Statistical Precursors to the “New” Predictive Analytics:
Description, Prediction, Prescription

Stanley L. Sclove[*]

Abstract
A paradigm providing a context for Statistical Analysis is that we want to proceed from Data to Information to Knowledge to Decisions, with Statistical Analysis occurring primarily between Data and Information. A paradigm, found in papers and textbooks, for the new “Predictive Analytics”, is: Description / Prediction / Prescription. In the discussion itself, we mentioned some precursors to the elements of the predictive-analytics paradigm. In this proceedings paper, we go into a bit more detail and include some material on Compound Models and Predictive Distributions.

Key Words: ASA Section on Risk Analysis; Mean squared error, Mean squared error of prediction; Compound models; prior distribution, conditional distribution; marginal distribution, predictive distribution

Contents

1 Introduction and Background **2**
1.1 Introduction: Predictive Analytics . . . 2
1.2 Description . . . 2
1.3 From Description to Prediction . . . 2
1.4 Prescription . . . 4

2 Levels of Granularity **4**
2.1 An Example . . . 4
2.2 Comparison of Models by Model-Selection Criteria . . . 5

3 An Example with Continuous Data **8**

4 Compound Models **9**
4.1 Probability Function Notation . . . 9
4.2 Compound Models . . . 9

5 The Gamma-Poisson Model **11**
5.1 Probability Functions for the Gamma-Poisson Model . . . 11
5.2 Gamma Family of Distributions . . . 11
5.3 Exponential Family of Distributions . . . 11
5.4 Development of the Gamma-Poisson Model . . . 11
5.4.1 The Joint Distribution of X and Λ . . . 12
5.4.2 Posterior Distribution of Λ . . . 13
5.4.3 Marginal Distribution of X . . . 13
5.4.4 Moments . . . 13
5.5 Empirical Bayes estimation . . . 13
5.5.1 The hyperparameters in terms of the moments of the marginal . . . 13
5.5.2 Estimating the prior parameters from the marginal . . . 13
5.6 Application to the Days Ill dataset . . . 13

6 Some Other Compound Models: Beta-Binomial; Normal-Normal **14**
6.1 Beta-Binomial Model . . . 14
6.2 Normal-Normal Model . . . 15

[*] Department of Information & Decision Sciences, University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607-7124. Writing of this paper was initiated on April 14, 2016.

1. Introduction and Background

This Roundtable Discussion was sponsored by the Section on Risk Analysis. I was invited by Professor Susan Simmons (North Carolina State University), former section chair, and Professor Yishi Wang (University of North Carolina - Wilmington), an organizer for the section's program, to present a roundtable discussion. I've recently been active in curriculum development and teaching in our new MS in Business Analytics at UIC and so decided to discuss some aspects of the related field of Predictive Analytics.

Background on the Risk Analysis Section. The Risk Analysis section grew out of a review boad of the U.S. Nuclear Regulatory Commission) on LPHR – low probability, high risk – events. Among the members of the committee were Bernie Harris, Lee Abrahamson, Harry Martz and Lisa Weissfeld.

When their work was done, rather than disband altogether, they decided to petition to form a new section of ASA, the Section on Risk Analysis. Bernie Harris and I were well acquainted through previous statistically-related activities. Bernie asked me to sign the petition for the formation of the new section. He said that they needed someone from a business school (which I am – though I also have taught in the Math department and the Division of Epidemiology & Biostatistics). Bernie also said that he wanted someone who wasn't a Bayesian. (I'm not "not a Bayesian" ! Bayesian methods are just fine with me. But anyway, ignoring Bernie's wish for non-Bayesianism, I agreed to Bernie's request to join the list asking for the new section. I later served as section Chair.)

1.1 Introduction: Predictive Analytics

Predictive Analytics is often described as having Descriptive, Predictive and Prescriptive aspects:

$$Description \longrightarrow Prediction \longrightarrow Prescription$$

One of the purposes of this presentation is to discuss some of the precursors to Predictive Analytics appearing in the statistics literature decades ago. Another is to show how Predictive Distributions provide an alternative model to some other statistical models that might be used in particular situations.

1.2 Description

Description, of course, includes the calculation and presentation of the usual descriptive statistics, such as the five-point summary (min, max, quartiles), the mean, standard deviation, skewness, kurtosis, and, for two variables at a time, correlations and scatterplots. We do not here dwell on Description *per se,* although it will be illustrated in an example to follow later. For now, let's move on to Prediction. A couple of seminal papers illustrated some concepts and steps in moving from Description to Prediction.

1.3 From Description to Prediction

Nicholson (1960), among others over the years, emphasized "shrinkage" of R^2, the fact that R^2 using the predicting function from the training sample on a new sample is less than the original, within-sample R^2. This is obvious, because the original R^2 results from an optimization (minimization of residual sum of squares) in the original sample. A comparable value would be obtained only by the same procedure applied to a complete test sample.

Stein (1960) considered out-of-sample prediction in the regression context. The conditional variance of Y given $X_1, X_2, \ldots, X_p$ is a parameter $\sigma^2_{y \cdot x}$, estimated by MSE, the error (residual) mean square. MSE is the in-sample mean squared error, SSE$/(n - p - 1)$, SSE being the sum of squared errors. One chooses the predicting function (subset of explanatory variables) that minimizes MSE. But this is merely choosing the best *descriptor.* What if the *prediction* problem is explicitly formulated?

Now consider predicting the values of Y for a new sample of $\boldsymbol{x}$s, using a predicting function estimated from the first (training) sample. That is, we now have

$$\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+2}, \ldots, \boldsymbol{x}_{n+m}$$

for m new individuals and have to predict the corresponding Ys.

This can be analyzed by considering predicting the value Y_{n+1}, given $\boldsymbol{x}_{n+1}$, the value of $\boldsymbol{x}$ for a new individual, $n+1$. The Mean Squared Error of Prediction, or MSEP, is

$$\text{MSEP} \;=\; \mathcal{E}[(Y_{n+1} - \hat{Y}_{n+1})^2].$$

Explicit calculations can be done in the multivariate Gaussian (MVN) case (Stein 1960). The MSEP turns out to be of the form

$$\text{MSEP} \;=\; \sigma^2_{y.x}\, C(n,p),$$

where $\sigma^2_{y.x}$ is the error variance, that is, the conditional variance of Y given $\boldsymbol{x}$, and $C(n,p)$ is a constant depending upon n and p, namely

$$C(n,p) \;=\; \frac{n(p+1)-2}{n(n-p-2)}.$$

The best predicting function is the one minimizing MSEP, not just MSE. Note that in perhaps simpler form,

$$C(n,p) = \frac{p+1-2/n}{n-p-2}.$$

Note also that $C(n,p)$ is an increasing function of p, so, other things being equal, that is, for equal values of $\sigma^2_{y.x}$, models with a smaller value of p, that is, with fewer predictors, are favored.

The computation proceeds by obtaining the conditional expectation, given the training sample of n $(\boldsymbol{x}, y)$ pairs and X_{n+1}, then taking the expectation over the training sample, and then over X_{n+1}. to obtain the full MSEP. An important step in the calculation relates to taking the expectation of the inverse sample sum-of-products matrix. If $\boldsymbol{V}$ is the sum-of-product matrix of the Xs, and $\boldsymbol{\Sigma}_{xx}$ is their true covariance matrix, then

$$\mathcal{E}[\text{tr}\, \boldsymbol{\Sigma}_{xx}\boldsymbol{V}^{-1}] \;=\; p\,\mathcal{E}[1\,/\,\chi^2_{n-p}] \;=\; \frac{p}{n-p-2}.$$

So, what we have seen is that, at least in the MVN case, alternative predicting functions can be evaluated in terms of an estimate of MSEP.

It would be nice to be able to extend this beyond the MVN case. However, Lukacs and Laha (*Applications of Characteristic Functions,* 1964) showed that linearity of regression and homoscedasticity imply joint Normality.

But, in terms of the theme of the current presentation, we have shown an instance in which Description was formally and expertly extended to Prediction. Next, a brief word on Prescription, and then we shall return to Description / Prediction.

1.4 Prescription

"Prescription" means optimization, such as by linear, quadratic and mathematical programming and other methods of Operations Research (OR) and applied mathematics.

As a predictive example, we can consider that, having fit a regression function, **prescription** would involve finding what values of the vector $\boldsymbol{x}$ of the explanatory variables optimize the predicted $y,$ possibly subject to constraints giving the region of permissible values of the xs.

We do not here pursue Prescription further, but, focusing more on Prediction, we move on to discuss early statistical precursors of "predictive distributions", as they are now called in Predictive Analytics. In particular, we have in mind the **Yule-Greenwood model** (1920), from almost a century ago. (See also Parzen (1962), p. 57.) It is Bayesian and predates Wald (1950), and Savage (1954), by decades.

2. Levels of Granularity

Levels of granularity for analyzing data include: histograms, distributions, mixtures of distributions, predictive distributions.

We start with histograms and then move on to consider other levels of granularity. Given a dataset, histograms with a few different bin widths can be made. A distribution can be fit, using the method of moments, maximum likelihood, or a combination. The finite mixture model can be employed.

2.1 An Example

Kenkel (1984) considered a hypothetical dataset of days ill in a year of n = 50 miners. The days ill are of course integer values. They range from 0 to 18.

Table 1: Frequencies of days ill

days	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
freq	2	3	5	5	2	5	5	4	6	3	0	1	4	1	2	0	0	1	1

The histogram (with bins 0, 1-2, 3-4, . . . ,17-18) suggests bimodality, with modes at about 7 days and 12 days.

The sample mean is about $\bar{x} = 6.6$ days and the sample variance is $s^2 = 19.07,$ that is, the sample standard deviation is $s = 4.37$ days. (By the way, I use a guideline of reporting the mean to one more decimal than what's in the data, and the standard deviation to two more decimals. Here, the data are integers, so that means one decimal for the mean and two decimals for the standard deviation.)

A single Poisson would not provide a good fit: for a Poisson distribution, the mean and variance are equal, but here the variance is much larger than the mean.

A mixture of two Poissons was fit. The mixture model has p.m.f. (probability mass function) $p(x) = \pi_1 \, p_1(x) + \pi_2 \, p_2(x),$ where $p_1(\cdot)$ is the p.m.f. of a Poisson distribution with parameter λ_1 and $p_2(\cdot)$ is the p.m.f. of a Poisson distribution with parameter $\lambda_2.$ The estimates were $\hat{\lambda_1} = 2.8$ days, $\hat{\pi_1} = .40,$ $\hat{\lambda_2} = 9.10$ days, $\hat{\pi_2} = .60.$ The results were obtained by finding the method-of–moments estimates and doing a grid search in their vicinity to maximize the likelihood, and also by the EM (Expectation-Maximization) algorithm, giving $\hat{\lambda_1} = 2.84$ days, $\hat{\lambda_2} = 9.20$ days, $\hat{\pi_1} = .41,$ $\hat{\pi_2} = .59.$

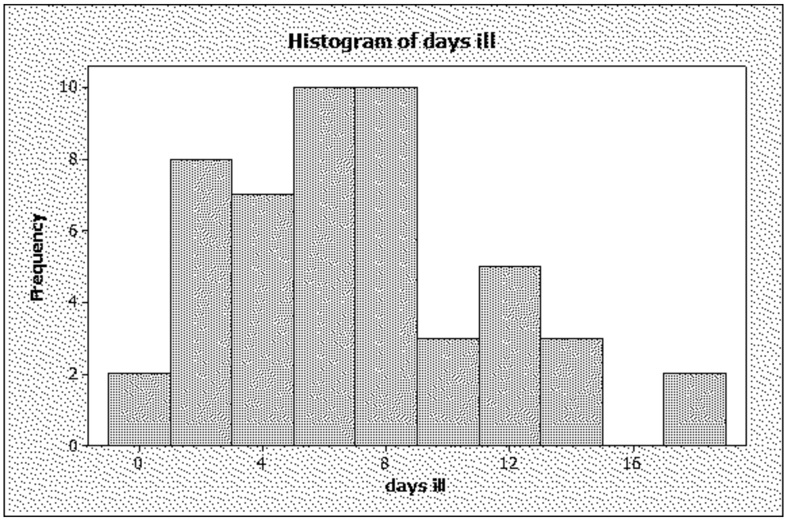

Figure 1: Histogram of days ill

2.2 Comparison of Models by Model-Selection Criteria

The two fits, by histogram and by Poisson mixture, were compared by means of model-selection criteria. Given K alternative models, indexed by $k = 1, 2, \ldots, K,$ penalized-likelihood model-selection criteria are smaller-is-better criteria that take the form

$$\text{MSC}_k \;=\; -2\,LL_k \;+\; a(n)\,m_k,$$

where m_k is the number of free parameters used in fitting Model $k,$ LL_k is the log maximum likelihood of Model k, and $a(n) = \ln n$ for BIC (Bayesian Information Criterion; Schwarz 1978) and = 2 for AIC (Akaike's Information Criterion; Akaike 1974; Kashyap 1982; Sakamoto 1992). That is, for $k = 1, 2, \ldots, K$ alternative models,

$$\text{AIC}_k \;=\; -2\,\text{LL}_k \;+\; 2\,m_k,$$

and

$$\text{BIC}_k \;=\; -2\,\text{LL}_k \;+\; (\ln n)\,m_k.$$

The number of parameters for the Poisson mixture is two means plus 2 mixing probabilities, less 1 because the probabilities must add to 1. That is 3 free parameters for the Poisson mixture. The number of parameters for the histogram, scored by the multinomial distribution with 17 categories (0 through 18, but 15 and 16 are missing), less 1 because the multinomial probabilities must add to 1, leaving 16 free parameters.

The results are in the next table. The histogram wins by a bit according to AIC, but the Poisson mixture wins by a wide margin according to BIC. To see this, note that BIC is derived (Schwarz 1978) as the first terms in the Taylor series expansion of (-2 times) the posterior probability of Model $k,$ $\Pr(\text{Model } k \,|\, \text{data}) = \text{pp}_k$, say. That is,

$$-2 \ln \text{pp}_k \;\approx\; \text{Const.} + \text{BIC}_k, \;\text{ or }\; \text{BIC}_k \approx C\,\exp(-\text{BIC}_k/2).$$

To calculate the posterior probabilities, one subtracts a large constant from each, divides by 2, exponeniates the negative of this, and sums these, dividing by the sum to normalize.

Table 2: Comparison of two models

Model, k	- 2 LL$_k$	m_k	AIC$_k$	BIC$_k$	pp$_k$
$k = 1$: histogram	261.6	16	293.6	324.2	5.0 $\times 10^{-7}$
$k = 2$: Poisson mixture	283.5	3	289.5	295.2	≈ 1

Table 3: Calculation of posterior probabilties of alternative models

Model, k	BIC$_k$	same - 295	$\exp(-\text{same}/2)$	pp$_k$
1	324.2	29.2	4.49×10^{-7}	4.98×10^{-7}
2	295.2	0.2	0.90085	1.000
		sum =	0.90085	

Different bin widths. How should the likelihood for histograms be computed? Let the sample be indexed by $i =, 2, \dots, n$. Given data points $x_1, x_2, \dots, x_n$, the likelihood for a given p.m.f. $p(\cdot)$ is

$$L \;=\; \Pi_{i=1}^{n}\, p(x_i).$$

Here $p(x_i)$ is the p.m.f. at the data point x_i. (For continuous data, we would write the p.d.f., $f(x_i)$.) But in the context of histograms what we can take $p(x_i)$ to be?
Denote the number of bins by J. Let the bin width be denoted by h . This is an increment along the x-axis.

Let the bins be indexed by $j,\ j = 1, 2, \dots, J$. The class limits are

$$x_0, x_0 + h, x_0 + 2h, \dots,\ x_0 + Jh.$$

The class intervals (bins) are

$$[x_0, x_0 + h), [x_0 + h, x_0 + 2h), \dots, [x_0 + (J-1)h,\ [x_0 + Jh).$$

The value x_0 is the “location”, often the sample minimum. In the present application, $x_0 = 0$. Now, let $j(x_i)$ denote the bin containing x_i and $n_{j(x_i}$ be the frequency in that bin. To approximate $f(x_i)$, motivated by $f(x_1) \approx [F(x_2) - F(x_1)] \,/\, (x_2\, -\, x_1)$, write

$$\begin{aligned}
f(x_i) \;&=\; \text{probability density at } x_i \\
&\approx\; \text{probability in bin containing } x_i \text{ / width of bin} \\
&=\; [n_{j(x_i)}\,/n]/h \\
&=\; n_{j(x_i)}/nh.
\end{aligned}$$

That is, the concept involved is that probability density is probability per unit length along the x axis. Thus the likelihood is

$$\begin{aligned}
L \;&=\; \Pi_{i=1}^{n}\, p(x_i) \\
&=\; \Pi_{i=1}^{n}\, (n_{j(x_i)}\,/\,nh) \\
&=\; (1/h^n)\, \Pi_{i=1}^{n}\, (n_{j(x_i)}/n) \\
&=\; (1/h^n)\, \Pi_{j=1}^{J}\, (n_j/n)^{n_j}.
\end{aligned}$$

Note that $p(x_1, x_2, \ldots, x_n) = \Pi_{i=1}^{n} p_{j(x_i)} = \Pi_{j=1}^{J} p_j^{n_j}$ is a **multinomial** p.m.f. with probabilities p_j and frequencies n_j for the J categories. The maximized likelihood L is this multinomial (with p_j estimated as n_j/n), divided by h^n, which may be viewed as an adjustment to the likelihood due to the bin width h. In computing the likelihood, the probability *density* is to be used, where "density" is probability / bin width. Note that with a continuous variable we would compute probability density as $f(x_i)/h$, that is, $f(x_i)/$(Lebesgue measure of the bin interval), whereas with a discrete variable we are really computing probability density as $p(x_i)/h$, where now h is the *counting measure* of the bin interval.

Number of parameters for fitting histograms. I do want to introduce a word of caution relating to the computation of the number of parameters for fitting histograms. The number m_k is supposed to be the number of *free* (independent) parameters used in fitting model k. But in fitting histograms, there is a connection between the bin width h and the number K of bins, in that $range/h = K$. I am investigating this further.

Varying bin widths. In the case of non-constant bin widths, with a bin-width of h_j for the j-th interval, take the probability density at x_i to be $n_{j(i)}/h_{j(i)}$, where $h_{j(i)}$ is the width of the interval in which x_i falls and $n_{j(i)}$ (short for $n_{j(x_i)}$) is the frequency (count) in that interval. The likelihood is

$$L = \Pi_{i=1}^{n} p(x_i) = \Pi_{i=1}^{n} (n_{j(i)} / n) / h_{j(i)} = (1/n^n) \Pi_{i=1}^{n} n_j / h_{j(i)}.$$

Table 4: Sample distribution with a bin width of 2

days	0 -1	2-3	4-5	6-7	8- 9	10- 11	12- 13	14- 15	16-17	18-19
freq	5	10	7	9	9	1	5	2	1	1

Table 5: Sample distribution with varying bin widths: bins 0, 1, 2-3,4-5.6-7, . . . , 16-17, 18

days	0	1	2-3	4-5	6-7	8- 9	10- 11	12- 13	14- 15	16-17	18
bin width	1	1	2	2	2	2	2	2	2	2	1
freq	2	3	10	7	9	9	1	5	2	1	1

According to AIC, the histogram with varying bin widths wins, the Poisson mixture coming in second. According to BIC (and, equivalentlly, posterior probability), the Poisson mixture scores the best, by far. But the point is not just which model wins, but that such a comparison can be made.

Levels of Granularity, cont'd. Perhaps another level of granularity is approached by *predictive distributions,* which may be viewed as getting to the individual level of granularity. Predictive distributions may be viewed in the light of compound distributions resulting from a prior distribution on the parameter at the individual level. From the viewpoint of modern statistics, a predictive distribution is merely the marginal distribution of the observable random variable, having integrated out the prior on the parameter. (Details to follow.)

Table 6: Comparison of models

Model, k	- 2 LL$_k$	m_k	AIC$_k$	BIC$_k$	pp$_k$
histogram, bin width h=1	261.6	16	293.6	324.2	.000
histogram, bin width h=2	273.2	9	291.2	308.4	.001
histogram, varying bin widths	267.8	9	285.8	303.0	.020
Poisson mixture	283.5	3	289.5	295.2	.978

The Yule-Greenwood model approaches modeling at the individual level, stating that each individual may have his or her own accident rate λ and so is an example of a *compound model.* In terms of granularity, the Yule-Greenwood model is a classical example at the level of the individual in that it employs a Poisson model for each individual's accident rate λ and then puts a (Gamma) distribution over the population of values of λ. The model is the Gamma-Poisson model (sometimes called the Poisson-Gamma model) and is a prime example of a **compound model.** The Gamma is a **conjugate** prior distribution for the Poisson, meaning that the posterior distribution of λ is also a member of the Gamma family.

3. An Example with Continuous Data

The variable in the next example will be expenditure in a week (£) of n = 60 English families on fruits and vegetables (Connor and Morrell 1977, data from the British Institute of Cost and Management Accountants). The data are reported to two decimals. The minimum is 0.21 £; the maximum, 2.13 £. The frequency distribution suggests possible bimodality.

The distribution above has a bin width h of 0.10. We consider also the results for $h =$ 0.2, for fitting a single Gamma and also for fitting a mixture of two (Gaussian) distributions. We compare these four fits by means of AIC and BIC.

The sample mean is $\bar{x} =$ 1.022 £, the sample standard deviation, $s =$ 0.4562 £(sample variance $s^2 =$ 0.2081).

The estimates of the Gamma parameters of the Gamma p.d.f.

$$f(x) \;=\; \lambda^{m-1} e^{-x/\beta} / \, \Gamma(m) \, \beta^m, \; x > 0.$$

The mean is $m\beta$. The variance is $m\beta^2$. Method-of-moments estimates are, for the scale parameter $\beta = \sigma^2/\mu, \; \hat{\beta} = \; s^2/\bar{x} = 0.2081/1.022 = 0.2035$. and for the shape parameter $m = \mu/\beta$, so $\hat{m} = \bar{x}/\hat{\beta} = \; 1.022/0.2035 = 5.0246$.

The mixture model has p.d.f. $f(x) \;=\; \pi_1 \, f_1(x) \;+\; \pi_2 \, f_2(x)$, where $f_1(\cdot)$ is the p.d.f. of a Gaussian distribution with mean μ_1 and variance σ_1^2 and $f_2(\cdot)$ is the pdf of a Gaussian with mean μ_2 and variance σ_2^2. The estimates are $\hat{\mu_1} \;=\; 0.72\pounds, \; \hat{\mu_2} \;=\; 1.46\pounds, \; \hat{\sigma_1} \;=\; 0.23\pounds, \;\; \hat{\sigma_2} \;\;=\;\; 0.27\pounds, \; \hat{\pi_1} \;=\; .62, \;\; \hat{\pi_2} \;=\; .38.$ The results were obtained by appoximate maximization of the likelihood doing an EM (Expectation-Maximizatiion) iteration. Note that the estimates of σ_1 and σ_2 are somewhat different; the ratio of variances is $(0.27/0.23)^2 = 0.075/0.051 = 1.46$.

The table summarizes the results. The Gamma wins, both according to AIC and BIC. The Gaussian mixture comes in second according to both criteria.

Table 7: Frequency distribution of weekly expenditure (£)

lower limit	upper limit	Frequency
0.21	0.3	1
0.31	0.4	3
0.41	0.5	4
0.51	0.6	4
0.61	0.7	4
0.71	0.8	6
0.81	0.9	5
0.91	1.0	5
1.01	1.1	4
1.11	1.2	4
1.21	1.3	5
1.31	1.4	2
1.41	1.5	2
1.51	1.6	3
1.61	1.7	2
1.71	1.8	3
1.81	1.9	1
1.91	2.0	1
2.01	2.1	0
2.11	2.2	1

4. Compound Models

4.1 Probability Function Notation

First, notation notation for probability functions will be reviewed.

The probability density function (p.d.f.) of a continuous random variable (r.v.) X, evaluated at x, will be denoted by $f_X(x)$. The p.d.f. of a continuous random variable Y, evaluated at y, is similarly denoted by $f_Y(y)$.

Now consider a bivariate variable $\boldsymbol{x} = (y,\, z)$. The joint p.d.f. of the r.v.s Y and Z, evaluated at (y, z) is $f_{Y,Z}(y, z)$. Example: $Y = WT, X = HT,$ the value of the joint p.d.f. at y = 80 kg and z = 170 cm is $f_{WT,HT}(80, 170)$.

Other notations include:

$f_{Y|X}(y|x)$: **conditional** probability density function of the r.v. Y, given that the value of the r.v. X is x. Example: $f_{WT|HT}(\text{wt}\,|HT = 170\text{cm})$. This represents the bell-shaped curve of weights for men of height 170 cm.

$f_{Y,Z}(y\,|\,z) = f_{Y|Z}(y|z)\, f_Z(z)$: This is the joint p.d.f. expressed as the product of the conditional of Y given Z and the marginal of Z

$f_Y(y) = \int f_{Y,Z}(y,z)\, dz = \int f_{Y|Z}(y|z)\, f_Z(z)\, dz$: marginal pdf of Y

In the development that follows, $f_Z(z)$ plays the role of the prior probability function on the parameter. That is, denoting the parameter by $\theta,$ the function $f_Z(z)$ will become $f_\Theta(\theta)$.

4.2 Compound Models

The elements of **compound models** are:

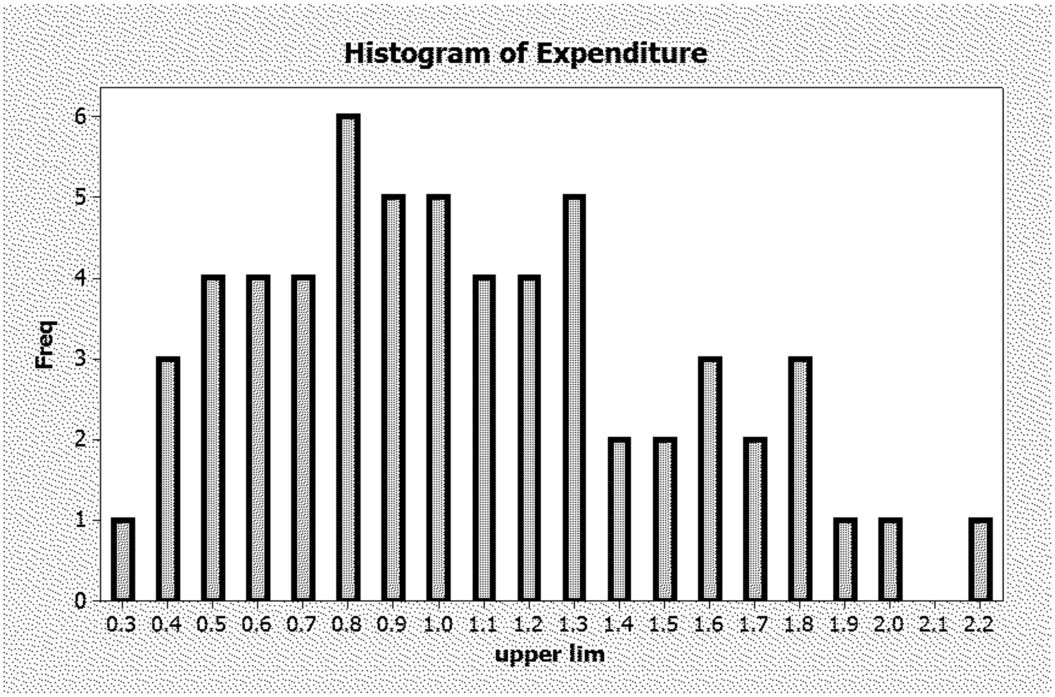

Figure 2: Histogram of Expenditure

Table 8: Comparison of results

Model, k	- 2 LL$_k$	m_k	AIC$_k$	BIC$_k$	pp$_k$
histogram, bin width h=0.1	61.68	18	97.68	135.38	.000
histogram, bin width h=0.2	66.25	9	84.25	103.10	.000
Gamma	71.75	2	75.75	79.94	.999
Gaussian mixture	72.88	5	82.88	93.35	.001

- The distribution of the observable r.v., given the parameter(s), that is, the conditional distribution of X, given the parameter(s); and the marginal distribution (predictive distribution). The marginal distribution will have the hyperparameters among its parameters.

- the prior distribution. Its parameters are called *hyperparameters.*

A generic symbol for the parameter(s) of the conditional distribution of X is the conventional $\boldsymbol{\theta}$. As a generic symbol for the hyperparameters, one could use $\boldsymbol{\alpha}$, since the prior comes first in the model when one thinks of the parameter value being given first, and then the value of the variable being observed.

For use in compound models, the probability functions include the following:

The conditional distribution of the observable r.v. X, given the value of the parameter, is $f_{X|\Theta}(x\,|\,\theta)$; pdf of X for given θ.

The prior distribution on the parameter $\boldsymbol{\theta}$ with hyperparameter vector $\boldsymbol{\alpha}$, $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta};\boldsymbol{\alpha})$.

The naming of compound models takes the form, Prior distribution – conditional distribution. In the Gamma-Poisson model, the conditional distribution of X given λ is Poisson(λ) and the prior distribution on λ is Gamma. In the Beta-Binomial mode, the conditional distribution of X given p is Binomial with success probability p and the prior distribution on p is Beta.

5. The Gamma-Poisson Model

5.1 Probability Functions for the Gamma-Poisson Model

In the Gamma-Poisson model, the distribution of X is Poisson with parameter usually called λ. The p.m.f. is

$$p(k) \;=\; e^{-\lambda}\,\lambda^k\,/\,k!,\;\; k = 0, 1, 2, \dots\,.$$

The mean and variance are both equal to λ.

Such a distribution can be considered, say, for the number of accidents per individual per year. For the days ill dataset (days ill in a year for a sample of N = 50 miners), we have fit a single Poisson (with mean 6.58 days per year). We looked at histograms and observed bimodality. Further, the fact that the sample variance of 19.06 was considerably larger than the sample mean was a hint of inadequacy of a single Poisson. A mixture of Poissons was fitted, with mixing probabilities about .6 and .4 and means about 3 days and 9 days. A finer level of granularity would be obtained by saying that *each person* has his own value of λ and putting a distributon on these over the population.

5.2 Gamma Family of Distributions

A gamma distribution could be a good choice. It is non-restrictive in that the family can achieve a wide variety of shapes. The single-parameter gamma has a shape parameter m; the two-parameter gamma family has, in addition, a scale parameter, β. (The reciprocal of β is the rate parameter.) A Gamma distribution with parameter m, has p.d.f.

$$f(\lambda) \;=\; \text{Const.}\,\lambda^{m-1}\,e^{-\lambda},\;\; \lambda > 0.$$

The constant is $1/\Gamma m$. More generally, the two-parameter Gamma can be used: the p.d.f. is

$$f(\lambda) \;=\; \lambda^{m-1}e^{-\lambda/\beta}/\,\Gamma(m)\,\beta^m,\;\; \lambda > 0.$$

The mean is $m\beta$. The variance is $m\beta^2$.

5.3 Exponential Family of Distributions

The special case of $m = 1$ in the Gamma family gives the negative exponential family of distributions. So the

$$f(\lambda) \;=\; e^{-\lambda/\beta}/\,\Gamma(m),\;\; \lambda > 0.$$

The mean is β. The variance is β^2.

5.4 Development of the Gamma-Poisson Model

Putting a population distribution over a parameter can be a very helpful way of modeling. The resulting model is called a **compound model.** In a compound model, the random variable X is considered as the result of sampling that yields an individual and that individual's value of a parameter, and then the individual's value of X is observed, from a distribution with that value of the parameter.

In this discussion, focus is on a couple of particular compound models, the Gamma-Poisson, and later, the Beta-Binomial.

The Yule-Greenwood model, from a modern viewpoint, is an application of the Gamma-Poisson model to a financial, in fact, actuarial, situation. It is in terms of a model for *accident rates* in a population. Suppose that the yearly number of accidents of any given individual i in a population is distributed according to a Poisson distribution with parameter λ_i accidents per year. (This is *count data,* similar to the days ill data.) Then the probability that individual i, with parameter value λ_i, has exactly k accidents in a year, k = 0, 1, 2, ..., is

$$e^{\lambda_i}\,\lambda_i^k\,/k!,\ k = 0, 1, 2, \ldots, .$$

Some individuals are more accident prone (have a higher accident rate) than others, so different individuals have different values of λ. A distribution can be put on λ to deal with this. This is the Yule-Greenwood model, dating from 1920; a precursor of the Predictive Distributions of the new Predictive Analytics, predating even Abraham Wald as a founder of modern mathematical statistics and decision theory and Jimmie Savage as a founder of modern Bayesian Statistics.

The standard choice of s prior distribution on λ is a Gamma distribution.

5.4.1 The Joint Distribution of X and Λ

The joint probability function of X and Λ is

$$f_{X,\Lambda}(x, \lambda) = f_\Lambda(\lambda)\, p_{X|\Lambda}(x\,|\lambda),\ x = 0, 1, 2, \ldots,\ \lambda > 0.$$

The expressions for the Gamma and Poisson are put into this. That is, the weight assigned to $p_{X\,||\Lambda}(x\,|\,\lambda)$ is $f_\Lambda(\lambda)$.

The joint probability function is is used to obtain

- the marginal distribution of X, by integrating out λ, and
- then the posterior distribution of Λ given x, by dividing the joint probability function by the marginal probability mass function of X.

.

Putting in the expressions for the Gamma and Poisson, it is seen that the marginal distribution of X, the number of accidents that a randomly selected individual has in a year, is of the form

$$f_X(x) \;=\; \int_0^\infty f_{X,\Lambda}(x, \lambda)\, d\lambda \;=\; \int f_{X|\Lambda}(x|\lambda)\, f_\Lambda(\lambda)\, d\lambda.$$

When the prior is Gamma and the conditional is Poisson, this marginal distribution can be shown to be *negative binomial.* Its parameters are m and $p = 1/(1 + \beta)$.

In the Bayesian model, the parameter of the conditional distribution of X, say θ, is treated as a random variable Θ.

In the Gamma-Poisson model, θ is the Poisson parameter λ.

The conditional distribution of X given that $\Theta \;=\; \theta$ is Poisson(λ). The probability mass function is

$$p_{X|\Lambda}(x; \lambda) \;=\; e^{-\lambda}\,\lambda^x\,/\,x!,\ x = 0, 1, 2,\ \ldots,\ .$$

The joint pdf of X and Λ can be written as $p_{X\,||\Lambda}(x\,|\,\lambda)$ which is $f_\Lambda(\lambda)(x, \lambda) = f_\Lambda\, p_{X|\Lambda}(x|\lambda$.

As mentioned above, from this, the posterior distribution of Λ, that is, the distribution of Λ given x, can be computed, and the marginal distribution of X can be computed.

5.4.2 *Posterior Distribution of* Λ

Analogous to $\Pr(B|A) = \Pr(A \cap B) / \Pr(A)$, the pdf of the posterior distribution is the joint pdf, divided by the marginal pdf of X :

$$f_{\Lambda|X}(\lambda|x) \;=\; f_{X,\Lambda}(x,\lambda) \,/\, f_X(x).$$

This will turn out to be a Gamma distribution, that is, it is in the same family as the prior. The Gamma is a *conjugate prior* for the Poisson.

5.4.3 *Marginal Distribution of* X

In **Predictive Analytics,** the marginal distribution of X is computed as a model of a future observation or observations of X.

The *marginal distribution* of X is obtained by integrating the joint distribution with respect to the parameter. Note that this computation combines information, by weighting the conditional distribution of X given λ with the prior on λ. This computationof the p.d.f. is, as stated above, $f_X(x) = \int_0^\infty t\, f(x|\,\lambda)\, f_\Lambda(\lambda)\, d\lambda$.

5.4.4 *Moments*

The mean of the marginal distribution of X is $mq/p = m\,\beta$. The variance of the marginal distribution of X is $mq/p^2 = m\,\beta(1+\beta)$.

5.5 Empirical Bayes estimation

Empirical Bayes estimation, at least in the present context, means estimating the parameters of the prior using observations from the marginal distribution.

5.5.1 *The hyperparameters in terms of the moments of the marginal*

The parameters of the prior are called *hyperparameters.* In this case, they are λ and β. Suppose we solve for them in terms of the first two moments of the marginal.

5.5.2 *Estimating the prior parameters from the marginal*

Estimates of the prior parameters m and β can be obtained by, for example, taking the expressions for the hyperparameters m and β in terms of the first two raw moments and plugging in estimates m_1' and m_2'. Given a sample $X_1, X_2, \ldots, X_N$, we have $m_1' = \bar{X} = \sum_{i=1}^N X_i \,/\, N$ and $m_2' \;=\; \sum_{i=1}^N X_i^2 \,/\, N$.

5.6 Application to the Days Ill dataset

Kenkel (1984) considered a hypothetical dataset of days ill of n = 50 miners. The days ill in a year ranged from 0 to 18; the distribution seems to be bimodal.

The p.m.f. of the Negative Binomial distribution with parameters m and p is where k is the number of trials in excess of m required to get m Heads. In the Gamma-Poisson model, the marginal distribution of X is Negative Binomial with parameters with parameters m and $p = 1/(1+\beta)$.

Given that the true mean of the marginal ("predictive distribution") Negative Binomial is $\mu = mq/p = m\beta$ and the true variance is $\sigma^2 \ = \ mq/p^2 = m\beta(1+\beta)$, and the sample mean $\bar{x} = 6.58$ and the sample variance $s^2 = 19.07,$ one can set up two equations and solve for method of moments estimates of the hyperparameters m and β in the Gamma prior for λ.
The equations are $[1] : m\,\beta = 6.58$; $[2] : \ m\,\beta(1+\beta) = 19.07$.
Putting [1] in [2] gives $6.58(1+\beta) = 19.07,\ 1+\beta \ = \ 19.07/6.58 \approx 2.898,\ \hat{beta} \approx 1.898.$
Then $m \approx 6.58/\beta = 6.58/1.898 \approx 3.467$. Now, $\mu = m(1-p)/p = m/p - m,\ \ \mu + m = m/p,\ \ p = m/(\mu + m)$ or, estimating $p = 3.467/(6.58 + 3.467) = 3.467/10.05 = 0.345.$ So now we have estimates of the hyperparameters.

To estimate the mean and variance of the Gamma prior, one can proceed as follows. The mean of the prior is $m\beta$, estimated as 6.58 days ill per year. The variance of the prior is $m\beta^2$, estimated as $6.58(1.898) \ \approx \ 12.49$. The standard deviation is thus estimated as $\sqrt{12.49} \approx 3.03$ days ill per year.

Maximum likelihood estimates are not in closed form but numerical values for them could be obtained by numerical maximization of the likelihood function. It is helpful to use the method of moments as a quick and simple method to get an idea of the values of the parameters.

The table includes the marginal negative binomial with $m \ = \ 3$ and $p \ = \ .344$ in the comparison.

Table 9: Comparison of models, cont'd

Model, k	- 2 LL$_k$	m_k	AIC$_k$	BIC$_k$	pp$_k$
histogram, bin width h=1	261.6	16	293.6	324.2	.000
histogram, bin width h=2	273.2	9	291.2	308.4	.000
histogram, varying bin widths	267.8	9	285.8	303.0	.002
Poisson mixture	283.5	3	289.5	295.2	.100
marginal Negative Binomial	283.0	2	286.0	290.8	.898

According to AIC, the histogram with varying bin widths still wins, the Negative Binomial coming in second. According to BIC (and, equivalentlly, posterior probability), the Negative Binomial scores the best, by far. This Negative Binomial is unimodal with a mode of .115 at 3 days. Because it is unimodal, it perhaps does not capture the flavor of the original data, which is reflected better by the Poisson mixture.

6. Some Other Compound Models: Beta-Binomial; Normal-Normal

6.1 Beta-Binomial Model

Another compound model is the Beta-Binomial model.

In this model, the conditional distribution of X given p is Binomial(n,p). The prior on p is Beta(α, β).

The posterior distribution of p given x is Beta$(\alpha + x,\ \beta + n - x)$. It is as if there had been a first round of $\alpha + \beta$ trials, with α successes, followed by a second round of n trials, with x successes.

Method of Moments estimates of the parameters of the prior can be relatively easilty obtained. So can the Bayes estimates.

6.2 Normal-Normal Model

We have considered the Gamma-Poisson model and, briefly, another prominent compound model, the Beta-Binomial model, with a Beta prior on the Binomial success probability parameters. Still another compound model is the Normal-Normal model.

In the Normal-Normal model, X is distributed according to $\mathcal{N}(\mu, \sigma^2),$ the Gaussian distribution with mean μ and variance σ^2. The prior on μ can be taken to be $\mathcal{N}(\mu_0,\, \sigma_0^2)$, or perhaps a Gaussian with a different mean if there is some particular reason to do this.

The posterior distribution is again Normal. That is, the Normal family is the conjugate family for the Normal distribution. The marginal distribution is also Normal, with mean $\mathcal{E}[X] = \mathcal{E}[\mathcal{E}[X|\mu] = \mathcal{E}[\mu] = \mu_0$. The variance of the marginal distribution is the mean of the conditional variance plus the variance of the conditional mean, $\mathcal{V}[X] = \mathcal{E}[\sigma^2] + \mathcal{V}[\mathcal{E}[\mu] = \sigma^2 + \mathcal{V}[\mu] = \sigma^2 + \sigma_0^2$. These two terms are the “components of variance”. The decomposition of the variance can be obtained also by doing the requisite algebra on the product of the prior and conditional.

This model is similar to a Random Effects model (Model II) in ANOVA. The parameter σ^2 is the error variance, and σ_0^2 is the variance of the random effects.

Multivariate generalizations could be interesting.

Comments on References

The books mentioned on predictive analytics, those of Murphy and Bishop, do not discuss the Gamma-Poisson model explicitly. Those who wish to consult these books may however refer to them to find

- Murphy, p. 41 on the Gamma family of dsitributions and/or
- Bishop, p. 688 on the Gamma family.

As mentioned, an original paper, anticipating the subject, is that of Greenwood and Yule (1920). To review background in Probability Theory in general, see, for example, Parzen (1992) or Ross (2014). To review background in probability models, see also Parzen, *Stochastic Processes* (1962) or Ross (1970, 1992).

REFERENCES

Major Greenwood and G. Udny Yule (1920). ”An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents”. *Journal of the Royal Statistical Societ,* **83:** 255-279. JSTOR 2341080.

Eugene Lukacs and R. G. Laha. *Applications of Characteristic Functions.* (Griffin’s Statistical Monographs & Courses, No. 14). New York: T. Hafner Pub. Co.

George Nicholson (1960). “Prediction in Future Samples”. Page 322–330 in Ingram Olkin, Sudhish G. Ghurye, Wassily Hoeffding, William G. Madow, and Henry B. Mann, eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford University Press, Stanford, CA. 517 pages.

Leonard J. Savage (1954). *The Foundations of Statistics.* John Wiley and Sons, New York.

Charles Stein (1960). “Multiple Regression”. Pages 424–443 in in Ingram Olkin, Sudhish G. Ghurye, Wassily Hoeffding, William G. Madow, and Henry B. Mann, eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford University Press, Stanford, CA. 517 pages.

Abraham Wald (1950). *Statistical Decision Functions*. John Wiley and Sons, New York; Chapman and Hall, London.

References for Data Examples

L. R. Connor and A.J.H. Morrell (1977). *Statistics in Theory and Practice. 7th ed.* London: Pitman.

James Kenkel (1984) *Introductory Statistics for Management and Economics. 2nd ed.* Duxbury Press, Boston, MA. Exercise 4, p. 31.

References on Model Selection Criteria

Hirotugu Akaike (1974). "A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control,* **19** (6): 716723.

Rangasami Kashyap (1982). "Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models". *IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume:PAMI-4 , Issue: 2)*, 99 – 104.

Yosiyuki Sakamoto (1992). *Categorical Data Analysis by AIC.* Springer, New York.

Gideon Schwarz (1978). "Estimating the Dimesnion of a Model". *Annals of Statistics,* **6,** 461-464.

References for Background Reading on Predictive Analytics

Christoper M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

Kevin P. Murphy. *Machine Learning: a Probabilistic Perspective.* The MIT Press, 2012.

References for Background Reading on Probability

Emanuel Parzen (1960). *Modern Probability Theory and its Applications.* Wiley. (Reprinted in 1992 as a Wiley Classics Edition.)

Sheldon M. Ross (2014). *Introduction to Probability Models. 11th ed.* Elsevier, Amsterdam, The Netherlands; Waltham, MA; San Diego, CA.

Refereences for Background Reading on Stochastic Processes

Emanuel Parzen (1962). *Stochastic Processes.* Wiley, New York. Dover Publications (Reprint of the original, published by Wiley.)

Sheldon M. Ross (1970). *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco. Reprinted, Dover, New York, 1992.