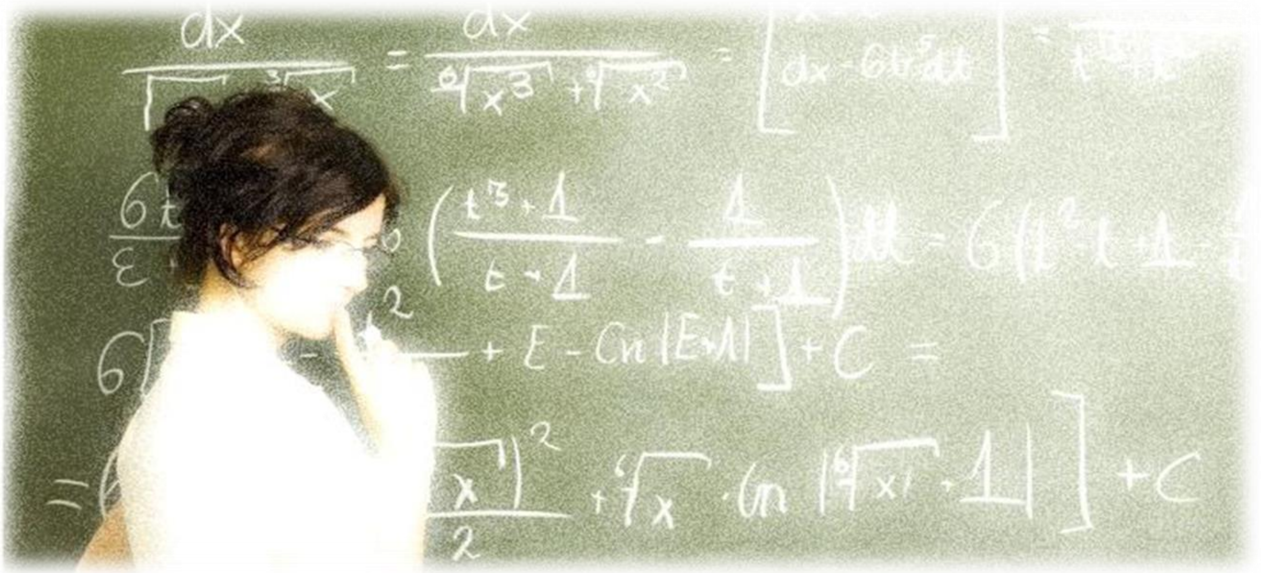
Welcome

- All lines will be muted during the presentation.
  - There will be interaction times when the lines are unmuted.
  - You can enter questions in the Q&A Box located on the right side of the screen at any time, and they will be addressed at the next break.

- This session is being recorded.
- We will send a link with the URL of the recording, and the PowerPoint slides to all participants after the presentation.

CMS Measure Development Informational Series



Reliability Testing – Concepts

January 22, 2014

Presenters:

Introduction by:

Karen Pace

National Quality Forum

Special Guest Speaker:

John L. Adams

Kaiser Permanente

Agenda

1. NQF Endorsement Criteria - Reliability

2. Reliability in measuring quality

3. How to choose a method of reliability testing

NQF Endorsement Criteria - Reliability

Karen Pace, PhD, MSN
Senior Director, Performance Measurement
National Quality Forum

NATIONAL QUALITY FORUM

Contact:
kpace@qualityforum.org

Conditions for Consideration

A. Measure Steward Agreement
   - All non-government organizations

B. Entity and process to maintain and update the measure as needed/at least every 3 years

C. Intended use of the measure includes accountability/public reporting as well as performance improvement

D. Measure is fully specified and tested for reliability and validity

E. Attests that harmonization and competing measures considered & addressed

F. Measure submission is complete – this is developer's presentation of the measure

Endorsement Criteria

- Major criteria describe desirable characteristics of quality performance measures for endorsement
- Hierarchy and Rationale
  - **Importance to measure and report** – measure those aspects with greatest potential of driving improvements; if not important, the other criteria less meaningful (must-pass)
  - **Scientific acceptability of measure properties** – goal is to make valid conclusions about quality; if not reliable and valid, risk of misclassification and improper interpretation (must-pass)
  - **Feasibility** – ideally, cause as little burden as possible; if not feasible, consider alternative approaches
  - **Usability and Use** – goal is to use endorsed measures for decisions related to accountability and improvement
  - If **competing measures**, select "best-in-class"
    If **related measures**, should be harmonized

2. Scientific Acceptability of Measure Properties
Must-pass criterion - must meet all subcriteria

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

2a. Reliability

- **2a1.** Precise specifications
- **2a2.** Empirical reliability testing

2b. Validity (and threats to validity)

- 2b1. Specifications consistent with evidence
- 2b2.Validity testing
- 2b3.-2b7.Testing/analysis related to threats to validity, e.g., exclusions, risk adjustment for outcomes)

2d. Composite performance measure – analysis of composite construction

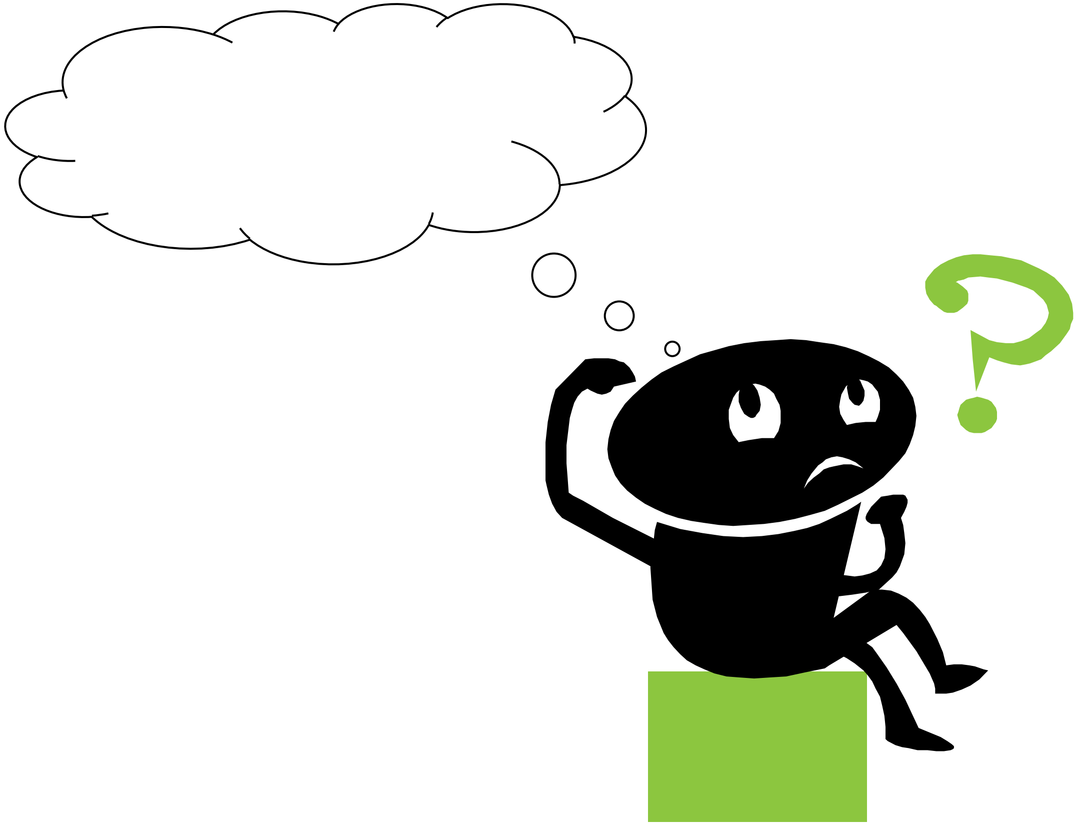
2a2. Reliability Testing

- Empirical testing conducted at level of either:
  - data elements used in the performance measure (e.g., inter-rater agreement on data elements used in the measure such as diagnosis, clinical value, intervention); or
  - computed performance scores for an accountable entity (e.g., signal-to-noise analysis of computed score such as percentage of patients who received the influenza vaccination)
- Updated evaluation guidance accepts testing at either level but testing at level of data elements only eligible for moderate rating; testing at level of performance score eligible for high rating
- Final evaluation rating depends on appropriate method, adequacy of sample, and result of testing

Resources

- NQF web pages submitting standards and measure evaluation
  - Document combining criteria plus guidance for evaluation
  - Examples of "what good looks like" for responses to measure submission items for evidence and measure testing
  - Measure Testing Task Force Report
  - Update of guidance for evaluating evidence, reliability, validity

Questions?



Measure Reliability Testing

Reliability in Provider Profiling

John L. Adams, Ph.D.

Principal Senior Statistician

Kaiser Permanente Center for Effectiveness & Safety Research

22 January 2014

What is provider profiling?

- Characterizing the quality of providers' service delivery:
    - How are individual physicians doing at making sure the patients they see are getting the care they need?
    - Which hospitals are best at avoiding readmissions?
    - How good is the quality of care at my health plan?

Some key references

- Reliability
  - Fleiss J, Levin B, Paik M. Statistical Methods for Rates & Proportions. Indianapolis, IN: Wiley-Interscience; 2003.
  - Hays RD, Revicki D. Reliability and validity (including responsiveness). In: Fayers P, Hays R, eds. Assessing Quality of Life In Clinical Trials. New York: Oxford University Press Inc.; 2005.
  - Shrout, PE, and Fleiss JL. (1979). "Intraclass correlations: Uses in assessing rater reliability". Psychol Bul 86 (2): 420–428. doi:10.1037//0033-2909.86.2.420.
  - Brennan RL, Generalizability theory. Springer-Verlag, 2001.
- HLM
  - Raudenbush SW, Bryk AS. Hierarchical Linear Models. Applications and Data Analysis Methods. Newbury Park, CA: Sage, 2nd ed., 2002.
- The reliability tutorial
  - Adams JL. The Reliability of Provider Profiling: A Tutorial. TR-653-NCQA. Santa Monica, CA: RAND, 2009. http://www.rand.org/pubs/technical_reports/TR653.html

Plan for the talk

- Defining reliability
- The primary importance of validity
- Reliability and other statistical measures
- Approach 1: The beta-binomial approach to calculating reliability
- Approach 2: The normal hierarchical modeling approach to calculating reliability
- Summary and questions

The fundamental definition

- Reliability: The squared correlation between a measurement and the truth
- Math notation:

$$\rho^2(measurement, truth)$$

- This would be easy to calculate if only we knew the truth!
- Most of the complications of reliability calculations come from various work arounds for not knowing the truth

A regression analogue

- If you could fit the regression model:

$$measurement = \beta_0 + \beta_1 truth + \varepsilon$$

- The R-squared from this regression would be the reliability

An equivalent definition that we will use

- The definition I find most useful is:

$$reliability = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

- Or with a more intuitive labeling:

$$reliability = \frac{\sigma^2_{Signal}}{\sigma^2_{Signal} + \sigma^2_{Noise}}$$

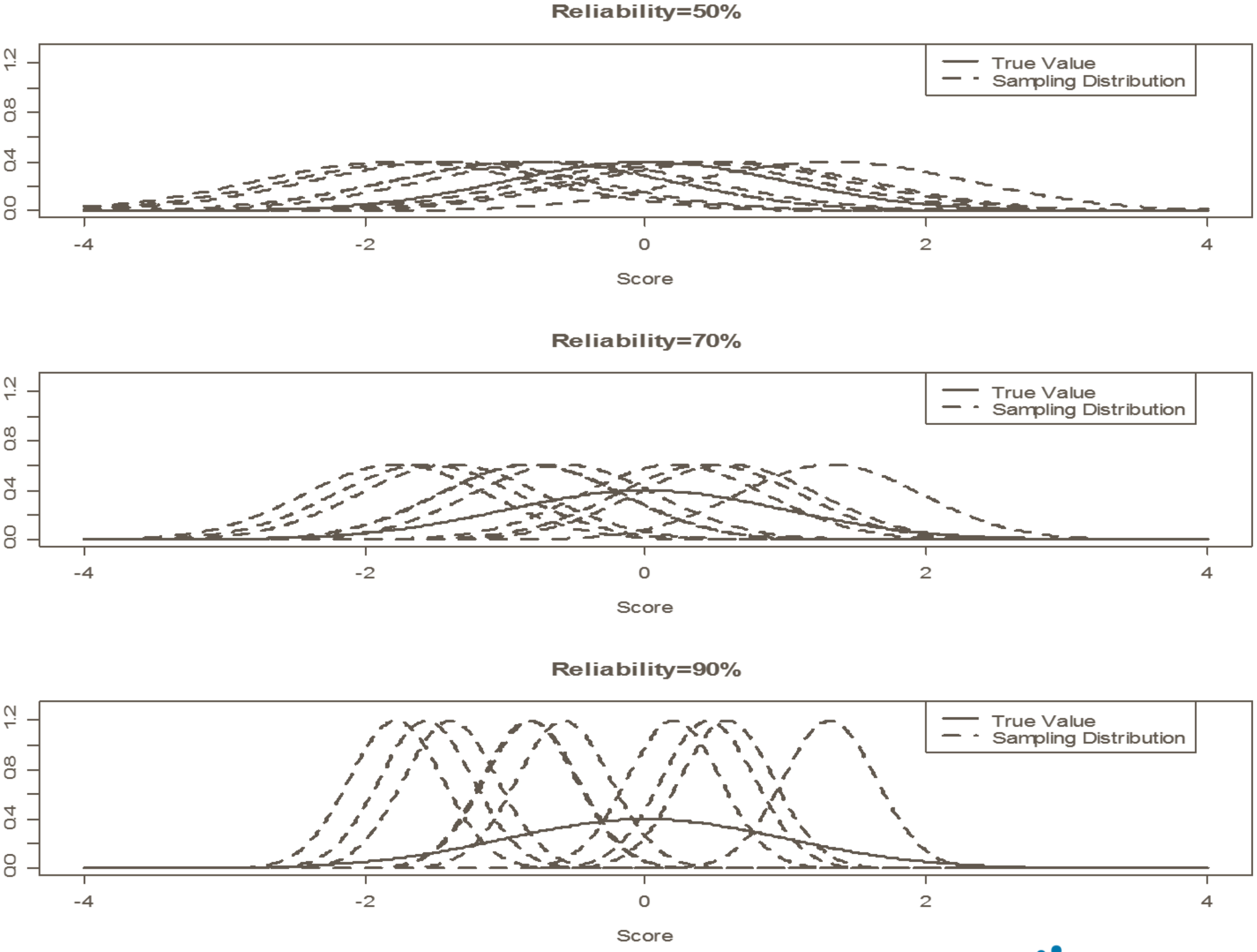
- Or made more specific to our setting:

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \sigma^2_{error}}$$

Here is a more detailed version for discussion

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \frac{\sigma^2_{error}}{n}}$$

What do different levels of reliability look like?



How do we get the reliability?

- We need a way to decompose the provider scores into provider-to-provider variation (signal) and noise
- This is usually done with something like an ANOVA model (old school) or a hierarchical model of some sort (new wave)
- Fit with mixed model (SAS) or specialty hierarchical (HLM) software
- This model can be extended in many ways
  - Fixed effects (e.g. case mix adjustment)
  - Hierarchy (MD within group within geography)

Why Should You Care About Reliability?

- Higher reliability increases the likelihood that you will assign a provider to the “right” group in a report card
  - Using low reliability information to drive behavior change could have undesirable consequences
- Sample size or standard errors, while often used as a proxy for reliability, may not be enough
  - So, minimum sample size or confidence interval requirements may not solve this problem

Is There a Minimum Level of Reliability?

- Psychometricians use a rule of thumb of 90% for drawing conclusions about individuals
- Lower levels (70-80%) are considered acceptable for drawing conclusions about groups
- Choice of level raises questions about the tradeoff between feasibility and scientific soundness

Some observations

- Reliability is often mistakenly thought of as a property of a measurement system (e.g. the SF-12 survey)
- The reason this common misunderstanding hasn’t made much trouble in other applications is that the other things that affect reliability are often held constant
- But reliability is a function of:
  - Provider-to-provider variation
    - And therefore depends on the population of providers!
  - Sample size
    - Which in many problems does vary from provider to provider

Why did this reliability stuff suddenly become important?

- Reliability is the measure of whether you can tell one provider from another
- There has recently been more interest in public reporting and pay for performance
- The focus has been on putting providers into categories
  - High performance networks
  - 1-5 star public reporting systems
  - Pay for performance programs
- Reliability tells you most of what you need to know about misclassification in these systems

What is different here from simpler reliability I learned in school?

- There are two features that are now different
  - Lack of balance
  - Heterogeneity
- Balance
  - In a typical survey measure (e.g. SF-12) everyone answers the same questions, each question only once
  - Here we don’t have balance because the number of observations can vary wildly from provider to provider
- Heterogeneity
  - This is different variances for each provider
  - Aggregate data often has different variances for each provider

Plan for the talk

- Defining reliability
- The primary importance of validity
- Reliability and other statistical measures
- Approach 1: The beta-binomial approach to calculating reliability
- Approach 2: The normal hierarchical modeling approach to calculating reliability
- Summary and questions

Validity

- Does the measurement measure what it claims to measure?
- If the answer is yes, the measure is valid
- Possible important questions in this context:
  - Is the measure controllable by the provider?
  - What about patient behavior?
  - Should the measures be case-mix adjusted?
  - Is it partially controlled by some other level of the system?
- Reliability ASSUMES validity

Getting the science right

- In the large validity is about getting the science right
- In empirical work this is often about building a defensible model

Consider what would happen if case-mix were not accounted for properly

- This formula would apply:

$$reliability = \frac{\sigma^2_{provider-to-provider} + \sigma^2_{case-mix}}{\sigma^2_{provider-to-provider} + \sigma^2_{case-mix} + \sigma^2_{error}}$$

- And reliability would appear to go up!
- This is a bad thing!
- This is why reliability depends critically on validity

CESR

KAISER PERMANENTE®

Plan for the talk

- Defining reliability
- The primary importance of validity
- Reliability and other statistical measures
- Approach 1: The beta-binomial approach to calculating reliability
- Approach 2: The normal hierarchical modeling approach to calculating reliability
- Summary and questions

Other reliability measures

- Test reliability
- Test-retest reliability
- Inter-rater reliability
- Cohen’s Kappa
- The intra-class correlation
- Cronbach’s alpha

Test-retest reliability

- Test-retest reliability compares a test and a retest separated in time
- This gives the world time to change between the measurements
  - Test conditions can change (e.g. different years)
  - Test subjects can change (e.g. practice evolves)
- Generally this will be an even lower bound for reliability
- This is an example of adding a facet (Brennan)

The intra-class correlation

- Simple measures like Kappa don’t generalize well to continuous measures
- Some measures are challenged by multiple raters and multiple scales
- Although there are several ways you could go the ICC is the most flexible generalization
- There is a famous ICC macro in SAS that calculates lots of ICCs
  - Think about correlation vs. squared correlation
  - Think about one item vs. the average of items at the provider level

Plan for the talk

- Defining reliability
- The primary importance of validity
- Reliability and other statistical measures
- Approach 1: The beta-binomial approach to calculating reliability
- Approach 2: The normal hierarchical modeling approach to calculating reliability
- Summary and questions

We will only consider simple pass/fail measures today

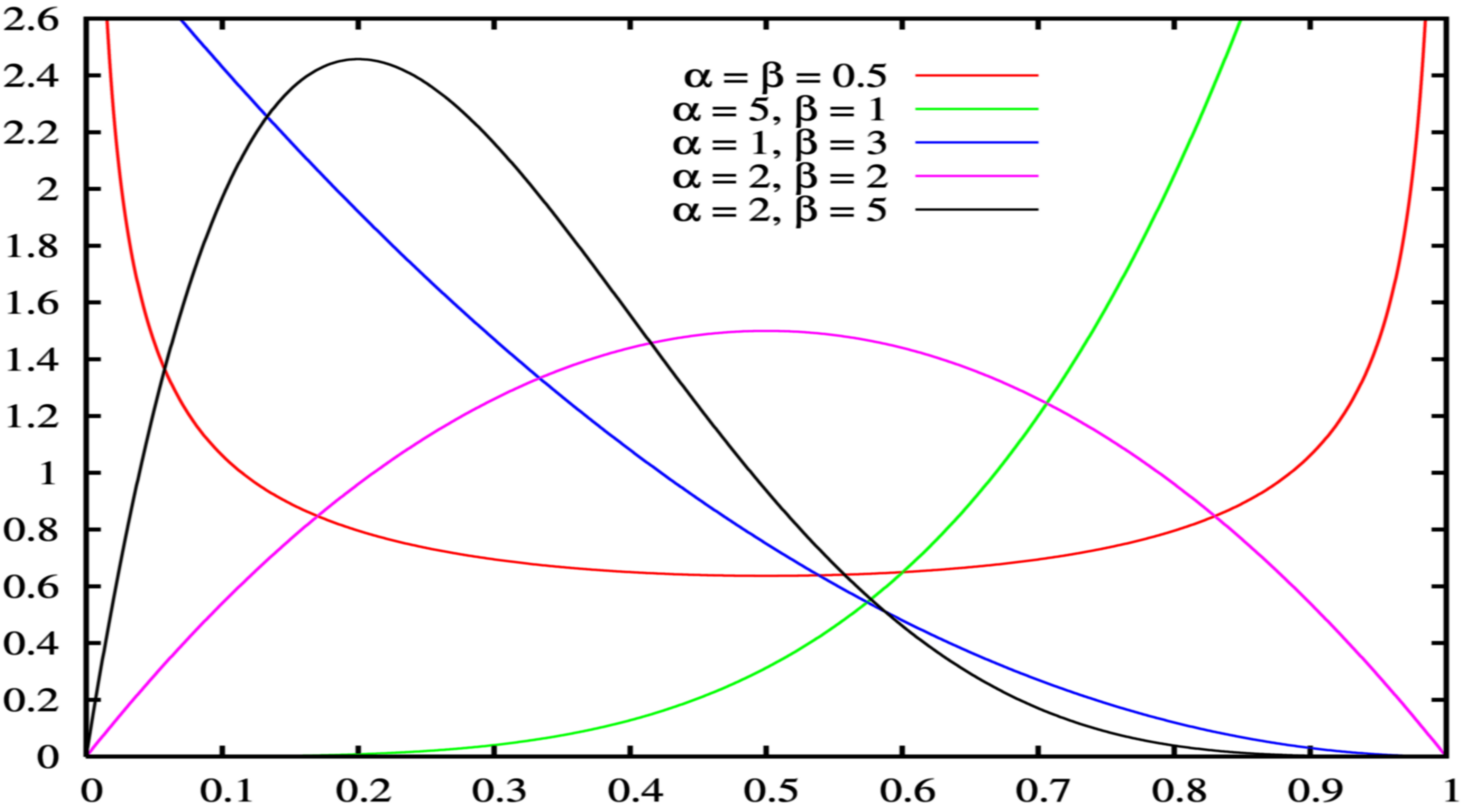
- Think of things like binary HEDIS© measures
  - Breast cancer screening
  - HbA1c testing for diabetics
- We will not talk today about how to case-mix adjust these measures
  - Could be important for things like readmission rates or measures with adherence drivers
- Everything here can be found in more detail in the reliability tutorial paper:
  - Adams JL. The Reliability of Provider Profiling: A Tutorial. TR-653-NCQA. Santa Monica, CA: RAND, 2009. http://www.rand.org/pubs/technical_reports/TR653.html

The beta-binomial model

- This is the most natural model for the reliability of pass/fail measures (e.g. HEDIS measures)
- The beta distribution
  - A distribution on the interval (0-1)
  - A very flexible 2 parameter distribution
  - Alpha and beta

$$\mu = \frac{\alpha}{(\alpha + \beta)} \qquad \sigma^2_{provider-to-provider} = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$$

What does the beta distribution look like?



How do you calculate the reliability from the beta-binomial?

- First you need to get the alpha and beta
  - From the fitting macro
- Then you need to calculate the provider variance:

$$\sigma^2_{provider\text{–}to-provider} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

- Then you need the usual binomial variance for the error:

$$\sigma^2_{error} = \frac{p(1-p)}{n}$$

Calculating the reliability

- The first step is to get an estimate of the provider-to-provider variance
- The best way I have found so far is:
  - MACRO BETABIN Version 2.2 March 2005
  - SUMMARY: Fits a Beta Binomial Model.
  - AUTHOR: Ian Wakeling - Qi Statistics
- As with all free software Caveate Emptor!
- I tested this by simulating datasets like those in the tutorial
- There is an example in the tutorial of a measure with a mean pass rate of 50% and all providers have a sample size of 10, I'll use that example in the next few slides.

Using the betabin macro output

- Remember the formula for the provider-to-provider variation:

$$\sigma^2_{provider\text{–}to-provider} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

- Then just plug in the numbers from the SAS output:

$$\sigma^2_{provider\text{–}to-provider} = \frac{4.5865 * 4.3862}{(4.5865+4.3862+1)(4.5865+4.3862)^2} = 0.025$$

- Just plug this and the provider’s error variance in the reliability formula

So the reliability depends on p!

This is different from the usual scale development situation

- No simple answer to the question: ”What is the reliability of my score?”
- The error variance depends on both the provider’s pass rate and the provider’s sample size
- Some cases:

Provider-to-provider variance	n	p	reliability
0.023	10	0.5	0.48
0.023	10	0.2	0.59
0.023	10	0.8	0.59
0.023	10	0.9	0.72

Plan for the Talk

- Defining reliability
- The primary importance of validity
- Reliability and other statistical measures
- Approach 1: The beta-binomial approach to calculating reliability
- Approach 2: The normal hierarchical modeling approach to calculating reliability
- Summary and questions

Start with a simple normal hierarchical model

- The usual HLM equation:

$$Score_i = P_i + \varepsilon_i$$

$$P_i \sim Normal(\mu, \sigma^2_{provider-to-provider})$$

- Where P_i is the true provider mean and ε_i is a normal error term with the provider variance (possibly heteroskedastic)

One way to fit this model is in SAS's proc mixed

- It can be a pretty ordinary problem if every provider has the same error variance (standard error) of their score
- It can be a tricky problem if the providers' have different error variances (and they often do)
  - You can use the GDATA trick in the tutorial
  - A knowledgeable SAS programmer or analyst can figure out other ways to do this
- But if you invest in learning how to do this the extension to case-mix adjustment or non-normal models is possible
- Similar models can be fit in Stata, Mplus, HLM, R, or other software

Just use the estimates from SAS

- Output from proc mixed:

```
Cov Parm      Estimate
PROVIDER      0.02507
Residual      0.2248
```

- This used the same data from the tutorial that was used with the beta-binomial example
  - Violates the normality assumptions
  - Gives about the same estimate of the provider-to-provider variance
- Reliability is calculated similarly to the beta-binomial example and results are very similar

Plan for the talk

- Defining reliability
- The primary importance of validity
- Reliability and other statistical measures
- Example 1 : The beta-binomial approach to calculating reliability
- Example 2: The normal hierarchical modeling approach to calculating reliability
- Summary and questions

So what should we use?

- The beta-binomial approach
  - Pros
    - Does the right thing in the unbalanced case
    - Is pretty fast compared to trying to get proc mixed to do the right thing
  - Cons
    - Not an everyday thing for most analysts
    - Does not extend to more complicated problems
- The normal HLM approach
  - Pros
    - Can be generalized to more complicated problems
    - Is more familiar to some analysts and programmers
  - Cons
    - Can be computationally intensive and a hassle

Questions?

In Summary

today we have covered:

1. NQF Endorsement Criteria - Reliability

2. Reliability in measuring quality

3. How to choose a method of reliability testing

Announcements

Joint Education Webinar with the Health Services Advisory Group (HSAG)

NATIONAL QUALITY FORUM

Contact:
emunthali@qualityforum.org

NQF Announcements

- **Upcoming Measure Submission Deadlines***
  - Admissions and Readmissions (February 5, 2014)
  - Health and Well-being (February 17, 2014)
  - Musculoskeletal (March 3, 2014)
  - Person and Family Centered Care-Phase 1 (March 14, 2014)
  - Surgery (March 17, 2014)
- **NEW! NQF's Measure Inventory Pipeline**
  - Links to submit a concept and to view submissions are available on *Submitting Standards* and *NQF Projects*
  - Contact measurepipeline@qualityforum.org
- **General information,** contact measuremaintenance@qualitforum.org

*Additional information about each project is available on *NQF Projects* page.

*All Measure Stewards must submit a fully-executed Measure Steward Agreement on/before the submission deadline.

Contact Information

Tennille Brown tennille.brown@cms.hhs.gov
410-786-5878

Katie Figueroa kfigueroa@hsag.com
602-801-6761

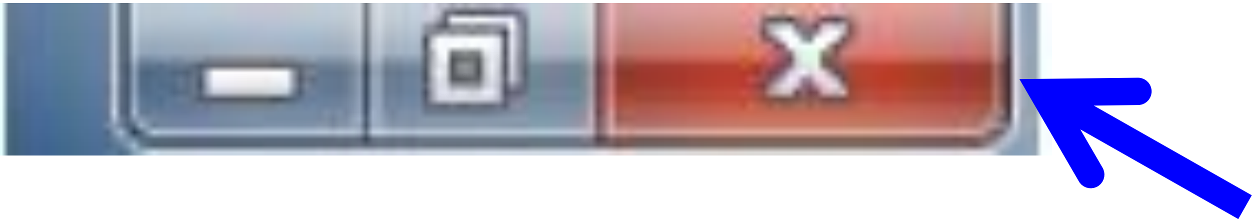
Beth Gualtieri bgualtieri@hsag.com
602-801-6756

Melba Hinojosa mhinojosa@hsag.com
602-801-6763

Elisa Munthali
emunthali@qualityforum.org

Karen Pace
kpace@qualityforum.org

Thank You



After you close the session by clicking **once** on the "X" in the upper right hand corner of your screen, the WebEx will present you with a Post-Activity Evaluation.

We value your input!