

# Comment

Persi Diaconis                      Erich Lehmann  
*Stanford University*                      *UC Berkeley*

February 1, 2008

Sandy Zabell gives us a good feel for Student and his times. We supplement this with later developments showing that Student's intuitive test is optimal in various senses (with some serious caveats). We also show how Fisher's and Hotelling's geometric interpretation of Student's test percolates to a wide variety of statistical applications. These developments bring the historical record up to the present.

## 1 Some History

Student's discovery of the  $t$ -distribution did not come out of a vacuum. As he writes in the introduction to his 1908 paper:

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to  $S/\sqrt{n}$  where  $S$  is the standard deviation of the sample, and to use the tables of the probability integral (i.e., the standard normal distribution).

A few paragraphs later, he adds:

Although it is well known that the method of using the normal curve is only trustworthy when the sample is "large", no one has yet told us very clearly where the limit between "large" and "small" samples is to be drawn.

The use of the  $t$ -statistic for the problem at hand was thus in consistent use in Student's time. A detailed discussion of Laplace's work on its distribution is provided by Hald (1998), Section 20.8, who raises the question why Laplace did not find the  $t$ -distribution, with the answer: "he fell for the easy normal approximation."

## 2 Optimality

The use of the sample mean  $\bar{x}$  to estimate the population mean  $\mu$ , with division by the estimated standard deviation for scaling purposes, seems intuitively so clearly the right thing to do (assuming that the sample comes from a normal distribution) that one might expect the one-sided  $t$ -test to be uniformly most powerful (UMP), i.e., for no improvement to be possible.

Unfortunately this is not the case. It turns out (Lehmann and Stein 1948) that at levels  $< \frac{1}{2}$  it is possible to improve on the power of the  $t$ -test against any particular alternative, the improvement being tailored to that alternative (with a relative loss of power against other alternatives). However, the situation changes when the condition of unbiasedness is imposed. Both the one-sided and two-sided  $t$ -test are UMP among all unbiased tests (Neyman and Pearson 1936, 1938).

Another optimality property is defined by noting that both the hypothesis  $H : \mu = 0$  and the class of alternatives  $\mu > 0$  remain invariant (i.e., unchanged) if all the observations are multiplied by a common positive constant  $c$  since then both  $\mu$  and the standard deviation  $\sigma$  are also multiplied by  $c$ . It is then natural to restrict attention to tests that are invariant under the group  $G$  of all such scale changes. Among all such invariant tests, the  $t$ -test is UMP. (The corresponding result holds for the two-sided  $t$ -test by allowing both positive and negative  $c$ .)

The fact that the  $t$ -test is UMP invariant implies another very different optimality result. The group  $G$  is an amenable group (a class of groups containing compact or abelian groups and their extensions), thus it follows from the Hunt-Stein theorem that the one-sided  $t$ -test maximizes the minimum power over the alternatives  $\mu/\sigma = k$  for any positive  $k$ , and hence also, for example, over the alternatives  $0 < \mu/\sigma < k$  or  $\mu/\sigma > k$ . For the two-sided case  $\mu/\sigma$  has to be replaced by  $|\mu|/\sigma$ . See Lehmann and Romano (2005, Sect 8.5) for more on the Hunt-Stein theorem.

The optimality results above do not require knowledge of the power function. Such knowledge is needed however if one wants to determine the sample size that achieves a given power. This power is given by the noncentral  $t$ -distribution with  $n - 1$  degrees of freedom and non-centrality parameter  $\sqrt{n}\mu/\sigma$ . The family of noncentral  $t$ -distributions for varying noncentrality parameters has monotone likelihood ratio, from which it follows that the  $t$ -test is UMP invariant for testing  $\mu/\sigma \leq c$  against  $\mu/\sigma > c$  for any  $c > 0$ . Further optimality properties (e.g., admissibility) are discussed in Anderson (2004, Sect. 5.6).

### 3 Two Caveats

The robustness properties discussed in Sandy Zabell's paper, together with the optimality properties mentioned in the preceding section, present the  $t$ -test as a very desirable solution to the testing problem in question. However, both must be taken with a grain of salt.

Consider first the test's robustness against non-normality. Evidence for this is provided by some empirical studies by E. S. Pearson (and later by a very extensive and careful investigation by Posten 1979). A different type of robustness occurs because the distribution of the  $t$ -statistic is asymptotically normal for all distributions which lie in the domain of attraction of the normal distribution. For those distributions therefore the asymptotic level of the  $t$ -test is equal to the nominal level.

Note however that it was shown by Bahadur and Savage (1956) that for any fixed sample size, no matter how large, there exist distributions with finite variance for which the level of the  $t$ -test is arbitrarily close to 1. This of course does not contradict the point-wise convergence to the nominal level but only shows that this convergence is not uniform. (Uniform and non-uniform convergence of the level of the  $t$ -test over various classes of distributions is further investigated in Romano 2004.) These results suggest that a certain amount of caution regarding robustness is required for finite sample sizes.

Let us now turn to the optimality results of Section 2. They hold under the assumption that the underlying distribution is normal, but even under slight deviations from normality the  $t$ -test can be far from optimal. The poor performance of the  $t$ -test, particularly for distributions with heavy tails, can be seen by comparison with nonparametric tests such as the Wilcoxon or Normal Scores tests. The asymptotic relative efficiency of  $t$  to either of those tests can be arbitrarily close to zero for sufficiently heavy-tailed distributions with finite variance. On the other hand, for all distributions with finite variance, the asymptotic efficiency relative to  $t$  is  $\geq .864$  for Wilcoxon and  $\geq 1$  for Normal Scores (Hodges and Lehmann 1956 or Chernoff and Savage 1958). Of course, there are other possible breakdowns; for a useful discussion of how dependence affects things, see Miller (1997, Chap. 1).

### 4 The Cost of Not Knowing $\sigma$

For testing the mean of a normal distribution with known  $\sigma$ , one would divide the sample mean  $\bar{x}$  by  $\sigma$  rather than by its estimate  $S$ . The resulting

test would then be more powerful than the  $t$ -test. To see how much is lost by using  $t$  instead of  $\bar{x}/\sigma$ , one can note that the asymptotic relative efficiency of the latter test to  $t$  is 1, which suggests that the loss is small.

More detailed information can be obtained by computing deficiency, i.e., the limit (as the sample size  $\rightarrow \infty$ ) of the difference between the number of observations required by the two tests to achieve the same forms (Walsh 1949 and Hodges and Lehmann 1970). This limit depends on  $\alpha$  and, for  $.01 \leq \alpha \leq .1$ , lies between .8 and 2.7. Thus the number of “wasted observations” is quite small even for large samples.

## 5 The Permutation $t$ -Test

As discussed in Section 4.7.2 of Zabell’s article, Fisher proposed a permutation version of the  $t$ -test which has an exact level without the assumption of normality, provided the underlying distribution is symmetric, or even without any distributional assumptions if the data are obtained through suitable randomization.

Fisher conjectured that the  $t$ -test and its permutation version are nearly equivalent, and confirmed this in an example. The two tests differ only in the critical value for the  $t$ -statistic, this value being constant for the  $t$ -test and random for its permutation alternative. Hoeffding (1952) showed that these two values tend to the same limit, and that asymptotically the two tests have the same power. These results justify Fisher’s suggestion that the usual  $t$ -test could be viewed as an approximation to its distribution-free permutation version.

Since the  $t$ -distribution tends to the normal as the sample size  $n$  tends to  $\infty$ , a third asymptotically-equivalent test consists in using the normal critical value. The problem of when the normal or  $t$  critical value provides the better approximation is solved in Diaconis and Holmes (1994), who also provide algorithms for the exact evaluation of the permutation distribution.

## 6 Some Extensions

So far, we have reviewed various properties of the  $t$ -test. For the remainder of these comments we shall consider some of the many extensions and applications of this test which have made Student’s work so influential.

In his 1908 paper, Student considered only the paired-comparisons (or equivalently, the one-sample) problem. In 1922, Fisher pointed out that the  $t$ -distribution also provided the solution to the two-sample problem (with

equal variances) and that of testing a regression coefficient. A crucial further extension (Fisher 1924) was to the  $F$ -distribution, which allowed the testing of several means or regression coefficients, or more generally testing the general univariate linear hypothesis.

The  $t$ -test was generalized further to the testing of a multivariate normal mean vector (Hotelling 1931) and later also to the sequential case (Wald 1947). In the remainder of these comments, we shall consider in more detail still another generalization, which is less well-known.

## 7 Geometric Understanding and the Volume of Tubes

Fisher's (1915) geometric proof of the distribution of the  $t$ -statistic has been extended by Hotelling, Efron, Eaton and others. It leads to a substantial part of modern differential geometry (Weyl's volume of tubes formulae).

Suppose  $X_1, X_2, \dots, X_n$  are independent with a normal  $(\mu, \sigma^2)$  distribution. The two-sided  $t$ -test for  $\mu = 0$  rejects if  $|T|$  is large, where

$$|T| = \frac{\sqrt{n}|\bar{X}|}{S_n}, \quad \bar{X} = \frac{1}{n} \sum_1^n X_n, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Fisher came to a geometric understanding of both the statistic and its null distribution as follows. Consider the vector  $X = (X_1, \dots, X_n)$ . If  $\mu \neq 0$ , the vector will be close to the vector  $(\mu, \mu, \dots, \mu) = \mu \mathbf{1}$ , with  $\mathbf{1} = (\mathbf{1}, \mathbf{1}, \dots, \mathbf{1})$ . Thus the angle between  $X$  and  $\mathbf{1}$  will tend to be small. On the other hand, if  $\mu = 0$ ,  $X$ , projected onto the unit sphere, is uniformly distributed. Thus, a simple, sensible test of  $\mu = 0$  is this: project  $X$  onto the unit sphere. Look at the distance to the vector  $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ . Reject if this distance is small. To calibrate this test, take a cap around  $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$  with (spherical) area .05% of the total area of the unit sphere. The test rejects if the projection of  $X$  is in the cap. Fisher shows that this is precisely the  $t$ -test with the cap-area the null distribution of the  $t$ -statistic.

Hotelling (1939) considered paired data  $(X_1, Y_1), \dots, (X_n, Y_n)$  modeled as

$$Y_i = ae^{bX_i} + \epsilon_i$$

with  $\epsilon_i$  independent  $n(0, \sigma^2)$  errors. To test  $a = 0$ , Hotelling suggested the following geometric test. Consider the vector  $\nu_b = (e^{bX_1}, \dots, e^{bX_n})$ . As  $b$  varies, this vector, projected onto the  $n$ -dimensional unit sphere, spans a

curve. If  $a = 0$ , the projection of  $Y$  onto the sphere is uniformly distributed. Hence one may test ‘ $a = 0$ ’ by seeing if the projection of  $Y$  is unusually close to the curve. For this, form a tube around the curve with area 5% of the total and reject by assessing if the projection of  $Y$  falls in this tube.

Thus Hotelling needed to find the volume of a tube around the curve. Consider (along with Hotelling) a curve in the plane and a tube of small radius  $r$  around this. If the curve is a straight line, of course the volume is  $2r \times \text{length}$ . Now consider bending the line. Infinitesimally, the volume of a wedge of one side of the line gets bigger and on the opposite side of the line gets smaller. Hotelling shows that there is exact cancellation! This result isn’t only for  $r$  infinitesimally small. As long as there is no global self-intersection, the volume of a tube around a curve is  $2r \times \text{length}$ . The analogous result holds for curves in high-dimensions, as well as for curves on the sphere. See Johansen and Johnstone (1990), Johnstone and Siegmund (1989) for background, details and further applications of Weyl’s formula to statistics.

To carry out such tests for two-parameter models, e.g., Fisher’s test for periodicity where  $y_i = a + b \cos(cx_i + d) + \epsilon_i$ , Hotelling needed to solve the higher-dimensional version of the problem: what is the volume of a tube around a surface? Of course, if the surface is a part of a flat plane in  $\mathbb{R}^n$ , the volume is

$$2r \times \text{surface area.}$$

If a flat surface is bent, say upward along parallel lines, this result continues to hold by the one-dimensional theorem. It is hard to picture what happens if the surface is twisted in space. Hotelling couldn’t solve this problem but gave a talk at the Princeton Mathematics Club on the subject in the late 1930s. Herman Weyl was in the audience and solved the problem. The volume of a tube of radius  $r$  about a surface is

$$2r \times \text{surface area} + cr^3 \times \text{average Gaussian curvature}$$

with  $c$  a universal constant. He found similar formulae for the volume of a tube around  $k$ -dimensional surfaces inside constant curvature spaces such as  $\mathbb{R}^n$  or the  $n$ -sphere. In all cases the volume is a polynomial of degree  $1 + \lfloor \frac{k}{2} \rfloor$  in  $r$  with intrinsically-defined coefficients (e.g., the coefficients do not depend on how the surface is embedded).

Weyl’s “higher-order intrinsic curvatures” were new in geometry. They lead Chern to his development of the modern version of the Gauss-Bonnet theorem and much else. The story is well told in Osserman (1978) or Gray

(2004). It has had application in statistics, in particular to the original problem of Hotelling, for periodicity in a time series; see Knowles and Siegmund (1989). For confidence sets in nonlinear regression, see Naiman (1990). For applications to projection pursuit, see Sun (1993). For the maximum of Gaussian processes, see Adler and Taylor (2007).

## 8 From Volume Tests to Algebraic Statistics

Some of the appeal and robustness properties of volume tests comes from their geometric description (see Efron 1969, Eaton and Efron 1970, and the discussion in Eaton 1989, pp. 63–67). This led Diaconis and Efron (1985) to suggest volume tests in a variety of other problems. For example, consider tests of independence for an  $I \times J$  contingency table. The surface of independence sits in the  $(IJ - 1)$ -dimensional simplex  $S$  of all probabilities on  $IJ$  cells. The observed table can be mapped into  $S$  (by dividing  $N_{ij}$  by the total  $N$ ). A test of independence rejects if the observed table is close to the surface. This requires a tube around the surface of volume .5% of the total volume of  $S$ . Computing this volume is difficult, as is the justification of uniform measure on  $S$ . A second volume test involves conditioning on the row and column sums, and sampling tables uniformly with these sums fixed. This algorithmic task has led to a healthy development which brings techniques of algebraic geometry into the picture.

Consider the task of generating a random  $I \times J$  table with fixed margins  $N_{i\cdot}, N_{\cdot j}$ ,  $1 \leq i \leq I$ ,  $1 \leq j \leq J$ . A Markov chain-Monte Carlo algorithm proceeds as follows. From one table  $T$  satisfying the constraints, pick a pair of rows and a pair of columns randomly. Change the entries of  $T$  in the four determined cells by adding and subtracting one in the pattern (choose with probability 1/2)

$$\begin{array}{ccc} + & - & \text{or} & - & + \\ - & + & & + & - \end{array}$$

If this results in a table with non-negative entries, the change is made; otherwise, the walk stays at  $T$ . It is not hard to see that this walk gives a connected Markov chain with a uniform stationary distribution on the set of tables with the fixed row and column sums.

Finding similar walks for higher-way tables or related problems in logistic regression defeated all comers. Diaconis and Sturmfels (1998) managed to cast this as a problem in computational algebra. The computer successfully finds such moves which are nowadays in routine use, via programs such as Latte (“Lattice point Enumeration”; available from [www.ucdavis.edu/~latte/](http://www.ucdavis.edu/~latte/)).

The algebraic geometric techniques then found application in design (Pistone et al. 2001) and in determining the behavior of complex singularities for maximum likelihood in log-linear models (Hosten and Meek 2006).

While far from the source, there is a clear trail back through Hotelling's volume test, Fisher's geometrical view and hence back to Student's original 1908 paper.

## References

- Adler, R. and Taylor, J. (2007). *Random Fields and Geometry*. Springer, New York.
- Anderson, T. W. (2004). *An Introduction to Multivariate Statistical Analysis*, 3rd edition. John Wiley & Sons, New York.
- Bahadur, R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27**, 1115–1122.
- Chernoff, H. and Savage, R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29**, 972–994.
- Diaconis, P. and Efron, B. (1985). Testing for independence in a two way table: New interpretations of the chi-square statistic. *Ann. Statist.* **13**, 845–913.
- Diaconis, P. and Holmes, S. (1994). Gray codes for randomization procedures. *Stat. Comput.* **4**, 287–302.
- Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26**, 363–397.
- Eaton, J. (1989). *Group Invariance Applications in Statistics*. IMS, Hayward, CA.
- Eaton, J. and Efron, B. (1970). Hotelling's  $T$ -test under symmetry conditions. *J. Amer. Statist. Assoc.* **65**, 702–711.
- Efron, B. (1969). Student's  $t$ -test under symmetry conditions. *J. Amer. Statist. Assoc.* **64**, 1278–1302.

- Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507–521.
- Fisher, R. (1924). On a distribution yielding the error functions of several well known statistics. In *Proc. Int. Cong. Math., Toronto* **2**, 805–813.
- Gray, A. (2004). *Tubes*, 2nd ed. Birkhäuser Verlag, Basel, Switzerland.
- Hald, H. (1998). *A History of Mathematical Statistics from 1750 to 1930*. John Wiley & Sons, New York.
- Hodges, J. and Lehmann, E. (1956). The efficiency of some nonparametric competitors of the  $t$ -test. *Ann. Math. Statist.* **27**, 324–335.
- Hodges, J. and Lehmann, E. (1970). Deficiency. *Ann. Math. Statist.* **41**, 783–801.
- Hoeffding, W. (1952). The large sample power of tests based on permutations of the observations. *Ann. Math. Statist.* **23**, 169–192.
- Hosten, S. and Meek, C. (2006). Computational algebraic statistics. *J. Symb. Comput.* **41**, 123–254.
- Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Statist.* **2**, 360–378.
- Hotelling, H. (1939). Tubes and spheres in  $n$ -spaces and a class of statistical problems. *Amer. J. Math.* **61**, 440–460.
- Johansen, S. and Johnstone, I. (1990). Hotelling's theorem on the volume of tubes: Some illustrations in simultaneous inference and data analysis. *Ann. Statist.* **18**, 652–684.
- Johnstone, I. and Siegmund, D. (1989). On Hotelling's formula for the volume of tubes and Naiman's inequality. *Ann. Statist.* **17**, 184–194.
- Knowles, M. and Siegmund, D. (1989). On Hotelling's approach to testing for a nonlinear parameter in regression. *Int. Statist. Rev.* **57** 205–220.
- Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York.
- Lehmann, E. and Stein, C. (1948). Most powerful tests of composite hypotheses. I. Normal distributions. *Ann. Math. Statist.* **19**, 495–516.

- Miller, R. G., Jr. (1997) *Beyond ANOVA: Basics of Applied Statistics*. Chapman and Hall/CRC Press, Boca Raton, FL.
- Naiman, D. (1990). Volumes of tubular neighborhoods of spherical polyhedra and statistical inference. *Ann. Statist* **18**, 685–716.
- Neyman, J. and Pearson, E. (1936). Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Statist. Res. Mem.* **1**, 113–137.
- Neyman, J. and Pearson, E. (1936, 1938). Contributions to the theory of testing statistical hypotheses. *Statist. Res. Mem.* **1**, 1–37; **2**, 25–57.
- Osserman, R. (1978). The isoperimetric inequality. *Bull. Amer. Math. Soc.* **84**, 1182–1238.
- Pistone, G., Riccomagno, E. and Wynn, H. P. (2001). *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman and Hall/CRC Press, Boca Raton, FL.
- Posten, H. (1979). The robustness of the one sample  $t$ -test over the Pearson system. *J. Stat. Comput. Sim.* **9**, 133–149.
- Romano, J. (2004). On nonparametric testing, the uniform behavior of the  $t$ -test, and related problems. *Scand. J. Statist.* **31**, 567–584.
- Sun, J. (1993). Tail probabilities of the maxima of Gaussian random fields. *Ann. Probab.* **21**, 34–71.
- Wald, A. (1947). *Sequential Analysis*. John Wiley & Sons, New York.
- Walsh, J. (1949). Some significance tests for the median which are valid under very general conditions. *Ann. Math. Statist.* **20**, 64–81.
- Weyl, H. (1939). On the volume of tubes. *Amer. J. Math.* **61**, 461–472.