Chapter 2

Statistics

In the present chapter, I will briefly review some statistical distributions that are used often in this book. I will also discuss some statistical techniques that are important in this book, but that may not be very well known. Good introductions to practically all the statistical techniques used here can be found in, for example, Lindgren [38], or Casella and Berger [10]. The group of techniques that are used most often is centered on the likelihood function, but in some instances bootstrapping will be used as well. They will be described briefly.

Many chapters in this book rely strongly on the difference between random variables and model parameters. To accentuate this difference, the general custom will be followed of labeling random variables with upper case letters, and parameters with lower case ones.

1 STATISTICAL DISTRIBUTIONS

The number of distributions used in this book is small, basically the binomial and Poisson distributions, and some variations on them.

1.1 Binomial and multinomial distributions

The binomial distribution is that of the number of fails in a given number of attempts, given the fail probability. To simplify notation, I will use Feller's one [22] for the probability density function of the binomial distribution. The probability that n fails will be observed in N tries if the fail probability is p is

$$b(n;N,p) = \binom{N}{n} p^n (1-p)^{N-n} \quad . \tag{2.1}$$

The expected value of n is Np, and its variance is Np(1 - p).

When p is very close to 0 or 1, the relationship between the expected value of n and its fluctuations becomes very simple. When p is very small, it can be neglected with respect to 1. The standard deviation of n is then roughly equal to the square root of its expected value. Likewise, when p is very close to 1, the standard deviation of N - n is roughly equal to the square root of that number. In other words, when p is either very small or very large, the typical size

of the variations in the number of the rarer events (failures with very low fail probability, passes otherwise,) is roughly equal to the square root of the number of those events, and does not depend on the number of the more common events.

The binomial distribution can be generalized by compounding [12]. In that case, the binomial parameter p is a random variable itself, with a probability distribution h(p). The expected value of p will be indicated by

$$\langle p \rangle = \int h(p) p \, dp \,, \tag{2.2}$$

and its variance by $\sigma^2(p)$.

The expected value of the number of fails in the compounded distribution equals N<p>, and its variance is equal to

$$N\langle p \rangle (1 - \langle p \rangle) + N^2 \sigma^2(p) \,. \tag{2.3}$$

The first term in this variance is the standard binomial one, the second one is the contribution from the finite width of h(p). It has the important consequence that, when N becomes large, the ratio of the standard deviation of the number of fails to its expected value does not go to 0, as in a pure binomial distribution, but, instead, to the finite ratio $\sigma(p)/\langle p \rangle$. Even with large N, therefore, the variability in the number of fails cannot be ignored, and can, in fact, be substantial.

Another extension of the binomial distribution is the multinomial [22] one, in which more than two outcomes are possible, each with their own probability of occurrence. There is no standard notation for this distribution. The one that will be used here was inspired by that for the binomial distribution. If there are k choices, with probabilities p_i for i = 1, ..., K, the probability $P(n_1, ..., n_k)$ of n_i occurrences of choice i is given by the multinomial probability

$$m(\{n_i\};N,\{p_i\}) = \frac{N!}{\prod n_i!} \prod p_i^{\,n_i} \,, \tag{2.4}$$

where n! stands for the factorial of n, and all products are from i = 1 to k. The sets $\{p_i\}$ and $\{n_i\}$ obey the obvious sum rules $\sum_i p_i = 1$, and $\sum_i n_i = N$.

By summing over all n_i except one, say n_j, we find that the probability of n_j occurrences out of N trials equals $b(n_j;N, p_j)$ Consequently, the expected value of any n_i equals Np_i.

1.2 Poisson and compound Poisson distributions

The Poisson distribution is that of the number of occurrences of some event in a given space, given the probability of an occurrence in a unit amount of space, and given that occurrences are independent. Typical examples are the number of events in a given amount of time or the number of defects in a given area. The latter example is the important one in this book.

The probability of an occurrence in a unit amount of space is also called the strength of the Poisson distribution. When the strength is ν, the probability of n occurrences in a unit amount of space equals

$$\frac{\nu^n}{n!}e^{-\nu}. \tag{2.5}$$

The expected value and variance of n are both equal to ν. The probability of no occurrence is $e^{-\nu}$.

A more general version of the Poisson distribution is the compound Poisson distribution, in which the strength ν is itself a random variable with some distribution $h(\nu)$ [12]. The probability of n occurrences is then equal to

$$\int h(\nu)\frac{\nu^n}{n!}e^{-\nu}\,d\nu. \tag{2.6}$$

It is easy to show, by interchanging integration and summation, that the expected value μ of n is now equal to $\langle\nu\rangle = \int h(\nu)\nu d\nu$, and that its variance equals $\mu + \sigma^2(\nu)$, in which $\sigma^2(\nu)$ is the variance of the Poisson strength ν. Compounding, therefore, always increases the variance of the observed yields.

Another effect of compounding is to increase the probability of no occurrences at all, at least when $\langle\nu\rangle$, the expected number of occurrences stays the same. This probability equals

$$p_0 = \int h(\nu)\,e^{-\nu}\,d\nu. \tag{2.7}$$

That compounding always increases p_0 compared to its Poisson value can be proven as follows. It is easy to see that $e^{-\nu} \geq e^{-\lambda} - (\nu - \lambda)e^{-\lambda}$, because $-(\nu - \lambda)e^{-\lambda}$ describes the tangent to $e^{-\nu}$ at $\nu = \lambda$, and because $e^{-\nu}$ curves upwards. The constant λ in the inequality can be any number, but is taken here as the mean of ν. In the compound model, we then find that $p_0 \geq e^{-\langle \nu \rangle}$.

1.3 Negative binomial distribution

An important example of a compound Poisson distribution is the negative binomial one. It emerges when the compounding function is the gamma distribution. In other words, when

$$h(\nu) = \frac{\nu^{\alpha - 1} e^{-\nu\alpha/\mu}}{\Gamma(\alpha)(\mu/\alpha)^{\alpha}}, \tag{2.8}$$

in which α is a positive parameter, called the cluster coefficient, μ is the mean of ν, and $\Gamma(x)$ is the gamma function of x. The negative binomial distribution can be generated in other ways than by compounding a Poisson distribution [12], but compounding is a very convenient one.

The probability of n occurrences in the negative binomial distribution equals

$$\frac{\Gamma(\alpha + n)}{n!\Gamma(\alpha)} \left(\frac{\mu}{\alpha}\right)^{n} \left(1 + \frac{\mu}{\alpha}\right)^{-\alpha - n}. \tag{2.9}$$

The expected value of n is μ, and its variance is $\mu + \mu^2/\alpha$. The probability of no occurrences equals

$$(1 + \mu/\alpha)^{-\alpha}. \tag{2.10}$$

The cluster coefficient functions as a sort of scale that separates the region $\alpha \gg \mu$ in which the negative binomial distribution is very similar to a Poisson distribution, from that in which the two are very different.

The cluster coefficient is related to the distributional parameters of the compounded Poisson distribution through

$$\frac{1}{\alpha} = \frac{\sigma^2(n) - \mu}{\mu^2}, \tag{2.11}$$

which suggests a rough estimate of the inverse of the cluster coefficient from actual data. Using the inverse of α rather than α itself is more meaningful, for the former vanishes in the limit of a pure Poisson distribution. It will be indicated by γ.

Equation (2.11) can also be seen as a generalized definition of a cluster coefficient, one that goes beyond its definition in the gamma function. As such, the estimate obtained from equation (2.11) need not be positive, even though α is in Equation (2.8). There is in fact no reason why the generalized cluster coefficient should always be positive, and we will find in Chapter 4 that it oftentimes is not.

Large values of γ correspond to strong clustering, and small values to little clustering. For example, when we calculate γ for the compound binomial distribution, it equals $-1/N$ in the case of no compounding, but then increases smoothly to positive values. It can become arbitrarily large when $\sigma(p)$, the width of the compounder, becomes large.

2 LIKELIHOOD

In many situations, the data that are collected have some known statistical properties, except that some parameters of the underlying distribution are not known. One of the goals of collecting the data is to estimate those parameters. An example is the passes and fails of an embedded SRAM on the chips. It is assumed to fail with a probability that may depend on the wafer column in which the chip is located. The numbers of passing and failing SRAMs per column have Binomial distributions, and one statistical analysis that can be done is estimating the fail probabilities of those distributions, and determining whether they are column dependent or not.

A standard way of constructing estimators for the parameters of a distribution is the maximum likelihood method. It relies on the so called likelihood function. This approach is described in some detail in the statistics books mentioned previously [10, 38], and in more detail in the book by Edwards [20].

The likelihood function is numerically proportional to the probability that the observed data would have been obtained, given a specific set of distributional parameters. By considering the likelihood function as a function of the parameters, with the observed data as fixed values, the probability is trans-

formed into a function of the parameters. The likelihood function is proportional to it, for factors that do not depend on the parameters turn out to be irrelevant.

In the example given above, the probability that the given numbers of passes and fails in the various columns would have been observed is equal to the product of a number of binomial probabilities, one for each column, and each one with its own fail probability. With the actual observations fixed, this product is a function of the column fail probabilities. It will vary when the fail probabilities are varied.

2.1 Maximum likelihood

The maximum likelihood method is based on the assumption that the best estimate of the physical fail probabilities, the ones that govern the actual passes and the fails on the physical wafers, is that set of probabilities that maximizes the likelihood function. It obviously depends on the observed data, because different sets of data will put the maximum of the likelihood function in different places.

The likelihood function is generally indicated by L. If we continue the example, L is function of the column fail probabilities p_i. To make the dependence on the observed data explicit, they are sometimes added to L as a condition:

$$L = L(p_1, \ldots, p_k | \text{data}) \,. \tag{2.12}$$

Given the data, the first step in the analysis is estimating the fail probabilities. As mentioned above, this is done by maximizing L, and entails two steps. First, the extrema of L have to be found, which can be done by solving

$$\frac{\partial L}{\partial p_i} = 0 \tag{2.13}$$

for each i (column in the example). Second, the maximum has to be found among the extrema. A maximum corresponds to an extremum where the matrix with elements

$$\frac{\partial \, \partial L}{\partial p_i \partial p_j} \tag{2.14}$$

is negative definite. In most cases, Equations (2.13) have only one solution, and that solution can trivially be shown to correspond to a maximum. In some cases, however, multiple solutions may have to be considered, and the negative definiteness of the matrix of second derivatives of L has to be established using numerical methods.

Strictly speaking, the found maximum should also be compared to values of L on the boundary of the range of the parameters of the distribution, for maxima on those boundaries usually do not obey Equation (2.13). In most cases encountered in this book, L trivially vanishes on this boundary, and is positive in the interior region of the range, so the question of maxima on the boundary does not occur.

There are in fact situations in which Equations (2.13) are so complex that they cannot be solved even with moderate effort. If all else fails, the maximum of L can always be found by reliable, but numerically more demanding maximization routines [44].

The estimates of the parameters are random variables, for they depend solely on the observations, and not on the parameters to the underlying distributions. These estimates, therefore, have a distribution, but that distribution is usually not known. Fortunately, for large sample - that is, large wafers in the example - the distribution of the estimates is approximately normal with a covariance matrix equal to minus the inverse of the matrix of second derivatives. The latter matrix is therefore not only important for establishing maximality of extrema, but also for gauging the accuracy of the estimates.

2.2 Likelihood ratio

The likelihood function is used not only for estimating parameters, but also for deciding whether one particular statistical model is better suited to explain the data than some other potential model. The manner in which that will be done in this book can be demonstrated with the example that we have been using in this section.

In the running example, there are two reasonable models. The first one, called the heterogeneous model, is the one that we have been using: one fail probability per column. The second one is called the homogeneous model, and is a simplification of the first: one fail probability for all columns. The heterogeneous model is always more accurate, for it has more adjustable parameters. The homogeneous one is more parsimonious, and may be preferred for that reason.

Even when the homogeneous model is correct, the numbers of fails on any given column will not always be equal to the mean, but will fluctuate around it. Small deviations of the numbers of fails around their respective means will not necessarily invalidate this model, therefore; only large deviations can do

that. The question is, "how large should the deviations be before we should discard the homogeneous model and assume the validity of the heterogeneous one ?"

This question can be answered to some extent with the likelihood ratio

$$\Lambda = \frac{L(\hat{p}\,|\,\text{data})}{L(\hat{p}_1, \ldots, \hat{p}_k\,|\,\text{data})}, \tag{2.15}$$

in which a carrot ($\hat{}$) over a variable indicates the maximum likelihood estimates of that variable, and $\hat{p}$ is the maximum likelihood estimate of the single fail probability in the homogeneous model.

Λ will never exceed 1, for both numerator and denominator are maximized, and the space of the p values is a subset of the space of the p_i values. Therefore, if $L(\hat{p})$ were larger than $L(\hat{p}_1, \ldots, \hat{p}_k)$, the latter could be increased by replacing the estimates of p_i by the estimate of p, contrary to the assumption that it is maximal.

Λ is a convenient measure of the extent to which the observed deviations match the expected ones; in other words, it is a good indicator of column similarity. If the homogeneous model reflects the true state of affairs, it will be close to 1, but not equal to it, because of statistical fluctuations. If this model is not the correct one, Λ will be much smaller than 1.

How much Λ should differ from its maximum value before the homogeneous model can be rejected depends of course on the size of the expected statistical fluctuations, which depend on the numbers of columns and chips per column through N_{DF}, the number of degrees of freedom. This number equals, in this case, $\sum(m_i - 1)$, in which the sum is over all the columns, and m_i is the number of chips in column i.

Under the null hypothesis that all columns have the same fail probability, $-2\ln\Lambda$ has approximately the chi-squared distribution with N_{DF} degrees of freedom [10]. Consequently, under the null hypothesis, the expected value of $-2\ln\Lambda$ equals N_{DF}, and its variance $2N_{DF}$.

If the null hypothesis is correct, the actual value of $-2\ln\Lambda$ is expected to be within a few standard deviations of its mean. A more convenient measure of column similarity, therefore, is the ratio

$$\rho = \frac{-2\ln\Lambda - N_{DF}}{\sqrt{2N_{DF}}}. \tag{2.16}$$

Any significant deviation of Λ from its mean leads to a large value of ρ, and indicates that one or more columns differ significantly from the others. Moreover, when the number of degrees of freedom is large, as it typically is, the chi-square distribution can be replaced by a normal one with the same mean and variance.

3 BOOTSTRAPPING

When estimating the values of distributional parameters or other distribution related quantities, we often would like to the know the accuracy of those estimates, in addition to the estimates themselves. When the statistical distribution of the estimator is known, the accuracy of the estimate can be obtained from the variance of the estimator. Oftentimes, however, the distribution is not known, or, if known, is valid only in the limit of very large samples. In such cases, other means have to be employed to get a sense of the accuracy of the estimators.

The variance of an estimator could also be estimated, and trivially so, if many samples were available. For then we could estimate whatever quantity we are interested in in each sample, and compare the results. Unfortunately, there is only one sample. It is possible, however, to create artificial samples, with many of the same statistical properties as real samples, and use these artificial samples as substitutes for the latter. This technique is called bootstrapping [41].

In bootstrapping, a large number of secondary samples are generated from the original one, called the primary sample. The secondary samples have the same size as the primary one, and are formed by randomly selecting the units of the sample (embedded SRAMs in our running example) from the original sample. The selection is done sequentially, and with replacement (so the same unit can be selected multiple times.)

The bootstrap assumption is that the statistical properties of primary samples are approximately the same as those of the secondary samples, based on a single primary one. For example, a single fail probability for the embedded SRAMs, valid for all columns, can be calculated for each secondary sample, and the distribution of these fail probabilities is assumed to approximate that of the maximum likelihood estimate of the fail probability in the primary sample.