Revisiting the Two-Sample Problem with Measurement Errors

Samiran Sinha
Texas A&M University

Joint Statistical Meetings 2024, Portland, OR

Collaborators: DongHyuk Lee and Soumendra Lahiri

Two-sample problem

- $X_1, \ldots, X_{n_x}$ are iid observations from a latent distribution F_x
- $Y_1, \ldots, Y_{n_y}$ are iid observations from a latent distribution F_y
- Seek to infer $Ho : F_x = F_y$ versus $Ha : F_x \neq F_y$
- Existing approaches are Kolmogorov-Smirnov (KS) test and Anderson-Darling (AD) test

What if the observed data are convoluted

- $(W_{1,1} \ldots, W_{1,m_x}), \ldots, (W_{n_x,1} \ldots, W_{n_x,m_x})$ are iid sets from a latent distribution F_{x+u_x}
- Assume observed, $W_{i,k} = X_i + U_{x,k}$
- $(V_{1,1} \ldots, V_{1,m_y}), \ldots, (V_{n_y,1} \ldots, V_{n_y,m_y})$ are iid sets from a latent distribution F_{y+u_y}
- Assume observed $V_{j,l} = Y_j + U_{y,l}$
- Seek to infer $Ho : F_x = F_y$ versus $Ha : F_x \neq F_y$
- $F_{x+u_x} = F_x \circ F_{u_x}$ is a convoluted distribution

$$F_{x+u_x}(r) = \int dF_x(r-s)dF_{u_x}(s).$$

- $F_{y+u_y} = F_y \circ F_{u_y}$ is also a convoluted distribution.

Motivating example: 2009-2010 NHANES study

- We seek to test if the distribution of systolic blood pressure is the same between no-alcohol and the alcohol groups.
- However, blood pressure inherently varies in a body; it changes over the 24-hour period. It is impossible to measure a person's actual blood pressure.
- NHANES collected at least three measurements of the surveyed individuals.
- F_x : Dist. of SBP among non-alcohol group
- F_y : Dist. of SBP among alcohol group
- $m_x = m_y = 3$ for the majority of the subjects, a glimpse of the data
- Seek to infer $Ho : F_x = F_y$ versus $Ha : F_x \neq F_y$
- However, observed data are not directly from F_x and F_y.

A glimpse of the data

Code

```
# Alcohol group          # Non-alcohol group
  129 134 126              132 138 135
  128 128 130              131 137 141
  131 131 132              122 125 139
  131 128 132              131 148 144
  137 129 132              137 134 143
  129 127 123              130 130 128
```

Taking the average of three measurements, and then applying the KS or AD test does nor preserve the level.

Our proposal (assumptions)

- $W_{i,k} = X_i + U_{x,k}$ and $V_{j,l} = Y_j + U_{y,l}$
- X_i's are iid F_x and Y_j's are iid F_y
- X and U_x are independent
- Y and U_y are independent
- Assume that F_{u_x} and F_{u_y} are symmetric about zero.

Our proposal

- Estimate the characteristic function (CF) of U_x
- Estimate the CF of X
- Estimate the characteristic function (CF) of U_y
- Estimate the CF of Y
- Compare the estimated characteristic functions of X and Y through an integrated, weighted, squared difference

Test statistic

- Estimated CF of X: $\widehat{a}_x(t) + i\widehat{b}_x(t)$, $i = \sqrt{-1}$
- Estimated CF of Y: $\widehat{a}_y(t) + i\widehat{b}_y(t)$
- Test statistic

$$T = \int_{-\infty}^{\infty} n_x \left[\{\widehat{a}_x(t) - \widehat{a}_y(t)\}^2 + \{\widehat{b}_x(t) - \widehat{b}_y(t)\}^2 \right] \omega(t) dt$$

- We took $\omega(t) > 0$ over $[t_1, t_2]$; interval $[t_1, t_2]$ includes zero, and $\omega(t) = 0$ for $t \in \overline{[t_1, t_2]}$
- **Theorem 1.** Under all stated assumptions, H_0, and when $n_x/n_y \to \rho \in (0, \infty)$, the test statistic T converges to $\int \{\zeta_1^2(t) + \zeta_2^2(t)\}\omega(t)dt$, where $\zeta_1(t)$ and $\zeta_2(t)$ are two mean-zero Gaussian processes with a complex covariance kernel.

notation continues

$\widehat{a}_x(t) = \sum_{j=1}^{n_x} \cos(t\overline{W}_j)/a_{2x}(t)$, $\widehat{b}_x(t) = \sum_{j=1}^{n_x} \sin(t\overline{W}_j)/a_{2x}(t)$,

$$a_{2x}(t) = \left| n_x^{-1} \sum_{j=1}^{n_x} M_x^{-1} \sum_{(l_1,l_2)\in\mathcal{S}_x} \cos\{(t/m_x)(W_{j,l_1} - W_{j,l_2})\} \right|^{m_x/2}$$

$\widehat{a}_y(t) = \sum_{j=1}^{n_y} \cos(t\overline{V}_j)/a_{2y}(t)$, $\widehat{b}_y(t) = \sum_{j=1}^{n_y} \sin(t\overline{V}_j)/a_{2y}(t)$,

$$a_{2y}(t) = \left| n_y^{-1} \sum_{j=1}^{n_y} M_y^{-1} \sum_{(l_1,l_2)\in\mathcal{S}_y} \cos\{(t/m_y)(V_{j,l_1} - V_{j,l_2})\} \right|^{m_y/2}$$

Novelties

- No one seems to have considered this problem earlier.
  - There are many works on density deconvolution (Delaigle et al., 2008, 2015; Stefanski and Carroll, 1988).
  - There are many work on covariate measurement errors or misclassification (Carroll et al., 2006; Gustafson, 2003)
- No assumption is made about F_x and F_y
- Other than symmetric about zero, no other assumption is made about the measurement errors' distribution.
- The method would work as long as $m_x \geq 2$ and $m_y \geq 2$ (at least two SBP measurements for every subject).

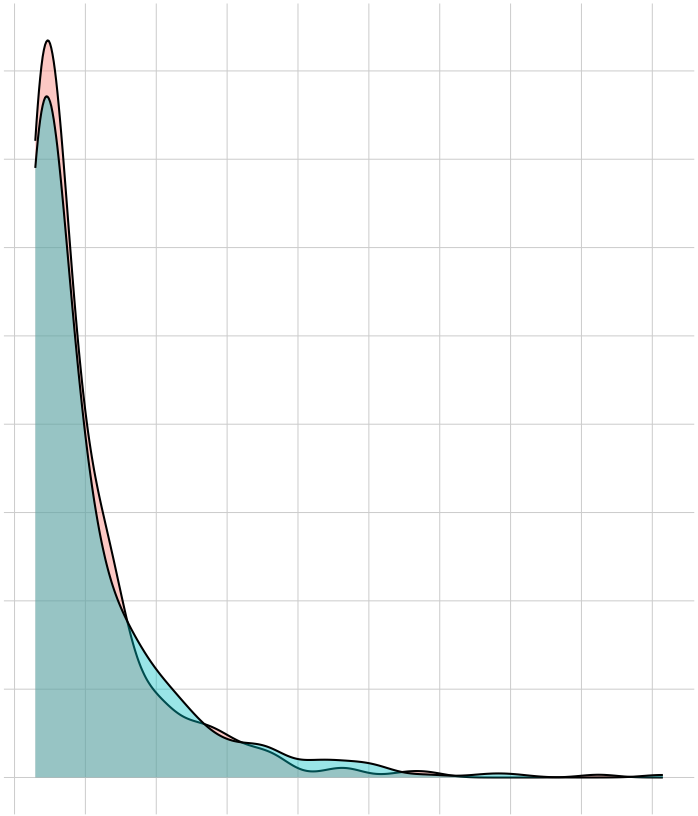
The test- continues

- It has been shown that as the sample size increases while ratio of two sample sizes converges to a positive constant, the power of the test conveges to one.
- Enumeration of the null distribution of the test statistic was next to impossible. So, we used a bootstrap method to assess the null distribution, or to calculate the p-value of the test.

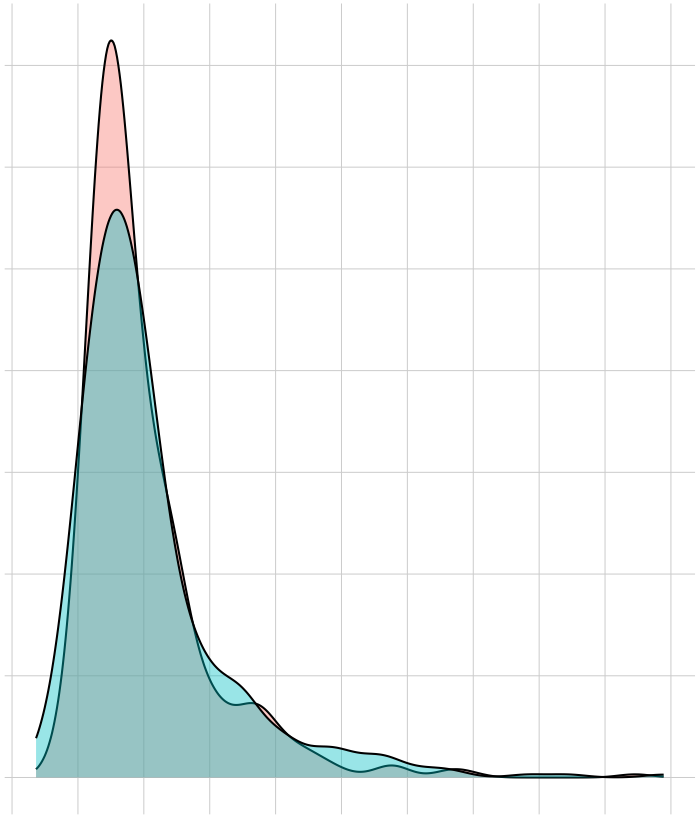
Simulation study

- Sample size for both groups $n = n_x = n_y$
- Repeated measurements for every subject, $m_x = m_y = 2$
- $X, Y \sim (\chi^2 - 1)/\sqrt{2}$ (i.e., H_0 holds)
- $U_x \sim DE(0, 0.35)$, $U_y \sim DE(0, 0.2)$
- Test carried out at the nominal level of 5%
- KS: Kolmogorov-Smirnov test, AD: Anderson-Darling test
- # replications: 5000

			proposed			
n	KS	AD	$unif_{0.99}$	$unif_{0.8}$	$norm_{0.99}$	$norm_{0.8}$
50	0.041	0.056	0.010	0.038	0.035	0.037
200	0.052	0.083	0.033	0.036	0.037	0.036
500	0.12	0.198	0.040	0.036	0.039	0.035



Plot of X and Y



Plot of convoluted data, $\overline{W}$ and $\overline{V}$

Simulation study

The choice of the weight function

- $norm_{0.99}$:

$$\omega(t) = \begin{cases} \exp(-t^2/2) & \text{for } t \in [t_1, t_2] \\ 0 & \text{otherwise} \end{cases}$$

$t_1 = \min\{F_x^{-1}(0.005), F_y^{-1}(0.005)\}$ $t_2 = \max\{F_x^{-1}(0.995), F_y^{-1}(0.995)\}$

- $norm_{0.8}$: $\omega(t) = \exp(-t^2/2)$ for $t \in [t_1, t_2]$ and zero otherwise;
$t_1 = \min\{F_x^{-1}(0.1), F_y^{-1}(0.1)\}$ $t_2 = \max\{F_x^{-1}(0.9), F_y^{-1}(0.9)\}$
- $unif_{0.99}$: $\omega(t) = 1$ for $t \in [t_1, t_2]$ and zero otherwise;
$t_1 = \min\{F_x^{-1}(0.005), F_y^{-1}(0.005)\}$ $t_2 = \max\{F_x^{-1}(0.995), F_y^{-1}(0.995)\}$
- $unif_{0.8}$: $\omega(t) = 1$ for $t \in [t_1, t_2]$ and zero otherwise;
$t_1 = \min\{F_x^{-1}(0.1), F_y^{-1}(0.1)\}$ $t_2 = \max\{F_x^{-1}(0.9), F_y^{-1}(0.9)\}$

Simulation study

- $X \sim Normal(0,1)$, $Y \sim Normal(0.2,1)$, $U_x, U_y \sim DE(0,0.35)$ (i.e., H_a holds)
- Test carried out at the nominal level of 5%
- KS: Kolmogorov-Smirnov test, AD: Anderson-Darling test
- # replications: 5000

			proposed			
n	KS	AD	$unif_{0.99}$	$unif_{0.8}$	$norm_{0.99}$	$norm_{0.8}$
50	0.099	0.137	0.053	0.108	0.092	0.115
200	0.327	0.443	0.156	0.369	0.322	0.393
500	0.728	0.820	0.401	0.749	0.695	0.775

Simulation study

- $X \sim Normal(0, 1)$, $Y \sim DE(0, 0.7)$, and $U_x, U_y \sim DE(0, 0.35)$ (i.e., H_a holds)
- Test carried out the nominal level of 5%
- KS: Kolmogorov-Smirnov test, AD: Anderson-Darling test
- # replications: 5000

			proposed			
n	KS	AD	$unif_{0.99}$	$unif_{0.8}$	$norm_{0.99}$	$norm_{0.8}$
50	0.063	0.069	0.095	0.053	0.085	0.051
200	0.147	0.154	0.439	0.110	0.315	0.095
500	0.423	0.470	0.863	0.275	0.745	0.222

Analysis of 2009-2010 NHANES data

- We focus on non-Hispanic white males ages 35-55 (homogeneous demographic).
- Alcohol consumption data are collected through two 24-hour recall interviews.
- non-alcoholic: both measurements ≤ 14 grams, alcoholic: otherwise
- $n_x = 207$ (non-alcoholic) and $n_y = 126$ (alcoholic).
- For both groups, we regressed the average of the SBP on BMI and income (ordinal); then applied the proposed tests on the *residuals*.

Analysis continues

- $\omega(t)$ of the proposed test: In the absence of any specific knowledge about the CF of the underlying distributions, in our opinion, the $unif_{0.99}$ weight is preferable as it covers a wide range of t-values and gives equal importance to the difference between the two CFs at any t.
- KS p-value$=0.058$, AD p-value$=0.001$, proposed method p-value$=0.001$
- Based on the p-value of the proposed test, there is evidence that the two distributions are different. This difference can be attributed to the difference in alcohol consumption.

Discussion

- An R package METest is available on CRAN (`https://CRAN.R-project.org/package=METest`).
- The symmetry assumption on the measurement errors U_x and U_y most likely be relaxed if the number of replications is more than or equal to three.
- We have not looked at the role of the survey weights. One student is looking into this issue.

Some references

- Carroll et al. (2006). Measurement Error in Nonlinear Models: A Modern Perspective. (**Book**)
- Delaigle et al. (2015). Methodology for nonparametric deconvolution when the error distribution is unknown. Journal of the Royal Statistical Society, Series B.
- Delaigle et al.(2008). On deconvolution with repeated measurements. The Annals of Statistics.
- Gustafson, P. (2003). Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. (**Book**)
- Hall et al. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. The Annals of Statistics.
- Lee et al. (2020). A test of homogeneity of distributions when observations are subject to measurement errors. Biometrics.
- Stefanski et al. (1990) Deconvoluting kernel density estimators. Statistics.
- Pettitt. (1976). A two-sample Anderson-Darling rank statistic. Biometrika.