Tech talk:

Probability from "C" to "G"

Robert E. Gladd, CQE

Item 1: *On an episode of NBC's "ER" earlier this year, two physicians—one a senior clinical researcher, the other his resident subordinate—are scurrying through the hospital halls discussing the latest results of a study they are working on. The mentor says, in a tone of enthusiastic anticipation, "the chi-square of our latest trial yielded a p-value of 0.06; we're just one percentage point away from being able to publish significant findings! Get right on it!"*

Item 2: *A CQI trainer with impressive academic and professional pedigrees addresses a QI seminar of health care employees on the topic of statistical fundamentals for process improvement. Discussing the nature of data distributions, he opines that "statisticians like to use big words like 'skew' and 'kurtosis,' but most things in nature are normally distributed. The normal distribution is a law of nature."*

Item 3: *A widely-used teaching text for radiation lab chemists publishes a monograph from a radioimmunoassay method development study for quantifying serum barbiturate concentrations. The researcher reports running a set of 30 serum blanks to obtain a background baseline and set a cut-off limit, and the following parameter estimates are given: $\mu = 15$ ng/ml., $\sigma = 27$ ng/ml. Consequently, a cut-off of 100 ng/ml. is established "to assure a <1% chance of a false positive."*

"Just gimme the p-value!!!"

I used to serve on a health care QI projects committee with a doc—call him Dr. I.M. Harried—who once precipitously burst out with *"just gimme the p-value"* in the midst of a teleconference call while we were discussing project data. As if the knowledge that "p=0.084" or "p<0.01" utterly settled the decision to be made concerning the clinical process issue at hand. Dr. Harried's inductive innocence is unfortunately all too common, and ends up in part being reflected in cutesy but inaccurate dialogue written for supposedly clinically astute television productions such as *"ER."*

A little statistics can be a dangerous thing. Recall Disraeli's lament. Better yet, recall Dr. Shewhart's admonition:

"...applied science...is even more exacting than pure science in certain matters of accuracy and precision. For example, a pure scientist makes a series of measurements and upon the basis of these makes what he [she? -Ed] considers to be the best estimates of accuracy and precision, regardless of how few measurements he may have. He may readily admit that future studies may prove such estimates to be in error...But now let us look at the applied scientist. He knows that if he were to act upon the meagre evidence sometimes available to the pure scientist, he would make the same mistakes as the pure scientist...He also knows that through his mistakes someone may lose a lot of money or suffer physical injury, or both. [3]

Forget for a moment the broader range of issues implicit in statistical reasoning—data dispersional characteristics, random and representative sampling, temporal stability, etc. We will address them shortly, but first of all, one fundamental and prevalent oversight needs attention: "p-values" or probability findings are themselves *estimates*, and, as such, also have distributional characteristics—minima, maxima, range, central tendencies, (a)symmetries. Such should be obvious: run a quantifiable experiment 100 times using 100 different random samples drawn from the same population, you get—in addition to a fairly Gaussian distribution of μ estimates (sample means)—*a distribution of probability estimates (used to assess your nul and alternative hypotheses).*

When someone makes an empirical assertion as to what "the probability *is*" I instinctively recall Deming's warning that "[P]robability has use; tests of significance do not. There is no true value of anything."[4] That one always gets 'em going, but it's true. My knee-jerk reply is always, "you mean your probability *estimate* is..."

Probability 101 revisited

So what? What's the point? Well, recall that the "expected value" of any measured phenomenon is the probability (0<=p<=1) times the "payoff" (i.e., the total benefit or penalty of an outcome). A thorough analysis that seeks to minimize risk *must* take into account the range and distribution of calculable probability estimates, which is why we have such tools as decision theory minmax criteria, loss functions, and Bayesian risk estimators.

More big words. Jeez, I think I'm gonna be ill. Back up a minute, OK; you wanna run that 'standard deviation' thing by me one more time?

OK: σ = RMS, the Root-Mean-Squared deviation, a.k.a. the square root of the mean squared deviation, a.k.a. the *expected variability* around the mean value of your data. So; now we've got "μ" and "σ." Let's head to Normal Distribution Land:

$$(n;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-0.5\left[\frac{(x-\mu)}{\sigma}\right]^2}$$

The Gaussian probability density function that gives us our familiar Bell Curve, which, when subjected to the mechanics of integral calculus, provides the "areas under the curve" that are analogous to the probability of our experimental numerical data points and μ estimates being "x" distance ± the target mean while still belonging to the population (i.e., no identifiable and remediable causes).

Remember (dimly) our friends H_0 and H_a, the "nul" and "alternative" hypotheses, respectively? Recall "setting the rejection criterion for H_0"? "Alpha" and "Beta" errors? Remember? *"Set $\alpha = 0.05$, if $Z>1.96$, reject H_0."* (recall that, by convention H_0, the "nul hypothesis" properly means that, going in, your experimental data characteristics—principally the means—are assumed to be "no different" statistically from the population or reference parameters).

So, under the assumptions of distributional normality, "set $\alpha = 0.05$, if $Z \geq 1.96$, reject H_0" is interpreted as a 5% chance of a "false positive" determination: attributing a "significant" (identifiably casual?) difference where none exists (the "Type I," or "alpha error"). Conversely, rejecting H_a (the alternative hypothesis) when it is in fact true is known as the β (beta) error: a "false negative," concluding that things are OK when they are not ("Type II" error).

Well, being right or wrong either way (α *or* β) has, for the most part, quantifiable consequences. The ranges of our win/loss scenarios ("expected values"), however, will depend on our implicit empirical assumptions. To this point we are talking about the "G" referred to in the title of this essay: *Gauss*—the probability estimates derived from "normally distributed" data.

Chebychev's Theorem

On to the "C" in our title. The Russian mathematician Chebychev derived a proof revealing that, for *any* data set with calculable mean and standard deviation estimates, the proportion ("probability") of data within ± "k" standard deviations is *always at least*

$$\left(1 - \frac{1}{k^2}\right)$$

So, at least 75% must be within ± 2σ [1-(1÷2^2)], 89% [1-(1÷3^2)] within ±3σ, and so on, irrespective of any distributional characteristics. Chebychev's Theorem provides us with "worst case" probability estimates, and serves as a lower bound for probability estimate distributions, with the "perfectly" normal, Gaussian distribution at the other end. Probability from "C" to "G."

Recall that a "z-score" (a.k.a. "standard score") represents the departure of a value from the mean, expressed as a decimal fraction or multiple of the standard deviation:

$$Z = \frac{(x-\mu)}{\sigma}$$

[3] Shewhart, Walter A., Statistical method from the viewpoint of quality control, (New York, Dover Publications, 1986), pp. 120-1. Originally published in 1939 by the U.S. Department of Agriculture. A watershed resource and delightful book.

[4] ibid., Forward, pg. i.

In a *perfectly* smooth Gaussian distribution (extant in theory alone), 68.2% of the area under the curve lies within $\pm 1\ \sigma$ ("one sigma"), 95.4% within $\pm 2\ \sigma$, and 99.7% within $\pm 3\ \sigma$. When we recognize that real-world data sets merely approximate the "normal" to varying degrees, a bit of circumspection would seem to be appropriate when making probability assertions. Comparing Chebychev area limits to the corresponding Gaussian areas for sigmas from one to five (i.e., the z-values) illuminates the gaps:

z-value	p(Chebychev)	p(Gauss)
1.0	0.000	0.682
1.5	0.556	0.866
2.0	0.750	0.954
2.5	0.840	0.988
3.0	0.889	0.997
3.5	0.918	>0.999
4.0	0.938	>0.999
4.5	0.951	>0.999
5.0	0.960	>0.999

Notice the 20.4% and 10.8% differences at 2 and 3 z, respectively between the Chebychev limits and the Normal Assumption.[5] Anyone who crunches some numbers, derives a z-score of 2.0, and blissfully concludes that they have found "significance, with <5% chance of an α error" is being perhaps unduly optimistic. Note that under Chebychev, "<5% α error" doesn't kick in until a z of about 4.5.

Let's re-visit Item 3 from the beginning of this article. Obviously, the researcher concluded that 100 ng/ml. was a cut-off point equivalent to a z-score of about 3.15 [$\mu = 15$, plus $3\sigma = 96$, ergo at 100, $p(\alpha \text{ error})<0.01$]. What's wrong with this picture?

First of all, these are *ratio-level* data, i.e., you can't have less than zero nanograms concentration of anything. For a ratio-level distribution to be "normal," the sigma cannot be greater than roughly 33% of the mean ($\mu - 3\sigma >\cong 0$). This fellow reports a mean of 15 with a sigma of 27, a relative standard deviation[6] $[(\sigma \div \mu) \times 100]$ of 180%!

It's Chebychev time, baby. These data—typical of matrix blank results, I might add—are highly skewed, and worthy of considerably more inductive caution than displayed by this researcher. There's a false precision/"significant figures" problem with this example also. Eschewing for a moment the dearth of support for statistical normality in the data as reported, had this chemist run 1,000 matrix blanks (or, minimally, 100) and not gotten a result > 100 ng/ml., one might buy off on an assertion of "$p(\alpha \text{ error}) < 0.01$" irrespective of any statistical skewness. But, where n = 30, one treads on shaky ground.[7] For instance: $1 \div 8 = 0.125$. If I merely sample eight widgets and find one defective, does that justify concluding that "12.5%" of my lot are defectives? If you express any finding as a percentage after examining less than 100 elements, you are *extrapolating.*

"...a Law of Nature."

Back to Item 2: The instructor was absolutely correct in one sense; many measurement distributions of natural phenomena closely approximate the Gaussian model. And, it is also the case that "higher moment" parameter estimates such as Skew and Kurtosis coefficients are unreliable when computed using small a "n," and should be interpreted carefully.

But quality improvement mainly involves the measurement of human organizational activities: manufacturing operations and service processes. One cannot just assume pure dispersional normality in such environments. To do so elevates the risk of error. Perhaps the increase in risk is trivial, perhaps not.[8] Depends on the consequences—the reasonable ranges in the matrix of "payoffs" and "payouts."

Our instructor that day was just trying to provide warm fuzzies to an audience comprised primarily of paradigm-shift-anxious and befuddled statistical naifs, in an effort to get some fundamental principles of variation across. And that is just fine; he had a lot to cover in a limited time. Problems loom, however, when subsequent lessons delve no further into the fine print. All too often, there *are* no ensuing lessons, given that training expenditures typically continue to be posted to "overhead expenses" rather than to "capital investments" on the chart of accounts.

In the Downsizing Era, we're all expected to become plug-'n-play one-person support-staff / CQI information-processing analytical departments, adept with "measurable objectives," databases, spreadsheets, and statistics. Excel and Quattro Pro serve up neat-o color (rotating 3-D!) graphics, replete with histograms, trend lines, and 95% confidence intervals after a few mouse clicks and drags. Never before has it been so easy to assemble such quantities of aesthetically pleasing misinformation. Process tampering (or neglect) via Presentation-Plus, 24-bit color, 720 x 360 dpi drop-shadowed Doo-Doo.

Disraeli's lieutenant

Style over substance: Been there, done that. I once presented a paper on an aspect of laboratory QC at an EPA radiochemistry conference. During the post-delivery Q&A a curmudgeonly research eminence from Canada rose in rebuke: "Well, that just goes to show once again that you can 'prove' *anything* with statistics." It was a valuable, if public and painful lesson in statistical humility. Given my misleading title—*Improved Evaluation of Environmental Radiochemical Inorganic Solid Matrix Replicate Precision: Normalized Range Analysis Revisited*—this man had come expecting information on how to enhance lab precision by reducing variability in certain problematic replicate QC samples, and he was not at all amused with what he regarded as a bunch of sophistic post-hoc p-value apologies for why our external PE inorganic solid replicates were so volatile.

Ouch! He was right. My paper was eloquently written and peppered with cool graphs of intersecting bell curves, shaded regions thereunder, and tables of z's and p's. It had been reviewed and approved by no fewer than three of my laboratory Ph.D. superiors (one of whom was my "co-author"). Our Gaussian sanguinity aside, it was technically accurate, as far as it went—which was *nowhere,* in terms of adding value by prescribing a useful method for reducing operational variability.

The censure that day is forever stamped in my synaptic inventory. Now, when I am to evaluate QI data, I focus on three questions:

1. Have the model assumptions been adequately addressed, including stability over time? (after all, we're *predicting,* right?)
2. What are the likely practical consquences of being wrong, either through process tampering (α) *or* neglect (β)? And, *most importantly,*
3. How are these results helping me locate remediable causes so we can improve things?

All inductive methodology, however simple or sophisticated, must work cautiously and critically toward this last end. The "statistically significant p-value" estimate *may* be a harbinger of a casual and correctible phenomenon; it may be a mere wake-up call to look more closely at process variation, *or;* it may be a misleading probabilistic apparition arising out of natural sampling variability and/or a haze of unsustainable assumptions.

You get what you inspect, not what you expect. ■

[5] Assuming a 2-tail assessment. See a problem?

[6] RSD, also referred to as the "CV" or coefficient of variation. Sometimes called the "percent standard deviation." Merely the ratio of sigma to mean.

[7] For an interesting discussion on the problem of computing confidence intervals when experiments or study results turn up zero "positives" (however operationally defined), see *If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators* by Hanley & Lippman-Hand, PhDs, in JAMA, 4-1-83, Vol. 249, No. 13. The authors examine the (Poisson process) mathematical basis for applying, among other things, a "rule of 4.6/n" with which to calculate the upper bound of a 99% confidence interval ($0 <= p <= 4.6/n$) when zero positives are found in a sample of size "n." By this rule, our chemist would need a run of at least 461 serum blanks, all < 100 ng/ml., to sustain his false-positive probability claim, given the obvious non-normality in his initial data. Otherwise he's stuck with Chebychev and about a 10% F.P. risk in the absence of a defensible fit with some other type of known probability density function. The "4.6" rule would be even less forgiving: $4.6 \div 30 = 0.153$.

[8] Another 'big word' bandied about is "robustness." It is argued—usually by computer-modeling academics with no money at risk—that various parametric assumptions can be violated with no harm done to outcomes. Someone forgot to inform several of the laboratory auditors I've had to face.