Tests for Mixed Paired and Two-Sample Designs/Classification

My Research I: New Statistics

- Develop new and improved statistical methodology
  - Nonparametric and Robust Methods

Nonparametric and Robust Methods

- e.g., t-test
  - Mathematically exact if data come from normal distribution
  - Otherwise, other methods can work better
    - I try to develop such nethods

My Research II: Interdisciplinary Research

- Develop new statistical methodology
- Novel application of existing methodology

Two Topics Today

- New Statistics: *Tests for Mixed Paired and Two-Sample Designs*
- Interdisciplinary: *Classification for zooarchaeology*

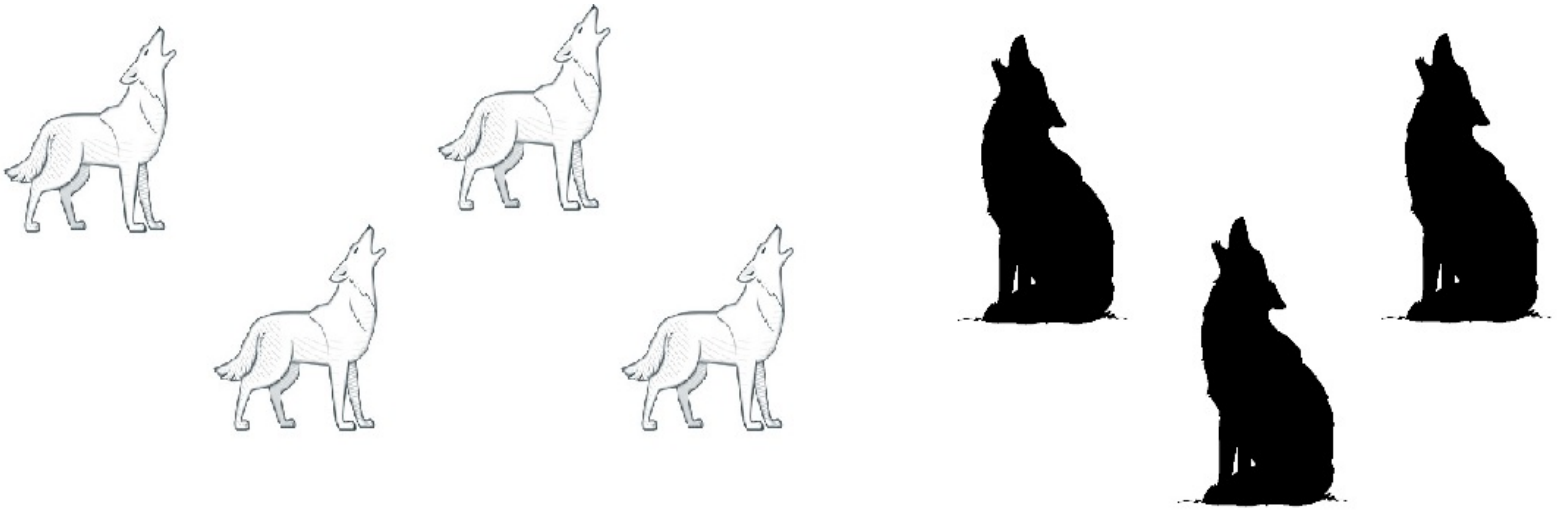
Mixed Paired and Two-Sample Designs

Coyotes.

- It was desired to compare two methods (QIAGEN DNeasy Kit, traditional chloroform isoamyl alcohol method) for extracting DNA from coyote blood samples.
- The response variable (Y) was mean concentration of DNA.
- A total of 30 coyotes were available for the study.
- Ideally, both methods would be used on each coyote
  - Reduced variability, as differences between treatments would be on the same subject
  - Fewer subject (coyotes) required

Mixed Paired and Two-Sample Designs-Coyotes

Due to constraints, however, both methods were used on only 6 coyotes (randomly selected). The kit was randomly assigned to be used for 8 of the remaining coyotes and the traditional method for the remaining 16 coyotes.



Mixed Paired and Two-Sample Designs-Laser Eye Surgery

Clinical trial to compare two methods of laser eye surgery.

- Patients with both eyes eligible have one eye randomly assigned to each treatment (dependent samples)
- Patients with only one eye eligible will each have one eye randomly assigned to one treatment (independent samples)

Analysis of Paired Data: Paired t-test

Under the assumptions

- The treatment response differences for the 6 coyotes receiving both treatments are independent
- The response differences for the population of all coyotes are normally distributed

The statistic

$$t_{paired} = \frac{\bar{d}}{s/\sqrt{n}} \backsim t(n-1)$$

where

- $n = 6$ is the number of paired data points
- $\bar{d} = \bar{x} - \bar{y}$ is the mean of the paired differences
- s is the standard deviation of the paired differences
- $t(n-1)$ is the Student's t-distribution with $n-1$ degrees of freedom

Analysis of Unpaired Data: Pooled t-test

Under the assumptions

- The responses for the 24 "unpaired" coyotes are independent
- The populations of responses are both normally distributed
- The populations of treated and untreated responses have the same variance

The statistic

$$t_{unpaired} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \backsim t(n_1 + n_2 - 2)$$

where

- $n_1 = 8, n_2 = 16$ are the sizes of the unpaired samples
- $\bar{X} - \bar{Y}$ is the treatment mean difference for the unpaired data
- s_1, s_2 are the standard deviations of the unpaired samples
- $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ is the pooled variance
- $t(n_1 + n_2 - 2)$ is the **Student's *t*-distribution** with $n_1 + n_2 - 2$ degrees of freedom

Nonparametric Alternative: Randomization/Permutation tests

Idea: Compute p-value based on the results of all hypothetical experiments

- **Assume there is no treatment effect**
- "Rerandomize (permute)" the observed data and compute the resulting test statistic
- Repeat for all possible rerandomizations
- Compare the test statistic of the actual randomization to the set all possible results
  - If the result looks "unusual", it is likely because the initial assumption is wrong, and there is a treatment effect

Nonparametric Alternative: Rank-based tests

Rank-based analogues to t_{paired} and $t_{unpaired}$ tests:

- Wilcoxon signed-ranks test (W_{paired})
- Wilcoxon rank-sum/Mann-Whitney test ($W_{unpaired}$)
-
- Rank-based methods are resistant to outliers

Can the separate tests be combined to use all of the data in a single test?

Consider the statistic

$$t_{combined} = \lambda t_{paired} + (1 - \lambda) t_{unpaired}$$

Note that $t_{combined} = t_{paired}$ when $\lambda = 1$ and $t_{combined} = t_{unpaired}$ when $\lambda = 0$.

- Much of the previous research has been focused on trying to approximate the distribution of $t_{combined}$.

Previous Research: Combined t-tests

- Lin & Stivers (1974)
- Bhoj (1978)
- Bhoj (1989)

All of these assumed normally distributed data and and derived approximate t-distributions.

Previous Research: Randomization tests

- Einsporn & Habtzghi (2013).
- Investigated randomization/permutation version of Bhoj (1978) test.

Previous Research: Nonparametric tests

- Dubnicka, et. al. (2002); Magel & Fu (2014).
- Proposed combined test based on Wilcoxon statistics.
- Johnson (2018). Proposed permutation versions of Dubnicka and Magel & Fu tests; performed extensive simulation study to compare various tests.
- Wang (2020(REU)). Investigated confidence intervals based on combined statistics.

Comparison of Approaches

- **None of the tests has been shown to be superior under all conditions**
- Bhoj (1989) method best for normally distributed data
- Einsporn & Habtzghi (2013) method best for nonnormal, "light-tailed" data
- Rank methods best for nonnormal "heavy-tailed" data
  - **Neither of the rank tests was best in all situations**


AND NOW FOR
SOMETHING
COMPLETELY
DIFFERENT

Zooarchaeology

- Zooarchaeology: studies the remains of animals from archaeological sites

Problem

Predict the percent of human vs. carnivore presence based on bone samples from a site

Solution?

- Compiled database of bone characteristics from experiments
  - Human only
  - Carnivore only
- Create a classification algorithm from these data to predict percent of human vs. carnivore presence for a sample from a site with unknown percents

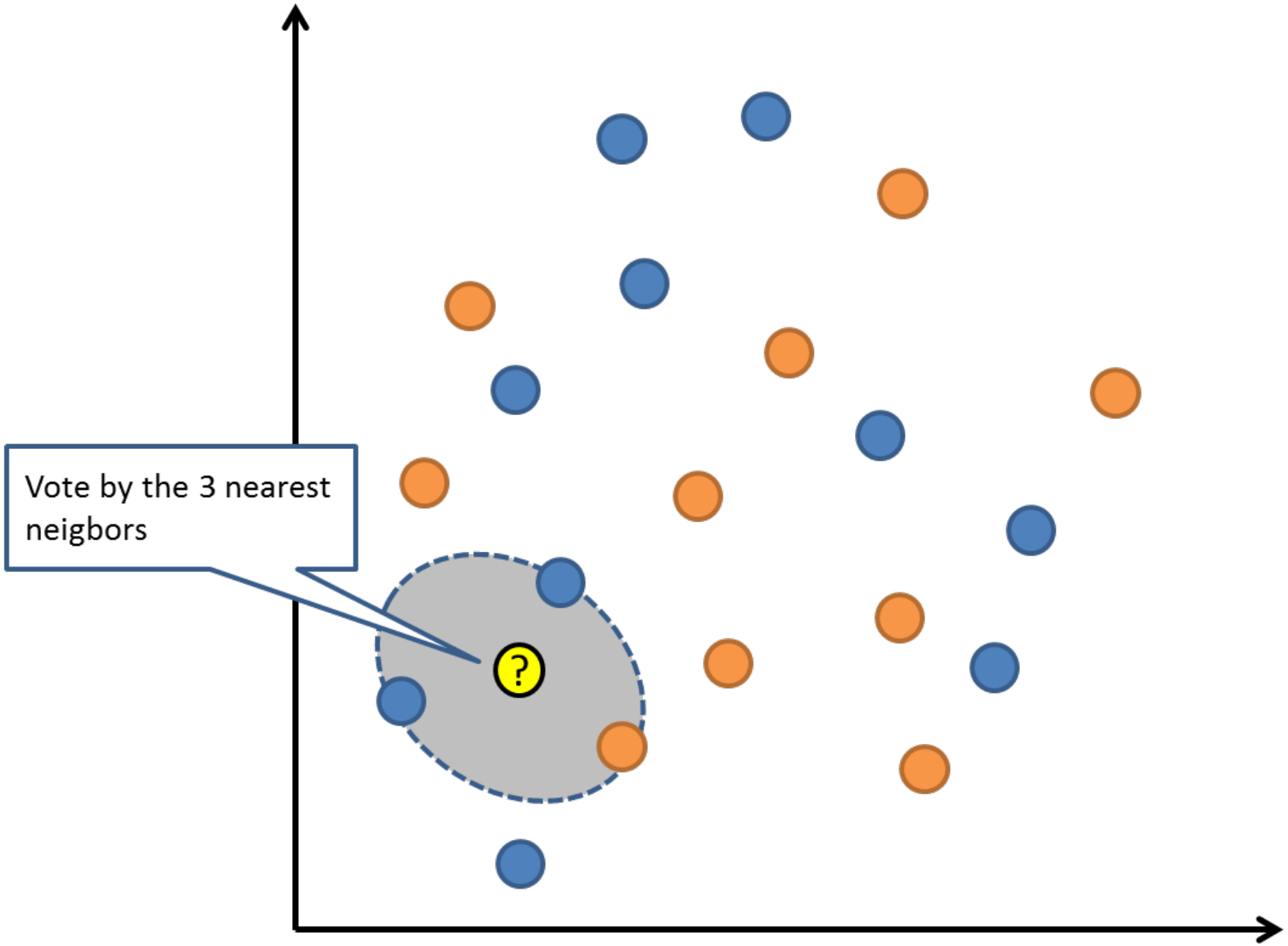
The Classification Problem

- Outcome: Known group label (e.g., human/carnivore)
- Features/Predictors: Characteristics of individuals in groups
-
- Want to be able to "predict" Group membership based on Features

Example: Spam emails

- Email service compiles data on emails reported as spam vs. OK
- Uses these data to "predict" whether an email is spam

Example algorithm: k-Nearest Neighbor Classifier



References

- Bhoj, D.S. (1978). Testing equality of means of correlated variates with missing data on both responses. *Biometrika*, 65:225-228.
- Bhoj, D.S. (1989). On comparing correlated means in the presence of incomplete data. *Biometrical Journal*, 31:279–288.
- Dubnicka, S.R., Blair, R.C., Hettmansperger, T.P. (2002). Rank-based procedures for mixed paired and two-sample designs. *Journal of Modern Applied Statistical Methods*, 1:32-41.
- Einsporn, R.L., Habtzghi, D. (2013). Combining paired and two-sample data using a permutation test. *Journal of Data Science*, 11:767-779.
- Johnson, E. N. (2018). Permutation Tests for Mixed Paired and Two-Sample Designs.[Unpublished masters' thesis]. University of North Carolina, Greensboro.

References

- Lin and Stivers (1974)
- Magel, R.C., Fu, R. (2014). Proposed nonparametric test for the mixed two-sample design. *Journal of Statistical Theory and Practice*, 8:221-237.