


The analysis of event rates using intervals

Jim Lemon , a

^a Statistical consultant, Australia

Abstract ■ Event interval analysis had historical antecedents in the past century, but the analysis of rates of events has been largely performed using counts of events. When the information content of intervals and counts of the same events are compared, it is clear that the information content of counts is sensitive to the number of events in a counting interval. The reduced information content of counts where the number of events in a counting interval is small may affect the analysis of event rates. Both simulated and historical data are used to illustrate such effects. It is concluded that event interval analysis may be more appropriate for the analysis of event rates when the events in question are few in the counting intervals.

Keywords ■ Uncommon events, information content, event history analysis, statistical model

 jim@bitwrit.com.au

Introduction

The rate at which specified events occur is often of interest to researchers. Whether this concerns the number of fatal road crashes in Iceland (Directorate of Health - Iceland, 2013) or the number of people who perish attempting to climb Mount Everest (Wikipedia, 2013), both the estimate of rate over time and changes in the rate are often studied. Rates of events are usually expressed in frequency per unit time, and the typical method of statistical analysis of rates uses such counts as the basic data. As events are often recorded as the time or date at which the event occurred, it is possible to use the intervals between events to estimate rates, and this method achieved some popularity around the middle of the last century (Maguire, Pearson & Wynn, 1952). Event history analysis, as it was then known, seemed to present a viable method for the analysis of event rates, but has received little attention since that time.

The Sequential Event Model

A common model for studying the rate of certain events specifies that the occurrence of each event is independent of other such events and that the distribution for all events is the same. In the paper cited above (Maguire, Pearson & Wynn, 1952), fatal industrial accidents resulting in a specified minimum number of deaths were studied. Each accident was independent and assumed to be distributed uniformly in time. The temporal resolution of the occurrence of these accidents was such that no inter-accident intervals of zero were encountered. Sequences of events appropriate for analysis by intervals should

consist in uniformly distributed independent events with no two events occurring at the same time. In practice, even if two or more events occur within the same time increment, typically a day, it can be assumed that the events did not occur simultaneously and equal fractions of a day separated them.

Information content of counts and intervals

When studying the rate of events, the times of occurrence are typically known to a certain accuracy. The information content of each datum can be specified as the minimum number of bits necessary to encode the event timing (Brillouin, 2004). Thus the number of time increments within the period of observation determines the minimum number of bits to encode each time.

$$nbits_{interval} = \lceil \log_2(n \text{ measurement increments}) \rceil$$

For example, if the period of observation is ten years and the accuracy of measurement is one day, each datum could be specified as one of the 3652 days in that interval. This would require twelve bits.

When the events are transformed to counts, the number of counting intervals and the maximum count per interval determine the information content.

$$nbits_{count} = \lceil \log_2(n \text{ counting intervals}) + \log_2(\text{maximum count per interval}) \rceil$$

In the above example, with ten counting intervals of

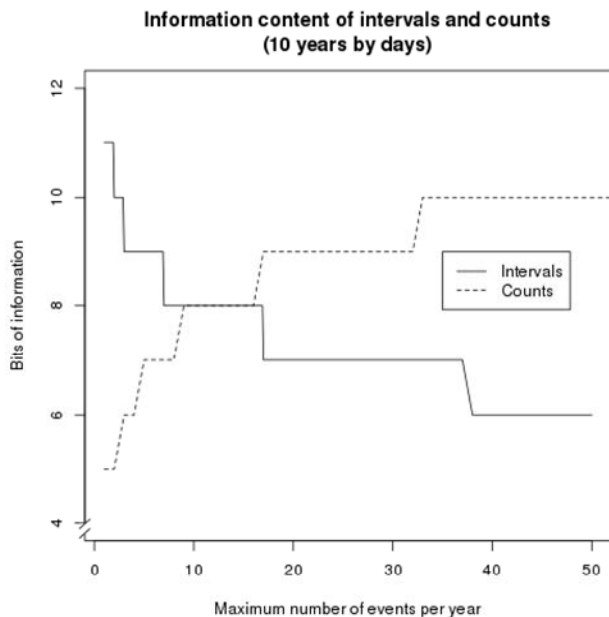


Figure 1 ■ Comparison of the information content of intervals and counts for ten years by number of events per year.

years and a maximum count per year of one hundred, eleven bits would be required to encode each datum. When the counts are small, however, the information content will be reduced. If a maximum of three events occurred per year, only six bits would be required to encode each datum. In general, the number of bits required to represent a number is the ceiling of the log of that number to the base 2. The relationship between the information content of the intervals and counts can be represented as a ratio:

$$nbits_{count}/nbits_{interval}$$

This ratio encapsulates the proportion of information available in the intervals that is expressed in the counts. It is possible for the ratio to exceed 1, if the number of events occurring within the counting intervals is greater than the number of time increments. It is clear from Figure 1 that the information content of the two types of data for ten years by days becomes equal at about nine events per year. However, it is in cases where the number of events per counting interval is small that is of concern, and as will be shown below, very large counts present a different problem.

Figure 1 illustrates the relationship between the information content of intervals measured at a resolution of days over ten years and the information

content of counts per year for the same data for maximum counts per year from one to 100. While the information content of intervals is determined by the number of measurement increments within the period of observation, the information content of counts is strongly influenced by the number of events in a counting interval.

Comparison of count and interval analyses

When examining changes in event rates, the typical null hypothesis is that the events are uniformly distributed within the period of observation. The number of events that occur in a given period of observation is inversely related to the mean interval between the events.

$$\text{mean interval} = \frac{\text{period of observation}}{\text{number of events}}$$

To test for changes with counts, the count of all events observed is divided into counts for equal intervals of time. These counts are distributed as Poisson variates (Haight, 1967). The intervals between events can also be used, and these follow an exponential distribution (Whitworth, 1951). The major difference between these two approaches is that the sequence of counts loses a great deal of information about the variance of the inter-event intervals, which are proportional to the square of the range of values. As shown in Figure 1, the smaller the number of events per counting interval, the greater the loss of information. Generalized linear modeling (GLM) can deal with a number of distributions by using a link function to ensure that linear changes in the transformed response variable correspond to linear changes in the predictor variables (Dobson, 1999). For the Poisson distribution, the link function is the natural logarithm, and for the exponential, the inverse (negative of the reciprocal). The inverse link function cannot accept zeros, thus it is necessary to separate events occurring in the same time increment by fractions of that increment. These link functions will be referred to as “poisson” and “Gamma” in the conventional notation. The following examples have all been created using the R statistical language (R Core Team, 2013).

Change in event rate with simulated data

The information loss described above is one unavoidable consequence of applying a counting process to uncommon events. The first example is specifically constructed to demonstrate what can

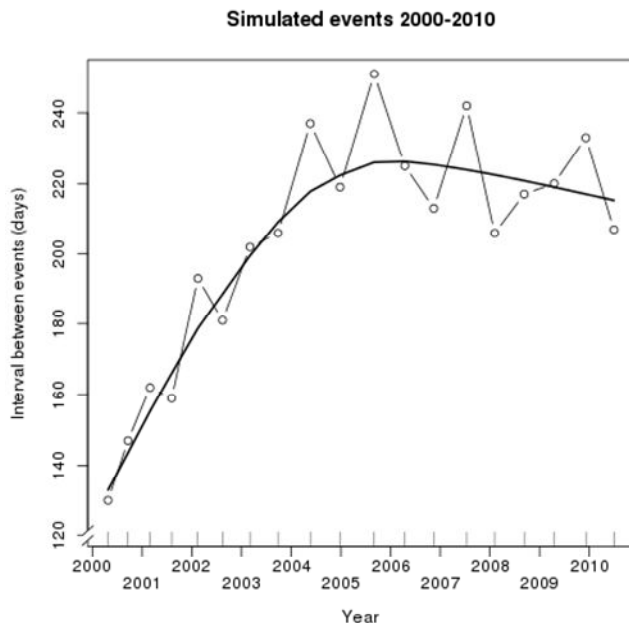


Figure 2 ■ Intervals between simulated events, 2000-2010

happen with uncommon events (Appendix, Listing 1). These data simulate events that happen about twice in a year. Using the method described above, we can calculate the information content of the intervals as twelve bits and the information content of the counts as six bits. As can be seen from the plot of the distribution of these events in Figure 2, the intervals gradually increase. The bold line is a smoothed estimate of the distribution of intervals (Friedman, 1984). However, the number of events per year remains almost the same.

Using a fairly standard GLM approach to test for a change in event rate, there appears to be no effect of time with a poisson link function ($z = -0.141$, $p = 0.89$; see Appendix, Listing and output 2).

However, these data are underdispersed (variance much smaller than the mean) with an index of dispersion of 0.09 and do not fit the assumptions of the Poisson distribution very well. Relaxing the assumptions by using a quasipoisson link may provide a better answer.

While the underdispersion is detected, the result is virtually the same, with no change in rate evident ($z = -0.45$, $p = 0.66$; see Appendix, Listing and output 3). Listing and output 4 shows the result of testing the intervals rather than the counts. Note that in all interval analyses, the first date is dropped, as the interval for that date is unknown.

This test reveals the increase in the intervals between events over time that is apparent in Figure 2

($z = 3.93$, $p = 0.001$; see Appendix, Listing and output 4)). This change does not appear to be linear, but increases and then levels out. A test for quadratic trend using the squared number of days from the final observation shows an even stronger effect ($z = 5.79$, $p = 0.00002$; see Appendix, Listing and output 5).

The decrease in the probability of the relationship over time given a constant rate as well as the corresponding decrease in the Akaike Information Criterion (AIC) value (Akaike, 1974) indicates that the change in rate is better modeled as non-linear. This example with simulated data illustrates the extent to which the loss of information incurred when using counts rather than event intervals can affect the outcome of a test for event rate change.

Change in event rate using historical data

The example above is somewhat contrived, using data that were chosen to have increasing intervals but a fairly constant rate per year. To demonstrate how the event intervals can be useful with realistic data, those of recorded hurricanes making landfall in the state of Florida, USA, during the twentieth century are employed. Florida is one of the states most likely to be struck by a hurricane, and there are good records for these events during the last century. Sixty five hurricanes made landfall in Florida between 1900 and 1999 according to Blake, Rappaport and Landsea, 2007 (see Appendix, Listing 6).

Figure 3 shows the intervals between these events over the twentieth century. While there is some clustering of hurricanes, the smoothed line (bold) appears to show a modest increase in the intervals over this time. Before looking for changes in the rate of hurricane landfalls, the distribution of these events can be examined. Figure 4 shows the distribution of intervals with a smoothed line of the actual distributions and a dotted line showing the theoretical exponential distribution for the estimated shape parameter of the observations. The fit appears to be acceptable.

As above, the analysis using counts will be compared with that using intervals. Using equations 2 and 3, the information content of the intervals in these data is sixteen bits, while that of the counts is nine bits. Listing and output 7 shows the result of the analysis of counts. There is no significant change in the rate ($z = -1.23$, $p = 0.24$ see Appendix, Listing 6).

In Listing and output 8, the same analysis is performed using the intervals. Here the apparent

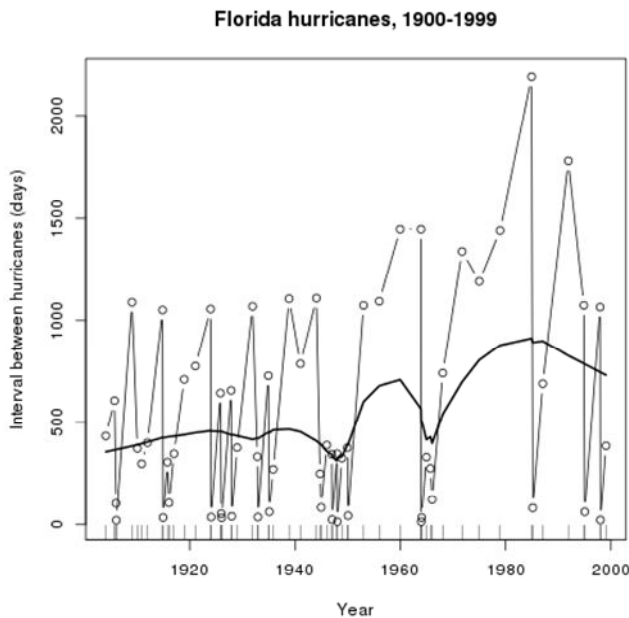


Figure 3 ■ Hurricanes making landfall in Florida 1900-1999, intervals between hurricanes by dates of hurricanes with smoothed interval curve.

increase in intervals between hurricanes emerges as a significant relationship between the dates of the hurricanes and the intervals between them, suggesting that there was a decrease in rate ($z = -2.51, p = 0.015$; see Appendix, Listing and output 7).

Loss of information and statistical power

To further illustrate the relationship between information content and event rates, a Monte Carlo simulation was conducted in which longer inter-event intervals and thus decreasing numbers of events occur during the period of observation. A function (Appendix, Listing 9) was written to uniformly distribute a specified number of events across a number of counting intervals (e.g. “years”) with a given time resolution (e.g. “days”). The function returns a list with three components, the time of occurrence of each event in units of the time resolution, the intervals between events and the number of events occurring in each of the counting intervals. A specified linear change can be added into the intervals to simulate an effect. One thousand repetitions of event generation and analysis using both counts and intervals were conducted for 20, 40, 80, 160, 320 and 640 events during the period of observation. The “effect” added was an approximately one third increase in the inter-event intervals across the period of observation.

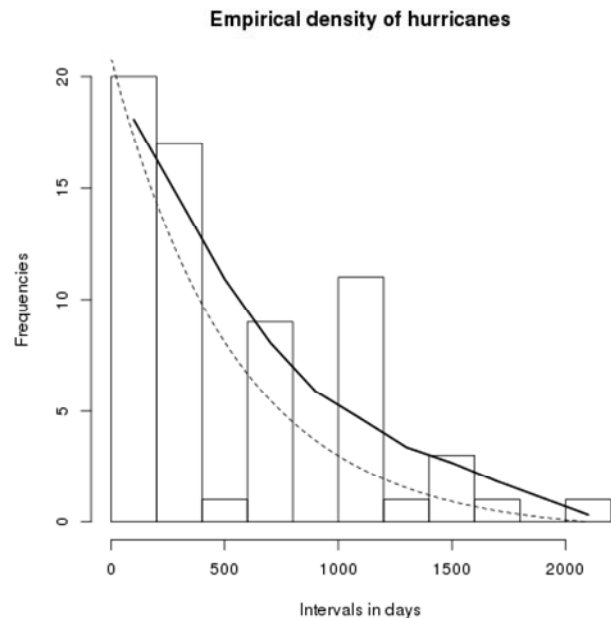


Figure 4 ■ Empirical density of hurricanes 1900-1999 with smoothed empirical density and best-fit exponential density curves

Figure 5 shows the proportion of significant ($p < 0.05$) tests for the count and interval models. As the counts of events increase, the information lost on the variance of intervals decreases. The index of dispersion for counts increases from about 0.3 to 2.4. The poisson distribution assumes an index of dispersion of 1, that is, the mean and variance should be approximately equal. Therefore the simulation with only two events per counting interval is very underdispersed, while the one with 640 events is overdispersed. The power of the count model to recognize a substantial linear change is very low with few events, but acceptable with many. The use of the negative binomial link function is typically recommended for dealing with overdispersion (Venables & Ripley, 2002)

Discussion

In both simulated and historical data, analysis by intervals rather than counts has revealed changes in rates that are apparent from graphical illustrations of the data. Both data sets were based on events that were uncommon, and thus led to small counts per counting interval. In this situation, the information content of intervals is considerably greater than that of counts per unit time.

Event interval analysis using the generalized linear model may provide a better method for studying

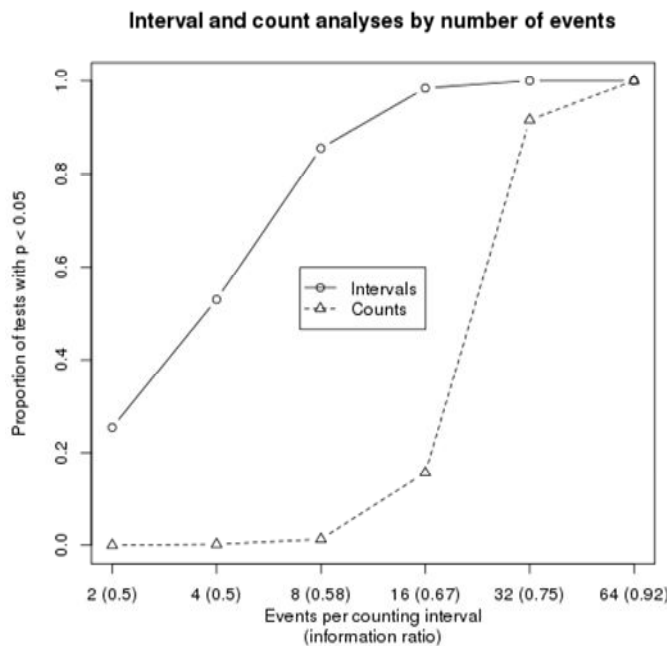


Figure 5 ■ Comparison of significant ($p < 0.05$) tests of simulated data by count and interval models

changes in the rate of sequential events when they are relatively uncommon. The loss of information inherent in transforming intervals to counts per unit time would be expected to reduce the power of statistical tests and appears to have done so in the two examples and the Monte Carlo simulation presented here. This should be particularly important when interactions between predictor variables are studied and argues for the wider use of event interval analysis in studying the rates of uncommon events.

A package for the R statistical language (eventInterval) has been created to demonstrate the methods for event interval analyses and to automate some of the procedures.

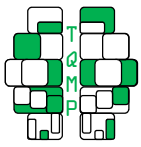
Appendix: Listing used in the article

Listing 1 ■ Dates, intervals and yearly counts for simulated events used in the first example.

```
bddates <- as.Date("14/6/2000", format="%d/%m/%Y")
bdints <- c(130,147,162,159,193,181,202,206,237,219,251,225,213,242,206,217,220,233,207)
bddates <- c(bddates,bddates+cumsum(bdints))
bdcunts <- table(format(bddates,"%Y"))
```

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6): 716–723.
- Blake, E.S., Rappaport, E.N., Landsea, C.W. (2007). The deadliest, costliest, and most intense United States tropical cyclones from 1851 to 2006 (and other frequently requested hurricane facts). NOAA Technical Memorandum NWS TPC-5, Miami: National Weather Service.
- Brillouin, L. (2004). *Science and information theory*. Mineola, NY: Dover Publications.
- Directorate of Health (2013). The Icelandic Accident Register. <http://www.landlaeknir.is/english/>, accessed 9 August 2013.
- Dobson, A.J. (1999) *An introduction to generalized linear models*. Boca Raton, Chapman & Hall.
- Friedman, J.H. (1984). *A variable span scatterplot smoother*. Stanford University Technical Report No. 5, Stanford: Laboratory for Computational Statistics.
- Haight, F.A. (1967). *Handbook of the Poisson Distribution*. New York: John Wiley & Sons.
- Maguire, B.A., Pearson, E.S. & Wynn, A.H.A. (1952). The time intervals between industrial accidents. *Biometrika*, 39(1/2): 168-180.
- R Core Team. (2013) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*. New York: Springer.
- Whitworth, W.A. (1951). *Choice and chance*. (reprinted) New York: Hafner.
- Wikipedia. (2013). List of people who died climbing Mount Everest. http://en.wikipedia.org/wiki/List_of_people_who_died_climbing_Mount_Everest, accessed 9 August 2013.



Listing and output 2 ■ Test for linear change in simulated event rate using the Poisson link.

```
summary(glm(bdcounsts~I(2000:2010), family="poisson"))

Call: glm(formula = bdcounsts ~ I(2000:2010), family = "poisson")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.66323  0.07123  0.10587  0.16066  0.20133
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  20.65141    141.81127   0.146   0.884
I(2000:2010) -0.01000     0.07073  -0.141   0.888
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1.0398  on 10  degrees of freedom
Residual deviance: 1.0198  on  9  degrees of freedom
AIC: 32.543
Number of Fisher Scoring iterations: 4
```

Listing and output 3 ■ Test for linear change in simulated event rate using the quasipoisson link

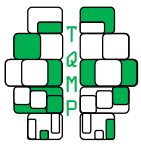
```
summary(glm(bdcounsts~I(2000:2010), family="quasipoisson"))

Call: glm(formula = bdcounsts ~ I(2000:2010), family = "quasipoisson")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.66323  0.07123  0.10587  0.16066  0.20133
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.65141    44.57510   0.463   0.654
I(2000:2010) -0.01000     0.02223  -0.450   0.663
(Dispersion parameter for quasipoisson family taken to be 0.09880144)
    Null deviance: 1.0398  on 10  degrees of freedom
Residual deviance: 1.0198  on  9  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 4
```

Listing and output 4 ■ Test for linear change in simulated event rate using the Gamma link

```
summary(glm(bdints~bddates[-1], family="Gamma"))

Call: glm(formula = bdints ~ bddates[-1], family = "Gamma")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.25998 -0.07860 -0.01345  0.07940  0.21579
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.135e-02  1.659e-03   6.845 2.85e-06
bddates[-1] -4.881e-07  1.244e-07  -3.925 0.00109
```

```
(Dispersion parameter for Gamma family taken to be 0.01662602)
Null deviance: 0.54105 on 18 degrees of freedom
Residual deviance: 0.28554 on 17 degrees of freedom
AIC: 181.5
Number of Fisher scoring iterations: 4
```

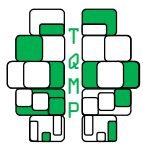
Listing and output 5 ■ Test for quadratic change in simulated event rate using the Gamma link

```
# calculate the time from the end of the observation interval
bddates2 <- as.numeric(bddates)-14972
# square that to test for quadratic trend
bddates2 <- bddates2*bddates2
summary(glm(bdints~bddates2[-1],family="Gamma"))

Call: glm(formula = bdints ~ bddates2[-1], family = "Gamma")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.17855 -0.07028 -0.01690  0.06403  0.17485
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.213e-03  1.579e-04  26.677 2.58e-15
bddates2[-1]  1.580e-10  2.729e-11   5.789 2.18e-05
(Dispersion parameter for Gamma family taken to be 0.01034407)
Null deviance: 0.54105 on 18 degrees of freedom
Residual deviance: 0.17441 on 17 degrees of freedom
AIC: 172.12
Number of Fisher Scoring iterations: 4
```

Listing 6 ■ Dates of hurricanes making landfall in Florida USA during the 20th century

```
fh_dates <- as.Date(c(
"1903-08-11", "1904-10-17", "1906-06-16", "1906-09-27", "1906-10-18", "1909-10-11",
"1910-10-18", "1911-08-11", "1912-09-14", "1915-08-01", "1915-09-04", "1916-07-05",
"1916-10-18", "1917-09-29", "1919-09-10", "1921-10-25", "1924-09-15", "1924-10-21",
"1926-07-27", "1926-09-18", "1926-10-21", "1928-08-08", "1928-09-17", "1929-09-28",
"1932-09-01", "1933-07-30", "1933-09-04", "1935-09-03", "1935-11-04", "1936-07-31",
"1939-08-11", "1941-10-06", "1944-10-19", "1945-06-24", "1945-09-15", "1946-10-08",
"1947-09-17", "1947-10-11", "1948-09-21", "1948-10-05", "1949-08-26", "1950-09-05",
"1950-10-18", "1953-09-26", "1956-09-24", "1960-09-10", "1964-08-27", "1964-09-10",
"1964-10-14", "1965-09-08", "1966-06-09", "1966-10-08", "1968-10-19", "1972-06-19",
"1975-09-23", "1979-09-03", "1985-09-01", "1985-11-21", "1987-10-12", "1992-08-24",
"1995-08-03", "1995-10-04", "1998-09-03", "1998-09-25", "1999-10-15"
), "%Y-%m-%d")
fh_days <- as.numeric(fh_dates)
fh_ints <- diff(fh_days)
fh_counts <- tabulate(as.numeric(factor(format(fh_dates, "%Y"),
levels=as.character(1900:1999))), nbins=100
)
```



Listing and output 7 ■ Test for linear change in hurricane landfall rate using the Poisson link

```
summary(glm(fh_counts~I(1900:1999), family="poisson"))

Call: glm(formula = fh_counts ~ I(1900:1999), family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3175  -1.1131  -0.9970   0.4776   2.2309

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.560841    8.434690   1.371    0.170
I(1900:1999) -0.006159    0.004337  -1.420    0.156

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 114.59  on 99  degrees of freedom
Residual deviance: 112.56  on 98  degrees of freedom
AIC: 218.13
Number of Fisher Scoring iterations: 5
```

Listing and output 8 ■ Test for linear change in hurricane landfall rate using the Gamma link

```
summary(glm(fh_ints~fh_days[-1], family="Gamma"))

Call: glm(formula = fh_ints~fh_days[-1], family = "Gamma")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3674  -1.4350  -0.2800   0.5792   1.2497

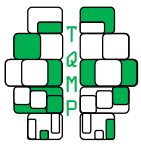
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.544e-03  1.977e-04   7.810 8.46e-11
fh_days -4.504e-08  1.798e-08  -2.505  0.0149

(Dispersion parameter for Gamma family taken to be 0.7422684)

Null deviance: 87.814  on 63  degrees of freedom
Residual deviance: 83.429  on 62  degrees of freedom
AIC: 937.51
Number of Fisher Scoring iterations: 6
```

Listing 9 ■ Function to generate interval and count data for Monte Carlo simulation.

```
# generate a sequence of simulated events for a period of observation
# nevents = number of events (e.g. 20)
# nci = number of counting intervals (e.g. 10 "years")
# incr_ci = time increments per counting interval (e.g. 365 "days")
# effect = change in mean interval from start to finish (e.g. 11 "days")
```

```
generate_event_seq <- function(nevents,nci,incr_ci,effect=0) {  
  # generate uniformly distributed events  
  # ranging from 1 (to avoid zeros) to the average interval  
  event_ints <- runif(nevents,1,nci*incr_ci/nevents)  
  # add in the effect if present  
  if(effect != 0) event_ints <- event_ints+seq(0,effect,length.out=nevents)  
  # adjust to fill the period of observation and  
  # remove fractions of time increments  
  event_ints <- round((event_ints*nci*incr_ci)/sum(event_ints),0)  
  # remove any zero intervals  
  event_ints[event_ints < 1] <- 1  
  # create the "times" of the events from the intervals  
  event_intc <- cumsum(event_ints)  
  # calculate the "times" of the ends of the counting intervals  
  ci_ends <- cumsum(rep(incr_ci,length.out=nci))  
  # initialize the counts  
  event_counts <- rep(0,nci)  
  # begin with the first counting interval  
  ci <- 1  
  # accumulate the number of events per counting interval  
  for(event in 1:nevents) {  
    # if the next event is beyond the current "end of counting interval",  
    # advance the counting interval until it is within it  
    while(event_intc[event] > ci_ends[ci] && ci < nci) if(ci < nci) ci <- ci+1  
    # add the current event to its counting interval  
    event_counts[ci] <- event_counts[ci]+1  
  }  
  return(list(times=event_intc,ints=event_ints,counts=event_counts))  
}
```

Citation

Lemon, J. (2014). The analysis of event rates using intervals. *The Quantitative Methods for Psychology*, 10(1), 68-76.

Copyright © 2014 Lemon. This is an open-access article distributed under the terms of the *Creative Commons Attribution License (CC BY)*. The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 20/09/13 ~ Accepted: 21/10/13