1 Computing the Standard Deviation of Sample Means

Quality control charts are based on sample means not on individual values within a sample. A sample is a group of items, which are considered all together for our analysis. Items within a sample lose their individual characteristics in the analysis. Rather a summary statistic, e.g. sample mean, is used to represent the information in the sample. See the examples of samples below:

1. A section of BA3352 students in the current semester is a sample of students. Then the sample size is the number of students in the section. Different sections constitute different samples. The number of sections offered in the current semester would be the number of samples.

2. Voters surveyed by a given polling agency on a single day is a sample. The sample size is the number of voters surveyed on that particular day. Polls made on different days constitute different samples. The number of the polls is the number of samples.

3. Customers buying a particular brand of perfume over a specified month can be considered as a sample. The sample size is the number of customers buying the perfume over the specified month. Another sample can be generated by considering customers buying another brand of perfume. If we consider four brands of perfumes, we end up with four samples.

The number of samples and the sample size can potentially be confusing. Sample size is the number of items within a group. Number of samples is the number of groups.

Example 1: After a midterm exam for a course that is given to five sections of a course, the average exam grade $\bar{x}_j$ in section j is computed and reported below.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5
Average grade	68	72	74	82	71

Suppose that there are 50 students in each section and use $x_{i,j}$ to denote the ith student's grade in Sec j. Then the average grades are computed by

$$\bar{x}_j = \frac{\sum_{i=1}^{50} x_{i,j}}{50} \qquad \text{for } j \in \{1,2,3,4,5\}.$$

Since all 50 grades within a section are reduced to a single summary statistic (the sample mean), all the students within a section are represented merely by the section's summary statistic (the sample mean); Individual student grades are immaterial for an analysis that checks if a certain sectin is performing better than the others. Clearly, the sample size is 50 and the number of samples is 5. □

There are two ways to compute the standard deviation $\sigma_{\bar{x}}$ of sample means. The first way requires the knowledge of the standard deviation σ_x of the individual values within a sample, the second way does not require σ_x.

1.1 Computing $\sigma_{\bar{x}}$ with known σ_x

In order to understand what we have and what we want, first recall that

$$Var(X) = \sigma_x^2 \ \ and \ \ Var(\bar{X}) = \sigma_{\bar{x}}^2.$$

Note that $Var(X)$ is known and we want to compute $Var(\bar{X})$.

In order to perform this computation, we need to recall the following proposition from statistics:

Proposition 1. *i) If X is a random variable and c is a constant, then $Var(c \cdot X) = c^2 \cdot Var(X)$.*
ii) If X_1 and X_2 are two independent random variables, then $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$.

Proof: i) First convince yourself that the mean of cX would be $c\bar{x}$ where $\bar{x}$ is the mean of X. We start with $Var(c \cdot X)$ and use the definition of variance

$$Var(cX) = \frac{1}{n}\sum_{i=1}^{n}(cx_i - c\bar{x})^2 = c^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = c^2 Var(X).$$

ii) Again by using the definition

$$\begin{aligned} Var(X_1 + X_2) &= \frac{1}{n}\sum_{i=1}^{n}(x_{1,i} + x_{2,i} - \bar{x}_1 - \bar{x}_2)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\{(x_{1,i} - \bar{x}_1)^2 + (x_{2,i} - \bar{x}_2)^2 + 2(x_{1,i} - \bar{x}_1)(x_{2,i} - \bar{x}_2)\} \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_{2,i} - \bar{x}_2)^2 + 2\frac{1}{n}\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(x_{2,i} - \bar{x}_2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_{2,i} - \bar{x}_2)^2 + 0 \\ &= Var(X_1) + Var(X_2) \end{aligned}$$

The fourth equality is due to the fact that X_1 and X_2 are independent so the sum of the cross products is zero. This sum would be the covariance of X_1 and X_2, if X_1 and X_2 were not independent. □

Now Proposition 1 can be used to relate the variance of the sample mean to the variance of the observation within the samples. We start with the definition ofthe sample mean, proceed as follows

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) \\ &\overset{Prop.1.i}{=} \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n}X_i\right) \\ &\overset{Prop.1.ii}{=} \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}Var(X_i) \\ &= \frac{n}{n^2}Var(X) \\ &= \frac{1}{n}Var(X) \end{aligned} \tag{1}$$

where we use the fact that each individual observation has the same variance as the other individuals: $Var(X_1) = Var(X_2) = Var(X_i) = Var(X)$ where X stands for a generic observation and represents one of $X_1, X_2, \ldots X_n$. This fact is assumed when constructing samples; otherwise, we would be grouping “apples” with “oranges”.

Given (1) which relates varainces, relating the standard deviations is easy. Just take the square root of the both sides in (1) to arrive at

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{n}}\sigma_x. \tag{2}$$

Example 2: Refer to Example 1 and suppose that the indivudual scores has a standard deviation of 20, compute the standard deviation of the sample means.
Solution: We are given $\sigma = 20$, sample size is already known as $n = 50$. Then by using (2),

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{n}}\sigma_x = \frac{1}{\sqrt{50}}20. \ \square$$

1.2 Computing $\sigma_{\bar{x}}$ with unknown σ_x

This method is rather direct; Without σ_x, the only information available is the population of the sample means $\{\bar{x}_1, \bar{x}_2, \ldots \bar{x}_m\}$ where the number of samples is denoted by m. We could use this population to estimate the standard deviation of the sample means. First let us compute the variance:

$$Var(\bar{X}) = \frac{1}{m}\sum_{j=1}^{m}(\bar{x}_j - \bar{\bar{x}})^2$$

where $\bar{\bar{x}}$ is the grand mean which can be computed by

$$\bar{\bar{x}} = \frac{1}{m}\sum_{j=1}^{m}\bar{x}_j.$$

Finally the standard deviation of the sample mean is

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\bar{x}_j - \bar{\bar{x}})^2}. \tag{3}$$

Example 3: Refer to Example 1 and compute the standard deviation of the sample means from the population $\{68, 72, 74, 82, 71\}$.
Solution: First we compute the grand mean

$$\bar{\bar{x}} = \frac{1}{m}\sum_{j=1}^{m}\bar{x}_j = 73.4.$$

Then the standard deviation of the sample means by (3) is

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{5}\{(68-73.4)^2 + (72-73.4)^2 + (74-73.4)^2 + (82-73.4)^2 + (71-73.4)^2\}}.$$

1.3 Remark

When σ_x is unknown, you must use (3) to compute $\sigma_{\bar{x}}$. In this case, you do not have any choice. When σ_x is known, you have to choose between equations (2) and (3). Unless otherwise is specified, use (2) to find $\sigma_{\bar{x}}$. Rationale here is that the computation in (2) is exact whereas (3) gives you only an estimate. The general principle applies: use the information available to you as much as possible and refrain from estimation unless absolutely necessary.

2 Exercise Questions

1. Every year about 500 people apply for UTD's full time MBA program. Over the years it has been observed that GMAT score of each of these people are distributed normally with mean 600 and variance 300.
a) If UTD decides to accept all applicants whose GMAT score is above 620, on average how many people will be accepted per year?
b) If UTD decides to accept 50 students with highest GMAT scores every year, what should be the cut off GMAT score (lowest score among the 50 accepted students).

2. Draw an Ishikawa diagram listing the possible causes of your midterm grade. Include Environment, Materials, Method, Personnel, etc.

3. Read "Continuous Improvement on the Free-Throw Line" pp.412-414 of the textbook. In couple sentences explain a process from your own life, which you have improved by studying reasons for failure or substandard performance. Example processes are parallel parking, speaking in public, washing dishes, finding the closest parking spot to your office/class, etc.

4. The DFW passenger data below pertains to the first eight months of 2001. Suppose that every month has 30 days. Number of passengers flying out of DFW airport per day and the number of passengers who are searched per day are:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
	$\bar{y}_{Jan}$	$\bar{y}_{Feb}$	$\bar{y}_{Mar}$	$\bar{y}_{Apr}$	$\bar{y}_{May}$	$\bar{y}_{Jun}$	$\bar{y}_{Jul}$	$\bar{y}_{Aug}$
Average # of passengers/day	15000	14000	12600	13300	14700	14100	16800	17500
	$\bar{z}_{Jan}$	$\bar{z}_{Feb}$	$\bar{z}_{Mar}$	$\bar{z}_{Apr}$	$\bar{z}_{May}$	$\bar{z}_{Jun}$	$\bar{z}_{Jul}$	$\bar{z}_{Aug}$
Average # of searched passengers/day	47	53	61	41	42	44	51	43

The average number of passengers per day is computed as follows. Let $y_{i,j}$ be the number of the passengers on the ith day of month j. The average number of passengers per day for month j is $\bar{y}_j$ defined as

$$\bar{y}_j = \frac{\sum_{i=1}^{30} y_{i,j}}{30} \qquad \text{for } j \in \{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug\}.$$

The average number of passengers searched per day is computed similarly. Let $z_{i,j}$ be the number of the passengers searched on the ith day of month j. The average number of passengers searched per day for month j is $\bar{z}_j$ defined as

$$\bar{z}_j = \frac{\sum_{i=1}^{30} z_{i,j}}{30} \qquad \text{for } j \in \{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug\}.$$

a) What is the sample size n for computing averages in the table?
b) Suppose that the standard deviation of the number of passengers ($y_{i,j}$) flying out of DFW every day is 3000, what is the standard deviation of the average number of passengers ($\bar{y}_j$) flying out of DFW per day?
c) Assuming a Normal distribution for the number of passengers, how many sigmas (σ) will give you a Type I error of 20% for an $\bar{x}$-chart on the average number of passengers flying out of DFW per day?

5. Refer to question 4.
a) Find out 3-sigma UCL and LCL for an $\bar{x}$ chart on the average number of passengers flying out of DFW

per day.
b) Is the process in control during the first eight months? Explain.

6. Refer to question 4.
a) Compute the variance of the average number of passengers searched ($\bar{z}_j$) per day during the first eight months. In other words, find the variance of the population $\{\bar{z}_{Jan}, \bar{z}_{Feb}, \bar{z}_{Mar}, \bar{z}_{Apr}, \bar{z}_{May}, \bar{z}_{Jun}, \bar{z}_{Jul}, \bar{z}_{Aug}\}$ by using the data in the table. Let us call this variance $\sigma^2_{\bar{z}}$.
b) Compute the ratio of $\sigma^2_{\bar{z}}$ to the grand mean of the averages of the passengers searched per day during the first eight months. Looking at this ratio and considering the fact that the number of searches per day is an integer number, what distribution would be appropriate to study the number of searches?
c) What are UCL and LCL for a 2.5-sigma c-control chart for the number of passengers searched per day?

7. Refer to question 4.
a) Obtain the proportion $\bar{r}_j$ of passengers searched per day for each month. In other words, construct the population $\{\bar{r}_{Jan}, \bar{r}_{Feb}, \bar{r}_{Mar}, \bar{r}_{Apr}, \bar{r}_{May}, \bar{r}_{Jun}, \bar{r}_{Jul}, \bar{r}_{Aug}\}$ by using the data in the table.
b) Compute the grand mean and the variance $\sigma^2_{\bar{r}}$ of the population in a).
c) What are UCL and LCL for a 2.5-sigma p-control chart for the proportion of passengers searched?

8. Refer to questions 4,6 and 7. Below are average number of passengers and average the number of passengers searched in September and October 2001.

	Sep	Oct
Average number of passengers/day	9100	6200
Average number of searched passengers/day	57	63

Using c- and p-control charts obtained in questions 6 and 7 and the recent numbers above determine if
a) The number of passengers searched per day is in control?
b) The proportion of passengers searched per day is in control?
c) How can you reconcile your answers if you say “yes” to either a) or b) above, and “no” to the other?