# Meet the Overlapping Coefficient:

## A Measure for Elevator Speeches

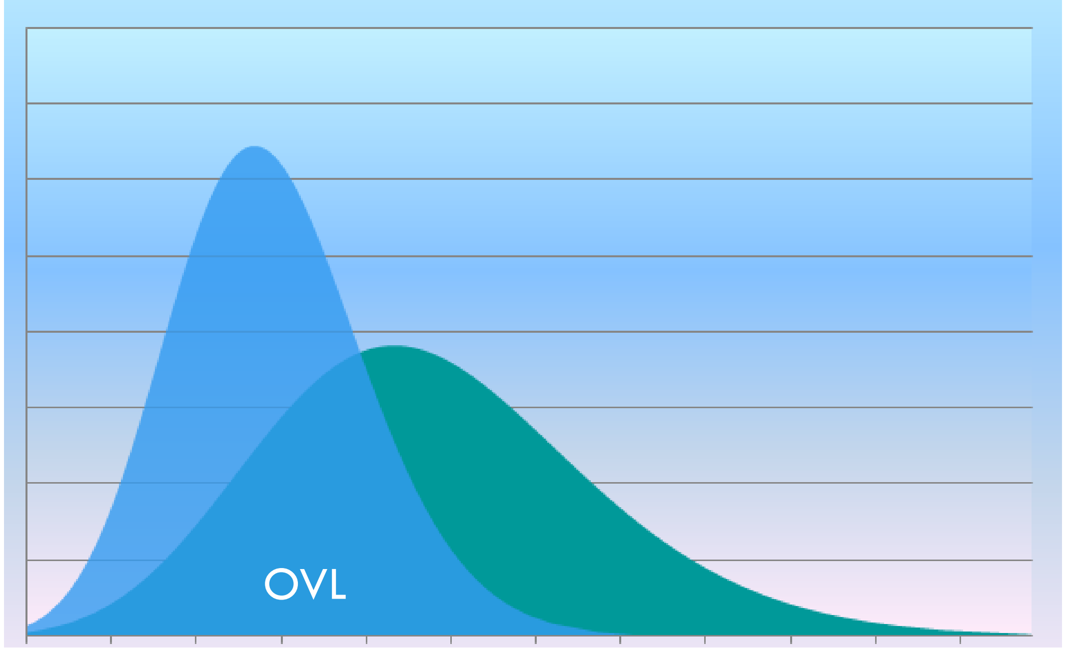
Brent Larson
larson@infinity.aero

# The Overlapping Coefficient

- What is it?
- Where did it come from?
- How might a cost analyst use it?
- How does one get the OVL?
- We want it now! I want it yesterday![1]

# What is this coefficient?

- The overlapping coefficient (OVL) refers to the area under two probability density functions simultaneously.[2]




For continuous distributions:

$$OVL = \int_{R_n} \min[\, f_1(\mathbf{x}), f_2(\mathbf{x})]d\mathbf{x}$$

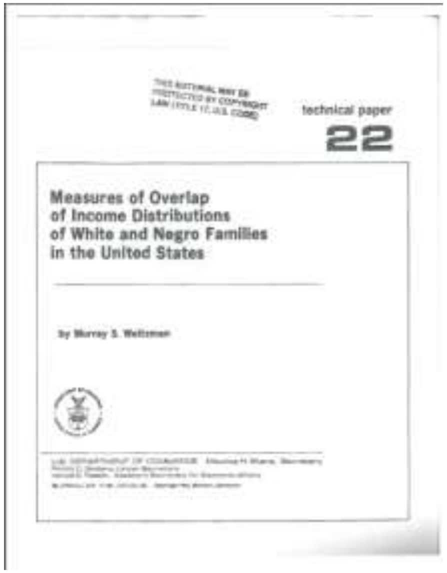
In discrete cases:

$$OVL = \sum_{\mathbf{x}} \min[\, f_1(\mathbf{x}), f_2(\mathbf{x})]$$

- The word "coefficient" means a measure of something
- Thus OVL is a measure of agreement or similarity[3]

# Where did the OVL come from?

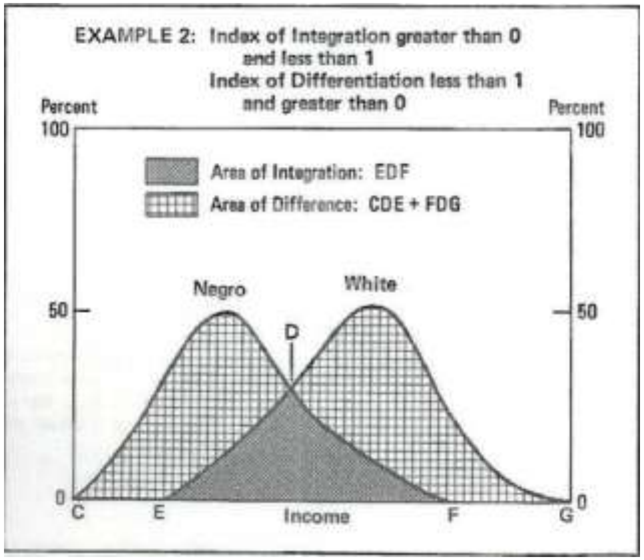
- In different form, OVL dates to the early days of Karl Pearson, ~ 1895
- Reportedly, explicit use begins in 1970[3] by economist Murray Weitzman to compare income distributions[4]



technical paper

22

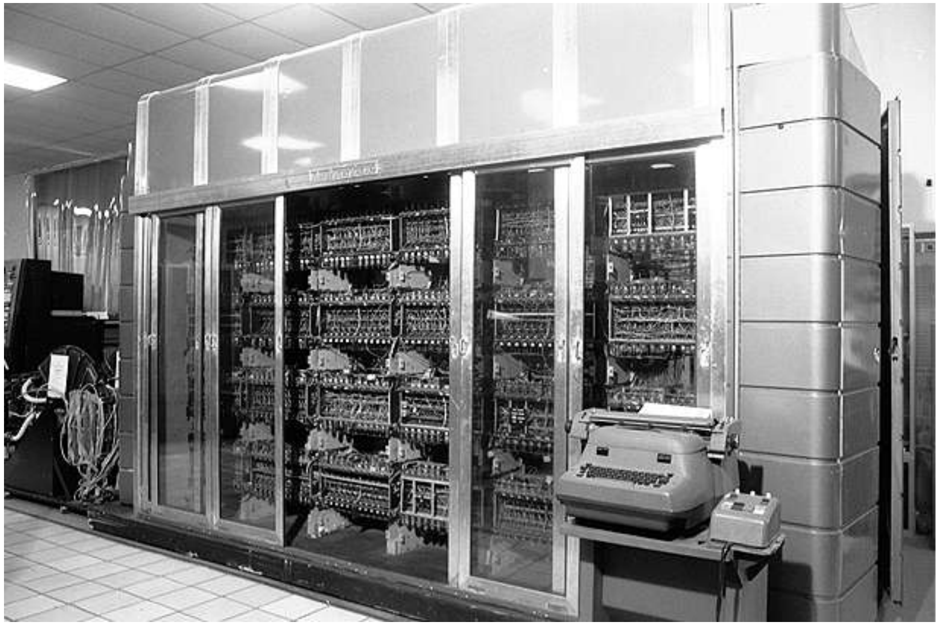
Measures of Overlap of Income Distributions of White and Negro Families in the United States




- Graphics from: Weitzman, M. S. (1970). *Measures of overlap of income distributions of white and Negro families in the United States*. Washington: U.S. Bureau of the Census; [for sale by the Supt. of Docs., U.S. Govt. Print. Off.

# Where did the OVL come from?

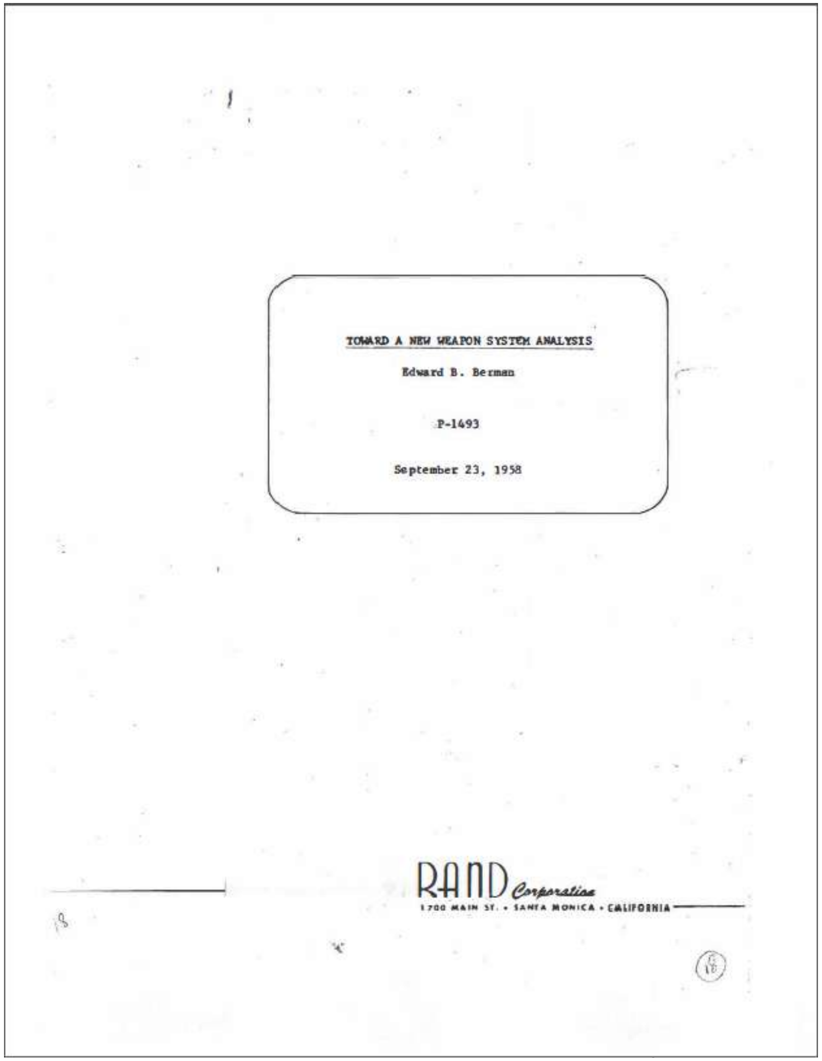
- Biostatisticians at UAB Huntsville develop & define OVL as currently used[3,6,7,8] ~ 1980's -1990's
- However. . . story is much richer – Guess who's involved?
- Here's a clue:
- Johnniac?



http://ed-thelen.org/comp-hist/Shustek/ShustekTour-02.html

# Where did the OVL come from?

- Yep. . . RAND Corporation!
- Modern, explicit use of the OVL in the continuous case may be found earlier – at the birthplace of Weapon Systems Analysis & Cost Analysis
- 1958 - Ed Berman, RAND consultant & Harvard trained economist uses overlapping distributions to compare weapon system alternatives[9]



TOWARD A NEW WEAPON SYSTEM ANALYSIS

Edward B. Berman

P-1493

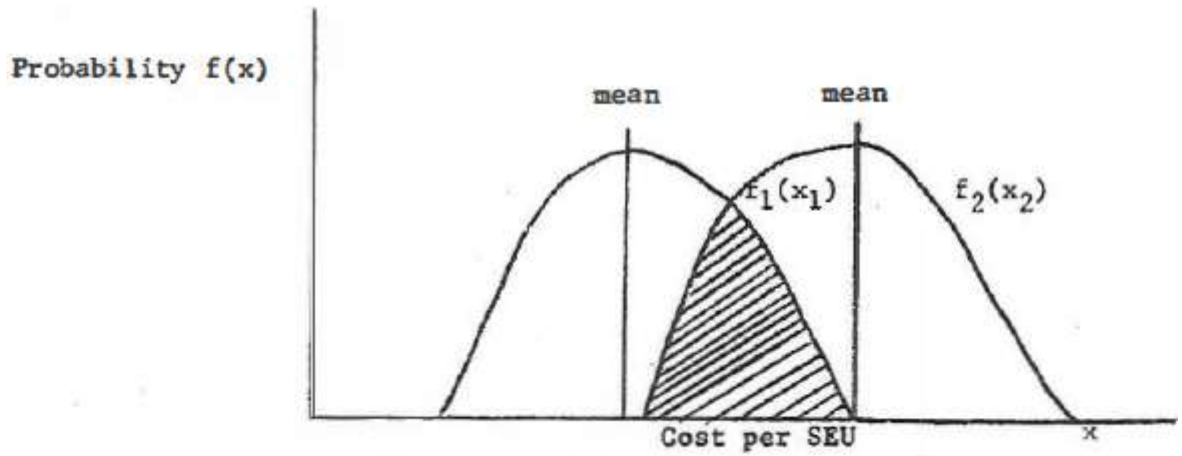
September 23, 1958

RAND Corporation

1700 MAIN ST. • SANTA MONICA • CALIFORNIA

# Where did the OVL come from?

- Here's the evidence. . .




Overlapping Cost-Probability Functions

- Here's Dr Berman's calculus. . .

$$V_2 = \int_{x_1=0}^{\infty} \int_{x_2=0}^{x_1} (x_1 - x_2)\, f_2\,(x_2)\, dx_2\, f_1(x_1)\, dx_1$$

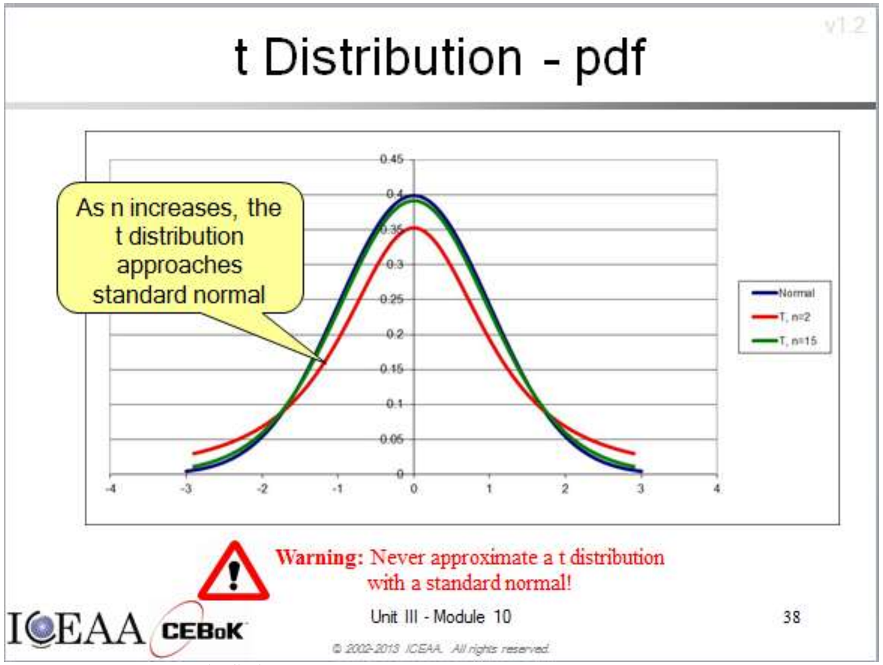
- Graphics from: Berman, E. B. (1958). Toward a new weapon system analysis. Santa Monica, Calif: Rand Corp.

# [OBTW. . . historical context] Where did the OVL come from?

- Berman's paper, written for David Novick[10] (the "father of cost analysis[11]"), is an earlier use of probability theory to model cost uncertainty than is commonly known
  - Berman modeled conceptually and at the total system cost level
- Appears to be lay groundwork for later developments in cost uncertainty analysis
  - Method of Moments – Steven Sobel, MITRE, 1965[12]
  - Monte Carlo simulation – Paul F. Dienemann, RAND 1966[13]
- Dr Paul Garvey credits Sobel for pioneering the method of moments technique to create a probability distribution of total system cost[14]
- Sobel worked for Berman at MITRE[15]
- . . . and Sobel cites Berman's work in his 1965 paper!

# How might a cost analyst use it?

- What's the OVL good for?
- Comparing theoretical weapon system models, etc.
- Also good for comparing probability models of different form - note these 3 overlapping distributions
- OVL ~ .86 for N(0,1), t(2)
- Models share 86% area
- Illustrates convergence of t to normal distribution

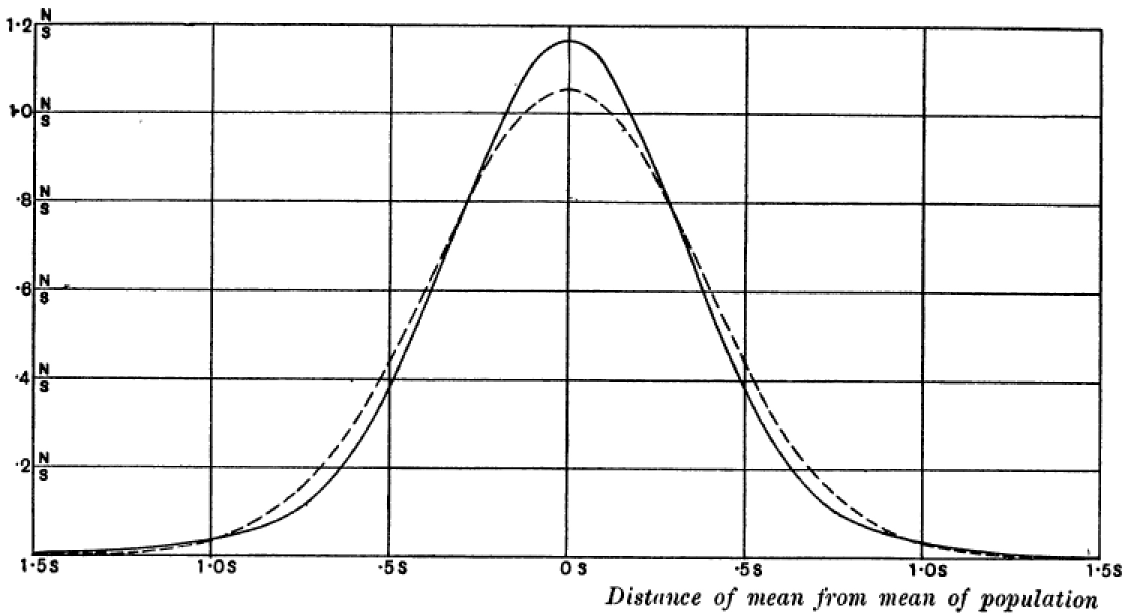


# More context. . .

- Look familiar?  - Ur case of previous graphic
- Would you believe that simulation was used?

DIAGRAM II. Solid curve $y = \frac{N}{S} \times \frac{8}{7} \cdot \frac{6}{5} \cdot \frac{4}{3} \cdot \frac{2}{\pi} \cos^{10}\theta$, $x/s' = \tan\theta$.

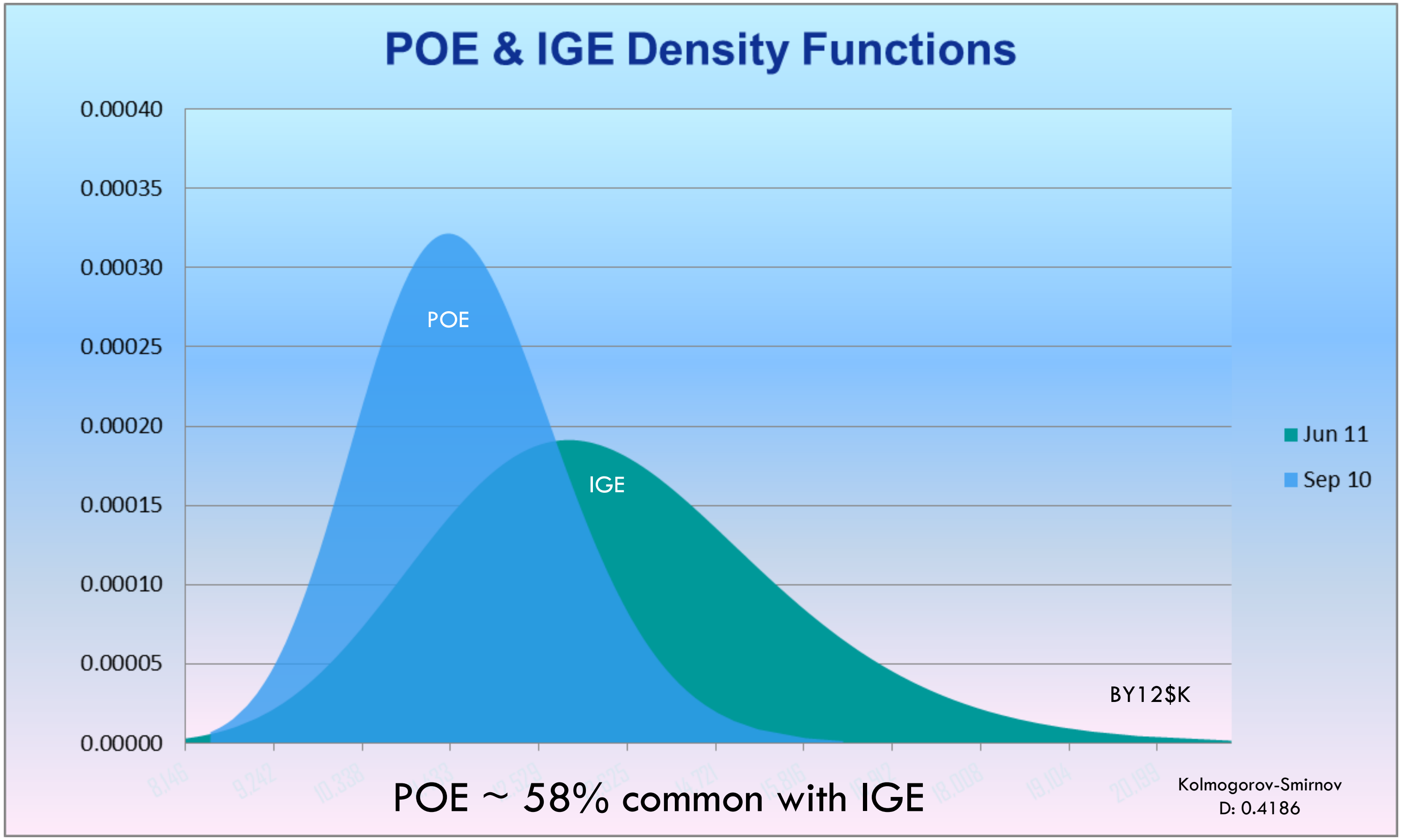
Broken line curve $y = \frac{\sqrt{7} \cdot N}{\sqrt{2\pi} \cdot s} e^{-\frac{7x^2}{2s^2}}$, the normal curve with the same S.D.




- Student (1908a). *The probable error of a mean*. Biometrika VI, 1-25.

# How might a cost analyst use it?

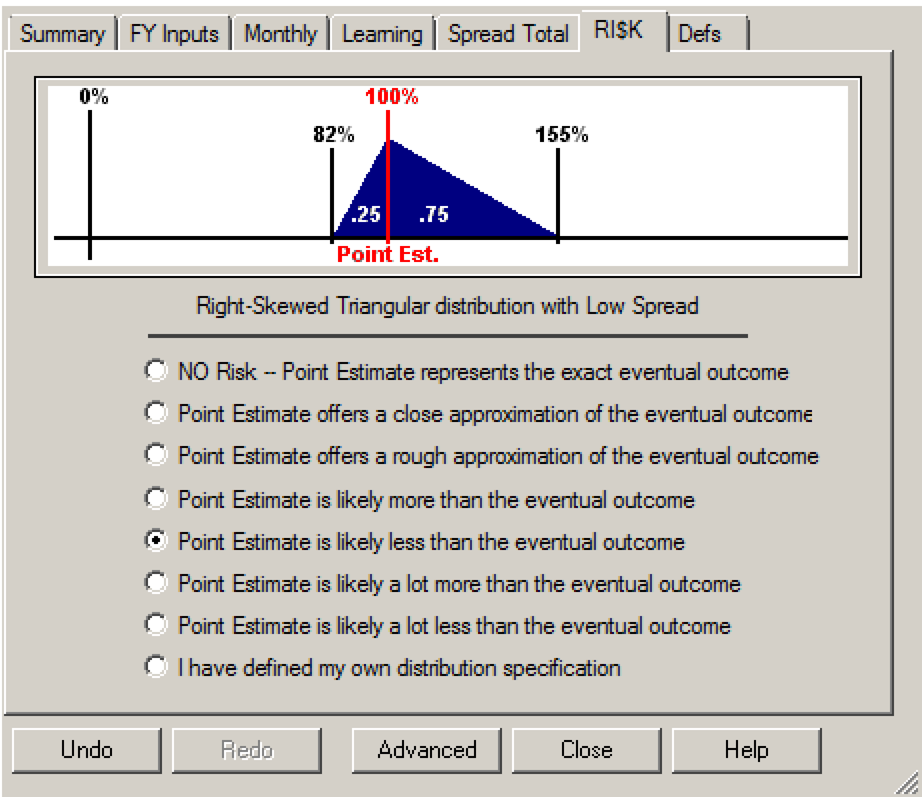
- Summarize change between estimates

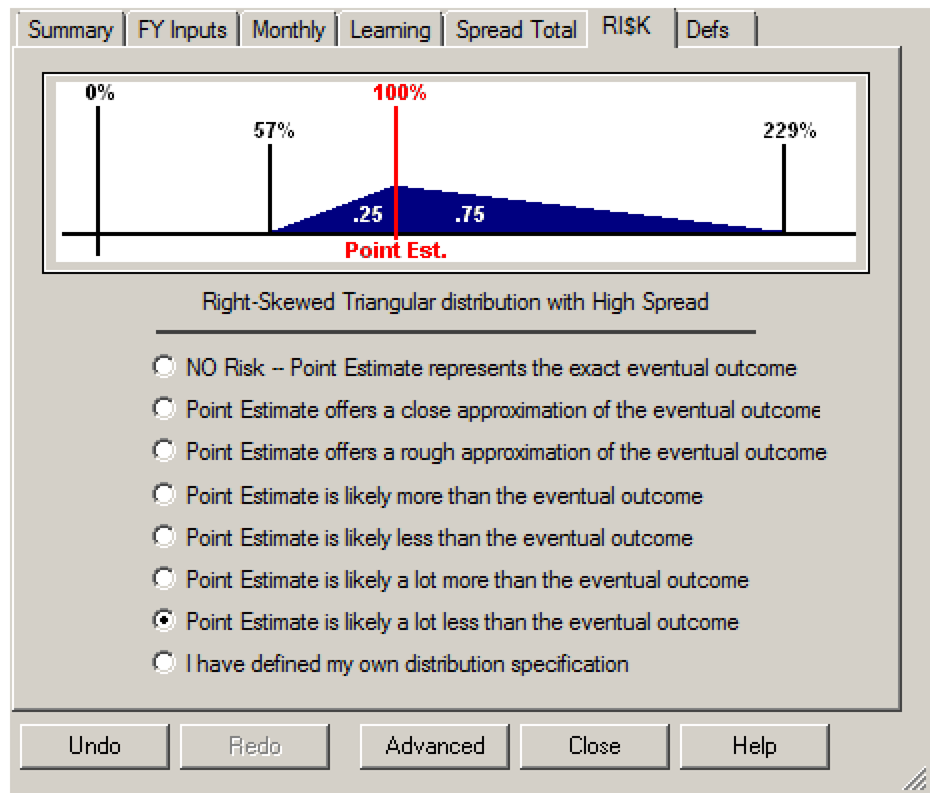


# How might a cost analyst use it?

- Find degree of similarity between input risk shapes

$$\widehat{OVL} \sim .60$$



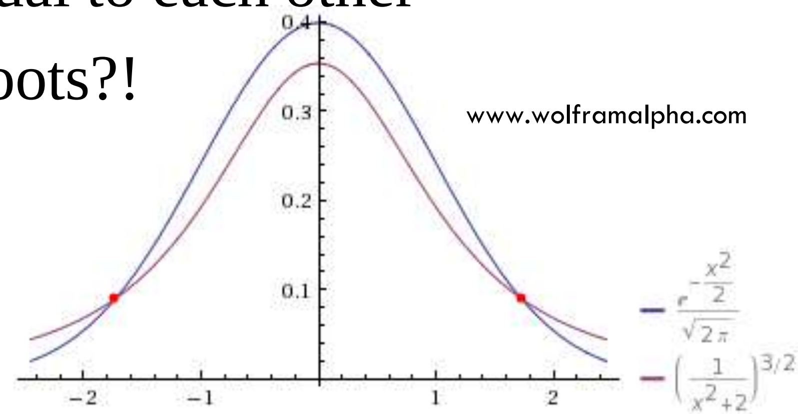



# How does one get the OVL?

- Compute area using intersecting points of overlapping distributions
  - Most distributions will intersect 0, 1 or 2 times
- Normal versus t example
  - Intersections may be determined analytically or numerically
- Risk shape example
  - Intersecting points found visually
- In the case of data without known distributional form
  - More work is required. . .

# How does one get the OVL?

- For parameterized models, e.g., for N(0,1), t(2)
- Step 1: WolframAlpha
  - Set equations for densities equal to each other
  - Click enter. . . and complex roots?!
- Step 2: Excel
  - Plug the real roots into NORM.S.DIST & T.DIST:




www.wolframalpha.com

```
=1-ABS(NORM.S.DIST(1.72511,TRUE)-T.DIST(1.72511,2,TRUE))-ABS(NORM.S.DIST(-1.72511,TRUE)-T.DIST(-1.72511,2,TRUE))
```

Solutions:

$x \approx -0.606179\, i$

$x \approx 0.606179\, i$

$x_1 \approx -1.72511$

$x_2 \approx 1.72511$

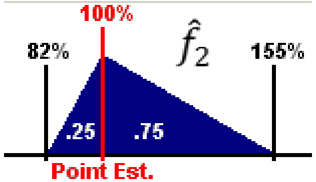
= 0.85786

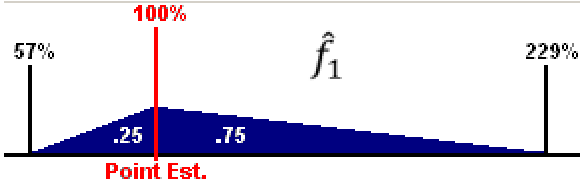
- Symbolically:

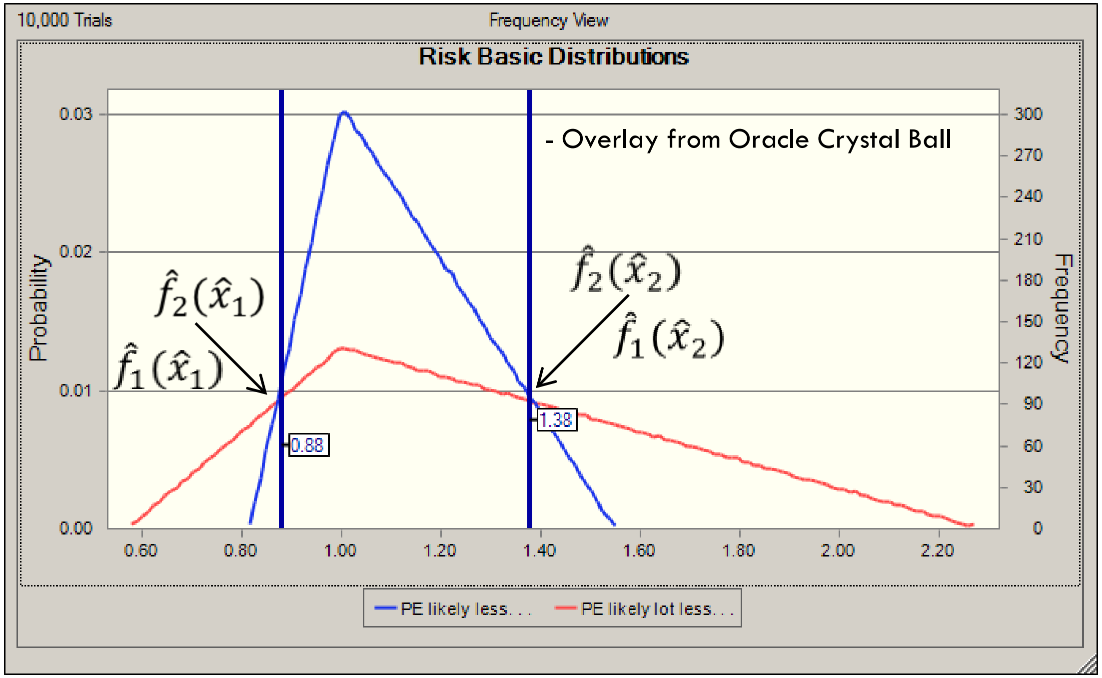
$$OVL = 1 - \left|\Phi(x_2) - F_2(x_2)\right| - \left|\Phi(x_1) - F_2(x_1)\right|$$

# How does one get the OVL?

- For risk shape example
- Step 1: Overlay chart
  - Eyeball roots
- Step 2: Excel
  - Calculate
  - No triangular distribution function in Excel
  - See backup



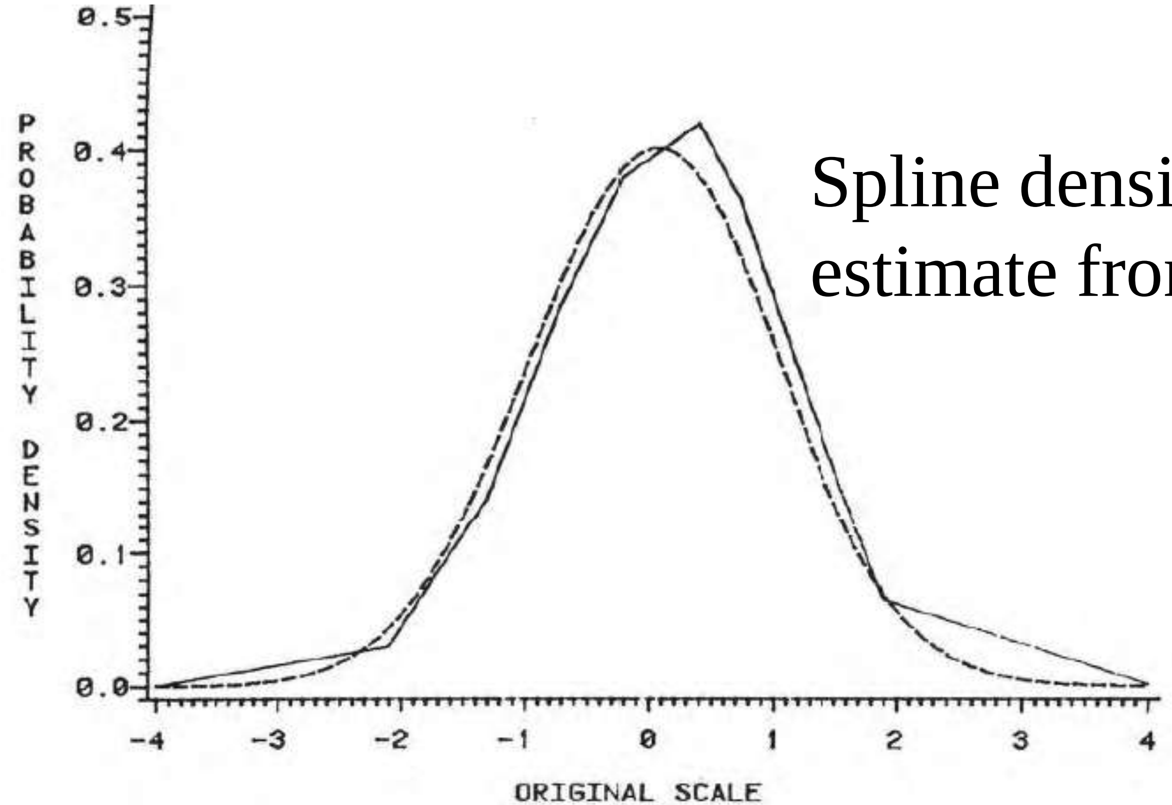

$$OVL = 1 - \left|\hat{F}_1(1.38) - \hat{F}_2(1.38)\right| - \left|\hat{F}_1(.88) - \hat{F}_2(.88)\right|$$

Triangular CDF math for Excel

= IF(cost<low,0,IF(cost<mode,(cost-low)^2/((high-low)*(mode-low)),IF(cost<=high,1-(high-cost)^2/((high-mode)*(high-low)),1)))

# How does one get the $\widehat{OVL}$?

- Got data? – Then historically with density estimation



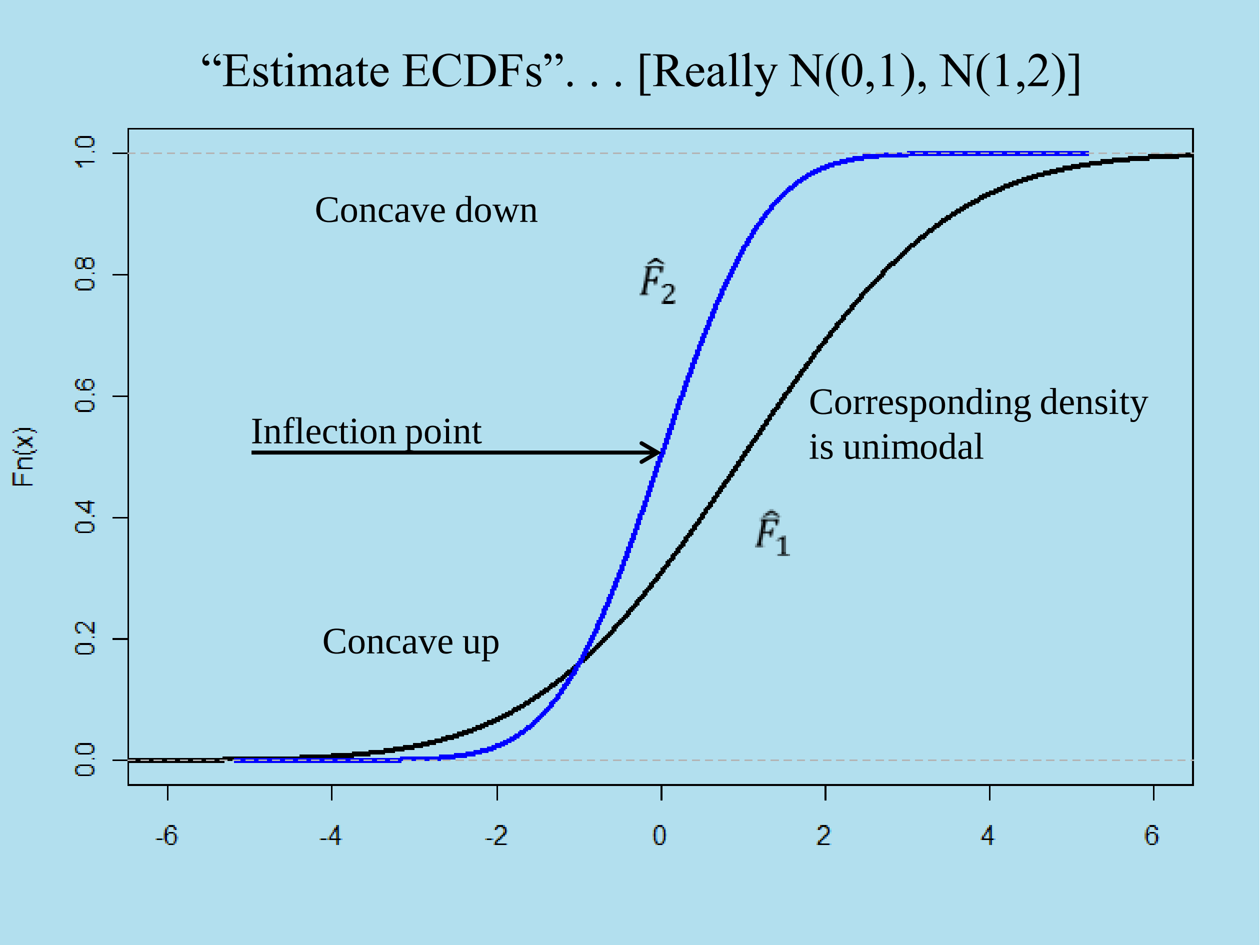

Spline density estimate from 1984

Figure 3.16 Spline density estimation: the spline-estimated density function obtained from a generated sample of 1000 standard-normal deviates. The estimated density is shown by the solid line, and the standard-normal density function is indicated by the broken line.

- Graphic From: Inman, H. F. (1984). *Behavior and properties of the overlapping coefficient as a measure of agreement between distributions*.

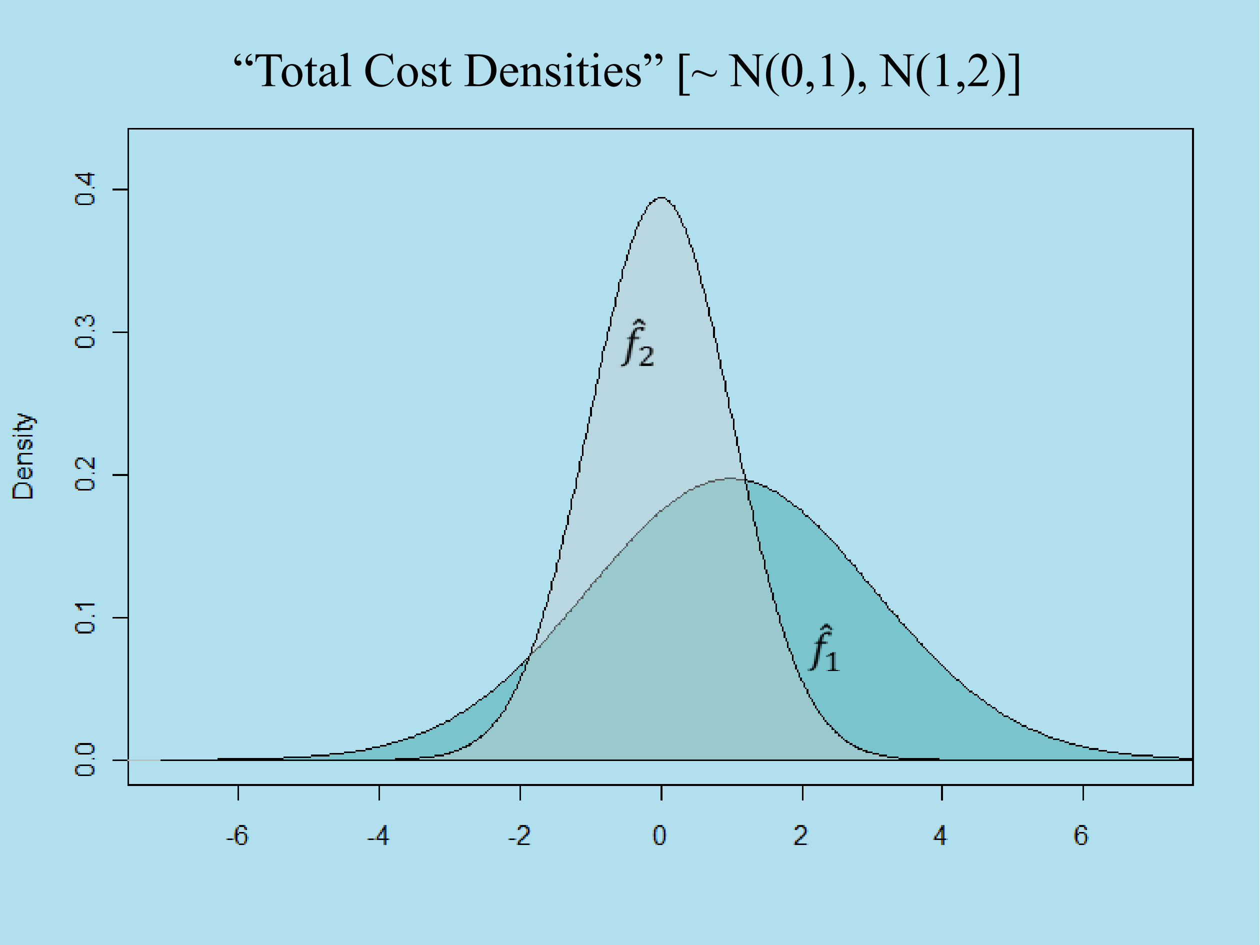
# How does one get the $\widehat{OVL}$?

- From S-Curves!  The story follows. .



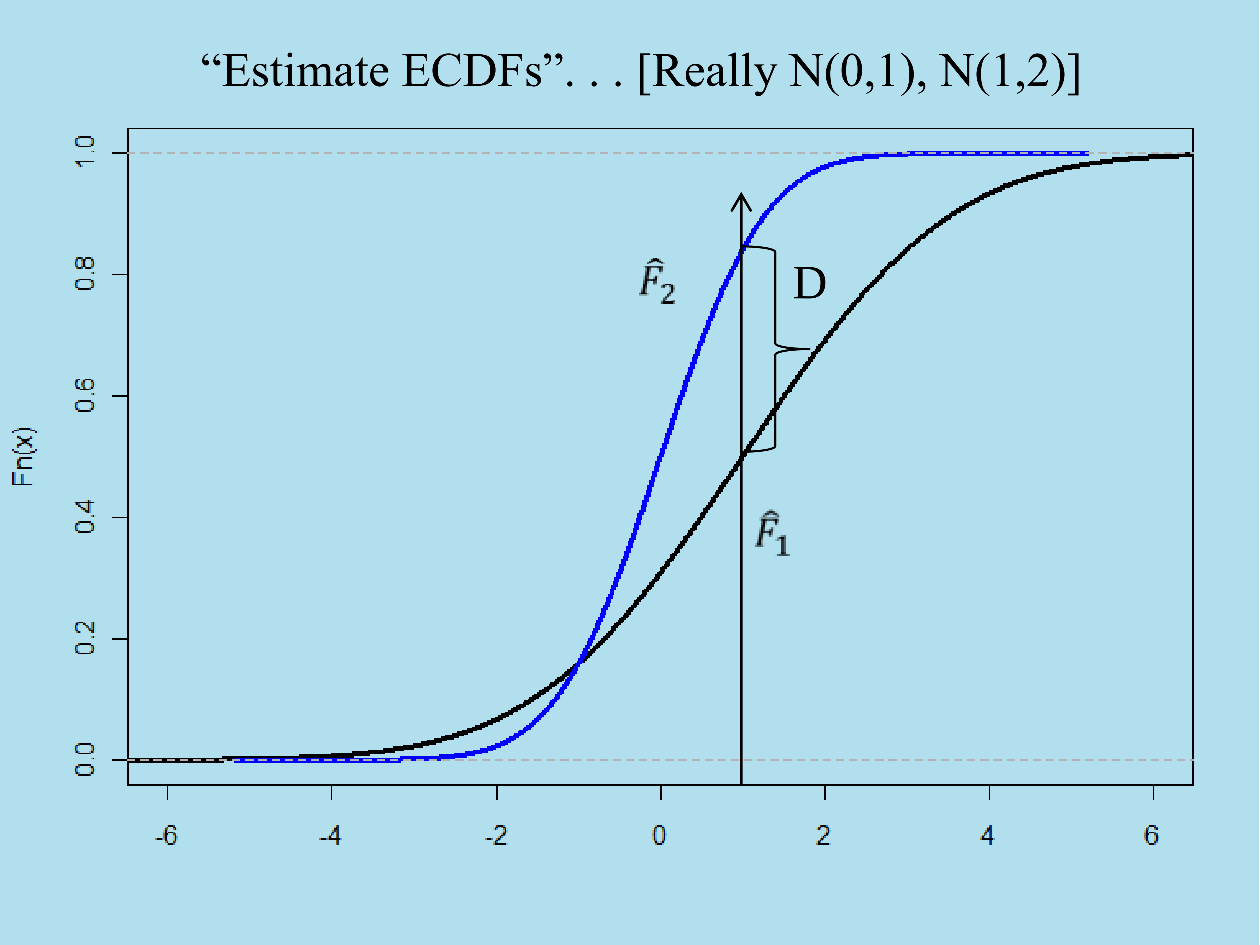
# How does one get the $\widehat{OVL}$?

- On flip side of the fundamental theorem of calculus. . .



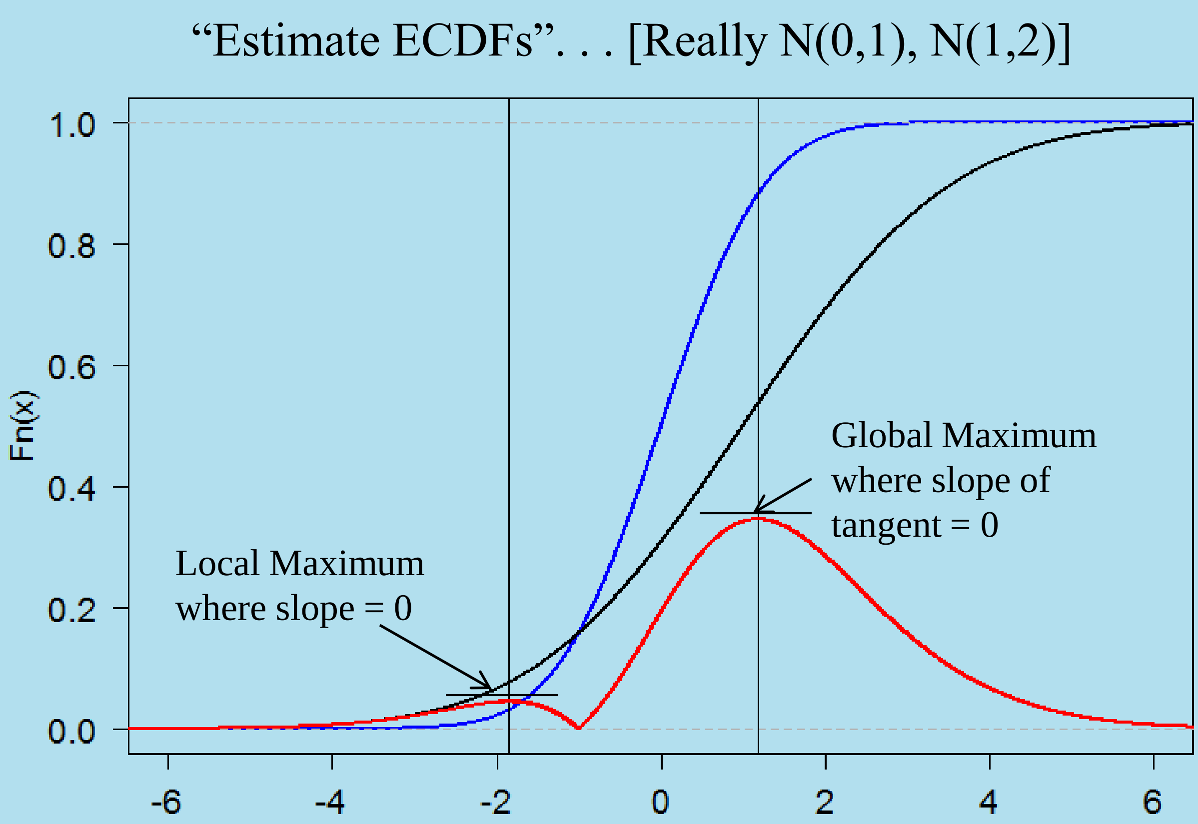
# How does one get the $\widehat{OVL}$?

- Curves share a distance between them



# How does one get the $\widehat{OVL}$?

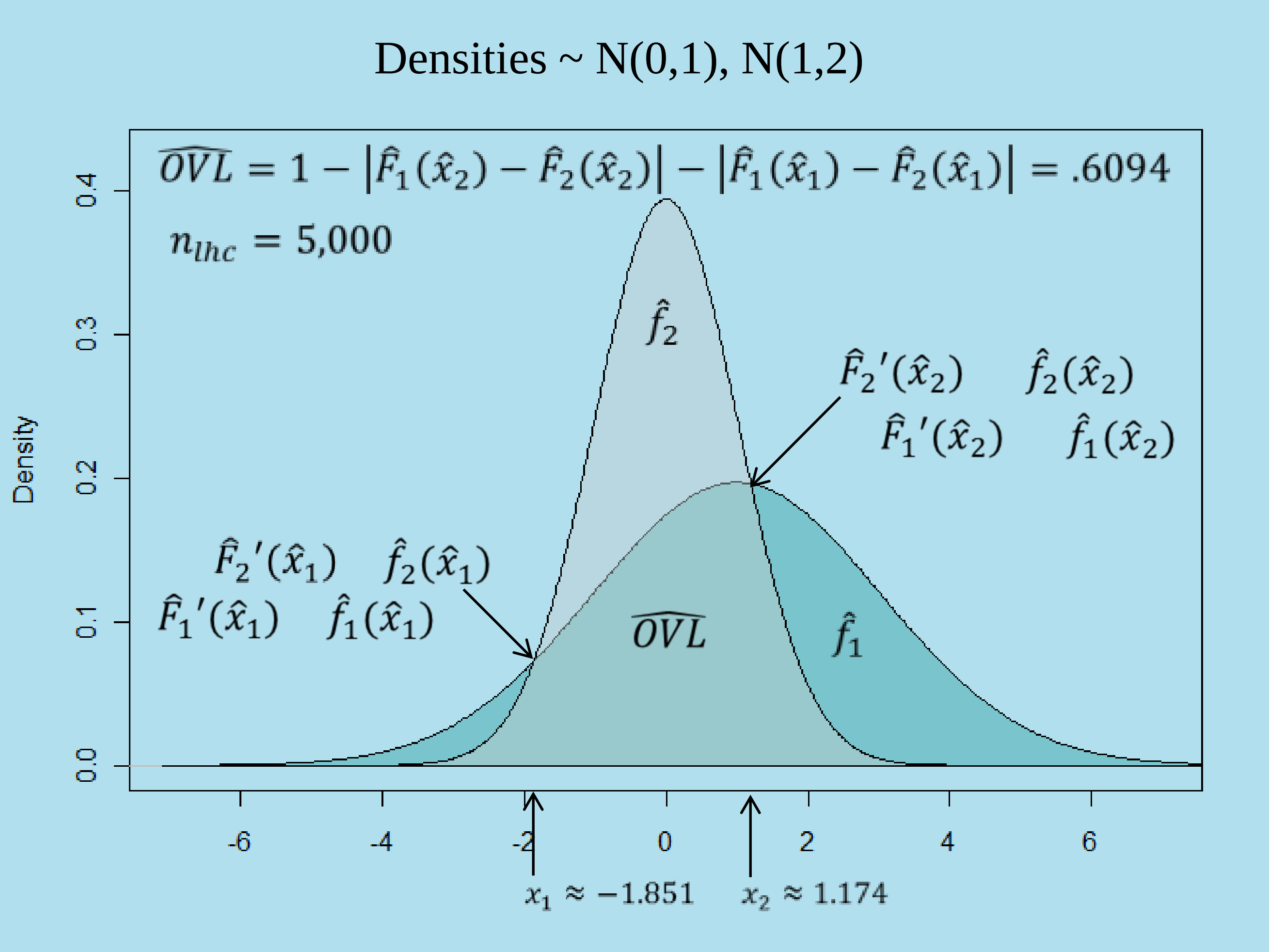
- Plotting every distance between S-Curves reveals. . .




- Graphic based on R code by COOLSerdash posted at http://stats.stackexchange.com/questions/59654/value-at-d-max-from-kolmogorov-smirnov-test-in-r

# How does one get away with this?

- Large sample size. . . 5,000 LHC trials



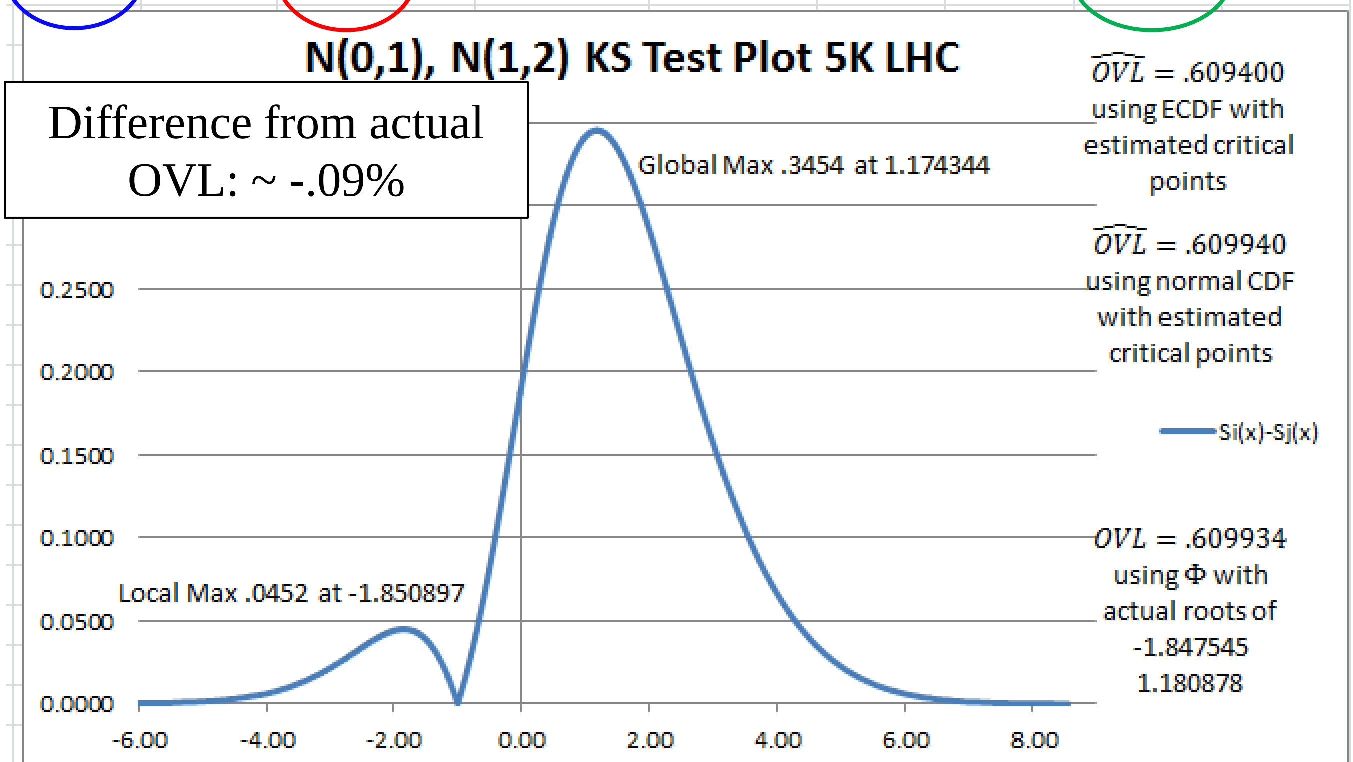
# How good is this OVL?

tions used in Inman (1984), Inman & Bradley (1989)

| Max | Crit Point | Crit Value | Row | Count Crit Vs | N(1,2)(x1) | N(0,1)x1 | N(1,2)(x2) | N(0,1)(x2) |
|---|---|---|---|---|---|---|---|---|
| Global | 1.174344 | 0.3454 | 7075 | 4 | 0.0772 | 0.0320 | 0.5346 | 0.8800 |
| Local | -1.850897 | 0.0452 | 548 | 5 | | | | |

| OVLhat calculations | | | % Change from actual OVL | | | Actual OVL |
|---|---|---|---|---|---|---|
| ECDF | Excel CDF | Inman (spline) | ECDF | Excel CDF | Inman (spline) | Excel CDF |
| 0.609400 | 0.609940 | 0.583812 | -0.000875505 | 9.21885E-06 | -0.0427 | 0.609934 |

Difference from actual OVL: ~ -.09%



# We want it now!

- "And if one has a method, its usefulness depends very much on whether it works quickly."
  - – The Princeton Companion to Mathematics
- Free CD?
  - Includes Excel file showing how to calculate the KS Two Sample Test and generate an $\widehat{OVL}$ from the data

# I want it yesterday!

- Special case for overlap with one intersection!
- Generate a couple hundred samples
- Paste into this web application and execute:
  - http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html

- Kirkman, T.W. (1996) Statistics to Use. http://www.physics.csbsju.edu/stats/ (May 15, 2014)

- Subtract the result from 1 and that's your $\widehat{OVL}$!
  - Remember the derivative of the max distance is where your probability density functions intersect
- More accuracy? Download the PAST tool for free
  - http://folk.uio.no/ohammer/past/

Hammer, Ø., Harper, D.A.T., Ryan, P.D. 2001. PAST: Paleontological Statistics software package for education and data analysis. Palaeontologia Electronica 4(1): 9 pp.(May 15, 2014)

# Overlap Wrap

- History of the common picture but obscure measure
  - Includes effort from the early days of cost analysis
- Application is wherever practical meaning is needed
  - In the context of comparing probability models or data
- Number quantifying OVL & $\widehat{OVL}$ is accessible
- Direct calculation from ECDF is the elegant method
  - But fitting distributions and using parameters could be quick
  - One intersection case yields quick answer with 1- D Statistic

# References

## Most citations pulled from WorldCat


1. Connolly, Billy. 1991 [Excerpt from] *Live at Hammersmith Odeon.* [aka, *Business Plan*.] http://www.youtube.com/watch?v=ggcZHdq6Jm0 Retrieved 15 May 2014
2. Bradley, E. L. 2006. Overlapping Coefficient. Encyclopedia of Statistical Sciences.
3. Inman, H. F., & Bradley, E. L. (January 01, 1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods, 18*, 10, 3851-3874.
4. Weitzman, M. S. (1970). *Measures of overlap of income distributions of white and Negro families in the United States*. Washington: U.S. Bureau of the Census; [for sale by the Supt. of Docs., U.S. Govt. Print. Off.
5. Madhuri, S. M., & Satya, N. M. (January 01, 1994). Overlap Coefficients of Two Normal Densities--Equal Means Case. *日本統計学会誌 / 日本統計学会 編, 24*, 2.)
6. Inman, H. F. (1984). *Behavior and properties of the overlapping coefficient as a measure of agreement between distributions*.
7. Clemons, T. E. (1997). *A nonparametric approach to estimating the overlapping coefficient using the kernel estimation technique*.
8. Clemons, T. E., & Bradley, E. L. (January 01, 2000). A nonparametric measure of the overlapping coefficient. *Computational Statistics and Data Analysis, 34*, 1, 51-61
9. Berman, E. B. (1958). *Toward a new weapon system analysis*. Santa Monica, Calif: Rand Corp.
10. E. B. Berman (personal communication, Sep 30, 2013)
11. Hough, P. G., & Rand Corporation. (1989). *Birth of a profession: Four decades of Military cost analysis*. Santa Monica, CA: Rand Corp.
12. Sobel, S. (1965). *A computerized technique to express uncertainty in advanced cost estimates*. Mitre Corp.
13. Dienemann, P. F. (1966). *Estimating cost uncertainty using Monte Carlo techniques*. Santa Monica, Calif: Rand Corp.
14. Garvey, P. R. (2000). *Probability methods for cost uncertainty analysis: A systems engineering perspective*. New York: M. Dekker.
15. E. B. Berman (personal communication, Oct 10, 2013)
16. Novick, D., & Rand Corporation. (1988). *Beginning of military cost analysis, 1950-1961*. Santa Monica, CA: Rand Corp
17. Student (1908a). *The probable error of a mean.* Biometrika VI, 1-25.
18. Mulekar, M. S. & Champanerkar, J. (2011). *Modeling Sampling Distributions Of Similarity Measures*. Section on Statistical Computing – JSM 2011
19. Conover, W. J. (1999). *Practical nonparametric statistics*. New York, NY [u.a.]: Wiley.
20. Sheskin, D. (2011). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Fla: Chapman & Hall/CRC.

# Backup

Contains:

- Excel for Risk Shape Example

# Risk Shape Example

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Calculations for OVLhat of ACE Risk Shapes** | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | |
| 4 | Low and High parameters are multipliers about the mode from ACE help file | | | | | | | | | | | | | |
| 5 | | Low | Mode | High | | | | | | | | | | |
| 6 | f2 | 0.816 | 1.000 | 1.551 | 0 | 0 | "PE likely less" | | | | | | | |
| 7 | f1 | 0.571 | 1.000 | 2.286 | 0 | 0 | "PE likely a lot less" | | | | | | | |
| 8 | | | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | | | |
| 10 | Visually determined intersections | | | | | | | | | | | | | |
| 11 | x1 | 0.88 | | | | | | | | | | | | |
| 12 | x2 | 1.38 | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | | | |
| 14 | | | f2(x1) | 0.030287 | = IF(B11<B6,0,IF(B11<C6,(B11-B6)^2/((D6-B6)*(C6-B6)),IF(B11<=D6,1-(D6-B11)^2/((D6-C6)*(D6-B6)),1))) | | | | | | | | | |
| 15 | | | f2(x2) | 0.927797 | = IF(B12<B6,0,IF(B12<C6,(B12-B6)^2/((D6-B6)*(C6-B6)),IF(B12<=D6,1-(D6-B12)^2/((D6-C6)*(D6-B6)),1))) | | | | | | | | | |
| 16 | | | f1(x1) | 0.129776 | = IF(B11<B7,0,IF(B11<C7,(B11-B7)^2/((D7-B7)*(C7-B7)),IF(B11<=D7,1-(D7-B11)^2/((D7-C7)*(D7-B7)),1))) | | | | | | | | | |
| 17 | | | f1(x2) | 0.627821 | = IF(B12<B7,0,IF(B12<C7,(B12-B7)^2/((D7-B7)*(C7-B7)),IF(B12<=D7,1-(D7-B12)^2/((D7-C7)*(D7-B7)),1))) | | | | | | | | | |
| 18 | | | OVLhat | 0.60053 | | | | | | | | | | |
| 19 | | | | | | | | | | | | | | |
| 20 | | | f2(x1) | 0.030326 | =CB.GetCertaintyFN(F6,B11)/100 | | | | | | | | | |
| 21 | | | f2(x2) | 0.927845 | =CB.GetCertaintyFN(F6,B12)/100 | | | | | | | | | |
| 22 | | | f1(x1) | 0.129846 | =CB.GetCertaintyFN(F7,B11)/100 | | | | | | | | | |
| 23 | | | f1(x2) | 0.627866 | =CB.GetCertaintyFN(F7,B12)/100 | | | | | | | | | |
| 24 | | | OVLhat | 0.60050 | =1-ABS(D21-D23)-ABS(D20-D22) | | | | | | | | | |