# Discrete distributions

- Benford Distribution
- Bernouilli distribution
- Binomial distribution
- Hypergeometric distribution
- Geometric distribution
- Multinomial distribution
- Negative binomial distribution
- Poisson distribution
- Zipf's law

# Continuous distributions

- Beta and Dirichlet distributions
- Cauchy distribution
- Chi Square distribution
- Exponential distribution
- Fisher-Snedecor distribution
- Gamma distribution
- Levy distribution
- Log-normal distribution
- Normal and related distributions
- Pareto Distributions
- Student's t distribution
- Uniform distribution
- Weibull distribution
- Extreme values and related distribution
- Distribution in circular statistics

# Compendium of Probability Distributions

VOSE SOFTWARE

RISK SOFTWARE SPECIALISTS

ModelRisk 3.0

The most advanced RISK MODELING SOFTWARE in the world

A COMPENDIUM OF DISTRIBUTIONS

# The Exponential Distribution

- The exponential distribution is often used to model the waiting time X between events occurring randomly and independently in time (or space).
- Because it is a continuous distribution, the height of the exponential curve at any X refers to probability density rather than probability.
- Probability is instead represented by area under the exponential curve.

# The Exponential Distribution

- The main assumption of the exponential distribution is that at any point in time, the probability of an event occurring in the next instant does not depend on how much time has already elapsed since the previous event.
- The parameter of the exponential distribution is the rate at which events occur the number of events per unit time.
- The mean of the exponential distribution is $1/\text{rate}$.

# The Exponential Distribution

Waiting time X to the next event has probability density

```
dexp(X, rate=1)
dexp(X, rate=1, log=TRUE)
```

The default value for the rate is 1, so you must alter to fit your circumstance.

# Exponential Distribution : Problem

Suppose the mean checkout time of a supermarket cashier is three minutes. Find the probability of a customer checkout being completed by the cashier in less than two minutes, three minutes and four minutes. (i.e. what percentage of “waiting times” are less than two, three and four minutes?)

# The Exponential Distribution

**Solution**

- The checkout processing rate is equals to one divided by the mean checkout completion time.
- Hence the processing rate is $1/3$ checkouts per minute.
- We then apply the function `pexp()` of the exponential distribution with rate$=1/3$.

# The Exponential Distribution

```
> pexp(2,rate=1/3)
[1] 0.4865829
>
> pexp(3,rate=1/3)
[1] 0.6321206
>
> pexp(4,rate=1/3)
[1] 0.7364029
>
> pexp(5,rate=1/3,lower=FALSE)
[1] 0.1888756
```

# The Exponential Distribution

- What is the median waiting time? To answer this question we would use the `qexp()` function.
- Recall that the median is value of $x$ such that $P(X \leq x) = 0.50$.
- Also determine the first and third quartiles $Q_1$ and $Q_3$.

```
> qexp(0.5,rate=1/3)
[1] 2.079442
> qexp(0.25,rate=1/3)
[1] 0.8630462
> qexp(0.75,rate=1/3)
[1] 4.158883
```

# The Geometric Probability Distribution

- Used for binary data (1 or 0; success or failure, etc).
- In a sequence of trials, the trial yielding the first success (all previous trials ending in failure) has a geometric distribution.
- The assumption required is that separate trials are independent and the probability of success p is the same in every trial.
- The probability of failure in each trial is $1 - p$.

# The Geometric Probability Distribution

The variable X counts the number of **failures** before the first **success**. In other words, X $= 0$ if success occurs on the 1st trial. X $= 1$ means that the 1st trial ended in failure and success occurred in the 2nd trial. X $= 2$ means that the 1st and 2nd trial ended in failure and that success occurred in the 3rd trial. And so on. $Pr[X]$ is calculated as

```
# p is the probability of success
dgeom(X, prob = p)
# log of the probability instead
dgeom(X, prob = p, log=TRUE)
```

# The Geometric Probability Distribution

The geometric distribution - Example R code

```
> N <- 10000
> x <- rgeom(N, .5)
> x <- rgeom(N, .01)
```

# Pseudo-Random Number Generation

- Mersenne Twister
-

## Random Number Generation

- Random Number Generation
- The Mersenne Twister and Diehard tests
- RDieharder R package

# Diehard Tests

### The Diehard Tests

- The diehard tests are a battery of statistical tests for measuring the quality of a random number generator.
- They were developed by George Marsaglia over several years and first published in 1995 on a CD-ROM of random numbers.

# Diehard Tests

Birthday spacings: Choose random points on a large interval. The spacings between the points should be asymptotically exponentially distributed. The name is based on the birthday paradox.

Overlapping permutations: Analyze sequences of five consecutive random numbers. The 120 possible orderings should occur with statistically equal probability.

# Diehard Tests

Ranks of matrices: Select some number of bits from some number of random numbers to form a matrix over 0,1, then determine the rank of the matrix. Count the ranks.

Monkey tests: Treat sequences of some number of bits as "words". Count the overlapping words in a stream. The number of "words" that don't appear should follow a known distribution. The name is based on the *infinite monkey* theorem.

# Diehard Tests

Count the 1s: Count the 1 bits in each of either successive or chosen bytes.
Convert the counts to "letters", and count the occurrences of five-letter "words".

Parking lot test: Randomly place unit circles in a $100 \times 100$ square. If the circle overlaps an existing one, try again.
After 12,000 tries, the number of successfully "parked" circles should follow a certain normal distribution.

# Diehard Tests

Minimum distance test: Randomly place 8,000 points in a 10,000 x 10,000 square, then find the minimum distance between the pairs.
The square of this distance should be exponentially distributed with a certain mean.

Random spheres test: Randomly choose 4,000 points in a cube of edge 1,000.
Center a sphere on each point, whose radius is the minimum distance to another point.
The smallest sphere's volume should be exponentially distributed with a certain mean.

# Diehard Tests

The squeeze test: Multiply 231 by random floats on (0,1) until you reach 1. Repeat this 100,000 times. The number of floats needed to reach 1 should follow a certain distribution.

Overlapping sums test: Generate a long sequence of random floats on (0,1).
Add sequences of 100 consecutive floats.
The sums should be normally distributed with characteristic mean and sigma.

# Diehard Tests

Runs test: Generate a long sequence of random floats on (0,1). Count ascending and descending runs. The counts should follow a certain distribution.

The craps test: Play 200,000 games of craps, counting the wins and the number of throws per game.
Each count should follow a certain distribution.

# Benford's Law

- Benford's law, also called the **First-Digit Law**, refers to the frequency distribution of digits in many (but not all) real-life sources of data.
- In this distribution, 1 occurs as the leading digit about 30% of the time, while larger digits occur in that position less frequently: 9 as the first digit less than 5% of the time.
- Benford's law also concerns the expected distribution for digits beyond the first, which approach a uniform distribution.

# Benford's Law

- It has been shown that this result applies to a wide variety of data sets, including electricity bills, street addresses, stock prices, population numbers, death rates, lengths of rivers, physical and mathematical constants, and processes described by power laws (which are very common in nature).
- It tends to be most accurate when values are distributed across multiple orders of magnitude.

# Benford's Law

- The graph here shows Benford's law for base 10.
- There is a generalization of the law to numbers expressed in other bases (for example, base 16), and also a generalization from leading 1 digit to leading n digits.
- It is named after physicist Frank Benford, who stated it in 1938, although it had been previously stated by Simon Newcomb in 1881.

# Benford Distribution

```
> # The Benford Distribution is the distribution of
> # the first digit of a number.
>
> library(VGAM)
>
> dbenf(c(1:9))
[1] 0.30103000 0.17609126 0.12493874
[4] 0.09691001 0.07918125 0.06694679
[7] 0.05799195 0.05115252 0.04575749
```

# Zipf Distribution

George Zipf popularized an ideaZipfs Lawthat approximates populations of cities, distribution of money in counties, and how frequently words are used. Nobel Prize-winning columnist Paul Krugmans wrote of Zipfs Law that

> *the usual complaint about economic theory is that our models are oversimplified  that they offer excessively neat views of complex, messy reality. [In the case of Zipfs law] the reverse is true: we have complex, messy models, yet reality is startlingly neat and simple.*

A Zipfian Distribution: How Often Words Appear

# Zipf Distribution

- A Zipfian distribution is a type of power law. A power law occurs when one event varies as a power of another. One application of Zipfs law states that in texts of natural language (e.g., books), each word is used twice as often as the next most commonly occuring word.
- The graph below applies the rule to word usage in 29 UK books below. The occurred 225,300 uses, and was the most commonly used word. Note that the graph is interactive; you can press the play with this data link to edit, embed, and share your own version.

# Zipf Distribution

Evaluating Power Laws

- We can test for a power law by plotting frequency (y-axis) against rank (x-axis) on a double log axis. Then check for a straight line.
- The graph below shows three attempts to fit a power law function to datasets. The plot on the left is a good fit. The plot in the middle is a decent fit.
- The plot on the right is not a good fit.

# Zipf Distribution

**Evaluating Zipfian Distributions For City Populations**
Another application of Zipfs law is for populations. Weve used ggplot2 to graph the population of cities (y-axis) and the rank of each city. In this dataset, New York has the highest population and is ranked first.

# Zipf Distribution

**GDP Of Nations**
We are approaching a Zipfians distribution for country GDP vs rank.

# Zipf Distribution

**Evaluating Power Laws For Many Datasets**

Researchers use power laws to determine how much inftrasture a city needs, examine the number of gas stations required in a city, and much more.

# Extreme Valeu Theory

- The Extreme Value Theory (EVT) is extensively used for modelling very large and/or very small events. Usually the focus of the analysis is the estimation of very extreme quantiles or tail probabilities.
- It is widely used in several areas, such as environment, insurance, whether and hydrology. Several R packages have been developed for fitting models in this framework.

# Extreme Value Theory

Extreme Value Theory is mathematical study of extreme values. Extreme value theory is a branch of statistics dealing with the extreme deviations from the median of probability distributions. The general theory sets out to assess the type of probability distributions generated by processes. Extreme value theory is important for assessing risk for highly unusual events, such as 100-year floods.

# Extreme Value Theory

- The field of extreme value theory was pioneered by Leonard Tippett (1902 – 1985).
- Tippett was employed by the British Cotton Industry Research Association, where he worked to make cotton thread stronger. In his studies, he realized that the strength of a thread was controlled by the strength of its weakest fibers.
- With the help of R. A. Fisher, Tippet obtained three asymptotic limits describing the distributions of extremes. The German mathematician Emil Julius Gumbel codified this theory in his 1958 book **Statistics of Extremes**, including the Gumbel distributions that bear his name.

# EVT: Frequency Analysis

- Frequency Analysis (*FA*): Probabilistic description of hydrological extremes
- Extraction of extreme variables from data
- Choice of a distribution
- Parameter estimation
- Quantiles + uncertainty

# Extreme Value Distribution

Extreme Value. The extreme value distribution is often used to model extreme events, such as the size of floods, gust velocities encountered by airplanes, maxima of stock market indices over a given year, etc.; it is also often used in reliability testing, for example in order to represent the distribution of failure times for electric circuits (see Hahn and Shapiro, 1967). The extreme value (Type I) distribution has the probability density function:

$$f(x) = 1/b \times e^{[-(x-a)/b]} \times e^{-e^{[-(x-a)/b]}}, for - 8 < x < 8, b > 0$$

where

a is the location parameter b is the scale parameter e is the base of the natural logarithm, sometimes called Euler's e (2.71...)

# Probability Distribution used for Extreme Value Theory

- Weibull law:

$$G(z) = \begin{cases} \exp\left\{-\left(-\left(\frac{z-b}{a}\right)\right)^{\alpha}\right\} & z < b \\ 1 & z \geq b \end{cases}$$

when the distribution of $M_n$ has a light tail with finite upper bound. Also known as Type 3.

- Gumbel law:

$$G(z) = \exp\left\{-\exp\left(-\left(\frac{z-b}{a}\right)\right)\right\} \text{ for } z \in \mathbb{R}.$$

when the distribution of $M_n$ has an exponential tail. Also known as Type 1

- Frchet Law:

$$G(z) = \begin{cases} 0 & z \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\} & z > b. \end{cases}$$

when the distribution of $M_n$ has a heavy tail (including polynomial decay). Also known as Type 2.

# Probability Distribution used for Extreme Value Theory

- The GEV distribution
- The Pareto distribution

# The Generalized Extreme Value (GEV) Distribution

- The generalized extreme value (GEV) distribution is a family of continuous probability distributions developed within extreme value theory to combine the Gumbel, Frchet and Weibull families also known as type I, II and III extreme value distributions.

By the extreme value theorem the GEV distribution is the limit distribution of properly normalized maxima of a sequence of independent and identically distributed random variables. Because of this, the GEV distribution is used as an approximation to model the maxima of long (finite) sequences of random variables.

In some fields of application the generalized extreme value distribution is known as the FisherTippett distribution, named after R. A. Fisher and L. H. C. Tippett. However usage of this name is sometimes restricted to mean the special case of the Gumbel distribution.

# The Pareto Distribution

In hydrology the Pareto distribution is applied to extreme events such as annually maximum one-day rainfalls and river discharges. The blue picture illustrates an example of fitting the Pareto distribution to ranked annually maximum one-day rainfalls showing also the 90% confidence belt based on the binomial distribution. The rainfall data are represented by plotting positions as part of the **cumulative frequency analysis**.

# The Gumbel Distribution

The extreme value type I distribution has two forms. One is based on the smallest extreme and the other is based on the largest extreme. We call these the minimum and maximum cases, respectively. Formulas and plots for both cases are given. The extreme value type I distribution is also referred to as the Gumbel distribution.

evd evd (extreme value distributions) is and add-on package for the R system. It extends simulation, distribution, quantile and density functions to univariate and multivariate parametric extreme value distributions, and provides fitting functions which calculate maximum likelihood estimates for univariate and bivariate models, and for univariate and bivariate threshold models.

evdbayes evdbayes is and add-on package for the R system. It provides functions for the bayesian analysis of extreme value models, using MCMC methods.

evir

evir is and add-on package for the R system. It is an R port (conversion) of Version 3 of Alexander McNeil's S library EVIS (Extreme Values in S). It contains functions for extreme value theory, which may be divided into the following groups; exploratory data analysis, block maxima, peaks over thresholds (univariate and bivariate), point processes, gev/gpd distributions.

# EVIR

Functions for extreme value theory, which may be divided into the following groups;

- exploratory data analysis,
- block maxima,
- peaks over thresholds (univariate and bivariate),
- point processes,
- gev/gpd distributions.

# Ocmulgee data set

The ocmulgee data frame has 40 rows and 2 columns. The columns contain maximum annual flood discharges, in units of 1000 cubed feet per second, from the Ocmulgee River in Georgia, USA at Hawkinsville (upstream) and Macon (downstream), for the years 1910 to 1949. The row names give the years of observation.

# ismev

ismev is and add-on package for the R system. It is an R port of S functions written by Stuart Coles to support (univariate) extreme value modelling (or modeling, if you're on the other side of the pond), including the computations carried out in Coles (2001). The functions may be divided into the following groups; maxima/minima, order statistics, peaks over thresholds and point processes. Coles (2001) is a textbook that provides an introduction to the topic at a relatively simple statistical level.

# extRemes

extRemes is and add-on package for the R system, written by Eric Gilleland, Rick Katz and Greg Young, and maintained by Eric Gilleland. It provides a windows GUI for the ismev package, allowing easy use of the tools that the ismev package provides, in addition to a few extra useful functions. See the extRemes web site for details on the package and version updates.

# extremevalues

Software package for detecting extreme values in one-dimensional data. According to the description section of the help file for the package (version 2.1), the package implements outlier detection methods introduced in a discussion paper by the package author, Mark van der Loo (M.P.J. van der Loo, Discussion paper 10003, Statistics Netherlands, The Hague, 2010; available at http://www.cbs.nl).

# fExtremes

fExtremes is and add-on package for the R system, maintained and primarily written by Diethelm Wuertz. The package contains functions for the exploratory data analysis of extreme values for insurance, economic and financial applications. It also brings together many of the elements of the packages evd, evir and ismev. The fExtremes package comprises part of the Rmetrics software collection. See the Rmetrics web site for details.

BSquare

BSquare is a package that models the quantile function using splines for non extremes and the generalized Pareto distribution for extremes. For more information, see its web page.

copula

copula is a package containing functions for exploring and modeling several commonly used copulas. MLE, pseudo-MLE and method of moments are all avialable. It does not allow for non-stationary regression, but does allow for multivariate modeling (as you would expect).

## `lmom` and `lmomRFA`

lmom Functions related to L-moments, includes functions to compute L-moment estimates for extreme value distribution parameters.

lmomRFA Package written and maintained by J.R.M. Hosking containing functions for regional frequency analysis (RFAA) using the methods of Hosking and Wallis (1997) Regional frequency analysis: an approach based on L-moments, Cambridge University Press (newest edition, 2005; 244 pp., ISBN-10: 0521019400).

# lmomco

Package written by William H. Asquith to compute L-moments, trimmed L-momes, L-comoments, and probability-weighted moment estimation for many distributions including extreme value distributions.

# POT

Functions written by Mathieu Ribatet for performing peak over threshold (POT) analysis for both univariate and bivariate cases.

# The SpatialExtremes Package

A new package by Mathieu Ribatet for performing multivariate and other spatial extremes methods, called SpatialExtremes, is now available via CRAN. It is still in a development stage, but check its web site for further information at http://spatialextremes.r-forge.r-project.org/

# texmex

A relatively new package with functions for performing the conditional EVA approach introduced in Heffernan and Tawn (2004), J. R. Statist. Soc. B, 66 (3), 497 - 546. Also contains good functions for fitting the GP df to data.

# VGAM

VGAM is a package for fitting vector generalized additive models. That is, it allows for modeling parameters as linear or smooth functions of covariates.