Continuous data – correlation

Fundamental Rule of Data Analysis

Different types of data require different statistical analyses.

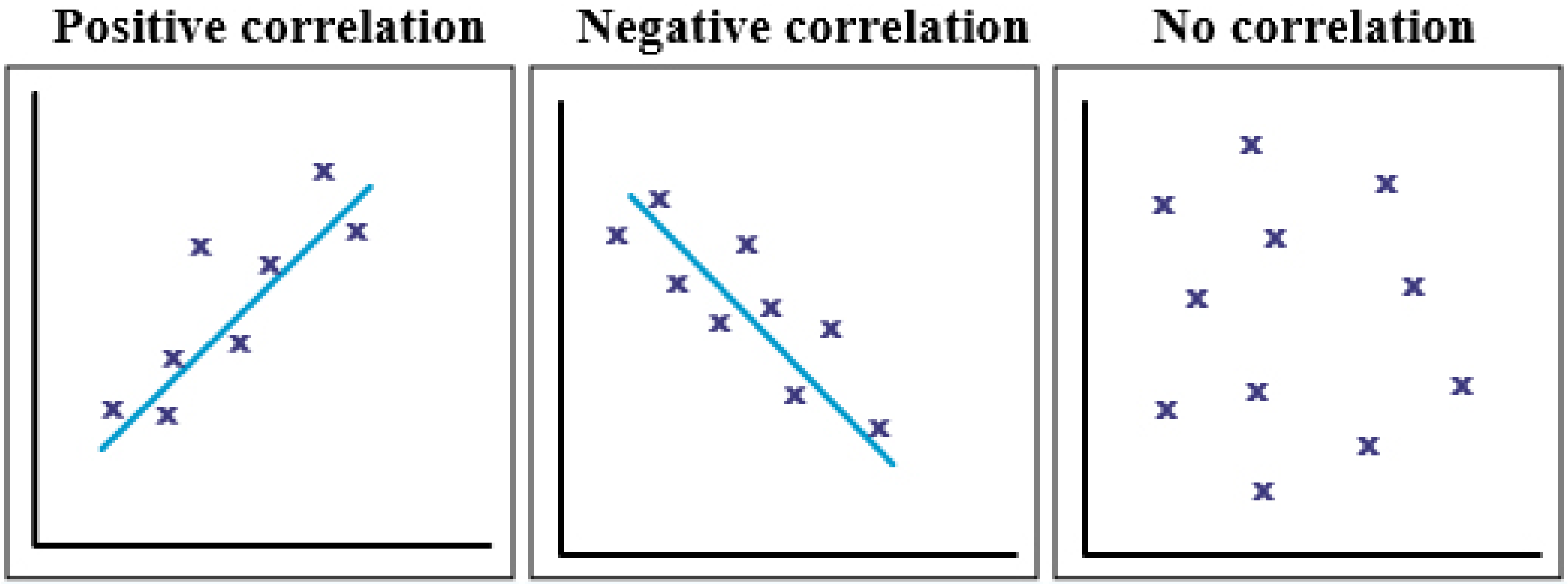
Types of Variables/Number of Groups

Variable(s)	Analysis
Continuous	One-group t-test
Continuous vs. Categorical (2 categories)	Two-group t-test
Continuous vs. Categorical (>2 categories)	ANOVA
Continuous vs. Continuous	Correlation/linear regression

Correlation

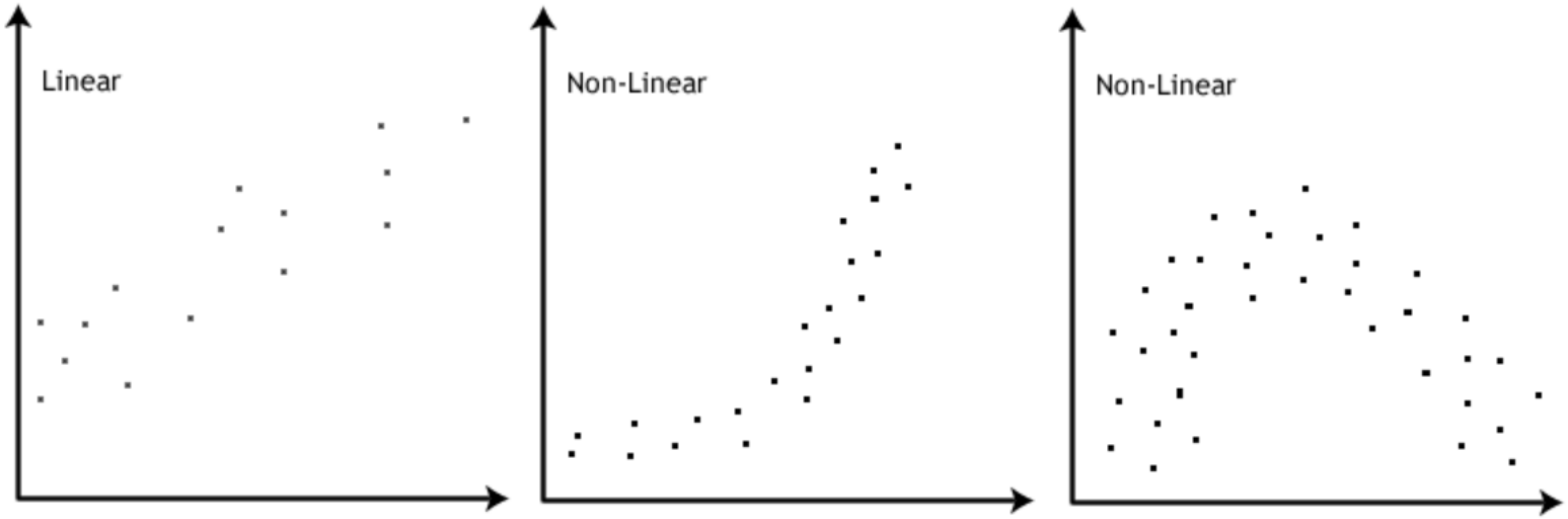
- **Correlation** is the degree to which two continuous variables are related, provided that relationship is linear
- A value of 1 or -1 indicates perfect correlation (all points fall exactly on a straight line)
- A value of 0 indicates no association between the two variables
- Measure of correlation has no units

Correlation



Linear Association

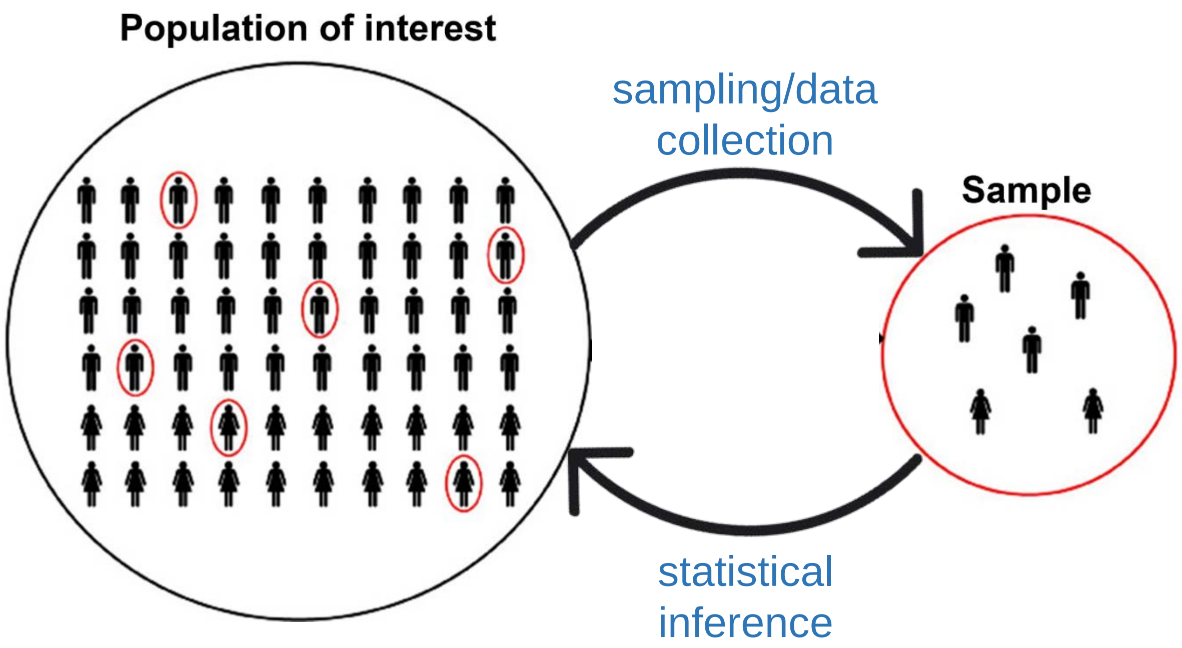
- Correlation only measures the strength of **linear** association
    - How close does it match up to a straight line
- Variables could be highly related, but if the relationship isn't linear, correlation may be low



Notation

- ρ = population correlation
- r = sample correlation

ρ is fixed



r varies (would be different with a different sample of the population)

Estimation and Testing

- Goal: Estimate the correlation between two variables in the population (ρ) and determine if the association is significant.
  - Best estimate of the population correlation is the sample correlation (r)
  - Can calculate confidence intervals for ρ
    - "We are 95% confident that the true correlation between ______ and ______ in the population is between ____ and _____."
  - Correlation test: hypothesis test to show if association is significant

$$H_0\colon\ \rho = 0 \qquad H_A\colon\ \rho \neq 0$$

Reject H_0 when p-value ≤ α
Fail to reject H_0 when p-value > α

Low Birth Weight Data

- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

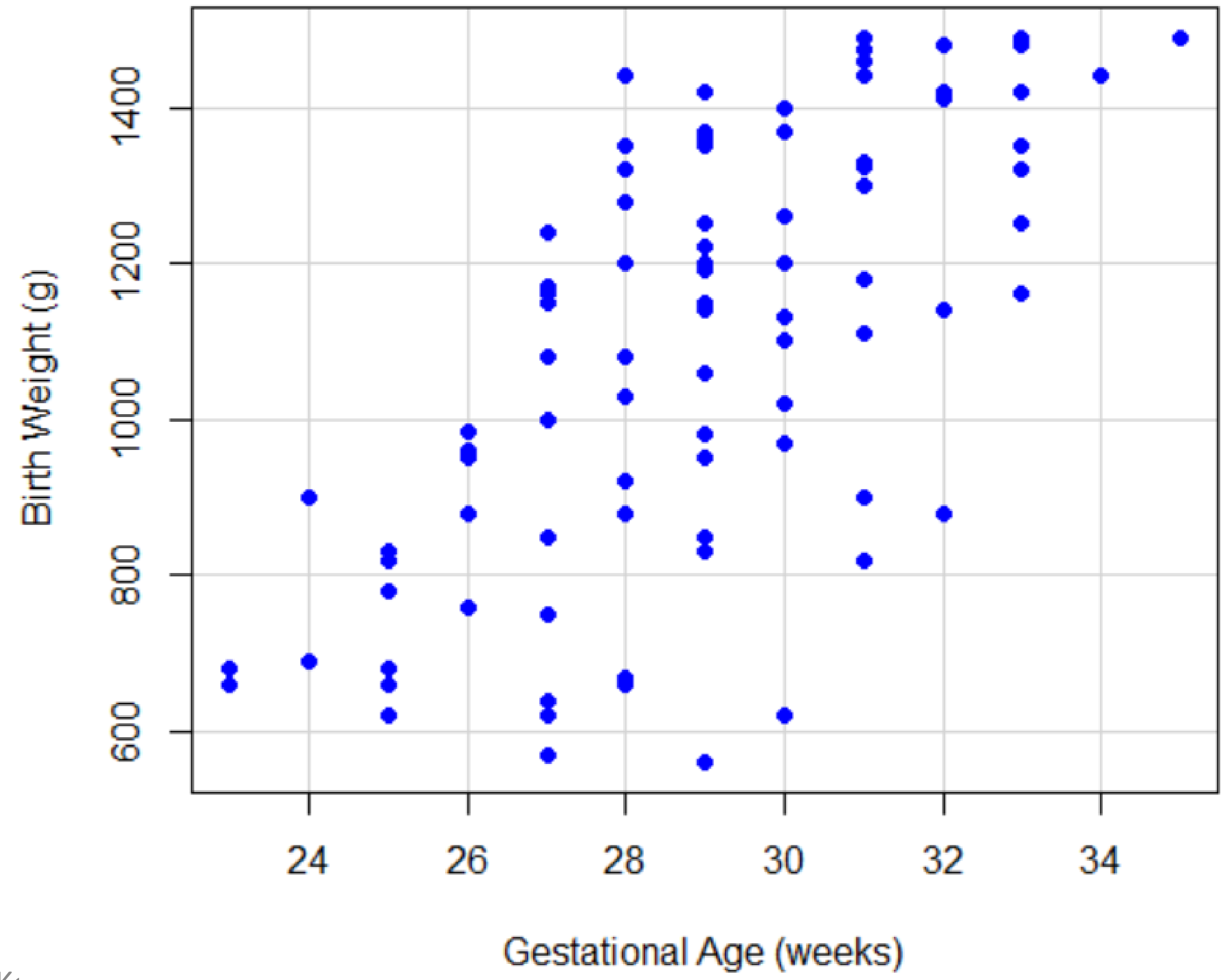
Variable	Description
gestage	Gestational age at time of birth (weeks)
momage	Mother's age (years)
birthwt	Birth weight of the baby (g)
length	Length of the baby (cm)
headcirc	Baby's head circumference (cm)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.

Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Example: Gestational Age/Birth Weight

- Estimate the correlation between gestational age and birth weight in low birth weight infants.
- Is there a significant association between gestational age and birth weight in low birth weight infants?

Example: Gestational Age/Birth Weight



Does the association appear to be linear?

Example: Gestational Age/Birth Weight

Correlation test for gestational age/birth weight:

```
        Pearson's product-moment correlation

data:  birthwt and gestage
t = 8.6954, df = 98, p-value = 8.149e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5325507 0.7580894
sample estimates:
      cor
0.6599376
```

$r = 0.66$
(strong positive association between gestational age and birth weight)

We are 95% confident that the true correlation between gestational age and birth weight in low birth weight infants is between 0.53 and 0.76.

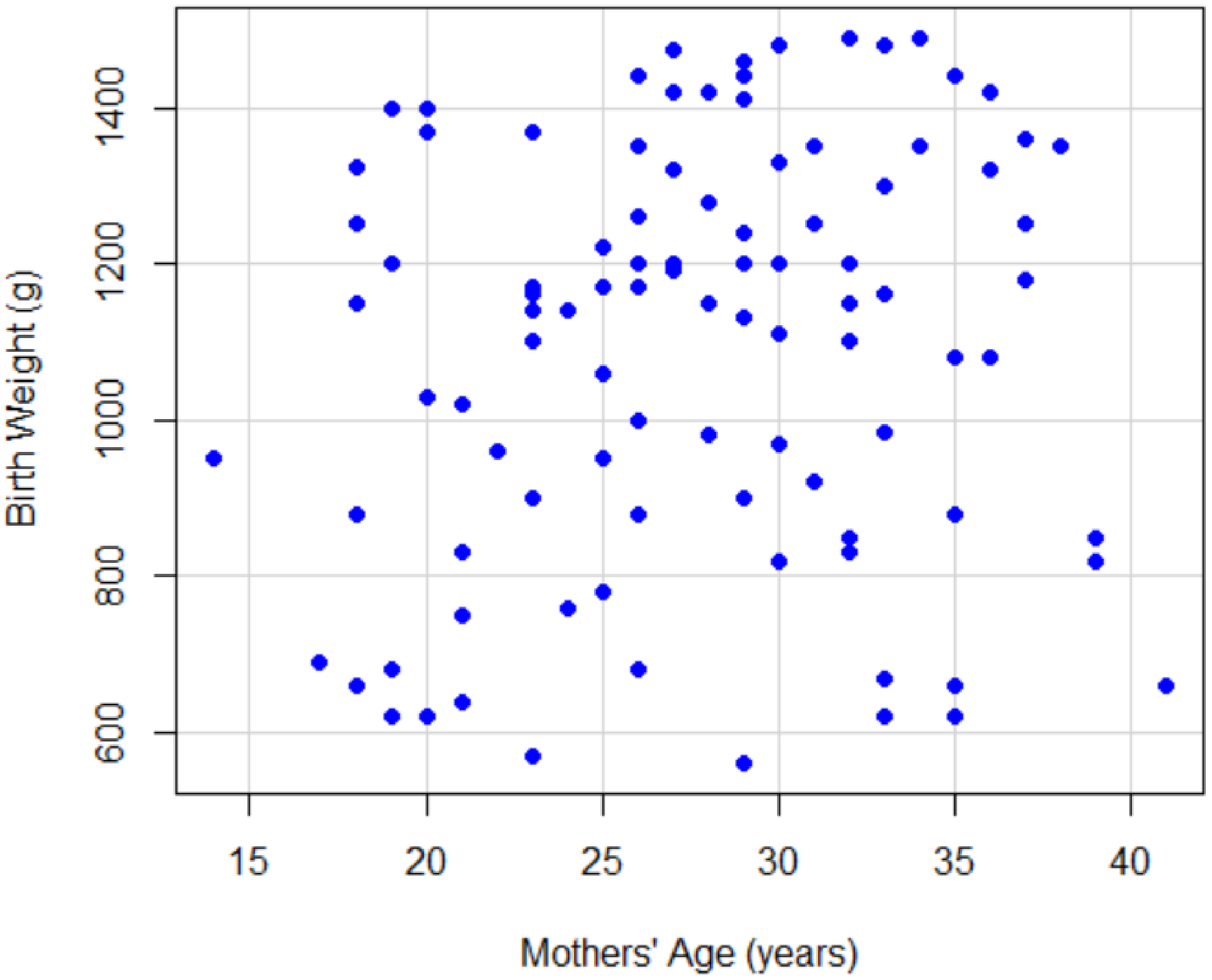
p-value < 0.001

Since the p-value is less than 0.05, we reject H_0 and conclude that there is sufficient evidence to suggest that there is an association between gestational age and birth weight in low birth weight infants.

Example: Mothers’ Age/Birth Weight

- Estimate the correlation between the age of the mother and the birth weight of the infant in low birth weight infants.
- Is there a significant association between mothers’ age and infants’ birth weight in low birth weight infants?

Example: Mothers' Age/Birth Weight



Does the association appear to be linear?

Example: Mothers' Age/Birth Weight

Correlation test for mothers' age/birth weight:

```
        Pearson's product-moment correlation

data:  birthwt and momage
t = 1.5488, df = 98, p-value = 0.1247
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04315656  0.34064763
sample estimates:
      cor
0.1545718
```

$r = 0.15$
(weak positive association between mothers' age and birth weight)

We are 95% confident that the true correlation between mothers' age and birth weight in low birth weight infants is between -0.04 and 0.34.

p-value $= 0.12$

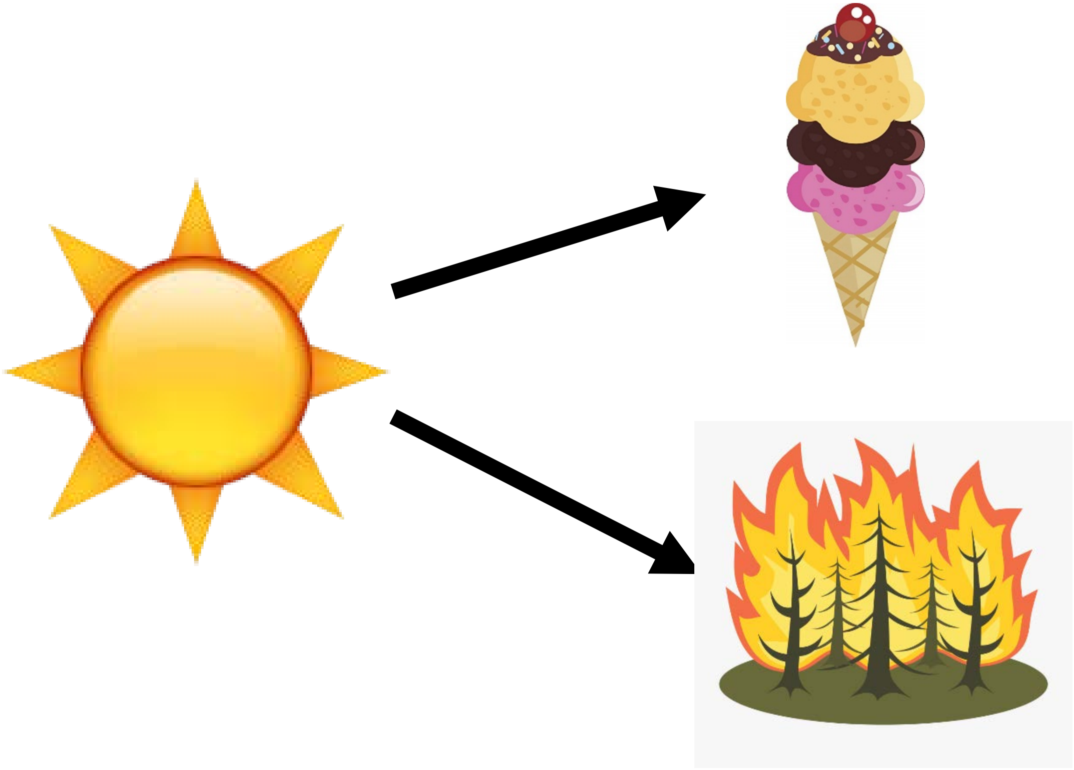
Since the p-value is greater than 0.05, we fail to reject H_0 and conclude that there is not sufficient evidence to say that there is an association between mothers' age and infants' birth weight in low birth weight infants.

Causation

- Correlation is not the same as causation
- Just because two variables are correlated does not necessarily mean that one causes the other
- Possible reasons for correlation between two variables:
    - One causes the other
    - Both are caused by a third variable (**confounding**)
    - Completely spurious

Example: Confounding

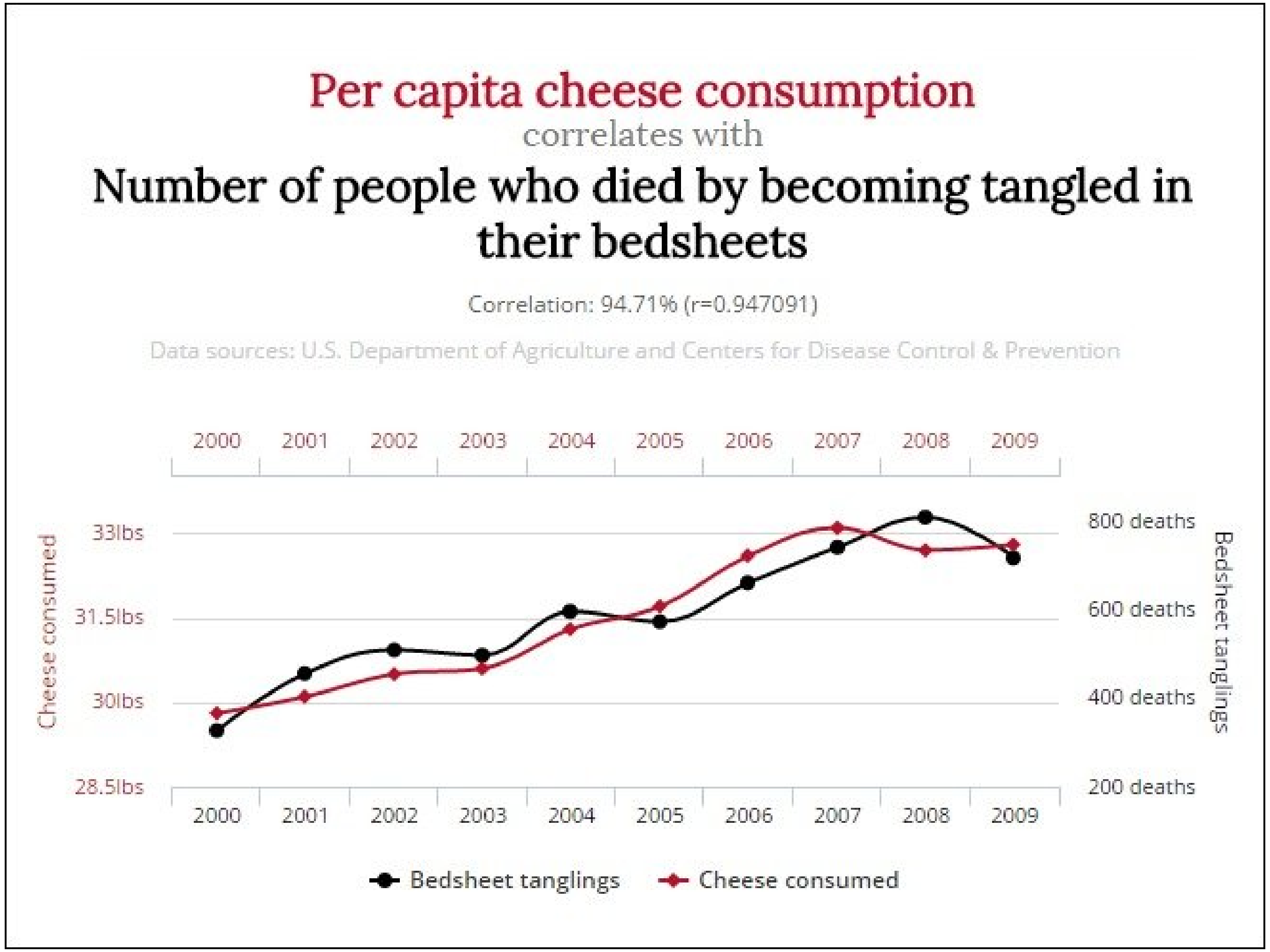
- Ice cream sales are correlated with forest fires
- Common cause = hot, dry weather



YouTube video “How Ice Cream Kills! Correlation vs. Causation”:

https://www.youtube.com/watch?v=VMUQSMFGBDo

Example: Spurious Correlation



Moving Toward Causation

- Experiments/randomized trials
  - Investigators manipulate the exposure variable
- Regression
  - Allows us to “control” for common causes or other variables (still not causal, but it helps)
- Causal inference
  - Advanced statistical technique that enables causal conclusions to be made when certain assumptions are met

Limitations of Correlation

- Only quantifies the strength of the linear relationship between two variables
- High correlation does not imply a cause-and-effect relationship
- Very sensitive to outliers/extreme values, and thus can sometimes be misleading
- Cannot be extrapolated beyond the observed ranges of the variables

Important Points

- Measure of correlation and test for significant correlation
- Concept of linear association
- When and why correlation is not causation
- Limitations of correlation