STATISTICS WORKSHEET

As you progress through this worksheet, you will learn how to:

Design a sampling experiment

Determine statistics of a sample
- Mean, standard deviation, confidence intervals of the mean

Check whether the data fit distributions, e.g.,
- Uniform
- Normal (or Gaussian)
- Poisson

You will achieve these aims by conducting an experiment and processing the data. The experiment you will conduct is to measure your heart rate. You will use the following as guidelines for measuring the data:

You will only measure your heat rate while resting or doing light physical activity (walking). Do not include data when you are doing "heavy" exercise.

You will measure your rate <u>over at least 30 seconds.</u> If you use an electronic pulse meter, wait long enough for the reading to stabilize. RECORD THE RATE PER MINUTE AS AN EVEN NUMBER – 71 is not allowed and you must decide whether to record it as 70 or as 72 – or take the measurement again!

You will record the date and time when you took the measurement, the rate in beats per minute, and any other pertinent information.

Leave at least 30 minutes between measurements. This is to ensure independence of measurements.

This worksheet includes a sample set of data, calculations and decisions. There are also sections marked **YOUR TURN.** When you reach these marks, complete the relevant section in the separate student worksheet. For simple calculations, include everything in the student worksheet. For more lengthy calculations use MathCAD and append the data and programs to the worksheet.

YOUR TURN: Turn to the student worksheet and complete the section - ***Fill in your name and other information:***

DESIGNING A SAMPLING EXPERIMENT

You want to find your day-long average heart rate. How many samples (measurements) should you take?

First, you need to take a few samples. But how many? This number will often be between 6 and 15, but these are definitely not absolute limits. In general:

> If it is expensive and/or time consuming to obtain a sample, you will probably keep your initial sample size smaller.
>
> If the population has a large deviation compared to the accuracy you want, you will need more samples.
>
> The number of samples goes UP as you ask for higher accuracy.

For this worked example we will choose to take 10 measurements for our initial sample.

OBSERVED DATA

$n_1 = 10$ 62, 58, 52, 66, 62, 74, 64, 68, 60, 72

YOUR TURN: Turn to the student worksheet and complete the section - ***Decide on the size of your small sample.*** Choose between 6 and 10.

CALCULATIONS

Calculate the mean and standard deviation of this small sample set. Remember to carry forward many more significant figures than we typically use for "normal" engineering.

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1} = \frac{62+58+52+66+62+74+64+68+60+72}{10} = \frac{638}{10} = 63.80000$$

$$S_1 = \sqrt{\frac{\sum_{i=1}^{n_1} \left(x_i - \bar{x}_1\right)^2}{\left(n_1 - 1\right)}} = 6.56252$$

YOUR TURN: Turn to the student worksheet and complete the section - ***Calculate the mean and standard deviation of your small sample.***

STANDARD ERROR OF THE MEAN

We need to choose a confidence interval (or level of significance). For this exercise we will say that we want to know the average heart rate within 2.5 beats per minute. In the example we will use 2%, 5% and 10% levels of significance and show the effect on our experiment.

We also need to know the number of degrees of freedom (dof). The dof is the number of samples less one for each estimated parameter (statistic). We are using the standard deviation to determine the standard error. Before we could calculate the standard deviation we had to calculate the mean. Therefore, we have (n-1) = (10-1) = 9 degrees of freedom.

We turn to the Student-t tables (or use MathCAD) and look up the value as follows for 9 degrees of freedom:

For a 2% level of significance (98% confidence) t = 2.82144
For a 5% level of significance (95% confidence) t = 2.26216
For a 10% level of significance (90% confidence) t = 1.83311

YOUR TURN: Turn to the student worksheet and complete the section - ***Degrees of freedom, level of significance and Student-t for your sample data.***

Now let's find the standard error of the mean for our sampled data. The standard error is calculated as:

$$\text{Standard error} = d = \frac{t.S_1}{\sqrt{n}}$$

For our data and the three different levels of significance:

Level of significance	Confidence Interval	Standard error (bpm)	Mean heart rate (bpm)
2%	98%	5.85519	63.8 ± 5.85519
5%	95%	4.69454	63.8 ± 4.69454
10%	90%	3.80417	63.8 ± 3. 80417

The achieved accuracy (5.9 b.p.m., 4.7 b.p.m., or 3.8 b.p.m.) is larger than the required accuracy (2.5 b.p.m.). Therefore we need more data samples.

If, at this stage, the achieved accuracy was better than the required accuracy, you would be done with data acquisition!

YOUR TURN: Turn to the student worksheet and complete the section - ***Calculate the standard error of your small sample.***

REQUIRED SAMPLE SIZE

We use the required accuracy and the calculated mean and standard deviation of our small sample to estimate the number of samples we require to get the necessary accuracy. Note the following:

> We assume the sample mean and standard deviation do not change as we sample and test. While the population mean and standard deviation are fixed (we just don't know what they are), the sample statistics will change. Hence, after sampling is complete we will have to re-check the achieved accuracy.
>
> Since we do not know how many samples to take, we cannot look up the appropriate Student-t statistic directly. This means our estimation procedure is iterative.

Here are the equations:

Accuracy, $d = \frac{t.S}{\sqrt{n}}$. Rearranging for the number of samples we get $n \approx \left(\frac{t.S}{d}\right)^2$

As we calculate the number of samples, in this worked example we only use the 5% level of significance information. At the end of this section, we give the final results for all three levels of significance, so that we can compare the effect the level of significance has on required sample size. Remember, the required confidence interval is $d = 2.5$ bpm.

First iteration:

$$n \approx \left(\frac{t.S}{d}\right)^2 = \left(\frac{2.26216 \times 6.56252}{2.5}\right)^2 = 35.262$$

Question – do you round this number up or down?

Second iteration:

The new Student-t for dof = 35 is 2.03011

$$n \approx \left(\frac{t.S}{d}\right)^2 = \left(\frac{2.03011 \times 6.56252}{2.5}\right)^2 = 28.40$$

Third iteration:

The new Student-t for dof = 28 is 2.04841

$$n \approx \left(\frac{t.S}{d}\right)^2 = \left(\frac{2.04841 \times 6.56252}{2.5}\right)^2 = 28.91$$

We have converged, and will need to take a total of about 29 samples.

Here is the comparison of the number of samples required to achieve an accuracy (standard error) of 2.5 beats per minute at different levels of significance:

Level of significance	Confidence Interval	Required number of samples required to get this accuracy
2%	98%	41
5%	95%	29
10%	90%	21

This example clearly shows that if you want more confidence in your result, you need to take more measurements.

YOUR TURN: Turn to the student worksheet and complete the section – ***Calculate the required sample size.***

ADDITIONAL SAMPLING

You already have a small sample set. You now need to take additional data. For this worked example, we choose to continue with the 5% level of significance. As we found above, we need 29 measurements in total. Since we already have 10 samples, we need an additional (29 – 10) = 19 samples. However, since the measurement is easy, quick and inexpensive to make, we elect to increase this to 37 samples in total (27 additional). Here are the advantages of 'over-sampling'

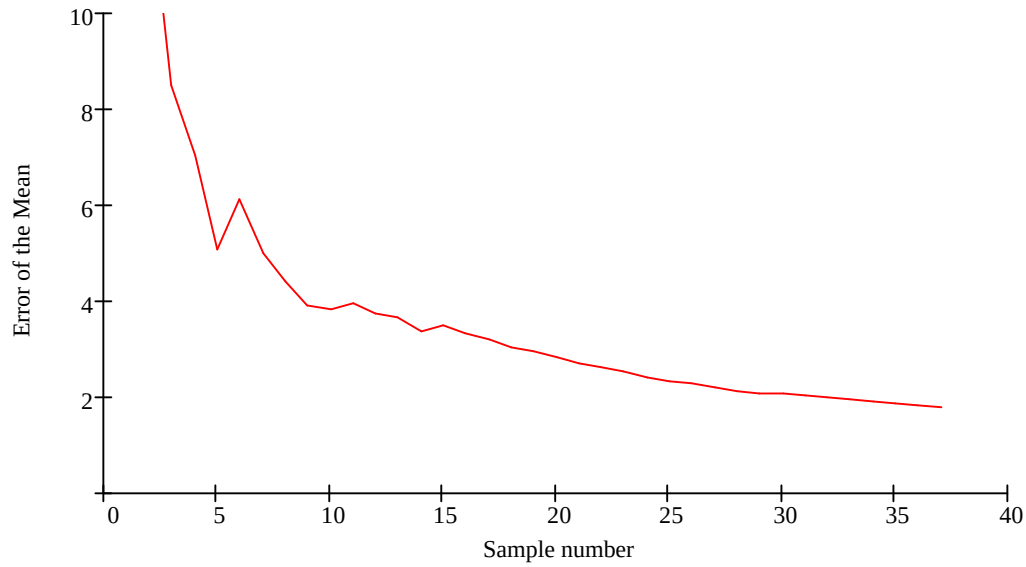
The final accuracy can be expected to be better than that specified.

Once you have completed this set of measurements, you are less likely to have to go back and take even more data.

Here are the additional 27 measurements:

76, 72, 56, 62, 78, 70, 74, 64, 58, 72, 66, 72, 60, 68,
62, 72, 64, 64, 72, 56, 60, 60, 58, 62, 64, 72, 66

It is worth noting that for an experiment that takes a lot of time or is expensive, it can be useful to check the achieved accuracy at different intervals during the data acquisition. Then, if the required accuracy is achieved before you have measured the "full" data set, you can stop recording data. As an example, the following graph shows how the error of the mean changes as we add each data sample above. Note that the error is not monotonic – that is, the error goes both down and up as the number of samples increases. Why might this be?



YOUR TURN: Turn to the student worksheet and complete the section - ***Why can the error of the mean sometimes go up (get worse) as we obtain more data samples?***

YOUR TURN: Turn to the student worksheet and complete the section - ***Acquire the additional data you need to complete your analysis.***

CHECK REQUIRED ACCURACY HAS BEEN ACHIEVED

We need to check that we have achieved the required accuracy. We do this in the same way that we did earlier for our small data set, only this time we use the full data set.

1) Calculate the mean and standard deviation of the full data set. Note that we do not show the full calculations here – just the results.

$$\bar{x}_{FULL} = \frac{\sum_{i=1}^{n_{FULL}} x_i}{n_1} = 65.35135$$

$$S_{FULL} = \sqrt{\frac{\sum_{i=1}^{n_{FULL}} \left(x_i - \bar{x}_1\right)^2}{\left(n_{FULL} - 1\right)}} = 6.377984$$

Recall that for this worked example we chose a 5% level of significance. We also have (37 – 1) = 36 degrees of freedom. We look up the Student-t value as:

t = 2.02809

The standard error of the mean is calculated as:

$$\text{Standard error} = d = \frac{t.S_{FULL}}{\sqrt{n_{FULL}}} = \frac{2.02809 \times 6.377984}{\sqrt{37}} = 2.12652$$

The average heart rate is therefore

65.35135 ± 2.12652 beats per minute

Or, with appropriate rounding,

65.3 ± 2.1 beats per minute

The error (2.1 b.p.m.) is less than the one we specified earlier in this worksheet (2.5 b.p.m.), so our experiment is complete.

Very often you will find that you did not achieve the required accuracy. This is quite normal. In this case, just go back to the very start of this worksheet and recalculate the required number of samples. But this time, instead of using a "small" sample, use the entire data set you have measured so far. If you used MathCAD for your calculations, it should be a quick exercise!

YOUR TURN: Turn to the student worksheet and complete the section - ***Calculate the mean and sample standard deviation for your full data set.***

GOODNESS OF FIT

So, you have found your average heart rate. We now try to determine what type of distribution best fits the data. We will do three "Goodness of Fit" tests:

Uniform distribution
Normal distribution
Poisson distribution

The general approach for all three Goodness of Fit tests is the same:

1) Make a hypothesis that a distribution is representative of the data sample.

2) Determine some statistics of the data set (e.g., mean, standard deviation).

3) Decide on a number of bins. Using the statistics, calculate an expected number of samples for each bin.

4) Compare the expected number of samples with the observed number using a Chi-squared test (Greek upper case symbol Chi^2, X^2).

5) If the Chi-squared test "passes" the conclusion is that "there is no strong evidence that the distribution *does not* represent the data." Note that we do not say a distribution fits the data – but that there is no evidence that it does NOT fit the data.

Goodness of Fit – Uniform Distribution

1) Make a hypothesis: The heart rate data can be modeled with a Uniform distribution.

YOUR TURN: Turn to the student worksheet and complete the section ***-Make your hypothesis about the Uniform distribution:***

2) Determine some sample statistics. For the uniform distribution we do not need any data set statistics.

YOUR TURN: Turn to the student worksheet and complete the section ***- Sample statistics for the Uniform distribution:***

3) Decide on bins, and number of samples in each bin Calculate expected number of samples for each bin.

Since the data range is 52 to 78 b.p.m., we choose to have 14 bins as follows:

51-53, 53-55, ... 77-79

Note that since we recorded out heartbeat as even numbers only, each measurement can only be assigned to one unique bin. This is one of the requirements for histogram and goodness of fit testing.

What could you have done to the bin ranges if you had recorded the data as all integers, e.g., 71bpm, 72bpm, etc?

To calculate the expected number of samples in each bin, E_k, we take the following approach, irrespective of the distribution we are considering:

Using the cumulative distribution function, determine the probability of a sample being *less than* the *lower limit* of each bin.

Using the cumulative distribution function, determine the probability of a sample being *less than* the *upper limit* of each bin.

The difference between these probabilities is the probability of a sample being in the bin. Multiply this probability by the total number of samples and we get the expected number of samples in the bin.

For a uniform distribution from 51 to 79, the following are the probabilities:

$P(x < 51) = 0$ (cannot happen)
$P(x < 79) = 1$ (must happen)
$P(51 < x < 79) = \frac{(x - 51)}{(79 - 51)}$

Bin limit	P(< bin limit) for Uniform distribution
51	0.00000
53	0.071429
55	0.142857
57	0.214286
59	0.285714
61	0.357143
63	0.428571
65	0.50000
67	0.571429
69	0.642857
71	0.714286
73	0.785714
75	0.857143
77	0.928571
79	1.00000

YOUR TURN: Turn to the student worksheet and complete the section - ***Decide on bin limits. Calculate the probability of a sample being less than each bin limit.***

We now calculate the expected number of samples in each bin. Note that since you only recorded your heart rate as an even number we do not have a problem with the heart rate hitting a bin limit. Note also, that we have included two extra bins to capture the very small and the very large values. Also note that fractional expected numbers are OK and quite normal! Recall that $n_{FULL} = 37$ for our demonstration data set.

Bin limits	P_i = P(upper)-P(lower)	Expected Number $= E_i = P_i.n_{FULL}$
0-51	0	0
51-53	0.071429	2.6429
53-55	0.071429	2.6429
55-57	0.071429	2.6429
57-59	0.071429	2.6429
59-61	0.071429	2.6429
61-63	0.071429	2.6429
63-65	0.071429	2.6429
65-67	0.071429	2.6429
67-69	0.071429	2.6429
69-71	0.071429	2.6429
71-73	0.071429	2.6429
73-75	0.071429	2.6429
75-77	0.071429	2.6429
77-79	0.071429	2.6429
79-infinity	0	0

Before we can use the X^2 Goodness of Fit test, we have to make some changes. The test only works well if the expected number of samples in each bin is at least 4 (some references suggest the minimum acceptable is 5 samples per bin). We achieve this minimum by combining some bins (bins do not have to be equal width for this analysis). For this particular example we will increase our bin width to 4 b.p.m. as follows:

Bin limits	Expected Number $= E_i = P_i.n_{FULL}$
0-51	0
51-55	5.285714
55-59	5.285714
59-63	5.285714
63-67	5.285714
67-71	5.285714
71-75	5.285714
75-79	5.285714
79-infinity	0

We now have k = 7 bins (we do not count the 'out-of-range' bins). Each bin now has at least 4 expected samples.

YOUR TURN: Turn to the student worksheet and complete the section - ***Calculate the expected number of samples in each bin. Modify the bins as necessary to ensure at least 4 samples are expected in each bin.***

4) Compare the expected number of samples with the observed number using a Chi-squared test.

The way we make the comparison is to calculate a test statistic as:

$$X_O^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where O_i is the observed number of samples in each bin and E_i is the expected number of samples based on the sample statistics. We then compare the test statistic with the X^2 distribution for (k – 1 – v) degrees of freedom, where v is the number of statistics we needed to determine for the distribution (zero for the Uniform distribution).

From tables, the X^2 statistic for 5% level of significance and (7 – 1 – 0) = 6 degrees of freedom is 12.592.

Bin limits	Expected Number, E_i	Observed number, O_i	$\frac{(O_i - E_i)^2}{E_i}$
51-55	5.285714	1	3.474903
55-59	5.285714	5	0.015444
59-63	5.285714	9	2.610039
63-67	5.285714	8	1.393822
67-71	5.285714	3	0.988417
71-75	5.285714	9	2.610039
75-79	5.285714	2	2.042471
TOTAL			13.13514

YOUR TURN: Turn to the student worksheet and complete the section - ***Calculate the test statistic.***

5) If the Chi-squared test "passes".
We now check the test statistic. How does the computed statistic (13.13514) compare with the tabulated X^2 statistic (12.592)? The test statistic is larger, so the test has *failed*.

> *Conclusion:* At the 5% level of significance there is strong evidence that a Uniform distribution is *not* representative of the heart rate data.

YOUR TURN: Turn to the student worksheet and complete the section – ***Check the test statistic for pass/fail.***

Goodness of Fit – Normal Distribution.
We will follow the same approach as we took for the Uniform distribution.

1) Hypothesis: The heart rate data can be modeled with a Normal distribution.

YOUR TURN: Turn to the student worksheet and complete the section –***Make your hypothesis about the Normal distribution:***

2) Determine some sample statistics.
For a Normal distribution we need the mean and standard deviation. We have already found those!

$$\bar{x} = 65.35135 \text{ bpm}$$
$$S = 6.377984 \text{ bpm}$$

YOUR TURN: Turn to the student worksheet and complete the section –***Sample statistics for the Normal distribution:***

3) Decide on bins, and number of samples in each bin. Calculate expected number of samples for each bin.
To start with we will use equal-spaced bins from 53 to 79 in intervals of 2 beats per minute. We will calculate the expected number of samples in each bin, E_k, using the same general approach as we took for the Uniform distribution, but using the different distribution equations/tables. That is:

> Using the cumulative distribution function, determine the probability of a sample being *less than* the *lower limit* of each bin.
>
> Using the cumulative distribution function, determine the probability of a sample being *less than* the *upper limit* of each bin.

The difference between these probabilities is the probability of a sample being in the bin. Multiply this probability by the total number of samples and we get the expected number of samples in the bin.

For the Normal distribution we calculate the z-statistic for the bin limits, and look up the probabilities in the Normal distribution table.

Bin limit	$z = \frac{(x - \bar{x})}{S}$	P(<z) from Normal distribution table
53	-1.9366	0.026399
55	-1.6230	0.052297
57	-1.3094	0.095199
59	-0.9958	0.159668
61	-0.6823	0.247542
63	-0.3687	0.35619
65	-0.0551	0.478034
67	+0.2585	0.601986
69	+0.5721	0.716363
71	+0.8856	0.812096
73	+1.1992	0.884780
75	+1.5128	0.934835
77	+1.8264	0.966104
79	+2.1400	0.983821
infinity	infinity	1.000000

YOUR TURN: Turn to the student worksheet and complete the section – ***Decide on bin limits. Calculate the probability of a sample being less than each bin limit.***

Now we calculate the expected number of samples in each bin. As before, we have included two extra bins to capture the very small and the very large values. These bins had little effect for the Uniform distribution, but are of importance for the Normal (and many other) distributions.

Bin limits	P_i = P(upper)-P(lower)	Expected Number $= E_i = P_i.n_{FULL}$
(-infinity)-53	0.026399	0.977
53-55	0.025897	0.958
55-57	0.042902	1.587
57-59	0.064469	2.385
59-61	0.087874	3.251
61-63	0.108646	4.020
63-65	0.121846	4.508
65-67	0.123952	4.586
67-69	0.114377	4.232
69-71	0.095734	3.542
71-73	0.072684	2.690
73-75	0.050056	1.852
75-77	0.031268	1.157
77-79	0.017717	0.656
79-infinity	0.016179	0.599

Again, we need to modify the bins to make sure we have at least 4 expectations in each bin. For this particular example we achieve this by combining some bins (bins do not have to be equal width for this analysis). Our reduced bin set is:

Bin limits	P_i = P(upper)-P(lower)	Expected Number $= E_i = P_i.n_{FULL}$	Observed number in bin, O_i
(-infinity)-59	0.159668	5.907707	6
59-63	0.19652	7.271252	9
63-65	0.121846	4.508301	5
65-67	0.123952	4.586214	3
67-69	0.114377	4.231939	2
69-73	0.168418	6.231448	8
73-infinity	0.11522	4.263139	4

YOUR TURN: Turn to the student worksheet and complete the section – ***Calculate the expected number of samples in each bin. Modify the bins as necessary to ensure at least 4 samples are expected in each bin.***

4) Compare the expected number of samples with the observed number using a Chi-squared test.

The test statistic is:

$$X_O^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where O_i is the observed number of samples in each bin and E_i is the expected number of samples based on the sample statistics.

We have k = 7 bins, and we had to calculate 2 statistics for the Normal distribution (mean and standard deviation). Therefore, there are (7 – 1 – 2) = 4 degrees of freedom.

From tables, the X^2 statistic for 5% level of significance and 4 degrees of freedom is 9.488.

Bin limits	Expected Number $= E_i = P_i.n_{FULL}$	Observed number, O_i	$\frac{(O_i - E_i)^2}{E_i}$
(-infinity)-59	5.907707	6	0.001442
59-63	7.271252	9	0.411012
63-65	4.508301	5	0.053627
65-67	4.586214	3	0.548617
67-69	4.231939	2	1.177133
69-73	6.231448	8	0.501934
73-infinity	4.263139	4	0.016242
TOTAL			2.710007

YOUR TURN: Turn to the student worksheet and complete the section – ***Calculate the test statistic.***

5) If the Chi-squared test "passes".

We now compare the computed statistic with the X^2 statistic. How does 2.710007 compare with 9.488?

Conclusion: At the 5% level of significance there is no strong evidence that a Normal distribution is not representative of the heart rate data.

Note the "double-negative" argument in the conclusion

YOUR TURN: Turn to the student worksheet and complete the section – ***Check the test statistic for pass/fail.***

Goodness of Fit – Poisson Distribution.

For this test, we use exactly that same approach as we did for both the Uniform and Normal distributions, except that we use the Poisson distribution function. This time, we will go straight through the exercise without a break, and then let you test your data set.

1) Hypothesis: The heart rate data can be modeled with a Poisson distribution.

2) Determine some sample statistics.
For a Poisson distribution we only need the sample mean. We have already have it as:

$$\bar{x} = 65.35135$$

3) Decide on bins, and number of samples in each bin Calculate expected number of samples for each bin.
To start with we will use equal-spaced bins from 53 to 79 in intervals of 2 beats per minute, and include the "very small" and "very large" bins. We then calculate the expected number of samples in each bin, E_k, using the tables of the cumulative distribution for the Poisson distribution.

Bin limits	P(in bin) = P(<upper limit)-P(<lower limit)	Expected Number $= E_i = P_i.n_{FULL}$	Observed number, O_i
0-53	0.067756	2.506955	1
53-55	0.041571	1.538111	0
55-57	0.056544	2.092145	2
57-59	0.071723	2.653766	3
59-61	0.085039	3.146444	4
61-63	0.094453	3.494761	5
63-65	0.09848	3.643748	5
65-67	0.096572	3.573153	3
67-69	0.089231	3.301539	2
69-71	0.077819	2.879298	1
71-73	0.06416	2.37391	7
73-75	0.050085	1.853163	2
75-77	0.037073	1.371714	1
77-79	0.026056	0.964074	1
79-infinity	0.043438	1.60722	0

Recall that we require the *expected* number in each bin to be at least 4. Therefore, we modify our bins as follows:

Bin limits	P(in bin) = P(<upper limit)-P(<lower limit)	Expected Number $= E_i = P_i.n_{FULL}$	Observed number, O_i	$\frac{(O_i - E_i)^2}{E_i}$
0-57	0.165871	6.13721	3	1.603675
55-61	0.156762	5.80021	7	0.24818
61-65	0.192933	7.138509	10	1.147036
65-69	0.185802	6.874692	5	0.511219
69-73	0.141979	5.253208	8	1.43624
73-infinity	0.156653	5.796171	4	0.556614
TOTAL				5.502963

4) Compare the expected number of samples with the observed number using a Chi-squared test.

The test statistic is:

$$X_O^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where O_i is the observed number of samples in each bin and E_i is the expected number of samples based on the sample statistics.

We have k = 6 bins, and we had to calculate one statistic for the Poisson distribution (the mean). Therefore, there are (6 – 1 – 1) = 4 degrees of freedom.

From tables, the X^2 statistic for 5% level of significance and 4 degrees of freedom is 9.488.

5) If the Chi-squared test "passes".

We now compare the computed statistic with the X^2 statistic. How does 5.502963 compare with 9.488?

> *Conclusion:* At the 5% level of significance there is no strong evidence that a Poisson distribution is not representative of the heart rate data.

Again, note the "double-negative" argument in the conclusion

YOUR TURN: Turn to the student worksheet and complete the section – ***This time, using the work you have already completed as examples, check your sample data against the Poisson distribution. Use separate paper and/or MathCAD worksheets (attach copies of all work).***

This completes the worksheet.