Final pre-publication version. Published version appeared as Chapter 36, pp 405- 412, In, Past, Present, and Future of Statistical Science, Lin, Genest, Banks, Molenberghs, Scott, and Wang Eds. CRC Press and COPSS.

Chapter 1

Buried treasures

Michael A. Newton
Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Keeping pace with the highly diversified research frontier of statistics is hard enough, but I suggest that we also pay ever closer attention to great works of the past. I offer no prescription for how to do this, but reflect instead on three cases from my own research where my solution involved realizing a new interpretation of an old, interesting but possibly uncelebrated result which had been developed in a different context.

1.1 Three short stories

1.1.1 Genomics meets sample surveys

Assessing differential expression patterns between cancer subtypes provides some insight into their biology and may direct further experimentation. On similar tissues cancer may follow distinct developmental pathways and thus produce distinct expression profiles. These differences may be captured by the sample variance statistic, which would be large when some members of a gene set (functional category) have high expression in one subtype compared to the other, and other members go the opposite way. A case in point is a collection of cell-cycle regulatory genes and their expression pattern in tumors related to human papilloma virus (HPV) infection. Pyeon *et al.* (2007) studied the transcriptional response in $n = 62$ head, neck and cervical cancer samples, some of which were positive for virus (HPV+) and some of which were not (HPV-). Gene-level analysis showed significant differential expression in both directions. Set-level analysis showed that one functional category stood out from the several thousands of known categories in having an especially large value of between-gene/within-

set sample variance. This category was detected using a standardized sample variance statistic; the detection itself launched a series of experiments on the involved genes, both in the same tissues under alternative measurement technology and on different tissues. The findings lead to a new hypothesis about how HPV+/- tumors differentially deregulate the cell-cycle processes during tumorigenesis as well as to biomarkers for HPV-associated cancers (Pyeon *et al.*, 2011). Figure 1 shows a summary of gene-level differential expression scores between HPV+ and HPV- cancers (so-called log fold changes), for all genes in the genome (left), as well as for $m = 99$ genes from a cell-cycle regulatory pathway.

A key statistical issue in this case was how to standardize a sample variance statistic. The gene-level data were first reduced to the log-scale fold change between HPV+ and HPV- cell types; these $\{x_g\}$, for genes g, were then considered fixed in subsequent calculations. For a known functional category $c \subset \{g = 1, 2, \ldots, G\}$ of size m, the statistic $u(x, c)$ measured the sample variance of the x_g's within c. This statistic was standardized by imagining the distribution of $u(x, C)$, for random sets C, considered to be drawn uniformly from among all $\binom{G}{m}$ possible size-m subsets of the genome. Well forgetting about all the genomics, the statistical question concerned the distribution of the sample variance in without-replacement finite-population sampling; in particular, I needed an expected value and variance of $u(x, C)$ under this sampling. Not being especially well versed in the findings of finite-population sampling, I approached these moment questions from first-principles and with a novice's vigor, figuring that something simple was bound to emerge. I did not make much progress on the variance of $u(x, C)$, but was delighted to discover a beautiful solution in Tukey (1950, page 517), which had been developed far from the context of genomics and which was not widely cited. Tukey's *buried treasure* used so-called K functions, which are set-level statistics whose expected value equals the same statistic computed on the whole population. Subsequently I learned that earlier R.A. Fisher had also derived this variance (see also Cho *et al.*, 2005), and, in any case, I was glad to have gained some insight from Tukey's general framework.

1.1.2 Bootstrapping and rank statistics

Researchers were actively probing the limits of bootstrap theory when I began my statistics career. A case of interest concerned generalized bootstrap means. From a real-valued random sample $X_1, X_2, \ldots, X_n$ one studied the conditional distribution of the randomized statistic

$$\bar{X}_n^W = \frac{1}{n} \sum_{i=1}^{n} W_{n,i} X_i,$$

conditional on the data $\{X_i\}$, and where the random weights $\{W_{n,i}\}$ were generated by the statistician to enable the conditional distribution of $\bar{X}_n^W$ to approximate the marginal sampling distribution of $\bar{X}_n$. Efron's bootstrap corresponds to weights having a certain multinomial distribution, but indications were that useful approximations were available for beyond the multinomial.

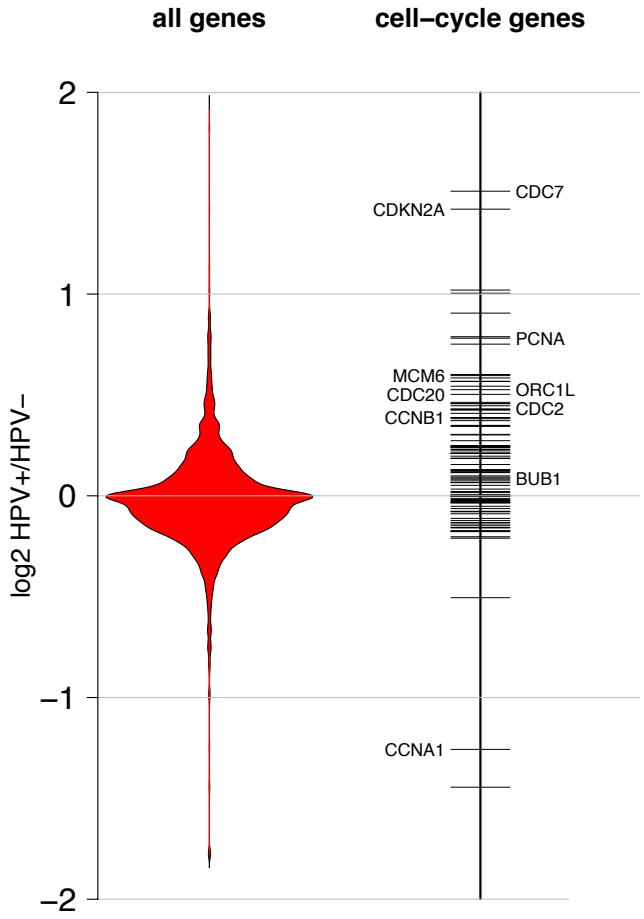

Figure 1.1: The relative positions of $m = 99$ cell-cycle genes (KEGG 04110) (right) are shown in the context of all measured genes (left) when genes are sorted by log fold change between HPV+ and HPV- tumors (vertical axis). Widths in the red *violin* plot indicate the empirical density. KEGG 04110 had higher standardized sample variance then any functional category in GO or KEGG. Based on this high variance, further experiments were performed on the 10 named genes (right) leading to a new hypothesis about how the HPV virus deregulates the control of cell cycle, and to biomarkers for HPV-associated cancer.

In a most rewarding collaboration, David Mason and I tackled the case where $\{W_{n,i}\}$ were exchangeable, making the seemingly superfluous observation that $\bar{X}_n^W$ must have the same conditional distribution, given data $\{X_i\}$, as the additionally randomized

$$T_n = \frac{1}{n}\sum_{i=1}^{n} W_{n,\pi_{n,i}} X_i$$

where, for each n, $\{\pi_{n,i}\}$ is a uniform random permutation of the integers $\{1, 2, \ldots, n\}$. While the usual boostrap statistic has two sources of randomness (one from the data and from the bootstrap weights), this T_n had yet a third source, neither generated by nature or the statistician, but just imagined owing to the exchangeability of the weights. Having all three sources allowed us to condition on both the data $\{X_i\}$ and the statistician-generated weights $\{W_{n,i}\}$, and still have some randomness in T_n.

A quite unconnected and somewhat amazing treasure from the theory of linear rank statistics now became relevant. Given two triangular arrays of constants, $\{a_{n,i}\}$ and $\{b_{n,i}\}$, the randomized mean

$$S_n = \sum_{i=1}^{n} a_{n,\pi_{n,i}} b_{n,i}$$

had been studied extensively in nonparametric testing, because this is the form of the linear rank statistic. Hájek (1961) presented weak conditions on the triangular arrays such that S_n is asymptotically normal, owing to the random shuffling caused by $\{\pi_{n,i}\}$. Thus, reconsidering Hájek's result in the new bootstrap context was the key to making progress on the weighted bootstrap problem (Mason and Newton, 1992).

1.1.3 Cancer genetics and stochastic geometry

A tumor is monoclonal in origin if all its cells trace by descent to a single initiated cell that is aberrant relative to the surrounding normal tissue (*e.g.*, incurs some critical genetic mutation). Tumors are well known to exhibit internal heterogeneity, but this does not preclude monoclonal origin, since mutation, clonal expansion, and selection are dynamic evolutionary processes occurring within a tumor that move the single initiated cell to a heterogeneous collection of descendants. Monoclonal origin is the accepted hypothesis for most cancers, but evidence is mounting that tumors may initiate through some form of molecular interaction between distinct clones. As advanced as biotechnology has become, the cellular events at the point of tumor initiation remain beyond our ability to observe directly, and so the question of monoclonal versus polyclonal origin has been difficult to resolve. I have been fortunate to work on the question in the context of intestinal cancer, in series of projects with W.F. Dove, A. Thliveris, and R. Halberg.

When measured at several months of age, intestinal tracts from mice used in the experiments were dotted with tumors. By some rather elaborate experimental techniques cell lineages could be marked by one of two colors: some tumors were pure in color, as one would expect under monoclonal origin, yet some contained cells of both colors, and were thus overtly polyclonal. The presence of such polyclonal tumors did not raise alarm bells, since it was possible that separate tumors were forming in close proximity, and that they had merged into a single tumor mass by the time of observation. If so, the polyclonality was merely a consequence of *random collision* of independently initiated clones, and did not represent a mechanistically important phenomenon. The investigators suspected, however, that the frequency of these overtly polyclonal (heterotypic) tumors was too high to be explained by random collision, especially considering the tumor size, the overall tumor frequency, and the lineage marker patterns. It may have been, and subsequent evidence has confirmed, that cellular interactions are critical in the initial stages of tumor development. The statistical task at hand was to assess available data in terms of evidence against the random collision hypothesis.

In modeling data on frequencies of various tumor types, it became necessary to calculate the expected number of monoclonal tumors, biclonal tumors, and triclonal tumors when initiation events occur randomly on the intestinal surface. This is a problem in stochastic geometry, as clones will collide if they are sufficiently close. Like in the gene-set-variance problem, I tackled the expected value using first principles and with hopes that a simple approximation might emerge. The monoclonal and biclonal expectations were not so hard, but the triclonal calculation gave me fits. And then I found Armitage (1949). In a problem on the overlap of dust particles on a sampling plate, Armitage had faced the same expected value calculation and had provided a rather thorough solution, with error bounds. If N particles land at random in a region of area A, and if they *clump* when they lie within δ units, then the expected numbers of singletons, clumps-of-two, and clumps-of-three particles are approximately

$$\mu_1 = Ne^{-4\psi}, \quad \mu_2 = 2N\left(\psi - \frac{4\pi + 3\sqrt{3}}{\pi}\psi^2\right), \quad \mu_3 = N\left(\frac{4(2\pi + 3\sqrt{3})}{3\pi}\psi^2\right),$$

where $\psi = N\pi\delta^2/(4A)$. Fortunately I could use the framework of stochastic geometry to link the quite different contexts (particle counting and tumor formation) and identify a path to testing the random collision hypothesis (Newton *et al.*, 2006). The biological consequences continue to be investigated.

1.2 Concluding remarks

I have found great utility in beautiful statistical findings that have been relatively uncelebrated by the field and that were developed in response to problems different than I was facing. I expect there are many such buried treasures, and I encourage statisticians to seek them out even as they push forward addressing

Figure 1.2: John Tukey, Jaraslav Hájek, and Peter Armitage.

all kinds of new statistical problems. Perhaps there is very little to what I'm saying. Had I been more prepared when launching into any of the three cases above I might have known right away how to use the available statistical results. But this seems like a lot to ask; our training programs are bursting with course work and cannot be expected to explain all of the discipline's treasures. You might also argue that the great thing about statistics and mathematics is that a single formalism works equally in all kinds of different contexts; my case studies do no more than express how the formalism is not dependent upon context. Perhaps my point is more that we must continue to exercise this formalism, continue to find analogies between distinct problems, and continue to support and develop tools that make these connections easier to identify. Thank goodness for archiving efforts like JSTOR and the modern search engines that help us find these treasures. All of us can help by continuing to support efforts, like open access, aiming to minimize barriers to information flow. Authors and journals can help by making a greater effort to cite key background references and suggest links to related problems. Instructors, especially of courses in mathematical statistics, can help by emphasizing the distinct contexts that enliven each statistical fact. Grant reviewers and tenure committees can help by recognizing that innovation comes not only in conjuring up new theory and methodology but also by the thoughtful development of existing statistical ideas in new and important contexts. Finally, thanks to John Tukey, Peter Armitage, and Jaraslav Hájek and others for the wonderful results they've left for us to find.

There is more treasure in books than in all the
pirate's loot on Treasure Island and best of all,
you can enjoy these riches every day of your life.

-Walt Disney

References

1. Armitage, P. (1949). An overlap problem arising in particle counting. *Biometrika*, 45, 501-519.

2. Cho, E., Cho, M.J., and Eltinge, J. (2005). The variance of the sample variance from a finite population. *Int. J. Pure Appl. Math.*, 21, 389-396.

3. Hájek, J (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *Ann. Math. Statist.*, 32, 506-523.

4. Mason, DM and Newton, MA (1992). A rank statistics approach to the consistency of a general bootstrap. *Annals of Statistics*, 20, 1611-1624.

5. Newton, MA, Clipson, L, Thliveris, AT and Halberg, RB (2006). A statistical test of the hypothesis that polyclonal intestinal tumors arise by random collision of initiated clones. *Biometrics*, *62*, 721-727.

6. Pyeon, D, Newton, MA, Lambert, PF, den Boon, JA, Sengupta, S, Marsit, CJ, Woodworth, CD, Connor, JP, Haugen, TH, Smith, EM, Kelsey, KT, Turek, LP, and Ahlquist, P. (2007). Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Research* 67, 4605-4619.

7. Pyeon, D, Lambert, P.F., Newton, MA, and Ahlquist, PG. (2011). Biomarkers for human papilloma virus-associated cancer. US Patent No. 8,012,678 B2.

8. Tukey, J.W. (1950). Some sampling simplified. *J. Amer. Statist. Assoc.*, 45, 501-519.

9. Images reproduced without permission from:
`http://en.wikipedia.org/wiki/John_Tukey` (Tukey);
`http://www.galaktia.com/matematia/` (Hájek);
`http://en.goldenmap.com/Peter_Armitage` (Armitage).