Day 02

Uncertainty, Error, Hypothesis testing

Uncertainty, error

- Standard deviation
- Standard error
- Confidence interval

Hypothesis testing

- Definition of steps
- Simulating the null hypothesis

Arjun Krishnan | **arjun**@msu.edu | the**krishnan**lab.org | @**comp**biologist

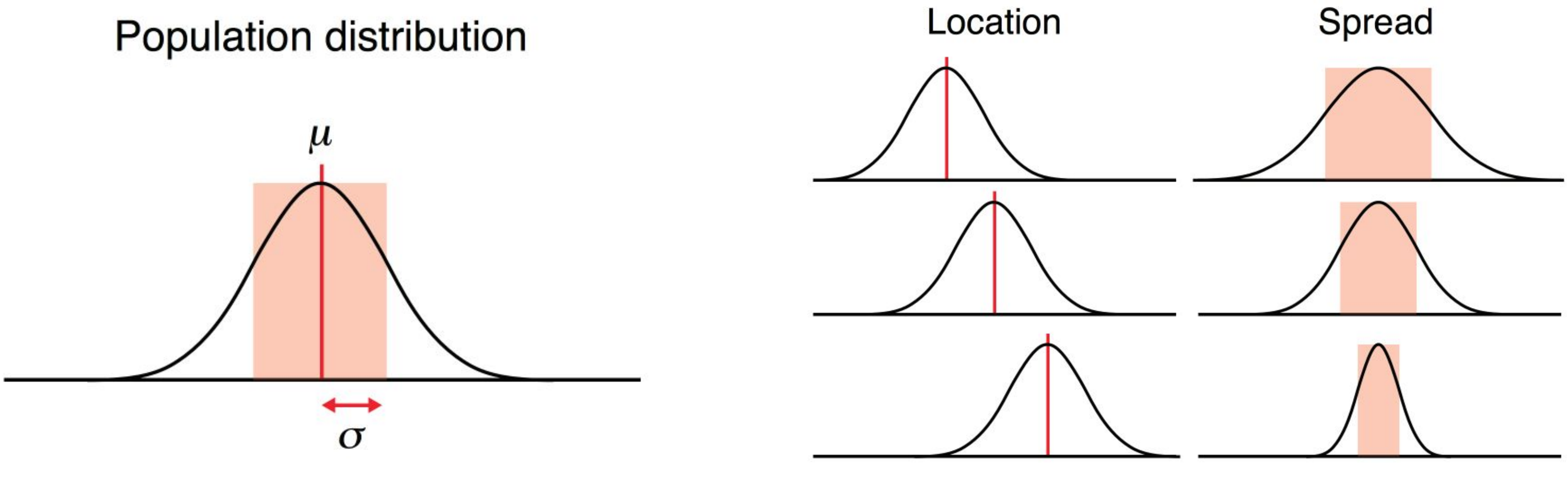
Whether we are right vs. the chances of being wrong

Repeated measurements → Range of values.

Statistics helps us by helping with:

- Modeling the role of chance
- Represent data as estimates with errors

Population distribution

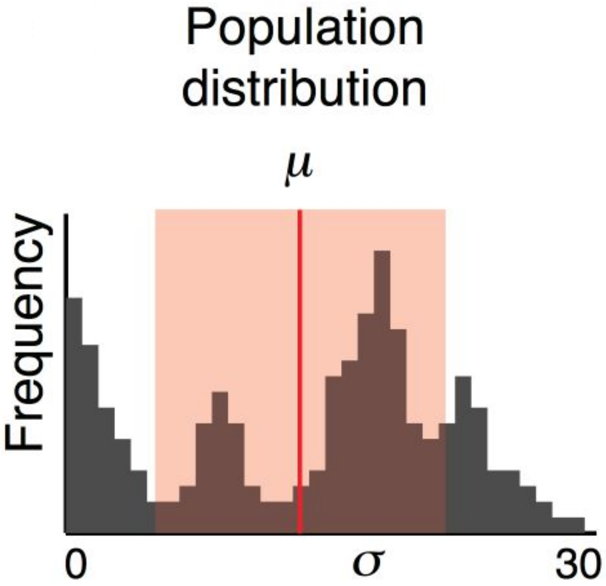


μ: Population mean | σ: Population standard deviation

These are, of course, hard to calculate because it is hard to collect data about the entire population.

https://www.nature.com/articles/nmeth.2613

Estimating population parameters by sampling



Samples

X_1 = [1,9,17,20,26]

X_2 = [8,11,16,24,25]

X_3 = [16,17,18,20,24]

...

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

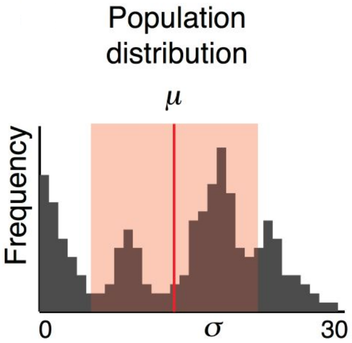
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

https://www.nature.com/articles/nmeth.2613

Standard deviation

- Error bars based on **s.d.** → spread of your data.
- Useful as predictors of the range of new samples.
- Only indirectly supports visual assessment of differences in values:
  - **s.d.** bars reflect the variation of the data
  - They do not reflect the error in your measurement.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Samples

X_1 = [1,9,17,20,26]

X_2 = [8,11,16,24,25]

X_3 = [16,17,18,20,24]

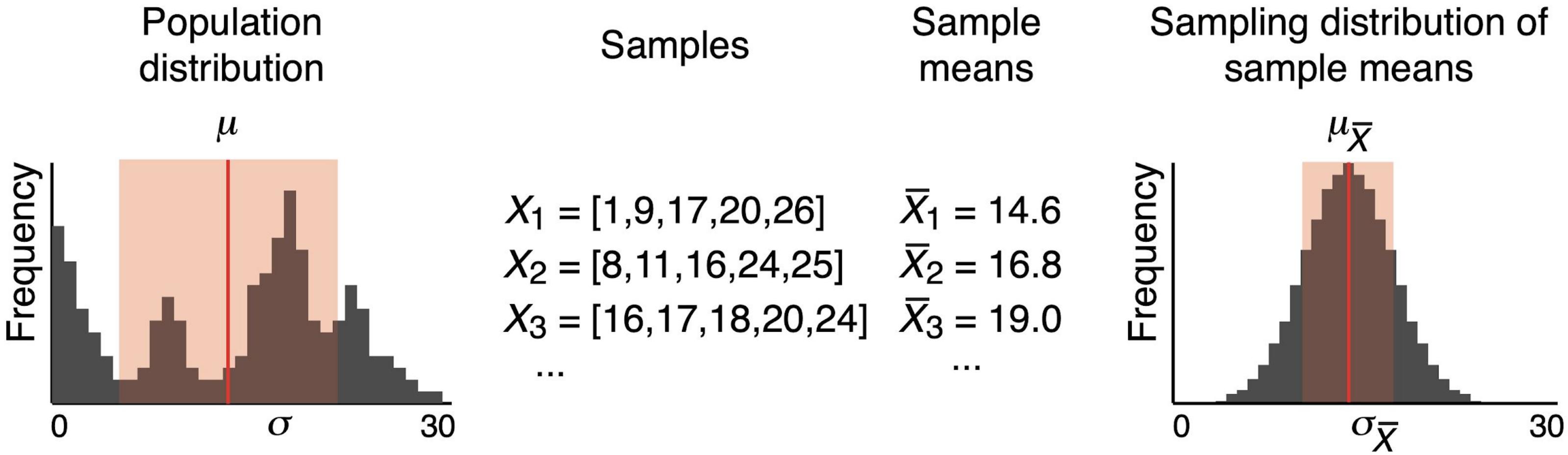
...

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

https://www.nature.com/articles/nmeth.2659

Standard error of the mean

- Error bars based on **s.e.m.** → spread of the ***means*** of independent measurement samples, not the sample you collected (your data).
- s.e.m. = standard deviation of the means



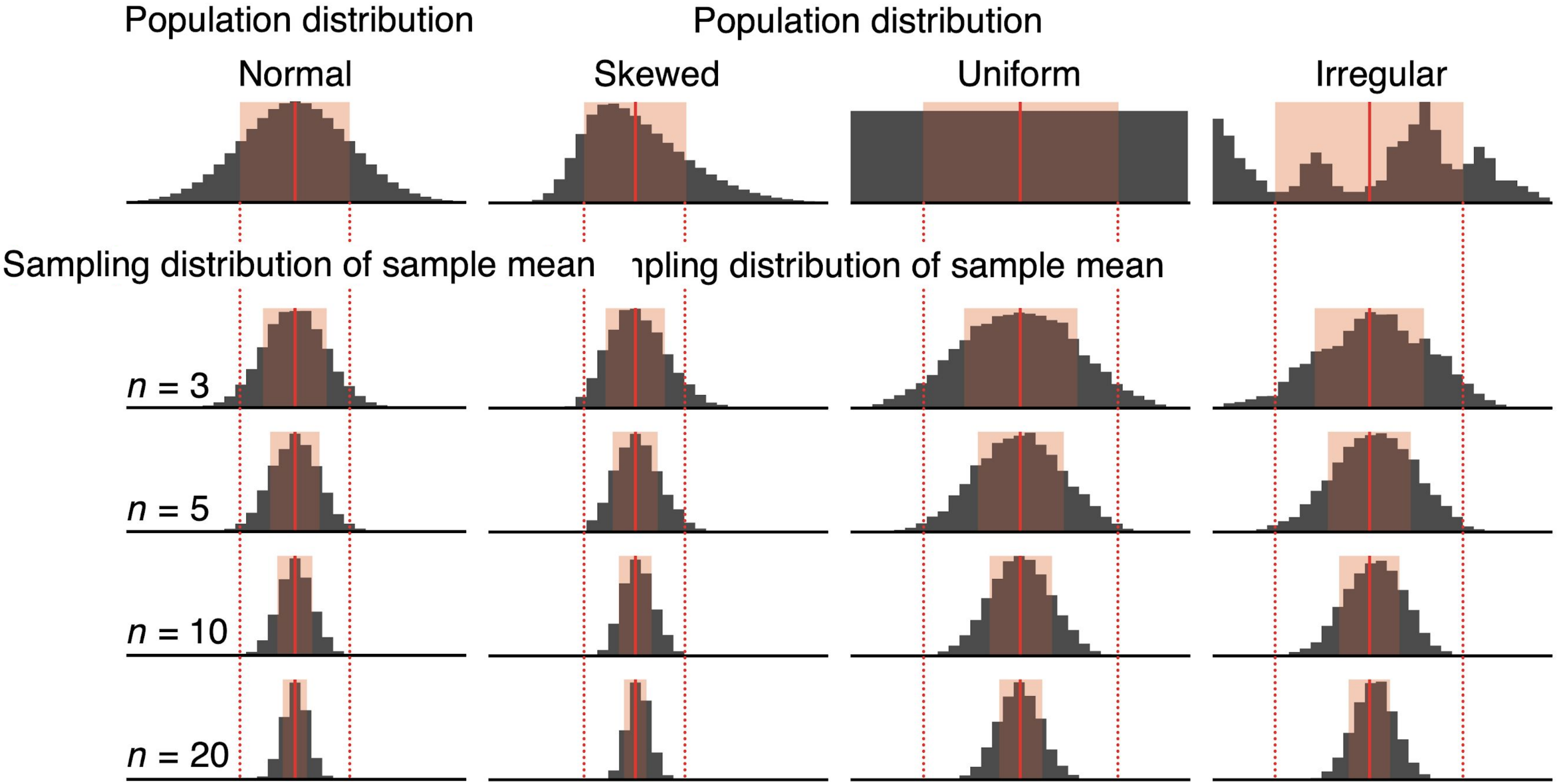
https://www.nature.com/articles/nmeth.2659

Standard error of the mean

- Error bars based on **s.e.m.** → spread of the ***means*** of independent measurement samples, not the sample you collected (your data).
- s.e.m. = standard deviation of the means
- s.e.m. << s.d. of individual samples
- In rare cases, can be estimated using a formula: s.e.m. = s.d. / √n
  - Rest of the times, use bootstrapping.
- Dependent on sample size:
  - Shrinks as we perform more measurements.

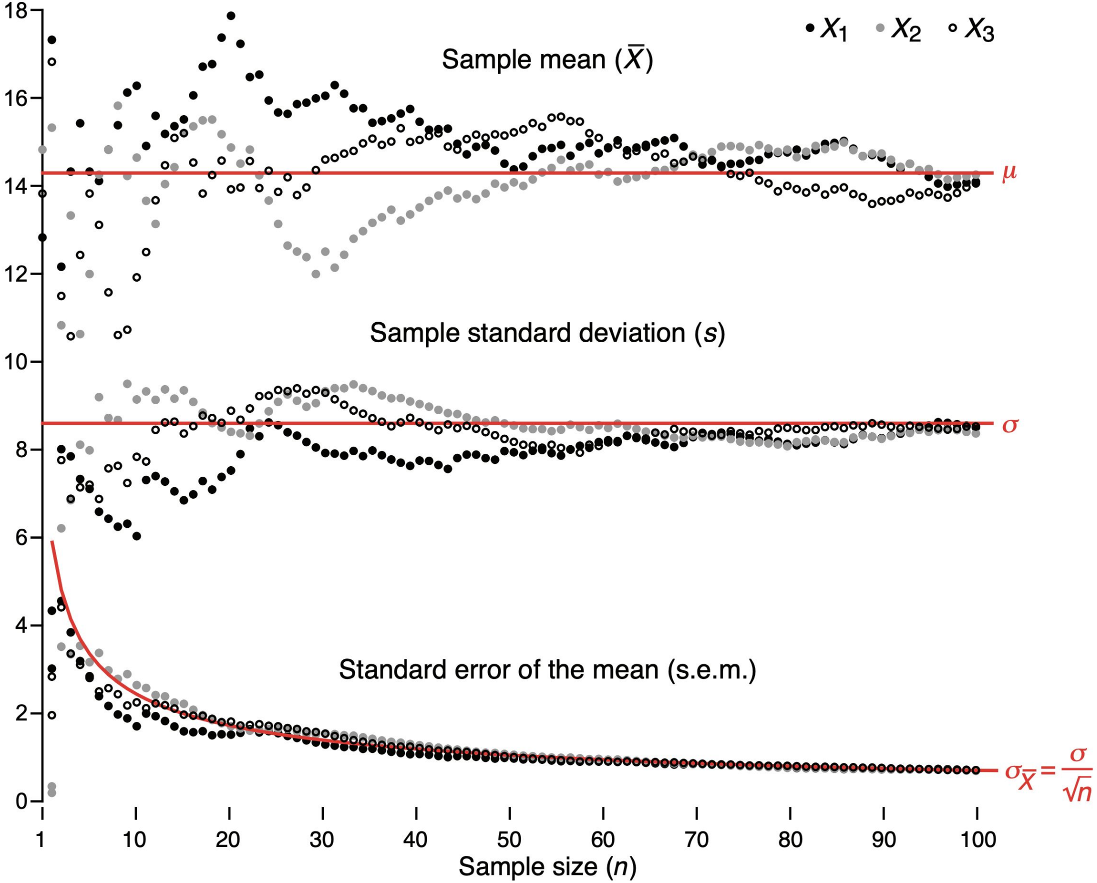
https://www.nature.com/articles/nmeth.2659

Standard error of the mean



https://www.nature.com/articles/nmeth.2659

Standard error of the mean



https://www.nature.com/articles/nmeth.2659

Let’s write code to calculate mean, s.d., and s.e.m. of a sample

Instructions

1. Generate 1000 random numbers from a normal distribution with mean = 0 & s.d. = 1.
   Let these 1000 numbers represent the population.
2. Randomly choose 10 numbers from these 1000.
   These 10 numbers represent a sample from the population.
3. Calculate the sample **mean**.
4. Calculate the sample **s.d.**
5. Calculate **s.e.m.** using the formula (s.d. / √n).

Let’s write code to empirically calculate s.e.m

Instructions

1. Generate 1000 random numbers from a normal distribution with mean = 0 & s.d. = 1
2. Repeat the following a 100 times:
    1. Randomly sample 10 numbers from the population of 1000 numbers
    2. Record their means
3. Calculate the s.d. of these 100 means.

What does this give you?

Recall: **s.e.m.** → spread of the ***means*** of independent measurement samples.

Let’s write code to empirically calculate s.e.m. of a sample

Given 10 numbers that represent a sample. We have no access to the entire population.

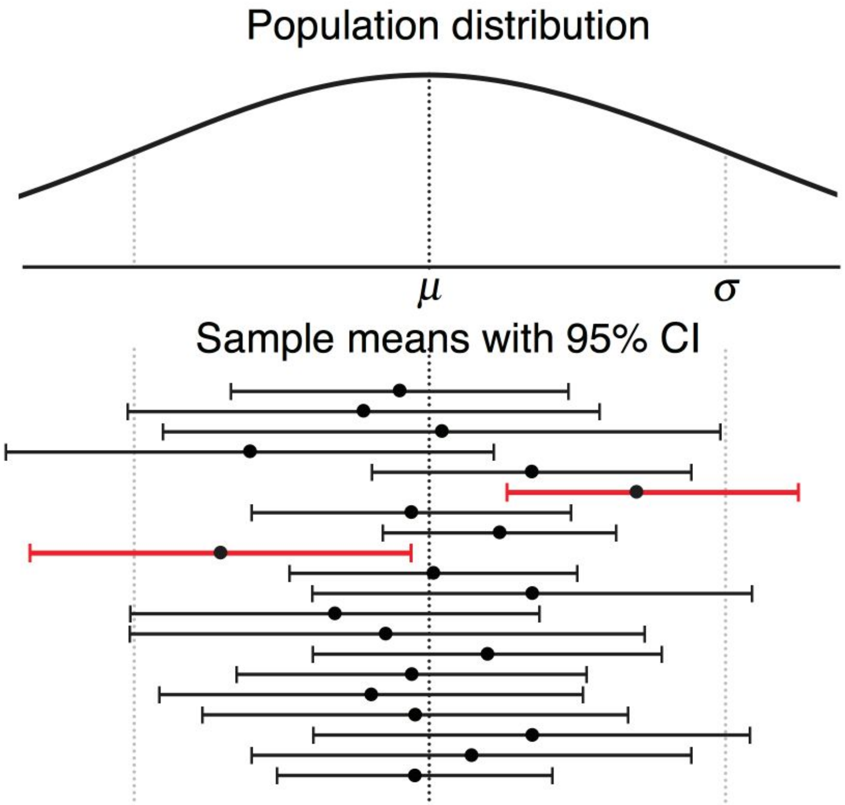
Instructions

1. Create 1000 *bootstrap* samples:
   a. Each time, sample 10 numbers *with replacement*
   b. Calculate the mean of each bootstrap sample
2. Calculate the **s.d.** of these means.

This is the **s.e.m.** of your sample estimated using bootstrapping!

Confidence interval

- **CI** is an interval estimate that indicates the reliability of a measurement.
  - The 95% CI bar captures the population mean 95% of the time.



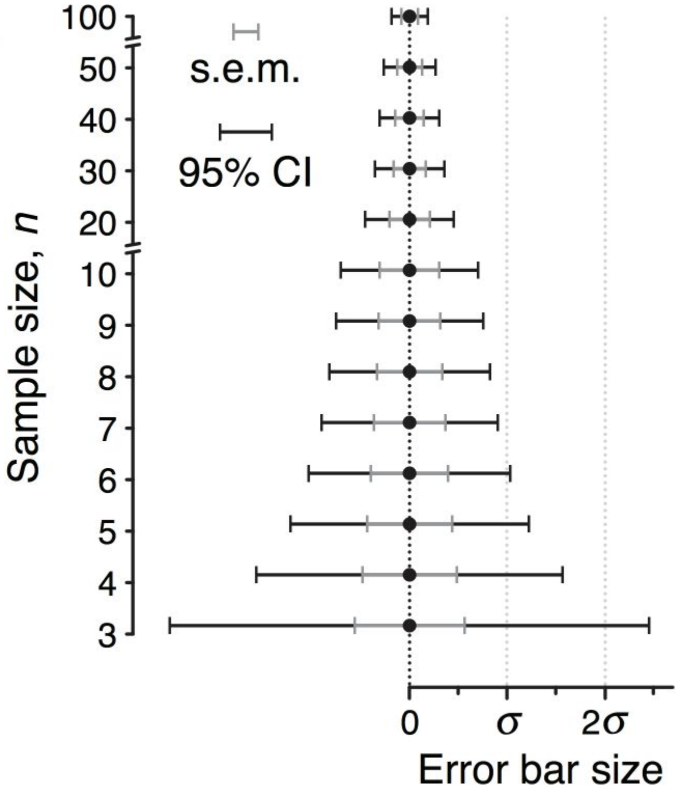
https://www.nature.com/articles/nmeth.2659

Confidence interval

- **CI** is an interval estimate that indicates the reliability of a measurement.
    - The 95% CI bar captures the population mean 95% of the time.
- Error bars based on CI → related to the standard error (s.e.m.)
    - Both can be calculated using a bootstrapping technique (works for $n \geq 10$).
        - Randomly sample *n* measurements from sample *with* replacement.
        - Calculate means
        - Calculate s.e.m. and 95% CI

https://www.nature.com/articles/nmeth.2659

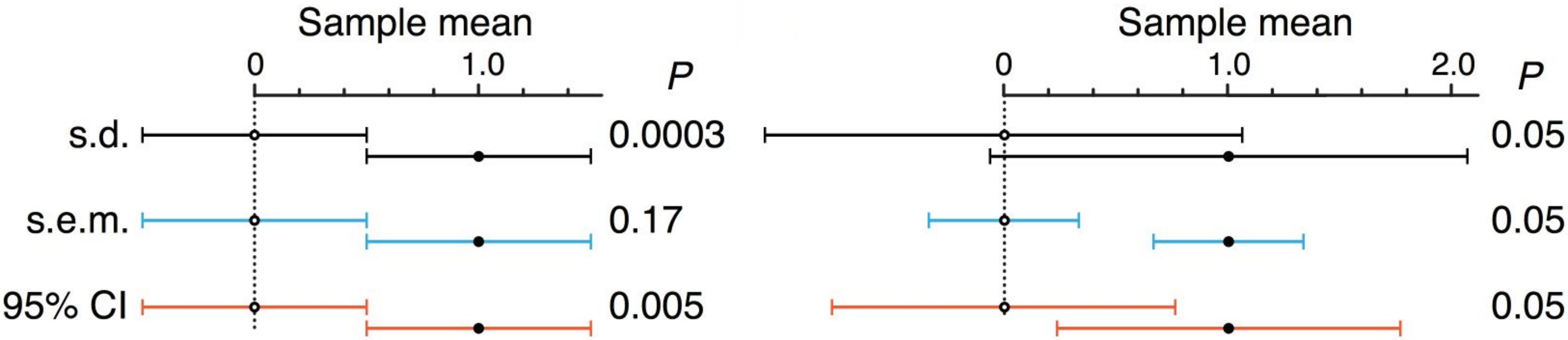
Confidence interval



https://www.nature.com/articles/nmeth.2659

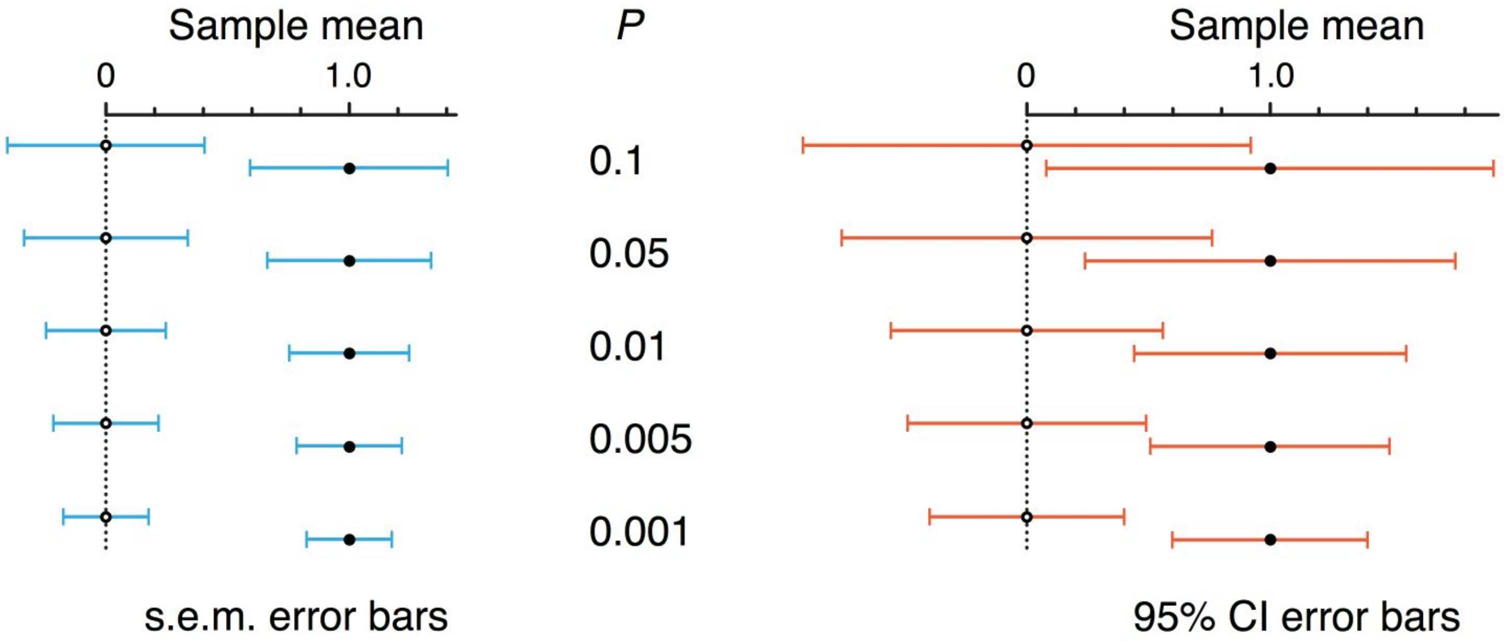
The type of “spread” matters

- Non-overlapping ≠ “significant” difference
- Overlapping ≠ not “significant” difference
- It depends on the type of the error bar.



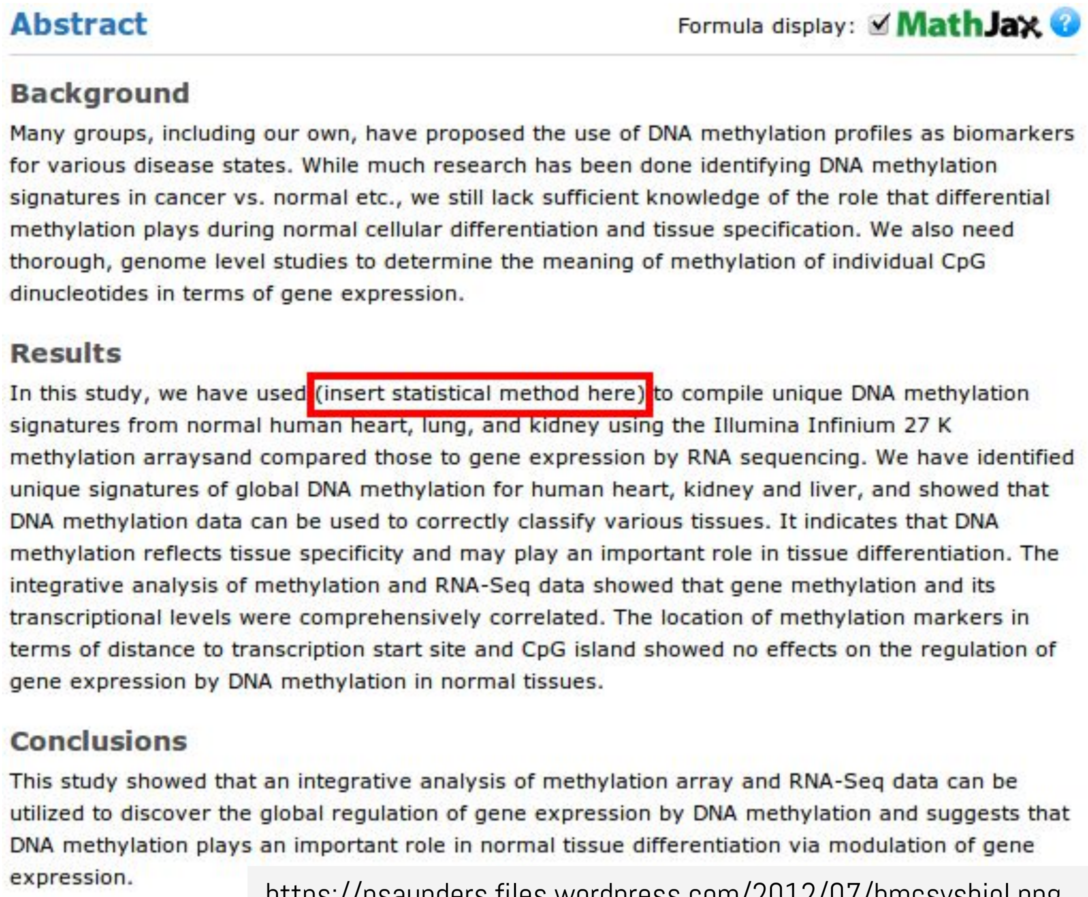
https://www.nature.com/articles/nmeth.2659

Standard error of the mean, Confidence interval



https://www.nature.com/articles/nmeth.2659

Statistical hypothesis testing



Abstract

Background

Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

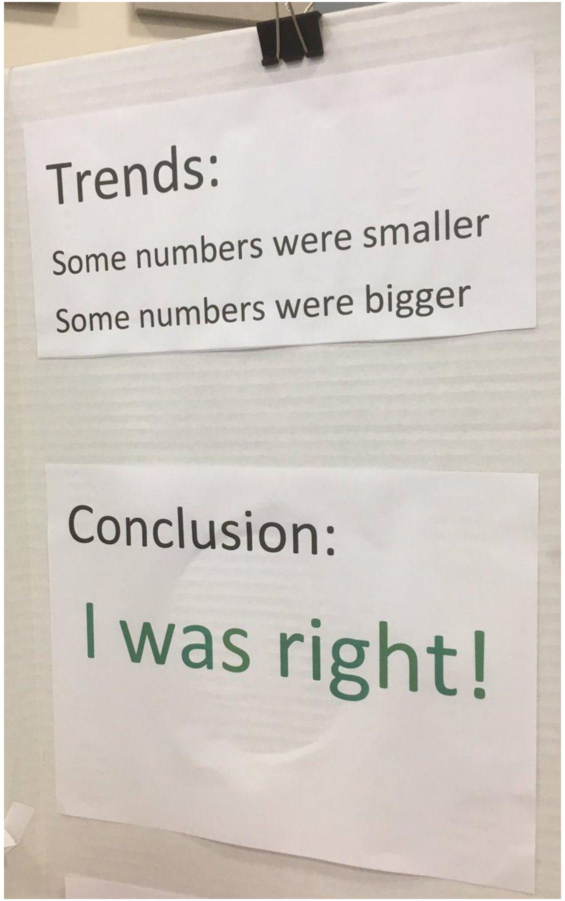
Results

In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arraysand compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

Conclusions

This study showed that an integrative analysis of methylation array and RNA-Seq data can be utilized to discover the global regulation of gene expression by DNA methylation and suggests that DNA methylation plays an important role in normal tissue differentiation via modulation of gene expression.

https://nsaunders.files.wordpress.com/2012/07/bmcsysbiol.png



Trends:

Some numbers were smaller

Some numbers were bigger

Conclusion:

I was right!

Statistical hypothesis testing

- Many scientific studies are interested in quantifying the difference in a particular parameter between two groups.
  - There's always some difference → Is it statistically significant difference?

- Say you're testing the efficacy of a cold medicine:
  - Two groups given placebo/medication
  - Followed-up: how long the cold lasted in each person in both groups
  - Null: Ineffective; Alternative: Effective

Statistical hypothesis testing

1. **Decide on the effect** that you are interested in, design a suitable experiment or study, pick a data summary function and test statistic.
2. **Set up a null hypothesis**, which is a simple, computationally tractable model of reality that lets you compute the null distribution, i.e., the possible outcomes of the test statistic and their probabilities under the assumption that the null hypothesis is true.
3. **Decide on the rejection region**, i.e., a subset of possible outcomes whose total probability is small.
4. **Do the experiment** and collect the data, compute the test statistic.
5. **Make a decision**: reject the null hypothesis – i.e. conclude that it is unlikely to be true – if the test statistic is in the rejection region.