

A Primer on Stochastic Process Analysis

Christian Breunig
University of Toronto
Department of Political Science

Presentation at the APSA Annual Meeting,
Seattle, August 31, 2011

Goals

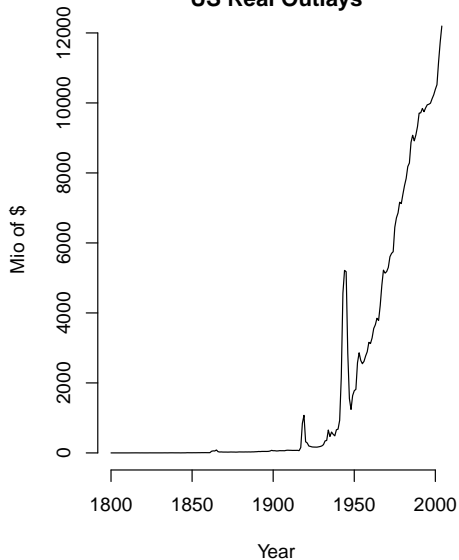
- survey important statistical tools used to assess data in situations where the entire distribution of values is of interest
- outline three broad conditions under which stochastic process methods are applicable
- discuss a variety of visual and analytical techniques, including kurtosis scores, distributional analysis, direct parameter estimates of probability density functions
- use US budget use as an example



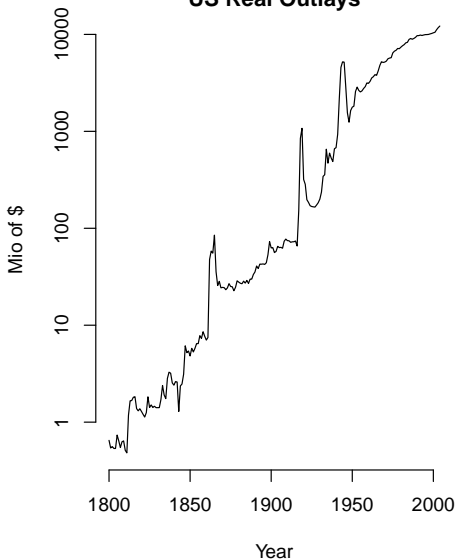
<http://individual.utoronto.ca/cbreunig/BreunigJones-2011-howto.zip>

Typical Agendas Data

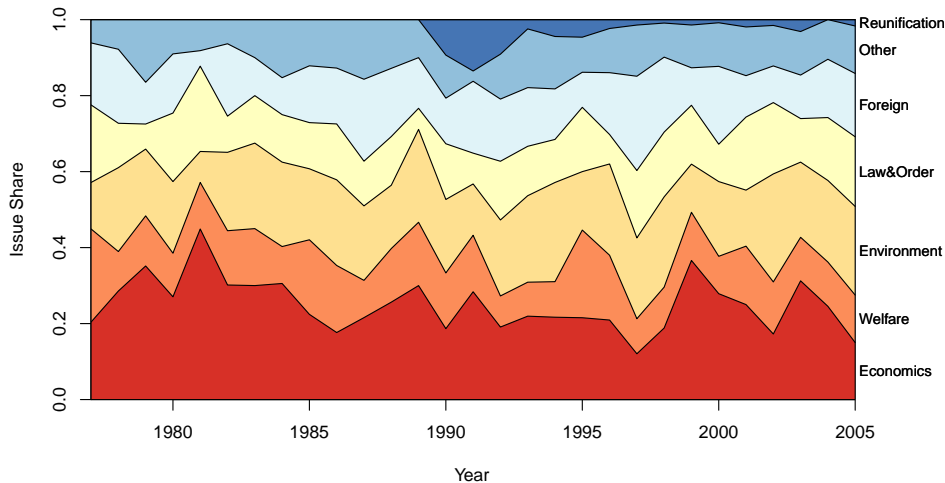
US Real Outlays



US Real Outlays



Typical Agendas Data

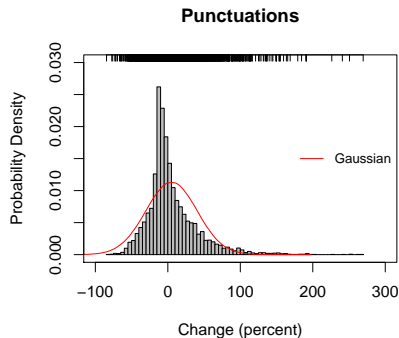
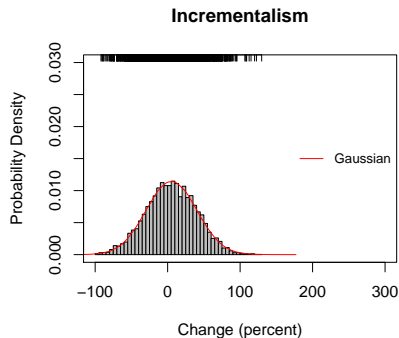


Stochastic Process Methods

- examine the full distribution of outcomes
- goal: map theoretical distributions to data
- empirical implication from theory but also good exploratory tool
- applicable when modelling not possible or specification not clear
 - 1 specification not possible due to complexity and uncertainty of the process
 - 2 establish broad empirical generalizations in situations with scarce prior knowledge
 - 3 alternative to regression when key measures are missing

Example: Budgets and Punctuations

- budgeting as stochastic process (Padgett 1980) and Jones and Baumgartner (2005) for policy
- **Incrementalism**: if decision-making is based on rational updating, budgets are Normally distributed.
- **Punctuations**: if decision-making is based on boundedly rational updating and institutional friction, budgets follow a power law.



First steps

- calculate policy change

- percentage-count method:

$$(count_t - count_{t-1}) / count_{t-1}$$

- percentage-percentage method:

$$(percentage_t - percentage_{t-1}) / percentage_{t-1}$$

Table: US Budget outlays

year	outlays
1800	0.649
1801	0.541
1802	0.560
1803	0.535
1804	0.535
⋮	⋮

Kurtosis

- assesses the “peakedness” of a distribution
- measures the “degree” of punctuation
- general kurtosis is sensitive to outliers
- alternative: L-kurtosis (Hosking 1989)

* In case any of my readers may be unfamiliar with the term “kurtosis” we may define mesokurtic as “having β_2 equal to 3,” while platykurtic curves have $\beta_2 < 3$ and leptokurtic > 3 . The important property which follows from this is that platykurtic curves have shorter “tails” than the



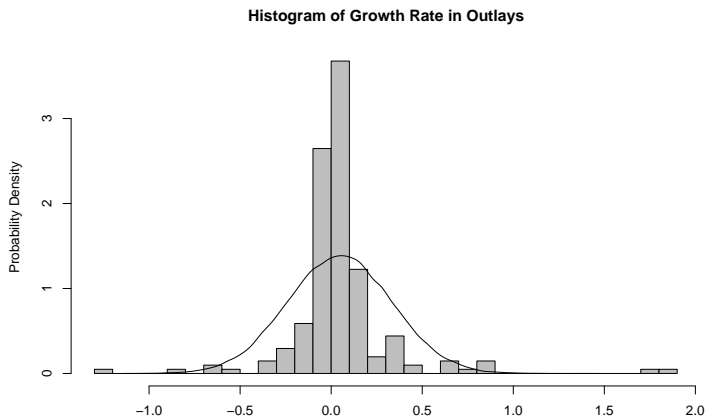
normal curve of error and leptokurtic longer “tails.” I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for “lepping,” though, perhaps, with equal reason they should be hares!

Table: US Budget outlays

Statistic	Value
Mean	0.05
Median	0.02
Variance	0.08
IQR	0.12
Skewness	1.92
Kurtosis	14.64
L-kurtosis	0.46
Min	-1.22
Max	1.85

Histograms and Goodness-of-Fit Tests

- simple visual tool for assessing distribution
- important to select bin width appropriately (minimizing the trade-off between bias and variance)
- add a normal distribution for easy comparison
- Kolmogorov-Smirnov (K-S) test assesses whether the probability distribution of the data comes from a specified distribution



Distributional Analysis

Background

Goal: identify whether the budget data follows an underlying probability law

- transform probability density function (pdf) into a linear function
- ⇒ rely on visual inspection and regression in order to judge data vs. hypothesized distribution
- pdf relates a range of the values of a variable of interest to the probability associated with that range

$$Pr(a < x < b) = \int_a^b f(x) dx$$

- exponential pdf

$$Pr(x) = \alpha \exp^{-\beta x} \Rightarrow \ln(Pr(x)) = \ln(\alpha) - \beta x$$

- Paretian pdf (power law)

$$Pr(x) = \alpha x^{-\beta} \Rightarrow \ln(Pr(x)) = \ln(\alpha) - \beta \ln(x)$$

Distributional Analysis

Scatter Plots and Direct Parameter Estimates

- in practice, rely on cumulative frequencies of the distribution in order to account for tail behavior better
- plot logarithm of cumulative frequencies vs. value of category midpoints (semi-log plot) or vs. logarithm value of category midpoints (log-log plot)
- visually assess if the data falls along a straight line with a scatter plot
- estimate the parameters associated with the plots and compare goodness-of-fit
- caution: fairly easy to confuse similar distributions (Clauset et al 2007)

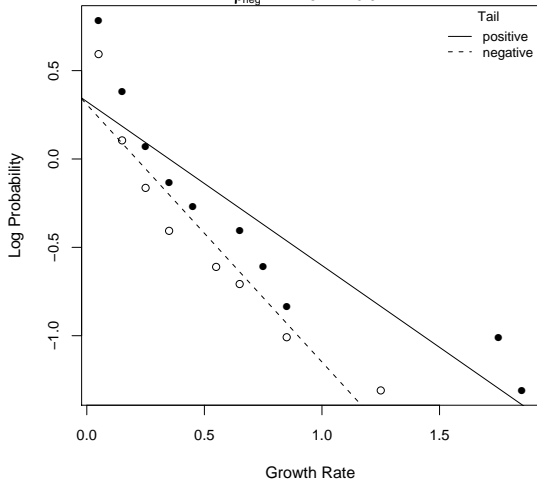
Distributional Analysis

Semi-log Plot

Semi-Log Plot

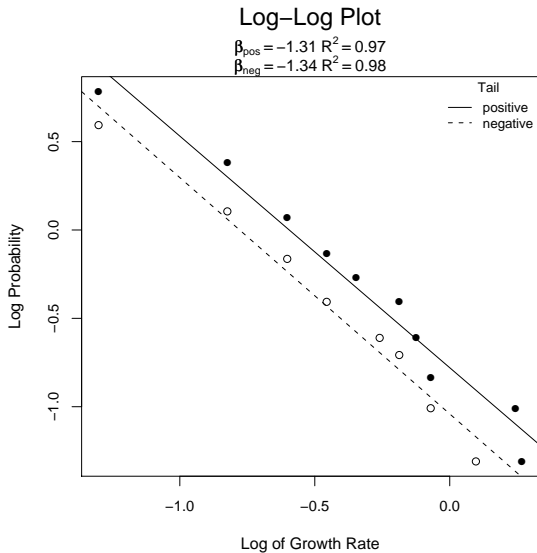
$$\beta_{\text{pos}} = -0.93 \quad R^2 = 0.83$$

$$\beta_{\text{neg}} = -1.46 \quad R^2 = 0.91$$



Distributional Analysis

Log-log Plot



Summary and Conclusion

- benefits of stochastic process methods
- kurtosis and histogram
- distributional analysis
- widely applicable, esp. for study of dynamic systems

Bonus Slide: Zero Counts

Q: How do I compute percent changes when there was no activity (i.e. zero counts)?

A: follow simple arithmetics (and recode as missing where necessary)

- do not add 1 to all agenda items (King and Zeng)

Table: Toy Agenda

year	topic	count	total	share
⋮	⋮	⋮	⋮	⋮
1987	400	16	28	0.571
1988	400	5	26	0.192
1989	400	0	2	0.000
1990	400	0	4	0.000
1991	400	0	7	0.000
1992	400	0	6	0.000
1993	400	18	24	0.750
1994	400	14	37	0.378