DRAFT

Glen Cowan
RHUL Physics
17 December, 2008

Likelihood ratio test to determine number of parameters

We discuss how to determine the optimal parametric function needed to describe a distribution by using a likelihood ratio test. The note developed out of discussions with Eilam Gross, Sascha Caron, Stephan Horner and from the talk by Horner [1] at the ATLAS Statistics Forum. Similar methods have been used in many other analyses, e.g., the $b \to s\gamma$ measurement by the BaBar Collaboration [2].

The basic idea is to consider a family of functions to fit the data with an increasing number of parameters and thus increasing flexibility. Beginning with the most inflexible model, one successively increases the number of parameters until the statistical test indicates that the model and data are in acceptable agreement. The statistical errors in the parameters of interest are increased as a result of their correlations with the nuisance parameters added to make the function sufficiently flexible. In this way the systematic uncertainty is incorporated directly into the statistical errors.

The difficulty with this procedure is in coming up with an acceptable family of parametric functions, as this choice will depend on the problem. In this note we explore an example that starts with the functional form as predicted by a Monte Carlo model, and then modifies it by multiplication with a linear combination of Bernstein basis polynomials. A goodness-of-fit test based on the profile likelihood ratio is used to determine the optimal order of the polynomials.

Suppose an analysis requires modeling a distribution of a variable x. In principle a Monte Carlo model can be used to give an absolute prediction for the number of entries in each bin of a histogram. In practice one does not regard the MC prediction as perfect, and there would be a systematic uncertainty connected with the assumption of this model.

Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ denote the mean number of entries predicted in the N bins of the histogram of x by a model, and let the number of entries observed in the data be $\mathbf{n} = (n_1, \ldots, n_N)$. Suppose that the true model, i.e., the one from which the data are generated, is $\boldsymbol{\nu}^{(\mathrm{t})}$, and the hypothesized models that we will consider to describe the data are denoted with a numerical superscript, e.g., $\boldsymbol{\nu}^{(0)}$, $\boldsymbol{\nu}^{(1)}$, etc. In this example we will model the number of entries in each bin of the histogram as an independent Poisson distributed value. That is, the probability for the data $\mathbf{n}$ is

$$P(\mathbf{n}; \boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i} \; . \tag{1}$$

Figure 1 shows possible distributions with a simple zeroth-order approximation in (a), the true (unknown) distribution in (b) and a possible data set in (c). The zeroth-order model could, for example, be based on Monte Carlo, which is in general systematically different from Nature. In addition, an MC prediction will in general have statistical fluctuations because of the limited number of events generated. Here these fluctuations have been suppressed as the point of the present example is to explore how to account for systematic effects. Incorporating the statistical errors from MC is straightforward and is discussed e.g., in Refs. [3, 4].

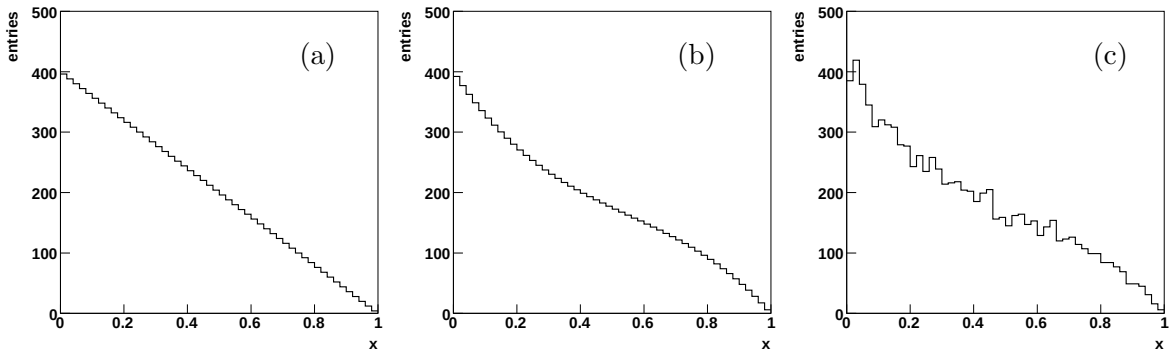

Figure 1: A sample distribution: (a) zeroth-order model (e.g., MC), (b) true (in general unknown), and (c) a possible data set generated from the true distribution.

The data in Fig. 1(c) clearly differ significantly from the zeroth-order model in (a). To quantify the level of compatibility one could compute Pearson's chi-square statistic,

$$\chi^2_{\rm P} = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \; . \tag{2}$$

An almost equivalent statistic is based on the likelihood ratio

$$\lambda(\boldsymbol{\nu}) = \frac{L(\boldsymbol{\nu})}{L(\hat{\boldsymbol{\nu}})} \tag{3}$$

where $L(\boldsymbol{\nu}) = P(\mathbf{n};\boldsymbol{\nu})$ is the likelihood of the hypothesized model $\boldsymbol{\nu}$, and $\hat{\boldsymbol{\nu}}$ is the maximum likelihood (ML) estimator for $\boldsymbol{\nu}$, i.e., the values of $\nu_1, \ldots, \nu_N$ which maximize the likelihood. By setting the derivative of $L(\boldsymbol{\nu})$ equal to zero and solving one easily finds

$$\hat{\nu}_i = n_i \tag{4}$$

for all i.

If the model $\boldsymbol{\nu}$ is correct, then Wilks' theorem [5] states that the distribution of the statistic

$$q_{\boldsymbol{\nu}} = -2 \ln \lambda(\boldsymbol{\nu}) = 2 \sum_{i=1}^{N} \left(n_i \ln \frac{n_i}{\nu_i} + \nu_i - n_i \right) \tag{5}$$

approaches a chi-square distribution for N degrees of freedom for a sufficiently large data sample.[1] In fact in many practical examples the chi-square approximation is extremely good even for moderate samples, e.g., n_i roughly a half dozen or more. Details on the regularity conditions required for Wilks' theorem to be valid are discussed in standard texts such as [6]. Pearson's $\chi^2_{\rm P}$ and the statistic $q_{\boldsymbol{\nu}}$ are for the present example very similar; here we will use $q_{\boldsymbol{\nu}}$.

For either goodness-of-fit statistic, $\chi^2_{\rm P}$ or $q_{\boldsymbol{\nu}}$, one would quantify the compatibility between data and model by giving the p-value. This is the probability, under assumption of the model

[1] In computing $q_{\boldsymbol{\nu}}$, the logarithmic term should be skipped if $n_i = 0$.

$\boldsymbol{\nu}$, to obtain a value of the statistic greater than or equal to that found with the actual data. That is,

$$p = \int_{q_{\boldsymbol{\nu},\mathrm{obs}}}^{\infty} f_{\chi^2}(z; N)\, dz \ , \tag{6}$$

where

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2} \tag{7}$$

is the chi-square distribution for N degrees of freedom, and Γ is the Euler gamma function.

A comparison between the data set in Fig. 1(c) and the model in (a) results in a value of $q_{\boldsymbol{\nu}} = 258.8$. There are $N = 50$ bins in the histogram, and using Eq. 6 gives a p-value of 6×10^{-30}. One would clearly reject the hypothesized model.

We can improve the level of agreement between the data and model by introducing further adjustable parameters. One way to do this is to scale the zeroth-order model $\boldsymbol{\nu}^{(0)}$ by a factor s, which depends on the value of the variable x (i.e., it depends on the bin number), and which contains a set of adjustable parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$. That is, the modified prediction for the mean number of entries in the ith bin is

$$\nu_i \to \nu_i s(x_i; \boldsymbol{\theta}) \ , \tag{8}$$

where x_i is the value of x in the centre of the ith bin.

In this note we examine using a superposition of Bernstein basis polynomials for the scaling function s. The set of $m + 1$ Bernstein basis polynomials of order m are defined as (see, e.g., [7]),

$$b_{k,m}(x) = \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k} \ . \tag{9}$$

These are nonzero in the range $[0, 1]$, which corresponds to the range of the variable x in the example of Fig. 1. Here we will assume that the variable in question has been translated and scaled to lie in this range. The Bernstein basis polynomials for orders 0 through 5 are shown in Fig. 2.

The scaling function $s(x)$ is taken as a linear combination of the basis polynomials (i.e., s is a Bernstein polynomial),

$$s(x) = \sum_{k=0}^{m} \beta_k b_{k,m}(x) \ . \tag{10}$$

For $\beta_k = 1$, $k = 0, \ldots, m$, one has $s(x) = 1$, so it is easy to identify the point in parameter space that corresponds to no scaling. An important property of Bernstein polynomials is that a basis polynomial of a given order $m - 1$ can always be written in terms of those of order m:

$$b_{k,m-1}(x) = \frac{m-k}{m} b_{k,m}(x) + \frac{k+1}{m} b_{k+1,m}(x) \ . \tag{11}$$

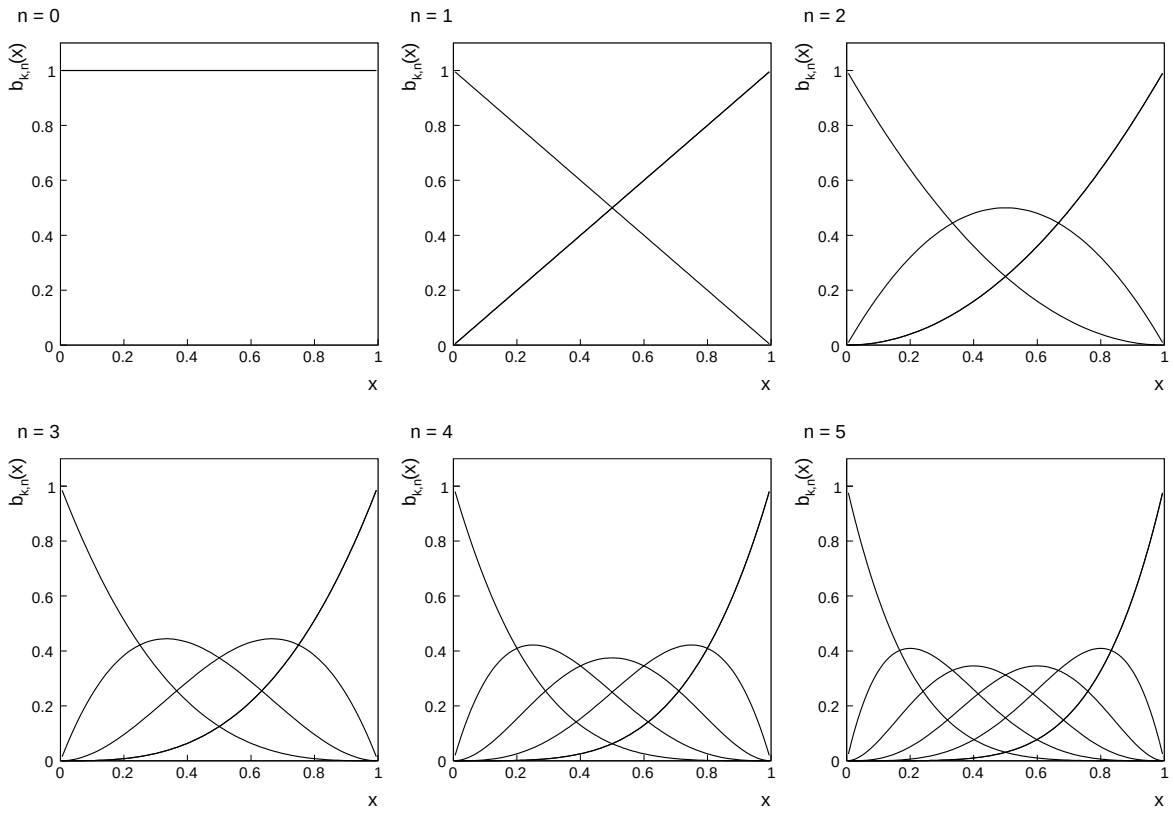

Figure 2: Bernstein basis polynomials of different orders n.

This means that the Bernstein polynomials defined using basis functions of successively increasing order form a nested family. That is, the model of order m contains as a special case the model of order $m-1$. This will be important in constructing the likelihood ratio test to determine whether it is necessary to increase the number of parameters in the model.

The strategy proposed in this note is to begin with the zeroth order model $\boldsymbol{\nu}^{(0)}$ and modify it by allowing for a simple scale factor, i.e.,

$$s(x;\beta_0) = \beta_0 b_{0,0}(x) = \beta_0 \ , \tag{12}$$

as $b_{0,0} = 1$. One can then construct a likelihood ratio test of the hypothesis $\boldsymbol{\nu}^{(0)}$ versus the alternative $\nu_i^{(1)} = \nu_i^{(0)} s(x_i;\beta_0)$ where β_0 is treated as an adjustable parameter. The hypothesis $\boldsymbol{\nu}^{(0)}$ corresponds to the special case $\beta_0 = 1$. The likelihood ratio that compares the two hypotheses is

$$\lambda = \frac{L(\beta_0 = 1)}{L(\hat{\beta}_0)} \ , \tag{13}$$

where $\hat{\beta}_0$ is the ML estimator of β_0. As before it is convenient to use the equivalent logarithmic variable

$$q = -2 \ln \lambda \ . \tag{14}$$

If the data favour the hypothesis $\beta_0 = 1$, then $\hat{\beta}_0$ will be close to one, λ will also be close to unity, and therefore q will be small. Larger values of q indicate that the $\beta_0 = 1$

hypothesis is in poor agreement with the data, and that some other value of β_0 is favoured. Wilks' theorem states that under the hypothesis of $\beta_0 = 1$, the sampling distribution of q should follow a chi-square distribution for one degree of freedom. Therefore the p-value of this hypothesis can be found from

$$p = \int_{q_{\rm obs}}^{\infty} f_{\chi^2}(z; 1)\, dz \ . \tag{15}$$

If the p-value is below a given threshold, say, $p_{\rm cut} = 0.1$ or 0.2, one would conclude that an adjustable value of β_0 is needed.

One then simply iterates this procedure, at each stage increasing the order of the Bernstein basis polynomials by one. That is, if one has decided that the data require modification by Bernstein basis polynomials of order m (i.e., $\boldsymbol{\beta}^{(m)} = (\beta_0, \ldots, \beta_m)$ as adjustable parameters), then one can test whether the data prefer to have the corresponding model of order $m + 1$, using the likelihood ratio

$$\lambda = \frac{L(\hat{\boldsymbol{\beta}}^{(m)})}{L(\hat{\boldsymbol{\beta}}^{(m+1)})} \ . \tag{16}$$

Under the assumption that the more restrictive model in the numerator is correct and providing the data sample is not too small, $q = -2 \ln \lambda$ will follow a chi-square distribution for one degree of freedom. At each stage one computes the likelihood ratio and from it obtains a p-value. If the p-value is below the threshold $p_{\rm cut}$, then one concludes that the additional parameter is required to describe the data.

In addition one would look at the overall goodness-of-fit using the statistic $q_{\boldsymbol{\nu}}$ given by Eq. (5), where here $\boldsymbol{\nu}$ represents the modified model from Eq. (8). If the model is correct then this should follow a chi-square distribution for $N - n_{\rm par}$ degrees of freedom, where $n_{\rm par} = m + 1$ is the order of the Bernstein basis polynomials.

In the example shown in Fig. 1, the "true" distribution in (b) was in fact produced by distorting the distribution in (a) with a 2nd order Bernstein polynomial using $\beta_0 = 1$, $\beta_1 = 0.5$ and $\beta_2 = 1.5$. Recall that the original undistorted distribution corresponds to $\beta_0 = \beta_1 = \beta_2 = 1$. So in this case one expects that to be well described the data will on average require a scale factor based on a 2nd-order basis function. This is in fact what one sees in the fits shown in Fig. 3; the fit with three adjustable parameters provides a good description of the data. Increasing the number of parameters further does not improve substantially the goodness-of-fit. This is also evident from the values of the test statistic q and $q_{\boldsymbol{\nu}}$ obtained using different numbers of free parameters as shown in Table 1.

Table 1: Values of the variables $q_{\boldsymbol{\nu}}$ and q and the corresponding p-values obtained from fits with different numbers of adjustable parameters.

$n_{\rm par}$	$q_{\boldsymbol{\nu}}$	$p_{\boldsymbol{\nu}}$	q	p
0	258.8	6.1×10^{-30}	98.9	2.6×10^{-23}
1	159.9	1.1×10^{-13}	15.4	8.9×10^{-05}
2	144.5	1.3×10^{-11}	112.0	3.5×10^{-26}
3	32.5	0.95	0.0013	0.97
4	32.5	0.93	0.26	0.61
5	32.2	0.92	0.37	0.54

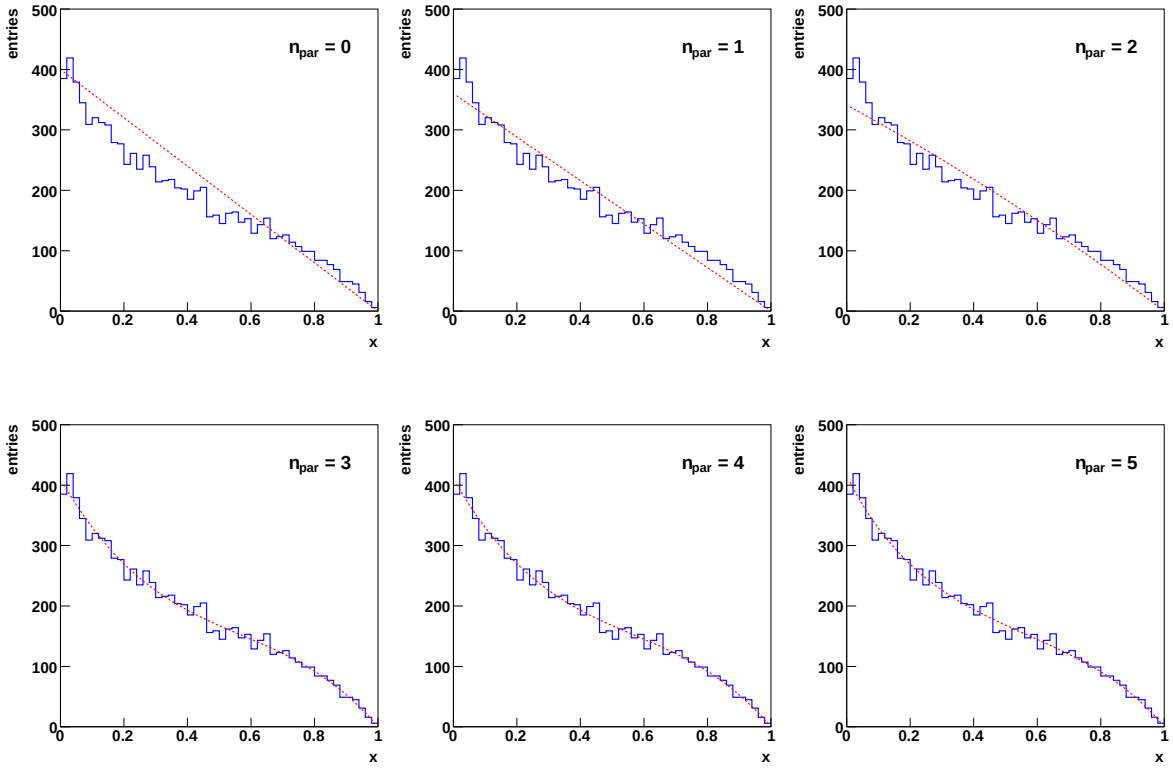

Figure 3: Fits for different numbers of adjustable parameters.

We can investigate this further by simulating the experiment many times. Figure 4 shows the distributions of the test variable q. For Fig. 4(a) this is the goodness-of-fit statistic for the zeroth-order model. For successive plots it shows the distribution of $q = -2 \ln \lambda$ where the likelihood ratio is based on the numbers of parameters indicated. From the plots one sees that distribution of q based on the comparison of the 3 and 4 parameter models is close to a chi-square distribution for one degree of freedom. So in most cases one would not reject the 3-parameter hypothesis.

Once the optimal number of parameters has been determine, then the correlations between the estimators of all of the parameters, including the new ones included in the enlarged model, will result in larger statistical errors for the parameters of interest. In this way the systematic uncertainty connected with the original model is taken into account.

References

[1] Stephan Horner, *Template fitting for background determination (SUSYFIT)*, presentation at the ATLAS Statistics Forum, 3 December, 2008.

[2] B. Aubert et al. (The BaBar Collaboration), *Measurement of the $B \to X_s\gamma$ branching fraction and photon energy spectrum using the recoil method*, Phys. Rev. D 77, 051103(R) (2008).

[3] G. Cowan, *MC Statistical Errors in ML Fits*, ATLAS Statistics Forum Note, 10 November 2008.

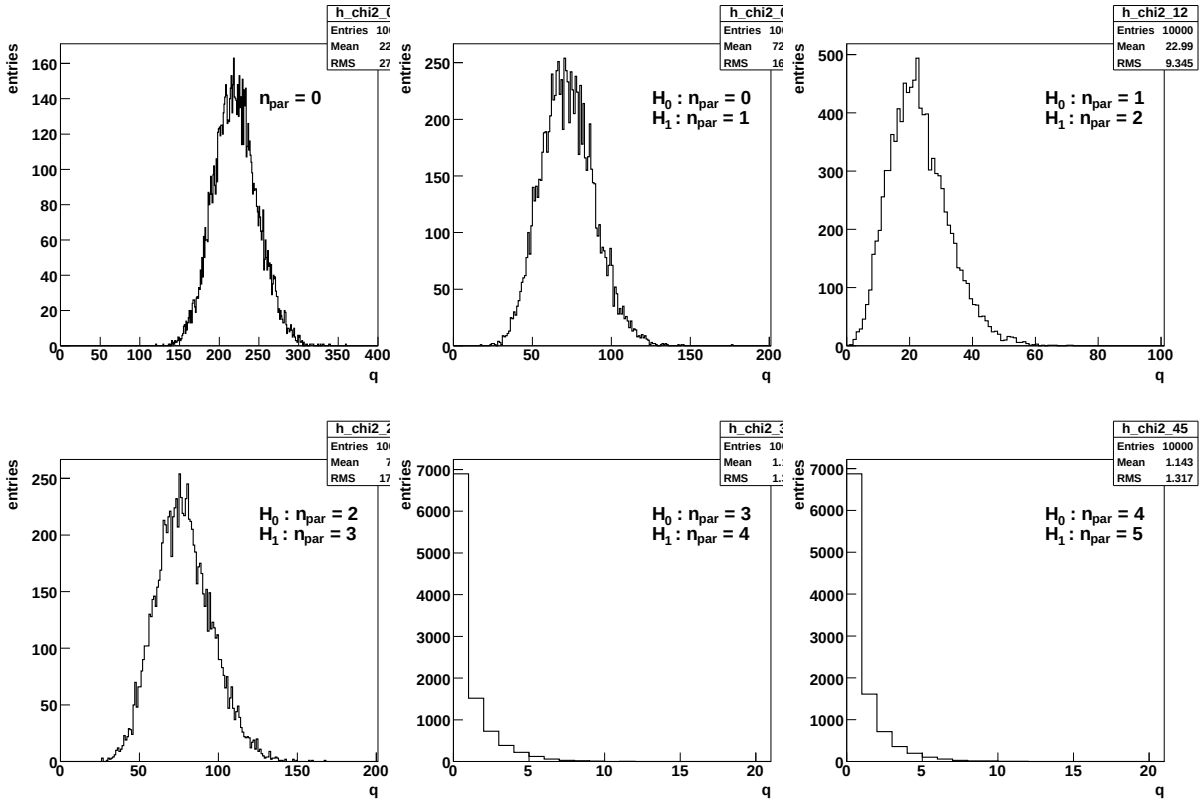

Figure 4: Distributions of the test variable q (a) for the zeroth-order model and (b)–(f) when comparing models with differing numbers of additional parameters.

[4] G. Cowan, E. Gross, *Discovery significance with statistical uncertainty in the background estimate*, ATLAS Statistics Forum Note, 8 May 2008.

[5] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

[6] A. Stuart, J.K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model* 6th Ed., Oxford Univ. Press (1999), and earlier editions by Kendall and Stuart.

[7] Wikipedia contributors, "Bernstein polynomial", *Wikipedia, The Free Encyclopedia*, `en.wikipedia.org/wiki/Bernstein_polynomial` [accessed 15 December, 2008].