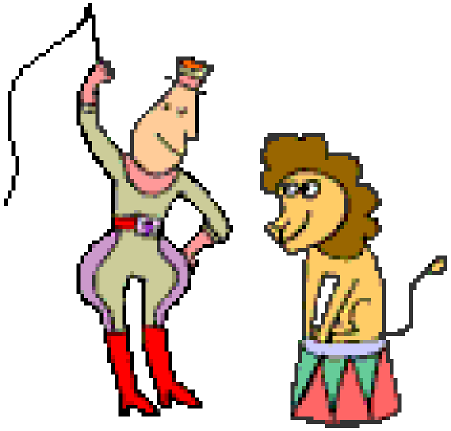
Taming Statistics with TamStat



Stephen Mansour, PhD
University of Scranton
Dyalog 18 Belfast, October 29, 2018

Statistical Tables are inconsistent

Z	0.01	0.02	0.03	0.04
0.0	0.500	0.504	0.508	0.512
0.1	0.540	0.544	0.548	0.552
0.2	0.579	0.583	0.587	0.591
0.3	0.618	0.622	0.626	0.629
0.4	0.655	0.659	0.663	0.666
0.5	0.691	0.695	0.698	0.702

Normal Table

D.F	.10	.05	.025	.01	.005
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.9	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17

Student t Table

Proliferation of Statistical Functions in Software

- Excel (4)
  - NORM.DIST,
  - NORM.INV,
  - NORMS.DIST,
  - NORMS.INV
- R(4)
  - dnorm,
  - pnorm,
  - qnorm,
  - rnorm
- TamStat(1)
  - normal

Normal Distribution

- Excel (6)

  T.DIST T.INV
  T.DIST.RT T.INV.2T
  T.DIST.2T T.TEST
- R(6)
  - dt
  - pt,
  - qt,
  - rt,
  - t.test
  - pairwise.t.test
- TamStat(1)
  - tDist

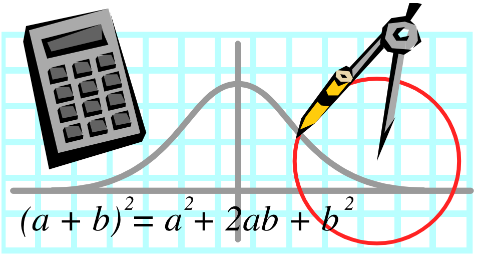
Student t Distribution

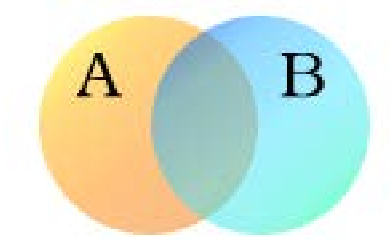
Data representation

- 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

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:

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

- 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

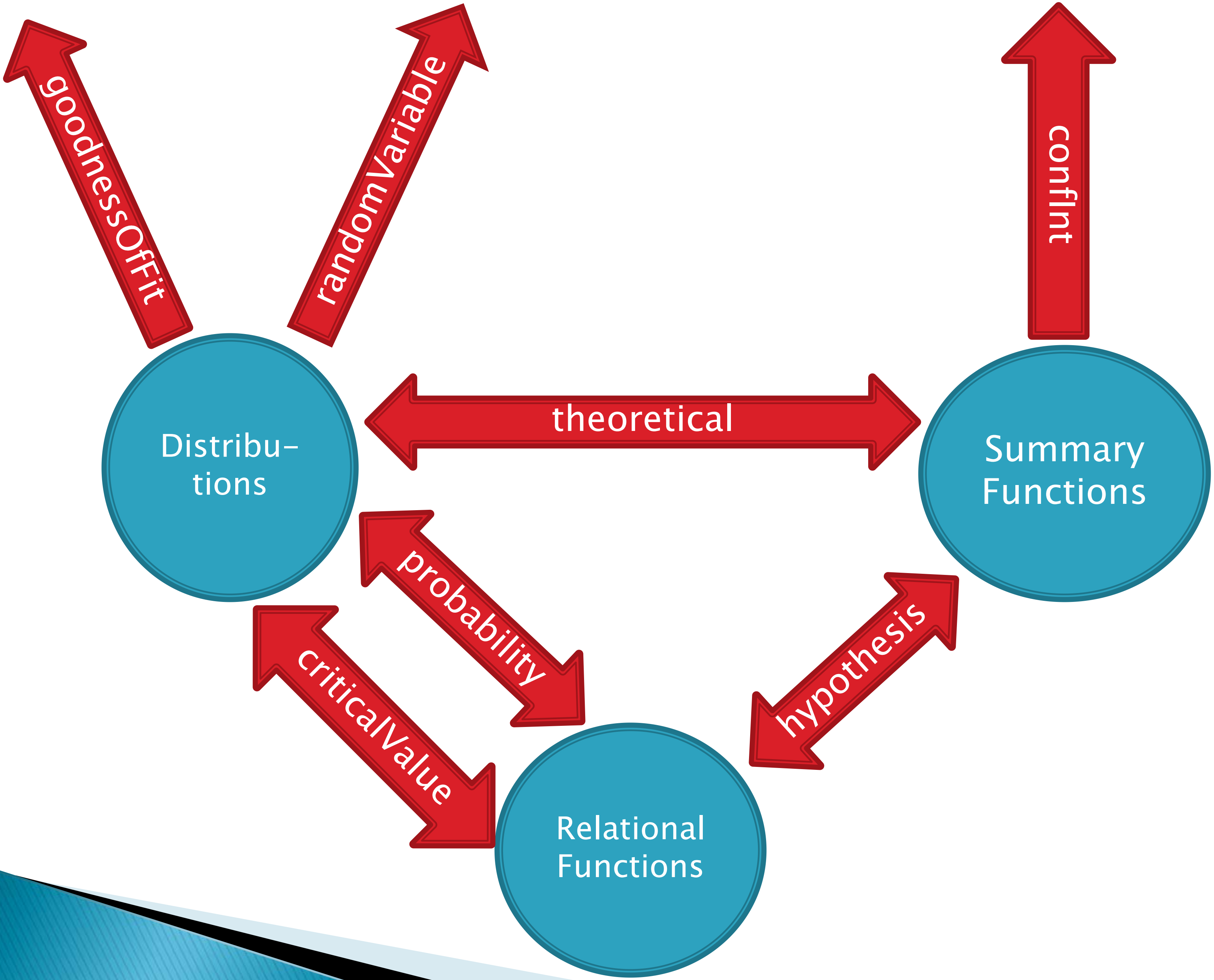
- Two types of distributions
  - Discrete
  - Continuous
- Discrete distributions are defined by the probability mass function
- Continuous distributions are defined by the density function
- The right argument is a Random Variable
- The left argument is a parameter list

Discrete Distributions

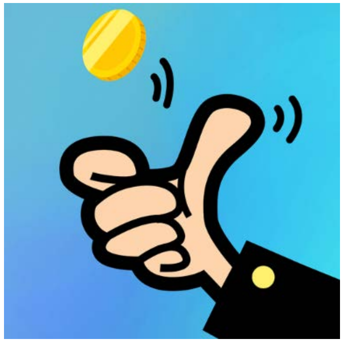
- A B uniform X
- N P binomial X
- P geometric X
- N P negativeBinomial X
- M poisson X
- K M N hyperGeometric X

Operators

- Operators modify or combine functions to do useful things.
- Some examples from mathematics:
- *Monadic:* *Dyadic:*
  - Derivative: $f'(x)$ Composition f∘g
  - Inverse $f^{-1}(x)$ Inner Product $\langle f, g \rangle$
- Using this concept, we define a probability operator to combine a distribution function with a relational function.


goodnessOfFit
randomVariable
confInt
Distribu-tions
theoretical
Summary Functions
probability
criticalValue
hypothesis
Relational Functions

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)`

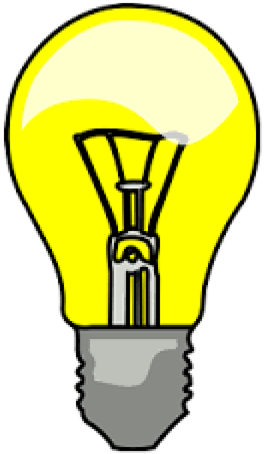
APL/TamStat:

```
7 0.5 binomial probability    >=    3
----- -------- -----------    --    -

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

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

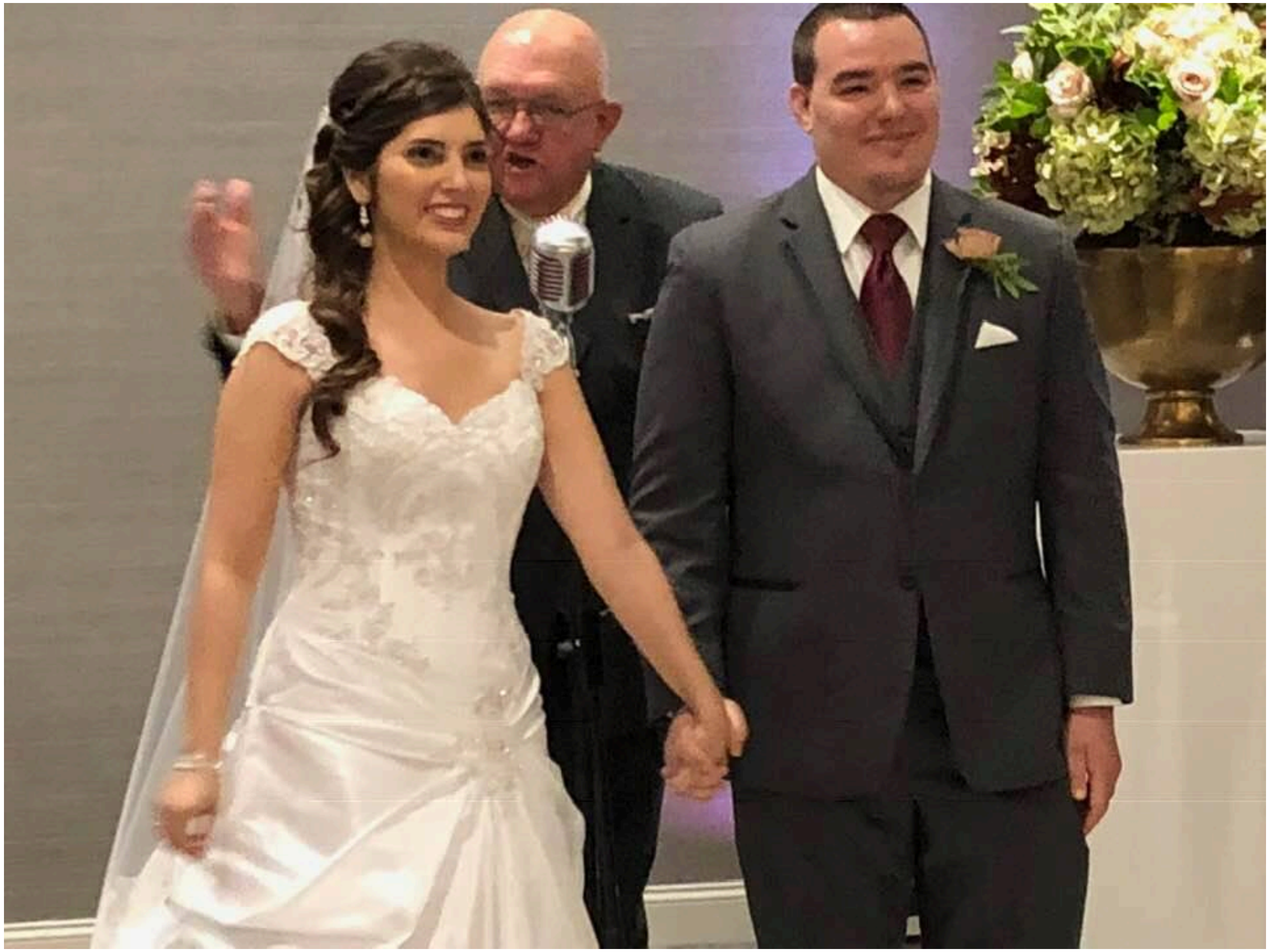
- Generate random data from any distribution
- Dyalog generates data from:
  - Uniform (Discrete): ?N
  - Rectangular(0,1) Continuous: ?0
- TamStat generates random data from all other distributions including normal, binomial, hypergeometric, etc.

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. $3000 = b_0 + b_1 25$

$5500 = b_0 + b_1 50$

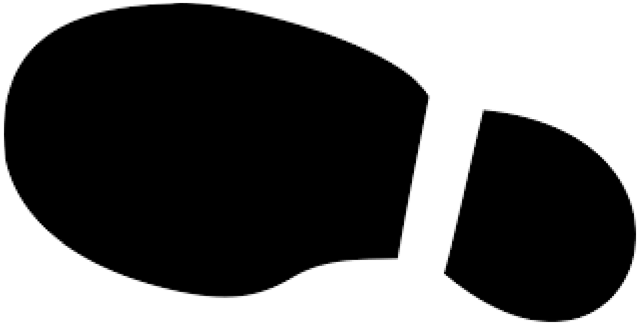
$b_0 = 500 \quad b_1 = 100$

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

```
    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
```

Weight Guesser

- The weight guesser at the county fair will give away a prize if his guess is more than 10 lbs. away from the customer's true weight.

- He observes that the customer's height is 6 feet and that his shoe size is 10-1/2. What is his best guess for the customer's weight?

Graphical User Interface

- Primarily for students of statistics
- Not designed for APL users
- Expression Builders
  - Summary Wizard
  - Distribution Wizard
  - Regression Wizard

Conclusion

- This is more about design and syntax, and less about implementation
- Most functions and operators can easily be written in APL.
- Internals not important to user
- R interface can be used if necessary for statistical calculations.
- Correct nomenclature and ease of use is critical.

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

THE UNIVERSITY OF
SCRANTON
A JESUIT UNIVERSITY
SCRANTON.EDU