given data set: 12, 29, 32, 34, 38, 49, 57

$$\text{Range} = 57 - 12 = 45$$

Variance

The variance uses the difference between each value and its arithmetic mean. The differences are squared to deal with positive and negative differences. The sample variance (s^2) is an unbiased estimator of the population variance (σ^2), with $n-1$ degrees of freedom.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

- Compute the variance of the sample data: 3, 5, 7.

Standard Deviation

The standard deviation is the square root of the variance (both population and sample). While the sample variance is the positive, unbiased estimator for the population variance, the units for the variance are squared. The standard deviation is a common method for numerically describing the distribution of a variable.

Sample standard deviation

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

- Compute the standard deviation of the sample data: 3, 5, 7

Standard Error of the Means

If we want to estimate the heights of eighty-year-old cherry trees, we can proceed as follows:

- Randomly select 100 trees
- Compute the sample mean of the 100 heights
- Use that as our estimate

We want to use this sample mean to estimate the true but unknown population mean.

- Sample 1—we compute sample mean \bar{x}
- Sample 2—we compute sample mean \bar{x}
- Sample 3—we compute sample mean \bar{x}

The sample mean (\bar{x}) is a random variable with its own probability distribution called the sampling distribution of the sample mean. The distribution of the sample mean will have a mean equal to μ and a standard deviation equal to $\frac{s}{\sqrt{n}}$

The standard error $\frac{s}{\sqrt{n}}$ is the standard deviation of all possible sample means

Example #1 :

5 Students in a college were selected at random and their ages were found to be 18, 21, 19, 20 and 26.

- a) calculate the standard deviation of the ages in the sample
- b) calculate the standard error

#2

In a certain university, the mean age of a student is 20.5 with a standard deviation of 0.8.

- a) calculate the standard error of the mean if a sample of 25 students were selected
- b) what would be the standard error of the mean be if a simple sample of 100 students were selected

Coefficient of Variation

To compare standard deviations between different populations or samples is difficult because the standard deviation depends on units of measure. The coefficient of variation expresses the standard deviation as a percentage of the sample or population mean.

$$\frac{\sigma}{\mu} * 100 \quad \frac{s}{\bar{x}} * 100$$

Ex: Store wait time in minutes

Store1	6.5	6.7	6.8	7.2	7.3	7.4	7.9
Store2	4.2	5.4	6.2	7.7.	8.4	9.2	9.8

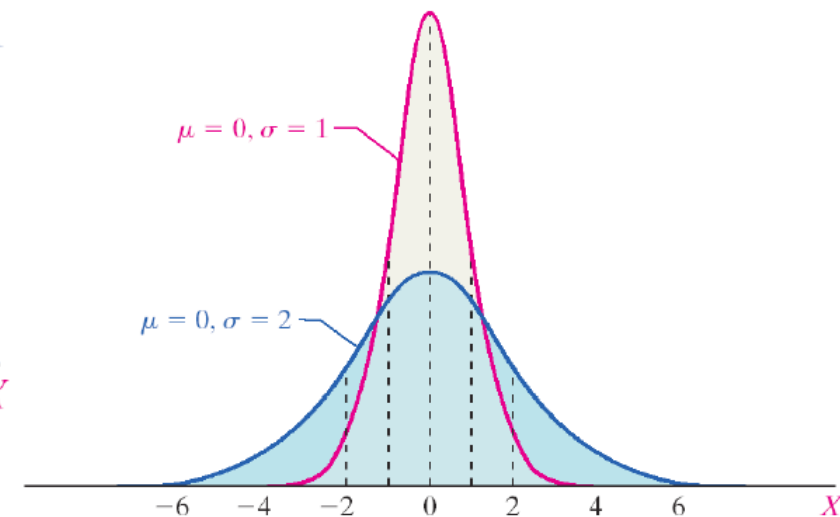
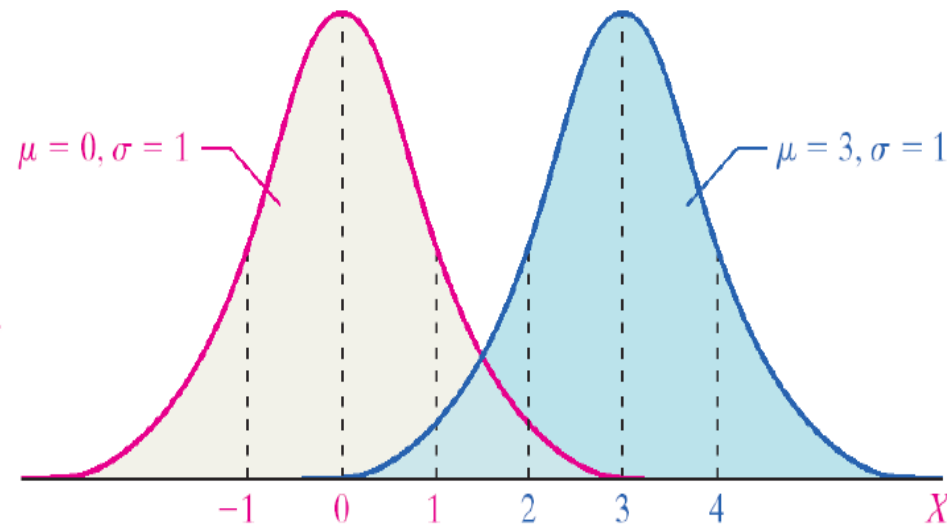
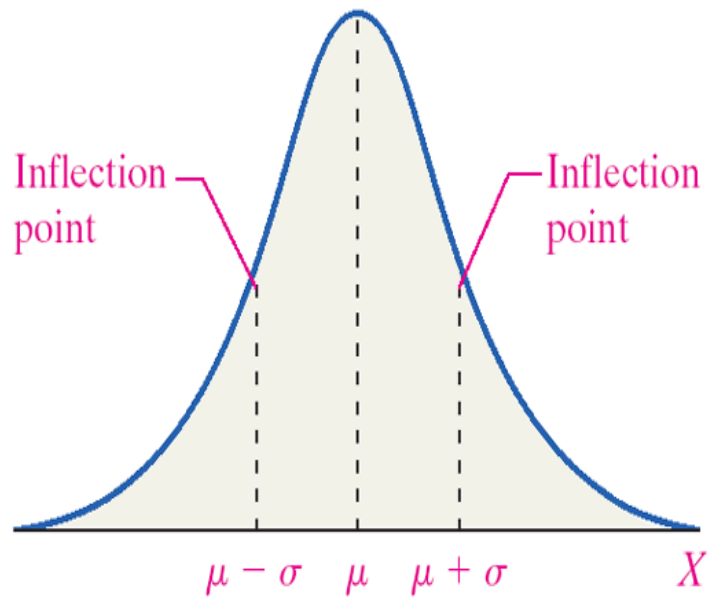
ID	X_i	X_i^2	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	25	625	-7.27	52.8529
2	35	1225	2.73	7.4529
3	55	3025	22.73	516.6529
4	15	225	-17.25	298.2529
5	40	1600	7.73	59.7529
6	25	625	-7.27	52.8529
7	55	3025	22.73	516.6529
8	35	1225	2.73	7.4529
9	45	2025	12.73	162.0529
10	5	25	-27.27	743.6529
11	20	400	-12.27	150.1819
Sum	355	14025	0.0	2568.1519
	$\sum_{i=1}^n X_i$	$\sum_{i=1}^n X_i^2$	$\sum_{i=1}^n (X_i - \bar{X})$	$\sum_{i=1}^n (X_i - \bar{X})^2$

Probability Distribution

- To find the probabilities associated with a continuous random variable, we use a probability density function (PDF).
- A PDF is an equation used to find probabilities for continuous random variables. The PDF must satisfy the following two rules:
 1. The area under the curve must equal one (over all possible values of the random variable).
 2. The probabilities must be equal to or greater than zero for all possible values of the random variable.

- **The Normal Distribution**

Many continuous random variables have a bell-shaped or somewhat symmetric distribution. The curve is bell-shaped, symmetric about the mean, and defined by μ and σ (the mean and standard deviation).



There are normal curves for every combination of μ and σ . The mean (μ) shifts the curve to the left or right. The standard deviation (σ) alters the spread of the curve.

The first pair of curves have different means but the same standard deviation.

The second pair of curves share the same mean (μ) but have different standard deviations. The pink curve has a smaller standard deviation. It is narrower and taller, and the probability is spread over a smaller range of values. The blue curve has a larger standard deviation. The curve is flatter, and the tails are thicker. The probability is spread over a larger range of values.

Properties of the normal curve:

- The mean is the center of this distribution and the highest point.
- The curve is symmetric about the mean. (The area to the left of the mean equals the area to the right of the mean.)
- The total area under the curve is equal to one.
- As x increases and decreases, the curve goes to zero but never touches.
- The PDF of an normal curve $y = \frac{1}{\sqrt{2\pi} \sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
- A normal curve can be used to estimate probabilities.
- A normal curve can be used to estimate proportions of a population that have certain x -values.

The Standard Normal Distribution

- There are millions of possible combinations of means and standard deviations for continuous random variables. Finding probabilities associated with these variables would require us to integrate the PDF over the range of values we are interested in. To avoid this, we can rely on the standard normal distribution.
- We can use the Z-score to standardize any normal random variable, converting the x-values to Z-scores.
- Mean = 120.
- Std. Dev = 12
- $Z = x_i - \text{mean} / \text{std. dev.}$

Application:

Standardization

Multiple feature : to scale down the feature values z score

Standard scaler in Sklearn.

Compare scores between different distribution.

Avg = 181

Std. dev = 12

Real value = 187

Avg = 182

Std. dev = 5

Real value = 185

Probability

Random Variables : A **random variable** is a variable that takes on different values determined by chance. In other words, it is a numerical quantity that varies at random.

Ex. Suppose we flip a fair coin three times and record if it shows a head or a tail. The outcome or sample space is $S=\{HHH,HHT,HTH,THH,TTT,TTH,THT,HTT\}$.

Discrete Random Variable: When the random variable can assume only a countable, sometimes infinite, number of values.

Continuous Random Variable: When the random variable can assume an uncountable number of values in a line interval.

Probability Functions:

- **Probability Mass Function (PMF)** for discrete random variable
- **Probability Density Function (PDF)** for continuous random variable
- **Cumulative Distribution Function (CDF)** for is a function that gives the probability that the random variable, X , is less than or equal to the value x .

Example : Consider the data set with the values : 0, 1, 2, 3, 4. If X is a random variable of a random draw from these values, what is the probability you select 2?

$$P(x=2) = ?$$

Find the CDF, in tabular form of the random variable, X , as defined above.

x	0	1	2	3	4
$F(x) = P(X \leq x)$	1/5	2/5	3/5	4/5	5/5=1

Hypothesis Testing

- The first step in hypothesis testing is to set up two competing hypotheses. The hypotheses are the most important aspect. If the hypotheses are incorrect, your conclusion will also be incorrect.
- The two hypotheses are named the null hypothesis and the alternative hypothesis.

Null hypothesis

The null hypothesis is typically denoted as H_0 . The null hypothesis states the "status quo". This hypothesis is assumed to be true until there is evidence to suggest otherwise.

Alternative hypothesis

The alternative hypothesis is typically denoted as H_a or H_1 . This is the statement that one wants to conclude. It is also called the research hypothesis.

The goal of hypothesis testing is to see if there is enough evidence against the null hypothesis. In other words, to see if there is enough evidence to reject the null hypothesis. If there is not enough evidence, then we fail to reject the null hypothesis.

- Ex: A man, Mr. XyZ, goes to trial and is tried for the murder of his ex-wife. He is either guilty or innocent. Set up the null and alternative hypotheses for this example.

The hypotheses being tested are:

- 1.The man is guilty
- 2.The man is innocent

Set-Up for One-Sample Hypotheses

One Sample Proportion

Research Question	Is the population proportion different from p_0 ?	Is the population proportion greater than p_0 ?	Is the population proportion less than p_0 ?
Null Hypothesis, H_0	$p = p_0$	$p = p_0$	$p = p_0$
Alternative Hypothesis, H_a	$p \neq p_0$	$p > p_0$	$p < p_0$

Chi-Square Test for Independence

- Chi square test is a hypothesis test that is used when you want to determine if there is a relationship between two categorical variables

Categorical variables

Gender

1 = male
2 = female

Preferred newspaper

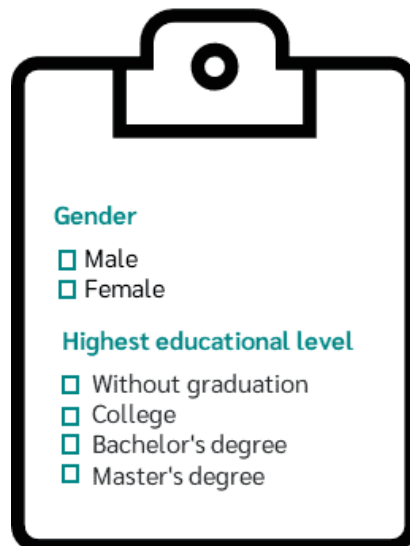
1 = The Washington Post
2 = The New York Times
3 = USA Today
4 = ...

Frequency of television

1 = daily
2 = several times per week
3 = more rarely
4 = never

Highest educational level

1 = Without graduation
2 = College
3 = Bachelor's degree
4 = Master's degree



Gender



☐ Male
☐ Female

Highest educational level

☐ Without graduation
☐ College
☐ Bachelor's degree
☐ Master's degree



Fall	Gender	Highest educational level
1	Male	College
2	Female	Without graduation
3	Male	Without graduation
4	Male	Bachelor's degree
5	Female	Master's degree
6	Male	Bachelor's degree
7	Female	Master's degree
...



	Female	Male
Without graduation	6	7
College	13	16
Bachelor's degree	16	15
Master's degree	8	11
Total	43	49

Is there a correlation
between gender and the
highest level of education?



Chi²- Test

- **Null hypothesis:** there is no relationship between gender and highest educational attainment.
- **Alternative hypothesis:** There is a correlation between gender and the highest educational attainment.

The chi-squared value is calculated via:

$$\chi^2 = \sum_{k=1}^n \frac{\overset{\text{Observed frequency}}{O_k} - \overset{\text{Expected frequency}}{E_k}}{E_k}^2$$

Observed frequency:

		Variable 2	
		Category A	Category B
Variable 1	Category A	10	13
	Category B	13	14

Expected frequency:

		Variable 2	
		Category A	Category B
Variable 1	Category A	9	11
	Category B	12	13

$$\chi^2 = \frac{(10 - 9)^2}{9} + \frac{(13 - 11)^2}{11} + \frac{(13 - 12)^2}{12} + \frac{(14 - 13)^2}{13} = 0.635$$

For a significance level of 5 %

$$df = (p - 1)(q - 1) = 1$$

For a significance level of 5 %, this results in 3.841

Since the calculated chi-squared value is smaller, there is no significant difference. As a **prerequisite** for this test, please note that all expected frequencies must be greater than 5.

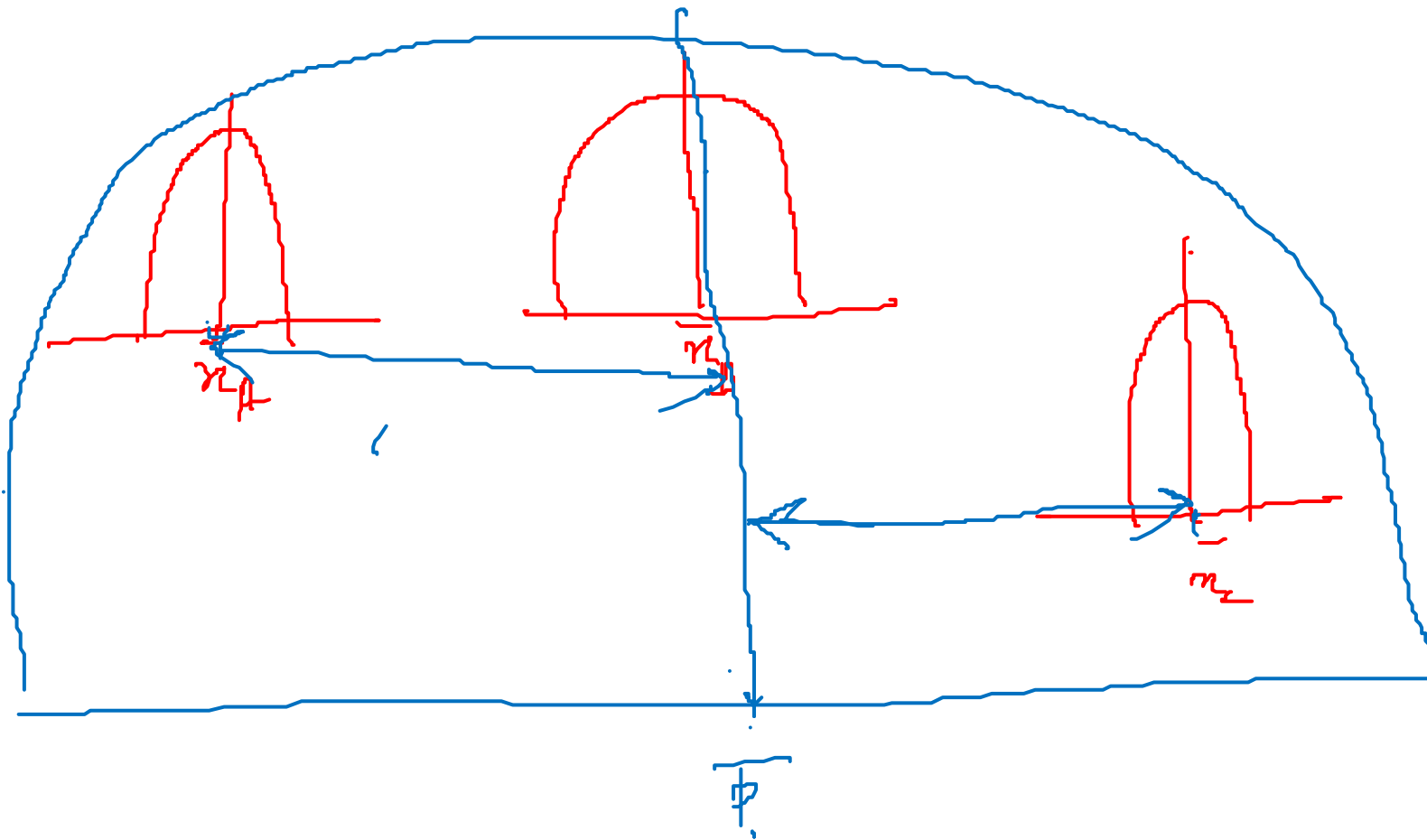
- A school principal would like to know which days of the week student are most likely to be absent . The principal expects that student will absent equally during the five-day school week the principal select a random sample of 100 teacher asking them which day of the week they had the highest number of student absence. The observed and expected result are shown in the table below based on this result to the days for the highest number of absences occur with equal frequency? (use a 5% significance level)

	Mon	Tue	Wed	Thu	Fri
Observed Absences	23	16	14	19	28
Expected Absences	20	20	20	20	20

- In an antimalarial campaign in India, quinine was administered to 500 person out of a total population of 2000. The number of fever cases is shown below :

Treatment	Fever	No fever	Total
Quinine	20	480	500
No Quinine	100	1400	1500
Total	120	1880	2000

Introduction to Analysis of Variance



3 samples

$$\mu_A = \mu_B = \mu_C$$

Null Hypothesis

Assumption

Make an analysis of the variance on given data:

- To assess the significance of possible variation in performance in a certain test between the convent school of a city a common test was given to several students taken at random from the 5th class of three schools concerned the result given below :

	A	B	C
9	13	14	
11	12	13	
13	10	17	
9	15	7	
8	5	9	

$H_0 = \mu_1 = \mu_2 = \mu_3$
Null hypothesis

F test

