31. COLLECTING AND PLOTTING DATA

31.1 Review/Background

So far we have discussed the physics of various reliability issues of modern transistor such as NBTI degradation, HCI, TDDB, etc. However, in practice, we need experimental data before making any statistical prediction about the reliability or lifetime of the chip. The goal of this chapter is to explore about the various aspects of data collection and the careful processing of the collected data for reliability prediction. In particular, we will discuss how a small set of data can be transformed into a population distribution function.

31.2 Origin of data and Statistical Inference

By the term experimental data points, we mean, for example, the failure times of a set of transistors in case of TDDB experiment. Clearly, we cannot perform this experiment on billions of transistors due to the cost and time constrains; still the prediction we make based on a limited amount of data points should hold for billions of transistors or the global population. It is then very important to construct the distribution function from a small set of experimental data.

There are some issues regarding data collection and also dealing with datasets. For example, in a certain situation, like acceleration-projection method, Figure. 31.1b, one would normally explore the tail of the distribution function. However, this tail is usually rather long and thereby entails a significant penalty in terms of cost and time if one were to explore the function by a substantial amount of data points. Moreover, if the initial data set is small, then any error present in the data can have serious consequences in the final prediction. For example look at the projection voltage acceleration in Figure. 31.1b for red and blue lines. Even though the two lines show a good agreement
for the higher acceleration voltages, using one versus the other will result in completely wrong prediction about the time to breakdown of the transistor, note the divergence of two lines for small voltages. Often the experimental dataset may be incomplete or non-uniform in quality, but still we need to make best decision possible from it. Hence data handling or the interpretation of the experimental data for making statistical inferences are highly non trivial issues and require profound statistical theories, which we will describe in the following sections.

31.3 Non-Parametric Information

The non-parametric information of a set of data includes quantities such as the mean and the variance of the dataset, without making any assumption regarding the underlying distribution function. These estimates are important to extract as much information as possible from the raw data before fitting those data to various parametric distribution functions. The various moments of the experimental data defined below are the non-parametric estimates generally people are interested in.
\[ \langle t \rangle = \left( \sum_{j=1,N} t_j \right) / N \]  

\[ s^2 = \left( \sum_{j=1,N} (t_j - \langle t \rangle)^2 \right) / (N - 1) \]  

\[ \delta_{T_K} = \sqrt{\frac{\sum_{j=1}^N (t_i - \langle t \rangle)^k}{N - k + 1}} \]

where \( N \) is the number of samples. It is interesting to note that a moment of degree \( n \), has \( (N-n) \) in its denominator, because estimation of each parameter uses one degree of freedom and the moments therefore must be normalized to the remaining set of independent samples (degree of freedom).

31.4 **Distribution of the Sample Statistics/Moment**

Assume here we have a reference population distribution represented in Figure. 31.2. This distribution is generated as following. Suppose we have a sample of \( n \) devices, \( n \) is 20 in Figure. 31.2, now we run the experiment for these 20 devices which gives us 20 data points. Then we find the average of these 20 data points, let us call this average \( \mu_1 \). Now we repeat this procedure \( m \) times, \( m \) is 10,000 in the example of Figure. 31.2. As a result, we will end up having 10,000 \( \mu \)'s, \( (\mu_1, \mu_2, ..., \mu_{10,000}) \). Figure. 31.2 shows the distribution of these average points.

Now given a random set of 20 samples depicted by its average value as the red arrow in Figure. 31.2, question is whether this dataset belongs to the above distribution or not. Generally speaking, further the arrow is from the population average of the distribution, less likely it is for the dataset to be a part of this distribution. Precisely, however, in order to answer this question one must calculate Z-parameters of equations 31.4 and 31.5 and also P-value which is the area below the distribution from the place where the arrow is to infinity. Then, based on the fact that larger values of P mean that
the red arrow is closer to the center of the distribution, one can decide whether this dataset is a part of that distribution or not.

Figure. 31.2 Distribution of sample statics for the sample of size 20. The above distribution was created by repeating the sampling for 10,000 times

\[
Z = \frac{X - \mu}{\sigma \sqrt{N}} \quad N > 30
\]

\[
Z = \frac{X - \mu}{s \sqrt{N}} \quad N < 30
\]

31.5 Dealing with Outliers

Sample moments are extremely sensitive to outliers and even a few such data points can distort the estimates. These outliers may arise from …Many times due to some unwanted reasons, (it may be noise or human error) some sample data points may
be erroneous. If possible, it is prudent to repeat the experiment corresponding to that data point; else we need to be extremely careful for the rejection of the measured data.

Main problem with averaging is that it is strongly influenced by outliers. Outliers are those points which are distant from the rest of the data. Generally the probability of outliers being a part of the distribution is small and the data point will usually lies at the very extreme tail of the underlying distribution. As a result of accounting for outliers, the average sometimes will be shift towards the outlier and give us rather inaccurate representation of the dataset. Hence, before any manipulation of data, one needs to exclude the outliers from parametric estimates. This can be done by using quartile approach, which is shown in Figure. 31.3. This approach is based on not the momentums of the dataset rather on the median therefore is not affected by the presence of outlier. In this approach, one first needs to divide the dataset based on the median into two sets and then divide each set, based on their median, into quartiles. Then they need to find the inter-quartile range, IQR, which is the distance between first and forth quintiles, median on the first and second sets. If the sample is outside of $1.5 \times IQR$ from second and third quintiles therefore that sample cannot be trusted and consequently cannot be used for further analysis. This does not mean that one can ignore that sample in their plots rather they should plot all data including outliers but don’t use them in their analysis. Remember that every data point contains information about our measurement and even though they should not be used for our proof of theory, they might shine light into some details of our experiment in the future.

![Figure. 31.3 Schematic showing how quintile approach works.](image-url)
Figure 31.4 schematically demonstrates why a box plot is a reliable method to indicate outliers without finding moments of data, remember that the moments are affected by the outliers. As can be seen in this figure, if a data point lies outside the box, the probability of that point being a part of the underlying distribution is so small and that is the reason why we can put that data point aside for our following analysis, i.e. finding average, standard deviation etc.

![Figure 31.4 Distribution of sample statics showing the logic behind 1.5 IQR in Box plot. As can be seen the probability of a sample being related to the distribution is small when it lies above ±1.5IQR [1].](image)

### 31.6 Removing Outliers based on Chauvenet’s Criteria

There is a well-defined prescription for the rejection of the erroneous data known as Chauvenet’s criteria [2]. We will illustrate these criteria with a simple example of breakdown times given in **Error! Reference source not found.**. According to this prescription, we need to perform the following calculations, called Chauvenet’s criteria, in order to decide on data rejection.
Table 31.1 A sample of breakdown time points.

<table>
<thead>
<tr>
<th>TBD</th>
<th>X(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>560</td>
<td></td>
</tr>
<tr>
<td>540</td>
<td></td>
</tr>
<tr>
<td>570</td>
<td></td>
</tr>
<tr>
<td>550</td>
<td></td>
</tr>
<tr>
<td>560</td>
<td></td>
</tr>
<tr>
<td>660 ??</td>
<td></td>
</tr>
<tr>
<td>580</td>
<td></td>
</tr>
<tr>
<td>570</td>
<td></td>
</tr>
<tr>
<td>550</td>
<td></td>
</tr>
</tbody>
</table>

1) Calculate the average value (μ) of the TBD (time to breakdown). In this example it is 571.1 s.
2) Calculate the sample variance, s = 35.51 s.
3) Find the normalized variable, \( z = \frac{x - \mu}{s} \). \( x \) is the sample TBD. In this case: \( z = \frac{660 - 571.1}{35.51} = 2.5 \)
4) Calculate the tails of the normal distribution with
5) \( p = P(-\infty < z < -2.5) + P(2.5 < z < \infty) = 0.01242 \).
6) Calculate \( N \times p = 9 \times 0.01242 = 0.1 \); \( N \) is the sample size.
7) If \( Np < 0.5 \), throw away the data points

As shown in Table 31.1, 9 measured data points for TBD are noted down. We want to know if the 6th data point (660) is correct or wrong, i.e., 660 is the suspicious datum. Based on how much the suspicious datum differs from the mean, use the normal distribution function to determine the probability that a given data point will be at the value of the suspect data point (step 4). Multiply this probability by the number of data points taken. If the result is less than 0.5, the suspicious data point may be discarded, i.e., a reading may be rejected if the probability of obtaining the particular deviation from the mean is less than \( \frac{1}{2N} \); where \( N \) is the no of data points.
31.7 Stem and Leaf Display: Pre-Histogram

In order to plot the data in the histogram, one must first divide the data points into two groups, stem and leaf. The range of stem as shown in Figure 31.5 can be derived from equations 31.6 and 31.7. After finding $h_n$ then we need to take it to the nearest power of 10 and use it as the bases for the stem. Be sure that after doing all these steps, your data precision should remain the same.

$$L = [10 \times \log_{10} n]$$  \hspace{1cm} 31.6

$$h_n = \left(\frac{\text{Range}}{L}\right) \text{ round to closest power of 10} \hspace{1cm} 31.7$$

Order data

```
44 46 47 49 63 64 66 68 68 72 72 75 76 81 84 88 106
```

<table>
<thead>
<tr>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>4679</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>634688</td>
</tr>
<tr>
<td>2256</td>
</tr>
<tr>
<td>148</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

Figure 31.5 An example of stem and leaf for an arbitrary set of data shown above. $L \sim 13$ and $h_n = 4.7$ and therefore is 10. That is the reason why stems are a products of 10.

31.8 Drawing Lines Resistant to Outliers

If your data points have outliers, then plotting a line using regular techniques will result in conclusion, which is corrupted by the outliers and can be misleading. In
order to draw a line resistant to outliers one needs to follow the method described in the following example.

Suppose we have a dataset, 9 data points in Figure. 31.6. In order to avoid the effect of outliers, we first divide data into three different groups following procedures in Table 31.2. Then, we will find the median of each group, for both data points in x-axis and also y-axis i.e. \((x_L, y_L), (x_M, y_M), (x_R, y_R)\). Now that we have the median of each group, we can draw a line between those three medians instead of original data points. This line, as a result, will not have any effect from outliers.

<table>
<thead>
<tr>
<th>Number of data points (n)</th>
<th>Dividing procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=3k</td>
<td>k, k, k</td>
</tr>
<tr>
<td>n=3k +1</td>
<td>k, k+1, k</td>
</tr>
<tr>
<td>n=3k+2</td>
<td>k+1, k, k+1</td>
</tr>
</tbody>
</table>

Table 31.2 Calculation with Hazen formula for censored and uncensored data.

Figure 31.6 Schematic representing an effective method to plot a line resistant to outliers.
31.9 Plotting of Data

Suppose we have 5 transistors and we have measured the 5 breakdown times for those transistors. Now if we want to plot the Cumulative distribution function (F) of the breakdown times, we can use either \( F_i = \frac{i}{n} \), or \( F_i = \frac{i}{n+1} \). We can immediately see there is a problem with first form. When \( i = n \), \( F_i = 1 \) and the Weibull distribution function \( W = \ln(-\ln(1-F_i)) \) blows up. Hence the second functional form is preferred for plotting the CDF.

![Figure 31.7 The discrete breakdown times for the sample of 5 transistors and the corresponding CDF.](image)

There is however one more problem associated with the arrival time distribution around each breakdown time point (shown as yellow region in Figure. 31.7). To explain this, let us assume we have a population of 25 transistors, and we need to represent the cumulative breakdown time distribution with a sample of 5 transistors only. This situation is illustrated in Figure. 31.8. If we use the simple formula \( F_i = \frac{i}{n} \) or \( F_i = \frac{i-1}{n} \), then some population data points will exactly match with the sample data points but the maximum error between the sample and the population data points will be very high. The
error can be minimized by using a new formula for the CDF based on median of the sample data points can be used. This formula is shown in equation 31.8 and it is also known as Hazen formula.

\[ F_{\text{Median},i} = \frac{i - \alpha}{n - 2\alpha + 1} \quad (\alpha = 0.3) \]  

In order to derive Hazen formula [2], let \( p \) be the probability that breakdown time is less than or equal to \( i^{th} \) sample time. If there is \( n \) sample data points, then the probability that population data value is less than \( i^{th} \) sample time is

\[ g = (\frac{n}{i})p^i(1-p)^{n-i} \]  

Hence
\[ g = \frac{dG}{dp} = \binom{n}{i} p^{i-1} (1-p)^{n-i} \]  

31.10

Since

\[ \int_0^{F_{\text{median},i}} gdp = \frac{1}{2} \]  

31.11

\[ F_{\text{median},i} = \frac{i - \alpha}{n - 2\alpha + 1} \quad (\alpha = 0.3) \]  

31.12

We see in Figure. 31.8 that the maximum error between the CDF of total population and the CDF of sample of 5 transistors is a minimum for the case of Hazen formula.

31.10 Handling Censored Data

So far we have discussed outliers in dataset and how to detect and exclude them from our analysis. Aside from outliers, there is another type of data called censored data which also needs to be identified and treated carefully during data analyzing step. Censored data, in this context, refers to a data point where due to human error, instrument malfunction, etc one experiment was not complete or was missing while the others were complete and accurate. In this case, throwing away the incomplete experiment result or data point will result in inaccurate projection of data. In order to understand what censored data is and how we should deal with them, let us look at the following example.

Suppose we have a sample of 5 transistors and we want to plot the CDF of the breakdown times of those transistors. If due to some reasons one sample data point is missing, how can we still make the best decision out of the remaining sample data points. In order to do so we need to be very careful in post processing of the remaining data (also
known as censored data). We will illustrate this with the following simple example shown in Figure. 31.9.

![Figure. 31.9 CDF of censored and uncensored data.](image)

<table>
<thead>
<tr>
<th>Uncensored Data (Hazen formula with $\alpha=0$)</th>
<th>Censored Data (third data point is missing) (Hazen formula with $\alpha=0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1 = 1/6$</td>
<td>$F_1 = 1/5$</td>
</tr>
<tr>
<td>$F_2 = 2/6$</td>
<td>$F_2 = 2/5$</td>
</tr>
<tr>
<td>$F_3 = 3/6$</td>
<td>$F_3 = 2/5$</td>
</tr>
<tr>
<td>$F_4 = 4/6$</td>
<td>$F_4 = 3/5$</td>
</tr>
<tr>
<td>$F_5 = 5/6$</td>
<td>$F_5 = 4/5$</td>
</tr>
</tbody>
</table>

Table 31.3 Calculation with Hazen formula for censored and uncensored data.

If all the 5 sample data points were present, then the CDF is plotted with the blue line in Figure. 31.9, with the Hazen formula ($\alpha = 0$) as shown in Table 31.1. Now,
let us assume the third sample is missing. In this case one can still use the Hazen formula with the 4 data points again presented in Table 31.3 and plotted in Figure. 31.9 with the red dashed line. We immediately see that there is a problem in this approach. Since the third sample is missing, that means the CDF for the first and second data points should not be affected. But with the Hazen formula, we see the first and second data points also get affected. In other words, if we use Hazen formula for censored data, then past data points are affected by the future missing data points, which is not desirable. To eradicate this problem, Kaplan-Meier Formula is used for plotting of the censored data and the formula is given by

\[
F_i = 1 - \left[ \frac{n - \alpha + 1}{n - 2\alpha + 1} \right] \prod_{i=1}^{f} \left[ \frac{n_{si} + 1 - \alpha}{n_{si} + 2 - \alpha} \right]
\]  

31.13

For \( \alpha = 0; \)

\[
F_i = 1 - \prod_{i=1}^{f} \left[ \frac{n_{si} + 1}{n_{si} + 2} \right]
\]  

31.14

This Formula is applicable for both the censored as well as uncensored data. We illustrate this with in the followings.

31.11 Kaplan–Meier Formula for Uncensored Data

<table>
<thead>
<tr>
<th>( n_{si} )</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>before ( t_i )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>after ( t_i )</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 31.4 value of \( n \) for set of data.

According to the Kaplan–Meier formula the CDF values are:

\[
F_1 = 1 - \frac{4 + 1}{4 + 2} = \frac{1}{6}
\]  

31.15
\[ F_2 = 1 - \frac{4 + 1 \cdot 3 + 1}{4 + 2 \cdot 3 + 2} = \frac{2}{6} \]

\[ F_3 = 1 - \frac{4 + 1 \cdot 3 + 1 \cdot 2 + 1}{4 + 2 \cdot 3 + 2 \cdot 2 + 2} = \frac{3}{6} \]

\[ F_4 = 1 - \frac{4 + 1 \cdot 3 + 1 \cdot 2 + 1 \cdot 1 + 1}{4 + 2 \cdot 3 + 2 \cdot 2 + 2 \cdot 1 + 2} = \frac{4}{6} \]

\[ F_5 = 1 - \frac{4 + 1 \cdot 3 + 1 \cdot 2 + 1 \cdot 1 + 1 \cdot 0 + 1}{4 + 2 \cdot 3 + 2 \cdot 2 + 2 \cdot 1 + 2 \cdot 0 + 2} = \frac{5}{6} \]

These values are different from those obtained by the Hazen formula for censored data. We note that in this case the first and second sample data points are not affected by that of the missing data point.

We can summarize the results for the CDF for the case of censored data with the following Table 31.5

<table>
<thead>
<tr>
<th>Method</th>
<th>T=1</th>
<th>T=2</th>
<th>T=3</th>
<th>T=4</th>
<th>T=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hazen Formula</td>
<td>1/6</td>
<td>2/6</td>
<td>3/6</td>
<td>4/6</td>
<td>5/6</td>
</tr>
<tr>
<td>Hazen formula with one sample missing</td>
<td>1/5</td>
<td>2/5</td>
<td>Missing</td>
<td>3/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Kaplan–Meier Formula</td>
<td>1/6</td>
<td>2/6</td>
<td>Missing</td>
<td>5/9</td>
<td>7/9</td>
</tr>
</tbody>
</table>

Table 31.5 Summary of CDFs derived using different formulas for censored data.

### 31.12 Conclusions

Generation of the experimental data is costly both in terms of time and equipment and hence careful treatment in analyzing the data is very important. Simple non-parametric estimates like mean, standard deviation, median, etc. are very useful
indicators for the corresponding quantities of the whole population space and they also help in selecting the appropriate distribution function. Non-parametric plotting of the distribution function gives important insight about the population but the plotting approaches for the censored and uncensored data are very different. Finally, a good parametric fit of the sample data allows one to estimate the tails of the distribution correctly. We will discuss the details of the fitting algorithms as well as the ‘goodness-of-fit’ of the distributions in next chapters.

References