# Nonparametrics: nonpareil?

Neuropsychology Brown Bag Lunch
EBIRE
Martinez, CA
15 May 2007

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

## 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 Rosenzweig, M. R., Bennett, E. L., & Diamond, M. C. (1972). Brain changes in response to experience., 1972. *Sci. Am., 226*, 22–29; See also Bennett, Rosenzweig and Diamond, 1969. Rat Brain: Effects of Environmental Enrichment on Wet and Dry Weights, *Science, 163*, 825–826; Freedman, Pisani and Purves, 2007. *Statistics.* W.W. Norton, 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 | $\pm 32$ | $\pm 33$ | $\pm 16$ | $\pm 6$ | $\pm 21$ | $\pm 17$ | $\pm 64$ | $\pm 7$ | $\pm 89$ | $\pm 2$ | $\pm 11$ |

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.

## 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, 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, 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?

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 $\chi^2$ 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.

## Quick overview of nonparametrics, robustness

### Parameters: related notions

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

- 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.

## References


Divenyi, P., P.B. Stark, and K. Haupt, 2005. Decline of Speech Understanding and Auditory Thresholds in the Elderly, *J. Acoustical Soc. Am., 118*, 1089–1100.

Freedman, D.A., R. Pisani and R. Purves, 2007. *Statistics, 4th ed.*, W.W. Norton, NY.

Lehmann, E.L., 1998. *Nonparametrics: Statistical Methods based on Ranks,* Prentice-Hall, NJ.

Rosenzweig, M. R., E.L. Bennett and M.C. Diamond, M. C., 1972. Brain changes in response to experience. *Scientific American, 226*, 22–29.

Bennett, E.L., M.R. Rosenzweig and M.C. Diamond, 1969. Rat Brain: Effects of Environmental Enrichment on Wet and Dry Weights, *Science, 163*, 825–826.

Stark, P.B., 2000-2006. Lecture notes for Statistics 240.

`www.stat.berkeley.edu/~stark/Teach/S240/Notes`.