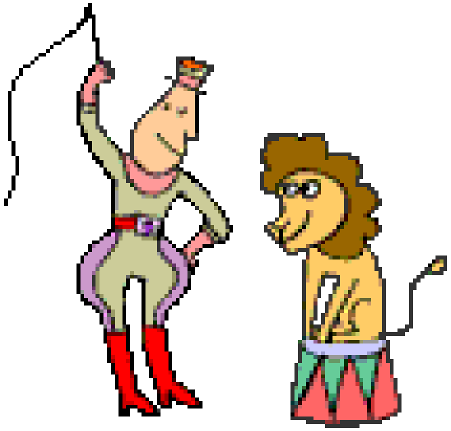
Taming Statistics with TamStat



Stephen Mansour, PhD
University of Scranton

Dyalog 18 Belfast, October 29, 2018

Overview

- TamStat framework
- Descriptive statistics including graphs, tables and summary functions
- Discrete and continuous probability distributions using the probability, criticalValue theoretical and randomVariable operators
- Regression models
- Inferential statistics using the confInt, sampleSize and hypothesis operators

Standards for naming variables, functions and operators

- Variables and namespaces always begin with a capital letter
  - e.g. Height, SEX, D.State
- TamStat functions and operators always begin with a lower-case character:
  - e.g. mean, randomVariable

Data representation

- Raw Data
  - Numeric vector
  - Character
    - Vector of character vectors
    - Comma delimited vector
    - Character matrix
- Frequency form – 2-column Matrix
  - 1st column: Value or midpoint
  - 2nd Column: integer
- Probability form – 2 – column Matrix
  - 1st column: Value or midpoint
  - 2nd Column: fraction
- Summary form – Namespace
  - count, mean, sdev

Database

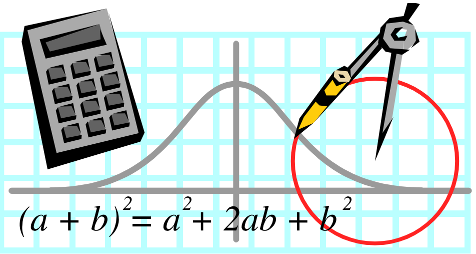
- A database is a namespace containing numeric and character data.
- Each variable must be all numeric or all character.
- Each variable must have the same length.
- A .csv file containing names in the first row and values in the succeeding rows can be imported as a database
- `D←import ''`
- `Variables D`
- `D.Height`

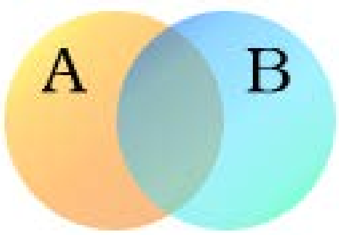
Exercise

- Import the Student Database
- Display a list of student heights
- Create a frequency distribution of heights
- Generate a histogram and a box plot
- Find the sample size, mean and standard deviation of each
- Create a summary namespace using the sample size (count), mean and standard deviation

Statistics deals primarily with four types of functions:

- Summary Functions
  - Descriptive Statistics
- Probability Distributions
  - Theoretical Models
- Relations
- Logic





Summary Functions

- Summary functions are of the form:

  $$y = f(x_1, x_2, \dots x_n)$$

- They produce a single value from a vector; similar to +/ (but not on higher order arrays)
- A statistic is a summary function of a sample; a parameter is a summary function of a population.
- Summary functions are all structurally equivalent
- Example: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Examples of Summary Functions

- Measures of Quantity
  - count, sum, sumSquares
- Measures of Center
  - mean, median, mode
- Measures of Spread
  - range, variance, sdev, iqr
- Measures of Position
  - percentile, quartile, percentileRange, zscore
- Measures of Shape
  - skewness, kurtosis

Probability Distributions

- Two types of distributions
  - Discrete
  - Continuous
- Discrete distributions are defined by the probability mass function
- Continuous distributions are defined by the density function
- The right argument is a Value
- The left argument is a parameter list

Discrete Distributions

- A B uniform X
- N P binomial X
- P geometric X
- N P negativeBinomial X
- M poisson X
- K M N hyperGeometric X

Continuous Distributions

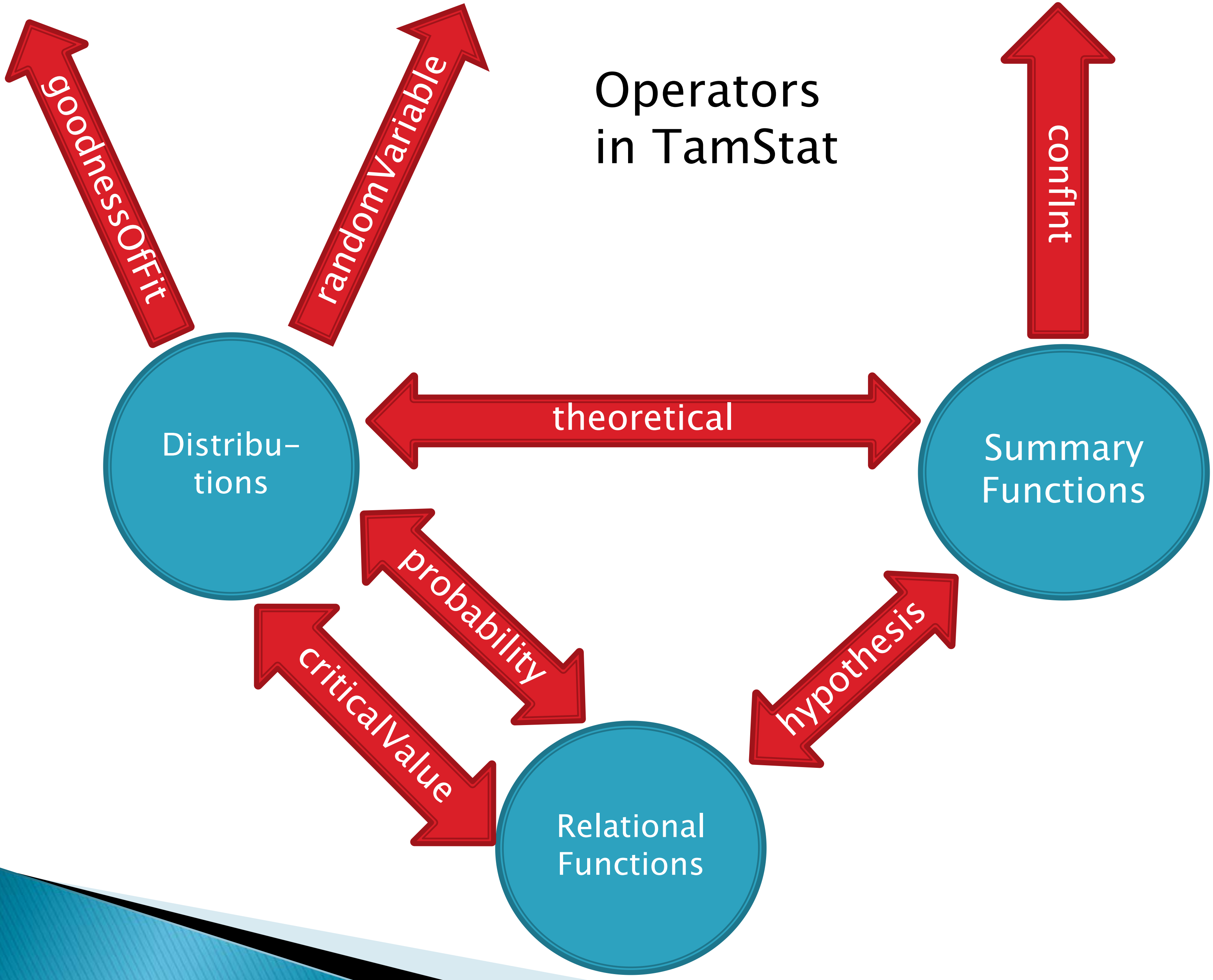
- A B rectangular X
- M exponential X
- M S normal X
- D chiSquare X
- D tDist X
- D1 D2 fDist X
- A M B triangular X
- M S logNormal X
- M S weibull X

~

Relational and Logical Functions

- Relational functions follow the usual definitions in APL
  - <, ≤, =, ≥, >, ≠, ϵ
- Additional relational functions include:
  - between, outside
- Logical functions also follow the usual definitions: ∨ ∧ ~ given

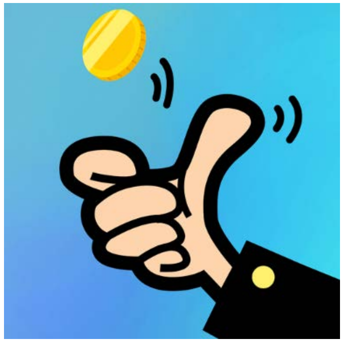
Operators in TamStat



Summary functions

- Using the student database, find the average height.
- Find a 95% confidence interval for the height
- Find a 99% confidence interval for the height
- Using the student database, find the proportion of students who are male.
- Find a 90% confidence interval for the proportion of male students.

Let’s look at an example:



What is the probability that you get at least 3 heads in seven coin tosses?

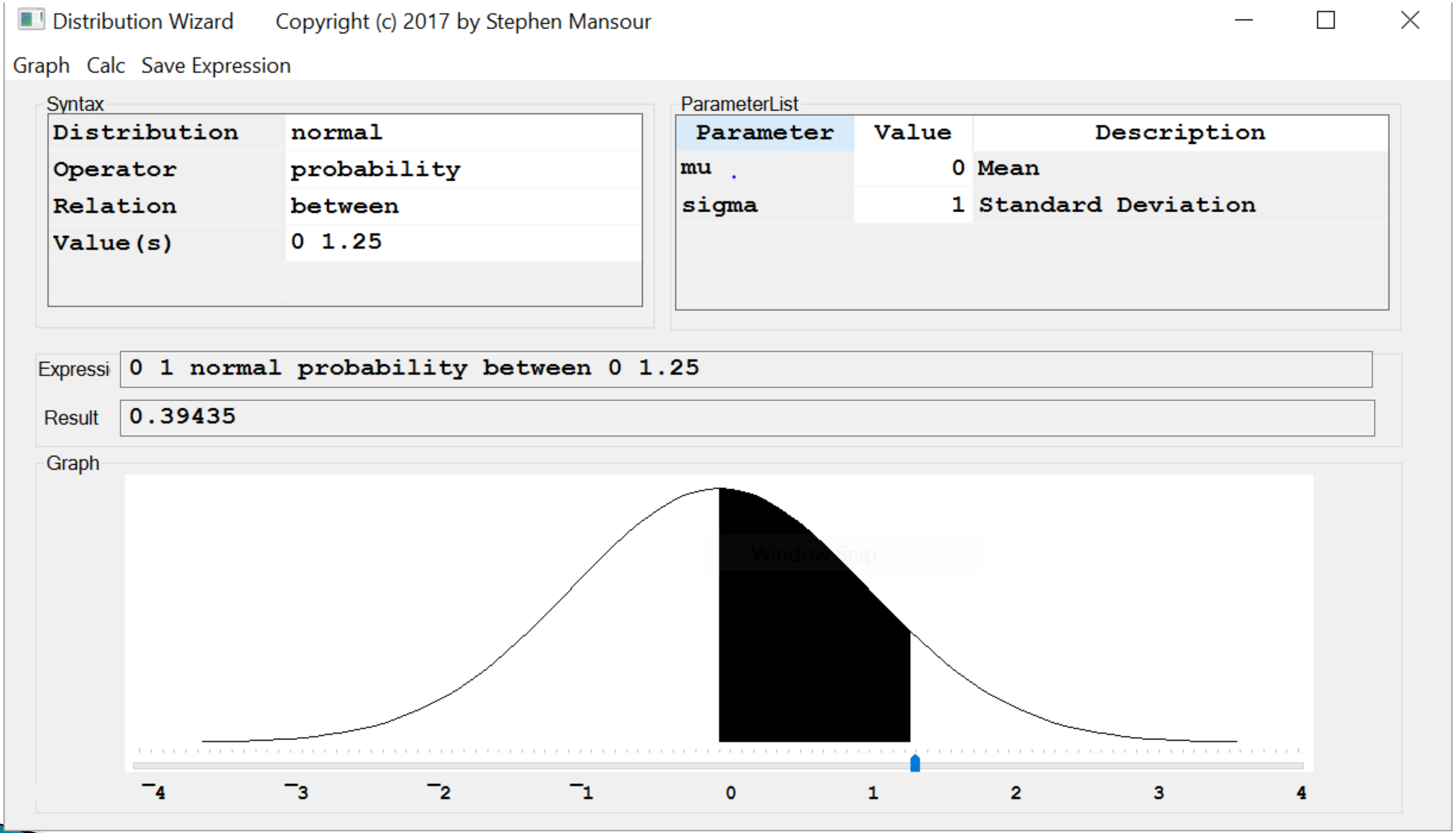
R: `pbinom(2,7,0.5,lower.tail=FALSE)`

APL/TamStat:

```
7 0.5 binomial probability    ≥    3
----- --------- ------------  -    -

   ↓     ↓           ↓        ↓      ↓
Left   Left      Operator    Right Right
 Arg   Operand               Oper  Arg
```

Distribution Wizard – Continous



Distribution	normal
Operator	probability
Relation	between
Value(s)	0 1.25

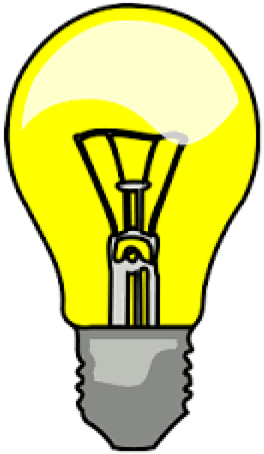
Parameter	Value	Description
mu	0	Mean
sigma	1	Standard Deviation

Expressi: 0 1 normal probability between 0 1.25

Result: 0.39435

A "Real–World" Reliability Example

- The failure rate for lightbulbs is 0.2% per hour.
- What is the mean time to fail?
- What is the probability that a lightbulb will last at least 750 hours?
- After how many hours will 90% of all light bulbs burn out?



Simulation

Generate random data from any distribution

Dyalog generates data from:

- Uniform (Discrete): ?N
- Rectangular(0,1) Continuous: ?0

TamStat generates random data from all other distributions including normal, binomial, hypergeometric, etc.

Simulation Problem

- You own an apartment house consisting of 40 flats.
- Each flat rents for £500 per month.
- Demand follows a discrete uniform distribution between 30 and 40 units.
- Your monthly expenses average £15000 per month with a standard deviation of £3000.
  - What is your expected profit?
  - What is the standard deviation?
  - What is the probability that you lose money?

Newsvendor problem

- A newsstand can buy newspapers for £1.50 and sell them for £2.00. Demand follows a poisson distribution with a mean of 35. How many newspapers should the owner of the newsstand purchase to maximize profit?
- $\Pi = E\left[p \min(q, D)\right] - cq$

where $\Pi =$ profit

$p =$ unit price $\quad c =$ unit cost

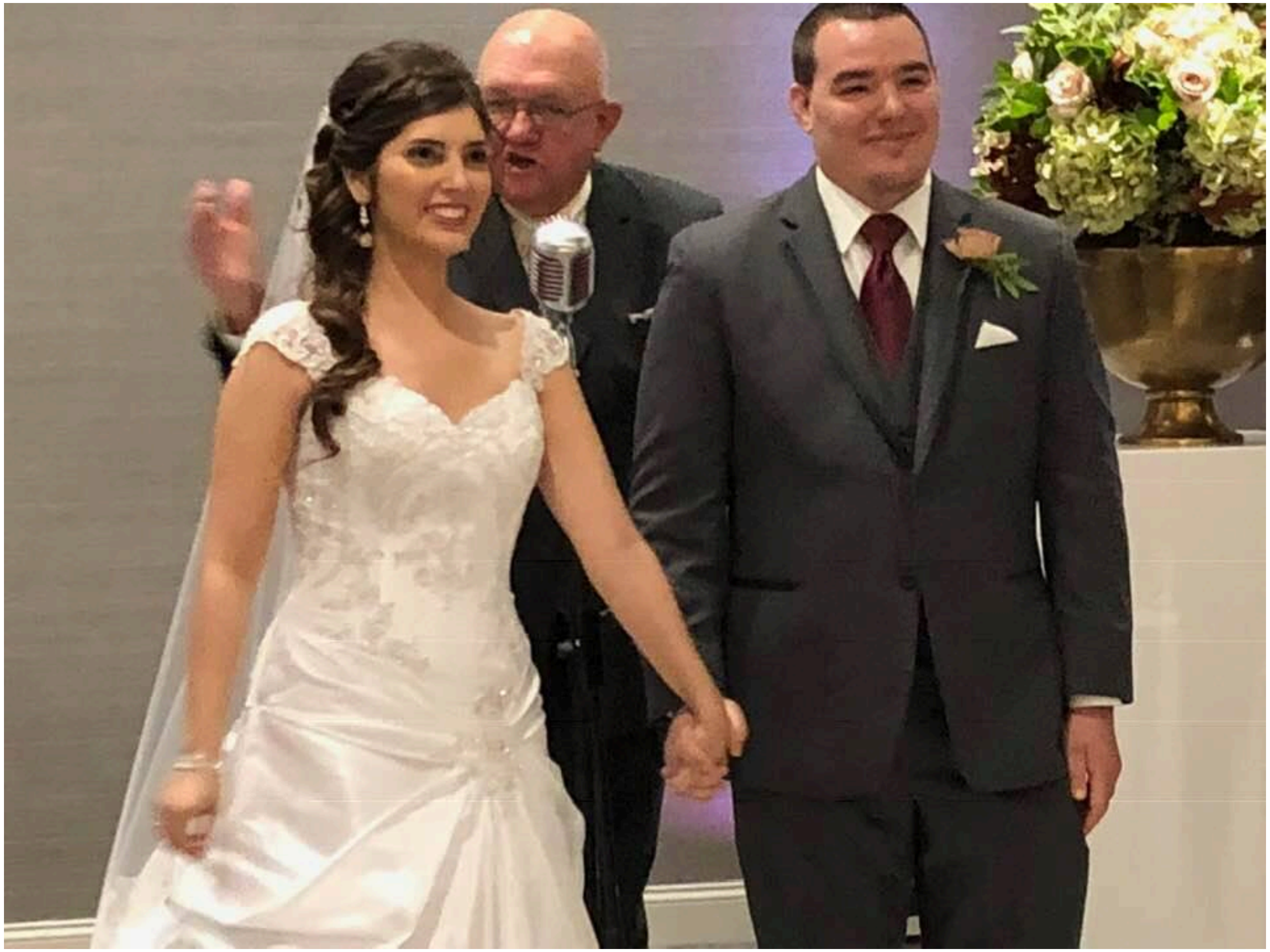
$q =$ quantity ordered $\quad D =$ demand

Inferential Statistics

- Confidence Intervals
  - Average height – point estimate, probably wrong
  - Height is somewhere between A and B

- Hypothesis tests
  - I think average height is x
  - Do the data support this?

Planning a Wedding



Planning a Wedding

- You are planning a wedding. Costs are
  - \$500 to rent the hall
  - \$100 per guest

1. You have 35 guests. What is the final cost?
2. You have a budget of \$8000 . How many guests can you invite?
3. Suppose the reception hall charges \$3000 for 25 guests and \$5500 for 50 guests. What are the fixed and variable costs?

Model:

$$f(x) = b_0 + b_1 x$$

$$f(x) = 500 + 100x$$

1. f(35) = \$4000

Arithmetic: $y = f(x)$

2. $f^{-1}(8000) = 75$

Algebra: $y = f(x)$

3. $3000 = b_0 + b_1 25$

$$5500 = b_0 + b_1 50$$

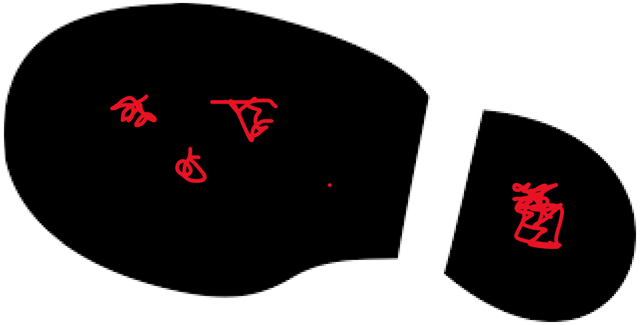
$$b_0 = 500 \quad b_1 = 100$$

3 or more equations: best fit

Regression: $y = f(x)$

CSI Scranton

You are investigating a murder. You find a bloody footprint size 9-1/2 near the body. What is the height of the suspect? If the suspect was known to be male, would that change anything?



Regression

- Draw a Scatter Plot
- Find the correlation between ShoeSize and Height
- Create a regression model
- Predict the height using MODEL.f
- Create a confidence interval
- Create a prediction interval
- Add D.Sex eq ‘M’
- Repeat the process

Regression

```
    D←import’’    ⍝ Import database as namespace
    D.Height      ⍝ Vector of Heights
    D.ShoeSize    ⍝ Vector of ShoeSizes
    MODEL←regress D.Height D.ShoeSize  ⍝ Simple Regression
    MODEL.B       ⍝ Intercept and Slope
50.77060572 1.771435553
       MODEL.RSq
68.37440979

 MODEL.
      MODEL.f 9.5 1
68.54922102
      MODEL.RSq
      MODEL.f confInt 9.5 1
67.45313462 69.64530743
      MODEL.f predInt 9.5 1
63.62800866 73.47043339
      .99 MODEL.f confInt 9.5 1
67.0785966 70.01984545
      .99 MODEL.f predInt 9.5 1
61.94640662 75.15203542
```

Hypothesis Test

- Using the student database, test the hypothesis that the average height is > 69 inches.

```
    report  D.Height mean hypothesis > 69
```

- Test the hypothesis that the percentage of students from Pennsylvania = 30%

```
H←(D.State eq 'PA') proportion hypothesis = .3

report H
```

Stephen M. Mansour, Ph.D.

- Adjunct Professor

  Operations and Information Management

  Kania School of Management
- Email: stephen.mansour@scranton.edu
- Website: www.tamstat.com

- Tel: (570)941-6278
- Address:

  University of Scranton
  Loyola Science Center 311D
  Monroe Ave and Linden St.
  Scranton, PA 18510