Chapter 1

Comparing data sets

Statistics is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics. Now I agree that the physicist, the chemist, the engineer, and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics, better scientific investigation. Whether in any given study this implies more or less mathematics is incidental.

GEORGE E. P. BOX

Important:

Typical task in statistics is answering one or more of the following questions:

1. Are the measured/provided datasets (often two or more) measuring/representing the same phenomenon/phenomena?
2. What is the distribution?
3. In the case of a single dataset, does it come from some known distribution, that is, does it sample a certain population (data)?
4. How confident can we be about the significance of the data and findings?

Comparing data sets is crucial for identifying and understanding *relationships and patterns* and identifying *anomalies and outliers* that might not be evident within a single data set. By juxtaposing data from different sources or experiments, one can discern correlations or causal relationships. For instance, in epidemiological studies, comparing data sets from different geographical regions or time periods can reveal trends and factors influencing the spread of diseases, guiding public health interventions. As a practical example, the unexpected observation of gravitational waves was confirmed by comparing data from multiple detectors, leading to a breakthrough in astrophysics[1]. [7]

In *any* scientific research, replication of results is fundamental. Comparing independent data sets allows researchers to confirm the reliability and generalizability of their findings. This practice is especially important in fields like psychology and medicine, where reproducibility concerns have

[1] The Nobel Prize in Physics 2017 was awarded to Rainer Weiss, Barry C. Barish and Kip S. Thorne "for decisive contributions to the LIGO detector and the observation of gravitational waves." LIGO is an abbreviation for Laser Interferometer Gravitational-Wave Observatory.

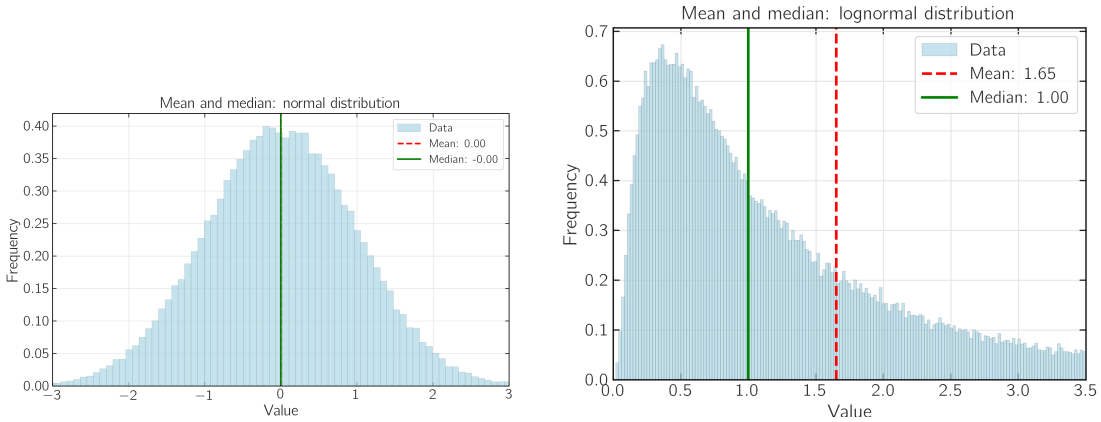

Figure 1.1: Illustration of the mean and median for normally distributed and lognormally distributed data. The mean and median agree with the theoretical predictions.

led to the so-called "replication crisis". [8] Independent of the field, reproducibility is absolutely fundamental as discussed in the Editorial "Trust but verify" in Nature Materials. [12]

Comparing data sets presents challenges. Issues such as data compatibility, standardization, and privacy concerns must be addressed. Ensuring data integrity and ethical data usage is critical to maintaining public trust and the validity of research findings.

There are many methods for comparing datasets. The most obvious one is *descriptive statistics*, that is, quantities such as mean, median, mode, variance and so on. In virtually all cases, data is *visualized* using, e.g., histograms, scatter plots or bar charts. The third approach, *inferential statistics*, aims to look deeper into the data. This involves comparisons of data sets using methods such as the t-test, χ^2-test, or ANOVA (Analysis of Variance). *Correlation analysis* produces measures for relationships between variables. Methods include the Pearson correlation coefficient and Spearman's rank correlation. Finally, *machine learning* methods can uncover hidden patterns that are not necessarily revealed by any of the above methods. The focus in this chapter is on inferential statistics and correlation analyses. The null hypothesis provides a specific claim that can be tested with statistical evidence.

1.1 Statistical hypothesis, significance and hypothesis testing

Before discussing some of the common methods and metrics, let's start with a brief discussion regarding *statistical hypothesis testing*. As the term implies, task is to determine of the data sufficiently support a particular hypothesis. Testing is done using one or several *test statistics*, such as the ones discussed in the sections below.

Statistical hypothesis testing consists of formulating a statement, i.e., a hypothesis, about a population parameter and then using statistical evidence from the sample to determine whether to reject or not that statement. The assumption is called *null hypothesis* and typically denoted by H_0. In the case H_0 holds, the conclusions is that any observed differences in the data are due torandomness. If the null hypothesis is rejected, then the *alternative hypothesis* (H_1) hold suggesting that there is a significant difference.

1. **One-sample tests.** Here a characteristic (chosen test statistic(s)) is compared to a known or hypothesized population characteristic. The purpose is to determine whether there is sufficient evidence to conclude that there is significant different from the known/hypothesized

characteristic(s). Example: Assessing the effectiveness of a new treatment by comparing to data from the known standard.

2. **Two-sample tests**. Here, the chosen statistic(s) of two independent or related samples are compared to determine if there is a significant difference between them. The canonical example is determining the effectiveness of drugs by comparing the treatment groups with control groups.
3. **Paired tests**. This type of testing is often done when some of the important variables cannot be controlled. In a paired test, members (the test subjects) are paired between the samples, and the difference between the members becomes the sample. The canonical example is testing the same subjects twice under different conditions, for example, before and after a treatment.

Statistical hypothesis testing one typically distinguishes between *Type I* and *Type II errors*.

1.1.1 Type I error and significance level α

A Type I error occurs when H_0 is wrongly rejected when it is actually true. This is commonly called *false positive*: it's the mistake of assuming an effect or difference when there isn't one. The probability of making a Type I error is denoted by α; it is also known as the significance level of the test. The significance level is chosen by the researcher before conducting the test and typically set at 0.05 (5%), indicating a 5% risk of committing a Type I error.

■ **Example 1.1** Type I error.

Consider testing a new drug for its effectiveness. A Type I error would occur if the test concludes that the drug is effective when, in fact, it is not any more effective than a placebo.

1.1.2 Type II Error and the probability of making it (β)

This is the case when H_0 is wrongly accepted when it is false, that is, he test fails to detect an actual difference. This is called a "false negative." The probability of making a Type II error is denoted by β. The *power of the test* $(1-\beta)$ is the probability of correctly rejecting a false H_o.

■ **Example 1.2** Type II error.

If the test on the effectiveness of a new drug falsely concludes that the drug is not effective when, in fact, it is effective. As a result, the company stop the development. This scenario highlights the importance of designing effective studies.

1.2 Quick recap: Basic descriptive statistics

As discussed above, descriptive statistics involve the usual characteristics, such as median, mean and variance. Below, we discuss this from the point of view of observations, or datasets.

1.2.1 Mode

Mode is the value that appears most frequently in a data set. Note that mode depends on binning of the data. For example, when analyzing market data, the mode (when binned properly) can indicate the most popular single product among consumers. Importantly, unlike the other measures that will be discussed below, the mode can be used as a qualitative measure with non-numerical (categorical)

data. In addition, a data set can have more than one mode, that is, it can be bimodal, or multimodal. When there is a significant difference between the mode and other measures of central tendency (mean and median, see below), it may indicate the presence of outliers or a skewed distribution.

1.2.2 Median

The **median** is the value separating the higher half from the lower half of a data sample, see Figure 1.1. Here, ones needs to pay attention if the dataset has an even or an odd number of observations. For notational simplicity, let's assume that we have dataset $\{x_1, x_2 \ldots, x_N\}$. For a dataset sorted in ascending order, the median is determined as follows:

1. **Odd number of observations:** the median is the value exactly in the middle of the dataset. If the dataset is $\{x_1, x_2, ..., x_N\}$ with observations sorted such that $x_1 \leq x_2 \leq ... \leq x_N$, and N is odd, then the median M is:
$$M = x_{\frac{N+1}{2}} \tag{1.1}$$
2. **Even number of observations:** the median is the average of the two middle numbers. For the same dataset sorted in ascending order, if N is even, then the median M is:
$$M = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \tag{1.2}$$

■ **Example 1.3** Python code for computing median, median standard deviation and variance.

```
import numpy as np  # let's us numpy There are also other options

# Example unsorted dataset
data = np.array([7, 2, 10, 4, 3, 1, 8, 5, 9, 6])

# Compute mean
mean = np.mean(data)
print(f"Mean: {mean}")

# Compute standard deviation
std_dev = np.std(data, ddof=1)  # Set ddof=1 for sample standard deviation
print(f"Standard Deviation: {std_dev}")

# Compute variance
variance = np.var(data, ddof=1)  # Set ddof=1 for sample variance
print(f"Variance: {variance}")

# Compute median
median = np.median(data)
print(f"Median: {median}")
```

1.2.3 Mean (average)

The **average (mean)** for dataset $\{x_1, x_2, ..., x_n\}$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i. \tag{1.3}$$

This is, of course, very familiar, but when we deal with data. See Figure 1.1 illustrating the mean and the median for two different distributions.

There is, however, an important subtlety that needs to be considered, namely, the difference between *population* and *sample means*. This will also have some implications regarding notation as will be discussed.

Population mean vs sample mean

First, note that the term 'population' refers to the data, not to a group of people. The **mean, or the population mean**, is the average over the whole data (population). It is typically denoted by the symbol μ.

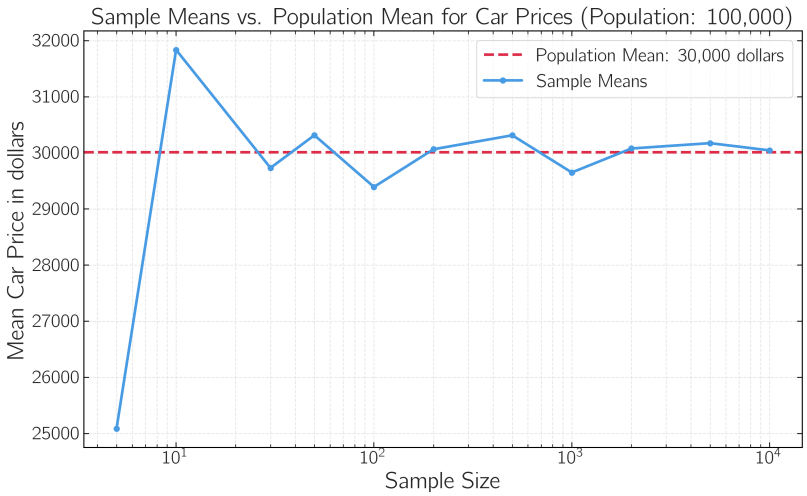

Figure 1.2: Illustration of population and sample means. Assume that the full dataset, that is, the population, consists of 100,000 data points. These data points represent car prices for car sold in one year in some given region. The car prices are assumed to follow a normal distribution with an average price of $30,000 and a standard deviation of $7,000 The dashed red line represents the *population mean*, calculated from the entire population. The blue line and markers illustrate how the *sample means* of various sizes (ranging from 5 to 10,000) drawn from this population *approach* the population mean. The x-axis is on a logarithmic scale to effectively display the progression across a wide range of sample sizes

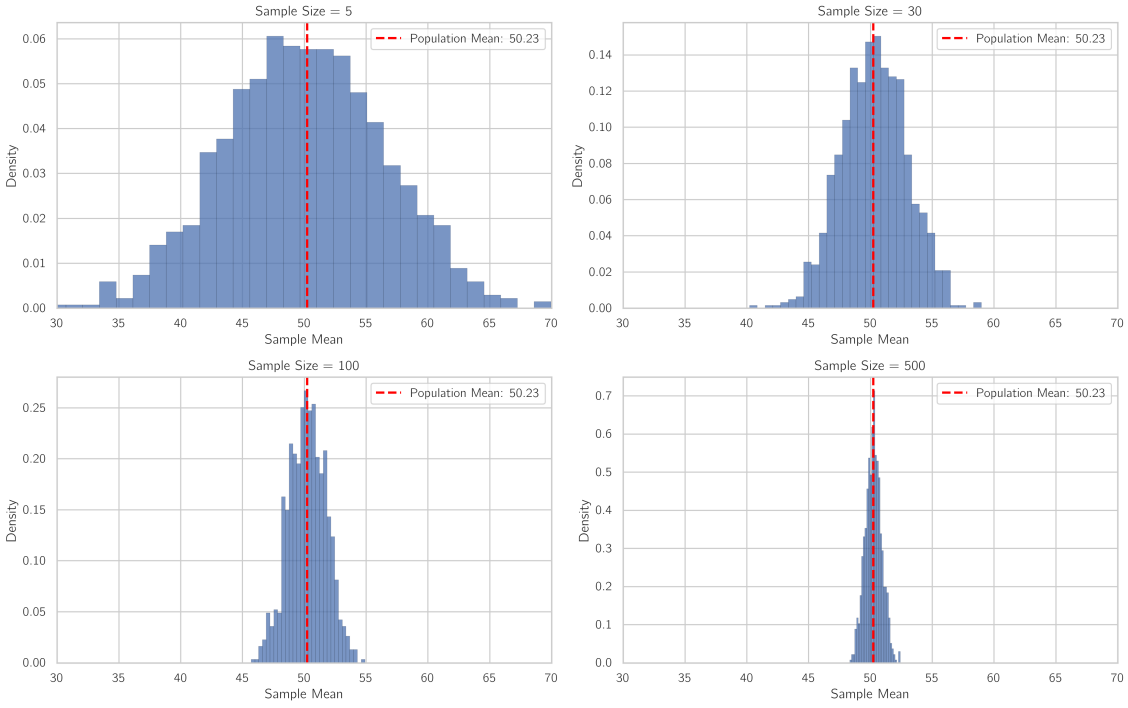

Figure 1.3: Illustration demonstrating the distribution of sample means for different sample sizes (computed using Equation 1.5. The population has 10,000 normally distributed data points with the population mean (Equation 1.4) of 50.23. Sample sizes of 5, 30, 100 and 100 were used, the number of samples in each case was 1,000. The x-axis is kept the same to show the how the distribution of $\bar{x}$ depends on the sample size.

Sample mean (sometimes also called *empirical mean*), on the other hand, is the average of all values in a *sample*. Sample is some subset of the full data (population). Sample mean is typically denoted as $\bar{x}$. The sample mean is used as an *estimate* of the population mean. Sample mean may be needed, for example, because direct calculation of the population mean may not be feasible due to the impracticality of measuring every individual in a large population. The concept is illustrated in Figure 1.2.

Assume that we have a population A of size N data points such that $A = \{x_1, x_2, ..., x_N\}$. Sample (subset) B is a subset of A (denoted as $B \subsetneq A$) and has less points (we denote the number of data points y_i in B as n) than the full population A, that is, $N > n$ and $\forall y_i \in B \Rightarrow y_i \in A$. With these, we define the population (μ) and sample ($\bar{x}$) averages are defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.4}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.5}$$

It is important to note that the sample mean is a random variable. Its calculated values (using Equation 1.5) differ depending on which members of the population have been included in the sample. This also means that the sample mean itself has its own distribution, Figure 1.3.

Since the sample mean is an estimator of the population mean, we need to be able to evaluate its reliability. The most common method is using the standard error of the mean. We will return to that after introducing a few more quantities for descriptive statistics.

1.2.4 Degrees of freedom

In statistics, the term *degrees of freedom* refers to the number of independent quantities that can vary in the analysis of data sets. The number of degrees of freedom may be limited by constraints. For example, this is the reason for the factor of $n-1$ (also known as Bessel's correction) when renormalizing *sample variance*. Let's illustrate this by an example.

Assume that we have a sample of independent random (normally) distributed datapoints $\{x_1, x_2, ..., x_n\}$. The sample average is given by Equation 1.5, that is,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

We can write the datapoints as a vector

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \bar{x}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}. \tag{1.6}$$

Since the sample average $\bar{x}$ is known, the vector of residuals has the *constraint*

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

This means that only $n-1$ of the datapoints are independent, that is, given $n-1$ values, the value of the n^{th} component is determined by the constraint: there are $n-1$ degrees of freedom.

1.2.5 Variance

Variance measures how far each number in the set is from the mean and thus from every other number in the set. As in the case of the mean, we need to separate between **population variance** and **sample variance**. As will be shown below, the two are described by slightly different formulas, the difference arising from different number of **degrees of freedom** in the two.

The formula for the **population variance**, denoted as σ^2, is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{1.7}$$

Note that σ^2 can also be calculated using

$$\sigma^2 = \frac{1}{N^2}\sum_{i<j}(x_i - x_j)^2. \tag{1.8}$$

Sample variance (s^2) is computed using the formula

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.9}$$

Here, the division is by a factor of $n-1$. This is called *Bessel's correction* and it arises from the reduced *degrees of freedom* (see the previous section) due to the presence of $\bar{x}$

■ **Example 1.4** Population and sample variance.

Consider a poll that rates pineapple as a pizza topping. The population (full dataset) is 10 people. The ratings are from 1 to 10. From this population, we randomly select a sample of 5 ratings the sample variances both with (Equation 1.9) and without (Equation 1.7) the Bessel correction. Here is the data:

- Population ratings (full data): [7, 4, 8, 5, 7, 10, 3, 7, 8, 5]
- Sample ratings (randomly selected from the full data): [7, 3, 5, 4, 7]
- Population Variance: 4.04
- Sample variance (n): 2.56
- Sample variance ($n-1$; Bessel correction): 3.20

The population variance is the true measure of variability for the full dataset is 4.04. Computing the sample variance without Bessel's correction, we get 2.56, which underestimates the true variability of the population. By applying Bessel's correction, the sample variance increases to 3.20. This is a closer estimate to the true population variance. The correction accounts for the loss of *degrees of freedom* due to estimating the mean from the sample data.

1.2.6 Standard deviation

The standard deviation is more commonly used than variance in practical settings because it is in the same units as the data, which makes it easier to interpret.

The **standard deviation** is the square root of the variance, providing a measure of the *dispersion* or spread of the data points in a dataset. The **population** and **sample** standard deviations are defined, respectively, as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}. \tag{1.10}$$

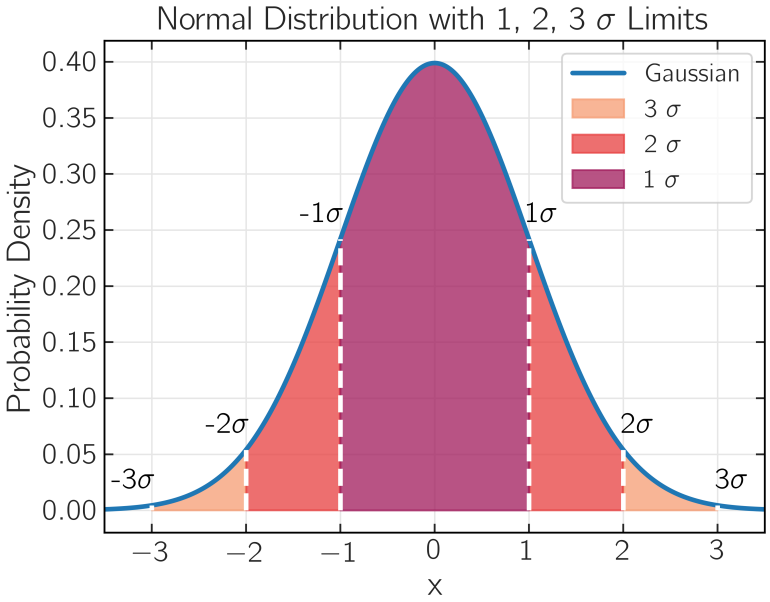

Figure 1.4: Standard normal distribution with the 1σ ($\approx 68\%$ of the data), 2σ ($\approx 95\%$), and 3σ ($\approx 99.7\%$) limits. Note that the height of the peak is given by the prefactor $1/(\sigma\sqrt{2\pi})$ and the location of the peak by μ in Equation 1.12. The $y-$axis gives the relative probability. The 1-2-3 σ limits are sometimes called as the 68–95–99.7 rule. In particle physics, however, 5σ, or 99.99994% confidence is required. See Why do physicists mention "five sigma" in their results? for a discussion.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{1.11}$$

1.3 Gaussian Distribution or the normal distribution

The Gaussian[2] distribution, also known as the normal distribution, is characterized by its bell-shaped curve, which is symmetric about its mean. It is defined by two parameters: the mean (μ) and the standard deviation (σ). The Gaussian distribution describes the distribution of a continuous variable and it probability distribution function (PDF) is given by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \tag{1.12}$$

where μ is the mean and σ^2 is the variance. Since the above equation is a PDF, it has to integrate to 1, that is,

$$\int_{\infty}^{\infty} dx\, f(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, e^{-(x-\mu)^2/(2\sigma^2)} = 1. \tag{1.13}$$

The Gaussian distribution has the property that about 68% of the data falls within one standard deviation of the mean, about 95% within two standard deviations, and about 99.7% within three standard deviations, Figure 1.4. As already mentioned above, the mean, mode and median are equal for a Gaussian distribution.

[2] Johann Carl Friedrich Gauss (1777-1823), often called the Prince of Mathematicians". Motto: *Pauca sed Matura* ("Few, but Ripe").

1.3.1 Standard normal distribution and the z-score

The special case called *standard normal distribution* or *unit normal distribution* has the properties

$$\begin{cases} \mu & = 0 \\ \sigma^2 & = 1. \end{cases} \tag{1.14}$$

The z-score, defined as

$$z = \frac{x - \mu}{\sigma}, \tag{1.15}$$

non-standardized raw scores can be compared to each other. Note that this involves the *population* mean and standard deviation that are often (typically) unknown. In such a case, the t-statistic is used as will be discussed in the following sections.

1.3.2 Gaussian distribution in machine learning

We will return this later, but here are a few quick points: Some machine learning methods, in particular linear and logistic regression (the residuals in them), and Gaussian Naive Bayes assume Gaussian distribution. Several statistical tests, such as the t-test and ANOVA assume that the data follows Gaussian distribution. It is also very common to transform data, when possible, to follow a Gaussian distribution. This is the case in feature engineering in machine learning. In addition, the commonly used technique of Principal Component Analysis (PCA) in machine learning assumes Gaussian distribution

1.4 The Central Limit Theorem (CLT)

Consider a random variable x. For example, rolling a die has 6 possible outcomes and for an unloaded die, each outcome has the probability of 1/6. Now, consider taking *multiple samples* of the random variable and adding them together, that is, for n samples one has

$$x_1 + x_2 + ... + x_n. \tag{1.16}$$

What the CLT states is that as n grows ($n \to \infty$ to be precise), the distribution of Equation 1.16 approaches the Gaussian distribution, Figure 1.5.

Requirements:

1. The random variables x_i must be independent of each other.
2. The x_i must come from the same distribution.
3. Finite variance (finite σ^2).

While almost all distributions satisfy the above, the Lorentz (or Cauchy) distribution does not obey rule 3). It has neither a well-defined mean nor a well-defined variance. Note that properties 1) and 2) are often combined and called using the term *independent and identically distributed*.

1.5 The t-distribution

The t-distribution, also known as the *Student's t-distribution*, is a continuous probability distribution similar in shape to the Gaussian distribution, that is, it is symmetric and has a bell shape, but has *heavier tails*, Figure 1.6. Importantly, and as the figure shows, the exact shape of the t-distribution depends on its *degrees of freedom* (ν). As Figure 1.6 shows, the distribution approaches the

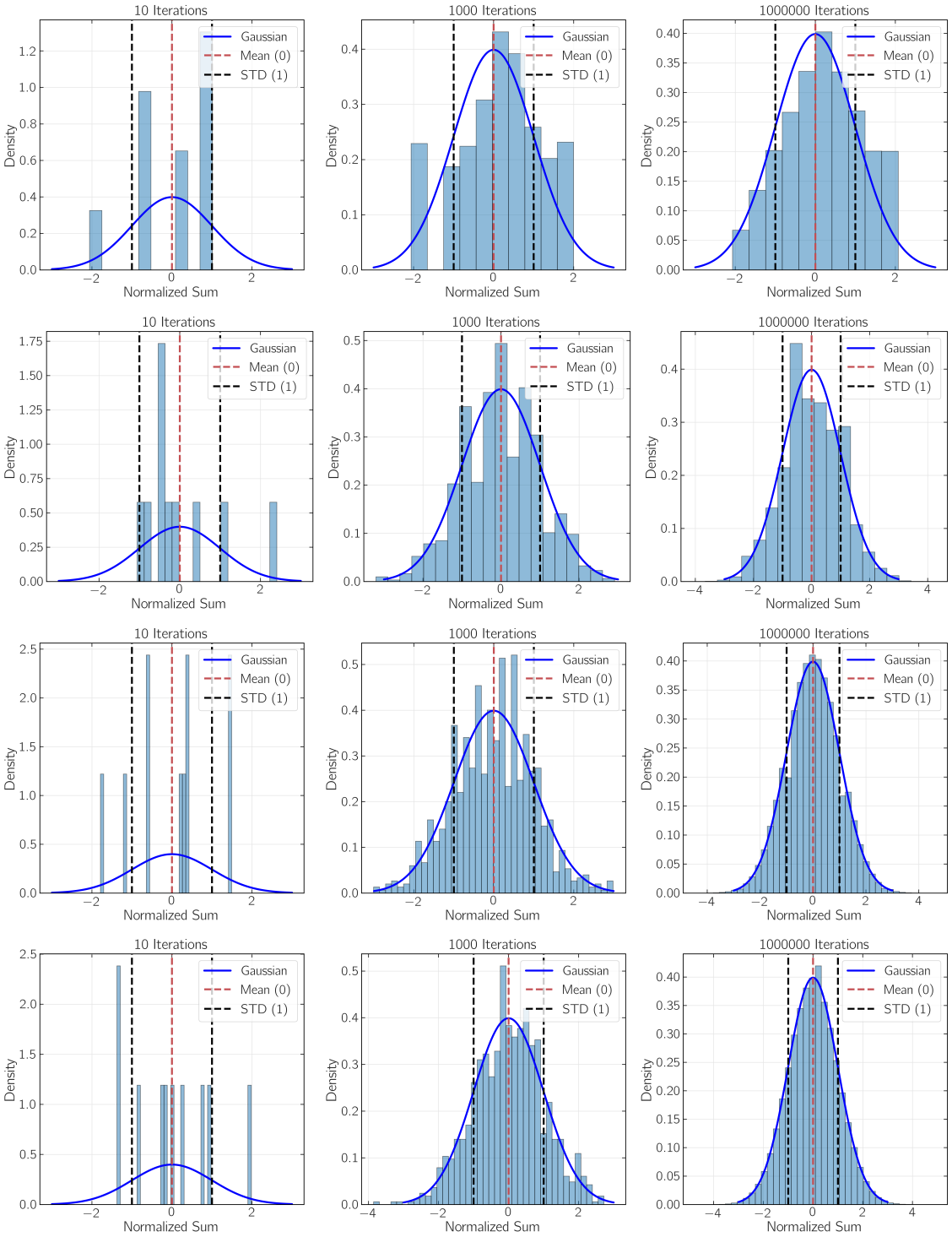

Figure 1.5: Approach toward a Gaussian distribution, experiments with a fair dice. Top row: Sum over 2 throws with 10, 1,000 and 1,000,000 samples. Second row: Sum over 10 throws with 10, 1,000 and 1,000,000 samples. Third row: Sum over 100 throws with 10, 1,000 and 1,000,000 samples. Last row: Sum over 1,000 throws with 10, 1,000 and 1,000,000 samples. The distributions have been renormalized to such that the mean and the variance are equal to one. Note that binning has also been adjusted as necessary.

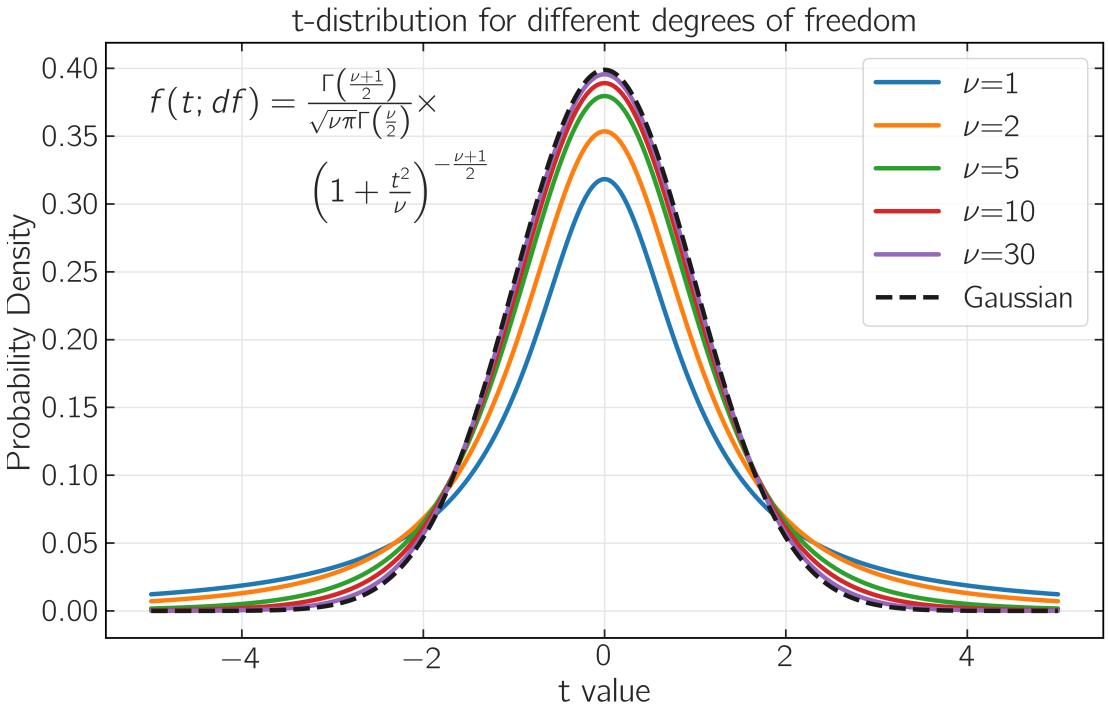

Figure 1.6: The t-distribution with different number of degrees of freedom (ν) and approach toward the Gaussian distribution as ν increases. When ν is small, the distribution has heavy tails. This reflects the increased uncertainty in estimates derived from small samples. For $\nu = 30$, the t-distribution is almost indistinguishable from the Gaussian distribution. This reflects the often used convention the for sample size $n > 30$ is enough for using the normal distribution.

Gaussian distribution as ν increases. For $\nu \to \infty$, it converges to the Gaussian distribution. Another special case is $\nu = 1$ when the distribution is the same as the Poisson, or the Lorentz, distribution (topic of the next section). When centered at zero, its mean is 0 (for $\nu > 1$); mode and median are also zero. Its variance depends on the degrees of freedom. For $\nu > 2$ it is $\frac{\nu}{\nu-2}$, ∞ for $1 < \nu \leq 2$, and otherwise it is undefined.

The t-distribution is used for estimating the mean of a normally distributed population in situations where the sample size is small, and the population standard deviation is unknown. This is because the estimate of the standard deviation from a small sample size is less reliable and tends to underestimate the true variability. The t-distribution, with its heavier tails, accounts for this extra uncertainty.

The PDF of the t-distribution for a given value t and ν degrees of freedom is given by the equation

$$f(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{1.17}$$

where Γ is the gamma function. For further details of the Γ function, see TP 1.5. For the special case of $\nu = 1$, Equation 1.17 takes the form

$$f(t;\nu = 1) = \frac{1}{\pi}\frac{1}{1+t^2}; \in \mathbb{R}. \tag{1.18}$$

This is also the same as the Cauchy, also called the Lorentz, distribution.

Technical Point 1.5.0

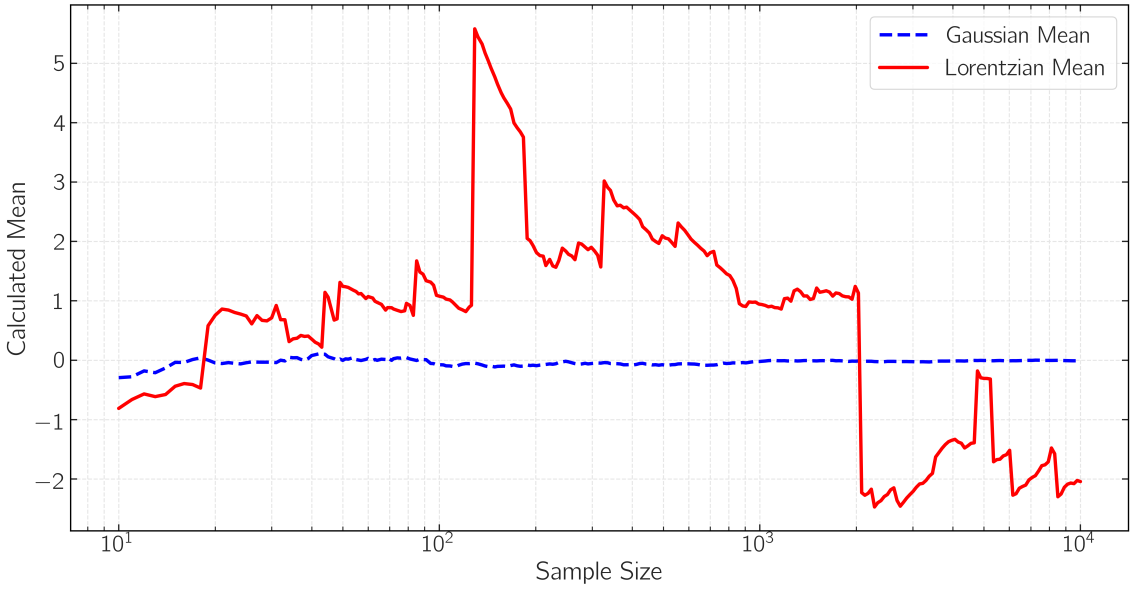

Figure 1.7: The mean of the Gaussian distribution converges as the sample size increases. The same is not true for the Lorentzian.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx, \tag{1.19}$$

This integral, known as the Euler integral of the second kind, converges for all complex numbers with a real part $\mathrm{Re}(z) > 0$. For positive integers (n), $\Gamma(n) = (n-1)!$.

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \quad \text{for } \Re(z) > 0$$

where $\mathrm{Re}(z)$ represents the real part of z.

1.6 Lorentzian distribution

The Lorentzian (or Cauchy) distribution is a continuous probability distribution. It is characterized by *heavy tails* and a peak that divides the distribution into two symmetric parts. Being heavy-tailed means that the distribution assigns higher probabilities to events far from the mean, it is able to describe phenomena with outliers or "black swan" events. Importantly, unlike almost all other distributions, the Lorentzian distribution does not have a well-defined mean or variance. This is also due to the heavy tails of the distribution, and it is manifested by the fact that as the integrals required to calculate them do not converge. For example, the mean as defined by the integral

$$\mu = \int_{-\infty}^{\infty} x f(x) dx$$

doesn't converge. Figure Note, however, that even though the mean doesn't exist, both the median and mode do. Because of that, the requirement for the applicability of CLT that the distribution has a finite variance is not fulfilled. The distribution is named after the Dutch physicist Hendrik Lorentz[3].

[3] Hendrik Antoon Lorentz (1853–1928), shared the 1902 Nobel Prize in Physics with Pieter Zeeman for the discovery and theoretical explanation of the Zeeman effect.

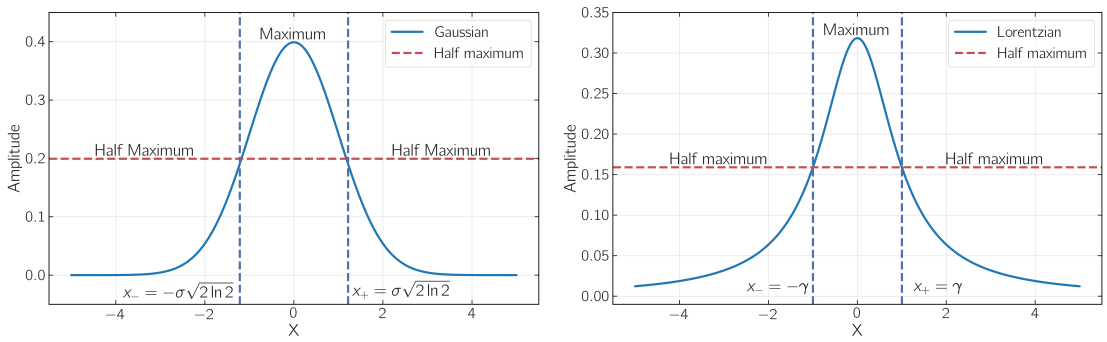

Figure 1.8: Full width at half maximum (FWHM) shown for the Gaussian and Lorentzian distributions. For the Gaussian distribution, the area within FWHM, $x \in [-\sigma\sqrt{2\ln 2}, \sigma\sqrt{2\ln 2}]$, is $\approx 78\%$. The peak of the Gaussian is $1/(\sigma\sqrt{2\pi})$. The peak of the Lorentzian is $1/(\pi\gamma)$.

The probability density function (PDF) of the Lorentzian distribution for a real variable x is given by

$$f(x;x_0,\gamma) = \frac{1}{\pi\gamma\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)} = \frac{1}{\pi}\left(\frac{\gamma}{\gamma^2+(x-x_0)^2}\right), \tag{1.20}$$

where x_0 indicates the peak's position on the x-axis, and γ controls the width (scale) of the distribution; γ is the *half the width at half maximum* (HWHM) of the peak. The parameter x_0 also defines the position of the mode and median.

One particular property the Lorentzian has is invariance under convolution, that is, that the convolution of two Lorentzian distributions is also a Lorentzian distribution. The Gaussian, Levy and exponential distributions also share that property. The Lorentzian has applications particularly in physics, signal processing and finance.

1.7 Digression: Full and half width at half maximum (FWHM and HWHM)

Full width at half maximum (FWHM) is the difference between the two values of the independent variable at which the dependent variable is equal to half of its maximum value. In other words, it is the width of a spectrum curve measured between those points on the y-axis which are half the maximum amplitude.

From the physical perspective, FWHM (and HWHM) are used in scattering. FWHM gives an estimate for the domain size. This is done with the help of the Scherrer equation.

1.8 Student's t-test

The t-test is a *statistical hypothesis test* used to determine if there is a significant difference between the means of two groups – it compares the means of two groups and determines whether they are from the same population. The t-test is typically used when the sample is less than 30, see Figure 1.6 and how it approaches the Gaussian distribution as the number of degrees of freedom increases. When the sample is larger, otehr tests such as the χ^2-test are used. The name, Student's t-distribution, has a bit of a curious origin:

Figure 1.9: Plaque of Gosset at Guinness in Dublin. He spent his whole career as

The t-test was developed in 1908 by William Sealy Gosset who used the pseudonym "Student" instead of his proper name [9]. The t-test is designed to address situations where the population standard deviation is unknown and the sample size is small [1, 5].

Assume the following situation: A company has two (or more) production lines for the same product and the company would like to know if the quality of their product from the two lines is the same or not. This is a situation in which the t-test can be applied.

There are different types of t-tests

1. One-sample t-test: Tests the mean of a single group against a known mean.
2. Two-sample t-test: There are two forms:
   (a) Independent t-test: Compares the means of two independent groups.
   (b) Paired t-test: Compares means from the same group at different times or under different conditions.

The t-test makes the following assumptions:

- The data should be approximately normally distributed.
- The data is continuous.
- The data is collected from a randomly selected portion of the total population.
- Equal or similar variances (for two-sample t-tests): Assumes homogeneity of variances.

1.8.1 Limitations and when the t-test can be used

The t-test, in its standard form, assumes that the data are approximately Gaussian distributed and that variances are equal between the groups that are being compared. The t-test can still be used in some circumstances where the data do not perfectly follow a Gaussian distribution. In particular, as discussed above, the CLT states that the distribution of sample means tends toward a Gaussian distribution regardless of the shape of the population distribution as long as the three conditions are fulfilled. Thus, for large sample sizes (usually $n > 30$), this property allows the t-test to be applied to data that are not strictly normal, as long as the sample size is sufficiently large.

Python code:

```
import numpy as np
from scipy import stats

# Generating random data
group1 = np.random.normal(100, 10, 100)
group2 = np.random.normal(110, 10, 100)

# Performing an Independent t-test
t_statistic, p_value = stats.ttest_ind(group1, group2)

print("t-statistic:", t_statistic)
print("p-value:", p_value)
```

```
def mean(data):
    return sum(data) / len(data)

def variance(data):
    n = len(data)
    mu = mean(data)
    return sum((xi - mu) ** 2 for xi in data) / (n - 1)

def stddev(data):
    return variance(data) ** 0.5

def t_test(data1, data2):
    # Calculate means
    mean1, mean2 = mean(data1), mean(data2)
```

```
    # Calculate standard deviations
    stddev1, stddev2 = stddev(data1), stddev(data2)

    # Calculate sample sizes
    n1, n2 = len(data1), len(data2)

    # Calculate pooled standard deviation
    sp = (((n1 - 1) * variance(data1) + (n2 - 1) * variance(data2)) / (n1 + n2 - 2)) ** 0.5

    # Calculate t-statistic
    t = (mean1 - mean2) / (sp * ((1/n1 + 1/n2) ** 0.5))

    # Degrees of freedom
    df = n1 + n2 - 2

    return t, df

# Example datasets
data1 = [1, 2, 3, 4, 5, 6, 7, 8, 9]
data2 = [2, 3, 4, 5, 6, 7, 8, 9, 10]

# Perform t-test
t_stat, df = t_test(data1, data2)

print(f"T-statistic: {t_stat}")
print(f"Degrees of Freedom: {df}")

# Simple approximation to estimate the p-value (not accurate for actual analysis)
# This part is highly simplified and should not be used for actual statistical analysis
# For a real analysis, you would use the t-distribution to find the p-value
p_value_approx = 2 * min((0.5 ** abs(t_stat)), 1 - (0.5 ** abs(t_stat)))
print(f"Approximated P-value (not accurate): {p_value_approx}")
```

■ **Example 1.5** t-test The independent two-sample t-test conducted on the two datasets, both generated with the same mean (50) and standard deviation (10), resulted in a T-statistic of approximately -0.15 and a P-value of approximately 0.879. Given that the P-value is much greater than the typical significance level of 0.05, we conclude that there is no significant difference between the two distributions. This outcome is expected since the two datasets were generated with identical parameters, illustrating a scenario where the t-test correctly identifies two distributions as being statistically similar.

1.9 χ^2 distribution and test

1.9.1 χ^2 distribution

The (χ^2) distribution is a continuous probability distribution that is widely used in statistical inference, particularly in hypothesis testing and in the construction of confidence intervals. Its mean is equal to k, the number of degrees of freedom, the mode is given by $\max(k-2,0)$ and the median is $\approx k(1-2/(9k))^3$. The difference between the previous distributions and the χ^2 distribution is that whereas the others are used in direct modelling of data, the χ^2 distribution is rather used for hypothesis testing, in particular the χ^2 test(s) as will be describe in the next section. Note also that the χ^2 distribution is not symmetric and it is defined only for non-negative values, the random variable is defined as

$$\chi^2(n) = \sum_{i=1}^{n} x_i^2. \tag{1.21}$$

The probability density function (PDF) of the chi-square distribution for a variable x and ν degrees of freedom is given by

$$f(x;\nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, \tag{1.22}$$

where Γ is the Gamma function, Equation 1.19.

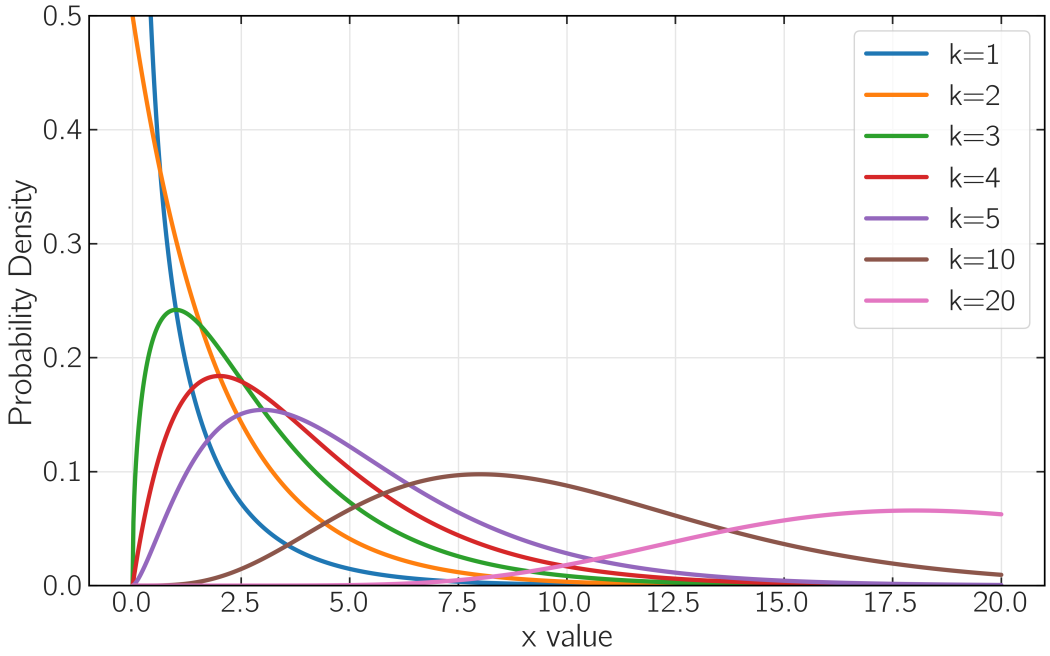

Figure 1.10: The χ^2 distribution for different number of degrees of freedom (k). As $k \to \infty$, the distribution approaches a Gaussian distribution.

1.9.2 χ^2-test

The χ^2-test is a statistical tool widely used for testing relationships between categorical variables. It s a non-parametric statistical test used to determine if there is a significant association between two categorical variables in a sample. The test compares the observed frequencies of the data to the expected frequencies under the *null hypothesis* of no association between the variables. Example uses include gene frequency studies and genetic cross analysis, consumer preferences and habits, and surveys and observational studies to understand relationships between different social factors.

The χ^2 statistic (the Pearson form) is calculated as

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{1.23}$$

where χ^2 is the test statistic, O_i is the number of observations, E_i is the expected frequency. The χ^2 test statistic measures how much the observed frequencies deviate from the expected frequencies.

The test has the following assumptions:

- Sample Size: A sufficient sample size is required to ensure the validity of the test.
- Independence: Data points must be independent.
- Frequency: Expected frequencies should be at least 5 for a valid test.

Python example:

```
import numpy as np
from scipy.stats import chi2_contingency

# Example data: Observed frequencies in a contingency table
data = np.array([[30, 10], [20, 40]])

# Performing the Chi-Squared Test
chi_square, p_value, degrees_freedom, expected = chi2_contingency(data)

print("Chi-Square Statistic:", chi_square)
print("p-value:", p_value)
```

```
print("Degrees of Freedom:", degrees_freedom)
print("Expected Frequencies:\n", expected)
```

1.10 Kolmogorov-Smirnov (K-S) test

The Kolmogorov-Smirnov (K-S) test is a *non-parametric* statistical test that compares the empirical distribution function of a sample with a reference probability distribution, or compares the empirical distribution functions of two samples. The K-S statistic measures the difference (distance) between the cumulative distribution function (CDF) of the reference and the measured distribution function.

1.10.1 One-Sample K-S Test:

In the one-sample K-S test, the test statistic is calculated as the maximum absolute difference between the reference CDF and the empirical cumulative distribution function (ECDF) of the sample. If $F_{n1}(x)$ is the ECDF of the sample and $F(x)$ is the CDF of the reference distribution, then

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{1.24}$$

where sup denotes the supremum.[4]

1.10.2 Two-Sample K-S Test:

In the two-sample K-S test, the test statistic is the maximum absolute difference between the ECDFs of the two samples (there is no reference CDF). If $F_{n1}(x)$ and $F_{n2}(x)$ are the ECDFs of the two samples, then

$$D_{n1,n2} = \sup_x |F_{n1}(x) - F_{n2}(x)|. \tag{1.25}$$

1.10.3 Interpretation and hypothesis testing

In both versions above, the null hypothesis states that the samples are drawn from the same distribution (or from a specified distribution in the one-sample test). A small value of D implies that the empirical distribution of the sample is close to the reference distribution, suggesting that the null hypothesis cannot be rejected at a given significance level.

■ **Example 1.6** One-Sample K-S Test. Suppose we have a sample data set and we want to test if it follows a normal distribution. We can use the `kstest` function from `scipy.stats`.

```
import numpy as np
from scipy import stats

# Sample data
data = np.random.normal(0, 1, 1000)

# Perform K-S test against a normal distribution
D, p_value = stats.kstest(data, 'norm')

print(f"K-S statistic: {D}")
print(f"P-value: {p_value}")
```

[4] Supremum is the least upper bound. The supremum of a subset S of a partially ordered set T is the least element in T that is greater than or equal to every element in S. In simpler terms, it's the smallest upper bound of S. Example: Consider the set $A = \{x \in R : x^2 < 2\}$, which consists of all real numbers whose square is less than 2. The set A has no maximum value because there is no largest real number that satisfies $x^2 < 2$. However, the supremum of A is $\sqrt{2}$, as $\sqrt{2}$ the least real number that is greater than every element in A.

■ **Example 1.7** To compare two independent samples and test if they come from the same distribution:

```
# Two independent samples
data1 = np.random.normal(0, 1, 1000)
data2 = np.random.normal(0.5, 1.5, 1000)

# Perform two-sample K-S test
D, p_value = stats.ks_2samp(data1, data2)

print(f"K-S statistic: {D}")
print(f"P-value: {p_value}")
```

1.11 Does sampling matter?

Let's examine a few examples:

■ **Example 1.8** Literary Digest Poll – 1936 U.S. Presidential Election [2, 6].

One of the most famous examples of bad sampling is the 1936 U.S. Presidential election. The Literary Digest, a popular magazine at the time, conducted a poll predicting that Alf Landon would defeat Franklin D. Roosevelt by a large margin. The poll was based on over 2 million returned questionnaires sent to magazine subscribers, car owners, and telephone users. This sampling method was biased towards wealthier Americans who were more likely to oppose Roosevelt. Roosevelt won the election in a landslide, showcasing the poll's failure due to its non-representative sample.

■ **Example 1.9** Early COVID-19 Infection Rates and Case Fatality Rates [10, 11].

During the early stages of the COVID-19 pandemic, many regions reported infection rates and case fatality rates based on available testing data. However, limited testing capacity and prioritization of testing for severe cases or healthcare workers meant that mild or asymptomatic cases were often underrepresented. This skewed sampling led to an overestimation of the case fatality rate and misunderstanding of the virus's spread in the general population.

■ **Example 1.10** The Volunteer's Dilemma. [3, 4]

Many psychological studies rely on volunteers, typically college students, due to convenience and accessibility. This practice has been criticized for producing results that may not generalize to the broader population, as volunteers might differ systematically from those who do not participate (e.g., in terms of personality, socioeconomic status, or health).

1.12 Problems

Problem 1.1 Do not use any Python libraries for this problem (=pure Python only): Write a code that computes the median for an arbitrary dataset.
Problem 1.2 Show that the alternative way to compute σ^2 given in Equation 1.8 is true.
Problem 1.3 Write a Python code that uses Equation 1.8 to compute the population variance.
Problem 1.4 Equation 1.9 computes the *sample variance*, that is division is made by $N_A - 1$ instead of N_A (entire dataset). What is the difference between the two cases and where does it come from. Write a Python code and plot the difference between the two as a function of the sample size.
Problem 1.5 Let $\mu = 170$ and $\sigma = 7$. Prove, showing all steps, that Equation 1.13 is true.

References

[1]Student, "The probable error of a mean", Biometrika **6**, 1–25 (1908).

[2]P. Squire, "Why the 1936 Literary Digest poll failed", Public Opin. Q. **52**, 125–133 (1988).

[3]J. Weesie, "Asymmetry and timing in the Volunteer's Dilemma", J. Conflict Resolut. **37**, 569–590 (1993).

[4]R. Manning, M. Levine, and A. Collins, "The Kitty Genovese murder and the social psychology of helping: the parable of the 38 witnesses", Am. Psychol. **62**, 555–562 (2007).

[5]S. L. Zabell, "On Student's 1908 article "the probable error of a mean"", J. Am. Stat. Assoc. **103**, 1–7 (2008).

[6]D. Lusinchi, ""President" Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?", Soc. Sci. Hist. **36**, 23–54 (2012).

[7]B. P. Abbott et al., "Observation of gravitational waves from a binary black hole merger", Phys. Rev. Lett. **116**, 061102 (2016).

[8]M. Baker, "1,500 scientists lift the lid on reproducibility", Nature **533**, 452–454 (2016).

[9]M. C. Wendl, "Pseudonymous fame", Science **351**, 1406 (2016).

[10]D. D. Rajgor, M. H. Lee, S. Archuleta, N. Bagdasarian, and S. C. Quek, "The many estimates of the COVID-19 case fatality rate", Lancet Infect. Dis. **20**, 776–777 (2020).

[11]COVID-19 Forecasting Team, "Variation in the COVID-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis", Lancet **399**, 1469–1488 (2022).

[12]"Trust but verify", Nat. Mater. **23**, 1 (2024).