Lecture 4 - **Statistical tests and Regression**

BIOENG-210 Course Notes
Prof. Gioele La Manno

March 2024

Contents

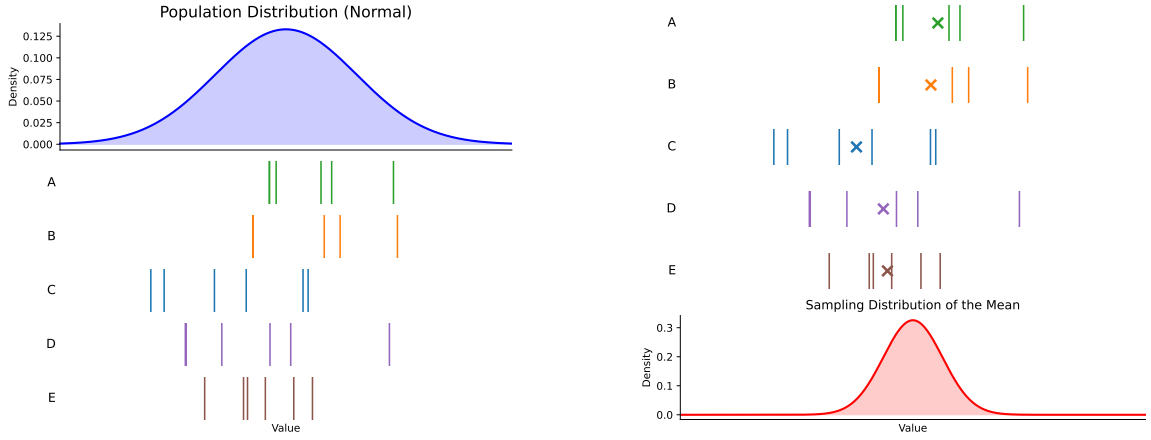
1 Introduction to Statistical Testing

In the previous lectures, we explored how to describe and model data through probability distributions. We learned methods to estimate parameters from samples and to quantify relationships between variables. However, a central challenge in scientific research is moving beyond description to draw substantive conclusions: Do two treatments produce different outcomes? Does a genetic variant affect disease risk? Is a cell type's gene expression pattern altered in a pathological state?

These questions require us to make inferences—to extend our reasoning beyond the immediate data at hand to broader scientific truths. This process lies at the heart of the scientific method, where we aim to test hypotheses and evaluate evidence systematically.

1.1 The Challenge of Drawing Conclusions from Data

When we collect biological data, we are typically observing only a small sample from a much larger population. For instance, we might analyze gene expression in a few dozen tumor samples, but we're interested in making claims about the biology of that cancer type in general. This creates a fundamental challenge: How do we use limited, noisy observations to draw reliable conclusions about broader phenomena? Inference is the branch of statistics that addresses this question.



(a) Samples drawn from a population

(b) Distribution of sample means of the samples

Figure 1: **The challenge of inference: From samples to populations**

The right panel of this plot displays the distribution of the sample mean. Since the mean is a function of random variables, it is itself a random variable and thus has its own distribution. The concept of sample distributions will be discussed later.The figures illustrate a key challenge in biological data analysis: different samples from the same population show natural variability. Even if there is no real biological effect, random chance alone will produce differences between samples. How do we know when an observed difference represents a genuine biological effect rather than random sampling variation? More precisely one could ask: How large must an observed difference be before we can confidently claim it represents a genuine biological effect rather than mere sampling variation?

1.2 Distinguishing Signal from Noise: A Practical Example

Before diving into formal statistical theory, let us consider a concrete biological example that illustrates our fundamental challenge:

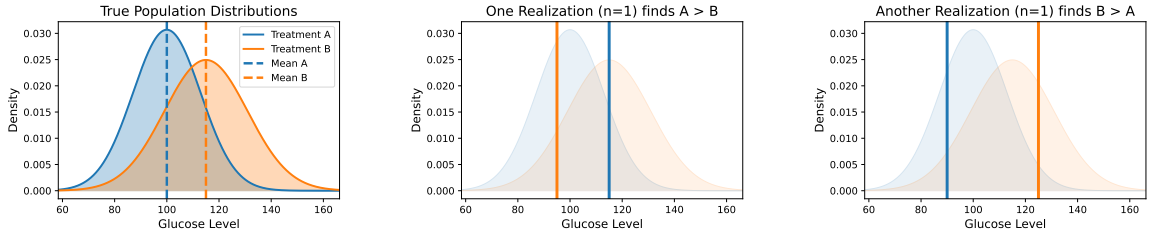
Imagine we are comparing two drug treatments for their effect on blood glucose levels in diabetic mice. For simplicity, let's call them Treatment A and Treatment B. We want to determine which treatment is more effective at lowering glucose levels.

If we measure just one mouse from each treatment group, we might observe that the mouse receiving Treatment A has lower glucose levels than the mouse receiving Treatment B. But should we conclude that Treatment A is generally superior?

This would be a dangerously hasty conclusion! There is chance that we have observed a particularly high or low value by chance.

Each individual mouse's response is affected by numerous factors beyond the treatment itself: genetic variation, initial health status, age, stress levels, and simple biological variability. What we have observed in our single samples might not represent the typical response to each treatment. More conceptually, "glucose levels of Treatment A" and "glucose levels of Treatment B" are random variables of which single measurements are just realizations. Let's recall this definition:

Definition 1.1 (Realization of a Random Variable)**.** A realization of a random variable is a specific observed value or outcome that the random variable takes in a single observation or experiment. While the random variable itself represents the entire set of possible outcomes along with their probabilities, a realization is just one concrete value from that set that occurs when we make a measurement or observation.



(a) Hidden True distributions of glucose levels

(b) Single sample from Experimenter 1

(c) Another single sample from Experimenter 2

Figure 2: **Realizations of random variables: Different conclusions based on single sample**

One could then try to be more mindful and have better plan to how many measurement make to lead to an answer. This is what we are going to call **experiment design**.

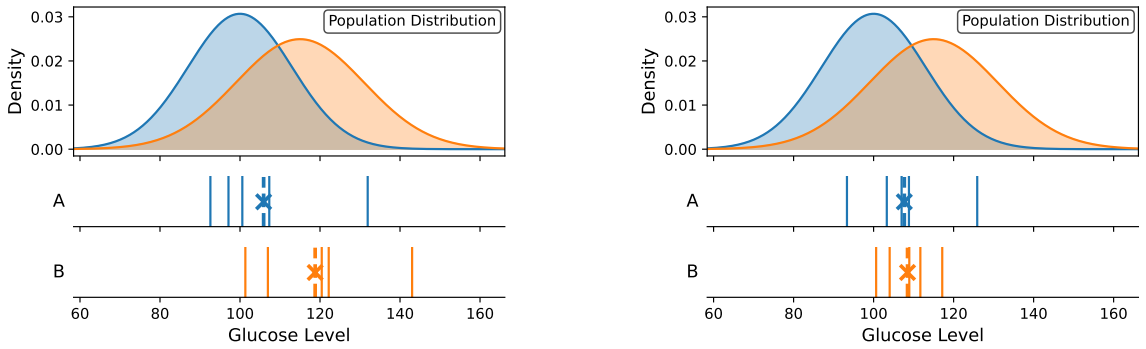
Definition 1.2 (Experiment Design)**.** Experiment design is the process of planning how to collect data in order to answer a research question or test a hypothesis effectively and efficiently. It involves determining the number of samples to collect, the type of measurements to make, and the overall structure of the study.

A basic idea of design is to collect more samples. Say we plan to measure glucose levels in 5 mice per treatment group, with the plan to compute the sample means and compare them. For example, if we collect measurements a_1, a_2, a_3, a_4, a_5 from mice receiving Treatment A and b_1, b_2, b_3, b_4, b_5 from mice receiving Treatment B, we would calculate:

$$\hat{\mu}_A = \frac{a_1 + a_2 + a_3 + a_4 + a_5}{5} \quad \text{and} \quad \hat{\mu}_B = \frac{b_1 + b_2 + b_3 + b_4 + b_5}{5}$$

The sample means provide better estimates, but random chance might still lead us astray. Why? Because $\hat{\mu}_A$ and $\hat{\mu}_B$, being sum of random variables, are themselves random variables, subject to sampling variability.

The fact that a sample mean constitutes a random variable may seem intuitive, but it forms a crucial foundation for statistical inference. When analyzing the distribution of such an estimator,



(a) Five sample from Treatment A and Treatment B

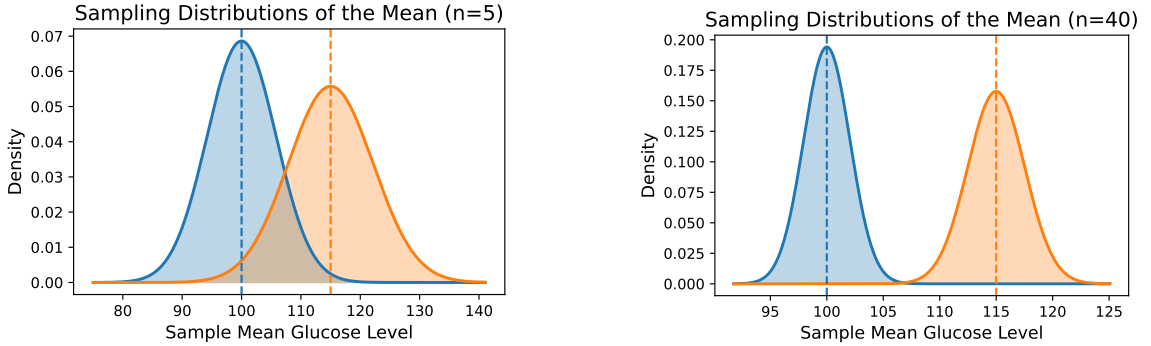
(b) Another five samples

Figure 3: **Realizations of random variables: Single samples from Treatment A and Treatment B**

we use the term "sampling distribution" to distinguish it from the fixed population parameter it estimates. This terminology emphasizes that the variability arises from the sampling process itself, not from uncertainty about the true parameter value.

Definition 1.3 (Sampling Distribution)**.** A sampling distribution is the probability distribution of a given statistic (such as a sample mean) based on a random sample. It describes how the statistic varies from sample to sample when repeatedly drawing samples of the same size from the same population.

Lastly, continuing on the example, we can realize the statistic of interest is the difference between the two means, $\hat{\Delta} = \hat{\mu}_B - \hat{\mu}_A$. This is also a random variable, since it is the difference of two random variables.



(a) Distribution of treatments A and B

(b) Distribution of treatments A and B with 25 samples

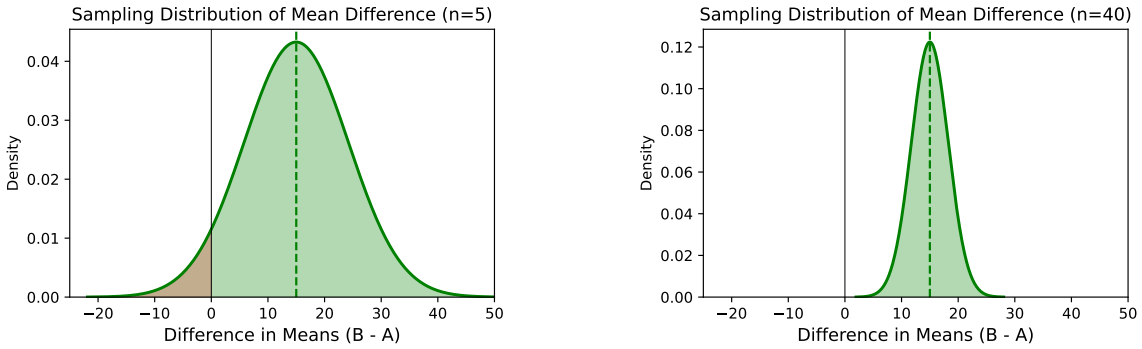
Figure 4: **Relation between mean sampling distributions and sample size**

The last intuition is that if we had a complete description of this random variable (i.e. its distribution and parameters would be identified), we could go pretty far. We could answer the question of how likely is that the difference between the two treatments is due to chance. How?

We have seen this when studying pdf and cdf. We could calculate:

$$P_{\hat{\theta}}(\Delta \leq 0)$$

This is just an integral of the distribution of Δ from 0 to ∞, where $\hat{\theta}$ is the estimated parameter of the distribution of Δ. In practice one uses the cumulative distribution function (CDF) of the random variable Δ to calculate this probability:



(a) Distribution of Δ with 5 samples

(b) Distribution of Δ with 25 samples

Figure 5: **Relation between mean sampling distributions and sample size**

$$P_{\hat{\theta}}(\Delta \leq 0) = CDF_{\Delta}(0|\theta)$$

This would give us the probability that the difference is just by chance. If this probability is very low, we could conclude that the difference is not due to chance and that the treatments are indeed different.

This example illuminates three fundamental insights:

- Individual observations can be misleading due to variability and measurement noise
- Increasing sample size enhances our ability to distinguish true effects from random variation
- The difference between group means is itself a random variable with a "sampling distribution" that depends on sample size.

This intuition points us toward a formal framework for drawing conclusions, but several questions remain unaddressed:

- What distributions best model our sample statistics in different experimental contexts?
- How does sample size quantitatively affect the precision of our estimates?
- How do we convert probabilities into binary decisions (effect exists/doesn't exist)?

1.3 The Philosophy of Statistical Testing

Statistical hypothesis testing provides a formal framework for distinguishing between genuine effects and random variation — a structured approach that forms the backbone of scientific inference. It allows us to move beyond subjective judgments on measurements and make rigorous, quantifiable statements about the evidence for biological effects.

We will begin with Fisher's approach to significance testing, which provides clear practical tools for evaluating evidence against a null hypothesis. At its core lies a philosophical stance about scientific knowledge: we cannot definitively "prove" our scientific hypotheses; we can only gather evidence that supports or contradicts them.

R.A. Fisher, one of the founders of modern statistics, developed significance testing as a framework for evaluating evidence against a specific hypothesis. In this approach, we focus on assessing the compatibility between our observed data and a reference hypothesis, rather than making definitive claims about truth.

Definition 1.4 (Null Hypothesis)**.** The null hypothesis (H_0) is a specific statement about a population parameter that represents the absence of the effect or phenomenon being studied. It serves as a reference point against which we evaluate evidence.

For example, if we are comparing the expression of a gene X between healthy and diseased tissues, our null hypothesis might be: "There is no difference in the mean expression level of gene X between healthy and diseased samples."

In Fisher's original framework, we do not explicitly formulate an alternative hypothesis. Instead, we focus on evaluating how compatible our observed data is with the null hypothesis, which represents a state of "no effect" or "no difference." By carefully evaluating evidence against the null hypothesis, we maintain scientific rigor while still allowing for discovery when the evidence warrants it.

1.4 Developing a Logical Approach to Testing

1.4.1 From Questions to Evidence

Our central challenge is to determine whether an observed difference is likely to be real or could be attributed to chance alone. In Fisher's framework, we ask: "How compatible is our observed data with the null hypothesis?"

To answer this question, we:

1. Define a precise null hypothesis (the "no effect" scenario)
2. Calculate a measure of how far our observed results deviate from what we would expect under this null hypothesis
3. Determine the probability of observing such a deviation (or more extreme) if the null hypothesis were true

1.5 From Test Statistics to Probability Calculations

For a concrete example, imagine we have measured gene expression in 30 control and 30 treatment samples, finding that the treatment group's mean is 2.5 units higher. To determine if this difference is meaningful, we need to calculate the probability of observing a difference this large or larger by random chance alone.

But this presents a computational challenge. How do we calculate the probability of all possible ways to observe a difference of 2.5 or greater? The space of possibilities is vast and complex.

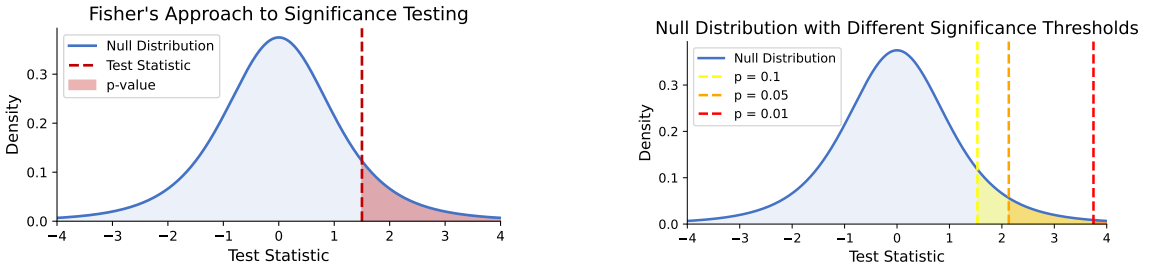
Fisher's insight was to develop a systematic approach:

1. Summarize the observed effect with a single number (a "test statistic")
2. Derive the probability distribution of this statistic under the null hypothesis
3. Compute the probability of observing a value as extreme or more extreme than the one we observed.

So we are ready to give this definition more formally.

Definition 1.5 (P-value)**.** The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the one actually observed, assuming that the null hypothesis is true.

It is important to note that this concept is frequently misunderstood and misinterpreted, even by experienced scientists. It is not the probability that the null hypothesis is true! Nor the probability that the observed effect is due to chance!



(a) Fisher's approach to significance testing

(b) Different significance thresholds and evidence levels

Figure 6: **Fisher's approach to significance testing and evidence interpretation**

1.5.1 The Role of Test Statistics and the P-value

In the early days of statistics, calculating probabilities for complex distributions or defining a sampling distribution was challenging. Statisticians needed a clever approach to make these calculations feasible. They developed test statistics that summarize observed effects into a standardized metric, accounting for both sample size and variability, while following analytically tractable probability distributions under the null hypothesis.

Historically, statisticians sought test statistics that were standardized, allowed for consistent interpretation across different studies. They studied their probability distributions analytically and tabulated critical values for common significance thresholds.

For our gene expression example, rather than working directly with the raw difference in means, we would work with the t-statistic:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

where $\bar{X}_A$ and $\bar{X}_B$ are the sample means, s_A^2 and s_B^2 are the standard deviations, and n_A and n_B are the sample sizes.

Under the null hypothesis (no difference between populations), this standardized difference follows a t-distribution, we are going to see this later. Researchers could look up critical values in statistical tables, eliminating the need for complex probability calculations. While modern computing has made these tables largely obsolete, the practice of using standardized test statistics remains deeply embedded in statistical methodology.

Once we have our test statistic and its null distribution, we can calculate the p-value, which is the cornerstone of Fisher's approach to evidence evaluation.

Importantly, Fisher emphasized that these thresholds are conventional rather than rigid decision boundaries, and the exact p-value should be reported to allow readers to form their own judgments about the strength of evidence.

1.6 The Complete Framework of Statistical Testing

Having built the components step by step, we can now summarize the Fisher's framework of significance testing:

1. Define the effect of interest and formulate a null hypothesis
2. Collect data and calculate an appropriate test statistic
3. Determine the distribution of this statistic under the null hypothesis
4. Calculate the p-value as the probability of observing a test statistic as extreme or more extreme, assuming the null hypothesis is true

5. Interpret the p-value as a measure of evidence against the null hypothesis

This framework is remarkably general—it applies to a wide range of testing scenarios, from simple comparisons to complex experimental designs. The specific details (which test statistic, which distribution) may vary, but the logical structure remains the same.

2 The Foundation: Sample Mean, Standard Error, and the Central Limit Theorem

2.1 The Foundation: Sample Mean and its Distribution

At the core of statistical testing lies our ability to estimate population parameters from sample data. The most fundamental estimator is the sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where $X_1, X_2, \ldots, X_n$ are independent observations from a population.

2.1.1 Sampling Distribution of the Mean

We already mentioned that when we calculate a sample mean, we are obtaining one realization of a random variable. The distribution of these potential sample means is called the sampling distribution.

Definition 2.1 (Sampling Distribution of the Mean)**.** The sampling distribution of the mean is the probability distribution of the sample mean $\bar{X}$ across all possible samples of size n from a population.

This distribution has several important properties:

Property 2.2 (Expected Value of Sample Mean)**.** The expected value of the sample mean equals the population mean:

$$E[\bar{X}] = \mu$$

This property makes the sample mean an unbiased estimator of the population mean.

Property 2.3 (Variance of Sample Mean)**.** The variance of the sample mean can be derived as follows:

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right)$$

Since the observations $X_1, X_2, \ldots, X_n$ are independent, the variance of their sum equals the sum of their variances:

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

This reveals a crucial insight: the variability of the sample mean decreases as sample size increases, making larger samples more reliable.

2.1.2 Standard Error of the Mean

The standard deviation of a sampling distribution is called the standard error. For the sample mean:

Definition 2.4 (Standard Error of the Mean)**.** The standard error of the mean (SEM) quantifies the precision of the sample mean as an estimate of the population mean:

$$\text{SEM} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size.

In practice, we rarely know the population standard deviation σ. Instead, we estimate it using the sample standard deviation s:

$$\hat{\text{SEM}} = \frac{s}{\sqrt{n}}$$

The standard error is fundamental to hypothesis testing because it allows us to quantify how much the sample mean is expected to vary from the true population mean by chance alone. This forms the basis for determining whether an observed difference is statistically significant.

2.1.3 Sampling Distribution of the Sample Standard Deviation

Definition 2.5 (Sample Standard Deviation)**.** The sample standard deviation s is an estimator of the population standard deviation σ, calculated as:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where $X_1, X_2, \ldots, X_n$ are the observed values and $\bar{X}$ is the sample mean.

Definition 2.6 (Chi-Square Distribution)**.** The chi-square distribution with k degrees of freedom, denoted χ^2_k, is the distribution of the sum of squares of k independent standard normal random variables. If $Z_1, Z_2, \ldots, Z_k$ are independent standard normal random variables, then the random variable $X = \sum_{i=1}^{k} Z_i^2$ follows a chi-square distribution with k degrees of freedom, written as:

$$X \sim \chi^2_k$$

and its probability density function is:

$$f_{\chi^2_k}(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \text{ for } x > 0$$

where Γ is the gamma function.

Why are we interested in such a specific, seemingly odd, and complex distribution? Summing independent squared values forms the foundation for constructing variances, standard errors, and similar measures. As a result, we will encounter this distribution frequently. Unlike the sample mean, the sampling distribution of the sample standard deviation is more complex. For samples from a normal distribution:

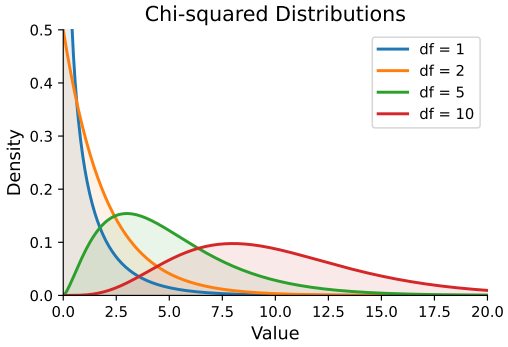
Property 2.7 (Sampling Distribution of Sample Variance)**.** If samples are drawn from a normal distribution with variance σ^2, then the quantity $\frac{(n-1)s^2}{\sigma^2}$ follows a chi-square distribution with $n-1$ degrees of freedom.

Let's try to demonstrate it. If samples $X_1, X_2, \ldots, X_n$ are drawn from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, we can show that $\frac{(n-1)s^2}{\sigma^2}$ follows a chi-square distribution with $(n-1)$ degrees of freedom.

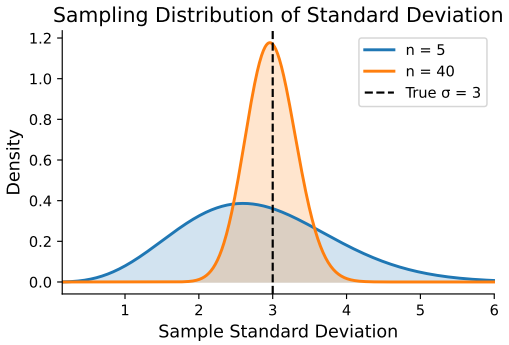
Let $Z_i = \frac{X_i - \mu}{\sigma}$, so each $Z_i \sim \mathcal{N}(0,1)$. Then:

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2 \quad (1)$$

Since $Z_i \sim \mathcal{N}(0,1)$, we know that $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$ by the definition of the chi-square distribution. Also, $\sqrt{n}\bar{Z} \sim \mathcal{N}(0,1)$, so $n\bar{Z}^2 \sim \chi_1^2$. Furthermore, the sum $\sum_{i=1}^{n} Z_i^2$ and the mean $\bar{Z}$ are independent. Therefore, $\sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2$ is the difference of two independent chi-square variables, with n and 1 degrees of freedom respectively, resulting in a chi-square distribution with $n-1$ degrees of freedom. Hence, $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$.



(a) Chi-squared distributions with different degrees of freedom



(b) Sampling distribution of standard deviation for n=5 and n=40 samples

Figure 7: **Chi-squared distributions and sampling distribution of standard deviation**

2.2 The Central Limit Theorem

One of the most important properties of the sample mean is that, regardless of the shape of the population distribution, the sampling distribution of the mean approaches a normal distribution as the sample size increases.

Theorem 2.8 (Central Limit Theorem)**.** For a sufficiently large sample size, the sampling distribution of the mean approaches a normal distribution regardless of the shape of the population distribution, as long as the population has a finite variance.

Specifically, as n increases:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

where $\xrightarrow{d}$ indicates convergence in distribution.

This remarkable result implies that even if our original data comes from a non-normal distribution (even a discrete or strictly positive one, as is often the case with biological measurements!) the sampling distribution of the mean will still approximate a normal distribution, provided the sample size is sufficiently large. The Central Limit Theorem justifies our use of normal-based inference in many situations, even when the underlying data might not be normally distributed. However, for small samples or when the population is highly skewed, we need to be more cautious about applying these methods.

3 Practical Testing Scenarios

With this framework established, we can now apply it to common testing scenarios in biological research. Rather than memorizing different tests for different situations, we will focus on understanding how the general framework adapts to specific contexts. For the following results to follow smoothly, we will review some basic concepts of probability and statistics.

3.1 Starting Simple: One-Sample Tests

The simplest statistical testing scenario occurs when we have a single sample from a distribution and want to make inferences about the population parameter. For example, imagine we've measured the expression levels of a particular gene in 30 tumor samples, and, informally speaking, we want to know if the mean expression level differs from a reference value observed in healthy tissue, assumed known (with certainty).

3.2 One-Sample Tests: If We Knew the Population Variance

Before diving into the realistic scenario where we must estimate the population variance, it's instructive to consider the simpler case where the population variance σ^2 is known.

While this is rarely the case in practice, understanding this scenario provides valuable insight into how statistical tests work and establishes a foundation for understanding the more practical case with estimated variance.

3.2.1 The Z-Test: Testing with Known Variance

To address this question, we need a test statistic that quantifies how far our sample mean $\bar{x}$ is from the reference value μ_0 that we want to compare our sample to. A natural choice would be the simple difference: $\bar{x} - \mu_0$. However, this raw difference doesn't account for:

- Sample size: The same raw difference is more meaningful with larger samples
- Variability: The same raw difference is more meaningful when data has less variability

Since the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a sum of random variables, we can apply the Central Limit Theorem. When the population variance σ^2 is known, the standardized sample mean follows a standard normal distribution:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Under the null hypothesis $H_0 : \mu = \mu_0$, our test statistic becomes:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

This z-statistic directly quantifies how many standard errors our observed mean deviates from the hypothesized value, with the sampling distribution being standard normal.

3.2.2 Calculating the P-value

We can consider a directional test (one-sided) or a non-directional test (two-sided) based on our research question. For a one-sided test, the p-value is the probability of observing a value as extreme as the one we observed in the direction specified by the alternative hypothesis. For a two-sided test, the p-value is the probability of observing a value as extreme in either direction (bigger or smaller than the reference value).

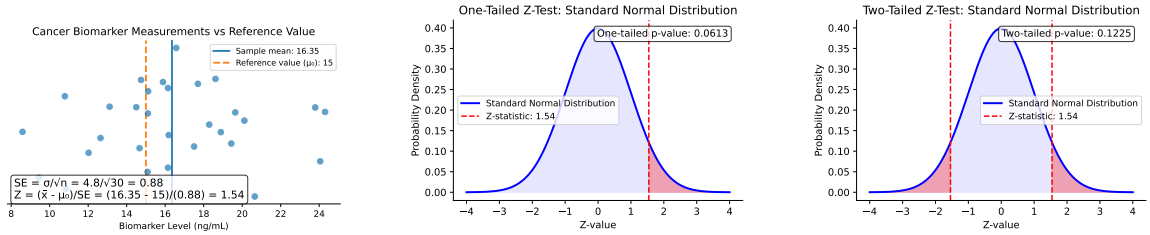
For one-sided tests:

- Lower-tailed test ($H_0 : \mu \geq \mu_0$ vs $H_a : \mu < \mu_0$): $p = \Phi(z)$
- Upper-tailed test ($H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$): $p = 1 - \Phi(z)$

where Φ is the cumulative distribution function of the standard normal distribution.

For a two-sided test (where we're testing whether the population mean differs from μ_0 in either direction), the p-value is:

$$p = 2 \times \min\{P(Z \leq z), P(Z \geq z)\} = 2 \times \min\{\Phi(z), 1 - \Phi(z)\}$$



(a) Sample mean vs reference value (b) One-tailed Z-test (c) Two-tailed Z-test

Figure 8: **One-sample Z-test visualization**

The first plot shows the data we want to test: a jittered set of points representing individual observations, with their sample mean clearly indicated. A horizontal reference line represents the reference value μ_0 that we're testing against. The second plot illustrates how the test statistic is calculated and how the p-value is determined from the standard normal distribution.

3.2.3 From Theory to Practice - Additional Uncertainty

In reality, we rarely know the true population variance σ^2 and must estimate it from our sample. This additional uncertainty means we need to modify our approach, leading to Student's t-test, which we'll explore next. Let's consider this simple substitution to the z-statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

It looks like almost the same as the z-statistic, but we have substituted a constant σ with a random variable s. s is a random variable because it is calculated from the data, and depends on the specific sample we have drawn.

The intuition of how the summary statistics is affected by this change is that we risks to make values more extreme than they should be sometimes. So the random variable t cannot follow a normal distribution, and we expect an overdispersed version of it.

We will derive which distribution exactly is needed in the next section. This is not a trivial derivation or one to memorize, however it is important to understand how these conclusions are reached in statistics.

3.3 A simplified derivation of the t-Distribution

To understand where the t-distribution comes from, we'll start with our test statistic and see how it relates to distributions we already know. We have:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Let's multiply and divide the numerator by σ:

$$t = \frac{\sigma(\bar{x} - \mu_0)/\sigma}{s/\sqrt{n}} = \frac{\sigma/\sqrt{n} \cdot (\bar{x} - \mu_0)/(\sigma/\sqrt{n})}{s/\sqrt{n}} = \frac{Z}{\frac{s}{\sigma}}$$

where $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ under the null hypothesis.

We know from earlier that $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$. Let's define $V = \frac{(n-1)s^2}{\sigma^2}$, so $\frac{s}{\sigma} = \sqrt{\frac{V}{n-1}}$.

This gives us:

$$t = \frac{Z}{\sqrt{\frac{V}{n-1}}}$$

Now we have t as a function of two random variables, Z and V, with known distributions:

- $Z \sim \mathcal{N}(0,1)$
- $V \sim \chi^2_{n-1}$
- Z and V are independent

To find the distribution of t, we need to apply a change of variables. Let's define a new variable $W = V$, so our transformation is:

$$T = \frac{Z}{\sqrt{W/(n-1)}}, \quad W = V$$

The inverse transformation is:

$$Z = T\sqrt{\frac{W}{n-1}}, \quad V = W$$

The Jacobian matrix for this transformation is:

$$J = \begin{pmatrix} \frac{\partial z}{\partial t} & \frac{\partial z}{\partial w} \\ \frac{\partial v}{\partial t} & \frac{\partial v}{\partial w} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{w}{n-1}} & \frac{t}{2\sqrt{w(n-1)}} \\ 0 & 1 \end{pmatrix}$$

The determinant of this Jacobian is:

$$|J| = \sqrt{\frac{w}{n-1}}$$

By the change of variables rule, the joint PDF of T and W is:

$$f_{T,W}(t,w) = f_{Z,V}(z(t,w), v(t,w)) \cdot |J|$$

Since Z and V are independent:

$$f_{Z,V}(z,v) = f_Z(z) \cdot f_V(v) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \cdot \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} v^{(n-1)/2-1} e^{-v/2}$$

Substituting our transformation:

$$f_{T,W}(t,w) = \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{\frac{w}{n-1}})^2/2} \cdot \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} w^{(n-1)/2-1} e^{-w/2} \cdot \sqrt{\frac{w}{n-1}} \tag{2}$$

$$= \frac{1}{\sqrt{2\pi(n-1)}} \cdot \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} \cdot w^{n/2-1} \cdot e^{-\frac{w}{2}(1+\frac{t^2}{n-1})} \tag{3}$$

To find the marginal distribution of T, we integrate over all possible values of W:

$$f_T(t) = \int_0^\infty f_{T,W}(t,w)\,dw$$

$$f_T(t) = \frac{1}{\sqrt{2\pi(n-1)}} \cdot \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} \int_0^\infty w^{n/2-1} \cdot e^{-\frac{w}{2}(1+\frac{t^2}{n-1})}\,dw \tag{4}$$

This integral takes the form of a gamma function:

$$\int_0^\infty x^{\alpha-1}e^{-\beta x}\,dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

With $\alpha = n/2$ and $\beta = \frac{1}{2}(1+\frac{t^2}{n-1})$:

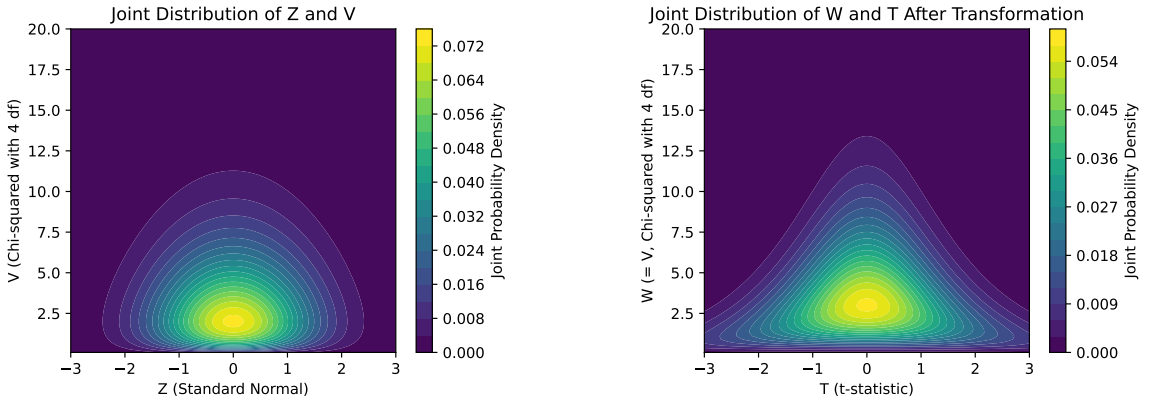
$$f_T(t) = \frac{1}{\sqrt{2\pi(n-1)}} \cdot \frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)} \cdot \frac{\Gamma(n/2)}{(\frac{1}{2}(1+\frac{t^2}{n-1}))^{n/2}} \tag{5}$$

$$= \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)} \cdot \left(1+\frac{t^2}{n-1}\right)^{-n/2} \tag{6}$$

After rearranging and using properties of the gamma function, we obtain the probability density function of Student's t-distribution with $\nu = n-1$ degrees of freedom:

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

The t-distribution approaches the standard normal distribution as the degrees of freedom increase, reflecting how our uncertainty about σ diminishes with larger sample sizes.



(a) Joint distribution of Z (normal) and V (chi-squared)

(b) Joint distribution of W and T after transformation

Figure 9: **Joint distributions in the derivation of the t-distribution**

The pdf of this distribution is given by:

$$f(t) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n-1}{2}}$$

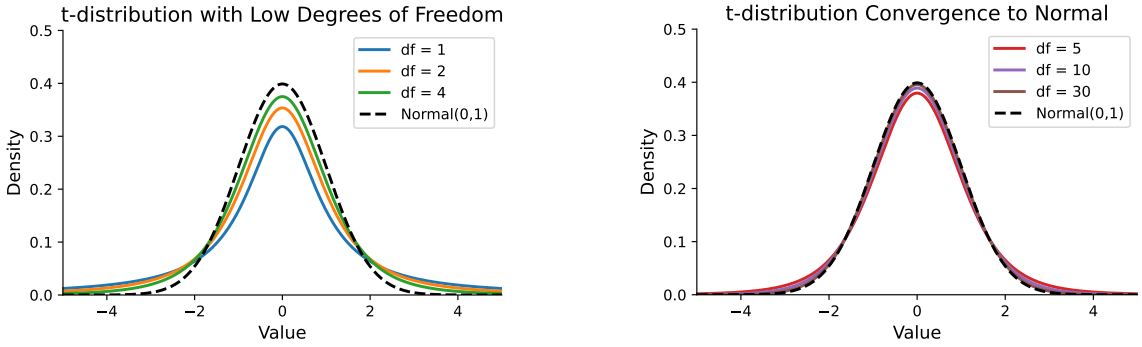
The t-distribution resembles the normal distribution but has heavier tails, reflecting the additional uncertainty from estimating the population variance.

As sample size increases, the t-distribution approaches the standard normal distribution, reflecting diminishing uncertainty in our variance estimate.

Theorem 3.1 (Student's t-distribution)**.** When sampling from a normally distributed population with unknown variance, the standardized sample mean follows a t-distribution with n-1 degrees of freedom:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

where $\bar{x}$ is the sample mean, μ_0 is the population mean, s is the sample standard deviation, and n is the sample size.



(a) t-distribution with varying degrees of freedom

(b) t-distribution convergence to normal distribution

Figure 10: **t-distribution and its convergence to the normal distribution**

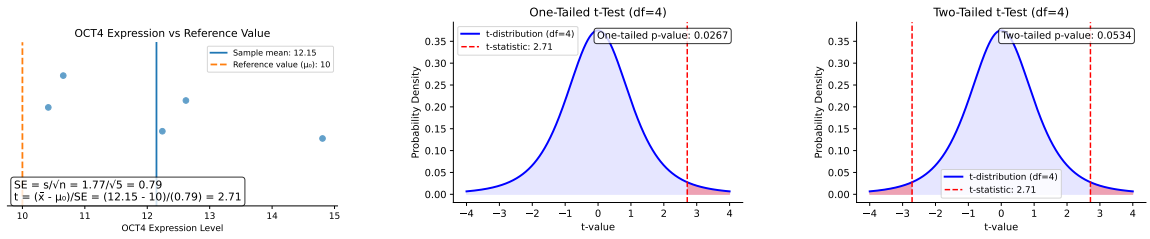
3.3.1 Performing the Test

In this framework, we can now formally test our hypothesis:

- Our parameter of interest is the population mean μ
- Null hypothesis: $\mu = \mu_0$ (reference value)
- Test statistic: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, which follows a t-distribution with $n-1$ degrees of freedom
- P-value: probability of observing a test statistic as extreme or more extreme than the one we calculated

The key insight is that even when the null hypothesis is true, we don't expect our sample mean to exactly equal μ_0 due to random variation. The test statistic quantifies "how many standard errors away" our sample mean is from the hypothesized value, and the t-distribution tells us how likely such deviations are to happen by chance.

Let's consider a biological example to illustrate the t-test in practice:



(a) OCT4 expression data

(b) One-tailed t-test (df=4)

(c) Two-tailed t-test (df=4)

Figure 11: **One-sample t-test visualization**

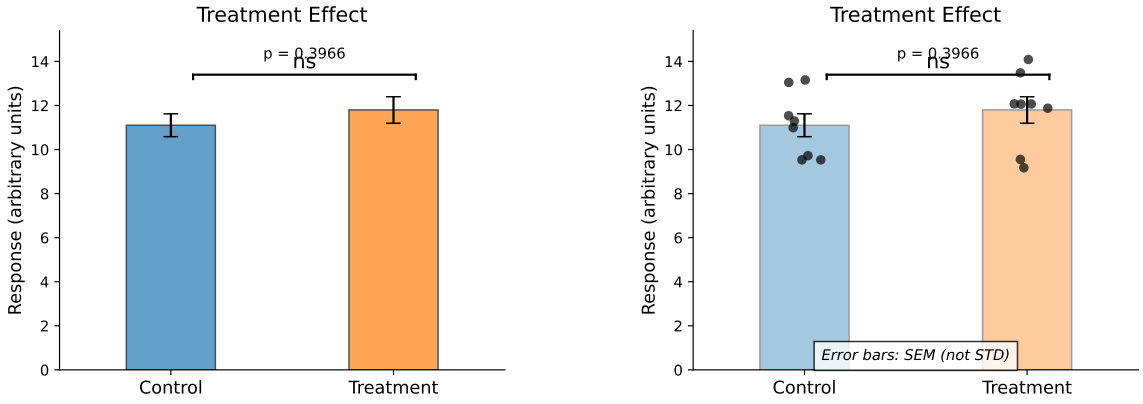
Suppose we have measured the expression of the stemness gene *OCT4* in 5 stem cell cultures, and we want to test whether the mean expression differs from a reference value we have established as normal value for this culture in the past. The above plots show the raw data, the one-tailed t-test and the two-tailed t-test.

3.4 Two-Sample Tests: Comparing Independent Groups

The two-sample t-test is one of the most common statistical tests found in biological research. In biological papers, this test is typically reported above bar plots showing means with standard error of the mean (SEM) error bars, allowing readers to visually assess the difference. These plots often feature a horizontal line connecting the bars with asterisks indicating the significance level (e.g., * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$).

Applying our framework with the Welch test, which doesn't assume equal variances between groups:

- Parameter of interest: difference between population means $\mu_1 - \mu_2$
- Null hypothesis: $\mu_1 - \mu_2 = 0$
- Test statistic: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, which accounts for potentially different variances in each group
- Null distribution: approximated by a t-distribution with degrees of freedom estimated from the data
- P-value: probability of observing a test statistic as extreme or more extreme, assuming no difference between the population means



(a) Common plot find in papers

(b) Recent papers require showing data

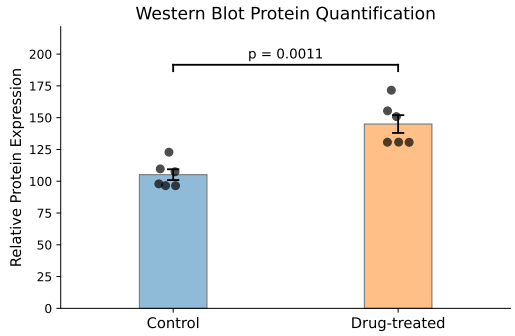
Figure 12: **Two-sample test visualization in biological research**

The Welch test is more general than tests that assume equal variances, making it appropriate for a wide range of biological data where variance homogeneity cannot be assumed.

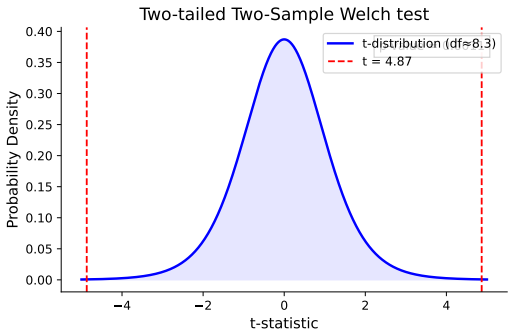
3.4.1 Biological Example: Protein Expression in Western Blot Analysis

Let's consider a concrete biological example. Suppose researchers are investigating the effect of a drug treatment on the expression of a protein involved in cell signaling. They perform western blot analysis on protein extracts from control and treated cell cultures, then quantify the band intensities to measure relative protein levels.

The question is whether the drug treatment significantly alters the expression of this protein. The researchers collect measurements from 6 independent control samples and 6 independent drug-treated samples.



(a) Western blot protein quantification data from control and drug-treated samples



(b) Two-tailed t-test for the protein expression data

Figure 13: **Two-sample t-test applied to western blot protein quantification data**

In this example, we calculate:

- Sample means: $\bar{x}_{control} = 1.0$ (normalized), $\bar{x}_{treated} = 1.75$
- Sample standard deviations: $s_{control} = 0.22$, $s_{treated} = 0.34$
- Sample sizes: $n_{control} = n_{treated} = 6$

Our test statistic is:

$$t = \frac{\bar{x}_{treated} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treated}}{n_{treated}} + \frac{s^2_{control}}{n_{control}}}} = \frac{1.75 - 1.0}{\sqrt{\frac{0.34^2}{6} + \frac{0.22^2}{6}}} \approx 4.78$$

Under the null hypothesis, this follows approximately a t-distribution with degrees of freedom calculated using the Welch-Satterthwaite equation (approximately 9 df in this example). For a two-sided test, the p-value is the probability of observing a t-statistic with absolute value as large or larger than 4.78, which is approximately p = 0.001.

The small p-value suggests strong evidence against the null hypothesis, indicating that the drug treatment significantly increases the expression of the target protein.

The Welch t-test is particularly valuable in biological research where equal variances between groups cannot be assumed.

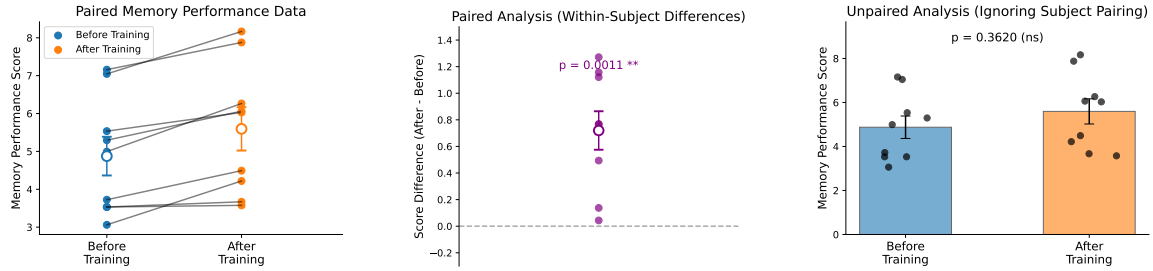
3.5 Working with Paired Samples

In many biological experiments, we have natural pairing between observations: before and after measurements on the same subjects, or on the same biological samples at different time points.

This pairing creates a dependency structure that must be accounted for in our analysis. Taking in consideration the pairing can change significantly our conclusions.

Let's consider a hypothetical example where we're testing the effect of a cognitive training program on memory performance in mice. We measure the memory scores of 9 mice before and after training, and we want to know if the training significantly improves memory. The between-subject variability is high, so it is difficult to identify the improvement from the mean and distribution of the pre and post training but we observe consistent improvements within each mouse.

The insight with paired designs is that we can convert the two-sample problem into a one-sample problem by analyzing the differences within each pair. Instead of comparing two separate sets of measurements, we:



(a) Paired data with before-after measurements connected by lines

(b) Paired analysis showing significant difference

(c) Unpaired analysis showing non-significant difference

Figure 14: **Paired vs. unpaired tests: How accounting for experimental design affects statistical power**

- Calculate the difference for each pair
- Test whether the mean difference is zero using a one-sample test

This approach:

- Accounts for the dependency between paired observations
- Reduces variability by controlling for subject-specific factors
- Typically provides greater statistical power than unpaired comparisons

4 Multiple Testing and False Discovery Control

4.1 From Evidence to Decisions: The Neyman-Pearson Framework

While Fisher's approach focuses on evaluating evidence, Jerzy Neyman and Egon Pearson later developed a complementary framework that emphasizes decision-making with controlled error rates. This extension introduces several additional concepts that are now widely used in statistical testing.

4.1.1 Alternative Hypotheses and Binary Decisions

The Neyman-Pearson framework explicitly introduces the concept of an alternative hypothesis and frames statistical testing as a binary decision problem.

Definition 4.1 (Alternative Hypothesis)**.** The alternative hypothesis (H_a or H_1) is a statement that contradicts the null hypothesis and represents the presence of the effect being studied.

In a gene expression example, the alternative hypothesis would be: "There is a difference in the mean expression level of gene X between healthy and diseased samples." The testing process now involves deciding between two competing hypotheses based on our observed data, rather than simply evaluating evidence against a single hypothesis.

4.1.2 Significance Level and Error Types

The Neyman-Pearson approach introduces the concept of a significance level as a decision threshold, along with a formal categorization of possible errors. Let's define these key concepts.

Definition 4.2 (Significance Level)**.** The significance level (α) is the probability threshold below which we reject the null hypothesis. It represents the maximum rate of false positives we are willing to accept.

When we compare our calculated p-value to this threshold:

- If $p < \alpha$, we reject the null hypothesis
- If $p \geq \alpha$, we fail to reject the null hypothesis

This decision process can lead to two types of errors:

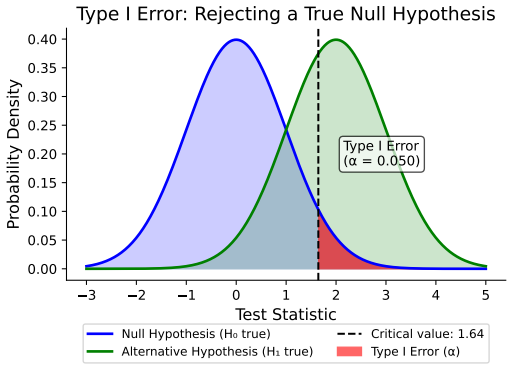
Definition 4.3 (Type I Error)**.** A Type I error occurs when we reject the null hypothesis when it is actually true (a false positive).

Definition 4.4 (Type II Error)**.** A Type II error occurs when we fail to reject the null hypothesis when it is actually false (a false negative).

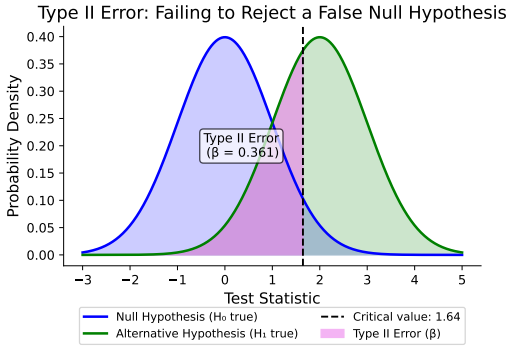
	H_0 **True**	H_0 **False**
Reject H_0	Type I Error	Correct Decision
Fail to Reject H_0	Correct Decision	Type II Error

The significance level α directly controls the Type I error rate: if we set $\alpha = 0.05$, we ensure that if the null hypothesis is true, we will incorrectly reject it no more than 5% of the time.

The probability of a Type II error is denoted by β, and the power of a test (the probability of correctly rejecting a false null hypothesis) is $1 - \beta$. Unlike the Type I error rate, the Type II error rate is not directly controlled in the testing procedure and depends on factors like sample size, effect size, and variability.



(a) Type I error: Rejecting a true null hypothesis (false positive)



(b) Type II error: Failing to reject a false null hypothesis (false negative)

Figure 15: **Type I and Type II errors in hypothesis testing**

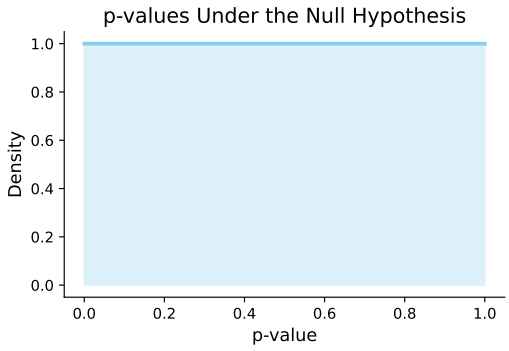
4.2 A Crucial Property: Uniformity Under the Null

For a valid statistical test, the distribution of p-values under the null hypothesis must be uniform between 0 and 1. This property is the foundation of error rate control and ensures that our significance threshold α directly controls the false positive rate.

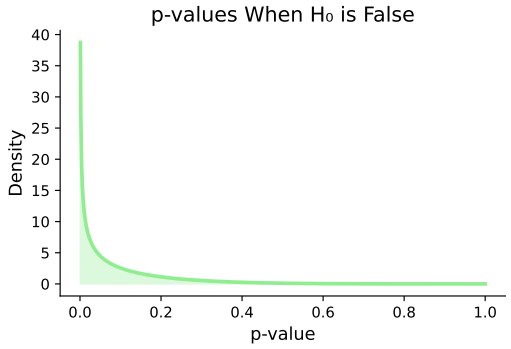
Theorem 4.5 (Uniformity of P-values Under the Null)**.** If the null hypothesis is true, and if the test statistic's distribution is continuous, then the p-value follows a uniform distribution on the interval [0,1].

This uniformity guarantees that when the null hypothesis is true:

- The probability of obtaining a p-value less than or equal to 0.05 is exactly 0.05
- The probability of obtaining a p-value less than or equal to any threshold α is exactly α



(a) Distribution of p-values under the null hypothesis



(b) A possible of p-values when the null hypothesis is false (in general is not known)

Figure 16: **Uniform distribution of p-values under the null hypothesis**

When the null hypothesis is false, the distribution of p-values shifts toward zero, with the extent of this shift depending on the effect size and sample size. This property becomes particularly important when we consider multiple testing scenarios.

4.3 The Multiple Testing Problem in High-Throughput Biology

Modern biological research often involves testing many hypotheses simultaneously, creating unique statistical challenges that extend beyond traditional hypothesis testing:

- Testing thousands of genes for differential expression in RNA sequencing
- Examining millions of genetic variants for disease association in genome-wide studies
- Analyzing hundreds of metabolites in a metabolomics study

This high-throughput nature creates a fundamental challenge: when we perform many tests, the probability of obtaining false positives increases dramatically, even when each individual test maintains a controlled false positive rate.

To understand why, consider testing 1,000 genes for differential expression when none are actually differentially expressed (i.e., all null hypotheses are true). If we use $\alpha = 0.05$ for each test:

- Each test has a 5% chance of producing a false positive
- Expected number of false positives $= 1,000 \times 0.05 = 50$

Without correction, we would expect about 50 "significant" results even when no real effects exist! This is a serious problem that can lead to spurious conclusions and wasted research efforts.

To illustrate this challenge, let's consider a typical RNA sequencing experiment comparing gene expression between two conditions (e.g., disease vs. healthy). For approximately 20,000 human protein-coding genes, we conduct 20,000 separate statistical tests asking essentially the same question for each gene: "Is this gene differentially expressed between conditions?"

4.4 The Problem of P-value Hacking and Multiple Testing

When we test a single hypothesis using a significance threshold of $\alpha = 0.05$, we accept a 5% risk of a Type I error (false positive). However, when we conduct multiple tests, this error rate applies to each individual test, causing the experiment-wide error rate to increase dramatically.

For example, if all 20,000 genes truly had identical expression between conditions (i.e., all null hypotheses are true), we would still expect $20,000 \times 0.05 = 1,000$ genes to show "significant" results by chance alone! This is why researchers performing high-throughput experiments often find "interesting" results even with random data.

This problem necessitates methods that account for the multiplicity of tests, particularly in biological research where false discoveries can lead to wasted resources in follow-up experiments or even misleading clinical applications.

4.5 Family-Wise Error Rate and the Bonferroni Correction

The most straightforward approach to multiple testing is to control the Family-Wise Error Rate (FWER) - the probability of making even one false discovery across all tests.

Definition 4.6 (Family-Wise Error Rate)**.** The Family-Wise Error Rate (FWER) is the probability of making at least one Type I error (false positive) among all the hypothesis tests conducted.

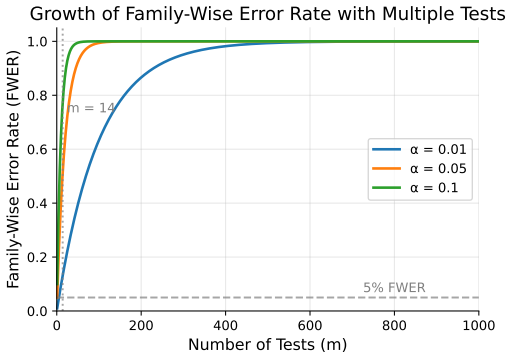
For m independent tests, each with significance level α, the probability of making at least one Type I error is:

$$\text{FWER} = 1 - (1 - \alpha)^m$$

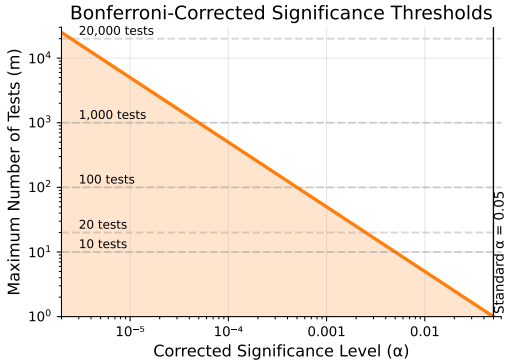
which grows rapidly with the number of tests. For example, with $\alpha = 0.05$ and $m = 100$ tests, FWER ≈ 0.994, meaning we're almost certain to get at least one false positive.

The simplest correction method is the Bonferroni correction, which divides the desired significance level by the number of tests:

$$\alpha_{\text{corrected}} = \frac{\alpha}{m}$$



(a) Growth of FWER with increasing number of tests



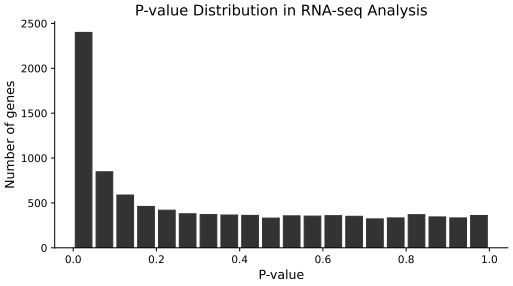
(b) Bonferroni-corrected significance thresholds for different numbers of tests

Figure 17: **Family-wise error rate and Bonferroni correction**

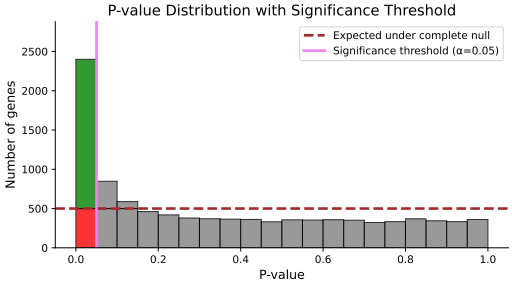
While mathematically sound, the Bonferroni correction becomes extremely conservative for large numbers of tests. For 20,000 genes, the corrected threshold would be $\alpha = 0.05/20,000 = 0.0000025$, which is so stringent that many true discoveries might be missed. This trade-off between Type I and Type II errors becomes increasingly problematic as the number of tests grows.

4.6 A Closer Look at P-value Distributions in Real Data

In actual biological datasets, p-values rarely follow a uniform distribution because some of the null hypotheses are indeed false (e.g., some genes are truly differentially expressed). Instead, we typically observe a mixture distribution:



(a) P-value distribution from RNA-seq data



(b) Its interpreation as mixture and relation to the FDR

Figure 18: **Empirical p-value distribution**

This observed p-value distribution provides valuable information. The relative heights of these two components (uniform vs. enriched near zero) give us insight into the proportion of true vs. false null hypotheses in our dataset. This insight forms the foundation for a more practical approach to multiple testing correction.

4.7 False Discovery Rate: A More Practical Approach

In many biological contexts, we're less concerned with avoiding any false positives and more concerned with maintaining a reasonable proportion of true discoveries among our significant results. This led to the development of False Discovery Rate (FDR) control by Benjamini and Hochberg.

Definition 4.7 (False Discovery Rate)**.** The False Discovery Rate (FDR) is the expected proportion of false positives among all rejected null hypotheses (among all "discoveries").

$$\text{FDR} = E\left[\frac{\text{Number of false positives}}{\text{Total number of rejections}}\right]$$

If we call 100 genes "significant" and expect an FDR of 0.1, then approximately 10 of these genes are likely false positives, while 90 represent true effects. This is often a more useful metric for biological research than the more stringent FWER.

With the FDR approach, we can understand the empirical p-value distribution as a mixture of two components:

- A uniform component representing true null hypotheses
- A component concentrated near zero representing false null hypotheses (true effects)

If we can estimate the proportion of true null hypotheses (π_0), we can better understand the expected proportion of false discoveries when using any particular significance threshold.

4.8 The Two-Groups Model and Local FDR

A deeper understanding of multiple testing comes from the "two-groups model" introduced by Efron. This model explicitly separates p-values into two categories:

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p)$$

where:

- π_0 is the proportion of true null hypotheses (tests where there is genuinely no effect)
- $f_0(p)$ is the density of p-values under the null hypothesis (uniform on [0,1])
- $f_1(p)$ is the density of p-values under the alternative hypothesis (concentrated near 0)
- $f(p)$ is the overall mixture density we observe in our data

This model allows us to directly estimate what proportion of significant results are false positives at any p-value threshold.

Definition 4.8 (Local False Discovery Rate)**.** The local false discovery rate at a specific p-value p is the probability that a test with exactly that p-value comes from the null hypothesis:

$$\text{local FDR}(p) = \frac{\pi_0 f_0(p)}{f(p)}$$

Unlike the traditional FDR, which applies to all tests with p-values below a threshold, the local FDR provides a test-specific measure of reliability. For example, a test with local FDR = 0.1 has approximately a 10% chance of being a false positive.

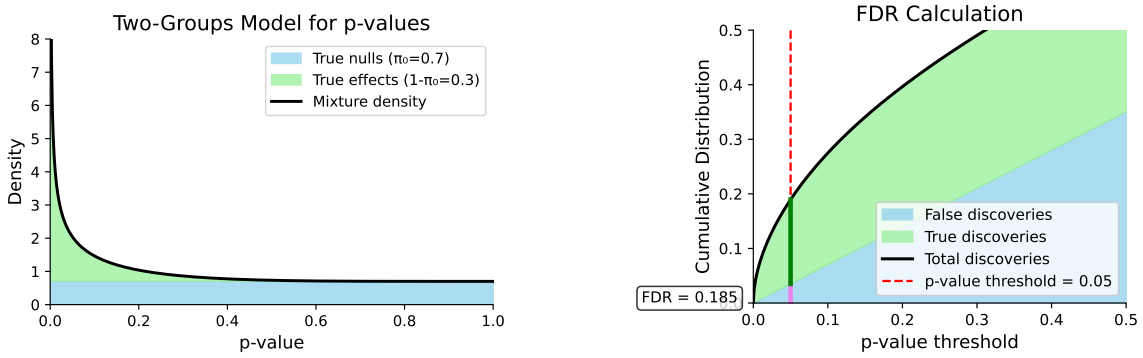
To extend from local FDR to the overall FDR for a significance threshold, we integrate over the p-value distribution:

$$F(p) = \int_0^p f(t)dt$$

The overall FDR for a threshold p is then:

$$\text{FDR}(p) = \frac{\pi_0 p}{F(p)}$$

This represents the expected proportion of false discoveries among all tests with p-values less than or equal to p.



(a) Two-groups model: local fdr

(b) FDR calculation

Figure 19: **Visualizing the Two-Groups Model and FDR Calculation**

4.9 The Benjamini-Hochberg Procedure

The most widely used method for controlling FDR is the Benjamini-Hochberg (BH) procedure:

1. Rank all p-values from smallest to largest: $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$
2. Find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$
3. Reject all null hypotheses with p-values less than or equal to $p_{(k)}$

This procedure guarantees that, under certain assumptions, the expected FDR will be at most α (also called the q-value in this context).

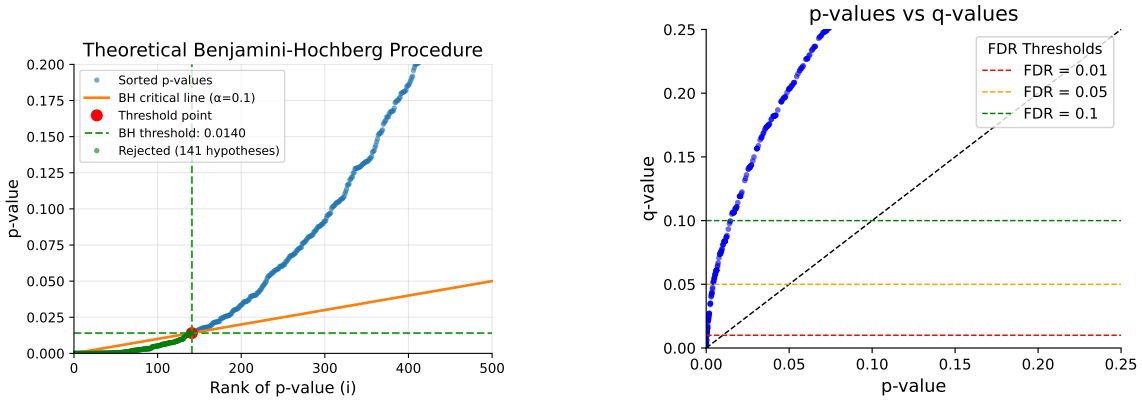
The BH procedure operates by identifying the significance threshold that balances the expected proportion of false positives. It's less conservative than FWER-controlling methods like Bonferroni, allowing for more discoveries while still maintaining a specified proportion of true findings.

4.10 Beyond BH: q-values and Additional FDR Methods

The q-value, introduced by John Storey, is the FDR analog of the p-value. For a given test, the q-value represents the minimum FDR that would be incurred if we called that test significant.

Definition 4.9 (q-value)**.** The q-value for a particular test is the expected proportion of false positives among all tests with equal or smaller p-values.

Unlike p-values, q-values directly account for multiple testing and provide a more interpretable measure in high-throughput settings. A gene with q-value = 0.05 means that 5% of genes considered significant at that threshold are expected to be false positives.



(a) Demonstration of BH procedure with ideal data

(b) q-value concept

Figure 20: **The Benjamini-Hochberg procedure for controlling FDR and q-values**

4.11 Practical Considerations in Biological Applications

When applying multiple testing procedures to biological data, several practical considerations arise:

1. **Independence assumptions:** Many correction methods assume independence between tests, which rarely holds in biology (e.g., correlated gene expression). More conservative corrections may be needed when tests are dependent.

2. **Pre-filtering:** Removing tests that are unlikely to yield significant results (e.g., filtering out low-expression genes before testing) can reduce the multiple testing burden.

3. **Effect size considerations:** Statistical significance does not imply biological significance. Tests with tiny p-values might represent biologically negligible effects in large datasets.

4. **The value of exploratory analysis:** In early-stage research, a higher FDR might be acceptable to generate hypotheses that will be validated in follow-up experiments.

The choice of multiple testing strategy should be guided by the specific research context, the number of tests performed, and the relative costs of Type I and Type II errors in the particular biological system under study.