Analysis of Patient-Level Cost Data
(With QALY Analysis Appendix)

Henry A. Glick, Ph.D.
Pharmacoeconomics
April 16, 2015
www.uphs.upenn.edu/dgimhsr

Outline

• Univariate analysis
  – Policy relevant parameter for CEA
  – Cost data 101
  – T-tests
  – Response to violation of normality
  – Primer on log cost
  – Why do different statistical tests lead to different inferences?
• Multivariable analysis
  – General linear models (GLM)

Policy Relevant Parameter for CEA

• In welfare economics, projects cost-beneficial if winners from any policy gain enough to be able to compensate losers and still be better off themselves
• Decision makers interested in total program cost/budget
• Policy relevant parameter quantifies how much losers lose, or cost, and how much winners win, or benefit
Policy Relevant Parameter for CEA (2)

- Whether or not data are skewed, sample mean * N provides unbiased estimate of population mean * N
  - Provides unbiased estimate of gains and losses
- When data are skewed, Median * N is a biased estimate of gains and losses

Initial advantage: sample mean (aka arithmetic mean)

Cost Data 101

- Cost data commonly right-skewed (i.e., long, heavy, right tails)
- Data tend to be skewed because:
  - Can have 0 costs, but not have negative costs
  - Most severe cases may require substantially more services than less severe cases
  - Certain events, which can be very expensive, occur in a relatively small number of patients
    - A minority of patients are responsible for a high proportion of health care costs

Typical Distribution Of Cost Data

Sk=1.04; Ku=4.9
Sk=1.52; Ku=9.2
Typical Distribution Of Cost Data (II)

• Heavy tails vs. “outliers”
  – Distributions with long, heavy, right tails will have larger sample means than medians

Problem Not Related Solely to “Outliers”

• Distribution when 5 observations with cost > 7200 (>3SD) are eliminated

Means and Medians When 5 Observations with Cost > 7200 are Eliminated

<table>
<thead>
<tr>
<th></th>
<th>Full Sample</th>
<th>Trimmmed *</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 0</td>
<td>Group 1</td>
</tr>
<tr>
<td>Mean</td>
<td>3015</td>
<td>3040</td>
</tr>
<tr>
<td>Median</td>
<td>2826</td>
<td>2901</td>
</tr>
</tbody>
</table>

* p = 0.003 and 0.000 for nonnormality of groups 0 and 1, respectively
“If the data are skewed, the mean doesn't tell us anything”

Do you agree?

Current wisdom about using parametric tests of means in cases where data are skewed??

??? Don’t analyze or report means ???
??? Analyze and report medians instead ???
What's rationale for analyzing and reporting medians instead of means?

Current Wisdom…

- Substitute nonparametric statistical tests for parametric tests because:
  - Data are skewed and Student’s t-test assumes normality
  - Data are skewed and OLS regression assumes normality of residuals?
  - In presence of skewness, distribution of mean likely to be much more variable than distribution of median
  - Others ???

Univariate Analysis: Parametric Tests Of Raw Means

- Usual starting point: T-tests and one way ANOVA
  - Used to test for differences in arithmetic means of total costs, QALYS, etc.
  - Makes assumption that costs are normally distributed
  - Normality assumption routinely violated for cost data, but t-tests have been shown to be robust to violations of this assumption when:
    - Samples moderately large
    - Samples are of similar size and skewness
    - Skewness is not too extreme
  - What is meant by “moderately large,” “similar size and skewness,” and “not too extreme”?
Univariate Analysis: Parametric Tests Of Raw Means

- Usual starting point: T-tests and one way ANOVA
  - Used to test for differences in arithmetic means of total costs, QALYS, etc.
  - Makes assumption that costs are normally distributed
  - Normality assumption routinely violated for cost data, but t-tests have been shown to be robust to violations of this assumption when:
    - Samples moderately large
    - Samples are of similar size and skewness
    - Skewness is not too extreme
  - What is meant by "moderately large," "similar size and skewness," and "not too extreme"?

Steps in Performing a T-test

- Evaluate whether or not outcome is normally distributed
  - sktest, joint test of skewness and kurtosis
  - Alternative tests:
    - swilk
    - sfrancia
- Evaluate whether or not standard deviations of costs for treatment groups are similar
- Perform t-test and interpret in relationship to prior two tests

Results of Tests of Normality and Equivalence of S.D. of Costs

<table>
<thead>
<tr>
<th>Test</th>
<th>p-value</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sktest, group 0</td>
<td>0.0</td>
<td>Failed</td>
</tr>
<tr>
<td>sktest, group 1</td>
<td>0.0</td>
<td>Failed</td>
</tr>
<tr>
<td>Equality of standard deviations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdtest</td>
<td>0.00</td>
<td>Failed</td>
</tr>
</tbody>
</table>
Results of T-Test

test cost, by(treat) unequal

Two-sample t test with unequal variances

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>95% Conf. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>250</td>
<td>3015</td>
<td>100.1052</td>
<td>1582.802</td>
<td>2817.839  3212.161</td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>3040</td>
<td>73.91742</td>
<td>1168.737</td>
<td>2894.417  3185.583</td>
</tr>
<tr>
<td>comb</td>
<td>500</td>
<td>3027.5</td>
<td>62.15917</td>
<td>1389.921</td>
<td>2905.374  3149.626</td>
</tr>
</tbody>
</table>

diff = mean(0) - mean(1)
So: diff = 0  Satterthwaite's degrees of freedom = 458.304

Pr(T < t) = 0.4204  Pr(|T| > |t|) = 0.8409  Pr(T > t) = 0.5796

Responses To Violation Of Normality Assumption

• Adopt nonparametric tests of other characteristics of distribution that are not as affected by nonnormality of distribution ("biostatistical" approach)
• Transform data to approximate normal distribution ("classic econometric" approach)
• Adopt tests of arithmetic means that avoid parametric assumptions (most recent development)

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

• Rationale: Can analyze characteristics that are not as affected by nonnormality of distribution
  – Wilcoxon rank-sum test
  – Kolmogorov-Smirnov test
Wilcoxon Rank-Sum

- Estimates probability that a randomly selected patient from one treatment group has a higher cost than a randomly selected patient from another treatment group.
- Referred to as a test of medians because frequency with which an Rx's patients have larger cost is unrelated to size of difference between patients' costs.
  - Rx 2 may be higher less of time, but when it is higher it may be much higher.

### Wilcoxon / Mann Whitney

<table>
<thead>
<tr>
<th>Group</th>
<th>Outcome</th>
<th>Rank</th>
<th>0 &gt; 1</th>
<th>1 &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>10</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>9</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>9</td>
<td>8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Means:
- 6.8 vs 8.0

Medians:
- 7 vs 5

Rank sum:
- 29 vs 26

Times greater:
- 14 vs 11

### Rank-Sum Test

`ranksum cost, by(treat)`

<table>
<thead>
<tr>
<th>treat</th>
<th>obs</th>
<th>rank sum</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>250</td>
<td>61183.5</td>
<td>62625</td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>64066.5</td>
<td>62625</td>
</tr>
<tr>
<td>combined</td>
<td>500</td>
<td>125250</td>
<td>125250</td>
</tr>
</tbody>
</table>

Unadjusted variance: 2609375.00
Adjustment for ties: 3.51
Adjusted variance: 2609371.49

Ho: cost(treat==0) = cost(treat==1)

z = -0.892

Prob > |z| = 0.3722
Kolmogorov-Smirnov

- Test of difference in cumulative distribution function
- Estimates whether maximum absolute difference between two cumulative distribution function estimates are significant

Cumulative Distribution

- Line 1 tests if group 0 has smaller values than group 1
- Line 2 tests if group 0 has larger values than group 1
- Line 3 provides a joint test

Kolmogorov-Smirnov Test

Two-sample Kolmogorov-Smirnov test for equality of distribution functions:

<table>
<thead>
<tr>
<th>Smaller group</th>
<th>D</th>
<th>P-value</th>
<th>Corrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:</td>
<td>0.1640</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>1:</td>
<td>-0.0640</td>
<td>0.359</td>
<td></td>
</tr>
<tr>
<td>Combined K-S:</td>
<td>0.1640</td>
<td>0.002</td>
<td>0.002</td>
</tr>
</tbody>
</table>

- ksmirnov cost, by(treat)
Potential Problem with Testing Other Characteristics of Distribution

- Tests indicate that some measure of cost distribution differs between treatment groups, such as its shape or location, but not necessarily that arithmetic means differ.
- Resulting p-values not necessarily applicable to arithmetic mean.

Rationales for Substituting Difference in Medians for Difference in Means?

- Can’t be because difference in sample means is a more biased estimate of difference in population means, because sample mean is unbiased while difference in sample medians is biased.
- What about fact that in presence of nonnormality, test of medians can be more efficient than test of means?
  - How important is efficiency of a biased estimator?
  - But...

Relative Bias Rationale

- Variability of difference in sample means is often larger than variability in difference in sample medians.
- Empirical question whether:
  \[
  \sum (\text{sample difference in means} - \text{true difference in means})^2 \\
  \sum (\text{sample difference in medians} - \text{true difference in means})^2
  \]
  
  - How important is efficiency of a biased estimator?
Are Sample Means Always Best Estimator?

- When cost data are sufficiently nonnormal, relative bias for median can be smaller than relative bias for arithmetic mean
  - e.g., can be shown in simulation that when log of cost is normally distributed, occurs only when sample sizes are small and true difference between mean and median is small
- Given that in actual data we never know truth, difficult to determine when other parameters will have lower relative bias than sample means
  - In part because degrees of both bias and skewness have to be taken into account

Response 2: Transform 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 Sample (Arithmetic) Mean

- Use these estimates and inferences from transformed costs to estimate and draw inferences about differences in untransformed costs
  - Estimation: Simple exponentiation of mean of log costs results in geometric mean, a downwardly biased estimate of arithmetic mean
  - Need to apply smearing factor
  - Inference: On retransformed scale, inferences about log of costs translate into inferences about differences in geometric mean, not arithmetic mean

Results of T Test of Log

ttest lcost, by(treat) unequal

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>250</td>
<td>7.863486</td>
<td>0.0364313</td>
<td>0.57603</td>
<td>7.791734 7.935239</td>
</tr>
<tr>
<td>1</td>
<td>250</td>
<td>7.95014</td>
<td>0.023952</td>
<td>0.3787148</td>
<td>7.902965 7.997314</td>
</tr>
</tbody>
</table>

comb | 500   | 7.906813| 0.0218642 | 0.4888993 | 7.863856 7.94977    |

diff | 0.0866533 | 0.0435998 | 0.1723483 | 0.009583  | t = -1.9875    |

Ho: diff = 0  Satterthwaite's degrees of freedom = 430.373

Ha: diff < 0  Pr(T < t) = 0.0238  Pr(|T| > |t|) = 0.0475  Pr(T > t) = 0.9762
Ha: diff > 0  Pr(T < t) = 0.9762  Pr(|T| > |t|) = 0.9525  Pr(T > t) = 0.0475

Primer On Log Transformation Of Costs
Log Transformation of Cost

<table>
<thead>
<tr>
<th>Raw Cost</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs:</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>87</td>
</tr>
<tr>
<td>Arith mean</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>Log of arithmetic mean</td>
<td>3.8918203</td>
<td>3.8918203</td>
</tr>
<tr>
<td>Geometric mean</td>
<td>38.8694</td>
<td>47.2554</td>
</tr>
<tr>
<td>Log Cost</td>
<td>Obs:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.708050</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.806663</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4.465908</td>
</tr>
<tr>
<td>Arithmetic mean of logs</td>
<td>3.660207</td>
<td>3.855568</td>
</tr>
<tr>
<td>( \exp(\text{mean of logs}) )</td>
<td>38.8694</td>
<td>47.2554</td>
</tr>
</tbody>
</table>

Downward Bias of Geometric Mean

- Exponentiation of mean of logs yields geometric mean
- In presence of variability in costs, geometric mean downwardly biased estimate of arithmetic mean
  - All else equal, greater variance, skewness, or kurtosis, greater downward bias
  - e.g., \((25 \times 30 \times 35)^{0.333} = 29.7196\)
    \((10 \times 30 \times 50)^{0.333} = 24.6621\)
    \((5 \times 30 \times 55)^{0.333} = 20.2062\)
    \((1 \times 30 \times 59)^{0.333} = 12.0664\)
- "Smearing" factor attempts to eliminate bias from exponentiation of mean of logs

Retransformation Of Log Of Cost (I)

- Duan's common smearing factor:
  \[ \phi = \frac{1}{N} \sum_{i=1}^{N} e^{\hat{z}_i - \bar{z}} \]
  where in univariate analysis, \( \hat{z}_i \) = group mean
- Most appropriate when treatment group variances are equivalent
Retransformation Of Log Of Cost (II)

<table>
<thead>
<tr>
<th>Group</th>
<th>Observ</th>
<th>ln</th>
<th>Z^2</th>
<th>e^(Z^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>2.708050</td>
<td>-0.9026834</td>
<td>0.4054801</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3.806663</td>
<td>0.1959289</td>
<td>1.216440</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4.317488</td>
<td>0.7067545</td>
<td>2.027401</td>
</tr>
<tr>
<td>Mean, 2</td>
<td>--</td>
<td>3.610734</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3.555348</td>
<td>-0.2344332</td>
<td>0.7910191</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.806663</td>
<td>0.0168812</td>
<td>1.017025</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4.007333</td>
<td>0.2175519</td>
<td>1.243030</td>
</tr>
<tr>
<td>Mean, 3</td>
<td>--</td>
<td>3.789781</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Smear: 1.116732

Common Smearing Retransformation (I)

- Retransformation formula
  
  \[ E(Y) = \Phi e^{\mu} \]

- Retransformation

<table>
<thead>
<tr>
<th>Group</th>
<th>( \Phi )</th>
<th>Predicted Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.116732</td>
<td>36.993</td>
</tr>
<tr>
<td>3</td>
<td>1.116732</td>
<td>44.247</td>
</tr>
</tbody>
</table>

Common Smearing Retransformation (II)

- Why are retransformed subgroup-specific means -- 41.3 and 49.4 -- so different from untransformed subgroup means of 45?
  - Because standard deviations of subgroups' logs are substantially different
    
    \( SD_2 = 0.8224; SD_3 = 0.2265 \)
  - Larger standard deviation for group 2 implies that compared with arithmetic mean, its geometric mean has greater downward bias than does geometric mean for group 3
  - Thus, multiplication of 2 groups' geometric means by a common smearing factor cannot give accurate estimates for both groups' arithmetic means
Subgroup-specific Smearing Factors (I)

- Manning has shown that in face of heteroscedasticity - i.e., differences in variance -- use of a common smearing factor in retransformation of predicted log of costs yields biased estimates of predicted costs
- We obtain unbiased estimates by use of subgroup-specific smearing factors
- Manning's subgroup-specific smearing factor:
  \[ \phi_j = \frac{1}{N_j} \sum_{i=1}^{N_j} e^{(z_{ij} - \hat{z}_j)} \]

Subgroup-specific Smearing Factors (II)

<table>
<thead>
<tr>
<th>Group</th>
<th>Observ</th>
<th>ln (z - (z_i))</th>
<th>(e^{(z - \hat{z}_i)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>2.708050</td>
<td>-0.9026834</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3.806663</td>
<td>0.1959289</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4.317488</td>
<td>0.7067545</td>
</tr>
<tr>
<td>Mean, 2</td>
<td>--</td>
<td>3.610734</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3.555348</td>
<td>-0.2344332</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3.806663</td>
<td>0.0168812</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4.007333</td>
<td>0.2175519</td>
</tr>
<tr>
<td>Mean, 3</td>
<td>--</td>
<td>3.789781</td>
<td>--</td>
</tr>
</tbody>
</table>

Subgroup-specific Smearing Retransformation (I)

- Retransformation formulas
  \[ E(\hat{Y}_i) = \phi_j e^{\hat{z}_j} \]
  \[ E(\hat{Y}_i) = \phi_j e^{\hat{z}_j} \]
- Retransformation

<table>
<thead>
<tr>
<th>Group</th>
<th>(\phi_j)</th>
<th>(e^{\hat{z}_j})</th>
<th>Predicted Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.21644</td>
<td>36.993</td>
<td>45.00</td>
</tr>
<tr>
<td>3</td>
<td>1.0170245</td>
<td>44.247</td>
<td>45.00</td>
</tr>
</tbody>
</table>
Subgroup-specific Smearing Retransformation (II)

- All else equal, in face of differences in variance (or skewness or kurtosis), use of subgroup-specific smearing factors yields unbiased estimates of subgroup means.
- Use of separate smearing factors eliminates efficiency gains from log transformation, because cannot assume p-value derived for log of cost applies to arithmetic mean of cost.

Potential Problems with Substituting Transformed Data for Raw Data (I)

- Log transformation doesn’t always result in normality.

Potential Problems with Substituting Transformed Data for Raw Data (II)

- P-value from t-test of log cost directly applies to difference in log of cost.
- Generally also applies to difference in geometric mean of cost.
  - Observe similar p-values for difference in log and difference in geometric mean.
- P-value for log may or may not be directly applicable to difference in arithmetic mean of untransformed cost.
Potential Problems with Substituting Transformed Data for Raw Data (III)

- Applicability of p-value for log to difference in arithmetic mean of untransformed cost depends on both distributions of log being normal and having equal variance and thus standard deviation
  - If log normally distributed and variances equal, inferences about difference in log generally applicable to difference in arithmetic mean
  - If log either not normally distributed or variances unequal, inferences about difference in log generally not applicable to difference in arithmetic mean

Response 3: Tests of Means that Avoid Parametric Assumptions

- Bootstrap estimates distribution of observed difference in arithmetic mean costs
- Yields a test of how likely it is that 0 is included in this distribution (by evaluating probability that observed difference in means is significantly different from 0)

Implementation of Bootstrap

- Random draw with replacement from each treatment group (thus creating multiple samples)
- Calculate difference in mean for each sample
- For percentile method: count number of replicates for which difference is above and below 0 (yielding a 1-tailed test of hypothesis of a cost difference)
- For parametric tests:
  - Because each bootstrap replicate represents a mean difference, when we summarize replicates, reported "standard deviation" is standard error
    - Difference in means / SE = z/t statistic
    - Difference in means + 1.96 SE = 95% CI
Nonparametric Bootstrap and Normality

- Nonparametric bootstrap does not depend on normality, so there is no violation of assumptions, but...
- If sample median has smaller relative bias than sample mean, may be better to use median whether sample mean is analyzed parametrically or nonparametrically

Example: Distribution of Costs, Chapter 5

<table>
<thead>
<tr>
<th></th>
<th>Group 0</th>
<th>Group 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arith Mean</td>
<td>3015</td>
<td>3040</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1582.802</td>
<td>1168.737</td>
</tr>
<tr>
<td>Quantiles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>899</td>
<td>1426</td>
</tr>
<tr>
<td>25%</td>
<td>1819</td>
<td>2226</td>
</tr>
<tr>
<td>50%</td>
<td>2825.5</td>
<td>2900.5</td>
</tr>
<tr>
<td>75%</td>
<td>3752</td>
<td>3604</td>
</tr>
<tr>
<td>95%</td>
<td>6103</td>
<td>5085</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.03501</td>
<td>1.525386</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>4.910192</td>
<td>9.234913</td>
</tr>
<tr>
<td>Geom Mean</td>
<td>2606.571</td>
<td>2836.971</td>
</tr>
<tr>
<td>Mean In</td>
<td>7.8634864</td>
<td>7.9501397</td>
</tr>
<tr>
<td>SD In</td>
<td>57602998</td>
<td>37871479</td>
</tr>
<tr>
<td>Obs</td>
<td>250</td>
<td>250</td>
</tr>
</tbody>
</table>

Example: P Values from 6 Univariate Tests of Difference in Cost

<table>
<thead>
<tr>
<th>SUMMARY TABLE</th>
<th>P-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIFFERENCE IN ARITHMETIC MEAN COST:</td>
<td>25.00</td>
<td>SE: 124.44</td>
</tr>
<tr>
<td>t-test, difference in means:</td>
<td>0.8409</td>
<td>-220 to 270</td>
</tr>
<tr>
<td>nonparametric BS, diff in means:</td>
<td>0.8600</td>
<td>-218 to 275</td>
</tr>
<tr>
<td>Wilcoxon rank-sum:</td>
<td>0.3722</td>
<td></td>
</tr>
<tr>
<td>Kolmogorov-Smirnov:</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>t-test, difference in logs:</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>transformation to normal:</td>
<td>Sqrt</td>
<td></td>
</tr>
<tr>
<td>t-test, transformed variable:</td>
<td>0.2907</td>
<td></td>
</tr>
<tr>
<td>test for heteroscedasticity:</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

Why Do Different Statistical Tests Lead To Different Inferences?

- Tests are evaluating differences in different statistics
  - T-test of untransformed costs: cannot infer that arithmetic means differ
  - Bootstrap: same (lack of) inference without normality assumption
  - Wilcoxon rank-sum test: same inference, but is accidental
    - Had medians differed, p-value would have been significant
  - T-test of log costs: can infer means of logs – and thus geometric means – differ
  - Kolmogorov-Smirnov test: can infer distributions differ (but not necessarily means or medians)

Summary, Univariate Analysis

- Want statistic that provides best estimate of population mean
  - Because mean * N is best estimate of what gainers gain and losers lose
- Best refers to a measure of error that incorporates both bias and variability
- In face of skewness:
  - Sample means less biased
  - Sample median often less variable
- Transformation/retransformation of limited value in presence of heteroscedasticity
Multivariable Analysis Of Economic Outcomes (I)

• Even if treatment is assigned in a randomized setting, use of multivariable analysis may have added benefits:
  – Improves power for tests of differences between groups (by explaining variation due to other causes)
  – Facilitates subgroup analyses for cost-effectiveness (e.g., more/less severe; different countries/centers)
  – Variations in economic conditions and practice pattern differences by provider, center, or country may have a large influence on costs and randomization may not account for all differences
  – Added advantage: Helps explain what is observed (e.g., coefficients for other variables should make sense economically)

Nonrandom Assignment (II)

• If treatment is not randomly assigned, multivariable analysis is necessary to adjust for observable imbalances between treatment groups, but it may NOT be sufficient

Common Multivariable Techniques Used for Analysis of Cost (I)

• Ordinary least squares regression predicting costs after randomization
• Ordinary least squares regression predicting log transformation of costs after randomization
• Generalized Linear Models
• On horizon:
  – Generalized Gamma regression (Manning et al., NBER technical working paper 293)
  – Extended estimating equations (Basu and Rathouz, Biostatistics 2005)
Current Standard: Generalized Linear Models (GLM)

- GLM models:
  - Don’t require normality or homoscedasticity.
  - Evaluate log of mean (or square root of mean or other transformation of mean), not mean of logs.
  - Thus, don’t raise problems related to retransformation from scale of estimation to raw scale.
- To build them, we must identify a "link function" and a "family" (based on data).

GLM Relaxes OLS Assumptions

- Ability to choose among different links relaxes assumption that $E(y|x) = \Sigma \beta_i X_i$ (OLS) or $E(\ln(y)|x) = \Sigma \beta_i X_i$ (Log OLS).
- Ability to choose among different families relaxes assumption of constant variance.
  - Gauss: constant variance.
  - Poisson: variance proportional to mean.
  - Gamma: variance proportional to square of mean.
  - Inverse gauss: variance proportional to cube of mean.

APPENDIX 1:
QALY Analysis
QALY Evaluation

• While substantial attention has been paid to models for evaluation of cost, substantially less has been paid to models for evaluation of QALYs
• QALY distribution shares certain complicating features with costs, but also has its own complicating features
  – Predictions should be confined to theoretical range of preference assessment instrument (e.g., –0.594 and 1.0 for EQ-5D)
  – Long, heavy LEFT tails
  – (Particularly for pre-scored instruments) Often multi-modal (see Figure on next slide)
  – (Commonly) Large fraction of 1s

Sample EQ-5D Scores

Multivariable Approaches

• There are beginnings of a literature on multivariable approaches
  – OLS (or GLM with identity link and gauss family) probably commonest
  – Alternatives
    • GLM with family (and link) diagnostics
    • GLM with a logit link and binomial 1 family or it's equivalent, beta regression (need specialized code for Stata, (Basu and Manca)
    • Adjusted limited dependent variable models (Alava et al.)
• While we demonstrate some of these methods, more work is required before we will be able to identify best practice
Implemented Models

• Start with GLM gauss/identity
  – Evaluate GLM diagnostics
  – If necessary, reestimate GLM with better fitting family
• Also assess GLM gamma/log
  – Evaluate GLM diagnostics
  – If necessary, reestimate GLM with better fitting family

Variance function:  \( V(u) = 1 \)
Link function: \[ g(u) = u \] \[ \text{[Gaussian]} \]
\[ \text{[Identity]} \]
Log likelihood = 85.080395  AIC = -32032.16  BIC = -3055.401

|          | Coef | Std Err | z    | P>|z|  | 95% CI       |
|----------|------|---------|------|------|----------------|
| 1.treat  | .0627749 | .0183515 | 3.42 | 0.001 | .0268067    .0987432 |
| dissev   | -.1512017 | .0831731 | -1.82 | 0.069 | -.314218    .0118147 |
| bcost    | -.0000359 | .0000121 | -2.96 | 0.003 | -.000060    -.0000122 |
| blqaly   | .207374 | .0633239 | 3.27 | 0.001 | .0832614    .3314867 |
| _cons    | .511092 | .0620345 | 8.24 | 0.000 | .3895067    .6326773 |

Common Starting Point:  GLM with Gauss/Identity

```stata
glm qaly i.treat dissev bcost blqaly , link(identity) family(gauss)
```

Variance function:  \( V(u) = 1 \)
Link function: \[ g(u) = u \] \[ \text{[Gaussian]} \]
\[ \text{[Identity]} \]
Log likelihood = 85.080395  AIC = -32032.16  BIC = -3055.401

|          | Coef | Std Err | z    | P>|z|  | 95% CI       |
|----------|------|---------|------|------|----------------|
| 1.treat  | .0627749 | .0183515 | 3.42 | 0.001 | .0268067    .0987432 |
| dissev   | -.1512017 | .0831731 | -1.82 | 0.069 | -.314218    .0118147 |
| bcost    | -.0000359 | .0000121 | -2.96 | 0.003 | -.000060    -.0000122 |
| blqaly   | .207374 | .0633239 | 3.27 | 0.001 | .0832614    .3314867 |
| _cons    | .511092 | .0620345 | 8.24 | 0.000 | .3895067    .6326773 |

GLM DIAGNOSTICS, Identity/Gauss

Fitted Model:  Link = Identity;  Family = Gaussian
Results, Modified Park Test (for Family)
Coefficient:  -.929485
Family, Chi2, and p-value in descending order of likelihood
Family   Chi2   P-value
Gaussian NLLS:  4.2582  0.0391
Poisson:  18.3496  0.0000
Gamma:  42.2987  0.0000
Inverse Gaussian or Wald  76.1054  0.0000
Results of tests of GLM Identity link
Pearson Correlation Test:  1
Pregibon Link Test:  .6741
Modified Hosmer and Lemeshow:  .8335
Troubling Findings

- Coefficient on modified Park test is negative (we don’t have any families that are negative) and p-value for named families are all significantly rejected.
- When confronted with coefficient < -0.5, consider subtracting all observations from maximum theoretically possible observation (e.g., 1.0 for most, if not all, instruments).

```stata
gen nqaly=1-qaly
sum qaly nqaly
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>qaly</td>
<td>500</td>
<td>0.594</td>
<td>0.212</td>
<td>0.057</td>
<td>0.969</td>
</tr>
<tr>
<td>nqaly</td>
<td>500</td>
<td>0.406</td>
<td>0.212</td>
<td>0.032</td>
<td>0.943</td>
</tr>
</tbody>
</table>

Variance function:  V(u) = 1
Link function:      g(u) = u  \[Gaussian\]
\[Identity\]
Log likelihood = 85.080395  AIC = -0.3203216  BIC = -3055.401

| nqaly  | Coef    | Std. Err. | Z      | P>|z| | 95% CI          |
|--------|---------|-----------|--------|------|----------------|
| 1.treat| 0.06275  | 0.01835   | -3.42  | 0.001| -.0987432, -.526807 |
| dissev | .15120   | .08317    | 1.82   | 0.069| -.0118147, .314218 |
| blcost | .0000359 | .0000121  | 2.96   | 0.003| -.0000122, .000060 |
| blqaly | -.20737  | .06332    | -3.27  | 0.001| -.3314867, -.0832614|
| _cons  | .48890   | .06203    | 7.88   | 0.000| .3673227, .6104933 |

Estimate NQALY, GLM with Gauss/Identity

```stata
glm nqaly i.treat dissev blcost blqaly, link(identity) family(gauss)
```

Identity/Gauss Recycled Predictions

```stata
glm nqaly i.treat dissev blcost blqaly,link(identity) family(gauss) margins treat
```

| Margin | Std. Err. | z      | P>|z| | [95% Conf. Interval] |
|--------|-----------|--------|------|---------------------|
| treat  |           |        |      |                     |
| 0      | 0.0130    | 33.70  | 0.000| 0.4118, 0.4627      |
| 1      | 0.0130    | 33.86  | 0.000| 0.3490, 0.3999      |

1-.4372 = .5628; 1-.3744 = .6256

DIFFERENCE: .0628
Fitted Model:  Link = Identity; Family = Gaussian

Results, Modified Park Test (for Family)

Coefficient: 0.686724

Family, Chi2, and p-value in descending order of likelihood

<table>
<thead>
<tr>
<th>Family</th>
<th>Chi2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>0.9443</td>
<td>0.3312</td>
</tr>
<tr>
<td>Gaussian</td>
<td>4.5374</td>
<td>0.0332</td>
</tr>
<tr>
<td>Gamma</td>
<td>16.5942</td>
<td>0.0000</td>
</tr>
<tr>
<td>Inverse Gaussian or Wald</td>
<td>51.4871</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Results of tests of GLM Identity link

Pearson Correlation Test: 1

Pregibon Link Test: 0.6741

Modified Hosmer and Lemeshow: 0.8335

Variance function: \( V(u) = u \)

Link function: \( g(u) = u \)

Log likelihood = -335.2046527  AIC = 1.360819

BIC = -3023.244

| nqaly | Coef   | Std Err | z     | P>|t| | 95% CI          |
|-------|--------|---------|-------|-----|----------------|
| 1.treat | -.06313 | .0566142 | -1.12 | .265 | -.1740918, .0478318 |
| dissev | .16252  | .269842  | 0.62  | .533 | -.3489997, .6740397 |
| blcost | .0000373| .0000387 | 0.96  | .335 | -.0000385, .0001132 |
| blqaly | -.199954| .1926091 | -1.04 | .299 | -.5774608, .1775532 |
| _cons | .477935 | .190924  | 2.50  | .012 | .1028309, .8512394 |

GLM DIAGNOSTICS, Identity/Gauss

Change Family to Poisson and Rerun Model

```
glm nqaly i.treat dissev blcost blqaly, link(identity)
family(poisson)
```

Variance function: \( V(u) = u \)

Link function: \( g(u) = u \)

Log likelihood = -335.2046527  AIC = 1.360819

BIC = -3023.244

| nqaly | Coef   | Std Err | z     | P>|t| | 95% CI          |
|-------|--------|---------|-------|-----|----------------|
| 1.treat | -.06313 | .0566142 | -1.12 | .265 | -.1740918, .0478318 |
| dissev | .16252  | .269842  | 0.62  | .533 | -.3489997, .6740397 |
| blcost | .0000373| .0000387 | 0.96  | .335 | -.0000385, .0001132 |
| blqaly | -.199954| .1926091 | -1.04 | .299 | -.5774608, .1775532 |
| _cons | .477935 | .190924  | 2.50  | .012 | .1028309, .8512394 |

Identity/Poisson Recycled Predictions

```
glm nqaly i.treat dissev blcost blqaly, link(identity)
family(poisson)
margins treat
```

```
Marginal Std. Err.  z  P>|z|  [95% Conf. Interval]
------------------- -----------------------------------------
 treat |        |
 |     0 | .4374  | 0.0417 | 10.49 | .000 | .3557, .5191 |
 |     1 | .3743  | 0.0386 | 9.71  | .000 | .2987, .4498 |

1-.4374 = .5626; 1-.3743 = .6257

DIFFERENCE: .0631
GLM DIAGNOSTICS, Identity/Poisson

Fitted Model: Link = Identity; Family = Poisson
Results, Modified Park Test (for Family)
Coefficient: 0.703074
Family, Chi2, and p-value in descending order of likelihood
Family   Chi2   P-value
Poisson  0.8796  0.3483
Gaussian NLLS: 4.9314  0.0264
Gamma: 16.7804  0.0000
Inverse Gaussian or Wald 52.6339  0.0000
Results of tests of GLM Identity link
Pearson Correlation Test: .9396
Pregibon Link Test: .6961
Modified Hosmer and Lemeshow: .8949

Can We Improve Link?
• Iteratively evaluate power links (in 0.1 intervals) between 1 and 2
  – Use modified Park test to select a family
  – Rerun GLM with power and preferred link
  – Evaluate fit statistics

Power 1.5 Link / Poisson Family

Variance function: V(u) = u
Link function: g(u) = u^(1.5)
Log likelihood = -335.199289 AIC 1.360797 BIC -3023.265

| nqaly   | Coef   | Std Err | z    | P>|z|  | 95% CI     |
|---------|--------|---------|------|------|-----------|
| 1.treat | -.059525 | .053554 | -1.11 | .266 | -.164488 .045439 |
| dissev  | .156198 | .244879 | .64  | .524 | -.323756 .636152 |
| blcost  | .000036 | .00037 | .97  | .331 | -.000037 .000109 |
| blqaly  | -.185844 | .180880 | -1.03 | .304 | -.540361 .168674 |
| cons    | .322960 | .180606 | 1.78 | .074 | .031021 .676941 |

Power 1.5 Link / Poisson Family

```
glm nqaly i.treat dissev blcost blqaly, link(power 1.5) family(poisson)
```
Power 1.5/Poisson Recycled Predictions

```
. glm nqaly i.treat dissev blcost blqaly,link(power 1.5)
   family(poisson)
```

```
. margins treat
```

```
|   | Margin | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---|--------|-----------|------|------|-----------------------|
|   | 0      | .4371     | 0.0415 | 10.53 | 0.000 | .3557 - .5186 |
|   | 1      | .3745     | 0.0384 | 9.75  | 0.000 | .2992 - .4498 |

1-.4371 = .5629; 1-.3745 = .6255

DIFFERENCE: .0626

GLM DIAGNOSTICS, Power 1.5/Poisson

```
. local max=1
. local min=0 (for EQ-5D, local min=-0.594)
. local a=-`min'/(`max'-`min')
. local b=1/(`max'-`min')
. gen bqaly=`a'+(`b'*qaly)

. sum qaly bqaly
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>qaly</td>
<td>500</td>
<td>.5941653</td>
<td>.2121148</td>
<td>.05679</td>
<td>.96822</td>
</tr>
<tr>
<td>bqaly</td>
<td>500</td>
<td>.5941653</td>
<td>.2121148</td>
<td>.05679</td>
<td>.96822</td>
</tr>
</tbody>
</table>

Logit Link, Binomial 1 Family

- Alternatively, we can transform QALY distribution so that it ranges between 0 and 1 and use a logit link and binomial 1 family (equivalent to beta regression)

```
local max=1
local min=0 (for EQ-5D, local min=-0.594)
local a=-`min'/(`max'-`min')
local b=1/(`max'-`min')
gen bqaly=`a'+(`b'*qaly)
```

```
. sum qaly bqaly
```
GLM with Binomial 1/Logit

\[ \text{glm bqaly i.treat dissev blcost blqaly, link(logit) family(binomial 1)} \]

Variance function: \( V(u) = u(1-u) \)  
Link function: \( g(u) = \ln(u/(1-u)) \)  
Log likelihood = -238.9699913  
AIC = 297.9888  
BIC = -2050.859

| Coef  | Std. Err. | z    | P>|z|  | 95% Conf. Interval |
|-------|-----------|------|------|-------------------|
| bqaly | .2626131  | .1834617 | 1.43 | 0.152 | -.0969653   .6221914 |
| dissev| -.6328458 | .832264  | -0.76 | 0.447 | -2.264053   .9983617 |
| blcost| -.0001494 | .0001208 | -1.24 | 0.216 | -.0003862    .0000975 |
| blqaly| .8675488  | .6338201 | 1.37  | 0.171 | -.3747157   2.109813 |
| _cons | .0373004  | .6190775 | 0.06  | 0.952 | -1.176069    1.250067 |

Logit/Binomial 1 Recycled Predictions

\[ \text{glm bqaly i.treat dissev blcost blqaly, link(logit) family(binomial 1)} \]
\[ \text{margins treat} \]

| Margin | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|--------|-----------|------|------|-------------------|
| treat  |           |      |      |                   |
| 0      | .5628     | .0312 | 18.02| 0.000 | .5016   .6441 |
| 1      | .6254     | .0305 | 20.53| 0.000 | .5657   .6852 |

DIFFERENCE: .0626

Run Link DIAGNOSTICS, Logit/Binomial 1

FITTED MODEL:  Link = Logit ; Family = Binomial  
Results of tests of GLM Identity link  
Pearson Correlation Test: .9914  
Pregibon Link Test: .5605  
Modified Hosmer and Lemeshow: .9242