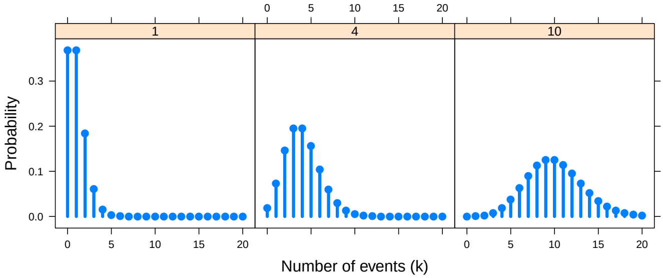
Discrete distributions

Michael Friendly

Psych 6136

September 17, 2017



Discrete distributions

Discrete distributions, such as the binomial, Poisson, negative binomial and others form building blocks for the analysis of categorical data (logistic regression, loglinear models, generalized linear models)
Such data consist of:

- **Counts of occurrences:** accidents, words in text, blood cells with some characteristic.
- **Data:** Basic outcome value, k , $k = 0, 1, \ldots$, and number of observations, n_k, with that value.

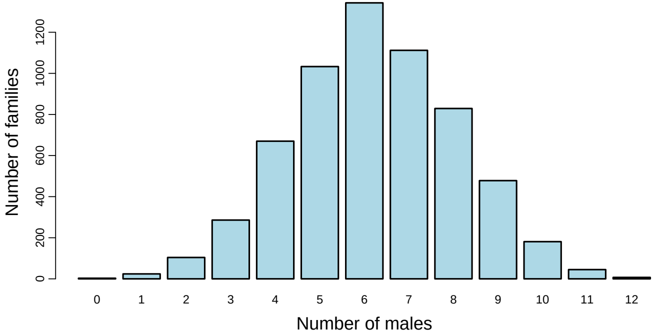
We distinguish between the count, k, and the frequency, n_k with which that count occurs.

Discrete distributions: Examples

Saxony families

Saxony families with 12 children having $k = 0, 1, \ldots 12$ sons.

k	0	1	2	3	4	5	6	7	8	9	10	11	12
n_k	3	24	104	286	670	1033	1343	1112	829	478	181	45	7



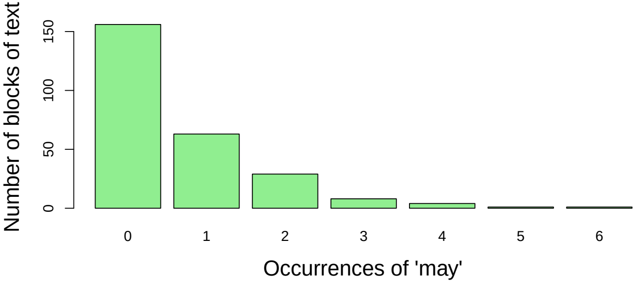
Discrete distributions: Examples I

Federalist papers— disputed authorship

- 77 essays by Hamilton, Jay & Madison: persuade NY voters to ratify Constitution, all signed with pseudonym (“Publius”)
- 65 known, 12 disputed (H & M both claimed sole authorship)
- Mosteller and Wallace (1984): Analysis of frequency distributions of key “marker” words: *from*, *may*, *whilst*, ...
- e.g., blocks of 200 words with *may*:

Occurrences (k)	0	1	2	3	4	5	6
Blocks (n_k)	156	63	29	8	4	1	1

Discrete distributions: Examples II



For each word,

- fit probability model (Poisson, NegBin)
- $\rightarrow$ estimate parameters $(\beta_1, \beta_2, \cdots)$
- $\rightarrow$ estimate log Odds (Hamilton vs. Madison)
- $\Longrightarrow$ All 12 of the disputed papers were attributed to Madison

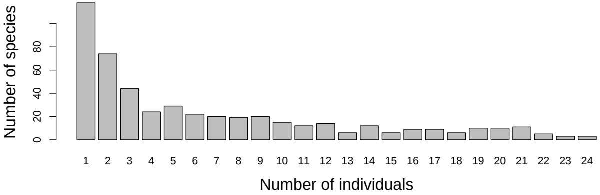
Type-token distributions I

- Basic count, k: number of "types"; frequency, n_k: number of instances observed
  - Frequencies of distinct words in a book or literary corpus
  - Number of subjects listing words as members of the semantic category "fruit"
  - Distinct species of animals caught in traps
- Differs from other distributions in that the frequency for $k = 0$ is *unobserved*
- Distribution is often extremely skewed (J-shaped)

Table: Number of butterfly species n_k for which k individuals were collected

Individuals (k)	1	2	3	4	5	6	7	8	9	10	11	12	
Species (n_k)	118	74	44	24	29	22	20	19	20	15	12	14	
Individuals (k)	13	14	15	16	17	18	19	20	21	22	23	24	Su
Species (n_k)	6	12	6	9	9	6	10	10	11	5	3	3	50

Type-token distributions II



Questions:

- What is the total population of butterflies in Malaya?
- How many wolves remain in Canada's Northwest territories?
- How many words did Shakespeare know?[a]

[a] In known works, Shakespeare used 31,534 distinct words (types), totaling 884,647 words (tokens). Answers depend on fitting a distribution, and estimating the probability for $k = 0$

Discrete distributions: Questions

General questions:

- What process gave rise to the distribution?
- Form of distribution: uniform, binomial, Poisson, negative binomial, geometric, etc.?
- Estimate parameters
- Visualize goodness of fit

For example:

- *Families in Saxony:* might expect a Bin(n, p) distribution with $n = 12$. Perhaps $p = 0.5$ as well.
- *Federalist Papers:* might expect a Poisson(λ) distribution.
- *Butterfly data:* perhaps a log-series distribution would be reasonable

Discrete distributions: Lack of fit

Lack of fit:

- Lack of fit tells us something about the process giving rise to the data
- Poisson: assumes constant small probability of the basic event
- Binomial: assumes constant probability and independent trials
- Negative binomal: allows for *overdispersion*, relative to Poisson

Motivation:

- Models for more complex categorical data use these basic discrete distributions
- Binomial (with predictors) $\rightarrow$ logistic regression
- Poisson (with predictors) $\rightarrow$ poisson regression, loglinear models
- $\Rightarrow$ many of these are special cases of *generalized linear models*

Common discrete distributions

Discrete distributions are all characterized by a probability function (or probability mass function), $\Pr(X = k) \equiv p(k)$ that the random variable X takes the value k.
The commonly used discrete distributions have the following forms:

Table: Discrete probability distributions

Discrete distribution	Probability function, $p(k)$	Parameters
Binomial	$\binom{n}{k} p^k (1-p)^{n-k}$	p = Pr (success); n = # trials
Poisson	$e^{-\lambda} \lambda^k / k!$	λ = mean
Negative binomial	$\binom{n+k-1}{k} p^n (1-p)^k$	p; n = # *successful* trials
Geometric	$p(1-p)^k$	p
Logarithmic series	$\theta^k / [-k \log(1-\theta)]$	θ

Binomial distribution

The binomial distribution, $\mathrm{Bin}(n, p)$,

$$\mathrm{Bin}(n,p) : \Pr\{X = k\} \equiv p(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n \ , \quad (1)$$

arises as the distribution of the number of events of interest ("successes") which occur in n *independent trials* when the probability of the event on any one trial is the *constant* value $p = \Pr(\text{event})$.
Examples:

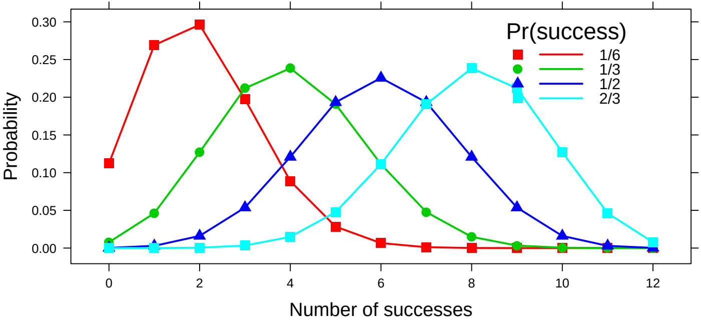
- Toss 10 fair coins— how many heads: $\mathrm{Bin}(10, \frac{1}{2})$
- Toss 12 fair dice— how many 5s or 6s: $\mathrm{Bin}(12, \frac{1}{3})$

Mean & variance:

$$\begin{aligned} \mathrm{Mean}[X] &= np \\ \mathrm{Var}[X] &= np(1-p) \end{aligned}$$

Binomial distribution

Binomial distributions for $k = 0, \ldots, 12$ successes in $n = 12$ trials, and four values of p



Poisson distribution

The Poisson distribution, Pois(λ),

$$\text{Pois}(\lambda) : \Pr\{X = k\} \equiv p(k) = \frac{e^{-\lambda}\,\lambda^k}{k!} \qquad k = 0, 1, \ldots \tag{2}$$

gives the probability of an event occurring $k = 0, 1, 2, \ldots$ times over a *large number of independent* trials, when the probability, p, that the event occurs on any one trial (in time or space) is *small and constant*.

Examples:

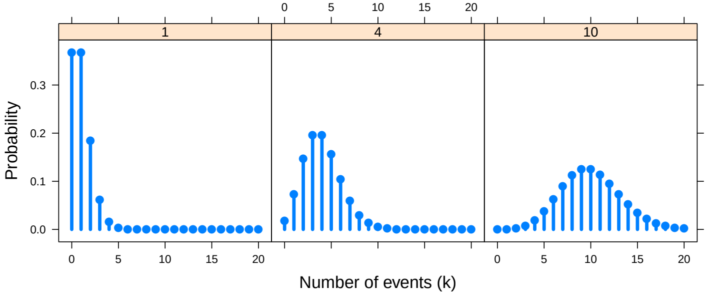
- Number of highway accidents at some given location
- Defects in a manufacturing process
- Number of goals scored in soccer games

Table: Total goals scored in 380 games in the Premier Football League, 1995/95 season

Total goals	0	1	2	3	4	5	6	7
Number of games	27	88	91	73	49	31	18	3

Poisson distribution

Poisson distributions for $\lambda = 1, 4, 10$



Mean, variance & skewness:

$$\begin{aligned} \text{Mean}[X] &= \lambda \\ \text{Var}[X] &= \lambda \\ \text{Skew}[X] &= \lambda^{-1/2} \end{aligned}$$

Negative binomial distribution

The Negative binomial distribution, $\text{NBin}(n, p)$,

$$\text{NBin}(n, p) : \Pr\{X = k\} \equiv p(k) = \binom{n+k-1}{k} p^n (1-p)^k \qquad k = 0, 1, \ldots, \infty$$

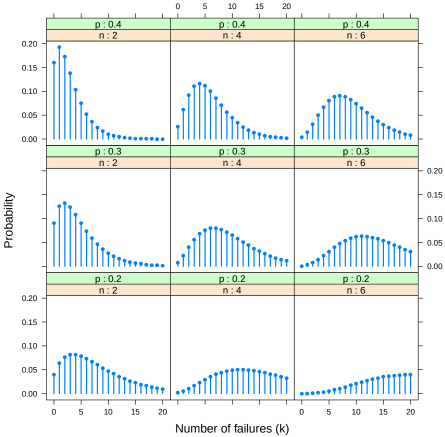
arises when a series of independent Bernoulli trials is observed with constant probability p of some event, and we ask how many non-events (failures), k, it takes to observe n successful events.
Example: Toss a coin; what is probability of getting $k = 0, 1, 2, \ldots$ tails before $n = 3$ heads?

This distribution is often used as an alternative to the Poisson when

- constant probability p or independence are violated
- variance is greater than the mean (overdispersion)

Negative binomial distribution

Negative binomial distributions for $n = 2, 4, 6$ and $p = 0.2, 0.3, 0.4$



Mean increases with n and decreases with p.

Fitting discrete distributions

Fitting a discrete distribution involves the following steps:

1. Estimate the parameter(s) from the data, e.g., p for binomial, λ for Poisson, etc. Typically done using maximum likelihood, but some distributions have simple expressions:
   - Binomial, $\hat{p} = \sum k n_k / (n \sum n_k)$ = mean / n
   - Poisson, $\hat{\lambda} = \sum k n_k / \sum n_k$ = mean
2. Calculate fitted probabilities, $\hat{p}(k)$ for the distribution, and then fitted frequencies, $N\hat{p}(k)$.
3. Assess Goodness of fit: Pearson X^2 or likelihood-ratio G^2

$$X^2 = \sum_{k=1}^{K} \frac{(n_k - N\hat{p}_k)^2}{N\hat{p}_k} \qquad G^2 = \sum_{k=1}^{K} n_k \log(\frac{n_k}{N\hat{p}_k})$$

Both have asymptotic chisquare distributions, χ^2_{K-s} with s estimated parameters, under the hypothesis that the data follows the chosen distribution.

Fitting and graphing discrete distributions

In R, the vcd and vcdExtra packages contain methods to fit, visualize, and diagnose discrete distributions:

- **Fitting:** `goodfit()` fits uniform, binomial, Poisson, negative binomial, geometric, logarithmic series distributions (or any specified multinomial)
- **Hanging rootograms:** Sensitively assess departure between Observed, Fitted counts (`rootogram()`)
- **Ord plots:** Diagnose form of a discrete distribution (`Ord_plot()`)
- **Robust distribution plots for various distributions** (`distplot()`)

Example: Saxony data

```
library(vcd)
data(Saxony)
Saxony

## nMales
##    0    1    2    3    4    5    6    7    8    9   10   11   12
##    3   24  104  286  670 1033 1343 1112  829  478  181   45    7
```

Use **`goodfit()`** to fit the binomial; test with **`summary()`**:

```
Sax.fit <- goodfit(Saxony, type="binomial")
summary(Sax.fit)

##
##   Goodness-of-fit test for binomial distribution
##
##                     X^2 df   P(> X^2)
## Likelihood Ratio 97.007 11 6.9782e-16
```

Example: Saxony data

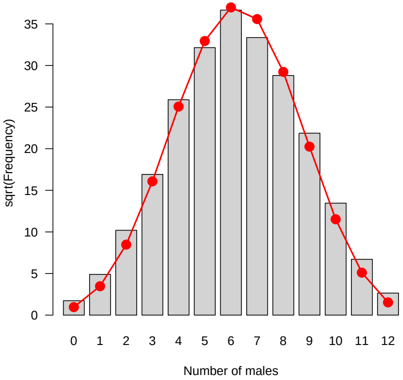
The `print()` method shows the details:

```
Sax.fit   # print

##
## Observed and fitted values for binomial distribution
## with parameters estimated by `ML'
##
##  count observed     fitted pearson residual
##      0        3    0.93284          2.14028
##      1       24   12.08884          3.42580
##      2      104   71.80317          3.79963
##      3      286  258.47513          1.71205
##      4      670  628.05501          1.67371
##      5     1033 1085.21070         -1.58490
##      6     1343 1367.27936         -0.65661
##      7     1112 1265.63031         -4.31841
##      8      829  854.24665         -0.86380
##      9      478  410.01256          3.35761
##     10      181  132.83570          4.17896
##     11       45   26.08246          3.70417
##     12        7    2.34727          3.03687
```

What's wrong with histograms?

Discrete distributions are often graphed as histograms, with a theoretical fitted distribution superimposed.

```
plot(Sax.fit,type="standing", xlab="Number of males")
```

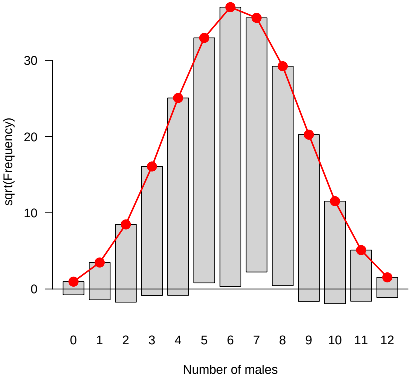


Problems:

- largest frequencies dominate display
- must assess deviations vs. a curve

Hang & root them → Hanging rootograms

```
plot(Sax.fit, xlab="Number of males")
```



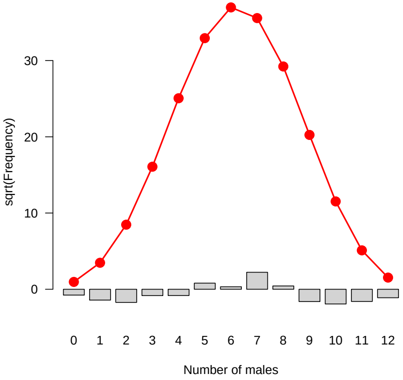
Tukey (1972, 1977):

- shift histogram bars to the fitted curve
- → judge deviations vs. horizontal line.
- plot $\sqrt{\text{freq}}$ → smaller frequencies are emphasized.

We can now see clearly where the binomial doesn’t fit

Highlight differences → Deviation rootograms

```
plot(Sax.fit, type="deviation", xlab="Number of males")
```



Deviation rootogram:

- emphasize differences between observed and fitted frequencies
- bars now show the residuals (gaps) directly

There are more families with very low or very high number of sons than the binomial predicts.

Q: Why is this so much better than the lack-of-fit test?

Example: Federalist papers

```
data(Federalist, package="vcd")
Federalist

## nMay
##   0   1   2   3   4   5   6
## 156  63  29   8   4   1   1
```

Fit the Poisson distribution:

```
Fed.fit0 <- goodfit(Federalist, type="poisson")
summary(Fed.fit0)

##
##   Goodness-of-fit test for poisson distribution
##
##                      X^2 df     P(> X^2)
## Likelihood Ratio 25.243  5 0.00012505
```

This fits very poorly!

Example: Federalist papers

Fit the Negative binomial distribution:

```
Fed.fit1 <- goodfit(Federalist, type="nbinomial")
summary(Fed.fit1)

##
##   Goodness-of-fit test for nbinomial distribution
##
##                     X^2 df P(> X^2)
## Likelihood Ratio 1.964  4  0.74238
```

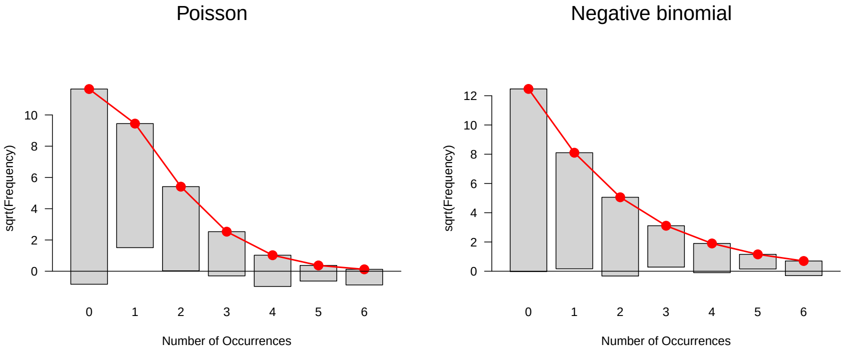
This now fits very well, indeed! Why?

- Poisson assumes that the probability of a given word (“may”) is constant across all blocks of text.
- Negative binomial allows the rate parameter λ to vary over blocks of text

Example: Federalist papers: Rootograms

Hanging rootograms for the Federalist Papers data, comparing the Poisson and negative binomial models:

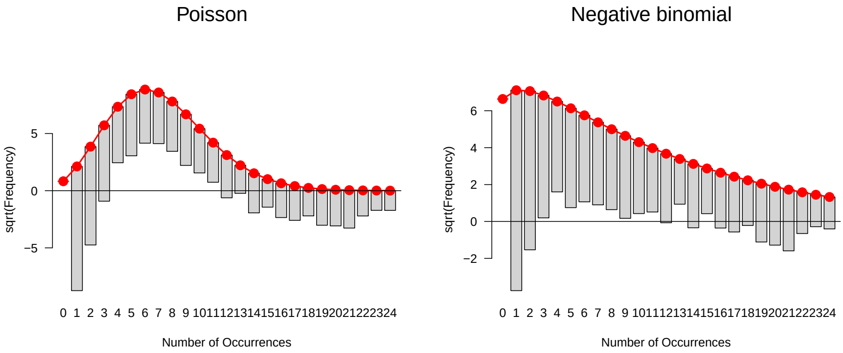
```
plot(Fed.fit0, main="Poisson")
plot(Fed.fit1, main="Negative binomial")
```



Example: Butterfly data

Butterfly data: neither Poisson or Negative binomial fit:

```
But.fit1 <- goodfit(Butterfly, type="poisson")
But.fit2 <- goodfit(Butterfly, type="nbinomial")
plot(But.fit1, main="Poisson")
plot(But.fit2, main="Negative binomial")
```



Ord plots: Diagnose form of discrete distribution

How to tell which discrete distributions are likely candidates?

- Ord (1967): for each of Poisson, Binomial, Negative binomial, and Logarithmic series distributions,
  - plot of kp_k/p_{k-1} against k is linear
  - signs of intercept and slope → determine the form, give rough estimates of parameters

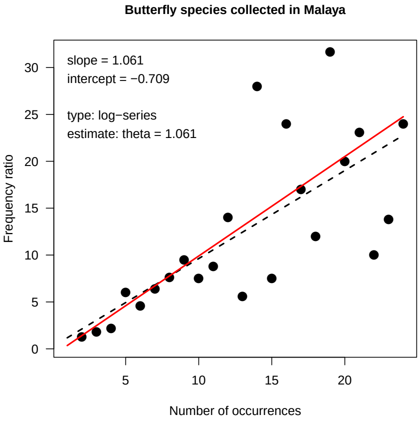
Slope (b)	Intercept (a)	Distribution (parameter)	Parameter estimate
0	$+$	Poisson (λ)	$\lambda = a$
$-$	$+$	Binomial (n, p)	$p = b/(b-1)$
$+$	$+$	Neg. binomial (n,p)	$p = 1 - b$
$+$	$-$	Log. series (θ)	$\theta = b$
			$\theta = -a$

- Fit line by WLS, using $\sqrt{n_k - 1}$ as weights
- A heuristic method: doesn't always work, but often a good start.

Ord plots: Examples

Ord plot for the Butterfly data. The slope and intercept in the plot correctly diagnoses the log-series distribution.

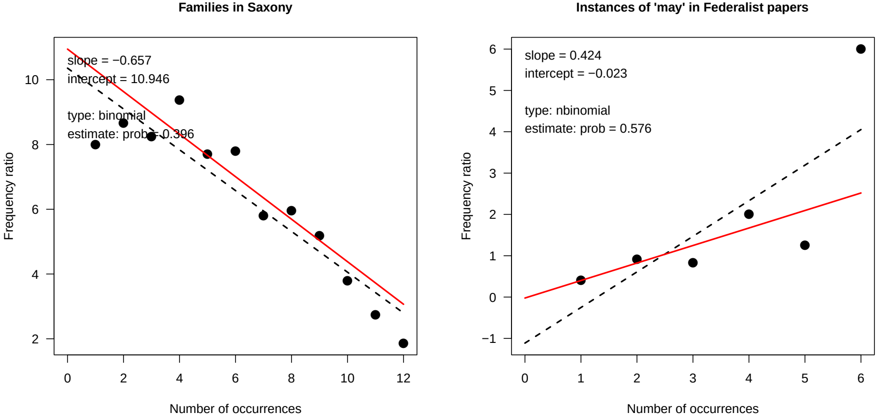
```
Ord_plot(Butterfly,
        main = "Butterfly species collected in Malaya", gp=gpar(c
```



Ord plots: Examples

Happily, these are all members of a family called the power series distributions. Ord plots for the Saxony and Federalist data sets:

```
Ord_plot(Saxony, main = "Families in Saxony", gp=gpar(cex=1), pch=16)
Ord_plot(Federalist, main = "Instances of 'may' in Federalist papers", gp=
```



Robust distribution plots: Poisson

- Ord plots lack robustness
  - one discrepant freqency, n_k affects points for both k and $k+1$
  - the use of WLS to fit the line is a small attempt to minimize this
- Robust plots for Poisson distribution (Hoaglin and Tukey, 1985)
  - For Poisson, plot ***count metameter*** = $\phi(n_k) = \log_e(k!\, n_k/N)$ vs. k
  - Linear relation $\Rightarrow$ Poisson, slope gives $\hat{\lambda}$
  - CI for points, diagnostic (influence) plot
  - Implemented in **`distplot()`** in the vcd package

Poissonness plots: Details

- If the distribution of n_k is Poisson(λ) for some fixed λ, then each observed frequency, $n_k \approx m_k = Np_k$.
- Then, setting $n_k = Np_k = e^{-\lambda} \, \lambda^k / k!$, and taking logs of both sides gives

$$\log(n_k) = \log\, N - \lambda + k\, \log\, \lambda - \log\, k!$$

which can be rearranged to

$$\phi\,(n_k) \equiv \log\left(\frac{k!\; n_k}{N}\right) = -\lambda + (\log\, \lambda)\, k$$

- $\Rightarrow$ if the distribution is Poisson, plotting $\phi(n_k)$ vs. k should give a line with
    - intercept = $-\lambda$
    - slope = log λ
- Nonlinear relation $\rightarrow$ distribution is *not* Poisson
- Hoaglin and Tukey (1985) give details on calculation of confidence intervals and influence measures.

Distribution plots: Other distributions

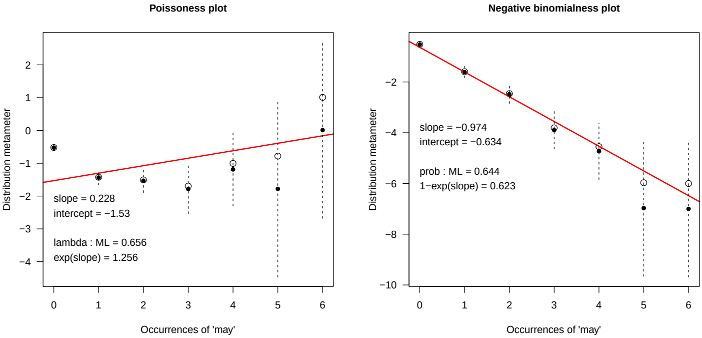
This idea extends readily to other discrete data distributions:

- The binomial, Poisson, negative binomial, geometric and logseries distributions are all members of a general power series family of discrete distributions. See: *DDAR*, Table 3.10 for details.
- This allows all of these to be represented in a plot of a suitable count metameter, $\phi(n_k)$ vs. k. See: *DDAR*, Table 3.12 for details.
- In these plots, a straight line confirms that the data follow the given distribution.
- Confidence intervals around the points indicate uncertainty for the count metameter.
- The slope and intercept of the line give estimates of the distribution parameters.

distplot: Example: Federalist

Diagnostic distribution plots for the Federalist papers data.

```
distplot(Federalist, type="poisson", xlab="Occurrences of 'may'")
distplot(Federalist, type="nbinomial", xlab="Occurrences of 'may'")
```

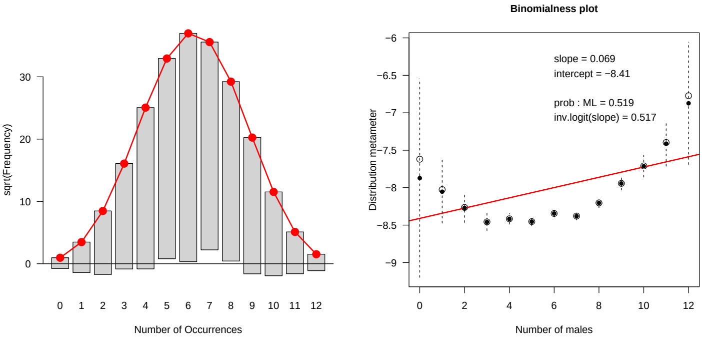


Again, the Poisson distribution is seen not to fit, while the Negative binomial appears reasonable.

distplot: Example: Saxony

For purported binomial distributions, the result is a “Binomialness” plot.

```
plot(goodfit(Saxony, type="binomial", par=list(size=12)))
distplot(Saxony, type="binomial", size=12, xlab="Number of males")
```



Both plots show heavier tails than in a binomial distribution.

What have we learned?

Main points:

- Discrete distributions involve basic *counts* of occurrences of some event occurring with varying *frequency*.
- The ideas and methods for one-way tables are building blocks for analysis of more complex data.
- Commonly used discrete distributions include the binomial, Poisson, negative binomial, and logarithmic series distributions, all members of a *power series* family.
- Fitting observed data to a distribution $\rightarrow$ fitted frequencies, $N\hat{p}_k$, $\rightarrow$ goodness-of-fit tests (Pearson X^2, LR G^2)
- R: **`goodfit()`** provides **`print()`**, **`summary()`** and **`plot()`** methods.
- Plotting with rootograms, Ord plots and generalized distribution plots can reveal *how* or*where* a distribution does not fit.

What have we learned?

Some explantions:

- The Saxony data were part of a much larger data set from Geissler (1889) (`Geissler` in vcdExtra).
  - For the binomial, with families of size $n = 12$, our analyses give $\hat{p} = \Pr(\mathit{male}) = 0.52$.
  - Other analyses (using more complex models) conclude that p varies among families with the same size.
  - One explanation is that family decisions to have another child are influenced by the boy–girl ratio in earlier children.
- As suggested earlier, the lack of fit of the Poisson distribution for words in the Federalist papers can be explained by *context* of the writing:
  - Given "marker" words appear more or less often over time and subject than predicted by constant rates (λ) for a given author (Madison or Hamilton)
  - The negative binomial distribution fit much better.
  - The estimated parameters for these texts allowed assigning all 12 disputed papers to Madison.

Looking ahead: PhdPubs data

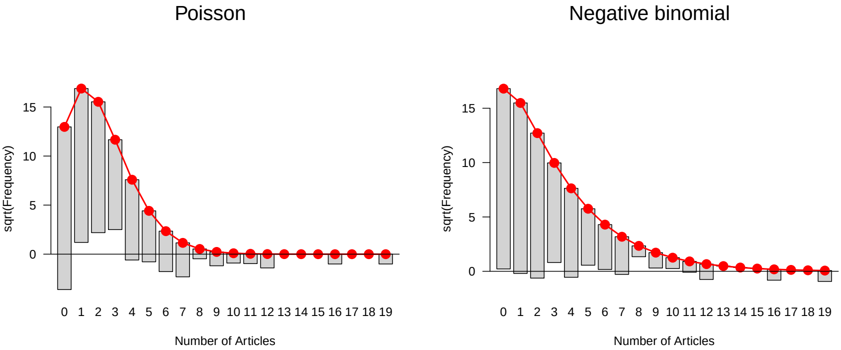
Example 3.24 in DDAR gives data on the number of publications by PhD candidates in the last 3 years of study

```
data("PhdPubs", package = "vcdExtra")
table(PhdPubs$articles)

##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  16  19
## 275 246 178  84  67  27  17  12   1   2   1   1   2   1   1
```

- There are a number of predictors: gender, marital status, number of young children, prestige of the doctoral department, and number of publications by the student’s mentor.
- When we fit a model (DDAR Example 11.1) using **`glm()`**, we need to specify the *form* of the distribution
- For now, ignore the predictors.
- Por the Poisson, equivalent to:
  `glm(articles ˜ 1, data=PhdPubs, family="poisson")`

Looking ahead: PhdPubs data

```
plot(goodfit(PhdPubs$articles), xlab = "Number of Articles",
     main = "Poisson")
plot(goodfit(PhdPubs$articles, type = "nbinomial"),
     xlab = "Number of Articles", main = "Negative binomial")
```



One reason the Poisson doesn’t fit: excess 0s (some never published)

Looking ahead: Count data models

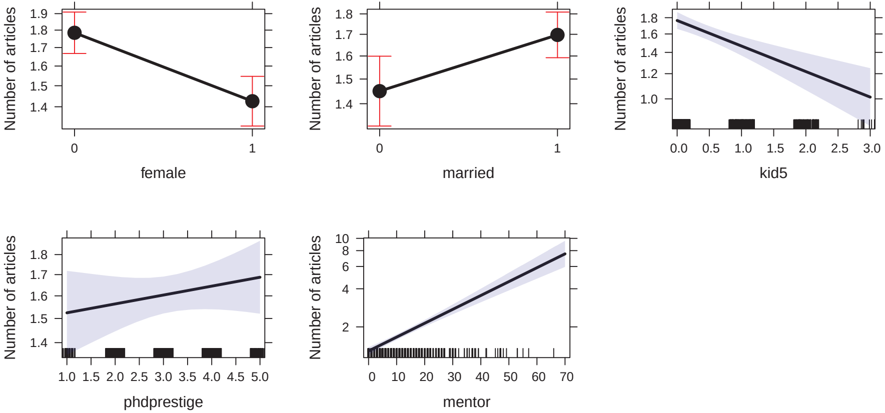
DDAR Chapter 11 describes fitting count data regression models.

```
# predictors: female, married, kid5, phdprestige, mentor
phd.pois  <- glm(articles ~ ., data=PhdPubs, family=poisson)
phd.nbin  <- glm.nb(articles ~ ., data=PhdPubs)

LRstats(phd.pois, phd.nbin)

## Likelihood summary table:
##           AIC  BIC LR Chisq  Df Pr(>Chisq)    
## phd.pois 3313 3342     1634 909     <2e-16 ***
## phd.nbin 3135 3169     1004 909      0.015 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Special models handle the problem of excess zeros: Zero-inflated (**`zeroinfl()`**) & Hurdle (**`hurdle()`**) models

Looking ahead: Effect plots

Effect plots show the predicted values for each term in a model, averaging over all other factors.



These are better visual summaries for a model than a table of coefficients.

References I

Geissler, A. Beitrage zur frage des geschlechts verhaltnisses der geborenen. *Z. K. Sachsischen Statistischen Bureaus*, 35(1):n.p., 1889.

Hoaglin, D. C. and Tukey, J. W. Checking the shape of discrete distributions. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Exploring Data Tables, Trends and Shapes*, chapter 9. John Wiley and Sons, New York, 1985.

Mosteller, F. and Wallace, D. L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, New York, NY, 1984.

Ord, J. K. Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society, Series A*, 130:232–238, 1967.

Tukey, J. W. Some graphic and semigraphic displays. In Bancroft, T. A., editor, *Statistical Papers in Honor of George W. Snedecor*, pp. 292–316. Iowa State University Press, Ames, IA, 1972.

Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977.