Risto B. Rushford
GSCM 451-Spr 2019
HW03

Short-Answer Problems

These concepts can appear on the optional short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1. Mean deviation
    a. What is a mean deviation?

The distance of the i^{th} value (Y_i) from the mean (m): $\text{deviation}_i = Y_i - m$

    b. What is the distinction between a positive and negative mean deviation?

A positive mean deviation is the distance between the mean and a value *greater* than the mean, while a negative mean deviation is the distance between the mean and a value *less* than the mean.

    c. In what context is a mean deviation considered an error?

Mean deviation is considered an error not because it is a miscalculation or wrong answer to a question, but because it is a measure of the difference between the real value and what it was forecasted to be.

2. How is the mean the 'balance point' of a set of numbers?

The mean is the 'balance point' of a set of numbers because the sum of all variances, both above and below the mean, will equal zero. This means that the mean is perfectly at the center or the data values.

3. To compute the standard deviation, we average the squared deviation scores? Why take the average instead of leaving just the sum?

The sum of mean deviations will always equal zero. This is a characteristic of the mean that must always be true, as such, the sum of mean deviations does not provide a good metric for evaluating the dataset. To remove the negative values from the deviations, the values are squared. To avoid confounding the variability with the sample size, the average of the squared deviations is taken instead of taking a sum.

4. Degrees of freedom.
    a. What is the "degrees of freedom" in the computation of the standard deviation?

The course defines degrees of freedom (*df*) of a statistic as: the number of data values *not* constrained by other statistical estimates previously calculated from the *same* data. *df* for the standard deviation is calculated as: $df = n - 1$

    b. Why use the degrees of freedom instead of the sample size?

Degrees of freedom are used in place of sample size because of the issue of data dependency, wherein calculations made from the same set of data are dependent upon each other and are not free to vary. *df* is considered to be the *effective* sample size to resolve the issue of data dependency.

5. How is the standard deviation related to the normal curve?

Standard deviation is a measure of how data values fluctuate. Assuming a normal distribution of those fluctuations, they will correspond to the "Normal Curve". This allows us to determine the range of expected variability about a predicted value (a *probability statement*).

6. What is the motivation for calling the normal curve: the 'normal curve'?

A normal curve is a bell-shaped curve showing the probability distribution of a continuous random variable. It is a representation of the "Normal Distribution", in which exactly half of the population values are above the mean and exactly half are above the mean. The mean, median, and mode have all converged to the same value, and about 68% of all data are within one standard deviation from the mean; 95% of data is within two standard deviations of the mean, and 99.7% of the data is within three standard deviations of the mean.

Worked Problems

1. Forecasting error for a stable process

Find the data for this problem are at: http://web.pdx.edu/~gerbing/451/Data/HW3_1.xlsx

This Excel file consists of four worksheets, named Y1, Y2, Y3, and Y4. Each worksheet consists of only the variable name, such as Y1, and 30 data values (based on last week's homework). Each set of data is from a stable process.

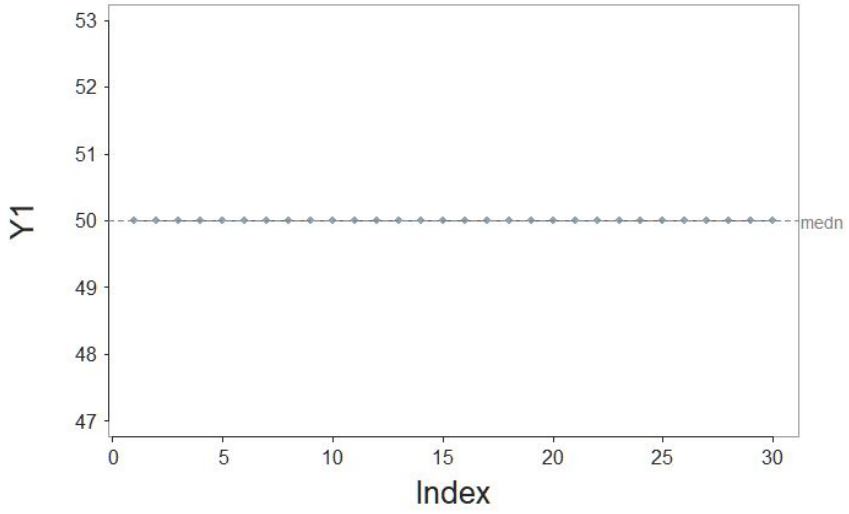
You can read the data directly from this Excel file on the web. To read the data will require four different Read statements, one for each of the following sub-problems. By default, Read reads the first (and often the only) worksheet in the file. To indicate other worksheets, add the `sheet` parameter, and set equal to either the name of the worksheet in quotes, such as `sheet="Y3"` or the ordinal number of the worksheet, such as `sheet=3`.

Remember, any Excel data file referenced on the web is just a data file. When on the job, you will need to understand your data. For example, you always look at it first before attempting anything. If you want to clarify what this data file looks like, take a look at it. Double-clicking on an Excel file usually downloads it, but your computer can also be configured to open directly into Excel, so what happens depends on your computer configuration. Either way, though, you can view the file.

To get the summary stats for each variable, either run `Histogram()` or `ss.brief()`, an abbreviation for brief summary statistics, which is also output by `Histogram()`.

a. Analysis of Y1.

i) Plot the run chart. So that all the plots are comparable, use the same y-axis for all four plots by adding the parameter `y.scale = c(47,53,5)`. That means start the axis from 47, go to 53, and show 5 intervals.



ii) Assuming a stable process, what is your forecast for the 31st value?

Assuming a *stable* process, the next value will be very close to 50.

iii) What is the assessment of the error of fit?

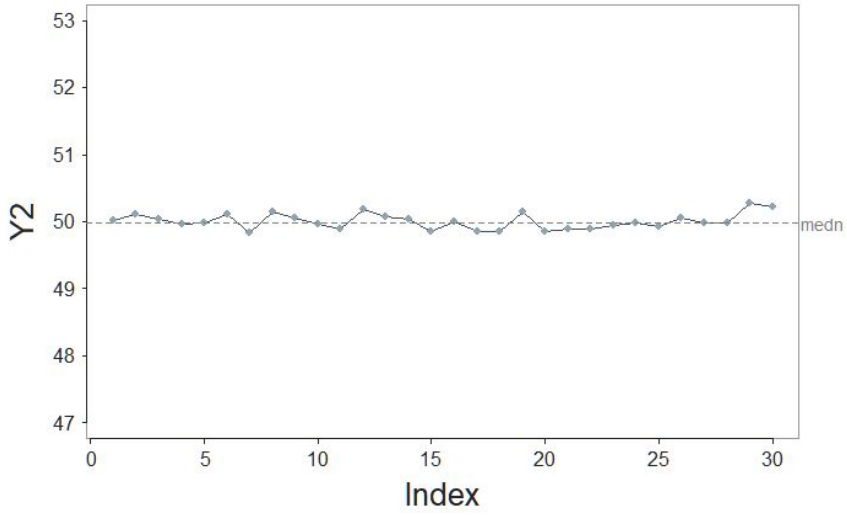
The standard deviation is 0.00098, meaning the error of fit for the values on this graph and the expected values is very small.

iv) Assuming a normal distribution of errors, what is the expected range for 95% of these errors?

Assuming a normal distribution, the expectation is that 95% of these errors fall within two standard deviations of the mean.

b. Analysis of Y2.

i) Plot the run chart. So that all the plots are comparable, use the same y-axis for all four plots by adding the parameter `y.scale = c(47,53,5)`. That means start the axis from 47, go to 53, and show 5 intervals.



ii) Assuming a stable process, what is your forecast for the 31st value?

My forecast for the 31st value of this dataset is 50.1

iii) What is the assessment of the error of fit?

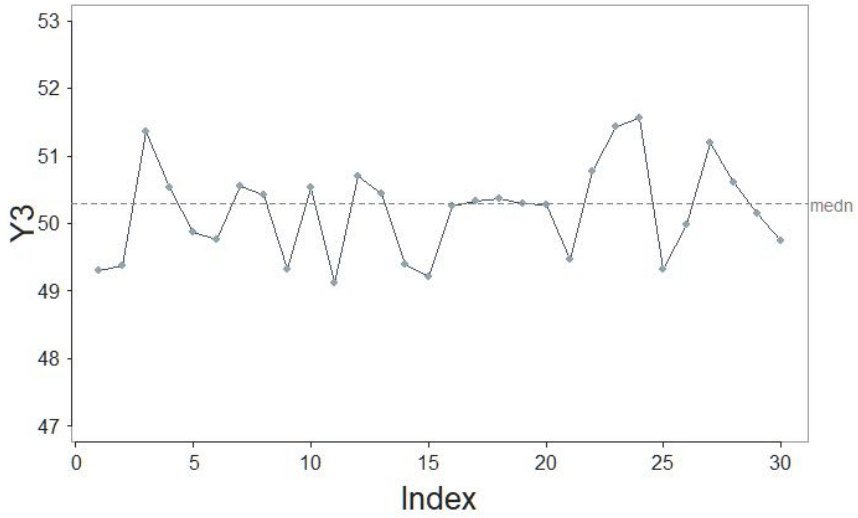
The standard deviation is 0.11672, the error of fit for these values is significantly more than for the previous data set.

iv) Assuming a normal distribution of errors, what is the expected range for 95% of these errors?

We would still expect that 95% of these errors would fall within two standard deviations of the mean.

c. Analysis of Y3.

i) Plot the run chart. So that all the plots are comparable, use the same y-axis for all four plots by adding the parameter `y.scale = c(47,53,5)`. That means start the axis from 47, go to 53, and show 5 intervals.



ii) Assuming a stable process, what is your forecast for the 31st value?

My forecast for the 31st value of this dataset is 51.

iii) What is the assessment of the error of fit?

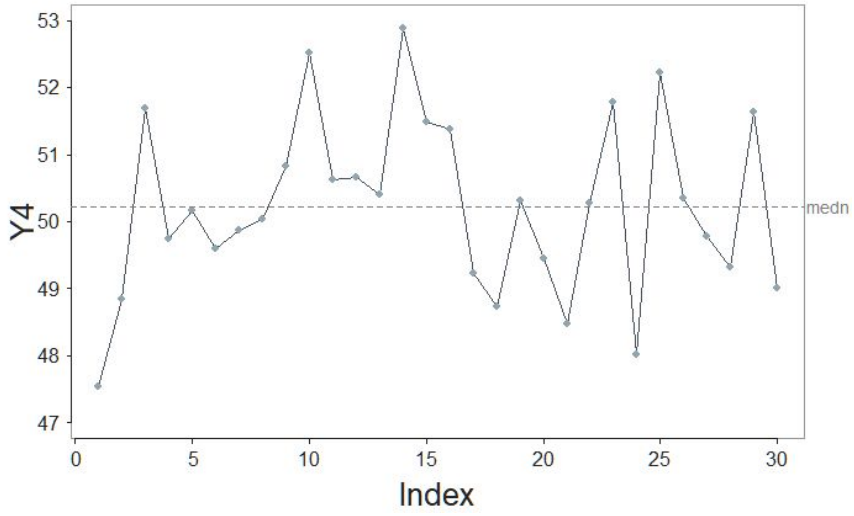
The standard deviation is 0.69139, which is significantly higher than for either of the previous two datasets, but still less than one. The data has a fairly poor error of fit.

iv) Assuming a normal distribution of errors, what is the expected range for 95% of these errors?

We would still expect that 95% of these errors would fall within two standard deviations of the mean.

Analysis of Y4.

Plot the run chart. So that all the plots are comparable, use the same y-axis for all four plots by adding the parameter `y.scale = c(47,53,5)`. That means start the axis from 47, go to 53, and show 5 intervals.



i) Assuming a stable process, what is your forecast for the 31st value?

My forecast for the 31st value of this dataset is 52.

ii) What is the assessment of the error of fit?

The standard deviation is 1.32304, which is almost double the previous dataset. Any forecast from this data would have a substantially high error of fit.

iii) Assuming a normal distribution of errors, what is the expected range for 95% of these errors?

We would still expect that 95% of these errors would fall within two standard deviations of the mean. The difference between this run chart and the first one is that the data points have a greater range, but the distribution will still be expected to approximate the normal curve.

d. Summarize the relationship between assessment of error and the overall shape of the corresponding plot.

The closer the plot resembles a flat line (according to the appropriate scale of examination), the higher the certainty that we can forecast values with a lower level of error. A more jagged line, especially with steep slopes between data points, can expect higher errors between the forecasted values and the corresponding measurements.

2. Understanding the Standard Deviation

Open the Y3 worksheet from the previous problem in Excel with the 30 values of Y3. Look at Slide #21 in Sec 2.1c.

a. Construct a similar Excel worksheet with Y3. That is, add columns for the mean, mean deviation, and squared mean deviation. Then sum the deviations and squared deviations. From that last sum, compute the variance and then standard deviation of Y3. Compare to the result of the Excel STDEV function.

Y3	mean	dev	dev^2
Sum		0.000	13.862
df			29
Mean			0.47801
St Dev			0.69139
STDEV()			0.69139

b. Now do the same procedure in R. You have already done a transformation of data value in R for the last homework, so you should be able to do this now as you just enter the formulas with each variable in the `d` data frame identified by a leading `d$`. But to help out I will get you started. Here we presume you read the data in the data frame named R. What is new here is the R `mean()` function. You could also do a histogram and grab the sample mean from that analysis as well. Do not worry about the column sums in the data table. Just show them separately using the `sum()` function.

```
d <- Read("http://web.pdx.edu/~gerbing/451/Data/HW3_1.xlsx", sheet=3)

d$Y3.mean <- mean(d$Y3)
d$Y3.dev <- d$Y3 - d$Y3.mean
d$Y3.sqdev <- (d$Y3.dev)^2
mean(d$Y3)
sum(d$Y3.dev)
SSY <- sum(d$Y3.sqdev)
SSY
df <- nrow(d) - 1
df
variance <- SSY / df
variance
stdeviation <- sqrt(variance)
stdeviation

[1] 50.18926
[1] -0.00000000000009237056
[1] 13.86242
[1] 29
[1] 0.4780145
```

```
[1] 0.6913859
```

c. Compare your manual calculation with the standard deviation from the `Histogram()` or `ss.brief()` functions.

```
ss.brief(Y3)
```

n	miss	mean	sd	min	mdn	max
30	0	50.18926	0.69139	49.12280	50.28970	51.56510

3. Normal curve and standard scores

If the mean is 50 and the standard deviation is 10:

a. What is the z-score of 60?

The z-score of 60 on such a curve is 1

b. If the distribution is normal, what is the probability that a random sample of a single element from that population is between 50 and 60? (easiest to use my `prob.norm()` function from Section 3.2c)

As the graph below shows, the probability of Y between 50 and 60 is 0.3413 or 34.13%