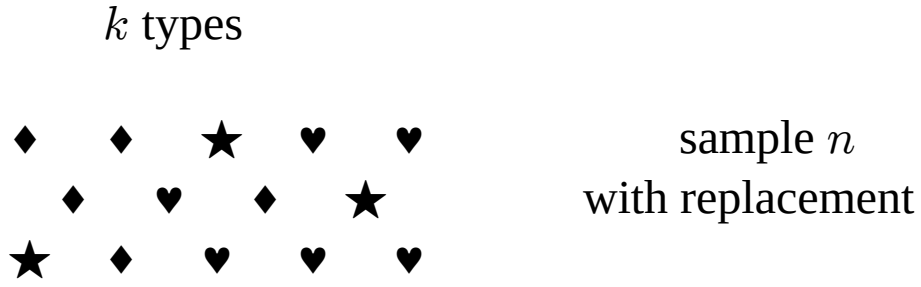
Multinomial models

The multinomial distribution is a generalization of the binomial distribution, for categorical variables with more than two response types. In a multinomial random experiment, each single trial results in one of k outcomes. For instance, sampling with replacement from a population with k types of individuals (like a bowl containing balls of k different colors) is a multinomial random experiment.



The multinomial distribution is a probability model for the counts resulting from such an experiment. Let:

$\pi_1 =$ proportion of type 1 (★) in the bowl
$\pi_2 =$ proportion of type 2 (♥) in the bowl
$\vdots$
$\pi_k =$ proportion of type k (♦) in the bowl

The π_j's are **constants** (parameters) in the distribution, and they sum to 1:

$$\pi_1 + \pi_2 + \cdots + \pi_k = 1.$$

A multinomial experiment occurs when we independently sample n objects from the bowl with replacement, and record the type of each one. Let:

$Y_1 =$ number of type 1 (★) in the **sample**
$Y_2 =$ number of type 2 (♥) in the **sample**
$\vdots$
$Y_k =$ number of type k (♦) in the **sample**

The counts Y_1, Y_2, …, Y_k are **random variables**. They take integer values between 0 and n inclusive, but they must also add to n:

$$Y_1 + Y_2 + \cdots + Y_k = n.$$

The Y_j's are **dependent**: the value of one affects the others. The multinomial distribution is the multivariate (or joint) probability distribution for the random variables Y_1, Y_2, …, Y_k:

$$P(Y_1 = y_1 \text{ and } Y_2 = y_2 \text{ and } \cdots \text{ and } Y_k = y_k)$$

$$= \frac{n!}{y_1! y_2! \cdots y_k!} \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k} ,$$

where y_1, y_2, ..., y_n are any nonnegative integers that add to n (all possible outcomes). We write

$Y_1, Y_2, \ldots, Y_k \sim$ multinomial(n, π_1, π_2, …, π_k) .

The binomial distribution is a special case corresponding to $k = 2$. In this case, the number of failures is $Y_2 = n - Y_1$.

Example. In political polling, each voter selected for the sample is categorized into one of k mutually exclusive categories. For instance, in a random sample of n voters in the USA,

$Y_1 =$ # democrats,
$Y_2 =$ # republicans,
$Y_3 =$ # independents (no party),
$Y_4 =$ # "other" (libertarians, greens, etc).

The counts Y_1, Y_2, Y_3, Y_4 are random variables, in that if the poll were to be repeated (a new sample of n voters selected) the counts would likely be somewhat different. Note that political polling and other such surveys are generally done by sampling without replacement. However, the multinomial model will be a good approximation to the distribution of the counts provided the changes in the proportions in the population (the π_j's) due to sampling are negligible. (The exact distribution of the counts for sampling without replacement is called the multivariate hypergeometric distribution).

Properties of the multinomial distribution

$\mathrm{E}(Y_j) = \pi_j n$

$\mathrm{V}(Y_j) = n\pi_j(1 - \pi_j)$

$$\text{Cov}(Y_i Y_j) \;=\; -\pi_i \pi_j \qquad (i \neq j)$$

Also, an important property of the multinomial distribution is that any subgroup of the Y_j's, conditional on their sum, has a multinomial distribution. For instance,

$$Y_1, Y_2, Y_3 \,|\, (Y_1 + Y_2 + Y_3 = m) \;\sim$$

$$\text{multinomial}(m, p_1, p_2, p_3)\ ,$$

where $p_i = \pi_i / (\pi_1 + \pi_2 + \pi_3)$.

Another important property is that any subgroup of the Y_j's can be pooled (summed) into one count, and the resulting collection of counts has a multinomial distribution, with the proportion for the new pooled category being the sum of the original probabilities for the counts being pooled. A special case of this property is that the marginal distribution of a single count is binomial:

$$Y_j \;\sim\; \text{binomial}(n, \pi_j)\ .$$

In this case, the number of failures is the sum of all the counts for the other categories.

Saturated parameter model

The multinomial distribution is called “saturated” if all of the population proportions $\pi_1, \pi_2, \ldots, \pi_k$ have unknown values. This translates to $k - 1$ free unknown parameters,

because knowing the values of $k-1$ of the proportions provides the value of the remaining proportion due to the constraint that the proportions sum to 1.

Suppose a multinomial experiment has been conducted. The data are the resulting counts, y_1, y_2, ..., y_k, and the "sample size" is the number of trials n. The likelihood function for the unknown parameters is

$$L(\pi_1, \pi_2, \ldots, \pi_k) = \frac{n!}{y_1!y_2!\cdots y_k!}\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_k^{y_k} ,$$

and the log-likelihood is

$$\log(L) = \log\left(\frac{n!}{y_1!y_2!\cdots y_k!}\right) + \sum_{j=1}^{k} y_j \log \pi_j .$$

The maximum likelihood (ML) estimates $\widehat{\pi}_1$, $\widehat{\pi}_2$, ..., $\widehat{\pi}_k$ are the values of π_1, π_2, ..., π_k that jointly maximize L or $\log(L)$, subject to the constraint that $\pi_1 + \pi_2 + \ldots + \pi_k = 1$. The constrained maximization has a symbolic solution that can be derived, for instance, with the Lagrange multiplier technique from calculus. The resulting solution is intuitive; the ML estimates are simply the sample proportions of the categories:

$$\widehat{\pi}_j = \frac{y_j}{n} .$$

The ML estimate for the mean of Y_j is found from $\widehat{\pi}_j$:

$$\widehat{\mathrm{E}}(Y_j) = n\widehat{\pi}_j = y_j .$$

Now $\widehat{\mathrm{E}}(Y_j)$ is the prediction for a new value of Y_j under the model, and the predictions fit all the observed counts y_j perfectly. The model has as many unknown parameters as free data points, hence the characterization of the model as "saturated."

One must be careful of terminology with the multinomial model when speaking about "sample size." The number of raw observations (each a single event, categorized into one of k categories) is n. The number of counts is k, of which $k-1$ are nonsingular random variables.

Reduced-parameter models

The saturated multinomial model has $k-1$ unknown free parameters for describing $k-1$ free counts. An interesting, and scientifically useful, class of multinomial models arise when the π_j's are constructed as functions of fewer underlying parameters. Many different functions arise out of various scientific mechanisms.

Examples of reduced-parameter models

1. *Hardy-Weinberg equilibrium*

Genes of two types (alleles), denoted A, a, present in a population at a location (locus) on a chromosome pair. Possible genotypes are AA, Aa, aa.

Bucket of gametes:

A A a
a A a A
a A A a A A

Gene proportions: p = proportion of A
$(1-p)$ = proportion of a

Random mating with no selection is like drawing two alleles at random (with replacement) from the bucket to make a new individual:

$P(\text{AA}) = p^2$
$P(\text{Aa}) = 2p(1-p)$ "Hardy-Weinberg proportions"
$P(\text{aa}) = (1-p)^2$

Draw sample of n individuals in population; determine their genotypes (AA, Aa, or aa). Y_1 = #AA, Y_2 = #Aa, Y_3 = #aa in sample.

Model: $Y_1, Y_2, Y_3 \sim$ multinomial(n, π_1, π_2, π_3)

where $\pi_1 = p^2$
$\pi_2 = 2p(1-p)$
$\pi_3 = (1-p)^2$

This is a multinomial model with *one* unknown parameter given by p.

Question for thought: suppose the counts resulting from sampling n individuals are denoted by y_1, y_2, y_3. What is the ML estimate for p?

2. *Bird-banding*

Capture (usually by mist netting) and band n adult birds, and release them back into the wild.

$r =$ prob. of surviving a given yr
$s =$ prob. that band is found & returned (given the bird dies) in the year of death

Study lasts for three years; $Y_1 =$ # band returns in first yr, $Y_2 =$ #band returns in second yr, $Y_3 =$ # band returns in third yr, $Y_4 =$ # birds with unreturned bands.

Model: $Y_1, Y_2, Y_3, Y_4 \sim$ multinomial$(n, \pi_1, \pi_2, \pi_3, \pi_4)$

$$\pi_1 = (1-r)s$$
$$\pi_2 = r(1-r)s$$
$$\pi_3 = r^2(1-r)s$$
$$\pi_4 = 1 - [(1-r)s + r(1-r)s + r^2(1-r)s]$$

Model has *two* unknown parameters r and s

3. *Fitting a probability distribution to grouped data*

The multinomial distribution is used for fitting probability distributions to data grouped into frequency counts. Suppose the set of real numbers is separated into k

intervals, with boundaries of known values given by s_1, s_2, ..., s_{k-1}. The boundaries are typically chosen by the investigator, but occasionally the boundaries are given by the scientific problem.

As an example, suppose X_1, X_2, ..., X_n represent a random sample drawn from a from a normal(μ, σ^2) distribution. Suppose also that these raw observations are grouped into interval categories, with the frequency counts represented by $Y_1, Y_2, \ldots, Y_k$:

$Y_1 =$ # of observations in $(-\infty, s_1)$
$Y_2 =$ # of observations in (s_1, s_2)
$Y_3 =$ # of observations in (s_2, s_3)
$\vdots$
$Y_{k-1} =$ # of observations in (s_{k-2}, s_{k-1})
$Y_k =$ # of observations in $(s_{k-1}, +\infty)$

The probabilities corresponding to these frequency counts are given by areas under a normal(μ, σ^2) curve:

$\pi_1 =$ area under normal curve from $-\infty$ to s_1
$\pi_2 =$ area under normal curve from s_1 to s_2
$\vdots$
$\pi_k =$ area under normal curve from s_{k-1} to $+\infty$

The multinomial model for the Y_j's has *two* unknown parameters: μ and σ^2.

Notes on picking boundaries: Setting the boundaries at values that cannot actually occur in the sample avoids the

need for specifying rules for categorizing an observation that lands on a border. For instance, one can define boundary values that have a greater number of decimal places than the precision of the raw observations. For discrete distributions defined on the integers, one can set the boundaries at half-integer values. Also, for adequate asymptotics (for the parameter estimates and the chisquare goodness of fit test), the intervals should be formed so as not to have low expected counts. The intervals do not have to be equal in width; rather, intervals roughly equal in expected frequency are better. Finally, to assure that the frequency counts have a multinomial distribution, the intervals should partition the entire sample space, so that every potential raw observation can be categorized into one of k mutually exclusive categories.

Maximum likelihood estimation for reduced parameter multinomial models

We write the general reduced parameter model as

$Y_1, Y_2, \ldots, Y_k \sim$ multinomial$(n, \pi_1, \pi_2, \ldots, \pi_k)$,

where the π_j's can be represented as functions of fewer underlying parameters (let's call the underlying parameters $\theta_1, \theta_2, ..., \theta_l$, with $l \leq k - 1$):

$$\begin{aligned}
\pi_1 &= g_1(\theta_1, \theta_2, ..., \theta_l) \\
\pi_2 &= g_2(\theta_1, \theta_2, ..., \theta_l) \\
&\vdots \\
\pi_k &= g_k(\theta_1, \theta_2, ..., \theta_l)
\end{aligned}$$

The data are the counts y_1, y_2, ..., y_k resulting from a sample of size n. The ML estimates of θ_1, θ_2, ..., θ_l are the values, say $\widehat{\theta}_1$, $\widehat{\theta}_2$, ..., $\widehat{\theta}_l$, that jointly maximize the multinomial log-likelihood function given by

$$\log(L) = \log\left(\frac{n!}{y_1!y_2!\cdots y_k!}\right) + \sum_{j=1}^{k} y_j \log g_j(\theta_1, \theta_2, ..., \theta_l) .$$

Numerical maximization is required for all but the simplest applications (such as the genetic H-W equilibrium, above). Interestingly, many computer packages for nonlinear least squares (SAS PROC NLIN, etc.) can be "tricked" into maximizing a multinomial likelihood instead of minimizing a sum of squares. The trick uses the counts y_1, y_2, ..., y_k as the observations of the "response variable" and the corresponding expected counts given by $\mathrm{E}(Y_j) = ng_j(\theta_1, \theta_2, ..., \theta_l)$, $j = 1, 2, ..., k$ as the regression model to be fit. The ordinary least squares estimates are *not* the ML estimates for the multinomial; the trick is to weight each observation by $1/[ng_j(\theta_1, \theta_2, ..., \theta_l)]$ and update the value of the weight with the new parameter values each iteration of the Gauss-Newton sum of squares algorithm. The "_WEIGHT_" statement in SAS PROC NLIN was designed for such use. Because the multinomial expected values typically are functionally different for each observation, the regression model form has to be specified separately for each observation by means of a series of "IF" statements. This "iteratively reweighted least squares" algorithm converges to the ML estimates for the

multinomial model, and the corresponding "weighted sum of squares" is the Pearson chisquare goodness of fit statistic. Furthermore, if the asymptotic variance-covariance matrix from the nonlinear regression is evaluated using $\sigma^2 = 1$ (where σ^2 is the variance parameter in the nonlinear regression), the result is the correct asymptotic variance-covariance matrix for the parameters of the multinomial model that arise from the Fisher information matrix. The "SIGSQ = 1" option in the PROC NLIN statement of SAS accomplishes the appropriate scaling of the variance covariance matrix for multinomial models.

Goodness of fit test for a reduced parameter multinomial model

The multinomial goodness of fit test is a statistical test of a specified reduced parameter model against a saturated model. The null hypothesis is a reduced parameter multinomial model for describing the Y_j's, that is, the π_j's can be represented in terms of fewer underlying parameters (θ_1, θ_2, ..., θ_l):

$$
\begin{aligned}
\text{H}_0\text{:}\quad & \pi_1 = g_1(\theta_1, \theta_2, ..., \theta_l) \\
& \pi_2 = g_2(\theta_1, \theta_2, ..., \theta_l) \\
& \qquad \vdots \\
& \pi_k = g_k(\theta_1, \theta_2, ..., \theta_l)
\end{aligned}
$$

Note: in goodness of fit tests, the null hypothesis is often the "research model" of interest.

The alternative hypothesis is that a more complex, unspecified multinomial model is necessary to describe the Y_j's, that is, all $k-1$ free parameters π_1, π_2, ..., π_{k-1} are needed to describe the data adequately:

$$\begin{aligned} \text{H}_1\text{:}\ \ & \pi_1 = \pi_1 \\ & \pi_2 = \pi_2 \\ & \quad \vdots \\ & \pi_{k-1} = \pi_{k-1} \end{aligned}$$

Here H_1 is the ordinary saturated multinomial. In situations where the null model is the research hypothesis (i.e. the model for which the researcher seeks to convince a skeptic of its merits), the alternative hypothesis becomes the skeptic's hypothesis. The skeptic would claim that the null hypothesis model is too simple a mechanism for producing the data.

The data are the counts y_1, y_2, ..., y_k. To perform the statistical test, one must calculate ML estimates for the null hypothesis model (usually using numerical maximization). The resulting maximized likelihood value, $\widehat{L}_0$, is evaluated as

$$\widehat{L}_0 = \frac{n!}{y_1!y_2!\cdots y_k!}\widetilde{\pi}_1^{y_1}\widetilde{\pi}_2^{y_2}\cdots\widetilde{\pi}_k^{y_k} \ .$$

where $\widetilde{\pi}_j = g_j\left(\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_l\right)$.

Similarly, the likelihood evaluated at the ML estimates for the alternative hypothesis is

$$\widehat{L}_1 = \frac{n!}{y_1!y_2!\cdots y_k!}\widehat{\pi}_1^{y_1}\widehat{\pi}_2^{y_2}\cdots\widehat{\pi}_k^{y_k}$$

where $\widehat{\pi}_j = \frac{y_j}{n}$. The likelihood ratio statistic for testing H_0 versus H_1 becomes

$$G^2 = -2\log\left(\frac{\widehat{L}_0}{\widehat{L}_1}\right).$$

For these multinomial models, the likelihood ratio statistic algebraically reduces to a simpler form, requiring calculation only of the ML estimates and expected values under H_0. Let $\widehat{E}_j = n\widetilde{\pi}_j$ (ML-estimated expected value of Y_j under the null hypothesis). The likelihood ratio goodness of fit statistic for the multinomial reduces to

$$G^2 = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{\widehat{E}_j}\right).$$

This is in the form 2 Σ observed $\log\left(\frac{\text{observed}}{\text{estimated expected}}\right)$.
Also, one can show by asymptotic expansion that

$$G^2 \approx \sum_{j=1}^{k} \frac{\left(y_j - \widehat{E}_j\right)^2}{\widehat{E}_j} = X^2,$$

that is, G^2 (likelihood ratio statistic) and Pearson's chisquare statistic are asymptotically similar.

If the data arise from the null hypothesis model, then G^2 and X^2 both have approximate chisquare distributions with $k-1-l$ df (# parameters estimated in alternative model minus # parameters estimated in null model). One rejects H_0 in favor of H_1 if $G^2 \geq \chi_\alpha^2$, where the chisquare percentile corresponds to $k-l-1$ df. One concludes that the null model "fits" if $G^2 < \chi_\alpha^2$.

In ordinary analysis of variance tests, the skeptic's hypothesis is usually the null model and the research hypothesis is the alternative. In goodness of fit tests, the null/alternative configuration of the skeptic's and research hypotheses is often reversed. As such, support for the research hypothesis in goodness of fit takes the form of "failure to reject the null hypothesis" and is thereby a weaker statement than rejection of the null in favor of the alternative. Acceptance of goodness of fit means only that the data are a plausible realization of the null model; it does not mean that the null model (or some approximation thereof) necessarily generated the data.

Goodness of fit is an important ingredient of model evaluation, but it is not the only ingredient. A statistical model ultimately is a scientific hypothesis that purports to explain how numerical observations arise, and evaluating the model's reliability as a predictive tool and as a supporting strand in a web of other scientific hypotheses requires further evidence. One might think of goodness of fit as scientifically necessary, but not scientifically sufficient.

Examples of goodness of fit tests

1. *Are human blood types in H-W proportions?*

3 alleles A, B, O with proportions a, b, o in population (where $a + b + o = 1$; there are *two* unknown parameters)

genotypes	phenotypes	frequencies (H-W)
AA, AO	type A	$\pi_1 = a^2 + 2ao$
BB, BO	type B	$\pi_2 = b^2 + 2bo$
AB	type AB	$\pi_3 = 2ab$
OO	type O	$\pi_4 = o^2$

Data (from Rao CR. 1973. Linear Statistical Inference and its Applications. Wiley): $y_1 = 182,\ y_2 = 60,\ y_3 = 17,\ y_4 = 176,\ n = 435$

ML estimates under H_0: H-W proportions (computer maximization of L_0):

$\widehat{a} = 0.2644485$

$\widehat{b} = 0.09319721$

$\widehat{o} = 0.6423543$

$\log\left(\widehat{L}_0\right) = -9.096694$

	observed
$\widehat{E}_1 = n\widetilde{\pi}_1 = n\left(\widehat{a}^2 + 2\widehat{a}\widehat{o}\right) = 178.20741$	182
$\widehat{E}_2 = n\widetilde{\pi}_2 = n\left(\widehat{b}^2 + 2\widehat{b}\widehat{o}\right) = 55.86139$	60
$\widehat{E}_3 = n\widetilde{\pi}_3 = n\left(2\widehat{a}\widehat{b}\right) = 21.44190$	17
$\widehat{E}_4 = n\widetilde{\pi}_4 = n\left(\widehat{o}^2\right) = 179.48931$	176

$$G^2 = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{\widehat{E}_j}\right) = 1.438994$$

$\mathrm{df} = 4 - 2 - 1 = 1$

$p = 0.2303022$

$(X^2 = 1.375346)$

$p > 0.05$ so do not reject H_0

```
#  R program to calculate maximum likelihood (ML) estimates for parameters
#  a b and o (allele frequencies) in the multinomial model for
#  Hardy-Weinberg proportions in human blood types.  The model is
#  Y1,Y2,Y3,Y4 ~ multinomial(n,p1,p2,p3,p4) where
#  p1=a+a + 2*a*o  blood type A proportion
#  p2=b*b + 2*b*o   blood type B proportion
#  p3=2*a*b          blood type AB proportion
#  p4=o*o            blood type O proportion
#
#  Here 0<a<1, 0<b<1, o=1-a-b.

#  Count frequencies are entered into the vector yy here.
yy=c(182,60,17,176);

#  Data in example are from:  Rao, CR 1973. Linear statistical inference and
#  its applications.  Wiley.
```

```
#  Set initial patameter values here.
a0=.33;
b0=.33;

#  ML objective function "negloglike.ml" is negative of log-likelihood;
#  the Nelder-Mead optimization routine in R, "optim", is a minimization
#  routine.  The two function arguments are:  theta = vector of
#  parameters (transformed to real line), ys = vector of frequencies.

negloglike.ml=function(theta,ys)
{
  a=exp(-exp(theta[1]));     #  Constrains 0 < a < 1.
  b=exp(-exp(theta[2]));     #  Constrains 0 < b < 1.
  o=1-a-b;
  n=sum(ys);
  k=length(ys);
  p=c( a*a+2*a*o,
       b*b+2*b*o,
       2*a*b,
       o*o );             #  H-W model for the probabilities.

  ofn=-sum(ys*log(p));     #  No need to calculate all the factorials.
  return(ofn);
}

# The ML estimates.
MULTML=optim(par=c(log(-log(a0)),log(-log(b0))),
  negloglike.ml,NULL,method="Nelder-Mead",ys=yy);
reslts=c(exp(-exp(MULTML$par[1])),exp(-exp(MULTML$par[2])),-MULTML$val);
a.ml=reslts[1];           # These are the ML estimates.
b.ml=reslts[2];           #         --
o.ml=1-a.ml-b.ml;      #         --

nn=sum(yy);
loglike.ml=reslts[3]+lfactorial(nn)-sum(lfactorial(yy));

# Calculate expected values, LR statistic, etc.
pp=c( a.ml*a.ml+2*a.ml*o.ml,
       b.ml*b.ml+2*b.ml*o.ml,
       2*a.ml*b.ml,
```

```
      o.ml*o.ml );
EE=nn*pp;
y1=yy;
y1[y1==0]=1;                          # Guard against log(0) in G-squared.
Gsq=2*sum(yy*log(y1/EE));   # G-squared goodness of fit statistic.
pval=1-pchisq(Gsq,1);

#  Print the results.
a.ml;
b.ml;
o.ml;
loglike.ml;
Gsq;
pval;
cbind(EE,yy);
```

2. *Do lodgepole pines have a Poisson spatial distribution?*

100 quadrats placed at random in a lodgepole pine forest.
$X =$ # trees in a quadrat.

Poisson model: $\mathrm{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, ...$

Frequency counts:
$Y_1 =$ # quadrats with 0 trees
$Y_2 =$ # quadrats with 1 tree
$Y_3 =$ # quadrats with 2 trees
$\vdots$
$Y_{k-1} =$ # quadrats with $k-2$ trees
$Y_k =$ # quadrats with $k-1$ or greater trees

Multinomial distribution of frequency counts:

$Y_1, Y_2, ..., Y_k \sim$ multinomial($n, \pi_1, \pi_2, ..., \pi_k$)

where

$$\pi_1 = \mathrm{P}(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} \quad (= g_1(\lambda))$$
$$\pi_2 = \mathrm{P}(X = 1) = \frac{e^{-\lambda}\lambda^1}{1!} \quad (= g_2(\lambda) \text{ etc.})$$
$$\pi_3 = \mathrm{P}(X = 2) = \frac{e^{-\lambda}\lambda^2}{2!}$$
$$\vdots$$
$$\pi_{k-1} = \mathrm{P}(X = k-2) = \frac{e^{-\lambda}\lambda^{k-2}}{(k-2)!}$$
$$\pi_k = \mathrm{P}(X \geq k-1) = 1 - (\pi_1 + \pi_2 + \cdots + \pi_{k-1})$$

# trees per quadrat (x)	frequency count (y_j)	estimated expected frequency ($ng_j\left(\widehat{\lambda}\right)$)
0	7	5.728991
1	16	16.382799
2	20	23.424378
3	24	22.328357
4	17	15.962714
5	9	9.129494
6	5	4.351164
≥ 7	2	2.692104

$\widehat{\lambda} = 2.859631$
$\log\left(\widehat{L}_0\right) = -13.97714$
$G^2 = 1.276895 \;\; p = 0.2584772$
$X^2 = 1.260605 \;\; p = 0.2615366$

```
#  R program to calculate maximum likelihood (ML) estimates for parameter
#  in the Poisson distribution using a multinomial likelihood for goodness of
#  fit test.
#  Here P[X=x] = exp(-lambda)(lambda^x)/(x!), x=0,1,2,... .
#  Data are pooled frequency counts:
#              y1 = #{X=0}, y2 = #{X=1},..., yk = #{X>=(k-1)}.

#  Count frequencies are entered into the vector yy here.  The last frequency
#  is the pooled tail counts.
yy=c(7,16,20,24,17,9,5,2);

#  Initial patameter value is calculated from the approximate sample mean.
kk=length(yy);
nn=sum(yy);
xx=0:(kk-1);
lambda0=sum(xx*yy)/nn;

#  ML objective function "negloglike.ml" is negative of log-likelihood;
#  the optimization routine in R, "optim", is a minimization
#  routine.  The two function arguments are:  theta = parameter
#  (transformed to real line), ys = vector of frequencies.

negloglike.ml=function(theta,ys)
{
  lambda=exp(theta);     #  Constrains 0 < lambda.
  k=length(ys);
  x=0:(k-1);
  x1=x[1:(k-1)];
  p=rep(0,k);
  p[1:(k-1)]=exp(-lambda+x1*log(lambda)-lfactorial(x1));
  p[k]=1-sum(p[1:(k-1)]);
  ofn=-sum(ys*log(p));     #  No need to calculate all the factorials.
  return(ofn);
}

# The ML estimate.
MULTML=optim(par=log(lambda0),
  negloglike.ml,NULL,method="BFGS",ys=yy);  #  Nelder-Mead algorithm is not
                                                    #  reliable for 1-D problems.
reslts=c(exp(MULTML$par[1]),-MULTML$val);
```

```
lambda.ml=reslts[1];           # This is the ML estimate.
nn=sum(yy);
loglike.ml=reslts[2]+lfactorial(nn)-sum(lfactorial(yy));   #  Log-likelihood.

# Calculate expected values, LR statistic, etc.
xx1=xx[1:(kk-1)];
pp=rep(0,kk);
pp[1:(kk-1)]=exp(-lambda.ml+xx1*log(lambda.ml)-lfactorial(xx1));
pp[kk]=1-sum(pp[1:(kk-1)]);
EE=nn*pp;
y1=yy;
y1[y1==0]=1;                      # Guard against log(0) in G-squared.
Gsq=2*sum(yy*log(y1/EE));  # G-squared goodness of fit statistic.
pvalG=1-pchisq(Gsq,1);         # p-value (chisquare distribution) for G-squared.
Xsq=sum((yy-EE)^2/EE);      # Pearson goodness of fit statistic.
pvalX=1-pchisq(Xsq,1);         # p-value (chisquare distribution) for Pearson.

#  Print the results.
lambda.ml;
loglike.ml;
Gsq;
pvalG;
Xsq;
pvalX;
cbind(EE,yy);
```

Confidence intervals

1. *Wald confidence intervals*. The curvature of the log-likelihood function near its peak describes how well the data distinguish among different nearby parameter values. If the log-likelihood function is a narrow, steep peak, then curvature is large and the data are providing good information for estimating the parameter values. If the log-likelihood function slopes down from its peak only gently, a large range of parameter values provide a log-likelihood value almost as high as that of the ML estimate, and the

data do not provide good information for parameter estimation.

Commonly used measures of the quality of estimation are based on the log-likelihood curvature. The log-likelihood curvature is defined by the Hessian matrix of second derivatives. The element in the ith row and the mth column of the Hessian matrix is:

$$\frac{\partial^2 \log L(\theta)}{\partial\theta_i \partial\theta_m} = \sum_{j=1}^{k} \left\{ \frac{y_j}{g_j(\theta)} \frac{\partial^2 g_j(\theta)}{\partial\theta_i \partial\theta_m} - \frac{y_j}{g_j^2(\theta)} \left(\frac{\partial g_j(\theta)}{\partial\theta_i} \right) \left(\frac{\partial g_j(\theta)}{\partial\theta_m} \right) \right\} .$$

Here $\theta = (\theta_1, \theta_2, ..., \theta_l)$. In hypothetically repeated samples, $\log L(\theta)$ and its derivatives are random variables. In particular, the elements of the Hessian matrix are linear functions of the mulinomial observations $y_1, y_2, ..., y_k$, and their expected values can be found by substituting the expected values $ng_1(\theta), ng_2(\theta), ..., ng_k(\theta)$ of the observations. The element in the ith row and the jth column in the matrix of expected values of the second derivatives is

$$\mathrm{E}\left[\frac{\partial^2 \log L(\theta)}{\partial\theta_i \partial\theta_m} \right] = -n \sum_{j=1}^{k} \left\{ \frac{1}{g_j(\theta)} \left(\frac{\partial g_j(\theta)}{\partial\theta_i} \right) \left(\frac{\partial g_j(\theta)}{\partial\theta_m} \right) \right\} .$$

The expression results from the fact that $\Sigma g_j(\theta) = 1$ thereby producing $\Sigma \partial^2 g_j(\theta) / (\partial\theta_i \partial\theta_m) = 0$.

The essential curvature of the log-likelihood near its maximum is negative; a measure of the amount of

information in the data toward parameter estimation is better scaled in the positive direction. Such a measure of estimation quality is provided by the "Fisher information matrix," consisting of the expected values of the Hessian matrix elements multiplied by -1:

$$I(\theta) = \left\{ -\mathrm{E}\left[\frac{\partial^2 \log L(\theta)}{\partial \theta_i \partial \theta_m} \right] \right\}.$$

An important result from statistical theory states that the ML parameter estimates in $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_l)$ have, under hypothetical repeated sampling, an asymptotic multivariate normal distribution with mean vector θ and variance-covariance matrix given by

$$V(\theta) = [I(\theta)]^{-1}.$$

Asymptotically valid confidence intervals, termed "Wald intervals," are constructed with this multivariate normal distribution. The variance covariance matrix $V(\theta)$ can be estimated by substituting any statistically consistent estimate of the Fisher information matrix $I(\theta)$. One such estimate is simply the ML estimate $I\left(\widehat{\theta}\right)$. Alternatively the information matrix can be estimated by using the Hessian matrix itself. The Hessian evaluated at the ML estimates $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_l)$ and multiplied by -1 is called the "observed information matrix":

$$J\left(\widehat{\theta}\right) = \left\{ - \frac{\partial^2 \log L\left(\widehat{\theta}\right)}{\partial\theta_i \partial\theta_m} \right\}.$$

Either $\left[I\left(\widehat{\theta}\right)\right]^{-1}$ or $\left[J\left(\widehat{\theta}\right)\right]^{-1}$ are statistically consistent estimates of $V(\theta)$.

A Wald interval is an asymptotic $100(1-\alpha)\%$ confidence interval for θ_i formed by

$$\widehat{\theta}_i \pm z_{\alpha/2}\sqrt{\widehat{v}_{ii}}\,,$$

where $z_{\alpha/2}$ is the $100[1-(\alpha/2)]$th percentile of a standard normal distribution and $\widehat{v}_{ii}$ is the ith element in the main diagonal of the estimated variance-covariance matrix $\widehat{V}(\theta)$.

In practice, sample sizes are frequently not large enough to attain "asymptopia," and actual coverage probabilities of asymptotic confidence intervals can be considerably different from the claimed probabilities. Wald intervals are known to be often too small. Simulations suggest that intervals using the observed information matrix $J\left(\widehat{\theta}\right)$ tend to have slightly better properties than those using the ML Fisher infomation matrix $I\left(\widehat{\theta}\right)$, but actual details differ from model to model.

2. *Bootstrap confidence intervals*. The key problem for constricting a confidence interval for a parameter θ_i is to

estimate how variable its ML estimate $\hat{\theta}_i$ is under hypothetical repeated sampling (the so-called sampling distribution of $\hat{\theta}_i$). Bootstrapping is a straightforward, computer-intensive appproach to "estimating the variability of estimation," and is the statistical version of "pulling yourself up by your bootstraps." The basic principle is easy: obtain an estimate of the model, simulate thousands of new data sets from the estimated model, and refit the model to each of the new data sets. The resulting collection of thousands of parameter values forms a statistically consistent estimate of the sampling distribution.

The estimated model in the present context is the multinomial distribution evaluated at the ML parameter estimates $\widehat{\theta}$. One computer-generates b multinomial data sets (say, $b = 10,000$) from the estimated model and re-calculates ML parameter estimates for each data set. Each bootstrap data set should have the same sample size as the original data. The resulting bootstrap parameter estimates $\widehat{\theta}^{(1)}$, $\widehat{\theta}^{(2)}$, ..., $\widehat{\theta}^{(b)}$ can be treated as a huge sample from the sampling distribution of $\widehat{\theta}$. An asymptotic 95% confidence interval for θ_i, for instance, is given by the empirical 2.5th and 97.5th percentiles of the bootstrap estimates of θ_i.

3. *Profile likelihood confidence intervals.* A confidence interval can be constructed by inverting a two-sided hypothesis test concerning the parameter in question. The null hypothesis is that the parameter is equal to a constant known value:

H_0: $\theta_i = \theta_{i0}$.

The alternative hypothesis is that the parameter is an unknown constant:

H_1: $\theta_i \neq \theta_{i0}$.

A valid $100(1-\alpha)$% confidence interval is the set of all values of θ_{i0} for which the null hypothesis would not be rejected, using a significance level of α.

A profile likelihood confidence interval for θ_i uses the generalized likelihood ratio test and the asymptotic chisquare distribution of the test statistic to form the confidence interval. For a range of fixed values of θ_i, one maximizes the log-likelihood with respect to the remaining unknown parameters. Usually the process requires many numerical maximizations, one for each new value of θ_i. Typically the maximized values of the log-likelihood are plotted versus the corresponding θ_i values, ideally resulting in a unimodal curve, like a rounded mountain. The summit of the curve represents $\log \widehat{L}_1$, the log-likelihood maximized under the alternative hypothesis (full ML estimates of all parameters including θ_i). The confidence interval is all the θ_i values for which the maximized log-likelihood is above a certain altitude on the peak. The threshold is given by the asymptotic chisquare distribution of the likelihood ratio test statistic. A valid asymptotic $100(1-\alpha)$% confidence interval is the set of θ_i values for which

$$G^2 \;=\; 2\Big(\log \widehat{L}_1 \;-\; \log \widehat{L}_0\Big) \;\leq\; \chi^2_\alpha\,,$$

where the chisquare distribution has 1 df. For example, a 95% confidence interval is the set of θ_i values for which the height of the log $\widehat{L}_0$ curve is within a vertical distance of $\chi^2_\alpha/2 = \; 3.84/2 = 1.92$ from the summit.

Beyond its use for constructing confidence intervals, a profile log-likelihood plot is a recommended ingredient of model evaluation. A nearly flat or multimodal profile can warn of estimability problems; different sets of parameters are producing similarly high log-likelihoods. An ideal profile looks parabolic. Many tough-to-estimate parameters (such as population size in mark recapture models) produce asymmetric profiles, with a steep decline on one side and gentle decline on the other, leading to confidence intervals that are highly asymmetric around the ML point estimate.

In general, decades of simulations of many models in the statistical literature suggests that both bootstrap and profile likelihood confidence intervals tend to perform better than Wald intervals. Neither bootstrap nor profile likelihood intervals produce interval boundaries outside the range of the parameter, a phenomenon which can occur with Wald intervals (estimated survival probabilities less than zero, etc.). Bootstrap and profile likelihood intervals tend to have actual coverage probabilities for moderate-sized samples that are adequately close to the claimed coverage

probabilities, while the actual Wald coverage probabilities tend to be too small. Nonetheless, there are situations for which all three can be bad. Statistical theory gives a few warnings: a parameter at or near the edge of its range, a random variable with a range that depends on an unknown model parameter, and sparse data are situations that can cause estimation problems. However, statistical theory provides no sweeping guidance for confidence intervals across model families; basically, each new model for each new application has to have its confidence intervals tested by simulation, for the conditions and sample sizes likely to be encountered in practice. There will never be a shortage of topics for statistical masters degrees.