(3) Graphical Data Analysis Techniques-Frequency Distributions

Graphical techniques are a very useful way of analyzing the way in which random measurement errors are distributed. The simplest way of doing this is to draw a ***histogram***, in which bands of equal width across the range of measurement values are defined and the number of measurements within each band is counted. The bands are often given the name ***data bins***. There are two alternative rules for calculating the best number of data bins to use:

The Sturges rule calculates the number of bands as follows:

$$\text{Number of bands} = 1 + 3.3 \log_{10}(n),$$

where **n** is the number of measurement values.

The Rice rule calculates the number of bands as $\mathbf{2}\boldsymbol{n}^{\mathbf{1/3}}$.

Obviously the result produced has to be rounded to the nearest integer in both cases.

When **n** is relatively small, the two rules suggest the same number of bins. However, for larger values of n, the Rice rule calculates a larger number of bins than the Sturges rule.

This is summarized in the table below:

Number of Measurements	Number of Bins Calculated by Sturges Rule	Number of Bins by Sturges (Rounded)	Number of Bins Calculated by Rice Rule	Number of Bins by Rice (Rounded)
10	4.3	4	4.3	4
15	4.9	5	4.9	5
20	5.3	5	5.4	5
25	5.6	6	5.8	6
30	5.9	6	6.2	7
50	6.6	7	7.4	7
100	7.6	8	9.3	9
200	8.6	9	11.7	12

Example

Draw a histogram for the 23 measurements in set C of the length measurement data given in the last lecture.

Solution

For 23 measurements, the recommended number of bands calculated according to the Sturges rule is

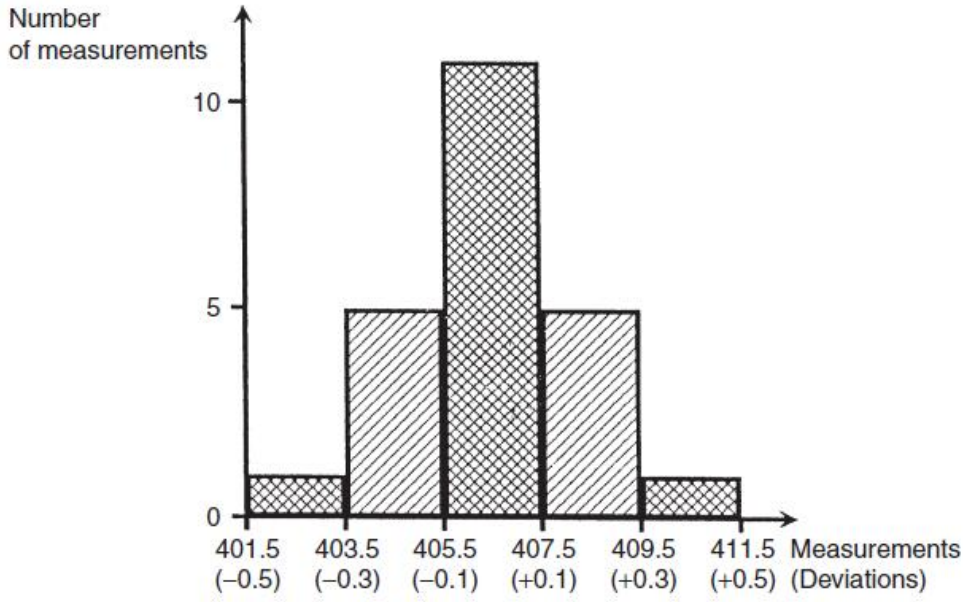
$$1+3.3 \log_{10}(23) = 5.49$$

This rounds to 5, since the number of bands must be an integer number.

To cover the span of measurements in data set C with 5 bands, the data bands need to be 2-mm wide. The boundaries of these bands must be carefully chosen so that no measurements fall on the boundary between different bands and cause ambiguity about which band to put them in. Since the measurements are integer numbers, this can be easily accomplished by defining the range of the first band as 401.5 to 403.5
and so on.
A histogram can now be drawn as in Figure below by counting the number of measurements in each band.

In the first band from 401.5 to 403.5, there is just 1 measurement and so the height of the histogram in this band is 1 unit.

In the next band from 403.5 to 405.5 there are 5 measurements and so the height of the histogram in this band is 5 units.
The rest of the histogram is completed in a similar fashion.



Histogram of measurements and deviations.

As it is the actual value of measurement error that is usually of most concern, it is often more useful to draw a histogram of the deviations of the measurements from the mean value rather than to draw a histogram of the measurements themselves. The starting point for this is to calculate the deviation of each measurement away from the calculated mean value. Then a histogram of deviations can be drawn by defining deviation bands of equal width and counting the number of deviation values in each band. This histogram has exactly the same shape as the histogram of the raw measurements except that the scaling of the horizontal axis has to be redefined in terms of the deviation values (these units are shown in brackets on figure above).

(4) Standard Gaussian Tables (z-Distribution)

A standard Gaussian table (sometimes called the z-distribution), such as that shown in table below, tabulates the area under the Gaussian curve F(z) for various values of z, where F(z) is given by

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} e^{(-z^2/2)} dz$$

For negative values of z, we can make use of the following relationship because the frequency distribution curve is normalized:

$$F(-z) = 1 - F(z)$$

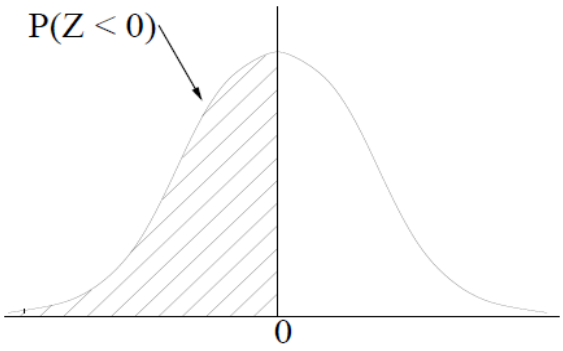
F(-z) is the area under the curve to the left of (-z), that is, it represents the proportion of data values ≤ -Z.

If Z is normally distributed with mean μ and standard deviation σ, we write

$$Z \sim N(\mu, \sigma^2)$$

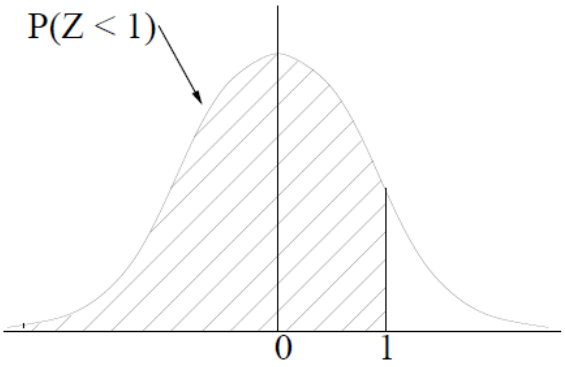
μ and σ are the **parameters** of the distribution.

Suppose $Z \sim N(0, 1)$, what is $P(Z < 0)$?



Symmetry $\Rightarrow P(Z < 0) = 0.5$

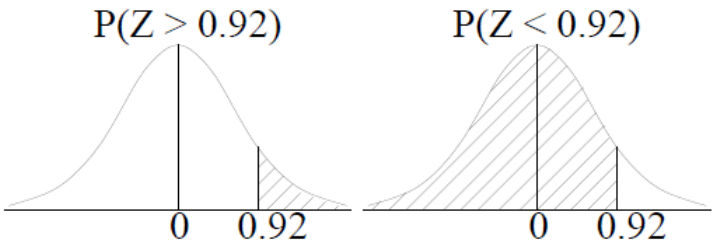
What about $P(Z < 1.0)$?



From this table we can identify that P(Z < 1.0) = 0.8413

Example

If $Z \sim \text{N}(0, 1)$ what is $P(Z > 0.92)$?

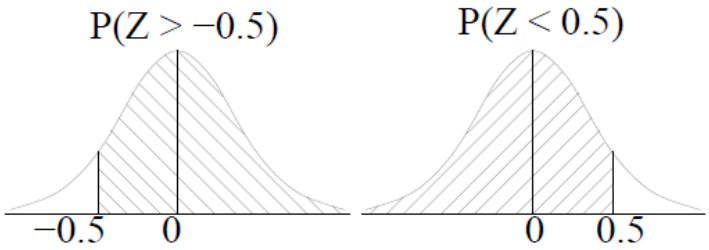


We know that $P(Z > 0.92) = 1 - P(Z < 0.92)$ and we can calculate $P(Z < 0.92)$ from the tables.

Thus, $P(Z > 0.92) = 1 - 0.8212 = 0.1788$

Example

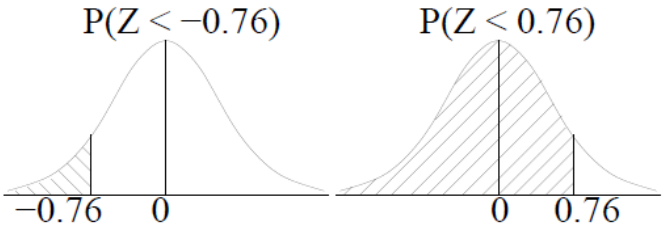
If $Z \sim \text{N}(0, 1)$ what is $P(Z > -0.5)$?



The Normal distribution is symmetric so we know that $P(Z > -0.5) = P(Z < 0.5) = 0.6915$

Example

If $Z \sim \text{N}(0, 1)$ what is $P(Z < -0.76)$?



By symmetry

$$\begin{aligned} P(Z < -0.76) = P(Z > 0.76) &= 1 - P(Z < 0.76) \\ &= 1 - 0.7764 \\ &= 0.2236 \end{aligned}$$

Error function table (area under a Gaussian curve or *z*-distribution)

	F(*z*)									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5000	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8906	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9926	0.9928	0.9930	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9986	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Example

How many measurements in a data set subject to random errors lie outside deviation boundaries of $+\sigma$ and $-\sigma$, that is, how many measurements have a deviation greater than $|\sigma|$?

Solution

The required number is represented by the sum of the two shaded areas in figure below. This can be expressed mathematically as $P(E < -\sigma \text{ or } E > +\sigma) = P(E < -\sigma)$ $P(E > +\sigma)$.

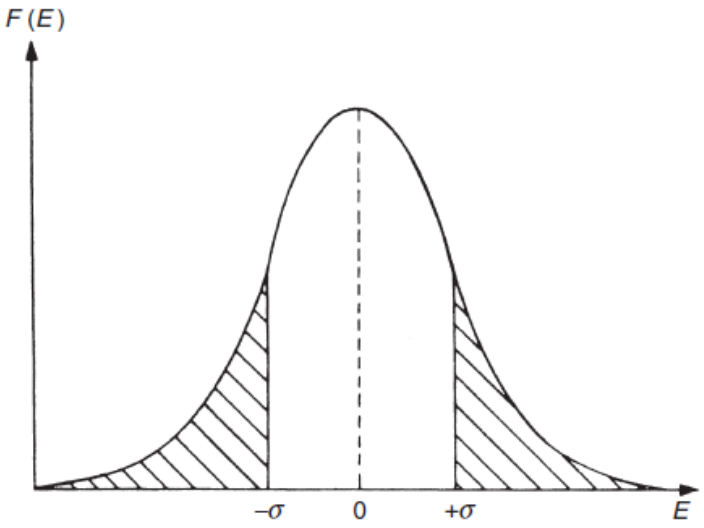
For $E = -\sigma$, $z = -1.0$.

Using Table above $P(E < -\sigma) = F(-1) = 1-F(1) = 1-0.8413 = 0.1587$.

Similarly, for $E = +\sigma$, $z = 1.0$, Table above gives

$P(E > \sigma) = 1 - P(E < \sigma) = 1 - F(1) = 1 - 0.8413 = 0.1587$. (This last step is valid because the frequency distribution curve is normalized such that the total area under it is unity.)

Thus, $P[E < -\sigma] + P[E > \sigma] = 0.1587 + 0.1587 = 0.3174 \sim 32\%$, that is, 32% of the measurements lie outside the $\mp\sigma$ boundaries, then 68% of the measurements lie inside.



$\mp\sigma$ boundaries

(5) ***Standard Error of the Mean***

The standard deviation of the mean values of a series of finite sets of measurements relative to the true mean (the mean of the infinite population that the finite set of measurements is drawn from) is defined as the standard error of the mean, α. This is calculated as

$$\alpha = \sigma/\sqrt{n}$$

Clearly, α tends toward zero as the number of measurements (n) in the data set expands toward infinity.

We also know that a range of $\pm$ one standard deviation (i.e., $\pm\alpha$) encompasses 68% of the deviations of sample means either side of the true value. Thus we can say that the measurement value obtained by calculating the mean of a set of n measurements, x1, x2…xn, can be

$$x = x_{\text{mean}} \pm \alpha$$

with 68% certainty that the magnitude of the error does not exceed $|\alpha|$. For the data set C of length measurements used earlier, $n = 23$, $\sigma = 1.88$, and $\alpha = 0.39$. The length can therefore be expressed as 406.5 ± 0.4 (68% confidence limit).

The problem of expressing the error with 68% certainty is that there is a 32% chance that the error is greater than α. Such a high probability of the error being greater than α may not be acceptable in many situations. If this is the case, we can use the fact that a range of $\pm$ two standard deviations, that is, $\pm 2\alpha$ encompasses 95.4% of the deviations of sample means either side of the true value. Thus, we can express the measurement value as

$$x = x_{\text{mean}} \pm 2\alpha$$

with 95.4% certainty that the magnitude of the error does not exceed $|2\alpha|$. This means that there is only a 4.6% chance that the error exceeds 2α. Referring again to set C of length measurements, $2\sigma = 3.76$, $2\alpha = 0.78$ and the length can be expressed as 406.5 ± 0.8 (95.4% confidence limits).

If we wish to express the maximum error with even greater probability that the value is correct, we could use $\pm 3\alpha$ limits (99.7% confidence). In this case, for the length measurements again, $3\sigma = 5.64$, $3\alpha = 1.17$ and the length should be expressed as 406.5 ± 1.2 (99.7% confidence limits). There is now only a 0.3% chance (3 in 1000) that the error exceeds this value of 1.2.

Example:

In a practical exercise to determine the freezing point of a metal alloy, the following measurements of the freezing point temperature were obtained:

519.5 521.7 518.9 520.3 521.4 520.1 519.8 520.2 518.6 521.5

Express the mean value and the error boundaries expressed to (a) 68% confidence limits, (b) 95.4% confidence limits, and (c) 99.7% confidence limits.

Solution:

First calculate the mean value of the measurements.

$$\text{Mean} = \sum(519.5 \;\; 521.7 \;\; 518.9 \;\; 520.3 \;\; 521.4 \;\; 520.1 \;\; 519.8 \;\; 520.2 \;\; 518.6 \;\; 521.5)$$
$$= 520.2$$

Next, calculate the deviations of the measurements from the mean, and hence the standard deviation.

Measurement	519.5	521.7	518.9	520.3	521.4	520.0	519.8	520.3	518.6	521.5
Deviation from mean	−0.7	+1.5	−1.3	+0.1	+1.2	−0.2	−0.4	+0.1	−1.6	+1.3
$(\text{deviations})^2$	0.49	2.25	1.69	0.01	1.44	0.04	0.16	0.01	2.56	1.69

$\sum(\text{deviations})^2 = 10.34$; n = number of measurements = 10.

Then, from (4.6) and (4.7), $\sigma = \sqrt{\frac{\sum(\text{deviations})^2}{n-1}} = 10.34/9 = 1.149$.

The standard error of the mean is given by

$$\alpha = \sigma \Big/ \sqrt{n} = 1.149 \Big/ \sqrt{9} = 0.383$$

The mean of the measurements expressed to 68% confidence limits is given by

$$x = x_{\text{mean}} \pm \alpha = 520.2 \pm 0.4$$

The mean of the measurements expressed to 95.4% confidence limits is given by

$$x = x_{\text{mean}} \pm 2\alpha = 520.2 \pm 0.8$$

The mean of the measurements expressed to 99.7% confidence limits is given by

$$x = x_{\text{mean}} \pm 3\alpha = 520.2 \pm 1.2$$

(6) *Estimation of Random Error in a Single Measurement*

In many situations, where measurements are subject to random errors, it is not practical to take repeated measurements and find the average value. Also, the averaging process becomes invalid if the measured quantity does not remain at a constant value, as is usually the case when process variables are being measured. Thus, if only one measurement can be made, some means of estimating the likely magnitude of error in it is required. The normal approach to this is to calculate the error within 95% confidence limits, that is, to calculate the value of the deviation D such that 95% of the area under the probability curve lies within limits of ±D. These limits correspond to a deviation of ±1.96σ. Thus, the maximum likely error in a single measurement can be expressed as

$$\text{Error} = \pm 1.96(\sigma + \alpha)$$