References

[BergKirkpatrickBurkettKlein12empirical] ✓ Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In Jun'ichi Tsujii, James Henderson, and Marius Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

TOPICS: statistics LOCALFILE: downloaded
URL: `http://www.aclweb.org/anthology/D12-1091`

NOTES: **Wikipedia on some important notions relevant to reading this paper:**
i.i.d. = independent and identically distributed: Used about a sequence or other collection of random variables in which each random variable has the same probability distribution as the others and all are mutually independent. One of the simplest statistical tests, the z-test, is used to test hypotheses about means of random variables. When using the z-test, one assumes (requires) that all observations are i.i.d. in order to satisfy the conditions of the central limit theorem, which states that the probability distribution of the sum (or average) of i.i.d. variables with finite variance approaches a normal distribution. Note that sampling without replacement is not independent, but is *exchangeable*—the joint probability distribution is invariant under permutation.
paired difference test compares two sets of measurements to assess whether their population means differ, using additional information about the sample that is not present in an ordinary unpaired testing situation, either to increase the statistical power, or to reduce the effects of confounders—usually, it is a "repeated measures" test that compares measurements within subjects (i.e. before and after treatment or, in this paper, under two different treatments) or a test in which subjects in both groups are paired by similar characteristics.

$$\begin{aligned}\overline{D} &= \frac{\Sigma\,(X_{i2}-X_{i1})}{n} = \overline{X}_2-\overline{X}_1,\\ \mathrm{var}(\bar{D}) &= \mathrm{var}(\bar{X}_2-\bar{X}_1)\\ &= \mathrm{var}(\bar{X}_2)+\mathrm{var}(\bar{X}_1)-2\mathrm{cov}(\bar{X}_1,\bar{X}_2)\\ &= \tfrac{\sigma_1^2}{n}+\tfrac{\sigma_2^2}{n}-\tfrac{2\sigma_1\sigma_2\mathrm{corr}(X_{i1},X_{i2})}{n},\end{aligned},$$

where n is the sample size and σ_1 and σ_2 are the population standard deviations of the X_{i1} and X_{i2} data, respectively. Note that the variance of D is lower if there is positive correlation within each pair. This explaines why, in the paper, a metric gain is more significant between systems with correlated output.
Types of tests for the difference of the means include:

- *Paired tests* basically reduce the question to single sample tests for the variable $X_i = X_{i1} - X_{i2}$:
  - *paired Z-test:* for normally distributed differences where the population deviation of difference is known to be s (and hence the Z-statistic (approximately) follows a normal distribution)

    $$Z = \frac{\overline{X}_1-\overline{X}_2}{\frac{s}{\sqrt{n}}}$$

    The Z test can also be used if the sample size is large (n>50). In that case, the sample variance is used as an estimate of the population variance:

    $$s^2 \approx \frac{\Sigma(X_i-\overline{X})^2}{n-1}.$$

  - *paired Student's t-test:* for normally distributed differences where the population standard deviation of difference is not known. The t-statistic is compared to the t-distribution with df degrees of freedom.

    $$t = \frac{\overline{X}_1-\overline{X}_2}{\frac{s}{\sqrt{n}}}, \qquad df = n-1,$$

    where s is the sample standard deviation of the differences.
  - *Wilcoxon signed-rank test* for differences that may not be normally distributed but are symmetrically distributed around the median. We assume that for all pairs there is a non-zero difference, order these differences by their absolute values and assign rank $R = i$ to the i-th pair; ties receive

a rank equal to the average of the ranks they span. The test statistic is calculated as

$$W = \left|\Sigma[\mathrm{sgn}(X_{i1} - X_{i2}) \cdot R_i]\right|.$$

The p-value can be calculated from the enumeration of all possible combinations of W given n; alternatively, W is approximately normally distributed for $n > 10$, so that a Z-test can be performed, using

$$Z = \frac{W - 0.5}{\sigma_W},$$
$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{6}}.$$

- *Unpaired tests* that assume *homogeneity of variance* (i.e. that the variance of the two samples is the same).
  - *Unpaired t-test* for samples from normal distributions with a same yet unknown variance.

    $$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$
    $$df = n_1 + n_2 - 2,$$

    where s_p is the *pooled standard deviation* and the *pooled variance* is calculated as

    $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

- *Unpaired tests* that do not assume *homogeneity of variance*, i.e. can cope with samples taken from populations with different variances:
  - *Welch's unequal variance t-test* is an adaptation of Student's t-test which makes use of unpooled variance (in the denominator of the formula for t) and is robust to violation of homogeneity of variance across samples. It assumes that both samples are taken from normal distributions. It computes the t statistic and the degrees of freedom df differently, but uses the same family of t-distributions to obtain a p-value and test the null hypothesis (using a one-tailed or a two-tailed test). [Ruxton06unequal] suggests to use this test even if the homogeneity of variance assumption would possibly hold (because testing for it can only introduce further errors).

    $$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$
    $$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}},$$

    where $\overline{X}_i, s_1^2$ and n_i are the sample mean, sample variance and sample size and $\nu_i = n_i - 1$ is the degrees of freedom associated with the ith variance estimate. Note that df does not have to be an integer;

    $$\min(n_1 - 1, n_2 - 1) < df < n_1 + n_2 - 2$$

    so the df is smaller than it would be in the usual "pooled" t-test.

Notes from the article itself

[1] For the following analyses, pairs are compared so that $\delta(x)$ is always positive.

- *hypothesis testing*: when a system A outperforms system B on a sample $x = x_1, \dots x_n$ by $\delta(x)$, we want to estimate the probability that on a different test set x' shows a similar victory assuming that in fact A is no better than B, i.e. we want to estimate $p(\delta(X) > \delta(x)|H_0) = p(x)$, where X ranges over all test sets of size n. To estimate the p-value, methods such as paired Student's t-test or paired bootstrap can be used.
- *bootstrap*: having a sample x of the population, we draw many new samples $x^{(i)}$ of size n (with replacement, so that the sampling is i.i.d.). We estimate $p(\delta(X) > \delta(x)|H_0)$ as the percentage of $x^{(i)}$ for which $\delta(x^{(i)}) > 2 * \delta(x)$. (Under H_0 that A does not in fact outperform B, we assume that the mean bootstrap gain is equal to the original test gain because we are subsampling a set on which A accidentally outperforms B.) In case the distribution is symmetric about $\delta(x)$, and the mean of $\delta(x^{(i)})$ is equal to $\delta(x)$),

this is equivalent to considering the proportion of bootstrap samples that give $\delta(x^{(i)}) < 0$. In particular, $E(\delta(x^{(i)})) = \delta(x)$ whenever the metric linearly decomposes over sample items (sentences), and by the central limit theorem, the distribution is symmetric whenever the sample size n is large.

- A major benefit of the bootstrap is that any evaluation metric can be used to compute $\delta(x)$—it does not even have to decompose over sentences (which we would need for any variation of the t-test approach).
- In designing a shared task it is important to know how large the test set must be in order for significance tests to be sensitive to small gains in the performance metric.
- For a fixed test size, the domain has only a small effect on the shape of the curve.
- *conventional wisdom*: a certain metric gain is roughly the point of significance for a given task (i.e. 0.4 F1 in parsing or 0.5 BLEU in MT)—the authors confirm that, when such rules of thumb are determined for a given test set (with fixed size and domain), they are fairly accurate. However, the larger the corpus size, the lower the threshold for $p < 0.05$.
- More similar systems tend to achieve significance with smaller metric gains, because their output is more correlated and hence the variance of the differences is smaller (and thus the testing statistics has larger absolute value and p is smaller).
- The authors propose a simple method for automatically generating arbitrary numbers of comparable system outputs: create new training sets by sampling the original training set with replacement. They validate the trends revealed by the synthetic method against data from public competitions.
- Extent to which statistical significance on a test corpus is predictive of performance on other test corpora—when the test set is i.i.d. drawn from the same distribution that generates the new data, the significance levels are well-calibrated, but as the domain of the new data diverges, the predictive ability of significance drops dramatically. E.g., for constituency parsing, when testing on section 23 of the WSJ corpus, $p < 0.00125$ is required to reasonably predict performance on the Brown corpus.
- Rules of thumb obtained by the authors:

task	metric	threshold for related systems	threshold for unrelated systems	used test set from	sample size
summarization	ROUGE		1.10	TAC 2008 t.s.	48 document collections
dep. parsing	unlabeled dep. acc.	1.20	1.51	CoNLL 2007 Chinese t.s.	690 sentences
MT	BLEU	0.28	0.37	German-English WMT 2010 news t.s.	2034 sentences
word alignment	AER	0.50	1.12	part of Hansard t.s.	100 sentences
constituency parsing	F1	0.47	0.57	section 23 of WSJ	

Questions for the article authors

- They said that GIZA++ failed to produce reasonable output when trained with some of these training sets (20 training sets among 1.1M sentences). Why?
- Explain footnote 5.

[Ruxton06unequal] Graeme D Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

Topics: statistics Localfile: downloaded
URL: `http://beheco.oxfordjournals.org/content/17/4/688.full`