Statistical inference & testing playground

June 2023

Francesco

TIER

Agenda

1. **Key terms: statistical inference & testing**

2. **Practice!**

TIER

Key terms

For statistical inference & testing

Statistic, Population vs Sample, Estimators, Inference

A ***statistic*** is any quantity computed from values **in a sample** which is considered for a statistical purpose (eg. estimating a population parameter, describing a sample, or evaluating a hypothesis). Its distribution is called ***sampling distribution***. For ex., the average (or mean) of sample values is a statistic. The term statistic is used both for the function and for the value of the function on a given sample.

When a statistic is used for estimating a population parameter, the statistic is called an ***estimator***. A ***population parameter*** is any characteristic of a population under study, but when it is not feasible to directly measure the value of a population parameter, statistical methods are used to **infer** the likely value of the parameter on the basis of a statistic computed from just a sample (***inference***).
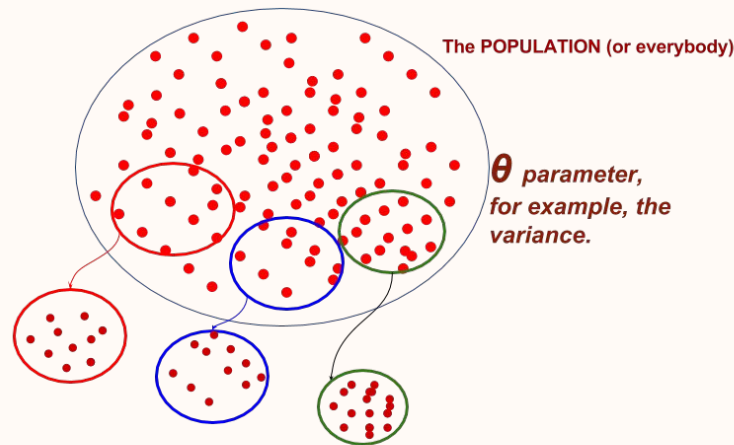

Population

Sampling

Sample

Inference

TIER
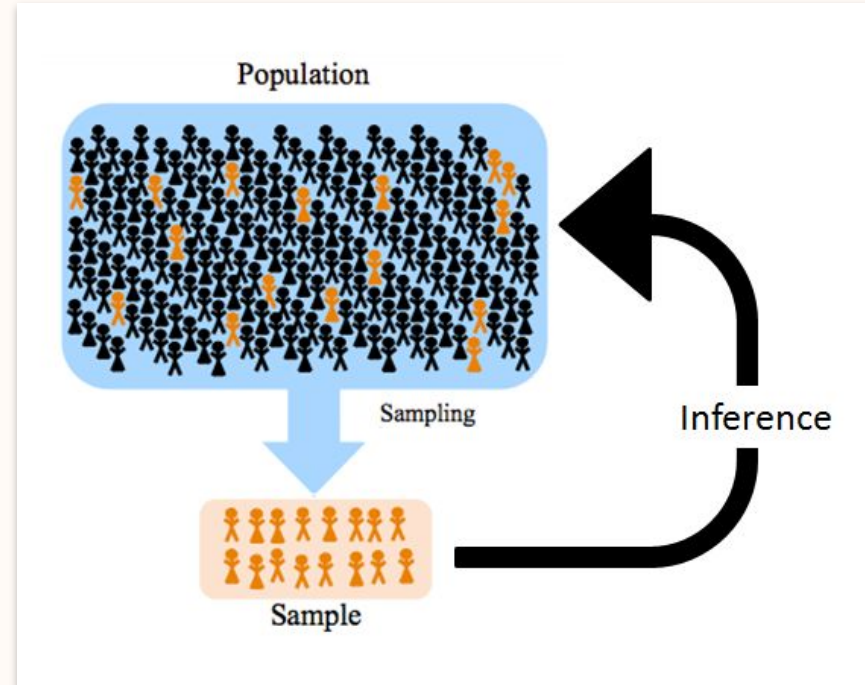
Problems when estimating parameters: Bias

The **_bias of an estimator_** is the difference between an estimator's expected value and the true value of the parameter being estimated**. Although an unbiased estimator is theoretically preferable to a biased estimator, in practice, biased estimators with small biases are frequently used.

A biased estimator may be more useful for several reasons. First, an unbiased estimator may not exist without further assumptions. Second, sometimes an unbiased estimator is hard to compute (eg. unbiased est of st.dev). Third, a biased estimator may have a lower value of mean squared error (eg. shrinkage estimators).

BIASED ESTIMATOR:

The POPULATION (or everybody)

θ parameter, for example, the variance.

The sample variances $\bar{\theta}$ will NOT reflect θ accurately, so that the $E[\bar{\theta}]$, or expectation (mean) of the sample variance is different from the population variance:

$$Bias = E[\bar{\theta}] - \theta \neq 0$$

Note: The OLS regression **beta** coefficient is also an example of an estimator (and btw - it's **BLUE**!)
The Gauss-Markov Theorem and BLUE OLS Coefficient Estimates - Statistics By Jim

TIER

Problems when estimating parameters: Standard error

The *standard error* of a statistic is the standard deviation of its sampling distribution, or an estimate of that standard deviation. If the statistic is the sample mean, it is called the **standard error of the mean (SEM)**.

The relationship between the standard error of the mean and the standard deviation is such that, for a given sample size, the standard error of the mean equals the standard deviation divided by the square root of the sample size (if samples are uncorrelated). **In other words, the standard error of the mean is a measure of the dispersion of sample means around the population mean.**

An important implication of this formula, due to the sqrt(n) factor, is that **reducing the error on the estimate by a factor of ten requires a hundred times (100x) as many observations**!

Standard Deviation….

Vs.

…Standard Error!!!

TIER

Problems when estimating parameters: Confidence intervals

A *confidence interval* (CI) is a range of estimates for an unknown parameter. A confidence interval is computed at a designated confidence level (95% being the most common). **The confidence level represents the long-run proportion of CIs** (at the given confidence level) **that theoretically contain the true value of the parameter**. For example, out of all intervals computed at the 95% level, 95% of them should contain the parameter's true value.

All else being the same, a **larger sample** produces a narrower confidence interval, **greater variability** in the sample produces a wider confidence interval, and a **higher confidence level** produces a wider confidence interval.
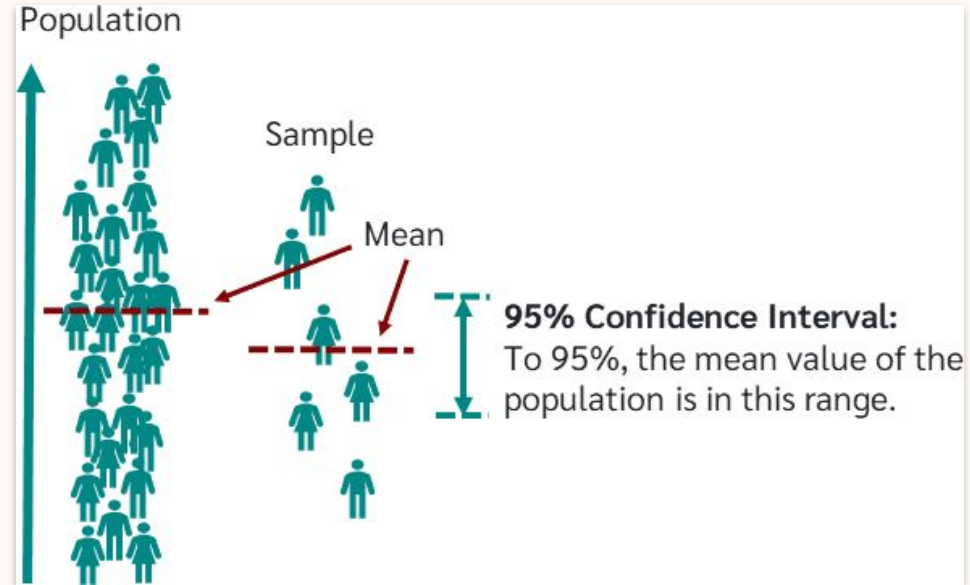
Note: if you want to know more about the true meaning of confidence intervals, see: *#Caution when using confidence intervals* or see *example from Scipy's docs*

TIER

How to make inference more robust: Bootstrapping

- Bootstrapping is, first of all, a **practical way of finding the sampling distribution**, at least approximately, from just one sample, instead of drawing multiple random samples from a population

- **It's very useful especially for inference**, to compute measures of accuracy for sample estimates (bias, standard error, confidence intervals, prediction errors, …)

- It simulates drawing multiple random samples by using just one sample from a population (*sampling with replacement*), so it very much relies on this single sample being as random as possible.

- Bootstrap testing **does not impose the normality restriction** on the distribution of the population(s) that the actual sample(s) is(are) drawn from, unlike the traditional t-test, and **it can be done for any statistic**!
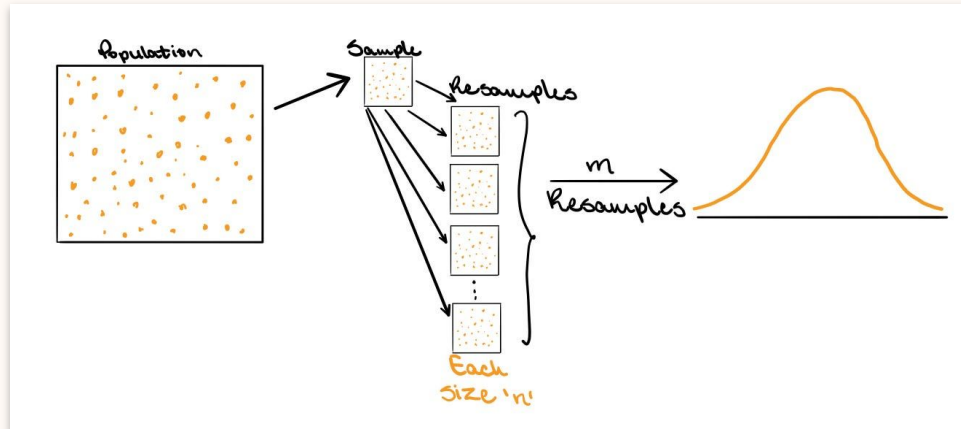

Bonus! *Is there a minimum sample size required for bootstrapping?*
In short: not really, it can work even for small sample sizes (eg. 20 obs); if they're not unreasonably small (eg. 4) - see answer here from the author of this book (or here).

TIER

Hypothesis testing, p-values, statistical significance

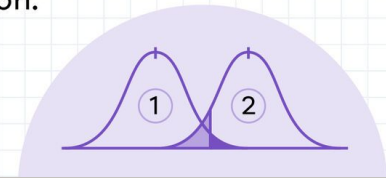A ***statistical hypothesis test*** is a method of statistical inference used to decide **whether the data at hand sufficiently support a particular hypothesis**. Hypothesis testing allows us to make probabilistic statements about population parameters.
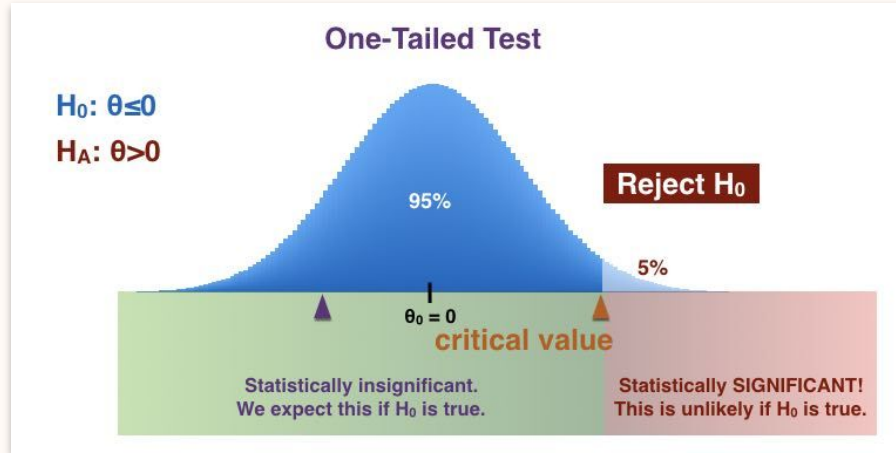
So many tests...and so many assumptions!
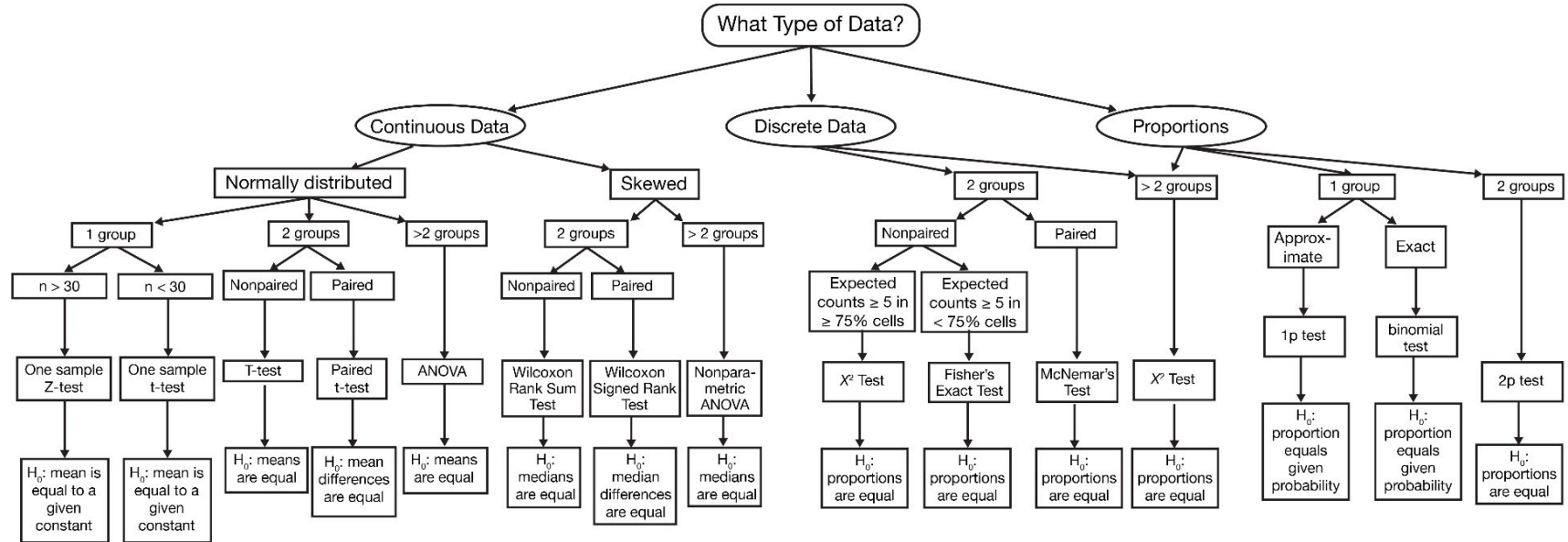

Flow chart: which test statistic should you use?

Problems of traditional hypothesis testing

Distribution assumptions that traditional tests do are not always accurate in reality.

4. In observational studies, the distribution of the test statistic under the null hypothesis is not known.

We often, mistakenly, hope/think that the distribution is that same nominal distribution as if a true experiment had been conducted (e.g., $F, t, z, x^{\#}$). If hypotheses are formed after looking at the data (data dredging) then the ability to make valid inference is severely compromised (e.g., model-based standard errors are not a valid measure of precision).

TIER

But in reality...There's only one test

*"**All tests try to answer the same question: "Is the apparent effect real, or is it due to chance?"** To answer that question, we formulate two hypotheses: the null hypothesis, H0, is a model of the system if the effect is due to chance; the alternate hypothesis, HA, is a model where the effect is real.*

*Ideally we should compute the probability of seeing the effect (E) under both hypotheses; that is P(E | H0) and P(E | HA). But formulating HA is not always easy, **so in conventional hypothesis testing, we just compute P(E | H0), which is the p-value**. If the p-value is small, we conclude that the effect is unlikely to have occurred by chance, which suggests that it is real.*

*That's hypothesis testing. All of the so-called tests you learn in statistics class are just ways to compute p-values efficiently. When computation was expensive, these shortcuts were important, **but now that computation is virtually free, they are not**.*

*And the shortcuts are **often based on simplifying assumptions and approximations**. If you violate the assumptions, the results can be misleading, which is why statistics classes are filled with dire warnings and students end up paralyzed with fear.*

***Fortunately, there is a simple alternative: simulation**. If you can construct a model of the null hypothesis, you can estimate p-values just by counting."*

TIER

How to build our own test

We just need two ingredients:

1. The **distribution of the possible outcomes** under the null hypothesis (which we can build via simulation).

2. A **measure of the "unexpectedness" of any outcome** (basically a test statistic).

Finally, the **p-value** will simply be the percentage of "null" scores that are higher than the observed score.


Article by S. Mazzanti

TIER
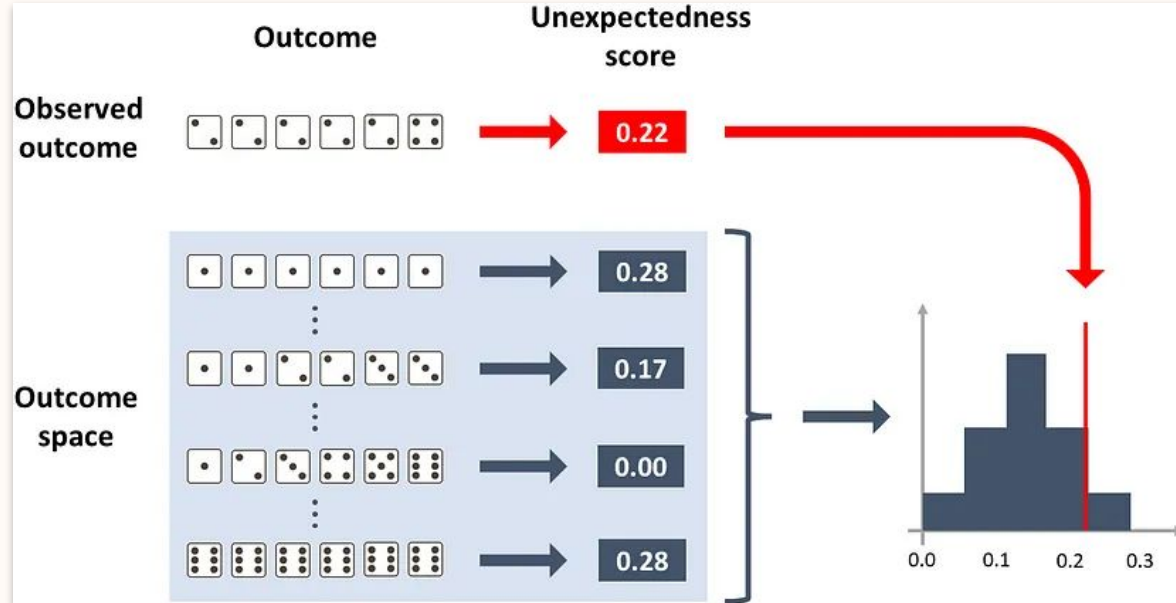
Enters Permutation Testing

A **permutation test** (also called re-randomization test) is an exact statistical hypothesis test making use of the proof by contradiction.

A permutation test involves two or more samples. The null hypothesis is that all samples come from the same distribution. **Under the null hypothesis, the distribution of the test statistic is obtained by calculating all possible values of the test statistic under possible rearrangements of the observed data** ([animation of process](#)).

Permutation tests exist in many situations where **parametric tests do not** and **exist for any test statistic**, regardless of whether or not its distribution is known.

TIER

Histogram of Null Distribution
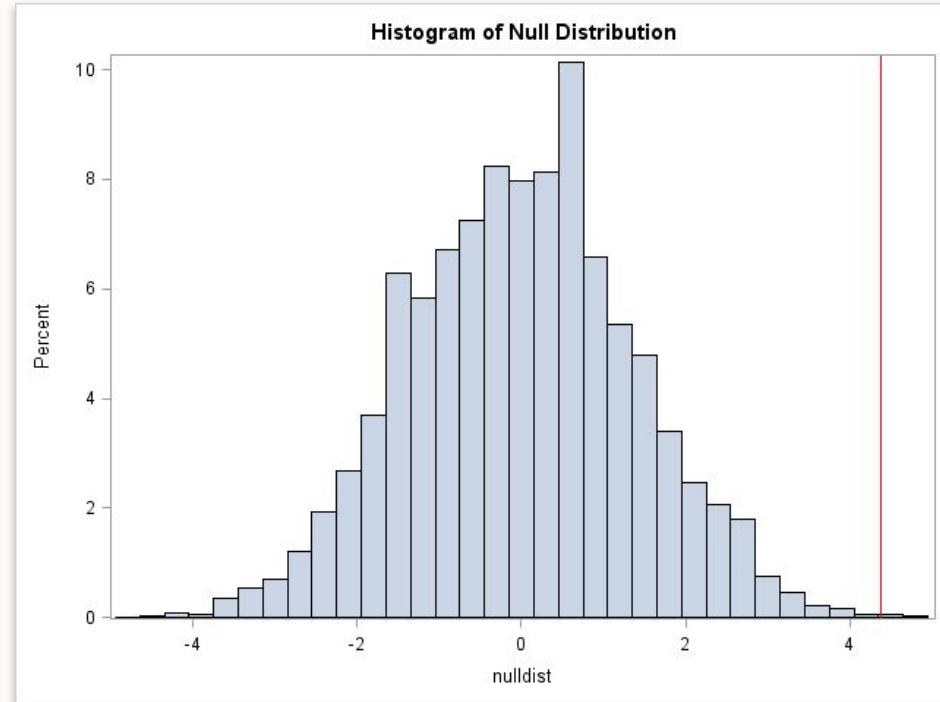
Note: Permutations are also very useful in ML settings, eg. for **feature importance**: [4.2. Permutation feature importance](#)

*For a **concrete example** of permutation feature importance, see **section 3.1** here: [Availability metrics with biggest impact on Missed Demand.pdf](#)*

Permutation tests & parametric tests

Note that **all simple and many relatively complex parametric tests have a corresponding permutation test version** that is defined by using the same test statistic as the parametric test, but obtains the p-value from the sample-specific permutation distribution of that statistic rather than from the theoretical distribution derived from the parametric assumption (eg. a permutation t-test, a permutation χ2 test of association, a permutation version of Aly's test for comparing variances, etc.)

Permutation tests are also already more common that one might think (eg. Fisher's exact test and Mann-Whitney/Wilcoxon's test are actually special cases of a permutation test).

Histogram of permuted t-statistic, log-transformed

TIER
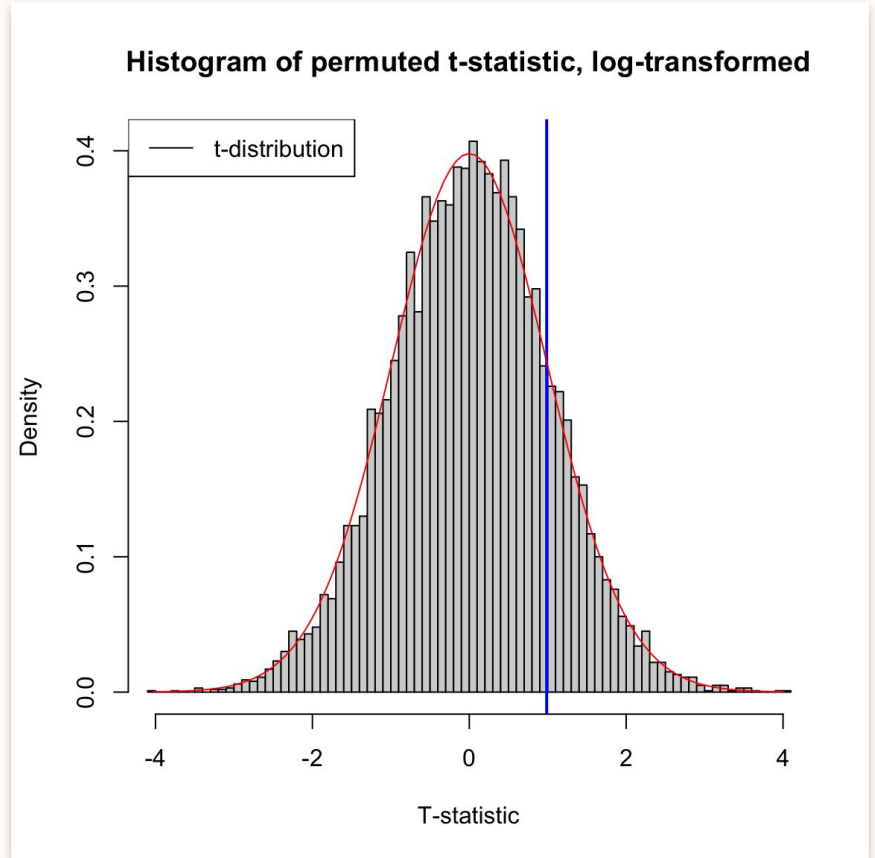
For more details on traditional parametric testing...

Other important concepts that aren't currently covered here but relevant for our practice session:

- **T distribution and T-test** (test of means, 2 groups)
- **F distribution and ANOVA; Kruskal–Wallis and Alexander-Govern** (tests of means, 3+ groups)
- **Z-test** (test of proportions)
- **Chi-square distribution and chi-square contingency tests** (test of proportions)
- **Exact tests: Fisher, Barnard, Boschloo** (test of proportions)
- **Power analysis & sample size calculation**

TIER
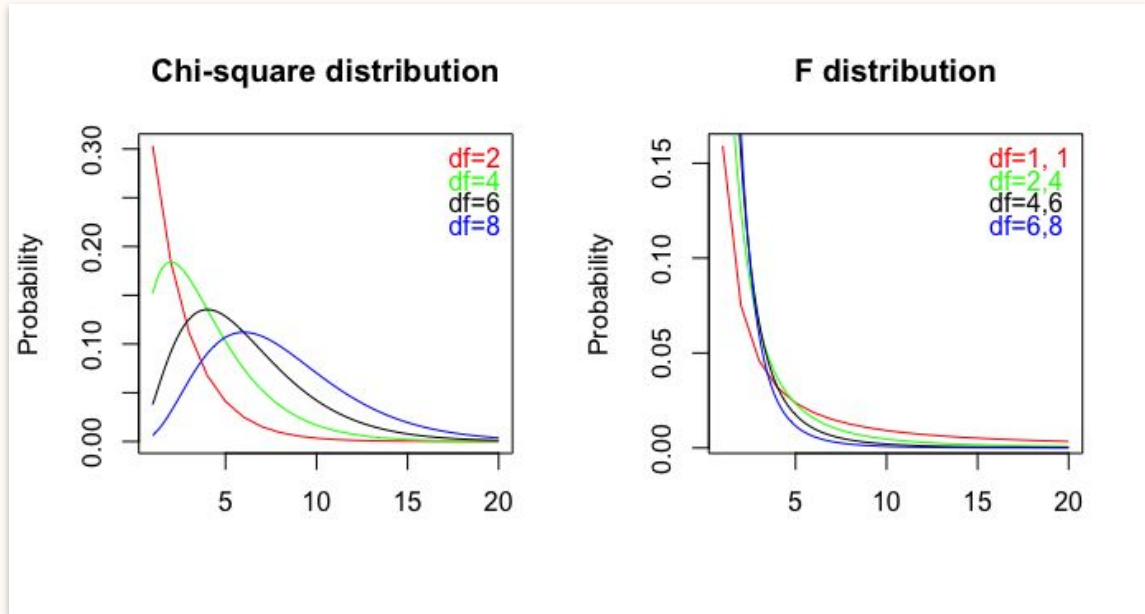
Practice!

Statistics playground content (1/2)

What we will be practising by running some Python code:

- **Bootstrapping (for inference)**
  - Computing estimates, bias, standard errors and CI using bootstrapping for any statistics
  - Comparing results vs approx formulas for the "mean"
  - Looking into technical implementation & speed of different implementations
- **Permutation tests vs T-tests vs OLS (2 groups)**
  - Comparing p-values from T-tests vs permutation tests (testing diff in means between 2 groups)
  - Comparing different technical implementations & speed, incl vectorised function
  - Comparing against OLS model results
  - Computing power of the test
- **Permutation tests of other multi-sample statistics** (eg. Pearson correlation)
- **ANOVA & post-hoc tests (3+ groups)**
  - Simple ANOVA results vs. results of permutation ANOVA test
  - Comparison with Kruskal–Wallis and Alexander-Govern tests (relaxing some of ANOVA's assumptions) and OLS models results
  - Post-hoc tests (Tukey's HSD test)

TIER

Statistics playground content (2/2)

(cont.d)

- **Z-test (2 groups)**
  - Comparing proportions of 2 different groups
  - Checking results vs. permutation proportion test
- **Chi-square contingency test (3+ groups)**
  - Comparing proportions of 3+ different groups
- **Exact test of proportions (2 groups)**
  - Testing difference of proportions between groups of small sample size with exact tests: Fisher, Barnard, Boschloo
- **Sample size calculations**
  - Compute required sample size for a T-test (diff in means), given standardised effect size, power & significance level (SEF, P & SL)
  - Compute required sample size for a Z-test (diff in proportions), given SEF, P & SL

TIER