Analytical Quality in the Medical Laboratory – The ASAP Concept. Part 3: Measurand

Stepman H.[1], Stöckl D.[2]

[1]*Ghent University, Faculty of Pharmaceutical Sciences, Laboratory for Analytical Chemistry.*
[2]*STT Consulting, Abraham Hansstraat 11, B-9667 Horebeke (Belgium), dietmar@stt-consulting.com.*

SUMMARY

Because the topic of this essay is often considered boring, we introduce here 3 friendly people who we may also use in later essays. They are the curious layperson, the "All-Knowing" Clinical Biochemist, and the "earth-bound" secretary. While the concept of the measurand seems to be easy, the discipline has surprisingly different opinions about certain specific measurands. This holds true, in particular, for what concerns the measurement of component mixtures. Nevertheless, the unit for mixture analysis should be mol and tests should measure equimolar to the medically relevant extent.

Keywords: Measurand, component, mixture analysis, surrogate, homework.

SOUHRN

Stepman H., Stöckl D.: Analytická kvalita v klinické laboratoři – koncepce ASAP. Část 3: Measurand

Protože téma tohoto eseje je často považováno za nudné, uvádíme zde tři milé osoby, které můžeme také využít i v případě pozdějších esejů. Jsou to zvědavý laik, „vševědoucí" klinický biochemik, a sekretářka bez fantazie. Zatímco koncept pojmu měřené veličiny se zdá být snadný, vědní obor má překvapivě odlišné názory o některých konkrétních měřených veličinách. To platí obecně zejména při měření složek směsí. Nicméně jednotkou pro analýzu směsi by měl být mol a testy by měly měřit v poměru molů klinicky významného obsahu.

Klíčová slova: measurand, složka, analýza směsi, náhrada, domácí úkol.

Introduction

We will introduce in this essay three friendly people: i) the curious layperson, ii) the "All-Knowing" Clinical Biochemist, and iii) the "earth-bound" secretary (Fig.1).

Biochemist: Today it is about the measurand, an important term in metrology, and metrology is at the very basis of our discipline (sadly, often forgotten).

Layperson: So it is about rain and thunderstorms?

Biochemist: Please apologize my unclear pronunciation, it is metrology and NOT meteorology; but on second thought: indeed, it IS rain and thunderstorms!

Layperson: Uuuh, forgot my umbrella!

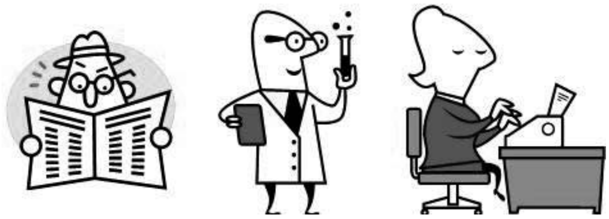

Fig. 1. The curious layperson, the "All-Knowing" Clinical Biochemist, and the "earth-bound" secretary.

Measurand

Biochemist: The definition of measurand = quantity intended to be measured [1, 2] and the preferred IUPAC-IFCC format for designations of quantities in laboratory medicine is "System-Component; kind-of-quantity"; example: "Plasma (Blood)-Sodium ion; amount-of-substance concentration equal to 143 mmol/l in a given person at a given time".

Layperson: So, if you intend to measure serum-creatinine, the component is creatinine?

Biochemist: Well, it depends (scratches his head); our discipline is somewhat divided on that. Some say, it's "alkaline-picrate-active substances" when measured by Jaffe-methods; some say it's "a color" when measured with Jaffe methods; most say it's creatinine.

Layperson: Oh I see, rain and thunderstorms!

Secretary: It's creatinine, I type.

Biochemist: And that was the easy one! In laboratory medicine, in fact, we measure very often component mixtures: total 25-hydroxy-vitamin D; total-triacylglycerides; thyroid stimulating hormone, and so one; the unit is, for example, mol/L and conversion to mass units (for example, g/L) is not possible (but, nevertheless done). A simple calculation clarifies that: 5 ng 25(OH)D_2 = 12.1 nmol/L, 5 µg/L 25(OH)D_3 = 12.5 nmol/L, thus 10 µg/L total 25(OH)D can either be 24.2 nmol/L, or 25 nmol/L or 24.6 nmol/L dependent on whether it represents the concentration of 25(OH)D_2, 25(OH)D_3 or a mixture of both.

Layperson: Uh, then your discipline is entirely lost?

Biochemist: Well (scratches his head), not entirely. We can create "surrogates" that we measure. For example: glycerol for total triacylglycerides; peptide fragments for proteins; virtual surrogates defined as epitopes present at certain regions of proteins. There is excellent reading available on that topic [3–5]!

Layperson: "Virtual surrogates"? Now I am lost!

Biochemist: Let's try to explain that by looking into what the brightest of them all has written [3]. The purist would say „*Insofar as the antigenic substances present in standards or test samples are dissimilar and/or molecularly heterogeneous, an immunoassay is invalid, and the results it yields have no universal significance. The only long-term solution to this problem is the development of assay systems measuring individual components of such heterogeneous mixtures*“ [3]. But the pragmatist may argue „*It is nevertheless possible to visualize circumstances in which an assay system, though analytically invalid in the strictest sense, responds only to a particular atomic group common to the molecules of substances differing in overall structure (for example, the protein moiety in TSH)*“ [3]. I take the pragmatist's point of view.

Layperson: Hear, hear: the pragmatist's point of view! But what problems do you encounter in practice?

Biochemist: First, we have to clearly relate the measurand to the diagnostic application of an assay. In the case of TSH, for example, a different measurand may be needed for TSH-secreting tumors (different glycosylation pattern). Second, it is important that each measured component contributes significantly to the diagnostic application of the assay. Third, small variations in mixture content in individual samples must be outweighed by the diagnostically important changes of the overall concentration of the mixture in healthy and sick persons [5].

Layperson: But what if the purists win?

Biochemist: Then we always have to include the respective antibody in the definition of the measurand and assays using different antibodies cannot be standardized and results cannot be compared. In turn, all immunoassay package inserts should include warnings such as those found in tumor-marker assays: *WARNING: CA 125 assay values obtained with different assay methods cannot be used interchangeably due to differences in assay methods and reagent specificity. The results reported by the laboratory to the physician must include the identity of the CA 125 assay used* [6].

Layperson (mumbling): Will they ever convince the purists?

Secretary: There is homework to do, I type.

Take-home messages

The VIM is FREELY available, download it!

The definition of a measurand includes system, component, and kind-of-quantity.

For what concerns the component, it is about what you intend to measure (“it's creatinine, I type”).

For what concerns component mixtures, there is homework to do; but the unit should be mol and tests should measure equimolar to the medically relevant extent.

Understand the purists (they provide the conceptual thinking); be pragmatic when possible (when it solves real life problems).

References

1. ISO/IEC Guide 99:2007. International vocabulary of metrology – Basic and general concepts and associated terms (VIM). International Organization for Standardization: Geneva, 2007 (VIM = Vocabulaire International de Métrologie).
2. JCGM 200:2012. International vocabulary of metrology – Basic and general concepts and associated terms (VIM). International Bureau of Weights and Measures (BIPM); Joint Committee for Guides in Metrology (JCGM): Paris, 2012. Freely available at: http://www.bipm.org/en/publications/guides/vim.html; See also, Clinical and Laboratory Standards Institute (CLSI): http://www.clsi.org/Content/NavigationMenu/Resources/HarmonizedTerminologyDatabase/Harmonized_Terminolo.htm.
3. **Ekins, R.** Immunoassay standardization. *Scand. J. Clin. Lab. Invest.* 1991, 51(Suppl 205), p. 33–46.
4. **Thienpont, L. M., Van Uytfanghe, K., De Leenheer, A. P.** Reference measurement systems in clinical chemistry [Review]. *Clin. Chim. Acta*, 2002, 323, p. 73-87.
5. **Thienpont, L. M., Van Houcke, S. K.** Traceability to a common standard for protein measurements by immunoassay for in-vitro diagnostic purposes. *Clin. Chim. Acta*, 2010, 411, p. 2058-61.
6. Package insert. Abbott Architect CA 125 II assay. Abbott laboratories: May 2007.

Do redakce došlo 29. 3. 2012

Adresa pro korespondenci
Dr. Dietmar Stöckl
STT Consulting
Abraham Hansstraat 11
B-9667 Horebeke
e-mail: dietmar@stt-consulting.com

Analytical Quality in the Medical Laboratory – The ASAP Concept. Part 4: Power

Stepman H.[1]**, Stöckl D.**[2]

[1]Ghent University, Faculty of Pharmaceutical Sciences, Laboratory for Analytical Chemistry.
[2]STT Consulting, Abraham Hansstraat 11, B-9667 Horebeke (Belgium), dietmar@stt-consulting.com

SUMMARY

In very basic terms, statistical power is the likelihood of achieving statistical significance. Three factors (effect-size, α, n), together with power, form a closed system – once any three are established, the fourth is completely determined. The goal of a power analysis is to find an appropriate balance among these factors by taking into account the substantive goals of a study. We exemplify the role of effect-size, α, and n on the power of a 1-sided F-test and give a general illustration of the power concept.

Keywords: Statistical power, sample size, effect size, α-error, imprecision.

SOUHRN

Stepman H., Stöckl D.: Analytická kvalita v klinické laboratoři – koncepce ASAP.
Část 4: statistická síla

V základním vyjádření je statistická síla pravděpodobnost dosažení statistické významnosti. Tři faktory (effect-size – velikost zamýšleného efektu), α, n), spolu se statistickou silou, tvoří uzavřený systém – pokud jsou všechny tři ustaveny, je tím zároveň i kompletně určen čtvrtý. Cílem silové analýzy je najít vhodnou rovnováhu mezi těmito faktory při zohlednění podstatných cílů studie. Ilustrujeme význam role velikosti zamýšleného efektu, α, a n na síle jednostranného F-testu, abychom poskytli celkový obrázek o pojetí konceptu statistické síly (testu).

Klíčová slova: statistická síla, velikost souboru, velikost účinku, α-chyba, nepřesnost

Introduction

In very basic terms, statistical power is the likelihood of achieving statistical significance [1]. In other words, statistical power is the probability of obtaining a p-value less than 0.05, for example. Obtaining $p < 0.05$ is exactly what many studies strive for, making the understanding of power calculations “a must”.

A power analysis is typically performed before a study is being planned. It is used to anticipate the likelihood that a study will yield a significant effect. Specifically, the larger the effect size, the larger the sample size, and/or the more liberal the criterion required for significance (α), the higher the expectation that the study will yield a statistically significant effect (= the higher the power will be) [1].

These three factors (effect-size, α, n), together with power, form a closed system – once any three are established, the fourth is completely determined. The goal of a power analysis is to find an appropriate balance among these factors by taking into account the substantive goals of the study, and the resources available to the researcher [1]. A more detailed explanation of the power concept can be found in a valuable internet resource [2].

Role of Effect Size, α, sample size, and imprecision

Effect size

The term „effect size“ refers to the magnitude of the effect under the alternate hypothesis. The nature of the effect size will vary from one statistical procedure to the next, but its function in power analysis is the same in all procedures.

The effect size should represent the smallest effect that would be of clinical, analytical, or other significance. In clinical trials for example, the selection of an effect size might take account of the severity of the illness being treated (a treatment effect that reduces mortality by one percent might be clinically important while a treatment effect that reduces transient asthma by 20% may be of little interest). It might take account of the existence of alternate treatments (if alternate treatments exist, a new treatment would need to surpass these other treatments to be important).

Alpha

Traditionally, researchers in some fields have accepted the notion that α should be set at 0.05 and power at 80% (corresponding to a β of 0.20). This notion is implicitly based on the assumption that a type I error is four times as harmful as a type II error (the ratio of β to α is 0.20/0.05 = 4), which has no basis in fact. Rather, it should fall to the researcher to strike a balance between α and β as befits the issues at hand. For example, if the study will be used to screen a new drug for further testing we might want to set α at 0.20 and power at 95%, to ensure that a potentially useful drug is not overlooked. On the other hand, if we were working with a drug that carried the risk of side effects and the study goal was to obtain FDA approval for use, we might want to set α at 0.01 while keeping power at 95%.

Sample Size

For any given effect size and α, increasing the sample size will increase the power.

Imprecision (variation in the data)

As always, high variation gives poor estimates (e.g., power), except the sample size is high. Note also, the standard deviation in power analysis is often taken from a pilot study. Therefore, it may be appropriate to calculate confidence intervals of power [3, 4].

Application (variance testing)

Assume you have an analytical method that samples volumetrically and you are not satisfied with its precision. You want to switch to gravimetrically-controlled sampling, because you want to IMPROVE precision. Typically, you work with 100 μl sample volume (=approximately 100 mg, for aqueous samples). You analyze 6 aliquots volumetrically sampled and 6 aliquots sampled with gravimetric control. The results are: SDgrav = 5, SDvol = 8 (note: this corresponds to the typically observed imprecision). The mean of both results is not relevant here. We investigate the results with a 1-sided F-test (you want to improve; you know gravimetric control should be better): n = 6; SDgrav = 5 (VARgrav = 25); SDvol = 8 (VARvol = 64). You obtain a *p*-value of 0.1627 for your 1-sided F-test. You are disappointed; gravimetric control may be better, but you could not demonstrate it.

Several questions may arise to get a significant test (α = 0.05):

1. What was the power of the initial experiment?
2. How many aliquots should have been measured at a given power (e.g., 0.9)?
3. How small should SDgrav have been with n = 6 at a given power (e.g., 0.9)?

We address these with the free G*Power software [5, 6] (other free software: [7, 8]).

1. What was the power of the initial experiment?

Choose Test Family: F-tests; Statistical test: Variance: Test of equality (two sample case); Type of power analysis: Post-hoc: Compute achieved power – given α, sample size, and effect size. Input needed is: Tails: One; Variance ratio (= 25/64 = 0.3906); Sample size (n_1 = n_2 = 6); α (= 0.05). The software calculates a power of 0.2369.

Conclusion: The power was roughly 24%.

2. How many aliquots should you have measured at a given power (e.g., 0.9)?

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size. Input needed is: Tails: One; Variance ratio (= 25/64 = 0.3906); power (=0.9); α (= 0.05). The software calculates a sample size of 41.

Conclusion: You should have measured 41 aliquots, each. Again, you are disappointed.

3. How small should SDgrav have been with n = 6 at a given power (e.g., 0.9)?

Type of power analysis: Sensitivity: Compute required effect size – given α, power, and sample size. Input needed is: Tails: One; power (=0.9); α (= 0.05); sample size (n1 = n2 = 6). The software calculates a variance RATIO of 0.0573. We assume that the variance of the old method was representative VARvol = 64, then VARgrav should have been 64 x 0.0573 = 3.67 and SDgrav should have been SQRT(3.67) = 1.92.

Conclusion: It is very difficult to demonstrate an improvement of imprecision with a low number of measurements (n <10)!

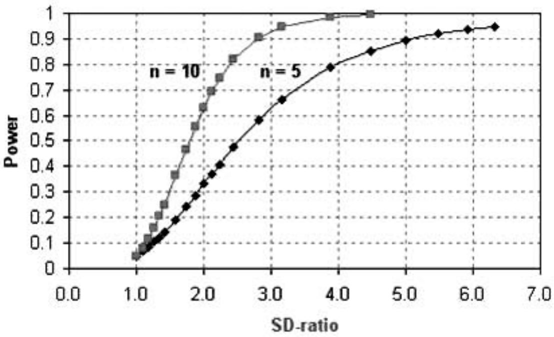

Fig. 1. shows that power curves for the F-test with low n are quite "flat" and reach a desirable power (e.g., p = 0.9) relatively late. For sufficient power of F-tests, sample sizes >10 are desirable.

Take-home messages

Statistical power is the likelihood of achieving statistical significance. Three factors (effect-size, α, n), together with power, form a closed system – once any three are established, the fourth is completely determined.

Power calculations are important BEFORE performing an experiment and should be part of the basic statistical knowledge in laboratory medicine.

When performing power analysis, we have to define the α-error (2-sided), first. Here, we define it at the 5%-level ($p = 0.05$, $z = 1.96$). Under null-hypothesis conditions, we get $p < 0.05$ in 5% of the cases, however, these are false positives (no effect introduced).

When we introduce an effect (here: shift of the population in k x standard error), the frequency of $p < 0.05$ increases:
50% at effect 1.96,
90% at effect 3.2415 (1.96 + 1.2816).
This is our power.
(1.2816 = z-value for 1-sided 90% probability)

We can graph the power versus the effect in a so-called power function. The corresponding power function is here the power function of the 2-sided z-test.

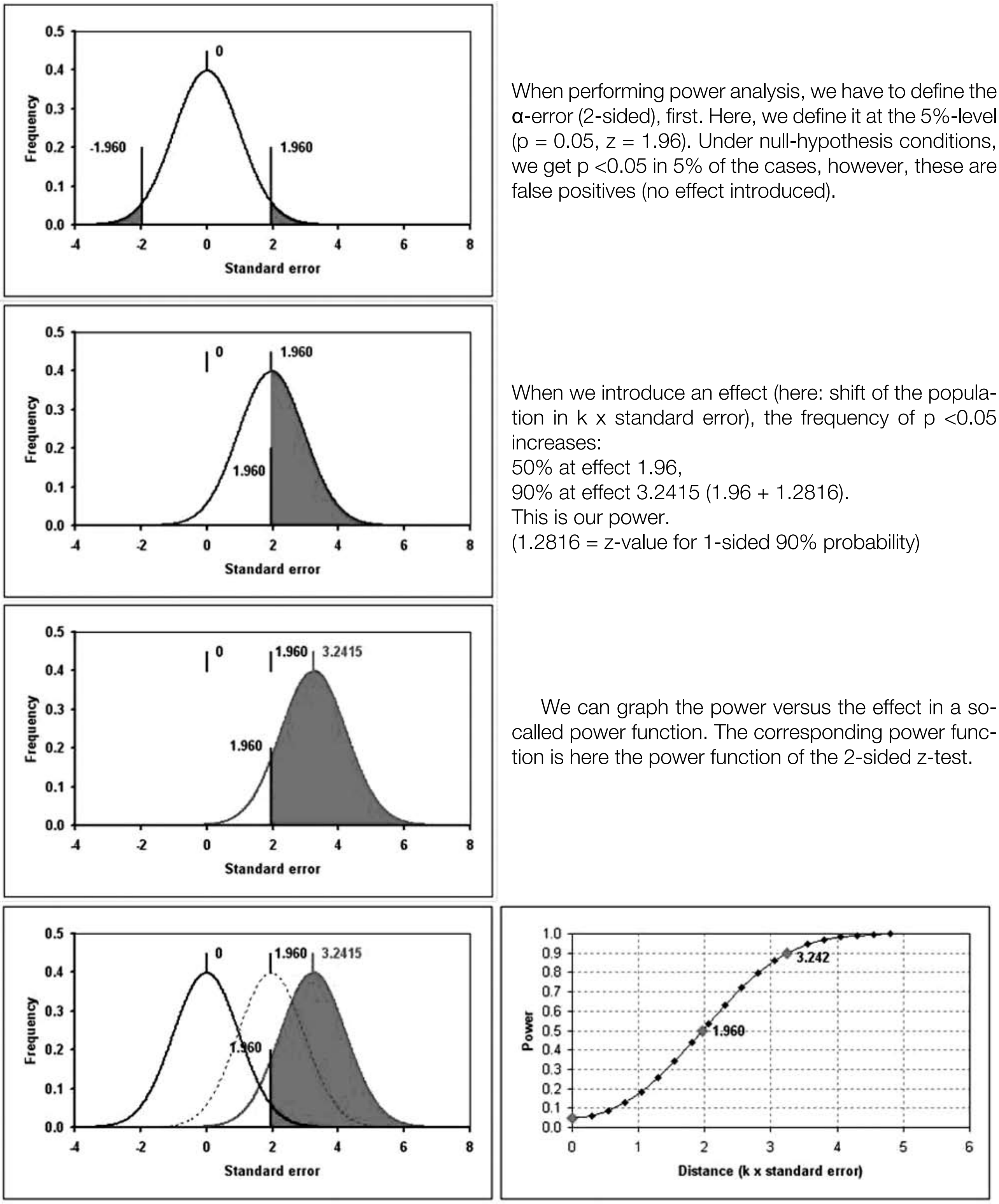

Fig. 2. Shows illustration of the power concept

References

1. http://www.power-analysis.com/power_analysis.htm
2. http://www.statsoft.com/textbook/power-analysis/
3. **Taylor, D. J., Muller, K. E.,** Computing confidence-bounds for power and sample-size of the general linear univariate model. *American Statistician*, 1995, 49, p. 43-7.
4. **Tarasinska, J.,** Confidence intervals for the power of Student's t-test. *Statistics & Probability Letters*, 2005, 73, p. 125-30.
5. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/
6. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html
7. http://www.statmethods.net/stats/power.html
8. http://www.cs.uiowa.edu/~rlenth/Power

Do redakce došlo 29. 3. 2012

Adresa pro korespondenci
Dr. Dietmar Stöckl
STT Consulting
Abraham Hansstraat 11
B-9667 Horebeke
e-mail: dietmar@stt-consulting.com