

Discrete distributions: Basic ideas

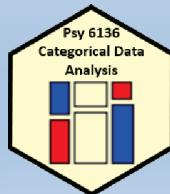
- Quantitative data: often assumed Normal (μ, σ^2) – **unreasonable** for CDA
- Binomial, Poisson, Negative binomial, ... are the building blocks for CDA
- They form the basis for modeling techniques
 - logistic regression, generalized linear models, Poisson regression
- Data: One-way tables
 - outcome variable ($k = 0, 1, 2, \dots$)
 - counts of occurrences (n_k): accidents, words in text, males in families of size k

Discrete distributions



Michael Friendly
Psych 6136

<https://friendly.github.io/6135/>



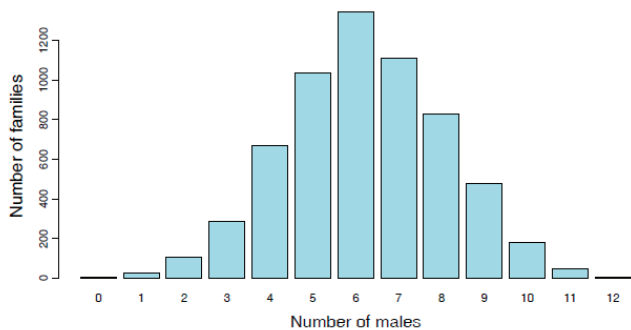
Examples: binomial?

Human sex ratio (Geissler, 1889): **Is there evidence that $\Pr(\text{male}) = 0.5$?**

Saxony families

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

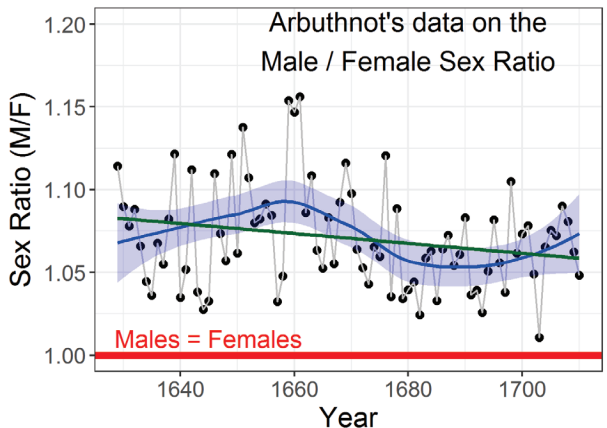
k	0	1	2	3	4	5	6	7	8	9	10	11	12
n_k	3	24	104	286	670	1033	1343	1112	829	478	181	45	7



Why did this matter? -- Arbuthot

John Arbuthot (1710) had gathered data on the ratios of male to female christenings in London from 1629-1710, 81 years

- He showed that in **EVERY** year Males > Females
- Calculated the **FIRST** significance test: like tossing 81 coins, and getting all heads
- Argued that the result must be due to divine providence: God likes more boys than girls!



Data: [HistData::Arbuthnot](#)

ggplot script:
<https://friendly.github.io/6135/R/arbutnot-gg.html>

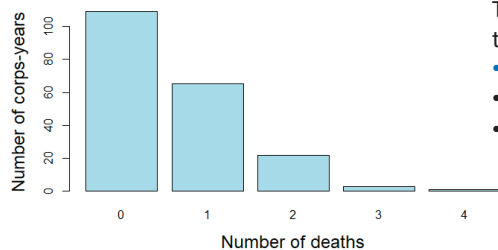
Example: Poisson

L. Von Bortkiewicz (1898) tallied the numbers of deaths by horse or mule kicks in 10 corps of the Prussian army over 20 years, → 200 corps-years

- In how many corps-years were there 0, 1, 2, ... deaths?
- This is among the earliest examples of a Poisson distribution



```
> data(HorseKicks, package="vcd")
> HorseKicks
nDeaths
 0    1    2    3    4
109  65  22   3   1
```



The Poisson distribution arises as that of the probability of 0, 1, 2, ...

- Rare events, that
- Occur with **constant** probability
- Forms the basis for `glm()` models with a count outcome.

5

Examples: count data

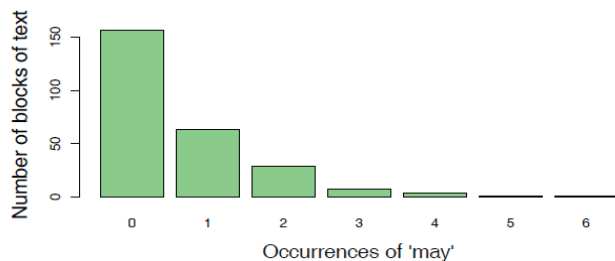
Federalist papers: Disputed authorship

- 77 essays by Alexander Hamilton, John Jay, James Madison to persuade voters to ratify the US constitution, all signed with pseudonym “Publius”
 - Who wrote each?
 - 65 known, 12 disputed (H & M both claimed sole authorship)
- Mosteller & Wallace (1984): analysis of frequency distⁿs of key “marker” words: from, **may**, whilst, upon, ...
- e.g., blocks of 200 words: occurrences (k) of “may” in how many blocks (n_k)

```
> data(Federalist, package = "vcd")
> Federalist
nMay
 0    1    2    3    4    5    6
156  63  29   8   4   1   1
```

6

Count data: models



For each word (“from”, “may”, “whilst”, ...)

- Fit a probability model [Poisson(λ), NegBin(λ, p)]
- Estimate parameters (λ, p)
- → Calculate log Odds (Hamilton vs. Madison)
- → All 12 disputed papers most likely written by **Madison**

(pioneered the use of cross-validation to assess model fit)

7

Example: Type-token distributions

- 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	St
Species (n_k)	6	12	6	9	9	6	10	10	11	5	3	3	5

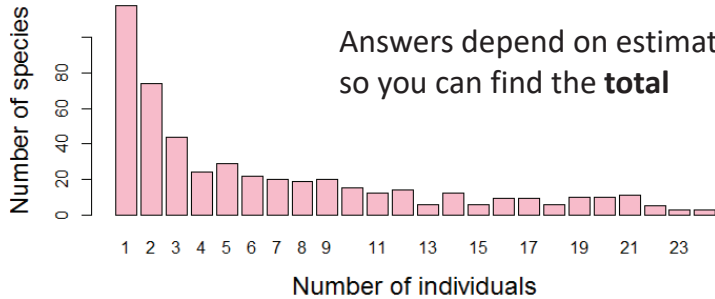
8

```
data(Butterfly, package="vcd")
barplot(Butterfly,
  xlab = "Number of individuals",
  ylab = "Number of species",
  col = "pink",
  cex.lab = 1.5)
```

Questions:

What is the total pop. of butterflies in Malaysia?
 How many wolves remain in Canada NWT?
 How many words did Shakespeare know?

Answers depend on estimating $\Pr(k=0)$,
 so you can find the **total**



Discrete distributions: Questions

- General questions
 - What process gave rise to the distribution?
 - What is the form: uniform, binomial, Poisson, negative binomial, ... ?
 - → Fit & estimate parameters
 - Visualize goodness of fit
 - → Use in some larger context to tell a story
- Examples
 - *Families in Saxony*: might expect $\text{Bin}(n=12, p)$; $p=0.5$?
 - *HorseKicks*: Poisson (λ); here, $\lambda = \text{mean} = 0.61$
 - *Federalist papers*: Perhaps Poisson(λ) or NegBin (λ, p)
 - *Butterfly data*: Perhaps a log-series distribution?

Fitting discrete distributions

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 binomial: allows for *overdispersion*, relative to Poisson

Motivation:

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

Common discrete distributions

Distribution	Counts, k	Values of X	$\Pr(X=k)$	Mean, $E(X)$	Var, $V(X)$
Bernoulli(p)	Success in 1 trial	$k=\{0, 1\}$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial(n,p)	# successes in n trials	$0, 1, \dots, n$	$\binom{n}{k} p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric(p)	# of trials to 1 st success	$0, 1, 2, \dots$	$p(1-p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Neg. binomial(k,p)	# of trials to k^{th} success	$0, 1, 2, \dots$	$\binom{n+k-1}{k} p^k(1-p)^{n-k}$	$\frac{k(1-p)}{p}$	$\frac{k(1-p)}{p^2}$
Poisson(λ)	# of events in interval	$0, 1, 2, \dots$	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ
Log series(p)	# of types observed	$0, 1, 2, \dots$	$\frac{p^k}{n \log(1-p)}$		

Discrete distributions: R

R functions: {d___, p___, q___, r___}

- d___ **density** function, $\Pr(X=k) = p(k)$
- p___ **cumulative** probability, $F(k) = \sum_{X \leq k} p(k)$
- q___ **quantile** function, find $k = F^{-1}(p)$, smallest value such that $F(k) \geq p$
- r___ **random** number generator

Discrete distribution	Density (pmf) function	Cumulative (CDF)	Quantile CDF ⁻¹	Random # generator
Binomial	dbinom()	pbinom()	qbinom()	rbinom()
Poisson	dpois()	ppois()	qpois()	rpois()
Negative binomial	dnbinom()	pnbinom()	qnbinom()	rnbinom()
Geometric	dgeom()	pgeom()	qgeom()	rgeom()
Logarithmic series	dlogseries()	plogseries()	qlogseries()	rlogseries()

```
e.g., > dbinom(0:4, size=4, p=1/2) # number of H in 4 coin tosses
[1] 0.0625 0.2500 0.3750 0.2500 0.0625
> dpois(0:4, lambda=3) # poisson, with λ = 3
[1] 0.0498 0.1494 0.2240 0.2240 0.1680
```

What is “binomial”

Bi-no-mi-al /bī¹nōmēāl/

- Taxonomy:** A **two-part name**, (genus, species) e.g., *Elephas maximus* for the Asian elephant; *Canis familiaris* for domestic dog
- Mathematics:** An algebraic expression of a sum of **two terms**, $(x + y)$ or expansion, $(x + y)^n$

$$\begin{aligned} (x+y)^0 &= 1 \\ (x+y)^1 &= 1x + 1y \\ (x+y)^2 &= 1x^2 + 2x^1y^1 + 1y^2 \\ (x+y)^3 &= 1x^3 + 3x^2y^1 + 3x^1y^2 + 1y^3 \\ (x+y)^4 &= 1x^4 + 4x^3y^1 + 6x^2y^2 + 4x^1y^3 + 1y^4 \\ (x+y)^5 &= 1x^5 + 5x^4y^1 + 10x^3y^2 + 10x^2y^3 + 5x^1y^4 + 1y^5 \end{aligned}$$

Coefficients of terms

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

(Pascal's triangle)

- Stats:** A distribution of number of “successes” in n Bernoulli trials w/ prob = p

Binomial distribution

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

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

ways to get k out of n
Pr(k events)
Pr(n-k non-events)

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? $\text{Bin}(10, 1/2)$
- Toss 12 fair dice— how many 5s or 6s? $\text{Bin}(12, 1/3)$

Mean, variance, skewness:

$$\text{Mean}[X] = n p$$

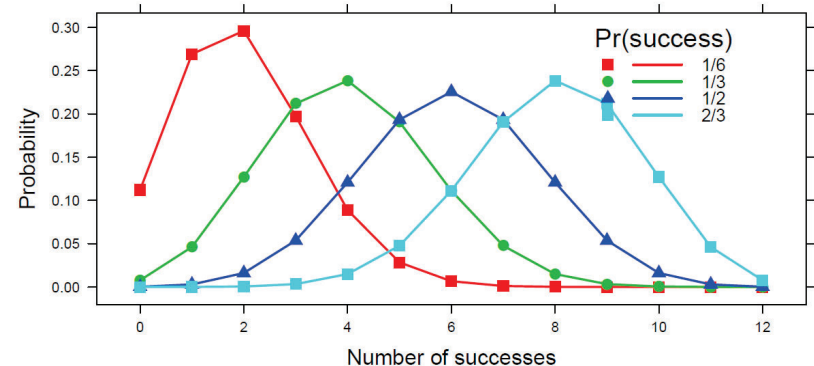
$$\text{MLE from data: } \hat{p} = \frac{\bar{x}}{n} = \frac{\sum_k k \times n k / \sum_k n_k}{n}$$

$$\text{Var}[X] = n p (1-p) = n p q$$

$$\text{Skew}[X] = n p q (q-p)$$

Binomial distribution

Binomial distributions for $k = 0, 1, 2, \dots, 12$ successes in $n=12$ trials, for 4 values of p



- Mean = $n p$
- Variance: $n p q$ is maximum when $p = 1/2$
- Skewed when $p \neq 1/2$

Code: [binomial.html](#)

DDAR Fig 3.9, pp 76-77

Poisson distribution

The Poisson distribution, $\text{Pois}(\lambda)$,

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

gives the probability of an event occurring $k = 0, 1, 2, \dots$ 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

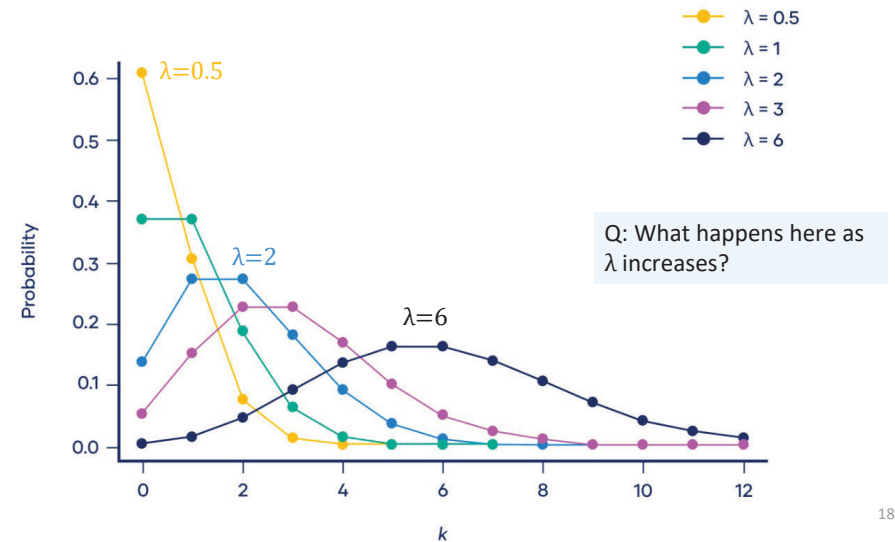
Table: Total goals scored in 380 games in the Premier Football League, 1995/95 season

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

17

Poisson distribution

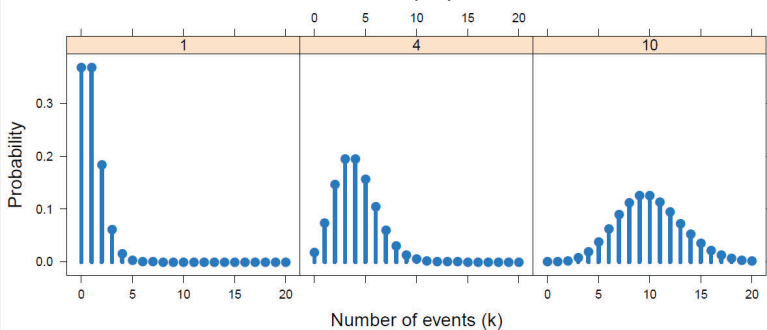
Poisson distributions for $\lambda = \frac{1}{2}, 1, 2, 3, 6$



18

Poisson distribution: Properties

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



DDAR Fig 3.10, p 81

Mean, variance, skewness:

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

$$\text{MLE: } \hat{\lambda} = \bar{x}$$

Properties:

Sum of $\text{Pois}(\lambda_1, \lambda_2, \lambda_3, \dots) = \text{Pois}(\sum \lambda_i)$
Approaches $N(\lambda, \lambda)$ as $n \rightarrow \infty$

19

History: Who discovered the "Poisson" distribution

Stigler's Law: No discovery in science is ever named for its primary originator

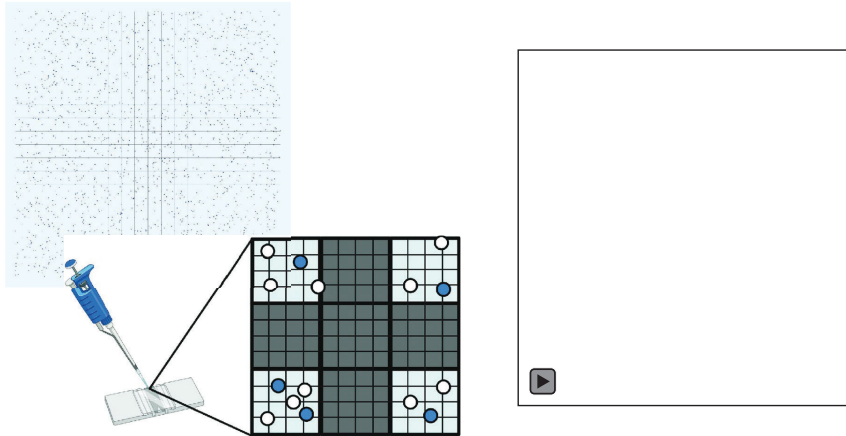
- De Moivre (1718) – Approximation to binomial as n gets largish
- **Poisson (1837)** – *Reserches sur la Probabilité des jugements en Matière criminelle...* -- Derives $e^{-\lambda} \lambda^k / k!$
 - Stigler says main result anticipated by De Moivre
- S. Newcomb (1860) – *Notes on a Theory of Probability*
 - First attempt at using this as a fit to **data**
 - Observations of stars: $\Pr(\text{any small space, } 1^\circ) \text{ contains } s \text{ stars, } s = 0, 1, 2, \dots$
- Von Bortkiewicz (1898) – *Law of Small Numbers*
 - Re-derives Poisson as limiting case of binomial
 - Several data sets (Horse kicks & others) – "agreement between theory and observation leave nothing to be desired"
- Gosset (1907): Heamacytometer Counts
 - "Student"'s first paper – first rigorous treatment of the Poisson for count data

See: Hanley & Bhatnagar (2022) The "Poisson" Distribution: History, Reenactments, Adaptations, *The American Statistician*, 76:4, 363-371, DOI: [10.1080/00031305.2022.2046159](https://doi.org/10.1080/00031305.2022.2046159)

20

Gosset: Hemacytometer Counts

Number of blood cells observed in a 20 x 20 grid on a slide



Source: <http://www.medicine.mcgill.ca/epidemiology/hanley/Gosset/>

21

Geometric distribution

- Distribution of the **number of trials before the first occurrence** of some event
 - Flip a coin: how many trials before 1st head?
 - Toss a die: number of tosses before 1st "6"?
 - Quality control: number of items inspected before finding a defective one

$$\text{Geom}(p): \Pr\{X=k\} = p(1-p)^k$$

1st head
k tails

$$\text{Mean: } E(X) = \frac{1-p}{p}$$

$$\text{Variance: } V(X) = \frac{1-p}{p^2}$$

Assumptions:

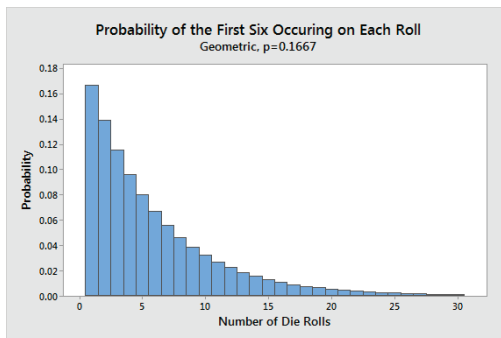
- Binary trials: Event / No event
- Independent trials
- Constant prob, p , on any trial

22

Geometric distribution: Tossing dice

```
> p <- 1/6
> x <- 0:10
> probs <- dgeom(x, prob = p)
> names(probs) <- x
> round(probs, 3)
  0      1      2      3      4      5      6      7      8      9     10
0.167 0.139 0.116 0.096 0.080 0.067 0.056 0.047 0.039 0.032 0.027
```

For `dgeom()`, x is failures before first success, so $k=1$ (success on trial 1) means $x=0$ failures



Q: Can you use this to test if dice are "fair", i.e., $p(k) = 1/6$?
How?

What distⁿ would give an easier test?

Ans: Binomial- obs. counts of 1:6

23

Negative binomial distribution

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

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

Pr(n success)
Pr(k failures)

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 $k = 0, 1, 2, \dots$ tails before $n = 3$ heads?

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

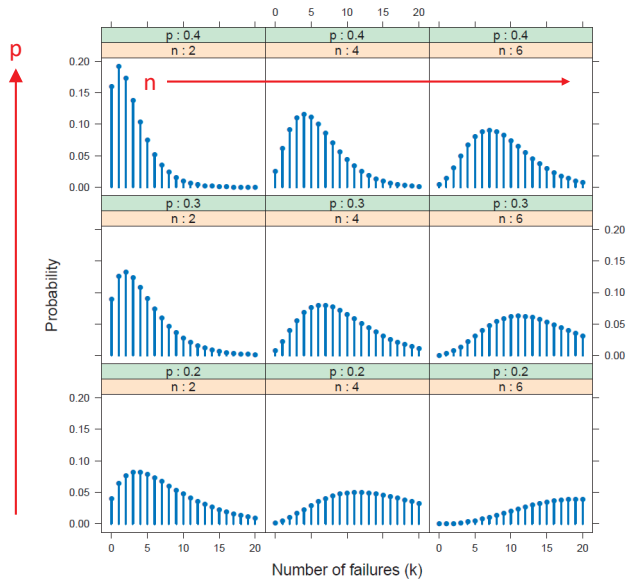
- constant probability p or independence are violated
- variance is greater than the mean (overdispersion)

Properties:

$$\text{Mean}(X) = \mu = \frac{n(1-p)}{p} \implies p = \frac{n}{n + \mu},$$

$$\text{Var}(X) = \frac{n(1-p)}{p^2} \implies \text{Var}(X) = \mu + \frac{\mu^2}{n}.$$

24



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

Mean:
 Increases with n
 Decreases with p

DDAR Fig 3.13, p 85

Quiz: Name that distribution

1. Weldon tossed 12 dice 26,306 times & tallied the number of times a **5 or 6** occurred

```
> WeldonDice
n56
  0   1   2   3   4   5   6   7   8   9  10+
185 1149 3265 5475 6114 5194 3067 1331 403 105 18
```

Bin($n=12, p=1/3$)

2. Pele practices penalty kicks for the upcoming 1958 FIFA World Cup. His average scoring has been $p=0.4$. What is the probability it will take him **1, 2, ... shots** to score a goal?

```
Nbin( $n=1, p=0.4$ )
> dnbinom(1:5, size=1, p=0.4)
[1] 0.240 0.144 0.086 0.052 0.031
```

3. A Geiger counter records the number of scintillations of α particles from a radioactive source, with an average rate of 20/msec. What is the probability of observing 40 in a 1 msec. interval?

Pois($\lambda=20$)

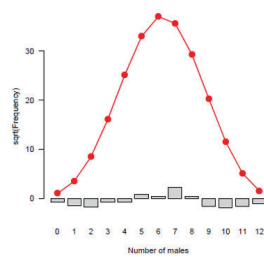
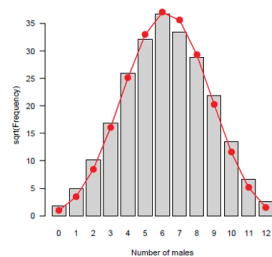
4. What is the distribution of the **time between** Geiger counter ticks?

Exponential distⁿ, $Pr(X=k) = \lambda e^{-\lambda k}$, mean = $1/\lambda$

Fitting discrete distributions

Fitting discrete distributions generally follows four main steps:

- **Visualize the data** to suggest candidate distributions.
- **Estimate the parameters** of the chosen distribution.
- **Evaluate the quality of the fit** graphically.
- **Perform goodness-of-fit statistical tests** to determine how well the chosen model represents the data



Goodness-of-fit test for
 χ^2 df $P(> \chi^2)$
 Likelihood Ratio 97 11 6.98e-16

Fitting discrete distributions: Details

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, $\hat{p} = \sum kn_k / (n \sum n_k) = \text{mean} / n$
 - Poisson, $\hat{\lambda} = \sum kn_k / \sum n_k = \text{mean}$
2. Calculate **fitted probabilities**, $\hat{p}(k)$ for the distribution, and then **fitted frequencies**, $N\hat{p}(k)$.
3. Assess **Goodness of fit**: Pearson χ^2 or likelihood-ratio G^2

$$\chi^2 = \sum_{k=1}^K \frac{(n_k - N\hat{p}_k)^2}{N\hat{p}_k} \quad G^2 = \sum_{k=1}^K n_k \log\left(\frac{n_k}{N\hat{p}_k}\right)$$

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

Fitting: Weldon's dice

Basic, naïve calculation of expected frequencies for a binomial distⁿ using `dbinom()`

```
> data(WeldonDice, package="vcd")
> Weldon.df <- as.data.frame(WeldonDice) # convert to data frame

> Prob <- dbinom(0:12, 12, 1/3) # binomial probabilities
> Prob <- c(Prob[1:10], sum(Prob[11:13])) # sum values for 10+
> Exp = round(sum(WeldonDice)*Prob) # expected frequencies
> Diff = Weldon.df[,"Freq"] - Exp # raw residuals
> Chisq = Diff^2 / Exp # contribution to chisquare
> data.frame(Weldon.df, Prob=round(Prob,5), Exp, Diff, Chisq)
  n56 Freq Prob Exp Diff Chisq
1    0  185 0.00771 203 -18 1.596
2    1 1149 0.04624 1216 -67 3.692
3    2 3265 0.12717 3345 -80 1.913
4    3 5475 0.21195 5576 -101 1.829
5    4 6114 0.23845 6273 -159 4.030
6    5 5194 0.19076 5018 176 6.173
7    6 3067 0.11127 2927 140 6.696
8    7 1331 0.04769 1255 76 4.602
9    8 403 0.01490 392 11 0.309
10   9 105 0.00331 87 18 3.724
11  10+ 18 0.00054 14 4 1.143
```

Doesn't calculate the MLE, \hat{p}
Manually sum $k \geq 10$

29

Fitting & graphing discrete distributions

In R, the `vcd` and `vcdExtra` packages provide functions to fit, visualize and diagnose discrete distributions

- **Fitting:** `goodfit()` fits uniform, binomial, Poisson, neg bin, geometric, logseries, ...
- **Graphing:** `rootogram()` assess departure between observed, fitted counts
- **Ord plot:** `Ordplot()` diagnose form of a discrete distribution
- **Robust plots:** `distplot()` handle problems with discrepant counts

30

Example: Saxony families

```
> data(Saxony, package="vcd")
> 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", par=list(size=12))
> summary(Sax.fit)
```

Goodness-of-fit test for binomial distribution

X ²	df	P(> X ²)	
Likelihood Ratio	97	11	6.98e-16

Doesn't fit. Why? Specify parameters

31

Example: Saxony families

The `print()` method for `goodfit` objects 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	0.933		2.140
1	24	12.089		3.426
2	104	71.803		3.800
3	286	258.475		1.712
4	670	628.055		1.674
5	1033	1085.211		-1.585
6	1343	1367.279		-0.657
7	1112	1265.630		-4.318
8	829	854.247		-0.864
9	478	410.013		3.358
10	181	132.836		4.179
11	45	26.082		3.704
12	7	2.347		3.037

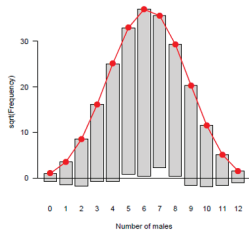
Pay attention to the signs & magnitudes of residuals, d_k

$$\text{Pearson } \chi^2 = \sum d_k^2$$

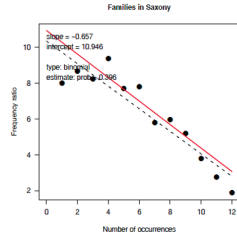
32

Graphing discrete distributions

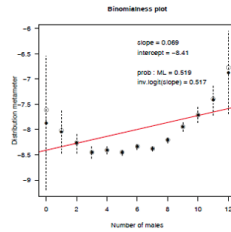
Rootograms



Ord plots



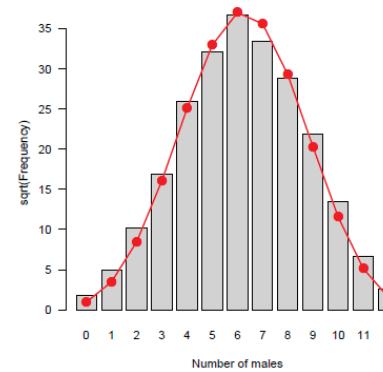
Robust distribution plots



What's wrong with simple histograms?

Discrete distributions are often graphed as histograms, with a theoretical fitted distribution superimposed
The plot() method for goodfit objects provides some alternatives

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

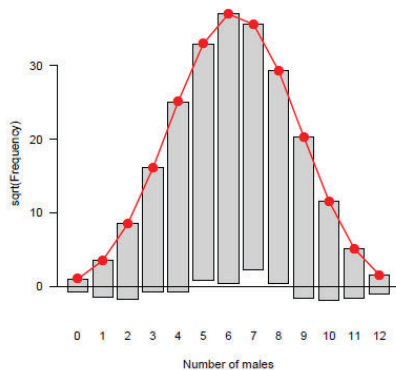


Problems:

- Largest frequencies dominate
- Must assess deviations vs. the fitted curve

Hanging rootograms

```
> plot(Sax.fit, type = "hanging", xlab = "Number of males") # default
```



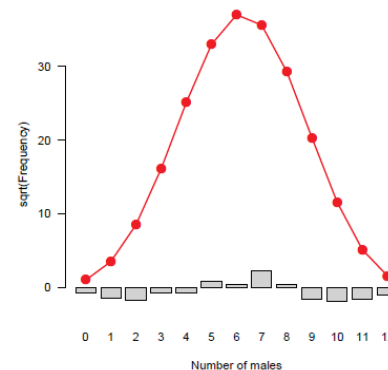
Tukey (1972, 1977):

- 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

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?

Q: What does this suggest about sex distribution of families in Saxony?

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 25.2 5 0.000125

This fits very poorly!
Why might this be the case?

Example: Federalist papers

Try 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 1.96 4 0.742

This now fits very well, indeed! Why?

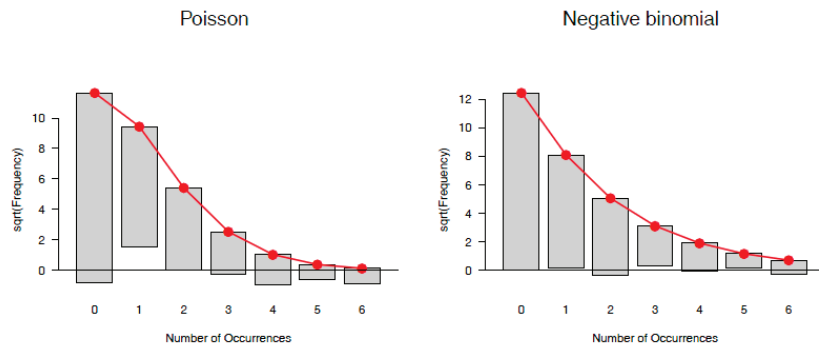
- Poisson assumes that the probability of a given word ("may") is constant across all blocks of text.
- Negative binomial allows the rate parameter λ to vary over blocks of text

In their landmark study, Mosteller & Wallace used a number of such words to distinguish between papers written by Madison & Hamilton and then assign the "unknown" author papers --- all to James Madison

Federalist papers: Rootograms

Hanging rootograms for the Federalist papers data, comparing Poisson and Negative binomial

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



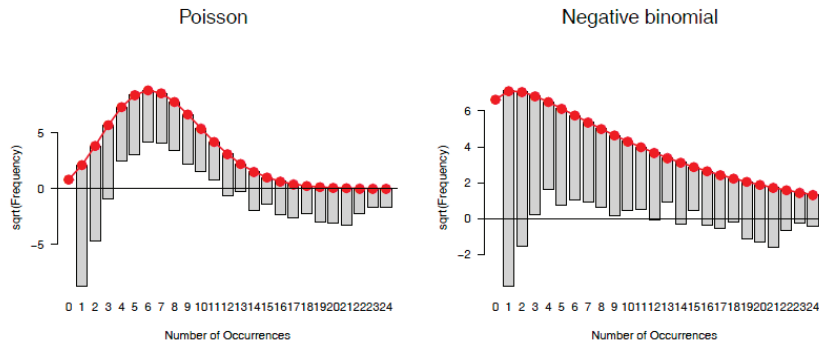
History corner: Why important?

- Resolved a long-standing historical debate
 - Authorship of 12 of 85 papers had been disputed: Madison or Hamilton?
 - Analysis of word frequencies → strong evidence all by Madison
- Established "stylometry" – statistical analysis of text for authorship
 - Word frequencies of "marker" words (by, upon, may) as reliable indicators
 - Did Shakespeare write all his plays? Or did Francis Bacon or Christopher Marlow write some?
- Pioneered Bayesian analysis: First large-scale Bayesian application
 - Prior prob: $p(\text{Hamilton}) = p(\text{Madison}) = \frac{1}{2}$
 - Update via parameters estimated from Nbin() distribution
- Introduced the idea of cross-validation
 - Leave-one-out → predict from the rest
- Heralded the age of computer-intensive methods

Butterfly data

Both Poisson and Negative binomial are terrible fits! What to do??

```
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 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 $k p_k / p_{k-1}$ against k is linear
 - signs of intercept and slope → determine the form, give rough estimates of parameters

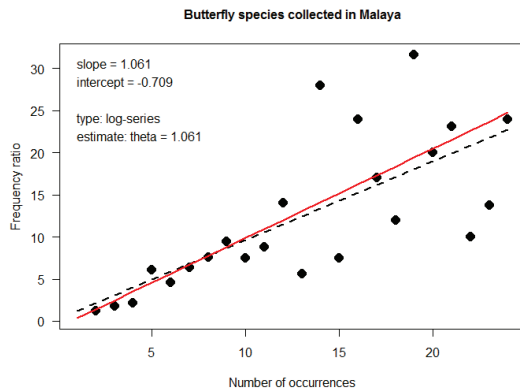
Slope (b)	Intercept (a)	Distribution (parameter)	Parameter estimate
0	+	Poisson (λ)	$\lambda = a$
-	+	Binomial (n, p)	$p = b / (b - 1)$
+	+	Neg. binomial (n, p)	$p = 1 - b$
+	-	Log. series (θ)	$\theta = b$ $\theta = -a$

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

Ord plot: Examples

Butterfly data: The slope and intercept correctly diagnoses the **log-series** distribution

```
> Ord_plot(Butterfly,
  main = "Butterfly species collected in Malaya",
  gp=gpar(cex=1), pch=16)
```



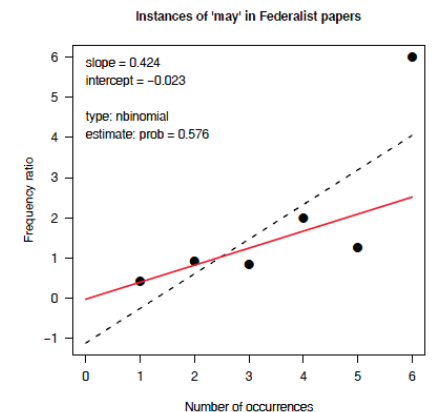
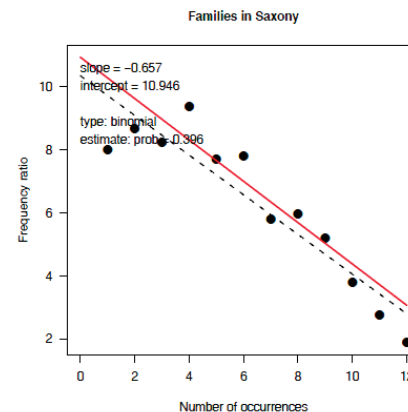
- + slope
- intercept
- log-series

OLS line shown in black
WLS line shown in red

Ord plots: Examples

Ord plots for the Saxony and Federalist data

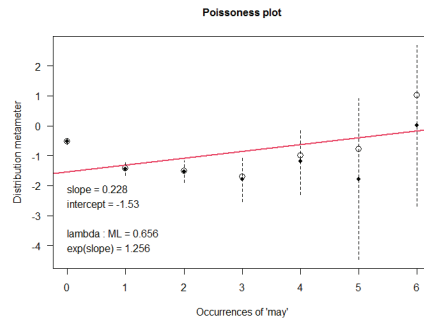
```
> Ord_plot(Saxony, main = "Families in Saxony", gp=gpar(cex=1), pch=16)
> Ord_plot(Federalist, main = "Instances of 'may' in Federalist papers", gp=gpar(cex=1), pch=16)
```



Robust distribution plots

- Ord plots lack robustness
 - one discrepant frequency, 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

For the Poisson distribution, this is called a "poissonness plot"



45

Poissonness plot: Details

- If the distribution of n_k is $\text{Poisson}(\lambda)$ 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 = $\log \lambda$
- Nonlinear relation \rightarrow distribution is *not* Poisson
- Hoaglin and Tukey (1985) give details on calculation of confidence intervals and influence measures.

46

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.

$$p(k) = a(k) \theta^k / f(\theta) \quad k = 0, 1, \dots$$

Table 3.10: The Power Series family of discrete distributions

Discrete Distribution	Probability function, $p(k)$	Series parameter, θ	Series function, $f(\theta)$	Series coefficient, $a(k)$
Poisson	$e^{-\lambda} \lambda^k / k!$	$\theta = \lambda$	e^θ	$1/k!$
Binomial	$\binom{n}{k} p^k (1-p)^{n-k}$	$\theta = p/(1-p)$	$(1+\theta)^n$	$\binom{n}{k}$
Negative binomial	$\binom{n+k-1}{k} p^n (1-p)^k$	$\theta = (1-p)$	$(1-\theta)^{-n}$	$\binom{n+k-1}{k}$
Geometric	$p(1-p)^k$	$\theta = (1-p)$	$(1-\theta)^{-1}$	1
Logarithmic series	$\theta^k / [-k \log(1-\theta)]$	$\theta = \theta$	$-\log(1-\theta)$	$1/k$

Math trick: This allows properties of the members of a family to be studied, related to each other and solve larger problems!

48

Other distributions

- This allows all of these to be represented in a plot of a suitable count metameter, $\phi(n_k)$ vs k . *DDAR* Table 3.12 gives details

Distribution	Probability function, $p(k)$	Count metameter, $\phi(n_k^*)$	Theoretical slope (b)	Theoretical intercept (a)
Poisson	$e^{-\lambda} \lambda^k / k!$	$\log(k! n_k^* / N)$	$\log(\lambda)$	$-\lambda$
Binomial	$\binom{n}{k} p^k (1-p)^{n-k}$	$\log(n_k^* / N \binom{n}{k})$	$\log\left(\frac{p}{1-p}\right)$	$n \log(1-p)$
Negative binomial	$\binom{n+k-1}{k} p^n (1-p)^k$	$\log(n_k^* / N \binom{n+k-1}{k})$	$\log(1-p)$	$n \log(p)$
Geometric	$p(1-p)^k$	$\log(n_k^* / N)$	$\log(1-p)$	$\log(p)$
Log series	$\theta^k / [-k \log(1-\theta)]$	$\log(k n_k^* / N)$	$\log(\theta)$	$-\log(-\log(1-\theta))$

Source: adapted from Hoaglin and Tukey (1985), Table 9-15.

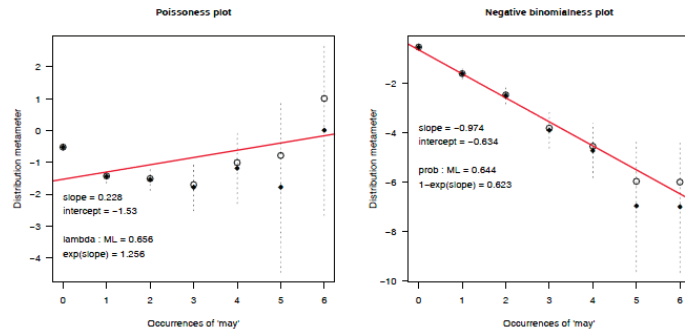
- In these plots, a straight line confirms that the data follow the given distribution.
- Confidence intervals around the points indicate **uncertainty** for the count metameter.
- The slope and intercept of the line give **estimates** of the distribution parameters.

49

distplot: Federalist

Try both Poisson & Negative binomial

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



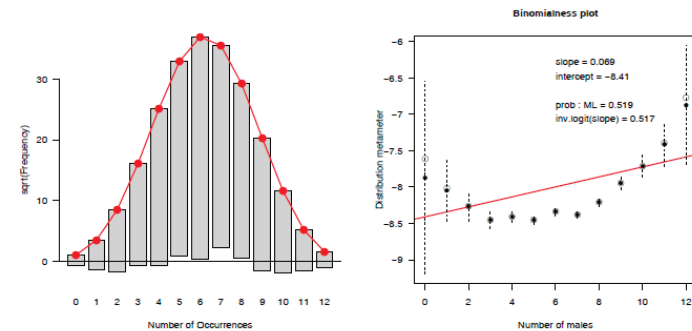
Again, the Poisson distribution is seen not to fit, while the Negative binomial appears reasonable.

50

distplot: 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 the binomial distribution. distplot() is more sensitive in diagnosing this

51

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.

52

What have we learned?

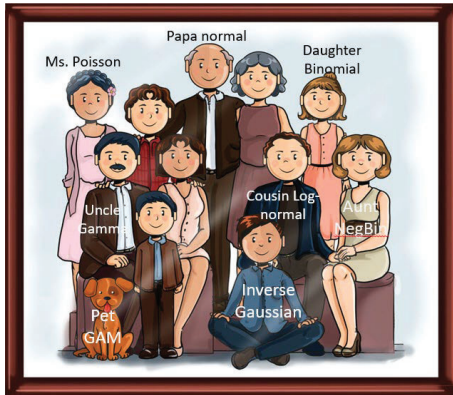
Some explanations:

- 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 $\hat{p} = \Pr(\text{male}) = 0.52$.
 - 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.

53

Looking ahead: The GLM family

These discrete distributions are also part of a larger family, the **exponential family**. This is the basis for the extension to generalized linear models, the **GLM family**.



Families share:

- **Same structure:** canonical form, “natural” parameters
- **Sufficient statistics:** all you need to estimate
- **Conjugate priors** in Bayesian analysis

Families allow:

- Generic **estimation**, like `glm()`
- Common **diagnostics** and **graphical** methods

Families provide: ❤️ mutual love, 👨‍👩‍👧 sense of belonging & identity, procreation!

54

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
```

N = 915, mean(articles) = 1.69

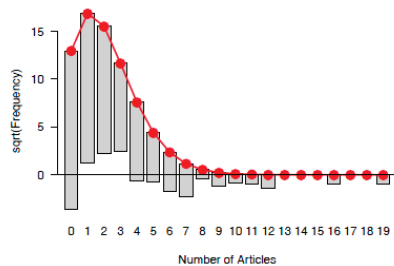
- There are **predictors**: gender, marital status, number of children, prestige of dept., # pubs by student’s mentor
- We fit such models with `glm()`, but need to specify the **form** of the distribution
- Ignoring the predictors for now, a baseline model could be `glm(articles ~ 1, data=PhdPubs, family = "poisson")`

55

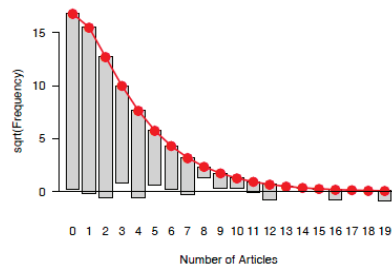
Looking ahead: PhdPubs

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

Poisson



Negative binomial



Poisson doesn’t fit: Need to account for **excess 0s** (some never published)
Neg binomial: Sort of OK, but should take predictors into account

56

Looking ahead: Count data models

Count data regression models (DDAR Ch 11)

- Include predictors
- Allow different distributions for **unexplained** variation
- Provide tests of one model vs. another
- Special models handle the problems of **excess zeros**: `zeroInfl()`, `hurdle()`

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

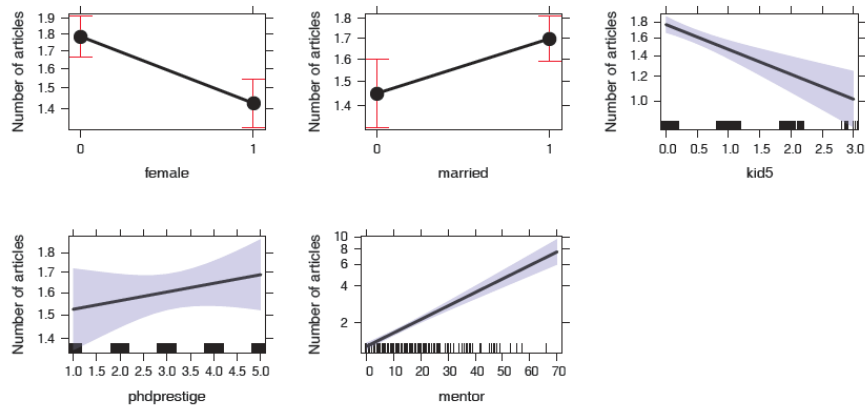
```
LRstats(phd.pois, phd.nbin)
```

```
## 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
```

57

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.

58

Summary

- Discrete distributions are the building blocks for categorical data analysis
 - Typically consist of basic counts of occurrences, with varying frequencies
 - Most common: binomial, Poisson, negative binomial
 - Others: geometric, log-series
- Fit with `goodfit()`; plot with `rootogram()`
 - Diagnostic plots: `Ord_plot()`, `distplot()`
- Models with predictors
 - Binomial → logistic regression
 - Poisson → poisson regression; logliner models
 - These are special cases of **generalized** linear models

59