

So, you can do bootstrapping?

And you think you don't need to bother collect more
individuals?

Background

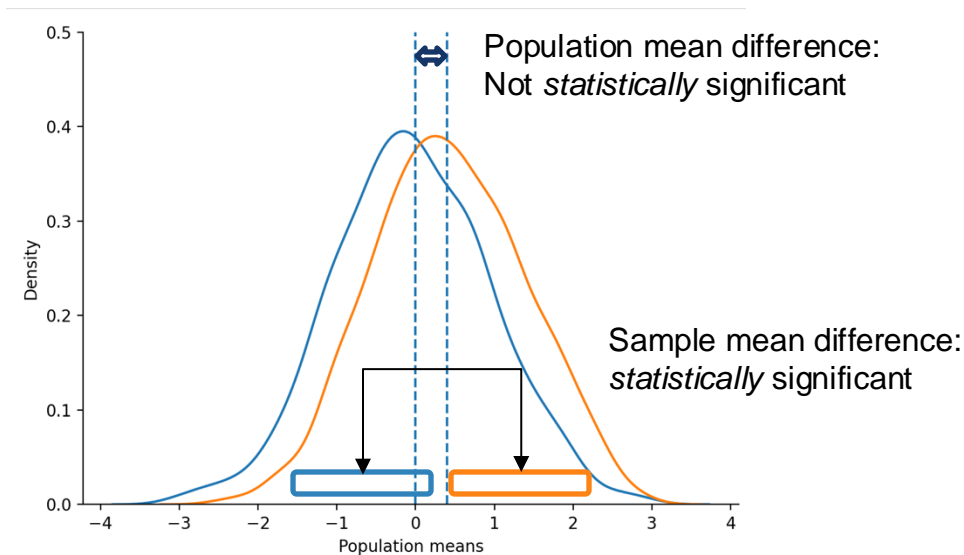
When you conduct a statistical test (ex. T-test, ANOVA), you calculate a statistic (ex. t for t-test, F for ANOVA). What exactly are these tests about?

Let's focus on **t-test** (two independent samples).

We test if two **population means** are significantly different. We cannot calculate population means, so we take **samples** of the two populations. Typically, you have access to YOUR SAMPLES - the samples you managed to recruit.

Background

How well do your samples represent corresponding populations? If your samples are not good representations, a *big* difference between YOUR sample means may be just a chance thing, not a true representation of the difference between population means.



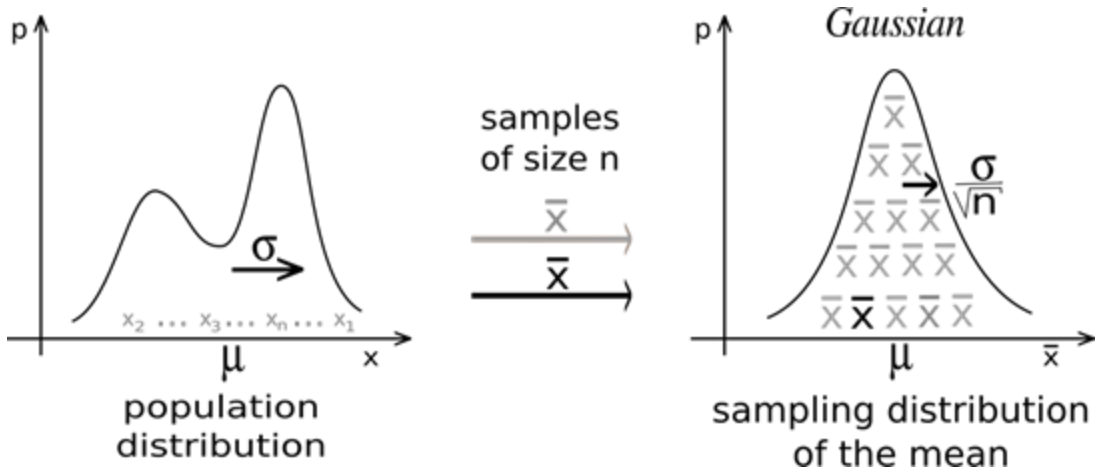
Background

So using your samples, you make an inference (hence the name: inferential statistics), wishing that your samples well represents their corresponding populations. Thankfully, those who devised hypothesis testing techniques (ex. R.A.Fisher, William Gosset, etc.) prepared buttress to your inference.

Background

There's a theorem named *Central Limit Theorem* (CLT). Simply put, it argues that if your sample size is large enough (typically $n = 30$), the sampling distribution of the sample mean made in a hypothetical scenario (where you can recruit N distinct samples whose sizes are all $n=30$; N is a LARGE number) will be **close to a normal distribution**.

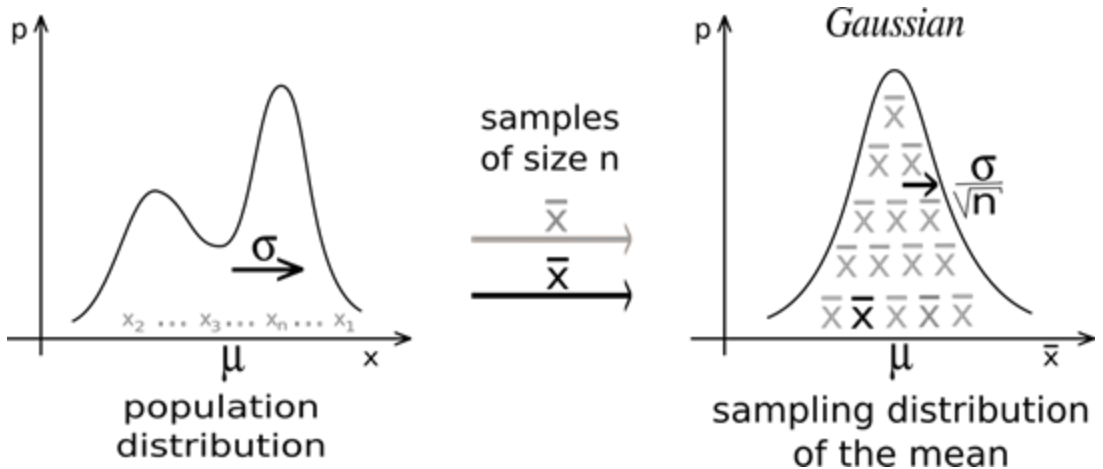
Image from wikipedia

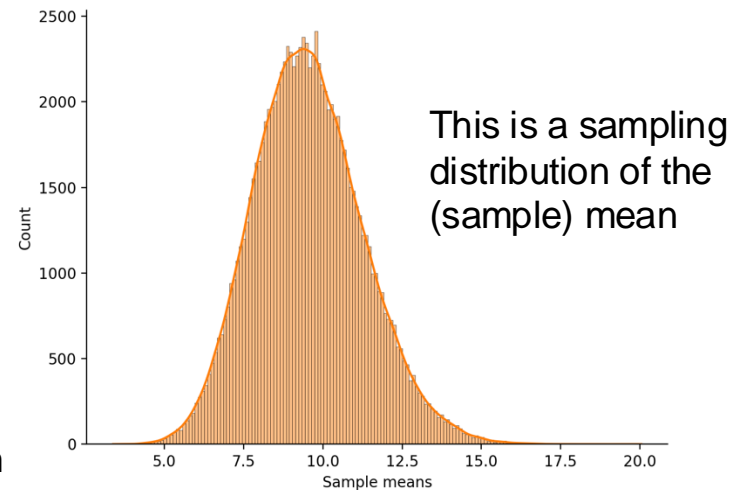
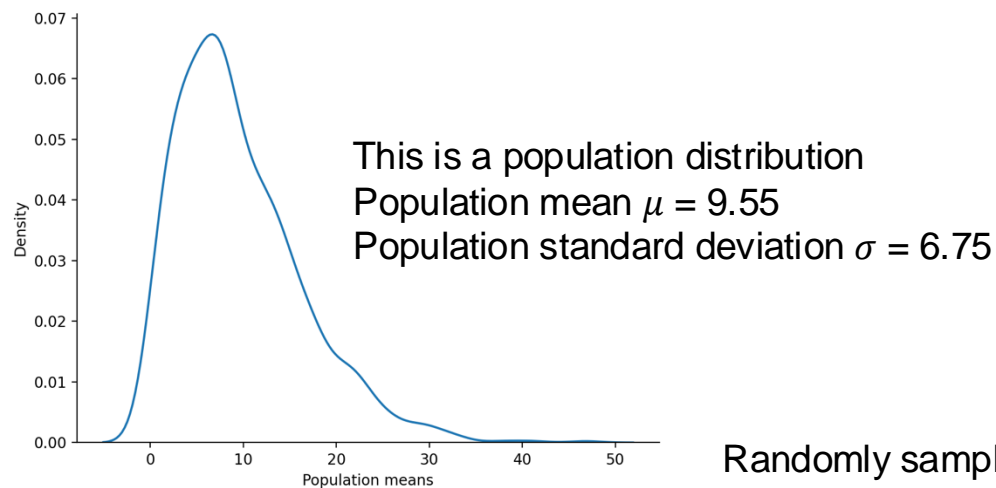


Background

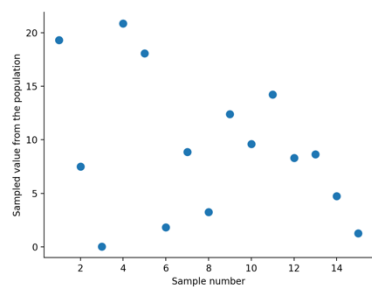
The mean of the sampling distribution is equal to the population mean (μ). The standard deviation of the sampling distribution is the population standard deviation, divided by the square root of n , the sample size. There is a mathematical proof to this, so doubt no more, please.

Image from wikipedia

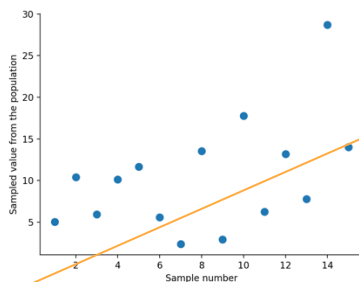




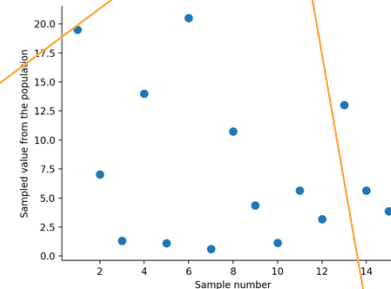
Randomly sample with
 $n = 15$, $N = 100,000$



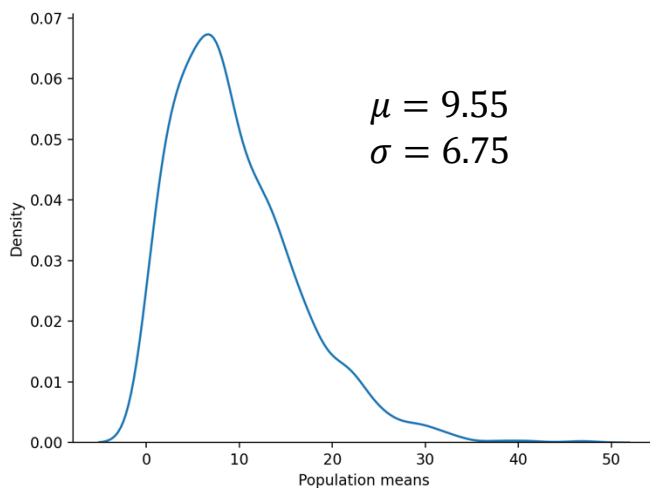
1st Sample
Sample mean: **9.26**



2nd Sample
Sample mean: **10.36**

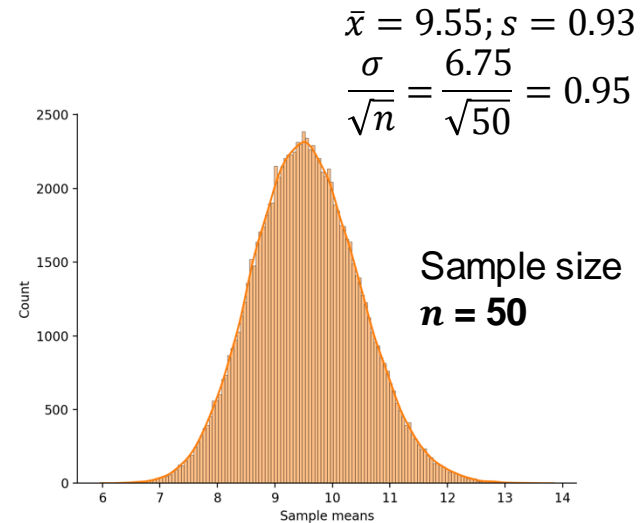
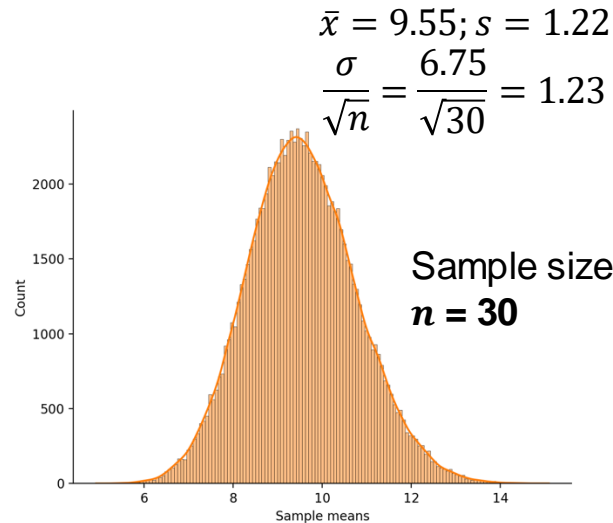
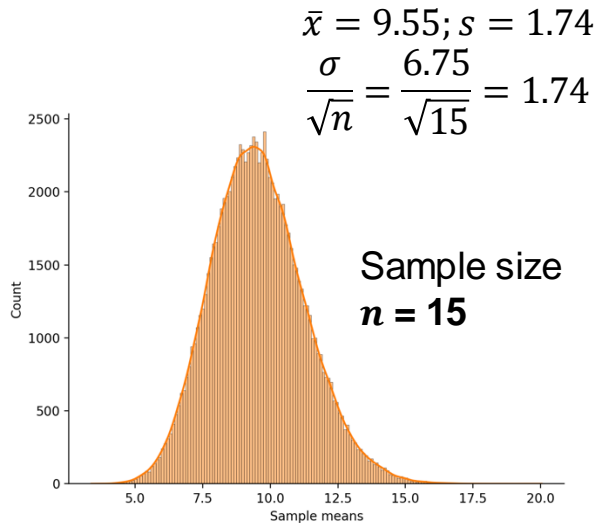


100,000th Sample
Sample mean: **7.44**



From the population distribution, three empirical sampling distributions with distinct sample size values are prepared. Check if the results match CLT. For each sample size:

1. Does a sampling distribution look *normal* to you?
2. Is the mean of a sampling distribution (\bar{x}) similar to the mean of the population distribution (μ)?
3. Is the standard deviation of a sampling distribution (s) similar to σ/\sqrt{n} ?



Background

Mathematically, it is:

More precisely, it states that as n gets larger, the distribution of the normalized mean $\sqrt{n}(X_n - \mu)$, i.e. the difference between the sample average \bar{X}_n and its limit μ , scaled by the factor \sqrt{n} , approaches the normal distribution with mean 0 and variance σ^2 . For large enough n , the distribution of \bar{X}_n gets arbitrarily close to the normal distribution with mean μ and variance σ^2/n .

Let $\{X_1, \dots, X_n\}$ be a sequence of i.i.d. random variables having a distribution with expected value given by μ and finite variance given by σ^2 .

Suppose we are interested in the sample average

$$\bar{X}_n \equiv \frac{X_1 + \dots + X_n}{n}.$$

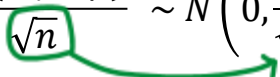
By the law of large numbers, the sample average converges almost surely (and therefore also converges in probability) to the expected value μ as $n \rightarrow \infty$.


This is why μ is called “limit”

Background

So in this case, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ will follow a standard normal distribution (mean 0, standard deviation = variance = 1).

$$\sqrt{n}(\bar{X}_n - \mu) \sim N(0, \sigma)$$

$$\rightarrow \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{n}} \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right)$$


$$\rightarrow \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N\left(0, \frac{\frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}}\right) = N(0, 1)$$


CLT, classical form // $N(0, \sigma)$ means a normal distribution with mean 0 and standard deviation σ

Divide the statistic by \sqrt{n} to make the form: $(\bar{X}_n - \mu) \sim N(0, \frac{\sigma}{\sqrt{n}})$

Divide the new statistic by σ/\sqrt{n} to check the approximate distribution of a variable: $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$

t-test

History [\[edit \]](#)

The term "t-statistic" is abbreviated from "hypothesis test statistic".^[1] In statistics, the *t*-distribution was first derived as a [posterior distribution](#) in 1876 by [Helmert](#)^{[2][3][4]} and [Lüroth](#).^{[5][6][7]} The *t*-distribution also appeared in a more general form as Pearson type [IV](#) distribution in [Karl Pearson's](#) 1895 paper.^[8] However, the *t*-distribution, also known as [Student's t-distribution](#), gets its name from [William Sealy Gosset](#), who first published it in English in 1908 in the scientific journal [Biometrika](#) using the pseudonym "Student"^{[9][10]} because his employer preferred staff to use [pen names](#) when publishing scientific papers.^[11] Gosset worked at the [Guinness Brewery](#) in [Dublin, Ireland](#), and was interested in the problems of small samples – for example, the chemical properties of barley with small sample sizes. Hence a second version of the etymology of the term Student is that Guinness did not want their competitors to know that they were using the *t*-test to determine the quality of raw material. Although it was William Gosset after whom the term "Student" is penned, it was actually through the work of [Ronald Fisher](#) that the distribution became well known as "Student's distribution"^[12] and "Student's *t*-test".

Gosset devised the *t*-test as an economical way to monitor the quality of [stout](#). The *t*-test work was submitted to and accepted in the journal [Biometrika](#) and published in 1908.^[9]

(One sample) t-test

T-statistic has the following formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$


“Standard error”


This formula is of one sample t-test. \bar{x} is the **sample mean**. μ_0 is the **population mean according to the null hypothesis** (0, usually). s is the **sample standard deviation**, and n is the **sample size**.

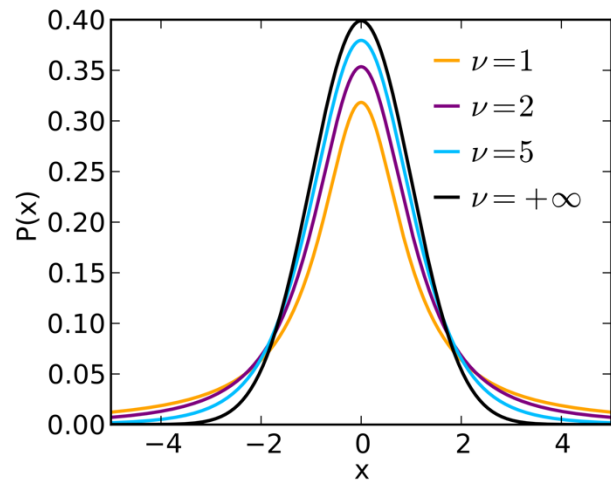
In slides 8&9, we saw that the Central Limit Theorem is about the variable $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$

Plot of the density function for several members of the Student t family (source: Wikipedia)

(One sample) t-test

See that t is very similar to Z, with σ replaced with s ?

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$




So t statistic is using an **estimate** of σ , s – your sample standard deviation.

t statistic follows a **t distribution** with one parameter: degrees of freedom (ν). This is a generalized form of standard normal distribution.

(Two samples) t-test

When there are two independent samples,

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Difference of the sample means

Difference of the sample means
according to the null hypothesis
(zero typically)

“pooled” standard error

1) Equal variance: $\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$

2) Unequal variance / Welch's method: $\sqrt{\left(\frac{s_1}{\sqrt{n_1}} \right)^2 + \left(\frac{s_2}{\sqrt{n_2}} \right)^2}$

t-test (One sample or two samples)

So either one sample or two independent samples t-test will calculate a t value. Then it will check the probability of observing that very t value or values greater than that in magnitude given a t distribution with a corresponding degrees of freedom.

t-test (One sample or two samples)

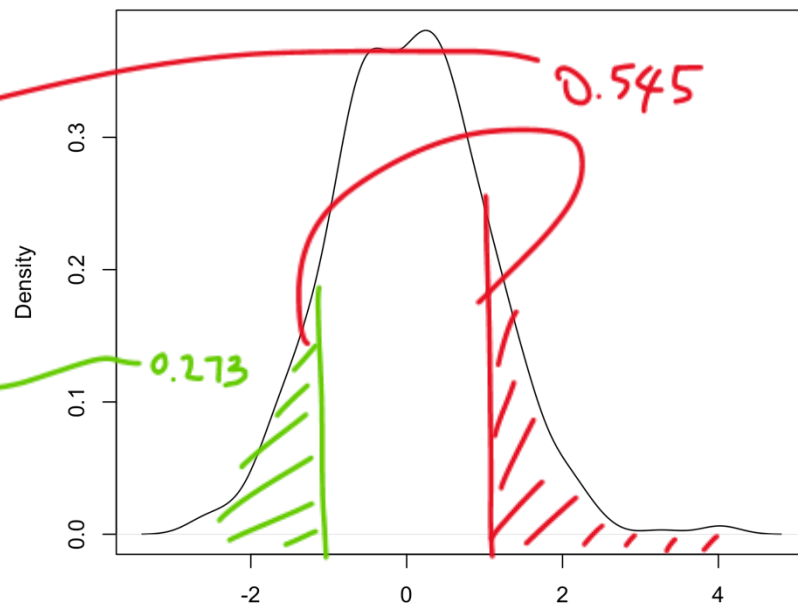
ex. T-test result introduced in hypothesis-testing_pt1:

T-Test

Group Statistics				
sex	N	Mean	Std. Deviation	Std. Error Mean
sbp 0	56	46.46	11.145	1.489
1	44	47.86	11.806	1.780

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference
sbp	Equal variances assumed	.079	.779	-.607	98	.273	.545	-1.399	2.305
	Equal variances not assumed			-.603	89.858	.274	.548	-1.399	2.321

t-distribution, df = 98



Ok, so bootstrapping, finally.

Because you're never convinced, read the word of the inventors of the bootstrapping method, Bradley Efron and Robert J. Tibshirani:

1. Efron & Tibshirani (1993) – *An Introduction to the Bootstrap*

"The bootstrap does not create new information; it merely resamples the existing data. The effective sample size remains the same as the original dataset, and any bias or limitations present in the original sample are carried over into the bootstrap replicates."

— (*An Introduction to the Bootstrap*, B. Efron & R. Tibshirani, 1993, p. 50)

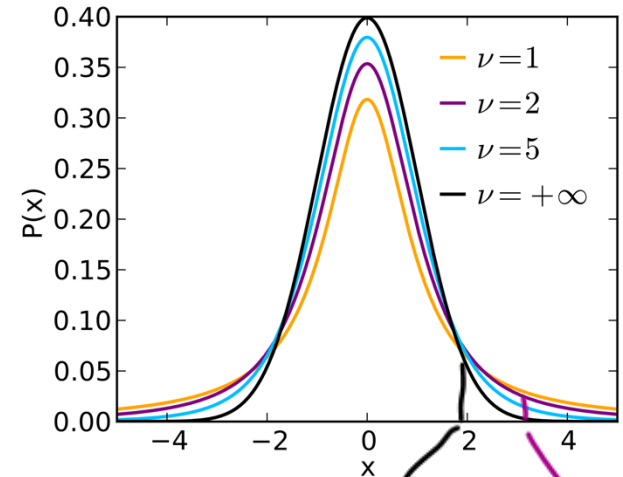
What does that entail?

If the effective sample size remains the same as the original dataset, your *statistical power* or the *effect size* remains the same. Your estimation can be more *accurate*, but it's not going to make it more *trustworthy*. To generalize your results (which is the ultimate goal of any hypothesis testing: making a claim about unseen population), you better aim to increase *power* by adding more independent observations.

Bootstrapping helps estimating the variability of a statistic

You may have a small sample size ($n < 30$ for one or each sample). Your estimate of the population mean (or the population mean difference, if two independent samples t-test) will have a wide *confidence interval*. Check the formula below: if n is *small*, $t_{\frac{\alpha}{2}, n-1}(\frac{s}{\sqrt{n}})$ will be much dependent on your sample's variability level (s). Also, $t_{\frac{\alpha}{2}, n-1}$ is bigger for smaller n .

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right)$$



$$-t_{\frac{0.05}{2}, n=\infty} = \sim 1.96$$

vs.

$$-t_{\frac{0.05}{2}, n=3} = 3.18$$

Bootstrapping helps estimating the variability of a statistic

You always want to be more accurate in terms of your estimation. In terms of the confidence interval, you want to provide a narrow one. In order to narrow down the confidence interval of your population mean estimate, you want to increase n . Bootstrapping will increase this value, because you resample from your original

sample. s will also decrease as n increases: $s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2}$

Back to our example...

Bootstrapping returns a **comparable, slightly narrower** confidence interval, so you can use this result to support that your estimate of the population mean difference based on your sample size is accurate / variability of the estimate is well estimated by the numbers you have. So if you have a small sample size, you can consider using bootstrapping to make such a claim.

Bootstrap for Independent Samples Test

Sampling = "Simple"

Sampling = “Simple”		Mean Difference	Bias	Std. Error	Bootstrap ^a	
					95% Confidence Interval Lower	95% Confidence Interval Upper
sbp	Equal variances assumed	1.399	−.024	2.325	−3.139	5.908
	Equal variances not assumed	1.399	−.024	2.325	−3.139	5.908

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Test

t-test for Equality of Means						95% Confidence Interval of the Difference		
t	df	Significance		Mean Difference	Std. Error Difference	Lower	Upper	
		One-Sided p	Two-Sided p					
9)	.607	98	.273	.545	1.399	2.305	-3.174	5.973
	.603	89.858	.274	.548	1.399	2.321	-3.211	6.010

Parametric test results

However, bootstrapping is limited for several reasons

Summary of Bootstrapping Limitations

Limitation	Why It Matters
Doesn't Increase Sample Size	Cannot fix small sample bias
Relies on Sample Representativeness	Biased samples → Biased bootstraps
Struggles with Skewed Distributions	May fail to capture rare extreme values
Computationally Intensive	Large-scale models take too long
Assumes Independence	Fails with time series, clustered data
Weak with Small Samples	Limited resampling makes it unreliable
Can Overestimate Certainty	May underestimate uncertainty in poorly sampled areas

However, bootstrapping is limited for several reasons

1. Bootstrapping Does Not Increase Sample Size

- It **reuses** the existing data but **does not generate** new information.
- The effective sample size remains the same, so it **cannot correct for small sample bias**.
- If your original sample is unrepresentative, bootstrapping will only **reinforce** that bias.

Example:

If you only have **10 subjects** in a study, bootstrapping **cannot magically make it feel like you had 100 subjects**.

However, bootstrapping is limited for several reasons

5. Does Not Work Well with Strongly Dependent Data

- Standard bootstrapping assumes **independent** observations.
- If data points are **correlated** (e.g., **time series**, **spatial data**, **repeated measures**), naive bootstrapping **breaks the structure** and can lead to misleading results.

Solution:

- Use **block bootstrapping** for time series data.
- Use **cluster bootstrapping** for hierarchical or grouped data.

6. Less Effective for Small Sample Sizes

- If your original sample is **very small** ($n < 30$), bootstrapping can be unreliable.
- The resamples may not fully capture the underlying variability.

Example:

Bootstrapping **5 data points** will just keep repeating those values, making the resampling pointless.

Thorpe et al. (2017)

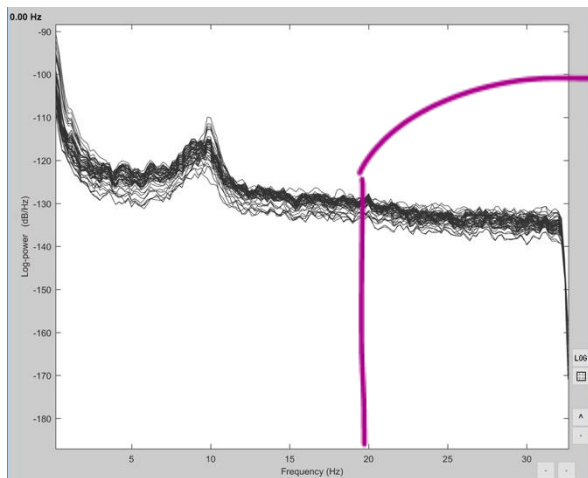
You always mention this paper. Let's read the exact sentence:

To assess the variability in the expected log PSD in each age group we computed empirical bootstrap distributions at each channel and frequency. Each bootstrap sample was computed

Is it clearer now? They also used bootstrapping method to assess **the variability of an estimate**: expected log PSD in each age group.

Thorpe et al. (2017)

For each subject, the resultant PSD were then log transformed (example plot below - totally unrelated; just for demonstration purpose)

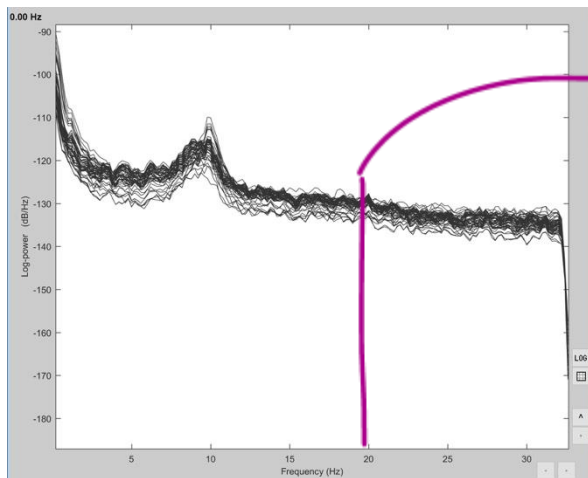


... we computed empirical bootstrap distributions at each channel and frequency.

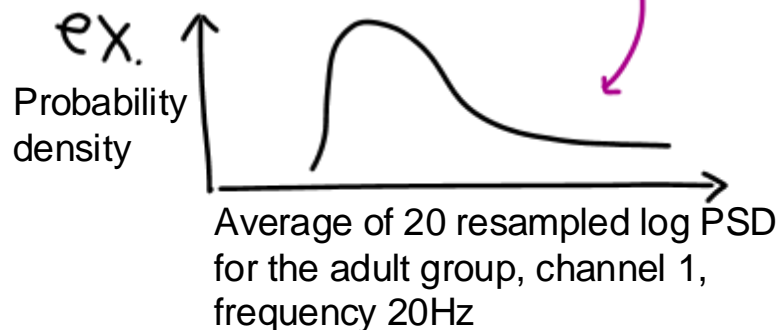
- 1) Log PSD distributions were sampled (with replacement) from all subjects a total of N_s times, where N_s is the number of subjects in each group.
>> so I assume for adult group, for each channel/frequency (ex. channel 1, frequency 20 Hz), values of 20 participants were resampled 20 times with replacement. The average of 20 resampled values was one estimate of log PSD of the adult group's channel 1, frequency 20 Hz

Thorpe et al. (2017)

For each subject, the resultant PSD were then log transformed (example plot below - totally unrelated; just for demonstration purpose)

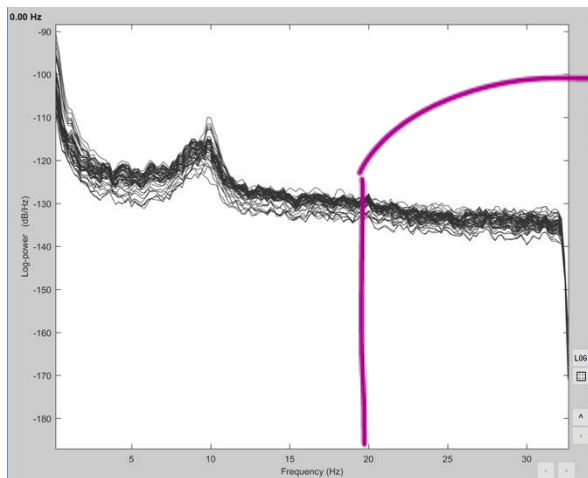


- 2) This resample -> averaging was repeated 10000 times, resulting in distributions describing the variability in expected log PSD for each channel/frequency pair.



Thorpe et al. (2017)

For each subject, the resultant PSD were then log transformed (example plot below - totally unrelated; just for demonstration purpose)

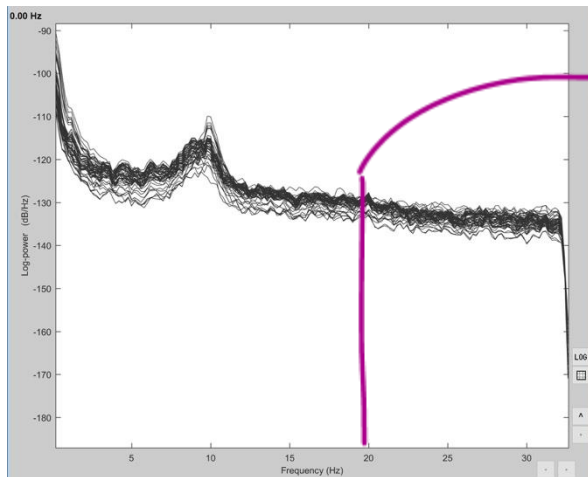


3) To determine alpha frequency bands of interest for each age group we averaged each bootstrap sample across channels, then defined an “alpha spectral peak” as any frequency in a broadly defined band of interest (set as 4-13 Hz for all age groups) which showed greater averaged log PSD than both its adjacent neighbors).

>> For each of 10,000 bootstrap samples – at each iteration, a single estimate of log PSD (average of 20 (N_s of the adult group) resampled log PSD values of the participants) is prepared for each channel/frequency. You average over channels, so that you have one estimate for each frequency value.

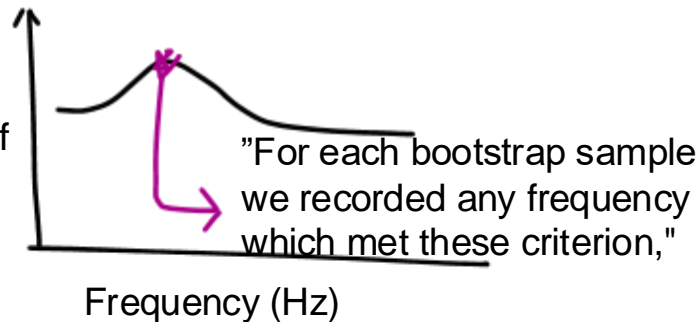
Thorpe et al. (2017)

For each subject, the resultant PSD were then log transformed (example plot below - totally unrelated; just for demonstration purpose)



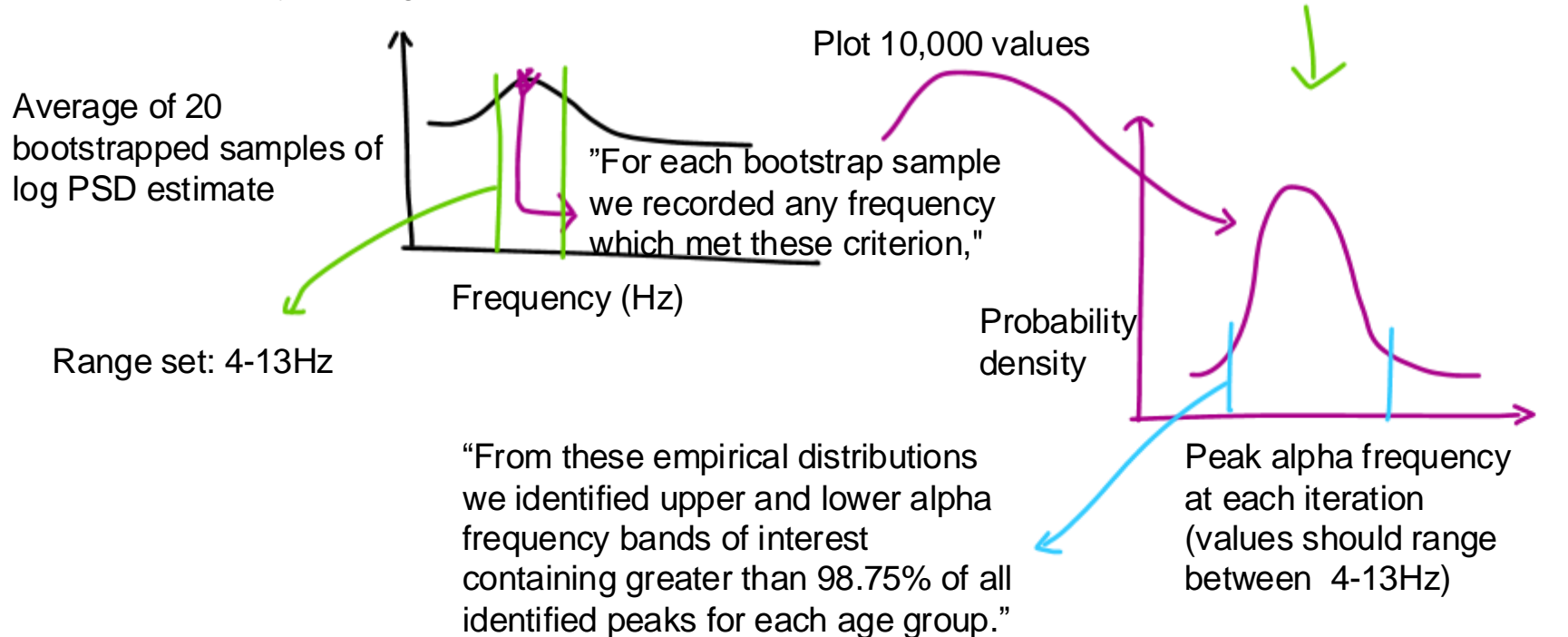
3) (cont'd) Averaging over channels at each iteration of 10,000 repetitions, you will get a distribution like below:

Average of 20
bootstrapped samples of
log PSD estimate



Thorpe et al. (2017)

3) (cont'd) Averaging over channels at each iteration of 10,000 repetitions, you will get a distribution like below:

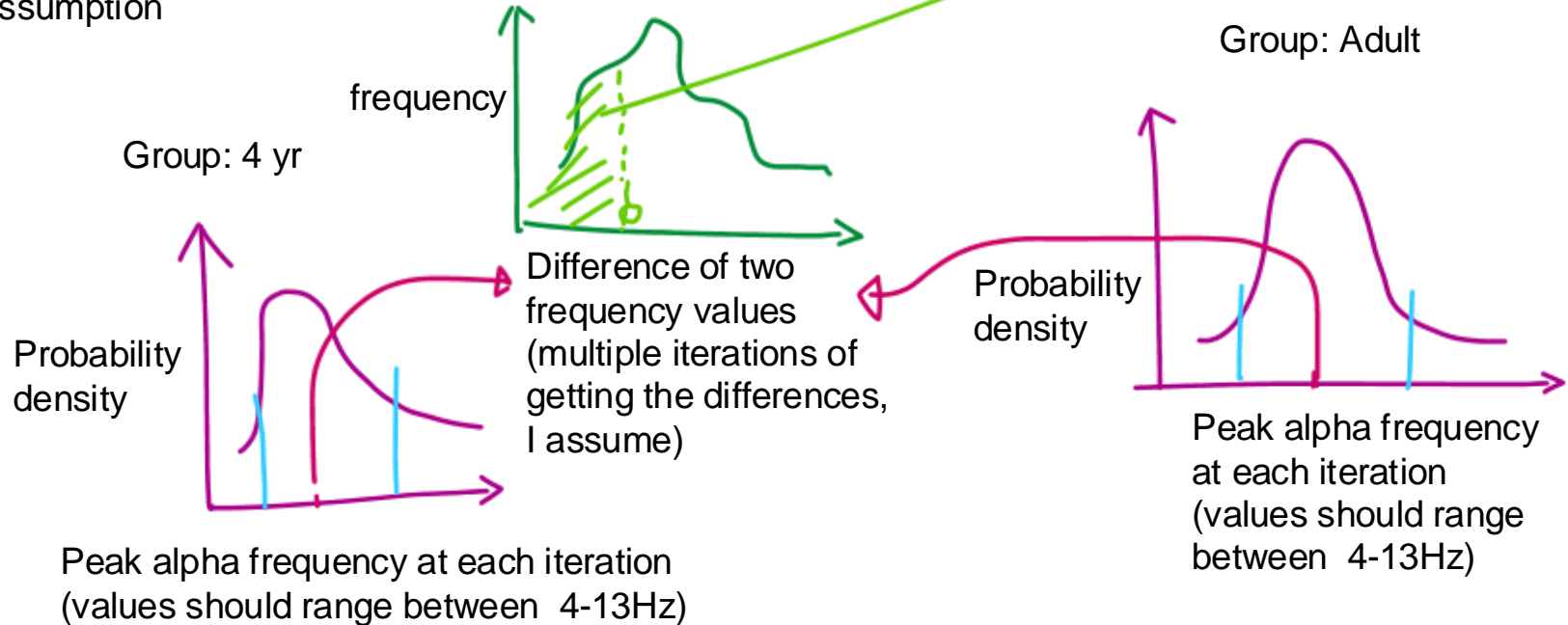


Thorpe et al. (2017)

p.9

4) To assess the significance of differences between the mean frequencies of these clusters across age groups, empirical p-values were computed (Not very clear, but here's my assumption)

Empirical p-value: the percentage of all pair-wise comparisons for which the difference was less than zero



Thorpe et al. (2017)

To summarize, the authors used bootstrap method to assess the variability in the expected log PSD in each group.

This is a feasible strategy, given that you cannot collect hundreds or thousands, not to mention 10,000 EEGs.

Again, it doesn't mean that there's no value in collecting more data from actual human beings. Using bootstrapping is encouraged – know what the tool is designed for and use it appropriately.