

Abstract

Statistical Significance

Beyond 0.05: Lessons from Genomics Research

Author: Kyle Kaufman

Genetic insights refining traditional statistical thresholds

Agenda

- **Slide 1: Clear Subject Identification**
 - Define the subject of the presentation.
 - Ensure the subject aligns with the audience's expectations and interests.
- **Slide 2: Presentation Topic Definition**
 - Clearly articulate the main topic of the presentation.
 - Use straightforward language to avoid confusion.
- **Slide 3: Clarification of Main Subject**
 - Break down the main subject into subtopics if necessary.
 - Ensure that all aspects of the main subject are covered in the presentation.
- **Slide 4: Highlighting Key Topics**
 - Identify and emphasize the key topics that are crucial for the audience to understand.
 - Use visual aids or bullet points to highlight important themes.
- **Slide 5: Specification of Presentation Focusd**
 - Clearly state the focus of the presentation to maintain a steady direction.
 - Ensure the content is relevant to the specified focus throughout the presentation.
- **Slide 6: Conclusion**
 - Recap the importance of having a clear agenda for an effective presentation.
 - Encourage questions or feedback on the agenda prior to finalizing the content.

Beyond $p < 0.05$: Genomics Insights



Challenges with Significance Thresholds

In genomic research, conventional p-value thresholds are inadequate due to the large number of simultaneous tests, necessitating revised methodologies.

The Importance of Statistical Power

In genomics, statistical power depends on both the sample size and the effect size, influencing the capacity to identify genuine associations.

Utilizing Power Analysis

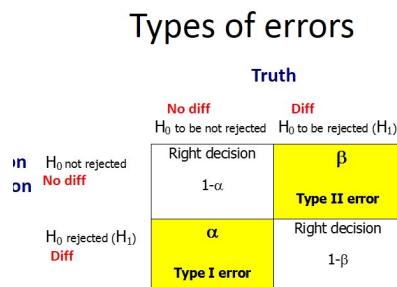
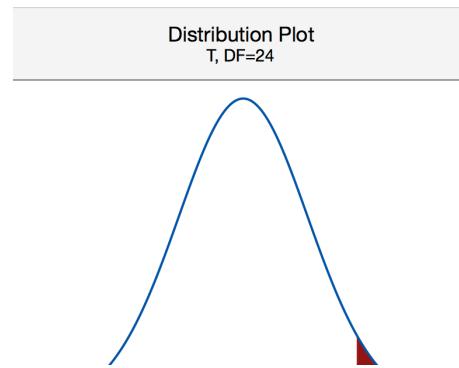
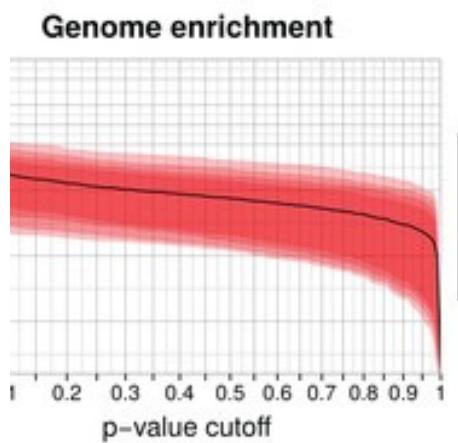
Power analysis tools assist researchers in designing genomic studies by determining the necessary sample sizes to achieve reliable genetic findings.

Additionally, the multiple testing issue is critical in genomic studies. Testing millions of genetic variants at once means that a 5% false positive rate can produce tens of thousands of false significant results. The adoption of genome-wide significance thresholds ($p < 5 \times 10^{-8}$) serves as an essential adjustment, albeit one that significantly raises sample size requirements to preserve adequate power.

These concepts apply broadly to any field involving multiple hypothesis testing, such as machine learning hyperparameter tuning, feature selection, and subgroup analyses. We introduce a framework for assessing statistical claims that takes into account the scale of testing, sample sufficiency, effect sizes, and the potential impact of findings.

Keywords: statistical power, multiple testing correction, genome-wide significance, p-value, Bonferroni correction, power analysis

Rethinking Statistical Significance



- H₀ is “true” but rejected: Type I or α error
- H₀ is “false” but not rejected: Type II or β error

Limitations of P-Value Threshold

The p-value of 0.05 alone may not ensure meaningful or reliable results in scientific studies.

Importance of Context and Sample Size

Interpreting statistical significance requires considering context, sample size, and the strength of observed effects.

Need for Critical Analysis

Relying solely on p-values can lead to errors; rigorous analysis and critical thinking are vital for accurate interpretation.

Sample Size Matters

Importance of Sample Size

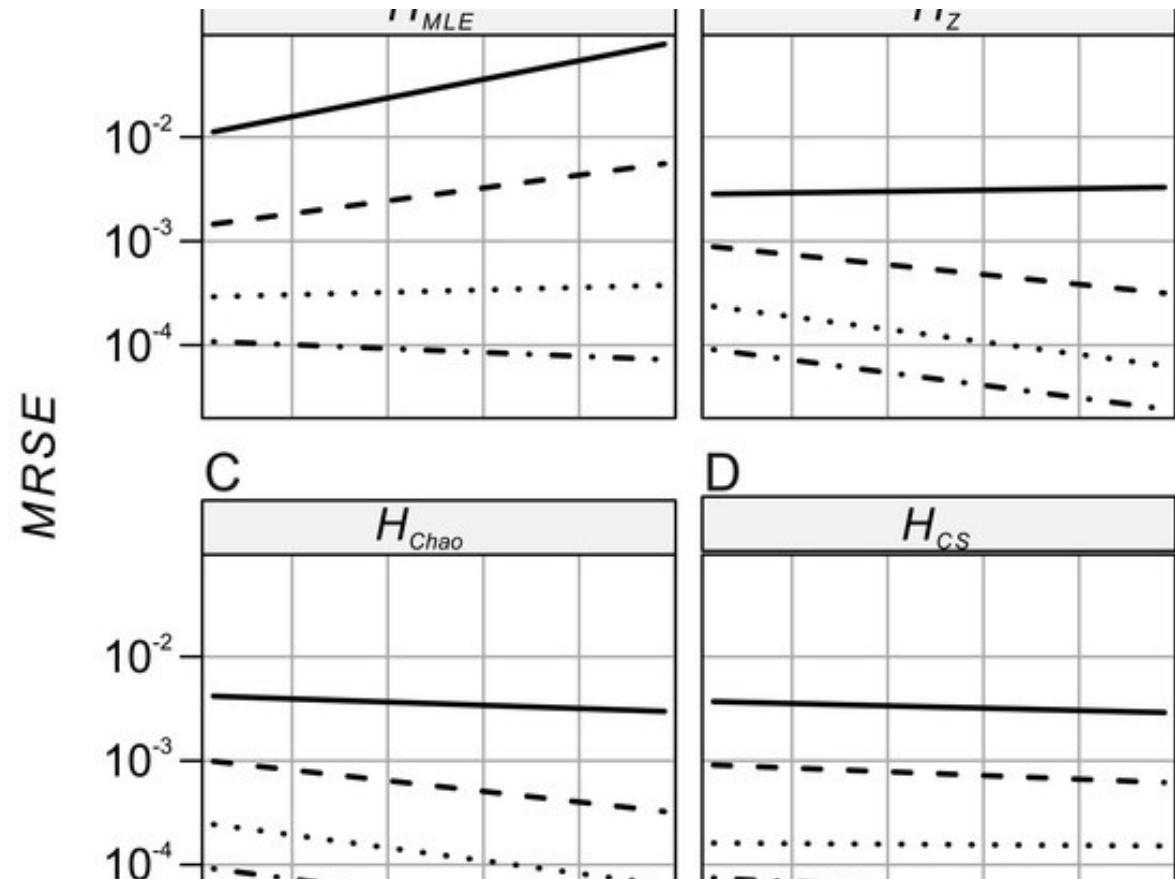
Selecting the right sample size is critical for achieving statistical power and accurately detecting real effects in studies.

Benefits of Larger Samples

Larger samples boost confidence in research results and lower the risk of missing significant findings.

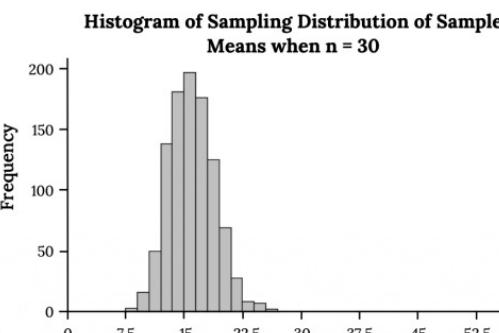
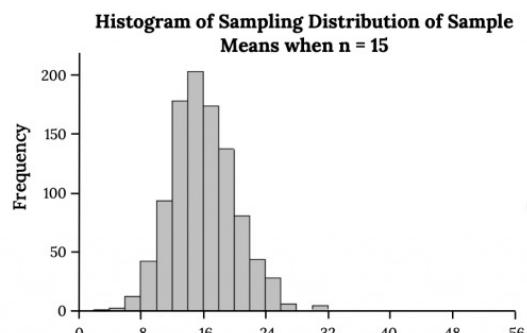
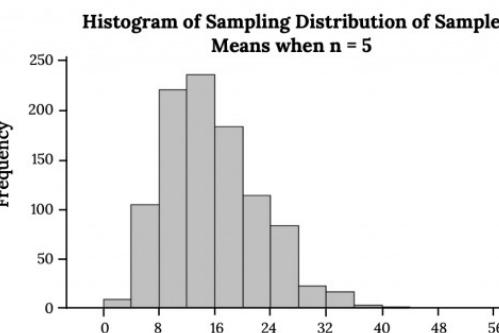
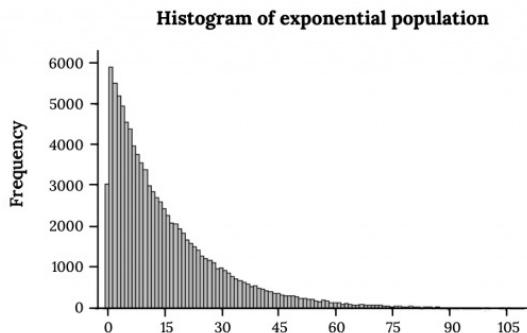
Ensuring Study Validity

Careful planning of sample size underpins the validity of studies and leads to meaningful, trustworthy conclusions.



[This Photo](#) by Unknown Author is licensed under [CC BY](#)

Why Sample Size Matters



Missing Real Discoveries

A sample size that is too small can overlook significant findings hidden in your data, leading to missed opportunities.

Wasted Time and Resources

Small samples produce unreliable outcomes, resulting in repeated experiments that waste valuable time and resources.

Replication and Credibility Issues

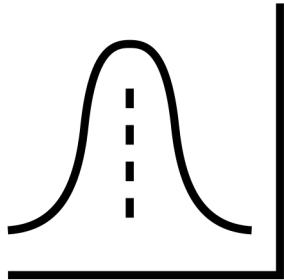
Findings from small samples often fail to replicate, undermining scientific credibility and the value of research.

Role of Power Analysis

Power analysis is a statistical tool that helps determine the correct sample size for valid and credible research.

[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Power Analysis Essentials

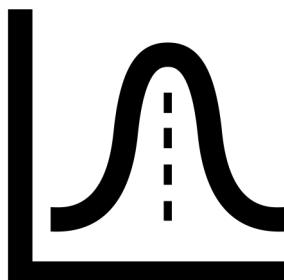


Determining Sample Size

Power analysis is used to calculate the minimum sample size required to reliably detect meaningful effects in a study.

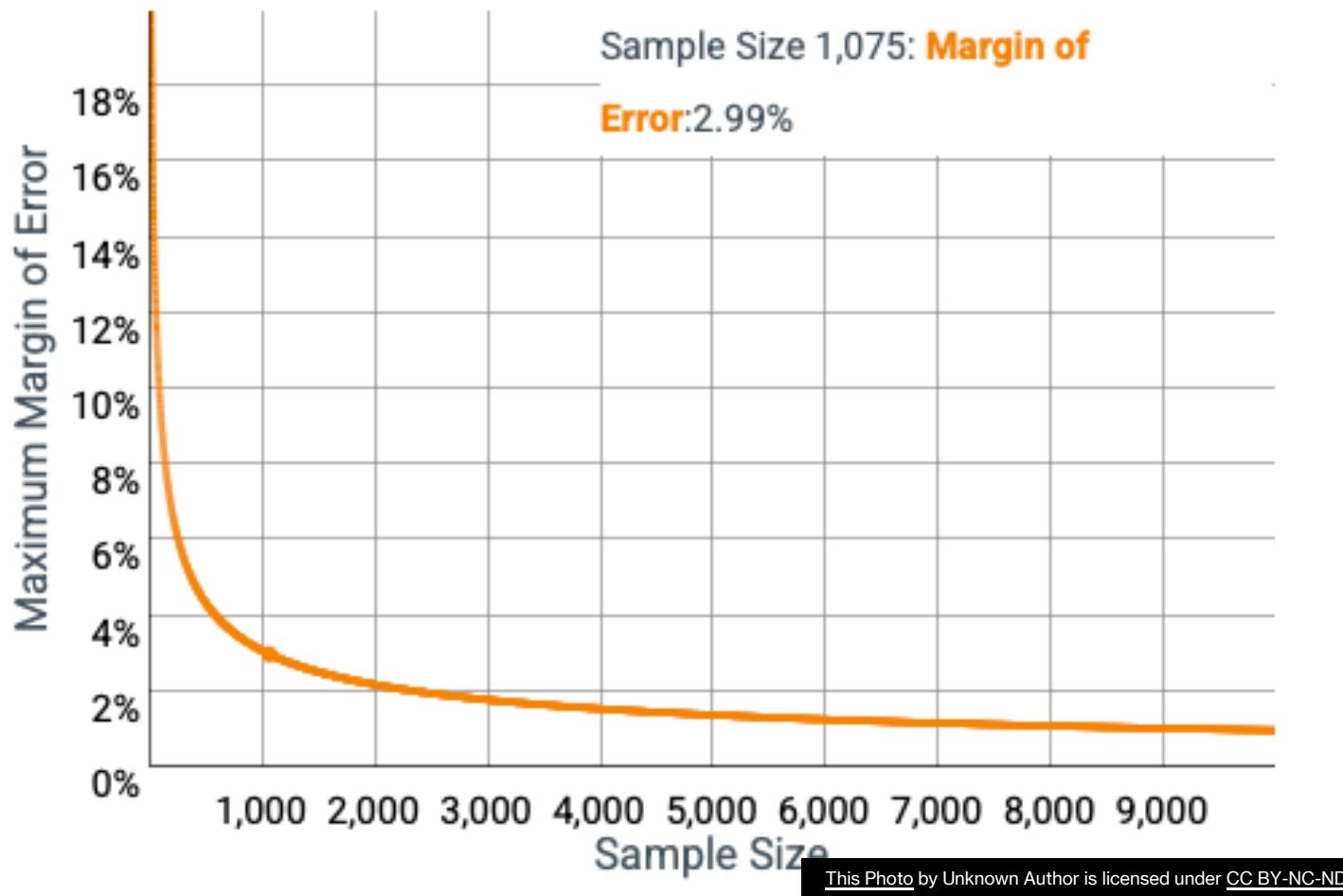
Key Factors Considered

It takes into account the alpha level, desired power, effect size, and data variance to ensure statistical validity.



Ensuring Study Reliability

Proper power analysis guarantees study design is strong enough to detect results that are truly significant.



Beyond Sample Size

Importance of Study Design

Robust study design ensures credible research outcomes and helps minimize biases, leading to valid scientific conclusions.

Role of Data Quality

High-quality data is essential for reliable results; poor data can compromise the integrity of research findings.

Analytical Methods Influence

Proper analytical methods allow accurate interpretation of data and strengthen the validity of scientific research.

The Multiple Testing Problem

Risk of False Positives

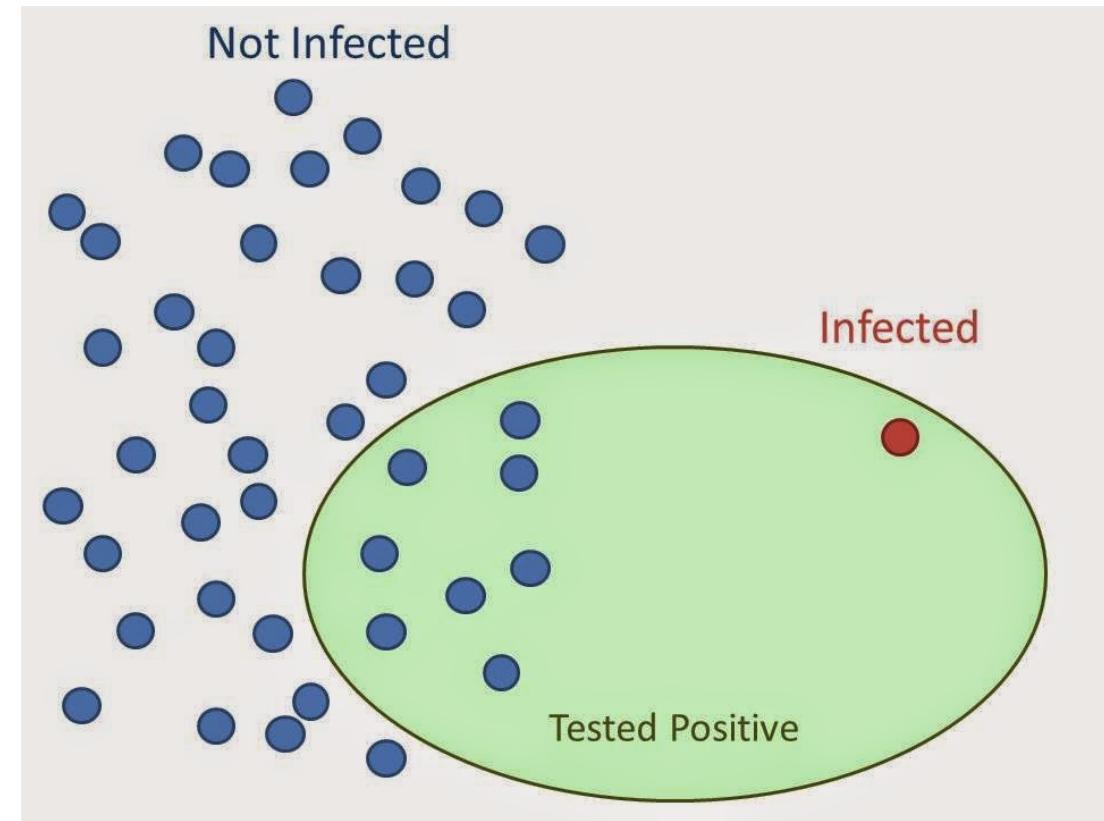
Performing multiple statistical tests increases the likelihood of false positive findings, which can mislead research conclusions.

Limitations of P-Value Thresholds

Using a traditional 0.05 p-value threshold may produce misleading results when many tests are run, raising cumulative error rates.

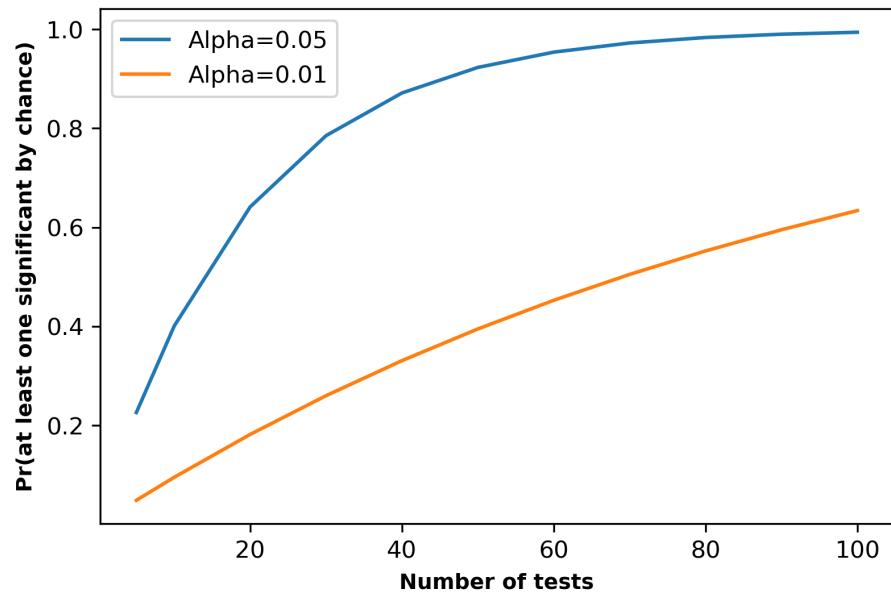
Correction Methods

Stricter significance thresholds or corrections like Bonferroni are used to minimize the risk of false discoveries in multiple testing.

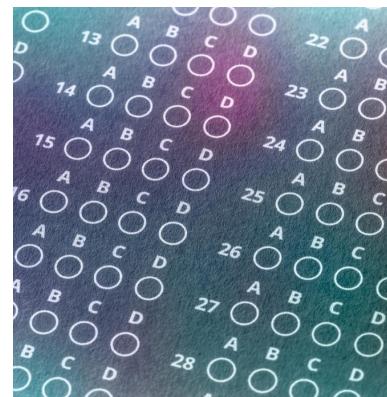


[This Photo](#) by Unknown Author is licensed under [CC BY-NC](#)

The Multiple Testing Problem



This Photo by Unknown Author is licensed under [CC BY](#)



A blackboard filled with mathematical calculations, including equations for variance, mean, and standard deviation, along with various handwritten notes and formulas, illustrating the complexity of statistical analysis.

Millions of Genetic Tests

DNA sequencing examines millions of genetic variants at the same time, vastly increasing statistical comparisons.

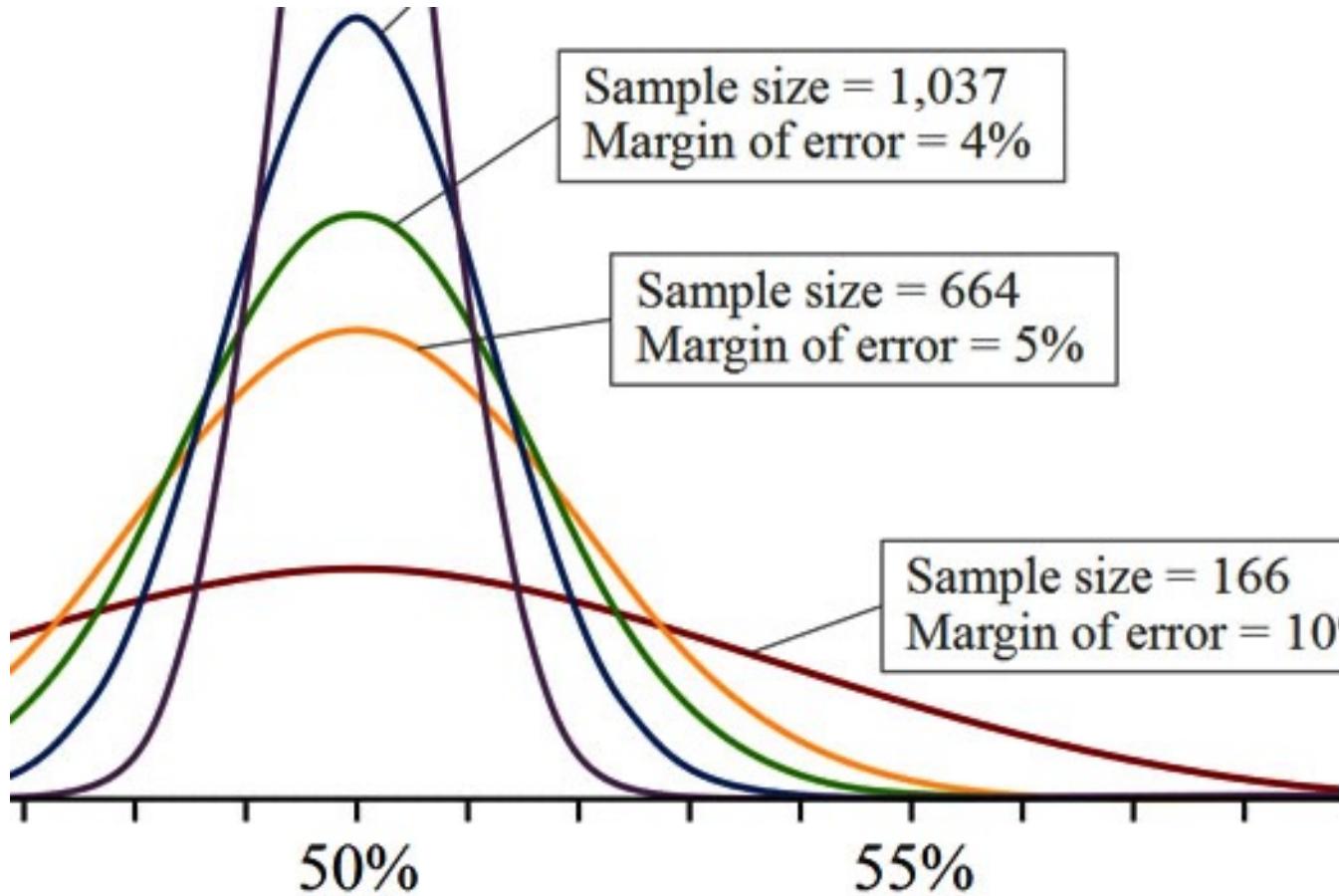
Increased False Positives

Testing so many variants raises the risk of false positives – results that appear significant purely by chance.

Need for Correction

Correcting for multiple comparisons is crucial to ensure findings are not misleading and truly significant.

Presentation Agenda



- Personal Experiences That Changed My Perspective On Statistics
- Lesson 1: The Critical Role Of Sample Size In Statistical Analysis
- Power Analysis: Ensuring Reliable Results In Research
- Lesson 2: Addressing The Multiple Testing Problem In Genetics
- Genome-Wide Significance: A More Stringent Approach To Statistical Testing

Personal Experiences That Changed My Perspective On Statistics



A Pivotal Conversation with My Dad, a Geneticist

Statistical Significance in Genetics

Understanding statistical significance is crucial for interpreting genetic study results accurately.

Limitations of P-value 0.05

A p-value threshold of 0.05 may be inadequate due to multiple testing and complex data in genetics.

Complexity of Genetic Data

Genetic data involves large-scale, complex datasets requiring advanced analysis methods.

Initial Assumptions and How They Evolved

Original Belief on Threshold

The 0.05 threshold was initially seen as a universal statistical standard.

Broadened Understanding

Genetics research revealed that significance thresholds vary depending on the context and study design.



Lesson 1: The Critical Role Of Sample Size In Statistical Analysis

TDM	729.89	915.51	185.62	▲ 25.43%
HUM	749.73	924.29	174.56	▲ 23.28%
DMW	833.72	1004.01	170.29	▲ 20.43%
YZJ	903.49	1127.46	223.97	▲ 24.79%
GLY	982.07	1219.39	237.32	▲ 24.17%
VDA	113.74	143.41	29.67	▲ 26.09%
UVV	468.08	535.41	67.33	▲ 14.38%
HJS	545.49	659.05	113.56	▲ 20.82%
EDR	592.70	699.00	87.34	▲ 17.20%
FLR	660.27	745.28	85.01	▲ 12.88%
UVG	155.59	181.57	25.98	▲ 16.70%
QUV	440.55	540.21	99.66	▲ 22.62%
HZT	285.51	344.93	59.47	▲ 20.83%
PCW	811.44	1029.66	218.22	▲ 26.89%
AIK	361.77	451.39	89.62	▲ 24.77%
ZJJ	858.36	994.57	136.21	▲ 15.87%
RHJ	694.79	1046.68	151.89	▲ 16.97%
WDV	84.08	109.90	84.87	▲ 12.97%

PPJ	912.63	1038.36	125.73	▲ 13.78%
UAQ	1309.55	1655.62	346.07	▲ 26.43%
DAQ	1295.17	1641.66	345.49	▲ 26.75%
PNR	654.33	775.84	121.51	▲ 18.57%
ZTM	201.52	231.21	30.69	▲ 15.68%
ZCK	391.59	491.48	99.89	▲ 25.51%
BNY	969.21	1130.65	161.44	▲ 16.66%
SDM	735.44	913.39	177.95	▲ 24.20%
TQO	1323.91	1646.42	322.51	▲ 24.36%
OJS	543.42	667.24	123.82	▲ 22.79%
ZIA	1402.01	1822.38	380.35	▲ 21.94%

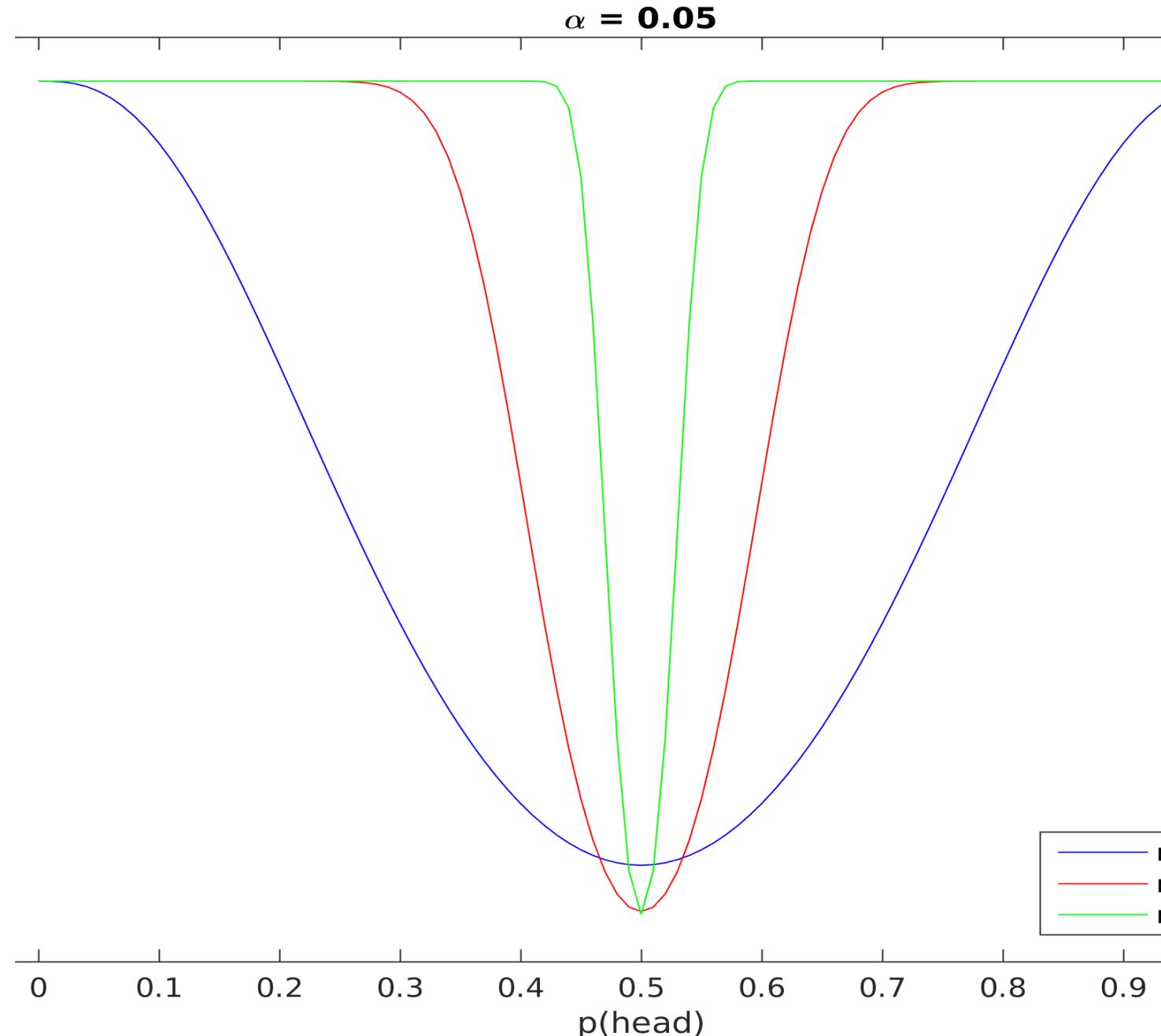
Understanding Statistical Power and Its Importance

Definition of Statistical Power

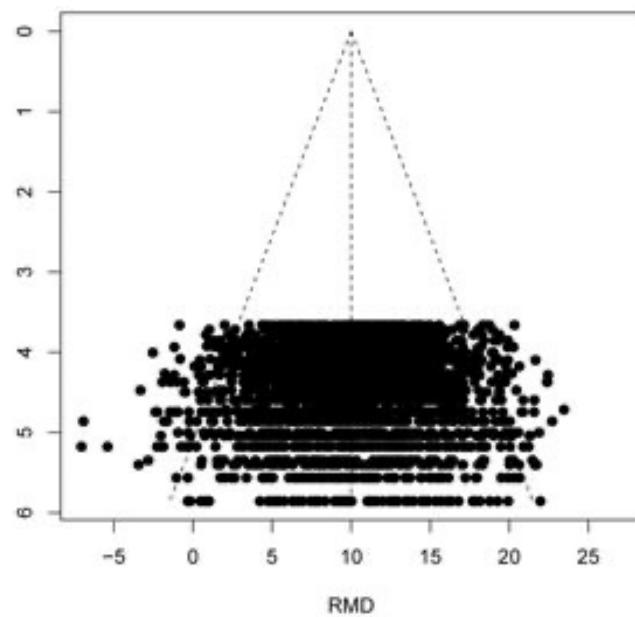
Statistical power measures the likelihood of detecting a true effect or correctly rejecting a false null hypothesis.

Importance in Study Design

Higher statistical power minimizes the chance of missing meaningful findings in research studies and experiments.

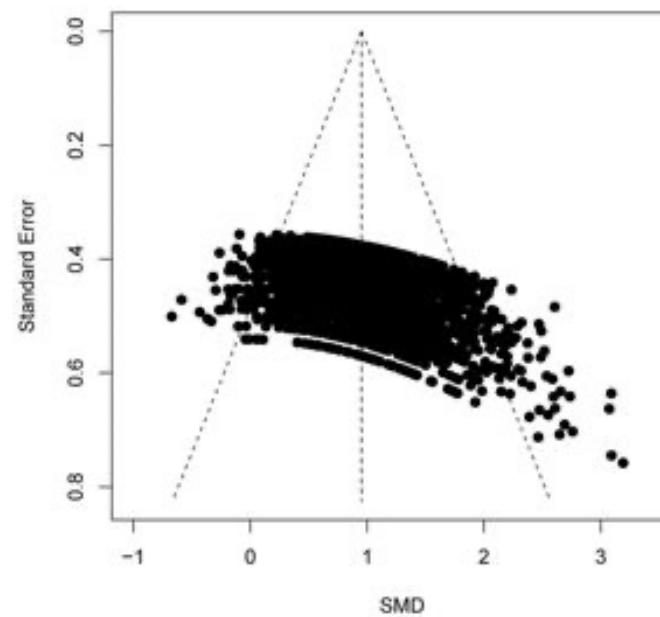


Simulation 2 (n=12-30)



B

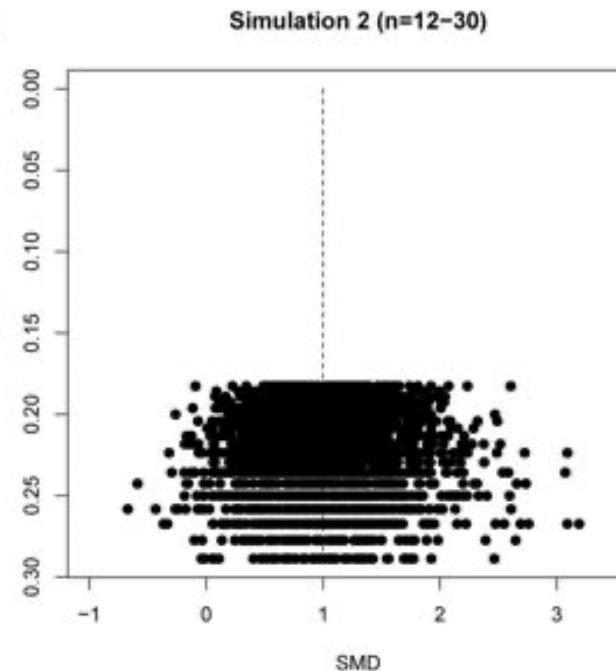
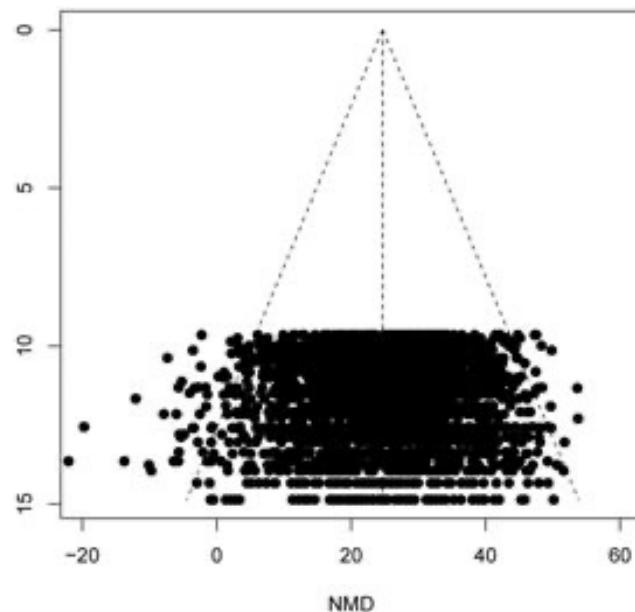
Simulation 2 (n=12-30)



Simulation 2 (n=12-30)

D

Simulation 2 (n=12-30)



Consequences of Small Sample Sizes: Missed Discoveries and Poor Replication

Increased Type II Errors

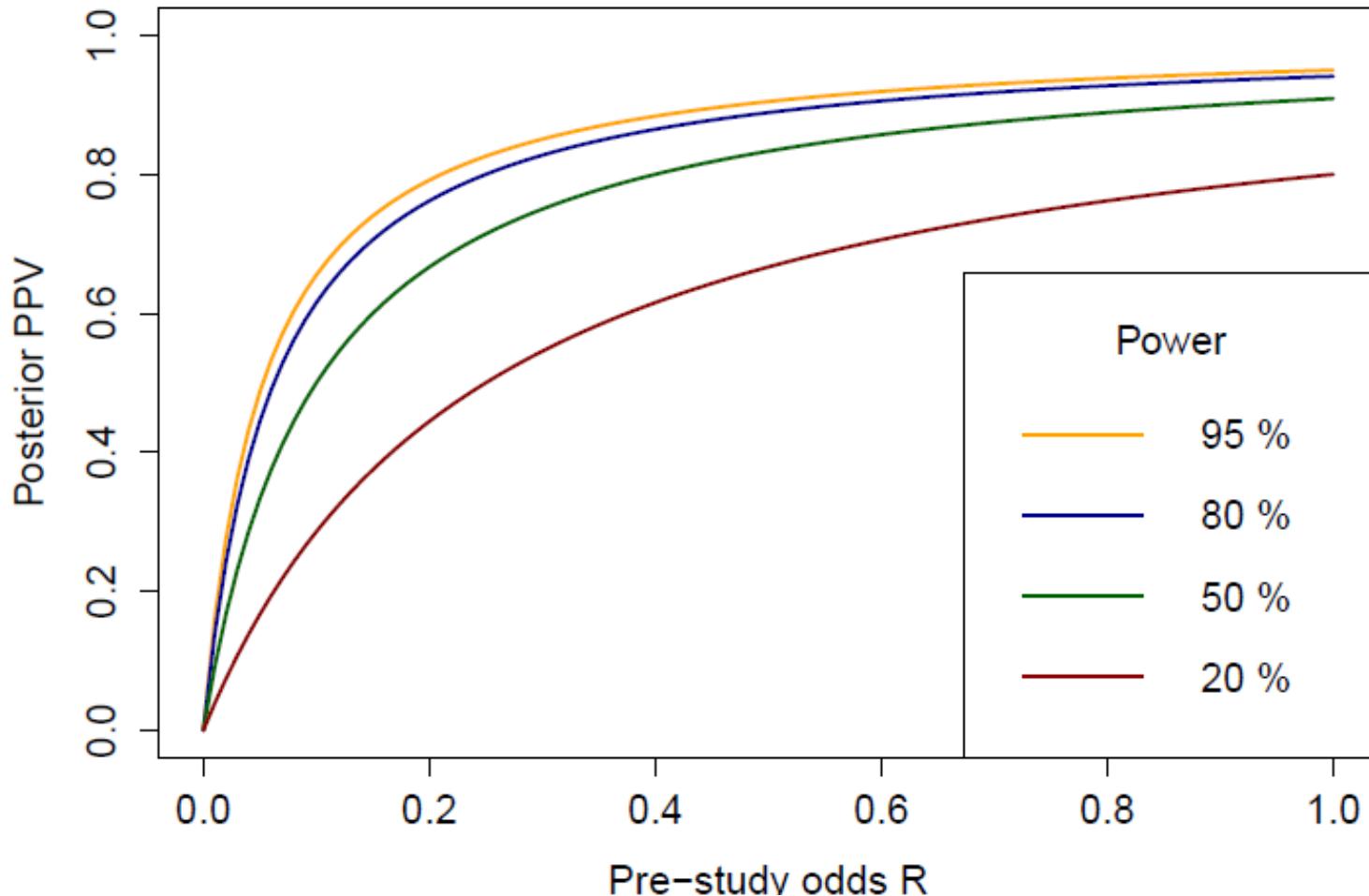
Small sample sizes raise the risk of failing to detect true effects, causing false negatives in research findings.

Missed Discoveries

Insufficient data reduces the chance of identifying significant effects, resulting in overlooked discoveries.

Replication Challenges

Small samples cause inconsistent results that are difficult to replicate, weakening scientific reliability.



Purpose of Power Analysis

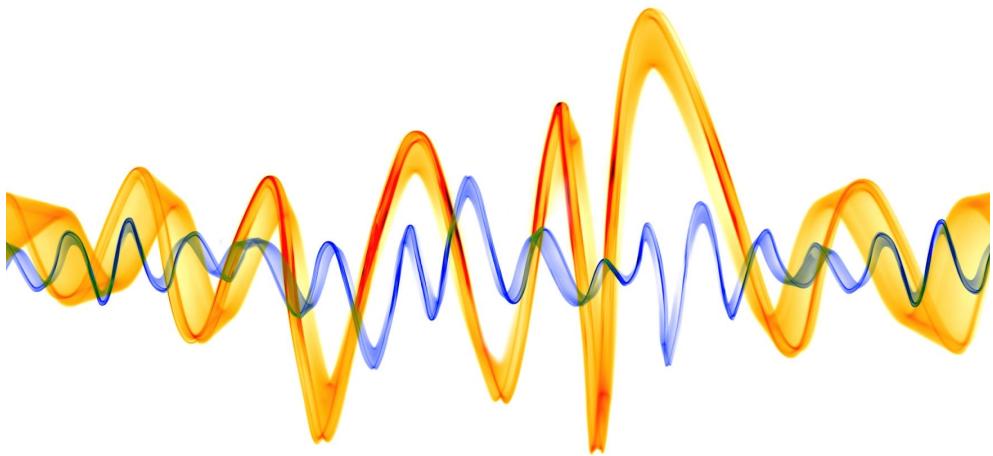
Power analysis estimates the minimum sample size needed to detect a meaningful effect confidently.

Balancing Resources and Rigor

It balances scientific rigor with efficient resource use by avoiding too small or excessively large samples.

Introduction to Power Analysis for Determining Minimum Sample Size

Power Analysis: Ensuring Reliable Results In Research



Key Components: Significance Threshold, Desired Power, Effect Size, Variance

Significance Threshold

Defines the probability of rejecting a true null hypothesis, usually set at 0.05 in studies.

Desired Power Level

Represents the likelihood of detecting an effect if it exists, commonly targeted at 80% or higher.

Effect Size

Measures the magnitude of the expected effect, impacting the required sample size for detection.

Data Variance

Describes the variability in the data, influencing the precision of estimates and sample size needs.

How Power Analysis Guides Proper Experimental Design

Determining Sample Size

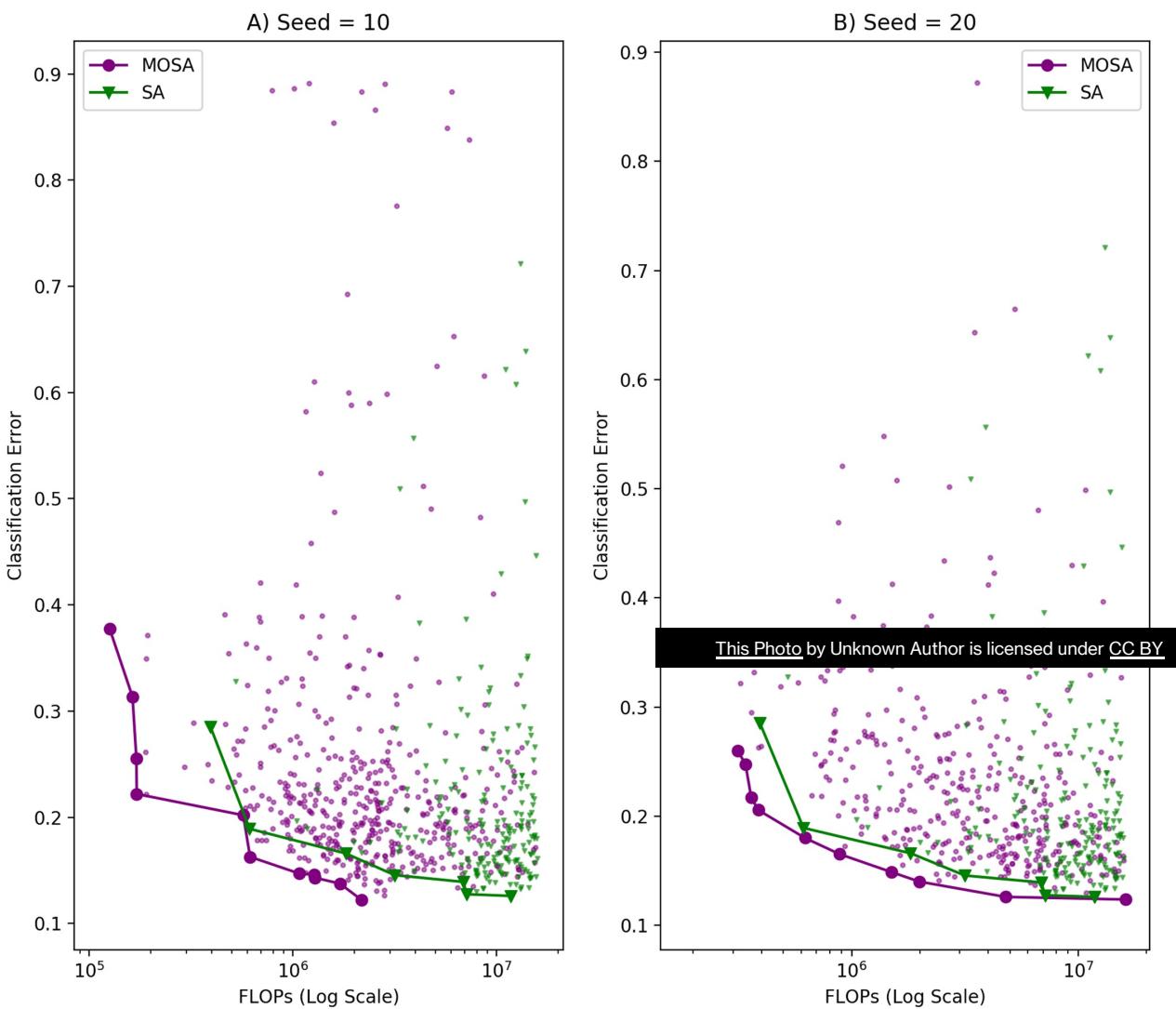
Power analysis helps determine the optimal sample size to detect meaningful effects with confidence.

Optimizing Study Parameters

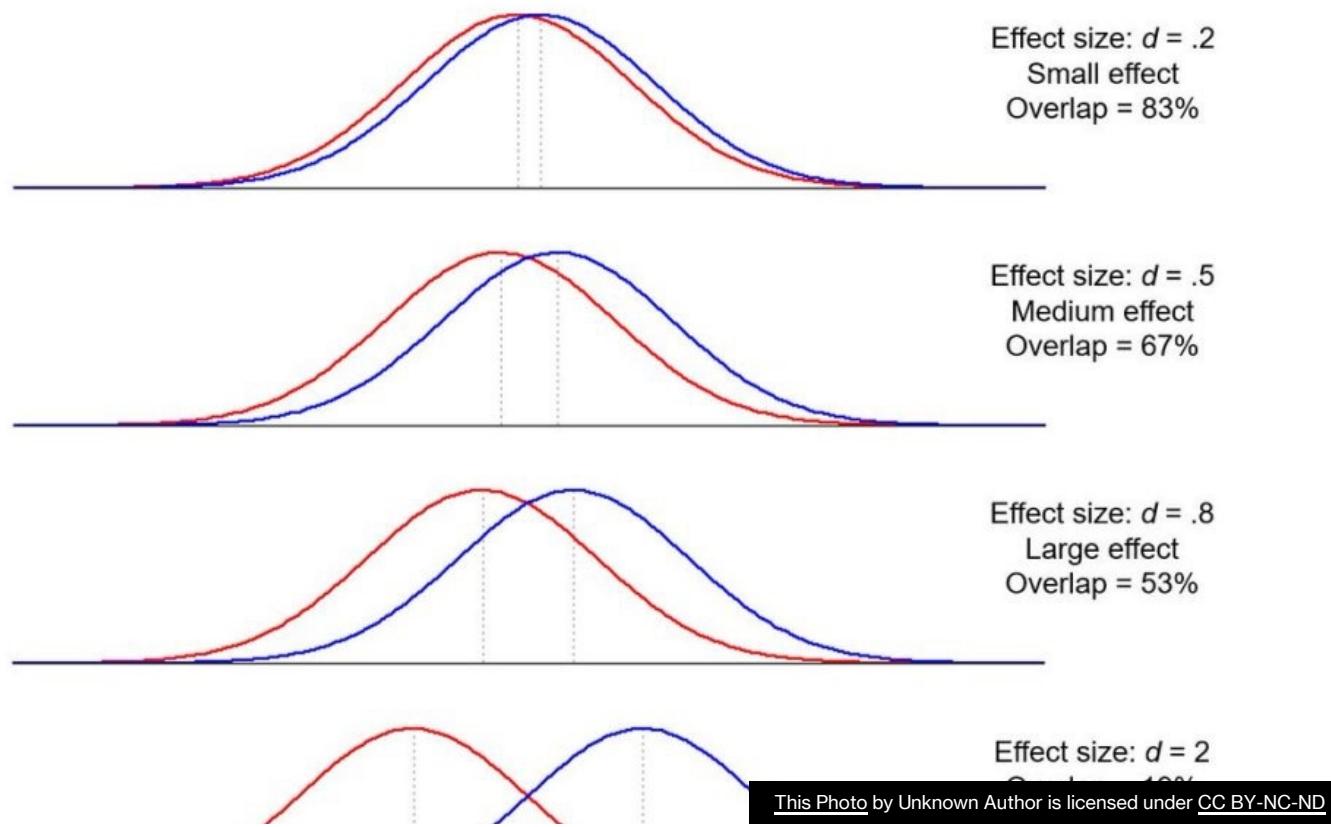
It guides the selection of study parameters to enhance experiment sensitivity and efficiency.

Resource Efficiency

Power analysis minimizes wasted resources by preventing underpowered or overpowered experiments.



Understanding Effect Sizes



Sample Size Importance

Sample size is crucial but insufficient by itself to ensure research validity and reliability.

Study Design Role

Robust study design significantly influences the accuracy and relevance of research findings.

Data Quality and Analysis

High-quality data and appropriate statistical methods are essential for valid and trustworthy results.

Limitations of Relying Solely on Sample Size

Lesson 2: Addressing The Multiple Testing Problem In Genetics

The Risks of Using a Standard 0.05 P-Value Threshold with Massive Datasets

High False Positive Rate

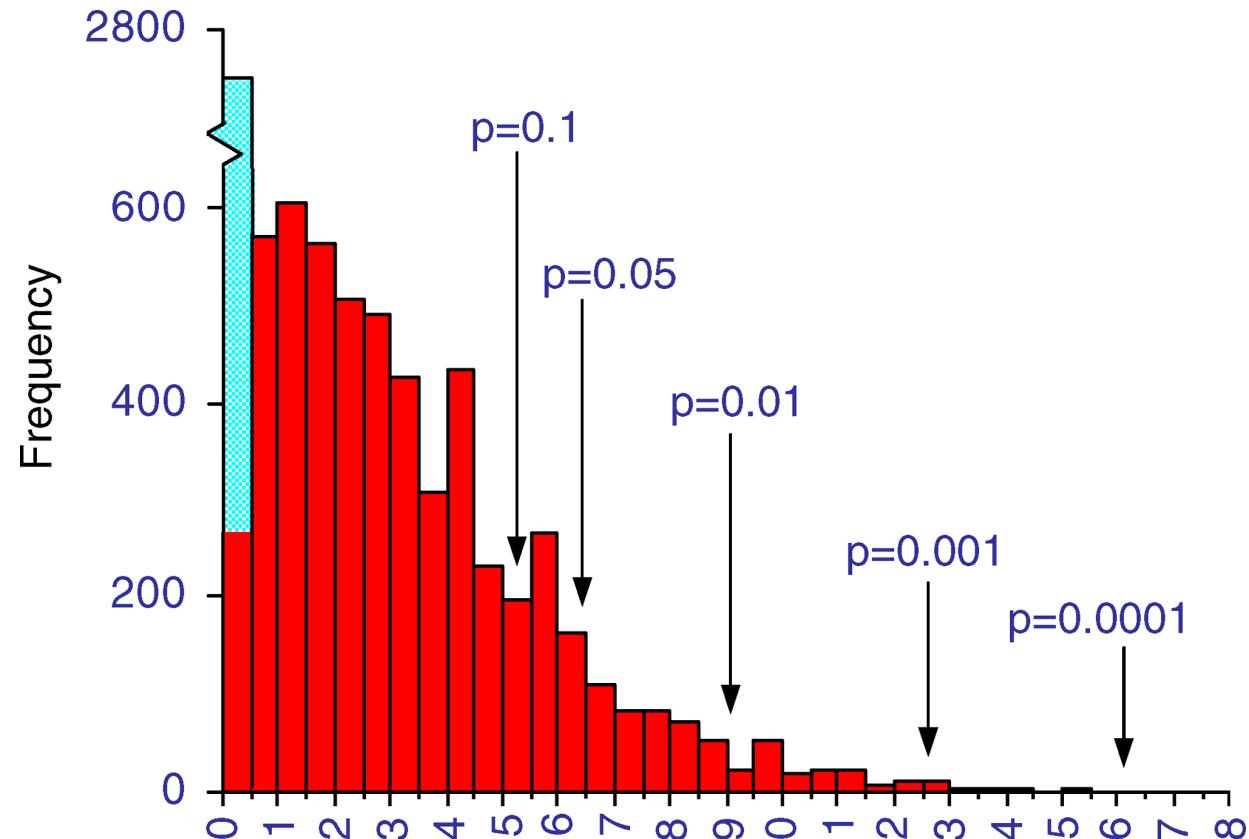
Using a 0.05 threshold with massive datasets increases false positives, causing unreliable findings.

Misleading Conclusions

False positives lead researchers to incorrect conclusions and flawed scientific claims.

Wasted Follow-up Efforts

Follow-up studies based on false positives waste time and resources in research.



LF This Photo by Unknown Author is licensed under CC BY-SA

Real-World Example: DNA Sequencing and Millions of Tests

High-Volume Statistical Testing

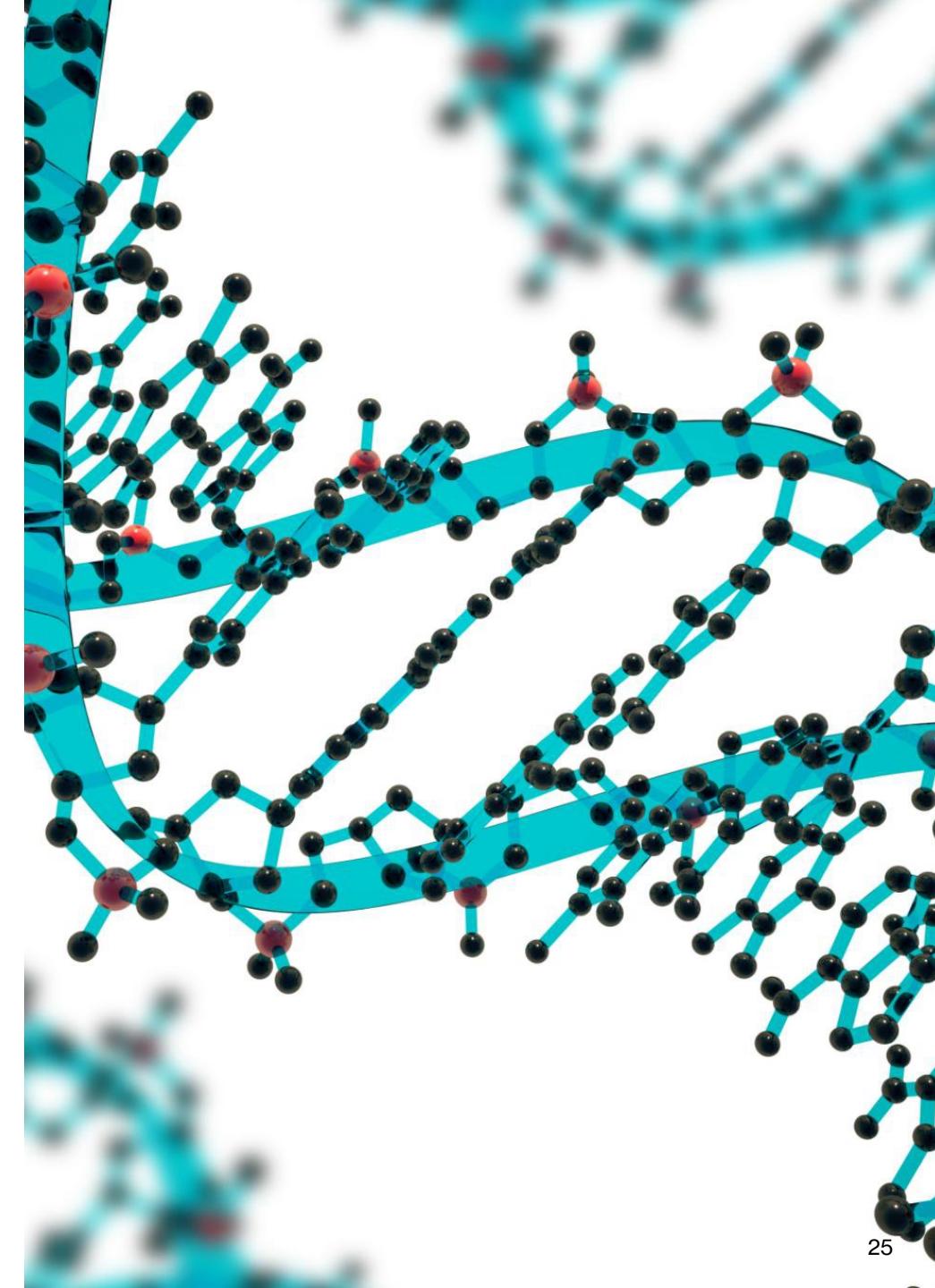
DNA sequencing involves conducting millions of statistical tests simultaneously to analyze genetic information.

Significance Criteria Importance

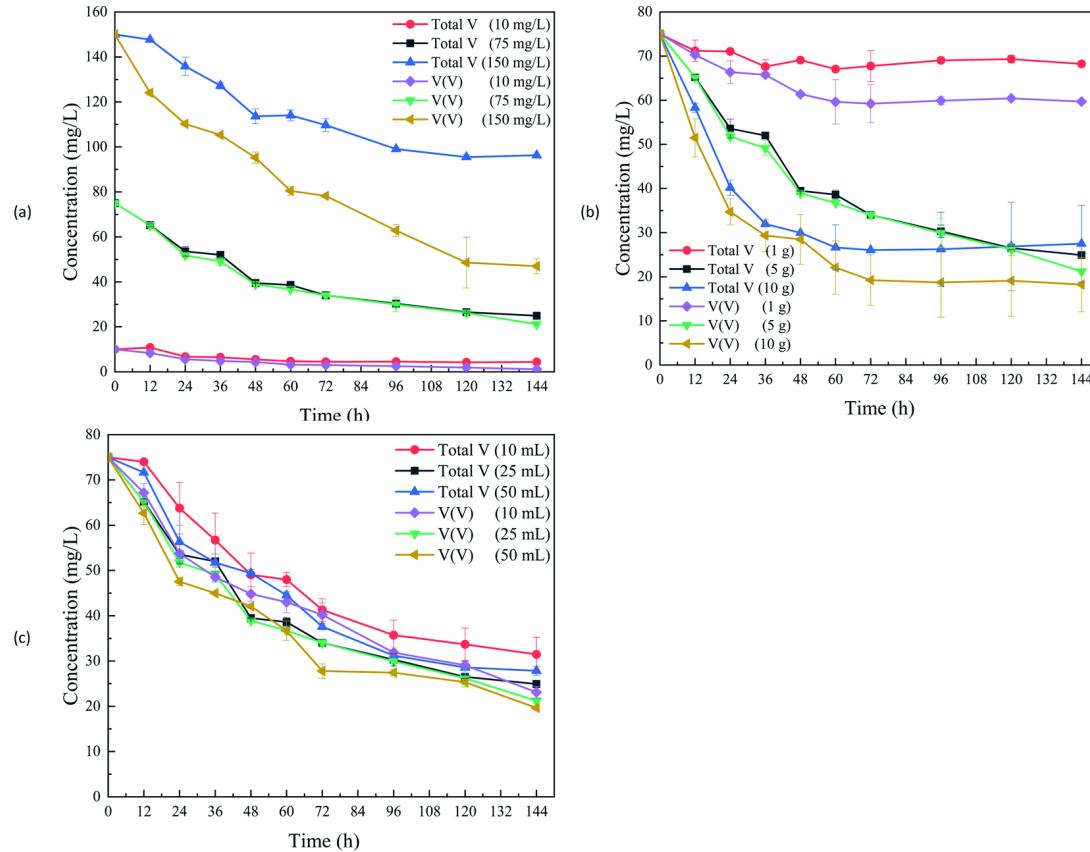
Stringent significance thresholds ensure the validity and reliability of results in large-scale DNA tests.

Ensuring Result Validity

Applying stricter criteria helps avoid false positives and maintains integrity in DNA sequencing studies.



Maximizing Study Efficiency



Set Clear Goals

Establishing specific objectives guides your study sessions and keeps you focused on what matters most.

Structure and Minimize Distractions

A well-organized schedule and a quiet study area significantly boost concentration and productivity.

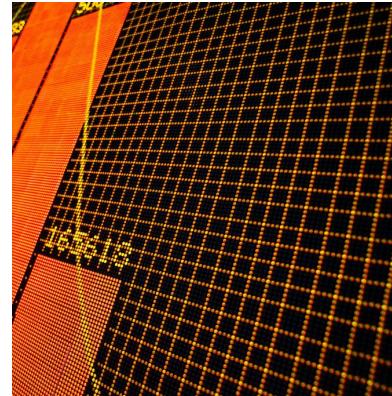
Optimize Timing and Environment

Adjusting study length, break intervals, and the environment can improve information retention and recall.

Evaluate and Refine Methods

Consistently reviewing and improving strategies ensures progress and keeps motivation high during learning.

The Prevalence and Impact of False Positives in Large-Scale Studies



		Disease present	Disease absent
Test result	True	A	B
	False	C	D
		A+C	B+D
Sensitivity = A/[A+C]			
Specificity = D/[B+D]			
PPV = A/[A+B]			
NPV = D/[C+D]			

Distortion of Scientific Understanding

False positives can mislead researchers, causing incorrect conclusions and flawed scientific theories.

Reduced Reproducibility

Studies affected by false positives have lower reproducibility, impacting the reliability of scientific findings.

Costly Misdirected Research

False positives lead to wasted resources and effort on invalid research directions, especially in genetics.

Genome-Wide Significance: A More Stringent Approach To Statistical Testing

Definition and Rationale for $P < 0.00000005$ in Genetics



Genome-wide Significance Threshold

The p-value threshold of 5×10^{-8} is used to determine genome-wide significance in genetic studies.

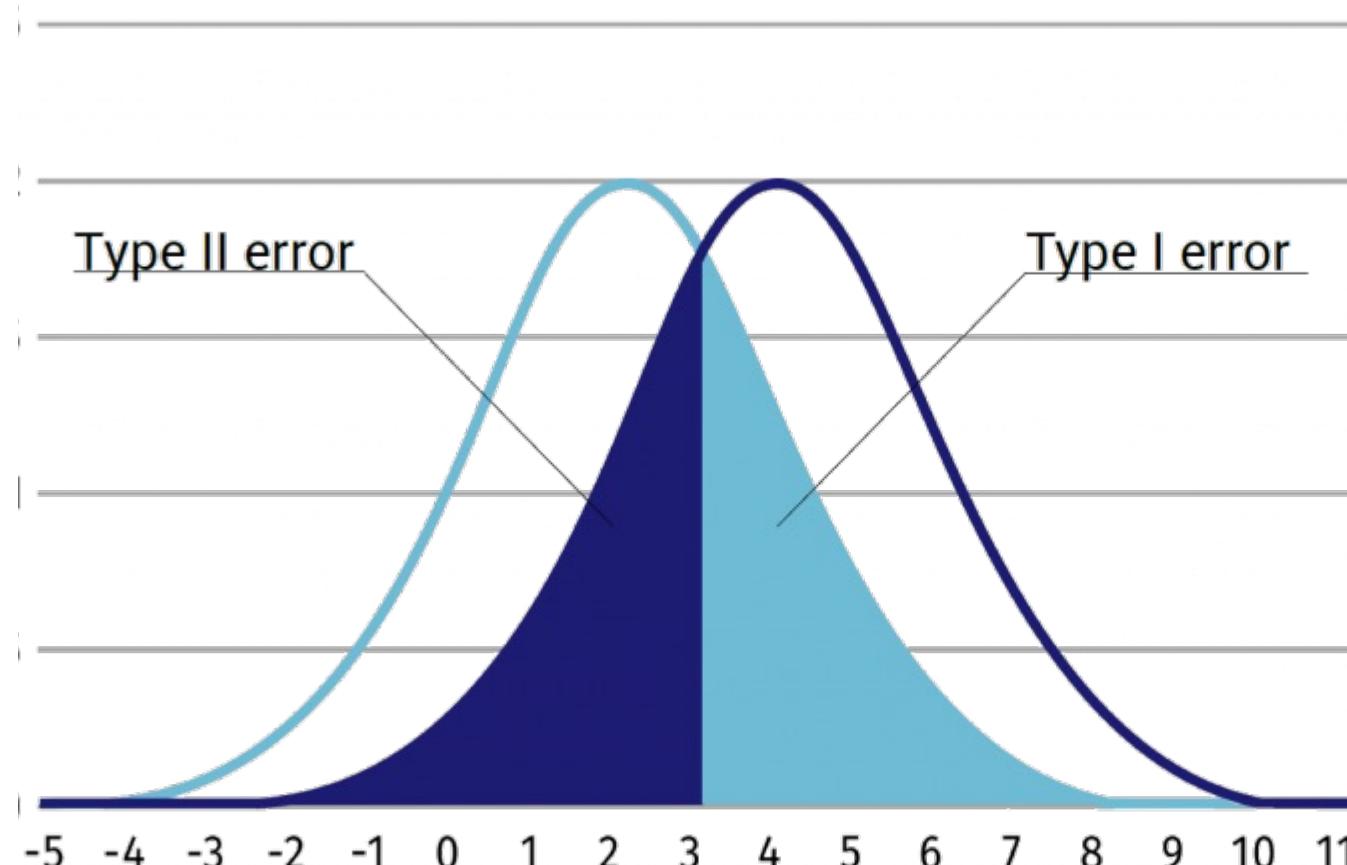
Multiple Testing Correction

This threshold accounts for multiple comparisons to reduce false positive discoveries in genetics.

Balancing Discovery and Reliability

Setting a stringent p-value ensures findings are reliable while allowing meaningful discoveries.

How Stricter Thresholds Reduce False Discoveries



This Photo by Unknown Author is licensed under CC BY-SA-NC

Minimizing False Positives

Stricter p-value thresholds reduce the chance of false positives in research findings, improving overall accuracy.

Improving Reproducibility

Lower p-value cutoffs help ensure findings are more likely to be true and reproducible across studies.

Implications for Research Integrity and Reproducibility



Stringent Significance Criteria

Applying strict significance thresholds enhances research quality and reduces false positives.

Fostering Trust

Higher research standards increase confidence in scientific findings among researchers and the public.

Meaningful Scientific Advances

Improved research integrity enables breakthroughs that drive scientific progress forward.

Conclusion

Beyond Universal Thresholds

Statistical significance is not solely about applying a fixed threshold but requires deeper contextual understanding.

Importance of Sample Size

Sample size critically affects the validity and reliability of statistical results in research.

Challenges in Multiple Testing

Multiple testing increases complexity, especially in genetics, requiring advanced statistical approaches.