Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Common problems with regression.
    a. Inferring causation.
    b. Extrapolation.
    c. Curvature.
2. Testing significance of correlation or slope.

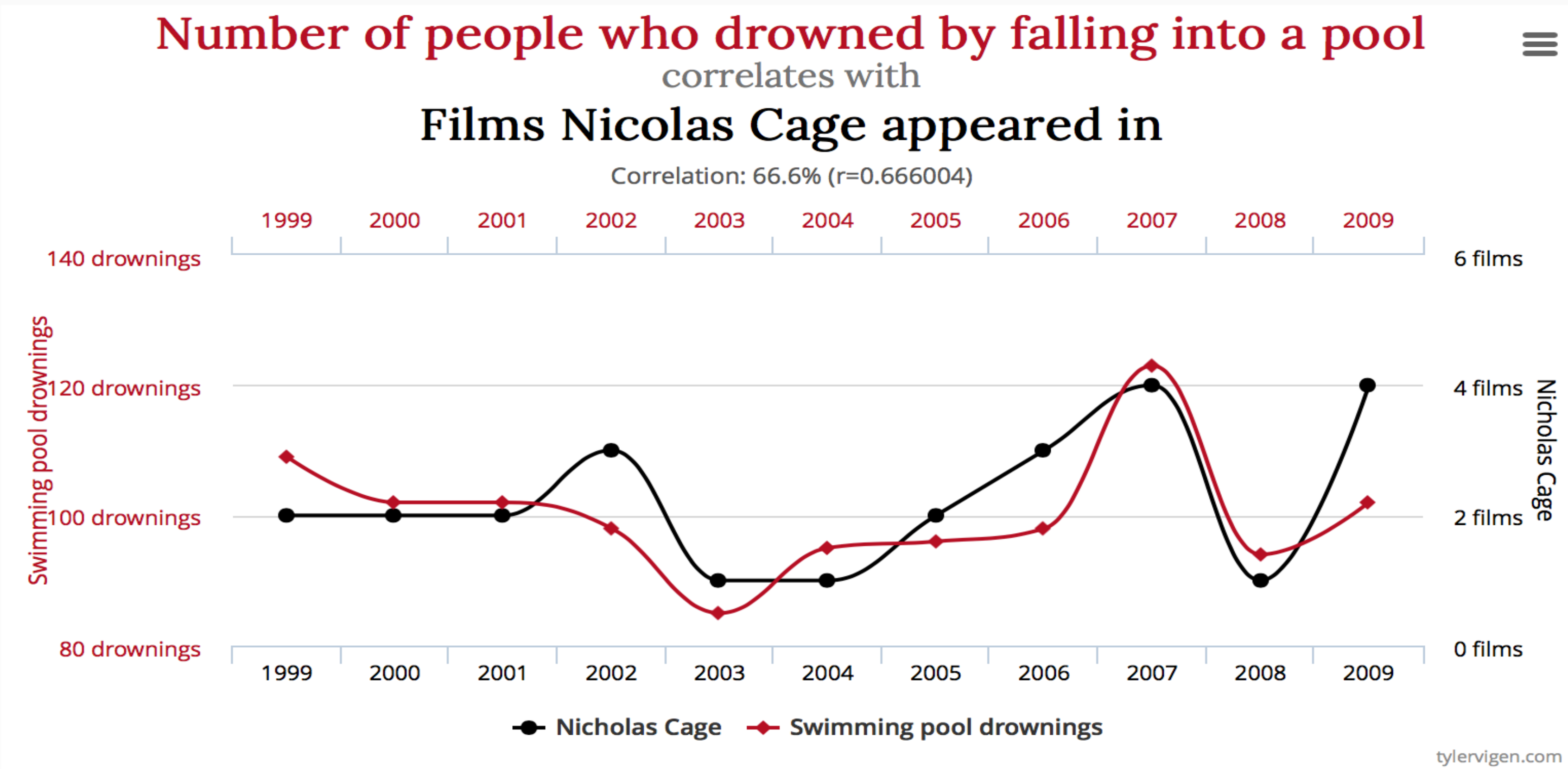
No class Thu Nov 24, Thanksgiving.
Read ch9.
Hw4 is 10.1.8, 10.3.14, 10.3.21, 10.4.11 and is due Tue Nov 29.

The final Fri Dec 9, 8am-11, right here, will be on ch1-10.
Bring a PENCIL and CALCULATOR and any books or notes you want. No computers.
http://www.stat.ucla.edu/~frederic/13/F16 .

# Common problems with regression.

- a. Correlation is not causation.

Especially with observational data.




Data sources: Centers for Disease Control & Prevention and Internet Movie Database

# Common problems with regression.

# Common problems with regression.

Holmes and Willett (2004) reviewed all prospective studies on fat consumption and breast cancer with at least 200 cases of breast cancer. "Not one study reported a significant positive association with total fat intake.... Overall, no association was observed between intake of total, saturated, monounsaturated, or polyunsaturated fat and risk for breast cancer."

They also state "The dietary fat hypothesis is largely based on the observation that national per capita fat consumption is highly correlated with breast cancer mortality rates. However, per capita fat consumption is highly correlated with economic development. Also, low parity and late age at first birth, greater body fat, and lower levels of physical activity are more prevalent in Western countries, and would be expected to confound the association with dietary fat."

# Common problems with regression.

- b. Extrapolation.

# Common problems with regression.

- b. Extrapolation.
- Often researchers extrapolate from high doses to low.

D.M. Odom et al.

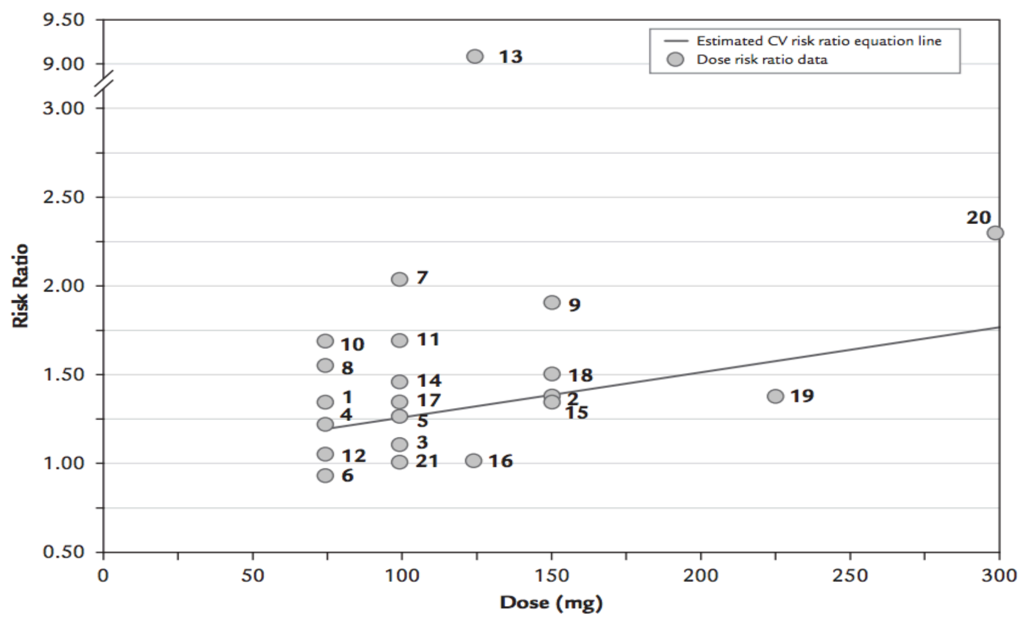



Figure 4. Relationship between diclofenac daily dose and the estimated risk ratio of a cardiovascular event. Numbers correspond to the observations in Table III.

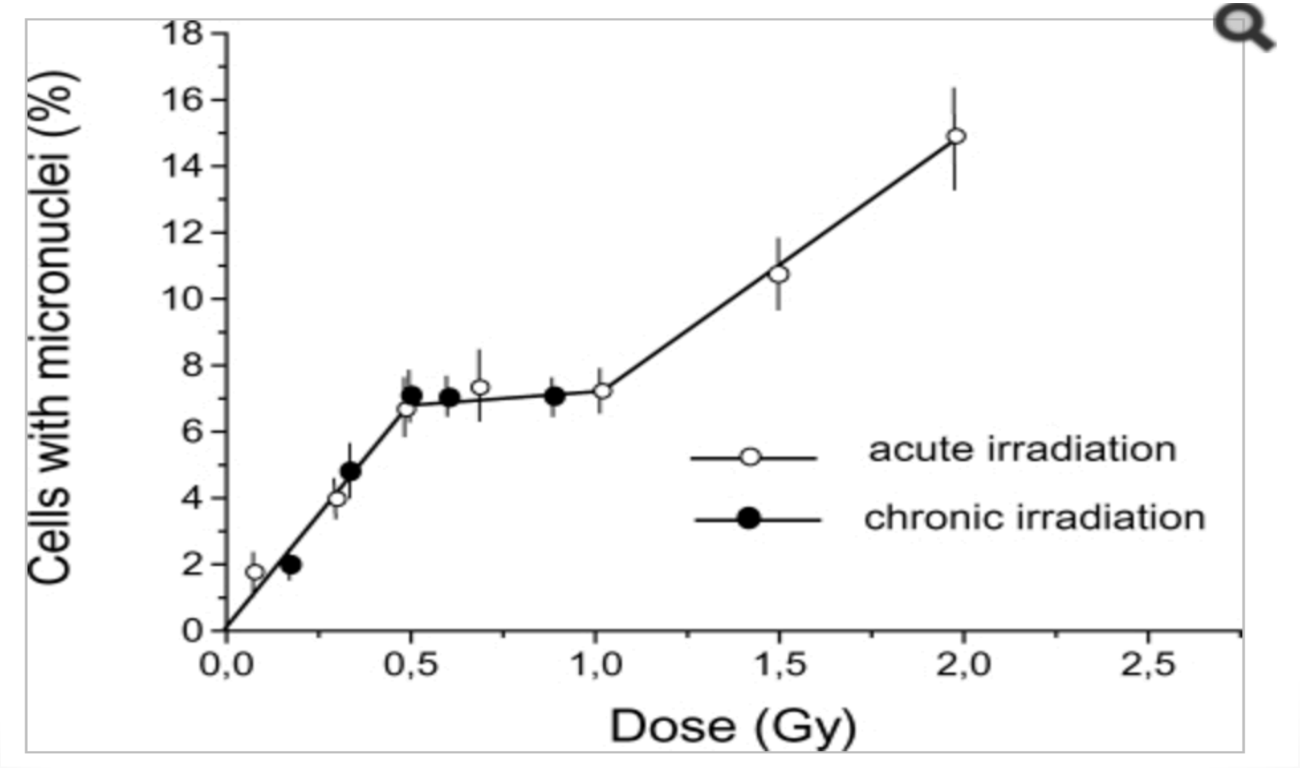
# Common problems with regression.

- b. Extrapolation.

The relationship can be nonlinear though.

Researchers also often extrapolate from animals to humans.

Zaichkina et al. (2004) on hamsters



# Common problems with regression.

- c. Curvature.

The best fitting line might fit poorly. Port et al. (2005).



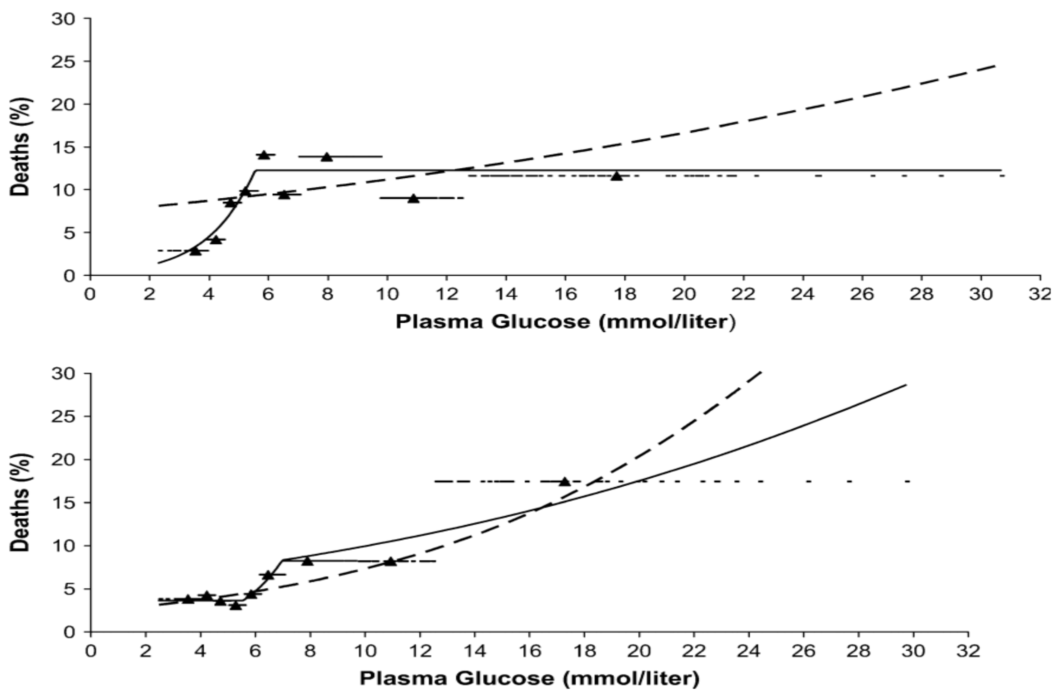
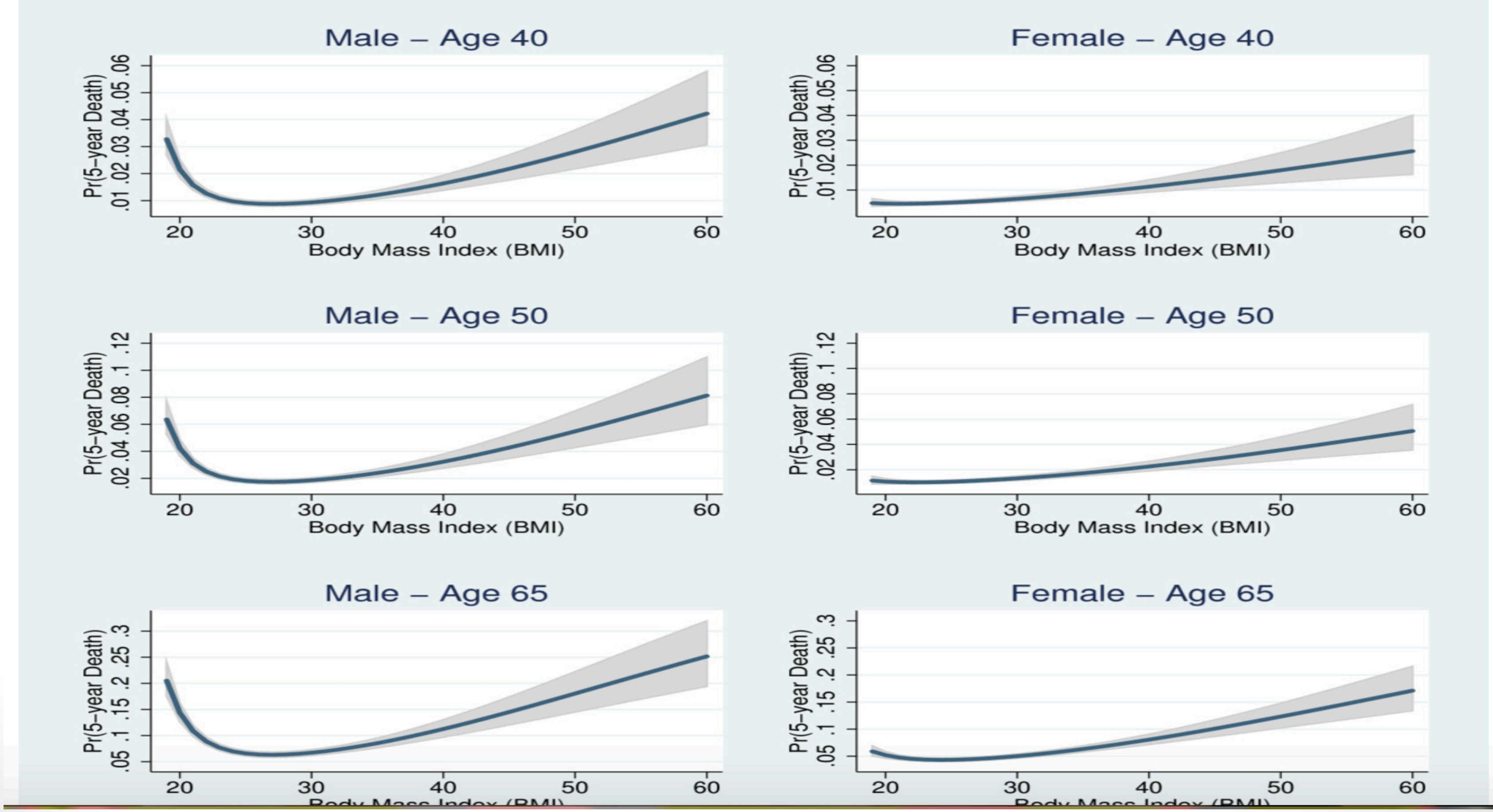



**FIGURE 4.** Adjusted 2-year rates of death from all causes for men (upper panel) and women (lower panel) separately, by glucose level, predicted by three models, Framingham Heart Study, 1948–1978. Linear model (dashed curve); optimal spline models (solid curve). The horizontal dashed

# Common problems with regression.

- c. Curvature.

The best fitting line might fit poorly. Wong et al. (2011).



# How well does the line fit?

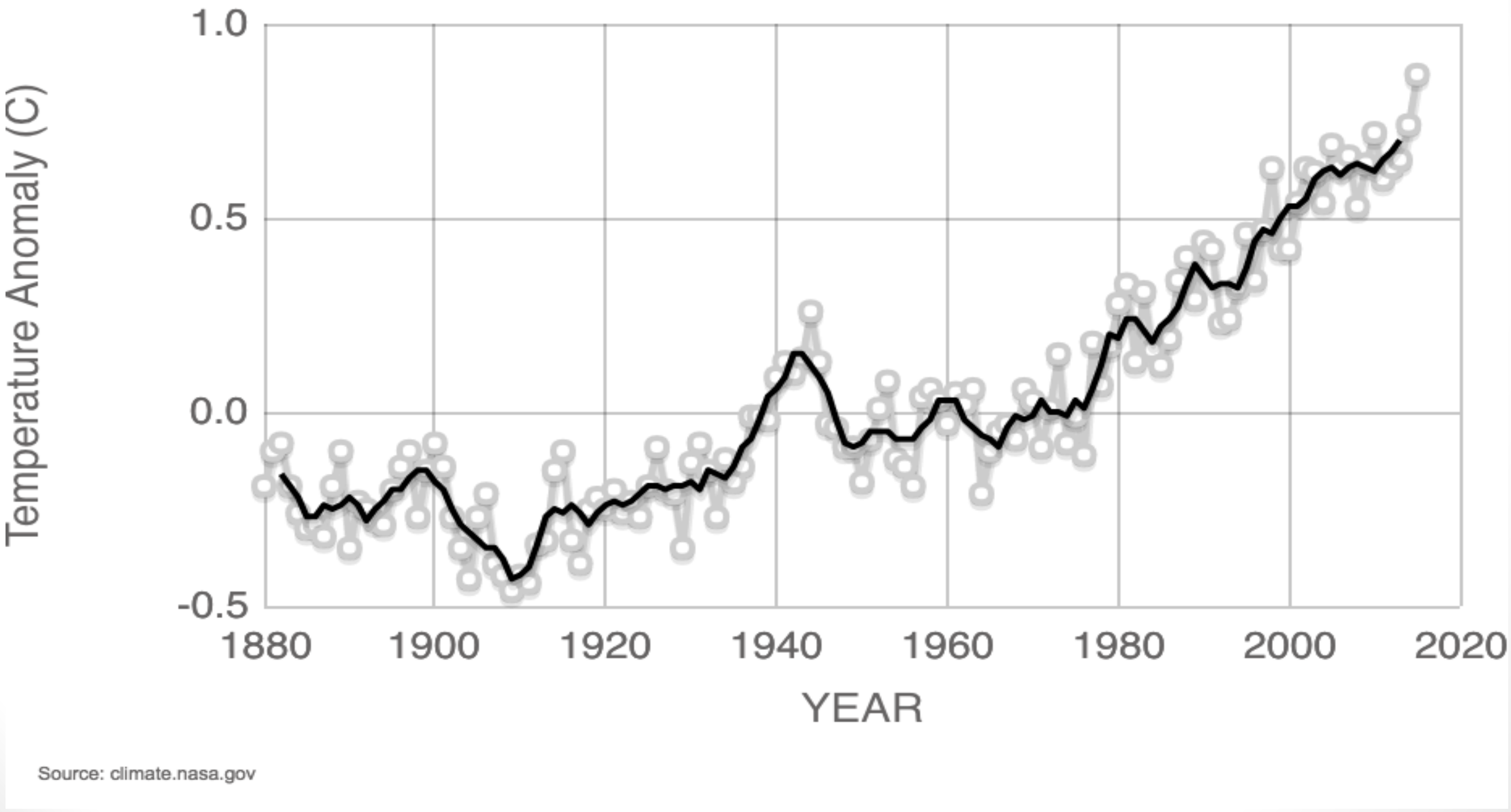
- $r^2$ is a measure of fit. It indicates the amount of scatter around the best fitting line.
- Residual plots can indicate curvature, outliers, or heteroskedasticity.
- $\sqrt{1-r^2}\, s_y$ is useful as a measure of how far off predictions would have been on average.

Note that regression residuals have mean zero, whether the line fits well or poorly.

# Common problems with regression.

- d. Statistical significance.

Could the observed correlation just be due to chance alone?



# Inference for the Regression Slope: Theory-Based Approach

Section 10.5

# Do students who spend more time in non-academic activities tend to have lower GPAs?

Example 10.4

# Do students who spend more time in non-academic activities tend to have lower GPAs?

- The subjects were 34 undergraduate students from the University of Minnesota.
- They were asked questions about how much time they spent in activities like work, watching TV, exercising, non-academic computer use, etc. as well as what their current GPA was.
- We are going to test to see if there is a negative association between the number of hours per week spent on nonacademic activities and GPA.

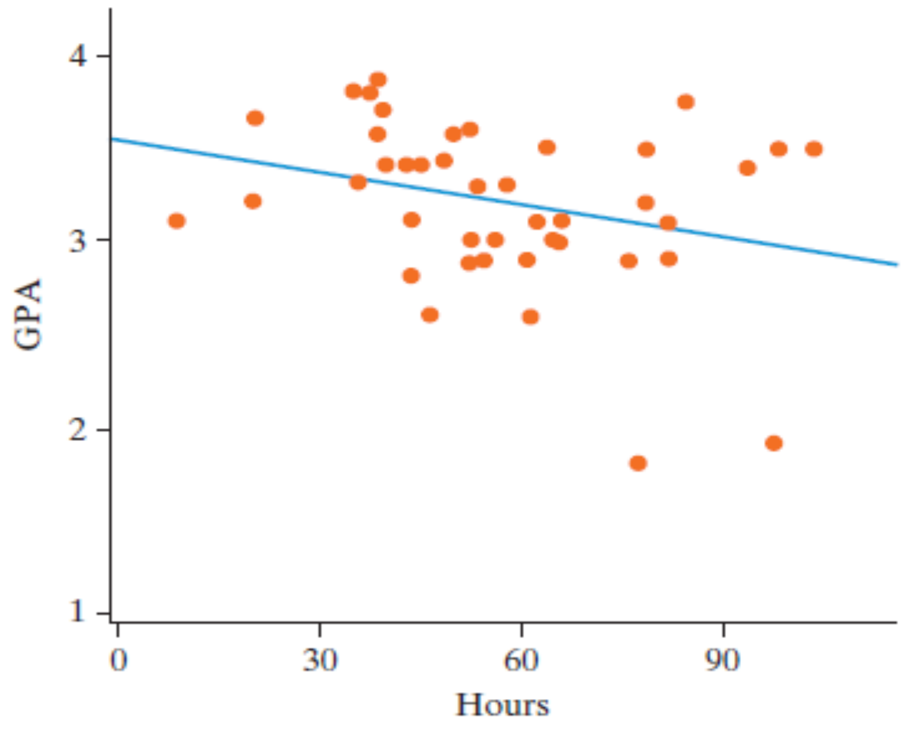
# Hypotheses

- Null Hypothesis: There is no association between the number of hours students spend on nonacademic activities and student GPA in the population.

- Alternative Hypothesis: There is a negative association between the number of hours students spend on nonacademic activities and student GPA in the population.

# Descriptive Statistics

- $\widehat{\text{GPA}} = 3.60 - 0.0059(\text{nonacademic hours}).$
- What do the slope and y-intercept mean?



# Shuffle to Develop Null Distribution

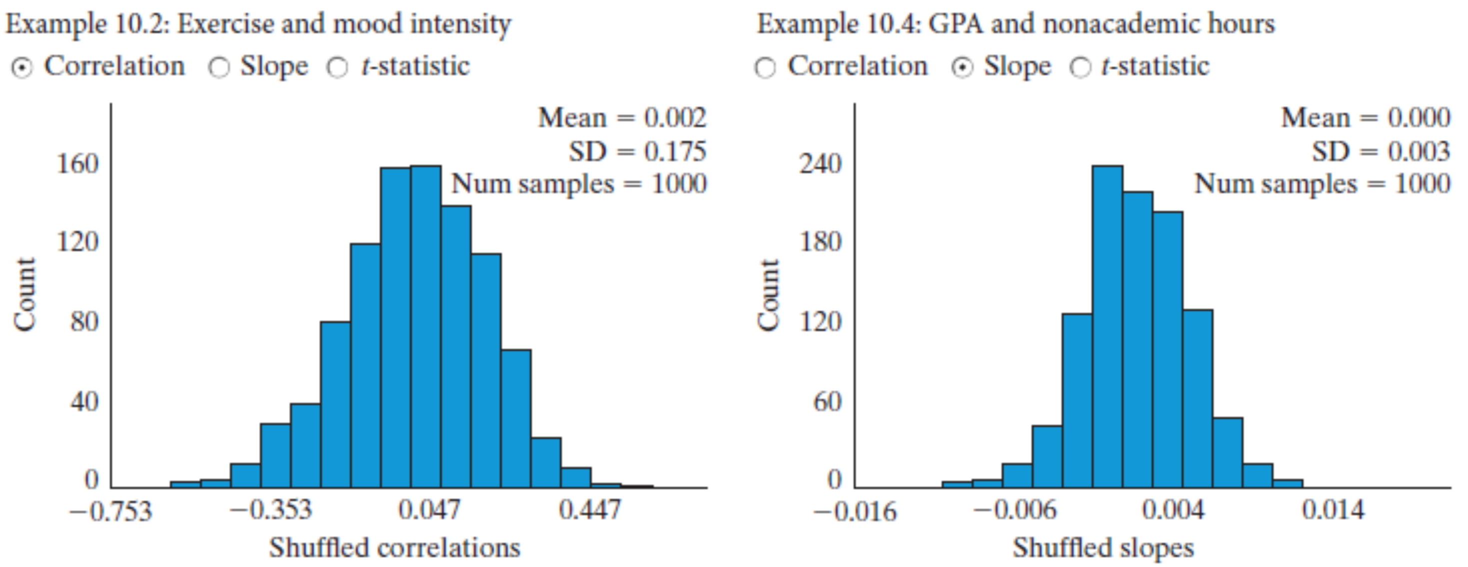
- We are going to shuffle just as we did with correlation to develop a null distribution.
- The only difference is that we will be calculating the slope each time and using that as our statistic.
- a test of association based on slope is equivalent to a test of association based on a correlation coefficient.

# Beta vs Rho

- Testing the slope of the regression line is equivalent to testing the correlation (same p-value, but obviously different confidence intervals since the statistics are different)
- Hence these hypotheses are equivalent.
  - $H_o$: $\beta = 0$ $H_a$: $\beta < 0$ (Slope)
  - $H_o$: $\rho = 0$ $H_a$: $\rho < 0$ (Correlation)
- Sample slope (b) Population (β: beta)
- Sample correlation (r) Population (ρ: rho)
- When we do the theory based test, we will be using the *t*-statistic which can be calculated from either the slope or correlation.

# Introduction

- Our null distributions are again bell-shaped and centered at 0 (for either correlation or slope as our statistic).




The book on p549 finds a p value of 3.3% by simulation.

# Validity Conditions

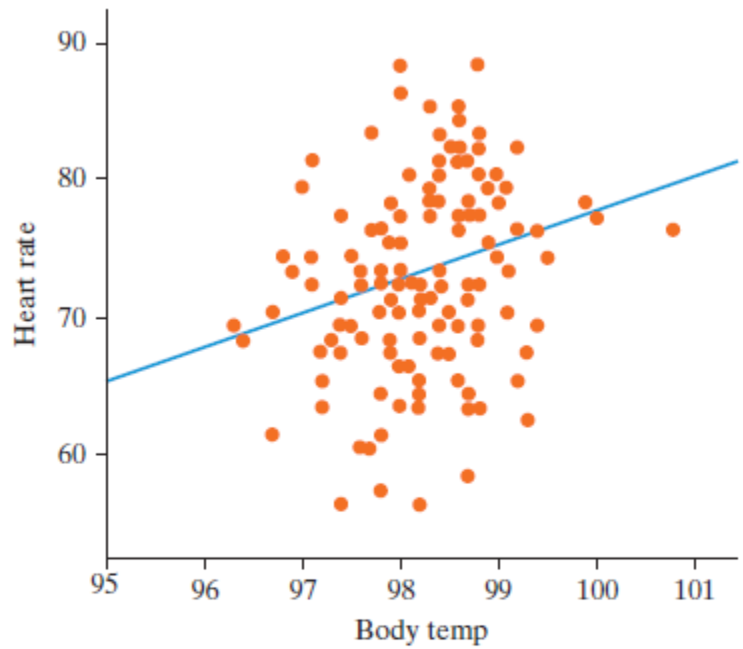
- Under certain conditions, theory-based inference for correlation or slope of the regression line use *t*-distributions.
- We could use simulations or the theory-based methods for the slope of the regression line.
- We would get the same p-value if we used correlation as our statistic.

# Predicting Heart Rate from Body Temperature

*Example 10.5A*

# Heart Rate and Body Temp

- Earlier we looked at the relationship between heart rate and body temperature with 130 healthy adults
- $\text{Predicted Heart Rate} = -166.3 + 2.44(\text{Temp})$
- r = 0.257

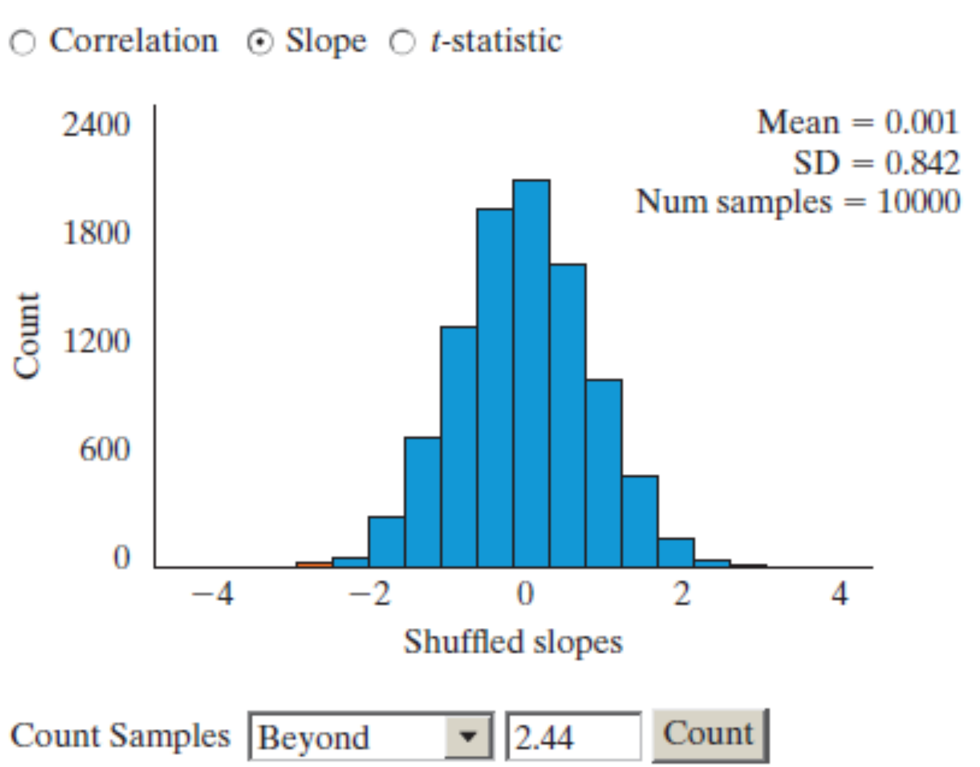


# Heart Rate and Body Temp

- We tested to see if we had convincing evidence that there is a positive association between heart rate and body temperature in the population using a simulation-based approach. (We will make it 2-sided this time.)
- **Null Hypothesis**: There is no association between heart rate and body temperature in the population. $\beta = 0$
- **Alternative Hypothesis**: There is an association between heart rate and body temperature in the population. $\beta \neq 0$

# Heart Rate and Body Temp

We get a very small p-value (0.0036). Anything as extreme as our observed slope of 2.44 happening by chance is very rare



# Heart Rate and Body Temp

- We can also approximate a 95% confidence interval

  observed statistic $\pm$ 2 SD of statistic

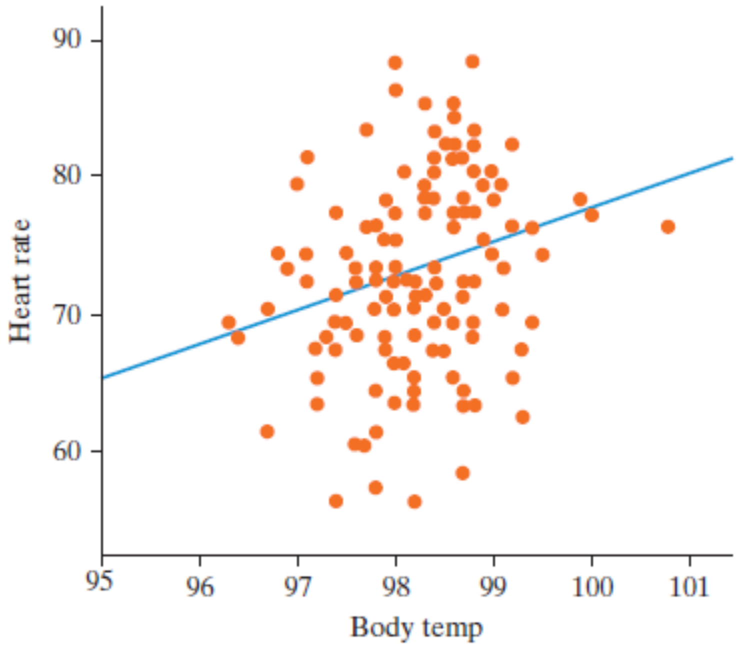
  $2.44 \pm 2(0.842) = 0.76$ to $4.12$

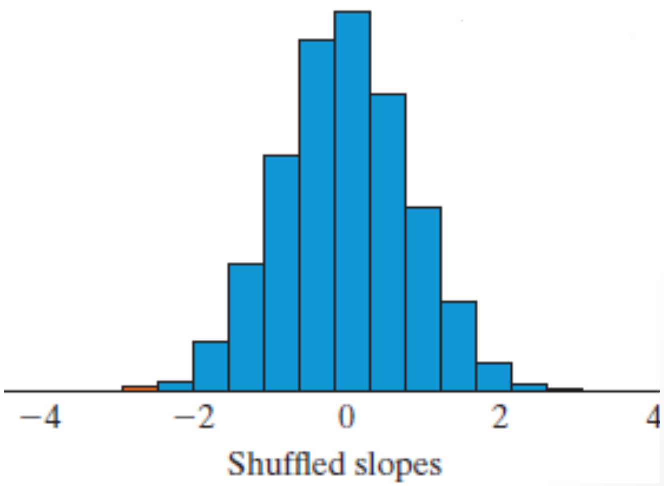
- What does this mean?

  We're 95% confident that, in the population of healthy adults, each 1° increase in body temp is associated with an increase in heart rate of between 0.76 to 4.12 beats per minute

# Heart Rate and Body Temp

- The theory-based approach should work well since the distribution has a nice bell shape
- Also check the scatterplot



# Heart Rate and Body Temp

- We will use the t-statistic to get our theory-based p-value.
- We will find a theory-based confidence interval for the slope.
- On p554, the book notes the formula
- t = $\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$.
- Here the t statistic is 2.97.
- The p-value is .36%. So the correlation is statistically significantly greater than zero.

# Smoking and Drinking

Example 10.5B

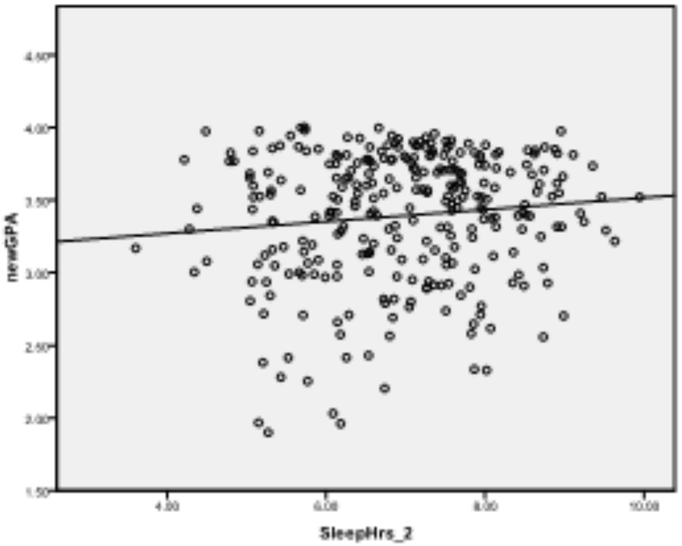
# Validity Conditions

Remember our validity conditions for theory-based inference for slope of the regression equation.

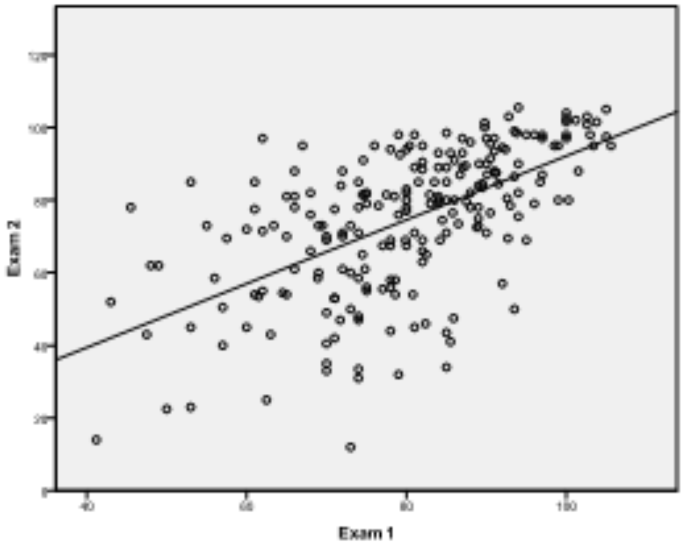
1. The scatterplot should follow a linear trend.
2. There should be approximately the same number of points above and below the regression line (symmetry).
3. The variability of vertical slices of the points should be similar. This is called homoskedasticity.

# Validity Conditions

- Let's look at some scatterplots that do not meet the requirements.




(a)



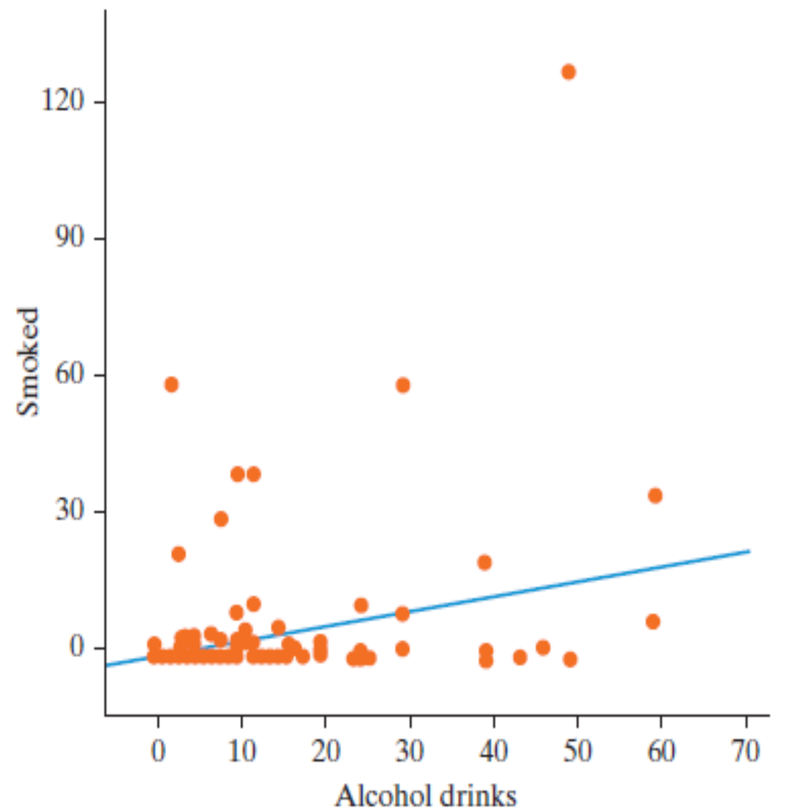

(b)

# Smoking and Drinking

The relationship between number of drinks and cigarettes per week for a random sample of students at Hope College.

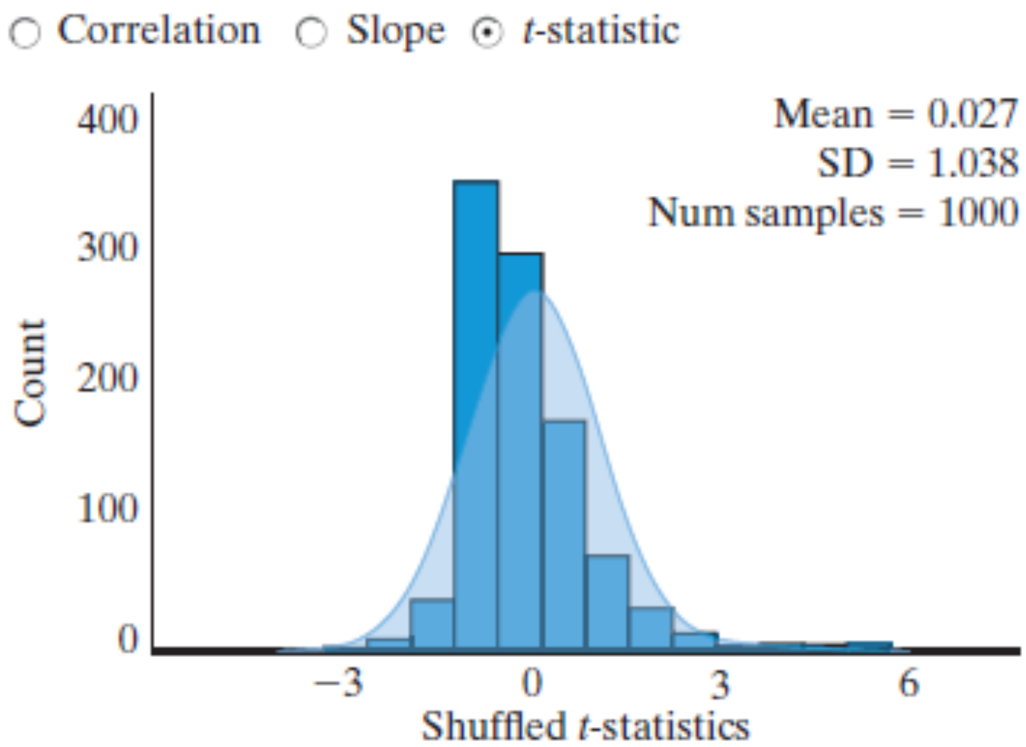
The dot at (0,0) represents 524 students

Are the conditions met? Hard to say. The book says no.



# Smoking and Drinking

- When the conditions are not met, applying simulation-based inference is preferable to theory-based t-tests and CIs.



# Validity Conditions

- What do you do when validity conditions aren't met for theory-based inference?
  - Use the simulated-based approach.
- Another strategy is to "transform" the data on a different scale so conditions are met.
  - The logarithmic scale is common.
- One can also fit a different curve, not necessarily a line.