Exercise 4 (Solution steps not posted at the time of class are in red)

Part 1

1. Open the file normtemp.mat that contains the measurements of normal body temperature taken on healthy people. The three columns in the dataset are temperature, gender, and heart rate. We will need only the first column.

2. If you are Fahrenheit-challenged like me, you can convert the temperature to degrees C:

   tempC = (normtemp(:,1) - 32) * 5/9

3. Plot the histogram. Contemplate if the data is normally-distributed. Note the total number of measurements N.

   You can view the total number of measurements (which is 130) by simply looking at the variable tempC in the Array Editor, or by running

   size(tempC)

   I would define a separate variable:
   N=130

4. Calculate the mean and the variance of the distribution. Remember to use the formula for variance that uses (N-1), rather than N. Does the calculated mean correspond to the value that you thought was supposed to be the normal body temperature? (The "supposed normal" actually may depend on the country where you were born! I personally always thought it was 36.6 C = 97.88 F.) We are going to calculate the probability that your calculated mean is consistent with the "supposed normal" temperature.

   meanC=mean(tempC)
   stdC=std(tempC)

5. Calculate the uncertainty on the mean, s_μ.

   smu=stdC/sqrt(N)

6. Now, calculate the probability that your calculated mean is consistent with the "supposed normal" temperature. First, calculate how many standard deviations (s_μ, not σ!) your mean value is away from the "supposed normal". For me, this would be something like t=abs((meanC-36.6)/smu), where variables meanC and smu are the mean and the error on the mean. This gives the value of *t* in the Student's distribution, i.e. how far away you are from the "true" mean, in units of s_μ.

7. Using t-distribution, calculate the probability of being this far away from the "true" mean. You will probably want to use the tcdf(t,N-1) function.

   ```
   nu=N-1                  %the number of degrees of freedom
   format long             %avoid roundoff errors in displaying of small probabilities
   tcdf(t,nu)-tcdf(-t,nu)
   ```

The last line calculates the probability of being between –t and t standard deviations (smu) of being away from the mean. (As always, the c.d.f. gives the probability of an event being between minus infinity and the specified number. By taking the difference of the two c.d.f. values, you get the probability of an event being between the two specified numbers, -t and t in this case.)

8. Verify your result by calculating the 99% confidence interval for the mean (i.e., the temperature range where the true mean should fall). You can use either *t*- or Gaussian distributions. They should be close for this large N.

   You can use, for example, disttool to see that 99% confidence interval corresponds to 2.36 standard deviations in Student's t with 129 degrees of freedom, or to 2.33 in standard Gaussian distribution. The confidence interval is therefore

   Tmin=meanC-2.36*smu
   Tmax=meanC+2.36*smu

Additional question 1: When someone's temperature should be considered abnormal? I.e., when the physician should give you a sick leave from the class?

This depends on your definition of 'abnormal'. For example, if you define it as exceeding the 99% confidence interval, then it is any temperature exceeding

Thigh=meanC+2.33*stdC

Note that, because the question is about an individual temperature measurement rather than the mean, I used the standard deviation of the distribution, not of the mean (stdC rather than smu). I also used the coefficient (2.33, see previous question) for the Gaussian distribution, because the temperature data is distributed normally.

Additional question 2: Is there a significant difference between males and females in terms of the normal body temperature? In the second column of your dataset, 1=male, 2=female.

By inspecting the variable normtemp you can see that the first 65 data points correspond to males while the rest are for females. Calculate means, variances, and errors on the means:

```
meanMale=mean(tempC(1:65))
meanFemale=mean(tempC(66:130))

stdMale=std(tempC(1:65))
stdFemale=std(tempC(66:130))

smuMale=stdMale/sqrt(65)
smuFemale=stdFemale/sqrt(130-65)
```

Check how many standard deviations (smu) the two means are apart:

```
(meanFemale-meanMale)/smuMale
(meanFemale-meanMale)/smuFemale
```

If the two means are more than 2.36 standard deviations (smu) apart, the difference is statistically significant with 99% confidence.

Additional question 3: You may have noticed that there are several temperature measurements that appear abnormally high or abnormally low for a healthy person. Is it possible that those persons were actually sick at the time of measurements? In other words, should you exclude these "abnormal" measurements from the statistics?

The maximum recorded temperature can be viewed by

```
max(tempC)
```

This is this many standard deviations away from the mean:

```
smax=(max(tempC)-meanC)/stdC
```

The probability of observing such or a more extreme temperature is

```
Prob = 1-normcdf(smax)
```

The expected number of events is then

```
Prob*N
```

According to the Chauvenet's criterion, if this number is less than ½ then the event can be discarded.

Exercise 3, Part 2:

9. We now want to investigate how the total number of measurements affects the confidence of your conclusions. Suppose that instead of 130 measurements, you only made 6. To simulate this, let's randomly pick 6 values from the dataset:

```
tempR(1:6)=0;
tempR(1)=tempC(floor(unifrnd(1,131)))
tempR(2)=tempC(floor(unifrnd(1,131)))
tempR(3)=tempC(floor(unifrnd(1,131)))
tempR(4)=tempC(floor(unifrnd(1,131)))
tempR(5)=tempC(floor(unifrnd(1,131)))
tempR(6)=tempC(floor(unifrnd(1,131)))
```

   This stores 6 randomly chosen values in the variable tempR. Plot them in a histogram.

10. Calculate the mean, variance, and the error on the mean.

11. With such a small sample, it is difficult to predict where your mean will be relative to the "true" mean, so instead of calculating the *t* value, let's use the t-distribution to calculate the 99% confidence interval. Table C.8 in Bevington's book tells me that, for N=6, the 99% confidence corresponds to exactly 4*t*. (Which value did I use for *v*?) You can verify this result using tcdf. So, for your 6-point dataset, what are the minimum and maximum

temperatures that correspond to this confidence interval? Is this interval for the value of the mean smaller or larger than the spread of your measured temperature values in the histogram?

The 95% confidence for N=6 approximately corresponds to t=2.6. See how this changes your result.

12. Compare the confidence limits that you obtained using Student's *t* to what you'd have obtained using Gaussian. Gaussian probability 99% corresponds to 2.6*t* (or $2.6\sigma_{\mu}$), whereas probability 95% approximately corresponds to 2*t* (regardless of the number of measurements!)