Development of Trend Analysis Guidelines at NCHS

Donald Malec
Associate Director of Science
Office of Research and Methodology
and
Spokesperson, NCHS Trends Analysis Workgroup

Presentation for the Board of Scientific Counselors
September 25, 2015

CDC

Guidance being developed by NCHS workgroup

- ► Members include analysts across NCHS
- ► Draft guidance document in two main parts: Guidance and Real Examples
  1. Overview of Key Trend Analysis Issues
  2. Choosing a Method for Trend Analysis
  3. Illustrative Examples of Trend Analysis with Alternative Comparisons
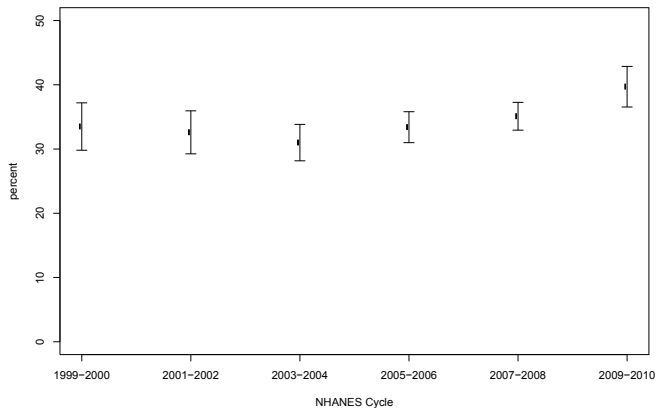 4. Technical Appendixes... with explanation of statistical guidance

An example of a trend

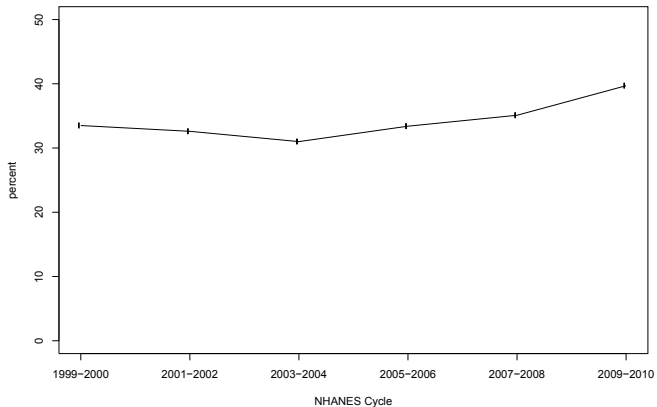Figure : Prevalence of Obesity among adults age 60 years and older, NHANES 1999-2010.
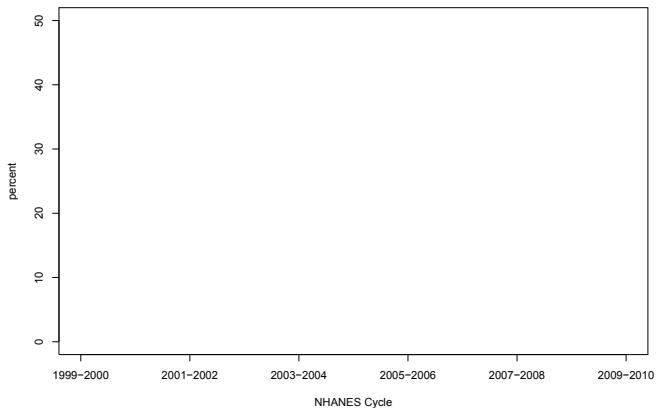
An example of a trend

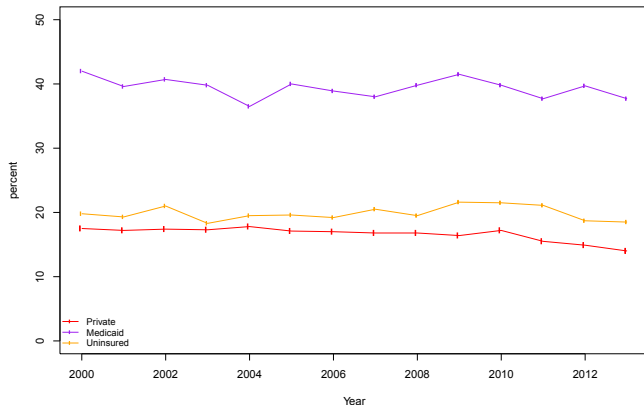Figure : Prevalence of Obesity among adults age 60 years and older, NHANES 1999-2010.

An example of a trend

Figure : Prevalence of Obesity among adults age 60 years and older, NHANES 1999-2010.

other examples of trends ...

Figure : Percent of persons with any emergency room use in the past 12 months among adults aged 18-64, by insurance status. NHIS 2000-2013.
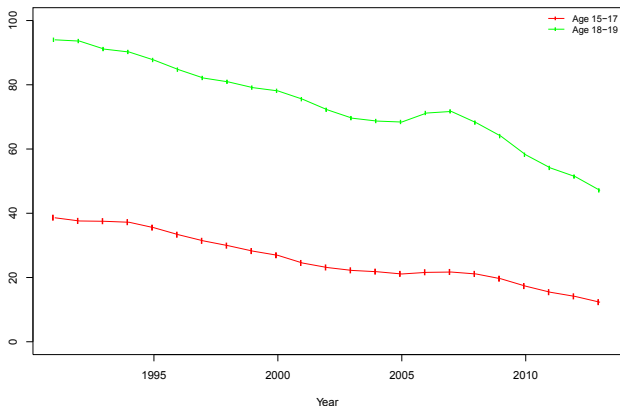
other examples of trends ...

Figure : Percentage of ED visits in which an EKG was ordered or provided: United States. NHAMCS 2005-2011.
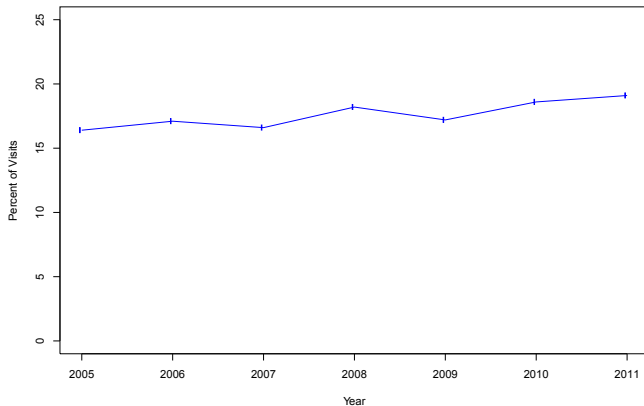
other examples of trends ...

Figure : Birth rates for women aged 15-17 and 18-19: United States, 1991-2013.

Aim of a Guidance Document for Trend Analysis

- **Focus on descriptive statistics, not prediction**
- **Document issues that may cause controversy - choice of study interval, choice of model, etc.**
- **Provide a summary of methods for reference**
- **When possible, suggest a preferred method**

Aim of a Guidance Document for Trend Analysis

- **Focus on descriptive statistics, not prediction**
- **Document issues that may cause controversy - choice of study interval, choice of model, etc.**
- **Provide a summary of methods for reference**
- **When possible, suggest a preferred method**

- **Recognize limitations of rules**
  - **Analyst will bring expertise into the analysis**
  - **Include "just enough detail"**

Why are trends analyzed at NCHS? .... here are a few

- **Topical ... trends in obesity prevalence**

- **Program changes ... trends in health insurance coverage**

- **Surveillance/ resource ... Health US and Healthy People**

- **Interesting changes (e.g. changes in slope) noted by analysts**

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**

- **Comparison to a benchmark year**

- **Linear (with curves) and non-linear regression**

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
  - least squares fit to population prevalence

- **Comparison to a benchmark year**

- **Linear (with curves) and non-linear regression**

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
  - least squares fit to population prevalence
  - population weighted least squares fit to population prevalence

- **Comparison to a benchmark year**

- **Linear (with curves) and non-linear regression**

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
    - least squares fit to population prevalence
    - population weighted least squares fit to population prevalence
    - sample variance weighted least squares fit to population prevalence
- **Comparison to a benchmark year**

- **Linear (with curves) and non-linear regression**

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
  - least squares fit to population prevalence
  - population weighted least squares fit to population prevalence
  - sample variance weighted least squares fit to population prevalence
- **Comparison to a benchmark year**
  - In 2013, 17.8% of noninstitutionalized adults aged 18 and over were current cigarette smokers, a decline from 2000 (2014 - HUS)
- **Linear (with curves) and non-linear regression**

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
  - least squares fit to population prevalence
  - population weighted least squares fit to population prevalence
  - sample variance weighted least squares fit to population prevalence
- **Comparison to a benchmark year**
  - In 2013, 17.8% of noninstitutionalized adults aged 18 and over were current cigarette smokers, a decline from 2000 (2014 - HUS)
- **Linear (with curves) and non-linear regression**
  - polynomial regression ... orthogonal polynomials

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
  - least squares fit to population prevalence
  - population weighted least squares fit to population prevalence
  - sample variance weighted least squares fit to population prevalence

- **Comparison to a benchmark year**
  - In 2013, 17.8% of noninstitutionalized adults aged 18 and over were current cigarette smokers, a decline from 2000 (2014 - HUS)

- **Linear (with curves) and non-linear regression**
  - polynomial regression ... orthogonal polynomials
  - linear splines ... joinpoint software (NCI)

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
  - least squares fit to population prevalence
  - population weighted least squares fit to population prevalence
  - sample variance weighted least squares fit to population prevalence
- **Comparison to a benchmark year**
  - In 2013, 17.8% of noninstitutionalized adults aged 18 and over were current cigarette smokers, a decline from 2000 (2014 - HUS)
- **Linear (with curves) and non-linear regression**
  - polynomial regression ... orthogonal polynomials
  - linear splines ... joinpoint software (NCI)
  - log transform, logistic transform of prevalence rates

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
  - least squares fit to population prevalence
  - population weighted least squares fit to population prevalence
  - sample variance weighted least squares fit to population prevalence
- **Comparison to a benchmark year**
  - In 2013, 17.8% of noninstitutionalized adults aged 18 and over were current cigarette smokers, a decline from 2000 (2014 - HUS)
- **Linear (with curves) and non-linear regression**
  - polynomial regression ... orthogonal polynomials
  - linear splines ... joinpoint software (NCI)
  - log transform, logistic transform of prevalence rates
 - smoothing by collapsing years, followed by additional analysis

First step: How are trends analyzed at NCHS now?

- **Linear regression (straight line)**
    - least squares fit to population prevalence
    - population weighted least squares fit to population prevalence
    - sample variance weighted least squares fit to population prevalence
- **Comparison to a benchmark year**
    - In 2013, 17.8% of noninstitutionalized adults aged 18 and over were current cigarette smokers, a decline from 2000 (2014 - HUS)
- **Linear (with curves) and non-linear regression**
    - polynomial regression ... orthogonal polynomials
    - linear splines ... joinpoint software (NCI)
    - log transform, logistic transform of prevalence rates
    - smoothing by collapsing years, followed by additional analysis
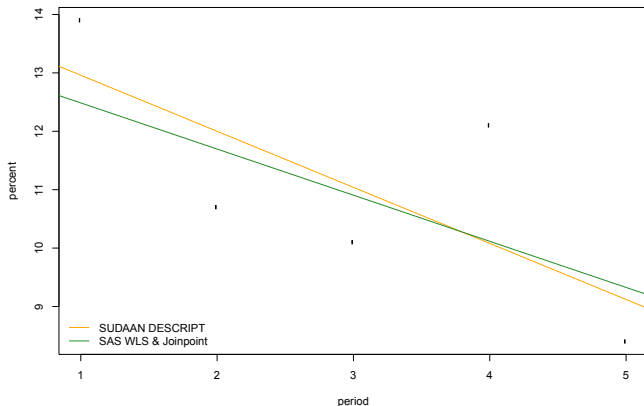   - pairwise comparisons

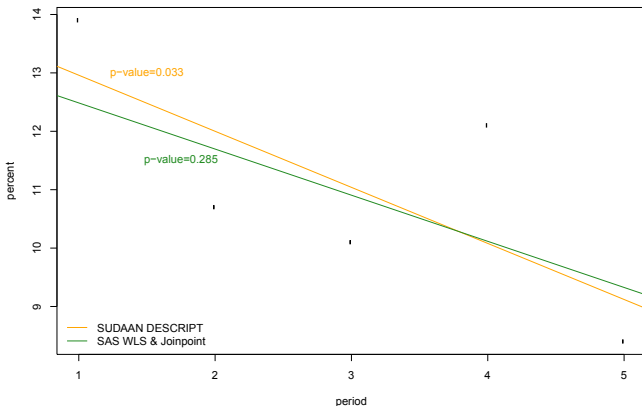# Methodological Issues: Design vs Model-based Approach
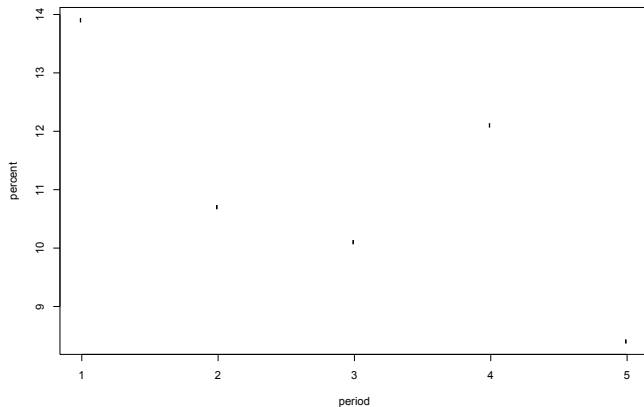
Methodological Issues: Design vs Model-based Approach

▶ Trend estimates are similar but significance can vary widely

Methodological Issues: Design vs Model-based Approach

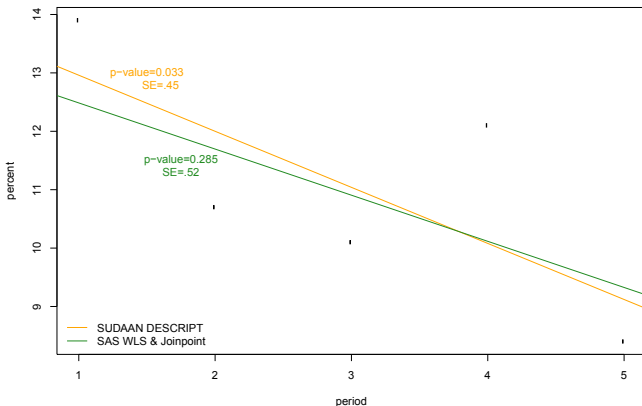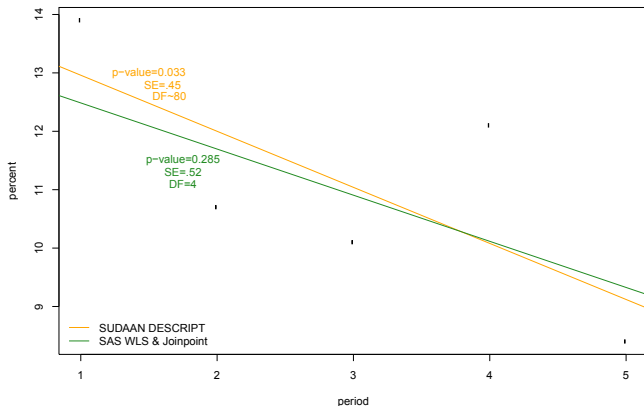▶ Trend estimates are similar but significance can vary widely

Methodological Issues: Design vs Model-based Approach

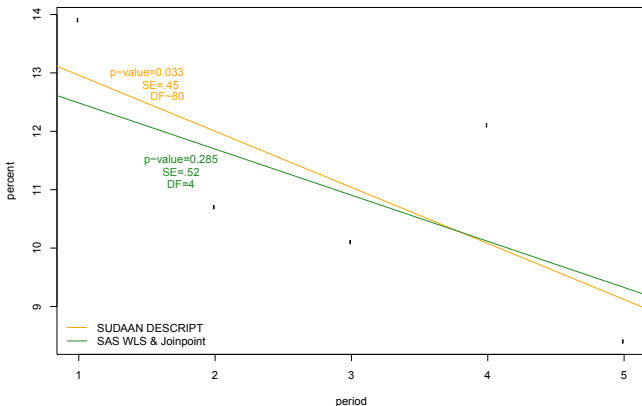- ▶ Trend estimates are similar but significance can vary widely

Methodological Issues: Design vs Model-based Approach

▶ Trend estimates are similar but significance can vary widely

Methodological Issues: Design vs Model-based Approach

- ▶ Trend estimates are similar but significance can vary widely

- ▶ Model-based includes "lack of model fit" as error.
  Design-based provides variance of each point estimate.

Methodological Issues: What Slope to Estimate?

- Not the usual issue of <u>how</u> to estimate (i.e. efficiency) but, rather, <u>what</u> to estimate
- When fitting a line to population values, NCHS uses (at least) three different targets:

Methodological Issues: What Slope to Estimate?

- ▶ Not the usual issue of <u>how</u> to estimate (i.e. efficiency) but, rather, <u>what</u> to estimate
- ▶ When fitting a line to population values, NCHS uses (at least) three different targets:
    - ▶ Unweighted Slope: slope $\propto \sum_{t=1}^{T} P_t \times year_t$

Methodological Issues: What Slope to Estimate?

- ▶ Not the usual issue of <u>how</u> to estimate (i.e. efficiency) but, rather, <u>what</u> to estimate
- ▶ When fitting a line to population values, NCHS uses (at least) three different targets:
  - ▶ Unweighted Slope: slope $\propto \sum_{t=1}^{T} P_t \times year_t$
  - ▶ Population weighted Slope: slope $\propto \sum_{t=1}^{T} N_t P_t \times year_t$

Methodological Issues: What Slope to Estimate?

- ▶ Not the usual issue of <u>how</u> to estimate (i.e. efficiency) but, rather, <u>what</u> to estimate
- ▶ When fitting a line to population values, NCHS uses (at least) three different targets:
  - ▶ Unweighted Slope: slope $\propto \sum_{t=1}^{T} P_t \times year_t$
  - ▶ Population weighted Slope: slope $\propto \sum_{t=1}^{T} N_t P_t \times year_t$
  - ▶ Sample variance weighted Slope: slope $\propto \sum_{t=1}^{T} \frac{P_t \times year_t}{var(\hat{P}_t)}$

# Methodological Issues: What Slope to Estimate?

- ▶ Not the usual issue of <u>how</u> to estimate (i.e. efficiency) but, rather, <u>what</u> to estimate
- ▶ When fitting a line to population values, NCHS uses (at least) three different targets:
  - ▶ Unweighted Slope: slope $\propto \sum_{t=1}^{T} P_t \times year_t$
  - ▶ Population weighted Slope: slope $\propto \sum_{t=1}^{T} N_t P_t \times year_t$
  - ▶ Sample variance weighted Slope: slope $\propto \sum_{t=1}^{T} \frac{P_t \times year_t}{var(\hat{P}_t)}$
- ▶ These three estimates will all be very similar if the underlying population size does not change and if the sample size/design doesn't change

# Methodological Issues: Survey estimates may be correlated between years

- ▶ Surveys estimates such as from NHIS are dependent between years
  - Software will take into account correlation but record-level data is needed
  - Many estimation operations work from table estimates (i.e., correlations are not typically constructed for all items)
    - ▶ Incorporating correlation into prevalence estimation usually results in small changes but there has been no systematic understanding
    - ▶ Requiring analysis to always include correlation will increase workload, could reduce output
    - ▶ Will recommend individual reports always incorporate dependencies but not clear on large pubs like HUS and HP

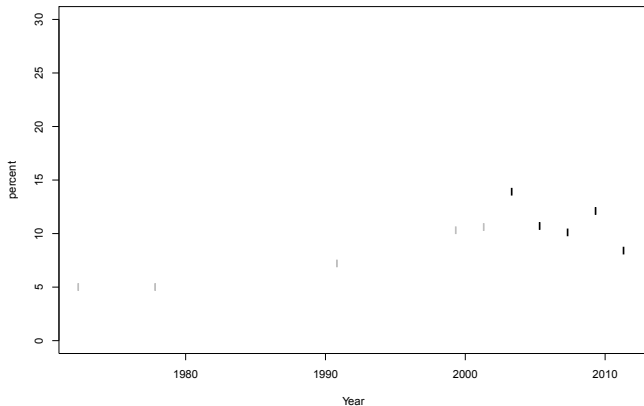# Methodological Issues: How to pick a time frame?
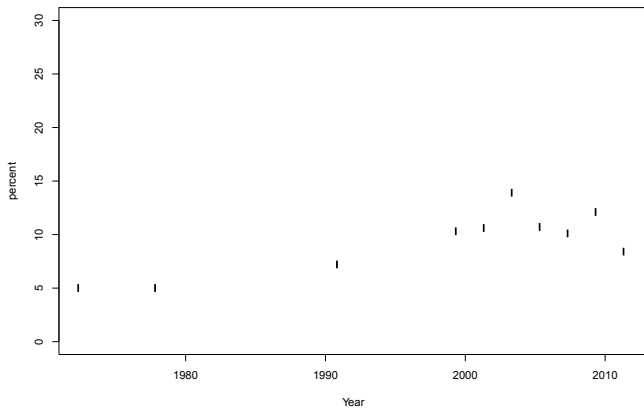
▶ Explaining an entire long trend may not be relevant

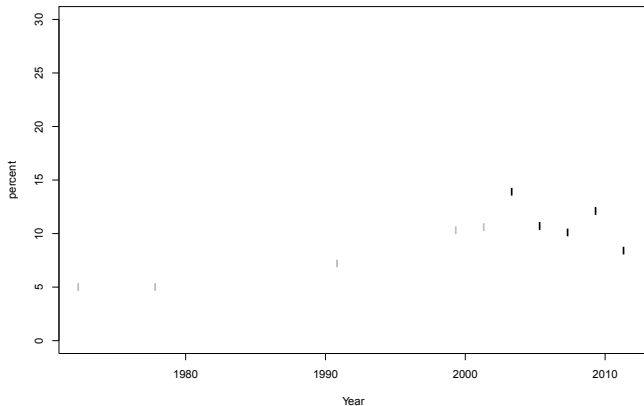# Methodological Issues: How to pick a time frame?
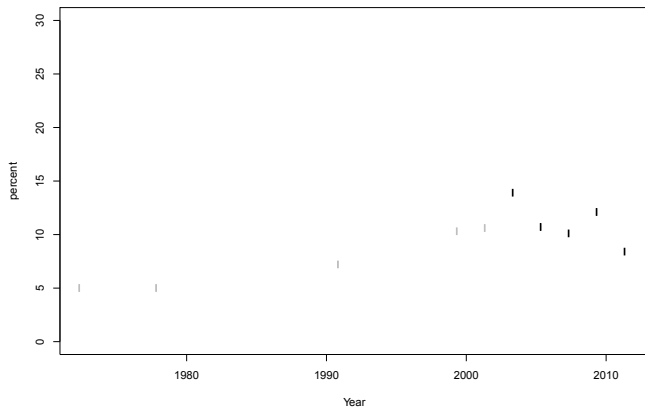
▶ Explaining an entire long trend may not be relevant

Methodological Issues: How to pick a time frame?

- Explaining an entire long trend may not be relevant
- Pick time period based on a specific policy change?
- Focus interest on past decade?
- Interested in what has been happening "recently"?

Methodological Issues: How to pick a time frame?

- ▶ Explaining an entire long trend may not be relevant
- ▶ Pick time period based on a specific policy change?
- ▶ Focus interest on past decade?
- ▶ Interested in what has been happening "recently"?
- ▶ issue: is it an interesting trend or a statistical anomaly?

Methodological Issues: Choice of transformation?

- View choice of transformation as flexible
  (" all transformations are wrong but ...")

- Case in point: logistic regression
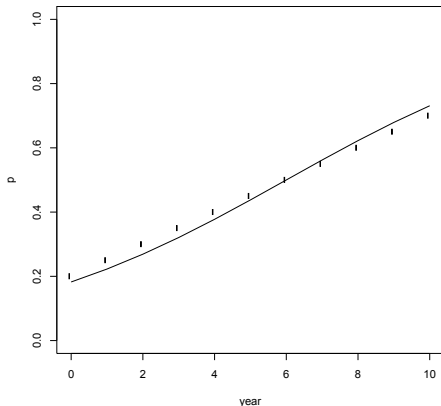
Methodological Issues: Choice of model?

- View choice of model as flexible ("all models are wrong ...")
- For trends, NCHS seems to rely on linear splines (joinpoint) and polynomial regression
  - joinpoint software:
    - <u>selection</u> is defensible from a design-based view
    - accounts for multiple testing in model choice
    - pinpoints an exact time point where a change takes place
  - polynomial regression model:
    - useful for checking deviations from linearity
    - relies on off-line separate multiple comparison model fitting
  - how important is it to always check the overall fit of the final trend model?
    - trade-off between false positives and false negatives?
    - is the complexity of a model obvious sometimes?

Methodological Issues: Other Issues

- Collapsing years together
    - individual estimates fail precision requirements for publication
    - group individual estimates into reliable groups of neighboring years
    - do a trend analysis on grouped data
    - theory states: regression estimates more precise if data not grouped
    - dilemma: present slope for ungrouped data with grouped individual estimates?

Methodological Issues: Other Issues

- Collapsing years together
    - individual estimates fail precision requirements for publication
    - group individual estimates into reliable groups of neighboring years
    - do a trend analysis on grouped data
    - theory states: regression estimates more precise if data not grouped
    - dilemma: present slope for ungrouped data with grouped individual estimates?

- Analysis with a small number of time points (3 or 4)
    - is trend analysis appropriate?
        - Yes - can still evaluate linearity or lack of
        - No - what does linearity mean for such a few points? - look at pairwise differences

Guidance Not Planned:

- ▶ Detailed guidance on software
- ▶ Time series methods
- ▶ Age period cohort models
- ▶ Determining underlying correlates of trend
- ▶ Causal analysis
- ▶ Superpopulation models
- ▶ Model fitting - new methods

What should guidance on trends consist of?

- What is the balance between subject matter expertise and statistical testing?
- How to guide multiple testing for balancing between false-positive and false negative conclusions?
- How much specific guidance should be provided to anticipate challenges to conclusions?
- How much guidance should be directed to researchers outside of NCHS?
- How much detail should reports include regarding methods/guidance used?