8 The Gaussian probability model

Until now we have been concerned only with the binary probability model. In this model there are two possible outcomes and the total probability of 1 is shared between them. It is an appropriate model when studying the occurrence of events, but not when studying a response for which there are many possible outcomes, such as blood pressure. For this the *Gaussian* or *normal* probability model is most commonly used.

In the Gaussian model the total probability of 1 is shared between many values. This is illustrated in the left panel of Fig. 8.1. When measurements are recorded to a fixed number of decimal places, there is a finite number of possible outcomes but, in principle, such measurements have infinitely many possible outcomes, so the probability attached to any one is effectively zero. For this reason it is the probability *density* per unit value which is specified by the model, not the probability of a given value. This is illustrated in the right panel of the figure. If π is the probability shared between values in a very narrow range, width h units, the probability density is π/h.

8.1 The standard Gaussian distribution

The standard Gaussian distribution has probability density centred at 0. The probability density at any value z (positive or negative) is given by

$$0.3989 \exp\left[-\frac{1}{2}(z)^2\right].$$

A graph of this probability density for different values of z is shown in Fig. 8.2. There is very little probability outside the range ±3.

Tables of the standard Gaussian distribution are widely available, and these readily allow calculation of the probability associated with specified ranges of z. For our purposes it is necessary only to record that the probability corresponding to the range $(-1.645, +1.645)$ is 0.90 and that for the range $(-1.960, +1.960)$ is 0.95.

If the probability model for z is a standard Gaussian distribution then the probability model for $(z)^2$ is called the *chi-squared* distribution on one degree of freedom. Tables of chi-squared distributions can be used to find

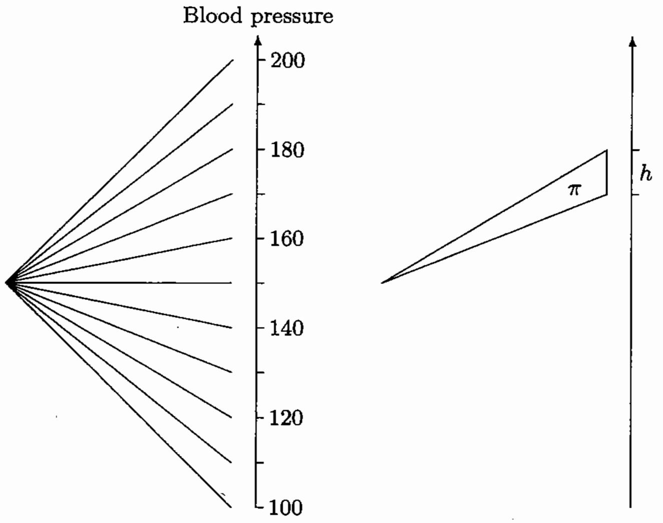


Fig. 8.1. Probability shared between many outcomes.

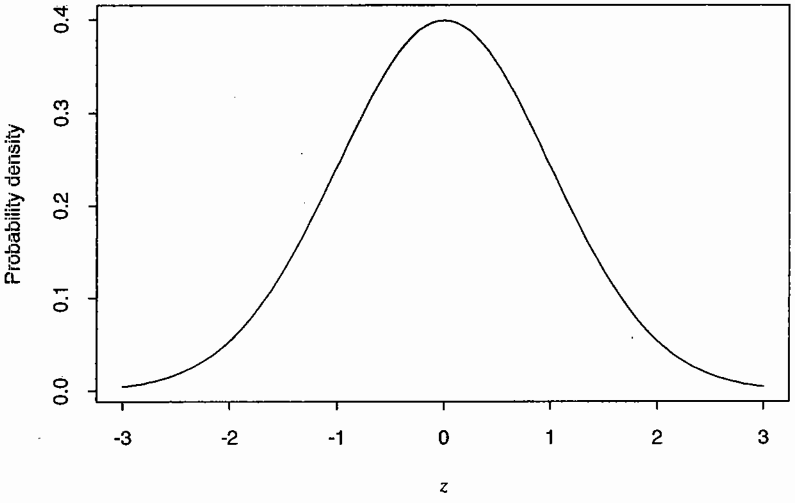


Fig. 8.2. The standard Gaussian distribution.

the probabilities of exceeding specified values of $(z)^2$ in the same way as tables of the standard Gaussian distribution are used to find probabilities of exceeding specified values of z.

Exercise 8.1. Use the tables in Appendix D to find the probability of exceeding the value 2.706 in a chi-squared distribution on one degree of freedom.

Note that, for $(z)^2$ to exceed 2.706, z must lie outside the range ± 1.645 of the standard normal distribution.

8.2 The general Gaussian model

It would be remarkable if the data we are analysing fell into the range -3 to $+3$, so for modelling the variability of real data, it is necessary to generalize the model to incorporate two parameters, one for the central value or *location*, and one for the spread or *scale* of the distribution. These are called the *mean* parameter and *standard deviation* parameter and are usually denoted by μ and σ respectively. A variable with such a distribution is derived by multiplying z by the scale factor and adding the location parameter. Thus

$$x = \mu + \sigma z.$$

has a distribution of the same general shape as the standard Gaussian distribution but centred around μ with most of its probability between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Exercise 8.2. If the mean and standard deviation of a general Gaussian distribution are 100 and 20 respectively, what ranges of values correspond to probabilities of 0.90 and 0.95 respectively?

Similarly, when x has a Gaussian distribution with mean μ and standard deviation σ then

$$z = \left(\frac{x - \mu}{\sigma}\right)$$

will have a *standard* Gaussian distribution. This fact can be used get the probability for a range of values of x using tables of z.

The probability density per unit of x when x has a Gaussian distribution with mean μ and standard deviation σ is

$$\frac{0.3989}{\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

This expression is obtained by substituting $(x - \mu)/\sigma$ for z in the probability density of a standard Gaussian distribution to obtain the probability density per σ units of x, and then dividing by σ to obtain the probability density per unit of x. Sometimes the distribution is described in terms of the square of σ, which is called the *variance*.

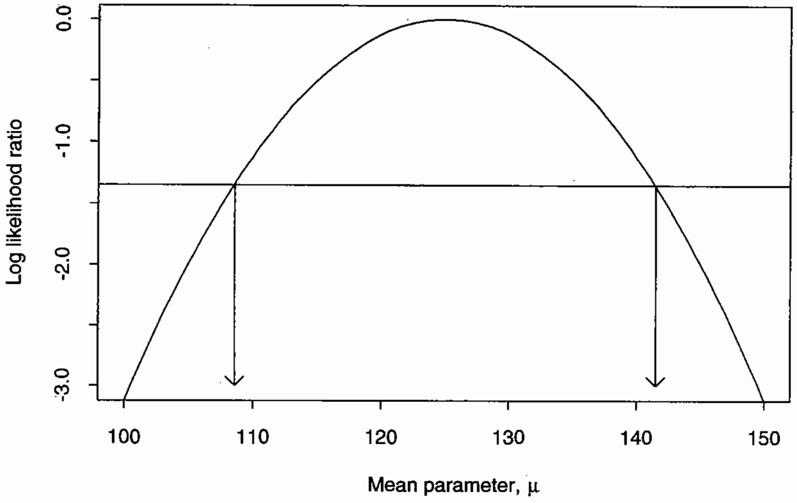


Fig. 8.3. The log likelihood ratio for the Gaussian mean, μ.

8.3 The Gaussian likelihood

Suppose a single value of x, say $x = 125$ is observed. Using the probability model that this is an observation from a Gaussian distribution with parameters μ and σ, the log likelihood for μ and σ is given by the log of the corresponding Gaussian probability density:

$$\log(0.3989) - \log(\sigma) - \frac{1}{2}\left(\frac{125-\mu}{\sigma}\right)^2 .$$

This log likelihood depends on two unknown parameters, but to keep things simple we shall assume that one of them, σ, is known from past experience to have the value 10. Omitting constant terms, the log likelihood for μ is then

$$-\frac{1}{2}\left(\frac{125-\mu}{10}\right)^2 .$$

The most likely value of μ is 125 and, since the above expression is zero at this point, this expression also gives the log likelihood *ratio* for μ. This is plotted in Fig. 8.3; curves with this shape are called *quadratic.*

We saw in Chapter 3 that we take the extremes of the supported range for a parameter to correspond to the value -1.353 for the log likelihood ratio. To find the limits of the supported range for μ we must therefore

solve the simple equation

$$-\frac{1}{2}\left(\frac{125-\mu}{10}\right)^2 = -1.353.$$

This takes only a few lines:

$$\begin{aligned}\left(\frac{125-\mu}{10}\right)^2 &= 2.706,\\ \left(\frac{125-\mu}{10}\right) &= \pm 1.645,\\ \mu &= 125 \pm 1.645 \times 10,\end{aligned}$$

so that supported values of μ are those between 108.6 and 141.5. In general, the log likelihood ratio for μ is

$$-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 ,$$

the most likely value of μ is the observation x, and the supported range for μ is

$$x \pm 1.645\sigma,$$

where σ is the standard deviation (which we assume to be known).

We saw in Exercise 8.1 that the probability of exceeding 2.706 in a chi-squared distribution is 0.10, and the probability corresponding to the range ±1.645 in the standard Gaussian distribution is 0.90. The fact that these numbers turn up in the above calculation is no accident and suggests that the log likelihood ratio criterion of -1.353 leads to supported ranges which have something to do with a probability of 0.90. This is indeed the case, but the relationship is not altogether straightforward and we shall defer this discussion to Chapter 10.

8.4 The likelihood with N observations

When there are N observations

$$x_1, x_2, \ldots, x_N,$$

the log likelihood for μ is obtained by adding the separate log likelihoods for each observation giving

$$\sum -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2 .$$

Let M refer to the mean of the observations,

$$M = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

It can be shown that the log likelihood can be rearranged as

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2 + \sum -\frac{1}{2}\left(\frac{x_i - M}{\sigma}\right)^2$$

where $S = \sigma/\sqrt{N}$, sometimes called the *standard error of the mean.* This rearrangement involves only elementary algebra and the details are omitted. The second part of this new expression for the log likelihood does not depend on μ and cancels in the log likelihood ratio for μ which is

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2,$$

The most likely value of μ is M, and setting the log likelihood ratio equal to -1.353 to obtain a supported range for μ gives

$$\mu = M \pm 1.645S.$$

As we would expect, with larger N, the value of S becomes smaller and the supported range narrower.

Exercise 8.3. The following measurements of systolic blood pressure were obtained from a sample of 20 men.

98	160	136	128	130	114	123	134	128	107
123	125	129	132	154	115	126	132	136	130

What is the most likely value for μ? Assuming that $\sigma = 14$, calculate the range of supported values for μ.

This exercise continues to make the unrealistic assumption, made throughout this chapter, that σ is *known.* In practice it must almost invariably be estimated from the data. We shall defer discussion of this until Chapter 34.

Solutions to the exercises

8.1 The probability of exceeding 2.706 in the chi-squared distribution with one degree of freedom is 0.10.

8.2 The range corresponding to a probability of 0.9 is

$$100 \pm 1.645 \times 20 = (67.1, 132.9)$$

and, for a probability of 0.95,

$$100 \pm 1.96 \times 20 = (60.8, 139.2).$$

8.3 The mean of the 20 measurements is 128.00 and this is the most likely value of μ. To calculate the supported range for μ, we first calculate

$$S = \frac{\sigma}{\sqrt{N}} = \frac{14}{\sqrt{20}} = 3.13$$

so that the range lies between

$$\mu = 128.00 \pm 1.645 \times 3.13$$

that is from 122.9 to 133.1 .

8 The Gaussian probability model

The above photograph is of the front of the 10 (zehn) Deutsche Mark banknote, before Germany and the other EC countries adopted the Euro.

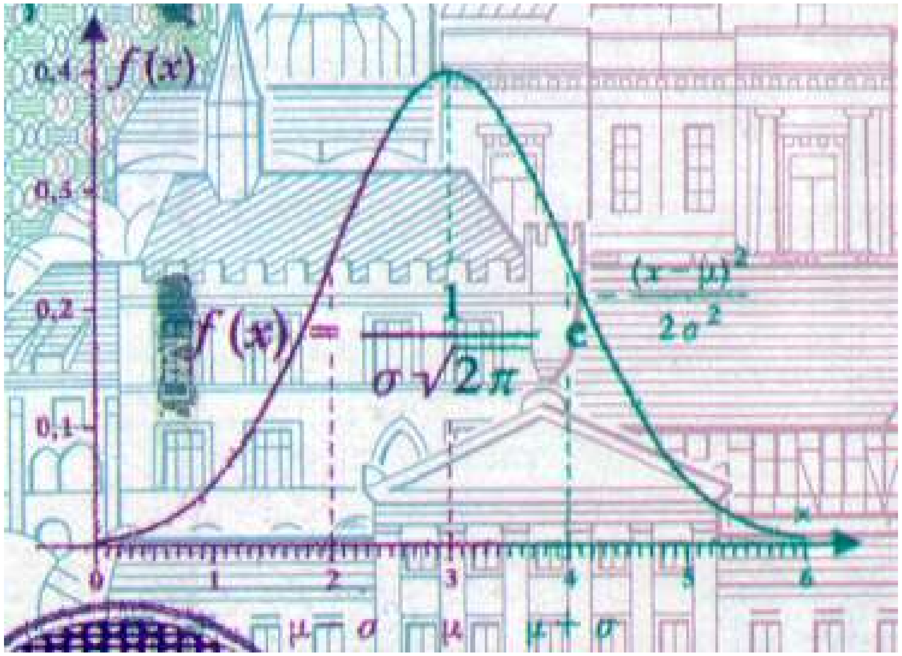
You can find other images by Googling 'Gauss 10dm image.'

For biographies of and historical notes on Gauss, see specialized sites such as http://www.mathunion.org/general/prizes/gauss/details/ or general ones such as http://en.wikipedia.org/wiki/Carl_Gauss. The story of the discovery of Ceres http://www.keplersdiscovery.com/Asteroid.html is interesting as it mentions the Method of Least Squares. There is a debate as to who first used the principle of *Least Squares*. It seems from this website that Gauss did, but it also seems he did not publish the method, so we cannot tell if he did in fact used it in the re-discovery of Ceres.

Legendre[1] published his method in 1806, and described the procedure very much the way it is taught today. Gauss seems to get the credit in this website http://en.wikipedia.org/wiki/Least_squares.

There is a chapter on 'Least Squares and the Combination of Observations' in Stephen Stigler's most readable and interesting book "The History of Statistics: The Measurement of Uncertainty before 1900."

Even though we generally credit Fisher with the development of Maximum Likelihood methods, it seems that Gauss used the principle.



Note that although C&H say "in the Gaussian model, the total probability of 1 is shared among many values," this statement applies to any model for a 'continuous' random variable.

'*Probability density per unit value*': this is a good description. This is how we should label the vertical axis of a pdf graph.

It is also why we can write likelihood contribution of an 'observed' y value, e.g., blood pressure of 90, as $pdf(y_{mid}, \theta)\ h$, where y_{mid} is the midpoint of the interval (of width h) that is reported as a '90'. Technically, we should use as $CDF(y_{upper}, \theta) - CDF(y_{lower}, \theta)$ but usually h is sufficiently narrow that the rectangular area $pdf(y_{mid}, \theta) \times h$ is a good approximation. Of course, since h is a constant, and does not involve θ, it is usually omitted from the likelihood contribution.

8.1 Standard Gaussian distribution

Because the pdf tends to be written with $1/\sqrt{2\pi}$ in front, then, unless we evaluate this expression, we don't get to see that it is indeed very close to 0.4.

Another point that gets overlooked is *where* (i.e., how far up the vertical axis) the '*point of inflection*' is to be found: i.e., imagine one were to rub one's

[1] http://en.wikipedia.org/wiki/Adrien-Marie_Legendre
http://www.nndb.com/people/891/000093612/
http://www.nndb.com/people/891/000093612/
http://www.britannica.com/EBchecked/topic/334063/least-squares-approximation

finger upwards along the curve: where does it change from concave upwards to convex upwards?

Supplementary Exercise 8.1. Determine, analytically/numerically, where on the z scale the point of inflexion is located, and at what height (express its vertical location as a fraction of the max. height of $1/\sqrt{2\pi} = 0.3989$). *The point of inflection helps one to draw a reasonably accurate pdf curve 'freehand.'*

The diagram on the next page shows how one could use a 'spinner' or 'roulette wheel' to generate **random numbers that follow a Gaussian distribution**, while the one on the page thereafter shows how to do so using the reverse CDF. In fact, both use the same idea. Francis Galton[2] proposed a method that used 6 dice.

Supplementary Exercise 8.2.
(i) Derive the 24 ordinate values used by Galton for his *Die I*.[3] Do so in R (or Excel), by dividing up the $(0, \infty)$ scale into 24 bins each containing equal 1/48ths of the total probability mass, and finding (as Galton did) the ordinate at the mid point of (the base of) each bin.
(ii) Out of curiosity, how many would change if he were to use the centres of gravity (mass) rather than the midpoints? Would the 'σ' be closer to 1?
(iii) Then, use R (or Excel) to simulate 12 throws of *Die I*, and the requisite number of throws of *Dies II* and *III*, to produce 12 values from a $N(\mu = 100,\ \sigma = 15)$ distribution.
(iv) In the study carried out at the University of Canterbury [cf Resources], the investigators say that "To determine the success of this experiment, we formulate the following question as a statistical hypothesis test: Are our sampled values taken independently and identically from an appropriate discrete distribution which approximates Galton's normal distribution?" To answer it, they used the data reported in their Appendix A.
If they had approached you about how to address their question, how would you have answered them? Would you have advised them to collect results of actual throws of the dice? Would you have calculate the optimal number of trials needed for the test in the same way that they did?

8.2 General Gaussian Model

Shouldn't textbooks use y rather than x when dealing with random variables? This notation is important when we come to regression: most applied work involves y's, each of which is the realization from a conditional-on-x distribution whose parameters are governed by a linear combination of a (possibly-vector-valued) θ and a (possibly-vector-valued) x, with x treated as fixed-by-design. JH notes that the designers of the German 10DM banknote to honour Gauss also used x rather than y.

The square of sigma is called the variance: In the very interesting website on Earliest Known Uses of Some of the Words of Mathematics[4] we find: MODULUS (in the Theory of Errors). In his first theory of least squares based on the normal distribution and presented in Gausss Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum (1809) Gauss used a measure of precision ("mensura praecisionis observationum" (p. 245) which he denoted by h: the reciprocal of h is $\sqrt(2)\sigma$, where σ is the standard deviation. Both h and its reciprocal have been called the modulus: the reciprocal in G. B Airy's On the Algebraical and Numerical Theory of Errors of Observation and the Combination of Observations (1861, p. 15) and h in E. T. Whittaker & G. Robinson's Calculus of Observations (1924, p. 175).

This website could keep one occupied for many hours, as on the same page of 'M' alone, there are entries for MARGIN OF ERROR, MARKOV CHAIN, MONTE CARLO, MARKOV CHAIN MONTE CARLO, MARTINGALE, MAXIMUM LIKELIHOOD, MEAN, MEDIAN, META-ANALYSIS, MINIMUM CHI-SQUARED, MODE, MOMENT, Moment generating function, MONTY HALL PROBLEM, MORAL EXPECTATION, MOVING AVERAGE, MULTICOLLINEARITY, MULTINOMIAL DISTRIBUTION, and MULTIVARIATE.

In the 'S' page (also full of other interesting terms) we can read that 'The term STANDARD DEVIATION was introduced by Karl Pearson (1857-1936) in 1893, "although the idea was by then nearly a century old" (Abbott; Stigler, page 328). According to the DSB: The term "standard deviation" was introduced in a lecture of 31 January 1893, as a convenient substitute for the cumbersome "root mean square error" and the older expressions "error of mean square" and "mean error."

8.3 The Gaussian Likelihood

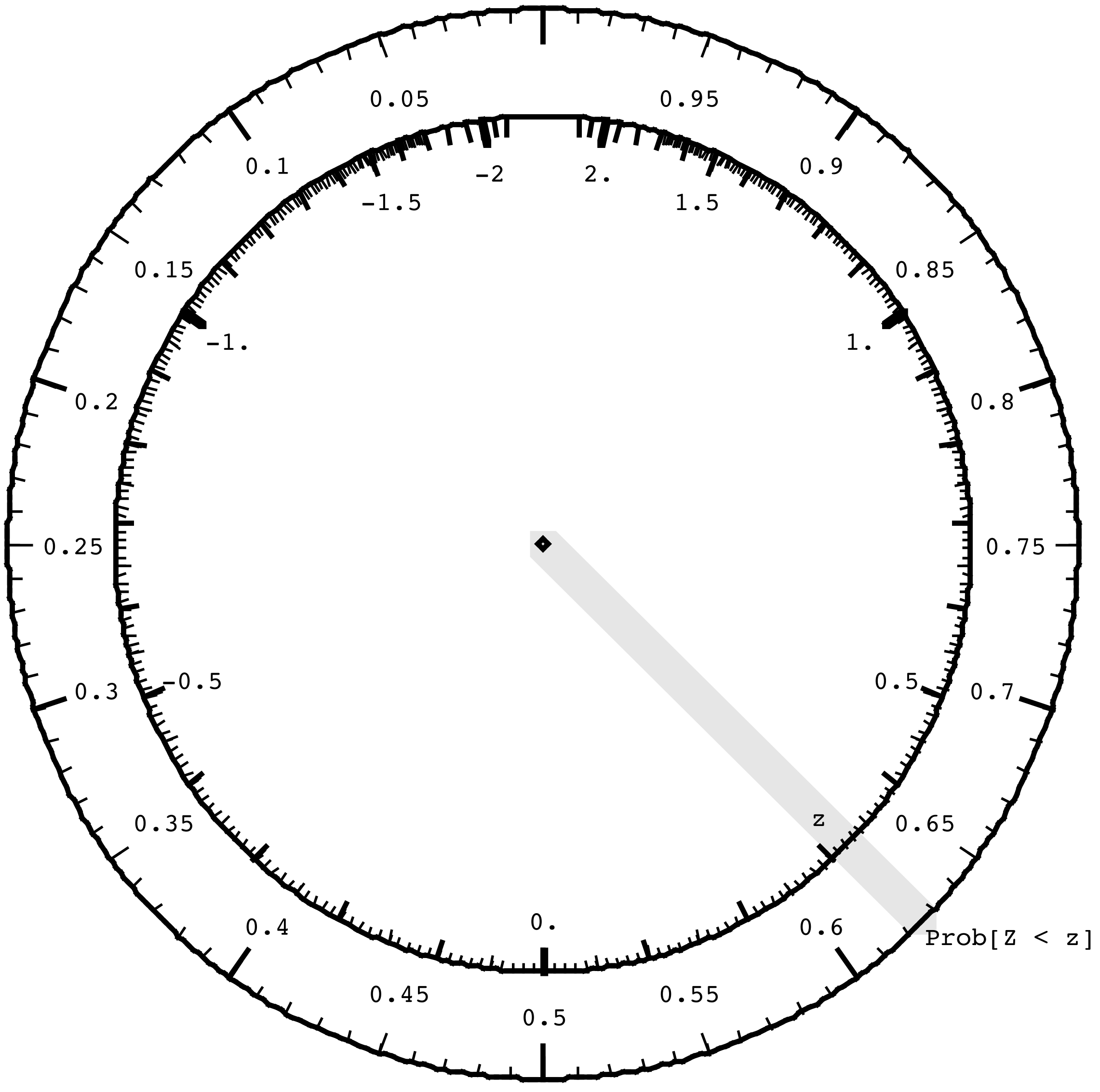
C&H finally explain why they adopted a cutoff on 1.353 back in chapter 3.

[2] Dice for Statistical Experiments. Nature(1890) 42 13-14 – in Resources, with related material.

[3] Note that whereas today we use the *Standard Deviation* (SD) as a measure of the spread of a Normal distribution [and use the '68-95-99.7 rule' for 1, 2 and 3 SD's – see `http://en.wikipedia.org/wiki/68-95-99.7_rule`] – Galton used the smaller '*Probable Error*' or 'PE'. The PE is approximately 2/3rds (0.6745) times the SD. The *Probable Error* gets its name from the fact that a deviate from the mean is just as likely (50%, 'as probable as not'), to be bigger than as smaller than the Probable Error: the interval $\mu \mp 1PE$ contains the middle 50% of the N(μ, PE) distribution, whereas the the interval $\mu \mp 1SD$ contains the middle 68%. Notice also that by a using a smaller measure of spread, Galton's grades or 'degrees' of 'extremeness' ran from -5° to +5°, where we might speak or write of Z values or Z scores from say -3 to +3.

[4] http://jeff560.tripod.com/m.html

Generating random numbers from a Gaussian Distribution / Connections with the Normal (Gaussian) Tables



Imagine a disk or "Spinner" with 2 concentric circles, and a spindle through the centre. Suppose that when spun it is equally likely to come to rest at any point on the outer circumference. This is reflected in markings of 0 to 1 (or, if you prefer, % to 100%) uniformly on the circumference of the outer circle.

Q: How should we mark the circumference of the inner circle so that repeated spins produce values with a Gaussian N(0,1) distribution? [see "spinner" in fig 4.9 page 317 of M&M]

A: Use the z values corresponding to the percentiles of the Gaussian Distribution!

Then, the spinner shown will produce Z values from minus to plus infinity..

IMPLICATIONS FOR MONTE CARLO (SIMULATION) WORK

1 Generate numbers with a Uniform Distribution on (0,1)

e.g. in Excel use the RAND() function

i.e. generate P = RAND()

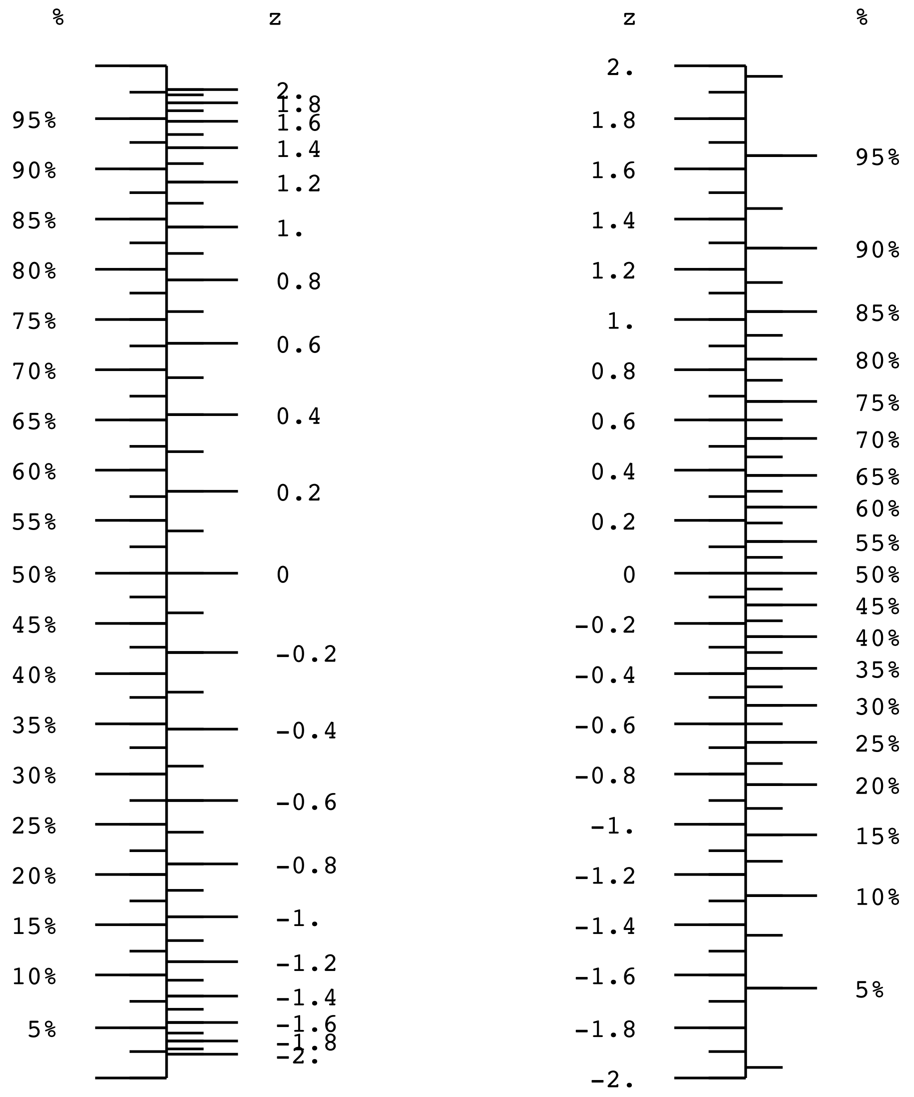
2 Calculate percentile corresponding to P

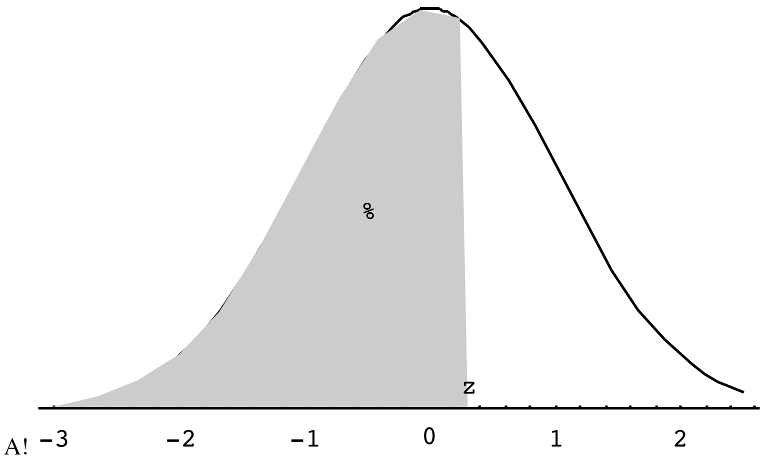
i.e. z = Z value such that Prob(Z < z) = P

in Excel, use NORMINV function,

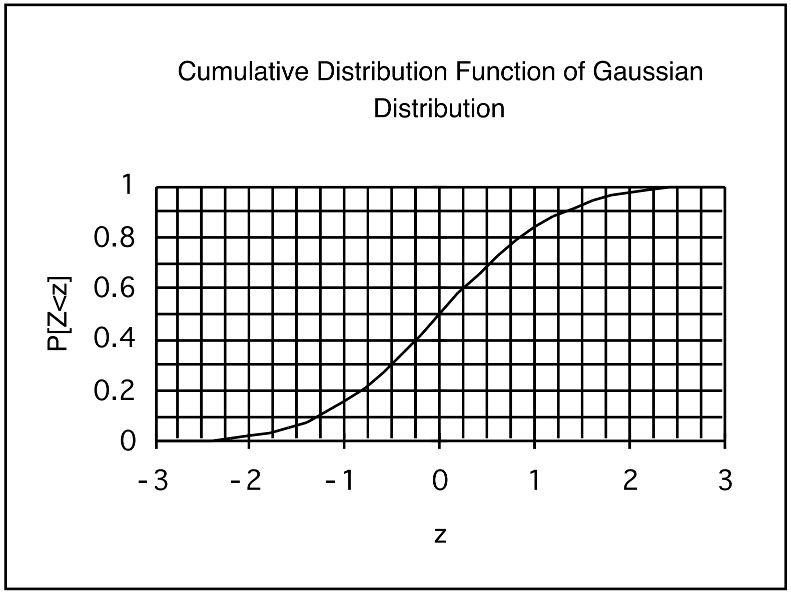
i.e.
calculate z = NORMINV(P,μ=0,σ=1)

Generating random numbers from a Gaussian Distribution / Connections with the Normal (Gaussian) Tables





Another way of visualizing the Table is given below. To generate a random Z, enter randomly at the vertical axis and find corresponding Z value!



T

The above **nomograms** illustrate the same idea: the function links the shaded area under the Gaussian curve with the corresponding z value. I t is shown , first with area or Percent or Pr(Z<z) as a function of z, and then vice-versa (as is done in Table A of M&M). Table A tabulates Prob[Z<z] as a function of z, but one can travel in either direction.

8.4 The Likelihood with N observations

.

JH prefers n **over** N. To him, N refers to the size of some universe of units, not to the size of a sample of units from it. In the social sciences, and particularly in the Publication Manual of the American Psychological Association[5], its JH's understanding that if the *overall* sample is $N = 20$ people, 9 men and 11 women, the *sub-sample* sizes of 9 and 11 would be referred to as $n = 9$ and $n = 11$. I understand from a reliable source that Clayton pursued a PhD in psychology, so that might explain his notation.

Likewise, JH is unsure why C&H use the letter M where we would normally write $\bar{x}$ [or $\bar{y}$!]

Supplementary Exercise 8.3. C&H say it requires only elementary algebra to rearrange the log likelihood. Do the algebra, and verify that that is indeed true.

[5] http://www.apastyle.org/manual/index.aspx

9
Approximate likelihoods

Because the Gaussian log likelihood for the mean parameter, μ, takes the simple form

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2$$

the supported range for μ also takes a simple form, namely

$$M \pm 1.645S.$$

For log likelihoods such as the Bernouilli and Poisson there is no simple algebraic expression for the supported range, and the values of the parameters at which the log likelihood is exactly -1.353 must be found by systematic trial and error. However, the shapes of these log likelihoods are *approximately* quadratic, and this fact can be used to derive simple formulae for approximate supported ranges. Methods based on quadratic approximation of the log likelihood are particularly important because the quadratic approximation becomes closer to the true log likelihood as the amount of data increases.

9.1 Approximating the log likelihood

Consider a general likelihood for the parameter, θ, of a probability model and let M be the most likely value of θ. Since the quadratic expression

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2$$

has a maximum value of zero when $\theta = M$ it can be used to to approximate the true log likelihood ratio, after an appropriate value of S has been chosen. Small values of S give quadratic curves with sharp peaks and large values of S give quadratic curves with broad peaks. We shall refer to S as the standard deviation of the estimate of θ. Alternatively, it is sometimes called the *standard error* of the estimate.

Once M has been found and S chosen, an approximate supported range for θ is found by solving the equation

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2 = -1.353,$$

to give

$$\theta = M \pm 1.645S.$$

Full details of how S is chosen are given later in the chapter, but for the moment we shall give formulae for S, without justification, and concentrate on how to use these in practice.

THE RISK PARAMETER

The log likelihood for π, the probability of failure, based on D failures and $N - D$ survivors is

$$D\log(\pi) + (N-D)\log(1-\pi).$$

The most likely value of π is D/N. To link with tradition we shall also refer to the most likely value of π as P (for proportion). The value of S which gives the best approximation to the log likelihood ratio is

$$S = \sqrt{\frac{P(1-P)}{N}}.$$

For the example we worked through in Chapter 3, $D = 4$ and $N = 10$ so that the value of P is 0.4 and

$$S = \sqrt{\frac{0.4 \times 0.6}{10}} = 0.1549.$$

An approximate supported range for π is given by

$$0.4 \pm 1.645 \times 0.1549$$

which is from 0.15 to 0.65, while the supported range obtained from the true curve lies from 0.17 to 0.65. The true and approximate log likelihood curves are shown in Fig. 9.1. The curve shown as a solid line is the true log likelihood ratio curve, while the broken line indicates the Gaussian approximation.

THE RATE PARAMETER

The log likelihood for a rate λ based on D cases and Y person years is

$$D\log(\lambda) - \lambda Y.$$

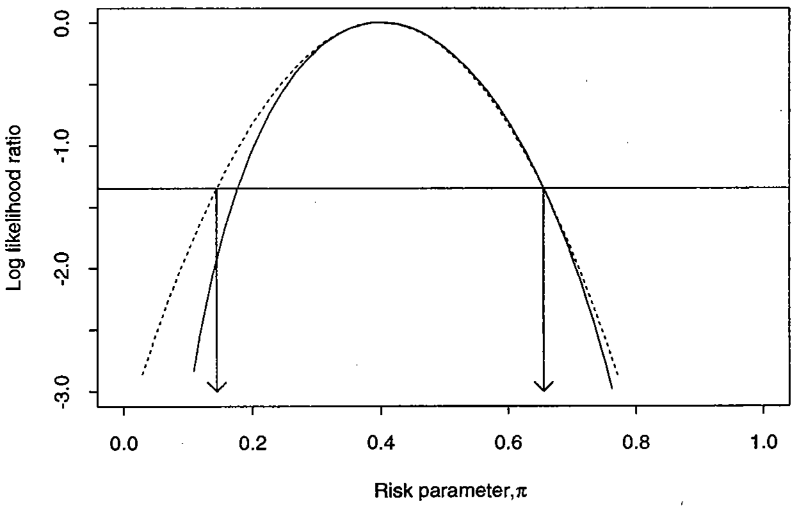


Fig. 9.1. True and approximate Bernouilli log likelihoods.

The most likely value of λ is D/Y and the value of S which gives the best approximation to the log likelihood ratio is

$$S = \frac{\sqrt{D}}{Y}.$$

For the example in Chapter 5, $D = 7$ and $Y = 500$. The most likely value of λ is 0.014 and

$$S = \sqrt{7}/500 = 0.00529.$$

An approximate supported range for λ is therefore

$$0.014 \pm 1.645 \times 0.00529$$

which is from 5.3/1000 to 22.7/1000. The true (solid line) and approximate (broken line) log likelihood ratio curves are shown in Fig. 9.2. The range of support obtained from the true curve spans from 7.0 to 24.6 per 1000.

Exercise 9.1. Find the approximate supported range for π, the probability of failure, based 7 failures and 93 survivors. Find also the approximate supported range for λ, the rate of failure, based on 30 failures over 1018 person-years.

9.2 Transforming the parameter

The Gaussian log likelihood curve for μ is symmetric about M and extends *indefinitely* to either side. However, the parameters of some probability models are not free to vary in this manner. For example, the rate parameter λ can take only positive values, and the risk parameter must lie between 0 and 1. Approximate supported ranges for such parameters calculated from the Gaussian approximation can, therefore, include impossible values.

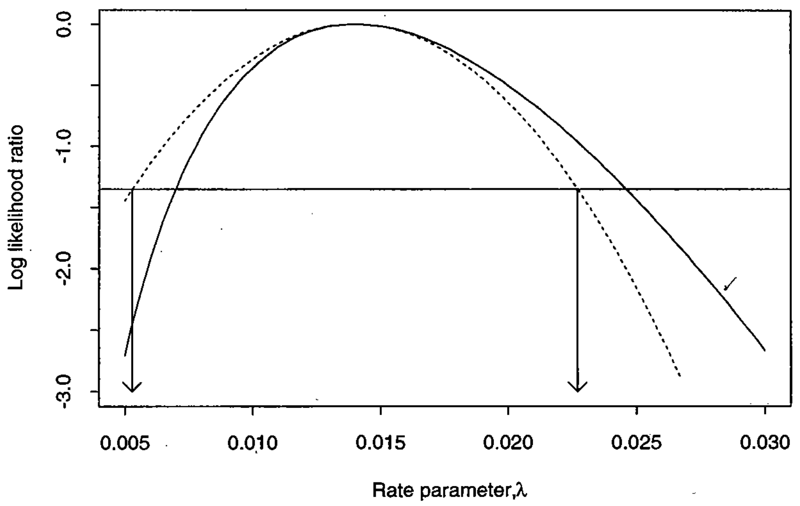


Fig. 9.2. True and approximate Poisson log likelihoods.

The solution to this problem is to find some function (or *transformation*) of the parameter which is unrestricted and to first find an approximate supported range for the transformed parameter.

THE LOG RATE PARAMETER

The rate parameter λ can take only positive values, but its logarithm is unrestricted. To calculate an approximate supported range for λ it is better, therefore, to first calculate a range for $\log(\lambda)$, and then to convert this back to a range for λ. Note that the range for $\log(\lambda)$ will always convert back to positive values for λ. To find the approximate range for $\log(\lambda)$ we need a new value of S — that which gives the best Gaussian approximation to the log likelihood ratio curve when plotted against $\log(\lambda)$. When a rate λ is estimated from D failures over Y person-years, this value of S is given by

$$S = \sqrt{1/D}.$$

Fig. 9.3 illustrates this new approximation for our example in which $D = 7$ and $Y = 500$ person-years. Here,

$$S = \sqrt{1/7} = 0.3780,$$

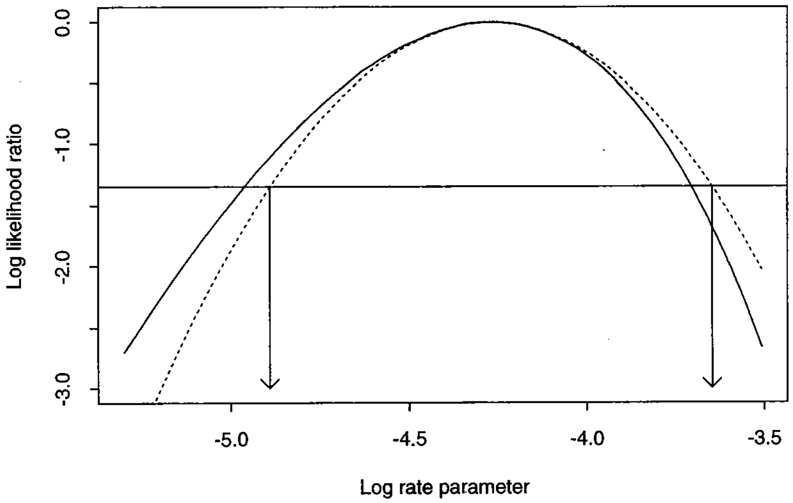


Fig. 9.3. Approximating the log likelihood for $\log(\lambda)$.

and an approximate supported range for $\log(\lambda)$ is

$$\log(7/500) \pm 1.645 \times \sqrt{1/7},$$

which is from -4.890 to -3.647. The range for λ is therefore from $\exp(-4.890)$ to $\exp(-3.647)$ which spans from 7.5/1000 to 26.1/1000.

A more convenient way of carrying out this calculation is suggested by noting that the limits of the range for λ are given by

$$\frac{7}{500} \overset{\times}{\div} \exp\left(1.645\sqrt{\frac{1}{7}}\right) = 0.014 \overset{\times}{\div} 1.862.$$

The range is then from $0.014/1.862 = 7.5/1000$ to $0.014 \times 1.862 = 26.1/1000$, as before. We shall refer to the quantity

$$\exp(1.645S)$$

as an *error factor*.

THE LOG ODDS PARAMETER

The same thing can be done when calculating a supported range for the risk parameter π based on D failures in N subjects. The value of π is restricted on both sides, by 0 on the left and by 1 on the right. The value of $\log(\pi)$ is still restricted on the right by zero because $\log(1) = 0$, but $\log(\Omega)$, where Ω is the odds corresponding to π, is not restricted at all. Hence we first find a range for $\log(\Omega)$ and then convert this back to a range for π. The most likely value of $\log(\Omega)$ is

$$M = \log\left(\frac{D}{N-D}\right)$$

and the value of S for approximating the log likelihood for $\log(\Omega)$ is

$$S = \sqrt{\frac{1}{D} + \frac{1}{N-D}}.$$

For the example where $D = 4$ and $N - D = 6$,

$$S = \sqrt{\frac{1}{4} + \frac{1}{6}} = 0.6455,$$

and an approximate supported range for $\log(\Omega)$ is given by

$$\log\left(\frac{4}{6}\right) \pm 1.645 \times 0.6455,$$

that is, from -1.4673 to 0.6564. This is a range for $\log(\Omega)$ and it is equivalent to a range for Ω from $\exp(-1.4673) = 0.231$ to $\exp(0.6564) = 1.928$. This can be calculated more easily by first calculating the error factor

$$\exp(1.645 \times 0.6455) = 2.892.$$

The most likely value of Ω is $4/6 = 0.667$, so that the supported range for Ω is

$$0.667 \overset{\times}{\div} 2.892$$

that is, from 0.231 to 1.928 as before. Finally, remembering that $\pi = \Omega/(1+\Omega)$, the range for π is given by

$$\frac{0.231}{1.231} \text{ to } \frac{1.928}{2.928}$$

which is from 0.19 to 0.66.

Some of the more commonly used values of S obtained by approximating the log likelihood are gathered together in Table 9.1.

Exercise 9.2. Repeat Exercise 9.1 by first finding 90% intervals for $\log(\Omega)$ and $\log(\lambda)$ respectively, and then converting these to intervals for π and λ.

Exercise 9.3. Repeat the above exercise using error factors.

Table 9.1. Some important Gaussian approximations

Parameter	M	S
π	$D/N = P$	$\sqrt{P(1-P)/N}$
λ	D/Y	$\sqrt{D}/Y$
$\log(\Omega)$	$\log[D/(N-D)]$	$\sqrt{1/D + 1/(N-D)}$
$\log(\lambda)$	$\log(D/Y)$	$\sqrt{1/D}$

★ 9.3 Finding the best quadratic approximation

We now return to the problem of how to determine the values for M and S. To do this we need some elementary ideas of calculus summarized in Appendix B. In particular, we need to be able to find the *gradient* (or *slope*) of the log-likelihood curve together with its *curvature*, which is defined as the rate of change of the gradient. The mathematical terms for these quantities are the first and second *derivatives* of the log likelihood function.

The value of M can be found by a direct search for that value of of θ which maximizes the log likelihood, but it is often easier to find the value of θ for which the gradient of the log likelihood is zero; this occurs when $\theta = M$.

The value of S is chosen to make the curvature of the quadratic approximation equal to that of the true log likelihood curve at M, thus ensuring that the true and approximate log likelihoods are very close to each other near $\theta = M$. The quadratic approximation to the log likelihood ratio is

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2,$$

and the rules summarized in Appendix B show that the curvature of this is constant and takes the value

$$-\frac{1}{(S)^2}.$$

We therefore choose the value of S to make $-1/(S)^2$ equal to the curvature of the true log likelihood curve at its peak.

THE RATE PARAMETER

The log likelihood for a rate λ is

$$D\log(\lambda) - \lambda Y.$$

Using the rules of calculus given in Appendix B the gradient of $\log(\lambda)$ is $1/\lambda$ and the gradient of λ is 1. Hence the gradient of the log likelihood is

$$\frac{D}{\lambda} - Y.$$

The maximum value of the log likelihood occurs when the gradient is zero, that is, when $\lambda = D/Y$, so the most likely value of λ is D/Y. The curvature of a graph at a point is defined as the rate of change of the gradient of the curve at that point. The rules of calculus show this to be

$$-\frac{D}{(\lambda)^2}.$$

The peak of the log likelihood occurs at $\lambda = D/Y$ so the curvature at the peak is found by replacing λ by D/Y in this expression to obtain

$$-\frac{(Y)^2}{D}.$$

Setting this equal to $-1/(S)^2$ gives

$$S = \sqrt{D}/Y,$$

which is the formula quoted earlier.

THE RISK PARAMETER

The log likelihood for the probability π based on D positive subjects out of a total of N is

$$D\log(\pi) + (N-D)\log(1-\pi).$$

The gradient of the log likelihood is

$$\frac{D}{\pi} - \frac{N-D}{1-\pi}$$

which is zero at $\pi = D/N$, also referred to as P. The gradient of the gradient is

$$-\frac{D}{(\pi)^2} - \frac{N-D}{(1-\pi)^2},$$

so the curvature at $\pi = P$ is

$$-\frac{D}{(P)^2} - \frac{N-D}{(1-P)^2}.$$

Replacing D by NP and $N-D$ by $N(1-P)$, this reduces to

$$-\frac{N}{P(1-P)}$$

so

$$S=\sqrt{\frac{P(1-P)}{N}}.$$

★ 9.4 Approximate likelihoods for transformed parameters

When the log likelihood for a parameter is plotted against the log of the parameter rather than the parameter itself, the curvature at the peak will be different. For example, the log likelihood for a rate parameter λ is

$$D\log(\lambda)-\lambda Y.$$

Plotting this against $\log(\lambda)$ is the same as expressing the log likelihood as a function of $\log(\lambda)$. To do this we introduce a new symbol β to stand for $\log(\lambda)$, so

$$\beta=\log(\lambda), \quad \lambda=\exp(\beta).$$

In terms of β the log likelihood is

$$D\beta-Y\exp(\beta).$$

The gradient of this with respect to β is

$$D-Y\exp(\beta)$$

and the curvature is

$$-Y\exp(\beta).$$

The most likely value of $\exp(\beta)$ (which equals λ) is D/Y, so the curvature at the peak is

$$-Y\times(D/Y)=-D.$$

It follows that

$$S=\sqrt{1/D}.$$

In general, derivations such as that above can be simplified considerably by using some further elementary calculus which provides a general rule for the relationship between the values of S on the two scales. In the case of the log transformation, this rule states that multiplying the value of S on the scale of λ by the gradient of $\log(\lambda)$ at $\lambda=M$ gives the value of S on the scale of $\log(\lambda)$. The rules of calculus tell us that, at $\lambda=M$, the gradient of the graph of $\log(\lambda)$ against λ is $1/M$. Since, on the λ scale, $M=D/Y$ and $S=\sqrt{D}/Y$, the rule tells us that the value of S for $\log(\lambda)$ is

$$\frac{\sqrt{D}}{Y}\times\frac{Y}{D}=\sqrt{\frac{1}{D}}.$$

This agrees with the expression obtained by the longer method.

A similar calculation shows that the curvature of the Bernouilli log likelihood, when plotted against $\log(\Omega)$, the log odds, is given by

$$S=\sqrt{\frac{1}{D}+\frac{1}{N-D}}.$$

Solutions to the exercises

9.1 An approximate supported range for π is given by

$$0.07\pm 1.645S$$

where $S=\sqrt{0.07\times 0.93/100}$. This gives a range from 0.028 to 0.112. An approximate supported range for λ is given by

$$30/1018\pm 1.645S$$

where $S=\sqrt{30}/1018$. This gives a range from 21/1000 to 38/1000.

9.2 The approximate supported range for $\log(\Omega)$ is given by

$$\log(7/93)\pm 1.645S$$

where

$$S=\sqrt{\frac{1}{7}+\frac{1}{93}}=0.3919.$$

This gives a range from -3.231 to -1.942. The range for Ω is from 0.040 to 0.143, and the range for π is from 0.038 to 0.125.
The approximate supported range for $\log(\lambda)$ is given by

$$\log(30/1018)\pm 1.645S$$

where

$$S=\sqrt{1/30}=0.1826.$$

This gives a range from -3.825 to -3.224. The range for λ is from 22/1000 to 40/1000.

9.3 The error factor for Ω is

$$\exp(1.645 \times 0.3919) = 1.905.$$

The most likely value for Ω is $7/93 = 0.075$ and the range for Ω is from $0.075/1.905 = 0.040$ to $0.075 \times 1.905 = 0.143$. The range for π is from 0.038 to 0.125.
The error factor for the rate is

$$\exp(1.645 \times 0.1826) = 1.350.$$

The most likely value of the rate is 29/1000 with range from $29/1.350 = 22$ per 1000 to $29 \times 1.350 = 40$ per 1000.

9 Approximate Likelihoods

This is a central chapter; you will use the Normal Approximation to the (sampling) distribution of ML parameter estimates throughout your career. The key is to work in the most appropriate parameter scale, the one where the sampling distribution is 'closest to Gaussian.'

Even though the title says approximate *likelihoods*, in fact the chapter is entirely about approximate *log likelihoods*. Indeed, just as R. A. Fisher did in his very first paper on this topic, exactly 100 years ago, we should always focus on the *log likelihood*. In that 1912 paper, Fisher never defined the likelihood, only its log. (In fact he did not call it the log *likelihood* ... that term came later. He just called it an absolute '*criterion*').

C&H write of there being no *simple algebraic expression* (or closed form) for the supported range for the parameters of the Binomial and Poisson models. In fact, the same is true *much more broadly*: this was also the case in just about all of the ML estimation problems we have dealt with so far (e.g. σ in Fisher's binned errors data, and in Tibshirani et al's 'accuracy of dart throws' data; μ and σ in Galton's data on the speeds of homing pigeons; the shape and rate/scale parameters of the gamma model for tumbler longevity; the HIV infection rate parameter λ in the circumcision studies; the proportion (π) of persons infected by West Nile virus; the parameters in the mutation rate function behind the genetic data from Iceland, etc. etc.. And this absence of a closed form will also be the norm for the parameters in many many *regression* models. Indeed, in most parameter-fitting applications, we do *not even have a closed form algebraic expression for the point estimates*, let alone their standard errors. So, it is important that we learn how to use a *more general approach.*

"*The quadratic approximation becomes closer to the true log likelihood as the amount of data increases*": The authors are referring to the role of the *Central Limit Theorem*, and to the near-Gaussian sampling distribution of $\hat{\theta}_{MLE}$.

9.1 Approximating the log likelihood

"*Consider a general likelihood for the parameter, θ, of a probability model and let M be the most likely value of θ. Since the quadratic expression*

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2$$

has a maximum value of zero when $\theta = M$, it can be used to approximate the true log likelihood ratio."

If JH were writing this, he would have said

"***Consider a general likelihood for the parameter, θ, of a probability model and let $\hat{\theta}_{ML}$ be the most likely value of θ. Since the quadratic expression***

$$-\frac{1}{2}\left(\frac{\hat{\theta}_{ML}-\theta}{S}\right)^2$$

has a maximum value of zero when $\theta = \hat{\theta}_{ML}$, it can be used to approximate the true log likelihood ratio."

"*We shall refer to S as the standard deviation of the estimate of θ. Alternatively, it is sometimes called the standard error of the estimate*"

Again, JH would have written:

"***We shall refer to $SE[\hat{\theta}_{ML}]$ as the standard deviation of the estimate of θ. Alternatively, it is sometimes called the standard error of the estimate***"

and (dropping the awkward *ML* subscript in the *SE*) that

"***Since the quadratic expression***

$$-\frac{1}{2}\left(\frac{\hat{\theta}_{ML}-\theta}{SE[\hat{\theta}]}\right)^2$$

has a maximum value of zero when $\theta = \hat{\theta}_{ML}$, it can be used to approximate the true log likelihood ratio."

Formulae for S [for now, without justification].

THE RISK [i.e., the PROPORTION, π] PARAMETER

As C&H illustrate in the next section, it is better to work in the logit[π] scale, unless the Normal approximation to the binomial itself is adequately accurate. Had the data been 40+ and 60−, rather than 4+ and 6−, the log-likelihood in the original, untransformed, (i.e., π) scale would have looked a lot closer to quadratic, i.e., the sampling distribution of a sample proportion would be close-enough-to-Normal for much more of the π range.

A common problem with using the Normal approximation in the case of limited-range parameters is that it works well enough at the less extreme limit (in this case, for values near the middle of the range) but poorly for parameter values near the more extreme limit (in this case, for values near the bottom of the 0-1 range). JH likes to say that there isn't enough room for a Gaussian distribution if we are at the lower end (0/10, 1/10, 2/10, ...) of the Binomial$(10, \pi)$ distribution, particularly if we entertain π values $<$ 0.5 (which case we would 'expect' to have $n \times \pi < 5$ in the sample with the characteristic/event of interest.

THE RATE [i.e., the INTENSITY, λ] PARAMETER[6]

"*The value of S (i.e., the value of $SE[\hat{\lambda}]$) which gives the best approximation to the log likelihood ratio is $S = \sqrt{D}/PT$* "

The reason for this form is because $\hat{\lambda}_{ML}$ has the form $\hat{\lambda}_{ML} = D/PT$, where D is the realization of a Poisson r.v., and PT is a known constant.

The lower limit of 5.3/1000 is the result of using the Normal approximation to a Poisson distribution. But a rate of 5.3/1000 means that with PT=500 person-time units, we would expect $\mu = (5.3/1000) \times 500 = 2.65$ 'events', and we know that the Poisson$[\mu = 2.65]$ distribution is quite skewed, with a lot of probability mass on its lower tail values of 0, 1 and 2, so it is not possible to approximate it by a Normal distribution − the lower tail of a $N[\mu = 2.65, \sigma = \sqrt{2.65}]$ distribution has an embarrassingly large amount its mass below 0!

Imagine what the lower limit would be if the observed count was $D = 2$: then the lower limit for the rate, based on the Normal approximation, would be $(2-1.645\times\sqrt{2})/500$, a negative rate! Using the form $(D\mp1.645\times\sqrt{D})/500$ (i.e. dealing first with the statistical uncertainty in the numerator, then dividing by the 'constant', 500) makes it much clearer where the 'weak link' is –it's the *numerator*!

Again, a parameter-transformation would help (although, with a μ this low, it is difficult to come up with any parameter scale on which the Normal distribution would be a good approximation – there just isn't enough '*granularity*').

9.2 Transforming the parameter

As was emphasized at the beginning, this is the key to more accurate approximations.

"*The solution to this problem is to find some function (or transformation) of the parameter which is unrestricted and to first find an approximate supported range for the transformed parameter.*"

Indeed! And it is often possible to use the range of the parameter to determine which transformation is likely to lead to a scale on which the log-likelihood is more Gaussian. For a parameter θ, such as a *proportion*, that takes values in the (0,1) range, the 'canonical' transformation is the *logit*, i.e., $\log\left[\frac{\theta}{1-\theta}\right]$, which takes on values all the way from $-\infty$ to $+\infty$.

For a parameter θ, such as a *rate*, that takes values in the $(0, \infty)$ range, the canonical transformation is the log, i.e., $\log[\theta]$. Indeed, for many of the examples to be considered from now on, JH may well use θ for the parameter measured on the full $(-\infty, +\infty)$ scale.

THE LOG RATE PARAMETER [with some editing by JH]

"The rate parameter λ can take only positive values, but its logarithm is unrestricted. To calculate an approximate supported range for λ, it is better, therefore, to first calculate a range for $\log[\lambda]$, and then to convert this back to a range for λ. (...) To find the approximate range for $\log[\lambda]$, we need a new value of S that which gives the best Gaussian approximation to the log likelihood ratio curve when plotted against $\log[\lambda]$. When a rate λ is estimated from D failures over PY person-years, i.e., as

$$\underline{\hat{\lambda} = D/PY},$$

this value of S $\underline{= SE[\hat{\lambda}] = \{\ Var[\hat{\lambda}]\ \}^{1/2}}$ is given by

$$S = \underline{SE = SE[\hat{\lambda}]} = \sqrt{\frac{1}{D}}.$$

(By ignoring the possibility of a Poisson count of zero, and using a Taylor series approximation often referred to as "the *Delta Method*"), at the beginning

[6] JH has changed the possibly confusing Y (for the Years of observation denominator) to PT (for amount of Population-Timd in which the events occurred). He left the 'morbid' term D (for Deaths') as is, even though he prefers to use C for 'number of Cases (instances) of', a *neutral* term that covers both 'bad' and 'good' types of events/characteristics.

of the course, when dealing with Poisson variation, we derived the **square root of the variance of the log of a Poisson r.v.,** Y, finding it to be (approximately)

$$\sqrt{\frac{1}{\mu_Y}},$$

where μ_Y is the expected value, $E[Y]$. And by using the observed value (D in C&H notation, y or c in JH's) as an estimate of its expectation, and plugging it into the square root of the expression for the variance, we have a good example of a standard error – the *estimated* standard deviation of a *statistic* or parameter-estimate.

"***A more convenient way of carrying out this calculation – using a (multiplicative) 'error factor.'*** "

JH encourages this Multiplicative Margin of Error (MME) i.e., using $\hat{\theta} \times \div MME$, rather than $\exp[log[\hat{\theta} \ \pm Z_{\alpha/2} SE[log[\hat{\theta}]]$ approach; think of it as expressing the uncertainty as X-fold (or, en français, 'X-fois'): the upper limit is X *times larger* than the point estimate; and the lower limit is X *times smaller* than the point estimate.

THE LOG ODDS PARAMETER [editing and notation-changes by JH]

Here we use the logit, or the log-odds, i.e., $\theta = \log[\Omega] = \log\left[\frac{\pi}{1-\pi}\right]$, so that $\hat{\theta} = \log\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right]$.

"When an odds, Ω, is estimated from $y+$ 'positives' and $y-$ 'negatives', i.e., as

$$\hat{\Omega} = \frac{y+}{y-}, \quad or \ \hat{\theta} = \widehat{\log[\Omega]} = \log\left[\frac{y+}{y-}\right],$$

the value of $S \ = SE[\hat{\theta}] = \{Var[\hat{\theta}]\}^{1/2}$ is given by

$$S = SE = SE[\hat{\theta}] = \sqrt{\frac{1}{y+} + \frac{1}{y-}}.$$

(By ignoring the possibility of a Binomial count of 0 or n, and using a Taylor series approximation often referred to as "the Delta Method"), at the beginning of the course, when dealing with Binomial variation, we derived the **square root of the variance of the log of the ratio of Binomial r.v.,** $Y+$, **to its complement** $Y-$, finding it to be (approximately)

$$\sqrt{\frac{1}{\mu_{Y+}} \ + \ \frac{1}{\mu_{Y-}}},$$

where μ_Y is the expected value. And by using the observed value (D in C&H notation, $y+$ in JH's) as an estimate of the one expectation, and the observed value of its complement ($N - D$ in C&H notation, $y-$ in JH's) and plugging it into the square root of the expression for the variance, we have a good example of a standard error – the *estimated* standard deviation of a *statistic* or parameter-estimate.

Epidemiologists are (overly) fond of 2×2 tables, with E ('Exposed') and $\bar{E}$ (not) and with D ('Diseased') and $\bar{D}$ (not) as the rows / columns and with $a, \ b, \ c, \ d$ as frequencies

	E	$\bar{E}$	
D	a	b	n_D
$\bar{D}$	c	d	$n_{\bar{D}}$
	n_E	$n_{\bar{E}}$	n

JH prefers this more neutral and more general notation (with X on x-axis, and going from 0 to 1):

	$X = 0$	$X = 1$
$Y = 1$	y_+	y_+
$Y = 0$	y_-	y_-
	$n_{X=0}$	$n_{X=1}$

and looking ahead to regression, with traditional X and Y axes, where the values need not be restricted to 0's and 1's, he would argue for this layout:

	1	y_+	y_+
Y			
	0	y_-	y_-
		0	1
		X	

In any event, at this stage, where all subjects have the same X (say X=0), our only interest is in the left (X=0) column of the table, and in the y_+ to y_- ratio. So, for the logit, when we substitute observed for expected values, we have the biostatistician's expression

$$SE[\ \hat{\text{logit}}[\pi]\ ] = SE[\ \widehat{\log[\Omega]}\ ] = SE[\hat{\theta}] = \sqrt{\frac{1}{y_+} + \frac{1}{y_-}}.$$

or the epidemiologist's expression, with a and c as the focus for now (or a and b if they happened to – or were taught to – exchange the row and columns labels) :

$$SE[\ \text{logit}\ ] = SE[\ \log[a/b]\ ] = \sqrt{1/a + 1/b}.$$

Supplementary Exercise 9.1. Suppose we measured the sex ratio in a sample of $n = 20$, obtaining the ratio ♂ : ♀ = 14:6. Calculate a 90% CI for the true ♂ : ♀ ratio, Ω, by first calculating a CI for $\log[\Omega]$.

9.3 Finding the best quadratic approximation

"*To do this we need some elementary ideas of calculus summarized in Appendix B. In particular, we need to be able to find the gradient (or slope) of the log-likelihood curve together with its curvature, which is defined as the rate of change of the gradient. The mathematical terms for these quantities are the first and second derivatives of the log likelihood function.*"

JH would add that "the *statistical* terms for these quantities are the ***Score*** and (with the sign changed to have a positive value) the ***Information***."

"*The value of $\hat{\theta}$ can be found by a direct search for that value of θ which maximizes the log likelihood, but it is often easier to find the value of θ for which the gradient of the log likelihood is zero; this occurs when $\theta = \hat{\theta}$. The value of S is chosen to make the curvature of the quadratic approximation equal to that of the true log likelihood curve at M, thus ensuring that the true and approximate log likelihoods are very close to each other near $\theta = \hat{\theta}$.*

We therefore choose the value of S to make $-1/(S)^2$ equal to the curvature of the true log likelihood curve at its peak.

THE RATE [i.e., λ] PARAMETER (estimated as '*event count*'$/PT = y/PT$)

$$\text{Lik}[\lambda] = \exp[-\mu]\ \mu^y, \quad \text{where } \mu = \lambda \times PT$$

$$\text{LogLik}[\lambda] = -\mu + y\log[\mu] = -\lambda \times PT + y\log[\lambda] - y\log[PT]$$

$$\text{Score}[\lambda] = \frac{d\ \text{LogLik}[\lambda]}{d\lambda} = \frac{y}{\lambda} - PT$$

$$\text{I}[\lambda] = -E\left[\frac{d^2\ \text{LogLik}[\lambda]}{d\lambda^2}\right] = E\left[\frac{y}{\lambda^2}\right] = \frac{\lambda \times PT}{\lambda^2} = \frac{PT}{\lambda}$$

Substituting $\hat{\lambda} = y/PT$ for λ yields

$$\text{I}[\lambda]_{|\hat{\lambda}} = \frac{PT^2}{y}$$

and using $\{\text{I}[\lambda]\}^{-1} = \frac{y}{PT^2}$ as $SE[\hat{\lambda}]^2$ yields

$$SE[\hat{\lambda}] = \frac{\sqrt{y}}{PT},$$

just as we had established (directly!) at the beginning of the course, treating $\hat{\lambda} = \frac{y}{PT} \sim \frac{Poisson(\mu)}{PT}$.

The *point of this long exercise* is that we can check in simple cases that the two approaches lead to the same $SE[\hat{\lambda}]$, but that there are many instances where there is no closed form, and where the 'curvature' approach is the only way to calculate a SE.

THE RISK (π) PARAMETER

C&H go through the same curvature exercise for this parameter, and arrive at the familiar 'binomial' SE of $\sqrt{\frac{\hat{\pi}\times(1-\hat{\pi})}{n}}$.

9.4 Approximate likelihoods for transformed parameters

First, JH applauds C&H for 'transforming' the *parameter* before transforming the random variable. See the American Statistician article "The PDF of a Function of a Random Variable: Teaching its Structure by Transforming Formalism into Intuition" by JH and D Teltsch – it is under Reprints/Talks on JH's website. Its more about a change of *scale* than the 'change of variable' that is usually taught. After all, the entity called 'Montreal temperatures' is the same (and temperatures in January are equally cold), whether we choose to measure them in $°F$ or $°C$!

Supplementary Exercise 9.2. C&H repeat the above SE calculations, but for the parameter $\beta = \log[\lambda]$ rather than λ. Fill in the steps they don't show. [*And note their wise choice of the letter β – they are thinking ahead to linear predictors in multiple regression. In this simple 1-sample example, think of it as a β_0 !*]

"*In general, derivations such as that above can be simplified considerably by using some further elementary calculus which provides a general rule for the relationship between the values of S (the SE) on the two scales. In the case of the log transformation, this rule states that multiplying the value of S on the scale of λ by the gradient of log(λ) at $\lambda = M$ gives the value of S on the scale of* $\log(\lambda)$. *The rules of calculus tell us that, at $\lambda = M$, the gradient of the graph of* $\log(\lambda)$ *against λ is* $1/M$. *(...)*"

Supplementary Exercise 9.3 Show that the 'simplification' that C&H describe is none other than the Delta Method, with its use of Jacobians to go from one scale to another. The answer you obtained by applying the 'Delta Method' to the transformed r.v. $\log[y/PT]$ is a check on their algebra.

"*This agrees with the expression obtained by the longer method.*"

By the *longer method*, they mean the use of the *curvature/information* (and its reciprocal) at $\theta = \hat{\theta}$. And the '*this*' refers to the *Delta Method*. To JH, the difference is that the *curvature/information* approach emphasizes the *parameter* scale, while the *Delta Method* emphasizes the *random variable* scale. The 'change-the-parameter-scale' is more general, and avoids having to worry about realizations of random variables (e.g., a count of zero) that do not map nicely to another (e.g., the log[*count*] or log[y]) scale.

9.5 Déjà vu: supp. exercises in Ch. 3 (Likelihood)

There was no *simple algebraic expression* (or closed form) for the supported range for the parameters of the models involving σ in Fisher's binned errors data, and in Tibshirani et al's 'accuracy of dart throws' data; μ and σ in Galton's data on the speeds of homing pigeons; the shape and rate/scale parameters of the gamma model for tumbler longevity; the HIV infection rate parameter λ in the circumcision studies; the proportion (π) of persons infected by (the sero-prevalence of) West Nile virus; the parameters in the mutation rate function behind the genetic data from Iceland, etc. etc.. So, ...

Supplementary Exercise 9.4 Revisit each these examples, and decide which, if any, parameter transformation might lead to a 'closer-to-Gaussian' i.e., 'closer-to-quadratic' log-likelihood. Then re-run the graphs on these new scales[7], and see if your intuition was borne out. Can you come up with any 'rule-of-thumb' to decide whether the extra work involved will be worth it?

[7] Once you have set up the LogLik function on one scale, it is usually easy to plot it on another. Or you can change the argument and insert the transformation at the beginning of the old function.