

Using Projection Splines to Explore Differences in Distributions

Nedra Whitehead, MS, PhD

Jun Liu, PhD

Lei Li, PhD

Jason Hsia, PhD

Outline of presentation

1. Introduction
2. Illustration of projection spline methodology
3. Case study
 - Exploring Racial Differences in Gestational Age Distribution among Very Low Risk Women

Background

- Compare distributions
 - Shape
 - Spread
 - Location
- Graphical methods
 - Visual assessment
- Statistical testing
 - Mostly location
 - Fewer spread
 - Very few shape

Methodology

- Rotated projection plots
- Projection splines
 - Shape, spread and location
 - Randomly sampled data

Jones CP. Living beyond our "means": new methods for comparing distributions. *American Journal of Epidemiology* 1997; 146(12):1056-1066.

Splines

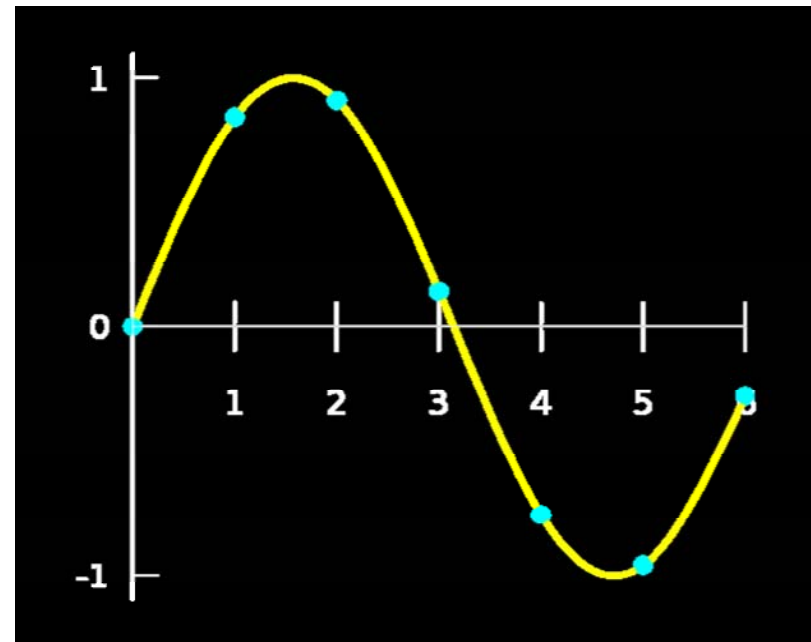
- Originated in drafting



Source: www.duckworksmagazine.com/.../splineDucks.htm

Splines in Statistics

- Mathematical spline
 - Series of functions joined together to fit a curve
 - Points of join are called knots
- Regression spline
 - Computed by regression model



Source: commons.wikimedia.org

Projection Spline Methodology

- Least squares regression
 - $y = \beta_0 + \beta_1 x + \beta_2 x k_1 + \beta_i x k_{(i-1)}$
 - SUDAAN sandwich variance estimator
- Violation of assumptions
 - Correlation between percentiles
 - Heteroscedasticity
- Statistical tests
 - Global differences
 - Specific changes in slope

Projection Spline Methodology

1. Initial decisions
 - a. Initial spacing of knots
 - b. Minimum number of quantiles between knots
 - c. 2-sided type 1 error (α) for maintaining the knot in the model
2. Fit initial model with all knots
3. Iter-1 test for global differences
 - H_0 : No difference in distributions ($\beta_{2,\dots}, \beta_i = 0$)
 - H_A : Difference in shape, spread, or location (at least 1 $\beta_{2,\dots}, \beta_i \neq 0$)
4. Iteratively refit

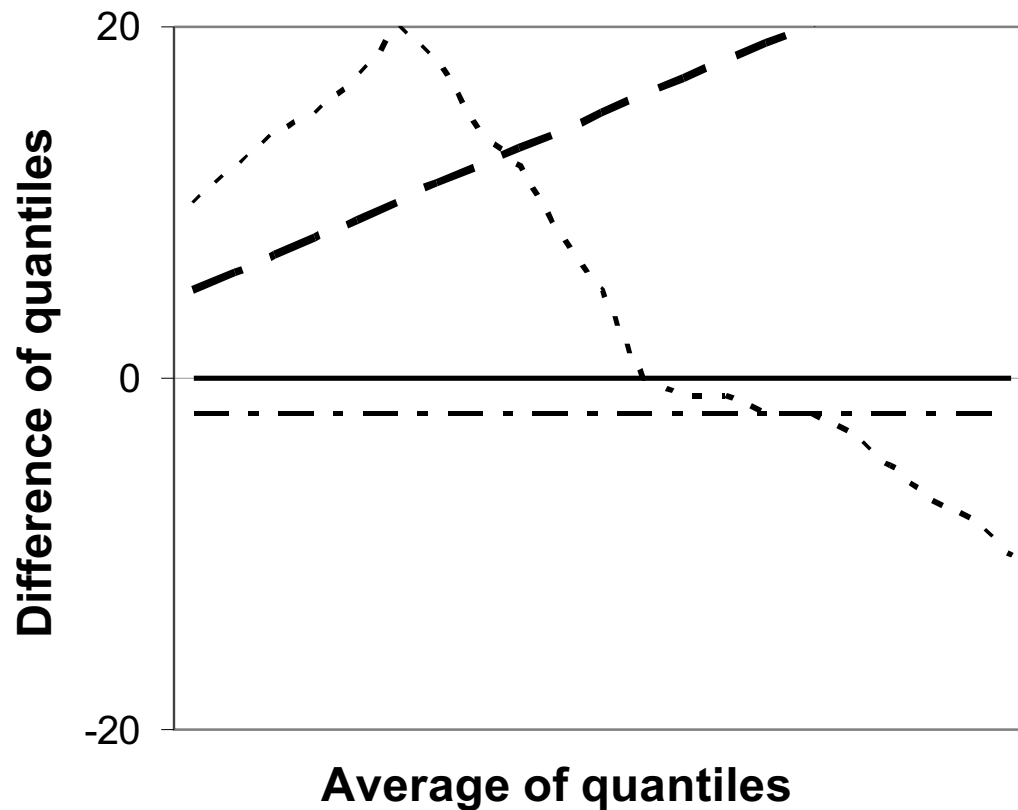
Projection splines methodology

- a) For each parameter
 - a) Wald test statistic
 - b) Associated p -value
- b) Remove least significant knot
- c) Refit model
- d) Repeat until no nonsignificant knots

Parameters	Wald test p-value
$\beta_2 \times k_1$.64
$\beta_3 \times k_2$.10
Parameters	Wald test p-value
$\beta_3 \times k_2$.01

Interpretation

Interpretation of Projection Plots and Splines



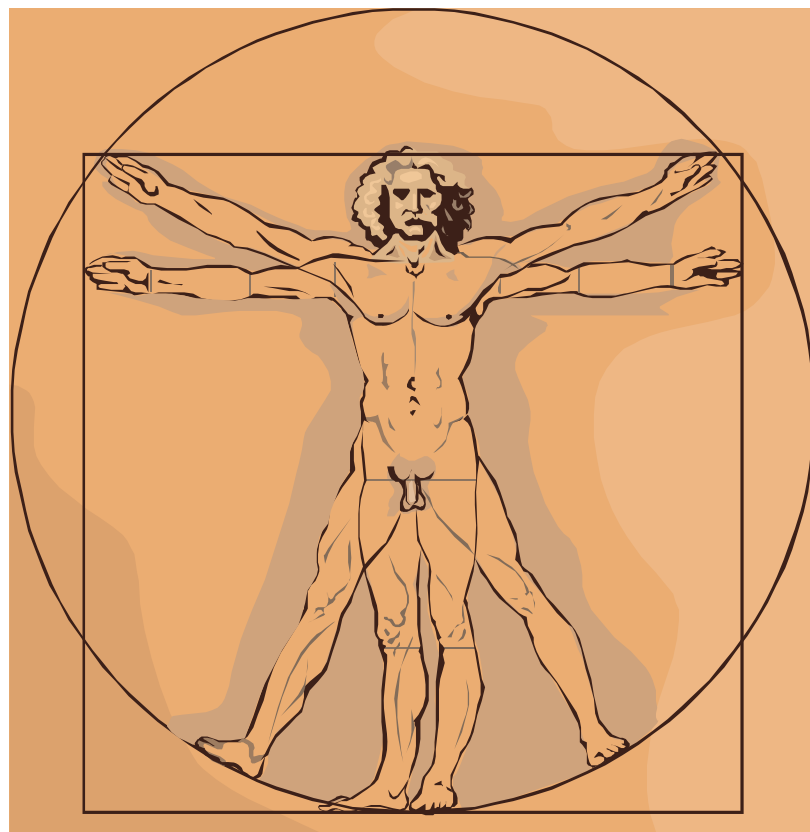
— Linear (no significant knots);
Slope = 0, Intercept = 0.
Distributions are the same.

- - - Linear (no significant knots);
Slope = 0, Intercept $\neq 0$:
Distributions differ in location.

- . - Linear (no significant knots);
Slope $\neq 0$, Distributions differ in
spread.

. . . Not linear ($>.1$ significant knots)
Distributions differ in shape.

Illustration of Methodology



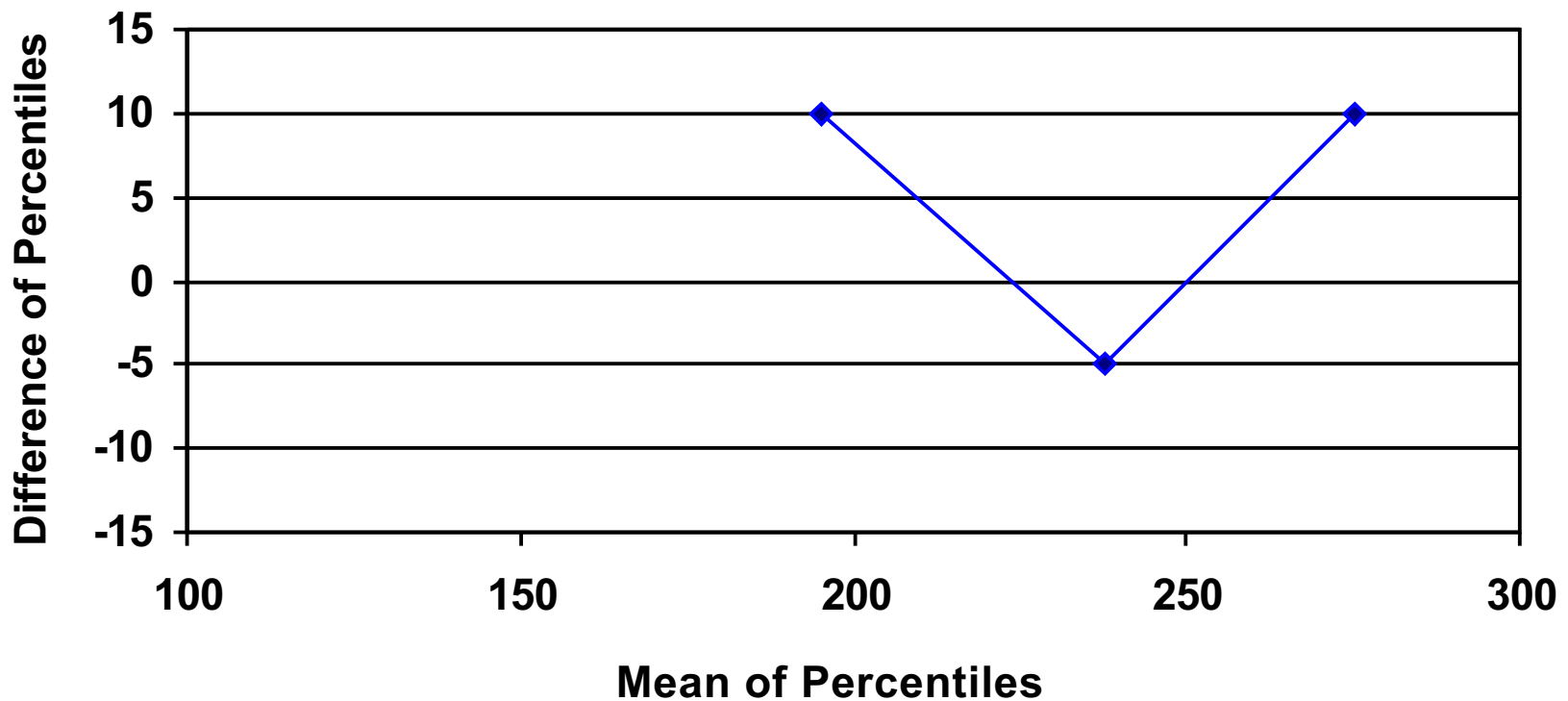
Data

Percentiles of Two Populations and Their Mean and Difference

Quartile	Population1	Population2	$x_i = (q_{1i} + q_{2i}) / 2$	$y_i = q_{1i} - q_{2i}$
1	190	200	195	10
2	240	235	237.5	-5
3	280	270	275	10

Rotated projection plot

Plot difference against ordered mean



Initial Setup of Projection Splines

- Calculate the covariance matrix of y
- Calculate initial knots
 - $x_{k_i} = 0$ where $x = k$
 - $x_{k_i} = (x-k)$ where $x >$
- Collapse knots if < 3 quantile values between knots

Value of x_{k_1} and x_{k_2} at selected values of x

	Values of x		
Initial knots	175	210	260
168	7	42	92
240	0	0	20

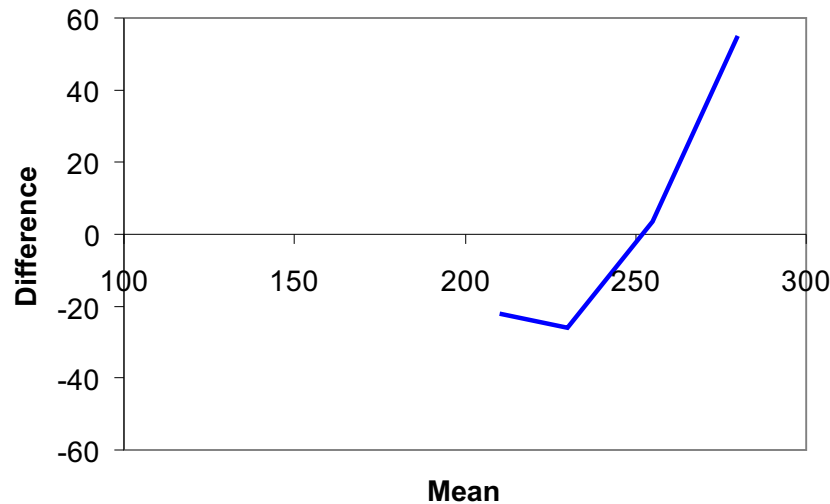
First Iteration

- Fit initial model
 - $y = \beta_0 + \beta_1x + \beta_2xk_1 + \beta_3xk_2$
- Iter-1 test
 - Wald F test statistic for H_0 : 175.10
 - Wald F test statistic for H_A : 266.0
 - iter-1 test statistic : 90.9
 - Degrees of freedom: 2
 - p-value: 0.01

Model term	Beta	p-value
Intercept	12.55	0.003
x	-0.14	0.01
xk ₁	0.003	0.64
xk ₂	2.05	0.10

Iterative Refitting

1. Drop the least significant knot
 - Refit the model
 - $y = \beta_0 + \beta_1x + \beta_2xk_2$



Model term	Beta	p-value
Intercept	15.55	0.37
x	-0.18	0.005
xk ₂	2.25	0.01

Case Study

- Compare
 - Gestational age
 - By race
 - Very low risk women
 - Demographic
 - Behavioral



Background

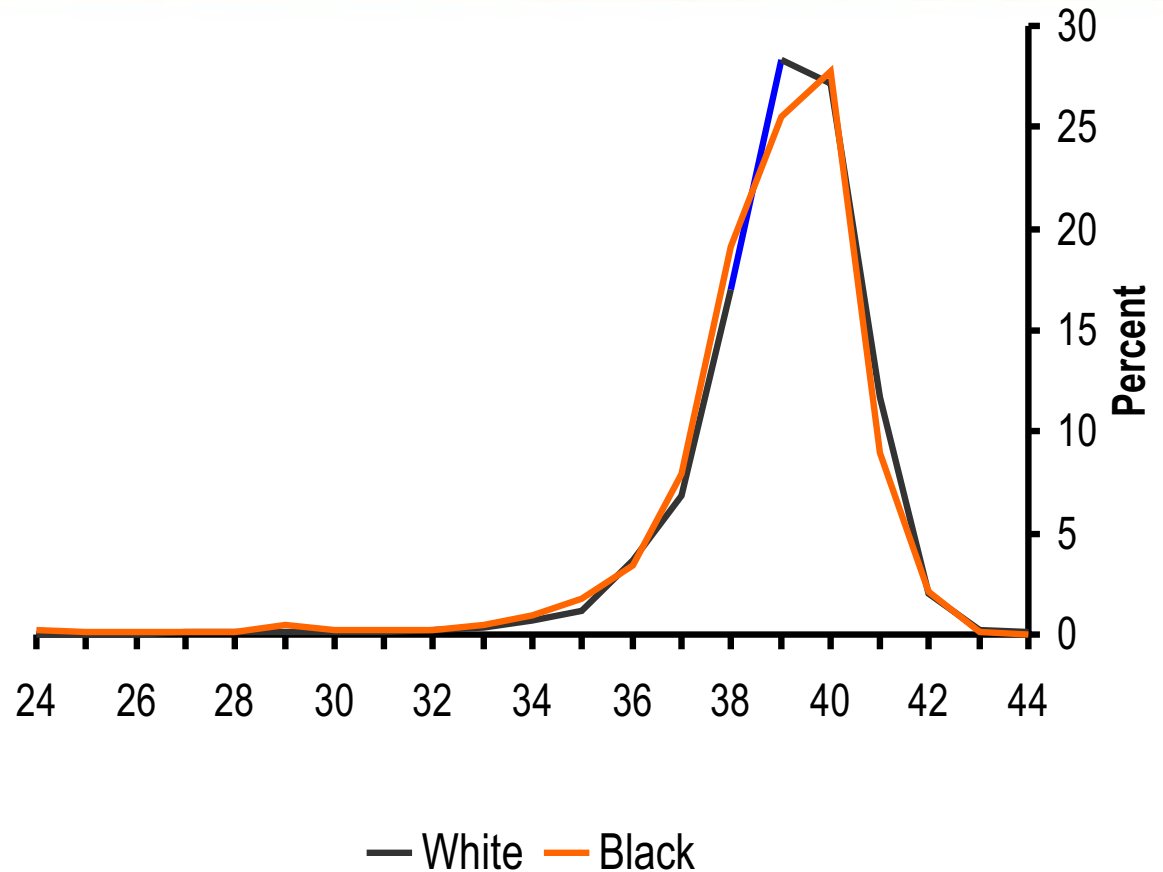
- Preterm infants
 - ? Neonatal death
- Infants of black women
 - ? Preterm
 - ? Death
- Naturally shifted distribution?



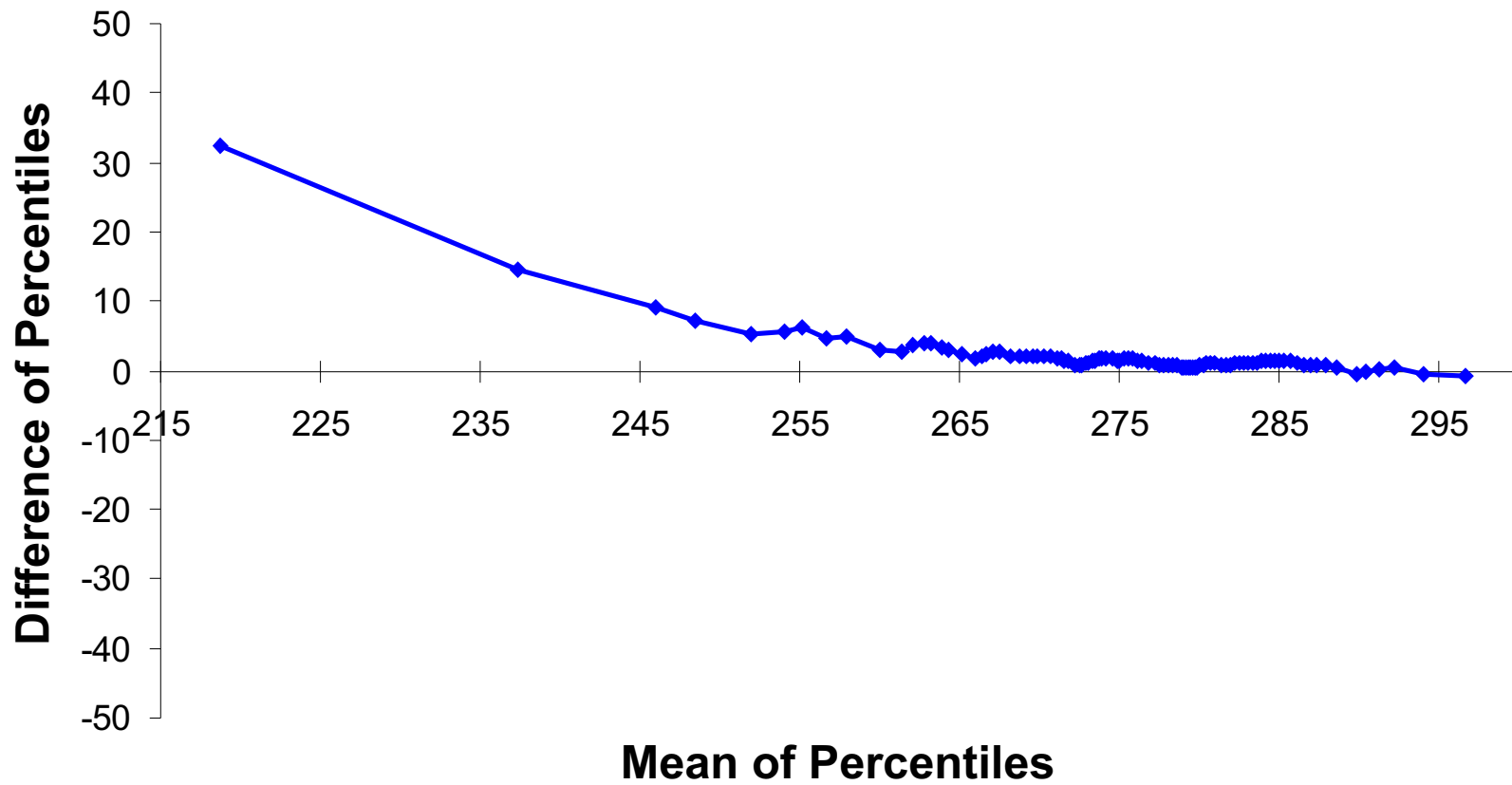
Gestational Age Distribution

Mean gestational age at delivery, by race

	Mean	SE
White	275.5	0.12
Black	273.4	0.51
Difference	-2.06	
P-value	<0.01	



Fitted Projection Spline



Projection Splines

- Setup
 - Knots
 - Every 3.5 days
 - > 3 between knots
 - 11 knots
- $\alpha = < 0.01$

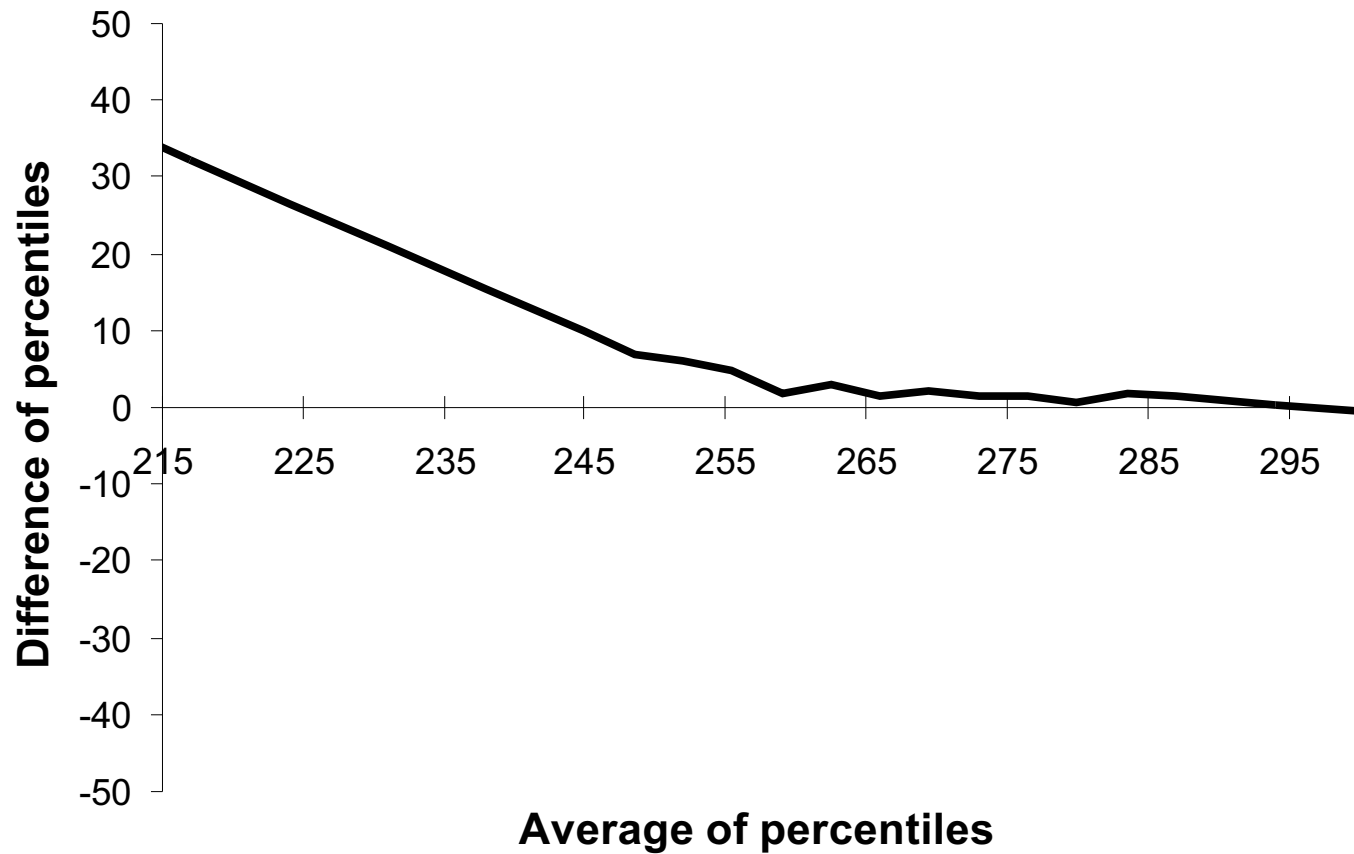
Initial model

- Knots from 248.5 – 287 days
- Iter-1 test
 - Wald F Test
 - Test Statistic – 1081.65
 - Degrees of freedom – 13
 - $p < 0.0001$

Final Model

Variable	Parameter (SE)	P-value
Intercept	205.84 (3.80)	< 0.0001
Mean (x)	-0.80 (0.02)	< 0.0001
Knot 1 (248.5)	0.50 (0.08)	< 0.0001
Knot 2 (255.5)	-0.61 (0.20)	0.0034
Knot 3 (259)	1.31 (0.27)	< 0.0001
Knot 4 (262.5)	-0.87 (0.19)	< 0.0001
Knot 5 (266)	0.61 (0.13)	< 0.0001
Knot 6 (269.5)	-0.30 (0.08)	0.0003
Knot 7 (273)	0.22 (0.07)	0.0021
Knot 8 (276.5)	-0.35 (0.06)	< 0.0001
Knot 9 (280)	0.65 (0.06)	< 0.0001
Knot 10 (283.5)	-0.51 (0.05)	< 0.0001

Fitted Projection Spline



Discussion

- Summary of Findings
 - Gestational age distributions differ
 - Greatest at preterm gestational ages
 - Not simply shifted
- Strengths
 - Women without most known risk factors
- Limitations
 - No information on infections

Discussion - Methodology

- Strengths
 - Useful information
 - Statistical test for differences
 - Adaptable to nonrandomly sampled data
- Limitations
 - Ignores correlation between percentiles
 - Sensitive to small differences

More Information

Visit us in booth 513/612

Nedra Whitehead
2951 Flowers Road
Suite 119
Atlanta, Georgia 30341
Phone: 770-986-5051
Email: nwhitehead@rti.org