Day 02

Uncertainty, Error, Hypothesis testing

Uncertainty, error

- Standard deviation
- Standard error
- Confidence interval

Hypothesis testing

- Definition of steps
- Simulating the null hypothesis

Before we being...

- Look out for messages on all channels: stagaps2019.slack.com
  - Instructions for next week
  - Blog/Newsletter sign-up sheet – PLEASE SIGN-UP

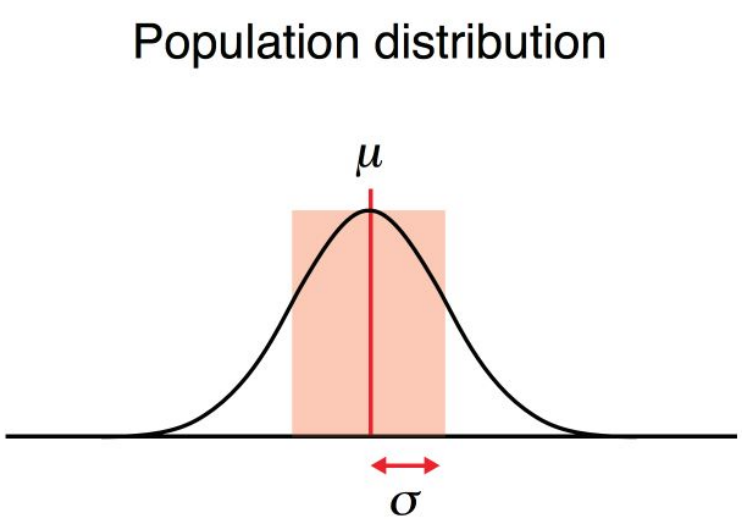
- Fill-in the incoming + self-assessment survey: bit.ly/statgaps2019_incoming
  - If you haven't done so already, fill-in the interest survey: bit.ly/statgaps2019_signup

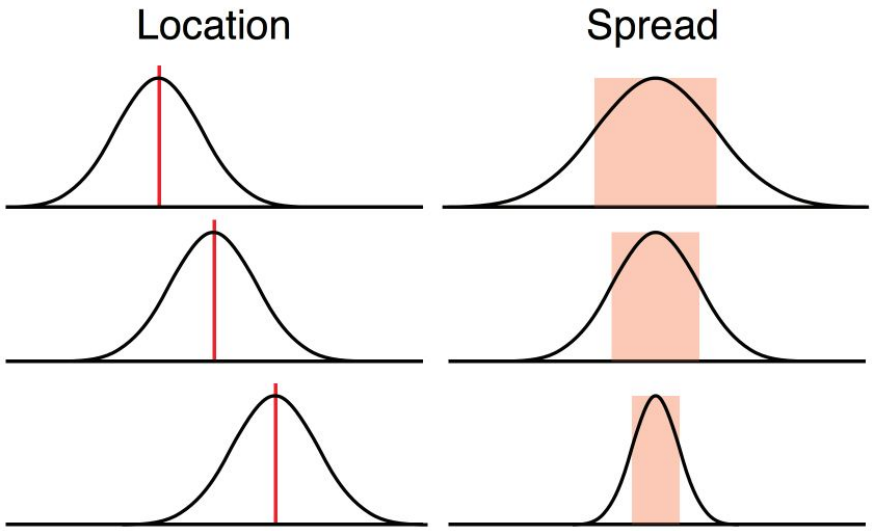
Whether we are right vs. the chances of being wrong

Repeated measurements → Range of values. Statistics helps us by helping with:

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

Population distribution



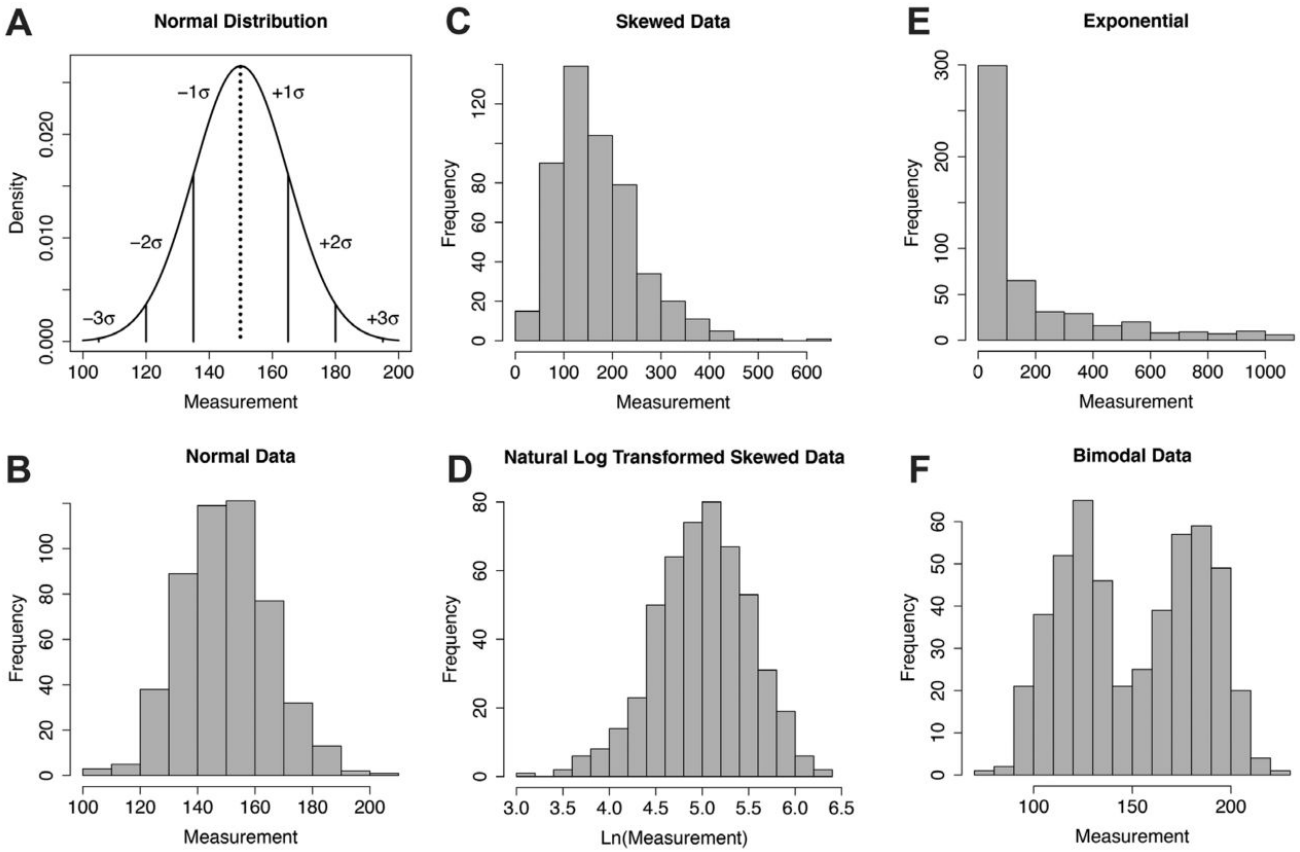


μ: Population mean | σ: Population standard deviation

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

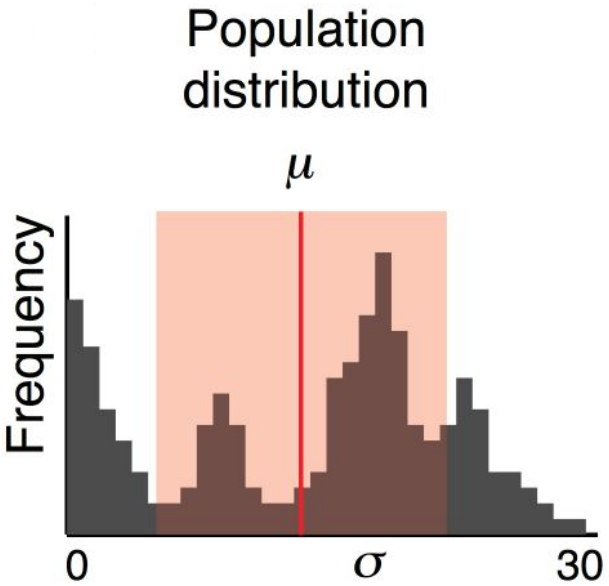
https://www.nature.com/articles/nmeth.2613

Population distribution



https://www.molbiolcell.org/doi/10.1091/mbc.E15-02-0076

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]
...

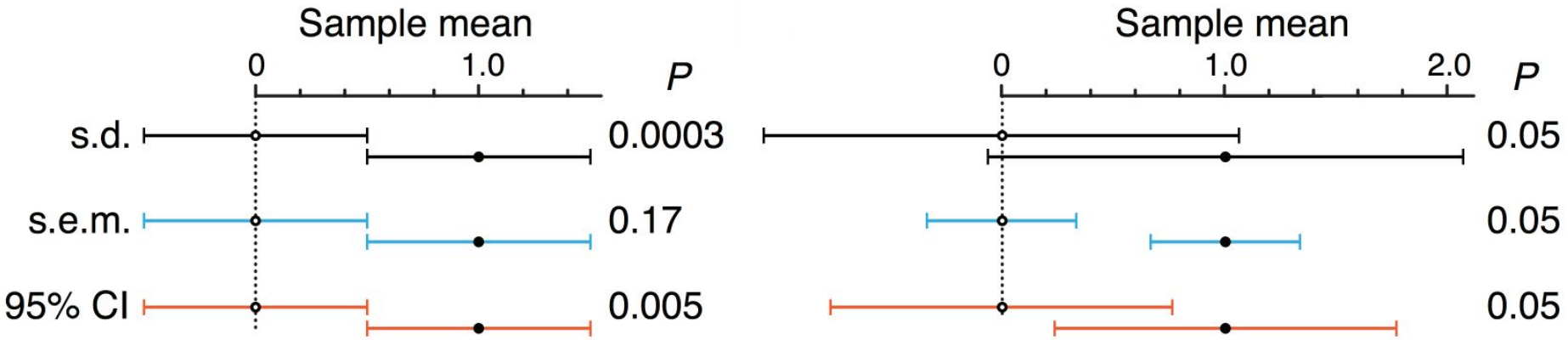
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

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

The type of “spread” matters

- Non-overlapping ≠ “significant” difference
- Overlapping ≠ not “significant” difference



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

Standard deviation

- Error bars based on s.d. → spread of your data.
- Useful as predictors of the range of new samples.
- Only indirectly support visual assessment of differences in values:
  - s.d. bars reflect the variation of the data
  - not the error in your measurement.
- Can be overlapping and the means can still be significantly different.

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

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
- Will be much smaller than the s.d. of individual samples
- In rare cases, can be estimated using a formula: s.e.m. = s.d. / $\sqrt{n}$
    - Rest of the times, use bootstrapping.
- Shrink as we perform more measurements.
- Non-overlapping ≠ “significant” difference

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

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

- R and Python (instructions will be posted on Slack)
  a. Install R, RStudio, and Tidyverse (package); Get familiar with R Notebooks
  b. Install Anaconda, Python 3.7, Jupyter Notebooks

- Instructions
  a. Generate 1000 random numbers from a normal distribution with mean = 0 and s.d. = 1
  b. Randomly sample 10 numbers from these 1000 and calculate their mean & s.d..

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. Randomly sample 10 numbers from these 1000 and calculate their mean & s.d..
3. Calculate s.e.m. using the formula (s.d. / $\sqrt{n}$).
4. Write a loop (100 iterations):
    a. Randomly sample 10 numbers
    b. Keep track of their means
5. Calculate the s.d. of these 100 means

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

Instructions

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

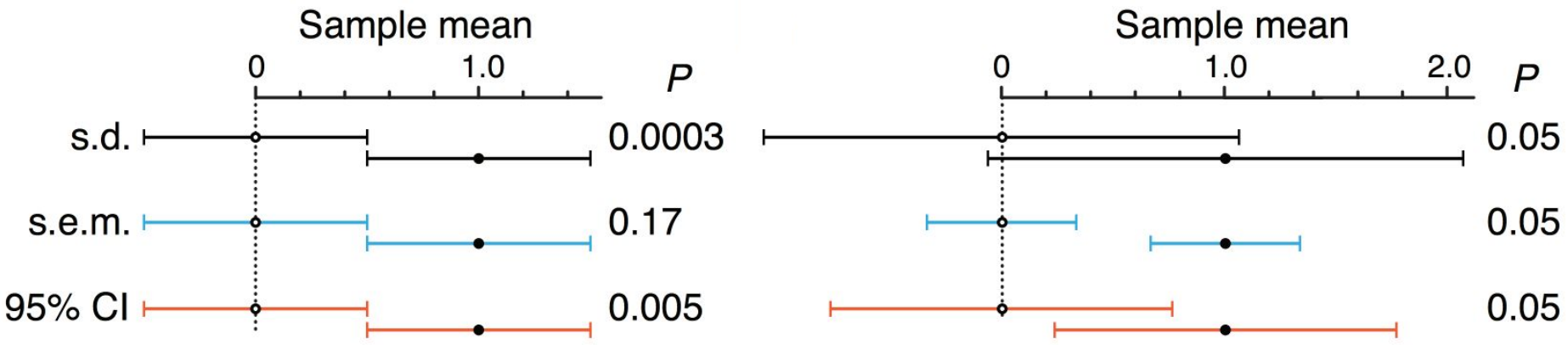
This is your s.e.m.

Confidence interval

- Range of values for a pop. parameter (e.g. mean) that has a high probability of containing the true value based on a sample of measurements.
    - 95% CI for a normally distributed value is expected to contain the true rate in approximately 95 out of 100 repetitions of the experiment.
- Error bars based on CI → related to the standard error (s.e.m)
    - Both can be calculated using a bootstrapping technique (works for n ≥ 10).
        - Randomly sample n measurements from sample *with* replacement.
        - Calculate means
        - Calculate s.e.m. and 95% CI

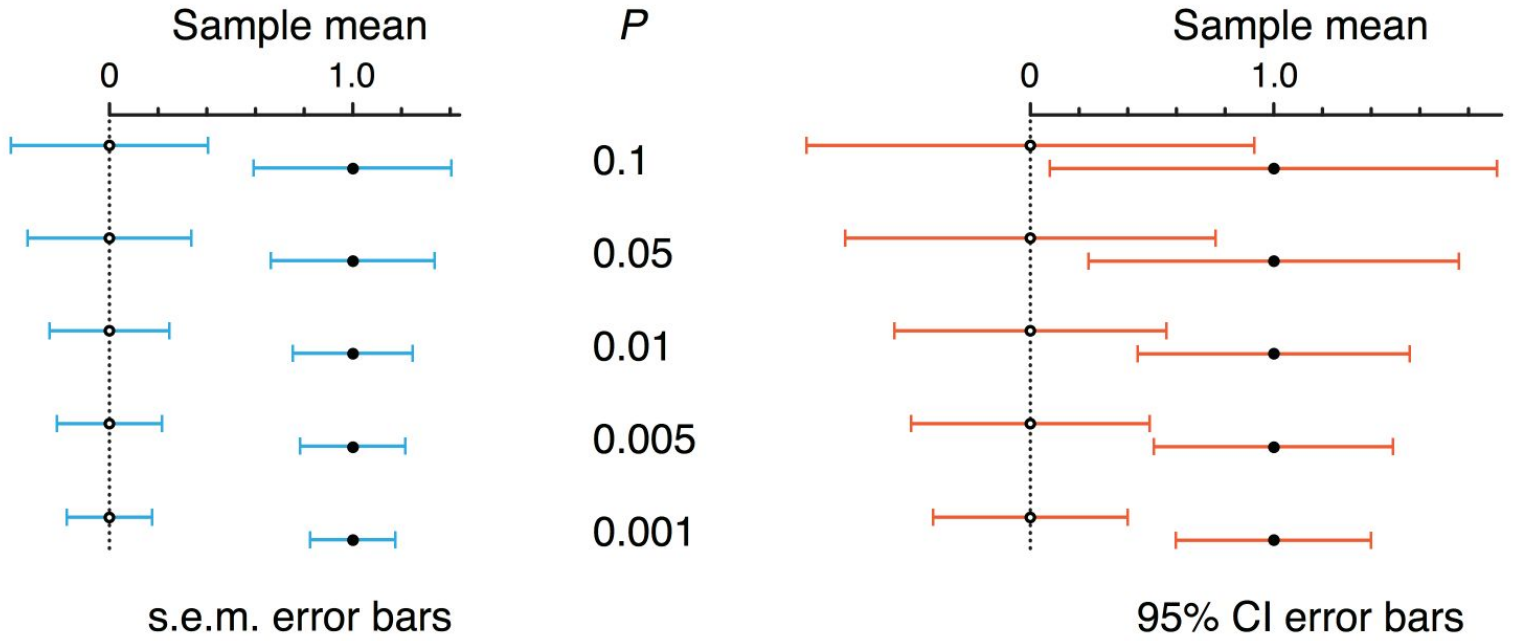
https://www.nature.com/articles/nmeth.2659

Standard error of the mean, Confidence interval



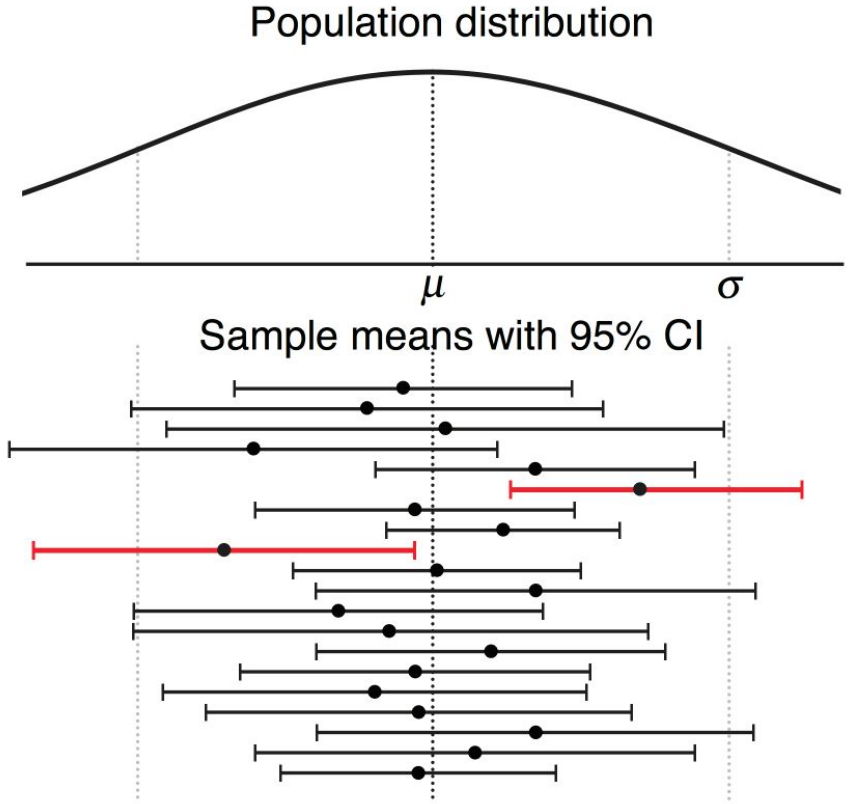
https://www.nature.com/articles/nmeth.2659

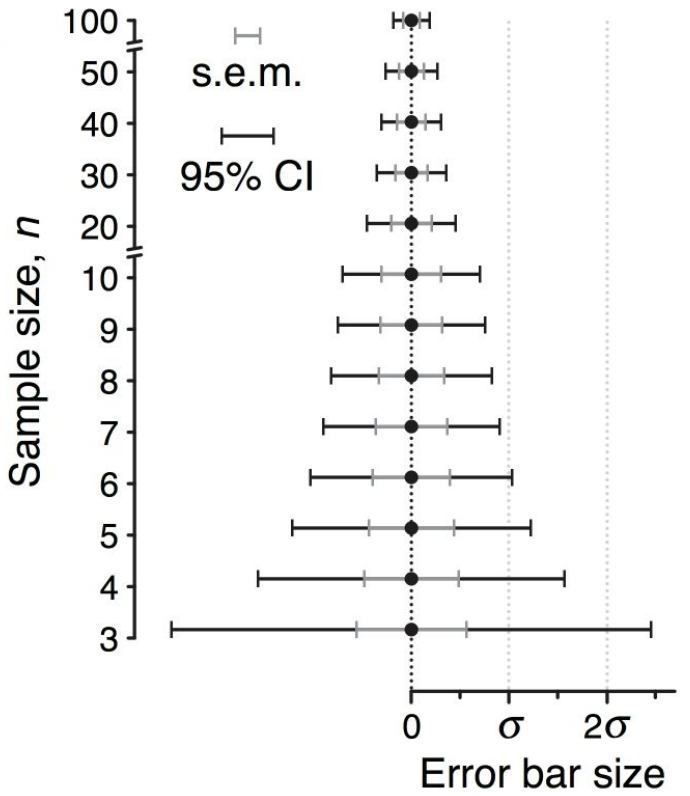
Standard error of the mean, Confidence interval



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

Confidence interval





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

Statistical hypothesis testing

Abstract

Formula display: MathJax

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.

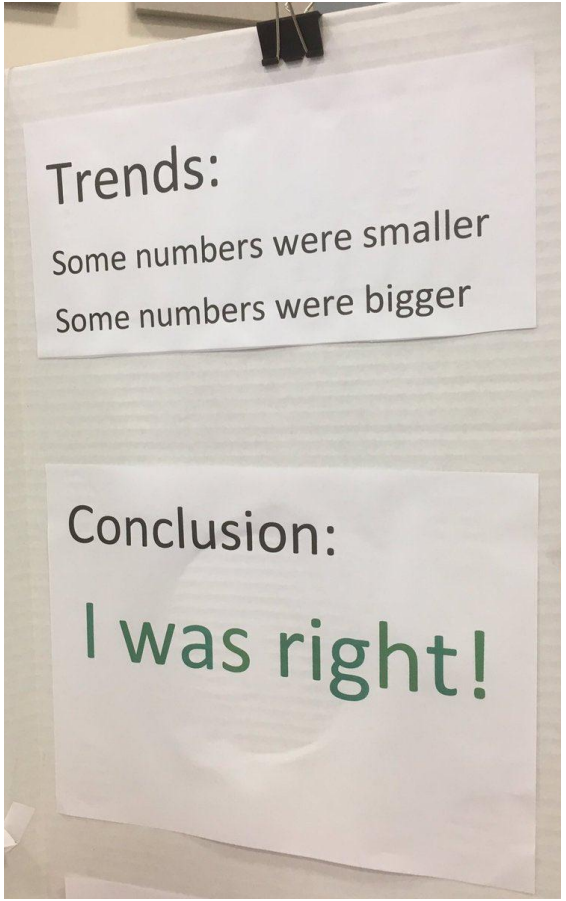
Results

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



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.

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.

Today: Write code to implement these two steps.

1. Effect = Difference in the mean of two distributions:
   a. Two normal distributions: (0, 0.5) and (0.2, 0.5)
2. Set up the null hypothesis for this test statistic.