part 1/4
Fundamental concepts of frequentist statistics

Oussama Abdoun (MEng, PhD) – oussama.abdoun@pm.me

Bayesian Statistics – CRNL – dec 2024

On the menu

Objectives of session 1

- develop a fine understanding of frequentist statistical inference, its interpretations & limitations
- understand the fundamental divergence between frequentist and bayesian statistics

Concepts

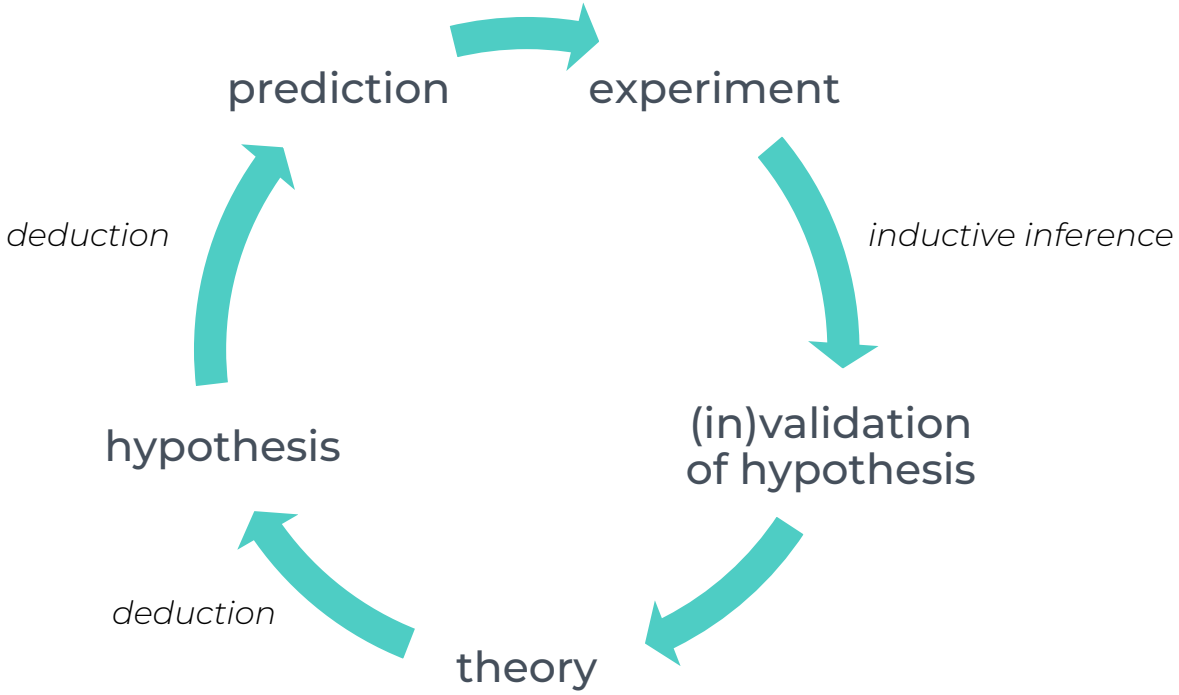
- Statistical inference
- Statistical model
- Parameter estimation
- Uncertainty
- Definitions of probability
- Confidence interval
- Null hypothesis testing & *p*-values
- History of frequentist statistics

Associated code at: **link**

1.

Inference

A model of the scientific method



Induction

uncertain and
particular premisses
⇓
uncertain and general
conclusions

Deduction

(True) statements
⇓
necessary conclusions

The chief guiding principle of both scientific and everyday knowledge [is] that it is possible to learn from experience and to make inferences from it ***beyond*** *the data directly known by sensation. (...)*

Inference is the use of sensations already experienced to derive information about sensations not yet experienced, to construct physical objects, and to describe the past and future of these physical objects.

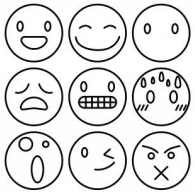
Harold Jeffreys
Scientific Inference (1931)

Everyday (perceptual) inference

inference

internal model

sensory input



inference

internal model

sensory input

inference

internal model

sensory input

Inference in science

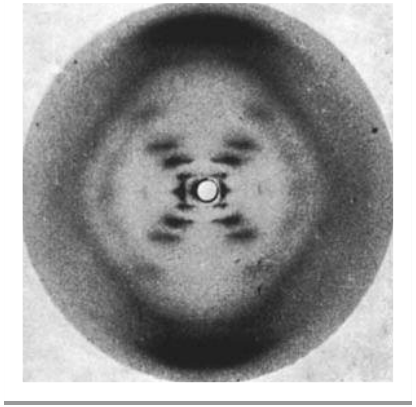
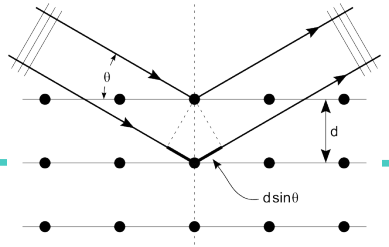
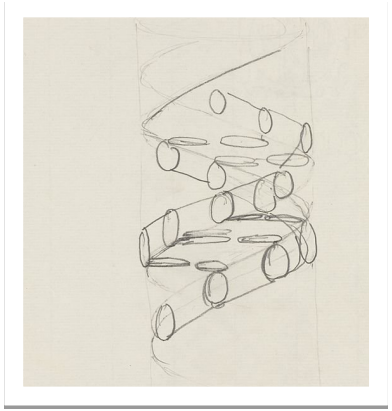


Photo 51, X-ray diffraction, May 1952
Rosalind Franklin & Raymond Gosling



Atomistic model of matter
Wave model of light
Diffraction laws



DNA structure, pencil sketch, 1953
Francis Crick

Inference in science

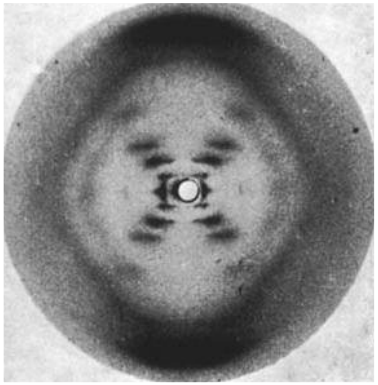
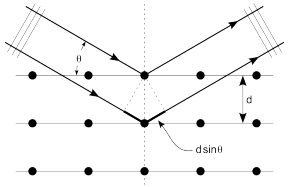
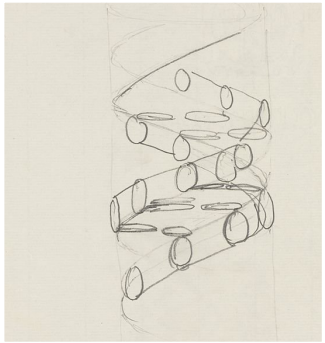


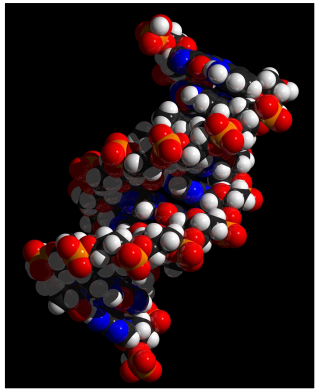
Photo 51, X-ray diffraction, May 1952
Rosalind Franklin & Raymond Gosling

Inference
(“inversion” of the model)

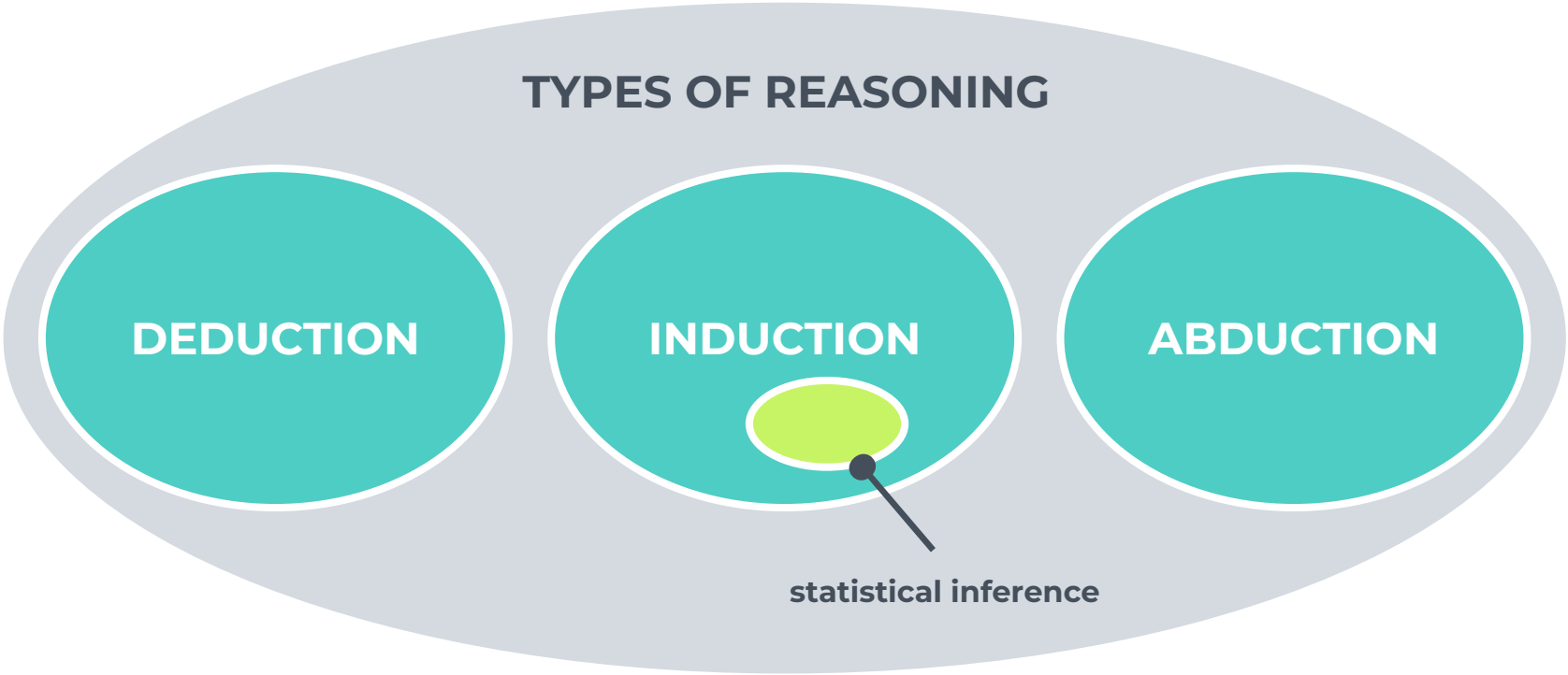


prediction
(“forward” application of the model)



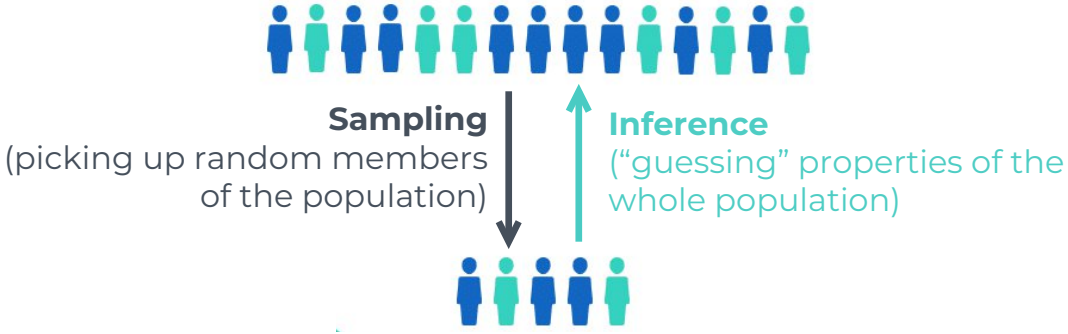


The place of statistical inference in science



Statistical inference

The process of estimating the properties of an underlying population from a sample of data.

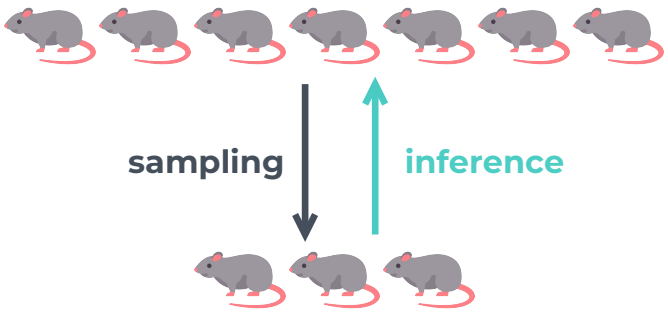


Statistical inference = type of induction
where uncertainty arises from having only *partial data*

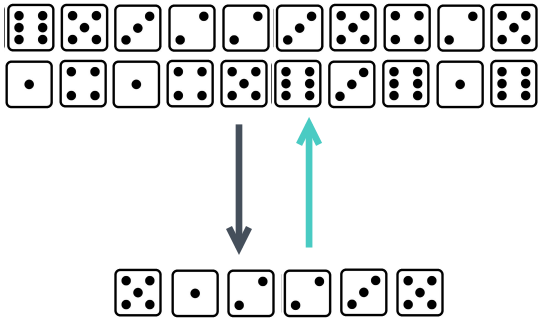
Statistical inference

Sampling and inference are required whenever observing the entire population of interest is impractical, either because:

the population is too **large**



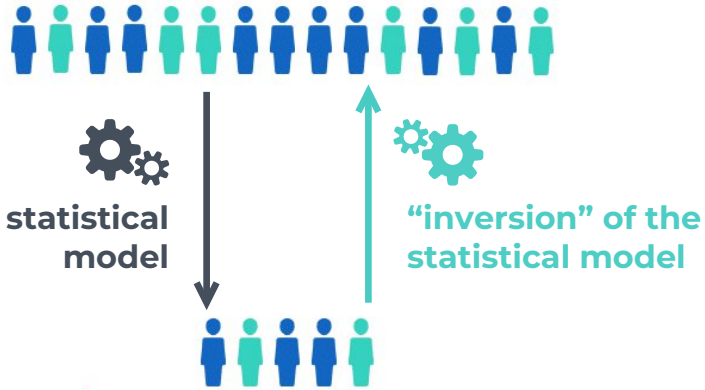
the population is **virtual** (events)



Statistical model

The model that allows statistical inference is called the **statistical model**. It is a **generative model** and makes use of **probability distributions**.

assumption on the
population distribution
+
assumption on the
sampling process*



*usually considered as random, but see Tobit models for truncated samples

2.

Parameter estimation

Notations

***True* population distribution parameters:** θ

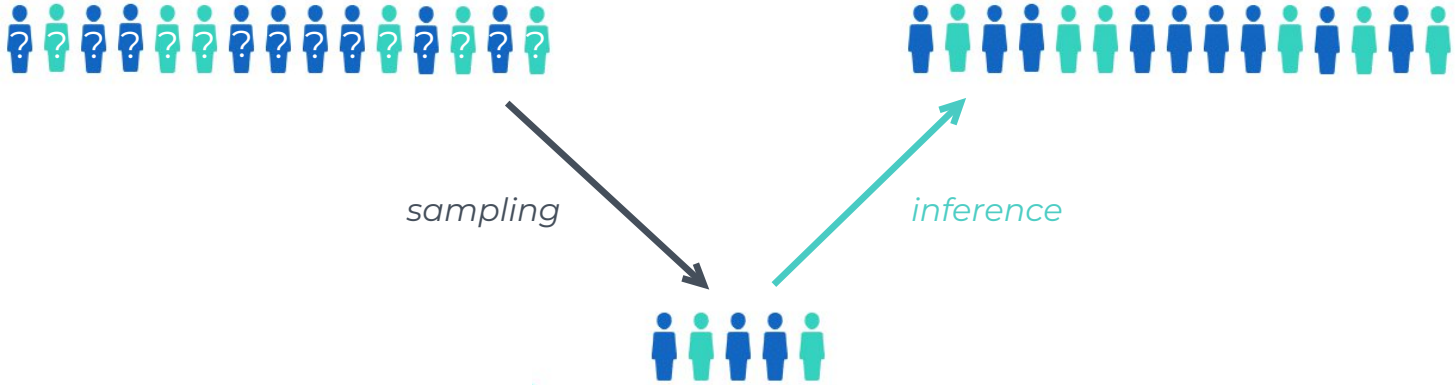
For the normal distribution: $\theta = (\mu, \sigma)$

(unknown FOREVER)

***Estimated* population distribution parameters:** $\hat{\theta}$

For the normal distribution: $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$

(calculated from available data)



Sample values: x_i

Sample statistics: $\overline{X}, s_n^2$

Point estimates: an example

You find 17 specimens of a curious species, the planarian worms. Laser measurements yielded the following data for the length of the planarians (in mm):

14.976, 7.04, 11.588, 14.46, 6.876, 9.258, 10.444, 5.425, 12.508, 11.844, 7.695, 4.642, 6.916, 2.947, 10.423, 9.558, 10.783

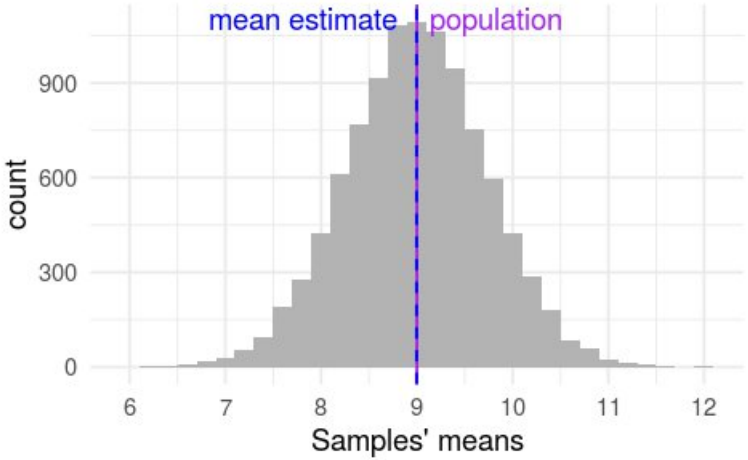
What should we do now?

- Visualize the data
- Describe the statistical properties of the sample
- Specify the statistical model: $X \sim N(\mu, \sigma^2)$
- Estimate the parameters of the model (i.e. the properties of the underlying population)

Estimating the population mean

Intuitively, what is the best estimate of μ ?

- The **sample mean**: $\hat{\mu} = \overline{X}$
- A good estimator should converge towards the true population.
- We can evaluate how good our estimator is using simulations ("God's view").



(**sampling distribution**)

On the left, we have generated 10,000 random samples of 17 planarians each. We have fixed the true parameters: $\mu = 9$ and $\sigma = 3$. This is equivalent to 10,000 replications of the same study with the same sample size, on the exact same population.

Even though some samples' means are really far from the population mean, as a whole samples converge toward the true value, indicating that this is a good estimator.

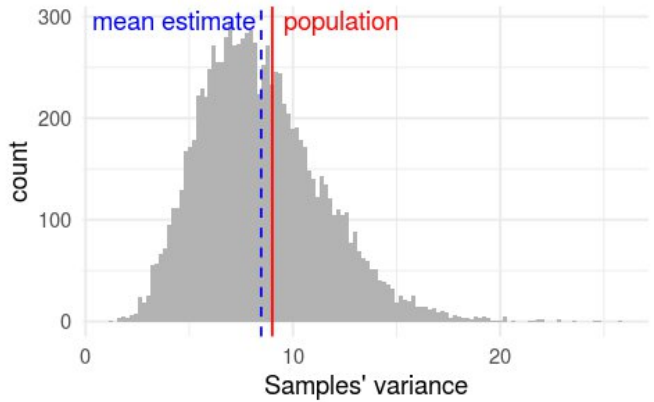
Estimating the population variance

What about the population variance σ^2

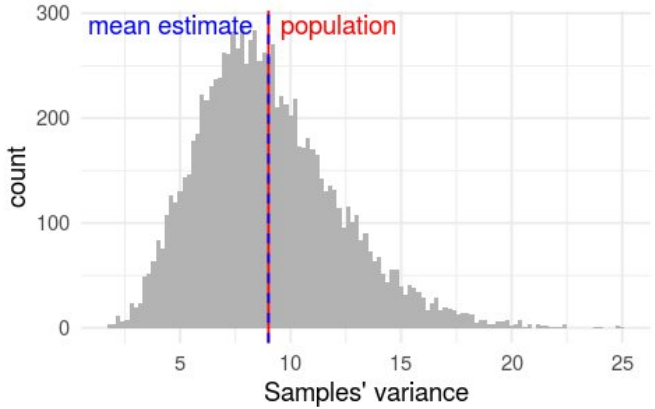
- "Intuitive" estimator: the sample variance s_n^2

  => not too bad but systematically underestimates the true value!

- **Unbiased** estimator (Bessel correction): $\frac{n}{n-1}s_n^2$



"Intuitive" estimator of variance

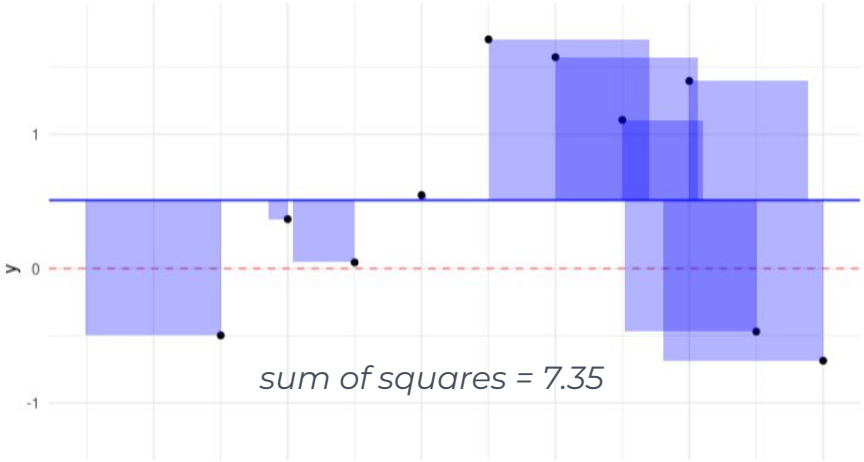


Unbiased estimator of variance

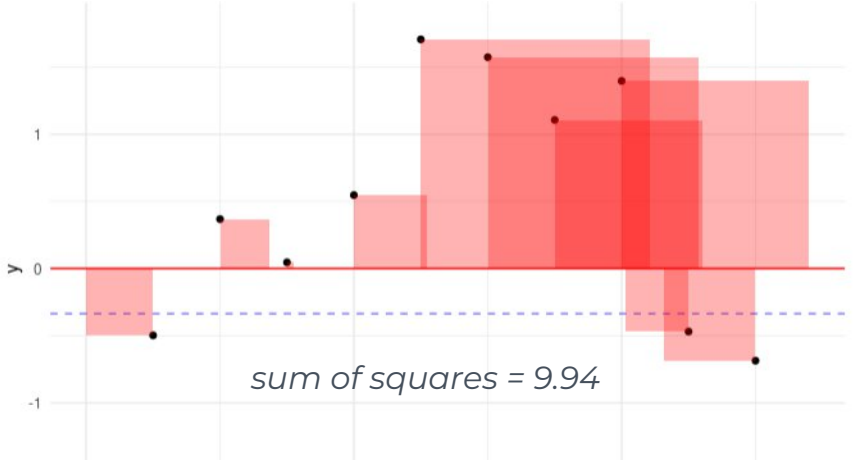
Estimating the population variance

Why is the sample variance a biased (under)estimate of the true population variance?

*Sample variance, based on the **sample** mean*



*Sample variance, based on the **true** mean*



Using the sample mean to estimate variance tends to "pull" the calculated deviations closer to zero than they would be if we were able to use the true population mean. That's because the sample mean, being calculated from the sample points, is always closer to them than the population mean..

Maximum likelihood estimation (MLE)

Mathematically, where do these estimators come from?

Conceptually, statistical inference consists in finding the **most plausible/likely** values for the population parameters, given the observed data. Mathematically, this means finding the population parameters values θ that **maximize** the following probability:

$$p(\theta|x_1,x_2,...,x_n) \quad \Leftarrow \textbf{bayesian statistics}$$

This is difficult. Instead, we can go for a simpler alternative: finding the values of population parameters for which the observed data are most likely, i.e. maximizing:

$$p(x_1,x_2,...,x_n|\theta) \quad \Leftarrow \textbf{frequentist statistics}$$

i.e., **assuming that the n observations are independent**, to maximize the **likelihood function**:

$$\mathscr{L}(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

where f is the (hypothesized) probability density function of the population.

For an application on parameter estimation of normal distributions, see on Wikipedia: tinyurl.com/y2jwpxx2

Maximum likelihood estimation (MLE)

MLE is the dominant estimation method in frequentist statistics. Why?

- **flexibility:** can be applied to a wide range of statistical models and distributions
- **computational tractability:** *relatively* easy/fast to calculate (using numerical optimization algorithms)

Limiting properties:

- **consistency:** it converges to the true population value
- **efficiency:** it produces estimates with the lowest possible variance

Bayesian statistics use entirely different estimation methods.
But there are mathematical equivalences:

MLE ⇔ maximum a posteriori (**MAP**) with a **uniform prior**

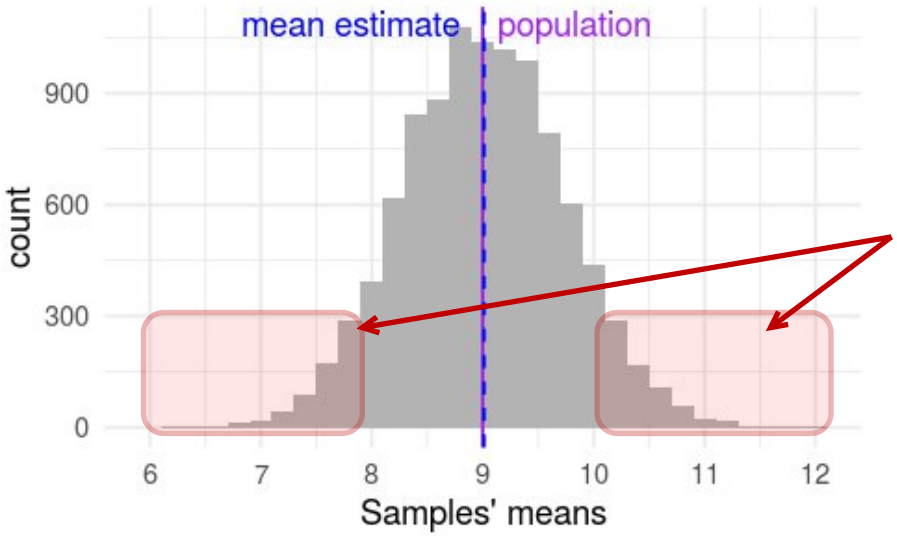
3.

Confidence intervals

Precision of population mean estimates

Our planarian sample is small (17 specimens only).

How much can we trust its population estimates?



(sampling distribution)

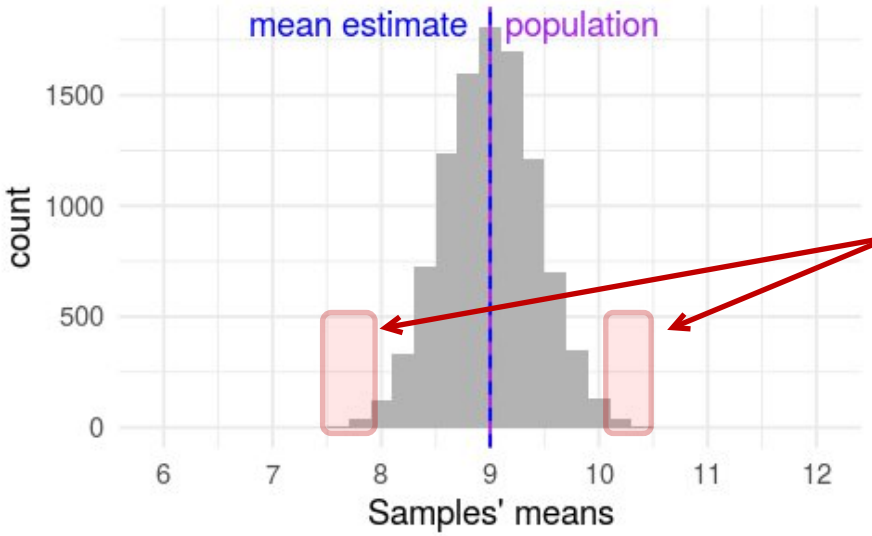
Many samples are way off! (~13%)

And your specific sample could be one of them… You can't know!

Precision of population mean estimates

A larger sample would be more precise!

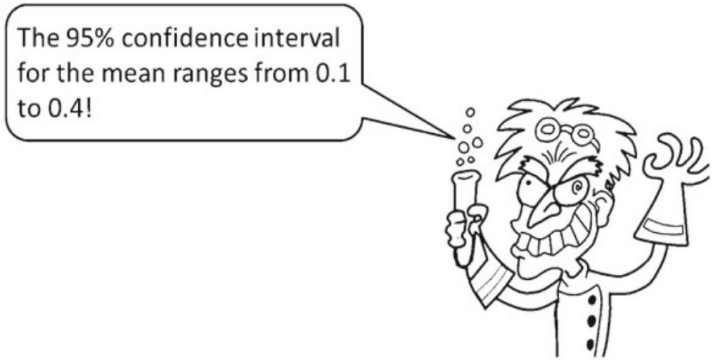
For example, with n = 50:



Fewer extreme samples (~0.5%)

Test

Professor Bumbledorf conducts an experiment, analyzes the data, and reports:

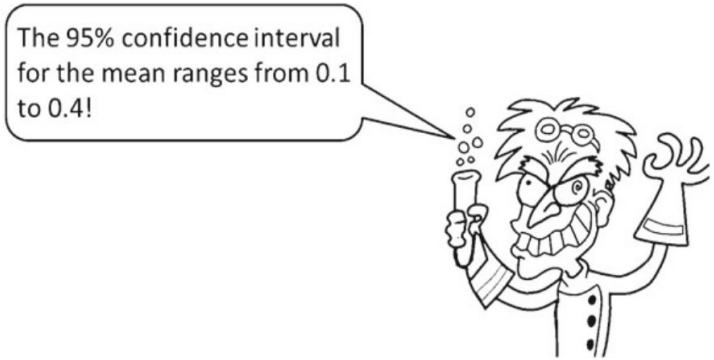


1. The probability that the true mean is greater than 0 is at least 95%.
2. The probability that the true mean equals 0 is smaller than 5%.
3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect.
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.

[ANSWERS IN THE NEXT SLIDE]

Test

Professor Bumbledorf conducts an experiment, analyzes the data, and reports:



A CI does **not** make a probability statement about population parameters. The confidence level (e.g., 95%) refers to the ***long-term success*** *rate* of the method, that is, how often this type of interval will capture the parameter of interest.

1. The probability that the true mean is greater than 0 is at least 95%.
2. The probability that the true mean equals 0 is smaller than 5%.
3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect.
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.

If we were to repeat the experiment over and over, then, in the long run, 95% of the CIs we would obtain would contain the true mean.

Everybody get confidence intervals wrong

Hoekstra, Morey, Rouder & Wagenmakers (2014). *Robust misinterpretation of confidence intervals*. Psychonomic Bulletin & Review (doi:10.3758/s13423-013-0572-3)

Study. 476 students and 120 researchers (including PhD students) were surveyed about their knowledge of confidence intervals using the test in the previous slide.

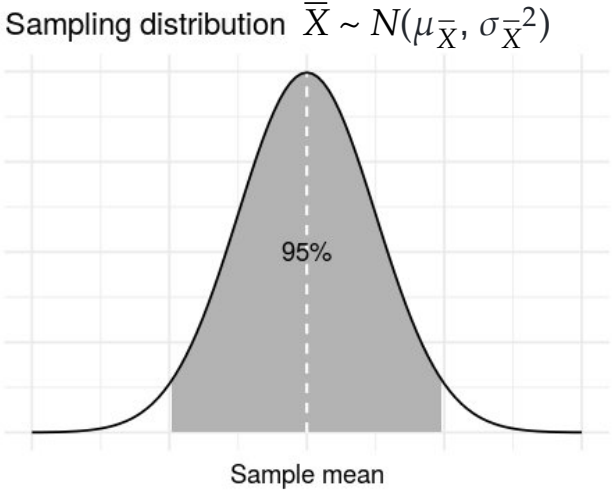
Conclusion from the authors. "*The APA Manual strongly encourages the use of CIs. Consequently, one might expect that most researchers in psychology would be well-informed about the interpretation of this* ***rather basic inferential outcome****. Our data, however, suggest that the opposite is true:* ***Both researchers and students in psychology have no reliable knowledge about the correct interpretation of CIs. Surprisingly, researchers' self-reported experience in statistics did not predict the[ir accuracy]. Even worse, researchers scored about as well as first-year students without any training in statistics****."*

Two views of probability

	Frequentist	Bayesian
Definition of probability	long-run frequency of events	degree of belief / certainty
View on model parameters	fixed (but unknown) constants	random variables
Probability apply to	data only	data & parameters

CIs & sampling distribution

Let's calculate the parameters of the sampling distribution.



Mean $\mu_{\overline{X}}$

We know that $\overline{X}$ is a good estimator of μ, so: $\mu_{\overline{X}} = \mu$

Mathematical derivation:

$$\mu_{\overline{X}} = E\left[\overline{X}\right] = E\left[\frac{\sum X_i}{n}\right] = \frac{1}{n}\sum E[X] = \frac{1}{n} n\mu = \mu$$

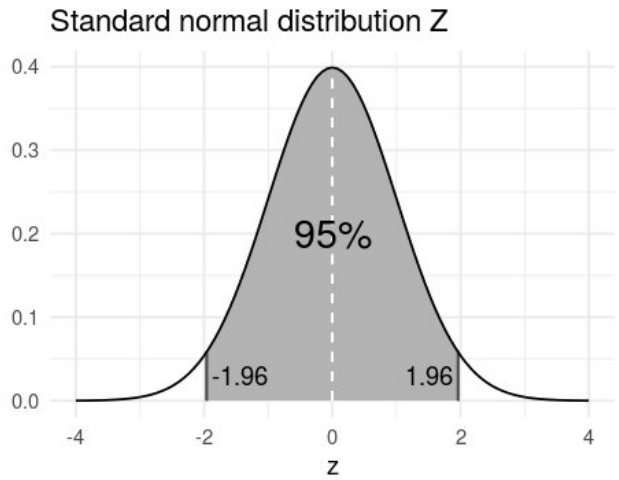
Variance $\sigma_{\overline{X}}^2$

Mathematical derivation:

$$\sigma_{\overline{X}}^2 = Var\left(\overline{X}\right) = Var\left(\frac{\sum X_i}{n}\right) = \ldots = \frac{\sigma^2}{n}$$

Therefore: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. This is consistent with our previous simulations showing that the distribution of sample means narrows down when we increase the sample size n.

CIs & sampling distribution



If σ **is known**, we can **standardize*** $\overline{X}$:

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) = Z$$

The middle 95% of the Z distribution lies between -1.96 and 1.96. Therefore, this will be true for the samples that fall in the middle 95% of the sampling distribution:

$$\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+1.96\frac{\sigma}{\sqrt{n}}$$

** i.e. subtract the mean and divide by the standard deviation*

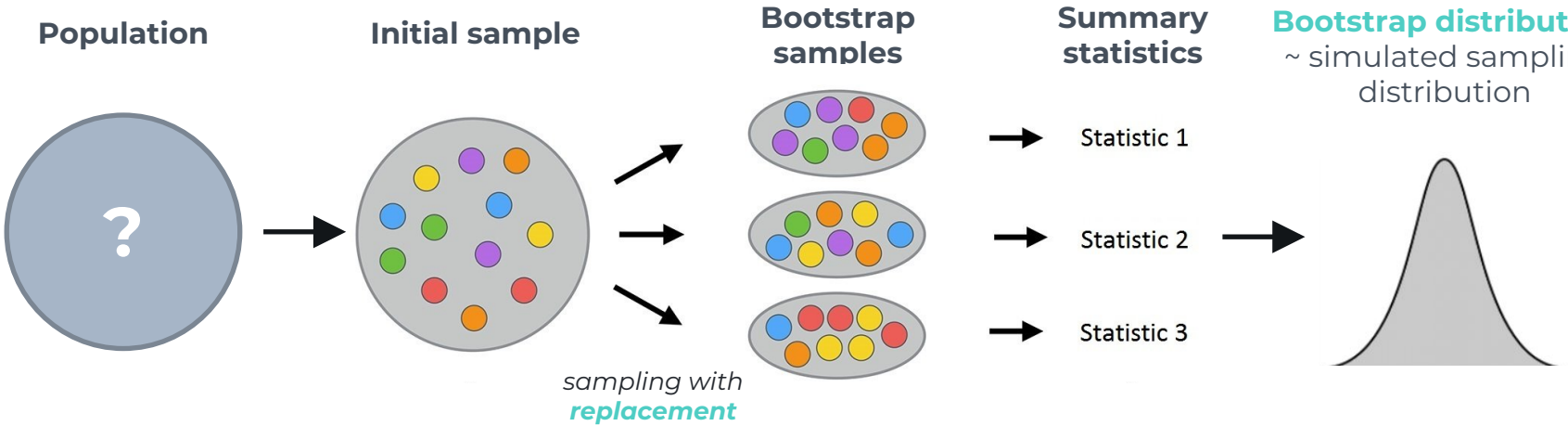
formula of the confidence interval!

⇒ confidence intervals are constructed using properties of the sampling distribution

⇒ confidence intervals should be interpreted in the context of repeated sampling, not in terms of the single sample at hand

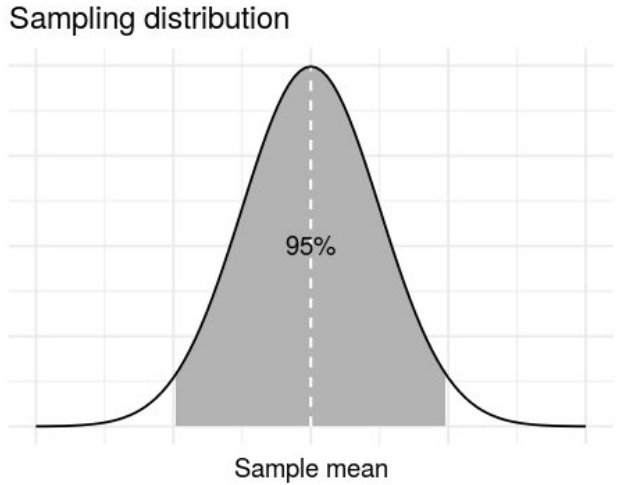
CIs & sampling distribution

Bootstrapping estimation



The **middle interval** that contains X% of the bootstrap distribution is a valid X% CI.

Confidence intervals
Interpretation



Remember how we constructed the confidence interval: for samples that are in the middle 95% of the sampling distribution, we have:

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

So the **correct interpretation** is: 95% of *95% confidence intervals* of all possible samples (that have been or will be or could be collected), contain the true population mean.

You can't know whether your particular sample belongs to this middle 95%. And once the data is collected, the confidence interval of a sample either *does* or *does not* contain the true population mean. It is meaningless to say it has 95% probability of containing the true population mean!

see visualization at rpsychologist.com/d3/ci/

A quick test

Right or wrong?

Context. The planarians have been cut, and fully regenerated. The new lengths have been measured, and the 95% CI is [7.2, 11.4] mm.

1.

If we cut them again and let them regenerate, the average new final length has a 95% probability to fall between 7.2 and 11.4

2.

Planarians have been bred, and samples of 17 specimens have been sent to many labs across the world for regeneration experiments. All scientific collaborators have agreed to publish results in the form of 95% CIs. Therefore, we can expect 95% of the future reported CIs to:

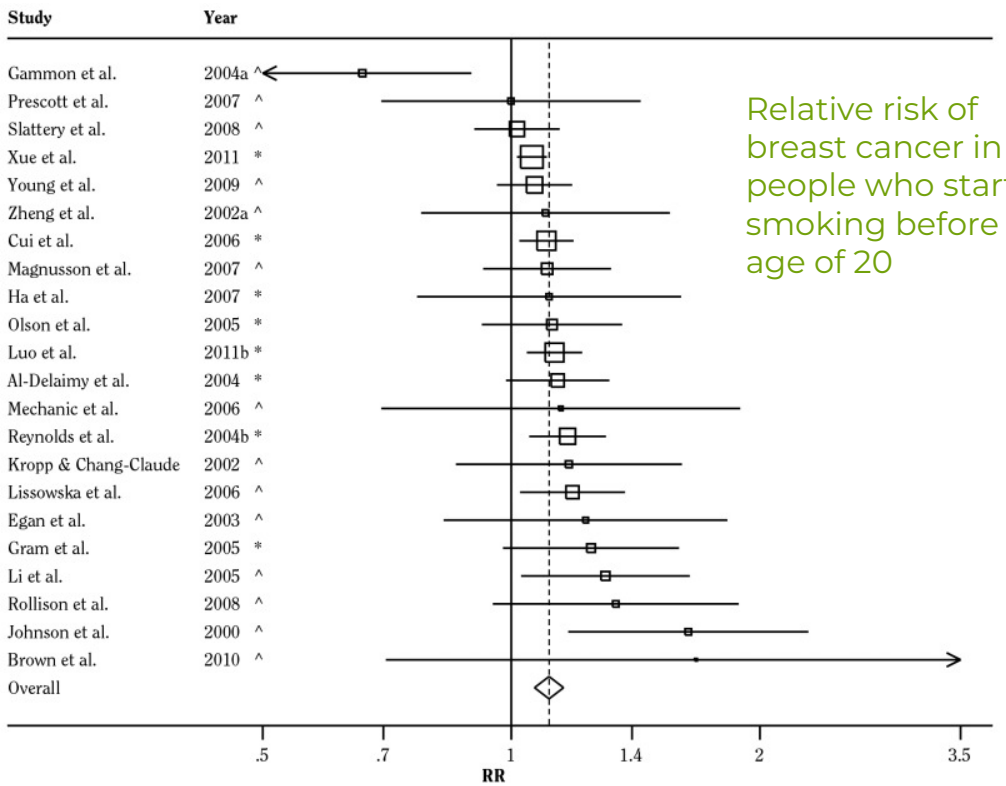
(a) overlap with the original [7.2, 11.4] CI 

(b) contain the true population value 

Why CIs are useful nonetheless

If all intuitive interpretations of CIs are wrong, **what's the point** of CIs?

1) Multiple CIs can be **combined** to produce a more **precise estimate**



Relative risk of breast cancer in people who started smoking before the age of 20

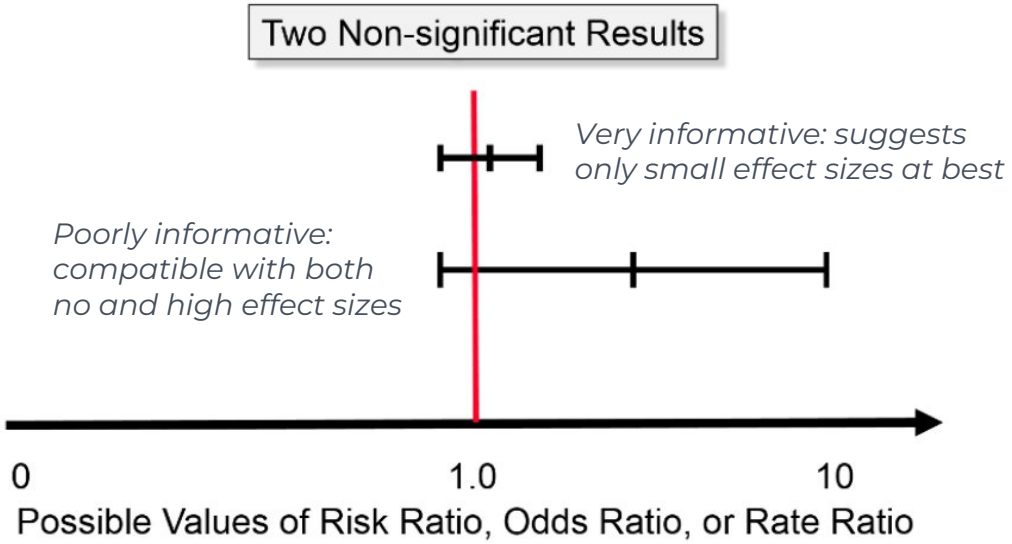
The Health Consequences of Smoking—50 Years of Progress (2014)

Why CIs are useful nonetheless

If all intuitive interpretations of CIs are wrong, **what's the point** of CIs?

1) Multiple CIs can be **combined** to produce a more **precise estimate**

2) The width of the CI tells us **how informative is the data**



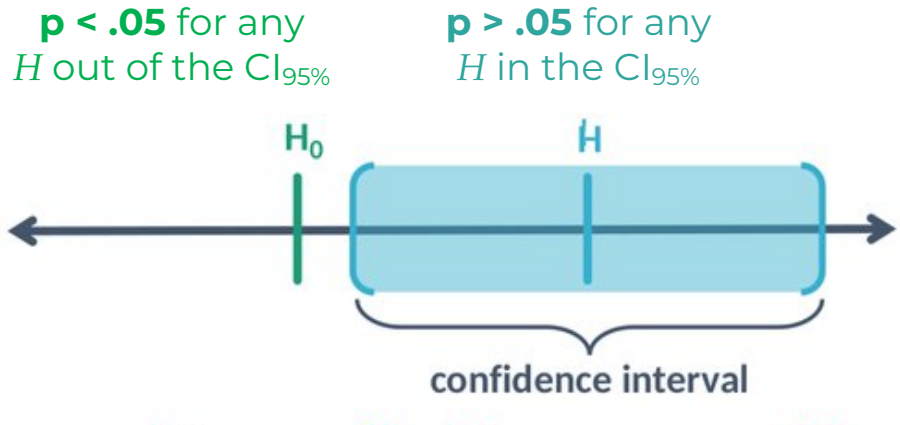
Why CIs are useful nonetheless

If all intuitive interpretations of CIs are wrong, **what's the point** of CIs?

1) Multiple CIs can be **combined** to produce a more **precise estimate**

2) The width of the CI tells us **how informative is the data**

3) CIs and *p* values are related, such that the CI gives a range of values **compatible** with the data (in the sense of significant testing)



Frequentist vs. Bayesian CIs

The correct interpretation of confidence intervals is **highly un-intuitive**.

It requires us to conceive of our sample as only one realization of an experiment that could be repeated a large number of times. The interpretation of confidence intervals is based on the **frequentist** view of probability as the ***long-run expected frequency of occurrence***.

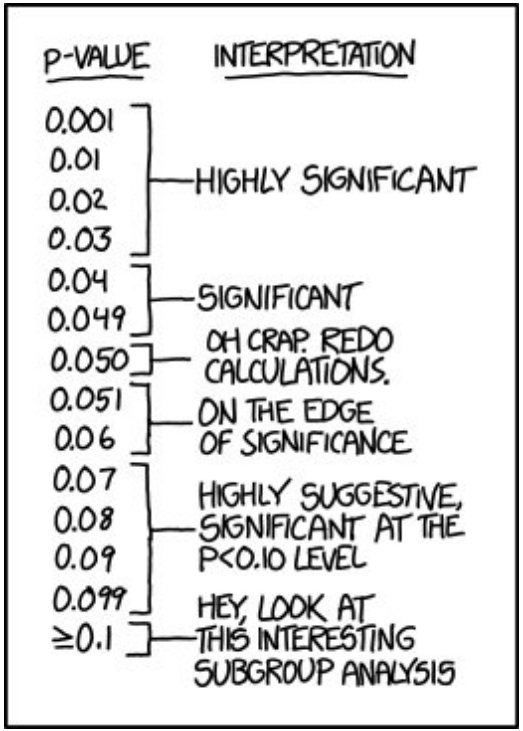
Typically, we misinterpret the confidence interval in a way that matches what we would like to get: an interval for which we are 95% confident (or there is 95% chance) that it contains the true population parameter.

Such intervals exist, they are called **credible intervals**. They rest on the **Bayesian** view of probability as level of certainty, or degree of plausibility or belief. However, credible intervals are dependent on **prior information** to be specified by the human analyst.

The word *confidence* is a terrible choice for confidence intervals. Some authors have proposed to replace it by **uncertainty intervals** or **compatibility intervals** (Gelman & Greenland 2019)

4.

Null hypothesis significance testing (NHST)



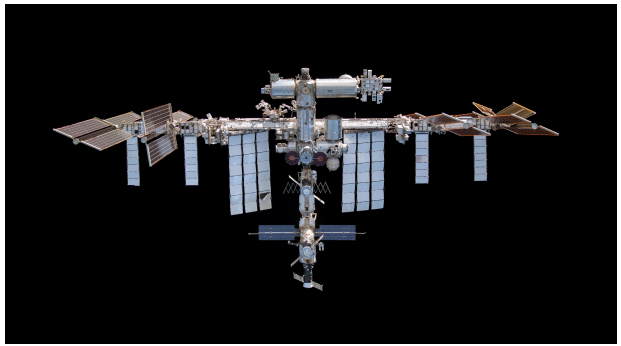
“If all else fails, use "significant at a p>0.05 level" and hope no one notices."

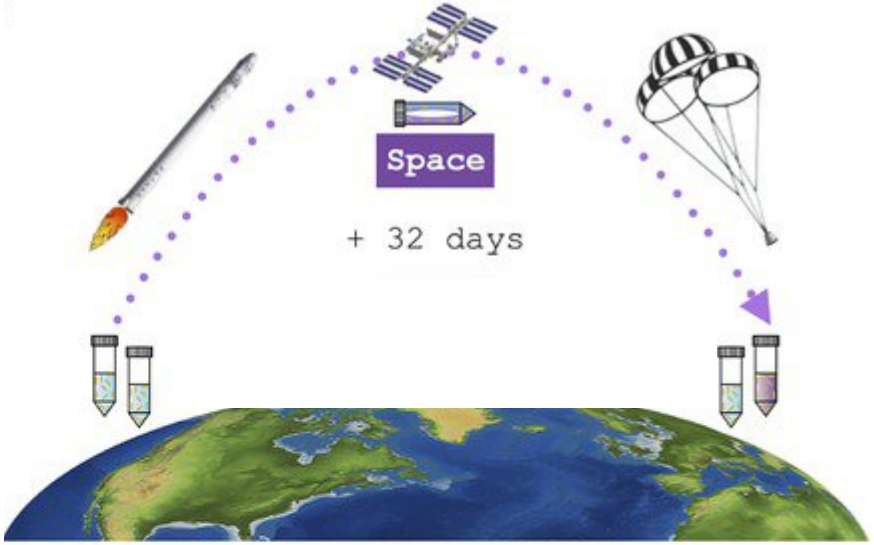
xkcd #1478

The modern NHST procedure

1. Collect data.
2. Specify a statistical model (i.e. a distribution for the population) with parameters θ.
3. Specify the null hypothesis H_0 over one of the parameters.
4. Find a way to summarize the data in one value, the **test statistic**, such that its sampling distribution is defined and calculable under the null hypothesis. This is a job for statisticians (phew!)
5. Calculate $p\left(sample\ statistic \middle| H_0\right)$, *aka* the ***p* value**: the probability of observing the data assuming the null hypothesis

Example scenario





Morokuma et al. 2017. *Planarian regeneration in space: Persistent anatomical, behavioral, and bacteriological changes induced by space travel.* Regeneration (doi:10.1002/reg2.79)

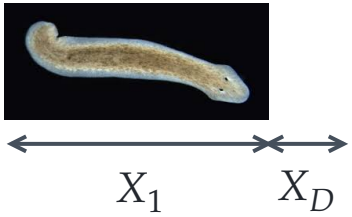
Example: mean of the normal distribution

Scenario A (the paired two-sample case)

1

data collection

Prior to being sent on the ISS, the 17 planarians have been measured. Biologists wonder whether they have grown or shrinked during their space journey.



2

statistical model

Trick: we work with $X_D = X_2 - X_1$, the **paired** difference between the two sets of data (the first and second length measurements), i.e. the *amount of growth*.

We assume $X_D \sim \mathcal{N}(\mu_D, \sigma_D{}^2)$

Example: mean of the normal distribution

3

null hypothesis

The null hypothesis is that the worms have neither grown nor shrank on average:

$$H_0: \mu_D = 0$$

4

test statistic with calculable sampling distribution

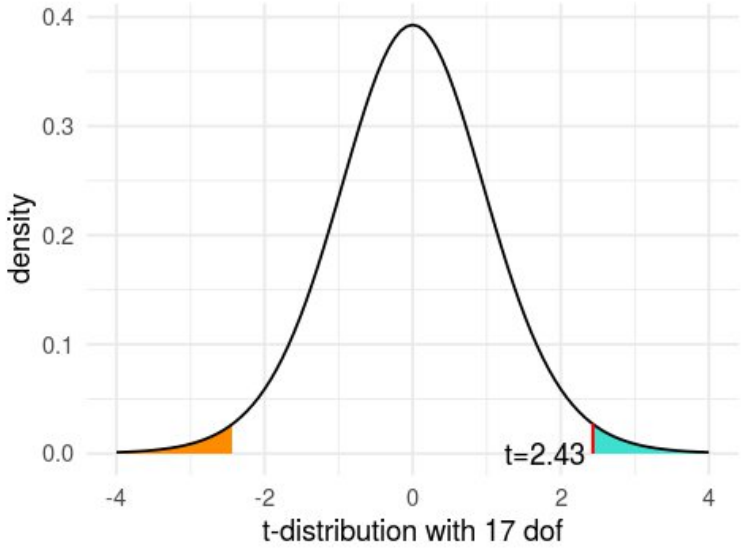
When we constructed confidence intervals, we have already shown that:

$$\frac{\overline{X_D} - \mu_D}{s/\sqrt{n}} \sim t_{n-1}$$

Therefore, under H_0, the quantity $\frac{\overline{X_D}}{s/\sqrt{n}}$ is a test statistic whose sampling distribution is known: it is the t-distribution with $n-1$ *degrees of freedom*.

Example: mean of the normal distribution

5



$$p\left(abs\left(t\right) \geq 2.43 \,\middle|\, \mu_D = 0\right) = 0.027$$

Significance testing

It appears that the probability of observing data as extreme as ours under the null hypothesis is quite small (2.7%). So small that we might be tempted to rule out the null hypothesis as a plausible source of data.

6

decision

One further (but optional) step is to use a **significance level α** (chosen before data collection!) as a criteria to decide whether or not we have sufficient evidence against the null hypothesis:

$p < \alpha$: we reject H_0

$p \geq \alpha$: we do not reject H_0

A glaring gap

What we intuitively desire:

most plausible values of θ given the data y

$a(\theta|y)$

⇒ offered by **Bayesian statistics**, using the Bayes law:

$$p(\theta|y) = \frac{p(y|\theta).p(\theta)}{p(y)}$$

A frequentist alternative:

$$p(Y \geq y|H_{range})$$

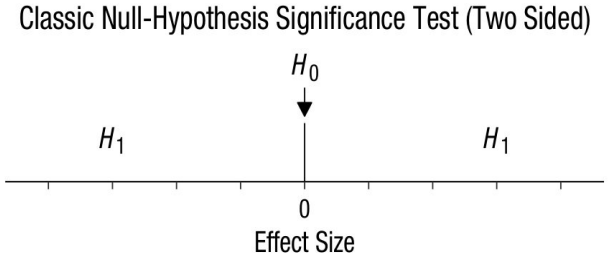
- ✓ Can test a hypothesis of practical significance
- ✗ Based on data we have not actually observed

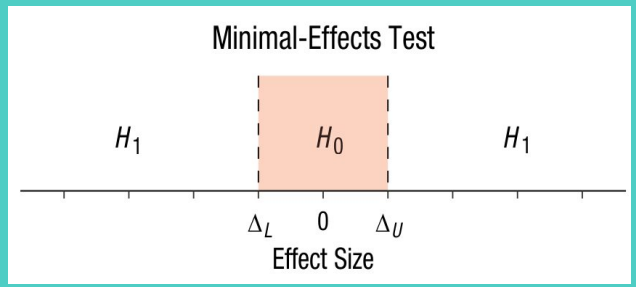
What frequentist NHST gives us:

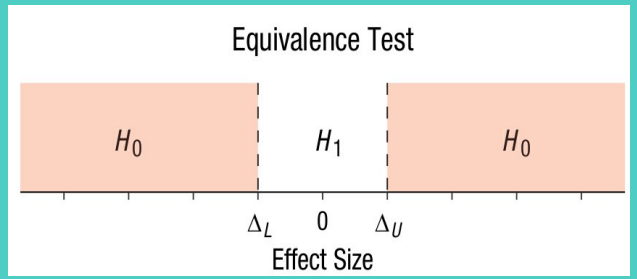
$$p(Y \geq y|H_0{:}\theta = \theta_0)$$

- ✗ Tests a hypothesis that we are not interested in
- ✗ Based on data we have not actually observed

Equivalence tests *aka* TOST







see Lakens, Scheel & Isager 2018 (doi.org/10.1177/2515245918770963)

Misinterpretations of p-values

Cite 6 common misinterpretations of *p*-values

Misinterpretations of p-values

1. The *p*-value is the probability of the null hypothesis.

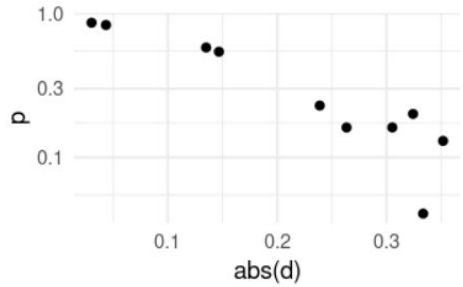
2. A *p*-value below the significance level α is evidence for the experimental hypothesis (e.g., that there is a difference between two populations' means).

3. A very high *p*-value is a strong evidence that there is no reliable effect.

4. The probability of the experimental hypothesis can be deduced from the *p*-value.

5. Smaller p-values indicate larger effect sizes

6. If the *p*-value is very small, most replications will be statistically significant too.

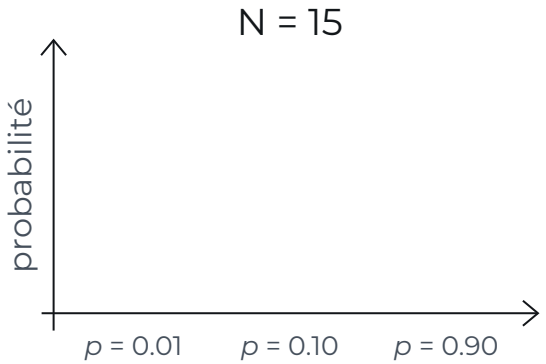


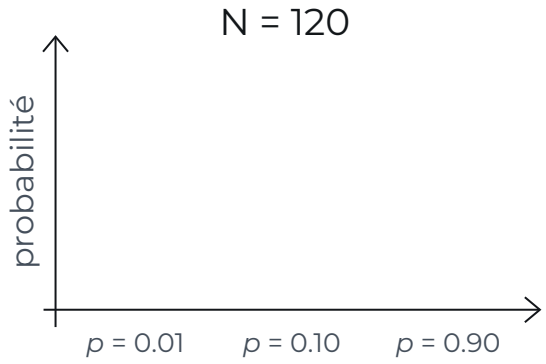
The p-value is not a perfect predictor of effect size

See also this excellent post from Daniel Lakens, discussing several misinterpretations and debunking them with simulations and graphs: *daniellakens.blogspot.com/2017/12/understanding-common-misconceptions.html*

Let’s play !

Supposons que l’on réalise un test sur un jeu de données, pour lequel on sait que l’hypothèse nulle est vraie. Selon vous, quelles sont les probabilités relatives d’obtenir une p-value de 0.01 ? 0.10 ? 0.90 ? Comment ces probabilités relatives changeraient-elles en augmentant la taille d’échantillon, par exemple de 15 à 120 ?



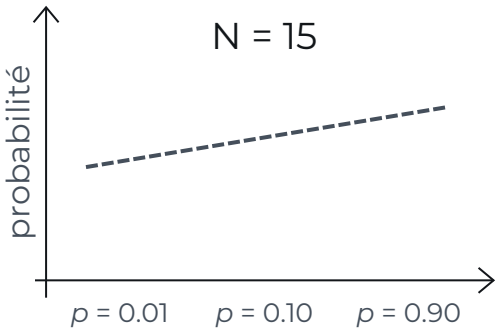


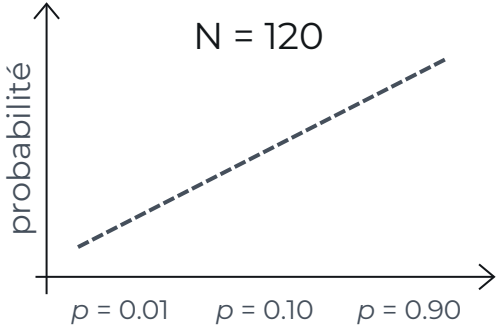
Let’s play !

The intuitive expectation:

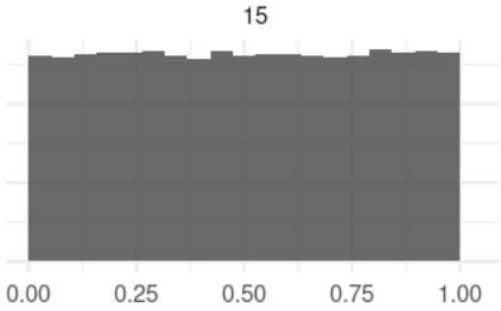
— Large p-values are more likely than small ones under H0

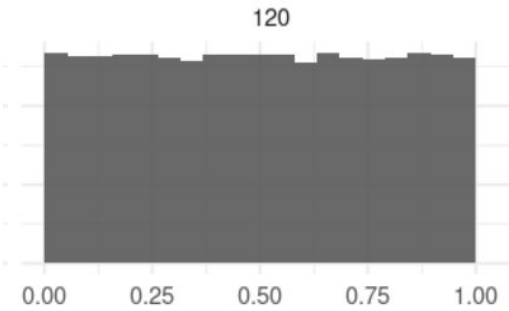
— This pattern is exacerbated with increasing sample size





Actual results from simulations:

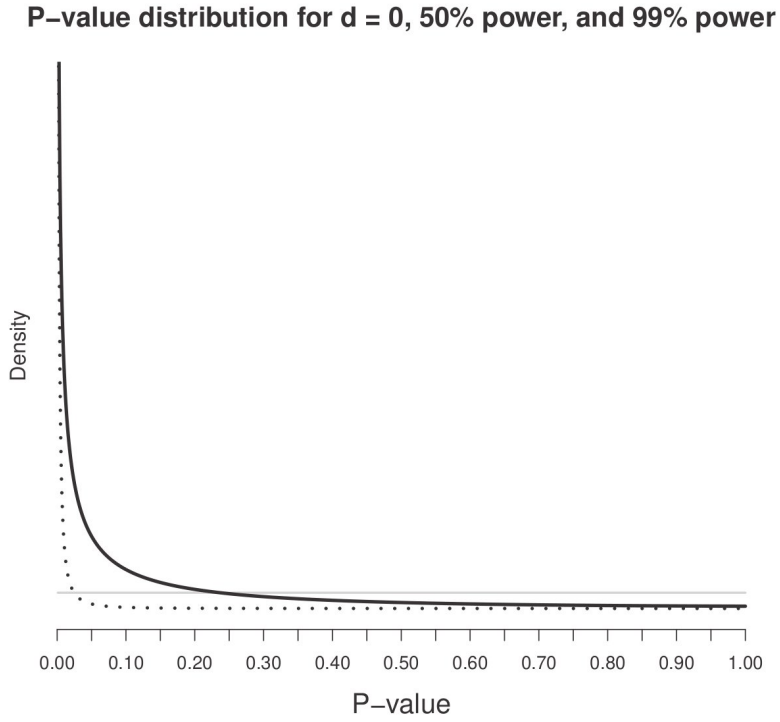




⇒ the distribution of *p*-values is always uniform when the null is true.

Let’s play !

Let’s play !



The distribution of p-values gets sharper when statistical power increases.

Let’s play !

Une étude teste un effet en lequel vous avez de bonnes raison théoriques (pensez-vous) de ne pas croire. La méthodologie de l’étude est irréprochable.

L’étude a utilisé un échantillon de 21 participants, une taille un peu plus faible que la moyenne du domaine.

L’étude a utilisé un échantillon de 210 participants, une taille beaucoup plus importante que la moyenne du domaine.

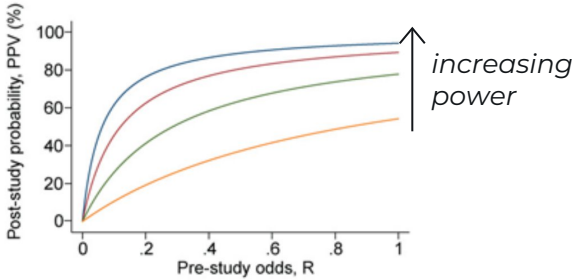
Pour le test statistique central de l’étude, les auteurs rapportent $p = 0.005$.

Comment réagissez-vous à ce résultat ?

A: je n’y crois toujours pas, ils ont simplement eu de la chance !

B: Ok, je dois admettre que j’ai peut-être tort.

C: Je ne sais plus quoi penser...



Limited interpretation of the *p*-value

- **Sensitivity to sample size:** even trivial effects can produce significant *p*-values, given sufficiently large sample sizes
- **Sensitivity to statistical power:** the evidential value of a single *p*-value depends on the statistical power of the test (Lakens 2022)
- **Sensitivity to base rate:** the evidential value of a single *p*-value depends on the frequency of true hypotheses among all hypotheses tested in the scientific field (compare confirmatory RCTs with exploratory high-throughput genomics) [Ioannidis 2005]
- **Arbitrary cutoff:** The difference between $p = .049$ and $p = .051$ is not significant! (Gelman & Stern 2006, doi:10.1198/000313006X152649)
  ⇒ risk of dichotomous thinking and overlooking meaningful results just above the threshold.

A cascade of consequences

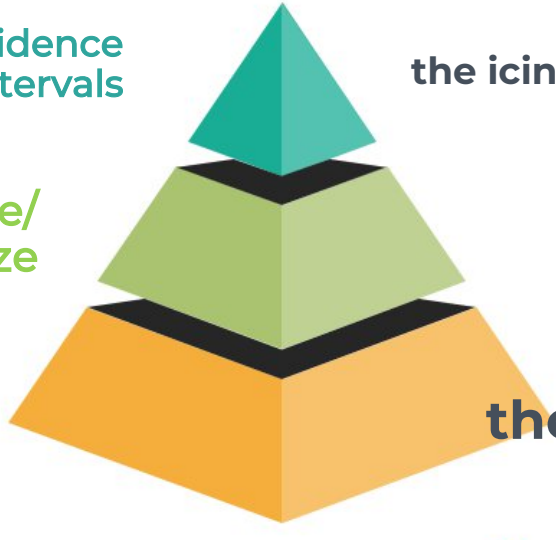
- **publication bias**
  ⇒ when compounded with low statistical power, leads to **overestimation of effect sizes**

- ***p*-hacking:** deliberate or involuntary
  ⇒ increases false positives, i.e. decreases replicability

- **obsessive overreliance on *p*-values**
  ⇒ occludes other sources of scientific evidence (prior information, plausible mechanism, consistency, etc.)
  ⇒ detrimental to **theorization efforts** (Meehl 1978, doi:10.1016/j.appsy.2004.02.001)
  ⇒ hinders the improvement of **statistical literacy**

The pyramid of importance in statistics

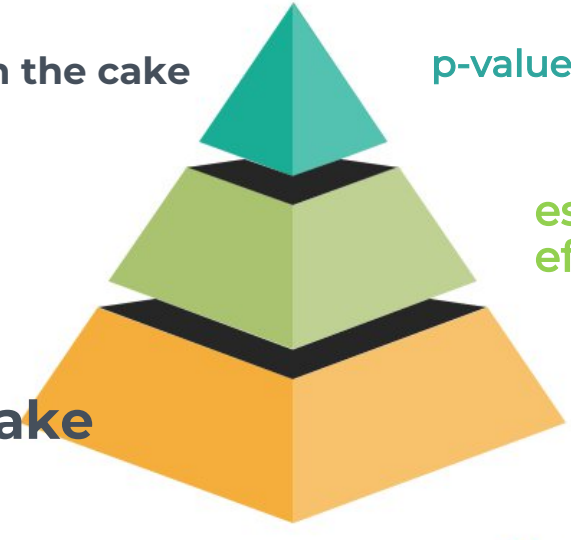
How we have learned...



the icing on the cake

the cake

...How we should do it



How did we get there?

1925+ : Ronald Fischer develops “significance testing” to provide a mathematically rigorous treatment of randomness in data, so as to extract useful information from **small samples**. He restricted his procedure to **exploratory research** (when we have no theory and no hypothesis), insisted that the results made sense only for **well controlled designs**, and viewed individual *p* values as mere data for **meta-analysis**.

1928+ : Neyman and Pearson set up to improve on Fisher’s approach and ended up inventing an entirely new procedure to be able to **decide between two competing hypotheses**, based on **effect size** and **type I and II error rates**. It was developed for quality control of standardized manufactured objects.

1940 – 1942 : Statistical textbooks for psychologists blend Fisher and Neyman-Pearson procedures, creating inconsistent chimeras that have been passed from generation to generation, until us.

Gigerenzer (2004). *Mindless statistics*. The Journal of Socio-Economics (doi:10.1016/j.socec.2004.09.033)
Nuzzo (2014). *Scientific method: Statistical errors.* Nature (doi:10.1038/506150a)
Perezgonzalez (2015). *Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing* (doi.org/10.3389/fpsyg.2015.00223)