Sample Size and Power for Cost-Effectiveness Analysis

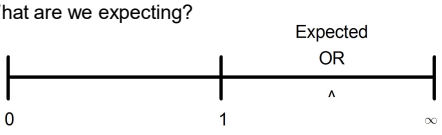
Henry Glick, Ph.D.

Outline

- Goals of sample size and power analysis
- Sample size
  - Formula
  - Similarities to and differences from clinical SS&PA
    - Correlation of difference
    - Willingness to pay
    - Standard deviations of cost and effect
- Power
  - Formula
  - Patterns of power
- (Briefly) Where to obtain the data
- Appendix: Sampling Uncertainty for Cost-Effectiveness

Goal of Clinical Sample Size Calculation

- Suppose we've estimated sample size for a change in odds of occurrence of a clinical outcome using $\alpha = 0.05$ and power = 0.8
- What are we expecting?

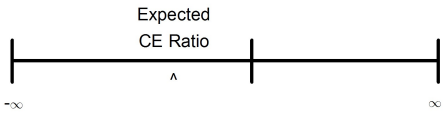


Goal of Clinical Sample Size Calculation (2)

- Expectation: Among repeated experiments, ORs' 95% CI exclude 1 from above in 80% of experiments and either include 1 or exclude 1 from below in 20% of experiments

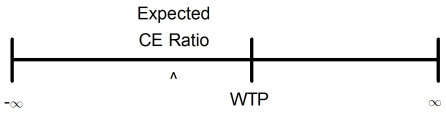
Goal of Cost-Effectiveness Sample Size Calculation

- Suppose we've estimated sample size for assessing cost-effectiveness using α = 0.05 and power = 0.8
- What are we expecting?



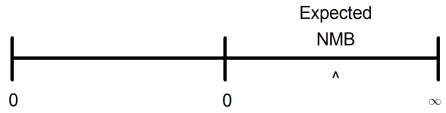
Cost Effectiveness Ratio

- Expectation: Among repeated experiments, ICERs' 95% CI exclude WTP/W from below in 80% of experiments and either include W or exclude W from above in 20%



Net Monetary Benefit

- Expectation: Among repeated experiments, NMBs' 95% CI exclude 0 from above in 80% of experiments and either include 0 or exclude 0 from below in 20%



Goal of Sample Size and Power Calculation

- More generally, sample size and power calculations allow us to conduct experiments with an expected likelihood that at conclusion of experiment we will be able to be confident in resulting comparison of costs and effects
  - e.g., may hypothesize that point estimate for cost-effectiveness ratio for therapy A will be 20,000 per QALY
  - May want to design an experiment that provides an 80% chance (i.e., power) of concluding with 95% confidence that therapy A is good value when we are willing to pay at most 75,000 per QALY

Sample Size / Power

- Sample size calculation
  - Given a desired alpha (α) and power (1-β), proactively manages probability of saying a difference exists when none does
    - Type 1 error; False positive; alpha; confidence
- Power Analysis
  - Given a desired alpha and a known sample size, proactively manages probability of saying no difference exists when one does
    - Type 2 error; False negative; (1-Beta); power
- "provides an 80% chance (power) of concluding with 95% confidence (alpha) that therapy is good value"

Other Cost-Effectiveness Sample Size Traditions

- Sample size approach described here comes out of frequentist statistical tradition
- Other approaches that have been discussed in cost-effectiveness literature include:
  - Bayesian (O'Hagan and Stevens)
  - Value of information (Koerkamp et al.)
  - Opportunity cost (Gafni et al.)
  - Decision model (Willan and O'Brien)

Sample Size Formula, Common SDs

- Assuming equal SDs and sample sizes, sample size formula is:

$$n = \frac{2\left(z_{\alpha}+z_{\beta}\right)^{2}\left(sd_{c}^{2} + \left(W\, sd_{q}\right)^{2} - \left(2\, W\, \rho\, sd_{c}\, sd_{q}\right)\right)}{\left(W\Delta Q - \Delta C\right)^{2}}$$

where n = sample size/group; z_{α} and z_{β} = z-statistics for α (e.g., 1.96) and β (e.g., 0.84) errors; sd = standard deviation for cost (c) and effect (q); W = maximum willingness to pay we wish to rule out; and ρ = correlation of difference in cost and effect; and (WΔQ-ΔC) = NMB

cess1i 200 .01 447.845 .01326715 -.71015 75000 .05 .8

Assumptions

Difference in costs:	200
Difference in effects:	0.01
Standard deviation, costs:	447.845
Standard deviation, effects:	0.01326715
Correlation, difference in costs and effects:	-0.71015
Willingness to pay:	75000
Two-tailed alpha level:	0.05
One-tailed beta level:	0.8
*** SAMPLE SIZE PER GROUP ***	95

Sample Size Supports Other α / β Pairs

- Yes, 95 participants per group support α=0.05 and power=0.8
- But what enters formula is sum of z_α and z_β $(z_\alpha + z_\beta)^2$
  - E.g., for α=0.05 and 1-β=0.8, 2.8016 (1.96 + .8416)
- 95 participants per group supports any z_α/z_β pair whose z-scores sum to 2.8016, e.g.,:

Alpha	Z_α	Power	z_β
0.01	2.5758	0.589	0.2258
0.03	2.1701	0.736	0.6315
0.05	1.9600	0.80	08416
0.075	1.7805	0.846	1.0211
0.10	1.6449	0.876	1.1567

Null Hypothesis, NMB

- Formula identifies a sample size that provides a 1-β% chance to have 1−α% confidence for rejection of null hypothesis that NMB (NMB = WQ − C) calculated by use of W equals 0
  - If assumptions about C, Q, sd_c, sd_q, and ρ are correct and if α=0.05 and 1-β=0.8, then with a sample size of 95 per group:
    - In approximately 800 of 1000 repeated experiments, lower limit of 95% confidence interval for difference in NMB will be greater than 0 (therapy represents good value)
    - In approximately 200, 95% confidence intervals will either include 0 or have an upper limit less than 0 (no difference in or bad value)

Null Hypothesis, CER and Acceptability

- Formula also identifies a sample size that provides a 1-β% chance to have 1−α% confidence for rejection of null hypothesis that cost-effectiveness ratio equals W (i.e., that 1−α% confidence interval for cost-effectiveness ratio excludes W)
- Or equivalently, identifies a sample size that provides a 1-β% chance for rejection of null hypothesis that at W, fraction of joint distribution of difference in cost and effect that is acceptable is greater than α/2% and less than 1−(α/2)%

Similarities With Clinical Sample Size Formulas

Error Rate

NMB Variance

$$n = \frac{2 \quad (z_\alpha + z_\beta)^2 \quad \left(sd_c^2 + \left(W\, sd_q\right)^2 - 2\, W\, \rho\, sd_c\, sd_q\right)}{\left(W\Delta Q - \Delta C\right)^2}$$

$$n = \frac{2 \quad (z_\alpha + z_\beta)^2 \quad \left(\quad sd_q^2 \quad \right)}{\Delta Q^2}$$

Difference^2

Differences in Formulas

$$Var_{NMB} = sd_c^2 + (W\, sd_q)^2 - (2\, W\, \rho\, sd_c\, sd_q)$$

- Variance of NMB more complicated than variance for usual continuous clinical differences
  - Includes ρ, correlation of difference between cost and effect
  - Includes W, decision threshold we are trying to rule out

$$\left(sd_c^2 + (W\, sd_q)^2 - (2\, W\, \rho\, sd_c\, sd_q)\right)$$

- Correlation of difference in cost and effect indicates how changes in difference in cost are related to changes in difference in effect
  - Negative (win/win) correlation: larger differences in effects associated with smaller differences in costs
    - e.g., asthma care: reductions in exacerbations leads to lower costs
  - Positive (win/lose) correlation: larger differences in effects are associated with larger differences in costs
    - e.g., life-saving care: increases in stroke survival may lead to higher care costs
- If W is positive, all else equal, larger positive correlations require fewer participants; larger negative correlations require more participants

Effect of Correlation on Sample Size

- If ΔC=200, ΔQ=.01; SDc= 447.845; SDq=.01326715; W=75,000; α=0.05; and 1-β=0.8:

Correlation	Sample Size
-0.50	85
-0.25	74
0.00	62
0.25	51
0.50	39
0.75	28

Willingness to Pay and Identification of an Appropriate Outcome Measure

- Sample size calculations require stipulation of W for a unit of outcome
- In many medical specialties, researchers use disease specific outcomes
- Can calculate a cost-effectiveness ratio for any outcome (e.g., cost/case detected; cost/abstinence day)
- But to be informative, outcome must be one for which we have recognized benchmarks of cost-effectiveness
  - Argues against use of too disease-specific an outcome for economic assessment

W and Point Estimate

- When W is greater than expected point estimate, resulting sample size and power allows us to be confident that MORE EFFECTIVE THERAPY is good value
  - Because confidence statements from these trials will be that point estimate for more effective therapy is less than willingness to pay
- When W is less than expected point estimate, resulting sample size and power allows us to be confident that MORE EFFECTIVE THERAPY is bad value
  - Because confidence statements from these trials will be that point estimate is greater than willingness to pay

Effect of Willingness to Pay (W)

- As already shown, direction of effect of correlation of difference is known
  - all else equal, more positive correlation, smaller sample size
- For W, no such consistent relationship exists
- Sample size approaches infinity and power approaches α/2 as expected point estimate approaches W
  - e.g., if W = 75,000, expected ΔC=15,000, and expected ΔQ=0.2, NMB ($W\Delta Q$-C) in denominator of sample size equation approaches 0
- Sample size reaches a minimum at what I refer to as widest definable interval which is uniquely defined for an experiment based on ΔC, SE_c, ΔQ. Se_q. and ρ

"Expected" Sample Size Table, W

- Common expectation: all else equal, larger W yields smaller sample size

	Sample Size Per Group
W	Exp 1 *
20,000	2050
30,000	1050
50,000	485
75,000	296
100,000	228
500,000	144

* ΔC=25; ΔQ=-0.01; sd_c=2500; sd_q=.03; ρ=0.0; α=.05; 1-β=.8

Sample Size Can Increase With Larger W

- Experiments can exist in which sample size increases as W increases

	Sample Size Per Group	
W	Exp 1	Exp 2*
20,000	2050	16
30,000	1050	21
50,000	485	28
75,000	296	45
100,000	228	76
500,000	144	896

* ΔC=-5000; ΔQ=0.01; sd_c=5000; sd_q=.15; ρ=-0.07; α=.05; 1-β=.8

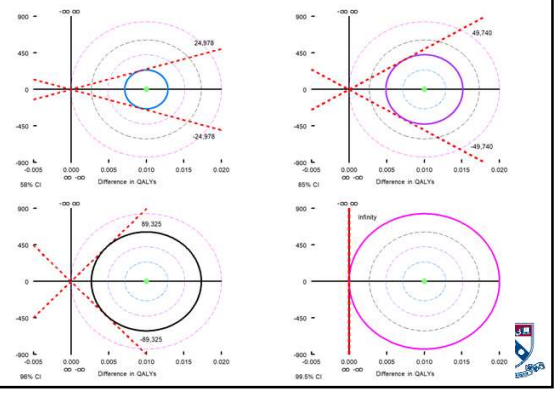
Sample Size Not Necessarily Monotonic With W

- Experiments can exist in which sample sizes decrease and then increase as W increases

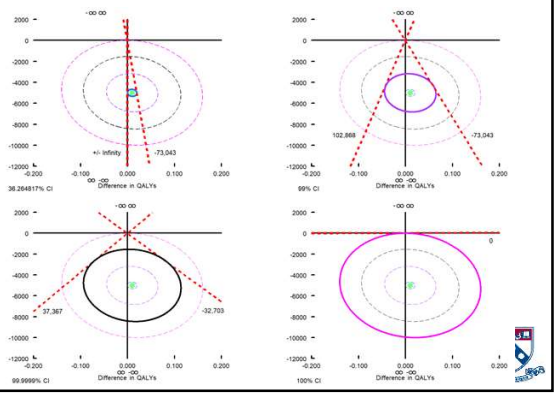
	Sample Size Per Group		
W	Exp 1	Exp 2	Exp 3*
20,000	2050	16	178
30,000	1050	21	158
50,000	485	28	***151***
75,000	296	45	153
100,000	228	76	156
500,000	144	896	170

* ΔC=-120; ΔQ=0.015; sd_c=1000; sd_q=.05; ρ=0.0; α=.05; 1-β=.8

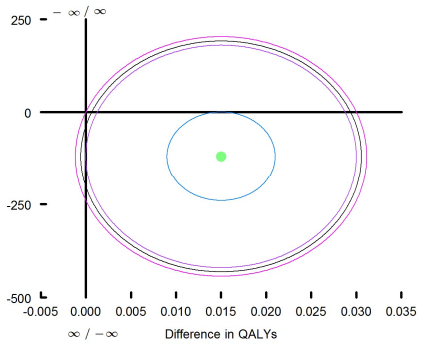
Experiment 1 Rationale (see samplesizelogic.pdf)



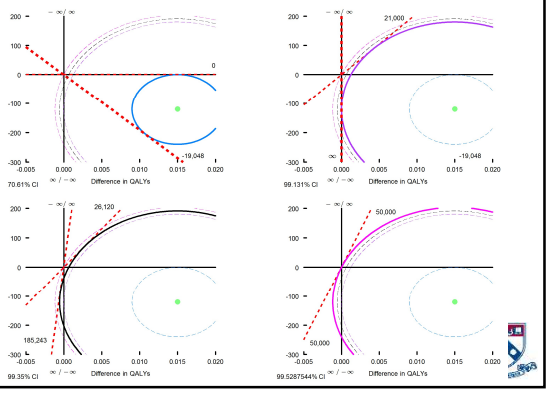
Experiment 2 Rationale (see samplesizelogic.pdf)



Experiment 3



Experiment 3 Rationale (see samplesizelogic.pdf)



Effect of SD_q VS SD_c on Sample Size

- All else equal, increases in SD_q or SD_c lead to increases in sample size
- Commonly thought that sample size for cost-effectiveness driven more by SD_c than by SD_q
  - If not, why do we commonly need a larger sample for cost-effectiveness outcome than for clinical outcome?
- However, if willingness to pay is substantially greater than SD_c, percentage changes in SD_q can have a substantially greater effect on sample size than will equivalent percentage changes in SD_c

Sample Size, SD_q, and SD_c

- Sample size formula is symmetric for SDs of cost and effect except for:

$$sd_c^2 + (W\ sd_q)^2$$

in numerator of equation

- Square of SD_q is weighted by square of W (e.g., $75{,}000^2$); square of SD_c is unweighted
  - So long as $W\ SD_q > SD_c$, SD_q will have a greater impact on sample size in face of equivalent percentage changes in SD_q and SD_c
  - E.g., if W=75,000 and SD_q=0.2; percentage changes in SD_q will have larger effect so long as SD_c<15000 (75,000*0.2)

Sample Size Tables, Relatively Large SD_c

- In this case with relatively larger SD_c's, equivalent percentage changes in SD_c and SD_q make no difference to required sample size for experiment

SD_c	N/Group	SD_q	N/Group
7500	390	0.1	390
15,000	635	0.2	635
22,500	1024	0.3	1024
30,000	1517	0.4	1517
45,000	3057	0.6	3057

ΔC=250; ΔQ=.0.05; unless otherwise specified, sd_c= 15,000 and sd_q=.2; ρ=-.1; W=75,000; α=.05; β=1-.8

Sample Size Tables, Relatively Small SD_c

- In this case with relatively smaller SD_c's, equivalent percentage changes in SD_c and SD_q yield substantially larger shifts in sample size given increases in SD_q

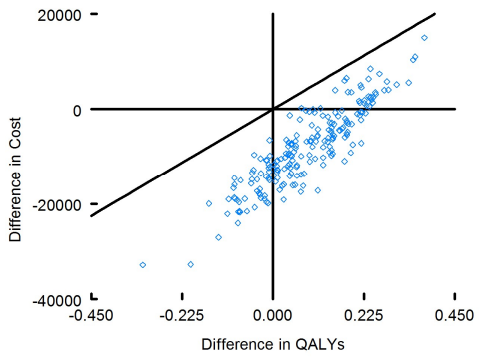
SD_c	N/Group	SD_q	N/Group
2500	306	0.1	114
5000	340	0.2	340
7500	389	0.3	710
10,000	455	0.4	1224
15,000	634	0.6	2685

ΔC=250; ΔQ=0.05; unless otherwise specified, sd_c= 5000 and sd_q=.2; ρ=-.1; w=75,000; α=.05; 1-β=.8

Economic Vs Clinical Sample Sizes

- Sample size required to answer economic questions typically considered larger than sample size required to answer clinical questions
  - But not necessarily in all cases
- ΔC and ΔQ are a joint outcome just as differences in nonfatal CVD events and all cause mortality are often combined into a joint outcome
- In same way that we can have more power for joint cardiovascular outcome than either individual outcome alone, we can have more power for cost-effectiveness than we do for costs or effects alone

What Can We Conclude About ΔC, ΔQ, Value?



What Can We Conclude?

- Difference in cost not significant
  - Because too large a fraction (>2.5%) of replicates above X axis ($0) and too small a fraction (<97.5%) below X axis
- Difference in effects not significant
  - Because too large a fraction (>2.5%) of replicates to left of Y axis (0 QALYs) and too small a fraction (<97.5%) to right of Y axis
- Can be 95% (100%) confident of value at specified WTP
  - Because all replicates fall below and to right of willingness to pay line
- There are some values of WTP where we can't be 95% confident (e.g., $0 and $∞)

Dropout and Sample Size

- Sample size estimates from formula appropriate if we expect no dropout from trial
- If we instead anticipate 10% dropout, divide sample size estimates by 0.9

Power Formula, Common SDs

- Assuming equal sds and sample sizes, power formula is:

$$z_\beta = \sqrt{\frac{n * (W\Delta Q - \Delta C)^2}{2\left(sd_c^2 + \left(W\, sd_q\right)^2 - 2\, W\, \rho\, sd_q\, sd_c\right)}} - z_\alpha$$

- Result is z_β, not power
- To estimate power, use normal distribution table to identify fraction of tail that is to left of z_β
  - Stata (V11+) code: power = normal(z_β)
  - E.g., -1.96 = 2.5% power; -0.84 = 20% power; 0 = 50% power; .84 = 80% power; 1.28 = 90%

cepow1i 200 .01 447.845 .01326715 -.71015 75000 .05 95

Assumptions

Difference in costs:	200
Difference in effects:	0.01
Standard deviation, costs:	447.845
Standard deviation, effects:	0.01326715
Correlation, difference in costs and effects:	-0.71015
Willingness to pay:	75000
Two-tailed alpha level:	0.05
Sample size per group:	95
*** POWER TO DETECT DIFFERENCE ***	0.802
z beta:	0.8471

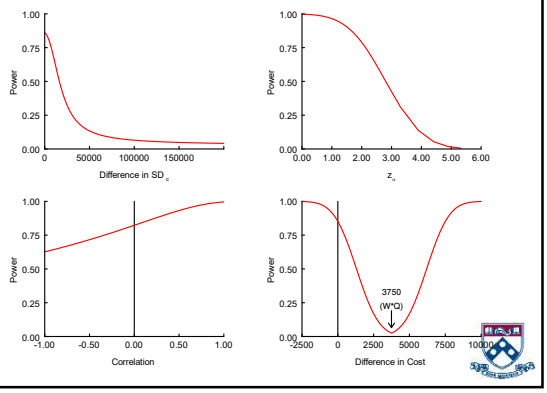
http://www.uphs.upenn.edu/dgimhsr/eeinct_ssandp.htm

Power Table for Varying Sample Sizes

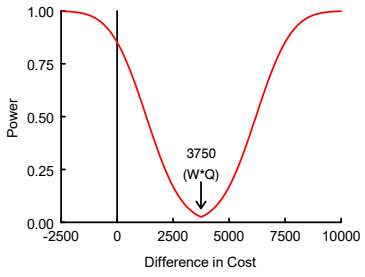
Sample Size	Power for W = 75,000
50	0.53
75	0.703
95	0.802
150	0.941
200	0.983

ΔC=200; ΔQ=0.01; sd_c=447.845; sd_q=.01326715; ρ=–.71015; w=75,000; and α=.05

Patterns of Power: SD, Z, ρ, ΔC (Q)



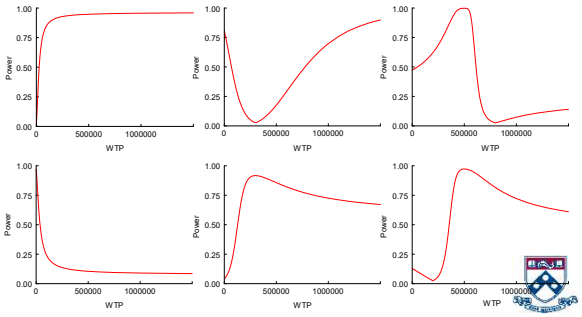
Power Doesn't Go to 0, No Matter Value of ΔC



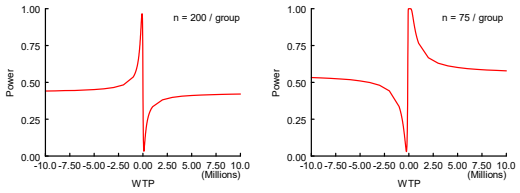
- Even when $\Delta C = W\Delta Q$ (i.e., NMB=0), we on average will still accidentally conclude the therapy is good value 2.5% ($\alpha/2$) of time

6 Patterns of Power for W>0

- Provided examples where sample size increased, decreased, or was v-shaped, but more patterns exist



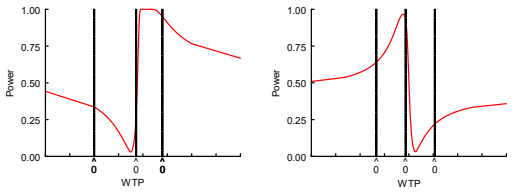
Two Underlying Patterns of Power for W



- When we plot power patterns for $-\infty \leq W \leq \infty$, there are 2 classes of patterns differentiated by whether value of W where power reaches a maximum is greater than or less than value of W where power reaches a minimum

Truncation at 0 Creates 6 Patterns

- Can see 6 patterns for values of W>0 because W can equal 0 at any point on each of 2 curves



- Pattern we observe depends on where 0 W falls

Dropout and Power

- If we anticipate 10% dropout, we will want to use "effective sample size" (e.g., 0.9 * 95) when we make our power calculations

Where to Obtain Necessary Data?

- When therapies are already in use: Expected differences in outcomes and standard deviations can be derived from feasibility studies or from records of patients
  - Potential sources
    - Medical charts of administrative data sets
    - Patient logs of their health care resource use
    - Asking patients and experts about kinds of care received by those with condition under study
  - Simple correlation between observed costs and effects may be an adequate proxy for measure of correlation used for estimating sample size

Obtaining Data for Novel Therapies

- For novel therapies, information about magnitude of incremental costs and outcomes may not be available
  - May need to be generated by assumption
  - Data on standard deviations for those who receive usual care/placebo may be obtained from feasibility studies or patient records
    - May want to assume sd from usual care (or a multiplier) will apply to new therapy, etc.

Summary

- Goal of sample size and power calculation for cost-effectiveness analysis is to identify likelihood that an experiment will allow us to be confident that a therapy is good or bad value when we adopt a particular willingness to pay
- Sample size and power depend on difference in cost and effect, SD of cost and effect, correlation of difference, our willingness to pay, and our target confidence level

Summary (2)

- When we estimate sample size or power, we often do so for varying levels of W
  - Sample size is undefined / power reaches a minimum when W equals point estimate for cost-effectiveness ratio (NMB=0)
- When W is substantially greater than SD for cost, changes in SD for effect generally have greater impact on sample size than do changes in SD for cost
- So long as W>0, positive correlations decrease sample size / increase power

Glick HA. Sample size and power for cost-effectiveness analysis (part 1). Pharmacoeconomics. 2011;29;189-98.

Glick HA. Sample size and power for cost-effectiveness analysis (part 2). The effect of maximum willingness to pay. Pharmacoeconomics. 2011;29:287-96.

Sample Size Formula, SDs Differ

- When SDs differ, formula becomes:

$$n = \frac{(z_{\alpha}+z_{\beta})^2 \left(\left(sd_{c0}^2 + sd_{c1}^2\right) + \left(W^2 \left(sd_{q0}^2 + sd_{q1}^2\right)\right) - \left(2\,W\,\rho \left(sd_{c0}^2 + sd_{c1}^2\right)^{0.5} \left(sd_{q0}^2 + sd_{q1}^2\right)^{0.5}\right)\right)}{(W\Delta Q - \Delta C)^2}$$

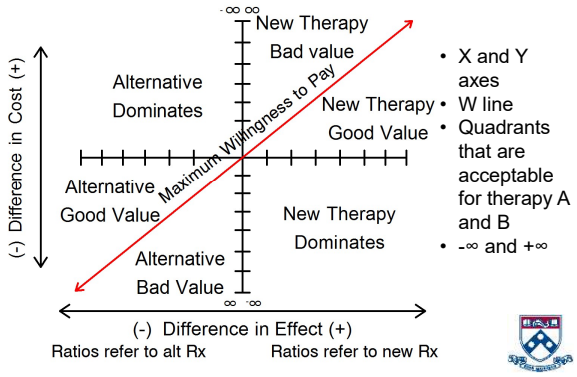
where n = n/group; tα/2 and tβ = t-statistics for α and β errors; sd = standard deviation for cost (c) and effect (q); W = maximum willingness to pay one wishes to rule out; and ρ = correlation of difference in cost and effect

Sampling Uncertainty Primer

Unidimensionality vs 2 Dimensionality

- Clinical outcomes typically are unidimentional and sampling uncertainty around these outcomes make sense on real number line
- Cost-effectiveness ratios are 2 dimensional and sampling uncertainty around these ratios can have unexpected properties on real number line
  - E.g., CI for ICER can include ∞ and -∞ but -∞ needn't represent lower bound of interval and ∞ needn't represent upper bound of interval
- Best to view results on 2D cost-effectiveness plane

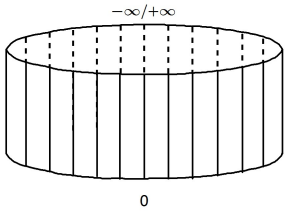
Cost-Effectiveness Plane and Value



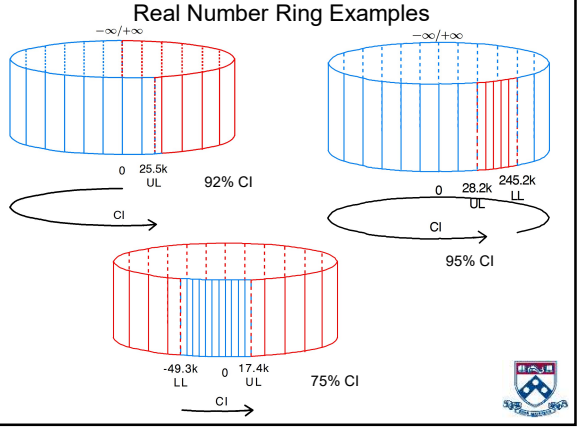
- X and Y axes
- W line
- Quadrants that are acceptable for therapy A and B
- -∞ and +∞

Real Number Ring

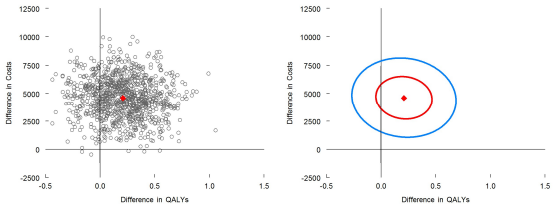
- If we want to move from CE plane to something like a real number line, overcome many problems if we map results onto real number ring



Real Number Ring Examples

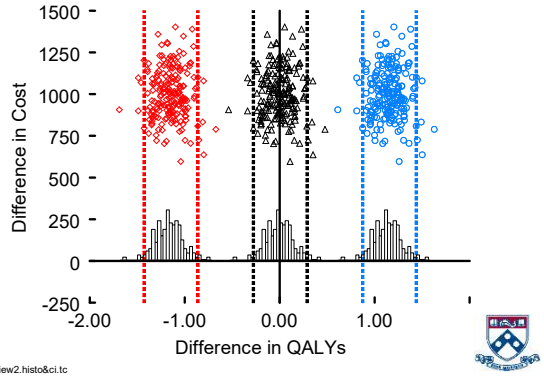


Results of Experiment on 2 Dimensional CE Plane



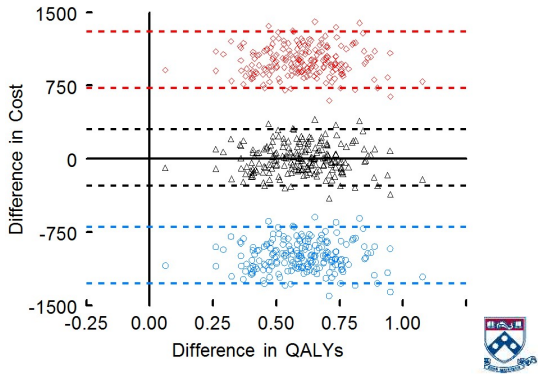
- Bootstrap of patient level data (left)
- Second order Monte Carlo (decision analysis with variables represented by distributions) (left)
- Bivariate normal curves (Δc, SEc, Δq, SEq, ρ) (right)

Plane Depicts Results of Effectiveness Analysis

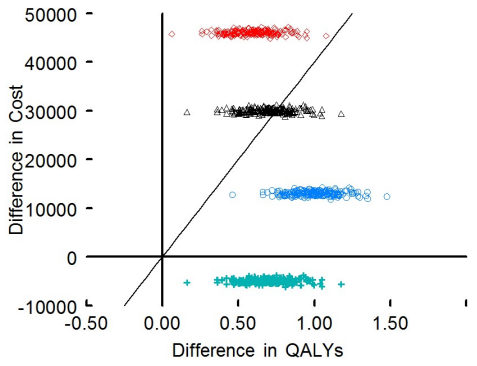


Overview2.histo&ci.tc

Plane Depicts Results of Cost Analysis



In Which Experiments Can We Be Confident of Value?



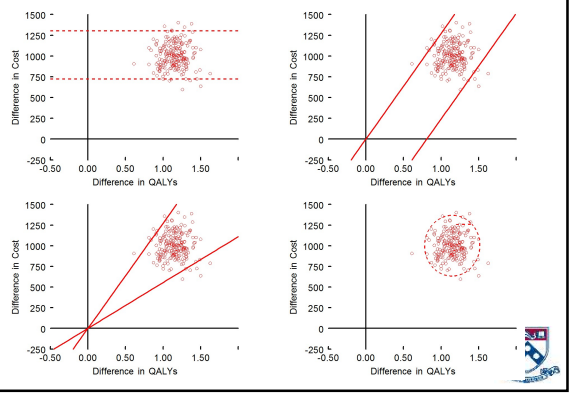
Red: confident of bad value
All points above W line

Black, not confident of value
Too many point on both sides of W line

Blue, confident of good value
All points below W line and above X axis

Cyan, confident of dominance
All points in lower right quadrant

95% CI for ICER?



95% CI

Upper left: CI for ΔC

Upper right: CI for NMB

Lower right: 95% confidence ellipse around point on C/E plane defined by ΔC and Δq (CE for point, not CI for ICER)

Lower left: 95% CI for ICER

Confidence Interval for ICER

- Because denominator of ratio can equal 0, there is no SE for ICER
- Thus CAN'T calculate ICER +/- 1.96 * SE_{ICER}
- CI for ICER defined as 0, 1, or 2 lines through origin of CE plane that exclude 2.5% of joint distribution of difference in cost and effect
- Fieller's theorem provides parametric equation for calculating these CI:

$$\frac{(\Delta C\Delta E - t_{\alpha/2}^2 \rho s_{\Delta C} s_{\Delta E}) \pm ([\Delta C\Delta E - t_{\alpha/2}^2 \rho s_{\Delta C} s_{\Delta E}]^2 - [\Delta E^2 - t_{\alpha/2}^2 s_{\Delta E}^2][\Delta C^2 - t_{\alpha/2}^2 s_{\Delta C}^2])^{0.5}}{\Delta E^2 - t_{\alpha/2}^2 s_{\Delta E}^2}$$

Net Monetary Benefit (NMB)

- NMB represents a transformation of ICER decision criterion ($W > \Delta C/\Delta Q$ becomes $W\Delta Q - \Delta C > 0$) which is linear and has a defined SE

$$SE_{NMB} = (SEc^2 + (W\ SEq^2) - 2\ W\ \rho\ SE_c\ SE_q)^{0.5}$$

- As with other differences, NMB is significant if it's CI excludes 0
- 95% CI_{NMB} = NMB +/- 1.96 * SE_{NMB}

$$NMB +/- 1.96\ (SEc^2 + (W\ SEq^2) - 2\ W\ \rho\ SE_c\ SE_q)^{0.5}$$

- If we set CI = 0, we can derive NMB CI equation from Fieller's theorem equation