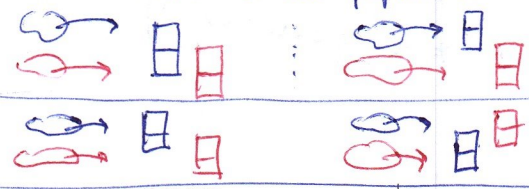
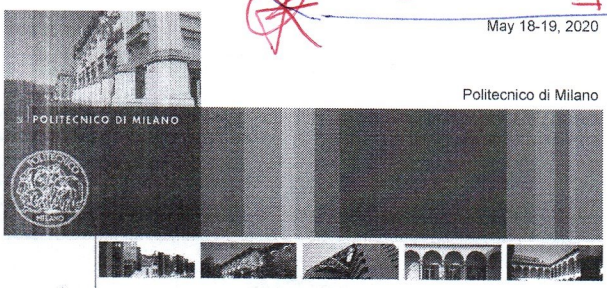
(*) this is how Fisher explained ~~the~~ how the Null hypothesis test works: when we compute the student t statistics and we compute the t value basically we're comparing our two samples with all the other samples that ~~can~~ could have been drawn from the two population



⇒ we have evidence that they differ

⇒ we don't have evidence that they differ



May 18-19, 2020

Politecnico di Milano

Applied Statistics 2019-2020

Permutation Tests → why? 2 problems:
- gaussianity
- sample dimensionality

Simone Vantini
MOX, Dept. of Mathematics, Politecnico di Milano
Leonardo Campus, Building 14, Floor VI,
02 2399 4584
simone.vantini@polimi.it

Gaussian Assumption and Parametric Tests for the Mean(s)

p = # random variables (features)

$1 = p < n = \infty$	Thanks to the Central Limit Theorem, Gaussianity is not a key point.
$1 = p < n \leq \infty$	The t-distribution is meant to model situations in which the sample size is not very large. So the Gaussianity of data is required for the t-test. Univariate Gaussianity is, anyhow, not difficult to assess (normality tests).
$1 \leq p < n \leq \infty$	Hotelling's T^2 test rely on multivariate Gaussianity of data. If p increases, multivariate Gaussianity can be difficult to assess (curse of dimensionality).
$1 \leq n < p \leq \infty$	High-dimensional tests rely on the multivariate Gaussianity of data, and they are not robust with respect to the violation of Gaussianity. Powerful Gaussianity tests are not available in the high-dimensional setting.
$1 \leq n < p = \infty$	In the functional case, normality is basically an unverifiable assumption.

We want to get rid of gaussianity assumption and develope tools that can achieve the same results as parametric tests by passing gaussianity.
→ that work with small samples or non-gaussian samples (or both)

we want to do something like the t-test:
- we compare the sample with all the samples that could have been drawn by the population that we model with a gaussian distribution

What Fisher was trying to say is: if we have 2 samples and we observe a difference of 1 inch in heights this is not enough to say that the two populations differ of 1 inch in heights. These are only 2 samples.

H_0: Englishmen taller of 1 inch wrt Frenchmen

Gaussian Assumption and Parametric Tests for the Mean(s)

All parametric tests (for the means) are exact either **asymptotically** or under the **Gaussianity assumption**

and

not exact otherwise

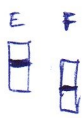
low sample size and no gaussianity creates problems

At the beginning of permutation tests

Let us suppose, for example, that we have measurements of the stature of ***a hundred Englishmen and a hundred Frenchmen****. It may be that* ***the first group are, on the average, an inch taller than the second****, although the two sets of heights will overlap widely. [...] The simplest way of understanding quite rigorously, yet without mathematics, what the calculations of the test of significance amount to, is to consider what would happen* ***if our two hundred actual measurements were written on cards, shuffled without regard to nationality, and divided at random into two new groups of a hundred each****. This division could be done in an enormous number of ways, but though the number is enormous it is a finite and a calculable number. We may suppose that* ***for each of these ways the difference between the two average statures is calculated****. Sometimes it will be less than an inch, sometimes greater. If* ***it is very seldom greater than an inch, in only one hundredth, for example****, of the ways in which the sub-division can possibly be made, the statistician will have been right in saying that* ***the samples differed significantly****. For* ***if****, in fact,* ***the two populations were homogeneous, there would be nothing to distinguish the particular subdivision*** *in which the Frenchmen are separated from the Englishmen* ***from among the aggregate of the other possible separations*** *which might have been made. Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method.*

Fisher, R. A. (1936). The coefficient of racial likeness and the future of craniometry, *Journal of the Anthropological Institute of Great Britain and Ireland*, pp. 57-63.

What we have:



we mix up heights in every way $\binom{200}{100}$ and we re-calculate the average: if the difference between averages is > 1 inch only few times (order of 100 units over $\binom{200}{100}$ possibilities) ⇒ the samples differs significantly

We shuffle nights:
if the original grouping is very extreme wrt the second → we reject H_0 (so that they're different)
If the original grouping is not very different from all the other grouping that we obtain mixing up all the heights
→ we don't reject H_0

this is a discrete-time distribution with 35 ~~values~~ possible values all with the same probability (1/35). Why?

Because all these permuted samples are constructed in the same way.

PERMUTATIONAL DISTRIBUTION (/ CONDITIONAL DISTR.) under null hypot.

this is the distribution of the test statistics under the null hypothesis conditionally on pooled set

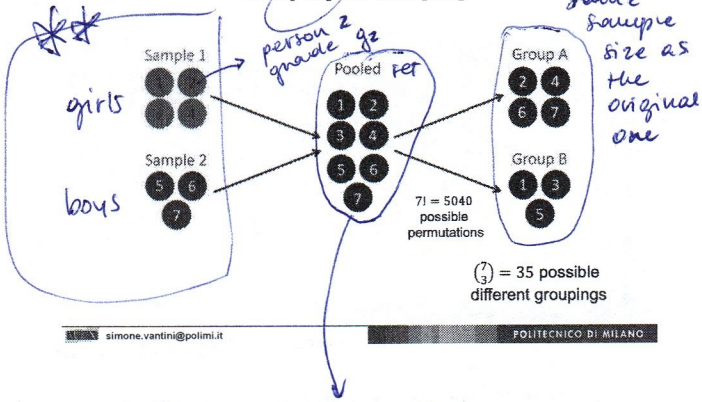
histogram build on the 35 possible values of the test statistics T

$\mathbb{E}[\text{grade(girls)}] \stackrel{?}{=} \mathbb{E}[\text{grade(boys)}]$

The two-population T-test in a glance

5

$H0: m_1 = m_2$ vs $H1: m_1 > m_2$



person 2 grade g_2

same sample size as the original one

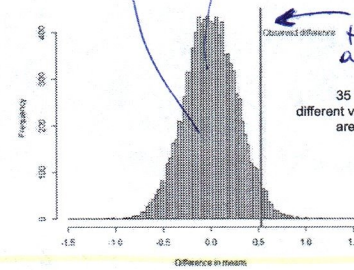
we don't care about the order; the Pooled sample is a set, not a vector

idea: compare the original couple of samples with the 35 possible pairs of groups we could obtain by permuting (the parametric test would compare (***) with all the possible couples of samples which could be drawn by sampling 4 times from a gaussian and other 3 times (independently) from the same gaussian). Notice that we have no idea on how the data are distributed

The two-population T-test in a glance

6

Permutational distribution of $T = \bar{x}_1 - \bar{x}_2$ **under the H0**



original value of the test statistic associate to the original grouping of the data

Conditional distribution of $T = \bar{x}_1 - \bar{x}_2$ given the pooled sample:
- Under the H0 all the 35 values of $\bar{x}_1 - \bar{x}_2$ are equally probable
- Under the H1 larger values of $\bar{x}_1 - \bar{x}_2$ are more probable

$$p = \frac{\sum_{k=1}^{35} I(T_k^* \geq T_0)}{35}$$

we have to compare the original sample with the permutated one ⇒ test statistics T

(***) ⇒ the permutation distribution of the test statistics is always a uniform discrete distribution. (under H_0)

when we compute the permutational distribution we can define the p-value as the number of permutations which has led to a sample whose value of the test statistic is larger or equal than the original one.

Permutation Tests

7

Some **corner-stones of permutational inference**:

- **Their aim is making fewer assumptions as possible on data distribution**
- **How they work**:
  - Likelihood-invariant transformations under the H0 and not invariant under H_1 (Conditional inference within induced equivalence classes)
  - Selection of the test statistic:
    - no a-priori optimal test statisticits distribution has to be stocastically larger under the «targeted» H1 than under the H0
    - possibility of working in purely metric spaces (i.e., complex data)
- **Inferential properties**:
  - Finite-sample exactness (differently from bootstrap)
  - Consistency (if the test statistic is properly chosen)
  - Asymptotic equivalence to parametric tests (when the same test statistic is used and the parametric assumptions hold)
- **Large computational costs** (Conditional Montecarlo)

because of $\binom{n}{m}$

⇒ we do not explore everything (we random sample permutations) (not the smartest method)

this guarantee that the conditional distr. is uniform described under H_0 and not uniform described under H_1

Likelihood-invariant transformations

8

Two-population test and 1-way ANOVA:
→ Value permutations (equivalent to group labels permutation)

One-population test and and paired two-popolution test:
→ Recentering in H0 and sign swaps (assuming simmetry)

Independence test:
→ Pair Recoupling

we have a point and we want to know if the center of the (symmetrical) distribution is equal to that point (we know nothing about the distribution (shape, ...))

"F-test" for linear models (linear regression and multi-way ANOVA)
→ Response permutations

"T-test" for linear models (linear regression and multi-way ANOVA)
→ Permutations of residuals of restricted model [*asymptotic*]
→ Permutations of residuals of complete model [*asymptotic*]

they're not exact because we would like to permute the errors but they're not observed, we only have residuals

March 2019 By Jared Wilber

THE PERMUTATION TEST

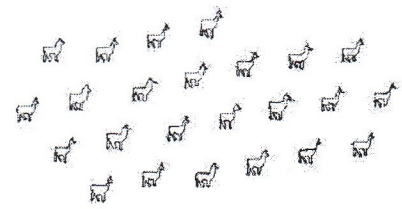
A Visual Explanation of Statistical Testing

Statistical tests, also known as hypothesis tests, are used in the design of experiments to measure the effect of some treatment(s) on experimental units. They are employed in a large number of contexts: Oncologists use them to measure the efficacy of new treatment options for cancer. Google uses them to determine which color of blue (e.g. this blue vs this blue) is most effective for outgoing links. And entomologists use them to study the sex habits of flies *.

Unfortunately, a lot of statistical tests require complex assumptions and convoluted formula. This is especially true of those methods taught in introductory courses, giving the false impression that experimental design is boring and unintuitive. But fret not, my valued reader - not all tests are so bad! In what follows, I present a visual explanation for the permutation test: an awesome nonparametric test that is light on assumptions, widely applicable, and very intuitive.

You're An Alpaca Shepherd Now

You've finally achieved your lifelong dream: you're an alpaca shepherd. And like any alpaca shepherd will tell you, your foremost concern is the wool quality of your herd.



Word on the street in Cusco is that a popular new shampoo increases the wool quality of your alpaca. But you're no sucker - you're going to find out for sure. You're going to test the difference with statistics.

In statistical testing, we structure experiments in terms of null & alternative hypotheses. Our test will have the following hypothesis schema:

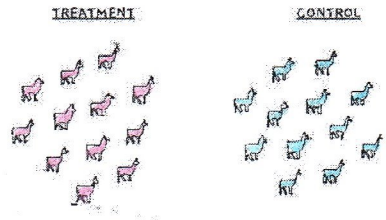
$$H_0: \mu_{treatment} <= \mu_{control}$$
$$H_A: \mu_{treatment} > \mu_{control}$$

Our null hypothesis claims that the new shampoo does *not* increase wool quality. The alternative hypothesis claims the opposite; new shampoo yields superior wool quality.

Randomization

As a first step, we randomly assign half of our sampled alpaca to the new shampoo, and half to the old. We say that the alpaca receiving the new shampoo belong to the *treatment* group, and the others to the *control* group. The assignment of an alpaca to a given diet is known as its *treatment assignment*.

Randomization of treatment assignment is very important. It removes bias and confounding from our results, and provides the basis for the theory underpinning our statistical test.



Response Values

After giving each alpaca its designated shampoo, we determine if the new shampoo has any effect on wool quality.

In statistics jargon, every experimental unit has a *response value*. For us, each alpaca is an experimental unit, and its measure of wool quality after taking its shampoo is its response value.

We can eyeball these values ourselves and get a feel for any perceived differences between the two shampoos. However, , we'll need a more rigourous method to determine if the differences are statistically significant.

Test Statistic

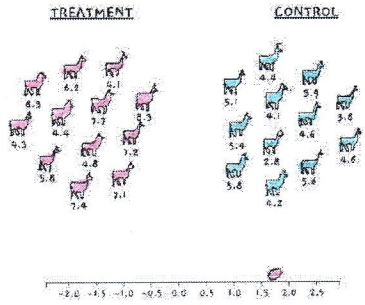
To determine whether or not the new shampoo really is effective, we need a way to quantify the difference between our null and alternative hypotheses.

Luckily for us, such a numerical summary exists: the test statistic.

A benefit of the permutation test is that it allows us to use any numerical value that we want for our test statistic.* Because our analysis is fairly straightforward, we'll simply use the difference in mean response values between the two shampoos:

$$\textit{Test Statistic} = \mu_{\text{Treatment}} - \mu_{\text{Control}}$$

To obtain our initial test statistic, we simply subtract the mean wool quality of the alpacas that used the new shampoo (*treatment group*) from the mean wool quality of the alpacas that did not use the new shampoo (*control group*).



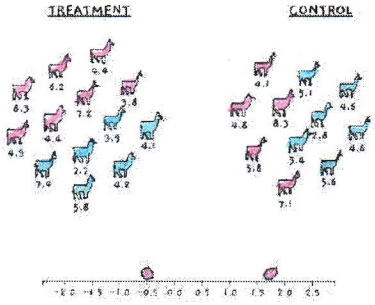
The 'P' in 'Permutation'

Enter the most important step of the permutation test, as well as its namesake.

While keeping the same response values we received earlier, we permute (shuffle) the treatment assignments of our alpaca, and re-calculate our test statistic.

We do this because we analyze the results of our experiment relative to the null hypothesis, which posits the new shampoo as having no benefit on wool quality.

While this may seem a bit odd, the logic is quite simple: if the new shampoo truly doesn't improve wool quality, shuffling the shampoo label of our alpaca and recalculating our test statistic won't matter - we'll obtain similar wool quality values for both groups.

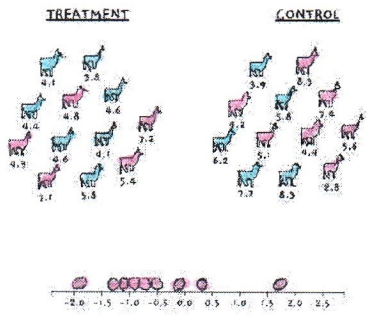


More Permutations

We repeat this process, permuting our data over and over again, and recalculate a test statistic at each iteration.

Ideally, we'd calculate a test statistic for *every* possible permutation of shampoo assignment among our alpaca. This would create an exact distribution of all possible test statistics under our null hypothesis.

Unfortunatley, calculating every permutation is often far too large for practicality. No worries! Instead we'll resample enough permutations to build an approximation to our distribution, as that'll work just as well.

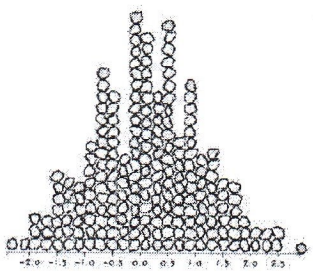


Test Statistic Distribution

Eventually, after some sufficient number of permutations, we create the approximate test statistic distribution.

This distribution approximates all possible test statistic values we could have seen *under the null hypothesis*. We can then use this distribution to obtain probabilities associated with different mean-difference values, where we assume that wool quality does not increase with the new shampoo.

By observing where our initial test statistic falls within this distribution, we obtain the final piece for our test: The magical p-value.

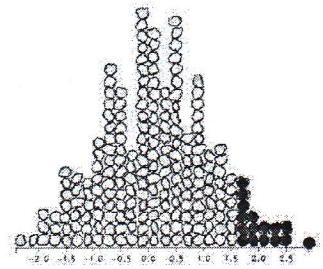


The P-Value

A p-value represents the probability of obtaining the observed values, assuming the null hypothesis is true. For us, it's the probability of obtaining the differences in wool quality we did, *assuming the new shampoo did not increase wool quality.*

To determine the outcome of our test, we compare our p-value to a significance level. This should be determined a priori, but we'll just say ours is 10%. If the p-value is less than or equal to the significance level, we reject the null hypothesis: the outcome is said to be statistically significant.

For us, a low p-value signals that, assuming the null hypothesis is true, the probability of obtaining our initial differences in wool quality occurs with a low probability. A high p-value signals the opposite, such an outcome is likely under the null hypothesis.



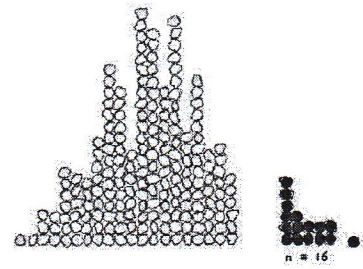
Our Results

To calculate the p-value for a permutation test, we simply count the number of test-statistics *as or more extreme* than our initial test statistic, and divide that number by the total number of test-statistics we calculated.

In our case, only sixteen out of our two-hundred test statistics were as or more extreme than our initial test statistic.

Thus, our p-value is:

P-Value = 16 / 200
= 0.08
= 8%



In other words, if it's truly the case that the new shampoo doesn't improve wool quality, then obtaining the initial difference in wool quality we did occurs with a probability of only 8%.

That's a fairly low probability. In fact, at our 10% level of significance, we reject our null hypothesis and accept our alternative: the new shampoo does appear to be increasing wool quality. Time to buy some more!

Permutation tests:

- the test statistic has to be sensitive to the violation of the null hypothesis.
- the idea is to find transformation of the sample (not only permutations) which are likelihood invariant under null hypothesis and are not likelihood invariant under H_1:
  this will guarantee that the conditional distribution is uniform discrete under H_0 while is not uniform discrete under H_1

there is no optimal test statistic
(like in parametric framework where we have optimal test statistics, where optimal intended as statistic that maximizes the power of the test)
because we should say something about the distribution of the data
(we can check the same test with different test statistics)

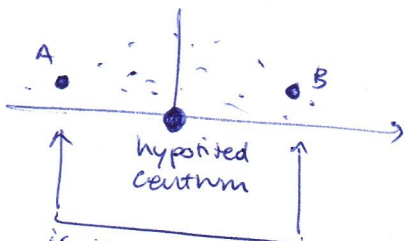
Permutation tests are asymptotic equivalent to parametric tests (with low n permutation tests and parametric tests could lead to different results (even in acceptance/rejecting)).

What tests can we do:

- two population 1-way ANOVA → permutations
- one population, paired two-population → assumptions: symmetry

assuming symmetry we're checking if the center is one given point or not

⟹ we go to the hypotised centrum and we start reflecting:



if H_0 is true (this is the centrum) then if we exchange A with B nothing should change

if center is 0:
we have to change all of the sign of the vector (all of the components)
(that's why they're called SIGN TESTS)

(even if we won't talk about them)

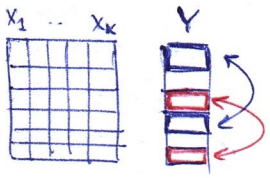
- **Indepence tests:**
  we have 2 variables: we dismount the pairs and under the assumptions of indepenedence we should obtain 2 new samples with the same value of the probability test function.
  If the pairs are dependent the two values are ≠.

- **linear models: F-tests**
  (linear regression and multi-way ANOVA)

  Test if all the regressors simultaneously are / are not significant.

  $$\begin{cases} H_0: \beta_j = 0 & j = 1, \dots, k \\ H_1: \exists\, \beta_j \neq 0 & j = 1, \dots, k \end{cases}$$

  → we permute the Y over the units →
  (if all the β_j are zero it will be the same if we permute)



- **linear models: t-tests**
  (linear regression and multi-way ANOVA)

  Test for the singular regressor / significance of one factor.

  → we permute the residuals under H_0

  → only family of permutation tests which are not exact (because we would like to permute the errors, but we have only the residuals)