Fondamenti di Analisi dei Dati

from data analysis to predictive techniques

Prof. Antonino Furnari (antonino.furnari@unict.it)
Corso di Studi in Informatica
Dip. di Matematica e Informatica
Università di Catania

Statistical Inference

Moving beyond describing samples to understanding entire populations through the power of statistical inference.

From S

In practice, we rarely have access to entire populations. Whether predicting election outcomes, monitoring manufacturing quality, or studying disease relationships, collecting complete population data is often unfeasible or impossible.

We have to resort to **analysing the sample**. However, how can we be sure that **the sample is representative of the population?** What if the conclusions we draw on samples do not generalise to the population?

Statistical inference provides the solution: we analyse carefully selected samples to draw reliable conclusions about populations.

Election Forecasting

Predicting vote percentages without interviewing every voter

Quality Control

Detecting defects without testing every product

Health Research

Understanding disease links in global populations

The Art of Sampling

Sampling is the foundation of inferential analysis. Whether working with pre-collected datasets or gathering new data, understanding sampling properties is crucial for drawing valid conclusions.

The easiest way to sample from a population is **randomly**. A simple random sample makes two assumptions:

01

Unbiasedness

Each population element has equal probability of selection, ensuring representative samples

02

Independence

Selecting one element doesn't influence selection of others, maintaining statistical validity

If we collect a large enough random sample, we usually have a good representative of the population.

When Sampling Goes Wrong

Common Pitfalls

Understanding sampling failures helps us design better studies and critically evaluate research claims.

1

Selection Bias

Example: Surveying only city centre visitors oversamples certain demographics whilst excluding peripheral residents entirely.

2

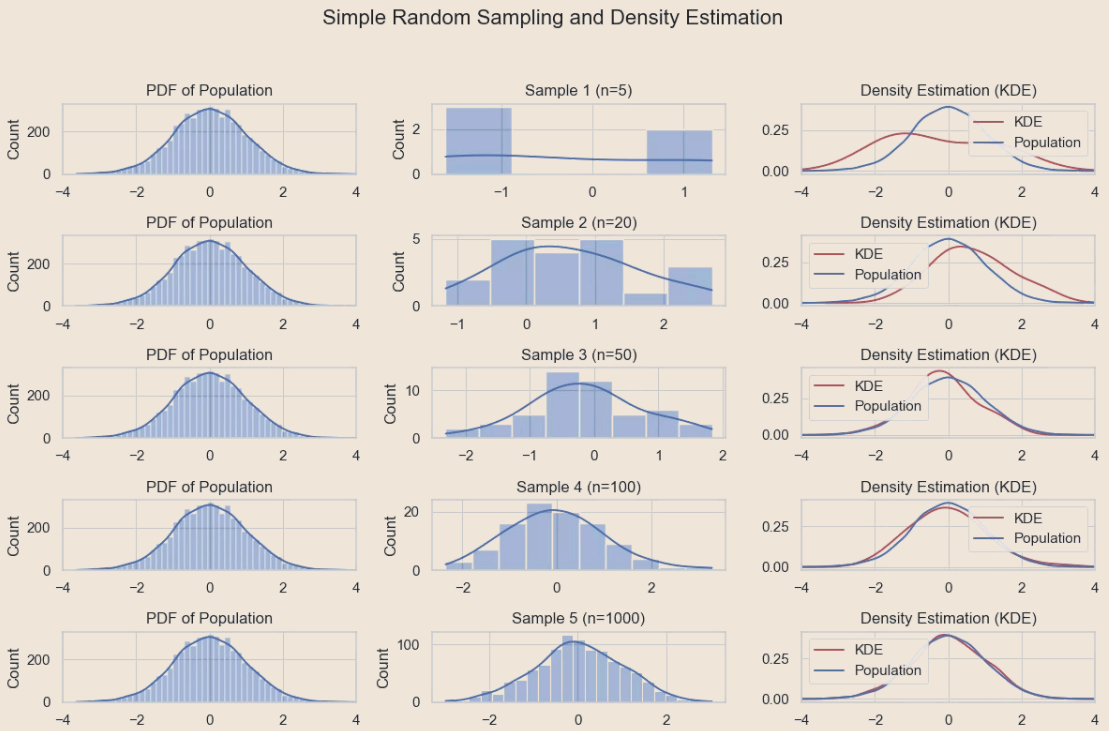
Response Bias

Example: Online surveys capture only highly motivated respondents, missing the silent majority's perspectives.

3

Coverage Bias

Example: Phone surveys miss people without phones and oversample those with multiple numbers.



Sample Size Matters

Larger samples provide more accurate estimates of population parameters. With small samples, estimates fluctuate wildly; with larger samples, they stabilise around the true population value.

The visualisation demonstrates this principle: as sample size increases from 5 to 1,000, the estimated distribution converges remarkably towards the population distribution.

Stratified Sampling: When Populations Are Complex

When populations contain distinct subgroups with different characteristics, simple random sampling may miss important diversity.

Stratified sampling addresses this by dividing the population into homogeneous strata (such as occupations) and sampling proportionally from each group.

This ensures all subgroups are adequately represented, though it requires prior knowledge of population structure.

The figure illustrates the ffect of stratified sampling on the total distribution, as compared to a simple random sample.

Sampling in Python

Python's NumPy and Pandas libraries make sampling straightforward. We can easily sample from known distributions or extract subsamples from existing datasets.

Sampling from Distributions

```
import numpy as np

# Normal distribution
samples = np.random.normal(
    mean=0, std=1, size=1000
)

# Exponential distribution
samples = np.random.exponential(
    scale=2, size=1000
)
```

Sampling from DataFrames

```
import pandas as pd

# Simple random sample
sample = data.sample(1000)

# Stratified sampling
stratified = (
 data.groupby('category')
 .apply(lambda x: x.sample(100))
)
```

The Sampling Distribution of the Mean

Consider a bakery producing 1kg cookie packages. Due to production variations, individual packages vary slightly in weight. If we repeatedly sample packages and calculate means, we get a **distribution of sample means**.

1000.2g	999.7g	1000.5g
Sample 1 Mean	**Sample 2 Mean**	**Sample 3 Mean**
First random sample	Second random sample	Third random sample

What's the "real mean" of packages?

A Random Variable for the Mean

We treat each package weight's as a random variable X_i, with:

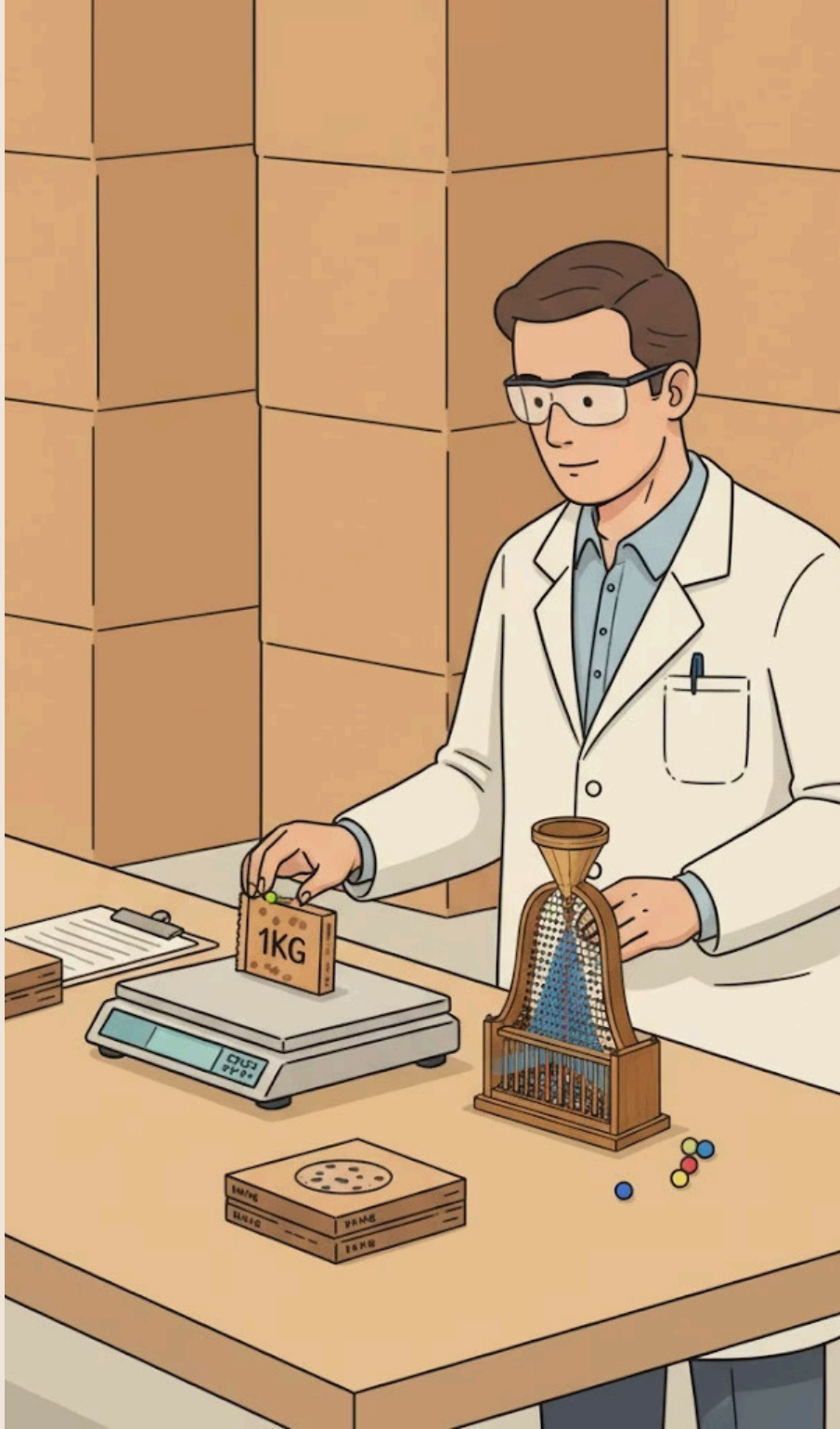
$$E[X_i] = \mu$$
$$Var[X_i] = \sigma^2$$

Where μ and σ^2 are the mean package's weight and variance **in the population**. Note that we don't have access to the population, so we don't have access to these values.

Given a sample of n elements, we can see the sample mean itself as a random variable:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

The random experiment associated with this variable consists in taking a random sample and measuring its means. As we change the sample, also the mean will change.



The Distribution of the Means

If we compute the mean multiple times, using different samples, this will vary, practically leading it to a **distribution of mean values**.

> Note that, while we cannot make assumptions on the distribution of X_i, **for the Central Limit Theorem we know that $\overline{X}$ will be Gaussian for large n** regardless of the shape of X_i.

We can easily compute the mean and variance of this distribution as follows:

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$$

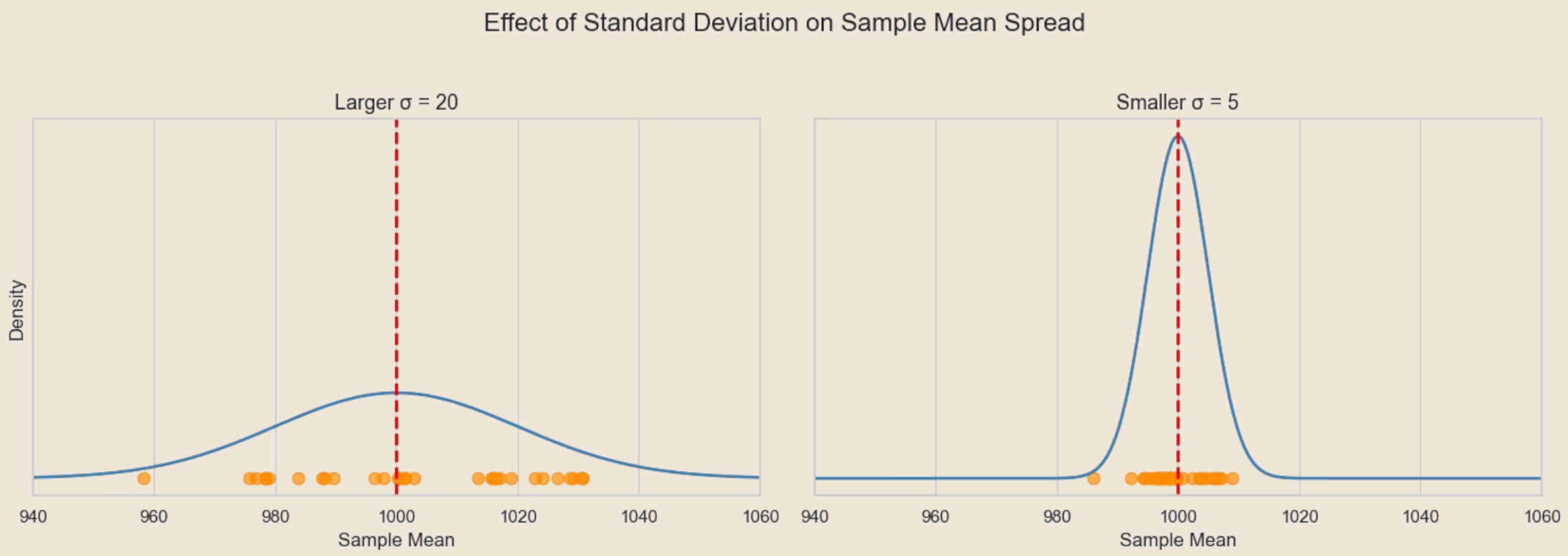
$$Var[\overline{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{\sigma^2}{n}$$

$$Std[\overline{X}] = \frac{\sigma}{\sqrt{n}}$$

We note that:

- The **expected value of the sample means** (the mean of the means) converges to the **mean of the population**. This is the value we want to estimate: the actual average weight of packages.
- The **standard deviation of the distribution** quantifies the precision according to which we can measure the mean. A small standard deviation indicates that sample means have small variability, so a single estimate will likely be closer to the true mean and hence more reliable.

This is exemplified by the following figure:



Standard Error: Quantifying Uncertainty

The standard deviation of the sampling distribution depends on the unknown population variance. In practice, we estimate it using the **standard error**:

$$SE(\overline{x}) = \frac{s_{n-1}}{\sqrt{n}}$$

1

Double the Sample

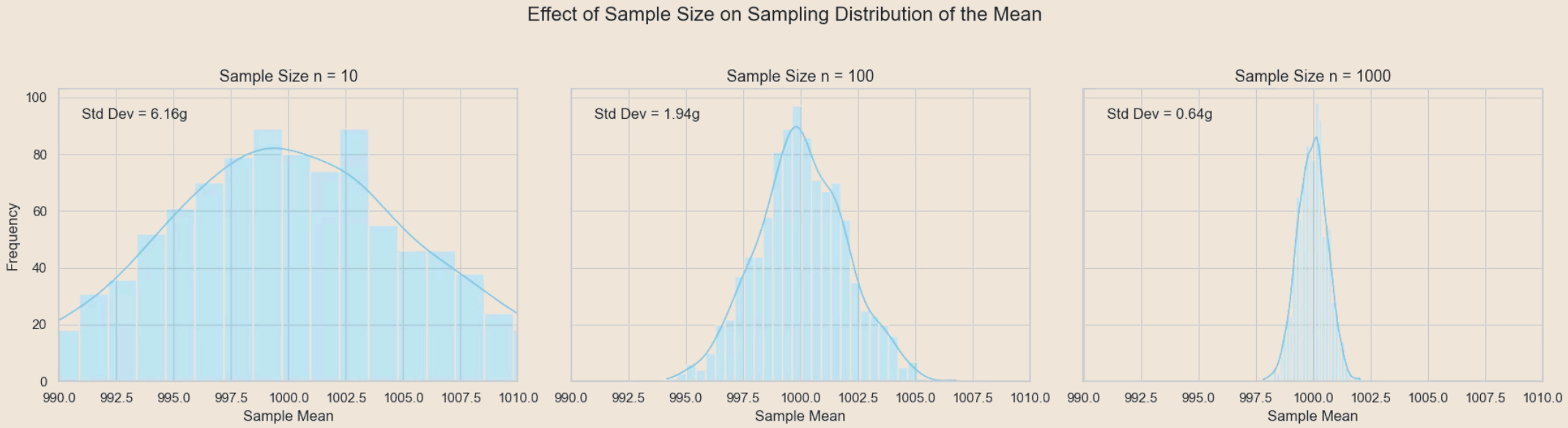
Error reduces by ~30%

2

Quadruple the Sample

Error reduces by 50%

The standard error tells us how close our sample mean likely is to the true population mean. Smaller values indicate greater precision and reliability.



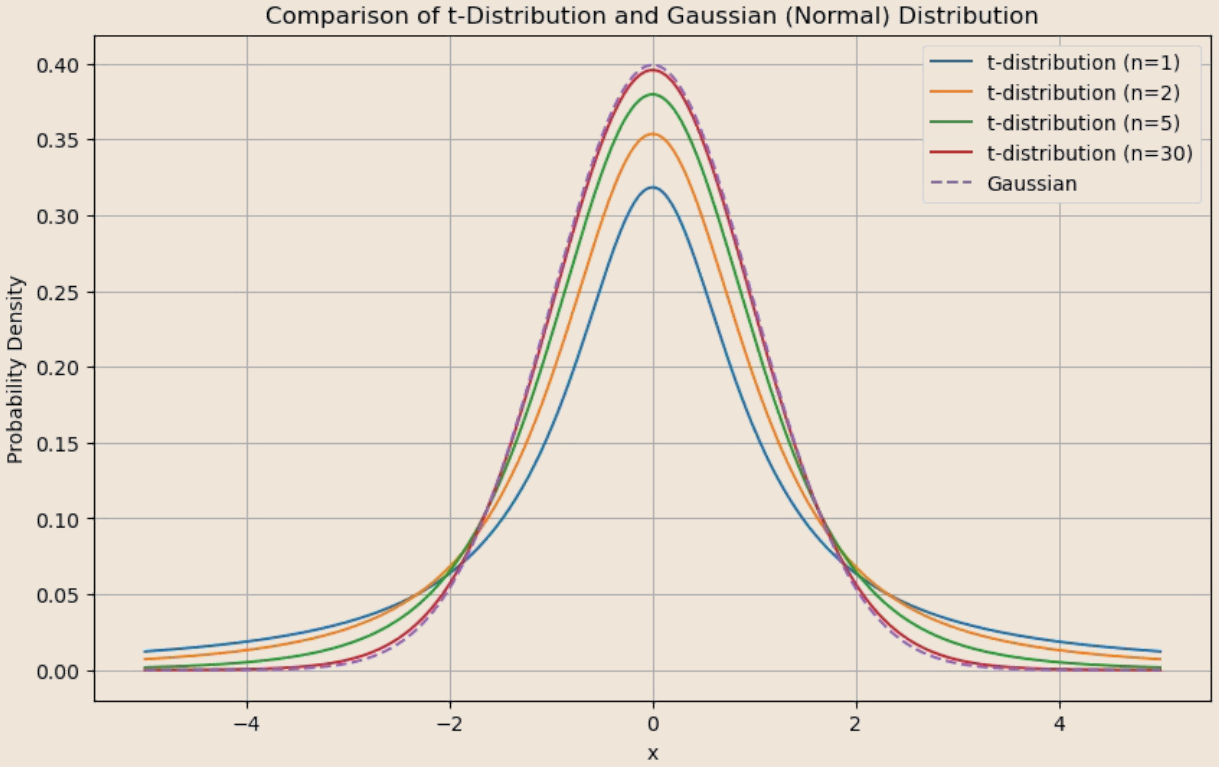
We often find $SE \propto \frac{1}{\sqrt{n}}$

The t-Student Distribution

For small samples, the sampling distribution isn't perfectly normal. The **t-Student distribution** accounts for additional uncertainty with heavier tails, allowing more extreme values.

As sample size increases, the t-distribution converges to the normal distribution. The distribution has n-1 degrees of freedom, where n is the sample size.

$$t_{n-1} = \frac{\overline{X} - \mu}{s_{n-1}/\sqrt{n}}$$



Note that the t-Student distribution is obtained by applying standardization using the standard error rather than the standard deviation (which is unknown).

Confidence Intervals: Quantifying Uncertainty

Let's get back to our example of mean package weights. Let's say we take a single sample of size $n = 1000$ and measure:

$$\overline{x} = 1000.2g$$

This is an estimate of our true mean. However, we know that if we repeat the sampling, the number will change. **How shall we deal with this variation**?

The idea is to **estimate the bounds of this variation** and, rather than computing a single estimate, return an **interval** within which we are confident enough that the true mean will be.

We will report these confidence intervals up to a confidence level, e.g., 95% and interpret them as follows:

> The true mean lies in the estimated interval with 95% confidence, i.e., if we were to repeat the sampling 100 times, 95 of these times, the true mean will be in the interval.

We also call $\alpha = 1 - \text{confidence level}$ the **significance level**. For a confidence level of 95%, we set $\alpha = 0.05$.

Defining Confidence Intervals

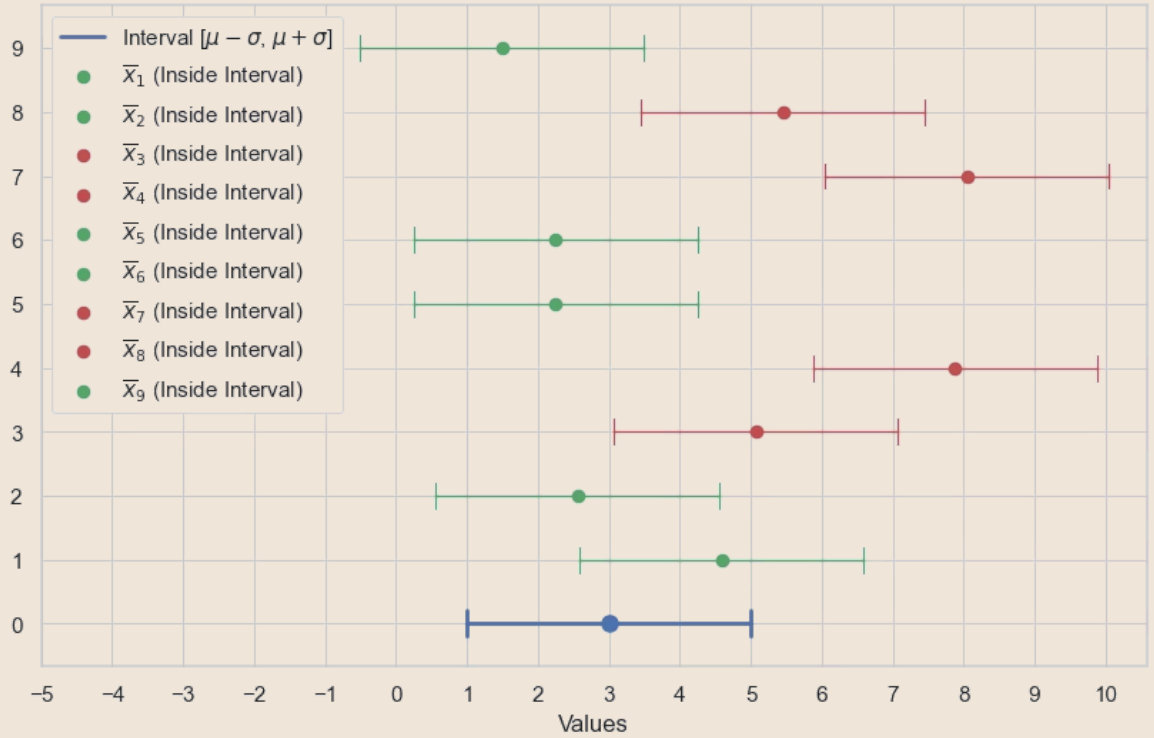
Since X is Gaussian, sampling $\overline{x}$ from X, we know that:

$$P(\mu - \sigma < \overline{x} < \mu + \sigma) = 0.683$$

Which is equivalent to:

$$P(\overline{x} - \sigma < \mu < \overline{x} + \sigma) = 0.683$$

> If we repeat sampling, 68.3% of the times, the true mean (the one we want to find) is in the confidence interval $[\overline{x} - \sigma, \overline{x} + \sigma]$



In practice, libraries will automatically find the appropriate β value (for Gaussians $\beta = 1.96$) such that:

$$P(\overline{x} - \beta\sigma < \mu < \overline{x} + \beta\sigma) = 1 - \alpha$$

Computing Confidence Intervals

We still need σ, the standard deviation of the population, to compute the bounds of the confidence interval. However, we do not have access to this value. Instead of it, we will use the Standard Error:

$$P(\overline{x} - \beta SE(\overline{x}) < \mu < \overline{x} + \beta SE(\overline{x})) = 1 - \alpha$$

Hence, we obtain the following confidence interval:

$$[\overline{x} - \beta SE(\overline{x}), \overline{x} + \beta SE(\overline{x})]$$

Let's get back to our cookie example. We measure:

$$\overline{x} = 1000.2g \qquad s_{n-1} = 0.1g \qquad SE(\overline{x}) = \frac{s_{n-1}}{\sqrt{n}} = 0.02g$$

We hence obtain the following confidence interval:

$$[1000.2 - 1.96 \cdot 0.02,\ 1000.2 + 1.96 \cdot 0.02] = [1000.18,\ 1000.22]$$

This means we are **95% confident** that the true average weight of cookie packages lies between **1000.18g and 1000.22g**.

Computing Confidence Intervals in Python

SciPy provides convenient functions for calculating confidence intervals across different scenarios.

Confidence Intervals for Means

```
from scipy import stats

interval = stats.norm.interval(
 confidence_level=0.95,
    loc=mean,
   scale=standard_error
)
```

Confidence Intervals for Variances

```
from scipy import stats

confidence_interval =
stats.chi2.interval(
 confidence_level=0.95,
    df=len(data) - 1
)
```

Confidence Intervals for Proportions

```
import statsmodels.api as sm
conf_interval =
sm.stats.proportion_confint(
 count, total,
 alpha=0.05
)
```

Bootstrapping: Inference Through Resampling

What if formulas don't exist for your statistic (e.g., median)? **Bootstrapping** uses computational power to estimate sampling distributions by repeatedly resampling from your original sample. This is also useful when the sample is very small and assuming Gaussianity is a bit more risky.

Start with Original Sample

Your dataset of size n

Resample with Replacement

Create new samples of size n

Calculate Statistic

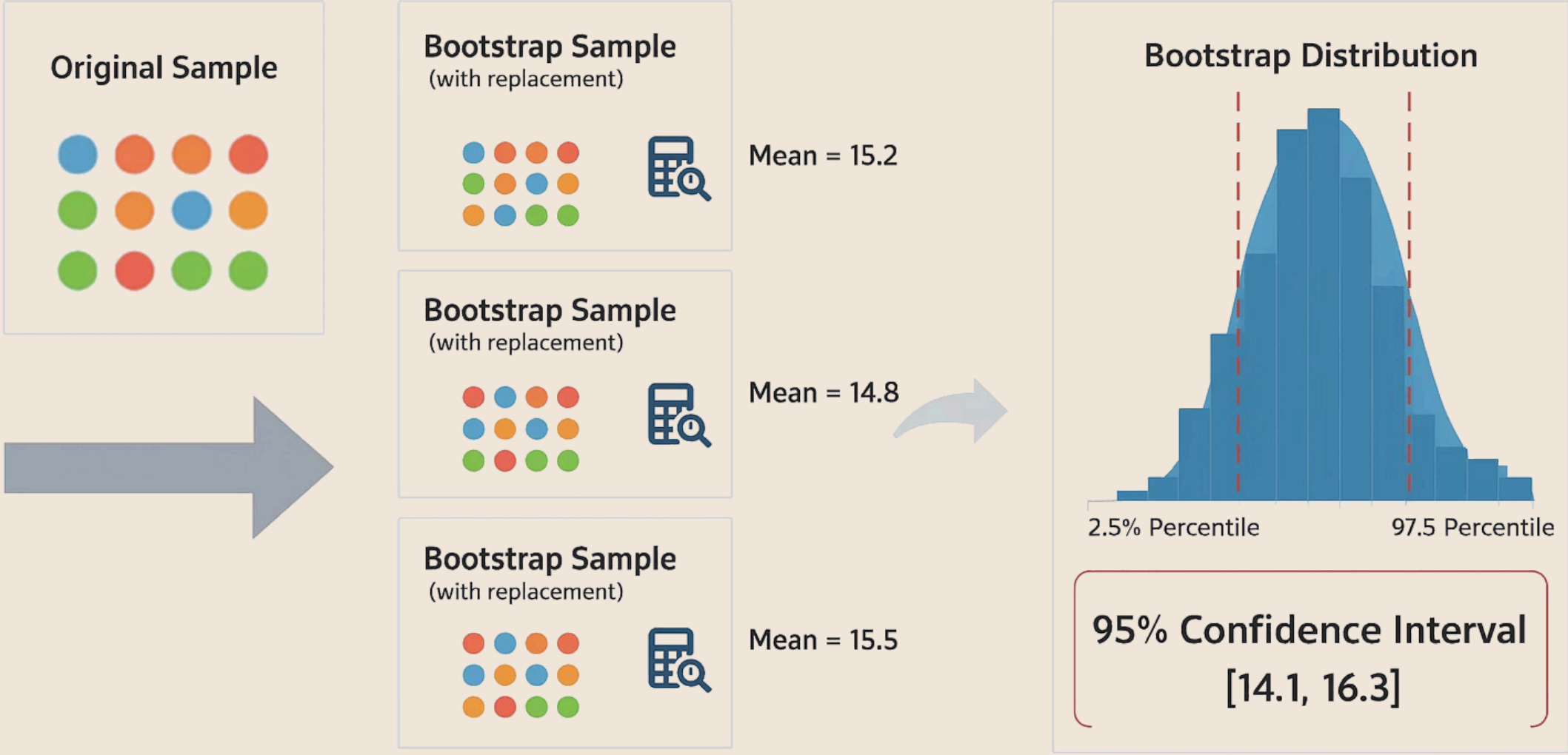
Compute your metric of interest

Repeat Thousands of Times

Build empirical distribution

Take percentiles to find interval

Take the 2.5% and 97.5% percentiles to find 95% confidence interval



Bias of an Estimator

Formulas used to compute a property of a population are called **estimators**. For instance, Let X be a random variable with mean ϕ and let $\{x_1, x_2, \ldots, x_n\}$ be a sample from the population, then:

$$T(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Is an estimator of the mean ϕ from the sample. Note that, since T depends on the sample, it is a **random variable**. Also, it can be **inaccurate**. To measure such inaccuracy, we define the bias of the estimator as follows:

$$Bias_{\phi}(T(X)) = E[T(X)] - \phi$$

The bias measures the average deviation of the estimate from the true value. The larger the bias, the more our estimates deviate from the true value. If we find that

$$E[T(X)] = \phi \qquad Bias_{\phi}(T(X)) = 0$$

Then we say that the estimator is **unbiased**. If the bias is different from zero, then our estimator **systematically underestimates or overestimates the true value**.

Unbiased Estimator of the Variance

To estimate the variance of cookie weight, we may use:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

This estimator **is biased** as it tends to overestimate the true variance. Indeed, it can be shown that:

$$E[s_n^2] = \frac{n-1}{n}\sigma^2$$

To correct this, we use the **unbiased estimator**:

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Variance of an Estimator

The **variance** of an estimator tells us how much the estimate fluctuates across different samples. It is defined as:

$$Var(T(X)) = E[(T(X) - E[T(X)])^2]$$

A low variance means that repeated samples give similar results. A high variance means that estimates are unstable and vary widely.

Bias-Variance Trade-Off

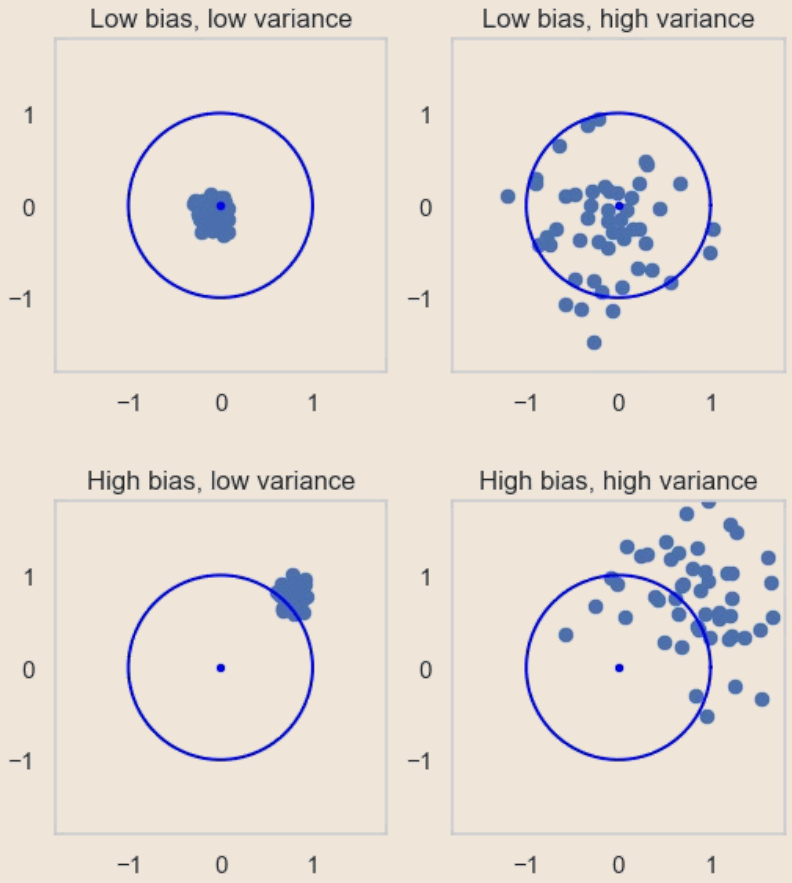
deally, we want an estimator with **low bias** and **low variance**. This means that:

- The estimates are close to each other (low variance),
- And they are close to the true value (low bias).

In practice, we can visualize four scenarios:

- **Low bias, low variance**: estimates are tightly clustered around the true value.
- **Low bias, high variance**: estimates are scattered but centered correctly.
- **High bias, low variance**: estimates are consistent but systematically wrong.
- **High bias, high variance**: estimates are scattered and off-target.

This is often illustrated as a target with darts: the true value is the bullseye, and each dart is an estimate. The goal is to hit close to the center, consistently.

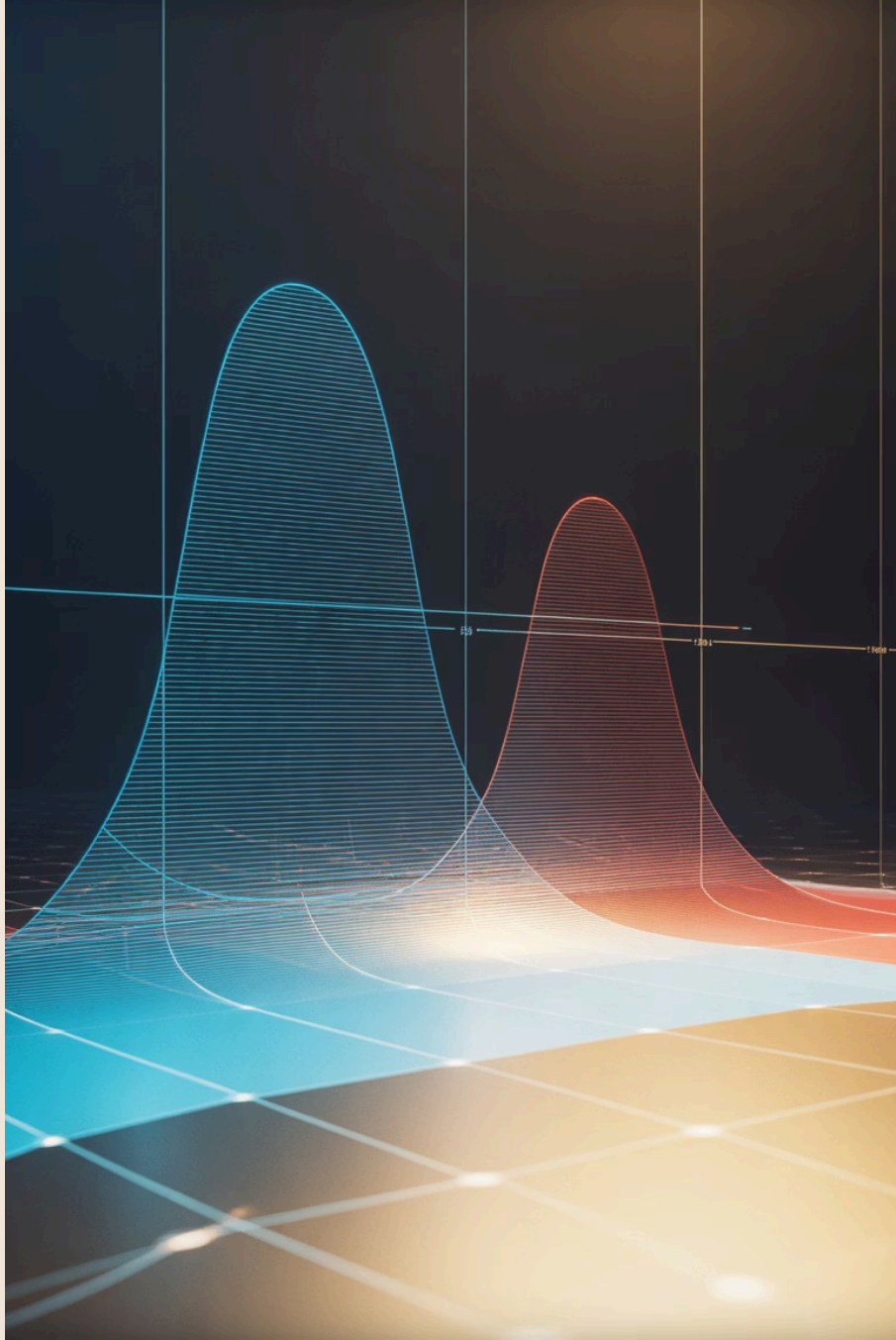


Hypothesis Testing

Confidence intervals provide a range of plausible values for a population parameter based on a sample. A **hypothesis test**, instead, is used to **challenge a specific claim** about that parameter. Examples of hypotheses we might want to test include:

- The average weight of cookie packages is exactly 1000g.
- Two different ovens produce cookies with the same average weight.
- The proportion of underweight packages is below a regulatory threshold.

If the hypothesis is rejected, we conclude that the claim is likely false. Otherwise, we do not have enough evidence to reject it, and we act as if it were true.



Hypothesis Testing and Cookies

In our cookie example, we set the machine to produce packages of 1Kg and measure:

$$\overline{x} = 1000.0005\text{g}, \quad s_{n-1} = 0.01\text{g}$$

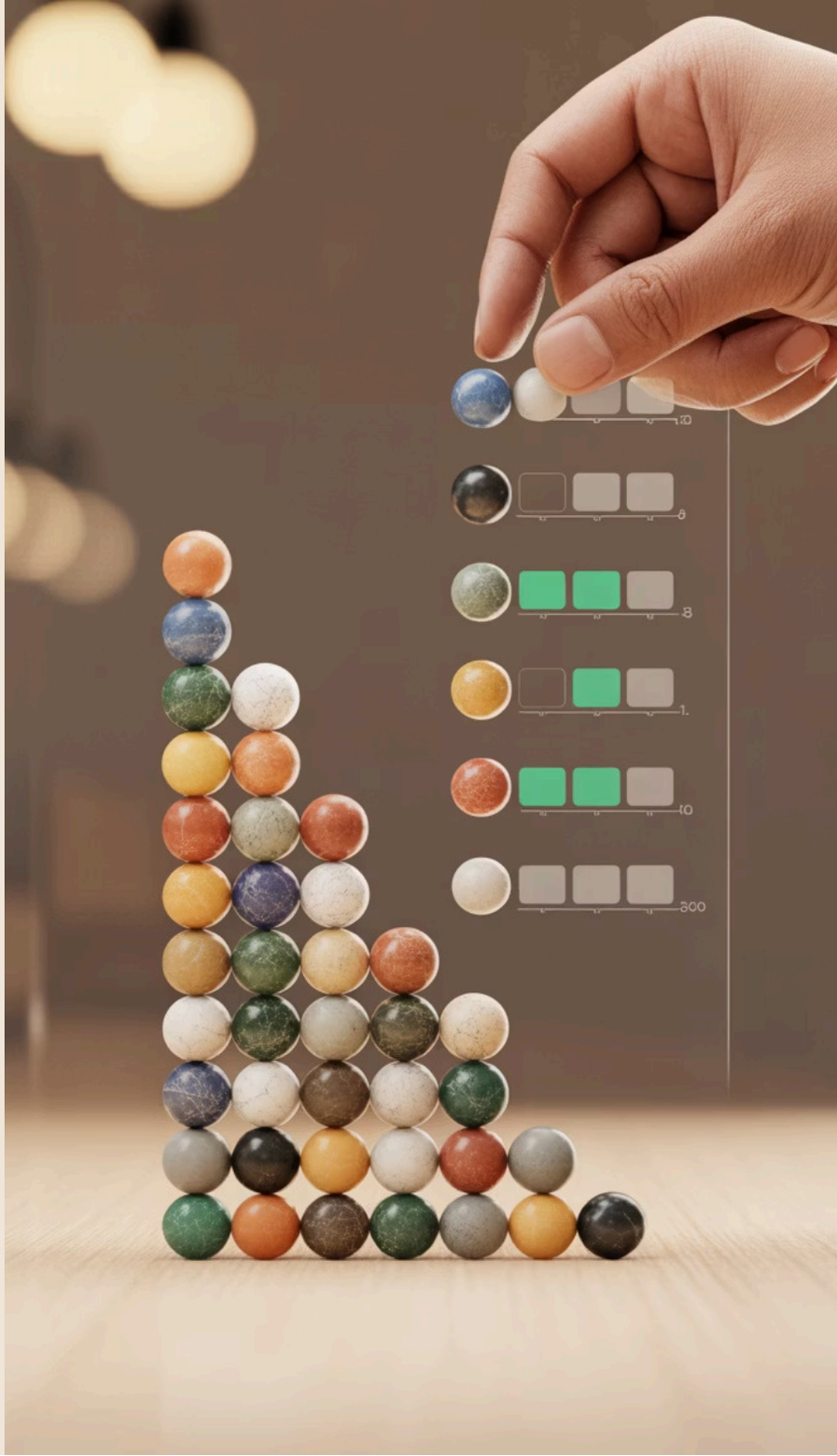
We assume the small deviation is due to measurement noise or natural variation, and we are inclined to believe that the **true mean is still** $\mu = 1000g$**.**

However, our quality control manager raises a concern: what if the population mean is **not** 1000g? She proposes a formal test to challenge the assumption. We define:

- **Null hypothesis** (H_0): the population mean is g
- **Alternative hypothesis** (H_a): the population mean is different from 1000g

The null hypothesis is the hypothesis we are trying to reject (what we are trying to prove). If we do so, then we embrace the alternative hypothesis.

Before proceeding, we ask what margin of error is acceptable. The manager says she can tolerate a **5% chance of wrongly rejecting** H_0. This defines our **significance level** $\alpha = 0.05$**.**



Test Statistic

We now ask: **how much does our sample mean $\overline{x} = 1000.1g$ deviate from the assumed mean $\mu_0 = 1000g$?** More precisely, **what is the probability of observing a difference this large (or larger) just by chance?**

To answer this, we compute the **test statistic** using the t-distribution:

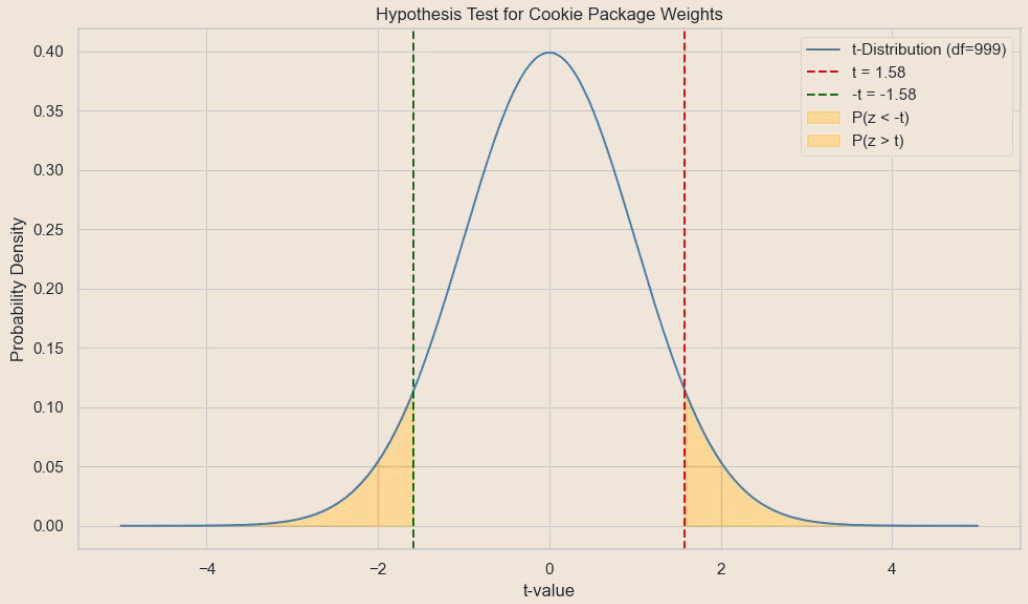
$$t = \frac{\overline{x} - \mu_0}{s_{n-1}/\sqrt{n}} = \frac{1000.0005 - 1000}{0.01/\sqrt{1000}} = 1.58$$

This tells us how many standard errors our estimate is away from the hypothesized mean.

We now ask: **what is the probability of observing a value this extreme or more extreme, assuming H_0 is true?** This is the **p-value**, defined as:

$$P(|z| > |t|)$$

This is the area under the tails of the t-distribution beyond t and $-t$, as shown on the right.



Since the t-Student distribution is symmetrical, we can easily compute the p-value as:

$$\text{p-value} = 2(1 - CDF_t(t))$$

In our case, we obtain:

$$p - value = 0.11$$

11% of the times, we obtain a deviation larger than the one we are observing. If we reject the null hypothesis, we make a mistake 11% of the times, which is larger than 5%, so we have to reject the test!

Does this mean that H_0 is confirmed? **No**, we just cannot say much - we have to collect more data and try again.

Understanding Test Errors

Hypothesis tests can produce two types of errors, each with different consequences depending on the application context.

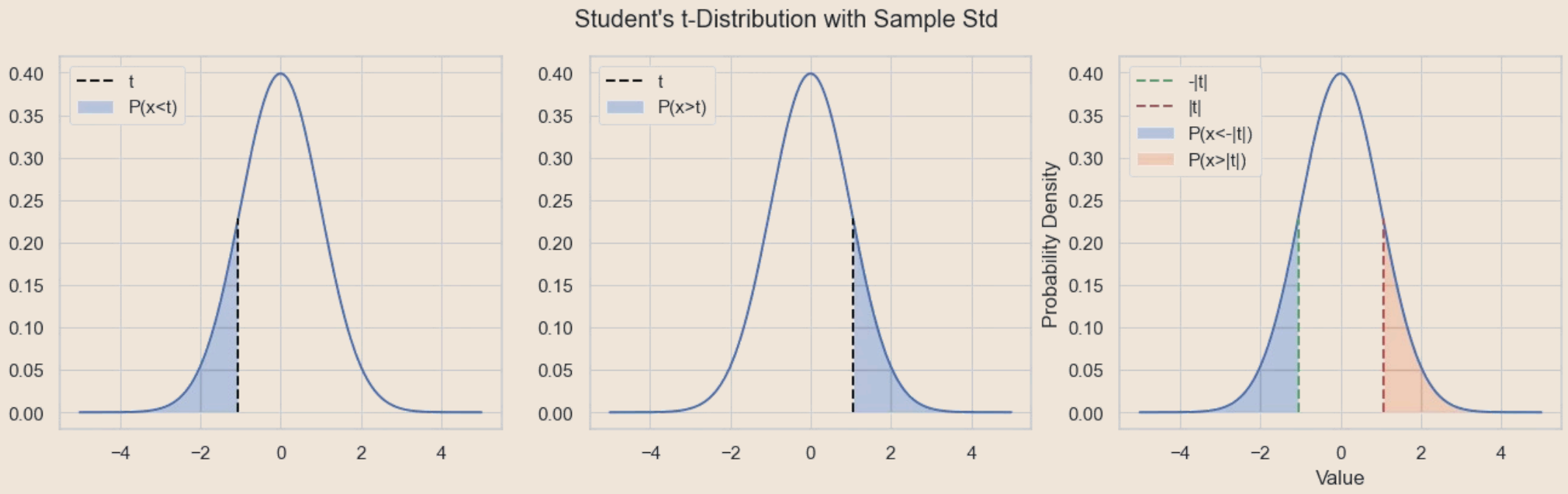
Reality	**Reject H_0**	**Fail to Reject H_0**
H_0 True	❌ Type I Error (False Positive) Probability: α	✅ Correct Decision
H_0 False	✅ Correct Decision	❌ Type II Error (False Negative)

One-tailed vs Two-tailed Tests

The type of hypothesis test performed depends on the alternative hypothesis. Previously, we examined a "**two-tailed test**" where we were interested in deviations in either direction from the null hypothesis. However, depending on the specific alternative hypothesis, we may use one-tailed tests:

- If the alternative hypothesis has the form $\mu \neq \mu_0$, we conduct a **two-tailed test**. This checks if the sample mean deviates significantly from the assumed population mean in either the positive or negative direction. The p-value is calculated as $P(|x| > |z|)$.
- If the alternative hypothesis has the form $\mu > \mu_0$, we conduct an **upper one-tailed test**. This checks if the sample mean is significantly greater than the assumed population mean. The p-value is calculated as $P(x > z)$.
- If the alternative hypothesis has the form $\mu < \mu_0$, we conduct a **lower one-tailed test**. This checks if the sample mean is significantly less than the assumed population mean. The p-value is calculated as $P(x < z)$.

This distinction significantly affects the computation of the p-value, as it determines which tail(s) of the distribution contribute to the probability calculation.



Hypothesis Testing in General

1 Define Hypotheses

Null (H_0) and alternative (H_a) hypotheses

Defines whether we have a one- or two-tailed test

2 Choose Significance Level

Typically α = 0.05 (5% error tolerance)

Probability of wrongly rejecting H_0

3 Calculate Test Statistic

Measure deviation from null hypothesis

The distribution may depend on the test (not aways t-Student)

4 Compute P-value

Uses CDF of the distribution

Probability of observing a deviation more extreme than the one we are observing if H_0 true

5 Make Decision

Reject H_0 if p-value < α

If failed, we cannot just accept H_0 - need more data

Common Statistical Tests

In practice, there are different tests. We won't see the mathematical details, but these are the main ones.

One-Sample t-Test

Tests whether a sample mean differs significantly from a hypothesised population mean.

Two-Sample t-Test

Compares means between two independent groups to assess significant differences.

Chi-Square Test

Examines relationships between categorical variables or tests goodness-of-fit.

Correlation Tests

Determines whether observed correlations are statistically significant or due to chance.

Assessing Normality

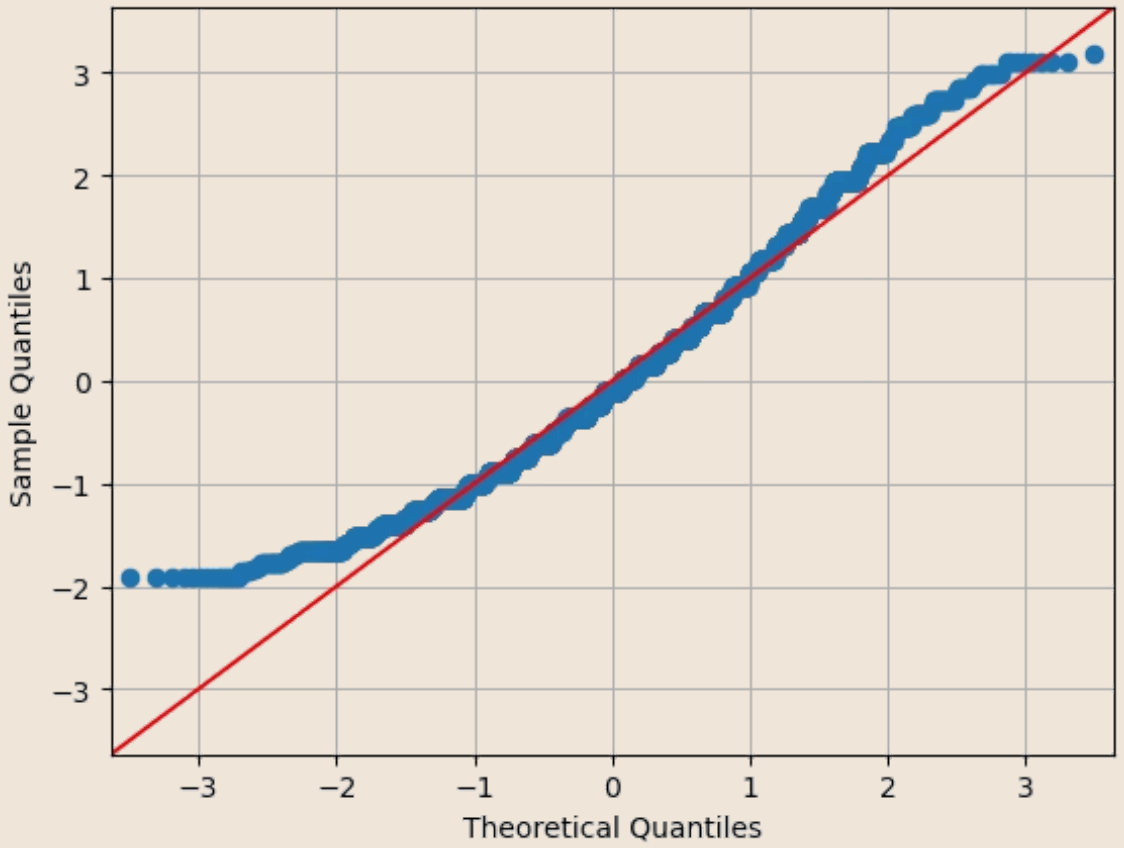
Many statistical methods assume normal distributions. Before applying these methods, we should verify this assumption using diagnostic tools.

Q-Q Plots

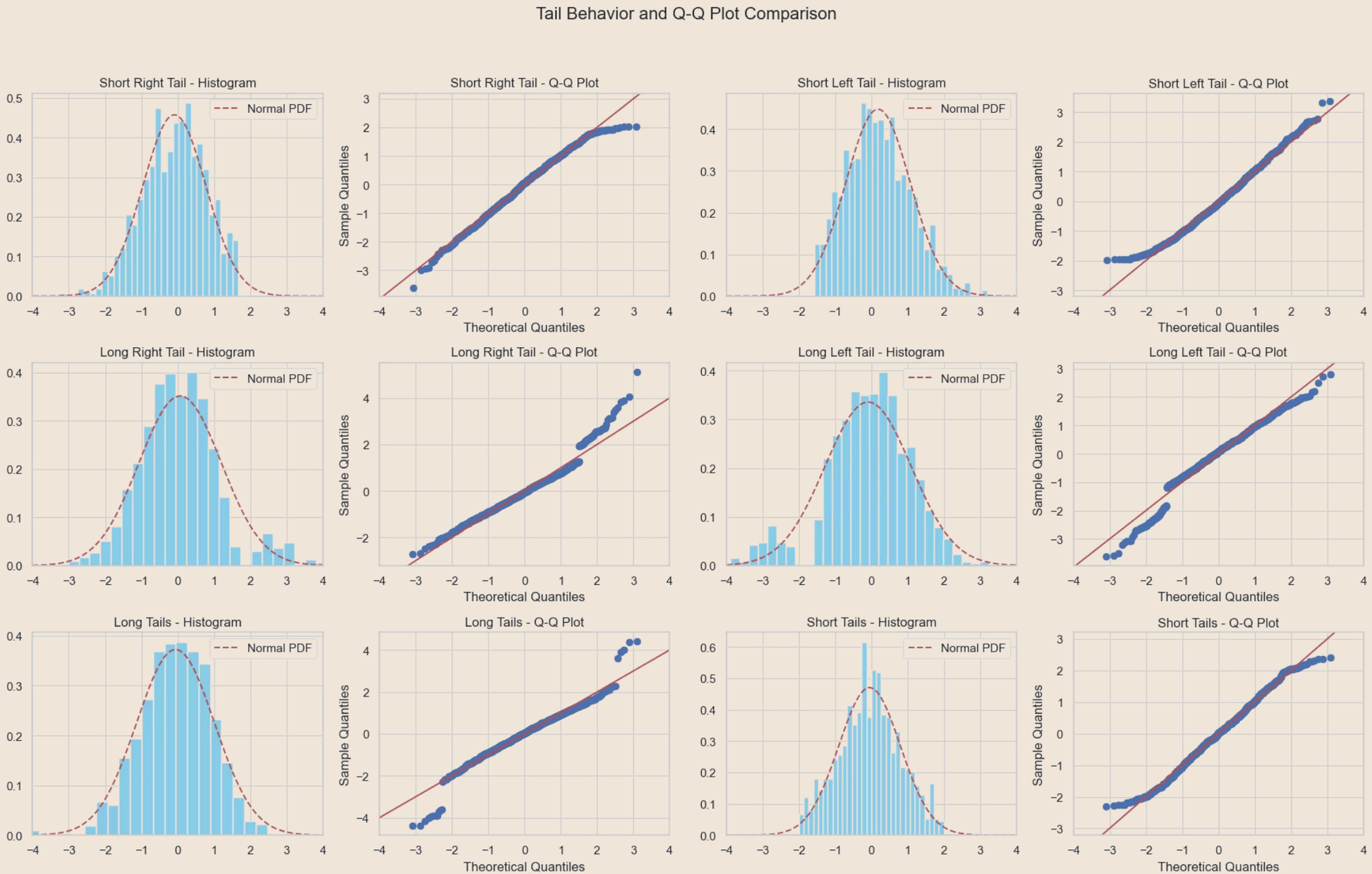
Compare sample quantiles against theoretical normal quantiles. Points falling along the diagonal line indicate normality.

Statistical Tests

- **Shapiro-Wilk:** For smaller samples ($n \leq 2{,}000$)
- **D'Agostino's K^2:** For larger samples ($n \geq 50$)



Q-Q Plots Examples



Conclusions and Next Steps

We Have Explored:

- **Sampling**
- **Sampling distribution of the mean**
- **Confidence intervals**
- **Bias and variance**
- **Hypothesis testing**
- **Main statistical tests**
- **Q-Q Plosts**

In next lectures, we will look at storytelling with data

References

- Chapters 6-8 of [1];
- Parts of chapter 9 of [2].

[1] Gonick, L., & Smith, W. (1993). The cartoon guide to statistics. HarperCollins Publishers, Inc.

[2] Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.