Basic Statistical Models

Dirk Grunwald
University of Colorado, Boulder

Basic Statistical Models

- What do we mean by “model”
- Random samples
- Statistical model for repeated measurements
- Example: disk drive failures
- Distribution features and statistical models
  - Sample expectation
  - Sample variance
- Linear regression models

What's a model?

- We've described various probability distributions and describe how some "real world" data tends to fit those distributions.

What's a model?

- We've described various probability distributions and describe how some "real world" data tends to fit those distributions.
- Now, we turn that on it's head – we have empirical data and we seek to understand if the data can be modeled by distributions.

What's a model?

- We've described various probability distributions and describe how some "real world" data tends to fit those distributions.
- Now, we turn that on it's head – we have empirical data and we seek to understand if the data can be <u>modeled</u> by distributions.
- To do this, we'll need to <u>estimate</u> parameters of the model

  from the empirical data.

What's a model?

- We've described various probability distributions and describe how some "real world" data tends to fit those distributions.
- Now, we turn that on it's head – we have empirical data and we seek to understand if the data can be <u>modeled</u> by distributions.
- To do this, we'll need to <u>estimate</u> parameters of the model

  from the empirical data.
- And, we'll need to determine how <u>confident</u> we are in our estimate of those parameters.

Random Sample

A random sample is a collection of random variables

$$X_1, X_2, \ldots, X_n$$

that have the same probability distribution and are mutually independent.

Random Sample

A random sample is a collection of random variables

$$X_1, X_2, \ldots, X_n$$

that have the same probability distribution and are mutually independent.

Example: Flip a coin 10 times, subsequent flip corresponds to $X_1, X_2, \ldots, X_{10}$

Random Sample

A random sample is a collection of random variables

$$X_1, X_2, \ldots, X_n$$

that have the same probability distribution and are mutually independent.

Example: Flip a coin 10 times, subsequent flip corresponds to $X_1, X_2, \ldots, X_{10}$

Example: Monitor a data center full of disk drives, count the number of failures per month.

Statistical Model for Repeated Measurements

Measured data are realization of random samples.

Statistical Model for Repeated Measurements

Measured data are realization of random samples.

Statistical Model for Repeated Measurements

A dataset consisting of values $x_1, x_2, \ldots, x_n$ of repeated measurements of the same quantity is modeled as the realization of a random sample $X_1, X_2, \ldots, X_n$. The model may include a partial specification of the probability distribution function for each X_i.

Statistical Model for Repeated Measurements

The probability of X_i is the model distribution, often a collection of distributions

Statistical Model for Repeated Measurements

The probability of X_i is the <u>model distribution</u>, often a collection of distributions that have <u>model parameters</u>.

Statistical Model for Repeated Measurements

The probability of X_i is the model distribution, often a collection of distributions that have model parameters.

We would believe the model distribution is derived from the specific true distribution.

Statistical Model for Repeated Measurements

The probability of X_i is the model distribution, often a collection of distributions that have model parameters.

We would believe the model distribution is derived from the specific true distribution.

Example: Manufactures rate disk drives using an exponential mean-time-to-failure (MTTF) model.

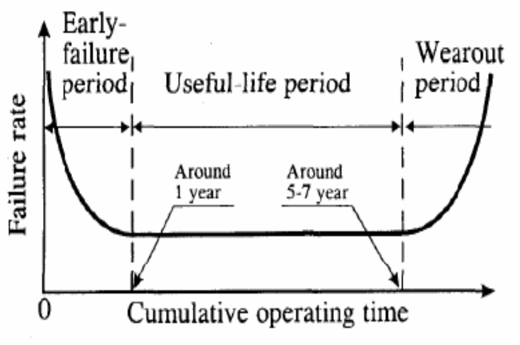
Statistical Model for Repeated Measurements

The probability of X_i is the model distribution, often a collection of distributions that have model parameters.

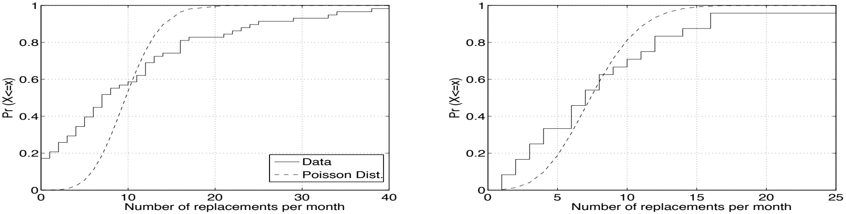
We would believe the model distribution is derived from the specific true distribution.

Example: Manufactures rate disk drives using an exponential mean-time-to-failure (MTTF) model. Accurate?

Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?, Bianca Schroeder, Garth A. Gibson (CMU), File systems and Storage Technology, 2007.



Example: disk drive failures II

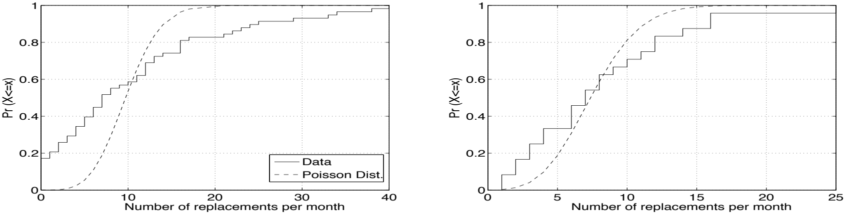


All years — Years 2-3

Figure 5: *CDF of number of disk replacements per month in HPC1*

- The statistical model includes three parts; birth, mid-life and death.

Example: disk drive failures II



All years   Years 2-3

Figure 5: *CDF of number of disk replacements per month in HPC1*

- The statistical model includes three parts; birth, mid-life and death.
- OK agreement with exponential assumption in mid-life phase.

Estimating parameters of the “true” distribution

- We’ve seen empirical estimators in EDA

Estimating parameters of the “true” distribution

- We’ve seen empirical estimators in EDA
- True μ estimated by <u>sample mean</u>

$$E[X] = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Estimating parameters of the "true" distribution

- We've seen empirical estimators in EDA
- True μ estimated by sample mean

$$E[X] = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- True σ^2 estimated by sample variance

$$s_n^2 = \frac{1}{n-1} \sum_{i=0}^{n} (x_i - \bar{x}_n)^2$$

  - Why $n - 1$?

Estimating parameters of the "true" distribution

- We've seen empirical estimators in EDA
- True μ estimated by sample mean

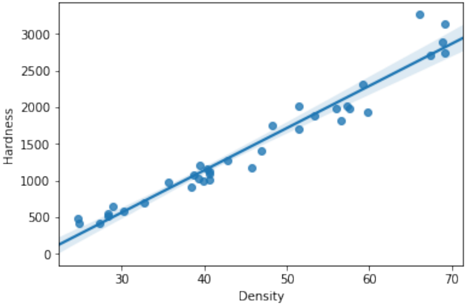
$$E[X] = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- True σ^2 estimated by sample variance

$$s_n^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x}_n)^2$$

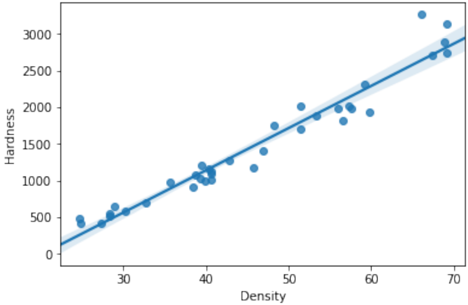
  - Why $n-1$? We'll discuss unbiased estimators later.

Linear Regression Models



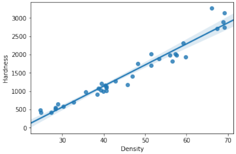
- Our model is hardness ~ density of timber

Linear Regression Models



- Our model is hardness $\sim$ density of timber
- A regression model would be
  hardness $\sim$ density of timber + noise

Linear Regression Models



Simple Linear Regression Model

A simple linear regression model for $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ assumes the x_i are non-random and the y_i are realizations of random variables Y_i satisfying

$$Y_i = \alpha + \beta x_i + R_i$$

where R_i are independent random variables with $E[U_i] = 0$ and $Var[U_i] = \sigma^2$.

True distribution sometimes not linear

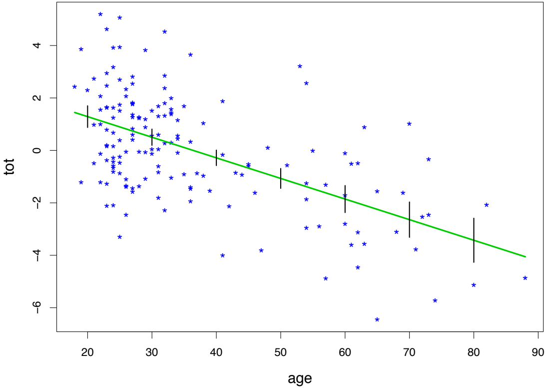


Figure 1.1 Kidney fitness `tot` vs `age` for 157 volunteers. The line is a linear regression fit, showing ± 2 standard errors at selected values of `age`.

Beyond simple linear regression

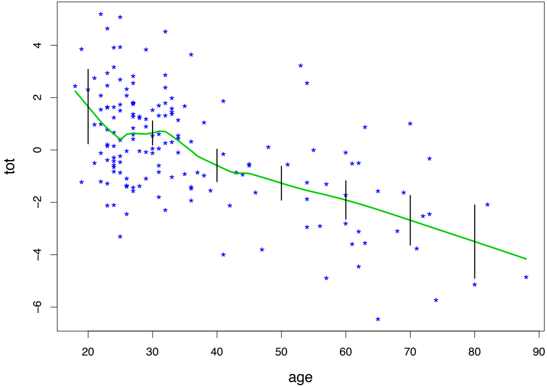


Figure 1.2 Local polynomial `lowess(x,y,1/3)` fit to the kidney-fitness data, with ± 2 bootstrap standard deviations.

Bootstrapped Models

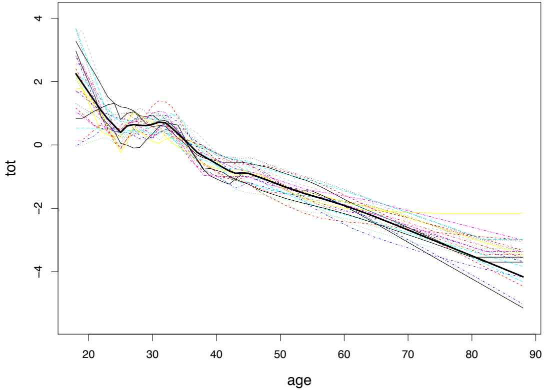


Figure 1.3 25 bootstrap replications of `lowess(x,y,1/3)`.