Class i	Center of Class	N_i			Cumulative frequency			Relative frequency, $p_i = N_i / \sum N_i$		
(mm)	(mm)	Location			Location			Location		
		A	B	C	A	B	C	A	B	C
0–0.125	0.0625	103	0	0	103	0	0	0.2032	0.0000	0.0000
0.125–0.375	0.2500	187	0	0	290	0	0	0.3688	0.0000	0.0000
0.375–0.625	0.5000	96	0	0	386	0	0	0.1893	0.0000	0.0000
0.625–0.875	0.7500	87	78	26	473	78	26	0.1716	0.1497	0.0478
0.875–1.125	1.0000	31	255	259	504	333	285	0.0611	0.4894	0.4761
1.125–1.375	1.2500	2	123	155	506	456	440	0.0039	0.2361	0.2849
1.375–1.625	1.5000	1	41	63	507	497	503	0.0020	0.0787	0.1158
1.625–1.875	1.7500	0	16	27	507	513	530	0.0000	0.0307	0.0496
1.875–2.125	2.0000	0	6	10	507	519	540	0.0000	0.0115	0.0184
2.125–2.375	2.2500	0	2	3	507	521	543	0.0000	0.0038	0.0055
2.375–2.625	2.5000	0	0	1	507	521	544	0	0	0.0018382
							Σ	1	1	1

Class i	Center of Class	$x_i p_i$			$x_i^2 p_i$			CDF(Σp_i)		
(mm)	(mm)	Location			Location			Location		
		A	B	C	A	B	C	A	B	C
0–0.125	0.0625	0.0127	0.0000	0.0000	0.0008	0.0000	0.0000	0.2032	0.0000	0.0000
0.125–0.375	0.2500	0.0922	0.0000	0.0000	0.0231	0.0000	0.0000	0.5720	0.0000	0.0000
0.375–0.625	0.5000	0.0947	0.0000	0.0000	0.0473	0.0000	0.0000	0.7613	0.0000	0.0000
0.625–0.875	0.7500	0.1287	0.1123	0.0358	0.0965	0.0842	0.0269	0.9329	0.1497	0.0478
0.875–1.125	1.0000	0.0611	0.4894	0.4761	0.0611	0.4894	0.4761	0.9941	0.6392	0.5239
1.125–1.375	1.2500	0.0049	0.2951	0.3562	0.0062	0.3689	0.4452	0.9980	0.8752	0.8088
1.375–1.625	1.5000	0.0030	0.1180	0.1737	0.0044	0.1771	0.2606	1.0000	0.9539	0.9246
1.625–1.875	1.7500	0.0000	0.0537	0.0869	0.0000	0.0940	0.1520	1.0000	0.9846	0.9743
1.875–2.125	2.0000	0.0000	0.0230	0.0368	0.0000	0.0461	0.0735	1.0000	0.9962	0.9926
2.125–2.375	2.2500	0.0000	0.0086	0.0124	0.0000	0.0194	0.0279	1.0000	1.0000	0.9982
2.375–2.625	2.5000	0.0000	0.0000	0.0046	0.0000	0.0000	0.0115	1.0000	1.0000	1.0000
	Σ	0.3973	1.1003	1.1824	0.2395	1.2792	1.4737			

a. Fill in the empty columns in the table.

Refer to the tables provided above.

b. Provide a point estimate for the mean bubble size of the population represented by the results of location A, B and C. What can you say about these values?

The mean can be found from the expected value as: $E(x) = \sum x_i p_i$ **From which the following values can be found for the three locations (in mm):**

$$\hat{\mu}_A = \overline{x}_A = 0.3973,$$
$$\hat{\mu}_B = \overline{x}_B = 1.1003,$$
$$\hat{\mu}_C = \overline{x}_C = 1.1824$$

These values suggest that there is little difference between the mean bubble size at locations B and C. However, both differ significantly from the mean bubble size at A.

c. Provide a point estimate for the standard deviation for the bubble size of the population represented by the results of location B.

The standard deviation can be found from the expected value as:

$$V(x)=\sum x_i^2 p_i-\mu^2=1.2792\text{-}1.1003^2=0.0685$$
$$s=\sqrt{V(x)}=0.2618$$

From which: $\hat{\sigma}_B=s_B=0.2618$ **mm.**

d. What is the standard error for the sample mean of B?

$$s_{\bar{x}}\approx s/\sqrt{n}=0.2618/\sqrt{11}=0.0789\text{ mm.}$$

e. What is the median for C?

The median is the value where 50% of the CDF is above it and 50% is below it. For location C, interpolation yields $\tilde{x}\approx 0.99$ **mm.**

f. What is the mode for C?

The mode is the most frequently occurring value. The most frequent value from location C is obtained at *N* = 259 which occurs at a bubble size of 1.00 mm.

g. Determine the CV for B.

$$CV=100s/\bar{x}=100(0.2618)/1.1003=23.8\%.$$

h. Determine coefficient of skewness for location C.

$$\text{Sk}=\frac{3(\text{mean}-\text{median})}{\text{standard deviation}}=\frac{3(\bar{x}-\tilde{x})}{s}$$
$$V(x)=\sum x_i^2 p_i-\mu^2=1.4737-1.1824^2=0.0755$$
$$s=\sqrt{V(x)}=0.2748$$
$$\text{Sk}=\frac{3(1.1824-0.99)}{0.2784}=2.1009.$$

Clearly, the distribution is skewed to the right.

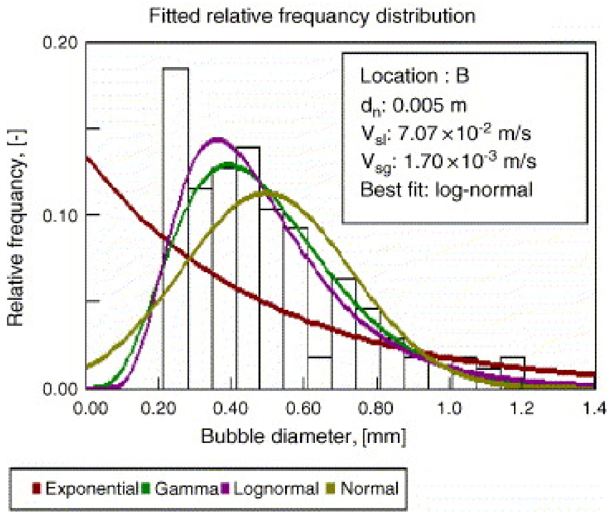
i. Determine the following probabilities

i. $P(X\leq 0.5)$ for location A = **0.7613**.

ii. $P(0.75\leq X\leq 1.25)$ for location C = **0.8088 – 0.0478 = 0.7610**.

iii. $P\,(X>1.25)$ for location B = **1 - *P* (*X* ≤ 1.25) = 1-0.8752 = 0.1248**.

j. The authors fitted four distributions to their results at location B. The results of fitting are shown in the figure provided. Which distribution in your opinion best describes the experimental data? Justify your answer.



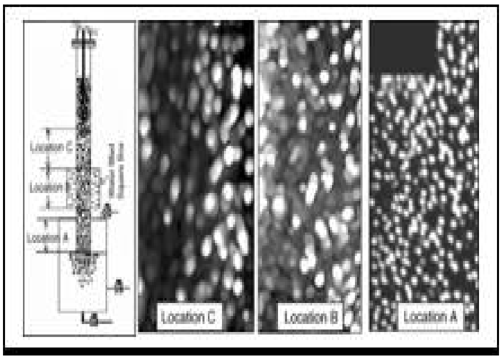


Figure 1 Fitted different distributions of bubble sizes at location B.

Lognormal distribution best fits the results. The distribution is skewed (exclude the symmetric normal distribution), and mono-modal which excludes the monotonically decreasing exponential distribution. Both gamma and lognormal distributions capture the basic features of the experimental results. However, the lognormal is closer to the experimental.

2. **(20 points)** The following data shows the per capita carbon dioxide emissions from the consumption and flaring of fossil fuels during the period 1997-2004 in metric tons of carbon dioxide (*Source: US DOE, Energy Information Administration, International Energy Annual 2004.*) Is there any strong evidence suggesting that the Israel's per capita emissions are higher than those of Jordan's per capita emissions? Comment on your conclusion.

Year	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
Israel	9.94	10.36	10.37	10.64	11.21	11.38	10.48	10.69
Jordan	3.02	3.14	3.10	3.10	2.95	2.99	3.11	3.32

Calculate the descriptive statistics for the two samples.

	Israel	Jordan
Mean	10.63375	3.09125
Variance	0.220227	0.012927
Observations	8	8

The two variances are not equal. Therefore, the Smith-Satterthwaite test is to be used. We wish to determine if there is any difference between the mean per capita emissions between Jordan and Israel. Apply the eight step procedure:

1. The parameters of interest are the per capita emissions for Jordan and Israel. We are interested in determining whether $\mu_1 - \mu_2 = 0$.
2. Null hypothesis. $H_0 : \mu_1 - \mu_2 = 0$
3. Alternate hypothesis. $H_1 : \mu_1 - \mu_2 \neq 0$
4. Level of significance is strong evidence $\alpha = 0.05$.
5. Test statistic is

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2 \text{ - } 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

6. Degrees of freedom

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left(0.2202/8 + 0.01293/8\right)^2}{\frac{(0.2202/8)^2}{7} + \frac{(0.01293/8)^2}{7}} = 8$$

Therefore, we would reject $H_0 : \mu_1 - \mu_2 = 0$ if $t_0 > t_{0.025,8} = 2.306$ or if $t_0 < -t_{0.025,8} = -2.306$.

7. Computation: using the sample data

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{10.63 - 3.09}{\sqrt{\frac{0.2202}{8} + \frac{0.01293}{8}}} = 44.18.$$

8. Conclusions: because $t_0 = 44.18 > t_{0.025,8} = 2.306$, we reject the null hypothesis.

   Therefore, there is strong evidence to conclude that the mean per capita emissions for Jordan are different than those for Israel. Furthermore, the mean per capita emissions for Israel are higher than those for Jordan!

3. (20 points) The following are the average population IQ in the states during the 2004 election between Bush and Kerry (*source: http://chrisevans3d.com/files/iq.htm*). Is there any strong evidence that the average population IQ for people voting for Kerry is different than those voting for Bush?

Population average IQ for states that **voted for Kerry**	*113*	*111*	*111*	*109*	*107*
	106	*105*	*105*	*104*	*103*
	102	*102*	*102*	*101*	*101*
	100	*100*	*99*	*99*	
Population average IQ for states that **voted for Bush**	*100*	*98*	*94*	*92*	*92*
	99	*98*	*94*	*92*	*90*
	99	*98*	*94*	*92*	*90*
	99	*96*	*93*	*92*	*90*
	99	*95*	*93*	*92*	*90*
	90	*89*	*89*	*87*	*87*
	85				

	Bush	Kerry
Mean	93.16	104.2
Variance	16.74	18.40
Observations	31	19

The solution here will be credited whether you use the normal distribution assuming the variances are those of the population or using the *t*-distribution assuming the variances are for a sample. Nevertheless, the conclusions arrived at should be the same between the two approaches. I will solve it using the *t*-distribution for convenience assuming equal variances.

We wish to determine if there is any difference between the mean IQ in states which voted for Bush and those which voted for Kerry. Apply the eight step procedure:

1. The parameters of interest are the average population IQ in the states during the 2004 election between Bush and Kerry. We are interested in determining whether $\mu_1 - \mu_2 = 0$.
2. Null hypothesis. $H_0 : \mu_1 - \mu_2 = 0$
3. Alternate hypothesis. $H_1 : \mu_1 - \mu_2 \neq 0$
4. Level of significance is strong evidence $\alpha = 0.05$.
5. Test statistic is
$$t_0 = \frac{\overline{x}_1 - \overline{x}_2 \text{ - } 0}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$
6. Therefore, we would reject $H_0 : \mu_1 - \mu_2 = 0$ if $t_0 > t_{0.025,48} = 1.678$ or if $t_0 < -t_{0.025,48} = -1.678$.

7. Computation: using the sample data, calculate the pooled standard deviation and the test statistic
$$s_p^2 = \frac{s_1^2\left(n_1 - 1\right) + s_2^2\left(n_2 - 1\right)}{n_1 + n_2 - 2} = \frac{16.74(30) + 18.40(18)}{31 + 19 - 2} = 17.36$$

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{93.16-104.2}{\sqrt{17.36\left(\frac{1}{31}+\frac{1}{19}\right)}} = -9.10$$

8. Conclusions: because $t_0 = -9.10 < -t_{0.025,48} = -1.678$, we reject the null hypothesis.

   Therefore, there is strong evidence to conclude that average population IQ in the states during the 2004 election between Bush and Kerry is different. Furthermore, average population IQ in the states during the 2004 election which voted for Bush is lower than those voting for Kerry!

The same calculations can be carried out using the assumption of the variance to be those of the population.

1. The parameters of interest are the average population IQ in the states during the 2004 election between Bush and Kerry. We are interested in determining whether $\mu_1 - \mu_2 = 0$.
2. Null hypothesis. $H_0 : \mu_1 - \mu_2 = 0$
3. Alternate hypothesis. $H_1 : \mu_1 - \mu_2 \neq 0$
4. Level of significance is strong evidence $\alpha = 0.05$.
5. Test statistic is

$$z_0 = \frac{\overline{x}_1 - \overline{x}_2 \text{ - } 0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

6. Therefore, we would reject $H_0 : \mu_1 - \mu_2 = 0$ if $z_0 > z_{0.025} = 1.96$ or if $z_0 < -z_{0.025} = -1.96$.
7. Computation: using the sample data, calculate the pooled standard deviation and the test statistic

$$z_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{93.16-104.2}{\sqrt{\frac{16.74}{31}+\frac{18.40}{19}}} = -9.00$$

8. Conclusions: because $z_0 = -9.00 < -z_{0.025} = -1.96$, we reject the null hypothesis.

   Therefore, there is strong evidence to conclude that average population IQ in the states during the 2004 election between Bush and Kerry is different. Furthermore, average population IQ in the states during the 2004 election which voted for Bush is lower than those voting for Kerry!

4. Let X: the number of graphite particles in a 1/4-inch-square area of casting. X is a Poisson with parameter $\lambda = 20(1/4) = 5$.

To answer the question, P(X < 2) = P(X ≤ 1) = 0.04043.

$$P(X < 2) = P(0) + P(1) = \frac{5^0 e^{-5}}{0!} + \frac{5^1 e^{-5}}{1!} = 0.04043$$

This is indeed a small probability. Therefore it is unusual to have a cast ion with fewer than two particles in a 1/4-inch-square area of casting.

5. (30 points) Safi, Nicolas, Neau, and Chevalier measured the diffusion coefficients of aromatic compounds at infinite dilution in binary mixtures at 298.15 K (*Source: Amor Safi, Christophe Nicolas, Evelyne Neau, and Jean-Louis Chevalier, Diffusion Coefficients of Aromatic Compounds at Infinite Dilution in Binary Mixtures at 298.15 K, J. Chem. Eng. Data, 52 (1), 126 -130, 2007*). An excerpt of their results of the infinite dilution diffusion coefficients of benzene (1) in mixtures of hexane (2) + ethanol (3) are given in below.

Table 1 infinite dilution diffusion coefficients of benzene (1) in mixtures of hexane (2) + ethanol (3)

i	$x = x_2$	$y =$ $10^5 D_{1,\mathrm{m}}^{\propto}$ $(\mathrm{cm}^2.\mathrm{s}^{-1})$	x^2	xy	y_{pred}	$y - y_{\mathrm{pred}}$	$(y-\bar{y})^2$	$(y - y_{\mathrm{pred}})^2$
1	0.0000	1.88	0.0000	0.0000	1.7011	0.1789	-0.786	0.03201
2	0.2024	2.13	0.04097	0.43111	2.1932	-0.0632	-0.536	0.003994
3	0.3942	2.45	0.15539	0.96579	2.6595	-0.2095	-0.216	0.04389
4	0.6082	3.05	0.36991	1.8550	3.1797	-0.1297	0.384	0.01682
5	0.7796	3.82	0.60778	2.9781	3.5964	0.2236	1.154	0.05000
Σ	**1.9844**	**13.33**	**1.174**	**6.23**				

a. Fit a simple linear regression model for the diffusion coefficient, $D_{1,\mathrm{m}}^{\propto}$, with the mole fraction, x_2.

$$\sum x_i = 1.9844, \quad \sum y_i = 13.33$$

$$\sum x_i^2 = 1.174, \quad \sum x_i y_i = 6.23$$

$$\bar{x} = 0.39688, \quad \bar{y} = 2.666$$

$$S_{xx} = \sum x_i^2 - \left(\sum x_i\right)/n = 0.38647$$

$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n = 0.93957$$

$$\beta_1 = S_{xy}/S_{xx} = 0.93957/0.38647 = 2.4311$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 2.666 - (2.4311)(0.39688) = 1.7011$$

$$y = \beta_0 + \beta_1 x = 1.7011 + 2.4311x$$

$$\mathrm{SSE} = \sum (y_i - y_{\mathrm{pred},i})^2 = 0.1467$$

$$\mathrm{SST} = \sum (y_i - \bar{y})^2 = 2.4309$$

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{0.1467}{2.4309} = 0.9396$$

$$R = \sqrt{R^2} = 0.9694$$

$$\hat{\sigma}^2 = \mathrm{SSE}/(n\text{-}2) = 0.1467/3 = 0.0489$$

$$\mathrm{RMSE} = \sqrt{\hat{\sigma}^2} = 0.2211$$

Find the confidence intervals for the intercept and slope. Use 95% confidence level.

$$t_{\alpha/2,n-2} = t_{0.025,3} = 3.1824$$

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$1.2991 \le \beta_1 \le 3.5631$$

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]}$$

$$1.1526 \le \beta_0 \le 2.2496$$

Predict the value at $x_2 = 1$ and compare it with the reported value of 4.70.

$$y(x=1) = 2.431 + 1.701(1) = 4.1323$$

Analyze the residuals and comment on model adequacy.

Calculate the cumulative probabilities after sorting the residuals. Subsequently, plot the probability versus the residual on the normal probability paper.

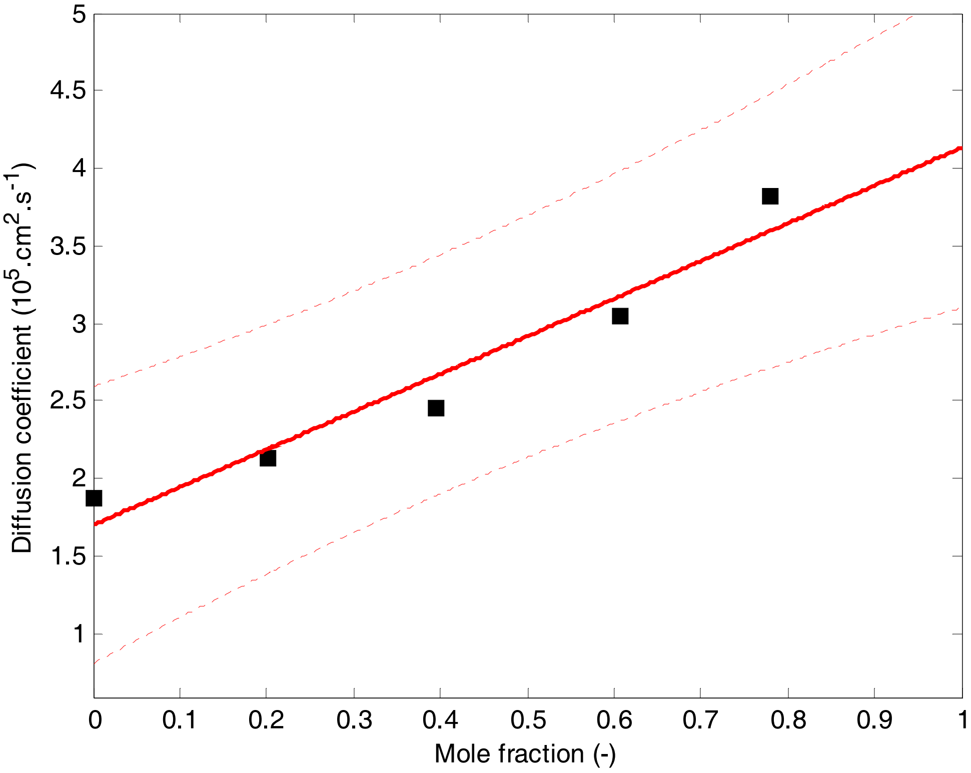
j	$y - y_{\text{pred}}$	$(j\text{-}0.5)/n$
1	-0.2095	0.1
2	-0.1297	0.3
3	-0.0632	0.5
4	0.1789	0.7
5	0.2236	0.9

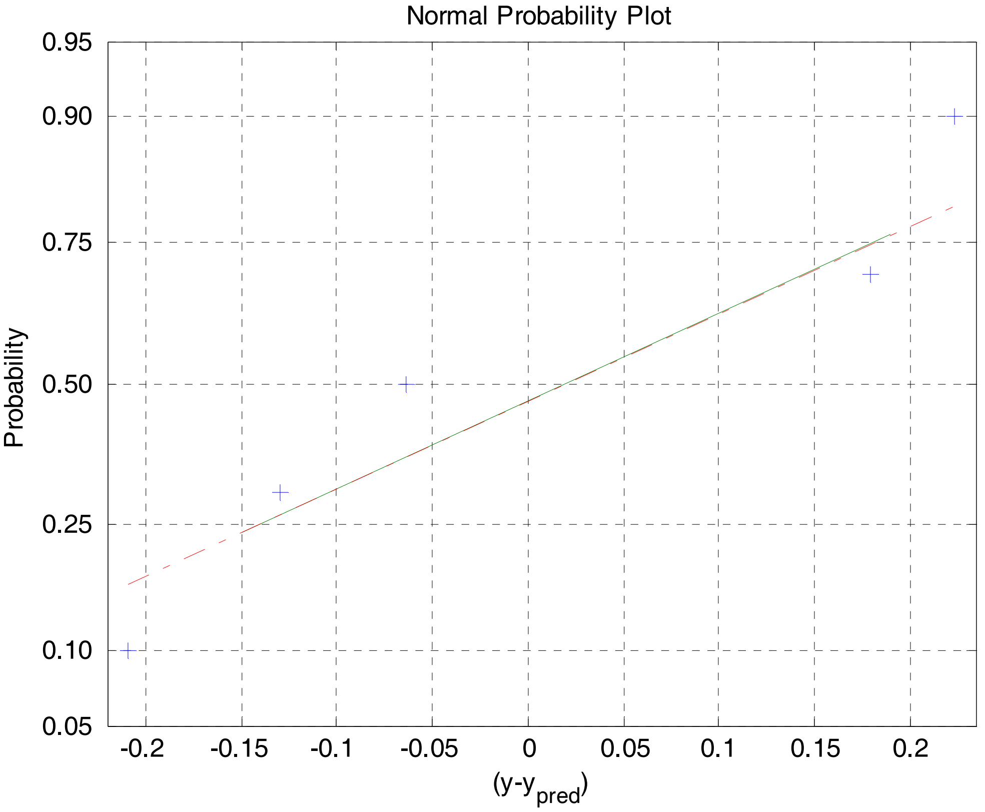
Clearly, the residuals do not follow a normal distribution. This is indicative of the inadequacy of the linear model for fitting the diffusion coefficient data.

Matlab output below for verification of the calculations.

```
Linear model Poly1:
       f(x) = p1*x + p2
Coefficients (with 95% confidence bounds):
       p1 =       2.431  (1.299, 3.563)
       p2 =       1.701  (1.153, 2.25)

Goodness of fit:
  SSE: 0.1467
  R-square: 0.9397
  Adjusted R-square: 0.9195
  RMSE: 0.2211
```


Diffusion coefficient (10^5.cm^2.s^{-1})
Mole fraction (-)
5
4.5
4
3.5
3
2.5
2
1.5
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Normal Probability Plot
Probability
(y-y$_{pred}$)
0.95
0.90
0.75
0.50
0.25
0.10
0.05
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2