Statistical Testing and Multiple test correction

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

March 2025

EPFL - BMI - UPLAMANNO

Contents

Drawing Conclusions from Data

In previous lectures, we described and modeled data through probability distributions. Now we face a greater challenge:

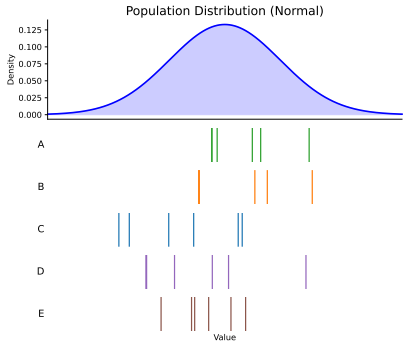
How do we use limited, noisy observations to draw reliable conclusions about broader phenomena?

Key questions in biological research:

- Do two treatments produce different outcomes?
- Does a genetic variant affect disease risk?
- Is a cell type's gene expression pattern altered in disease?

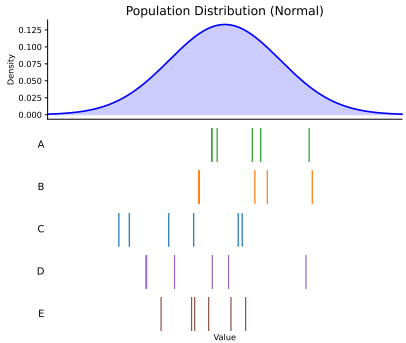
Statistical testing provides a systematic framework to address these questions.

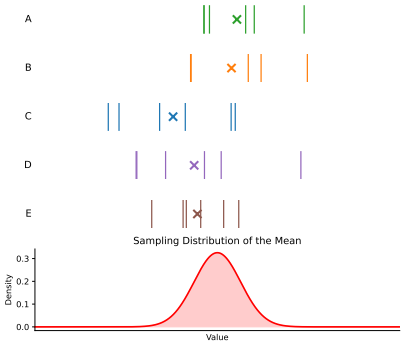
From Samples to Populations: The Key to Statistical Inference



How large must an observed difference be before we can claim it represents a genuine biological effect rather than sampling variation?

From Samples to Populations: The Key to Statistical Inference



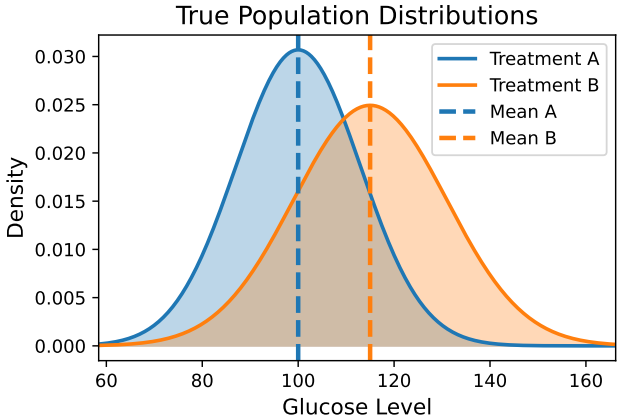


How large must an observed difference be before we can claim it represents a genuine biological effect rather than sampling variation?

Distinguishing Signal from Noise: A Practical Example

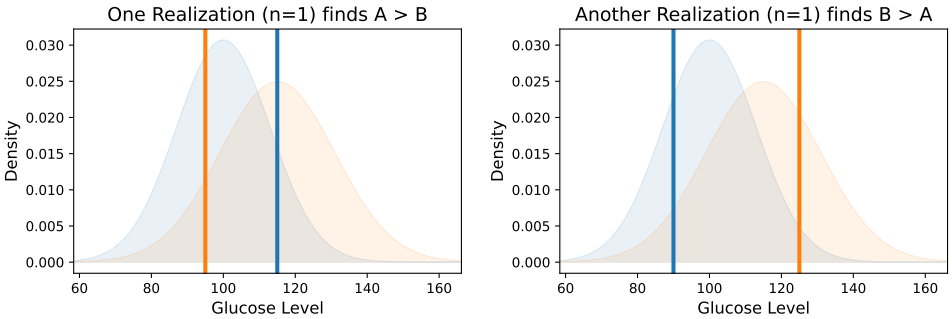
Comparing two drug treatments for their effect on blood glucose levels in diabetic mice.

The true distributions of glucose levels under each treatment remain hidden:



The Problem with Small Samples

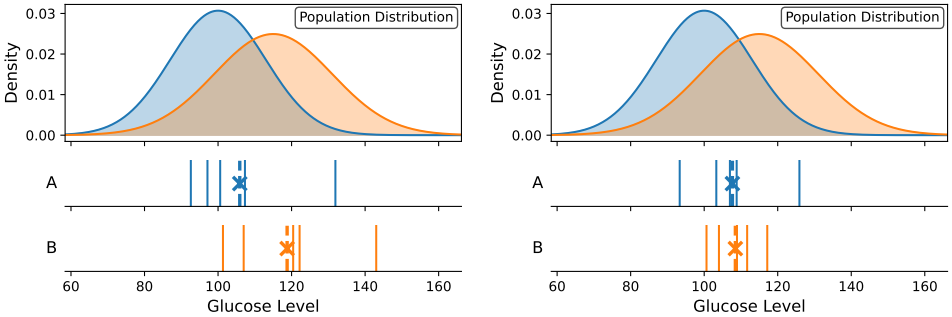
Different single samples can lead to contradictory conclusions:



Two experimenters might reach opposite conclusions based on different single samples from each treatment.

Improved Design: Collecting Multiple Samples

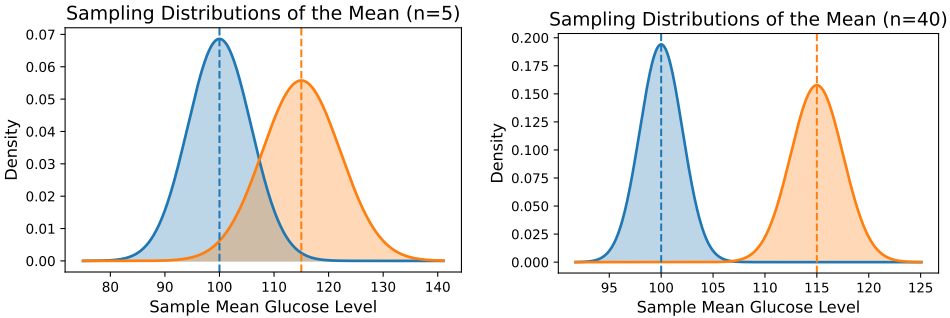
With five samples per treatment, we can calculate more reliable sample means:



Increasing sample size enhances our ability to detect true effects, but random variation remains.

Sample Size and Precision of Estimates

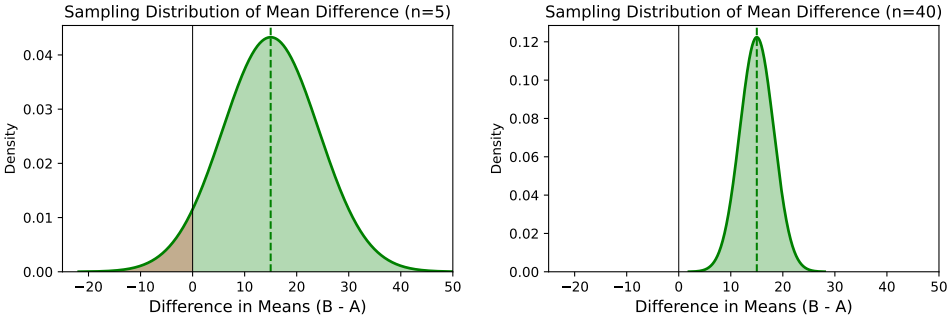
Sampling distributions become narrower with larger sample sizes:



As sample size increases from n=5 to n=25, sampling distributions become more distinct, making it easier to detect true differences.

Distribution of the Difference Between Means

The difference between sample means ($\Delta = \mu_B - \mu_A$) is itself a random variable:



With larger samples, the distribution of differences becomes narrower and shifts away from zero.

The Philosophy of Statistical Testing

Statistical hypothesis testing provides a structured approach to distinguish genuine effects from random variation.

Fisher's approach to significance testing focuses on:

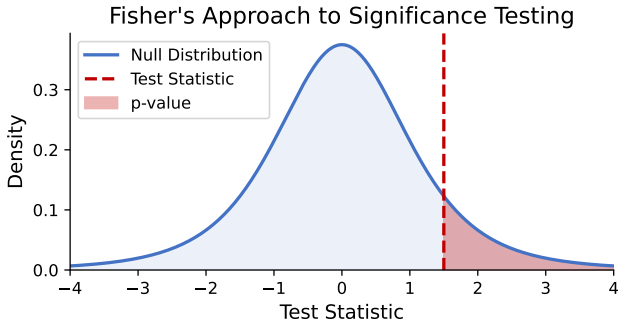
- Evaluating evidence against a null hypothesis
- Not making definitive claims about "truth"
- Quantifying compatibility between observed data and a reference hypothesis

Definition (Null Hypothesis)

The null hypothesis (H_0) is a specific statement about a population parameter that represents the absence of the effect being studied.

Fisher's Approach to Significance Testing

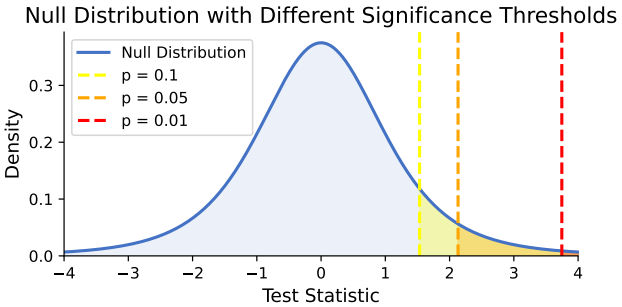
A systematic approach to evaluate evidence:



The p-value represents the probability of obtaining a test statistic as extreme or more extreme than observed, assuming the null hypothesis is true.

Interpreting Evidence: Significance Thresholds

Different significance thresholds represent varying levels of evidence against the null hypothesis:



Fisher emphasized these thresholds are conventions rather than rigid decision boundaries.

The Complete Framework of Statistical Testing

Fisher's statistical testing framework follows this process:

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

This framework applies to a wide range of testing scenarios, with varying test statistics and distributions.

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

The Sample Mean and its Distribution

The most fundamental estimator is the sample mean:

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

Properties of the sampling distribution of the mean:

Property (Expected Value of Sample Mean)

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

The sample mean is an unbiased estimator of the population mean.

Property (Variance of Sample Mean)

$Var(\bar{X}) = \frac{\sigma^2}{n}$

The variability of the sample mean decreases as sample size increases.

Standard Error of the Mean

Definition (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 estimate the SEM using the sample standard deviation:

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

The standard error is fundamental to hypothesis testing because it quantifies how much the sample mean is expected to vary from the true population mean by chance alone.

Sample Standard Deviation

Definition (Sample Standard Deviation)

The sample standard deviation s is an estimator of the population standard deviation σ:

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

Definition (Chi-Square Distribution)

If $Z_1, Z_2, \ldots, Z_k$ are independent standard normal random variables, then:

$$X = \sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$$

The chi-square distribution with k degrees of freedom has probability density:

Sampling Distribution of Sample Variance

Property (Sampling Distribution of Sample Variance)

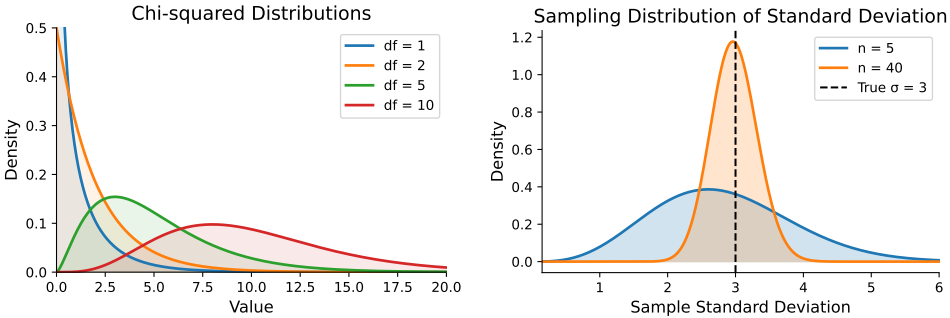
If samples are drawn from a normal distribution with variance σ^2, then:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

This property means that the sampling distribution of the sample variance is more complex than that of the sample mean.

Unlike the sample mean, which has a symmetric sampling distribution, the sampling distribution of the sample variance is right-skewed, especially for small sample sizes.

Chi-squared Distribution and Sampling Distribution of Standard Deviation



Left: Chi-squared distributions with different degrees of freedom Right: Sampling distribution of standard deviation for n=5 and n=40 samples

The Central Limit Theorem

Theorem (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.
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 means that even if our original data comes from a non-normal distribution (as is common in biological measurements), the sampling distribution of the mean will still approximate a normal distribution with sufficiently large samples.

Practical Testing Scenarios

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.

Example: We've measured gene expression levels in 30 tumor samples and want to know if the mean expression level differs from a reference value observed in healthy tissue.

This scenario helps us understand the fundamental structure of statistical tests before moving to more complex situations.

The Z-Test: Testing with Known Variance

When population variance σ^2 is known, we can use the z-test:

Under the Central Limit Theorem, the standardized sample mean follows a standard normal distribution:

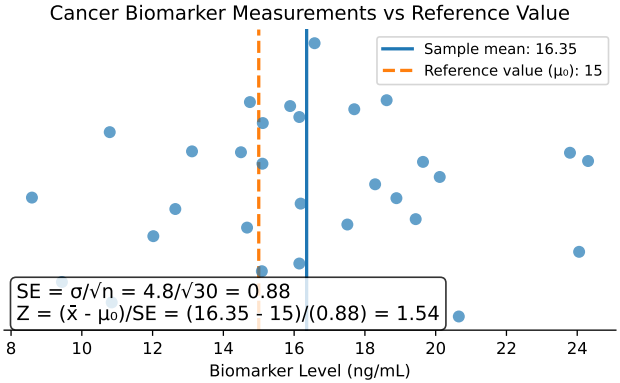
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

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.

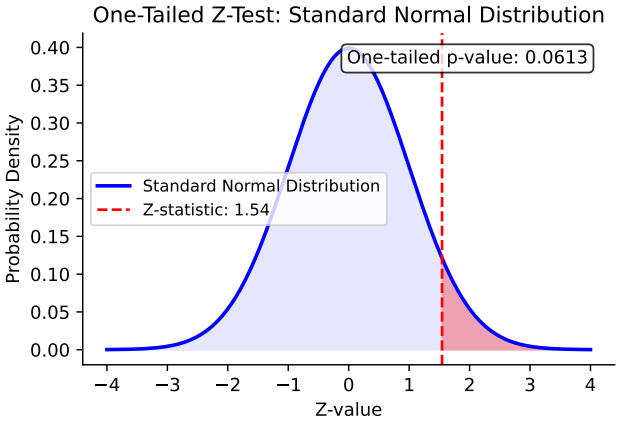
One-Sample Test Visualization



The sample mean (horizontal line) compared to the reference value μ_0 (dashed line).
The key question: Does our sample appear to come from a population with mean μ_0?

One-tailed Z-test

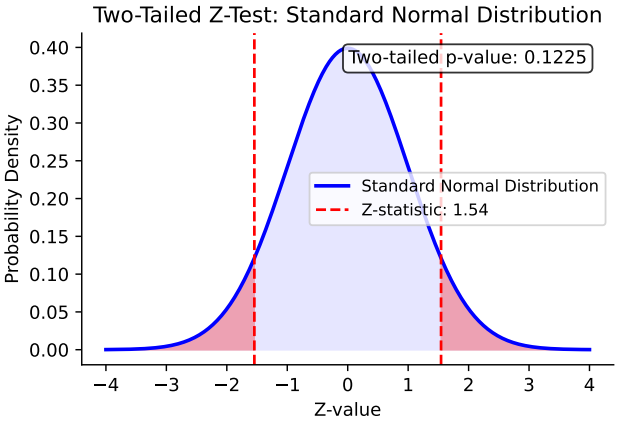
For a one-sided test, the p-value is the probability in one tail of the distribution:



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

Two-tailed Z-test

For a two-sided test, the p-value accounts for deviations in either direction:



Two-tailed test ($H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$):

From Theory to Practice - Additional Uncertainty

In reality, we rarely know the population variance σ^2 and must estimate it from our sample.

We substitute the estimated standard deviation into our test statistic:

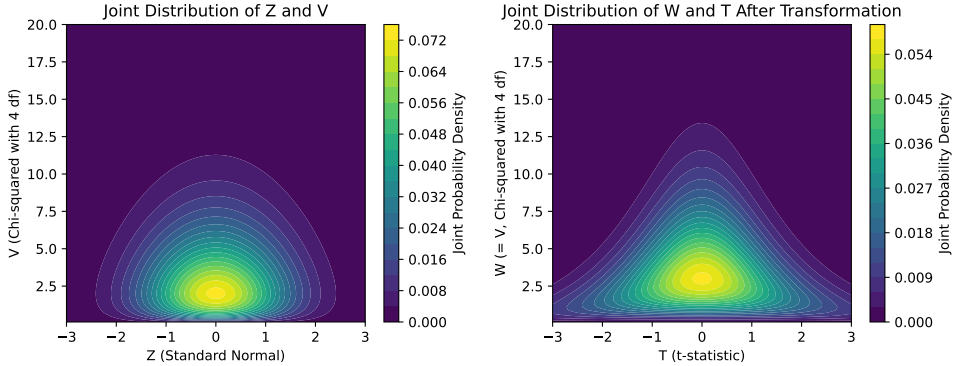
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

This introduces additional uncertainty: s is a random variable calculated from the sample, not a fixed parameter.

The resulting test statistic t cannot follow a normal distribution—we expect a more dispersed distribution with heavier tails.

Deriving the t-Distribution: Joint Distributions

The t-distribution emerges from the relationship between two random variables:



Left: Joint distribution of Z (standard normal) and V (chi-squared) Right: Joint distribution after transformation showing how the t-distribution emerges

Student's t-distribution

Theorem (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:

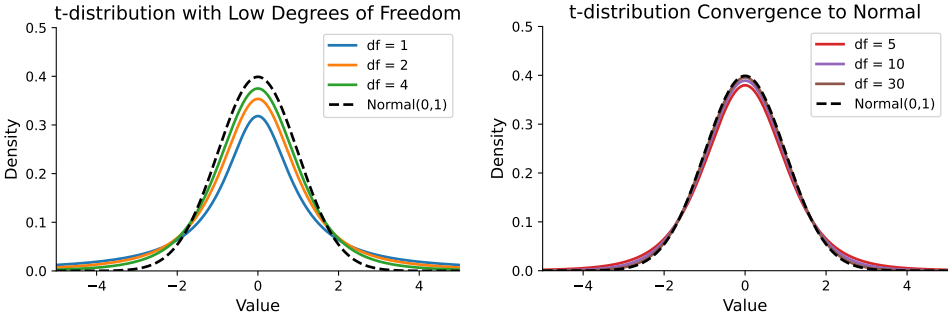
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

The probability density function of the t-distribution:

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

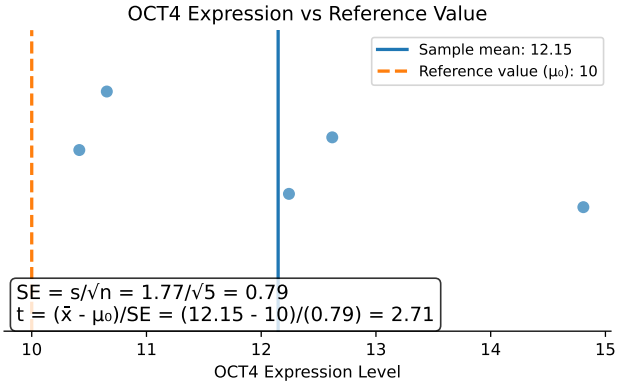
where $\nu = n - 1$ is the degrees of freedom.

The t-distribution and its Convergence to Normal



Left: t-distributions with small degrees of freedom have heavier tails than the normal distribution Right: As degrees of freedom increase, the t-distribution approaches the standard normal distribution

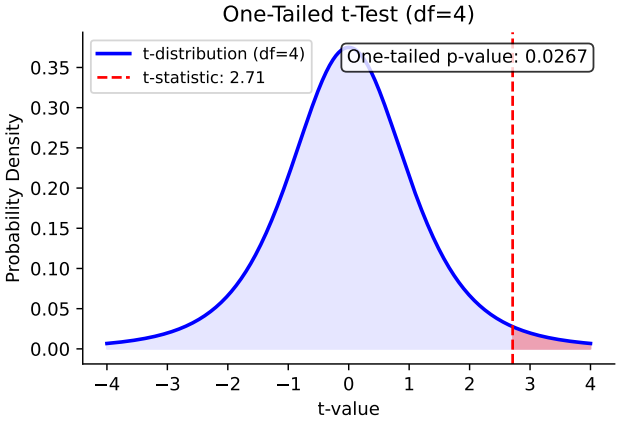
One-sample t-test Example: OCT4 Expression



Testing whether OCT4 expression in 5 stem cell cultures differs from an established reference value.

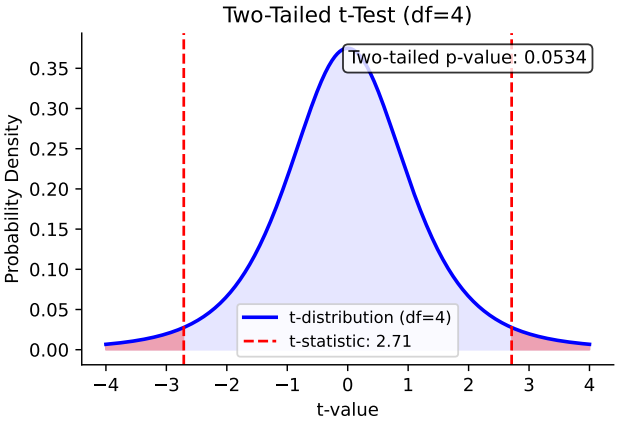
One-tailed t-test

With unknown variance and small sample size (n=5), we use the t-distribution with 4 degrees of freedom:



Two-tailed t-test

For testing whether OCT4 expression differs from the reference in either direction:



The p-value includes both tails of the t-distribution, capturing deviations in either

Two-Sample Tests: Comparing Independent Groups

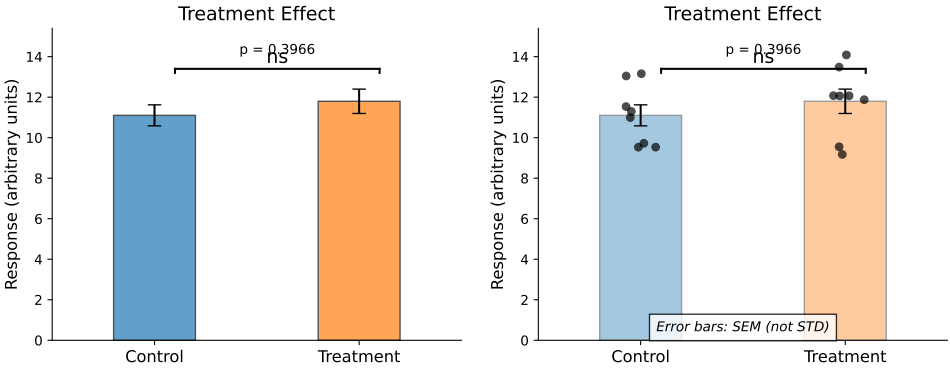
The two-sample t-test is one of the most common statistical tests in biological research.

In biological papers, this test is typically reported above bar plots showing means with standard error of the mean (SEM) error bars.

Key elements of the Welch test (which doesn't assume equal variances):

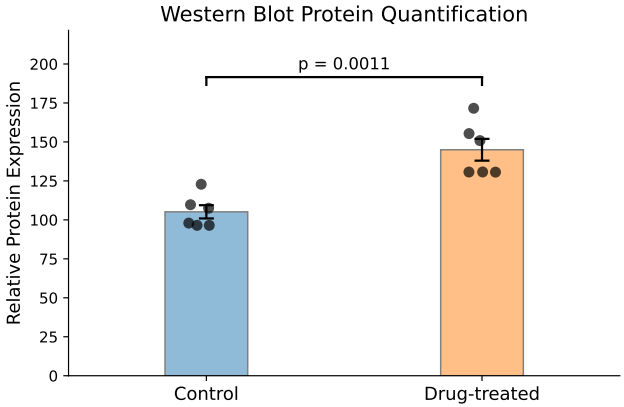
- Parameter of interest: difference between population means $\mu_1 - \mu_2$
- Null hypothesis: $\mu_1 - \mu_2 = 0$
- Test statistic: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Two-Sample Tests in Biological Research



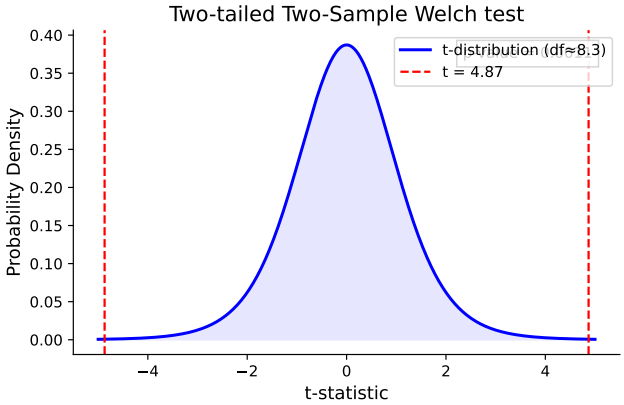
Left: Traditional bar plot with error bars found in papers Right: Modern presentation showing individual data points (increasingly required by journals)

Biological Example: Protein Expression in Western Blot Analysis



Investigating whether drug treatment alters protein expression by comparing 6 control and 6 treated samples.

Two-Sample t-test for Protein Expression Data



With $\bar{x}_{control} = 1.0$, $\bar{x}_{treated} = 1.75$, $s_{control} = 0.22$, $s_{treated} = 0.34$, and n=6 for each group, we calculate $t \approx 4.78$ with approximately 9 degrees of freedom, giving $p \approx 0.001$.

Working with Paired Samples

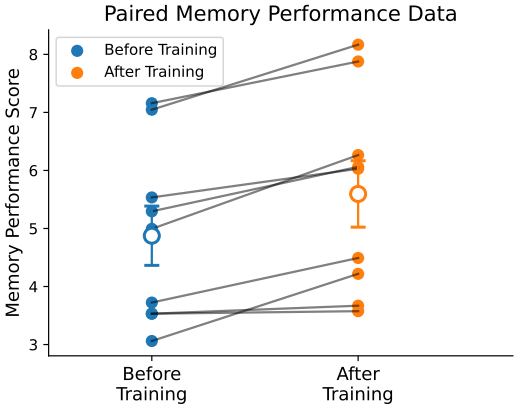
In many biological experiments, we have natural pairing between observations:

- Before and after measurements on the same subjects
- Same biological samples measured at different time points
- Matched pairs of subjects (e.g., twins, littermates)

When between-subject variability is high, accounting for pairing can dramatically increase statistical power.

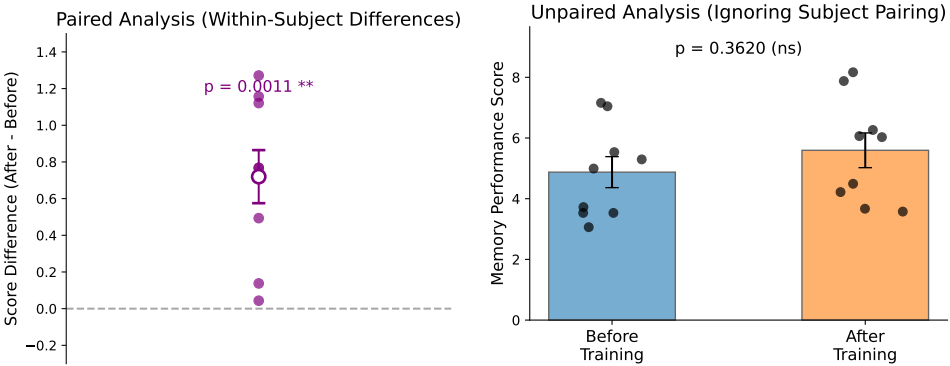
The key insight: Convert a two-sample problem into a one-sample problem by analyzing the differences within each pair.

Paired Data Visualization



Paired measurements from 9 mice before and after cognitive training, with connecting lines visualizing the within-subject relationships. Despite high between-subject variability, most subjects show consistent improvement.

Paired vs. Unpaired Analysis



Left: Paired analysis showing significant difference

Right: Unpaired analysis showing non-significant difference

Accounting for the experimental design (pairing) dramatically affects statistical power.

Question 1

What is the standard error of the mean (SEM)?

A) The standard deviation of the population

B) The standard deviation of the sample

C) The standard deviation of the sampling distribution of the mean

D) The variance of the sampling distribution of the mean

Question 2

A researcher conducts a drug trial and obtains a p-value of 0.03 for the difference in mean response between treatment and control groups. What is the correct interpretation of this p-value?

A) There is a 3% probability that the drug has no effect

B) There is a 3% probability that the observed difference occurred by chance

C) If the drug truly has no effect, there is a 3% probability of observing a difference as large or larger than what was observed

D) 97% of patients will respond positively to the drug

Question 3

A sample of size $n = 20$ is drawn from a normal population. The sample variance s^2 is calculated. Which of the following expressions correctly describes the sampling distribution of the quantity $\frac{(n-1)s^2}{\sigma^2}$?

A) A chi-square distribution with n degrees of freedom

B) A chi-square distribution with $n - 1$ degrees of freedom

C) A t-distribution with $n - 1$ degrees of freedom

D) A normal distribution with mean 0 and variance 1

Question 4

In a paired t-test with 15 pairs of observations, the degrees of freedom for the test statistic is:

A) 14

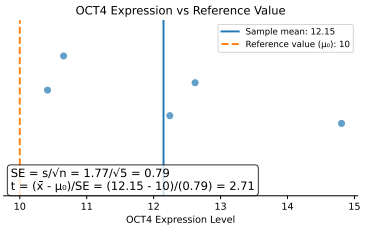
B) 15

C) 28

D) 29

Question 5

The figure below shows one-sample data with a reference value (dashed line). If you wanted to test whether the sample mean is significantly different from the reference value, what would be the most appropriate test?



A. Z-test, because the sample size is small

B. Two-sample t-test, because the reference represents a second group

C. One-sample t-test, because we deal with a fixed reference value

D. One-sample Paired t-test, to account for the experimental design

Multiple Testing and False Discovery Control

From Evidence to Decisions: The Neyman-Pearson Framework

While Fisher's approach focuses on evaluating evidence, Neyman and Pearson developed a complementary framework emphasizing decision-making with controlled error rates.

Definition (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.

Definition (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.

If $p < \alpha$, we reject H_0; if $p \geq \alpha$, we fail to reject H_0.

Type I and Type II Errors

This decision process can lead to two types of errors:

Definition (Type I Error)

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

Definition (Type II Error)

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

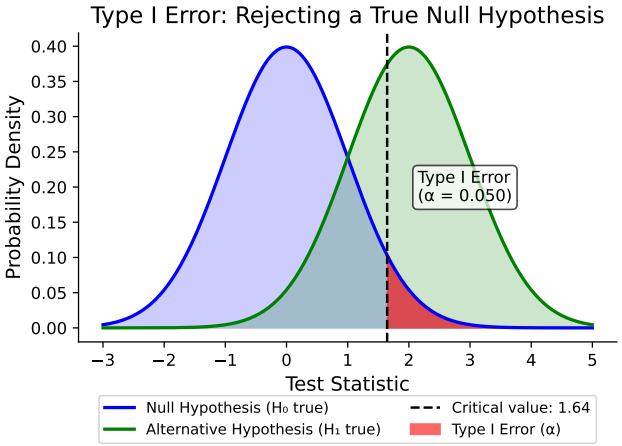
The Decision Matrix in Hypothesis Testing

	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 probability of a Type I error is controlled by the significance level α. The probability of a Type II error (β) depends on sample size, effect size, and variability.

Visualizing Type I Error

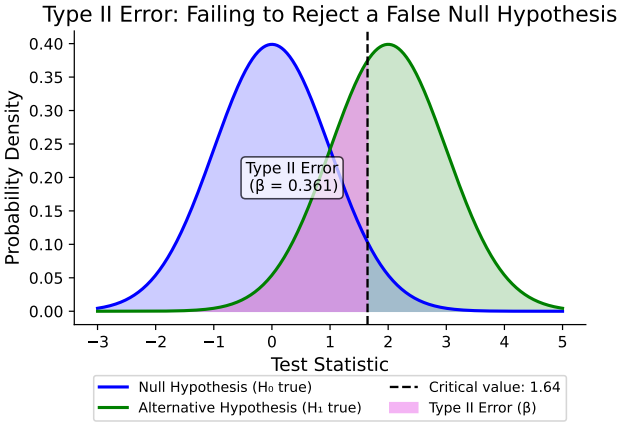
Type I error is the probability of rejecting a true null hypothesis (false positive):



The significance level α directly controls the Type I error rate: if $\alpha = 0.05$, we will

Visualizing Type II Error

Type II error is the probability of failing to reject a false null hypothesis (false negative):

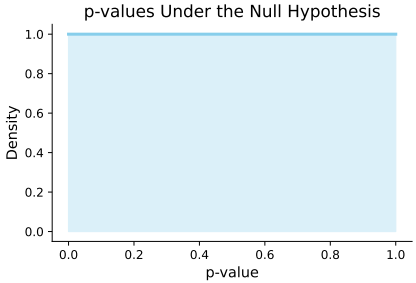


The Type II error rate (β) is not directly controlled in the testing procedure and

A Crucial Property: Uniformity Under the Null

Theorem (Uniformity of P-values Under the Null)

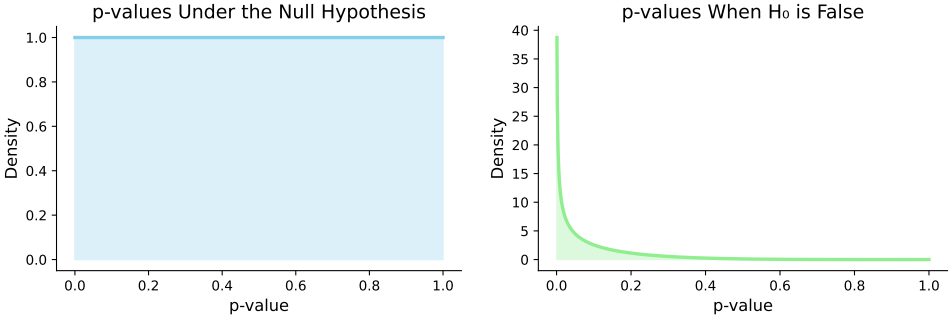
If the null hypothesis is true, and if the test statistic's distribution is continuous, then the p-value follows a uniform distribution on the interval [0,1].



This uniformity ensures that when the null hypothesis is true: The probability of obtaining a p-value ≤ 0.05 is exactly 0.05

P-values When the Null Hypothesis is False

When the null hypothesis is false, p-values tend to be smaller:



Left: Distribution of p-values under the null hypothesis (uniform)
Right: Distribution of p-values when the null hypothesis is false (shifted toward zero)
The extent of this shift depends on the effect size and sample size.

The Multiple Testing Problem in High-Throughput Biology

Modern biological research often involves testing many hypotheses simultaneously:

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

This creates a fundamental challenge: When we perform many tests, the probability of obtaining false positives increases dramatically.

Example: Testing 1,000 genes when none are differentially expressed

- 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!

Family-Wise Error Rate

Definition (Family-Wise Error Rate)

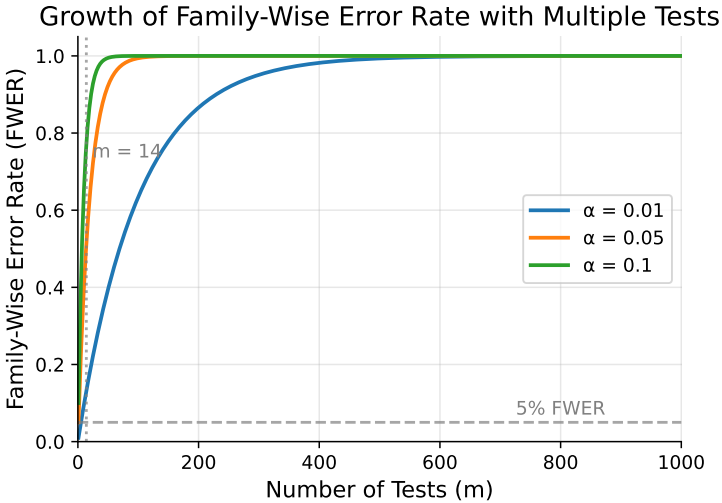
The Family-Wise Error Rate (FWER) is the probability of making at least one Type I error (false positive) among all the hypothesis tests conducted.

For m independent tests, each with significance level α:

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

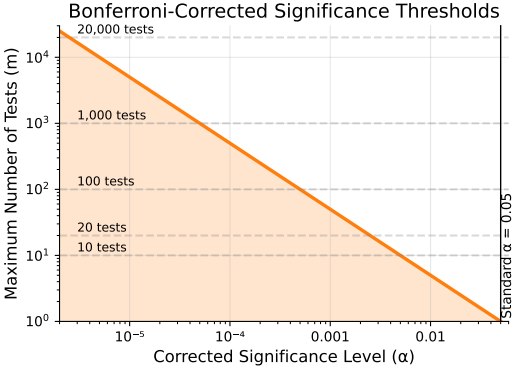
This grows rapidly with the number of tests. For example, with $\alpha = 0.05$ and $m = 100$ tests, FWER ≈ 0.994, meaning we're almost certain to get at least one false positive.

FWER Growth with Number of Tests



Bonferroni Correction

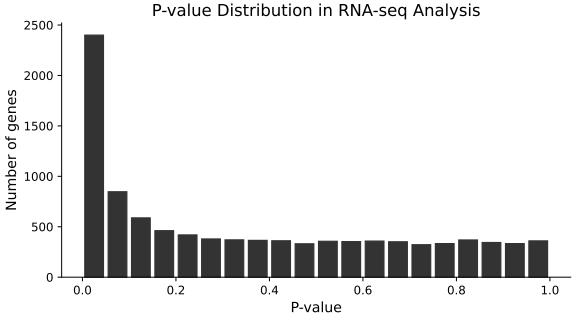
The Bonferroni correction is the simplest method to control FWER: $\alpha_{\text{corrected}} = \frac{\alpha}{m}$



For 20,000 genes, the corrected threshold becomes $\alpha = 0.05/20,000 = 0.0000025$, which is extremely stringent. The approach controls false positives but leads to many false negatives.

P-value Distributions in Real Data

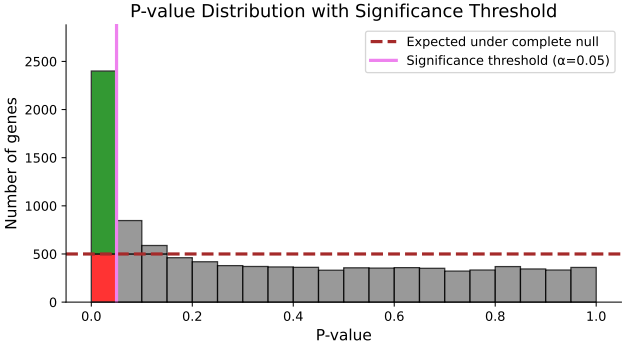
In actual biological datasets, p-values rarely follow a uniform distribution because some null hypotheses are false:



A typical p-value histogram from RNA-seq differential expression analysis shows enrichment of small p-values (genes that are truly differentially expressed) and a relatively flat distribution for larger p-values.

Interpreting P-value Distributions as Mixtures

The observed p-value distribution can be understood as a mixture:



The relative heights of the uniform component (non-differential genes) versus the enrichment near zero (differential genes) give insight into the proportion of true vs. false null hypotheses in our dataset.

False Discovery Rate: A More Practical Approach

Definition (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 are likely false positives, while 90 represent true effects.

This is often a more useful metric for biological research than the more stringent FWER, providing a balance between false positives and false negatives.

The Two-Groups Model

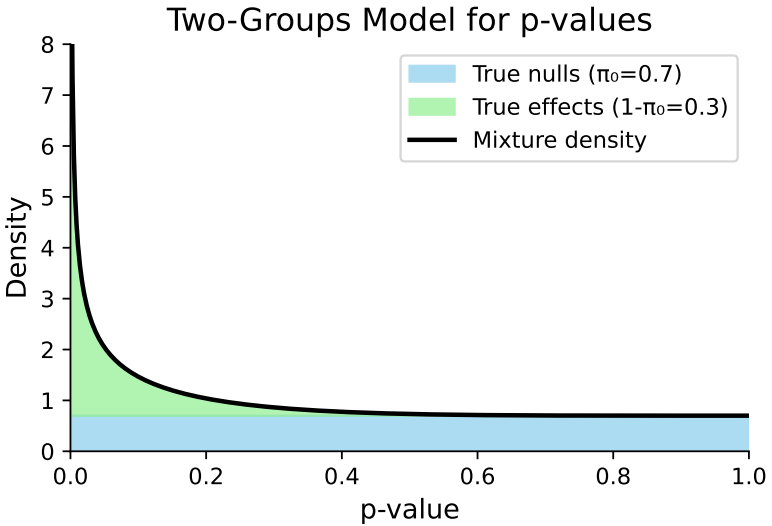
A deeper understanding comes from the "two-groups model" introduced by Efron:

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

where:

- π_0 is the proportion of true null hypotheses
- $f_0(p)$ is the density of p-values under the null (uniform)
- $f_1(p)$ is the density of p-values under the alternative (near 0)
- $f(p)$ is the overall mixture density observed in data

The Two-Groups Model: Visual Representation



Local False Discovery Rate

Definition (Local False Discovery Rate)

The local false discovery rate at a specific p-value p is:

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

This gives the probability that a test with exactly that p-value comes from the null hypothesis.

To extend from local FDR to the overall FDR for a significance threshold:

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

where $F(p) = \int_0^p f(t)dt$ is the cumulative distribution function.

Visualizing Local FDR and Overall FDR

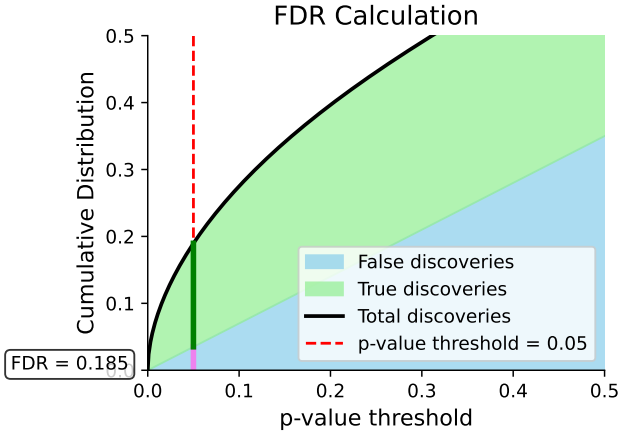


Figure: The local false discovery rate at each p-value (blue curve) can be integrated to determine the overall false discovery rate at any threshold.

The Benjamini-Hochberg Procedure

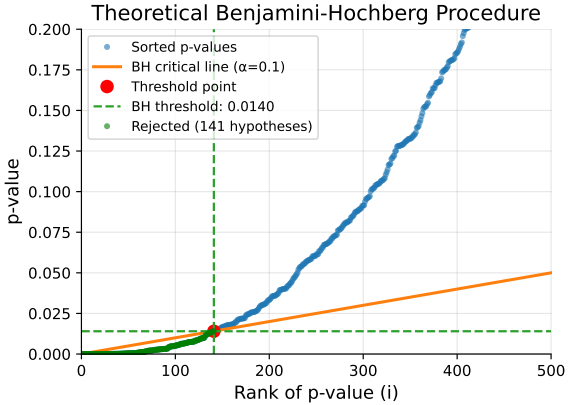
The most widely used method for controlling FDR is the Benjamini-Hochberg (BH) procedure:

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 $\leq p_{(k)}$

This guarantees that the expected FDR will be at most α.

The BH procedure effectively creates an adaptive threshold that becomes more stringent as the p-value increases, accounting for the expected proportion of false discoveries.

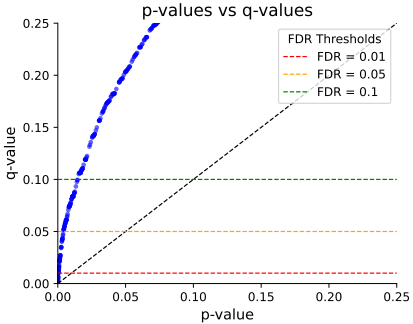
Visualizing the Benjamini-Hochberg Procedure



The procedure rejects all hypotheses with p-values falling below the BH line with slope α/m, ensuring FDR control while being less conservative than Bonferroni correction.

Definition (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. A gene with q-value = 0.05 means that 5% of genes considered significant at that threshold are expected to be false positives.

Practical Considerations in Biological Applications

When applying multiple testing procedures to biological data:

- **Independence assumptions:** Most correction methods assume independence between tests, which rarely holds in biology (e.g., correlated gene expression)
- **Pre-filtering:** Removing tests unlikely to yield significant results (e.g., filtering low-expression genes) can reduce the multiple testing burden
- **Effect size considerations:** Statistical significance does not imply biological significance
- **Exploratory vs. confirmatory analysis:** In early-stage research, a higher FDR might be acceptable for generating hypotheses to be validated in follow-up experiments

The choice of multiple testing strategy should be guided by the research context, the number of tests, and the relative costs of Type I and Type II errors.

Question 6

A biologist tests 2,000 genes for differential expression between healthy and diseased tissue using a significance threshold of $\alpha = 0.05$, and finds 200 significant genes. If they had not applied any multiple testing correction, approximately how many false positives would be expected among these 200 genes?

A) 10 false positives

B) 100 false positives

C) 200 false positives

D) It's impossible to estimate without knowing the true proportion of differentially expressed genes

Question 7

When controlling the Family-Wise Error Rate (FWER) at $\alpha = 0.05$ using the Bonferroni correction for 1,000 independent tests, what is the corrected significance threshold for each individual test?

A) $\alpha_{corrected} = 0.05$

B) $\alpha_{corrected} = 0.005$

C) $\alpha_{corrected} = 0.0005$

D) $\alpha_{corrected} = 0.00005$

Question 8

What is the key difference between controlling the Family-Wise Error Rate (FWER) and controlling the False Discovery Rate (FDR)?

A) FWER controls the probability of making at least one false discovery, while FDR controls the expected proportion of false discoveries among all rejected null hypotheses

B) FWER is applicable only to small numbers of tests, while FDR works for any number of tests

C) FWER requires independence between tests, while FDR does not make any assumptions about independence

D) FWER is always more powerful than FDR regardless of the number of tests

Question 9

A researcher calculates a test statistic $t = 2.8$ from a sample of $n = 10$ observations. Under the null hypothesis, this statistic follows a t-distribution with 9 degrees of freedom. If the p-value for a two-tailed test is 0.021, what would be the p-value if the researcher had inappropriately used a standard normal distribution instead of the t-distribution?

A) $p < 0.005$

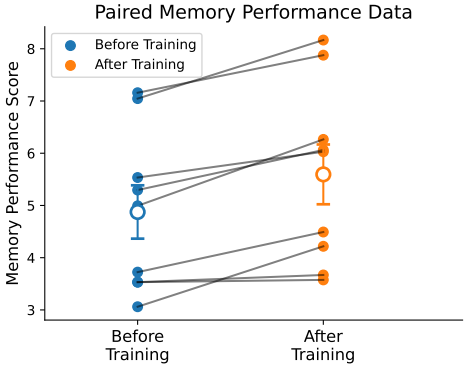
B) $0.005 < p < 0.01$

C) $0.01 < p < 0.02$

D) $p > 0.02$

Question 10

Referring to the figure showing paired versus unpaired analysis, why does the paired t-test detect a significant effect while the unpaired t-test does not?



A) The paired t-test uses more sophisticated statistical methods

B) The paired t-test accounts for the consistent within-subject differences despite high between-subject variability

C) The paired t-test has a different null hypothesis than the unpaired t-test

D) The paired t-test requires fewer assumptions about the data distribution