Statistics 240: Nonparametric and Robust Methods

Philip B. Stark
Department of Statistics
University of California, Berkeley
www.stat.berkeley.edu/∼stark

ROUGH DRAFT NOTES—WORK IN PROGRESS!

Last edited 10 November 2010

Course Website:

`http://statistics.berkeley.edu/~stark/Teach/S240/F10`

Example: Effect of treatment in a randomized controlled experiment

11 pairs of rats, each pair from the same litter.

Randomly—by coin tosses—put one of each pair into "enriched" environment; other sib gets "normal" environment.

After 65 days, measure cortical mass (mg).

treatment	689	656	668	660	679	663	664	647	694	633	653
control	657	623	652	654	658	646	600	640	605	635	642
difference	32	33	16	6	21	17	64	7	89	-2	11

How should we analyze the data?

(Cartoon of [**?**]. See also [**?**] and [**?**, pp. 498ff]. The experiment had 3 levels, not 2, and there were several trials.)

Informal Hypotheses

Null hypothesis: treatment has "no effect."

Alternative hypothesis: treatment increases cortical mass.

Suggests 1-sided test for an increase.

Test contenders

- 2-sample Student t-test:

$$\frac{\text{mean(treatment) - mean(control)}}{\text{pooled estimate of SD of difference of means}}$$

- 1-sample Student t-test on the differences:

$$\frac{\text{mean(differences)}}{\text{SD(differences)}/\sqrt{11}}$$

  Better, since littermates are presumably more homogeneous.

- Permutation test using t-statistic of differences: same statistic, different way to calculate P-value. Even better?

Strong null hypothesis

Treatment has no effect whatsoever—as if cortical mass were assigned to each rat before the randomization.

Then equally likely that the rat with the heavier cortex will be assigned to treatment or to control, independently across littermate pairs.

Gives $2^{11} = 2,048$ equally likely possibilities:

difference	± 32	± 33	± 16	± 6	± 21	± 17	± 64	± 7	± 89	± 2	± 11

For example, just as likely to observe original differences as

difference	-32	-33	-16	-6	-21	-17	-64	-7	-89	-2	-11

Weak null hypothesis

On average across pairs, treatment makes no difference.

Alternatives

Individual's response depends only on that individual's assignment

Special cases: shift, scale, etc.

Interactions/Interference: my response could depend on whether you are assigned to treatment or control.

Assumptions of the tests

- 2-sample t-test: masses are iid sample from normal distribution, same unknown variance, same unknown mean. Tests weak null hypothesis (plus normality, independence, non-interference, etc.).

- 1-sample t-test on the differences: mass differences are iid sample from normal distribution, unknown variance, zero mean. Tests weak null hypothesis (plus normality, independence, non-interference, etc.)

- Permutation test: Randomization fair, independent across pairs. Tests strong null hypothesis.

Assumptions of the permutation test are true by design: That's how treatment was assigned.

Student t-test calculations

Mean of differences: 26.73mg
Sample SD of differences: 27.33mg
t-statistic: $26.73/(27.33/\sqrt{11}) = 3.244$.

P-value for 1-sided t-test: 0.0044

Why do cortical weights have normal distribution?

Why is variance of the difference between treatment and control the same for different litters?

Treatment and control are *dependent* because assigning a rat to treatment excludes it from the control group, and vice versa.

Does P-value depend on assuming differences are iid sample from a normal distribution? If we reject the null, is that because there is a treatment effect, or because the other assumptions are wrong?

Permutation t-test calculations

Could enumerate all $2^{11} = 2,048$ equally likely possibilities.
Calculate t-statistic for each.

P-value is

$$P = \frac{\text{number of possibilities with } t \geq 3.244}{2{,}048}$$

(For mean instead of t, would be $2/2,048 = 0.00098$.)

For more pairs, impractical to enumerate, but can simulate:

Assign a random sign to each difference.
Compute t-statistic
Repeat 100,000 times

$$P \approx \frac{\text{number of simulations with } t \geq 3.244}{100{,}000}$$

Calculations

```
simPermuTP <- function(z, iter) {
        # P.B. Stark, www.stat.berkeley.edu/~stark 5/14/07
        # simulated P-value for 1-sided 1-sample t-test under the
        # randomization model.
            n  <- length(z)
            ts <- mean(z)/(sd(z)/sqrt(n))      # t test statistic
            sum(replicate(iter, {zp <- z*(2*floor(runif(n)+0.5)-1);
                                 tst <- mean(zp)/(sd(zp)/sqrt(n));
                                 (tst >= ts)
                                }
                          )
               )/iter
}
simPermuTP(diffr, 100000)
0.0011
```

(versus 0.0044 for Student's t distribution)

Other tests: sign test, Wilcoxon signed-rank test

Sign test: Count pairs where treated rat has heavier cortex, i.e., where difference is positive.

Under strong null, distribution of the number of positive differences is Binomial(11, 1/2). Like number of heads in 11 independent tosses of a fair coin. (Assumes no ties w/i pairs.)

P-value is chance of 10 or more heads in 11 tosses of a fair coin: 0.0059.

Only uses signs of differences, not information that only the smallest absolute difference was negative.

Wilcoxon signed-rank test uses information about the ordering of the differences: rank the absolute values of the differences, then give them the observed signs and sum them. Null distribution: assign signs at random and sum.

Still more tests, for other alternatives

All the tests we've seen here are sensitive to *shifts*–the alternative hypothesis is that treatment increases response (cortical mass).

There are also nonparametric tests that are sensitive to other treatment effects, e.g., treatment increases the variability of the response.

And there are tests for whether treatment has any effect at all on the distribution of the responses.

You can design a test statistic to be sensitive to any change that interests you, then use the permutation distribution to get a P-value (and simulation to approximate that P-value).

Silliness

Treat ordinal data (e.g., Likert scale) as if measured on a linear scale; use Student t-test.

Maybe not so silly for large samples...

t-test asymptotically distribution-free.

How big is big?

Back to Rosenzweig et al.

Actually had 3 treatments: enriched, standard, deprived.

Randomized 3 rats per litter into the 3 treatments, independently across n litters.

How should we analyze these data?

Test contenders

n litters, s treatments (sibs per litter).

- ANOVA–the F-test:

$$F = \frac{\text{BSS}/(s-1)}{\text{WSS}/(n-s)}$$

- Permutation F-test: use permutation distribution instead of F distribution to get P-value.

- Friedman test: Rank within litters. Mean rank for treatment i is $\bar{R}_i$.

$$Q = \frac{12n}{s(s+1)} \sum_{i=1}^{s} \left(\bar{R}_i - \frac{s+1}{2} \right)^2 .$$

  P-value from permutation distribution.

Strong null hypothesis

Treatment has no effect whatsoever—as if cortical mass were assigned to each rat before the randomization.

Then equally likely that each littermate is assigned to each treatment, independently across litters.

There are $3! = 6$ assignments of each triple to treatments.

Thus, 6^n equally likely assignments across all litters.

For 11 litters, that's 362,797,056 possibilities.

Weak null hypothesis

The average cortical weight for all three treatment groups are equal. On average across triples, treatment makes no difference.

Assumptions of the tests

- F-test: masses are iid sample from normal distribution, same unknown variance, same unknown mean for all litters and treatments. Tests weak null hypothesis.

- Permutation F-test: Randomization was as advertised: fair, independent across triples. Tests strong null hypothesis.

- Friedman test: Ditto.

Assumptions of the permutation test and Friedman test are true by design: that's how treatment was assigned.

Friedman test statistic has χ^2 distribution asymptotically. Ties are a complication.

F-test assumptions–reasonable?

Why do cortical weights have normal distribution for each litter and for each treatment?

Why is the variance of cortical weights the same for different litters?

Why is the variance of cortical weights the same for different treatments?

Is F a good statistic for this alternative?

F (and Friedman statistic) sensitive to differences among the mean responses for each treatment group, no matter what pattern the differences have.

But the treatments and the responses can be ordered: we hypothesize that more stimulation produces greater cortical mass.

deprived	$\Longrightarrow$	normal	$\Longrightarrow$	enriched
low mass	$\Longrightarrow$	medium mass	$\Longrightarrow$	high mass

Can we use that to make a more sensitive test?

A test against an ordered alternative

Within each litter triple, count pairs of responses that are "in order." Sum across litters.

E.g., if one triple had cortical masses

deprived	640
normal	660
enriched	650

that would contribute 2 to the sum: $660 \geq 640$, $650 \geq 640$, but $640 < 650$.

Each litter triple contributes between 0 and 3 to the overall sum.

Null distribution for the test based on the permutation distribution: 6 equally likely assignments per litter, independent across litters.

A different test against an ordered alternative

Within each litter triple, add differences that are "in order." Sum across litters.

E.g., if one triple had cortical masses

deprived	640
normal	660
enriched	650

that would contribute 30 to the sum: $660 - 640 = 20$ and $650 - 640 = 10$, but $640 < 650$, so that pair contributes zero.

Each litter triple contributes between 0 and 2 $\times$ range to the sum.

Null distribution for the test based on the permutation distribution: 6 equally likely assignments per litter, independent across litters.

Quick overview of nonparametrics, robustness

Parameters: related notions

- Constants that index a family of functions–e.g., the normal curve depends on μ and σ ($f(x) = (2\pi)^{1/2}\sigma^{-1}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$)

- A property of a probability distribution, e.g., 2nd moment, a percentile, etc.

Parametric statistics: assume a functional form for the probability distribution of the observations; worry perhaps about some parameters in that function.

Non-parametric statistics: fewer, weaker assumptions about the probability distribution. E.g., randomization model, or observations are iid.

Density estimation, nonparametric regression: Infinitely many parameters. Requires regularity assumptions to make inferences. Plus iid or something like it.

Semiparametrics: Underlying functional form unknown, but relationship between different groups is parametric. E.g., Cox proportional hazards model.

Robust statistics: assume a functional form for the probability distribution, but worry about whether the procedure is sensitive to "small" departures from that assumed form.

Groups A *group* is an ordered pair $(\mathcal{G}, \times)$, where $\mathcal{G}$ is a collection of objects (the elements of the group) and $\times$ is a mapping from $\mathcal{G} \bigotimes \mathcal{G}$ onto $\mathcal{G}$,

$$\begin{aligned} \times : \ \mathcal{G} \bigotimes \mathcal{G} \ &\to \mathcal{G} \\ (a, b) \ &\mapsto a \times b, \end{aligned}$$

satisfying the following axioms:

1. $\exists e \in \mathcal{G}$ s.t. $\forall a \in \mathcal{G}$, $e \times a = a$. The element e is called the *identity*.

2. For each $a \in \mathcal{G}$, $\exists a^{-1} \in \mathcal{G}$ s.t. $a^{-1} \times a = e$. (Every element has an inverse.)

3. If $a, b, c \in \mathcal{G}$, then $a \times (b \times c) = (a \times b) \times c$. (The group operation is associative.)

Abelian groups If, in addition, for every $a, b \in \mathcal{G}$, $a \times b = b \times a$ (if the group operation commutes), we say that $(\mathcal{G}, \times)$ is an *Abelian group* or *commutative group*.

The permutation group Consider a collection of n objects, numbered 1 to n. A *permutation* is an ordering of the objects. We can represent the permutation as a vector. The kth component of the vector is the number of the object that is kth in the ordering.

For instance, if we have 5 objects, the permutation

$$(1, 2, 3, 4, 5) \tag{1}$$

represents the objects in their numbered order, while

$$(1, 3, 4, 5, 2) \tag{2}$$

is the permutation that has item 1 first, item 3 second, item 4 third, item 5 fourth, and item 2 fifth.

Permutations as matrices. Associativity follows from associativity of matrix multiplication. [FIX ME!]

The permutation group is not Abelian

For instance, consider the permutation group on 3 objects. Let $\pi_1 \equiv (2,1,3)$ and $\pi_2 \equiv (1,3,2)$.

Then $\pi_1\pi_2(1,2,3) = (3,1,2)$, but $\pi_2\pi_1(1,2,3) = (2,3,1)$.

Simulation: pseudo-random number generation

Most computers cannot generate truly random numbers, although there is special equipment that can (usually, these rely on a physical source of "noise," such as a resistor or a radiation detector). Most so-called random numbers generated by computers are really "pseudo-random" numbers, sequences generated by a software algorithm called a pseudo-random number generator (PRNG) from a starting point, called a *seed*. Pseudo-random numbers behave much like random numbers for many purposes.

The seed of a pseudo-random number generator can be thought of as the initial state of the algorithm. Each time the algorithm produces a number, it alters its state—deterministically. If you start a given algorithm from the same seed, you will get the same sequence of pseudo-random numbers.

Each pseudo-random number generator has only finitely many states. Eventually—after the *period* of the generator, the generator gets back to its initial state and the sequence repeats. If the state of the PRNG is n bits long, the period of the PRNG is at most 2^n bits—but can be substantially shorter, depending on the algorithm.

Better generators have more states and longer periods, but that comes at a price: speed. There is a tradeoff between the computational efficiency of a pseudo-random number generator and the difficulty of telling that its output is not really random (measured, for example, by the number of bits one must examine).

Evaluating PRNGs

See `http://csrc.nist.gov/rng/` for a suite of tests of pseudo-random number generators. Tests can be based on statistics such as the number of zero and one bits in a block or sequence, the number of runs in sequences of differing lengths, the length of the longest run, spectral properties, compressibility (the less random a sequence is, the easier it is to compress), and so on.

You should check which PRNG is used by any software package you rely on for simulations. Linear Congruential Generators are of the form

$$x_i = ((ax_{i-1} + b) \mod m)/(m-1).$$

They used to be popular but are best avoided. (They tend to have a short period, and the sequences have underlying regularity that can spoil performance for many purposes. For instance, if the LCG is used to generate n-dimensional points, those points lie on at most $m^{1/n}$ hyperplanes in $\mathbb{R}^n$.

For statistical simulations, a particularly good, efficient pseudo-random number generator is the Mersenne Twister. The state of the Mersenne Twister is a 624-vector of 32-bit integers and a pointer to one of those vectors. It has a period of $2^{19937} - 1$, which is on the order of 10^{6001}. It is implemented in R (it's the default), Python, Perl, and many other languages. For cryptography, a higher level of randomness is needed than for most statistical simulations.

No pseudo-random number generator is best for all purposes. But some are truly terrible.

For instance, the PRNG in Microsoft Excel is a faulty implementation of an algorithm (the Wichmann-Hill algorithm, which combines four LCGs) that isn't good in the first place.

McCullough, B.D., Heiser, David A., 2008. On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics and Data Analysis 52* (10), 4570–4578.

http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6V8V-4S1S6F
cdi=5880&_user=4420&_orig=mlkt&_coverDate=06%2F15%2F2008&_sk=
999479989&view=c&wchp=dGLbVzb-zSkWb&md5=85d93a6c0700f2dbc483f5ed6b23
/sdarticle.pdf

Excerpt: Excel 2007, like its predecessors, fails a standard set of intermediate-level accuracy tests in three areas: statistical distributions, random number generation, and estimation. Additional errors in specific Excel procedures are discussed. Microsoft's continuing inability to correctly fix errors is discussed. No statistical procedure in Excel should be used until Microsoft documents that the procedure is correct; it is not safe to assume that Microsoft Excel's statistical procedures

give the correct answer. Persons who wish to conduct statistical analyses should use some other package.

If users could set the seeds, it would be an easy matter to compute successive values of the WH RNG and thus ascertain whether Excel is correctly generating WH RNGs. We pointedly note that Microsoft programmers obviously have the ability to set the seeds and to verify the output from the RNG; for some reason they did not do so. Given Microsoft's previous failure to implement correctly the WH RNG, that the Microsoft programmers did not take this easy and obvious opportunity to check their code for the patch is absolutely astounding.

McCullough, B.D., 2008. Microsoft's 'Not the Wichmann-Hill' random number generator. *Computational Statistics and Data Analysis 52* (10), 4587–4593.

http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6V8V-4S21TG
cdi=5880&_user=4420&_orig=search&_coverDate=06%2F15%2F2008&_
sk=999479989&view=c&wchp=dGLbVtz-zSkzk&md5=38238ccd25a60a408480df345
/sdarticle.pdf

Drawing (pseudo-)random samples using PRNGs

A standard technique for drawing a pseudo-random sample of size n from N item is to assign each of the N items a pseudo-random number, then take the sample to be the n items that were assigned the n smallest pseudo-random numbers.

Note that when N is large and n is a moderate fraction of N, PRNGs might not be able to generate all $\binom{N}{n}$ subsets.

Henceforth, will assume that the PRNG is "good enough" that its departure from randomness does not affect the accuracy our simulations enough to matter.

Bernoulli trials

A *Bernoulli trial* is a random experiment with two possible outcomes, success and failure. The probability of success is p; the probability of failure is $1 - p$.

Events A and B are *independent* if $P(AB) = P(A)P(B)$.

A collection of events is independent if the probability of the intersection of every subcollection is equal to the product of the probabilities of the members of that subcollection.

Two random variables are independent if every event determined by the first random variable is independent of every event determined by the second.

Binomial distribution

Consider a sequence of n independent Bernoulli trials with the same probability p of success in each trial. Let X be the total number of successes in the n trials.

Then X has a binomial probability distribution:

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{3}$$

Hypergeometric distribution

A *simple random sample of size* n from a finite population of N things is a random sample drawn without replacement in such a way that each of the $\binom{N}{n}$ subsets of size n from the population is equally likely to be the sample.

Consider drawing a simple random sample from a population of N objects of which G are good and $N-G$ are bad. Let X be the number of good objects in the sample.

Then X has a hypergeometric distribution:

$$P(X = x) = \frac{\binom{G}{x}\binom{N-G}{n-x}}{\binom{N}{n}}, \tag{4}$$

for $\max(0, n-(N-G)) \leq x \leq \min(n, G)$.

Hypothesis testing

Much of this course concerns hypothesis tests. We will think of a test as consisting of a set of possible outcomes (data), called the *acceptance region*. The complement of the acceptance region is the *rejection region*.

We reject the null hypothesis if the data are in the rejection region.

The *significance level* α is an upper bound on the chance that the outcome will turn out to be in the rejection region if the null hypothesis is true.

Conditional tests

The chance is sometimes a conditional probability rather than an unconditional probability. That is, we have a rule that generates an acceptance region that depends on some aspect of the data. We've already seen an example of that in the rat cortical mass experiment. There, we conditioned on the cortical masses, but not on the the randomization.

If we test to obtain conditional significance level α (or smaller) no matter what the data are, then the unconditional significance level is still α:

$$\begin{aligned} \Pr\{\text{Type I error}\} &= \int_x \Pr\{\text{Type I error}|X=x\}\mu(dx) \\ &\leq \sup_x \Pr\{\text{Type I error}|X=x\}. \end{aligned}$$

P-values

Suppose we have a family of hypothesis tests for testing a given null hypothesis at every significance level $\alpha \in (0, 1)$. Let A_α denote the acceptance region for the test at level α.

Suppose further that the tests *nest*, in the sense that if $\alpha_1 < \alpha_2$, then $A_{\alpha_1} \subset A_{\alpha_2}$

Then the P-value of the hypothesis (for data X) is

$$\inf\{\alpha : X \notin A_\alpha\} \tag{5}$$

Confidence sets

We have a collection of hypotheses $\mathcal{H}$. We know that some $H \in \mathcal{H}$ must be true—but we don't know which one.

A rule that uses the data to select a subset of $\mathcal{H}$ is a $1 - \alpha$ *confidence procedure* if the chance that it selects a subset that includes H is at least $1 - \alpha$.

The subset that the rule selects is called a $1 - \alpha$ *confidence set*.

The *coverage probability at* G is the chance that the rule selects a set that includes G if $G \in \mathcal{H}$ is the true hypothesis.

Duality between tests and confidence sets

Suppose that some hypothesis $H \in \mathcal{H}$ must be true. Suppose we have a family of significance-level α tests $\{A_G : G \in \mathcal{H}\}$ such that for each $G \in \mathcal{H}$,

$$\Pr_G\{X \notin A_G\} \leq \alpha. \tag{6}$$

Then the set

$$C(X) \equiv \{G \in \mathcal{H} : A_G \ni X\} \tag{7}$$

is a $1 - \alpha$ confidence set for the true H.

Tests and confidence sets for Bernoulli p

We have $X_j \sim$ Bernoulli(p). Can draw as big an iid sample $\{X_j\}_{j=1}^n$ as we like.

We want to test the hypothesis that $p \leq p_0$ at level α and we want to find a 1-sided upper confidence interval for p.

Or might want 2-sided confidence interval, or to test the hypothesis $p > p_0$, or a 1-sided lower confidence interval.

Tests for Bernoulli p: fixed n

Test hypothesis $p \geq p_0$ at level α based on number X of successes in n independent trials, n fixed. Then $X \sim$ Binomial(n, p) with n known and p not.

Reject when $X = x$ if

$$\alpha \geq \Pr\{X \leq x || p = p_0\} = \sum_{t=0}^{x} \binom{n}{t} p_0^t (1 - p_0)^{n-t}. \qquad (8)$$

Here, the notation $||$ means "computed on the assumption that." It's common to use a single vertical bar for this purpose, but single bars also denote conditioning; here, we have an assumption, not a conditional probability.

Upper confidence bound:

$$p_\alpha^+ = \max\{\pi : \sum_{t=0}^{x} \binom{n}{t} \pi^t (1 - \pi)^{n-t} > \alpha\}. \qquad (9)$$

Upper confidence bound for binomial p

```
binoUpperCL <- function(n, x, cl = 0.975, inc=0.000001, p=x/n) {
    if (x < n) {
            f <- pbinom(x, n, p, lower.tail = TRUE);
            while (f >= 1-cl) {
                p <- p + inc;
                f <- pbinom(x, n, p, lower.tail = TRUE)
            }
            p
    } else {
            1.0
    }
}
```

Lower confidence bound for binomial p

```
binoLowerCL <- function(n, x, cl = 0.975, inc=0.000001, p=x/n) {
    if (x > 0) {
            f <- pbinom(x-1, n, p, lower.tail = FALSE);
            while (f >= 1-cl) {
                p <- p - inc;
                f <- pbinom(x-1, n, p, lower.tail = FALSE)
            }
            p
    } else {
            0.0
    }
}
```

Lower confidence bound for "good" items from SRS

```
hyperLowerCL <- function(N, n, x, cl = 0.975, p=ceiling(N*x/n)) {
    if (x < n) {
            f <- phyper(x-1, p, N-p, n, lower.tail = FALSE);
            while (f >= 1-cl) {
                p <- p - 1;
                f <- phyper(x-1, p, N-p, n, lower.tail = FALSE);
            }
            p
    } else {
            0.0
    }
}
```

Upper confidence bound for "good" items from SRS

```
hyperUpperCL <- function(N, n, x, cl = 0.975, p=floor(N*x/n)) {
    if (x < n) {
            f <- phyper(x, p, N-p, n, lower.tail = TRUE);
            while (f >= 1-cl) {
                p <- p + 1;
                f <- phyper(x, p, N-p, n, lower.tail = TRUE);
            }
            p
    } else {
            1.0
    }
}
```

Sequential test for p

If generating each X_j is expensive (e.g., if it involves running a climate model on supercomputer clusters for months), might want to minimize the sample size. Sequential testing: draw until you have strong evidence that $p \leq p_0$ (or that $p > p_0$).

Null: $p > p_0$. Control the chance of type I error.

Two common criteria: expected sample size at fixed p and maximum expected sample size.

α: maximum chance of rejecting null when $p > p_0$.

β: maximum chance of not rejecting null when $p \leq p_1 < p_0$.

Let $T_m \equiv \sum_{j=1}^{m} X_j$ and

$$\frac{p_{1m}}{p_{0m}} \equiv \frac{p_1^{T_m}(1-p_1)^{m-T_m}}{p_0^{T_m}(1-p_0)^{m-T_m}}. \tag{10}$$

Ratio of binomial probability when $p = p_1$ to binomial probability when $p = p_0$ (binomial coefficients in the numerator and denominator cancel).

Wald's sequential probability ratio test (SPRT) for p

Conclude $p \leq p_0$ if

$$\frac{p_{1m}}{p_{0m}} \geq \frac{1-\beta}{\alpha}. \tag{11}$$

Conclude $p > p_0$ if

$$\frac{p_{1m}}{p_{0m}} \leq \frac{\beta}{1-\alpha}. \tag{12}$$

Otherwise, draw again.

The SPRT approximately minimizes the expected sample size when the true p is p_0 or $p_1 < p_0$. For values in (p_1, p_0), it can have larger sample sizes than fixed tests.

SPRT miracle

Don't need to know the distribution of the test statistic under the null hypothesis to find the critical values for the test.

Derivation of Wald's SPRT

Testing between two hypotheses, H_0 and H_1, on the basis of data $\{X_j\} \subset \mathcal{X}$, with $\mathcal{X}$ a measurable space. According to both hypotheses, the $\{X_j\}$ are iid. Each hypothesis specifies a probability distribution for the data.

Suppose those two distributions are absolutely continuous with respect to some dominating measure μ on $\mathcal{X}$. Let f_0 be the density (wrt μ) of the distribution of X_j if H_0 is true and let f_1 be the density (wrt μ) of the distribution of X_j if H_1 is true.

Neyman-Pearson Lemma

For testing H_0 against H_1 based on $\{X_j\}_{j=1}^n$, most powerful level-α test is of the form

$$\text{Reject if } \frac{\Pi_{j=1}^n f_1(X_j)}{\Pi_{j=1}^n f_0(X_j)} \geq t_\alpha, \tag{13}$$

with t_α chosen so that the test has level α; that is, to be the smallest value of t for which

$$\Pr\left(\frac{\Pi_{j=1}^n f_1(X_j)}{\Pi_{i=1}^n f_0(X_j)} \geq t || \{X_j\} \text{ iid } f_0\right) \leq \alpha. \tag{14}$$

(Randomization might be necessary to attain level α exactly.)

Derivation of SPRT (contd)

At stage m, divide outcome space into 3 disjoint regions, A_{0m}, A_{1m} and A_m.

Draw X_1. If $X_1 \in A_{01}$, accept H_0 and stop. If $X_1 \in A_{11}$, accept H_1 and stop. If $X_1 \in A_1$, draw X_2.

If you draw X_m, then: If $X_m \in A_{0m}$, accept H_0 and stop. If $X_m \in A_{1m}$, accept H_1 and stop. If $X_m \in A_m$, draw X_{m+1}.

Derivation (contd)

A fixed-n test is a special case: A_{0m} and A_{1m} are empty for $m < n$ (so A_m is the entire outcome space when $m < n$).

A_{0n} is the acceptance region of the test, and A_{1n} is the complement of A_{0n} (so A_m is empty when $m = n$).

Derivation (contd)

Suppose sequential procedure stops after drawing X_N. N is random.

Suppose S is a particular sequential test procedure. Let $\mathbb{E}(N||h) = \mathbb{E}(N||h, S)$ be the expected value of N if H_h is true, $h \in \{0, 1\}$, for test S.

Tests S and S' have the same strength if they have the same chances of type I errors and of type II errors (α and β).

If S and S' have the same strength, S is better than S' if $\mathbb{E}(N||h, S) \leq \mathbb{E}(N||h, S')$ for $h = 1, 2$, with strict inequality for either $h = 1$ or $h = 2$ (or both).

Admissible, best, and optimal sequential tests

A sequential test is *admissible* if there is no better test of the same strength.

A test of a given strength is *best* if, among all tests of that strength, it has the smallest values of both $\mathbb{E}(N||0)$ and $\mathbb{E}(N||1)$. (This is the analog of a most powerful test in the fixed-n setting.)

A test S^* is *optimal* if it is admissible and

$$\max_h \mathbb{E}(N||h, S^*) \leq \max_h \mathbb{E}(N||h, S) \tag{15}$$

for all admissible tests S with the same strength as S^*.

The *efficiency* of a sequential test S is

$$\frac{\max_h \mathbb{E}(N||h, S^*)}{\max_h \mathbb{E}(N||h, S)} \tag{16}$$

where S^* is an optimal test with the same strength as S.

Bayes decision

Suppose we had a prior on $\{H_0, H_1\}$:

$$\Pr\{H_0 \text{ is true }\} = \pi_0, \quad \Pr\{H_1 \text{ is true }\} = \pi_1 = 1 - \pi_0. \quad (17)$$

Let

$$p_{hm} \equiv \Pi_{j=1}^{m} f_h(X_j), \quad h \in \{0, 1\}. \quad (18)$$

After making m draws, the posterior probability of H_h given the data $\{X_j\}$ is

$$\pi_{hm} = \frac{\pi_h p_{hm}}{\pi_0 p_{0m} + \pi_1 p_{1m}}. \quad (19)$$

Bayes decision, contd

Let $\lambda_h \in (1/2, 1)$ be the desired posterior probability of accepting hypothesis H_h when it is true, $h \in \{0, 1\}$.

Then we could test by accepting H_h at stage m (and stopping) if $\pi_{hm} \geq \lambda_h$ at stage m, and drawing again if $\pi_{0m} < \lambda_0$ and $\pi_{1m} < \lambda_1$.

Implicitly defines

$$A_{hm} = \{x : \pi_{hm} \geq \lambda_h \text{ if } X_j = x_j, j = 1, \ldots, n\}. \qquad (20)$$

Need A_{0m}, A_{1m} to be disjoint. Suppose not: $\pi_{0m} \geq \lambda_0$ and $\pi_{1m} \geq \lambda_1$. Then $1 = \pi_{0m} + \pi_{1m} \geq \lambda_0 + \lambda_1 > 1$, a contradiction.

Bayes decision, contd

Re-write stopping rule:

Accept H_0 at stage m if

$$\frac{p_{1m}}{p_{0m}} \leq \frac{\pi_0}{\pi_1} \cdot \frac{1-\lambda_0}{\lambda_0}; \tag{21}$$

accept H_1 at stage m if

$$\frac{p_{1m}}{p_{0m}} \geq \frac{\pi_0}{\pi_1} \cdot \frac{\lambda_1}{1-\lambda_1}. \tag{22}$$

Right hand sides don't depend on m.

Even if $\{\pi_h\}$ do not exist, makes sense to use the rule

- accept H_0 if $\frac{p_{1m}}{p_{0m}} \leq a$
- accept H_1 if $\frac{p_{1m}}{p_{0m}} \geq b$
- draw again if $a < \frac{p_{1m}}{p_{0m}} < b$

for some $a < b$. This is the SPRT.

Cylindrical points

Have a potentially infinite sequence of observations, $(X_j)_{j=1}^{\infty}$. Each possible sequence is an element of $\mathbb{R}^{\infty}$.

Suppose we have a finite sequence $(x_j)_{j=1}^{m}$.

The *cylindrical point* defined by $(x_j)_{j=1}^{m}$ is

$$C((x_j)_{j=1}^{m}) \equiv \{y \in \mathbb{R}^{\infty} : y_j = x_j, j = 1, \ldots, m\}. \qquad (23)$$

Suppose $S \subset \mathbb{R}^{\infty}$. If there is some $(x_j)_{j=1}^{m}$ for which $S = C((x_j)_{j=1}^{m})$, then S is a *cylindrical point of order* m.

The cylindrical point $C((x_j)_{j=1}^m)$ is of type 0 iff

$$\frac{p_{1m}}{p_{0m}} \leq a \tag{24}$$

and

$$a < \frac{p_{1k}}{p_{0k}} < b, \quad k = 1, \ldots, m-1. \tag{25}$$

The cylindrical point $C((x_j)_{j=1}^m)$ is of type 1 iff

$$\frac{p_{1m}}{p_{0m}} \geq b \tag{26}$$

and

$$a < \frac{p_{1k}}{p_{0k}} < b, \quad k = 1, \ldots, m-1. \tag{27}$$

Let $\mathcal{C}_h$ be the union of all cylindrical points of type h, $h = 0, 1$. Then

$$\Pr\{\mathcal{C}_0 \cup \mathcal{C}_1 || H_h\} = 1, \quad h = 0, 1. \tag{28}$$

(Requires work to show–this is where iid assumption is used.) This means that the procedure terminates with probability 1, if H_0 is true or if H_1 is true.

For every element of $\mathcal{C}_1$, $\frac{p_{1m}}{p_{0m}} \geq b$, so

$$\Pr\{\mathcal{C}_1 || H_1\} \geq b \Pr\{\mathcal{C}_1 || H_0\} = b\alpha. \tag{29}$$

Similarly,

$$\beta = \Pr\{\mathcal{C}_0 || H_1\} \leq a \Pr\{\mathcal{C}_0 || H_0\}. \tag{30}$$

We also have

$$\Pr\{\mathcal{C}_0||H_0\} = 1-\alpha \text{ and } \Pr\{\mathcal{C}_1||H_1\} = 1-\beta. \tag{31}$$

Hence

$$1-\beta \geq b\alpha \tag{32}$$

and

$$\beta \leq a(1-\alpha). \tag{33}$$

Rearranging yields

$$\frac{\alpha}{1-\beta} \leq \frac{1}{b} \tag{34}$$

and

$$\frac{\beta}{1-\alpha} \leq a. \tag{35}$$

Notes:

- iid assumption can be weakened substantially. Only used to prove that $\Pr\{\mathcal{C}_0 \cup \mathcal{C}_1 || H_h\} = 1$.

- only need to know the likelihood function under the two hypotheses, α, and β

- can sharpen the choice of thresholds, but not by much if α and β are small.

References on sequential tests for Bernoulli p

[**?**, **?**, **?**]

References on sequential tests forMonte Carlo p

[**?**, **?**, **?**]

References on 2-SPRT, etc. [FIX ME!]

Assignment: comparing SPRT and fixed sample-size tests

Implement the SPRT in R from scratch.

Taking $p_0 = 0.05$, $p_1 = 0.045$, $\alpha = 0.001$, $\beta = 0.01$, estimate (by simulation) the expected number of samples that must be drawn when $p = 0.01, 0.02, \ldots, 0.1$.

Compare the expected sample sizes with sample sizes for a fixed-n test with the same α and β.

Justify your choice of the number of replications to use in your simulations.

For each of the 10 scenarios, report the empirical fraction of type I and type II errors and 99% confidence intervals for the probabilities of those errors.

Assignment hint

Order of operations matters for accuracy and stability.

Raising a small number to a high power will eventually give you zero (to machine precision). If you calculate the numerator and the denominator in SPRT separately, you will eventually get nonsense as m gets large.

Rather than take a ratio of large products that are each going to zero, it's much more stable to take a product of ratios that are all close to 1.

Assignment hint, contd.

So, compute

$$(p_1/p_0)^{T_m}((1-p_1)/(1-p_0))^{m-T_m} \tag{36}$$

rather than

$$[p_1^{T_m}(1-p_1)^{m-T_m}]/[p_0^{T_m}(1-p_0)^{m-T_m}]. \tag{37}$$

If you want to work with the log SPR, compute

$$T_m \log((p_1/p_0) + (m-T_m)\log((1-p_1)/(1-p_0)) \tag{38}$$

rather than

$$T_m \log p_1 + (m-T_m)\log(1-p_1) - T_m \log(p_0) - (m-T_m)\log(1-p_0). \tag{39}$$

Fisher's exact test, Fisher's "Lady Tasting Tea" experiment

Under what groups is the distribution of the the data invariant in those problems?

`http://statistics.berkeley.edu/~stark/SticiGui/Text/percentageTests.htm#fisher_dependent`

`http://statistics.berkeley.edu/~stark/Teach/S240/Notes/ch3.htm`

General test based on group invariance

Follow Romano (1990).

Data $X \sim P$ takes values in $\mathcal{X}$.

$\mathcal{G}$ is a finite group of transformations from $\mathcal{X}$ to $\mathcal{X}$. $\#\mathcal{G} = G$.

Want to test null hypothesis $H_0 : P \in \Omega_0$.

Suppose H_0 implies that P is invariant under $\mathcal{G}$:

$$\forall g \in \mathcal{G}, \;\; X \sim gX. \tag{40}$$

The *orbit of* x *(under* $\mathcal{G}$*)* is $\{gx : g \in \mathcal{G}\}$. (Does the orbit of x always contain G points?)

Test statistic T

Let $T : \mathcal{X} \to \Re$ be a test statistic.

We want to test H_0 at significance level α.

For each fixed x, let $T^{(k)}(x)$ be the kth smallest element of the multiset

$$[T(gx) : g \in \mathcal{G}]. \tag{41}$$

These are the G (not necessarily distinct) values T takes on the orbit of x.

Finding the rejection region

Let

$$r \equiv G - \lfloor \alpha G \rfloor. \tag{42}$$

Define

$$G^{+}(x) \equiv \#\{g \in \mathcal{G} : T(gx) > T^{(r)}(x)\} \tag{43}$$

and

$$G^{r}(x) \equiv \#\{g \in \mathcal{G} : T(gx) = T^{(r)}(x)\} \tag{44}$$

Finding the rejection region, contd.

Let

$$a(x) \equiv \frac{\alpha G - G^{+}(x)}{G^{r}(x)}. \tag{45}$$

Define

$$\phi(x) \equiv \begin{cases} 1, & T(x) > T^{(r)}(x), \\ a(x), & T(x) = T^{(r)}(x), \\ 0, & T(x) < T^{(r)}(x) \end{cases} \tag{46}$$

To test the hypothesis, generate $U \sim U[0,1]$ independent of X.

Reject H_0 if $\phi(x) \geq U$. (Randomized test.)

Test has level α unconditionally

For each $x \in \mathcal{X}$,

$$\sum_{g \in \mathcal{G}} \phi(gx) = G^+(x) + a(x) G^r(x) = \alpha G. \tag{47}$$

So if $X \sim gX \sim P$ for all $g \in \mathcal{G}$,

$$\begin{aligned} \alpha &= \mathbb{E}_P \frac{1}{G} \sum_{g \in \mathcal{G}} \phi(gX) \\ &= \frac{1}{G} \sum_{g \in \mathcal{G}} \mathbb{E}_P \phi(X) \\ &= \mathbb{E}_P \phi(X). \end{aligned} \tag{48}$$

The unconditional chance of a Type I error is exactly α.

Tests for the mean of a symmetric distribution

Data $X = (X_j)_{j=1}^N \in \mathcal{X} = \Re^n$.

$\{X_j\}$ iid P; $\mathbb{E}X_j = \mu$.

Suppose P is symmetric. Examples?

Reflection group

Let $\mathcal{G}_\mu$ be the group of reflections of coordinates about μ.

Let $x \in \Re^n$. Each $g \in \mathcal{G}_\mu$ is of the form

$$g(x) = (\mu + (-1)^{\gamma_j}(x_j - \mu))_{j=1}^n \tag{49}$$

for some $\gamma = (\gamma_j)_{j=1}^n \in \{0, 1\}^n$.

Is $\mathcal{G}_\mu$ really a group?

What's the identity element? What's the inverse of g? What γ corresponds to $g_1 g_2$?

What is G, the number of elements of $\mathcal{G}_\mu$?

What is the orbit of a point x under $\mathcal{G}_\mu$? Are there always 2^n distinct elements of the orbit?

Test statistic

$$T(X) = |\bar{X} - \mu_0| = \left| \frac{1}{n} \sum_{j=1}^{n} X_j - \mu_0 \right|. \tag{50}$$

If $\mathbb{E}X_j = \mu_0$, this is expected to be small—but how large a value would be surprising?

If the expected value of X_j is μ and P is symmetric (i.e., if H_0 is true), the 2^n potential data

$$\{gX : g \in \mathcal{G}_\mu\} \tag{51}$$

in the orbit of X under $\mathcal{G}$ are equally likely.

Hence, the values in the multiset

$$[T(gx) : g \in \mathcal{G}_\mu] \tag{52}$$

are equally likely, conditional on the event that X is in the orbit of x.

How to test H_0: $\mu = \mu_0$?

We observe $X = x$.

If fewer than αG values in $[T(gx) : g \in \mathcal{G}_{\mu_0}]$ are greater than or equal to $T(x)$, reject. If more than αG values are greater than $T(x)$, don't reject. Otherwise, randomize.

If n is big . . .

How can we sample at random from the orbit?

Toss fair coin n times in sequence, independently. Take $\gamma_j =$ 1 if the jth toss gives heads; $\gamma_j = 0$ if tails.

Amounts to sampling with replacement from the orbit of x.

Other test statistics?

Could use t-statistic, but calibrate critical value using the permutation distribution.

Could use a measure of dispersion around the hypothesized mean (the true mean minimizes expected RMS difference, assuming variance is finite).

What about counting the number of values that are above μ_0?

Define

$$T(x) \equiv \sum_{j=1}^{n} 1_{x \geq \mu_0}. \tag{53}$$

Sign test for the median

We are assuming P is symmetric, so the expected value and median are equal.

To avoid unhelpful complexity, suppose P is continuous. Then

$$\Pr\{X_j \geq \mu\} = 1/2, \tag{54}$$

$$\{1_{X_j \geq \mu}\}_{j=1}^{n} \text{ are iid Bernoulli}(1/2), \tag{55}$$

and

$$T(X) \sim \text{ Binomial}(n, 1/2). \tag{56}$$

This leads to the sign test: Reject the hypothesis that the median is μ_0 if $T(X)$ is too large or too small. Thresholds set to get level α test, using the fact that $T(X) \sim$ Binomial$(n, 1/2)$.

Is the sign test equivalent to the permutation test for the same test statistic?

Suppose we are using the test statistic $T(x) = \sum_{j=1}^{n} 1_{x \geq \mu_0}$. Do the permutation test and the sign test reject for the same values of $x \in \mathcal{X}$?

Suppose no component of x is equal to μ_0. According to the sign test, the chance that $T(x) = k$ is $\binom{n}{k} 2^{-n}$ if the null is true.

What's the chance that $T(x) = k$ according to the permutation test? There are $G = 2^n$ points in the orbit of x under $\mathcal{G}$. If the null is true, all have equal probability 2^{-n}. Of these points, $\binom{n}{k}$ have k components with positive deviations from μ_0. Hence, for the permutation test, the chance that $T(x) = k$ is also $\binom{n}{nk} 2^{-n}$: The two tests are equivalent.

Confidence intervals for μ for symmetric P

Invert two-sided tests.

What if P is not symmetric?

Romano [**?**]: the permutation test based on either the sample mean or Studentized sample mean still have the right level asymptotically.

Heuristic: if $\mathbf{Var}X_j$ is finite and n is large, the sample mean is approximately normal—and thus symmetric—even when the distribution of X_j is not symmetric. The permutation distribution of the sample mean or Studentized sample mean under the null approaches a normal with mean zero and the "right" variance.

Assignment

Implement the two-sided, one-sample permutation test using the sample mean as the test statistic, simulate results with and without symmetry.

Compare the power with the t-test for a normal.

Compare power and level under symmetry with P not normal.

Compare power and level when P is asymmetric (e.g., absolute value of a Normal, non-central χ^2 with a small number of degrees of freedom, triangular distribution, . . .)

Runs test for independence

Suppose we toss a biased coin N times, independently. The coin has chance p of landing heads in each toss and chance $1-p$ of landing tails in each toss. We do not know p.

Since the tosses are iid, they are exchangeable: The chance of any particular sequence of n heads and $N-n$ tails is $p^n(1-p)^{N-n}$. The probability distribution is invariant under permutations of the trials.

For any particular observed sequence of n heads and $N-n$ tails, the orbit of the data under the action of the permutation group consists of all $\binom{N}{n}$ sequences of n heads (H) and $N-n$ tails (T). That amounts to conditioning on the number n of heads in the (fixed) number N of tosses, but not on whether each toss resulted in H or T.

Runs test, contd.

A *run* is a sequence of H or T. The sequence HHTHTTTHTH has 7 runs: HH, T, H, TTT, H, T and H. If the tosses are independent, each arrangement of the n heads and $m \equiv N-n$ tails among the N tosses has probability $1/\binom{N}{n}$; there are $1/\binom{N}{n}$ equally likely elements in the orbit of the observed sequence under the permutation group. We will compute the (conditional) probability distribution of R given n, assuming independence of the trials.

If $n = N$ or if $n = 0$, $R \equiv 1$. If $k = 1$, there are only two possibilities: first all the heads, then all the tails, or first all the tails, then all the heads. I.e., the sequence is either

$$(\mathrm{HH}\ldots\mathrm{HTT}\ldots\mathrm{T}) \text{ or } (\mathrm{TT}\ldots\mathrm{THH}\ldots\mathrm{H})$$

The probability that $R = 1$ is thus $2/\binom{N}{n}$, if the null hypothesis is true.

Runs test, contd.

How can R be even, i.e., $R = 2k$? If the sequence starts with H, we need to choose where to break the sequence of H to insert T, then where to break that sequence of T to insert H, etc. If the sequence starts with T, we need to choose where to break the sequence of T to insert H, then where to break that sequence of H to insert T, etc.

We need to break the n heads into k groups, which means picking $k-1$ breakpoints, but the first breakpoint needs to come after the first H, and the last breakpoint needs to come before the nth H, so there are only $n-1$ places those $k-1$ breakpoints can be. And we need to break the m tails into k groups, which means picking $k-1$ breakpoints, but the first needs to be after the first T and the last needs to be before the mth T, so there are only $m-1$ places those $k-1$ breakpoints can be.

Runs test, contd.

The number of sequences with $R = 2k$ that start with H is thus

$$\binom{n-1}{k-1} \times \binom{m-1}{k-1}. \tag{57}$$

The number of sequences with $R = 2k$ that start with T is the same (just read right-to-left instead of left-to-right). Thus, if the tosses are independent and there are n heads in all,

$$P\{R = 2k\} = 2 \times \binom{n-1}{k-1} \times \binom{m-1}{k-1} / \binom{N}{n}. \tag{58}$$

Runs test, contd.

Now consider how we can have $R = 2k+1$ (odd). Either the sequence starts and ends with H or it starts and ends with T. Suppose it starts with H. Then we need to break the string of n heads in k places to form $k+1$ groups using k groups of tails formed by breaking the m tails in $k-1$ places. If the sequence starts with T, we need to break the m tails in k places to form $k+1$ groups using k groups of heads formed by breaking the n heads in $k-1$ places. Thus, if the tosses are independent and there are n heads in all,

$$P\{R = 2k+1\} = \frac{\binom{n-1}{k} \times \binom{m-1}{k-1} + \binom{n-1}{k-1} \times \binom{m-1}{k}}{\binom{N}{n}}. \qquad (59)$$

Runs test, contd.

Nothing in this derivation used the probability p of heads. The conditional distribution under the null hypothesis depends only on the fact that the tosses are iid, so that all arrangements with a given number of heads are equally likely.

Note the connection with the sign test for the median and the one-sample test for the mean of a symmetric istribution, discussed previously. There, under the null, we had exchangeability with respect to reflections and permutations. The alternative still had exchangeability with respect to permutations, but not reflections.

Here, under the null we have exchangeability with respect to permutations. Under the alternative, we don't.

Runs test, contd.

The runs test is also connected to Fisher's Exact Test—in both, we condition on the number of "successes" and look at whether those successes are distributed in the way we would expect if the null held.

Runs test, contd.

Let I_j be the indicator of the event that the outcome of the $j+1$st toss differs from the outcome of the jth toss, $j = 1, \ldots, N-1$. Then

$$R = 1 + \sum_{j=1}^{N-1} I_j. \tag{60}$$

Runs test, contd.

Under the null, conditional on n,

$$
\begin{aligned}
P\{I_j = 1\} &= P\{I_j = 1 | j\text{th toss lands H, } n\} \times \\
&\quad P\{j\text{th toss lands H} | n\} + \\
&\quad P\{I_j = 1 | j\text{th toss lands T, } n\} \times \\
&\quad P\{\ j\text{th toss lands T} | n\} \\
&= P\{j+1\text{st toss lands T } | j\text{th toss lands H}, n\} \times \\
&\quad P\{j\text{th toss lands H} | n\} + \\
&\quad P\{j+1\text{st toss lands H} | j\text{th toss lands T}, n\} \times \\
&\quad P\{j\text{th toss lands T} | n\} \\
&= [m/(N-1)] \times [n/N] + [n/(N-1)] \times [m/N] \\
&= 2nm/[N(N-1)]. \qquad (61)
\end{aligned}
$$

Runs test, contd.

The indicators I_j are identically distributed under the null hypothesis, so if the null holds,

$$\begin{aligned} \mathbb{E}R &= \mathbb{E}[1 + \sum_{j=1}^{N-1} I_j] \\ &= 1 + (N-1) \times 2nm/[N(N-1)] \\ &= 1 + 2nm/N. \end{aligned} \tag{62}$$

Example of Runs test

Air temperature is measured at noon in a climate-controlled room for 20 days in a row. We want to test the null hypothesis that temperatures on different days are independent and identically distributed.

Let T_j be the temperature on day j, $j = 1, \ldots, 20$. If the measurements were iid, whether each day's temperature is above or below a given temperature t is like a toss of a possibly biased coin, with tosses on different days independent of each other. We could consider a temperature above t to be a head and a temperature below t to be a tail.

Example, contd.

Take t to be the median of the 20 measurements. In this example, n=10, m=10, N=20. We will suppose that there are no ties among the measured temperatures. Under the null hypothesis, the expected number of runs is

$$\mathbb{E}R = 1 + 2mn/N = 11. \tag{63}$$

The minimum possible number of runs is 2 and the maximum is 20. Since we expect temperature on successive days to have positive serial correlation (think about it!), we might expect to see fewer runs than we would if temperatures on different days were independent. So, let's do a one-sided test that rejects if there are too few runs. We will aim for a test at significance level 5%.

Example, contd.

$$P\{R=2\} = \frac{2}{\binom{20}{10}} = 1.082509e-05. \tag{64}$$

$$P\{R=3\} = 2 \times \frac{\binom{9}{1}\binom{9}{0}}{\binom{20}{10}} = 9.74258e-05. \tag{65}$$

$$P\{R=4\} = 2 \times \frac{\binom{9}{1}\binom{9}{1}}{\binom{20}{10}} = 8.768321e-04. \tag{66}$$

Example, contd.

$$P\{R = 5\} = 2 \times \frac{\binom{9}{2}\binom{9}{1}}{\binom{20}{10}} = 0.003507329. \quad (67)$$

$$P\{R = 6\} = 2 \times \frac{\binom{9}{2}\binom{9}{2}}{\binom{20}{10}} = 0.01402931. \quad (68)$$

$$P\{R = 7\} = 2 \times \frac{\binom{9}{3}\binom{9}{2}}{\binom{20}{10}} = 0.03273507. \quad (69)$$

$$P\{R \leq 6\} = 2 \times \frac{2 + 9 + 81 + 324 + 1296}{\binom{20}{10}} \approx 0.0185. \quad (70)$$

Example, contd.

So, we should reject the null hypothesis if $R \leq 6$, which gives a significance level of 1.9%. Including $R = 7$ in the rejection region would make the significance level slightly too big: 5.1%.

Normal approximation to the null distribution of runs

When N, n and m are large, the combinatorics can be difficult to evaluate numerically. There are at least two options: asymptotic approximation and simulation. There is a normal approximation to the null distribution of R. As n and $m \to \infty$ and $m/n \to \gamma$,

$$[R - 2m/(1+\gamma)]/\sqrt{4\gamma m/(1+\gamma)^3} \to N(0,1) \qquad (71)$$

in distribution.

Code for runs test

Here is an R function to simulate the null distribution of the number R of runs, and evaluate the P-value of the observed value of R conditional on n, for a one-sided test against the alternative that the distribution produces fewer runs than independent trials would tend to produce. The input is a vector of length N; each element is equal to either "1" (heads) or "-1" (tails). The test statistic is calculated by finding $1 + \sum_{j=1}^{N-1} I_j$, as we did above in finding $\mathbb{E}R$.

```
simRunTest <- function(x, iter) {
    N <- length(x);
    ts <- 1 + sum(x != c(x[2:N],x[N]));
        # test statistic: count transitions I_j.
    sum(replicate(iter, {
                  xp <- sample(x);
                  ((1 + sum(xp != c(xp[2:N],xp[N]))) <= ts)
                        }
                 )
       )/iter
}
```

Numerical example

Suppose the observed sequence is $x = (-1, -1, 1, 1, 1, -1, 1)$, for which $N = 7$, $n = 4$, $m = 3$ and $R = 4$. In one trial with $iter = 10,000$, the simulated P-value using simRunTest was 0.5449. Exact calculation gives

$$P_0(R \leq 4) = (2 + 5 + 12)/35 = 19/35 \approx 0.5429. \qquad (72)$$

The standard error of the estimated probability is thus

$$SE = [(19/35 \times 16/35)/10000]^{1/2} \approx 0.005. \qquad (73)$$

The simulation was off by about

$$(0.5449 - 0.5429)/0.005 \approx 0.41\text{SE}. \qquad (74)$$

Two-sample Tests

We observe $\{X_j\}_{j=1}^n$ iid F_X and $\{Y_j\}_{j=1}^m$ iid F_Y.

Want to test the strong null hypothesis $H_0 : F_X = F_Y$.

Let $N = n + m$, $X = (X_j, \ldots, X_n, Y_1, \ldots, Y_m) \in \Re^N$.

Let $\pi = (\pi_j)_{j=1}^N$ be a permutation of $\{1, \ldots, N\}$. Let $\mathcal{G}$ be the permutation group on $\Re^N$. Note that $G = \#\mathcal{G} = N!$.

Under the null, the probability distribution of X is invariant under $\mathcal{G}$. The distribution of X and gX is the same: For any permutation π, we are just as likely to observe $X = gx = (x_{\pi_j})_{j=1}^N$ as we are to observe $X = x$.

Permutation group

$$G = \#\mathcal{G} = N! \tag{75}$$

$$\#\{gx : g \in \mathcal{G}\} \leq N! \tag{76}$$

(less than $N!$ if x has repeated components).

$$\#\{T(gx) : g \in \mathcal{G}\} \leq \#\{gx : g \in \mathcal{G}\} \leq N! \tag{77}$$

will be much less than $\binom{N}{n}$ if T only cares about sets assigned to the first n components and to the last m components, not the ordering within those sets.

"Impartial" use of the data: Arrangements

An "arrangement" of the data is a partition of it into two sets, n considered to be from the first (X) population and m considered to be from the second (Y) population. In an arrangement, the order of the values within each of those two sets does not matter.

Some statisticians require the test statistic to be "impartial": Since $\{X_j\}$ are iid and $\{Y_j\}$ are iid, statistic shouldn't privilege some X_j or Y_j over others. The labeling is considered to be arbitrary.

Not compelling for sequential methods, where the labeling could indicate the order in which the observations were made.

For the test statistics we consider, only the arrangement matters: They are impartial.

Unbiased tests

A test is *unbiased* if the chance it rejects the null is never smaller when the null is false than when the null is true.

Testing whether $\mathbb{E}F_X = \mathbb{E}F_Y$

$F_X = F_Y$ implies $\mathbb{E}F_X = \mathbb{E}F_Y$, but not *vice versa*. Weaker hypothesis.

Start with *strong* null that $F_X = F_Y$, but want test to be sensitive to differences between $\mathbb{E}F_X$ and $\mathbb{E}F_Y$. We are testing the strong null, but we want power against the alternative $\mathbb{E}F_X \neq \mathbb{E}F_Y$.

Test statistic. Let $\bar{X} \equiv \frac{1}{n}\sum_{j=1}^{n} X_j$ and $\bar{Y} \equiv \frac{1}{m}\sum_{j=1}^{m} Y_j$.

$$T_{n,m}(X) = T_{n,m}(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \equiv n^{1/2}(\bar{X} - \bar{Y}) \quad (78)$$

To test, if we observe $X = x$, look at the values of $\{T_{n,m}(gx) : g \in \mathcal{G}\}$. We reject if $T(x)$ is "extreme" compared with those values.

Distribution of $T_{n,m}$ without invariance

Romano [**?**] shows that if $\mathbb{E}F_X = \mathbb{E}F_Y = \mu$, if F_X and F_Y have finite variances, and if $m/N \to \lambda \in (0,1)$ as $n \to \infty$, then the asymptotic distribution of $T_{n,m}$ is normal with mean zero and variance

$$\sigma_p^2 = \lambda^{-1/2}(\lambda \mathbf{Var} F_Y + (1-\lambda)\mathbf{Var} F_X). \tag{79}$$

The permutation distribution and unconditional asymptotic distributions are equal only if $\mathbf{Var} F_X = \mathbf{Var} F_Y$ or $\lambda = 1/2$.

Hence, whether the test is asymptotically valid for the "weak" null hypothesis depends on the relative sample sizes.

Pitman's papers

[**?**, **?**, **?**]

Paper 1: Two-sample test for equality of two distribution based on sample mean. This is what we just looked at.

Paper 2: correlation by permutation. Exchangeability required. Issues for time series in particular.

Paper 3: ANOVA by permutation.

Permutation test for association

Observe $\{(X_j, Y_j)\}_{j=1}^n$. $\{X_j\}_{j=1}^n$ are iid and $\{Y_j\}_{j=1}^n$ are iid.

The pairs are independent of each other; the question is whether, within pairs, X and Y are independent.

If they were, the joint distribution of $\{(X_j, Y_j)\}_{j=1}^n$ would be the same as the joint distribution of $\{(X_j, Y_{\pi_j})\}_{j=1}^n$ for every permutation π of $\{1, \ldots, n\}$. The joint distribution of $\{(X_j, Y_j)\}_{j=1}^n$ would be invariant under the group of permutations of the indices of the Y variables.

The Ys would be *exchangeable* given the Xs. Exchangeability involves not just the independence of X and Y, but also the fact that the Ys are iid. If they had different distributions or were dependent, exchangeability would not hold.

Test Statistic

Testing the null hypothesis that $\{Y_j\}$ are exchangeable given $\{X_j\}$, which is implied by the assumption that $\{Y_j\}$ are iid, combined with the null that $\{X_j\}$ are independent of $\{Y_j\}$.

Can treat either $\{X_j\}$ or $\{Y_j\}$ as fixed numbers—not necessarily random. For instance, $\{X_j\}$ could index deterministic locations at which the observations $\{Y_j\}$ were made.

The issue is whether, conditional on one of the variables, all pairings with the other variable are equally likely.

We want to be sensitive to violations of the null for which there is dependence within pairs. One statistic that is sensi-

tive to linear association is the "ordinary" Pearson correlation coefficient:

$$r_{XY} \equiv \frac{1}{n} \sum_{j=1}^{n} \frac{(X_j - \bar{X})(Y_j - \bar{Y})}{\mathrm{SD}(X)\mathrm{SD}(Y)}, \tag{80}$$

where $\mathrm{SD}(X) \equiv (\sum_{j=1}^{n}(X_j - \bar{X})^2/n)^{1/2}$, and $\mathrm{SD}(Y)$ is defined analogously.

Example: acclamations and coinage symbols

Example from Norena [**?**].

Roman emperors were "acclaimed" with various honors during their reigns. Coins minted in their reigns have a variety of symbols, one of which is Victoria (symbolizing military victory). (Coins recovered from various caches; assumed to be a representative sample—which is not entirely plausible.) Look at emperors from the year 96 to 218.

Imperatorial acclamations and Victoria coinage

Emperor	relative frequency of Victoria coins	imperatorial acclamations
Nerva	0	2
Trajan	35	13
Hadrian	14	1
Antoninus Pius	3	0
Marcus Aurelius	16	10
Commodus	11	8
Septimius Severus	28	11
Caracalla	15	3
Macrinus	1	0

Correlation coefficient is $r = 0.844$.

Permutation test P-value

```
iter <- 10^6;  # iterations used in simulation

simPermuTest <- function(x, y, iter) {   # simulated permutation
distribution
           ts <- abs(cor(x,y))                  # test statistic
           sum(replicate(iter, (abs(cor(x,sample(y))) >= ts)))/iter
}

x <-  c(0, 35, 14, 3, 16, 11, 28, 15, 1);
y <-  c(2, 13, 1, 0, 10, 8, 11, 3, 0);

cor(x11, y11)   #  0.844
simPermuTest(x11, y11, iter)  #  0.003
```

The P-value is 0.003. Tests on four other symbols and acclamations gave P-values ranging from 0.011 to 0.048.

Spearman's Rank Correlation

Test statistic: replace each X_j by its rank and each Y_j by its rank. Test statistic is the correlation coefficient of those ranks. This is Spearman's rank correlation coefficient, denoted $r_S(X, Y)$.

Significance level from the distribution of the statistic under permutations, as above.

Can use *midranks* to deal with tied observations.

What if $\{Y_j\}$ are dependent or have different distributions?

In time series context where j indexes time, typical that $\{Y_j\}$ have different distributions and are dependent.

Trends, cycles, etc., correspond to different means at different times. Variances can depend on time, too. So can other aspects of the distribution.

Failure of exchangeability: serial correlation

The next few examples are from Walther [**?**, **?**].

Suppose $\{X_j\}_{j=1}^{100}$ and $\{Y_j\}_{j=1}^{100}$ are iid $N(0,1)$ and independent of each other. Let

$$S_k \equiv \sum_{j=1}^{k} X_j \quad \text{and} \quad T_k \equiv \sum_{j=1}^{k} Y_j. \tag{81}$$

Then

$$P(r_S(S,T) > c_{0.01}) \approx 0.67, \tag{82}$$

where $c_{0.01}$ is the critical value for a one-sided level 0.01 test against the alternative of positive association.

Serial correlation, contd.

Even though $\{S_j\}$ and $\{T_j\}$ are independent, the probability that their Spearman rank correlation coefficient exceeds the 0.01 critical value for the test is over 2/3.

That is because the two series S and T each have serial correlation: not all pairings (S_j, T_{π_j}) are equally likely—even though the two series are independent. The series are not conditionally exchangeable even though they are independent, because neither series is iid.

Failure of exchangeability: difference in variances

Serial correlation is not the only way that exchangeability can fail. For example, if the mean or the noise level varies with time, that violates the null hypothesis.

Take $X = (1, 2, 3, 4)$ fixed. Let (Y_1, Y_2, Y_3, Y_4) be independent, jointly Gaussian with zero mean, $\sigma(Y_j) = 1$, $j = 1, 2, 3$, and $\sigma(Y_4) = 2$. If $\{Y_j\}$ were exchangeable—which they are not—then

$$P_0(r_S(X, Y) = 1) = 1/4! = 1/24 \approx 4.17\%. \qquad (83)$$

($r_S = 1$ whenever $Y_1 < Y_2 < Y_3 < Y_4$.) Simulation shows that $P(r_S(X, Y) = 1) \approx 7\%$:

Simulation estimate of chance $r_S = 1$ for non-exchangeable data

```
iter <- 10000;
s <- c(1,1,1,4);
sum(replicate(iter, {
            x <- rnorm(length(s), sd=s);
            !is.unsorted(x)  # r_S=1 if x is ordered
                      }
              )
    )/iter
```

Failure of exchangeability: difference in variances

Take $X = (1, 2, 3, 4, 5)$ fixed, let $(Y_1, Y_2, Y_3, Y_4, Y_5)$ be independent, jointly Gaussian with zero mean and standard deviations 1, 1, 1, 3, and 5, respectively.

Under the (false) null hypothesis that all pairings $\{(X_j, Y_{\pi_j})\}$ are equally likely,

$$P_0 r_S(X, Y) = 1 = 1/5! \approx 0.83\%, \tag{84}$$

Simulation shows that the actual probability is about 2.1%.

In these examples, the null hypothesis is false, but not because $\{X_j\}$ and $\{Y_j\}$ are dependent. It is false because not all pairings $\{(X_j, Y_{\pi_j})$ are equally likely. The "identically distributed" part of the null hypothesis fails.

Failure of exchangeability: difference in means

$X_j = j$, ... [FIX ME!]

Permutation F-test

We have m batches of n subjects, not necessarily a sample from any larger population.

In each batch, the n subjects are assigned to n treatments at random. Let r_{jk} be the response of the subject in batch j who was assigned to treatment k, $j = 1, \ldots, m$, $k = 1, \ldots, n$.

Linear model:

$$r_{jk} = B_j + T_k + e_{jk}. \tag{85}$$

B_j is a "batch effect" that is the same for all subjects in batch j. T_k is the effect of treatment k. e_{jk} is an observational error for the individual in batch j assigned to treatment k.

Want to test the null hypothesis that $T_1 = T_2 = \ldots = T_n$.

ANOVA

Decompose total sum of squares:

$$S = S_B + S_T + S_e, \tag{86}$$

where S_B is independent of the treatments, S_T is independent of the batches, and S_e is the residual sum of squares, and is independent of batches and of treatments.

Test statistic

$$F = \frac{S_T}{S_T + S_e}. \tag{87}$$

Large if the "within-batch" variation can be accounted for primarily by an additive treatment effect.

In usual F-test, assume that $\{e_{jk}\}$ are iid normal with mean zero, common variance σ^2.

ANOVA "ticket model"

Each individual is represented by a ticket with n numbers on it. The kth number on ticket m is the response that individual m would have if assigned to treatment k. We assume non-interference: each individual's potential responses are summarized by those n numbers; they do not depend on which treatment any other subjects are given.

If treatment has no effect whatsoever, then, individual m's n numbers are all equal. That is, within each batch, the treatment label is arbitrary. Responses should be invariant under permutations of the treatment labels.

Let π^j be a permutation of $\{1, \ldots, n\}$, for $j = 1, \ldots, m$. If we observe $\{r_{jk}\}$ and the null is true, we might just as well have observed $\{r_{j\pi^j_k}\}$

Keep batches together, but permute the treatment labels within batches.

Permutation distribution

Condition on the event that the data are in the orbit of the observed data under the action of the permutation group. Then all points in that orbit are equally likely.

Find the P-value by comparing the observed value of F with the distribution of values in the orbit of the observed data under the group that permutes the labelings within each batch.

Reject the null hypothesis if the observed value of F is surprisingly large.

If the space of permutations is too big, can use randomly selected permutations.

How strong is the usual null for the F-test?

The permutation test just described tests the strong null that the n numbers on individual m's ticket are equal. A weaker null that might be of interest is whether the n means of the mn potential responses to each of the n treatments are equal—that is, whether on average there is any difference among the treatments.

Does the usual F-test test this weaker null? What hypothesis does the F-test actually address?

F-test null hypothesis

For F to have an F-distribution, the "noise" terms $\{e_{jk}\}$ have to be iid normal random variables with mean zero (that makes it the ratio of two chi-square variables if the treatment effects are equal). The distribution of $\{e_{jk}\}$ cannot depend on which batch the subject is in nor which treatment the subject receives.

Why should each e_{jk} have expected value zero? Why should e_{jk} be independent of the assignment to treatment? Why should e_{jk} have the same distribution for all individuals?

Why should the effect of a given treatment be the same for every individual?

Why is there no interaction between treatment and batch?

Why is the "batch effect" the same for all members of a batch?

Sharpening the description

To study whether different treatments have different effects, need to consider hypothetical counterfactuals.

What would the response of the individual in batch j assigned to treatment k have been, if the individual instead had been assigned to treatment $\ell \neq k$?

The model

$$r_{jk} = B_j + T_k + e_{jk} \tag{88}$$

doesn't say. We need a "response schedule" [**?**].

Response schedule

Think of the subjects within batches before they are assigned to treatment. The response schedule says that *if* subject ℓ in batch j is assigned to treatment k, his response will be

$$r_\ell = B_j + T_k + e_\ell. \tag{89}$$

If he is assigned to treatment k', his response will be

$$r_\ell = B_j + T_{k'} + e_\ell. \tag{90}$$

How is e_ℓ generated? Would e_ℓ be the same if subject ℓ were assigned to treatment k' instead of treatment k?

Two visions of $\{e_\ell\}$: Vision 1

The errors $\{e_\ell\}$ are generated (iid Normal, mean zero) before the subjects are assigned to treatments. Once generated, $\{e_\ell\}$ are intrinsic properties of the individuals—unaffected by assignment to treatment. And the assignment to treatment does not depend on $\{e_\ell\}$.

If this is true, then if the treatment effects $\{T_k\}$ are all equal, that implies *more* than the strong null: Within each batch, each individual's responses would be the same no matter which treatment was assigned—the n numbers on each individual's ticket are equal, just as in the strong null. But in addition, the expected responses of *all* subjects in each batch are equal.

Two visions of $\{e_\ell\}$: Vision 2

$\{e_\ell\}$ are generated after the assignment to batch and treatment, but their distribution does not depend on that assignment. If the assignment had been different, $\{e_\ell\}$ might have been different—but it is always Normal with zero mean and the same variance, and never depends on the assignment of that subject or any other subject.

If this is true, then the null that the treatment effects $\{T_k\}$ are all equal implies a weakening of the null: Within each batch, each individual's responses are the same in expectation no matter which treatment is assigned, but a given individual might have n different numbers on his ticket.

Two visions of $\{e_\ell\}$: Vision 2, contd.

However, the hypothesis says more than that: Within each batch, every subject's expected response is the same—the same as the expected response of all the other subjects, no matter which treatment is assigned to each of them.

This seems rather stronger than the "strong null" that the permutation test tests, since the strong null does not require different subjects to have the same expected responses.

Note that in neither vision 1 nor vision 2 is the null the "natural" weak null that the average of the responses (across subjects) for each treatment are equal, even though the numbers are not necessarily equal subject by subject.

Other test statistics

The F statistic makes sense for testing against the omnibus alternative that there is a difference among the treatment effects.

If there is more structure to the alternative, can devise tests with more power against the alternative.

For instance, if, under the alternative, the treatment effects are ordered, can use a test statistic that is sensitive to the order. Suppose that the treatments are ordered so if the alternative is true, $T_1 \le T_2 \le \cdots \le T_n$.

Pitman correlation

$$S \equiv \sum_{k=1}^{n} f(k) \sum_{j=1}^{m} r_{jk} \tag{91}$$

where f is a monotone increasing function.

Most powerful permutation tests

How should we select the rejection region?

If we have a simple null and a simple alternative, most powerful test is likelihood ratio test.

If the null completely specifies the chance of outcomes in the orbit of the data and the alternative does too, can find the likelihood ratio for each point in the orbit.

Problem: The orbits might not be the same under the null and under the alternative. If they were the same, the chance of each outcome in the orbit would be the same under the two hypotheses, since the elements of the orbit are equally likely.

Example: incompatible orbits

6 subjects; 3 assigned to treatment and 3 to control, at random.

Null: treatment has no effect at all.

Alternative: treatment increases the response by 1 unit.

Data: responses to control $\{1, 2, 3\}$. Responses to treatment $\{2, 3.5, 4\}$.

Under the null, the orbit consists of the 20 ways of partitioning $\{1, 2, 2, 3, 3.5, 4\}$ into two groups of 3; equivalently, the 6! permutations of $\{1, 2, 2, 3, 3.5, 4\}$.

Example: incompatible orbits, contd.

Under the alternative, the orbit consists of 20 points, but they are generated differently: Find the responses that would have been observed if none had been assigned to treatment by subtracting the hypothesized treatment effect. That gives $\{1, 1, 2, 2.5, 3, 3\}$.

Take $\{1, 1, 2, 2.5, 3, 3\}$ and partition it into two groups of 3; add 1 to each element of the second group. Equivalently, take each of the 6! permutations of $\{1, 1, 2, 2.5, 3, 3\}$ and add 1 to the last 3 elements.

Example: incompatible orbits, contd.

Under the null, one element of the orbit is $(1, 2, 2, 3, 3.5, 4)$. That point is not in the orbit under the alternative.

Under the alternative, one element of the orbit is $(1, 1, 2, 3.5, 4, 4)$. That point is not in the orbit under the null.

Most powerful test would have in its rejection region those points that are in the orbit under the null but not in the orbit under the alternative, since for those points, the likelihood ratio is infinite. Hot helpful.

What are we conditioning on? If we condition on the event that the data are in the orbit of the observed data, can't do much.

Most Powerful Permutation Tests

See Lehmann and Romano [**?**], pp. 177ff. [FIX ME!]

The Population Model

We've been taking N subjects as given; the assignment to treatment or control was random.

Now consider instead the N subjects to be random samples from two populations.

Observe $\{X_j\}_{j=1}^n$ iid F_X and $\{Y_j\}_{j=1}^m$ iid F_Y.

Want to test the hypothesis $F_X = F_Y$.

Many scenarios give rise to the same conditional null distribution for the permutation test:

- Randomization model for comparing two treatments. N subjects are given and fixed; n are assigned at random to treatment and $m = N - n$ to control.

- Population model for comparing two treatments. N subjects are a drawn as a simple random sample from a much larger population; n are assigned at random to treatment and $m = N - n$ to control.

- Comparing two sub-populations using a sample from each. A simple random sample of n subjects is drawn from one much larger population with many more than N members, and a simple random sample of m subjects is drawn from another population with many more than m members.

- Comparing two sub-populations using a sample from the pooled population. A simple random sample of N subjects is drawn from the pooled population, giving random samples from the two populations with random sample sizes. Condition on the sample sizes n and m.

- Comparing two sets of measurements. Independent sets of n and m measurements come from two sources. Test hypothesis that the two sources have the same distribution.

In all those scenarios, if the null hypothesis is true the data are exchangeable—the distribution is invariant under permutations.

Hence, a test with the right (conditional) level for one scenario can be applied in all the others.

Aside: stochastic ordering

Suppose X and Y are real-valued random variables. X is *stochastically larger* than Y, written $Y \preceq X$, if

$$\Pr\{X \geq x\} \geq \Pr\{Y \geq x\} \quad \forall x \in \Re. \tag{92}$$

Exercise: Show that $Y \preceq X$ if and only if for every monotonically increasing function u, $\mathbb{E}u(Y) \leq \mathbb{E}u(X)$.

Example of a power estimate: stochastic ordering

Observe $\{X_j\}_{j=1}^n$ iid F_X and $\{Y_j\}_{j=1}^m$ iid F_Y.

Under the null, $F_X = F_Y$.

Under the alternative F_X is stochastically than F_Y.

Test at level α using the sample mean as the test statistic.

What's the chance of rejecting the null if the alternative is true?

Aside: the Probability Transform

Let X be a real-valued random variable with continuous cdf F.

Then $F(X) \sim U[0,1]$.

Proof: Note that $F(X) \in [0,1]$. If F is continuous, for $p \in (0,1)$ there is a unique value $x_p \in \Re$ such that $F(x_p) = p$. (That is, F has an inverse function F^{-1} on $(0,1)$.)

For any $p \in (0,1)$,

$$\Pr\{F(X) \leq p\} = \Pr\{X \leq x_p\} = p. \tag{93}$$

But this is exactly what it means to have a uniform distribution.

Patch for non-necessarily continuous F? See below.

Aside: a useful stochastic ordering lemma (From [**?**]; see also [**?**].)

X is stochastically larger than Y iff there is a random variable W and monotone nondecreasing functions f and g such that $g(x) \leq f(x)$ for all x, $g(W) \sim Y$ and $f(W) \sim X$.

Proof: Suppose $Y \preceq X$. Let F_X be the cdf of X and let F_Y be the cdf of Y. Define

$$f(p) \equiv \inf\{x : F_X(x) \geq p\} \quad \text{and} \quad g(p) \equiv \inf\{x : F_Y(x) \geq p\}. \tag{94}$$

Then f and g are nondecreasing functions on $[0, 1]$. Since $F_Y(x) \leq F_X(x)$, it follows that $g(p) \leq f(p)$.

Proof of lemma, contd.

(a) If $p \leq F_X(x)$, then $f(p) \leq f(F_X(x)) \leq x$. (The first inequality follows from the monotonicity of f; the second from the fact that F_X is nondecreasing but can be flat on some intervals. Hence, the smallest y for which $F_X(y) \geq F_X(x)$ can be smaller than x.)

(b) If $f(p) \leq x$ then $F_X(f(p)) \leq F_X(x)$, and hence $p \leq F_X(x)$. (The first inequality follows from the monotonicity of F; the second from the fact that f is the infimum.)

Let $W \sim U[0,1]$. By (a),

$$F_X(x) = \mathsf{Pr}\{W \leq F_X(x)\} \leq \mathsf{Pr}\{f(W) \leq x\}. \qquad (95)$$

By (b),

$$\mathsf{Pr}\{f(W) \leq x\} \leq \mathsf{Pr}\{W \leq F_X(x)\} = F_X(x). \qquad (96)$$

Hence $\mathsf{Pr}\{f(W) \leq x\} = F_X(x)$: $f(W) \sim F_X$. Similarly, $g(W) \sim F_Y$.

Proof of lemma, contd.

In the other direction, suppose there is a random variable W and monotone nondecreasing functions f and g such that $g(x) \leq f(x)$ for all x, $g(W) \sim Y$ and $f(W) \sim X$.

Then

$$\Pr\{X \geq x\} = \Pr\{f(W) \geq x\} \geq \Pr\{g(W) \geq x\} = \Pr\{Y \geq x\}. \tag{97}$$

Consequence of the lemma

Let $\{X_j\}_{j=1}^n$ iid with cdf F_X and $\{Y_j\}_{j=1}^m$ iid with cdf F_Y, with F_X and F_Y continuous. Let $N = n + m$.

Let $X = (X_1, \ldots, X_n, Y_1, \ldots, Y_m) \in \Re^N$. Let $\phi : \Re^N \to [0, 1]$. Suppose that ϕ satisfies two constraints:

(a) If $F_X = F_Y$ then

$$\mathbb{E}\phi(X) = \alpha, \tag{98}$$

so that ϕ is the test function for a level α test of the hypothesis $F_X = F_Y$.

(b) If $x_j \leq x'_j$, $j = 1, \ldots, n$, then

$$\phi(x_1, \ldots, x_n, y_1, \ldots, y_m) \leq \phi(x'_1, \ldots, x'_n, y_1, \ldots, y_m). \tag{99}$$

Then $\mathbb{E}\phi \geq \alpha$ whenever $F_Y \preceq F_X$.

Proof of consequence

By the lemma, there are nondecreasing functions $g \leq f$ and iid random variables $\{W_j\}_{j=1}^N$ such that $f(W_j) \sim F_X$ and $g(W_j) \sim F_Y$.

By (a),

$$\mathbb{E}\phi(g(W_1), \ldots, g(W_n), g(W_{n+1}), \ldots, g(W_N)) = \alpha. \quad (100)$$

By (b),

$$\mathbb{E}\phi(f(W_1), \ldots, f(W_n), g(W_{n+1}), \ldots, g(W_N)) = \beta \geq \alpha. \quad (101)$$

The permutation test based on the sample mean is unbiased under the alternative of stochastic dominance

Suppose $\{X_j\}_{j=1}^n$ are iid with cdf F_X and $\{Y_j\}_{j=1}^m$ are iid with cdf F_Y. Define the test function ϕ so that $\phi(X) = 1$ iff $\sum_{j=1}^n X_j$ (or $\frac{1}{n}\sum_{j=1}^n X_j$) is greater than the sum (or mean, respectively) of the first n elements of $\alpha N!$ of the $N!$ (not necessarily distinguishable) permutations of the multiset $[X_1, \ldots, X_n, Y_1, \ldots, Y_m]$.

Define the power function

$$\beta(F_X, F_Y) = \mathbb{E}\phi(X_1, \ldots, X_n, Y_1, \ldots, Y_m). \tag{102}$$

Then $\beta(F_X, F_X) = \alpha$ and $\beta(F_X, F_Y) \geq \alpha$ for all pairs of distributions for which F_X is stochastically larger than F_Y.

Proof.

We have already proved that $\beta(F_X, F_X) = \alpha$: The permutation test is exact (possibly requiring randomization, depending on α).

To show that $\beta(F_X, F_Y) > \alpha$, suppose that the observed value of X_j is x_j, the observed value of Y_j is y_j, and let $w = (x_1, \ldots, x_n, y_1, \ldots, y_m) \in \Re^N$.

$\phi = 1$ if $\sum_{j=1}^n w_j$ is sufficiently large compared with $\sum_{j=1}^n w_{\pi_j}$ for enough permutations π.

Want to show that $\Pr\{\phi = 1\}$ is larger when $F_Y \preceq F_X$ than when $F_Y = F_X$.

By the lemma, suffices to show that if $x_j \le x'_j$, $j = 1, \ldots, n$, then

$$\phi(x_1, \ldots, x_n, y_1, \ldots, y_m) \le \phi(x'_1, \ldots, x'_n, y_1, \ldots, y_m). \quad (103)$$

I.e., suffices to show that if we would reject the null for data

$$w = (x_1, \ldots, x_n, y_1, \ldots, y_m), \quad (104)$$

we would also reject it for data

$$w' = (x'_1, \ldots, x'_n, y_1, \ldots, y_m). \quad (105)$$

Suppose $\phi(w) = 1$. Then $\sum_{j=1}^n w_j > \sum_{j=1}^n w_{\pi_j}$ for at least $\alpha N!$ of the permutations π of $\{1, \ldots, N\}$. Let π be one such permutation.

Suppose that r of the first n elements of π are not between 1 and n; that is, the permutation replaces r of the first n components of w with later components of w.

Let $\{\ell_k\}_{k=1}^r$ be the components that are moved out of the first n components by permutation π (the components that are "lost") and let $\{f_k\}_{k=1}^r$ be the components that are moved in (the components that are "found"). Then

$$\begin{aligned} 0 &< \sum_{j=1}^{n} w_j - \sum_{j=1}^{n} w_{\pi_j} \\ &= \sum_{k=1}^{r} w_{\ell_k} - \sum_{k=1}^{r} w_{f_k} \\ &\leq \sum_{k=1}^{r} w'_{\ell_k} - \sum_{k=1}^{r} w_{f_k} \\ &= \sum_{j=1}^{n} w'_j - \sum_{j=1}^{n} w'_{\pi_j}, \end{aligned}$$

so $\phi(w') = 1$. Hence $\phi(w') \geq \phi(w)$.

Estimating shifts:

If you know that $F_X(x) = F_Y(x - d)$, can construct a confidence interval for the shift d.

How?

Invert hypothesis tests.

Exercise:

By simulation, estimate the power of the level $\alpha = 0.05$ two-sample test based on the sample mean against the alternative that $F_Y \preceq F_X$ when

1. $Y \sim N(0,1)$ and $X \sim N(\mu,1)$, $\mu = 0.5, 1, 1.5$. Compare with the power of the usual t test.

2. $Y \sim$ Exponential(1) and $X \sim$ Exponential(λ), with $1/\lambda = 1.5, 2, 2.5$. Compare with the power of a likelihood ratio test.

3. $Y \sim$ Chi-square(2) and $X \sim$ Chi-square(b), with $b = 3, 4, 5$. Compare with the power of a likelihood ratio test.

Exercise, contd.

Use sample sizes $n = 10, 50, 100$ and $m = n/2, n$, and $2n$. Justify your choice of the number of replications to use in each simulation. Note: this involves 81 simulations. Please tabulate the results in some readable format.

For each of the scenarios above, *pretend* that $F_X(x) = F_Y(x - d)$ for some shift d that could be positive or negative. (This is true in the first scenario where the data are normal with different means, but not in the second and third, where the distributions have different shapes.) Invert two-sided tests to find 95% confidence intervals for d, which would be the difference in means if the shift model were true. By simulation, estimate the probability that the confidence intervals cover the true difference in means in the 81 cases. Justify your choice of the number of replications to use in each simulation. Discuss the results.

Smirnov Test

Consider a two-sample test against the *omnibus alternative* that there is any difference at all between the two groups. We want a test statistic that is sensitive to differences other than differences in the means.

The Smirnov test is based on the difference between the empirical cumulative distribution function (cdf) of the treatment responses and the empirical cdf of the control responses. It has some power against all kinds of violations of the strong null hypothesis.

Smirnov Test, contd. Let $F_{X,n}$ denote the empirical cdf of the treatment responses:

$$F_{X,n}(x) \equiv \#\{x_j : x_j \leq x\}/n, \tag{106}$$

and define $F_{Y,m}$ analogously as the empirical cdf of the control responses. If there are no ties within the treatment group, $F_{X,n}$ jumps by $1/n$ at each treatment response value. If there are no ties within the control group, $F_{Y,m}$ jumps by 1/m at each control response value. If k of the treatment responses are tied and equal to x_0, then $F_{X,n}$ jumps by k/n at x_0.

Smirnov Test, contd.

The Smirnov test statistic is

$$D_{m,n} \equiv \sup_x |F_{X,n}(x) - F_{Y,m}(x)|. \tag{107}$$

It is easy to see that the supremum is attained at one of the data values. We can also see that the supremum depends only on the ranks of the data, because the order of the jumps matters, but the precise values of x at which the jumps occur do not matter. Therefore, the test

$$\text{Reject null if } D_{m,n} > c, \tag{108}$$

for an appropriately chosen value of c, is a nonparametric test of the strong null hypothesis.

Smirnov Test, contd.

Null distribution of $D_{m,n}$ for $n = 3$, $m = 2$. There are $\binom{5}{3} = 10$ possible assignments of 3 of the subjects to treatment, each of which has probability 1/10 under the strong null hypothesis. Assume that the 5 data are distinct (no ties). Then the possible values of $D_{m,n}$ are

treatment ranks	control ranks	$D_{m,n}$
1, 2, 3	4, 5	1
1, 2, 4	3, 5	2/3
1, 2, 5	3, 4	2/3
1, 3, 4	2, 5	1/2
1, 3, 5	2, 4	1/3
1, 4, 5	2, 3	2/3
2, 3, 4	1, 5	1/2
2, 3, 5	1, 4	1/2
2, 4, 5	1, 3	2/3
3, 4, 5	1, 2	1

Smirnov Test, contd.

The null probability distribution of $D_{m,n}$ is

d	$\Pr\{D_{2,3} = d\}$
1/3	1/10
1/2	3/10
2/3	4/10
1	2/10

Thus a Smirnov test (against the omnibus alternative) with significance level 0.2 would reject the strong null hypothesis when $D_{2,3} = 1$; smaller significance levels are not attainable when n and m are so small.

Smirnov Test, contd.

Critical values can be calculated analytically fairly easily when $n = m$ (when the treatment and control groups are the same size) and the significance level is an integer multiple a of $1/n$. Let $k = \lfloor n/a \rfloor$. Then, under the strong null hypothesis, provided there are no ties,

$$\begin{aligned}\Pr\{D_{m,n} > a/n\} \;=\; & \left(\binom{2n}{n-a} - \binom{2n}{n-2a} + \right. \\ & \left. + \binom{2n}{n-3a} - \cdots + (-1)^{k-1}\binom{2n}{n-ka}\right) \times \\ & \times \left(2\binom{2n}{n}\right)^{-1}. \qquad (109)\end{aligned}$$

Smirnov Test, contd.

When there are ties, $D_{m,n}$ tends to be smaller, so tail probabilities from this expression still give conservative tests. (If both members of a tie are assigned to the same group (treatment or control) the tie does not change the value of $D_{m,n}$ from the value it would have had if the pair differed slightly. If one member of a tie is assigned to treatment and one to control, the tie can decrease the value of $D_{m,n}$ from the value it would have had if the observations differed slightly. Thus $D_{m,n}$ is stochastically smaller when there are ties.

```
simSmirnovTest <- function(x, y, iter) {
# P.B. Stark, statistics.berkeley.edu/~stark  9/8/2006
    ts <- smirnov(x,y)
    z <- c(x, y)                    # pooled responses
    sum(replicate(iter, {
                zp <- sample(z);
                (smirnov(zp[1:length(x)],zp[(length(x)+1):length(z)])
                  }
                 )
        )/iter
}
```

We can also estimate the power of the Smirnov statistic against a shift alternative by simulation. Suppose that the effect of treatment is to increase each subject's response by d, no matter what the response would have been in control. (Remember, this is a very restrictive assumption, but the results might be suggestive anyway.) Then the responses $x = (x_j)_{j=1}^n$ of the treated subjects would have been

$$x - d = (x_1 - d, \ldots, x_n - d) \tag{110}$$

had they been in the control group, and the responses y of the control group would have been $y + d$ had they been in the treatment group.

See old lecture notes, chapter 4.

Multidimensional analogues?

Test statistics: sample mean versus ...

Multidimensional analogue of the runs test: deleting edges from minimal spanning tree. Depends on metric.

Bickel [**?**] probability of lower-left quadrants.

Friedman and Rafsky [**?**] generalized "runs" test based in minimal spanning trees.

Ferger [**?**] change point at known point.

Hall and Tajvidi [**?**] inter-point differences to each point.

Baringhaus and Franz [**?**] sum of interpoint differences across samples minus half the sums of interpoint differences within samples.

Multidimensional analogues of the Smirnov test: VC classes, Romano's work.

Gretton et al. [**?**] generalizes from indicator functions of sets to expectation of more general functions (e.g., elements of a universal reproducing kernel Hilbert space). Good properties if the star of the set of functions is dense in C_0 in the L_∞ norm.

Median versus mean.

Testing the hypothesis of symmetry.

Testing exchangeability: earthquake aftershocks.

The two-sample problem

We observe $\{X_j\}_{j=1}^n$ iid P and $\{Y_j\}_{j=1}^m$ iid Q.

P and Q are measures on a common sigma-algebra $\mathcal{A}$.

Want to test the hypothesis $P = Q$.

(Recall that the math is essentially the same whether the two samples are random samples from two populations, measurements on a single group randomly divided into treatment and control, or—conditional on the sample sizes—a single random sample from a population with two types of elements.)

The two-sample problem in $\mathbb{R}^p$

For $x \in \mathbb{R}^p$, define the lower left quadrant w.r.t. x:

$$L_x \equiv \{y \in \mathbb{R}^p : y_j \leq x_j, j = 1, \ldots, p. \tag{111}$$

Let

$$\delta(P, Q) \equiv \sup_{x \in \mathbb{R}^p} |P(L_x) - Q(L_x)| \tag{112}$$

Bickel [**?**] shows that the permutation principle applied to the test statistic $\delta(\hat{P}, \hat{Q})$ gives a test with exact level α and asymptotic power 1 against every fixed alternative.

Generalizing "runs" test: Minimal spanning trees. Friedman and Rafsky [**?**]

Graph: nodes, pairs of nodes called "edges." An edge connects two nodes if it contains those two nodes. *Directed graph*: the edges are ordered pairs, not just pairs. *Edge-weighted graph*: edges are triples—a pair of nodes and a real number, the edge weight. A graph is *complete* if every pair of nodes has an edge that connects them. Degree of a node: number of edges that include it. *Subgraph*: graph with all nodes and edges among those in a given graph. If two subgraphs have no nodes in common, they are *disjoint*. If two subgraphs have no edges in common, they are *orthogonal*. *Spanning subgraph*: subgraph that contains all nodes in the graph.

Path between two nodes: alternating sequence of nodes and edges so that each edge includes the nodes that surround it; nodes must be distinct except possibly the first and last. *Connected* graph: path between every distinct pair of nodes. *Cycle*: path that starts and ends with the same node. *Tree*: connected graph with no cycles. Connected subgraph of a tree is also a tree, called a *subtree*.

Spanning tree is a spanning subgraph that is also a tree. The (first) minimal spanning tree (MST) of an edge-weighted graph is a spanning tree that minimizes the sum of the edge weights. The second MST is the spanning tree with minimal total weight that is orthogonal to the MST. The kth MST is the spanning tree with minimum total weight that is orthogonal to the 1st through $k-1$st MSTs. The MST connects very few close points; basing test on 1st through kth can improve power.

Eccentricity of a node: number of edges in a path with greatest length starting at that node. *Antipode* of a node: the node at the other end of a path with greatest length. *Diameter* of a graph: eccentricity of node with largest eccentricity. *Center of a graph*: node with minimal eccentricity.

Rooted tree: one node is designated "root." *Parent* of a node in a rooted tree: next-to-last node in a path from the root to the node. All nodes but the root have parents. *Daughter* of a node: all nodes connected to a node, except the node's parent. *Ancestor* of a node: node on the path that connects it to the root. *Descendant* of a node: all nodes for which that node is an ancestor. *Depth* of a node: depth of root is zero; depth of other nodes is number of edges on path that connects it to the root. *Height* of rooted graph: maximum depth of any node.

For data, think of the edge weight as a measure of dissimilarity, such as Euclidean distance. Every pair of data connected by an edge.

Two important facts:
(i) The MST has as a subgraph the "nearest neighbor graph" that links each point to its closest point. (ii) If an edge is deleted from a MST, creates two disjoint subgraphs.
The deleted edge has the smallest weight of all edges that could connect those two subgraphs.

Univariate runs test: Form MST of data. Delete all edges that connect points of different groups. Count disjoint subgraphs that remain. Those disjoint subgraphs are *runs*.

Multivariate analog: identical.

Null distribution of R

Recall that n is number of data in one group, m is the number in the other, and $N = n + m$. Complete edge-weighted graph for the pooled data has $\binom{N}{2=N(N-1)/2}$ edges. If there are no ties among the $N(N-)/2$ distances, the kth MST is unique.

MST has $N - 1$ edges connecting the N nodes.

Number those edges arbitrarily from 1 to $N-1$. Let $Z_j = 1$ if the jth edge connects nodes from different samples; $Z_j = 0$ if not.

The number of runs is $R \equiv 1 + \sum_{j=1}^{N-1} Z_j$.

Null distribution of R, contd.

What's the chance the jth edge connects nodes from different samples? Under the null, the points that the edge connects are a random sample of size 2 from the N points; the number of X points in that sample is hypergeometric. The chance that sample contains one X and one Y is

$$\Pr\{Z_j = 1\} = \frac{\binom{n}{1}\binom{m}{1}}{\binom{N}{2}} = \frac{2nm}{N(N-1)}. \tag{113}$$

Hence, under the null,

$$\mathbb{E}R = \mathbb{E}(1 + \sum_{j=1}^{N-1} Z_j) = 1 + \frac{2nm}{N}, \tag{114}$$

just as we found for the univariate runs test. Calculating the variance is a bit harder; the covariance of Z_i and Z_j depends on whether they share a node.

In any event, can condition on the edge weights and simulate the distribution of R under the null by making random permutations.

Combining orthogonal MSTs

As N increases, MST includes a smaller and smaller fraction of the $N(N-1)/2$ edges in the complete graph. Many "close" pairs of points are not linked by the MST: potential loss of power.

Proposal: Use all edges contained in the first k MSTs. Can approximate null distribution of R by simulation—random permutations.

MST analogue of the Smirnov test: ranking by MST

Root the MST at a node with maximum eccentricity. Height-directed pre-order (HDP) traversal of the tree: Recursive definition.

1. visit the root

2. HDP in ascending order of height the subtrees rooted at the daughters of the root. Break height ties by starting with the daughters closest (in edge weight) to the root.

"Rank" the nodes in the order visited by the HDP. Apply the Smirnov test to those ranks.

Univariate analogue is the sorted list of data: standard Smirnov test. Like standard Smirnov test, not very sensitive to differences in scale.

Can get analogue of the Siegel-Tukey by rooting at a center node and ranking according to node depth.

Tests based 2 by 2 contingency tables

Idea: label the nodes in some (binary) way that does not depend on the sample they come from. Under the null, the number of X nodes with that label should be hypergeometric.

MST connects the nodes. Each node has some degree. Proposed test statistic: number of X nodes with degree 1. Tends to be large if the X nodes are on the "periphery" of the tree.

Why not something like the sum of the degrees of the X nodes?

Tests based on inter-point distances

Hall and Tajvidi [**?**], Baringhaus and Franz [**?**].

Hall an Tajvidi use "distance" measure $D(x, y)$ on the outcome space $\mathcal{X}$. D is nonnegative and symmetric but doesn't necessarily satisfy triangle inequality.

Let Z be the pooled sample.

For $1 \leq i \leq m$, let $N_i(j)$ be the number of Xs among the set of Zs for which $D(Y_i, Z_k)$ is no larger than its jth smallest value for all $m + n - 1$ of the Zs other than Y_i.

Define $M_i(j)$ analogously. Under the null,

$$\mathbb{E}\{N_i(j)|\{Z_j\}\} = \frac{nj}{m+n-1} \quad \text{and} \quad \mathbb{E}\{M_i(j)|\{Z_j\}\} = \frac{mj}{m+n-1}. \tag{115}$$

Test statistic combines $|M_i(j) - \mathbb{E}M_i(j)|$ and $|N_i(j) - \mathbb{E}N_i(j)|$,
for instance a weighted average of powers of them:

$$T = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |M_i(j) - \mathbb{E}M_i(j)|^{\gamma} w_1(j) + \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} |N_i(j) - \mathbb{E}N_i(j)|^{\gamma} w_2(j). \tag{116}$$

Generalizing KS distance

Probability measures P and Q defined on common sigma-algebra $\mathcal{A}$.

Class $\mathcal{V}$ of sets, elements of $\mathcal{A}$.

$$\delta_{\mathcal{V}}(P,Q) \equiv \sup_{V \in \mathcal{V}} |P(V) - Q(V)|. \tag{117}$$

Class $\mathcal{F}$ of $\mathcal{A}$-measurable functions.

$$\delta_{\mathcal{F}}(P,Q) \equiv \sup_{f \in \mathcal{F}} |\mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{X \sim Q}(f(X))|. \tag{118}$$

Special case: $\mathcal{F}$ consists of indicator functions of elements of $\mathcal{V}$.

If $\mathcal{X} = \mathbb{R}$ and $\mathcal{F}$ is the set of functions with variation bounded by 1, $\delta_{\mathcal{F}}$ is the KS distance.

δ is a pseudo-metric on probability distributions on $\mathcal{X}$

Need to show that it is positive semi-definite and satisfies the triangle inequality.

Positive semi-definiteness and symmetry are immediate (from absolute value).

Triangle inequality: For probabilities P, Q, S,

$$
\begin{aligned}
\delta_{\mathcal{F}}(P,Q)+\delta_{\mathcal{F}}(Q,S) &= \sup_{f\in\mathcal{F}}|\mathbb{E}_{X\sim P}f(X)-\mathbb{E}_{X\sim Q}(f(X))| \\
&\quad +\sup_{g\in\mathcal{F}}|\mathbb{E}_{X\sim Q}f(X)-\mathbb{E}_{X\sim S}(f(X))| \\
&\leq \sup_{f\in\mathcal{F}}\Big(|\mathbb{E}_{X\sim P}f(X)-\mathbb{E}_{X\sim Q}(f(X))| \\
&\quad +|\mathbb{E}_{X\sim Q}f(X)-\mathbb{E}_{X\sim S}(f(X))|\Big) \\
&\leq \sup_{f\in\mathcal{F}}|\mathbb{E}_{X\sim P}f(X)-\mathbb{E}_{X\sim S}(f(X))| \\
&= \delta_{\mathcal{F}}(P,S). \qquad (119)
\end{aligned}
$$

δ is a metric for suitable $\mathcal{F}$

The *cone* or *star* of a subset $\mathcal{F}$ of a linear vector space is the set $\{\alpha f : f \in \mathcal{F}, \alpha \in [0, 1)\}$.

If the cone of $\mathcal{F}$ is dense in $C(\mathcal{X})$ in $L_\infty(\mathcal{X})$, then $\delta_\mathcal{F}$ is a metric on probability distributions. (Need to show that $\delta_\mathcal{F}(P, Q) = 0$ iff $P = Q$.)

(The cone generated by the rational numbers between -1 and 1 is dense in the reals in absolute-value distance, for example.)

Finding δ is easy if $\mathcal{F}$ is the unit RKHS ball

In a RKHS $\mathcal{H}$. Let the point-evaluator at x be k_x, so $f(x) = \langle f, k_x \rangle$. The kernel is $K(x,y) = \langle k_x, k_y \rangle$, which is k_x viewed as a function of x and of its argument.

Let

$$\mu_P \equiv \mathbb{E}_{X \sim P} k_X = \mathbb{E}_{X \sim P} K(X, y). \tag{120}$$

Then

$$\mathbb{E}_{X \sim P} f(X) = \mathbb{E}_{X \sim P} \langle k_x, f \rangle = \langle \mu_P, f \rangle. \tag{121}$$

(The function f is deterministic; we are averaging its values when its argument X is drawn at random according to P.)

Finding δ, contd. We are taking $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$.

$$\begin{aligned} \delta_{\mathcal{F}}(P, Q) &= \sup_{f:\|f\|_{\mathcal{H}} \leq 1} |\langle \mu_P, f \rangle - \langle \mu_Q, f \rangle| \\ &= \sup_{f:\|f\|_{\mathcal{H}} \leq 1} \langle \mu_P - \mu_Q, f \rangle \\ &= \|\mu_P - \mu_Q\|. \end{aligned} \tag{122}$$

The last step follows from the equivalence of the operator norm and the Hilbert-space norm, since $\mathcal{H}$ is reflexive (self-dual).

Finding δ, contd. Maximizing a continuous function over the closed unit ball in $\mathcal{H}$:

$$\begin{aligned} f^*(y) \equiv \arg \max_{f:\|f\|_{\mathcal{H}} \leq 1} \langle \mu_P - \mu_Q, f \rangle &= (\mu_P - \mu_Q)/\|\mu_P - \mu_Q\| \\ &= \frac{\mathbb{E}_{X \sim P} K(X, y) - \mathbb{E}_{X \sim Q} K(X, y)}{\|\mathbb{E}_{X \sim P} K(X, y) - \mathbb{E}_{X \sim Q} K(X, y)\|} \end{aligned} \tag{123}$$

The plug-in estimate is

$$\widehat{f}^*(y) \equiv \frac{1}{n} \sum_{j=1}^{n} K(X_j, y) - \frac{1}{m} \sum_{j=1}^{m} K(Y_j, y). \tag{124}$$

It is biased, but I'm not convinced that matters.

Gretton et al. [**?**] propose test statistic: Pick $\mathcal{F}$, a set of functions whose star is dense in $C(\mathcal{X})$ wrt $L_\infty(\mathcal{X})$. They use the unit ball in a universal RKHS.

Test statistic is a variant of $\delta_{\mathcal{F}}(\hat{P}, \hat{Q})$:

$$\mathrm{MMD}(\mathcal{F}, P, Q) = \mathbb{E}_{X,Y\sim P} K(X,Y) - 2\mathbb{E}_{X\sim P, Y\sim Q} K(X,Y) + \mathbb{E}_{X,Y\sim Q} K(X,Y \tag{125}$$

They calibrate by bootstrap. Why not calibrate by permutation?

Tests and Confidence Sets for percentiles of a continuous distribution

Test statistic?

Under what group is the distribution of the the test statistic invariant if the null is true?

How can we generate the orbit of the data under that group? How big is the orbit?

How can we sample at random from the orbit?

Abstract permutation tests

See old lecture notes, chapter 8.

Earthquake catalog declustering: see PSU talk.

Nonparametric Combinations of Tests (NPC)

Setting: multidimensional data, typically mixed types (some variables continuous, some [possibly ordered] categorical , etc.)

Use $k > 1$ partial tests.

Notation will follow [**?**]. Have C samples of V-dimensional data (each sample might correspond to a different treatment, e.g.). Sample space for each observation is $\mathcal{X}$. Samples may have unequal sizes.

NPC: Notation

Data

$$\mathbf{X} = \{\mathbf{X}_j\}_{j=1}^{C} = \{\mathbf{X}_{ji}\}_{i=1}^{n_j}{}_{j=1}^{C} = \{X_{hji}\}_{i=1}^{n_j}{}_{j=1}^{C}{}_{h=1}^{V}. \quad (126)$$

$$n \equiv \sum_{j=1}^{C} n_j. \quad (127)$$

For each j, $\{X_{ji}\}_{i=1}^{n_j}$ are iid $P_j \in \mathcal{P}$. Known class $\mathcal{P}$ of distributions on a common sigma algebra on $\mathcal{X}$.

Null hypothesis H_0: P_j are all equal.

Then have exchangeability w.r.t. C groups.

NPC: Notation, contd.

H_0 might be broken into sub-hypotheses $\{H_{0i}\}_{i=1}^k$, such that

$$H_0 = \cap_{i=1}^k H_{0i}. \tag{128}$$

H_0 is the overall null.

Alternative can be broken into

$$H_1 = \cup_{i=1}^k H_{1k}. \tag{129}$$

H_1 is overall alternative.

k-dimensional vector of test statistics $\mathbf{T}(\mathbf{X})$

Randomized experiment

There are N subjects.

The subjects are given; they are not necessarily a sample from some larger population.

Assign a simple random sample of size n of the N subjects to treatment, and the remaining $m = N - n$ subjects to control.

For each subject, we observe a (univariate) quantitative response. (More on multivariate responses later.)

No assumption about the values of that quantitative response; they need not follow any particular distribution.

The null hypothesis is that treatment "makes no difference."

What's the relevant group?

Under what group is the distribution of the data invariant if the null is true?

How can we generate the orbit of the data under that group? How big is the orbit?

How can we sample at random from the orbit?

Alternative hypotheses

Many alternatives are interesting. The most common are the *shift alternative*, the *dispersion alternative*, and the *omnibus alternative*.

The shift alternative is that treatment changes the mean response. (There are left-sided, right-sided and two-sided versions of the shift alternative.)

The dispersion alternative is that treatment changes the scatter of the responses.

The omnibus alternative is that treatment changes the response in some way—any way whatsoever.

Testing

The deliberate randomization makes it possible to test rigorously whether whether treatment affects response in the group of N subjects.

Up to sampling error—which can be quantified—differences between the responses of the treatment and control groups must be due to the effect of treatment: randomization tends to balance other factors that affect the responses, and that otherwise would lead to confounding.

However, conclusions about the effect of treatment among the N subjects cannot be extrapolated to any other population, because we do not know where the subjects came from (how they came to be part of the experiment).

The Neyman Model

We model the experiment as follows: Each of the N subjects is represented by a ticket with two numbers on it, a left and a right number.

The left number is the response the subject would have if assigned to the control group; the right number is the response the subject would have if assigned to the treatment group.

These numbers are written on the tickets before the experiment starts. Assigning the subject to treatment or control only determines whether we observe the left or the right number for that subject.

Let u_j be the left number on the jth ticket and let v_j be the right number on the jth ticket.

The experiment reveals either u_j (if subject j is assigned to treatment) or v_j (if subject j is assigned to control).

Non-interference

There are only two numbers on each ticket.

Whether u_j or v_j is revealed depends only on whether subject j is assigned to treatment or to control.

Let X_j be the indicator of whether subject j was treated. That is, $X_j = 0$ if subject j is in the control group, and $X_j = 1$ if subject j is in the treatment group.

If $X_j = 0$, u_j is revealed. If $X_j = 1$, v_j is revealed.

More generally, the observed response of subject j could depend on *all* the assignments $\{X_j\}_{j=1}^n$.

In the Neyman model, it depends only on X_j. This is the hypothesis of *non-interference*.

Strong Null Hypotheses

The strong null hypothesis is

$$u_j = v_j, \;\; j = 1, 2, \ldots, N. \tag{130}$$

That is, the strong null hypothesis is that the left and right numbers on each ticket are equal. Subject by subject, treatment makes no difference at all. The N observed responses will be the same, no matter which subjects are assigned to treatment.

Weak Null Hypotheses

The weak null hypothesis is that the average of the left numbers equals the average of the right numbers:

$$\sum_{j=1}^{N} u_j = \sum_{j=1}^{N} v_j. \tag{131}$$

In the weak null hypothesis, treatment makes no difference *on average*: treatment might increase the responses for some individuals, provided it decreases the responses of other individuals by a balancing amount.

Extending the Neyman Model to random responses

Generalize ticket model from two responses (one under treatment, the other under control) to two random variables.

Instead of thinking of subject j's (potential) responses as a fixed pair of numbers $\{(u_j, v_j)\}_{j=1}^{N}$, think of them as a pair of random variables (U_j, V_j).

A realization of U_j will be observed if subject j is assigned to control (i.e., if $X_j = 0$)

A realization of V_j will be observed if subject j is assigned to treatment (i.e., if $X_j = 1$)

The joint probability distributions of U_j and V_j are fixed before the randomization (but are unknown).

Assumption of no confounding

$\{X_j\}$ are in general dependent (e.g., if m subjects are selected for treatment at random).

But $\{X_j\}$ don't depend on characteristics of the subjects (including $\{(U_j, V_j)\}$).

The set $\{X_j\}_{j=1}^n$ is independent of the set $\{(U_j, V_j)\}_{j=1}^n$.

Assumption of Non-Interference

The random pairs

$$\{(U_j, V_j)\}_{j=1}^{N} \tag{132}$$

are independent across j. That is, if we knew the values of any subset of the pairs, it wouldn't tell us anything about the values of the rest of the pairs.

(Within each pair, U_j and V_j can be dependent.)

Nothing about subject j's potential responses depends on any other subjects.

Fine points

Think of the realizations of $\{(U_j, V_j)\}$ as being generated before $\{X_j\}$ are generated.

Once $\{(U_j, V_j)\}$ are generated, they are intrinsic properties of subject j.

Null hypotheses

- For each j, $U_j \sim V_j$
- For each j, $\mathbb{E}U_j = \mathbb{E}V_j$
- $\mathbb{E}\sum_{j=1}^{N} U_j = \mathbb{E}\sum_{j=1}^{N} V_j$
- qth quantile of U_j equals qth quantile of V_j (e.g., for survival time)
- what else?

Special case: Binary responses

{die, live}, {don't improve, improve}, {not harmed, harmed}.

Code responses as {0, 1}.

Then subject j is characterized by 4 numbers that specify joint distribution of (U_j, V_j).

$$\begin{aligned} p_j &= \Pr\{U_j = 0 \text{ and } V_j = 0\} \\ q_j &= \Pr\{U_j = 0 \text{ and } V_j = 1\} \\ r_j &= \Pr\{U_j = 1 \text{ and } V_j = 0\} \\ s_j &= \Pr\{U_j = 1 \text{ and } V_j = 1\} \end{aligned} \tag{133}$$

$p_j, q_j, r_j, s_j \geq 0$ and $p_j + q_j + r_j + s_j = 1$.

Causal inference

Think of treatment as exposure to something that might cause harm (e.g., a potential carcinogen).

A response of 1 means the subject was "harmed" (e.g., got cancer). A response of 0 means the subject was not harmed.

E.g., if $X_j = 1$ and $V_j = 1$, subject j was exposed and suffered harm. If $X_j = 0$ and $U_j = 1$, subject j was not exposed, but suffered harm.

In general, does exposure cause harm?

Would exposure cause harm to subject j?

If $X_j = 1$ and $V_j = 1$, did exposure harm subject j?

Relative risk

Generally, data are from epidemiological studies, not randomized experiments: confounding is a major concern.

Pretend we have controlled, randomized, double-blind experiment.

Recall that $n = \sum_j X_j$ is the number of exposed (treated) subjects and $m = N - n = \sum_j (1 - X_j)$ is the number of unexposed (control) subjects.

Common test for causation is based on relative risk (RR):

$$\text{RR} = \frac{\frac{1}{n}\sum X_j V_j}{\frac{1}{m}\sum (1 - X_j) U_j}. \tag{134}$$

Numerator is the rate of harm in the treated group; denominator is the rate of harm in the control group.

General and specific causation in the law

Plaintiff was exposed to something, was harmed, and sues the party responsible for his exposure.

To win, plaintiff must show by a preponderance of the evidence (i.e., that more likely than not):

General causation: exposure can cause harm

Specific causation: the exposure caused that individual's harm

Relative risk and specific causation

Common legal test for specific causation: RR > 2.

Claim: if RR > 2, "more likely than not" the exposure caused plaintiff's harm.

Thought experiment

2000 subjects. 1000 subjects selected at random and exposed; 1000 left unexposed.

Among unexposed, 2 are harmed. Among exposed, 20 are harmed. Then RR = 10.

Heuristic argument: but for the exposure, there would only have been 2 harmed among the exposed, so 18 of the 20 injuries were caused by the exposure.

Is the plaintiff selected at random?

Pick a subject at random from the 20 who were exposed and harmed.

$$\begin{aligned}\Pr\{\text{ that subject's harm was caused by exposure }\} &= 18/20 \\ &= 90\% \\ &= 1 - \frac{1}{\text{RR}} \\ &> 50\%.\end{aligned}$$

If the RR had been 4, the chance would have been 75%.

If RR had been 2, the chance would have been 50%: the threshold for "more likely than not."

Hypothetical Counterfactuals and Causation

Suppose $X_j = 1$ and $V_j = 1$ (subject j was exposed and harmed).

Exposure caused the harm if $U_j = 0$: But for the exposure, subject j would not have been harmed.

Involves counterfactual: Subject j was in fact exposed!

(If $U_j = 1$, subject would have been harmed whether or not the exposure happened: The exposure did not cause the harm.)

Probability that exposure causes harm

Chance subject j would be harmed if unexposed

$$\beta_j = \Pr\{U_j = 1\} = r_j + s_j. \tag{135}$$

Chance subject j would be harmed if exposed

$$\gamma_j = \Pr\{V_j = 1\} = q_j + s_j. \tag{136}$$

β_j and γ_j are identifiable, but (p_j, q_j, r_j, s_j) are not separately identifiable.

Even if β_j and γ_j are known, cannot determine q_j.

Overall rate of harm caused by exposure

Overall expected rate of harm if no subjects were exposed

$$\beta = \frac{1}{N}\sum_{j=1}^{N} \Pr\{U_j = 1\} = \frac{1}{N}\sum_{j=1}^{N}(r_j + s_j). \tag{137}$$

Overall expected rate of harm if all subjects were exposed

$$\gamma = \frac{1}{N}\sum_{j=1}^{N} \Pr\{V_j = 1\} = \frac{1}{N}\sum_{j=1}^{N}(q_j + s_j). \tag{138}$$

Difference in expected rate of harm if all were exposed versus if none were exposed

$$\gamma - \beta. \tag{139}$$

This is average causal effect of exposure on harm (among the N subjects in the study group).

Estimating rate of harm caused by exposure

γ, β, $\gamma - \beta$ are estimable:

$$\gamma = \mathbb{E}\frac{1}{n}\sum_{j=1}^{N} X_j V_j \tag{140}$$

$$\beta = \frac{1}{m}\sum_{j=1}^{N}(1 - X_j)U_j. \tag{141}$$

Focus on the algebra, not the statistics

Take γ and β to be known, with

$$0 < \beta < \gamma \tag{142}$$

(so exposure does, on average, cause harm: RR $= \gamma/\beta > 1$) and

$$\beta + \gamma < 1 \tag{143}$$

(the rate of harm, even with exposure, is not too large).

Probability of specific causation

Exposure and response are independent, so

$$\Pr\{U_j = 0 | V_j = 1, X_j = 1\} = \Pr\{U_j = 0 | V_j = 1\}. \quad (144)$$

Conditional chance that exposure caused subject j's harm is

$$\pi_j = \Pr\{U_j = 0 | V_j = 1\} = q_j/\gamma_j = q_j/(q_j + s_j). \quad (145)$$

Define $\pi_j = 0$ if $\gamma_j = 0$.

π_j is conditional probability that subject j would not have been harmed if left unexposed, given that subject j was exposed and injured.

Probability of specific causation, cont.

"More likely than not," exposure caused the harm if $\pi_j > 1/2$.

Since q_j is not identifiable, π_j is not identifiable. Cannot answer the question using epidemiological data.

To estimate π_j requires additional assumptions; generally not testable.

Average probability of causation: helpful lemmas

$$\bar{\pi} \equiv \frac{1}{N} \sum_{j=1}^{N} \pi_j, \tag{146}$$

$$\bar{q} \equiv \frac{1}{N} \sum_{j=1}^{N} q_j. \tag{147}$$

$$\bar{q} \leq \frac{1}{N} \sum_{j=1}^{N} (q_j + s_j) = \gamma; \quad \bar{q} = \gamma \quad \text{iff} \quad s_j = 0 \ \ \forall j. \tag{148}$$

$$\bar{q} \geq \gamma - \beta; \quad \bar{q} = \gamma - \beta \quad \text{iff} \quad r_j = 0 \ \ \forall j. \tag{149}$$

(Proof: $\bar{q} = \gamma - \beta + \bar{r}$.)

Average probability of causation: lower bound

$$\inf \bar{\pi} = \gamma - \beta. \tag{150}$$

Proof.

$$\bar{\pi} = \frac{1}{N}\sum_{j=1}^{N} \pi_j \geq \frac{1}{N}\sum_{j=1}^{N} (q_j + s_j)\pi_j = \frac{1}{N}\sum_{j=1}^{N} q_j = \bar{q} \geq \gamma - \beta. \tag{151}$$

Equality holds if $r_j = 0$ for all subjects, and $q_j = 0$ unless $q_j + s_j = 1$.

For instance, suppose there are two types of subjects: for one type, exposure does not change the chance of harm, while for the other, there is harm only if the subject is exposed.

Average probability of causation: upper bound

$$\sup \bar{\pi} = 1. \tag{152}$$

Proof. Take $s_j \equiv 0$, $r_j \equiv \beta$, $q_j \equiv \gamma$, and $p_j \equiv 1 - \beta - \gamma$ for all j.

Example: $\pi_j = \bar{\pi} = 1$

$p_j = 0.95$, $q_j = 0.04$, $r_j = 0.01$, $s_j = 0$. Then

$$\pi_j = q_j/(q_j + s_j) = 0.04/(0.04 + 0) = 1, \ \ \forall j. \tag{153}$$

Note that

$$\beta = \bar{r} + \bar{s} = 0.01 < \bar{q} + \bar{s} = \gamma = 0.04 \tag{154}$$

and $\beta + \gamma = 0.05 < 1$.

Example: $\pi_j = \bar{\pi} = 3/4$

$p_j = 0.96$, $q_j = 0.03$, $r_j = 0$, $s_j = 0.01$. Then

$$\pi_j = q_j/(q_j + s_j) = 0.03/(0.03 + 0.01) = 3/4, \ \ \forall j. \quad (155)$$

Note that

$$\beta = \bar{r} + \bar{s} = 0.01 < \bar{q} + \bar{s} = \gamma = 0.04 \quad (156)$$

and $\beta + \gamma = 0.05 < 1$.

Example: $\pi_j = \bar{\pi} = 0.03$

For 97% of subjects, $p_j = 96/97$, $q_j = r_j = 0$, $s_j = 1/97$.

For 3% of subjects, $p_j = 0$, $q_j = 1$, $r_j = s_j = 0$.

For the first group,

$$\pi_j = q_j/(q_j + s_j) = 0/(0 + 1/97) = 0. \tag{157}$$

For the second group,

$$\pi_j = q_j/(q_j + s_j) = 1/(1 + 0) = 1. \tag{158}$$

Hence

$$\bar{\pi} = 0.97 \times 0 + 0.03 \times 1 = 0.03. \tag{159}$$

Again $\beta = 0.01$, $\gamma = 0.04$.

All three examples have the same values of $\bar{p}$, $\bar{q}$, $\bar{r}$, and $\bar{s}$—and hence of β and γ—but the values of $\bar{\pi}$ are 1, 3/4, and 0.03.

Epidemiological studies can (at best) determine β and γ.

The claim that RR > 2 means causation is "more likely than not" ties to example 2.

Probability of specific causation: random selection

Plaintiff was exposed and harmed. What is the chance that exposure caused the harm?

Arguments generally assume plaintiff was chosen at random from the study group. But how?

1. Pick subject at random; condition that subject was exposed and harmed.

2. Divide subjects into 2 groups, exposed and not. Condition that at least one exposed was harmed. Pick one subject at random from the exposed and harmed.

3. Pick a subject at random; condition that subject was exposed and harmed and that subject sues.

In all three, n of N are assigned at random to exposure, exposure X_j is independent of response pair (U_j, V_j), and $\{(U_j, V_j)\}_{j=1}^{N}$ are independent random pairs.

Probability of specific causation: scenario 1

η is a random integer between 1 and N. Condition that $X_\eta = V_\eta = 1$.

Then

$$\Pr\{U_\eta = 0 | X_\eta = 1, V_\eta = 1\} \geq 1 - \frac{1}{\mathrm{RR}}. \qquad (160)$$

Can be as high as 1.

Probability of specific causation: scenario 2

$\mathcal{X}$ are the exposed; $\mathcal{R} \subset \mathcal{X}$ are the exposed and harmed. ρ uniform on $\mathcal{R}$ when $\mathcal{R}$ is nonempty.

Then $U_\rho = 0$ means that exposure caused harm to subject ρ, who was selected at random from the exposed, harmed.

Let $\mathcal{J}$ be a typical nonempty $\mathcal{R}$; $|\mathcal{J}|$ is the cardinality of $\mathcal{J}$, so $1 \leq |\mathcal{J}| \leq n$.

Then

$$\Pr\{U_\rho = 0 | \mathcal{R} \neq \emptyset\} = \frac{\sum_{\mathcal{J}} \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \pi_j \Pr\{\mathcal{R} = \mathcal{J}\}}{\sum_{\mathcal{J}} \Pr\{\mathcal{R} = \mathcal{J}\}} \qquad (161)$$

Comparing scenarios 1 and 2

$N = 3$, $n = 2$.

Scenario 1:

$$(q_1 + q_2 + q_3)/(\gamma_1 + \gamma_2 + \gamma_3). \tag{162}$$

Weighted average of π_1, π_2, π_3 with weights γ_1, γ_2, γ_3.

Scenario 2 also weighted average of π_1, π_2, π_3, but weights are

$$\gamma_1(2-3\gamma/2+\gamma_1/2),\ \gamma_2(2-3\gamma/2+\gamma_2/2),\ \gamma_3((2-3\gamma/2+\gamma_3/2). \tag{163}$$

Scenario 3

Your mileage may vary.

Suppose propensity to sue depends on individual characteristics. E.g., we also have $\{Y_j\}_{j=1}^N$. If $X_j = V_j = Y_j = 1$, subject j was exposed, harmed, and sues. Suppose $q_j + s_j > 0 \ \forall j$.

Suppose

$$\Pr\{Y_j = 1 | X_j = V_j = 1\} = \frac{\lambda}{q_j + s_j}. \tag{164}$$

Healthier subjects are more likely to sue.

Then

$$\Pr\{U_\eta = 0 | X_\eta = V_\eta = Y_\eta = 1\} = \bar{\pi}. \tag{165}$$

Average probability of causation, rather than relative risk, controls things in this scenario.

Bibliography

E.L. Bennett, M.R. Rosenzweig, and M.C. Diamond. Rat brain: Effects of environmental enrichment on wet and dry weights. *Science*, 163:825–826, 1969.

D. Freedman, R. Pisani, R. Purves, and A. Adhikari. *Statistics*. W.W. Norton and Co., New York, 4th edition, 2004.

D.A. Freedman and P.B. Stark. The swine flu vaccine and Guillain-Barré syndrome: a case study in relative risk and specific causation. *Evaluation Review*, 23:619–647, 1999.

E.L. Lehmann. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.

E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, 2nd edition, 1998.

E.L. Lehmann and H.J.M. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks*. McGraww Hill Text, 2nd edition, 1988.

E.J.G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to J. Roy. Statis. Soc.*, 4:119–130, 1937.

E.J.G. Pitman. Significance tests which may be applied to samples from any populations. II. the correlation coefficient test. *Supplement to J. Roy. Statis. Soc.*, 4:225–232, 1937.

E.J.G. Pitman. Significance tests which may be applied to samples from any populations. III. the analysis of variance test. *Supplement to J. Roy. Statis. Soc.*, 4:322–335, 1937.

J.P. Romano. A bootstrap revival of some nonparametric distance tests. *J. Am. Stat. Assoc.*, 83:698–708, 1988.

J.P. Romano. On the behavior of randomization tests without a group invariance assumption. *J. Am. Stat. Assoc.*, 85:686–692, 1990.

M.R. Rosenzweig, E.L. Bennett, and M.C. Diamond. Brain changes in response to experience. *Scientific American*, 226:22–29, 1972.