Validity

February 22, 2017; due Sunday, March 5, 2017 (11:59pm)

1 (everyone)

Establishing clear criteria by which an analysis is to be judged *valid* is one of the most important tasks facing you in data analysis. Pick any of the academic papers assigned throughout this course (i.e., any text except ML and NCM) and discuss the ways in which it establishes (or fails to establish) the nine types of validity outlined in Krippendorff (2004):

- Face validity
- Social validity
- Sampling validity
- Semantic validity
- Structural validity
- Functional validity
- Convergence validity
- Discriminant validity
- Predictive validity

Deliverable: one-page paper (single-spaced).

2 (choose either 2.1 or 2.2)

2.1 Implementation

The permutation test is a robust hypothesis test that doesn't require some of the assumptions of classical tests. Remember that hypothesis tests measure the likelihood of an observed test statistic under a null hypothesis; for example, let X denote a set of n_x items with mean $mean(X)$ and Y denote a set of n_y items with mean $mean(Y)$; if we want to measure the degree to which X and Y are different, we can ask how likely is the statistic $mean(X) - mean(Y)$ under the null hypothesis that $mean(X) = mean(Y)$.

In a permutation test, we begin by calculating an observed test statistic of interest t. In table 1, this would correspond to the observed $mean(X) - mean(Y)$ for the *true labels* (-3.86); this is the statistic whose probability we would like to assess under the null hypothesis. In a permutation test, we calculate the null hypothesis by randomly shuffling the labels (X and Y) for the dataset (this keeps the total number of items with label X and Y the same); under the null hypothesis that $mean(X) = mean(Y)$, it should not matter whether a given data point has label X or Y (because the null assumption is that they all come from the same underlying distribution).

A two-tailed permutation test establishes the probability of seeing a test statistic as extreme as t under the null by counting the number of times t is exceeded by the same statistic t_b calculated from a permuted dataset, in B trials:

$$\hat{p} = \frac{1}{B} \sum_{b=1}^{B} I[abs(t) \leq abs(t_b)] \tag{1}$$

$$\text{where } I[z] = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{if } z \text{ is false} \end{cases} \tag{2}$$

This counts the fraction of times we see a test statistic *more* extreme, simply by chance alone. In the example in table 1, $abs(t) \leq abs(t_b)$ in only one permutation ($perm_5$), leading to a $\hat{p}$-value of $\frac{1}{5} = 0.20$. A true permutation

	value	true labels	$perm_1$	$perm_2$	$perm_3$	$perm_4$	$perm_5$
x_1	6.7	X	X	Y	X	X	X
x_2	4.1	X	Y	X	Y	Y	X
x_3	6.0	X	X	Y	Y	Y	X
x_4	7.1	X	Y	X	X	Y	Y
x_5	3.9	X	Y	Y	X	X	X
x_6	7.8	Y	X	Y	Y	X	Y
x_7	10.2	Y	X	X	Y	Y	Y
x_8	6.7	Y	Y	Y	Y	X	X
x_9	13.1	Y	Y	X	X	Y	Y
x_{10}	9.3	Y	X	X	X	X	Y
mean(X)		5.56	8.0	8.76	8.02	6.88	5.48
mean(Y)		9.42	6.98	6.22	6.96	8.1	9.5
true difference		-3.86					
permuted difference			1.02	2.54	1.06	-1.22	-4.02
abs(true difference) $\leq$ abs(permuted difference)			False	False	False	False	True

Table 1: Permuted labels. The "true difference" is mean(X)-mean(Y) for the true labels (-3.86)

test would conduct one such trial for every possible reshuffling of labels; a commonly used approximation (which you should use here) is to simply shuffle the labels a total of B times ("Monte Carlo permutation test"); for the purposes of this homework, let $B = 10000$.[1]

The folder `http://courses.ischool.berkeley.edu/i290-dds/s17/hw/dds_s17_hw2/` contains a dataset mapping movies (featurized by their genres and major actors who performed in them) to a binary decision of whether or not it was among the 25% highest grossing movies in that set (a "box office hit"). One such feature is whether the movie stars *John Goodman*. Consider the following hypothesis:

> Movies that contain *John Goodman* have a significantly different number of box office hits (either higher or lower) than those that do not.

Code and execute a permutation test evaluating this hypothesis. Can the null hypothesis (that movies featuring *John Goodman* have the same proportion of box office hits as those that do not) be rejected with a significance level of $\alpha = 0.01$? If so, what is the size and direction (positive or negative) of the effect?

Deliverables: one paragraph containing your written answers to the question above, along with code to support it. (Note you must you code your own test here, and not rely on existing implementations.)

2.2 Critique

The nine forms of validity outlined above represent a detailed taxonomy of the different ways in which an analysis can be judged for the extent to which it is valid. What other possible forms of validity are missing from this taxonomy that should be represented within it? Present an argument for a single form of validity—a.) why it captures an important dimension that should be assessed, b.) why you believe it's not captured within Krippendorff's taxonomy, and c.) tangible ways in which an analysis could be assessed according to this dimension.

Deliverable: one-page paper (single-spaced)

[1] Note for reference (not needed for this homework): Rather than choosing a fixed B, a more correct method would establish a 99% confidence interval on that $\hat{p}$-value by plugging $\hat{p}$ and B into the Wilson score interval (which is preferable to the normal approximation to the binomial proportion test, which fails at $\hat{p} = 0$):

$$\frac{1}{1 + \frac{1}{B}2.58^2} \times \left(\hat{p} + \frac{1}{2B}2.58^2 \pm 2.58\sqrt{\frac{1}{B}\hat{p}(1-\hat{p}) + \frac{1}{4B^2}2.58^2} \right) \tag{3}$$

If the upper bound of the confidence interval in equation 3 is less than α, we can reject the null hypothesis at an α significance level.