**Sampling Distribution**

In real life, we often **don’t have access to the entire population**, so we take a sample and use it to **estimate population characteristics**. The results from samples follow different probability distributions, known as **sampling distributions**.

In statistics, a sampling distribution is the probability distribution of a statistic (such as the mean, variance, or proportion) that is obtained by drawing all possible samples of a fixed size (n) from a population. It describes how that statistic varies across different samples.

**Key Concepts**

* **Population:** The entire group of individuals, objects, or measurements of interest.
* **Sample:** A subset of the population selected for analysis.
* **Statistic:** A numerical value calculated from a sample (e.g., sample mean, sample variance).
* **Sampling Distribution:** The distribution of a statistic across all possible samples of the same size from a given population.

**T-Distribution**

* The t-distribution is a probability distribution that arises when estimating the mean of a normally distributed population when the sample size is small or the population standard deviation is unknown.
* It is similar to the standard normal distribution but has heavier tails, which accounts for the increased uncertainty when dealing with small samples or unknown population variance.
* The shape of the t-distribution depends on its degrees of freedom (df), which are related to the sample size (n). As the sample size increases, the t-distribution approaches the standard normal distribution.
* Formula:
* t = (sample\_mean - population\_mean) / (sample\_standard\_deviation / sqrt(sample\_size))

### **Real-World Applications**

* Used in **A/B testing** (e.g., testing whether a new marketing strategy increases sales).
* Used in **medical studies** to compare the **effectiveness of two treatments**.
* Helps in **small sample hypothesis testing** (e.g., determining if a new drug improves patient recovery time).

### **Key Properties**

* Symmetric and bell-shaped.
* As sample size **increases**, it **approaches the Normal distribution**.
* Defined by **degrees of freedom (df = sample size - 1)**.

**F-Distribution**

* The F-distribution is a probability distribution that arises in tests comparing variances.
* It is the distribution of the ratio of two independent chi-square variables, each divided by its degrees of freedom.
* The F-distribution is used in ANOVA to compare the variances between groups and within groups to determine if there are significant differences between group means.
* It is also used in regression analysis to test the overall significance of the model.
* The F distribution is defined by two types of degrees of freedom: numerator degrees of freedom (df1) and denominator degrees of freedom (df2).
* Formula:
* F = (variance\_between\_groups / df1) / (variance\_within\_groups / df2)

### **Real-World Applications**

* Used in **quality control** to check if two machines produce parts with the **same variance**.
* In **finance**, it helps test whether **different investment strategies have different risks**.
* In **marketing**, it checks if multiple advertising strategies lead to the same level of customer engagement.

**Chi-Square Distribution**

* The chi-square distribution is a probability distribution that arises in various statistical tests, particularly those involving categorical data or variances.
* It is the distribution of the sum of the squares of independent standard normal random variables.
* The shape of the chi-square distribution depends on its degrees of freedom (df), which are related to the number of independent pieces of information used to calculate the chi-square statistic.
* Formula:
* chi2 = sum((observed\_value - expected\_value)^2 / expected\_value)

### **Real-World Applications**

* Used in **healthcare** to check if smoking and lung disease are related.
* In **sports**, it tests if **player performance is independent of the team they play for**.
* In **survey analysis**, it helps check whether customer preferences are **random or follow a pattern**.

**Similarities**

* All three distributions are related to the normal distribution.
* They are all used in hypothesis testing.
* Their shapes are determined by degrees of freedom.
* The F and chi-square distributions are skewed to the right.

**Differences**

* The t-distribution is used for inference about a single mean, especially when the population variance is unknown.
* The F-distribution compares variances of two or more groups.
* The chi-square distribution is used for tests on categorical data and inferences about variance.
* The t-distribution can be symmetrical (like the normal distribution), while the F and chi-square distributions are skewed.
* The t-distribution has one parameter (degrees of freedom), while the F-distribution has two (numerator and denominator degrees of freedom), and the Chi-square distribution has one (degrees of freedom).

**When Each Distribution Is Applied (with Examples)**

* **T-Distribution**
  + **Application:**
    - Used to test hypotheses about a population mean when the population standard deviation is unknown and the sample size is small (typically n < 30).
    - Used to construct confidence intervals for a population mean.
  + **Example:**
    - A researcher wants to determine if the average IQ score of students at a particular university is significantly different from the national average of 100. They collect a sample of 25 students and find the sample mean and sample standard deviation. A t-test is used to compare the sample mean to the national average.
* **F-Distribution**
  + **Application:**
    - Used in Analysis of Variance (ANOVA) to compare the means of two or more groups.
    - Used to test the equality of variances of two populations.
    - Used in regression analysis to test the overall significance of the regression model.
  + **Example:**
    - A company wants to compare the effectiveness of three different teaching methods on student performance. Students are randomly assigned to one of the three methods, and their test scores are recorded. ANOVA, using the F-distribution, is used to determine if there are any significant differences in average test scores between the three groups.
* **Chi-Square Distribution**
  + **Application:**
    - Used in goodness-of-fit tests to determine if observed data fits a theoretical distribution.
    - Used in tests of independence to determine if there is a relationship between two categorical variables.
    - Used to test hypotheses about a population variance.
  + **Example:**
    - A market researcher wants to investigate if there is an association between gender (male/female) and preference for a particular brand of coffee (Brand A, B, or C). A chi-square test of independence is used to analyze the data from a sample of consumers to see if there is a statistically significant relationship between these two categorical variables.

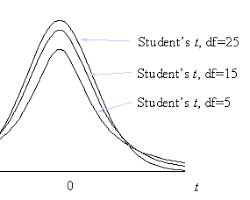
**Degrees of Freedom, and relation to the Distributions?**

Degrees of freedom (df) represent the number of independent pieces of information available in a sample to estimate a population parameter. In simpler terms, it's the number of values in the final calculation of a statistic that are free to vary.

**Relationship to the Distributions**

* **t-Distribution:** The t-distribution's shape is determined by its degrees of freedom. For a one-sample t-test, df = n - 1, where n is the sample size. As df increases, the t-distribution approaches the standard normal distribution.

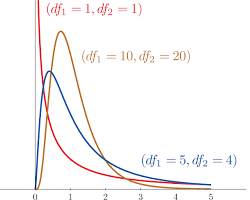
Graph:

[[https://encrypted-tbn2.gstatic.com/favicon-tbn?q=tbn:ANd9GcRyq6qcqiVx940Y-_ZLzCI-K36jMggvh9gWWQj3DtrXENdDc_wy3bkM15vx3GrKKFwuGY7wXAw9xd76sVHszvaVWgQt6MXjMSNaoRyxIZaD](https://www.statsdirect.co.uk/help/distributions/t.htm)www.statsdirect.co.uk](https://www.statsdirect.co.uk/help/distributions/t.htm)

Tdistribution with varying degrees of freedom

* **F-Distribution:** The F-distribution has two types of degrees of freedom: numerator degrees of freedom (df1) and denominator degrees of freedom (df2). These correspond to the degrees of freedom associated with the variance between groups and the variance within groups, respectively, in ANOVA.

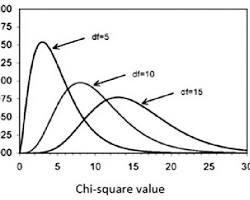
Graph:

[saylordotorg.github.io](https://saylordotorg.github.io/text_introductory-statistics/s15-03-f-tests-for-equality-of-two-va.html)

F-distribution with different pairs of degrees of freedom

* **Chi-Square Distribution:** The chi-square distribution's shape is also determined by its degrees of freedom. In a goodness-of-fit test, df is typically the number of categories minus the number of estimated parameters. In a test of independence, df = (number of rows - 1) \* (number of columns - 1) in the contingency table.

Graph:

[[https://encrypted-tbn0.gstatic.com/favicon-tbn?q=tbn:ANd9GcQHnHFqVlgn9GP_QNJZJqp5p4tCYLbsP3aagcGHhOG1R5B49YpECY11Ifpa36Xt9B7B7i5sptMSQYILbaYUaLXDbBCp19SJHEuOmPv3eRw](https://www.researchgate.net/figure/Distribution-of-chi-square-values-with-different-degrees-of-freedom-Modified-from_fig3_336060118)www.researchgate.net](https://www.researchgate.net/figure/Distribution-of-chi-square-values-with-different-degrees-of-freedom-Modified-from_fig3_336060118)

Chisquare distribution with varying degrees of freedom

In essence, degrees of freedom dictate the shape and spread of these distributions, which in turn affects how we make statistical inferences.

**Scenarios**

**Problem:** A botanist wants to determine if a new fertilizer increases the average height of sunflower plants. They take a random sample of 20 sunflower seedlings, apply the new fertilizer, and measure their heights after a month. The botanist assumes that the heights of sunflower plants are normally distributed, but the population variance is unknown.

* **Why t-distribution?**
  + The t-distribution is used to estimate the population mean when the population variance is unknown and the sample size is small (or moderate).
  + Here, the botanist has a sample of 20 plants (a relatively small sample size), the population of sunflower heights is assumed to be normally distributed, and the population variance of the heights is unknown.
  + The t-distribution allows us to make inferences about the population mean height, even though we don't know the population variance. We estimate it using the sample variance.
  + The t-test will help the botanist determine if the average height of the fertilized plants is significantly different from the average height of plants that do not use the fertilizer.

**Problem:** A company wants to assess the effectiveness of three different advertising campaigns on sales. They divide a sample of similar markets into three groups. Each group is exposed to one of the three advertising campaigns. After a certain period, the sales revenue for each group is recorded. The company assumes that sales revenue in each market follows a normal distribution.

* **Why F-distribution?**
  + The F-distribution is used in Analysis of Variance (ANOVA) to compare the means of two or more groups.
  + In this case, the company wants to determine if there's a significant difference in the average sales revenue generated by the three different advertising campaigns.
  + ANOVA calculates the F-statistic by comparing the variance between the groups (i.e., the variance in sales revenue due to the different campaigns) to the variance within the groups (i.e., the random variation in sales revenue within each group).
  + If the F-statistic is large, it suggests that the differences between the group means are significant.

**Problem:** A hospital administrator wants to investigate if the number of patient admissions is independent of the day of the week. They collect data on the number of patients admitted each day for a month.

* **Why Chi-square distribution?**
  + The Chi-square distribution is used to test the independence of categorical variables.
  + Here, the hospital administrator wants to determine if there's a relationship between the categorical variable "day of the week" and the categorical variable "number of patient admissions" (which can be categorized into ranges).
  + The Chi-square test compares the observed frequencies of admissions for each day of the week with the expected frequencies, assuming that admissions are independent of the day of the week.
  + A significant Chi-square statistic would suggest that patient admissions are not independent of the day of the week.

**Relationship Between the Three Distributions**

* The **t-distribution** is derived from the **Normal distribution** when estimating a population mean with a small sample.
* The **F-distribution** is derived from the ratio of two **Chi-Square distributed variables** divided by their respective degrees of freedom.
* The **Chi-Square distribution** itself is a special case of the **Gamma distribution** and is the sum of squared standard normal variables.
* The **t-distribution** and **Chi-Square distribution** both influence the F-distribution, as t-values are derived using sample variance estimates, which follow a Chi-Square distribution.
* As **sample size increases**, the **t-distribution approaches the Normal distribution**, and the **Chi-Square and F-distributions** become more symmetrical.

## ****6. Comparison of t-Test, F-Test, and Chi-Square Test****

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **t-Test** | **F-Test** | **Chi-Square Test** |
| **Purpose** | Compare means | Compare variances | Test independence/goodness of fit |
| **Key Assumptions** | Normality, equal variance (for independent samples) | Normality, independent groups | Observations are independent, expected frequency > 5 in each category |
| **Typical Application** | Comparing two sample means | Comparing multiple variances (ANOVA) | Testing categorical data relationships |
| **Example** | Testing if a new drug improves patient recovery time | Checking if different marketing campaigns have different engagement levels | Determining if gender affects brand preference |
| **Degrees of Freedom** | (n₁ + n₂ - 2) for two samples | (df1, df2) from variances | (Rows - 1) \* (Columns - 1) |
| **Shape of Distribution** | Symmetric, bell-shaped | Right-skewed | Right-skewed but normal-like for large df |

**Summary Table: When to Use t, F, and Chi-Square Distributions**

| **Distribution** | **When to Use** | **Example Scenario** |
| --- | --- | --- |
| **t-Distribution** | **Comparing means** in small samples (n < 30) with unknown variance | Testing if a new drug improves patient recovery time |
| **F-Distribution** | **Comparing variances** or multiple means (ANOVA, regression) | Checking if three marketing strategies have the same engagement rate |
| **Chi-Square Distribution** | **Testing independence, goodness-of-fit, or variance estimation** | Checking if smoking is linked to lung disease in a survey |

**Q1: A sample of size 20 is drawn from a normal population with an unknown variance. Which sampling distribution would be appropriate for estimating the population mean and why?**

Since the **sample size is 20** and the **population variance is unknown**, the appropriate sampling distribution for estimating the **population mean** is the **t-distribution (Student’s t-distribution)**.

**Reasoning:**

1. **Sample Size Consideration:**
   * When the sample size is **small (n < 30)**, the **t-distribution** is used instead of the **Normal distribution**.
2. **Unknown Population Variance:**
   * The **t-distribution** is used when the **population standard deviation (σ) is unknown**, and we estimate it using the **sample standard deviation (s)**.
3. **Assumption of Normality:**
   * The sample is drawn from a **Normal population**, so the **t-distribution applies** even for small sample sizes.
4. **Degrees of Freedom:**
   * The **t-distribution** is defined by **degrees of freedom (df = n - 1)**.
   * Here, **df = 20 - 1 = 19**.

**Conclusion:**

Since the **sample size is small (n = 20)** and **population variance is unknown**, the **t-distribution** is the appropriate choice for estimating the **population mean**.