DETECTING ANOMALIES IN NATIONAL MORTALITY DATA

Andrew Cairns

Heriot-Watt University, Scotland

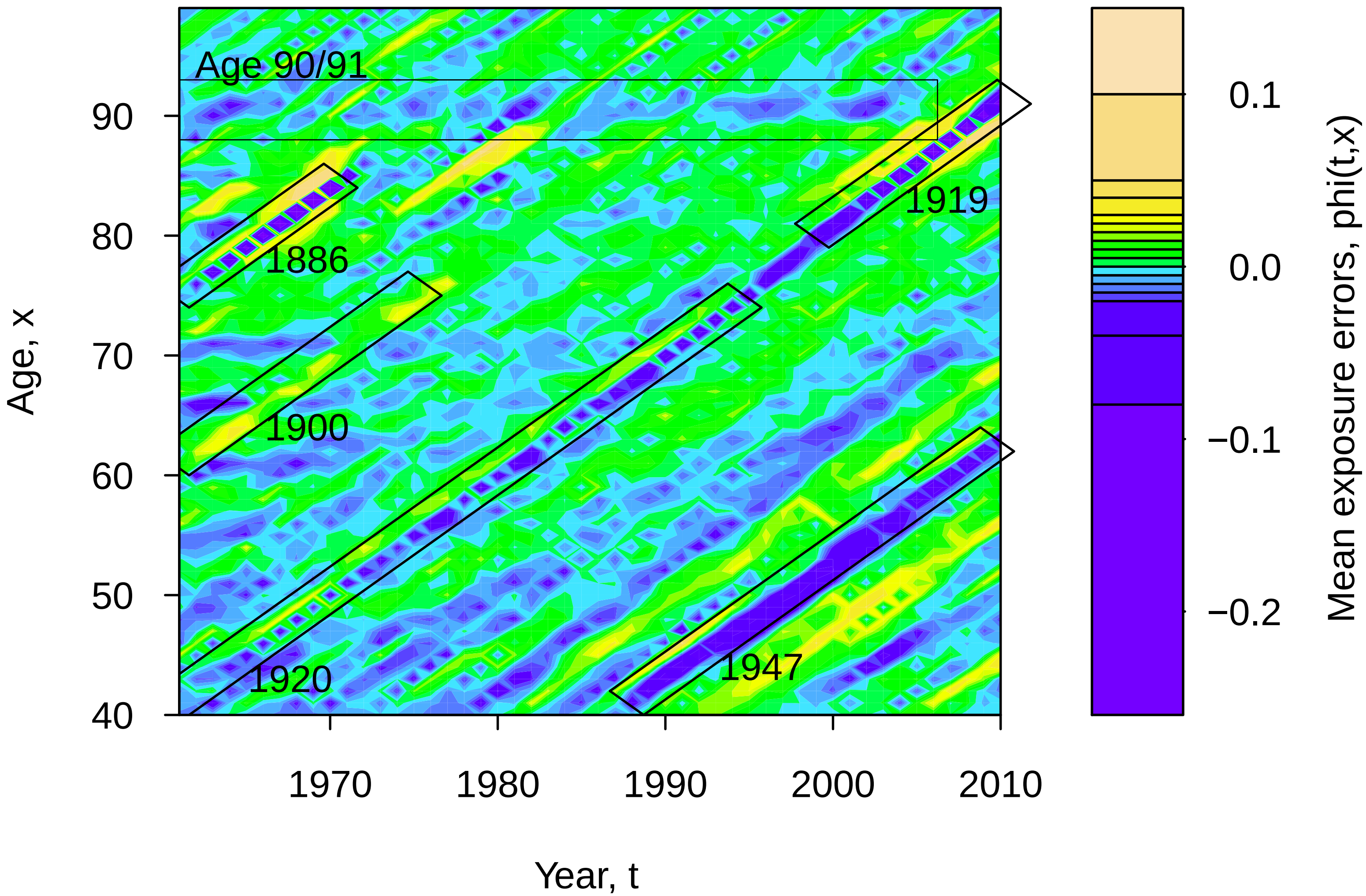
and

The Maxwell Institute, Edinburgh

Joint work with: David Blake, Kevin Dowd and Amy Kessler

International Mortality and Longevity Symposium 2014

Potential Errors in *post-2011* Population Estimates



Source data: ONS EW males deaths and *revised* population estimates.

Plan

1. Background and motivation

2. Data issues: deaths, population, exposures

3. Graphical diagnostics and signature plots

4. Model-based analysis of historical population data

5. Conclusions and next steps

1: Background and Motivation

- England and Wales data $+$ other countries
- $D(t, x)$: Death counts considered to be accurate
- $P(t + \frac{1}{2}, x)$ mid-year population is an *estimate*
- Crude $m(t, x) = D(t, x)/P(t + \frac{1}{2}, x)$ not $D(t, x)/E(t, x)$
- Post 2011 census revisions $\Rightarrow$ some big revisions
- Similar magnitude revisions after 2001 census

Why Do Errors in Population Data Matter?

Potential impact on

- Population mortality forecasts
- Forecasts of sub-population mortality
- Calibration of multi-population models
- Calculation of annuity liabilities and Value-at-Risk
- Assessed levels of uncertainty in the above
- Buyout pricing
- Assessment of basis risk in longevity hedges
- Assessment of hedges and hedging instruments

Aims

- How to identify anomalies in data
- How to *pre-whiten* your mortality data

  before modelling and forecasting

2: Population Estimates, Exposures, Death Rates

$$\text{Death rate } m(t,x) = \frac{D(t,x)}{E(t,x)}$$

- $E(t,x) =$ 'exposure' in year t (central exposed to risk)

  $= $ *average value* of $P(s,x)$ from t to $t+1$

  $P(s,x) =$ population at exact time s aged x last birthday

- England & Wales $\Rightarrow$ only $P(t+\frac{1}{2},x)$ reported

- Common assumption: $E(t,x) = P(t+\frac{1}{2},x)$

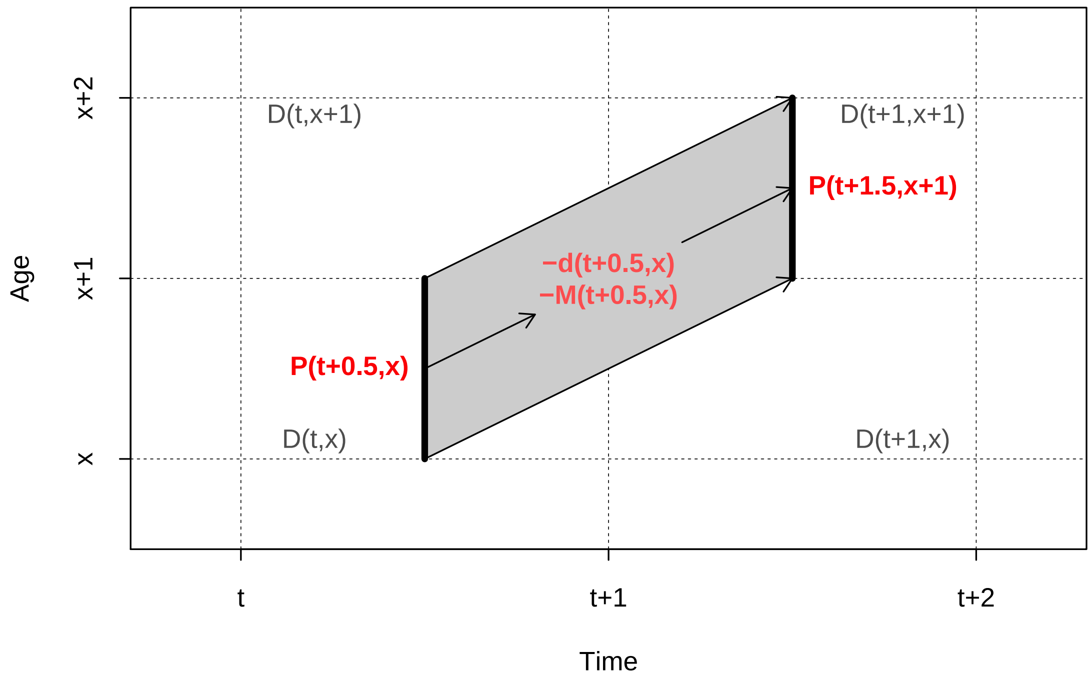
  - e.g. ONS reported death rates: $m(t,x) = D(t,x)/P(t+\frac{1}{2},x)$

2.1: Where Can Errors in $E(t, x)$ Occur?

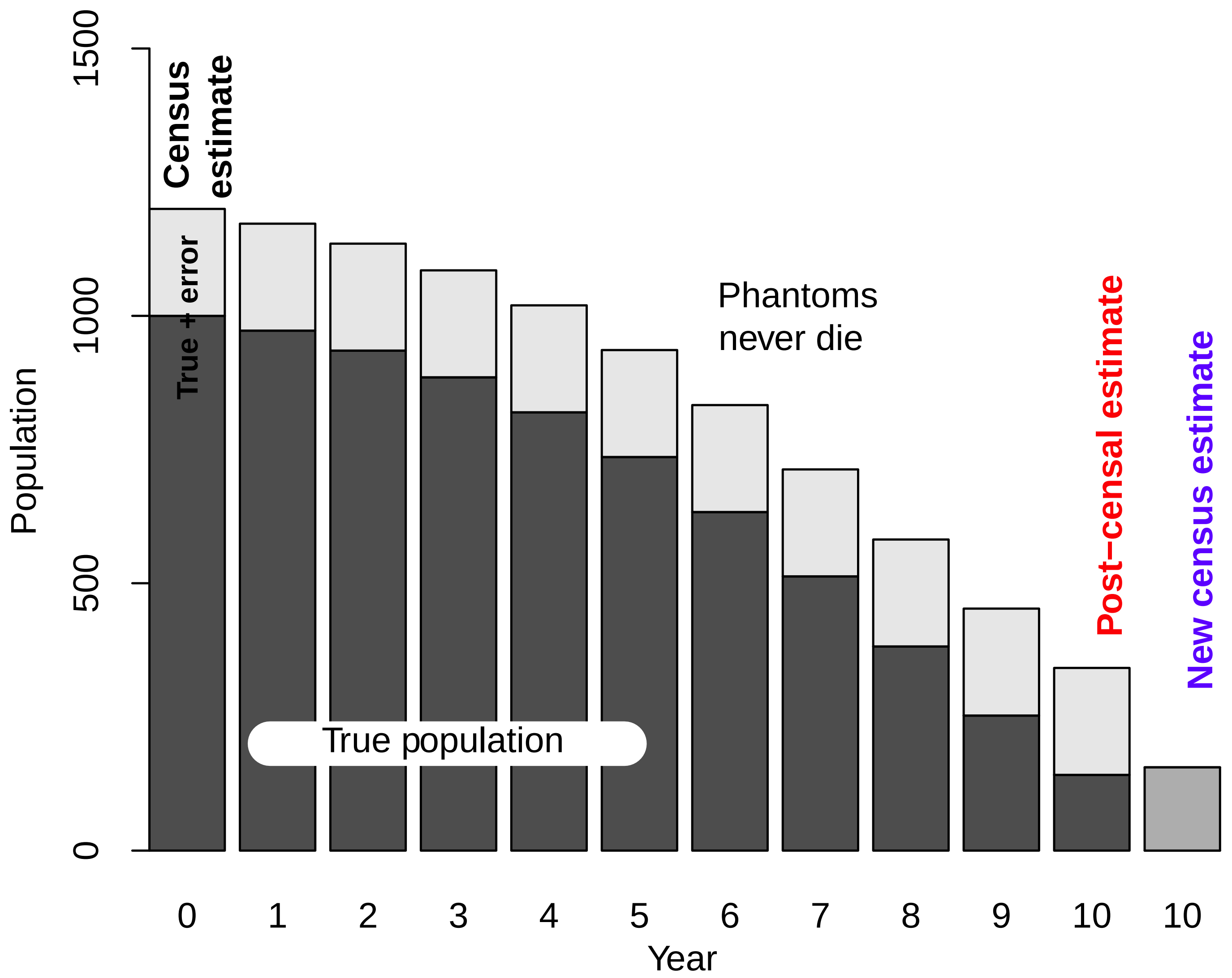
- Known errors: Inaccurate $P(t + \frac{1}{2}, x)$
  - no ID card system
  - infrequent censuses, under-enumeration
  - migration etc.
  - mis-reported age at census
- Lesser known errors:
  - inaccurate shift from census date to mid-year
  - assumption that $P(t + \frac{1}{2}, x) \approx E(t, x)$

2.1.1: Propagation of General Errors Through Time

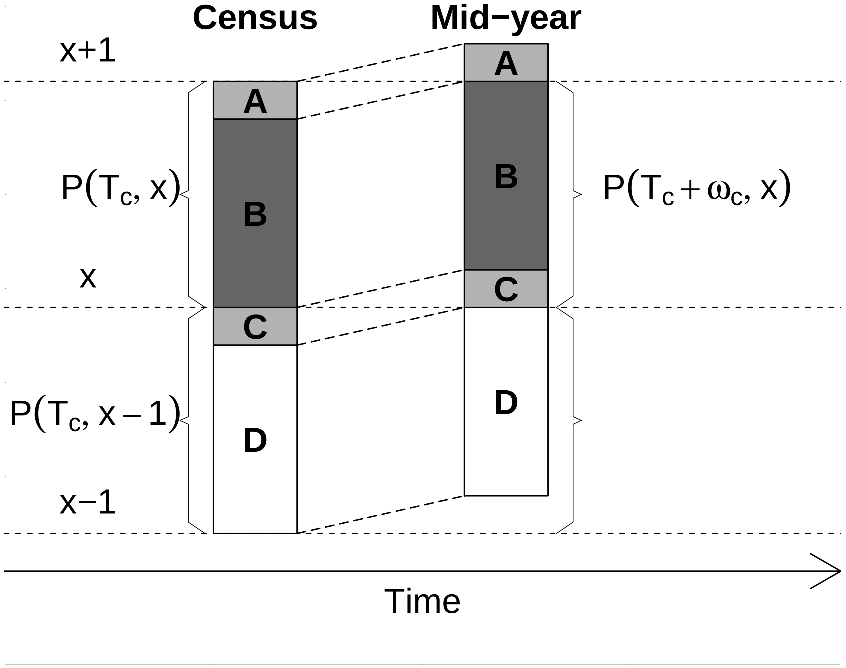
Errors follow cohorts $\Rightarrow$ “Phantoms never die”



Phantoms Never Die



2.2: Census to Mid-year Shift



ONS 2001 assumption: *birthdays spread evenly throughout the year*

Conjecture:

– different methodology used in earlier censuses and in 2011

Can We Improve on This Assumption?

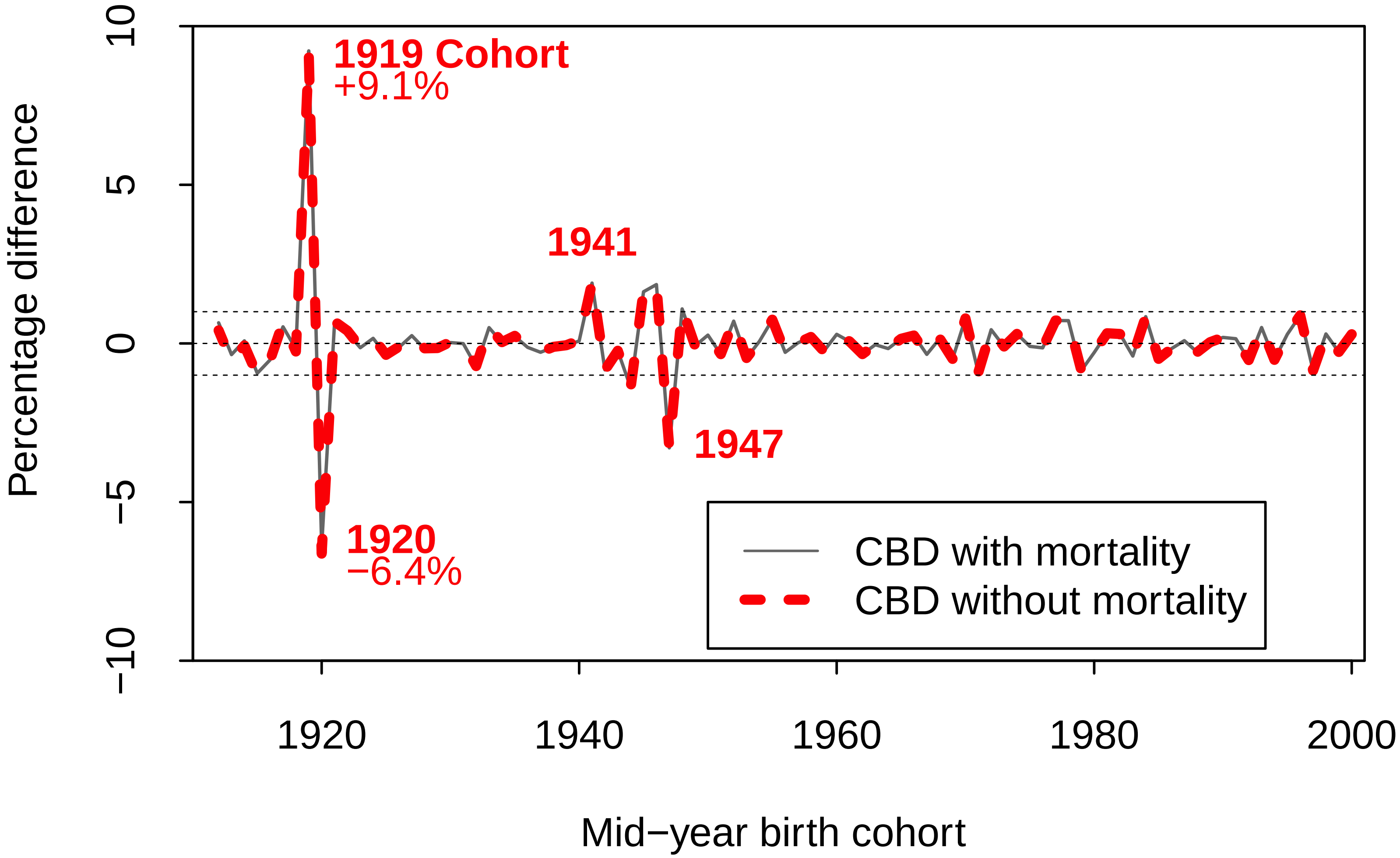
The Cohort Births/Deaths (CBD) Exposures Methodology

Underlying hypothesis:

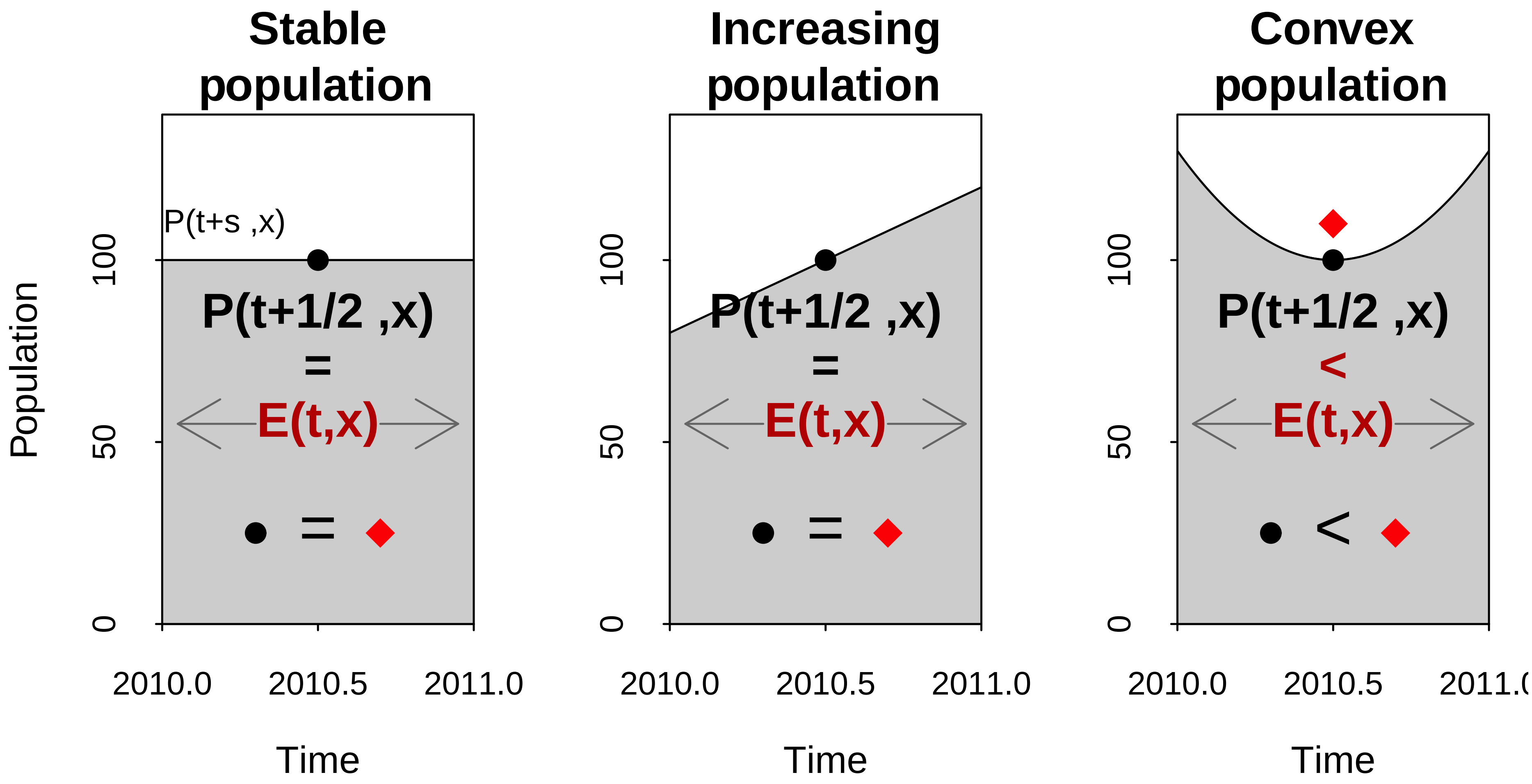
- At any point in time t, pattern of birthdays at t will reflect
  - actual pattern of births x years earlier
  - deaths (impact at high ages)
  - migration and birth patterns of immigrants
- Irregular pattern of births can lead to errors in census $\longrightarrow$ mid-year shift

Birth month	Age on 30/4/2001	Proportion	2001 census	ONS estimate	Age at mid-year	ONS mid-year
May-June 1918	82	2/12	72114	12019	83	
July 1918-April 1919	82	10/12		60095	82	79352
May-June 1919	81	2/12	115545	19257	82	
July 1919-April 1920	81	10/12		96288	81	

Birth month	No. of births	Age on 30/4/2001	Proportion	2001 census	CBD estimate	Age at mid-year	CBD mid-year
5-6/1918	113475	82	0.17785	72114	12825	83	
7/1918-4/1919	524566	82	0.82215		59289	82	72741
5-6/1919	99174	81	0.11642	115545	13452	82	
7/1919-4/1920	752725	81	0.88358		102093	81	


1919 Cohort
+9.1%
1941
1947
1920
−6.4%
CBD with mortality
CBD without mortality
Percentage difference
Mid−year birth cohort
10
5
0
−5
−10
1920
1940
1960
1980
2000

2.3: Proposal to Improve Estimates of Exposures



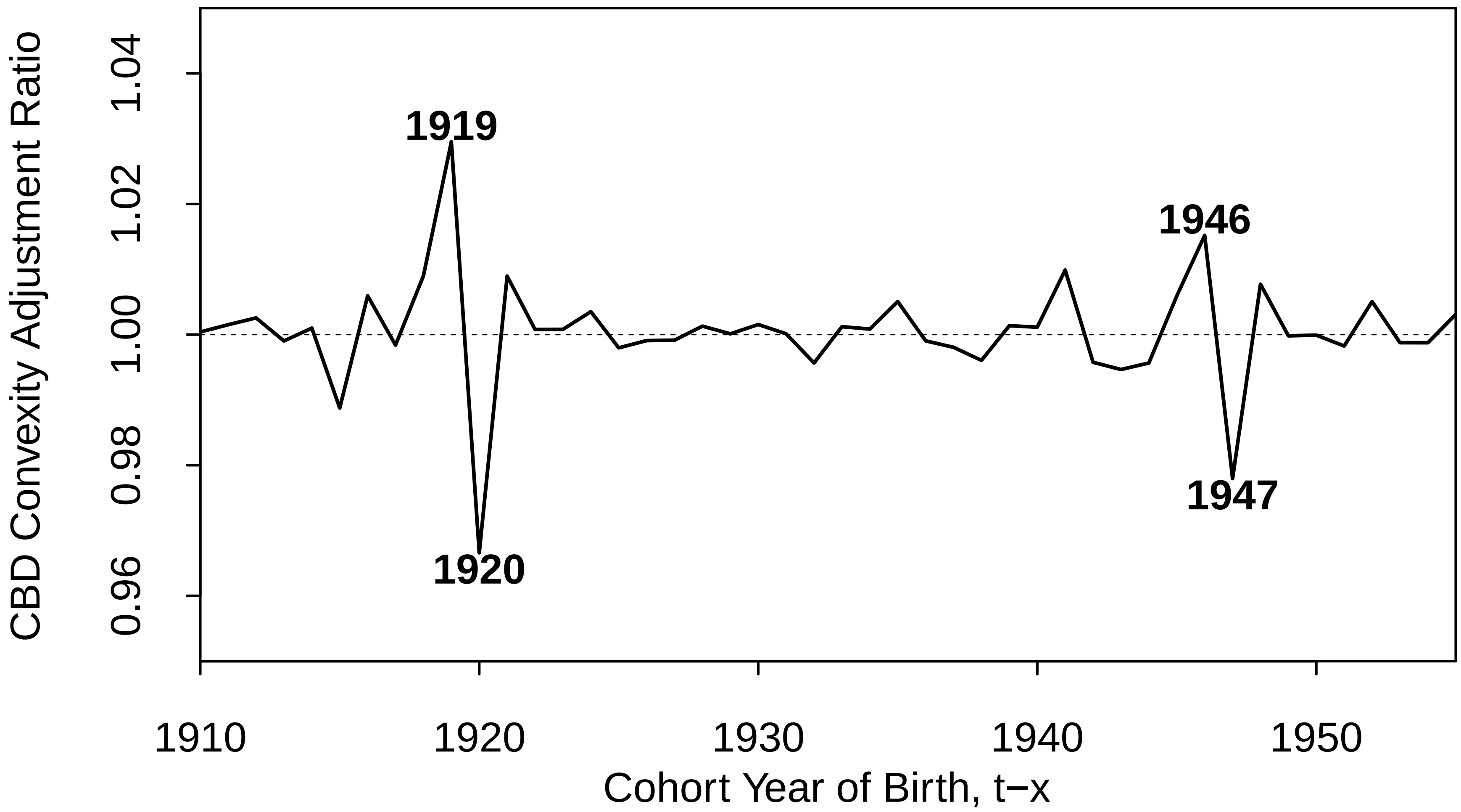
Proposal to Improve Estimates of Exposures

- Death rate $m(t,x) = D(t,x)/E(t,x)$
- Current assumption: $E(t,x) = P(t+\frac{1}{2},x)$
- CBD Exposures Methodology:

$$\text{Assume } E(t,x) = P(t+\frac{1}{2},x) \times \frac{E(t-x,0)}{P(t+\frac{1}{2}-x,0)}$$

- $E(t-x,0)/P(t+\frac{1}{2}-x,0) =$ Convexity Adjustment Ratio
- CAR based on monthly pattern of births over $t-x-1$ to $t-x+1$

CBD Exposures Methodology: Convexity Adjustment Ratio



2.4: High Age Methodology

- ONS reports
  - $P(t+\frac{1}{2}, 90+)$ only
  - $D(t,x)$ for $x = 90, 91, 92, \ldots$
- $P(t+\frac{1}{2}, x)$ for $x = 90, 91, \ldots$ derived using the Kannisto-Thatcher Method (extinct cohorts)
- Conjecture: Potential for inconsistencies at the boundary between ages 89 and $90+$

3: How to identify anomalies

Graphical Diagnostics and Signature Plots

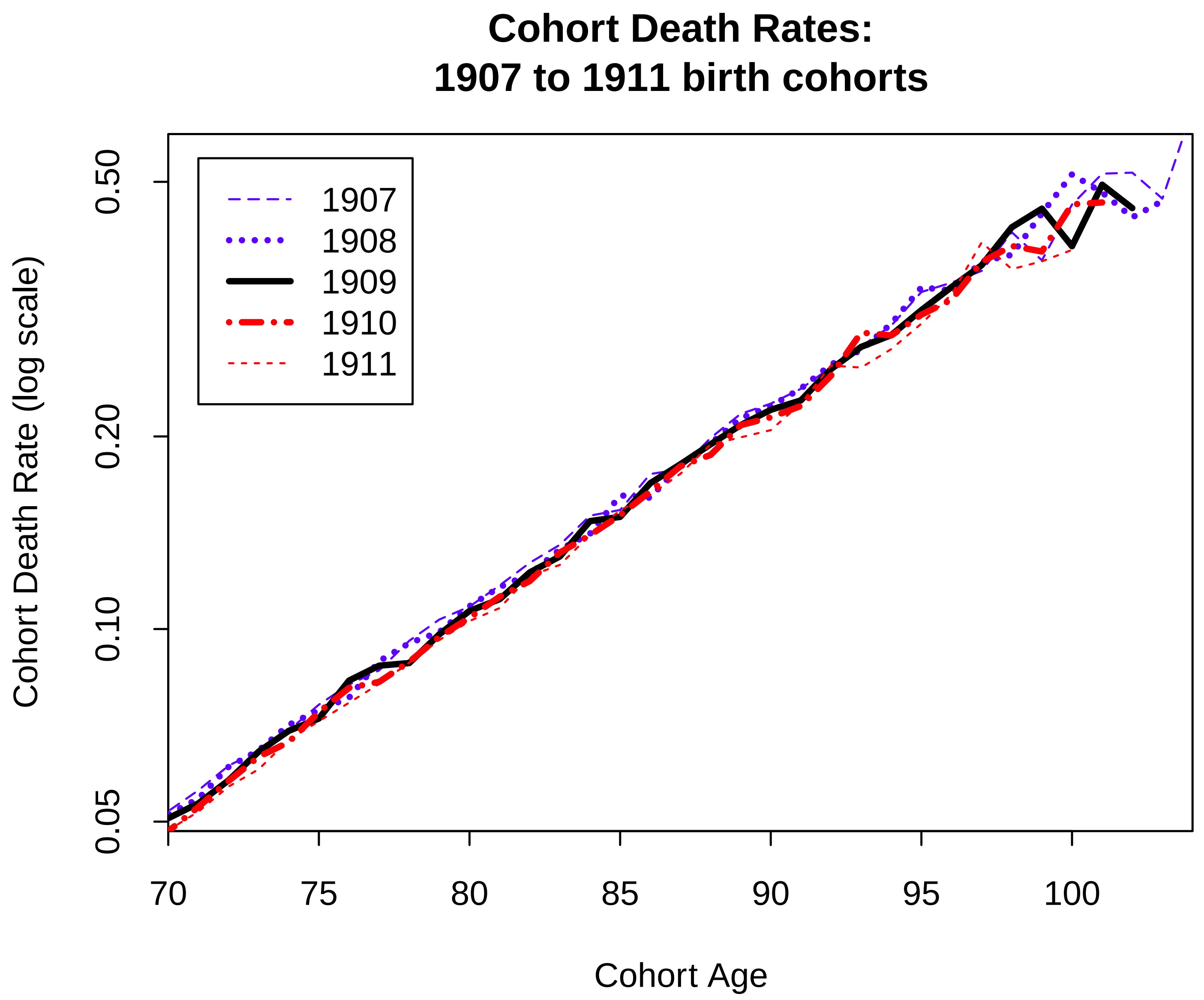
- Graphical diagnostics
  - hypothesis $\Rightarrow$

    plot should exhibit specific characteristics
- Signature plots
  - what if it does not?

3.1: Graphical Diagnostic 1

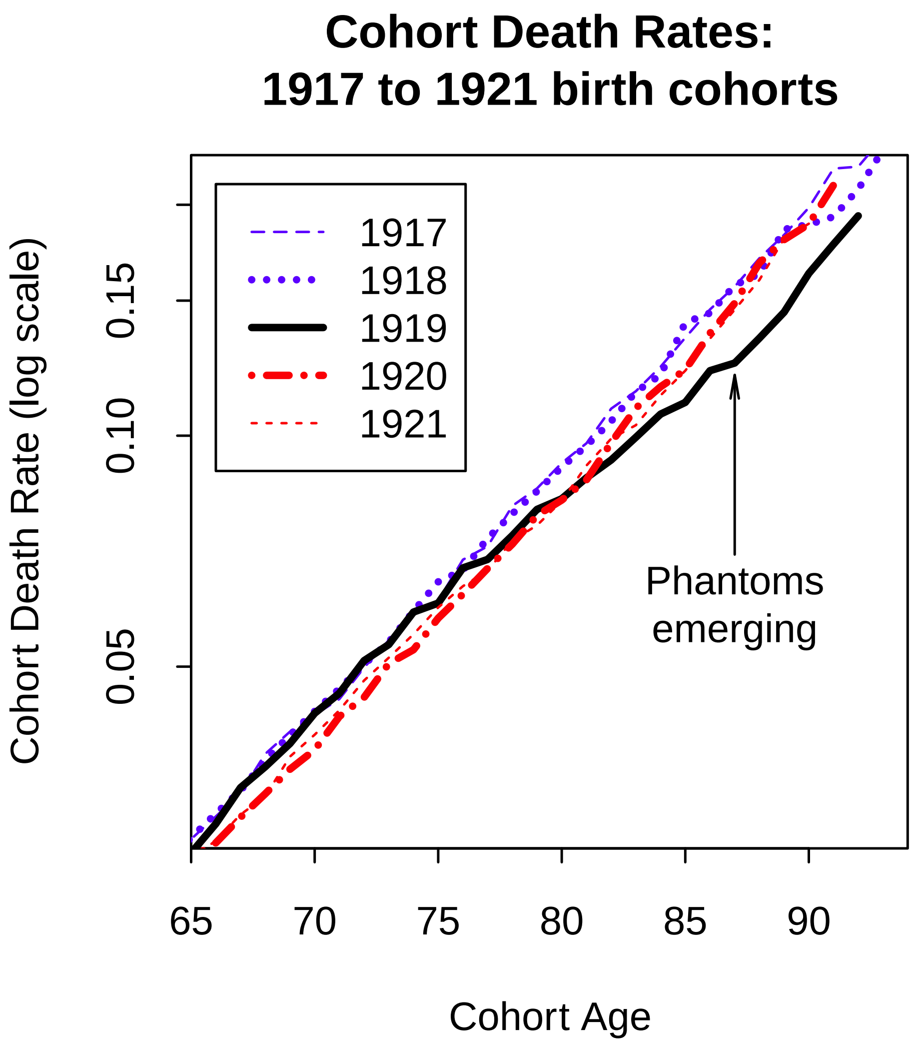
Hypothesis: *Crude death rates by age for successive cohorts should look similar.*

$\Rightarrow$ Plot crude death rates against age.



Cohort death rates by age for 1907 to 1911 cohorts. ONS revised EW males data up to 2011.

Signature Plot: Emergence of Phantoms



3.2: Graphical Diagnostic 2

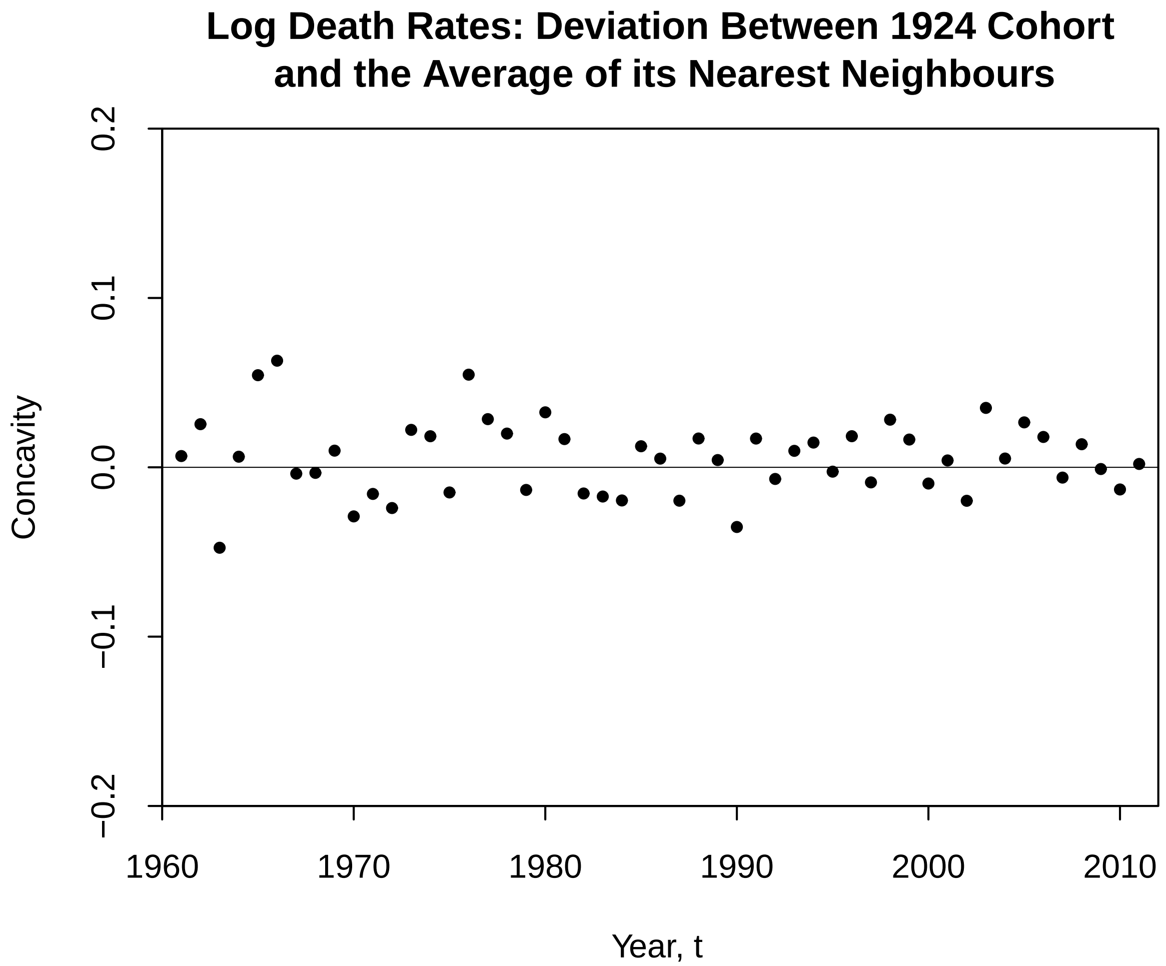
Hypothesis: *Underlying log death rates are approximately linear*

$\Rightarrow$ Plot concavity of log death rates: the difference between log of one death rate and the average of its immediate neighbours:

$$C(t, x_0)$$
$$= \log m(t, x_0 + t)$$
$$-\frac{1}{2}\Big(\log m(t, x_0 + t - 1) + \log m(t, x_0 + t + 1)\Big)$$

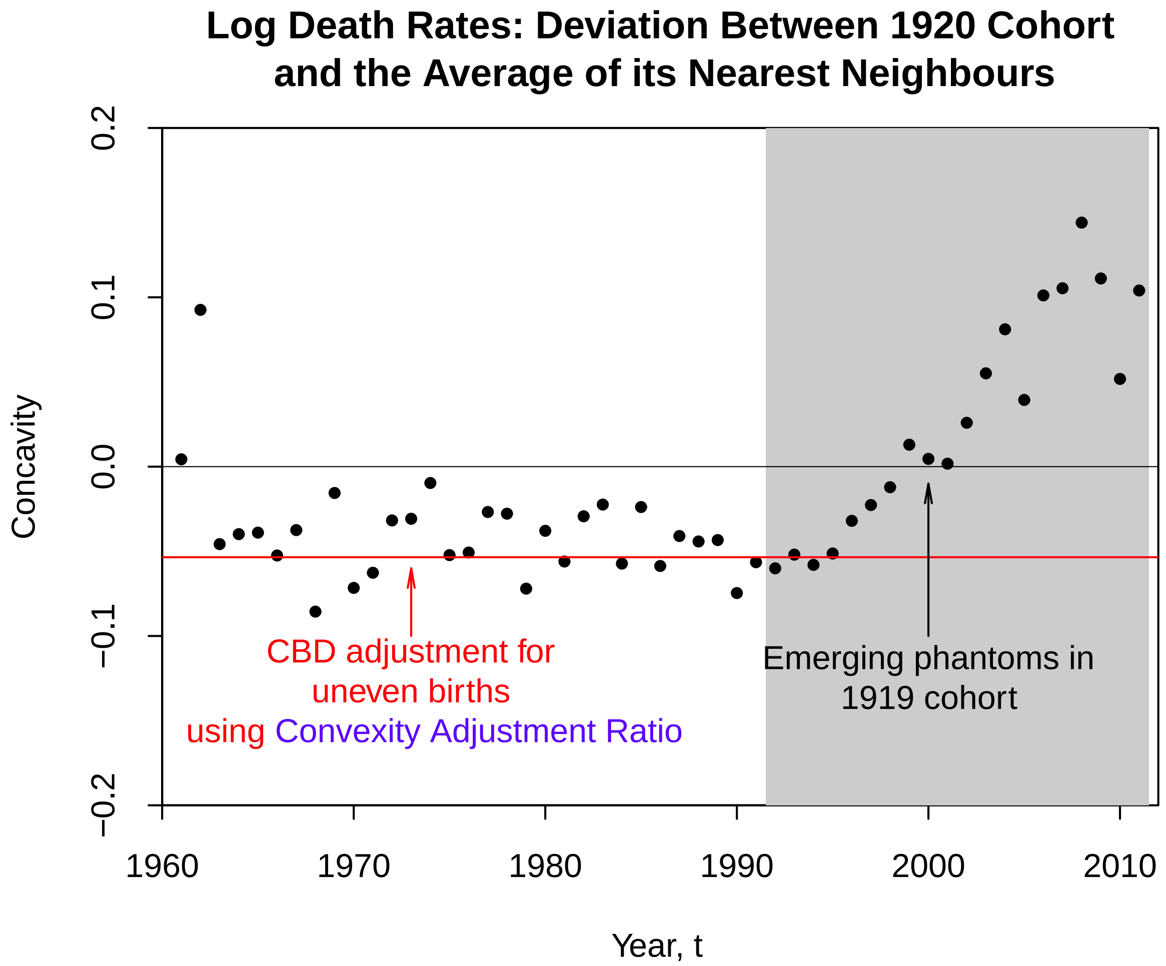
If log death rates are linear then this should be close to 0.

Concavity function: 1924 Cohort (age 37-87)



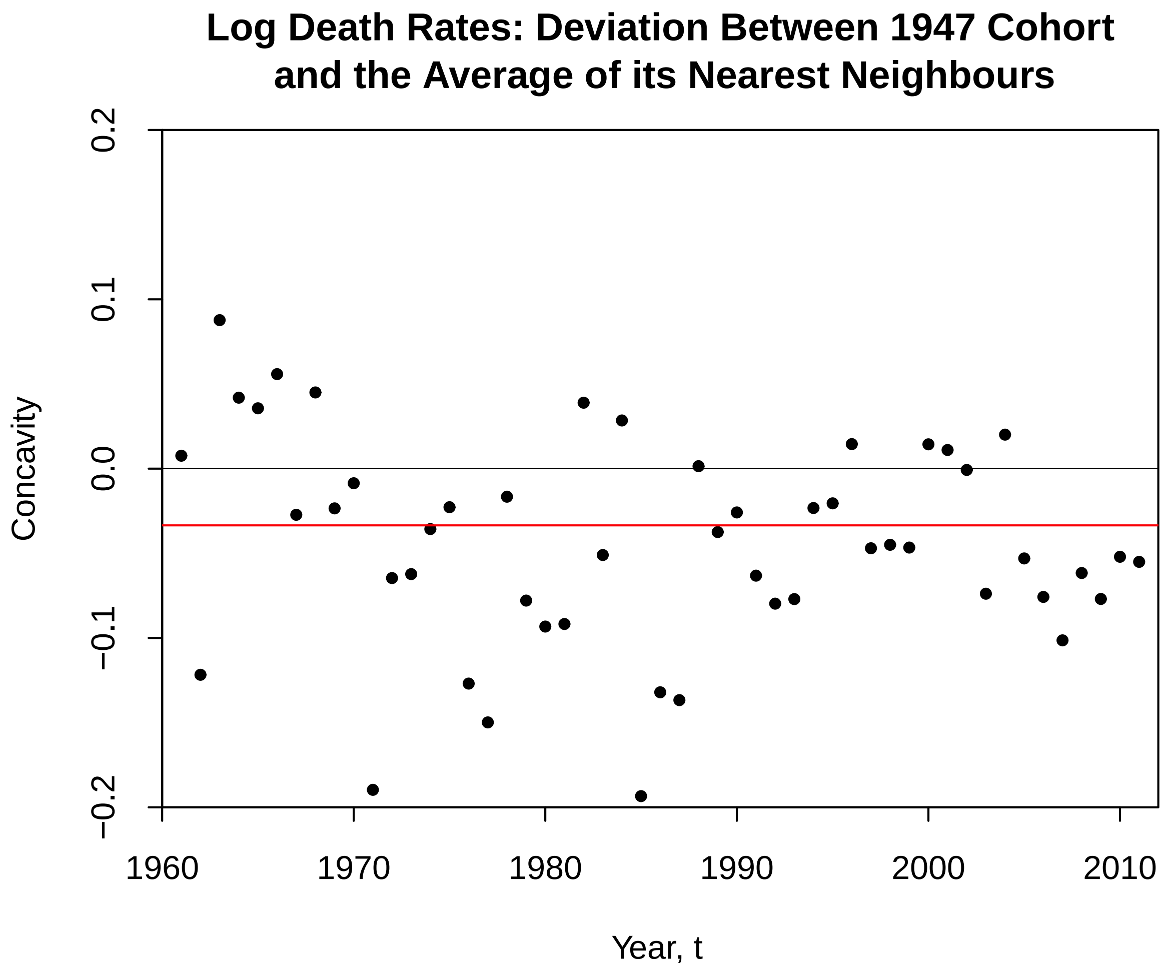
Dots are randomly above and below 0.

Concavity function: 1920 Cohort



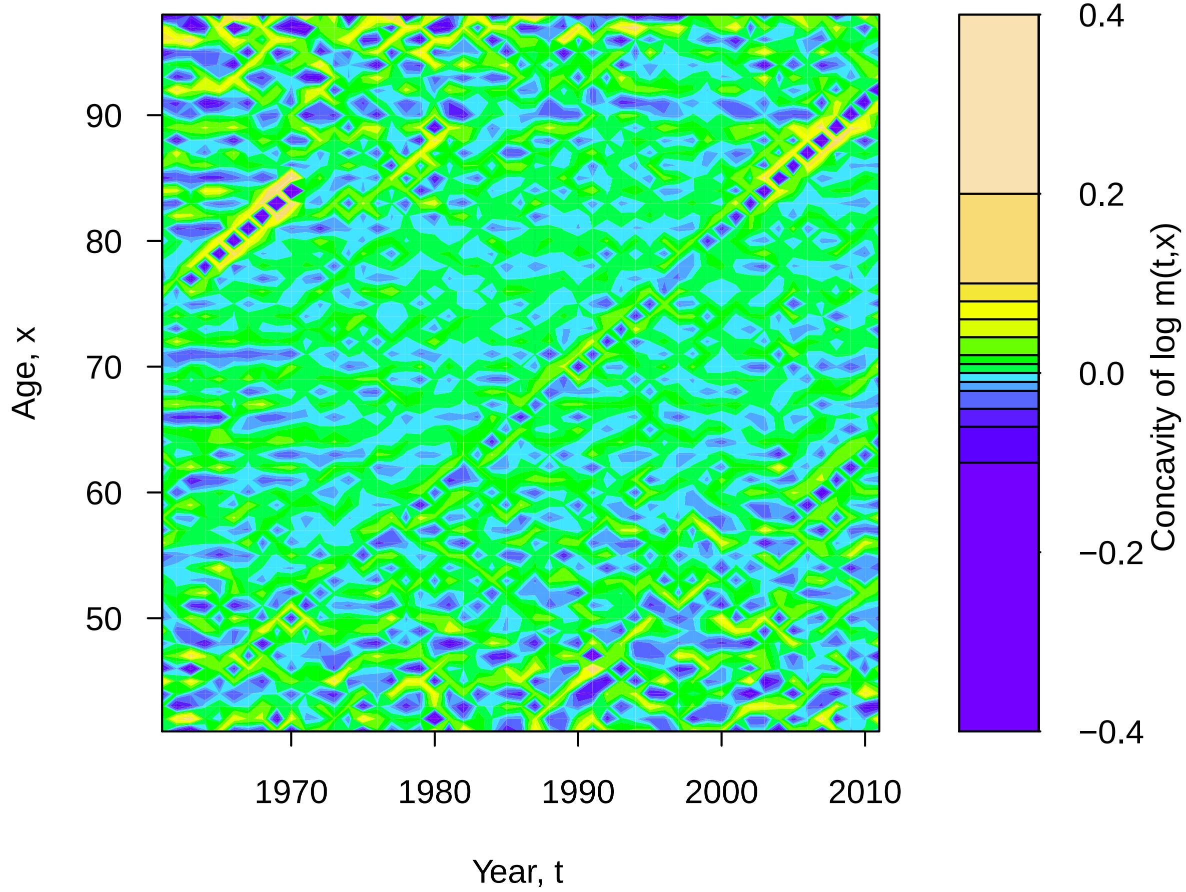
Signature plot: births pattern $\Rightarrow$ true $E(t,x) < P(t+\frac{1}{2},x)$

Concavity function: 1947 Cohort



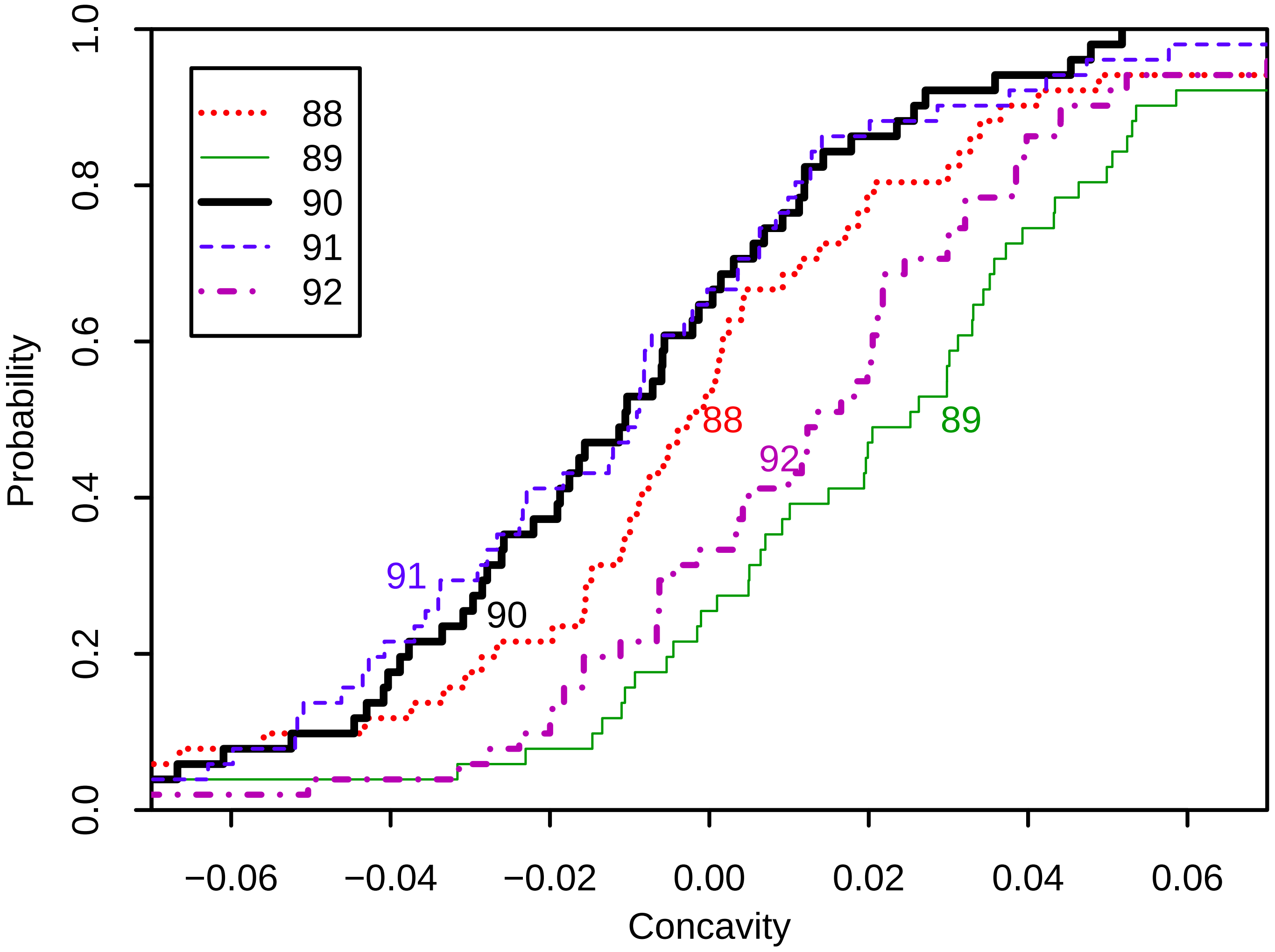
Dosts mostly below $0 \Rightarrow$ cause for concern

Concavity function in 2-Dimensions: Heat Map



Sampling variation $\Rightarrow$ more extremes < 50 and > 90

Concavity Function: Empirical CDF's by Age; 88-92



Heat Map: by Age and Calendar Year

Identifiable non-random patterns

Signatures:

- Diagonals $\Rightarrow$ issues with a cohort
- Horizontals $\Rightarrow$ anomalies in reported age at death ???
- Age at death errors are more plausible than systematic age-dependent errors in exposures.
- Except: Prominent horizontal anomaly around 89/90

3.3: Graphical Diagnostic 3

Hypothesis: *Changes in cohort population sizes should match pattern of reported deaths*

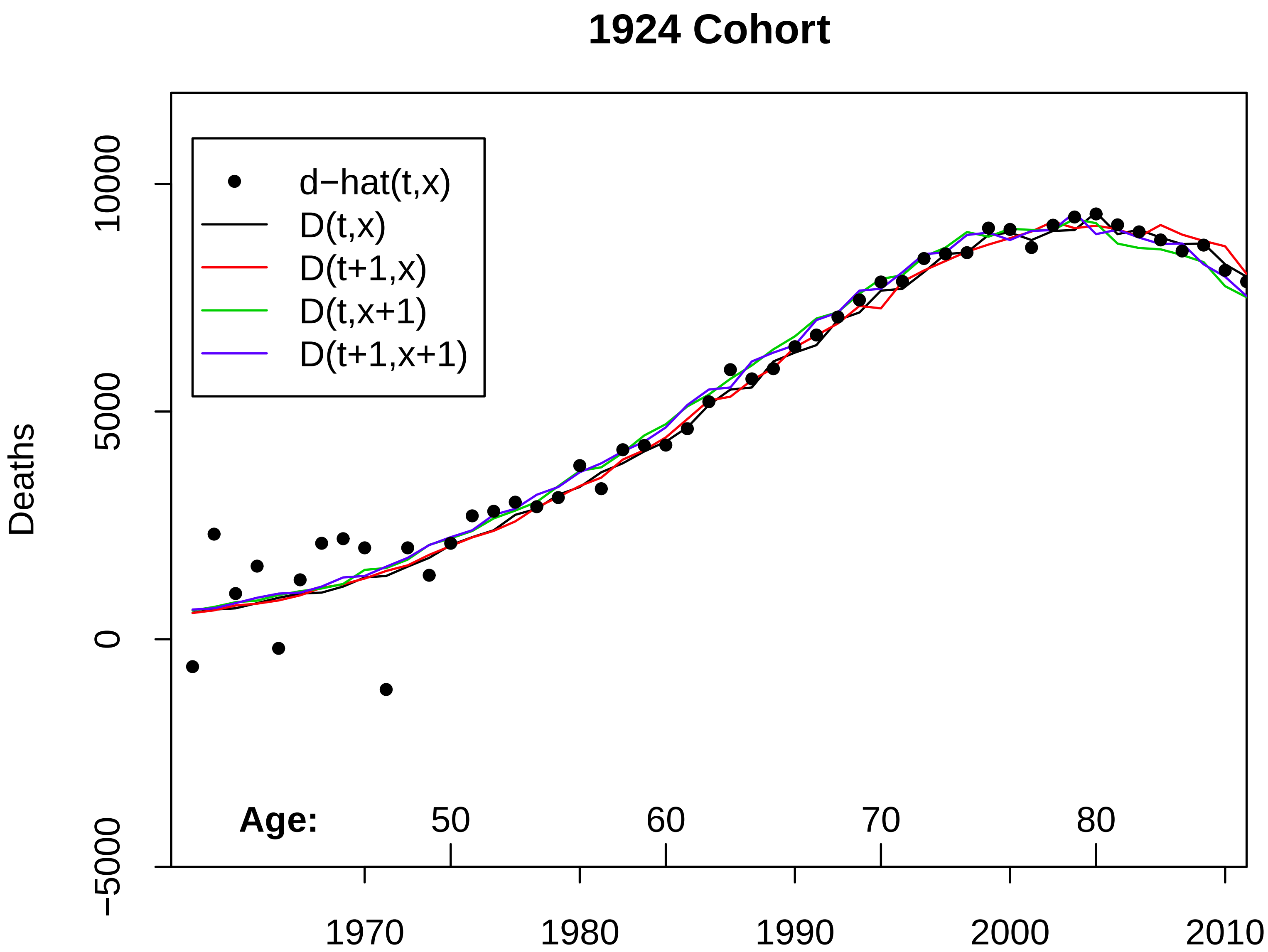
- Underlying data:
  - mid-year population, $P(t+\frac{1}{2},x)$
  - deaths in one calendar year, $D(t,x)$
- Define $\hat{d}(t+\frac{1}{2},x) = P(t+\frac{1}{2},x) - P(t+\frac{3}{2},x+1)$
- Plot $\hat{d}(t+\frac{1}{2},x)$ by cohort
- Compare with surrounding $D(t,x)$
- $\hat{d}$ and D should be similar if little or no net migration (e.g. high ages)

Prior adjustments

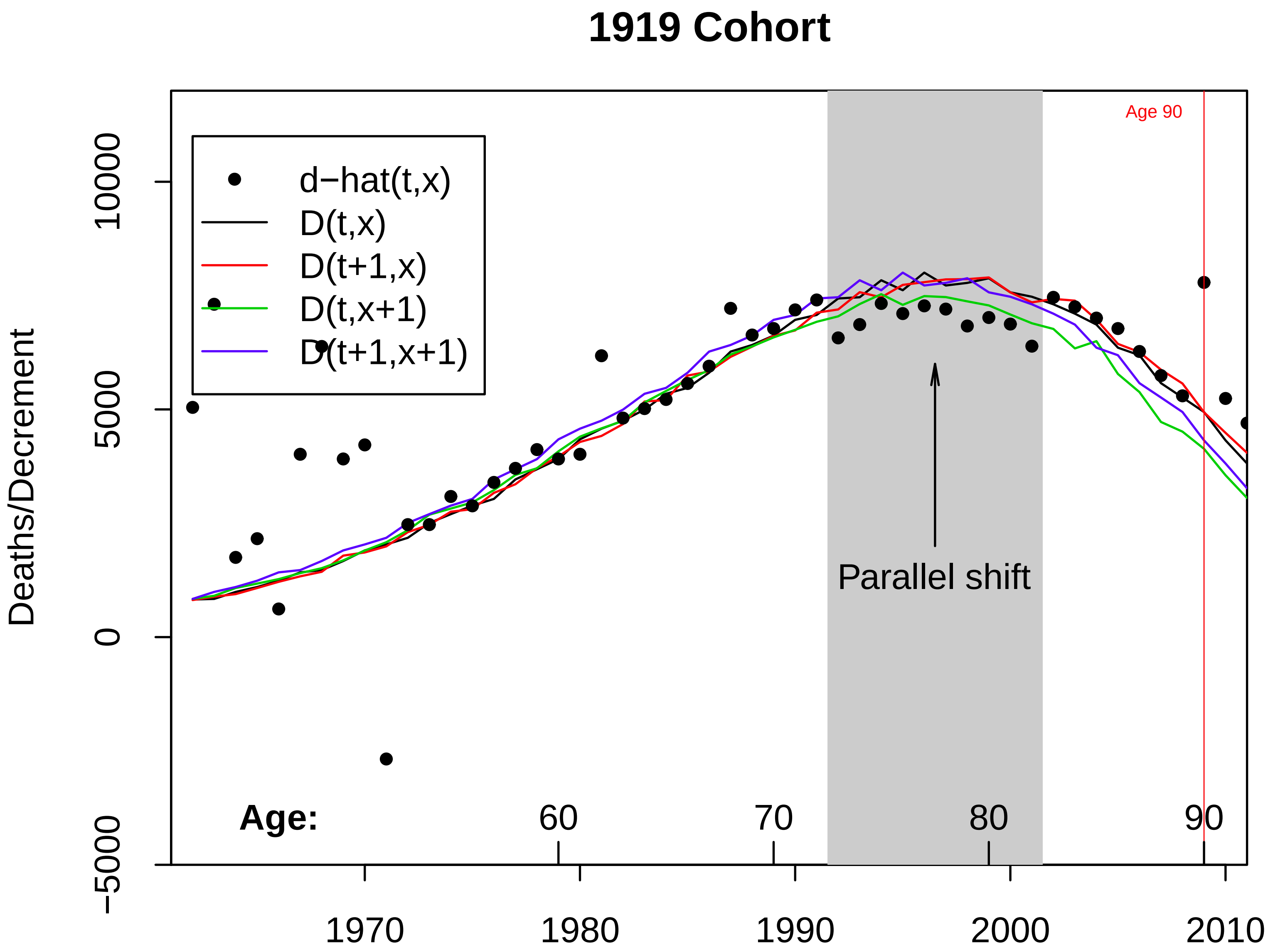
- Decrements: adjust for $E(t,x) \neq P(t+\frac{1}{2},x)$
  $\Rightarrow \hat{d}(t+\frac{1}{2},x)$ multiplied by $CAR(t-x)$

- Cohorts ± 1 year: adjust for different birth rates
  $D(t,x+1) \times E(t-x,0)/E(t-x-1,0)$
  $D(t+1,x) \times E(t-x,0)/E(t-x+1,0)$

Standard Graphical Diagnostic 3: 1924 Cohort, Deaths Curve

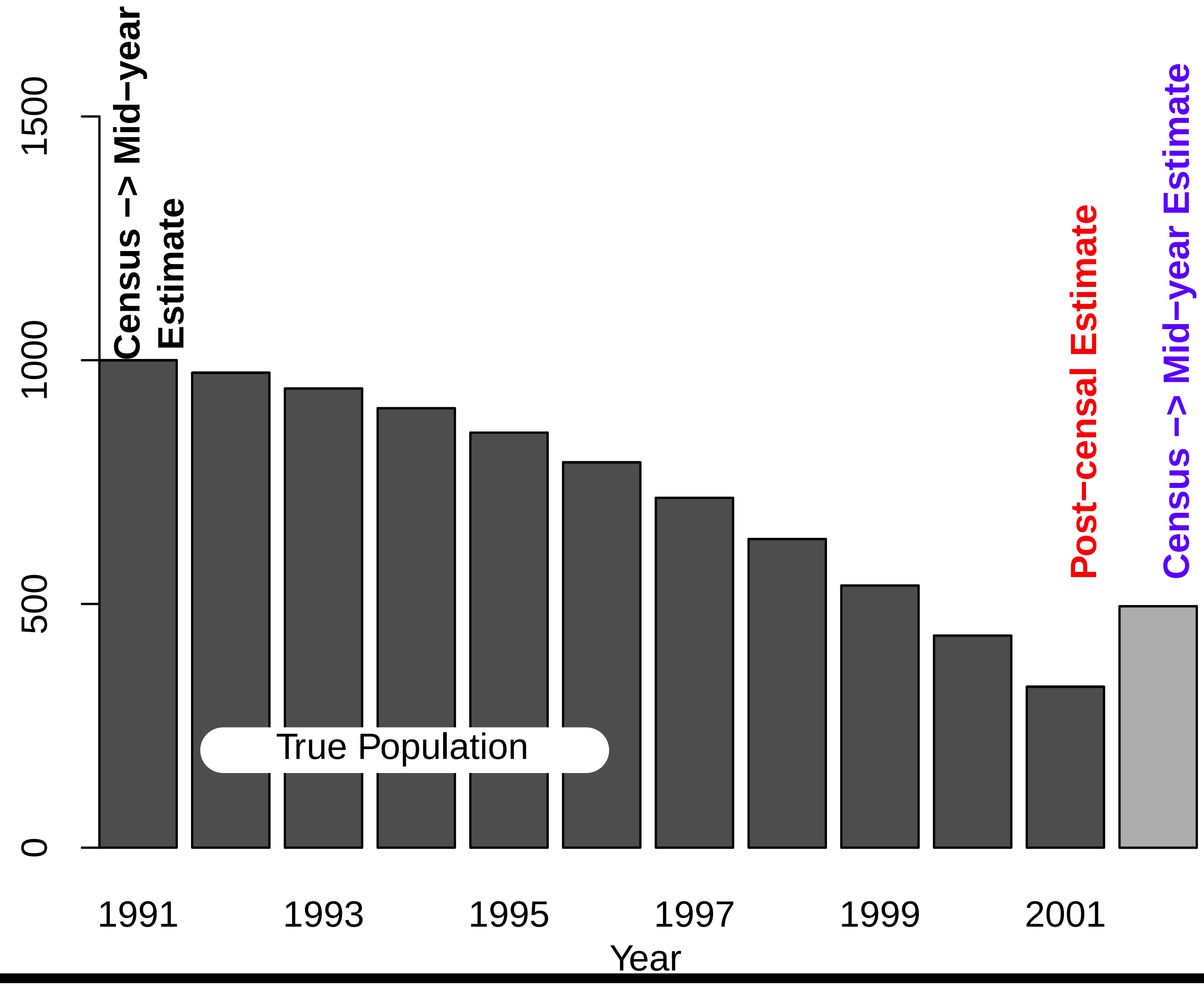


Signature Plot: Backfilling the 1919 Cohort by ONS



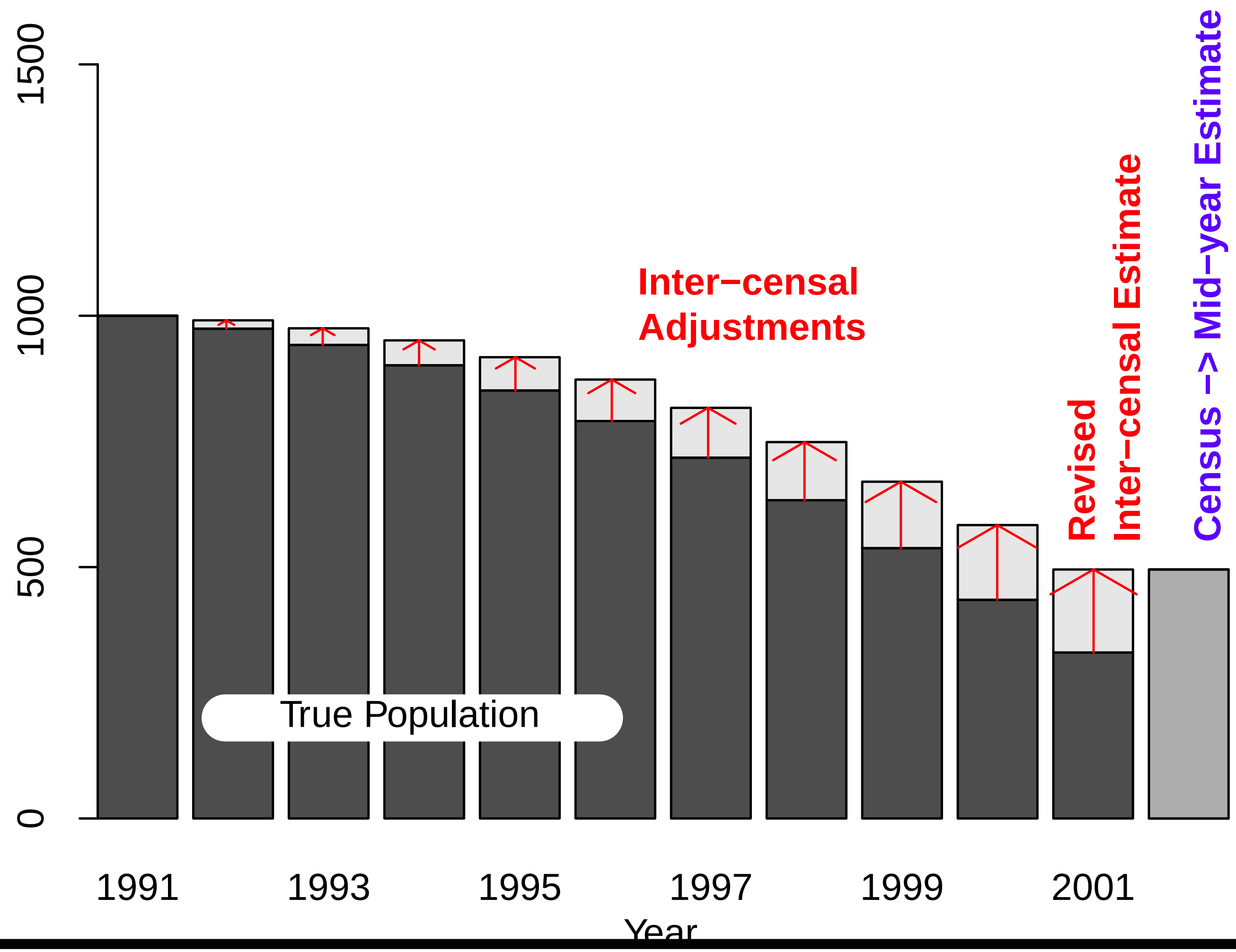
Possible Explanation: Census → Mid-year Pop Error

1919 cohort (stylized)

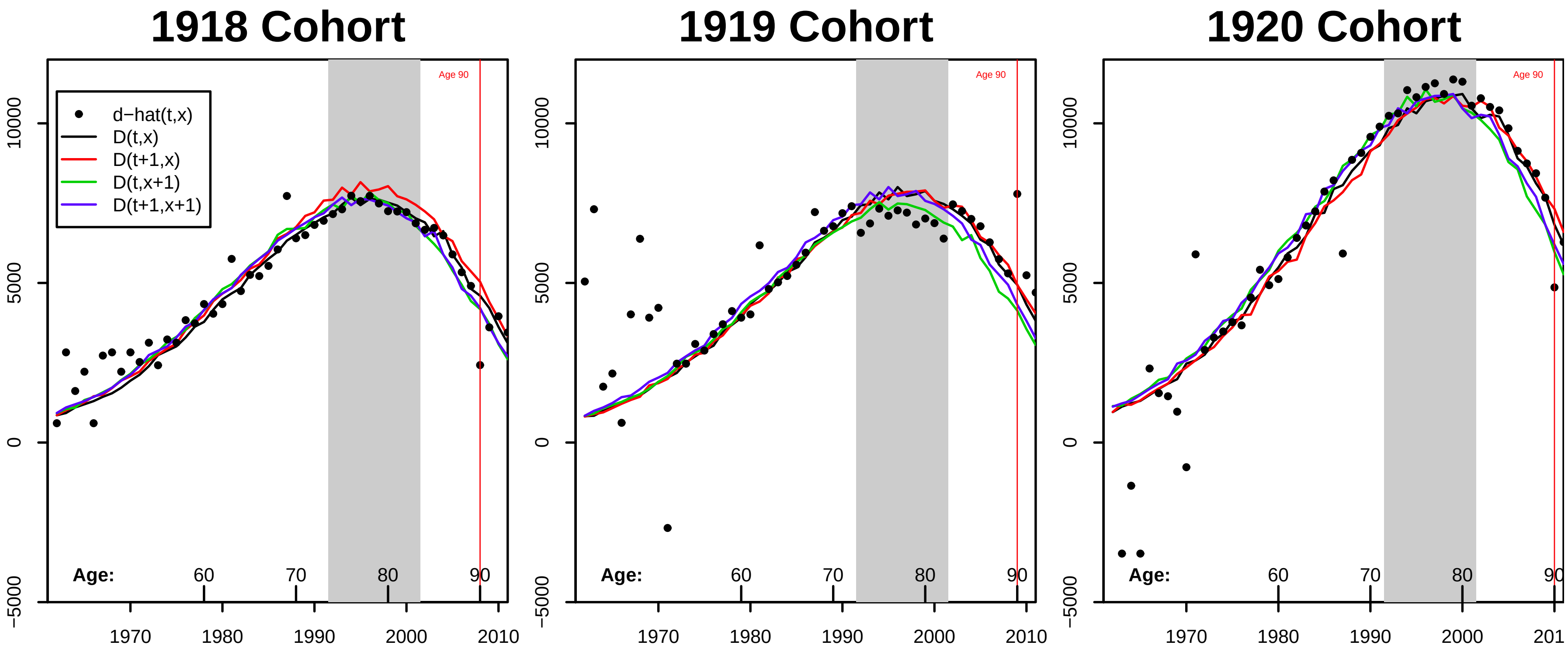


Factual Consquence: Backfilling (ONS Methodology)

1919 cohort (stylized)

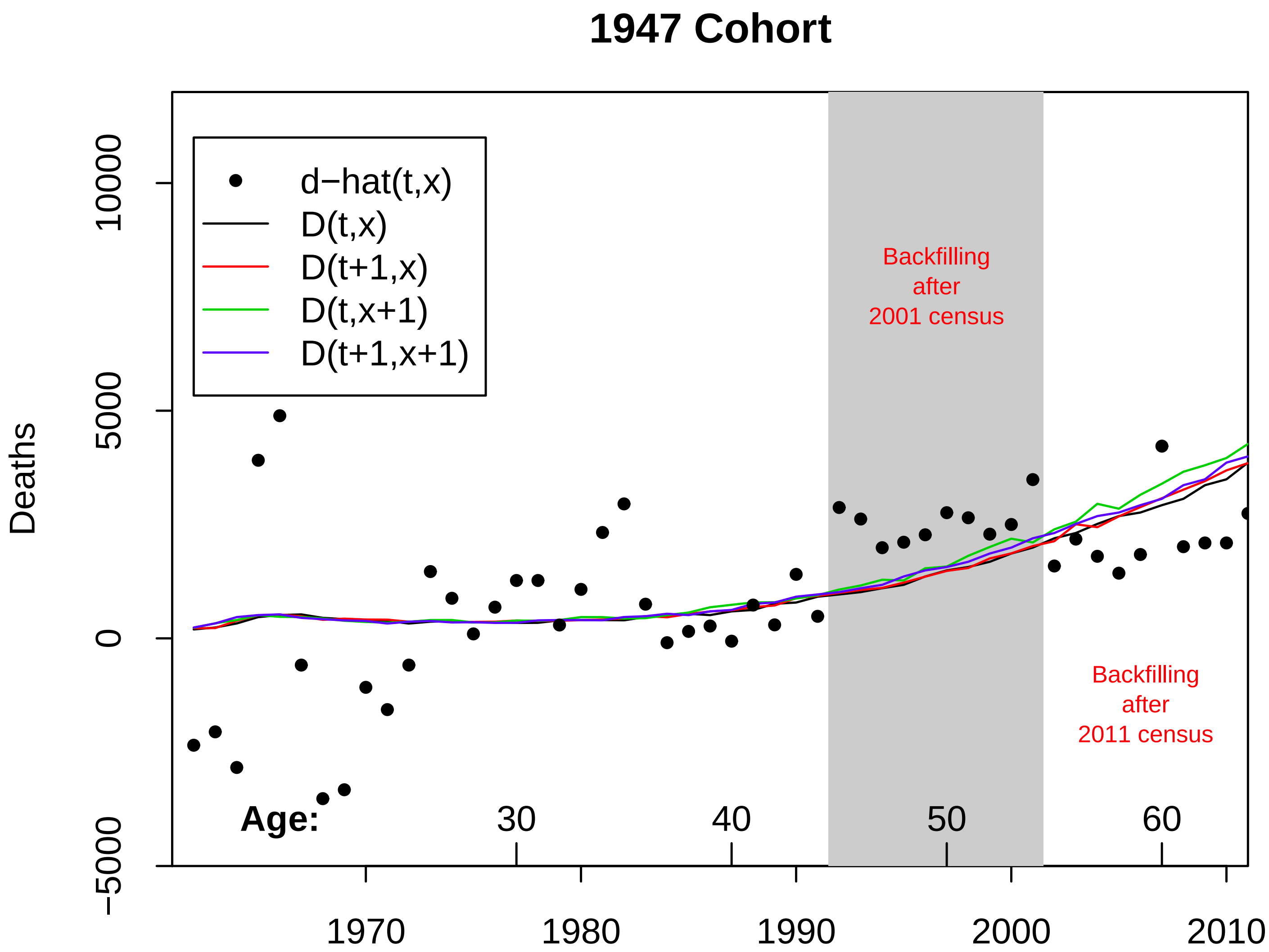


1918, 1919 and 1920 Cohorts, Deaths Curves



- 1920 cohort: similar shift in opposite direction
- Age 90 anomaly for all 3 cohorts ⇒ cause for concern

Signature Plot: Backfilling the 1947 Cohort



Again consistent with ONS versus CBD methodologies

3.4: Summary

- Errors remain in the ONS population data
- Combination of three graphical diagnostics highlight known anomalies (e.g.1919) and some unexpected discoveries (e.g. 1920, 1947 cohorts; age 89/90)
- Anomalies characterised by cohort and by age
- CBD Exposures Methodology can be used to improve estimates of exposures
- CBD Exposures Methodology explains the 1919 anomaly that has emerged since 1991

4: Model-Based Analysis of Historical Population Data

4.1: Proposed Solution: Bayesian Adjustment of Exposures

Bayesian prior hypotheses:

A: Death counts are accurate

B: Exposures are subject to errors

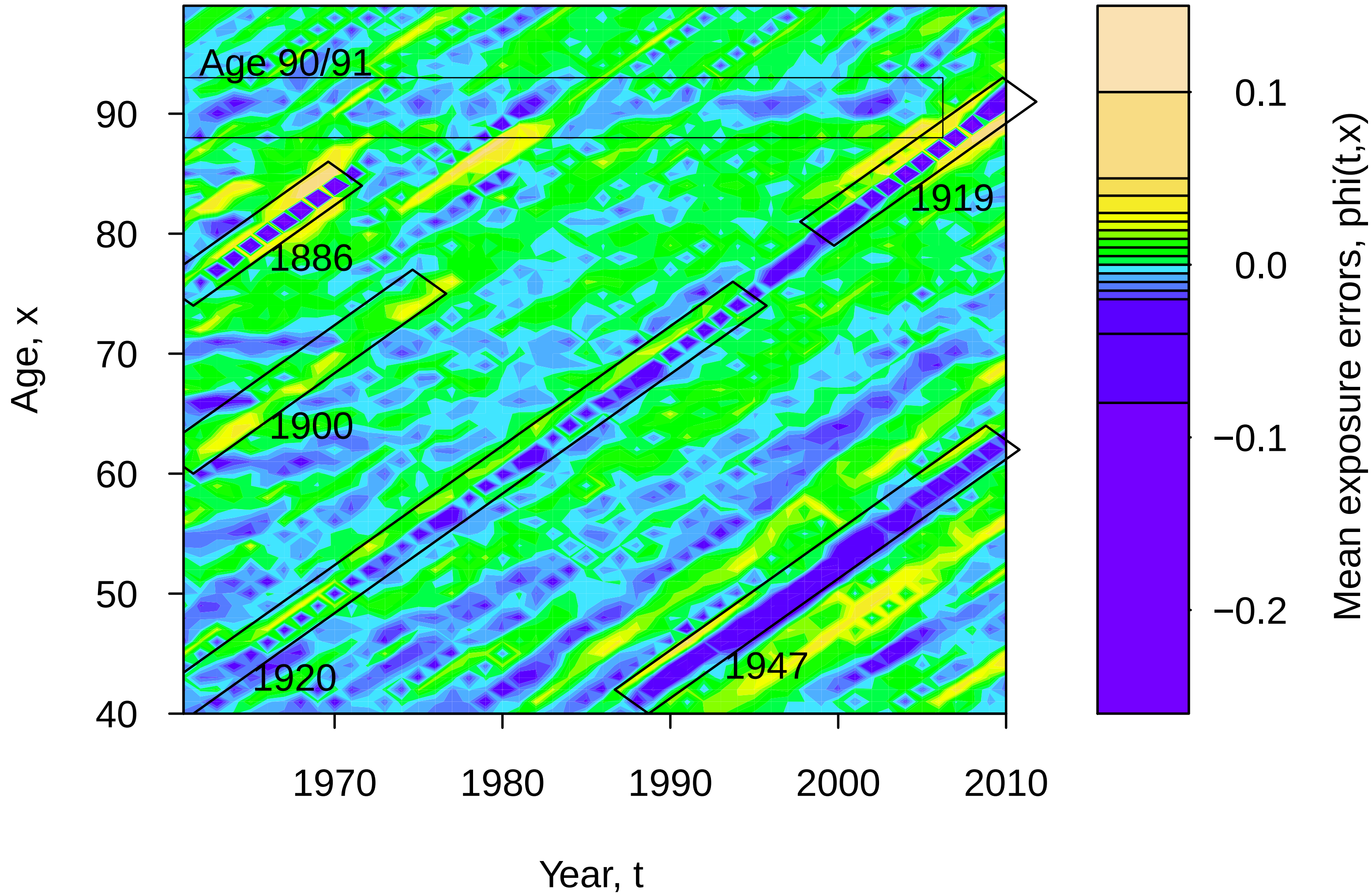
- errors following cohorts are correlated through time

C: Within each calendar year:

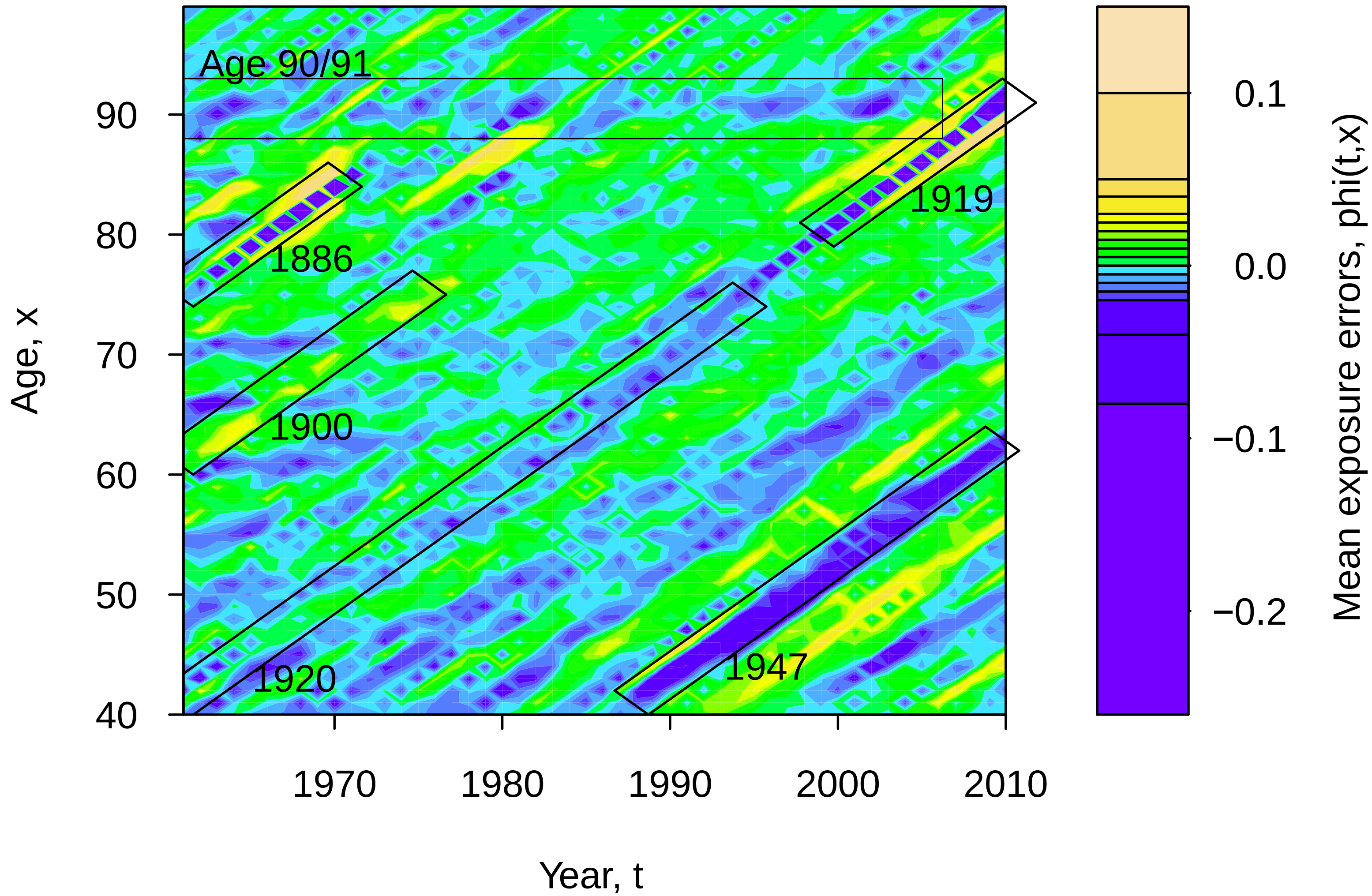
- curve of underlying death rates is “smooth”

Adjust exposures to achieve a balance between B and C

4.2: Results: Assume $E(t,x) = P(t+\frac{1}{2},x)$ Mid-year Population



Exposures, $E(t,x)$, Adjusted Using CBD Convexity Adjustment Ratio



$E(t,x) = P(t+\frac{1}{2},x) \times$ Convexity Adjustment Ratio

4.3: Results 1

- Results confirm conclusions based on graphical diagnostics (e.g. problems with 1919, 1947 cohorts; age 89/90 boundary)
- Bayesian approach allows us to *quantify rigorously* the size of the error

Results 2

- CBD Exposures Methodology:
  - convexity adjustment for $E(t,x) \neq P(t+\frac{1}{2},x)$ explains 1920 anomaly
  - CBD dampens other anomalies (e.g. 1947 cohort)
- Other anomalies remain but we have some explanations
  - 1919 cohort explained by 2001 census $+$ backfilling
  - age 89/90 $\Rightarrow$ issues with Kannisto-Thatcher methodology
  - e.g. ages 70, 80 $\Rightarrow$ potential bias in reporting of age at death
- 1947 (1940-1960) cohort(s) should be seen as an issue financially

6: Conclusions and Next Steps

- Significant errors remain in EW males data
- Similar issues with females data
- Errors will exist in data for many other countries
- CBD Exposures Methodology can be used to mitigate errors in exposures
  - census-to-mid-year adjustment
  - mid-year population to exposures: CAR
  - Use exact date of birth in the census questionnaire!
- Kannisto-Thatcher high age methodology needs revisiting
- Financial impact: post WW-2 cohorts need special consideration

Thank you!

Questions?

Paper online:

http://www.macs.hw.ac.uk/∼andrewc/papers/ajgc71.pdf

Bonus slides

Impact of Population Revisions on Mortality Rates

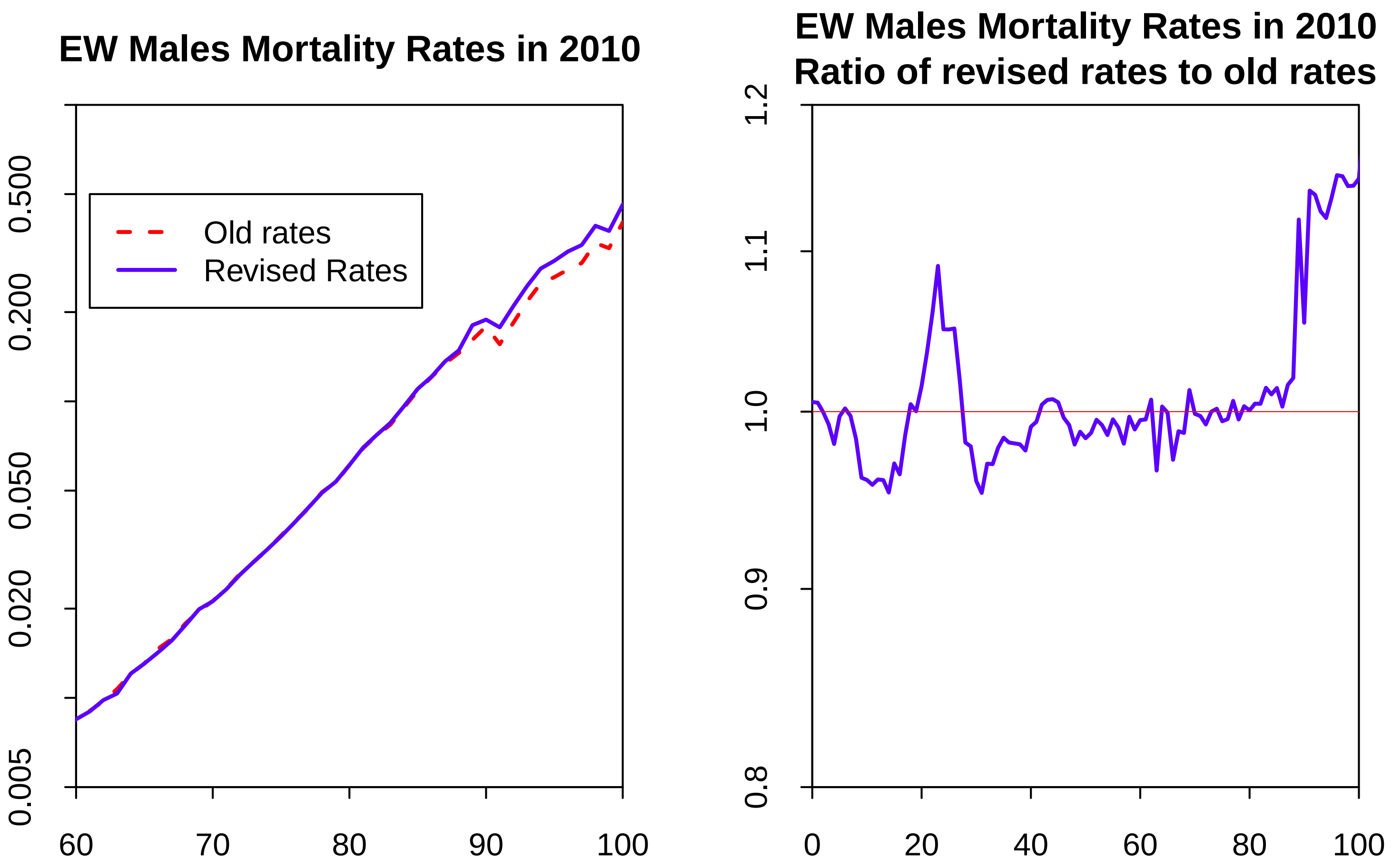
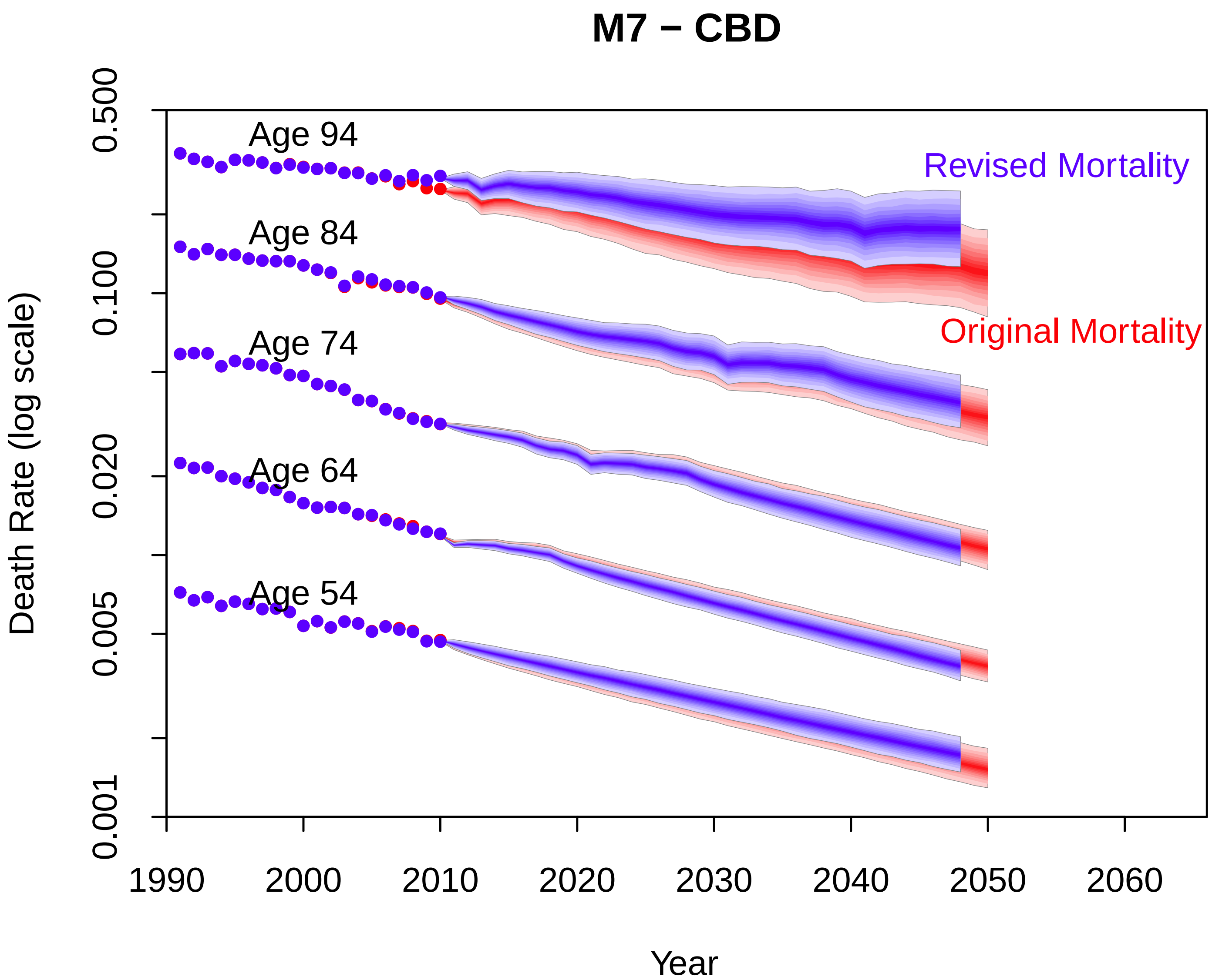


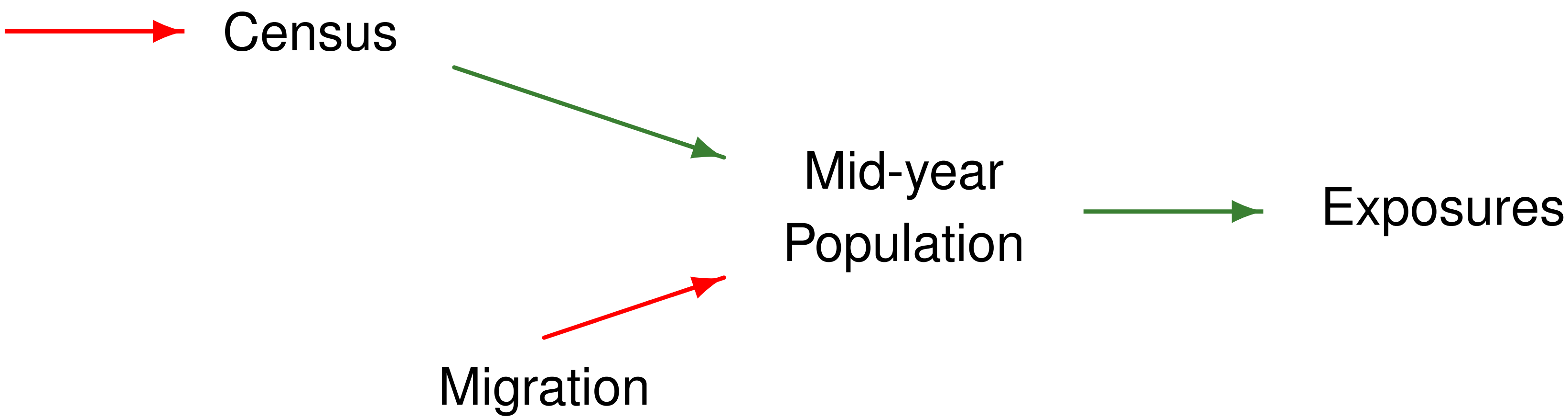
Figure 1:

Types of Impact: Base Table; Central Trend; Future Uncertainty



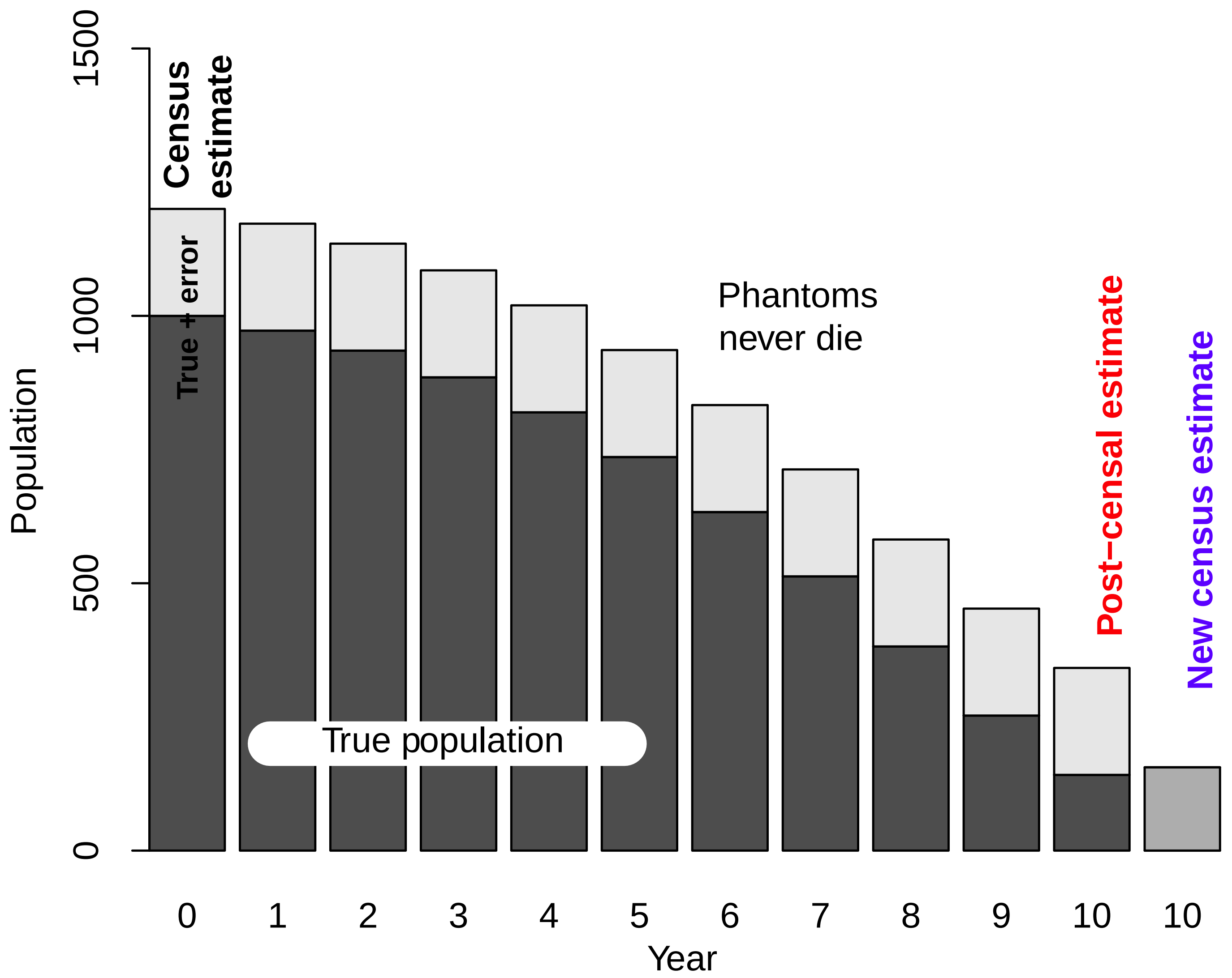
Where Can Errors in $E(t, x)$ Occur?

Four sources of error:

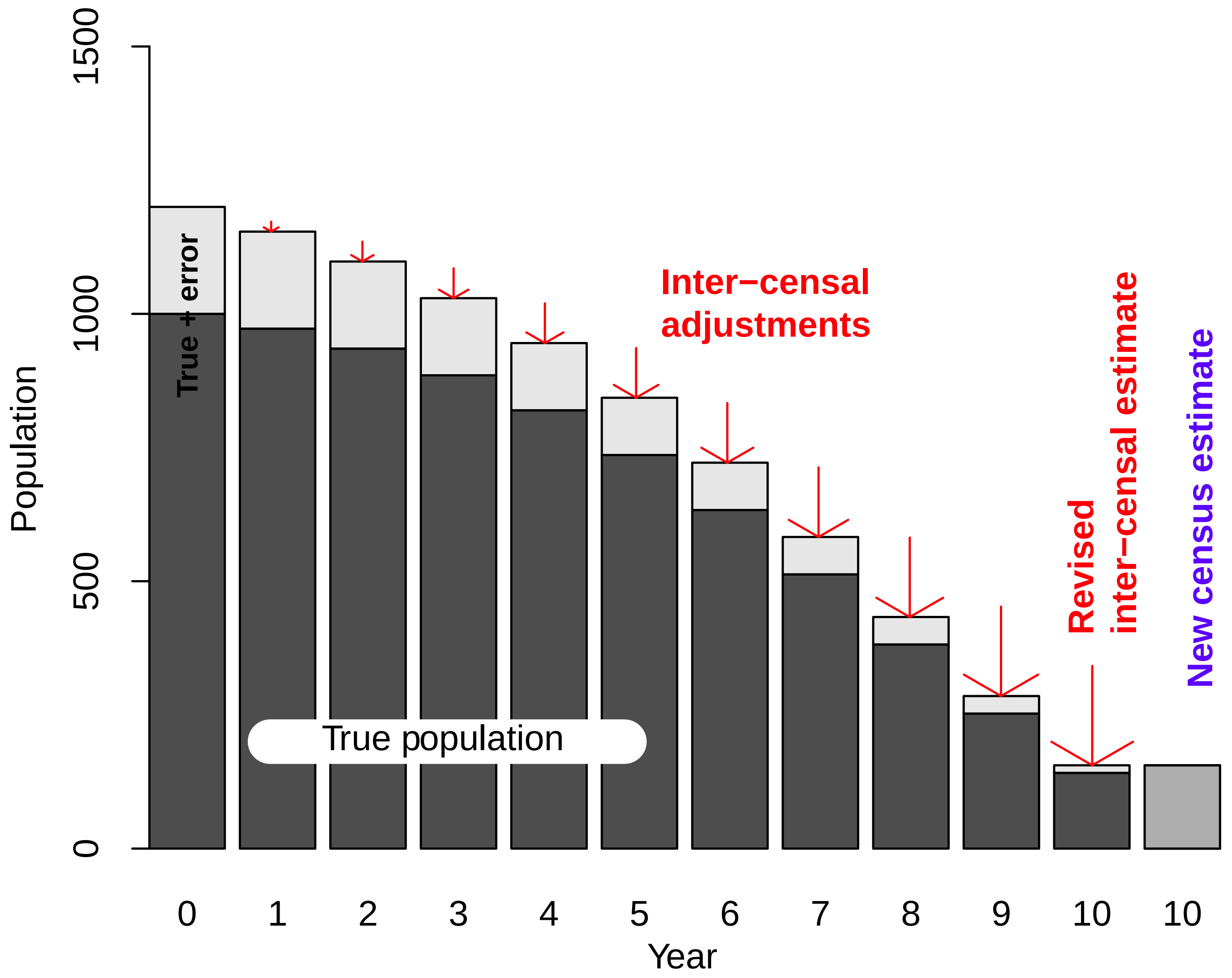


Errors that can be mitigated using CBD Exposures Methodology

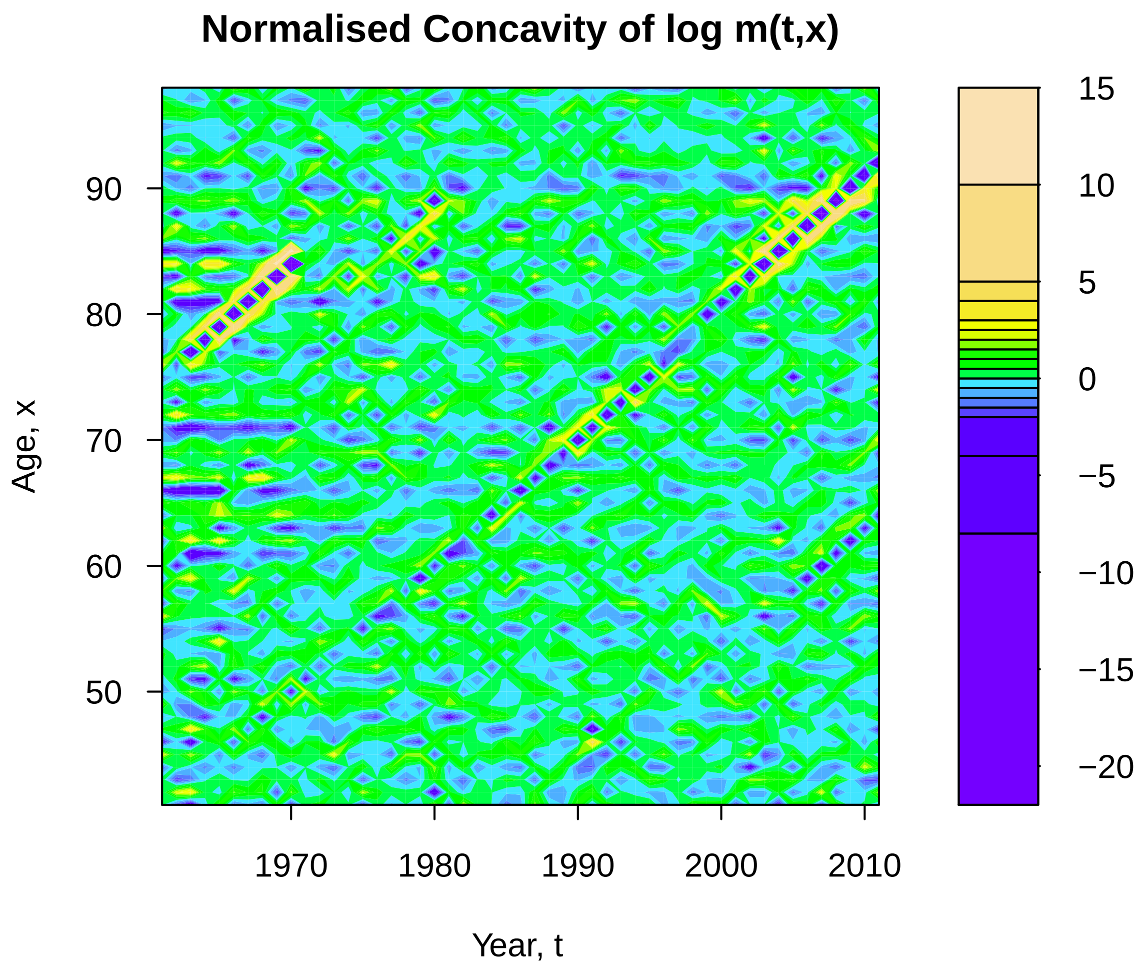
Phantoms Never Die



Factual Consquence: Backfilling (ONS Methodology)



Same Data in 2-Dimensions: Heat Map – Normalised



Sampling variation $\Rightarrow$ more extremes < 50 and > 90

Why Use a Bayesian Approach

- Coherent framework within which we can
  - build in prior beliefs (hypotheses A, B, C)
- Output $\Rightarrow$ straightforward to assess impact of parameter uncertainty