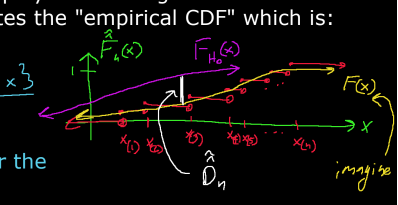
How about a general test for goodness of fit? For example, what if I have data x = <1.73, -0.49, 0.93, 2.16, 0.03> and I want to prove this is not realized from a specific DGP. For example, the DGP is not iid N(0, 1). The hypotheses then are:

$$H_0: X_1, \ldots, X_n \overset{iid}{\sim} N(0,1) \iff F(x) = \Phi(x) \text{ (std. normal)}$$
$$H_a: \text{not } H_0 \iff F(x) \neq \Phi(x)$$

For continuous rv X, we can employ the Kolmogorov-Smirnov test (KS test). This test first computes the "empirical CDF" which is:

$$\hat{F}_n(x) := \frac{\#\{X_i \leq x\}}{n}$$



F-hat is a function estimator for the true function F, the CDF.

We need a test statistic that gauges the difference between the empirical CDF and the CDF assumed by the null hypothesis. If that test statistic is large => reject the null.

$$\hat{D}_n\left(\hat{F}_n(x), F_{H_0}(x)\right) := \sup_x \left\{ \left| \hat{F}_n(x) - F_{H_0}(x) \right| \right\}$$

This is called the "supremum norm difference". You can think of it as maximum absolute distance.

Thm: $\hat{D}_n$ converges to 0 under the null hypothesis. (Glivenko-Cantelli, 1933)

This means the empirical CDF converges to the true CDF for all x. It also implies that it converges to a value > 0 if the null is not true. Thus power of this test should converge to 1 as n increases.

Kolmogorov then proved in 1933 that

$$\sqrt{n}\,\hat{D}_n \xrightarrow{d} K,$$ the "Kolmogorov distribution"

$$\Downarrow$$

$$\sqrt{n}\,\hat{D}_n \overset{\cdot}{\sim} K$$

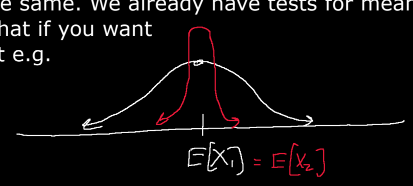
The 95% cutoff value is 1.36
The 99% cutoff value is 1.63

This is an amazing distribution-free result. This works for any F(x)!

Tables of critical values of K have been precomputed. But this distribution approximation is very crude and should only be trusted for n > 50. There are finite approximations but... we won't study them. They're likely distribution-dependent.

There is also a KS-test for discrete X which we won't study.

What if you have two samples. And you want to test if the DGP's are the same. We already have tests for means being the same. But what if you want to test e.g.



$$H_0: F_1(x) = F_2(x), \quad H_a: F_1(x) \neq F_2(x)$$

$$\hat{D}_{n_1,n_2} := \sup_x \left\{ \hat{F}_1(x) - \hat{F}_2(x) \right\}$$

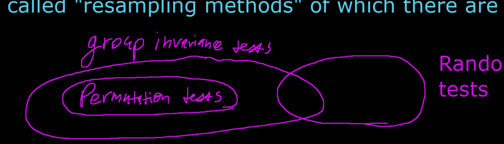
Kolmogorov also proved that:

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\hat{D}_{n_1,n_2} \xrightarrow{d} K$$ this yields the 2-sample KS test

The 2-sample Anderson-Darling (AD) test is very similar. The 2-sample Wilcoxon-Mann-Whitney U test does not have the restriction of continuous DGPs. We don't have time for these.

The KS, AD, U tests are all examples of "non-parametric tests" (so is the 2-sample Wald test for difference in means). "Non-parametric" means that we make no explicit assumptions about the functional forms of the DGPs that produce the samples' data. Parametric tests are more powerful (usually) because you have more information. But they come with assumptions. If the assumptions are unjustified, the tests may be invalid. These are also called "distribution-free" tests.

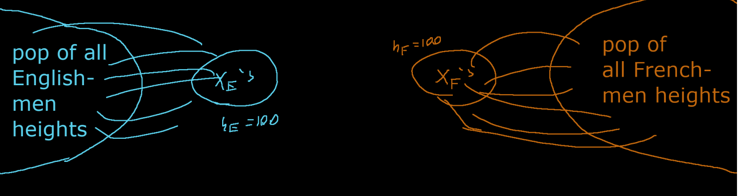
There is a totally different strategy to create nonparametric tests called "resampling methods" of which there are a few:



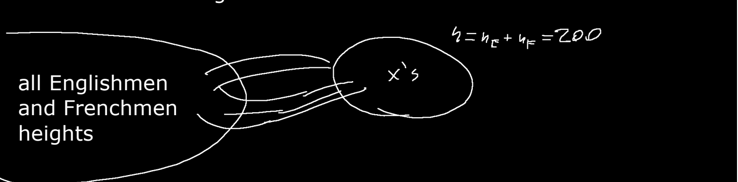
Let's assume we want to test the same null/alternative as the 2-sample KS test:

$$H_0: F_1(x) = F_2(x) \quad \text{vs.} \quad H_a: F_1(x) \neq F_2(x)$$

Fisher (1936) had the following thought experiment. Imagine n_E= 100 Englishmen and n_F= 100 Frenchmen and measure their heights: $x_{E1}, \ldots, x_{E100}, x_{F1}, \ldots, x_{F100}$



Under the null that Englishmen and Frenchmen heights are realized from the same DGP, we imagine just one giant population which includes Englishmen and Frenchmen



Imagine fake samples from the "giant population" that are arbitrarily divided into n_E Englishmen and n_F Frenchmen. These fake division are on arbitrary partitions of the original data.

Fake sample 1: randomly permute n_E + n_F = 200 observations and call the first n_E = 100 "Englishment heights" and the second n_F = 100 "Frenchmen heights"

$$I_{1,1} \subset \{1, 2, \ldots, n\}, \quad I_{2,1} \subset \{1, 2, \ldots, n\} = \{1, \ldots, n\} \setminus I_1$$
$$\text{s.t. } |I_{1,1}| = n_E, \ |I_{2,1}| = n_F, \ I_{1,1} \cup I_{2,1} = \{1, 2, \ldots, n\}$$

Doing this B many times will give you many permutation splits.

Now for each permutation split, compute a test statistic that gauges departure from the null (which is both CDF's are equal). Some common examples:

(a) $\bar{x}_{1,b} - \bar{x}_{2,b} = \frac{1}{n_E} \sum_{i \in I_{1,b}} x_i - \frac{1}{n_F} \sum_{i \in I_{2,b}} x_i$

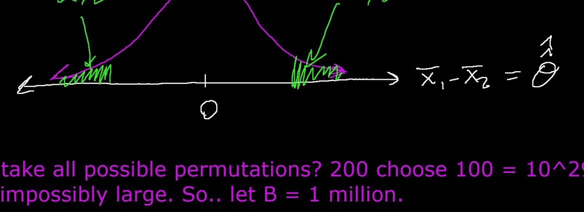
(b) $Med_{1,b} - Med_{2,b}$

(c) $\hat{D}_{n_1,n_2,b}$ from the KS test (the sup difference)

(d) $\frac{\bar{x}_{1,b}}{\bar{x}_{2,b}}$

There are also more. Each test statistic will yield a different permutation test with different power for different DGP's.

Let's consider (a), the difference in sample averages. What is the sampling distribution under the null hypothesis? Fisher says, just look at the test statistics over a large enough B.



Can you take all possible permutations? 200 choose 100 = 10^29 which is impossibly large. So.. let B = 1 million.

What is the RET region? Declare an alpha and put alpha/2 in each tail. For example, at alpha = 5%, you order all the sample averages. And then the lower RET cutoff is the 25,000th largest sample and the upper RET cutoff is the 975,000th largest sample:

$$RET = \left[\hat{\theta}_{(\frac{\alpha}{2}B)}, \ \hat{\theta}_{((1-\frac{\alpha}{2})B)}\right] \text{ in general}$$

Now calculate the true sample statistic and see if it falls in RET or not.

$$\hat{\theta} = \bar{x}_E - \bar{x}_F \overset{?}{\in} RET \quad \xrightarrow{\text{yes}} \text{Retain}, \quad \xrightarrow{\text{No}} \text{Reject}$$

You can also create a CI for the sample statistic using this permutation idea, but it is a bit more complicated so we won't cover it.

This permuation test is a "computational resampling" approach

need a computer   use dataset over and over

Here's one of the most famous computation resampling approaches: Efron's Boostrap (1979). Imagine you have a DGP f(x; θ_1, ..., θ_K) and you have an arbitrary function of the parameters you are interested in:

$$\phi = g(\theta_1, \ldots, \theta_K) \quad \text{estimated by} \quad \hat{\phi} = w(X_1, \ldots, X_n)$$

whose sampling distribution is totally unknown

For example, $\phi = Med[X]$

or $\phi := \frac{E[X] - r_{free}}{SD[X]}$ the "Sharpe ratio" used in finance

$$\hat{\phi}^{mm} = \frac{\bar{X} - r_{free}}{\hat{\sigma}} \sim ???$$

We need the distribution to do hypothesis testing and to generate confidence intervals.

The bootstrap solves this problem in most situations.