Analysis of costs for Cost-Effectiveness Analysis

Henry Glick
University of Pennsylvania
International Conference on Health Policy Statistics
Philadelphia, PA
01/18/2008

Outline
• Policy-relevant parameter for cost-effectiveness
• Problems posed by nonnormality of cost data
• Common responses to violation of normality
  – Problematic responses
    • Nonparametric tests of characteristics of the distribution other than the arithmetic or sample mean
    • Transformation of the data
  – More promising responses
    • Nonparametric tests of sample means
    • Adoption of more flexible models for testing means
• General comments

Other Parameters
• Other summary statistics -- e.g., the difference in the median cost or the difference in the mean of the log of cost -- may be useful in describing the data, but do not provide information about the incremental cost that will be incurred by treating all patients nor the cost saved by treating patients with one therapy versus another
  – They thus are not associated with social efficiency
• Lack of symmetry of the cost distribution does not change the summary statistic of interest
• Evaluating some other difference, be it in medians or geometric means, simply because the cost distribution satisfies the assumptions of the tests for these statistics, may be tempting, but does not answer the question we are asking

The Problem
• Common feature of cost data is right-skewness (i.e., long, heavy, right tails)

The Problem (cont.)
• Distributions with long, heavy, right tails will have a mean that differs from the median, independent of "outliers"
• Cost data also can’t be negative, and can have large fractions of observations with 0 cost
• IMPLICATION: Nonnormality of cost data can pose problems for common parametric tests such as t-Test, ANOVA, and OLS regression

Parametric Tests Of Arithmetic (Sample) Means
• Lack of normality of cost data does not necessarily rule out use of t-tests, ANOVA, and OLS regression
  – In large samples t-tests have been shown to be robust to violations of this assumption when:
    • Samples are of similar size and skewness
    • Skewness is not too extreme
Common (Bad and Good) Responses To Violation Of Normality

- Adopt nonparametric tests of other characteristics of the distribution that are not as affected by the nonnormality of the distribution ("biostatistical" approach)
- Transform the data so they approximate a normal distribution ("classic econometric" approach)
- Adopt tests of arithmetic means that avoid parametric assumptions
- Adopt more flexible parametric models
- Propensity scores and other matching methods (Not discussed further)
- Bayesian analysis (Not discussed further)

Response 1: Non-parametric Tests of Other Characteristics of the Distribution

- Proposes to analyze the characteristics that are not as affected by the nonnormality of the distribution, e.g.,
  - Wilcoxon rank-sum test
  - Kolmogorov-Smirnov test

Nonparametric Tests

- Wilcoxon tests difference in medians
  - Estimates the probability that a randomly selected patient from one treatment group has a higher cost than a randomly selected patient from another treatment group
- Kolmogorov-Smirnov tests difference in the cumulative distribution function
  - Estimates whether the maximum absolute difference between two cumulative distribution function estimates are significant
- Tests indicate that some measure of the cost distribution differs between the treatment groups, such as its shape or location, but resulting p-values need not be applicable to the arithmetic mean

Why Median May Seem Useful?

- Unlike with cost / cost-effectiveness analysis, the median is the characteristic of the distribution that describes a number of important epidemiologic variables
  - E.g., relative risk and odds ratio

Evaluation of OR and RR

- Truth
  - Underlying risk =10%; in presence of risk factor , risk = 20%
    - RR = 2; OR = 2.25
- Experiment
  - Draw repeated samples from these populations
  - Calculate and save the observed RR and OR from each sample
  - Evaluate the mean and median of the distribution of RR and OR

Results of Simulation of RR and OR *

<table>
<thead>
<tr>
<th>N / Group</th>
<th>Median</th>
<th>Mean</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR (Truth = 2)</td>
<td>2.33</td>
<td>2.25</td>
<td>2.99</td>
<td></td>
</tr>
<tr>
<td>OR (Truth = 2.25)</td>
<td>2.22</td>
<td>2.26</td>
<td>2.56</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.07</td>
<td>2.26</td>
<td>2.56</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>2.04</td>
<td>2.26</td>
<td>2.34</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>2.01</td>
<td>2.26</td>
<td>2.30</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>2.01</td>
<td>2.26</td>
<td>2.27</td>
<td></td>
</tr>
</tbody>
</table>

* Statistics based on 10,000 draws for each sample size
**Median Not as Relevant for Cost**

- **Simulation**
  - Draw normal distributions of the log of cost (known means and SDs) for 4 populations (differentiated by mean and SD of log) and exponentiate them
  - To identify truth on the cost scale for each population draw 1000 samples of 2,000,000 observations
  - To evaluate different experiments, draw 20,000 samples with N's per simulated treatment group of 25, 50, 100, 250, 500, or 1000
- Calculate difference in mean and median within each sample
  - Report mean of resulting differences
  - Compare "observed" differences in each sample with true difference in means

---

**Simulated Truth**

<table>
<thead>
<tr>
<th>Log of Cost</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
<td>Mean</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4.5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

* Log of cost is normally distributed  
† Based on 1000 draws with 2,000,000 observations / draw

---

**True Difference Versus Observed Difference**

<table>
<thead>
<tr>
<th>Pop</th>
<th>N = 25 / group</th>
<th>N = 50 / group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ΔMean</td>
<td>ΔMed</td>
</tr>
<tr>
<td>1 vs 2</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>1 vs 3</td>
<td>124</td>
<td>96</td>
</tr>
<tr>
<td>1 vs 4</td>
<td>593</td>
<td>361</td>
</tr>
<tr>
<td>2 vs 3</td>
<td>77</td>
<td>60</td>
</tr>
<tr>
<td>2 vs 4</td>
<td>546</td>
<td>324</td>
</tr>
<tr>
<td>3 vs 4</td>
<td>469</td>
<td>265</td>
</tr>
</tbody>
</table>

* Mean of parameter values from 20,000 draws

---

**Rationale?**

- What rationales have been proposed to justify nonparametric tests of other characteristics of the distribution?
- Can’t be because the difference in sample means is a more biased estimate of the parameter of interest, because this difference is unbiased while the difference in medians is biased
- Rationale 1: Some have argued that when the data are nonnormal, the test of medians can be more efficient than the test of means
  - How can an efficiency claim for a biased estimator be a reason for adoption?
  - “Sufi looking under the wrong lamp” joke

---

**Mean Squared Error (Σ(O-T)^2)**

<table>
<thead>
<tr>
<th></th>
<th>ΔMean –</th>
<th>N = 25 / group</th>
<th>N = 50 / group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop</td>
<td>ΔMedian</td>
<td>Mean</td>
<td>Med</td>
</tr>
<tr>
<td>1 vs 2</td>
<td>12</td>
<td>1184</td>
<td>1060</td>
</tr>
<tr>
<td>1 vs 3</td>
<td>31</td>
<td>2686</td>
<td>3476</td>
</tr>
<tr>
<td>1 vs 4</td>
<td>245</td>
<td>59,260</td>
<td>129,24</td>
</tr>
<tr>
<td>2 vs 3</td>
<td>19</td>
<td>3188</td>
<td>2856</td>
</tr>
<tr>
<td>2 vs 4</td>
<td>233</td>
<td>59798</td>
<td>120,174</td>
</tr>
<tr>
<td>3 vs 4</td>
<td>214</td>
<td>61026</td>
<td>106,108</td>
</tr>
</tbody>
</table>

* (x0,000) Mean from 20,000 draws
Findings, Nonparametric Tests of Other Parameters

- The difference in the sample means is an unbiased estimate of the parameter of interest, whereas the difference in sample medians is not.
- The variability of the difference in sample means is generally larger than the variability in the difference in sample medians.
- If $$\sum (\Delta \text{Median} - \text{Truth})^2$$ is smaller than $$\sum (\Delta \text{Mean} - \text{Truth})^2$$, there may be a justification for using $$\Delta \text{Median}$$.
- If the log of cost is normally distributed, occurs only when the sample sizes are small and the true difference between the mean and median is small.

Response 2: Transform the Data

- Transform costs so they approximate a normal distribution.
  - Common transformations:
    - Log (arbitrary additional transformations required if any observation equals 0).
    - Square root.
  - Estimate and draw inferences about differences in transformed costs.

Estimates and Inferences Not Necessarily Applicable to Arithmetic Mean

- Goal is to use these estimates and inferences to estimate and draw inferences about differences in untransformed costs.
  - Estimation: Simple exponentiation of mean of log costs results in geometric mean, which is a biased estimate of the arithmetic mean.
  - Need to apply smearing factor.
  - Inference: On the retransformed scale, inferences about the log of costs translate into inferences about differences in the geometric mean rather than the arithmetic mean.

Estimation: The Smearing Factor

- "Smearing" factor attempts to eliminate bias from simple exponentiation of the mean of the logs.
  - When log is normally distributed:
    - If variances are equivalent can use Duan's common smearing factor.
    - If variances differ between groups, must use Manning's subgroup-specific smearing factor.
  - When log isn't normally distributed, one cannot rely solely on tests of difference in log variance to determine whether to use common or subgroup-specific smearing retransformation.

Potential Problems with Testing Transformation of the Data (I)

1. Log transformation doesn't always result in normality.
2. Resulting p-value has direct applicability to the difference in the log of cost and also generally applies to the difference in the geometric mean of cost.
   - The p-value for the log may or may not be directly applicable to the difference in arithmetic (sample) mean of cost.

Potential Problems with Testing Transformation (II)

- Whether the p-value for the log is applicable to the difference in the arithmetic mean of untransformed cost depends on whether the two distributions of the log are normal and whether they have equal variance and thus standard deviation.
  - If log cost is normally distributed and if the variances are equal, inferences about the difference in log cost are generally applicable to the difference in arithmetic mean cost.
  - If log cost is normally distributed and if the variances are unequal, inferences about the difference in log cost generally will not be applicable to the difference in arithmetic mean cost.
Response 3: Tests of Means that Avoid Parametric Assumptions

- Nonparametric bootstrap estimates the distribution of the observed difference in arithmetic mean costs

- Inferences based on the probability that the observed difference in means is significantly different from 0

Response 4: Adopt More Flexible Models

- Generalized Linear Models
  - These models have the advantages of the log models, but (a direct transformation of) $\Delta C$ is estimated, so they do not require any smearing correction
  - To build them, one must identify a "link function" and a "family" (based on the data)

The Link Function (I)

- Link function directly characterizes how the linear combination of the predictors is related to the mean on the original scale
  - e.g. identity, log, power # (square root, etc.)
- For example, if one uses a log link, one is assuming:
  $$\ln(E(y/x))=X\beta$$
- GLM with a log link differs from log OLS in part because in log OLS, one is assuming:
  $$E(\ln(y)/x)=X\beta$$
- GLM with a log link differs from log OLS in part because in log OLS, one is assuming:
  $$E(\ln(y)/x)=X\beta$$
- $\ln(E(y/x)) \neq E(\ln(y)/x)$
  i.e. log of the mean $\neq$ mean of the log costs

The Link Function (II)

- Log link is most commonly used in literature but may not be the best in all cases
- No single test identifies the appropriate link
- Can employ multiple tests of fit:
  - Pregibon link test checks linearity of response on scale of estimation
  - Modified Hosmer Lemeshow test checks for systematic bias in fit on raw scale
  - Pearson’s correlation test checks for systematic bias in fit on raw scale
  - Ideally, all 3 tests will yield nonsignificant p-values

The Family

- Specifies the distribution that reflects the mean-variance relationship
  - Gaussian
    • constant variance
  - Poisson
    • variance is proportional to mean
  - Gamma
    • variance is proportional to square of mean
  - Inverse gaussian
    • variance is proportional to cube of mean

Modified Park Test

- A “constructive” test that recommends a family given a particular link function
- Implemented after GLM regression assuming a link
GLM Comments (I)

- Advantages
  - None of the retransformation problems faced by log OLS models
  - Gains in precision from estimator that matches data generating mechanism
  - Consistent even if not the correct family distribution
    - Choice of family affects efficiency only if link function and covariates are correctly specified

GLM Comments (II)

- Disadvantages
  - Can suffer substantial precision losses
    - If heavy-tailed (log) error term, i.e., log-scale residuals have high kurtosis (>3)
    - If family is misspecified

Extended Estimating Equations

- Basu and Rathouz (2005) have proposed use of extended estimating equations (EEE) which estimate the link function and family along with the coefficients and standard errors
- Tends to need a large number of observations (thousands not hundreds) to converge
- P-values?

General Comments

- To reiterate, the difference in the sample means is the parameter of interest in cost-effectiveness analysis
- The distribution of cost can pose problems for common parametric tests of cost
- Responses in the literature that suggest that we should evaluate something other than the difference in the sample mean (or a direct transformation of this difference) – e.g., nonparametric tests of other characteristics of the distribution or transformations of cost – generally create more problems than they solve
- Nonparametric evaluation of the difference in the sample mean or use of more flexible models that evaluate a direct transformation of this difference generally pose fewer problems

Simulated Truth, Log Nonnormally Distributed

<table>
<thead>
<tr>
<th>Pop</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>SK</th>
<th>KT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>0.75</td>
<td>75</td>
<td>79</td>
<td>52</td>
<td>10.3</td>
<td>781</td>
</tr>
<tr>
<td>2</td>
<td>4.5</td>
<td>0.75</td>
<td>123</td>
<td>134</td>
<td>86</td>
<td>10.1</td>
<td>687</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.75</td>
<td>203</td>
<td>223</td>
<td>141</td>
<td>10.3</td>
<td>727</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>1.0</td>
<td>722</td>
<td>1552</td>
<td>378</td>
<td>57.6</td>
<td>23K</td>
</tr>
</tbody>
</table>

* Skewness, log of cost = 0.3; kurtosis, log of cost = 3.3
† Based on 1000 draws with 2,000,000 observations / draw
### True Difference Versus Observed Difference*

<table>
<thead>
<tr>
<th>Pop</th>
<th>True Diff</th>
<th>N = 50 / group</th>
<th>N = 100 / group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 v 2</td>
<td>48</td>
<td>48</td>
<td>49</td>
</tr>
<tr>
<td>1 v 3</td>
<td>128</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>1 v 4</td>
<td>648</td>
<td>648</td>
<td>647</td>
</tr>
<tr>
<td>2 v 3</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>2 v 4</td>
<td>599</td>
<td>600</td>
<td>598</td>
</tr>
<tr>
<td>3 v 4</td>
<td>519</td>
<td>520</td>
<td>518</td>
</tr>
</tbody>
</table>

*Mean of parameter values from 20,000 draws

### Mean Squared Error ($\Sigma(O-T)^2$)

<table>
<thead>
<tr>
<th>Pop</th>
<th>$\Delta$Mean, $\Delta$Median</th>
<th>N = 50 / group</th>
<th>N = 100 / group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 v 2</td>
<td>$539$ $384$</td>
<td>260</td>
<td>298</td>
</tr>
<tr>
<td>1 v 3</td>
<td>$1148$ $1862$</td>
<td>588</td>
<td>1676</td>
</tr>
<tr>
<td>1 v 4</td>
<td>$321$ $47K$</td>
<td>24K</td>
<td>104K</td>
</tr>
<tr>
<td>2 v 3</td>
<td>$25$ $1394$ $1050$</td>
<td>705</td>
<td>806</td>
</tr>
<tr>
<td>2 v 4</td>
<td>$307$ $48K$</td>
<td>24K</td>
<td>95K</td>
</tr>
<tr>
<td>3 v 4</td>
<td>$282$ $48K$</td>
<td>25K</td>
<td>81K</td>
</tr>
</tbody>
</table>

* (x0,000) Mean from 20,000 draws