The Five Most Common Randomization Paradigms:

A) Correlation –

The population is a (bivariate) pair of random variables, (X, Y)

The data are n pairs of observations on the population, $(x_1, y_1), \ldots, (x_n, y_n)$

The null hypothesis is that the random variables X and Y are statistically independent, which implies that the x's can be fixed as is and all $n!$ orderings of the y's should be equally-likely

As an example, for usual numerical data we could obtain the correlation for each "way" (or, for an equivalent statistic such as the raw cross-product, $\sum x_i y_i$, that would give the same p-value)

We compare the observed correlation, r_{obs}, to the table to obtain the exact p-value, i.e., the probability of seeing a result as or more extreme than what was observed if the null hypothesis is true

This is called Pitman's test if the original data are used

If ranks are used, the Pearson correlation turns into the Spearman correlation, and the test is referred to as "Hotelling-Pabst"

This also justifies the test for "no association" based on Kendall's Tau and/or the Goodman-Kruskal Gamma coefficient (discussed elsewhere)

Fisher's Exact Test:

Consider a 2×2 contingency table for the observed data:

		Attribute 2	
		Present(1)	Absent(0)
Attribute 1	Present(1)	a	b
	Absent(0)	c	d

Again, the data are n pairs of observations on the population, $(x_1, y_1), \ldots, (x_n, y_n)$, where the x's are 0 or 1, and the y's are 0 or 1.

Choose as a test statistic, say, the value "a"

B) Two-dependent samples –

The population is a (bivariate) pair of dependent random variables, (X, Y)

The data are n pairs of observations on the population, $(x_1, y_1), \ldots, (x_n, y_n)$

The null hypothesis is that the random variables X and Y have the same distribution, which implies that all 2^n interchanges of the x's with the y's are equally-likely

Or if we consider the data to be the differences, $(x_1 - y_1), \ldots, (x_n - y_n)$, then all 2^n assignments of sign to the absolute values, $|(x_1 - y_1)|, \ldots, |(x_n - y_n)|$, are equally-likely

Obtain a statistic as the sum and table (or the sum of scores for the plus signs, say)

This is called Fisher's test if the original data are used

If we use ranks of the absolute values of the differences, this is called Wilcoxon's test

The idea of two-dependent sample also justifies the sign test and McNemar's test for correlated proportions

Suppose we let $d_i = |x_i - y_i|$ (or some function, $d_i = f(|x_i - y_i|)$

Let T = sum of d_i for the plus signs, and we reject if T is too extreme

Under the null hypothesis:

$E(T) = \frac{1}{2}\sum_{i=1}^{n} d_i$ ($= \frac{n(n+1)}{4}$ for untied ranks)

$V(T) = \frac{1}{4}\sum_{i=1}^{n} d_i^2$ ($= \frac{n(n+1)(2n+1)}{24}$ for untied ranks)

$\frac{T-E(T)}{\sqrt{V(T)}} \sim N(0,1)$

Wilcoxon Test: using ranks of the d_i

Sign Test: all $d_i = 1$; $E(T) = \frac{n}{2}$; $V(T) = \frac{n}{4}$

McNemar's test of correlated proportions:

		y_i	
		0	1
x_i	0	a	b
	1	c	d

We are interested in whether the proportion of 1's for the x's is the same as the proportions of 1's for the y's

$x_i - y_i \neq 0$ if $x_i = 1, y_i = 0$ or $x_i = 0, y_i = 1$

So, $T = c$ (which is the number of $x_i = 1, y_i = 0$)

$E(T) = \frac{1}{2} \sum_{i=1}^{n} d_i = \frac{b+c}{2}$

$V(T) = \frac{1}{4} \sum_{i=1}^{n} d_i^2 = \frac{b+c}{4}$

$\frac{c-((b+c)/2)}{\sqrt{(b+c)/4}} = \frac{c-b}{\sqrt{b+c}} \sim N(0,1)$

The Sign Test (and relatives):

A number of statistical procedures, primarily nonparametric ones, are based on the Binomial distribution.

The basic idea is to compare a particular Binomial distribution that is hypothesized theoretically (usually, a fair coin where $p = 1/2$) to the number of successes observed empirically. If the number of successes is too extreme, it casts doubt on the reasonableness of the hypothesized distribution.

This would be akin to replacing the absolute values, $|(x_1 - y_1)|, \ldots, |(x_n - y_n)|$ in the two-dependent sample context by 1's; then considering all 2^n assignments of sign to be equally-likely gets us to the Binomial with $p = 1/2$.

C) Two-independent samples –

The population is a pair of independent random variables, say X and X'

The data are n_1 observations on X (Group I): $x_1, \ldots, x_{n_1}$,

and $n - n_1 \equiv n_2$ observations on X' (Group II): $x_{n_1+1}, \ldots, x_n$

The null hypothesis is that the random variables X and X' have the same distribution, and implies that the $\binom{n}{n_1}$ ways of picking observations for Groups I and II are equally-likely

Use a statistic such as the mean difference and table

This is again called Fisher's test if the original data are used

In an experimental context we can justify the use of the usual two-independent sample t-test as an approximation to the randomization test (and for the two-dependent sample t-test as well)

Suppose we do an experiment with a performance measure and a drug/no drug condition (I and II)

We have $n = 50$ subjects, and randomly assign 25 to condition I and the remaining to II; the drugs are imposed (or not) and we now wish to assess the null hypothesis of "no difference" versus "some difference"

We choose some statistic, e.g., the difference between the two means: $M_I - M_{II}$, and if it is too large or small, reject the null hypothesis

Question: how to decide if $M_I - M_{II}$ is too extreme?

We could use the t-test but since we haven't sampled from any two populations, it is not clear whether this is justifiable

An alternative is based on randomization:

If the null hypothesis is true, and since the subjects were randomly assigned to begin with, the scores seen for Groups I and II should still look as if we took all the scores and merely assigned 25 to each group at random.

Given the pool of 50, there are $\binom{50}{25}$ ways of doing this; for each, get a mean difference and do a distribution. The p-value is the number of differences as extreme or more than my observed difference

The t-distribution is a good approximation to this, and thus is one justification for using the two-independent sample t-test.

Internal validity:

Is the difference you see due to the group structure unambiguously? If so, the experiment is said to be internally valid

Moreover, a causal implication can be made: the group structure "caused" the difference we see

If fact, random assignment is the only way to guarantee this

Otherwise, we slip into quasi-experimentation, threats to internal validity, and in general, "Campbell and Stanley"

We also have external (or ecological) validity: are the treatments similar to what occurs in the "real world"

If we let e_i be some score given to x_i, choose the test statistic T_I = sum of scores in group I

$E(T_I) = \frac{n_1}{n} \sum_{i=1}^{n} e_i$

$V(T_I) = \frac{n_1(n-n_1)}{n-1}(\frac{1}{n} \sum_{i=1}^{n} (e_i - \bar{e})^2)$

$\frac{T_I - E(T_I)}{\sqrt{V(T_I}} \sim N(0,1)$

Mann-Whitney Test:

Let $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_I$ (this is the number of observations in group II greater than in group I)

$U' = n_1 n_2 - U$ (this is the number of observations in group I greater than in group II)

$U + U' = n_1 n_2$ (and the probability of a randomly drawn observation from group II being greater than one randomly drawn one from group I is $\frac{U}{n_1 n_2}$)

Wald-Wolfowitz Runs Test:

In the two-independent sample context, we use a different test statistic

Arrange all the scores in a row ordered from smallest to largest, and attach the group designation (I and II) to each observation

If the two groups were different, we would expect higher scores in one group and lower scores in the other. In other words, we would expect few runs of all I's or II's if the groups were different

If R is the number of runs, we would reject the null hypothesis if we observe too few runs:

$$E(R) = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$V(R) = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

$$\frac{R-E(R)}{\sqrt{V(R)}} \sim N(0,1)$$

D) K-independent samples –

Extending the notation for two-independent samples, there are $\frac{n!}{n_1!...n_K!}$ equally likely ways the data could be used to form the K groups

Maybe use

$H = \frac{SSBetween}{MSTotal} \sim \chi^2_{K-1}$,

an asymptotic result that holds when the null hypothesis is true that all K random variables have the same distributions

When ranks are used, we have the Kruskal-Wallis analysis-of-variance by ranks

K-dependent samples –

Extending the structure for two-dependent sample, we have the data laid out as follows:

		1	2	$\cdots$	K
blocks	1				
	2				
	$\vdots$				
	n				

There are $(K!)^n$ equally-likely ways that data could be rearranged within blocks

Maybe use

$$\frac{SSTreatments}{(SSTreatments+SSInteraction)/(n(K-1)} \sim \chi^2_{K-1}$$

When ranks (within blocks) are used, we have (Milton) Friedman's test

If the data are 0 and 1, we have Cochran's Q test (this extends McNemar's test) – i.e., are the proportions of 1's the same over the K columns

If we are concerned with the degree to which the rankings are consistent within blocks, we can use Kendall's coefficient of concordance (discussed later)

General Randomization Paradigm for all Non-parametric Methods:

a) Under H_o, what is equally-likely?

b) Choose some statistic, and get its value for each "way", and table

c) compare the observed "way" to the table, and obtain a p-value

Mechanics:

1) Complete enumeration: some times you don't have to do it yourself and tables are available, particularly if you do (untied) ranks instead of the original data. Also, now (in SYSTAT, SPSS, etc) we have "exact" tests

2) Sample the distribution: get as close to the actual approximation as you want; a Monte Carlo p-value

3) Approximations by moments: usually this is the chi-square or the normal (asymptotically)

Remember, *equivalent statistics* are those that lead to exactly the same p values

Extensions to Other Randomization Paradigms:

To single-subject interventions (and designs), see

Edgington & Onghena, 2007, Chapter 11: "N-of-1 Designs" (Randomization Tests, 4th edition)

———————

To compare (in a correlational sense) data matrices such as

		Staff			
		1	2	···	K
Patients	1				
	2				
	⋮				
	n				

See: Hubert, Assignment Methods in Combinatorial Data Analysis (1987)

To use in fMRI analyses and solving the multiple comparison problem, see:

T. E. Nichols and A. P. Holmes

Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples

Human Brain Mapping, 2001, 15, 1–25.

Also, see the Matlab toobox SnPM (which goes with SPM – Statistical Parametric Mapping)