Statistics of RNA-seq analysis

Zeynep Kalender-Atak

Source: Dominique Laurent Couturier, CRUK-CI

```
> dds <- DESeqDataSetFromMatrix(cnts, DataFrame(cond), ~ cond)
> dds <- DESeq(dds)
> results(dds)

log2 fold change (MLE): cond 2 vs 1
Wald test p-value: cond 2 vs 1
DataFrame with 1000 rows and 6 columns
      baseMean log2FoldChange     lfcSE      stat    pvalue      padj
     <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
1      97.3140      -0.682067  0.344525 -1.979730 0.0477339  0.745842
2     109.9860      -0.228819  0.450720 -0.507676 0.6116808  0.944354
3      98.8111       0.104291  0.462113  0.225683 0.8214483  0.978382
4     103.2615       0.306400  0.297682  1.029284 0.3033460  0.944354
5      97.9406       0.316338  0.357242  0.885501 0.3758864  0.944354
...        ...            ...       ...       ...       ...       ...
996    86.8057      0.0467703  0.287042  0.162939 0.8705668  0.980044
997   101.4437     -0.2070806  0.339886 -0.609264 0.5423495  0.944354
998    78.1356     -0.6372790  0.369515 -1.724637 0.0845930  0.824310
999    89.2920      0.7554725  0.306192  2.467314 0.0136131  0.614613
1000  103.5569     -0.0728875  0.348655 -0.209053 0.8344065  0.978382
```

Outline

- Statistical Concepts - Bite size statistics
  - PPDAC Cycle
  - Hypothesis testing
  - Type I and II errors
  - Power Analysis
- Statistical aspects of bulk RNA-seq analysis
  - Generalized Linear Models
  - Negative Binomial Regression
  - Multiple Comparisons

Image credit: The Pennsylvania State University

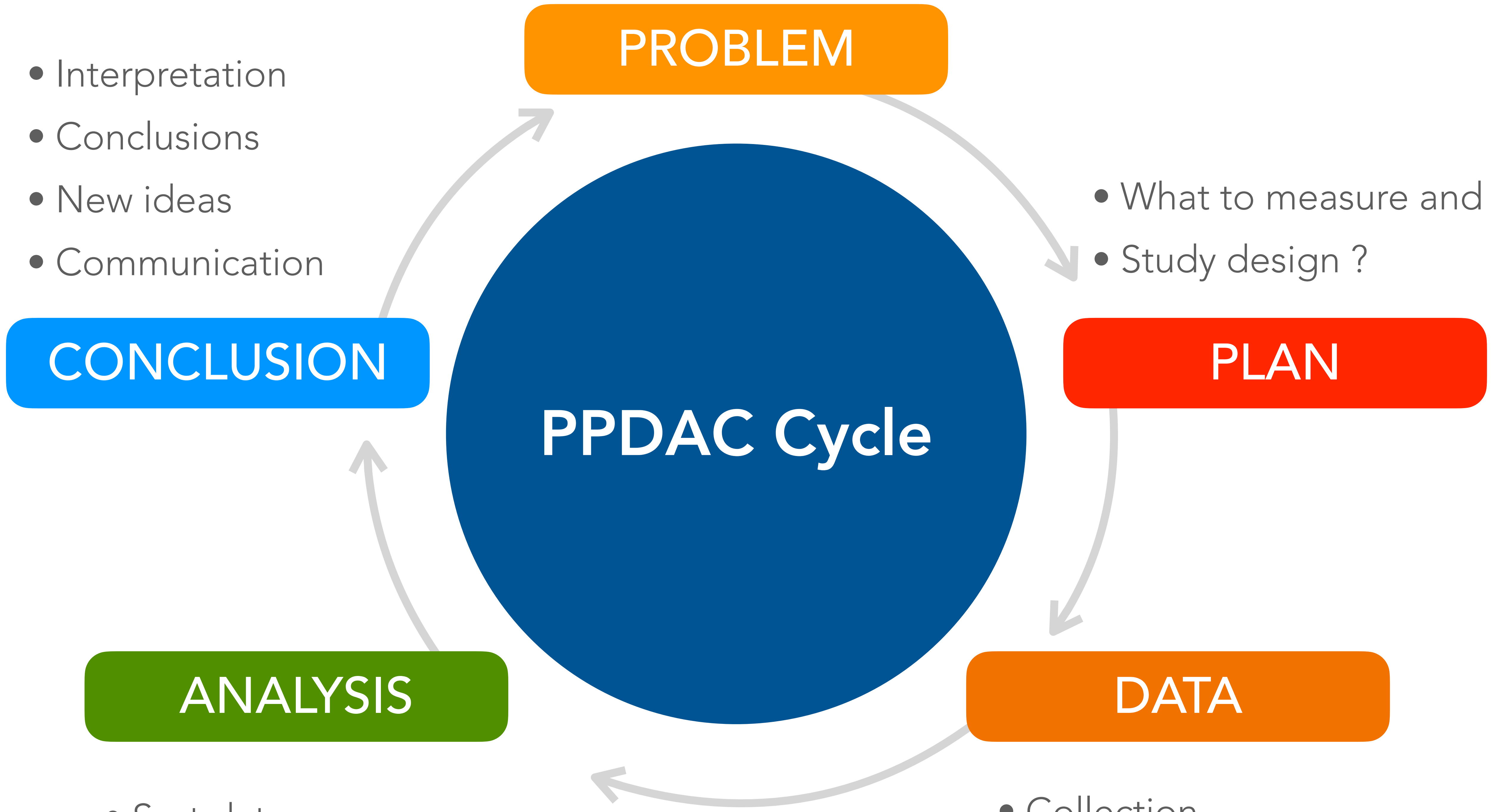
Image credit: The Pennsylvania State University

Data literacy

The ability to not only carry out statistical analysis on real-world problems, but also to understand and critique any conclusions drawn by others on the basis of statistics.

Statistics as an investigative process of problem-solving and decision-making

- Understanding and defining the problem.
- How do we go about answering this question?



- What to measure and how?
- Study design ?

- Collection
- Management
- Cleaning

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation

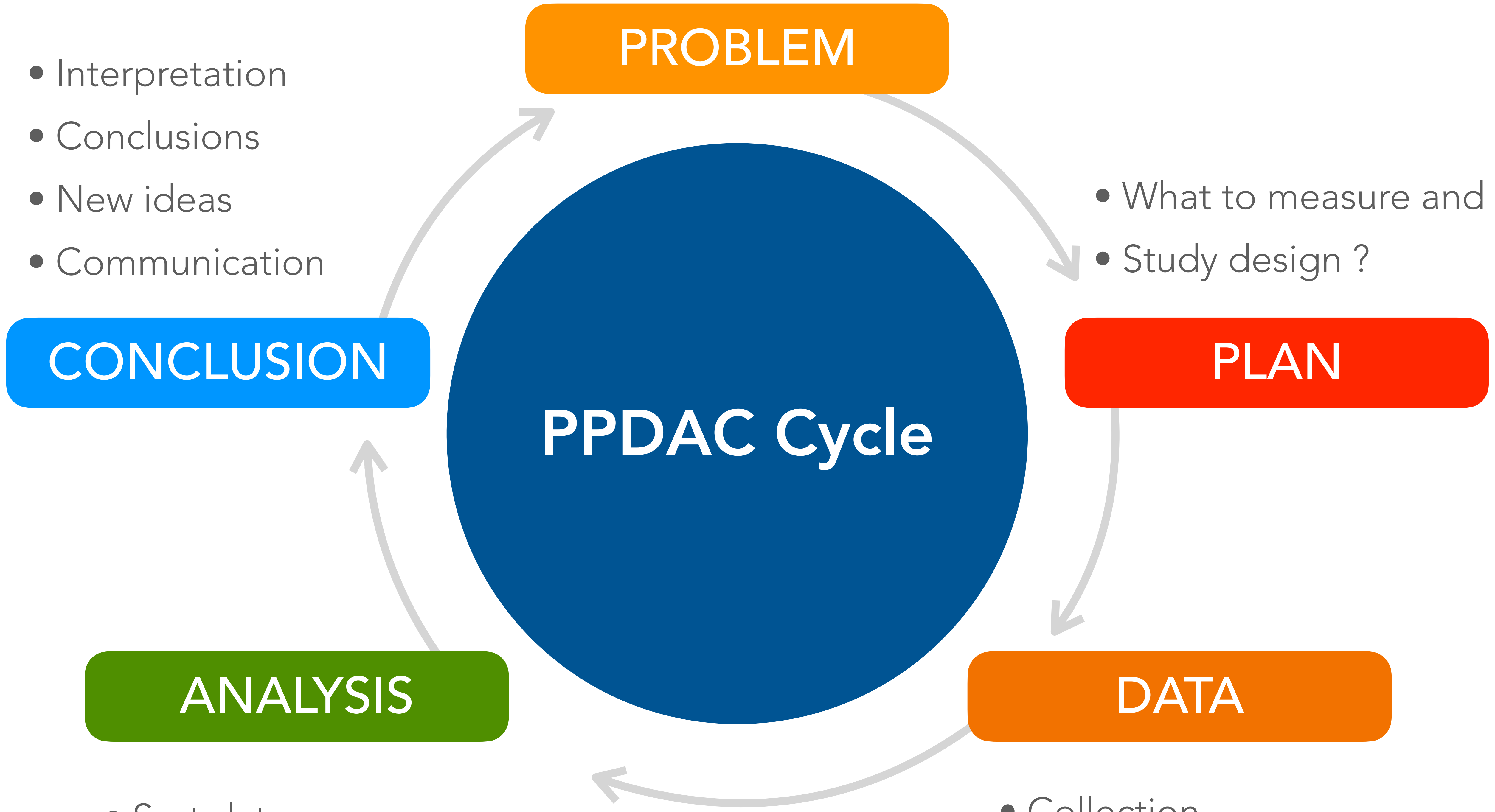
- Interpretation
- Conclusions
- New ideas
- Communication

R. J. MacKay. R. W. Oldford. "Scientific Method, Statistical Method and the Speed of Light." *Statist. Sci.* 15 (3) 254 - 278, August 2000.

Statistics as an investigative process of problem-solving and decision-making

Toxoplasma gondii infection causes a host of severe neurological disorders. Our understanding of the molecular mechanisms associated with infection is incomplete.

- Understanding and defining the problem.
- How do we go about answering this question?



- What to measure and how?
- Study design ?

- Collection
- Management
- Cleaning

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation

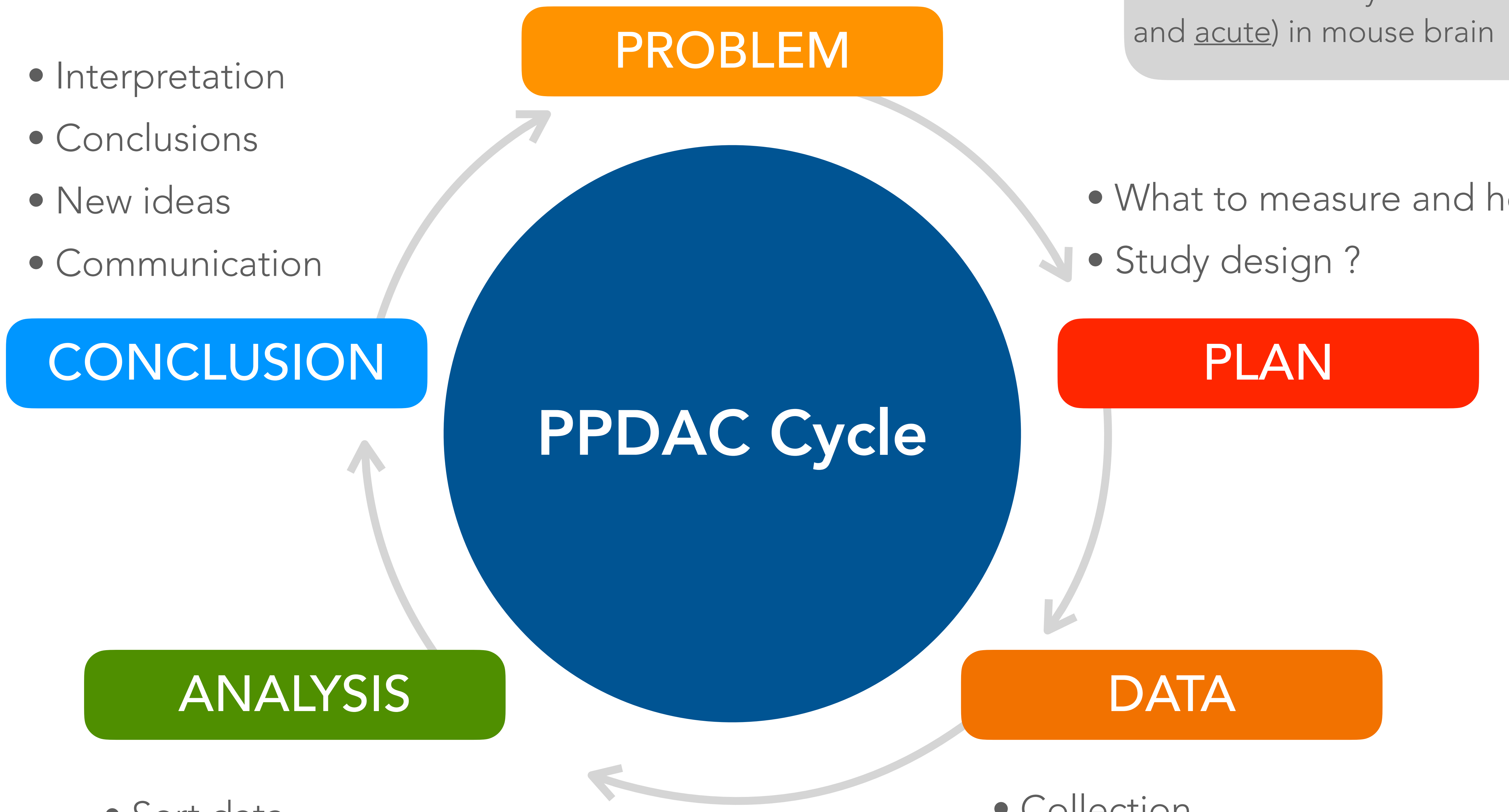
- Interpretation
- Conclusions
- New ideas
- Communication

Hu et al. Profiling of Mouse Brain During Acute and Chronic Infections by Toxoplasma gondii Oocysts. Front. Microbiol. 2020

Statistics as an investigative process of problem-solving and decision-making

Toxoplasma gondii infection causes a host of severe neurological disorders. Our understanding of the molecular mechanisms associated with infection is incomplete.

We want to study the effect of Toxoplasma gondii infection (chronic and acute) in mouse brain



PROBLEM
- Understanding and defining the problem.
- How do we go about answering this question?

PLAN
- What to measure and how?
- Study design ?

DATA
- Collection
- Management
- Cleaning

ANALYSIS
- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation

CONCLUSION
- Interpretation
- Conclusions
- New ideas
- Communication

Hu et al. Profiling of Mouse Brain During Acute and Chronic Infections by Toxoplasma gondii Oocysts. Front. Microbiol. 2020

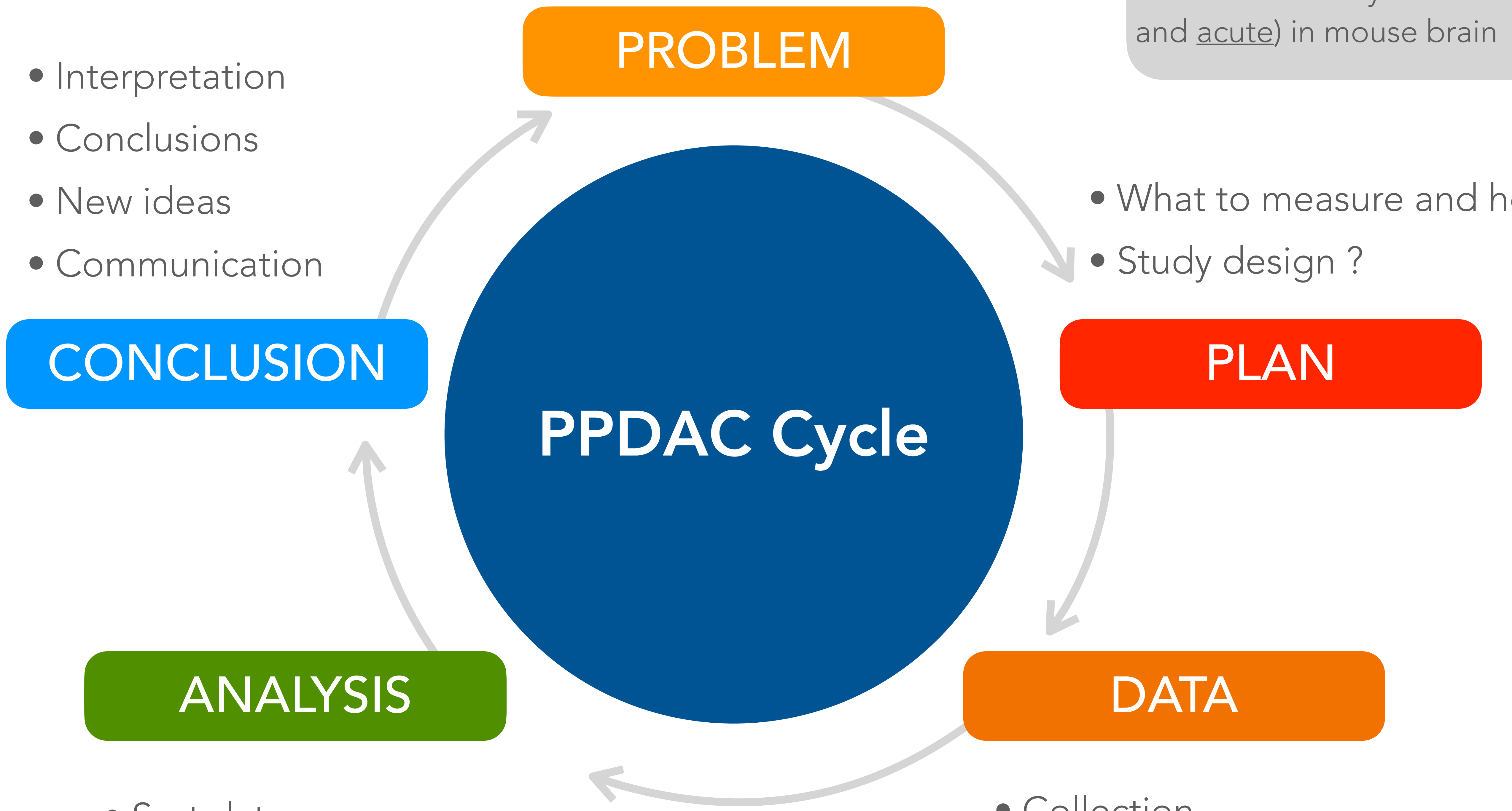
Statistics as an investigative process of problem-solving and decision-making

- Understanding and defining the problem.
- How do we go about answering this question?

Toxoplasma gondii infection causes a host of severe neurological disorders. Our understanding of the molecular mechanisms associated with infection is incomplete.

We want to study the effect of Toxoplasma gondii infection (chronic and acute) in mouse brain

- Interpretation
- Conclusions
- New ideas
- Communication



- What to measure and how?
- Study design ?

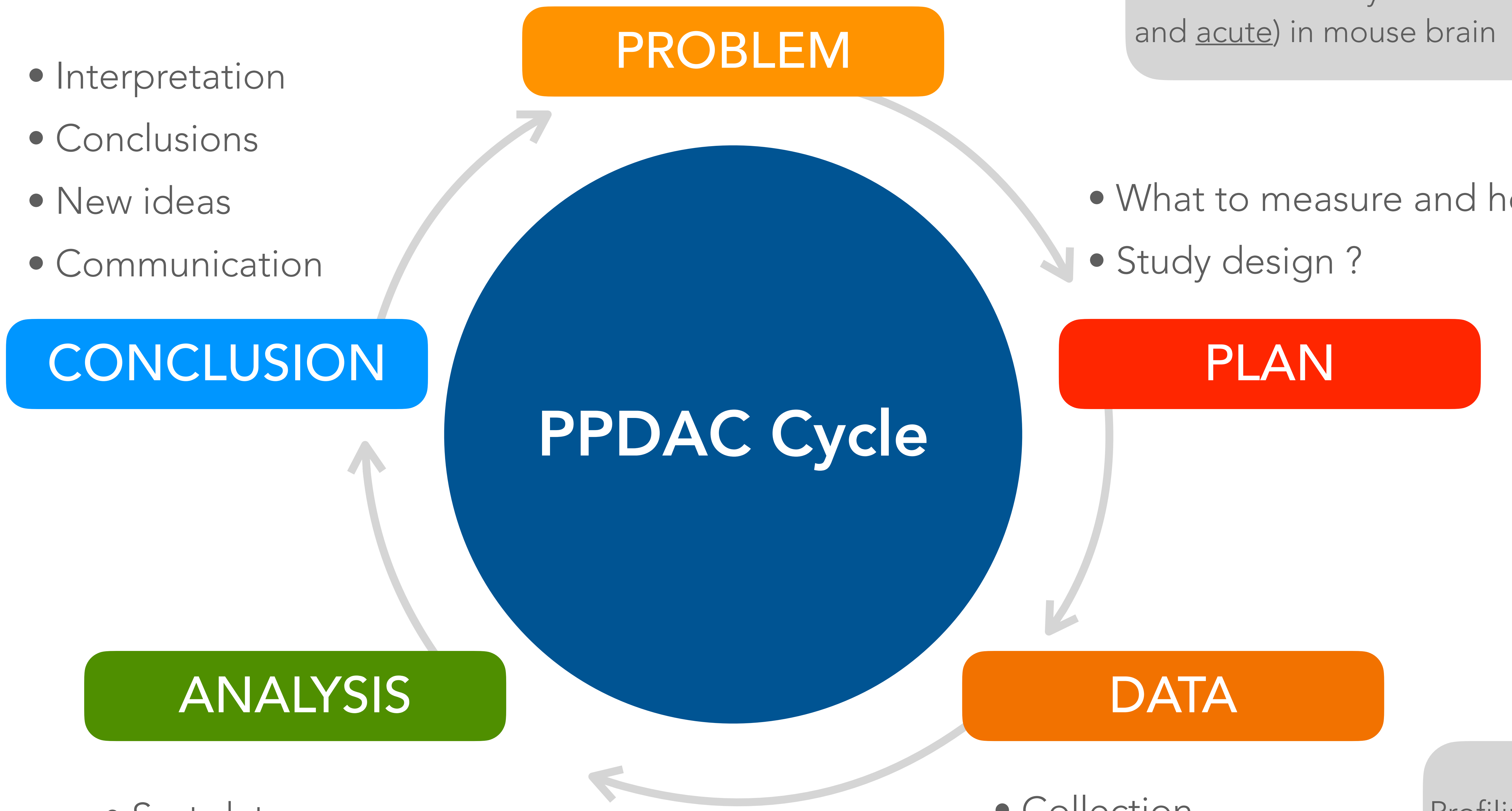
- Total gene expression profile of the brain in infection versus no-infection
- A two-factor study with three biological replicates in each group with matched controls

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation

- Collection
- Management
- Cleaning

Hu et al. Profiling of Mouse Brain During Acute and Chronic Infections by Toxoplasma gondii Oocysts. Front. Microbiol. 2020

Statistics as an investigative process of problem-solving and decision-making



PROBLEM

- Understanding and defining the problem.
- How do we go about answering this question?

Toxoplasma gondii infection causes a host of severe neurological disorders. Our understanding of the molecular mechanisms associated with infection is incomplete.

We want to study the effect of Toxoplasma gondii infection (chronic and acute) in mouse brain

PLAN

- What to measure and how?
- Study design ?

- Total gene expression profile of the brain in infection versus no-infection
- A two-factor study with three biological replicates in each group with matched controls

DATA

- Collection
- Management
- Cleaning

Profiling the total transcriptome with RNA-seq
Preprocessing and quality control

ANALYSIS

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation

CONCLUSION

- Interpretation
- Conclusions
- New ideas
- Communication

Hu et al. Profiling of Mouse Brain During Acute and Chronic Infections by Toxoplasma gondii Oocysts. Front. Microbiol. 2020

Statistics as an investigative process of problem-solving and decision-making

PPDAC Cycle

PROBLEM

- Understanding and defining the problem.
- How do we go about answering this question?

Toxoplasma gondii infection causes a host of severe neurological disorders. Our understanding of the molecular mechanisms associated with infection is incomplete.

We want to study the effect of Toxoplasma gondii infection (<u>chronic</u> and <u>acute</u>) in mouse brain

PLAN

- What to measure and how?
- Study design ?

- Total gene expression profile of the brain in infection versus no-infection
- A two-factor study with three biological replicates in each group with matched controls

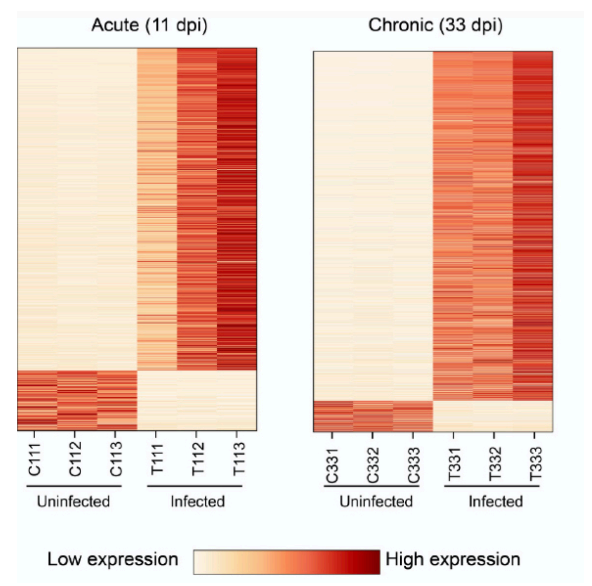
DATA

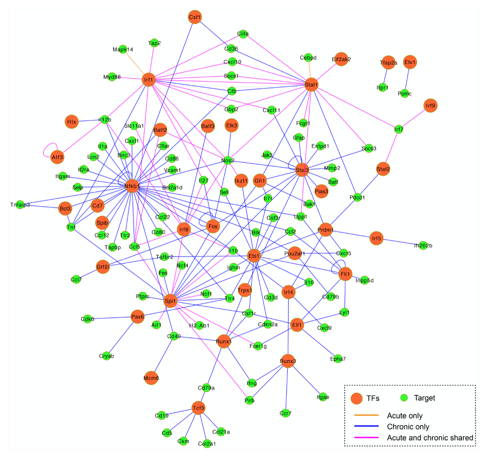
- Collection
- Management
- Cleaning

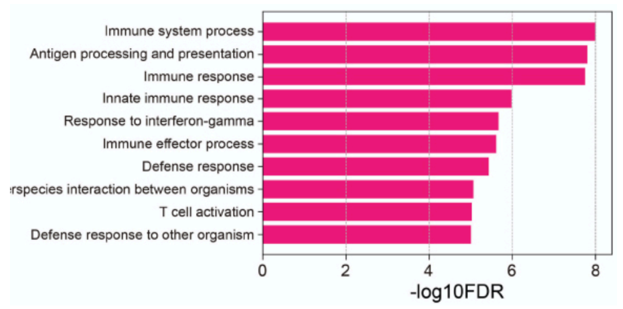
Profiling the total transcriptome with RNA-seq
Preprocessing and quality control

ANALYSIS

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation





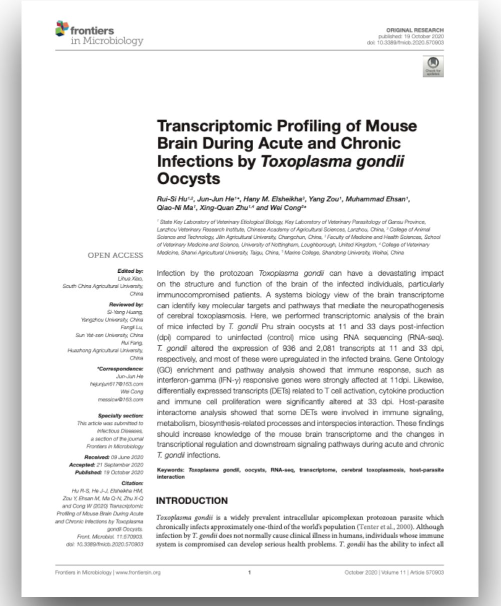


CONCLUSION

- Interpretation
- Conclusions
- New ideas
- Communication

Hu et al. Profiling of Mouse Brain During Acute and Chronic Infections by Toxoplasma gondii Oocysts. Front. Microbiol. 2020

Statistics as an investigative process of problem-solving and decision-making



PPDAC Cycle

PROBLEM

- Understanding and defining the problem.
- How do we go about answering this question?

Toxoplasma gondii infection causes a host of severe neurological disorders. Our understanding of the molecular mechanisms associated with infection is incomplete.

We want to study the effect of Toxoplasma gondii infection (chronic and acute) in mouse brain

PLAN

- What to measure and how?
- Study design ?

- Total gene expression profile of the brain in infection versus no-infection
- A two-factor study with three biological replicates in each group with matched controls

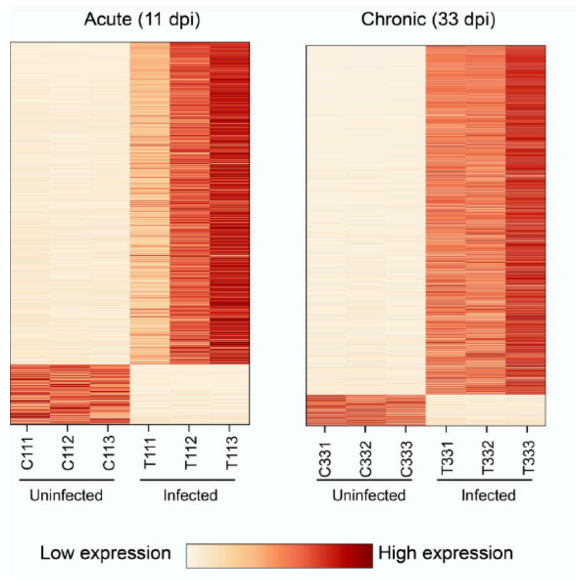
DATA

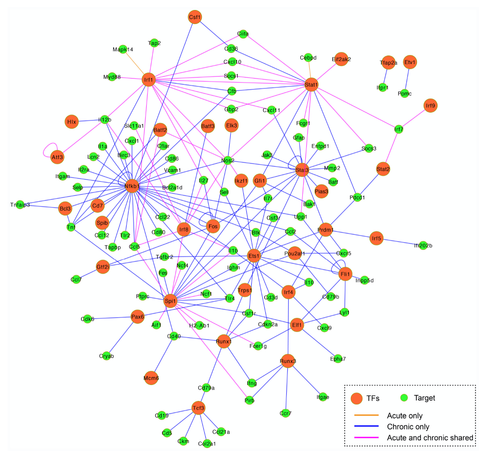
- Collection
- Management
- Cleaning

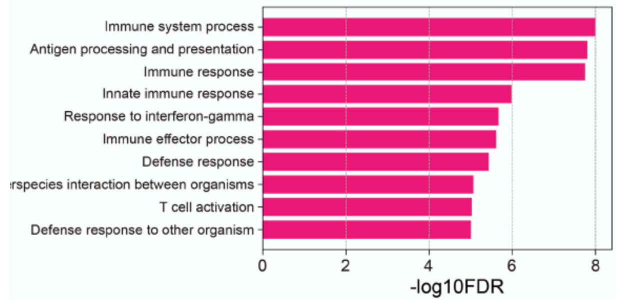
Profiling the total transcriptome with RNA-seq
Preprocessing and quality control

ANALYSIS

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation







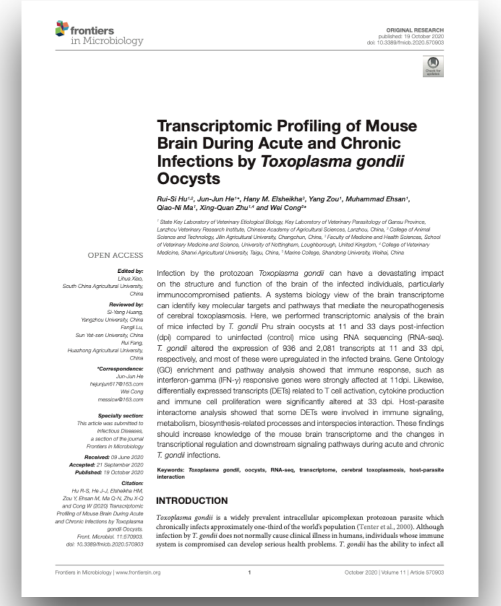
CONCLUSION

- Interpretation
- Conclusions
- New ideas
- Communication

- IFN-γ response increases as infection progresses
- Calcium response pathways are downregulated

Hu et al. Profiling of Mouse Brain During Acute and Chronic Infections by Toxoplasma gondii Oocysts. Front. Microbiol. 2020

Statistics as an investigative process of problem-solving and decision-making


Transcriptomic Profiling of Mouse Brain During Acute and Chronic Infections by *Toxoplasma gondii* Oocysts

PROBLEM

- Understanding and defining the problem.
- How do we go about answering this question?

Toxoplasma gondii infection causes a host of severe neurological disorders. Our understanding of the molecular mechanisms associated with infection is incomplete.

We want to study the effect of Toxoplasma gondii infection (chronic and acute) in mouse brain

PLAN

- What to measure and how?
- Study design ?

- Total gene expression profile of the brain in infection versus no-infection
- A two-factor study with three biological replicates in each group with matched controls

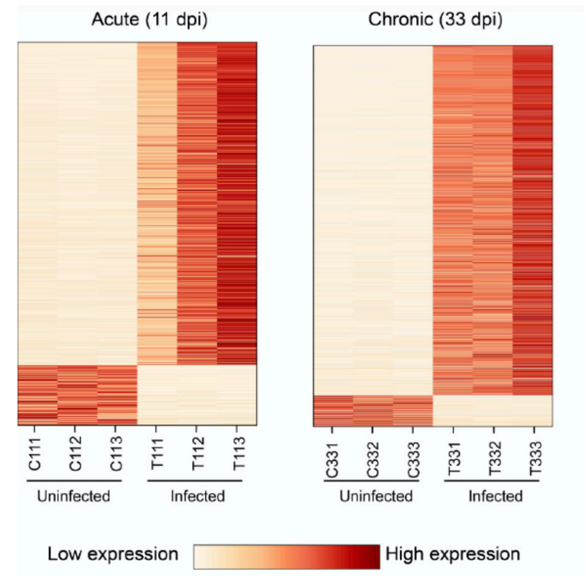
DATA

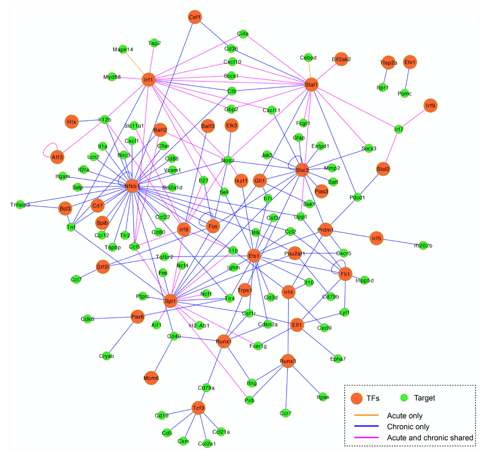
- Collection
- Management
- Cleaning

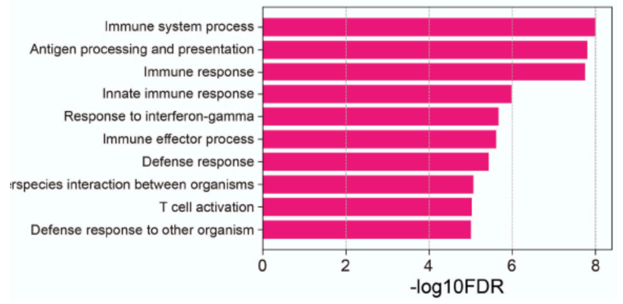
Profiling the total transcriptome with RNA-seq
Preprocessing and quality control

ANALYSIS

- Sort data
- Construct table, graphs
- Look for patterns
- Hypothesis generation







CONCLUSION

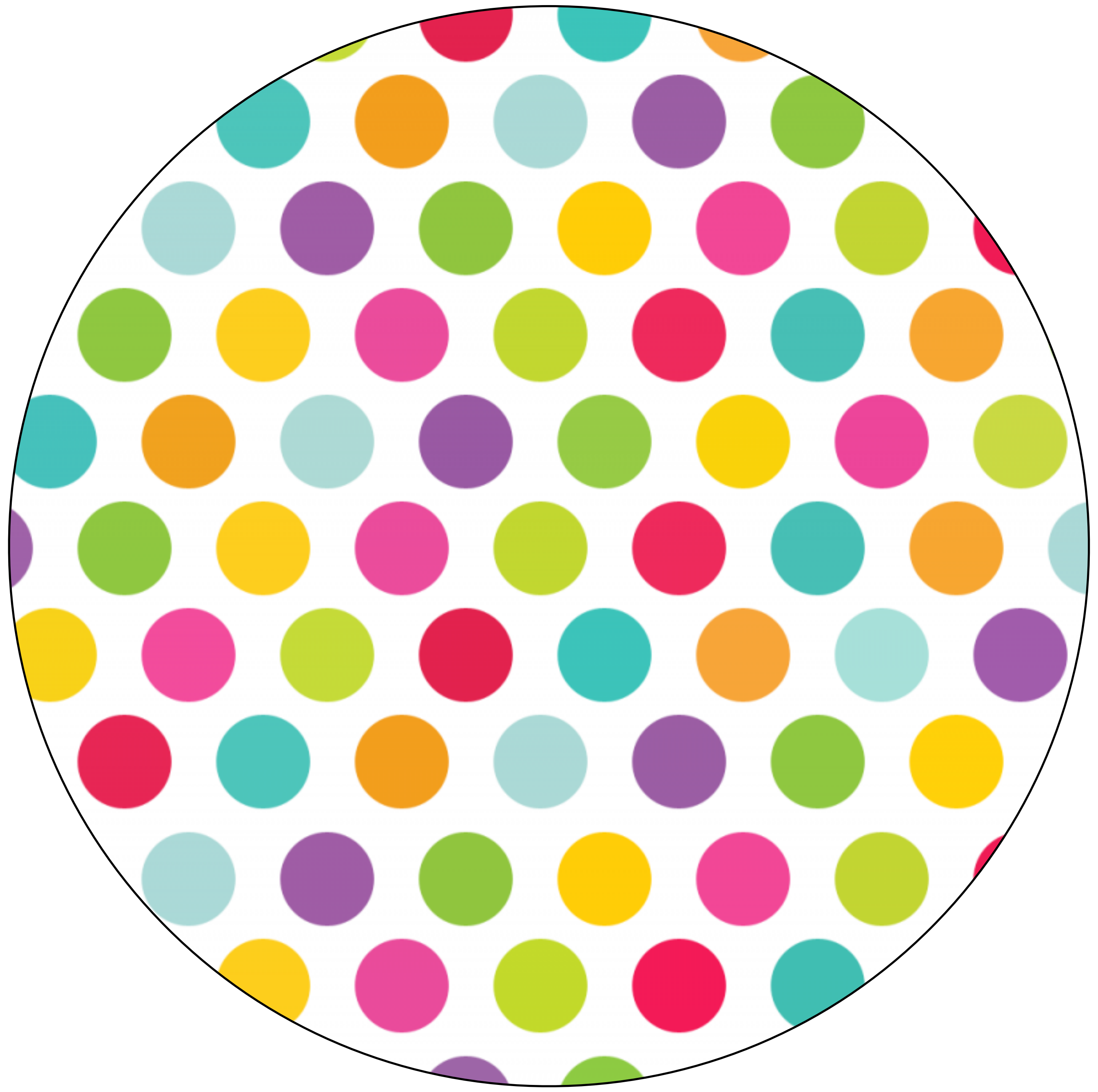
- Interpretation
- Conclusions
- New ideas
- Communication

- IFN-γ response increases as infection progresses
- Calcium response pathways are downregulated

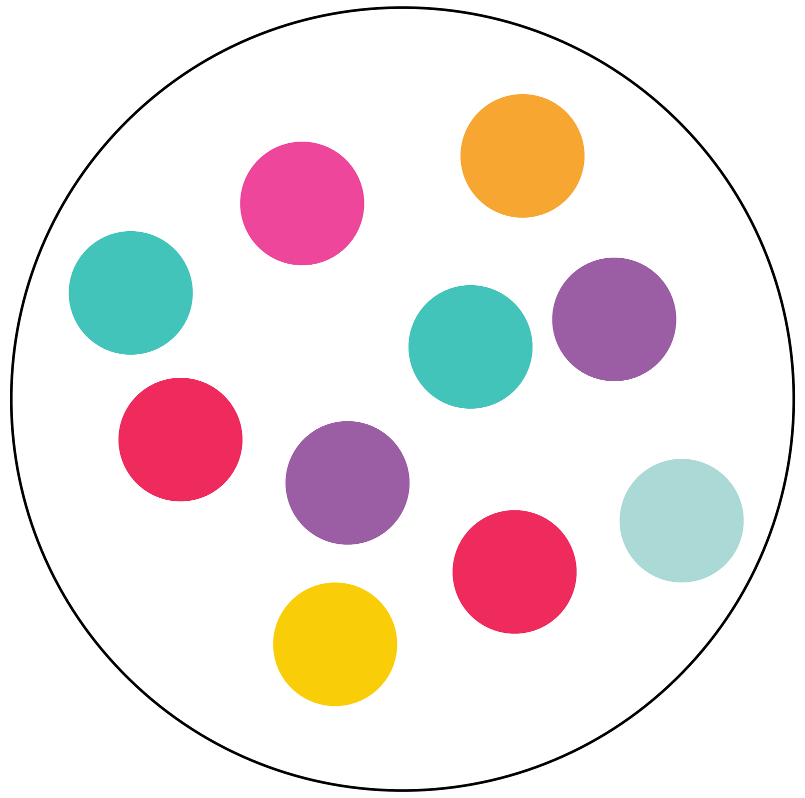
PPDAC Cycle

Hu et al. Profiling of Mouse Brain During Acute and Chronic Infections by Toxoplasma gondii Oocysts. Front. Microbiol. 2020

Basics on inferential statistics and hypothesis testing



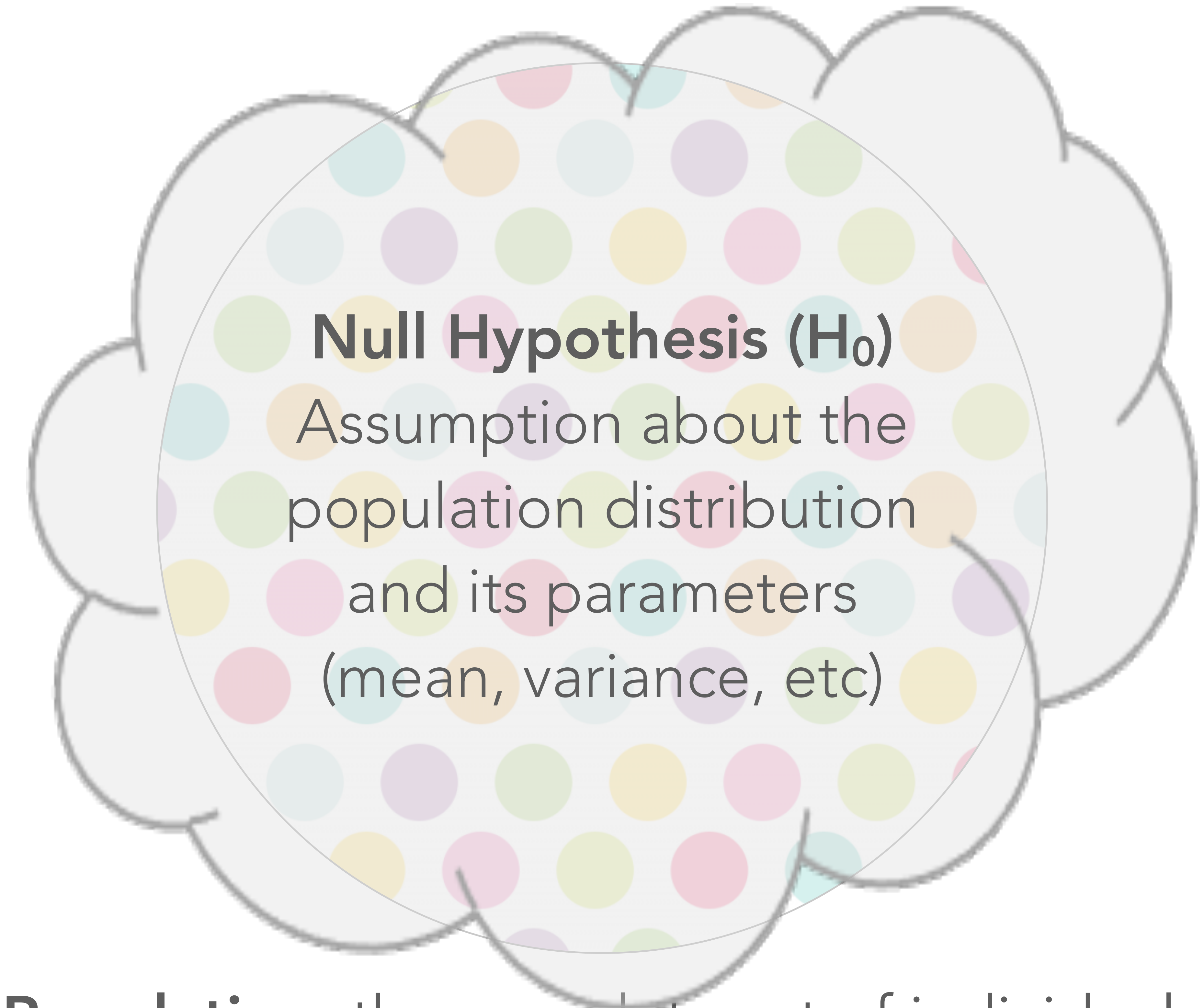
Population: the complete set of individuals that we are interested in



Sample: smaller set of individuals that is representative of the population

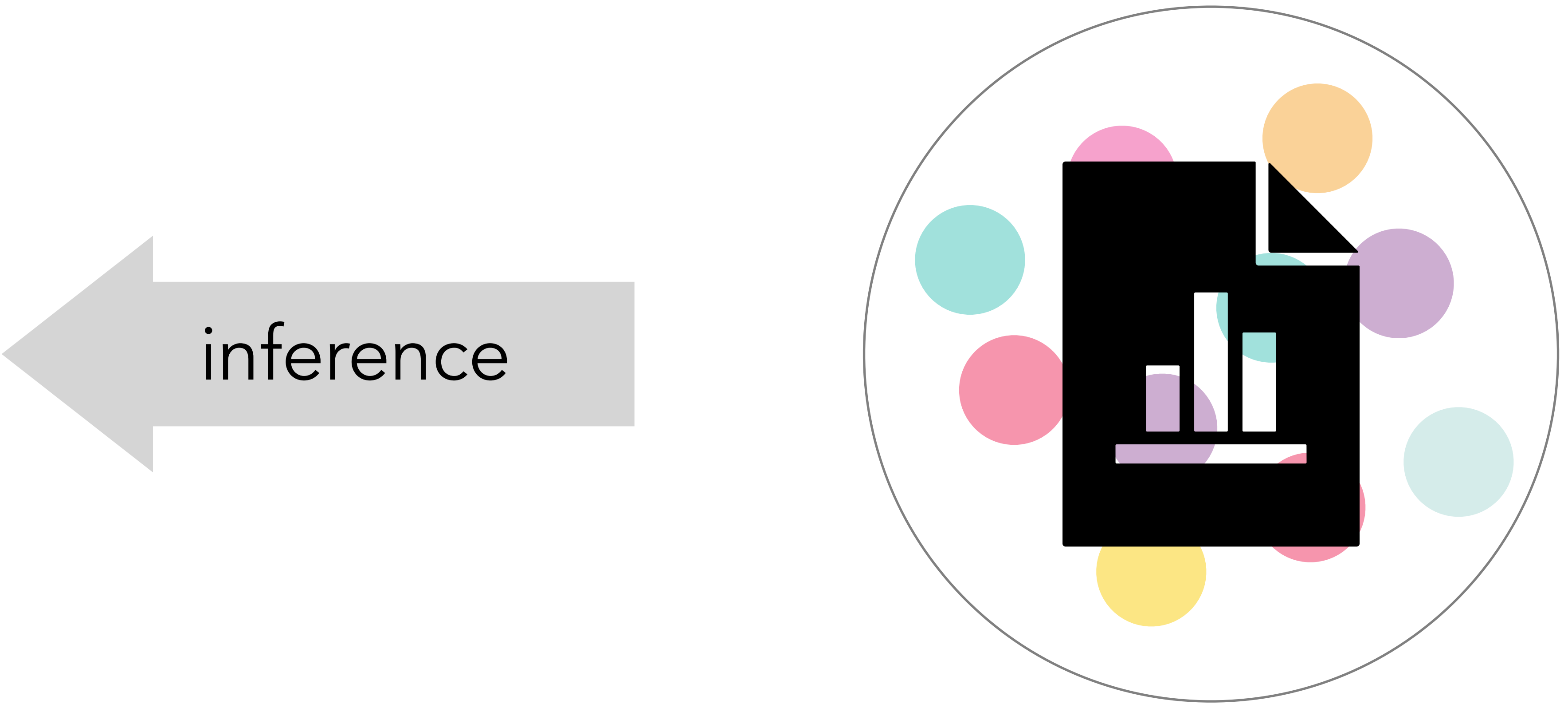
Variable: what we are interested in measuring

Basics on inferential statistics and hypothesis testing



Inference means two things:

1. Estimating population parameters
2. Testing hypothesis regarding the population distribution



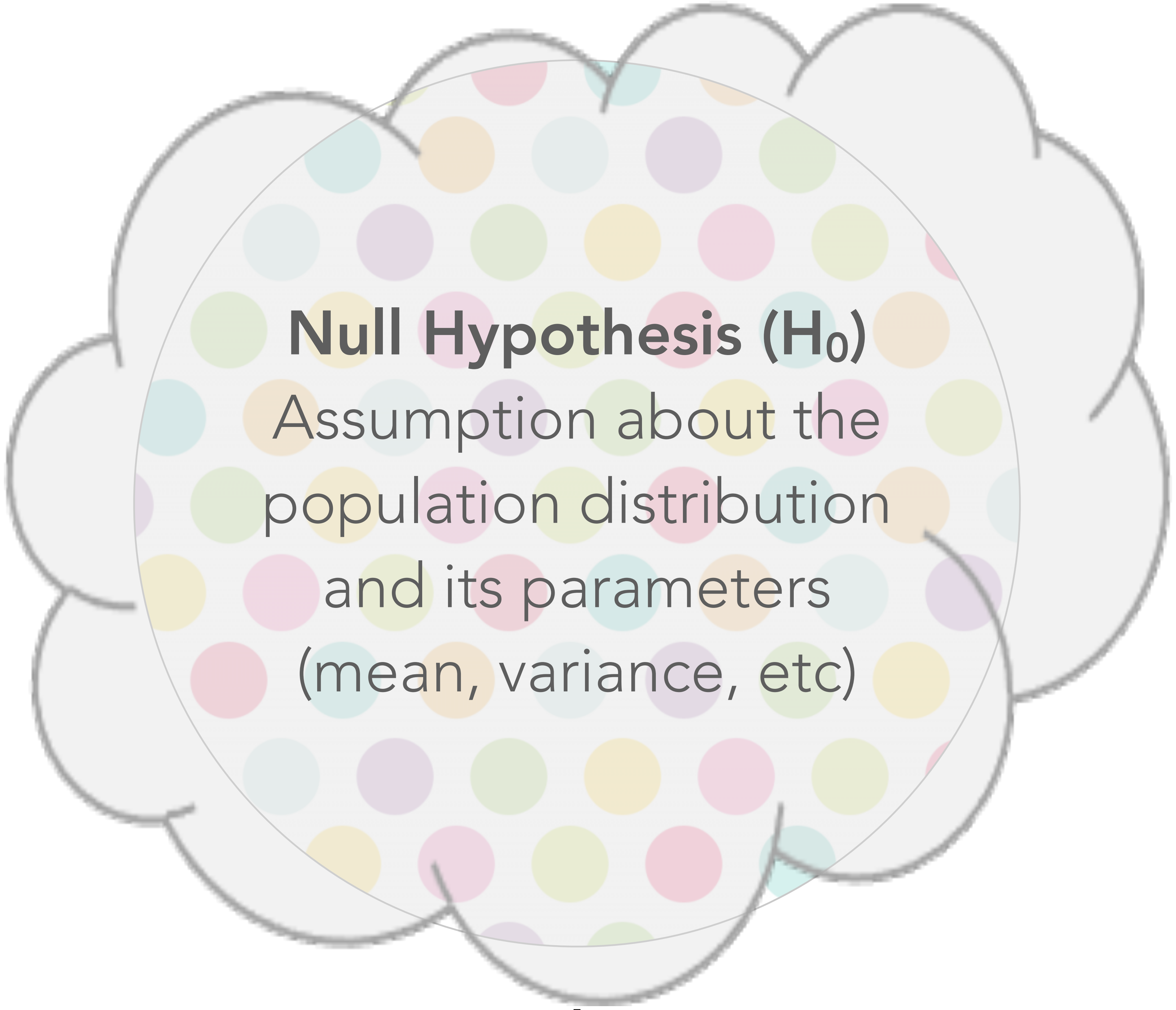
Population: the complete set of individuals that we are interested in

Sample: smaller set of individuals that is representative of the population

Variable: what we are interested in measuring

Slide content adapted from Matt Castle

Basics on inferential statistics and hypothesis testing



Choose test

Based on data and H_0

↓

Calculate test statistic, interpret results

Slide content adapted from Matt Castle

A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

Example adapted from Sal Khan

A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: Drug has no effect on response time
H_1: Drug has an effect on response time

A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Example adapted from Sal Khan

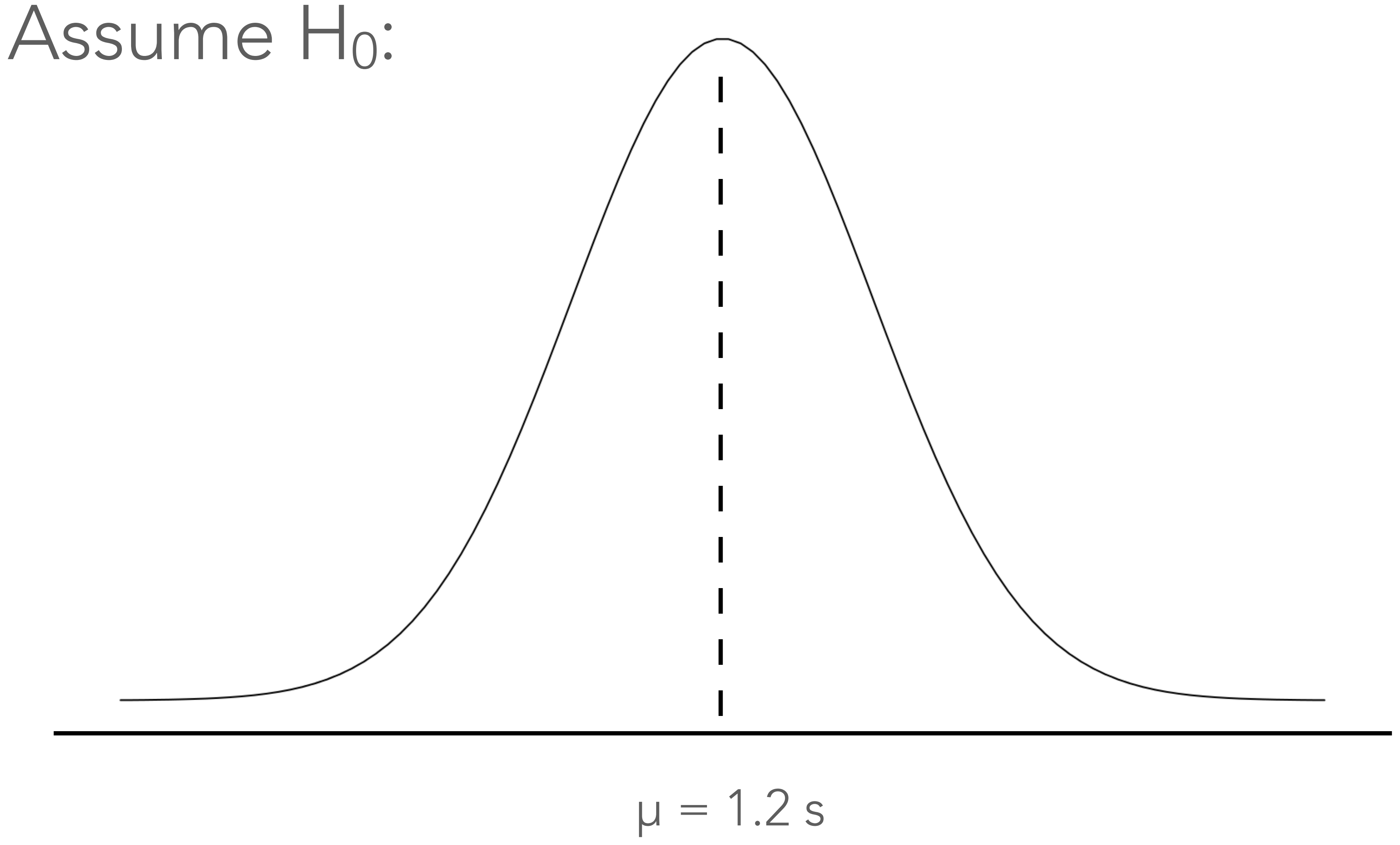
A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Assume H_0:



Example adapted from Sal Khan

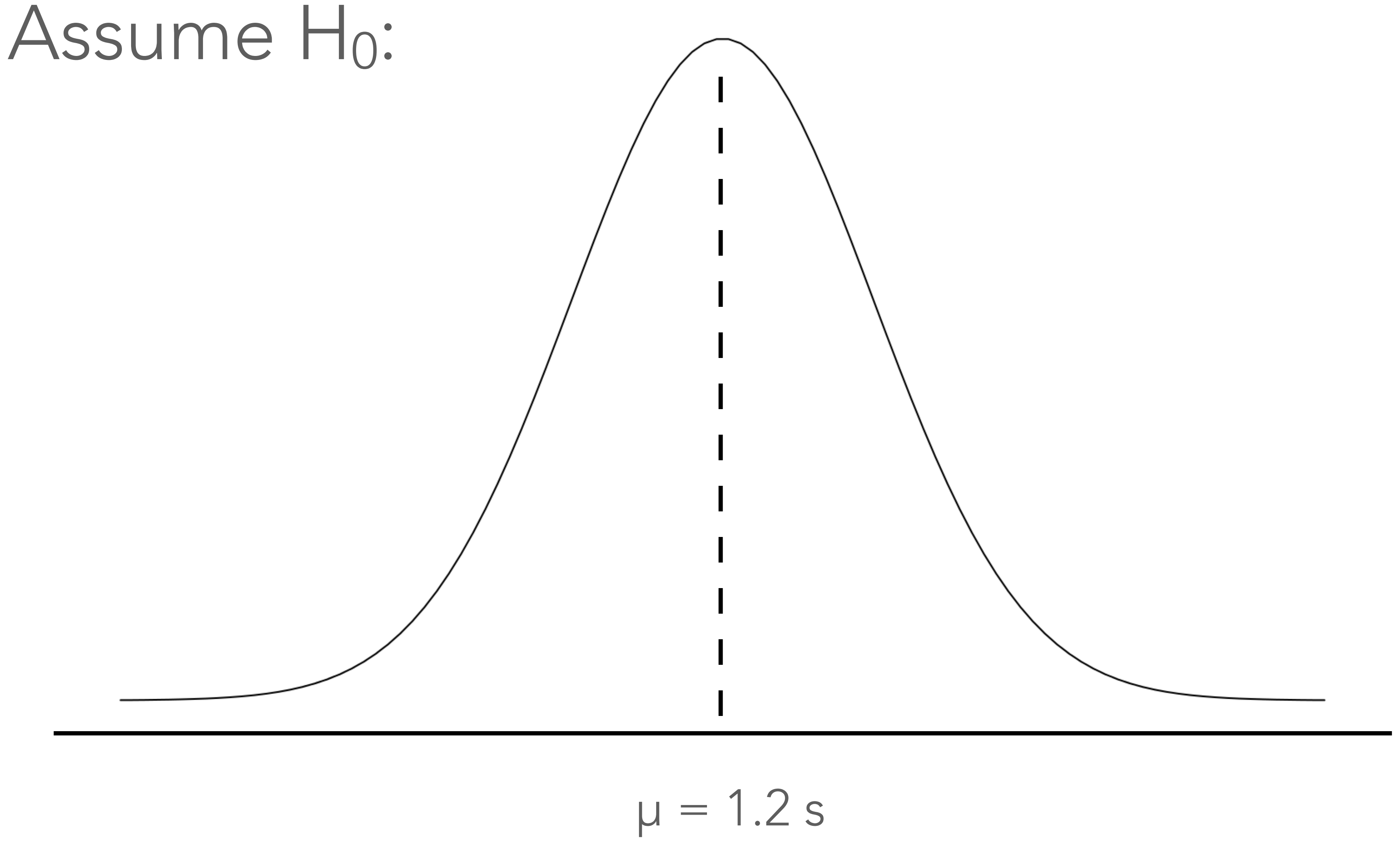
A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

Calculate test statistic $$t = \frac{m - \mu}{s/\sqrt{n}}$$

Assume H_0:



Example adapted from Sal Khan

A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

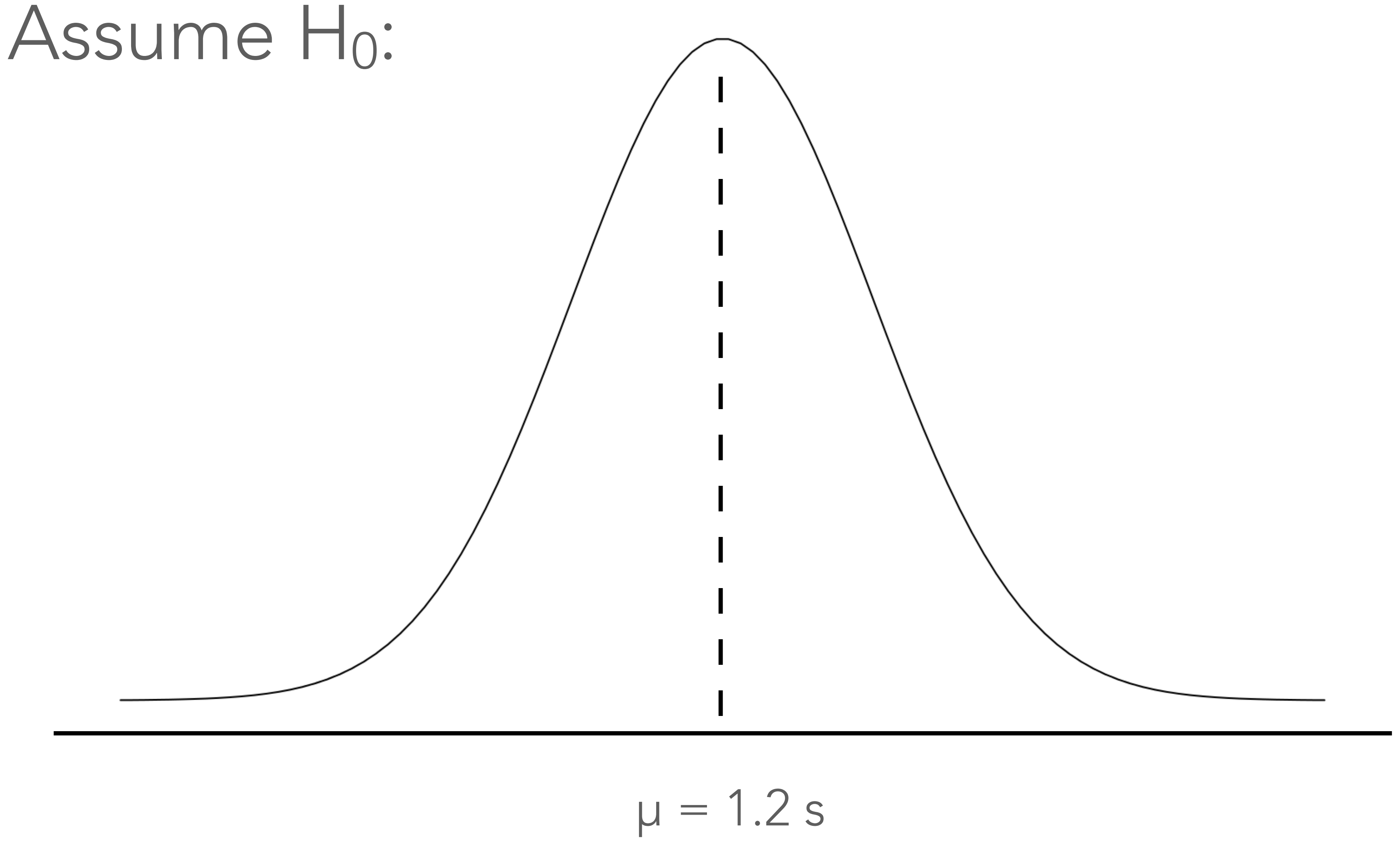
H_0: μ = 1.2 s

H_1: μ ≠ 1.2 s

Calculate test statistic $$t = \frac{m - \mu}{s/\sqrt{n}}$$

1.05 → m, 1.2 → μ, 0.5 → s, 100 → n

Assume H_0:



Example adapted from Sal Khan

A simple example

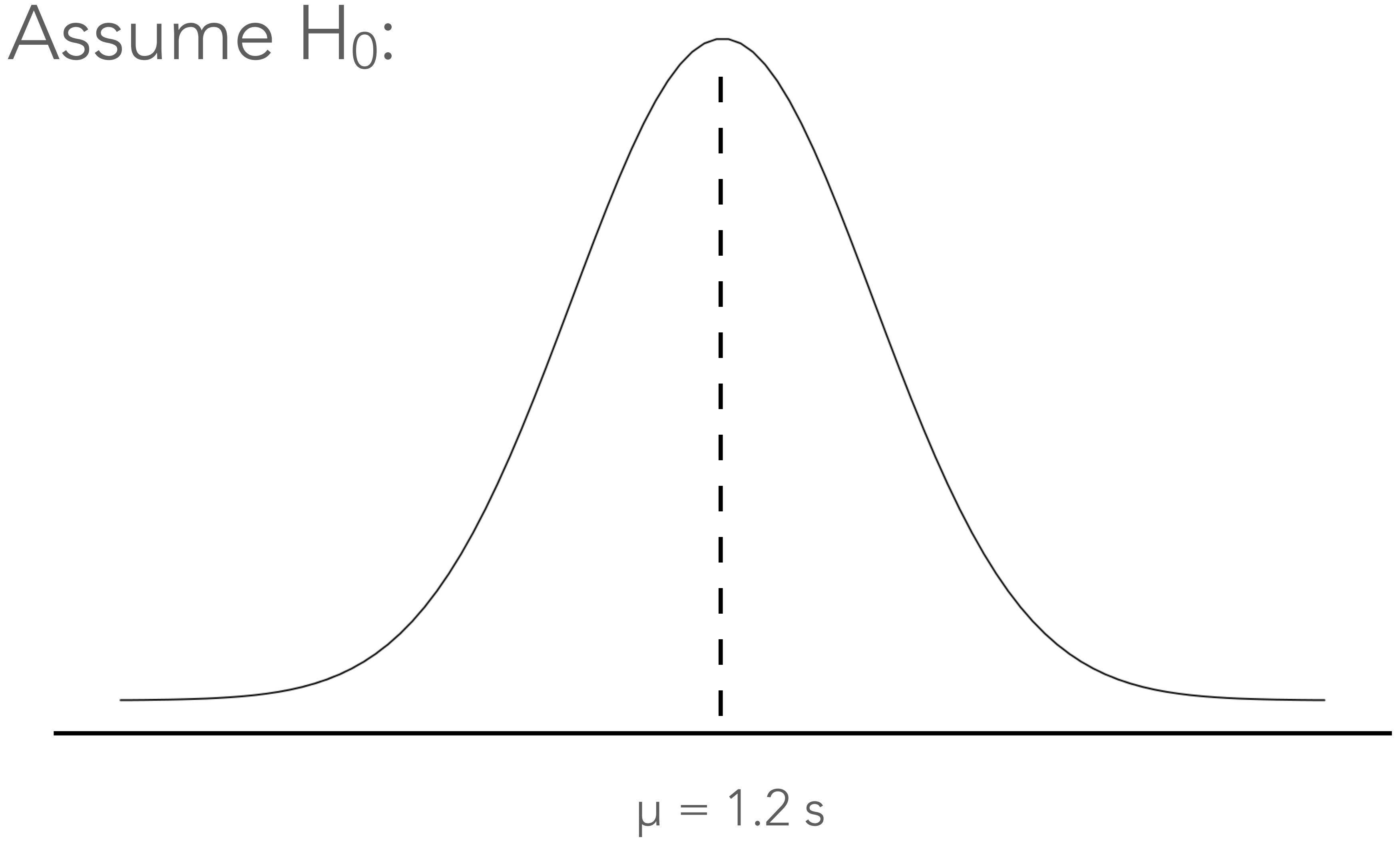
A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $$t = \frac{m - \mu}{s/\sqrt{n}} = -3$$

Assume H_0:



Example adapted from Sal Khan

A simple example

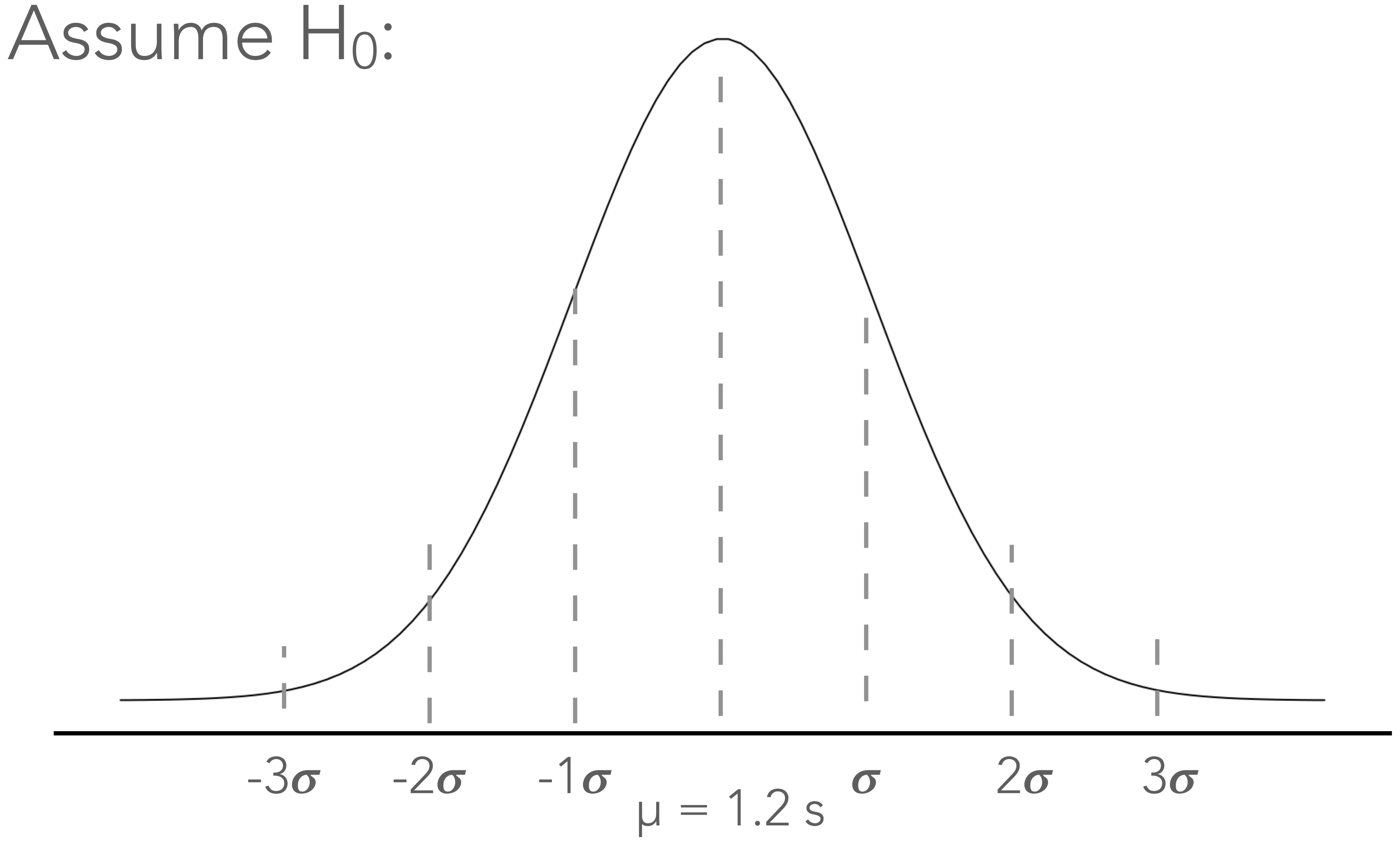
A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s

H_1: $\mu \neq 1.2$ s

Calculate test statistic $$t = \frac{m - \mu}{s/\sqrt{n}} = -3$$

Assume H_0:



This means that the sample mean (1.05) is 3 standard deviations away from the mean

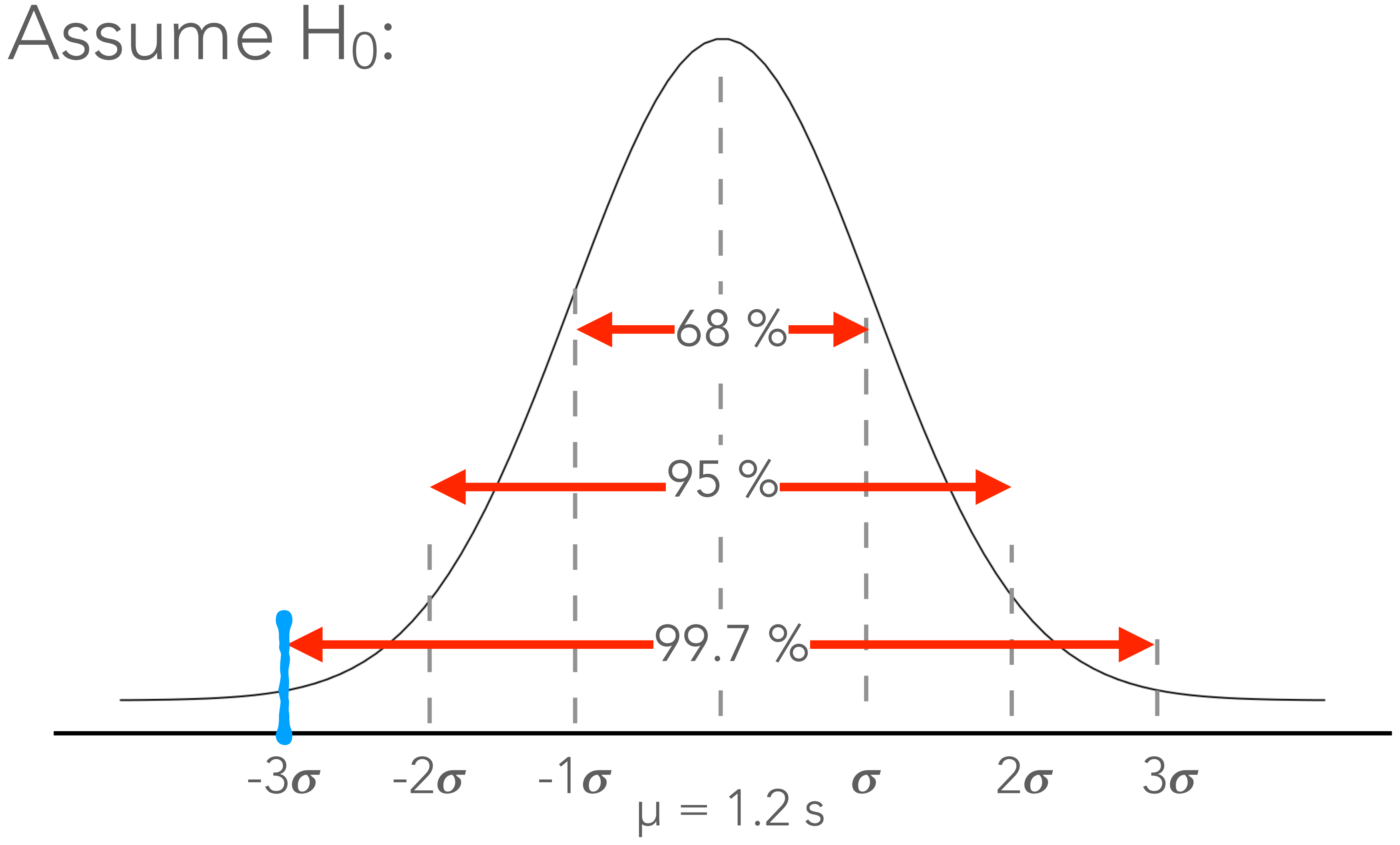
A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

Calculate test statistic $$t = \frac{m - \mu}{s/\sqrt{n}} = -3$$

Assume H_0:



This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

Example adapted from Sal Khan

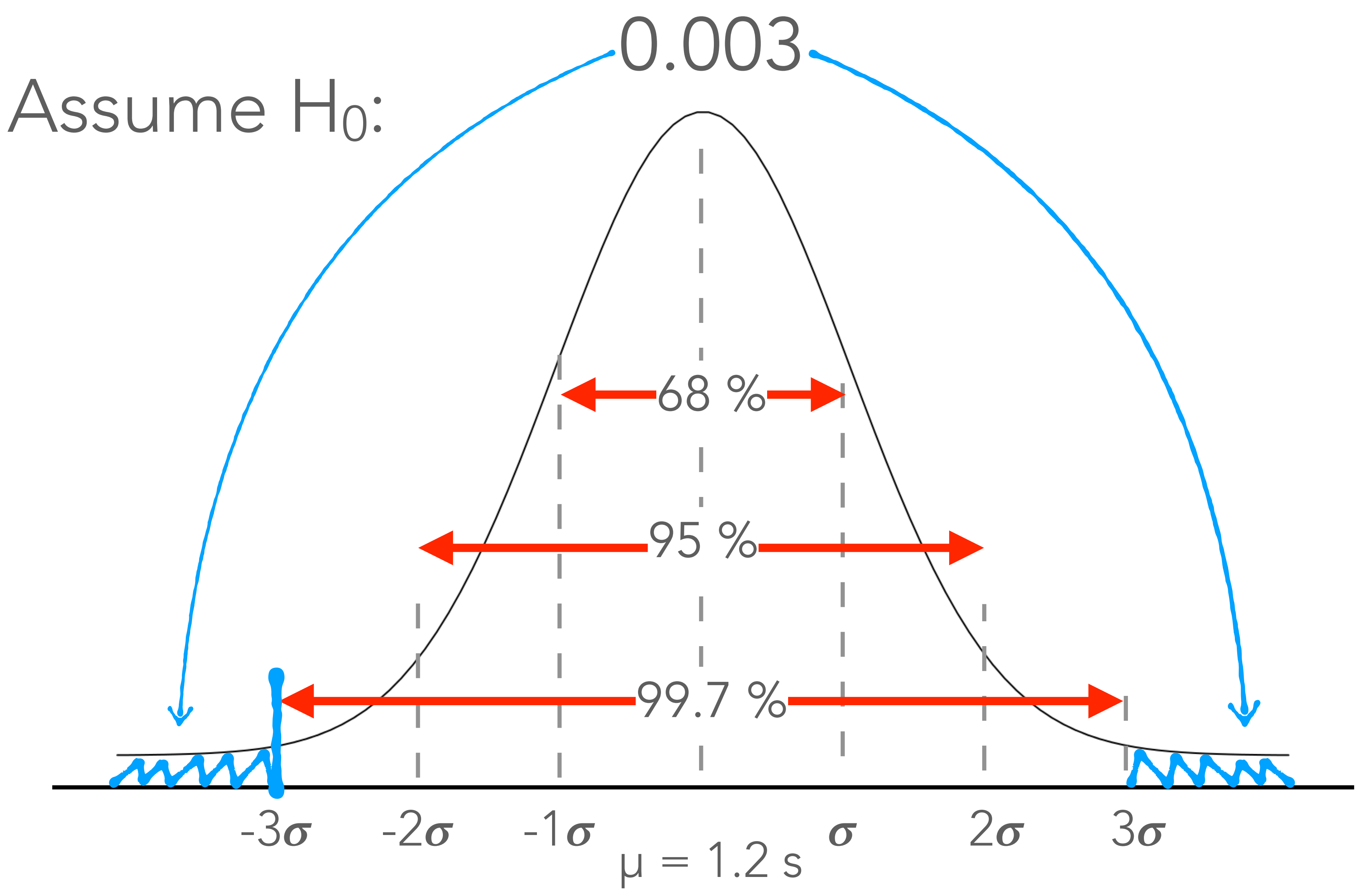
A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}}$ = - 3

Assume H_0:



This means that the sample mean (1.05) is 3 standard deviations away from the mean

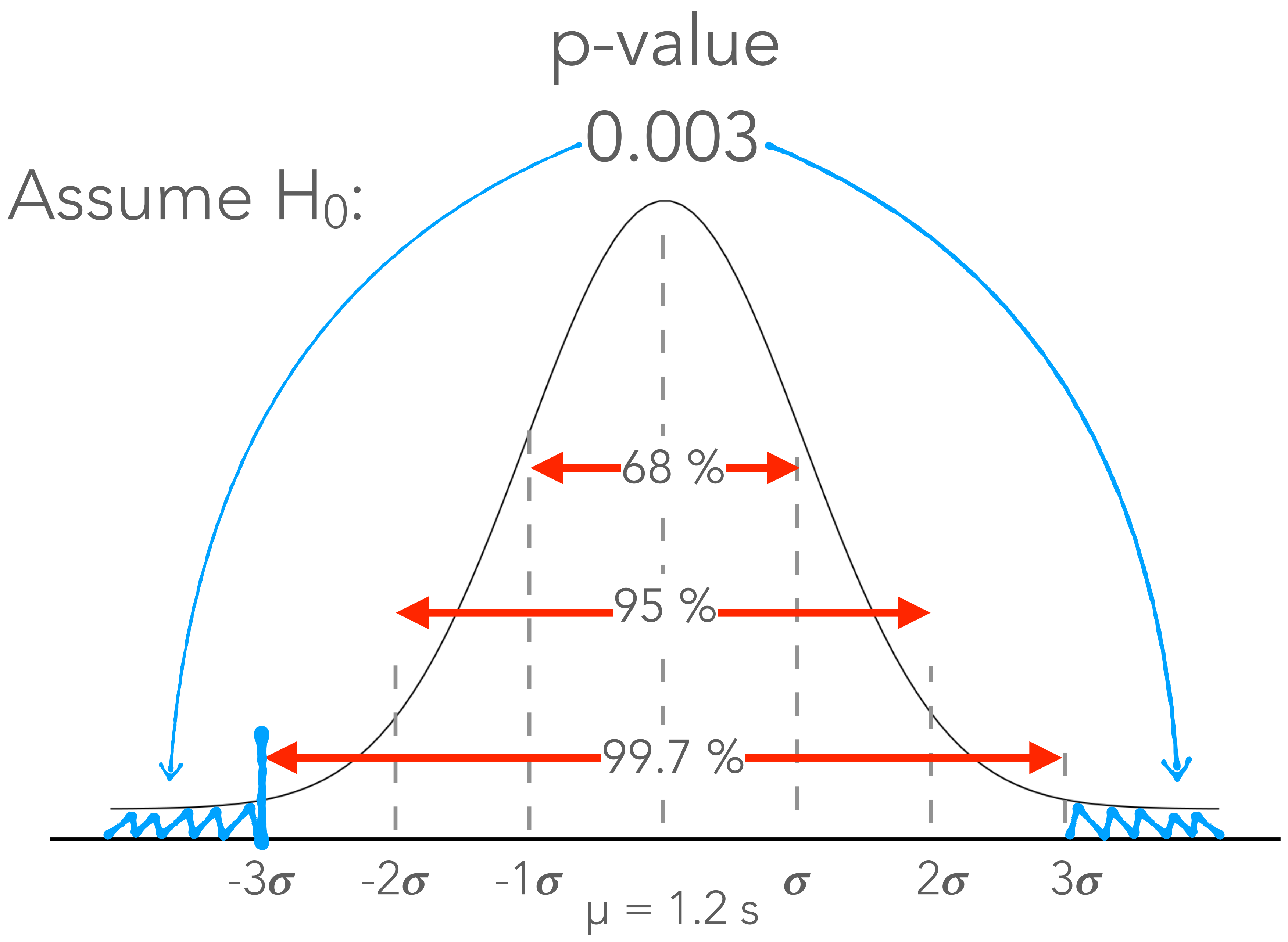
What is the probability of observing a test statistic as extreme as 1.05?

A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

Calculate test statistic $$t = \frac{m - \mu}{s/\sqrt{n}} = -3$$



This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

p-value = 2 min[P(t ≤ t_{obs}|H_0),P(t ≥ t_{obs}|H_0)]

Example adapted from Sal Khan

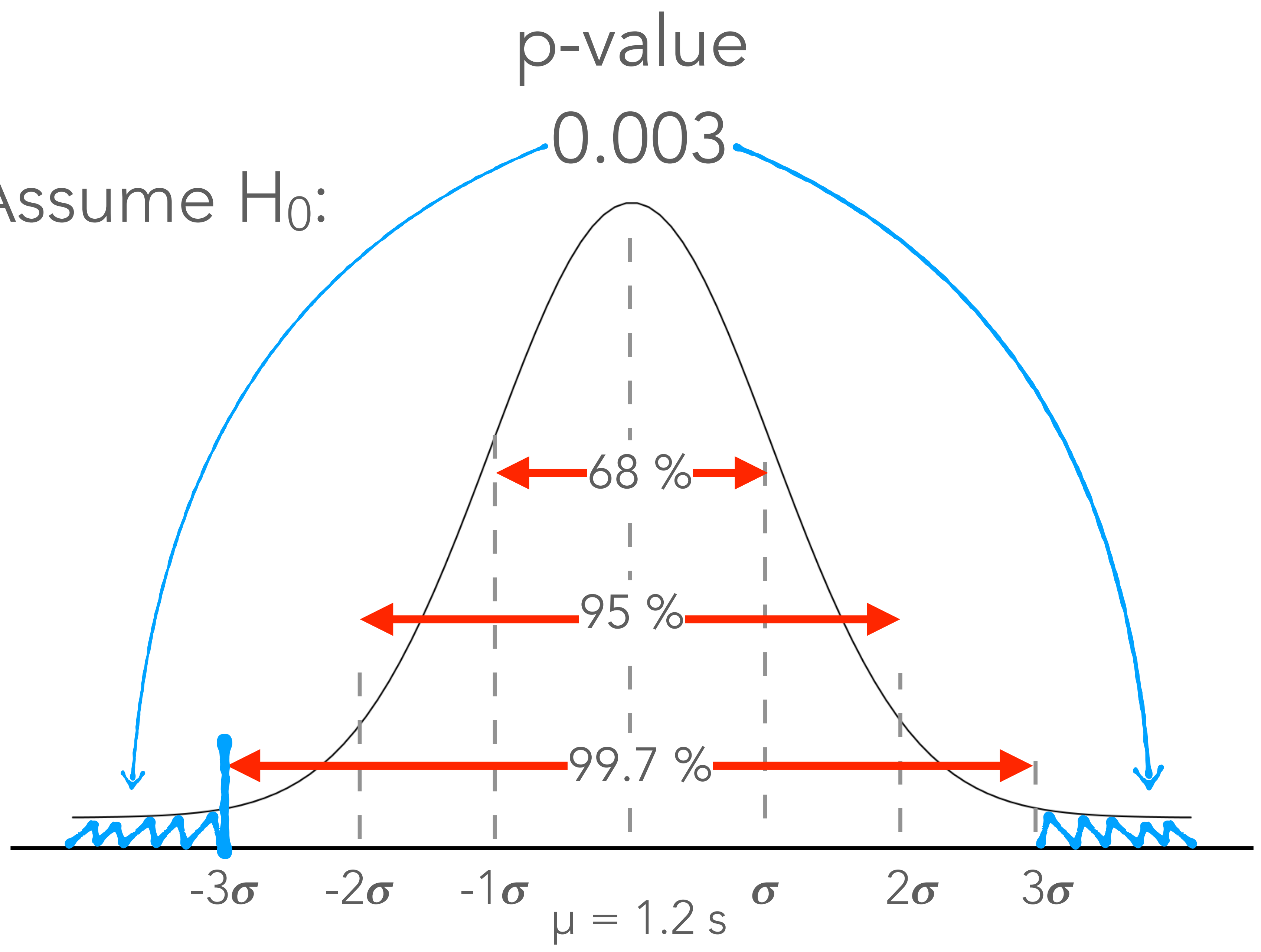
A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

Calculate test statistic $$t = \frac{m - \mu}{s/\sqrt{n}} = -3$$

Assume H_0:



This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

$$\text{p-value} = 2\min[P(t \leq t_{obs} | H_0), P(t \geq t_{obs} | H_0)]$$

We reject the null hypothesis!

Example adapted from Sal Khan

A simple example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with the sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time ?

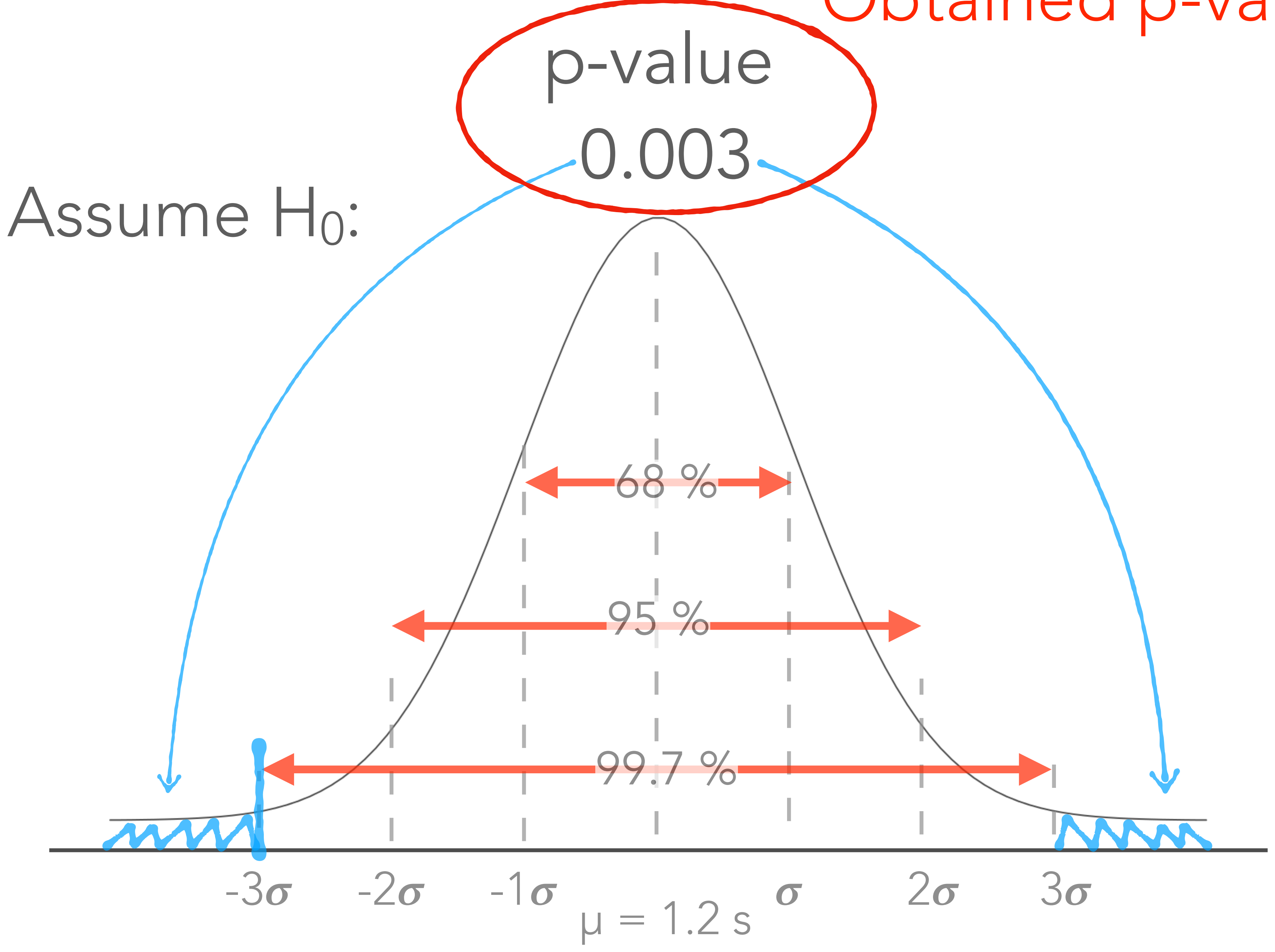
Constructed the null and alternative hypothesis about the population

H_0: μ = 1.2 s
H_1: μ ≠ 1.2 s

Calculate test statistic $t = \dfrac{m - \mu}{s/\sqrt{n}}$ = - 3

Calculated test statistic

Obtained p-value



This means that the sample mean (1.05) is 3 standard deviations away from the mean

What is the probability of observing a test statistic as extreme as 1.05?

$$\text{p-value} = 2\min[P(t \le t_{obs} \mid H_0), P(t \ge t_{obs} \mid H_0)]$$

Reached a conclusion

We reject the null hypothesis!

Example adapted from Sal Khan

Key Concepts - Hypothesis Testing

- All statistical tests are based on assumptions!
- All statistics can be wrong
- Statistical tests are probabilistic in nature
- There is always a chance that the result is wrong (even when all assumptions met perfectly):
  - Either significant result when no difference (Type I),
  - Or insignificant results when there is an actual difference (Type II)

Type I and Type II Errors

- All hypothesis tests involve making a decision:

Is this result significant or not?

Type I and Type II Errors

- All hypothesis tests involve making a decision:

Is this result significant or not?

- This decision can be wrong in two ways:

Slide content adapted from Matt Castle

Type I and Type II Errors

- All hypothesis tests involve making a decision:

Is this result significant or not?

- This decision can be wrong in two ways:

Type I error or False positive

This is when you reject the null hypothesis when it is true

"You're pregnant !"

Slide content adapted from Matt Castle

Type I and Type II Errors

- All hypothesis tests involve making a decision:

Is this result significant or not?

- This decision can be wrong in two ways:

Type I error or False positive

This is when you reject the null hypothesis when it is true

"You're pregnant ! "

Type II error or False negative

This is when you fail to reject the null hypothesis when it isn't true

"You're not pregnant"

Slide content adapted from Matt Castle

Type I and Type II Errors

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

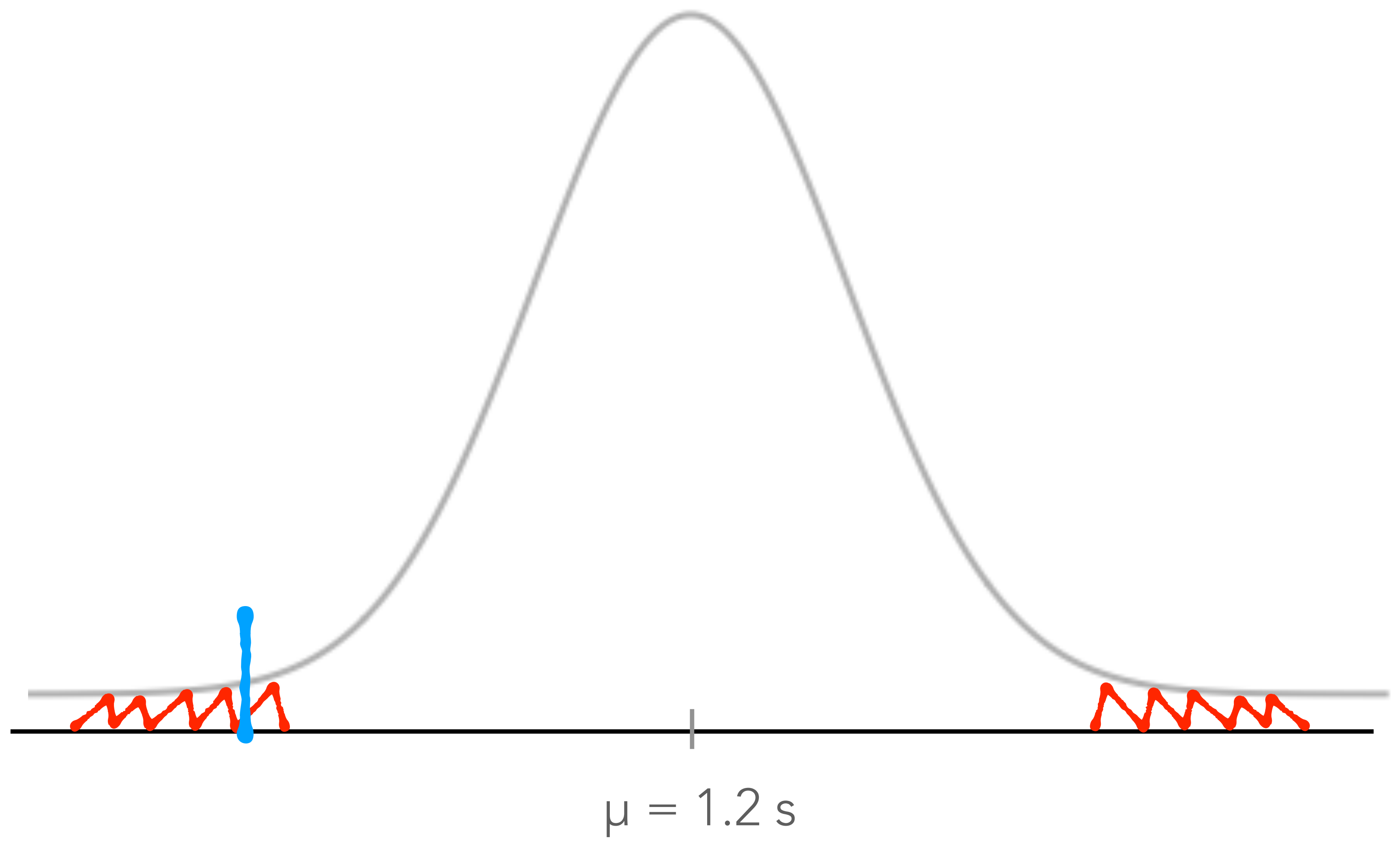
if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

Type I and Type II Errors

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct

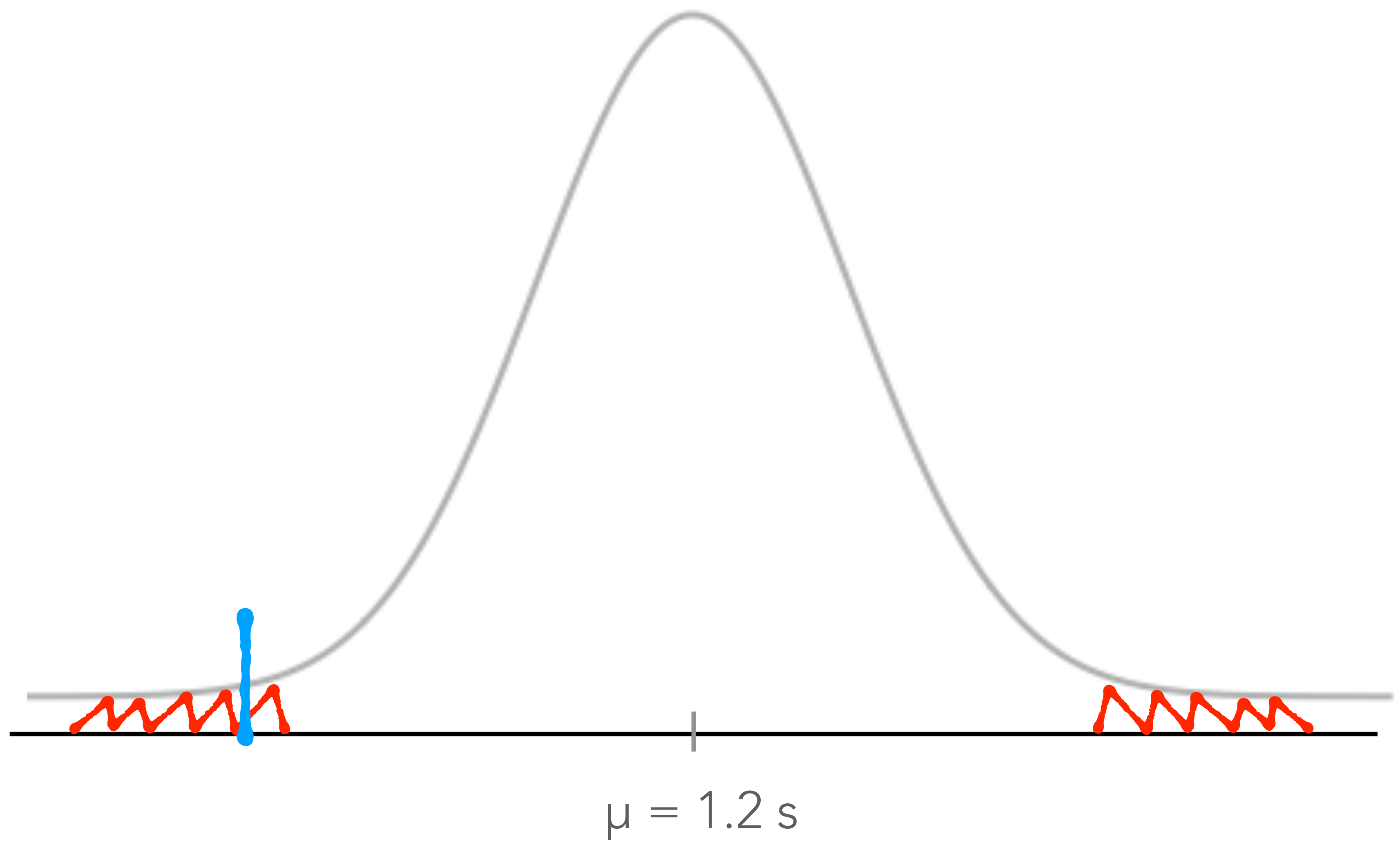


Type I and Type II Errors

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



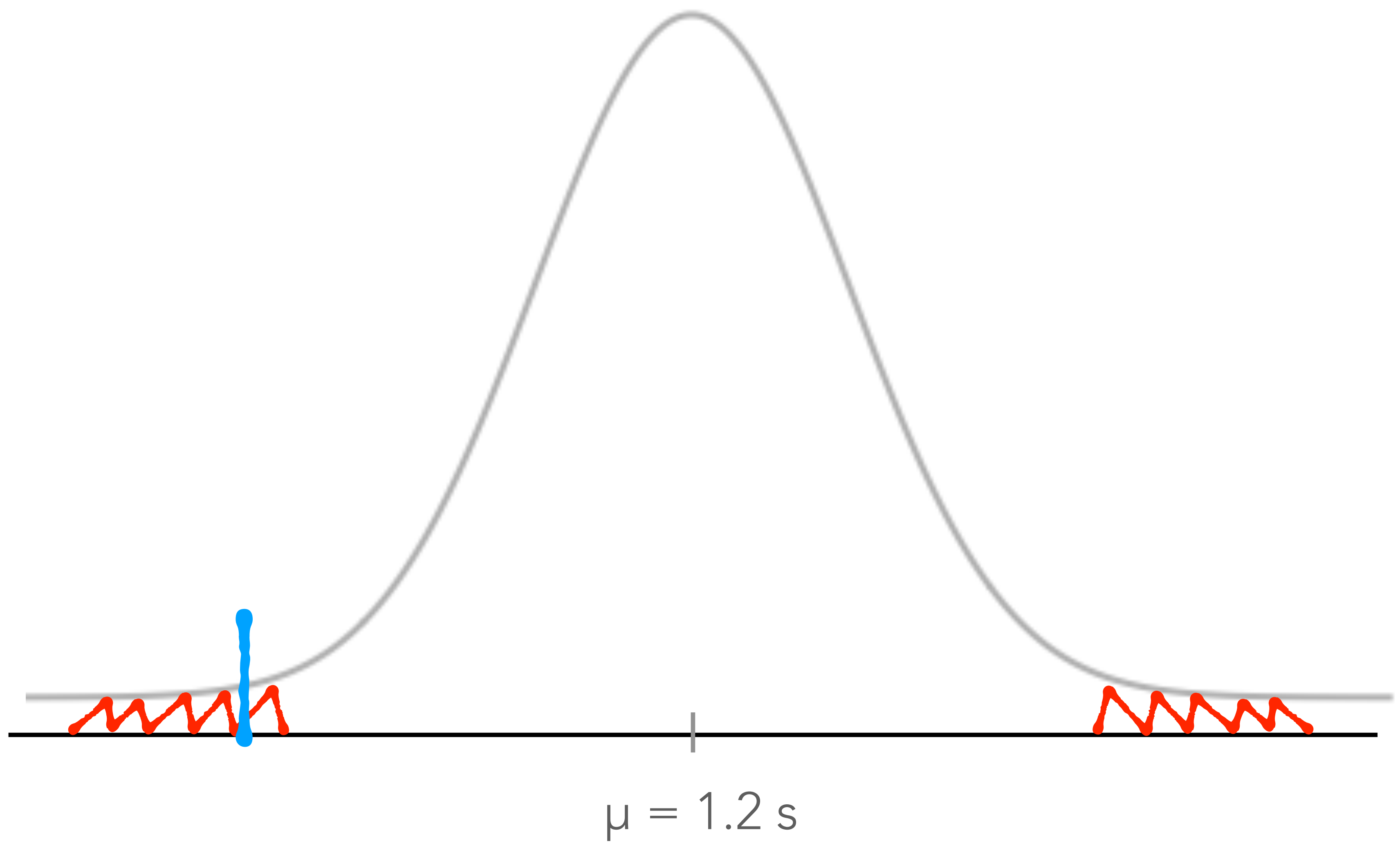
Suppose H_1 true:

Type I and Type II Errors

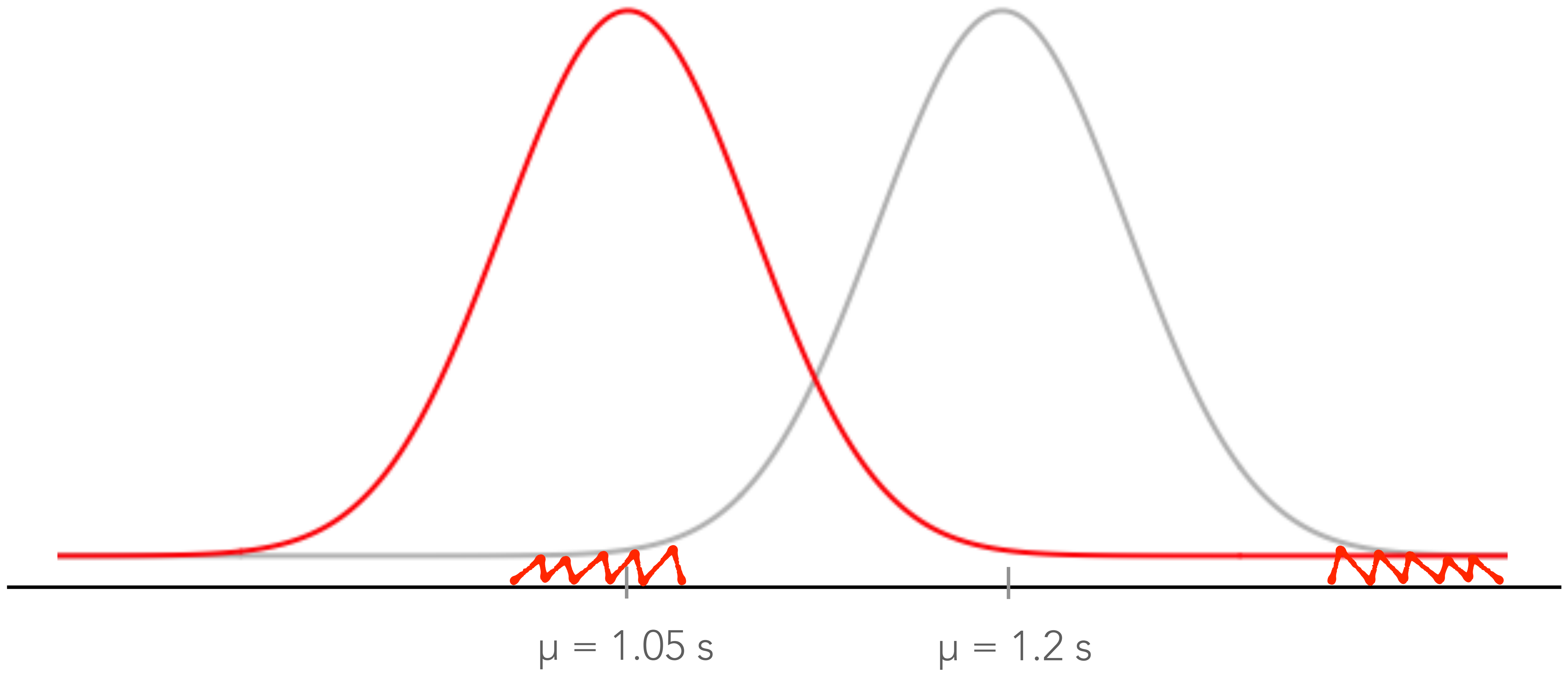
H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



Suppose H_1 true:

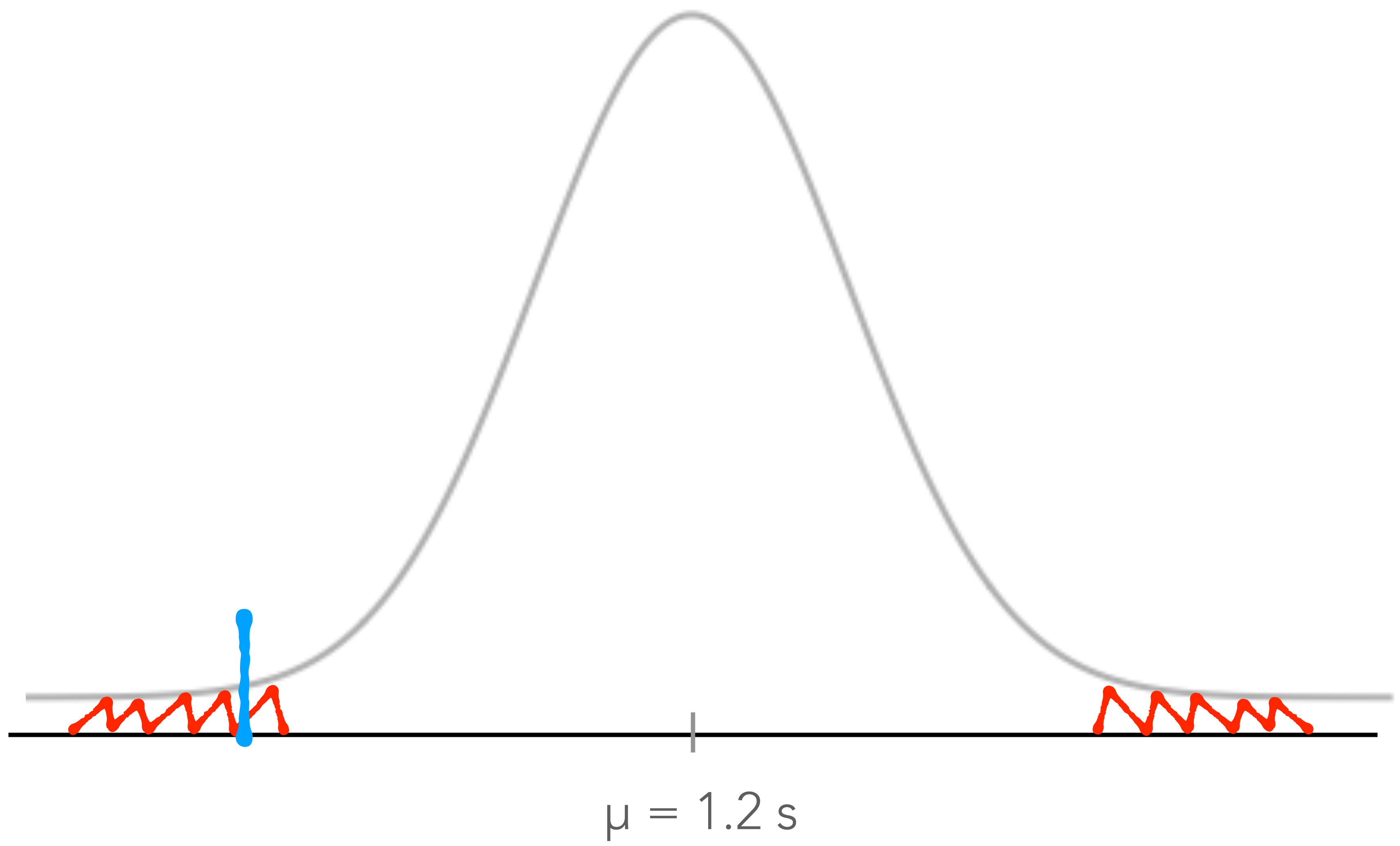


Type I and Type II Errors

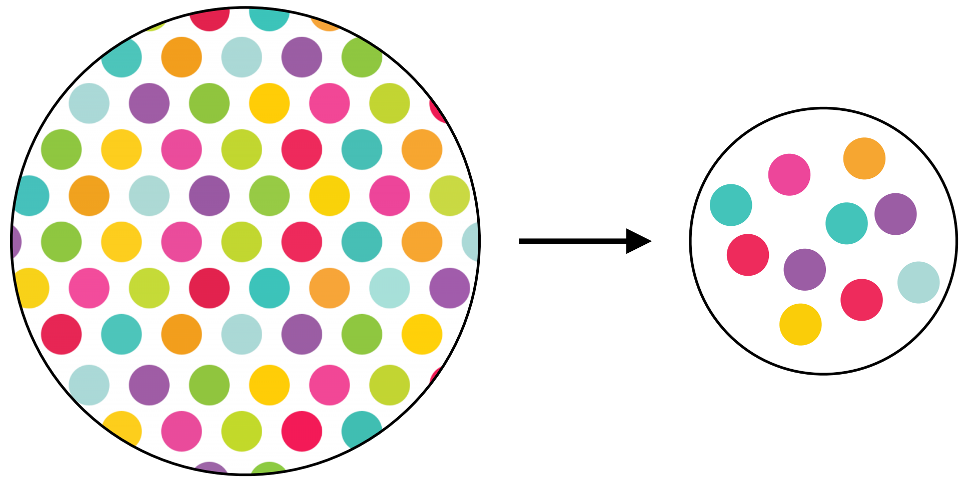
H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

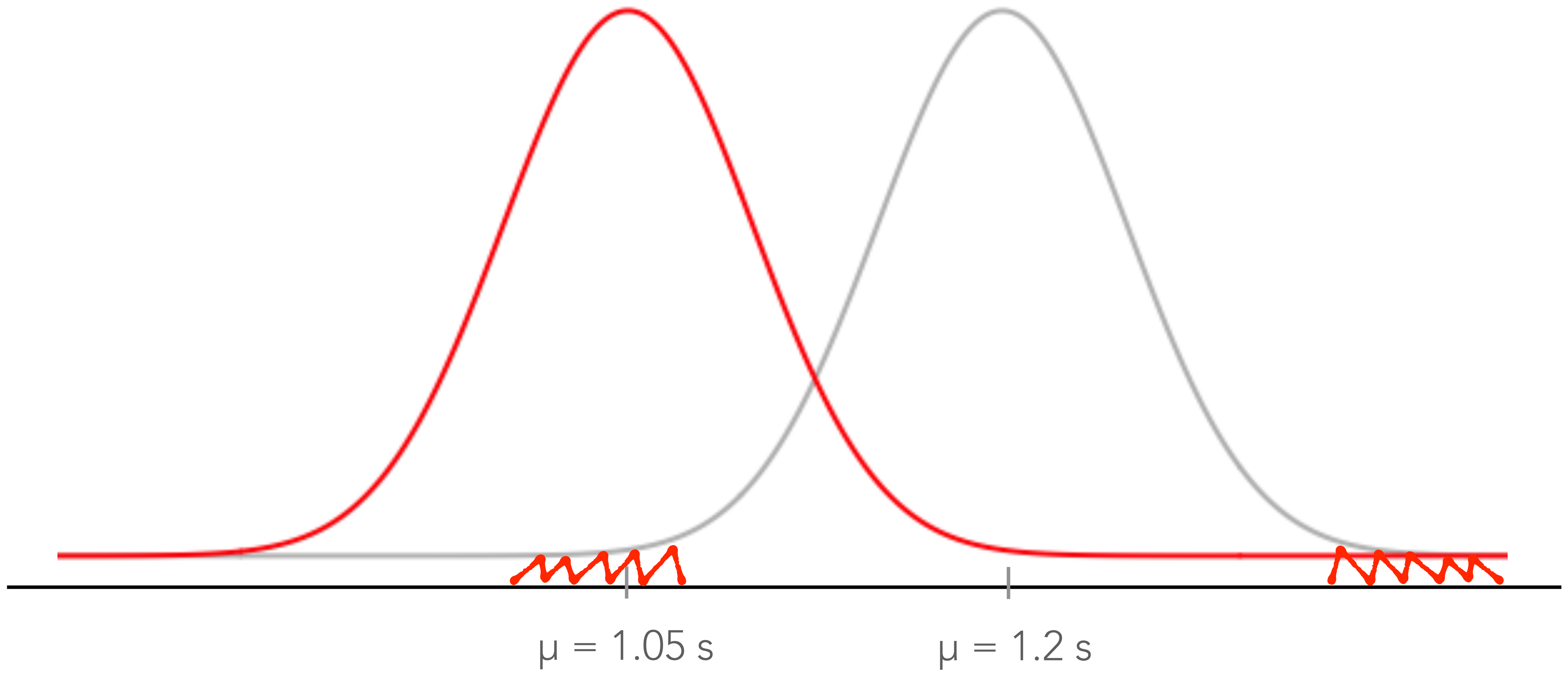
α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



Suppose H_1 true:



Depending on your sampling, you might fail to reject H_0

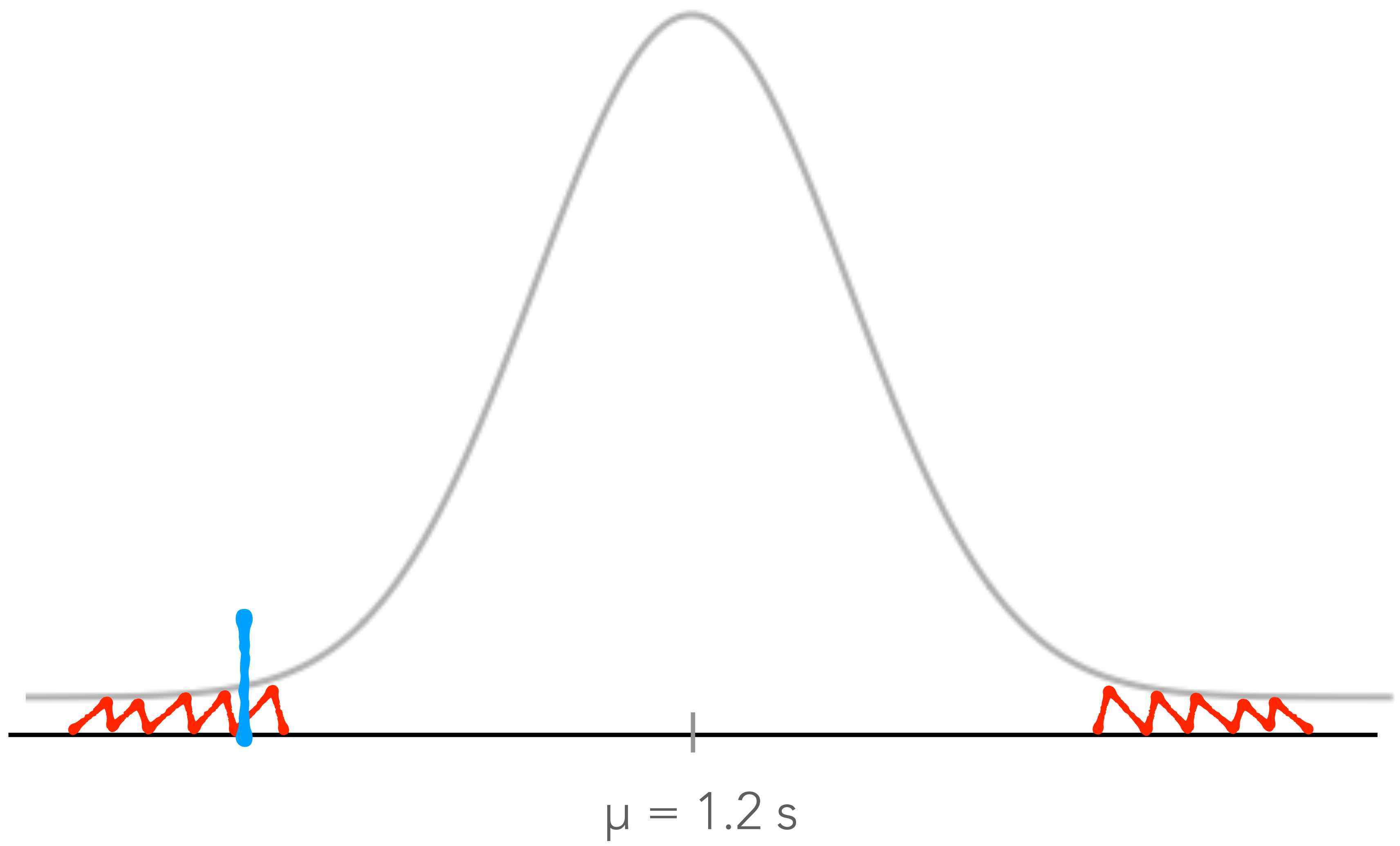


Type I and Type II Errors

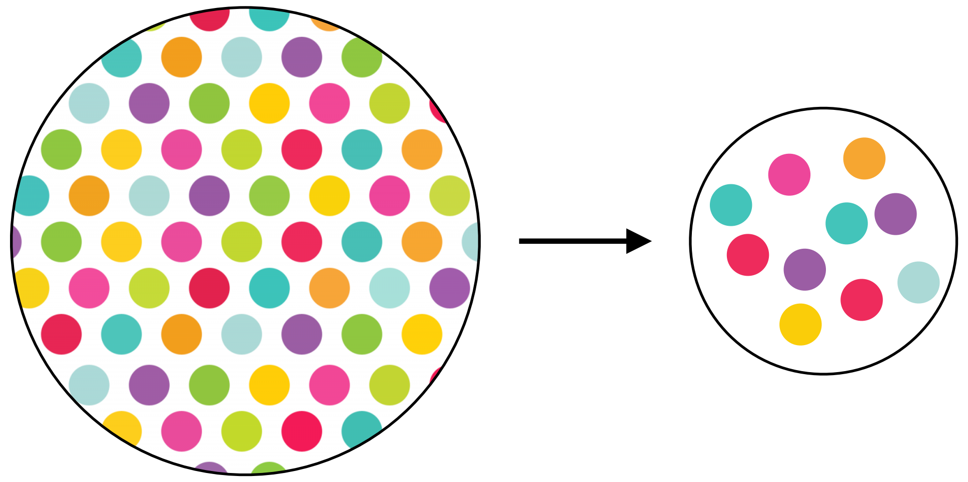
H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

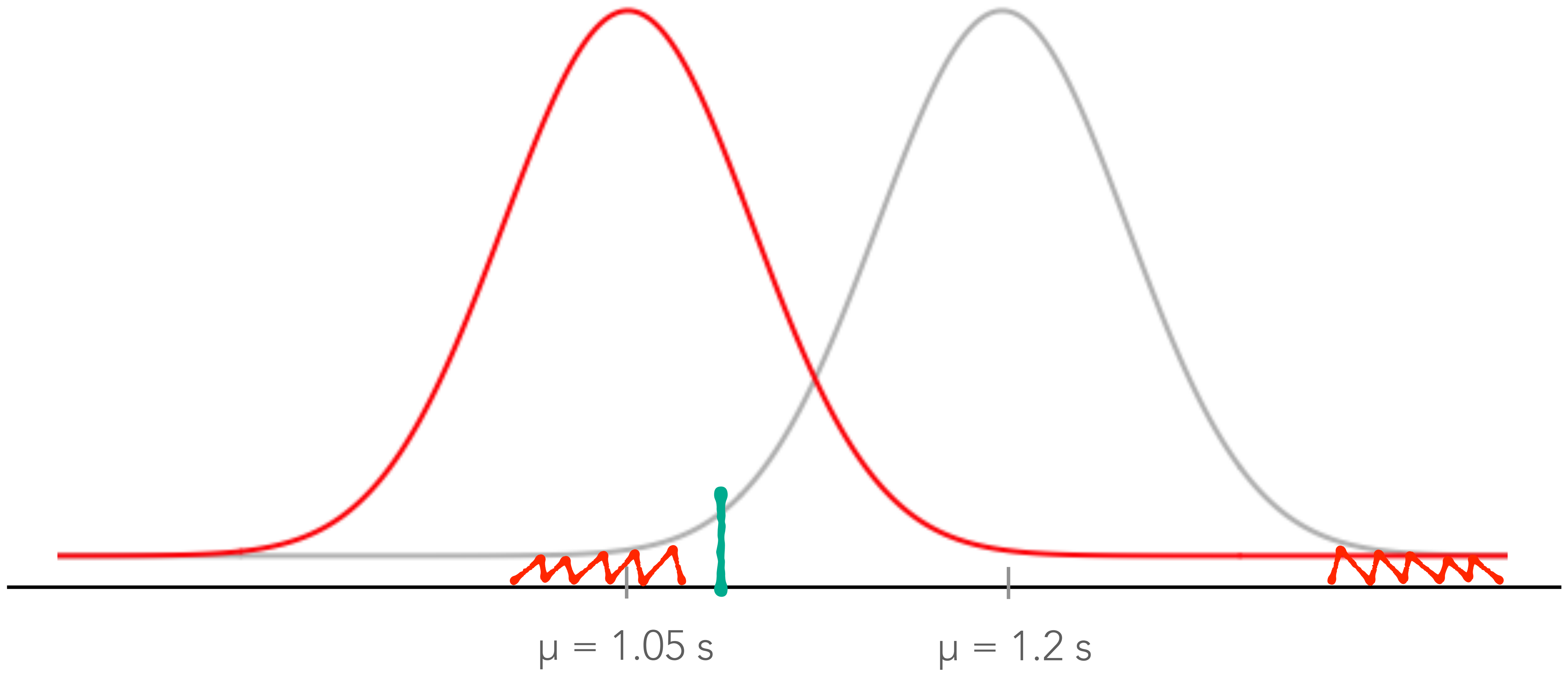
α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



Suppose H_1 true:



Depending on your sampling, you might fail to reject H_0

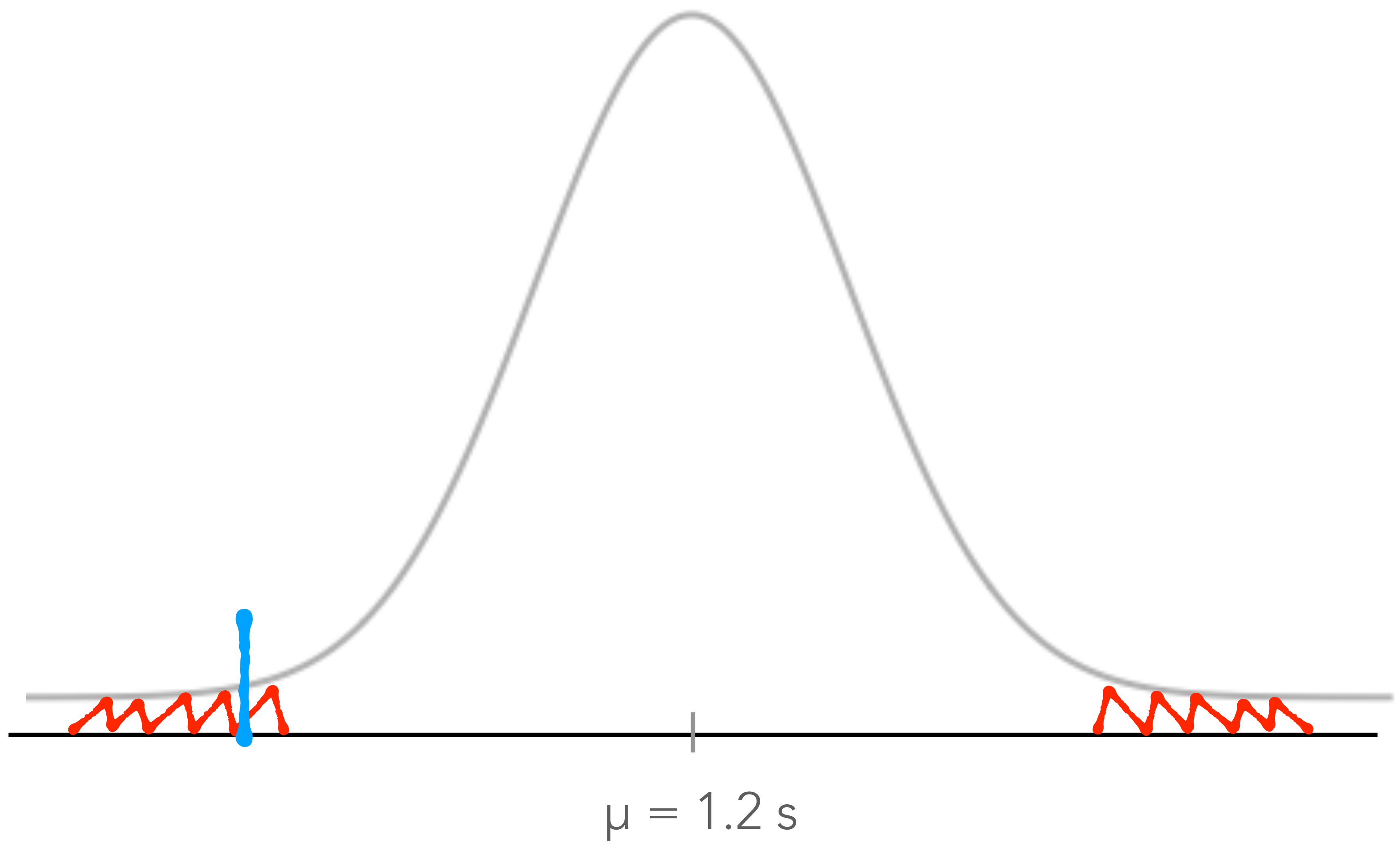


Type I and Type II Errors

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

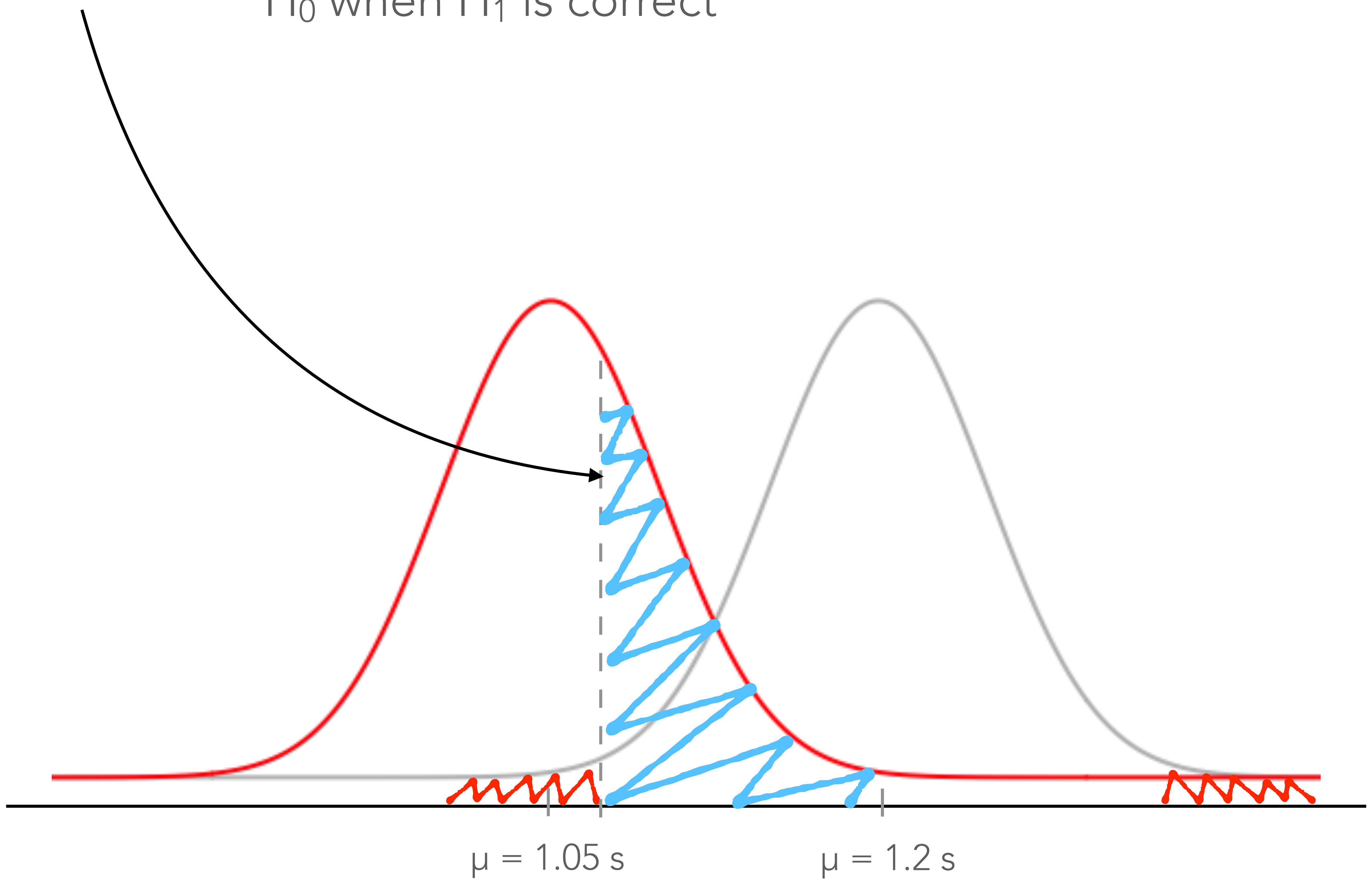
if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

$\alpha = 0.05$ → the type I error, the probability of rejecting H_0 when H_0 is correct



Suppose H_1 true:

θ → the type II error, the probability of not rejecting H_0 when H_1 is correct

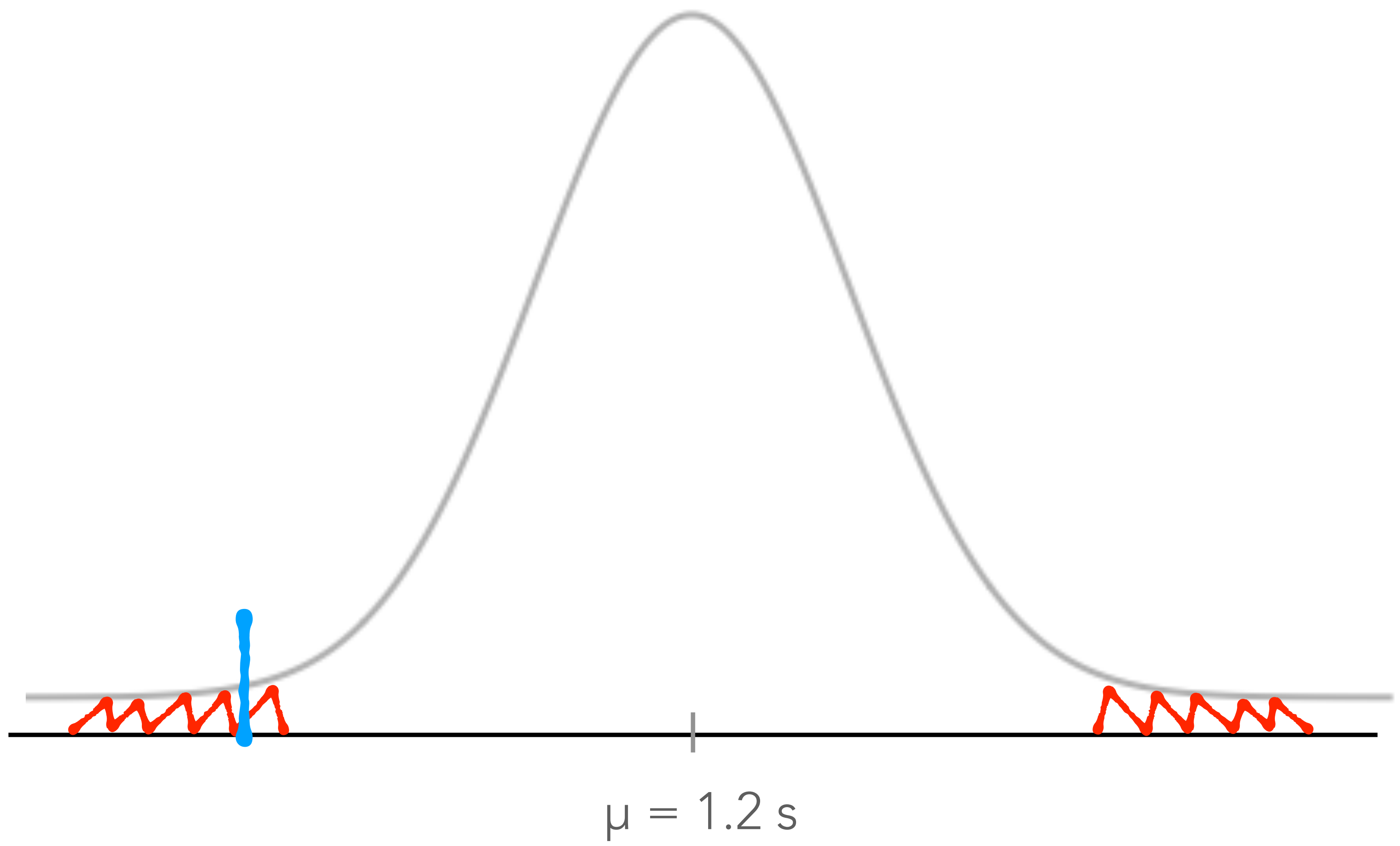


Type I and Type II Errors

H_0: $\mu = 1.2$ s
H_1: $\mu \neq 1.2$ s

if p-value > α → do not reject H_0
if p-value < α → reject H_0 in favour of H_1

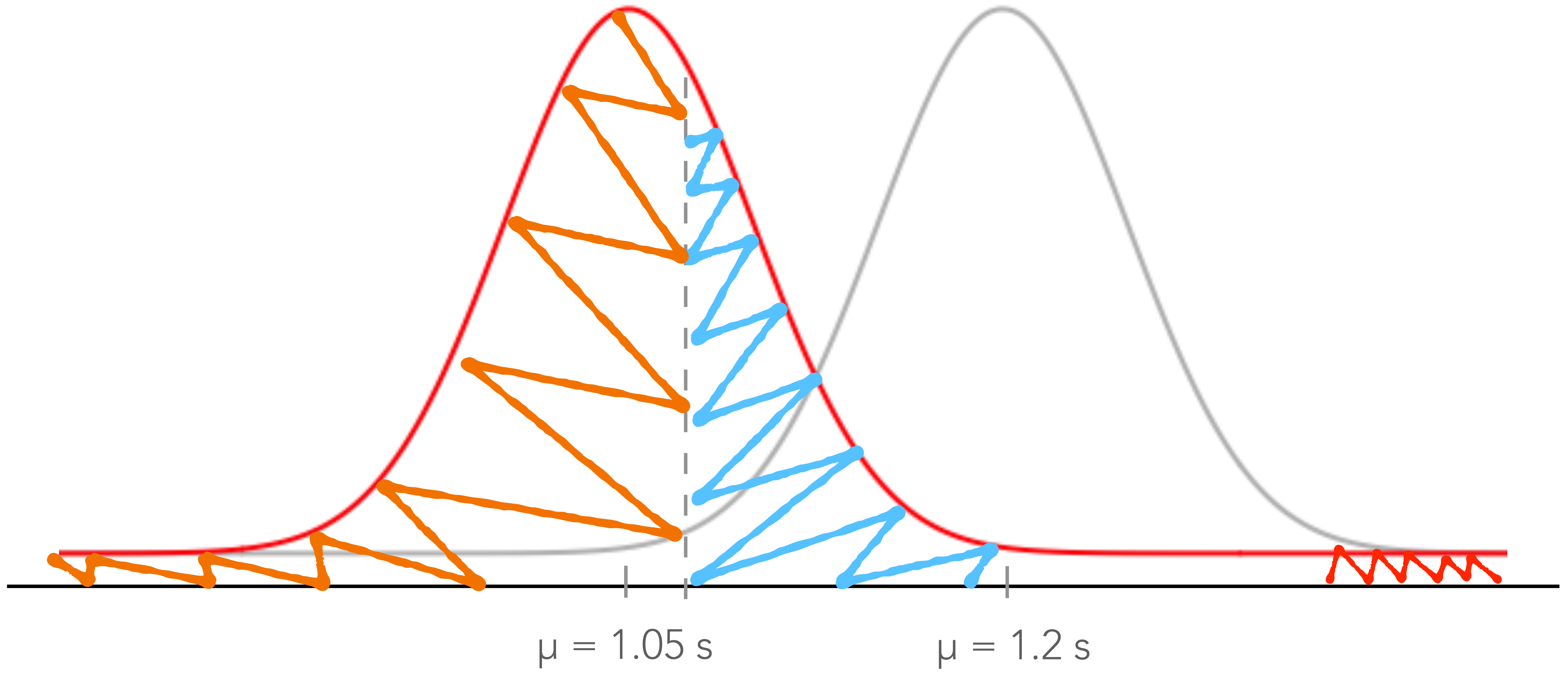
α=0.05 → the type I error, the probability of rejecting H_0 when H_0 is correct



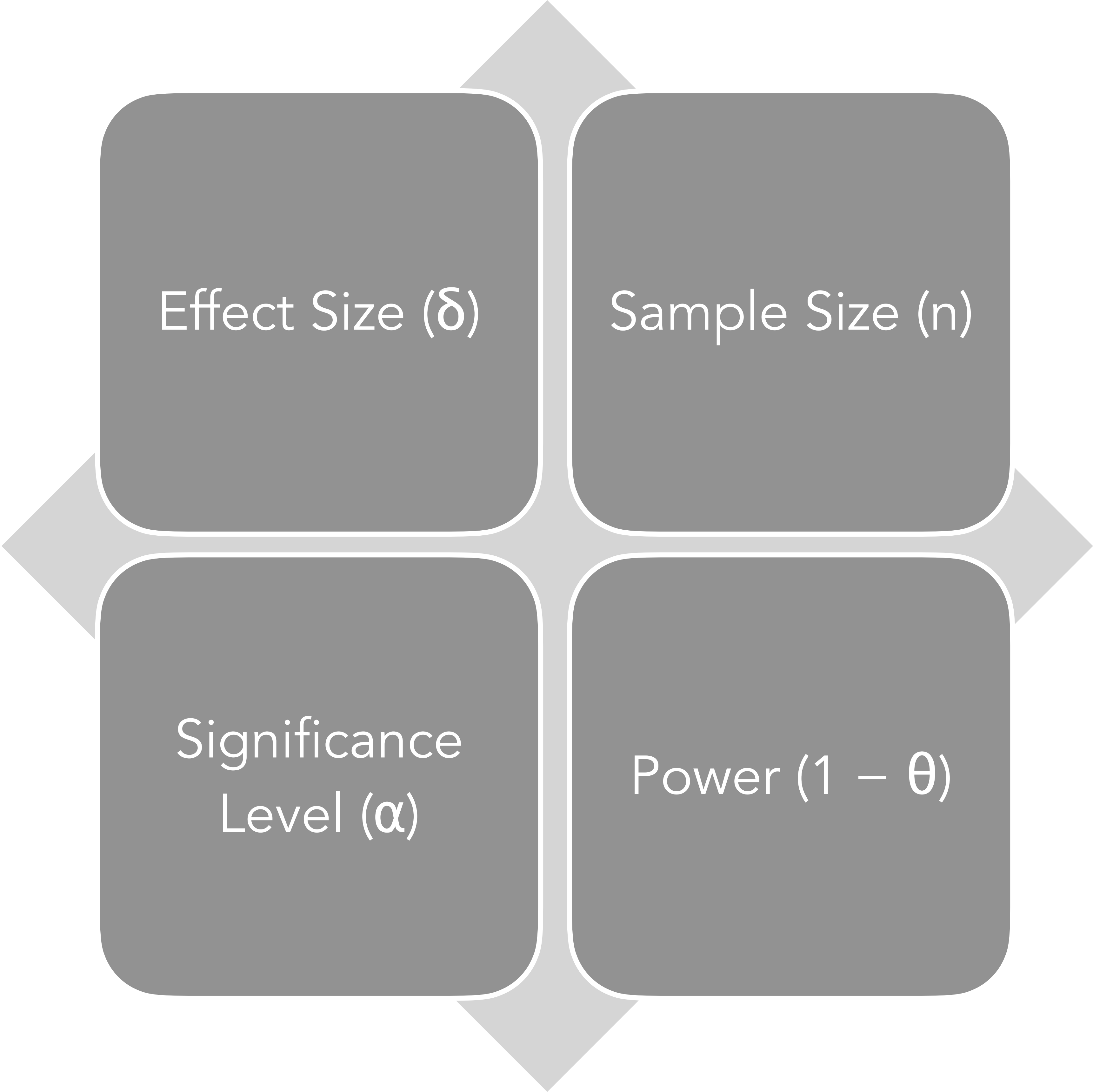
Suppose H_1 true:

θ → the type II error, the probability of not rejecting H_0 when H_1 is correct

1- θ → Power is the probability that we actually detect an effect that exists

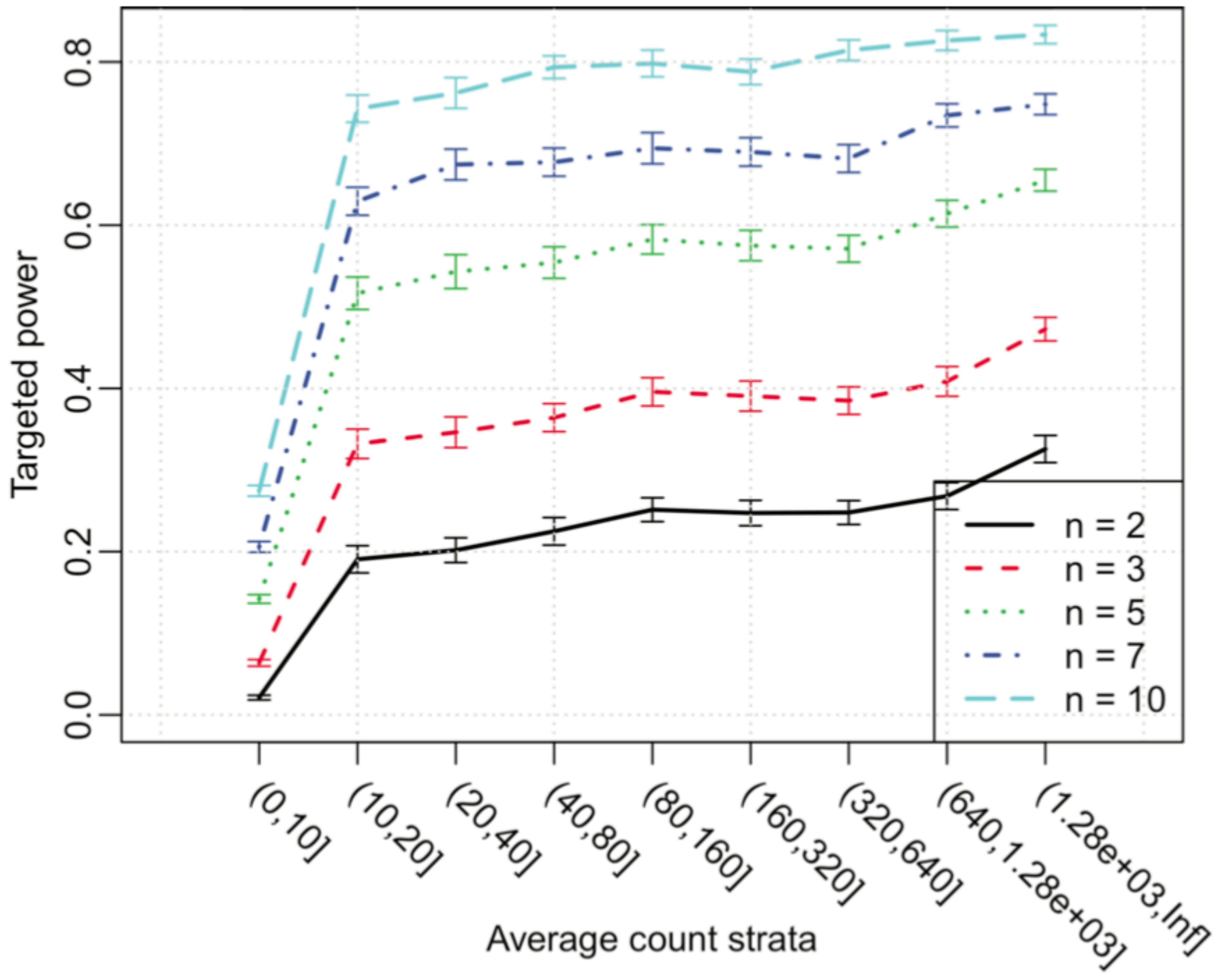


Power Analysis



- The four concepts are linked
- If we know three, we can work out the forth
- Power calculation: Aim is to define the probability (1-θ) to detect an effect size of interest (δ) at the α level with a sample size of *n* biological replicates
- Sample size calculation: Aim is to define the sample size (*n*) allowing to detect an effect size of interest (δ) at the α level with a given probability (1 – θ).

Power Analysis in Differential Expression Analysis



(Wu, Wang and Wu (2015))

Statistical Aspects of Differential Expression Analysis

Linear Modeling

Model the expression of each gene as linear combination of explanatory factors (eg. treatment, age, sex, etc.)

Statistical Aspects of Differential Expression Analysis

Linear Modeling

Model the expression of each gene as linear combination of explanatory factors (eg. treatment, age, sex, etc.)

$$y = a + (b * \text{treatment}) + (c * \text{age}) + (d * \text{sex}) + e$$

y = expression of gene

a, b, c, d = parameters estimated from the data

a = intercept (expression when factors are at reference level)

e = error term

Statistical Aspects of Differential Expression Analysis

Linear Modeling

Model the expression of each gene as linear combination of explanatory factors (eg. treatment, age, sex, etc.)

y = a + (b * treatment) + (c * age) + (d * sex) + e

y = expression of gene

a, b, c, d = parameters estimated from the data

a = intercept (expression when factors are at reference level)

e = error term

observation = deterministic model + residual error

Statistical Aspects of Differential Expression Analysis

Linear Modeling

Model the expression of each gene as linear combination of explanatory factors (eg. treatment, age, sex, etc.)

y = a + (b * treatment) + (c * age) + (d * sex) + e

y = expression of gene

a, b, c, d = parameters estimated from the data

a = intercept (expression when factors are at reference level)

e = error term

observation = deterministic model + residual error

$$y = \beta X + \varepsilon$$

Express the count data vector of a given gene, y, as a function parameter vector *(β)* times a design matrix *(X)* plus a stochastic error vector ε

Statistical Aspects of Differential Expression Analysis

Linear Modeling

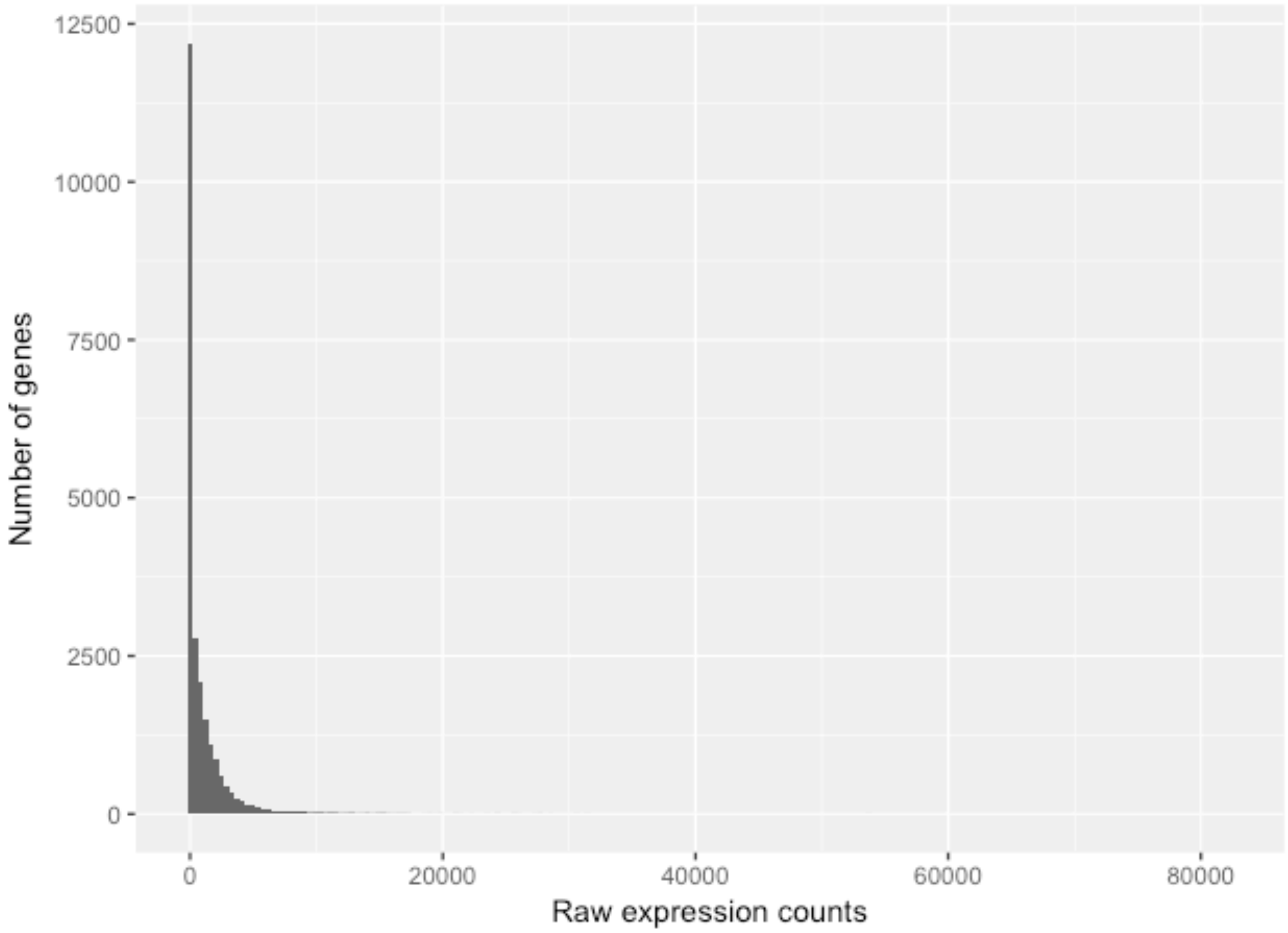
- Collect the information related to each sample for predictors of interest
- Define β, the sets of parameters we are interested in
- build the X matrix that relates the sample information with the β
- estimate the β and use statistical inference to assess significance (p-values)

Construction of Design Matrix

Next Session!

Statistical Aspects of Differential Expression Analysis

Characteristics of RNA-seq data

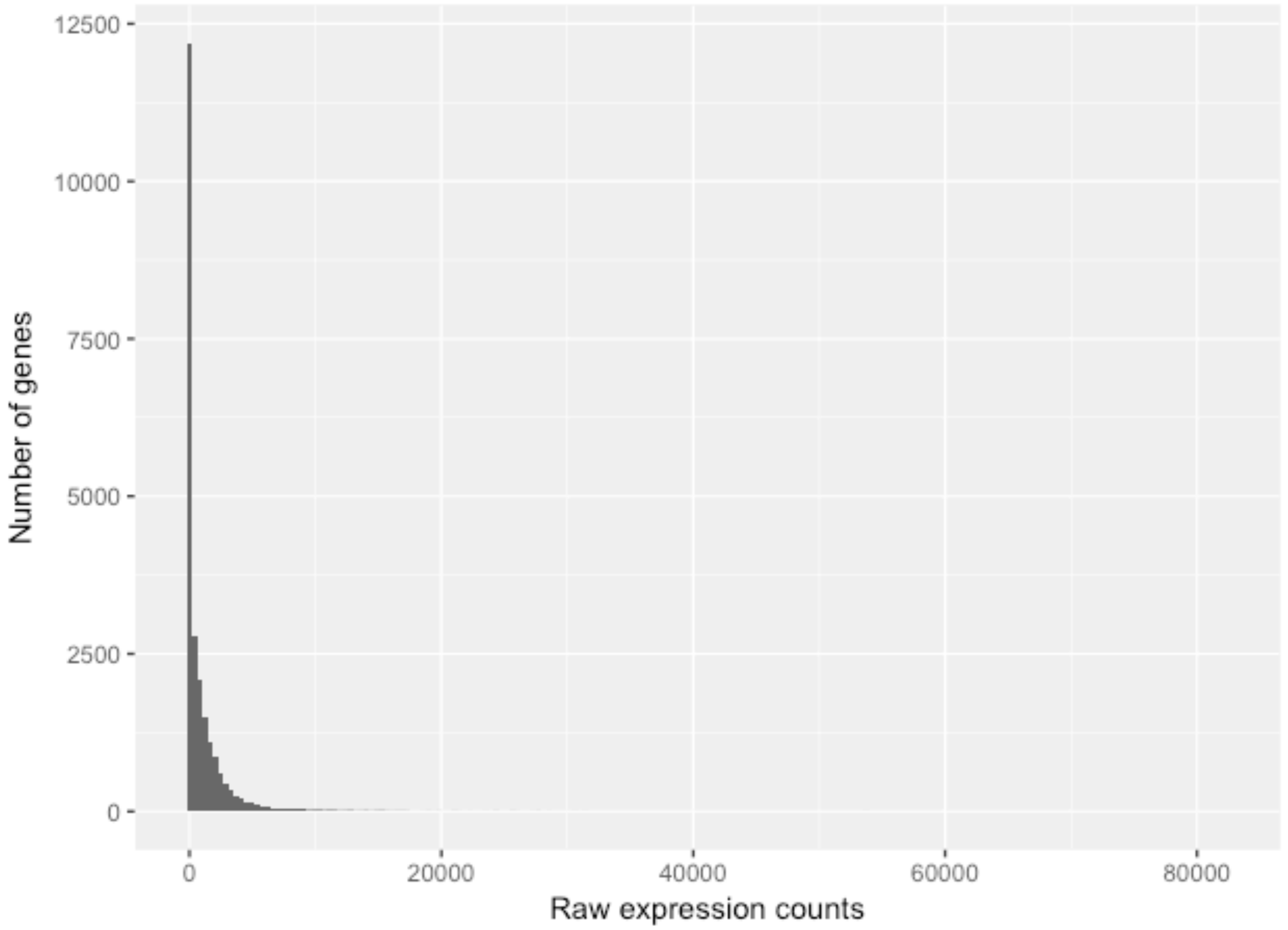


This plot illustrates some **common features** of RNA-seq count data:

- a low number of counts associated with a large proportion of genes
- a long right tail due to the lack of any upper limit for expression
- large dynamic range

Statistical Aspects of Differential Expression Analysis

Characteristics of RNA-seq data



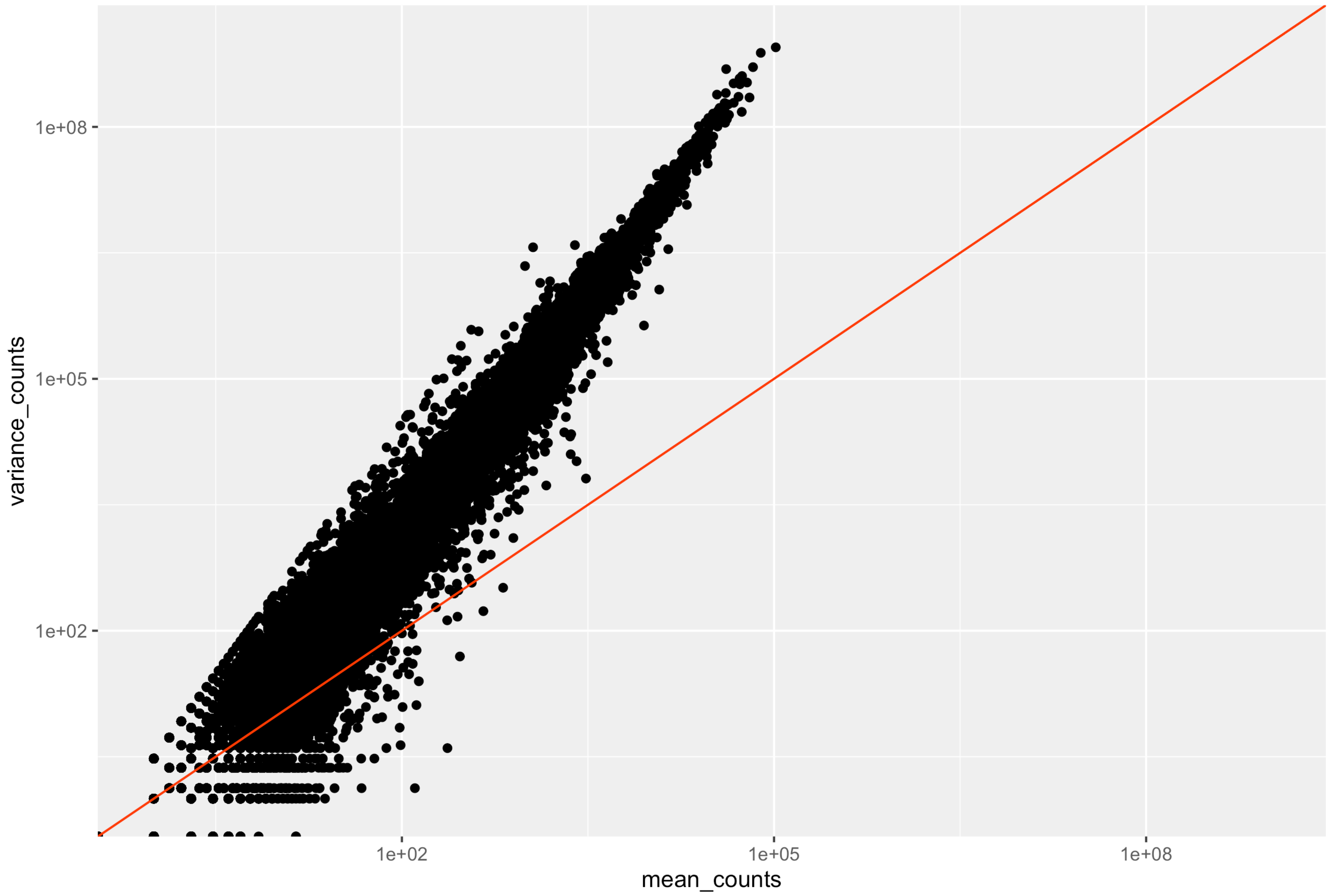
This plot illustrates some **common features** of RNA-seq count data:

- a low number of counts associated with a large proportion of genes
- a long right tail due to the lack of any upper limit for expression
- large dynamic range

Looking at the shape of the histogram, we see that it is *not normally distributed*.

Statistical Aspects of Differential Expression Analysis

Characteristics of RNA-seq data

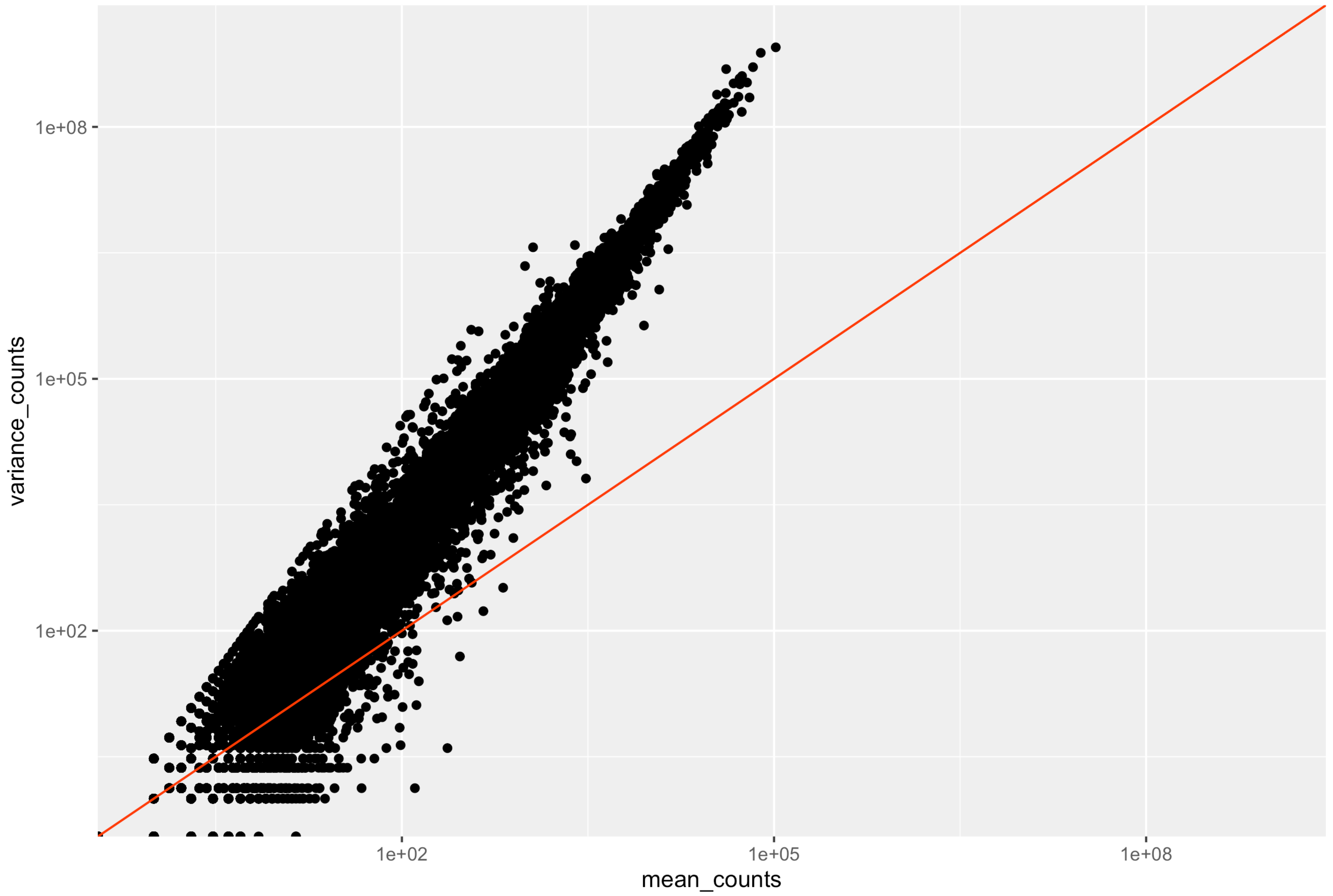


To assess the properties of the data we are working with, we can look at the mean-variance relationship.

For the genes with **high mean expression**, the variance across replicates tends to be greater than the mean (scatter is above the red line).

Statistical Aspects of Differential Expression Analysis

Characteristics of RNA-seq data



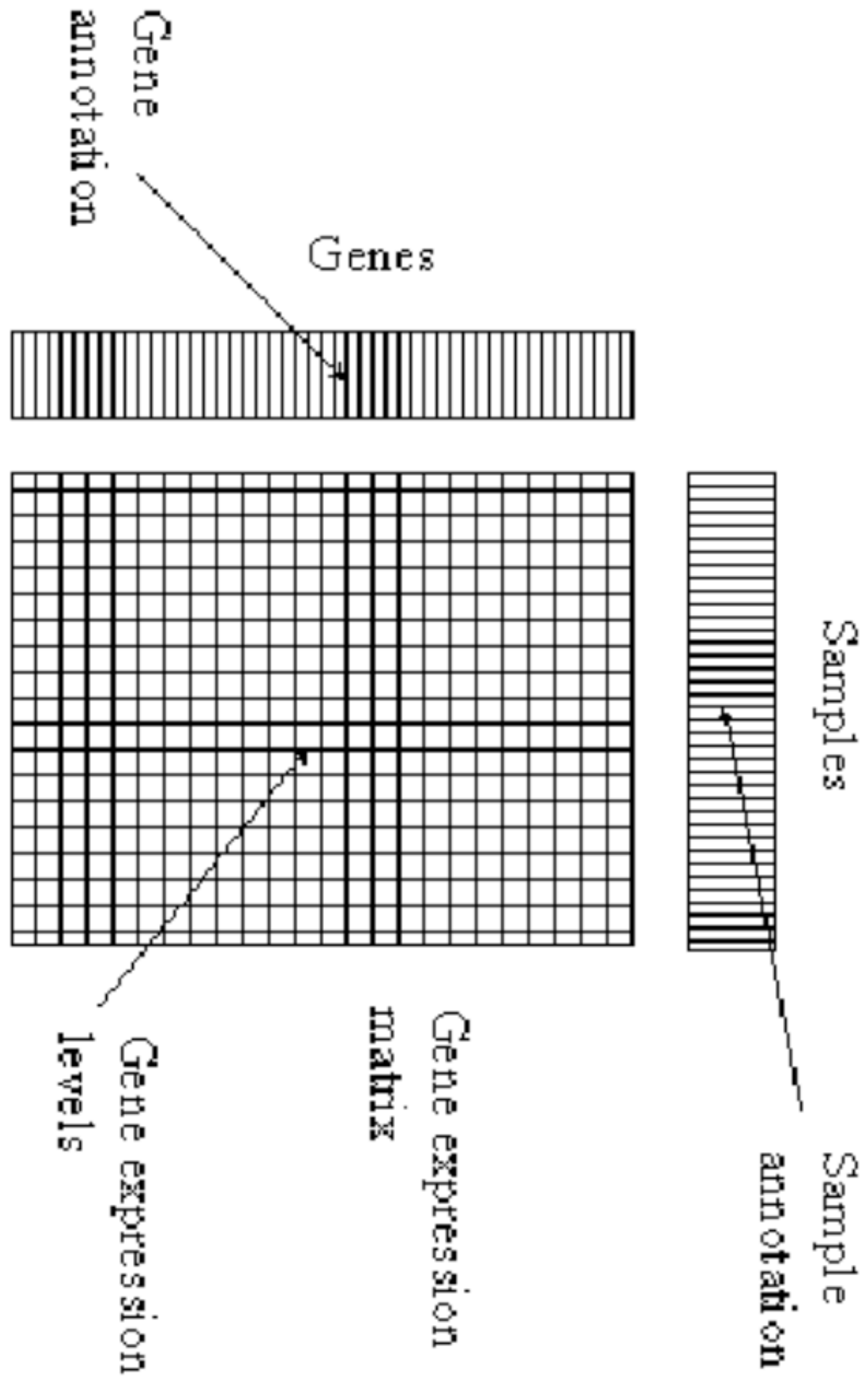
To assess the properties of the data we are working with, we can look at the mean-variance relationship.

For the genes with **high mean expression**, the variance across replicates tends to be greater than the mean (scatter is above the red line).

Essentially, the **Negative Binomial** is a good approximation for data where the mean < variance, as is the case with RNA-Seq count data.

Statistical Aspects of Differential Expression Analysis

Negative Binomial Regression



$$\mathbf{y} \sim \mathsf{NB}(\boldsymbol{\mu}, \phi)$$

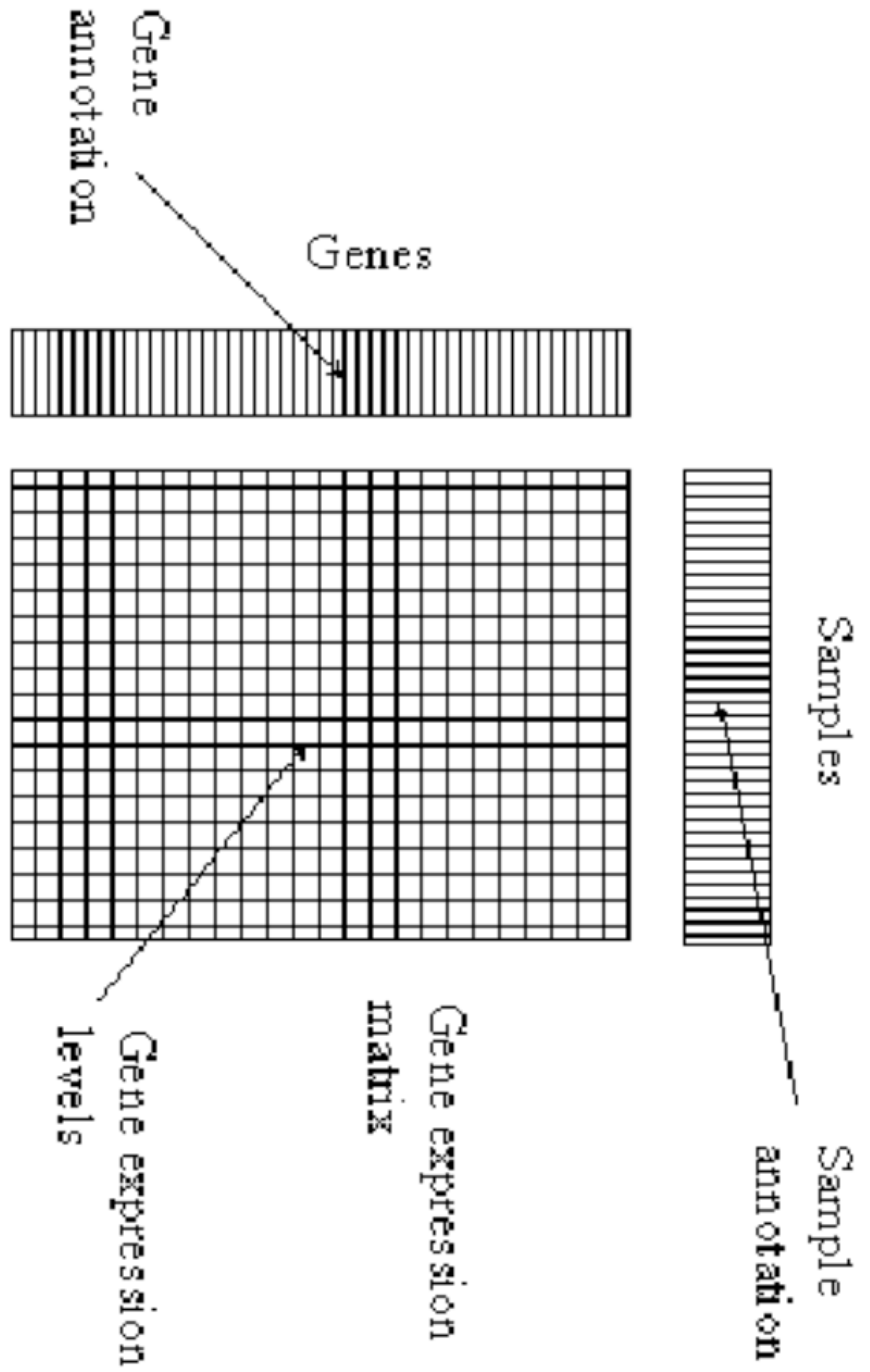
$$\mathsf{E}[\mathbf{y}] = \boldsymbol{\mu} = \mathbf{s}\ 2^{\mathbf{X}\boldsymbol{\beta}}$$

where

- $\mathbf{y}$ denotes the $(n \times 1)$ **count** vector of expression intensities of a given gene,
- $\mathbf{X}$ denotes the $(n \times p)$ design/predictor matrix,
- $\boldsymbol{\beta}$ denotes the $(p \times 1)$ parameter vector,
- ϕ denotes the dispersion parameter,
- $\mathbf{s}$ denotes the scaling factor vector (library size),
- $\mathsf{E}[\mathbf{y}] = \boldsymbol{\mu}$ denotes the expectation of $\mathbf{y}$

Statistical Aspects of Differential Expression Analysis

Negative Binomial Regression



$$\mathbf{y} \sim \mathrm{NB}(\boldsymbol{\mu}, \phi)$$

$$\mathrm{E}[\mathbf{y}] = \boldsymbol{\mu} = \mathbf{s}\ 2^{\mathbf{X}\boldsymbol{\beta}}$$

where

- $\mathbf{y}$ denotes the $(n \times 1)$ **count** vector of expression intensities of a given gene,
- $\mathbf{X}$ denotes the $(n \times p)$ design/predictor matrix,
- $\boldsymbol{\beta}$ denotes the $(p \times 1)$ parameter vector,
- ϕ denotes the dispersion parameter,
- $\mathbf{s}$ denotes the scaling factor vector (library size),
- $\mathrm{E}[\mathbf{y}] = \boldsymbol{\mu}$ denotes the expectation of $\mathbf{y}$

After the model is fit, coefficients are estimated for each sample group along with their standard error. The coefficents are the estimates for the log2 fold-changes, and will be used as input for hypothesis testing.

Statistical Aspects of Differential Expression Analysis

Negative Binomial Regression

Statistical Aspects of Differential Expression Analysis

Negative Binomial Regression

Recall the simple linear regression model for expression:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Statistical Aspects of Differential Expression Analysis

Negative Binomial Regression

Recall the simple linear regression model for expression:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

- where X=0 (untreated)

  or X=1 (treated)

- y is the observed "expression" of the gene
- ε is the measurement noise term
- The parameter of interest is $\beta 1$ (the treatment effect)

Statistical Aspects of Differential Expression Analysis

General Hypothesis

Statistical Aspects of Differential Expression Analysis

General Hypothesis

- Is the RNA abundance level for any of the *m* genes affected by treatment

Statistical Aspects of Differential Expression Analysis

General Hypothesis

- Is the RNA abundance level for any of the *m* genes affected by treatment
- Let H_{0j} denote the null hypothesis for gene *j*
  - H_{0j}: The RNA abundance level for gene *j* is not affected by treatment
  - H_{1j}: The RNA abundance level for gene *j* is affected by treatment

Statistical Aspects of Differential Expression Analysis

General Hypothesis

- Is the RNA abundance level for any of the *m* genes affected by treatment
- Let H_{0j} denote the null hypothesis for gene *j*
  - H_{0j}: The RNA abundance level for gene *j* is not affected by treatment
  - H_{1j}: The RNA abundance level for gene *j* is affected by treatment
- The global null hypothesis is H_{01} and H_{02} and .. and .. H_{0m} are all true

Statistical Aspects of Differential Expression Analysis

General Hypothesis

- Is the RNA abundance level for any of the *m* genes affected by treatment
- Let H_{0j} denote the null hypothesis for gene *j*
  - H_{0j}: The RNA abundance level for gene *j* is not affected by treatment
  - H_{1j}: The RNA abundance level for gene *j* is affected by treatment
- The global null hypothesis is H_{01} and H_{02} and .. and .. H_{0m} are all true
- The global alternative is H_{11} or H_{12} or .. or .. H_{1m} is true

Statistical Aspects of Differential Expression Analysis

General Hypothesis

- Is the RNA abundance level for any of the *m* genes affected by treatment
- Let H_{0j} denote the null hypothesis for gene *j*
    - H_{0j}: The RNA abundance level for gene *j* is not affected by treatment
    - H_{1j}: The RNA abundance level for gene *j* is affected by treatment
- The global null hypothesis is H_{01} and H_{02} and .. and .. H_{0m} are all true
- The global alternative is H_{11} or H_{12} or .. or .. H_{1m} is true
- In other words, under the alternative at least one of the alternative hypothesis is true

Statistical Aspects of Differential Expression Analysis

General Hypothesis

- Reformulation
- The global null hypothesis: $\beta_{11}=0$ and $\beta_{21}=0$ and $\beta m_1=0$
  - In other words, all of the β_{j1} are equal to zero
- The global alternative is $\beta_{11} \neq 0$ or $\beta_{21} \neq 0$ or ... or $\beta_{m1} \neq 0$
  - In other words, at least one of the β_{j1} is not equal to zero.

Multiplicity Correction

Multiplicity Correction

- A gene with a significance cut-off of $\alpha = 0.05$, means there is a 5% chance it is a false positive.

Multiplicity Correction

- A gene with a significance cut-off of $\alpha = 0.05$, means there is a 5% chance it is a false positive.
- If we test for 20,000 genes for differential expression at $\alpha = 0.05$, we would expect to find 1,000 genes by chance

Multiplicity Correction

- A gene with a significance cut-off of $\alpha = 0.05$, means there is a 5% chance it is a false positive.
- If we test for 20,000 genes for differential expression at $\alpha = 0.05$, we would expect to find 1,000 genes by chance
- If we found 3000 genes to be differentially expressed total, roughly one third of our genes are false positives!

Multiplicity Correction

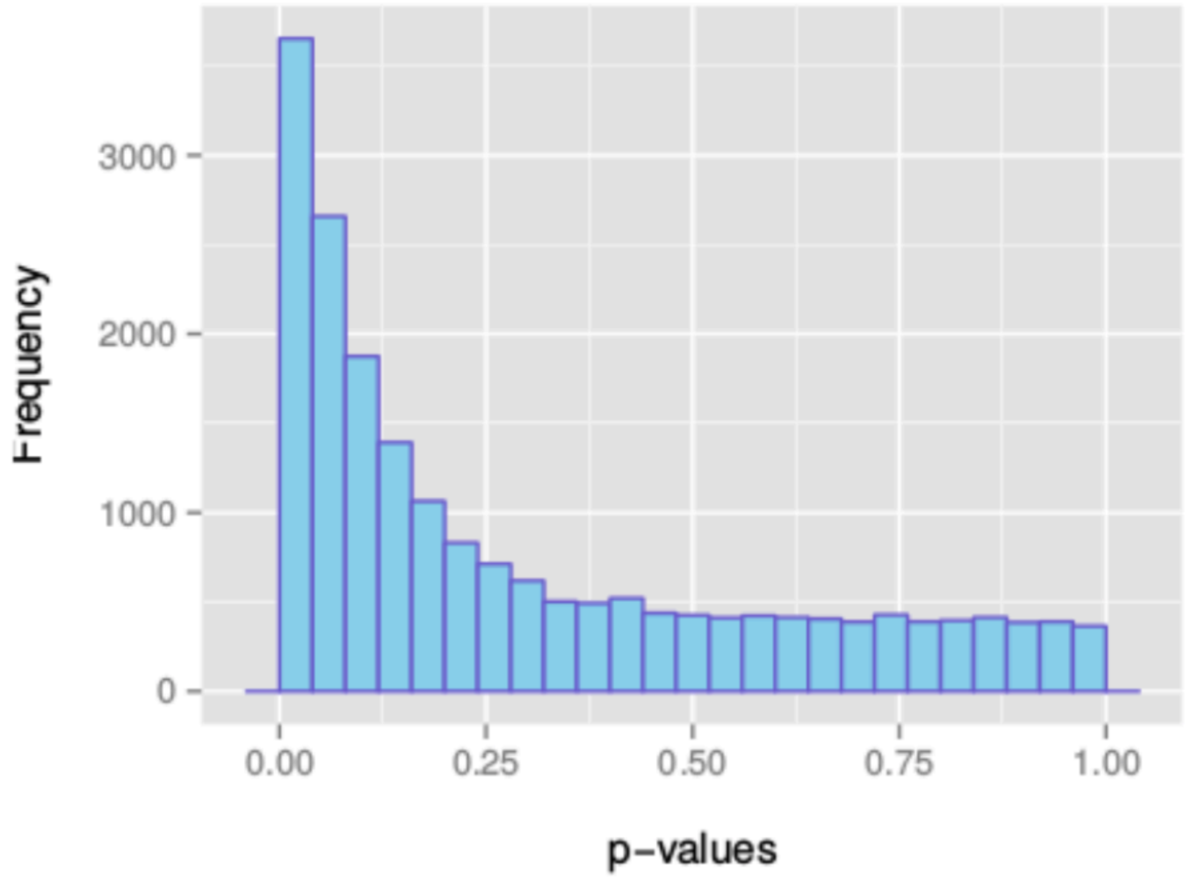
- A gene with a significance cut-off of $\alpha = 0.05$, means there is a 5% chance it is a false positive.
- If we test for 20,000 genes for differential expression at $\alpha = 0.05$, we would expect to find 1,000 genes by chance
- If we found 3000 genes to be differentially expressed total, roughly one third of our genes are false positives!
- The more genes we test, the more we inflate the false positive rate. This is the multiple testing problem.

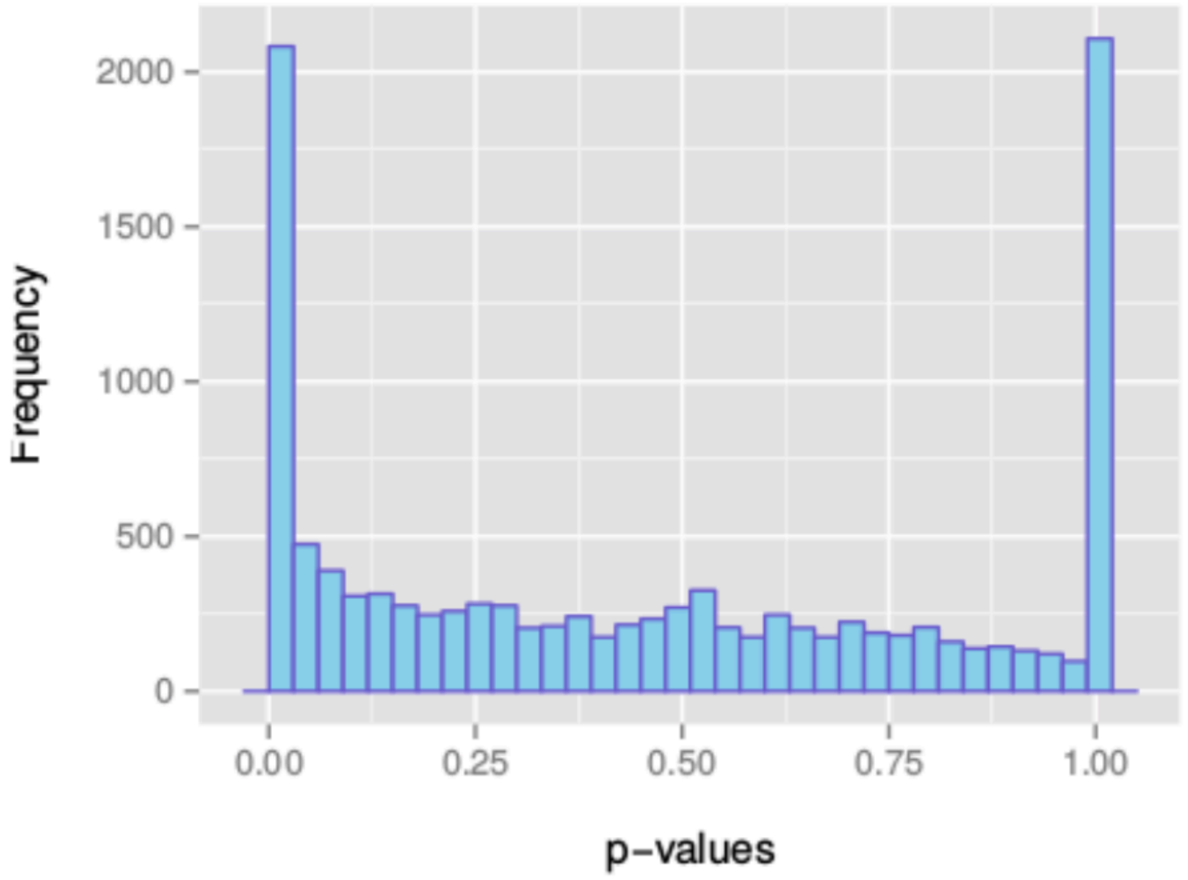
Multiplicity Correction

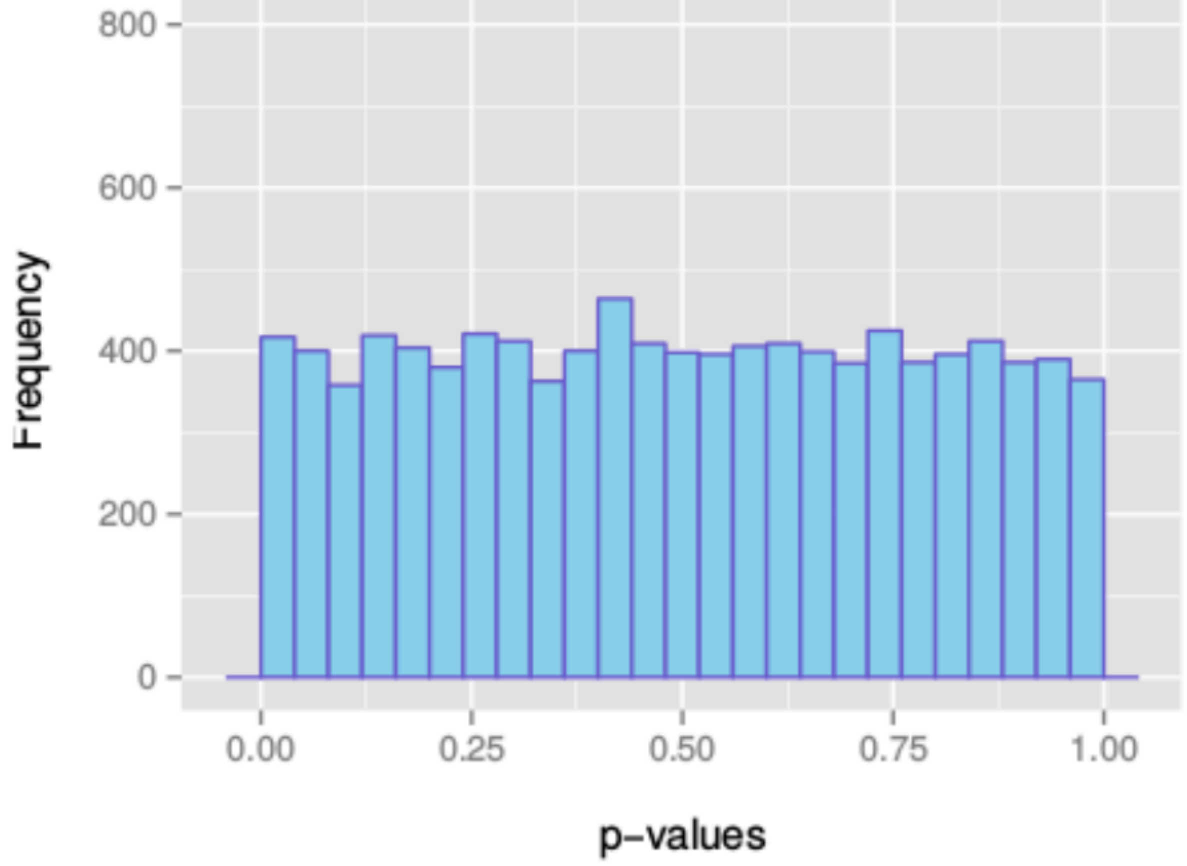
- **Bonferroni**: The adjusted p-value is calculated by: α * k (k = total number of tests). This is a very conservative approach
- **FDR/Benjamini-Hochberg:** Benjamini and Hochberg (1995) defined the concept of FDR and created an algorithm to control the expected FDR below a specified level given a list of independent p-values.

Multiplicity Correction

Examples of expected overall distribution







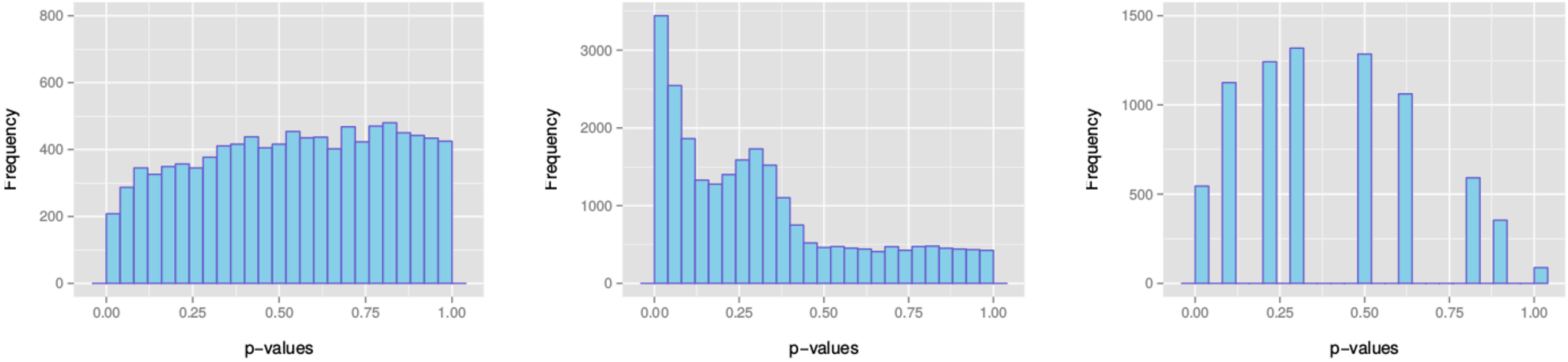
(a) : the most desirable shape

(b) : very low counts genes usually have large p-values

(c) : do not expect positive tests after correction

Multiplicity Correction

Examples of unexpected overall distribution



(a) : indicates a batch effect (confounding hidden variables)

(b) : the test statistics may be inappropriate (due to strong correlation structure for instance)

(c) : discrete distribution of p-values : unexpected

Conclusions

- Assumptions assumptions assumptions