

STAT 516 fa 2026 exam 01

75 minutes, one page of handwritten notes allowed, no calculators

1.

Arctic explorer Fridtjof Nansen published in his book *Farthest North* (Nansen and Sverdrup (1898)) a table of the names and weights, in pounds and kilograms, of the sled dogs he and his co-explorer took with them on their journey seeking the North Pole. The table is reproduced below:

dogs

	Name	Lbs	Kilos
1	Kvik	78.0	35.7
2	Freia	50.0	22.7
3	Barbara	49.5	22.5
4	Suggen	61.5	28.0
5	Flint	59.5	27.0
6	Barrabas	61.5	28.0
7	Gulen	60.5	27.5
8	Haren	61.5	28.0
9	Barnet	39.0	17.7
10	Sultan	68.0	31.0
11	Klapperslangen	59.5	27.0
12	Blok	59.0	26.8
13	Bjelki	38.0	17.3
14	Sjöliget	40.0	18.0
15	Katta	45.5	20.7
16	Narrifas	46.0	21.0
17	Livjaegeren	38.5	17.5
18	Potifar	57.0	26.0
19	Storaeven	70.0	31.8
20	Isbjön	61.5	28.0

21	Lilleraeven	59.0	26.7
22	Kvindfolket	37.0	26.0
23	Perpetuum	63.0	28.6
24	Baro	60.5	27.5
25	Russen	58.0	26.5
26	Kaifas	69.0	31.5
27	Ulenka	57.0	26.0
28	Pan	65.0	29.5

Below is some R code and output.

```
lm_out <- lm(Lbs ~ Kilos, data = dogs)
summary(lm_out)
```

Call:

```
lm(formula = Lbs ~ Kilos, data = dogs)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.43	0.58	0.71	0.84	1.04

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.210	4.229	-0.05	0.96
Kilos	2.179	0.161	13.54	2.8e-13 ***

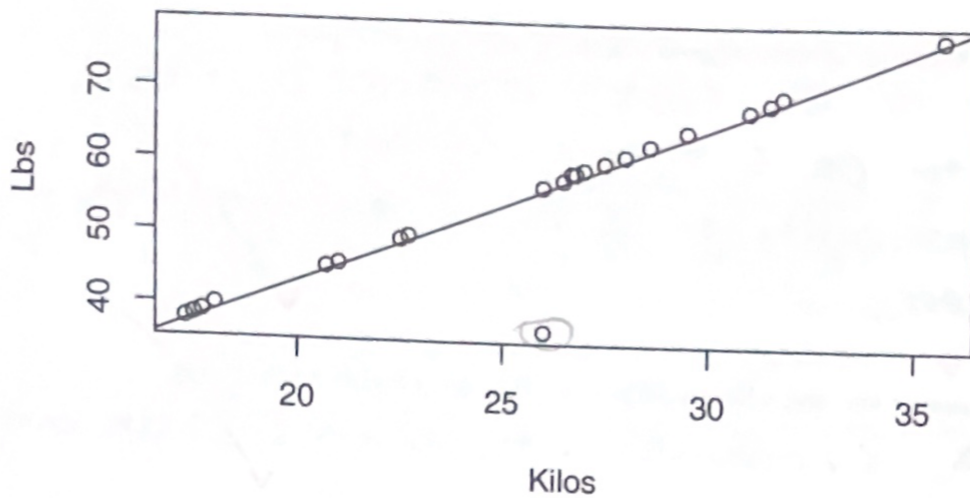
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.88 on 26 degrees of freedom

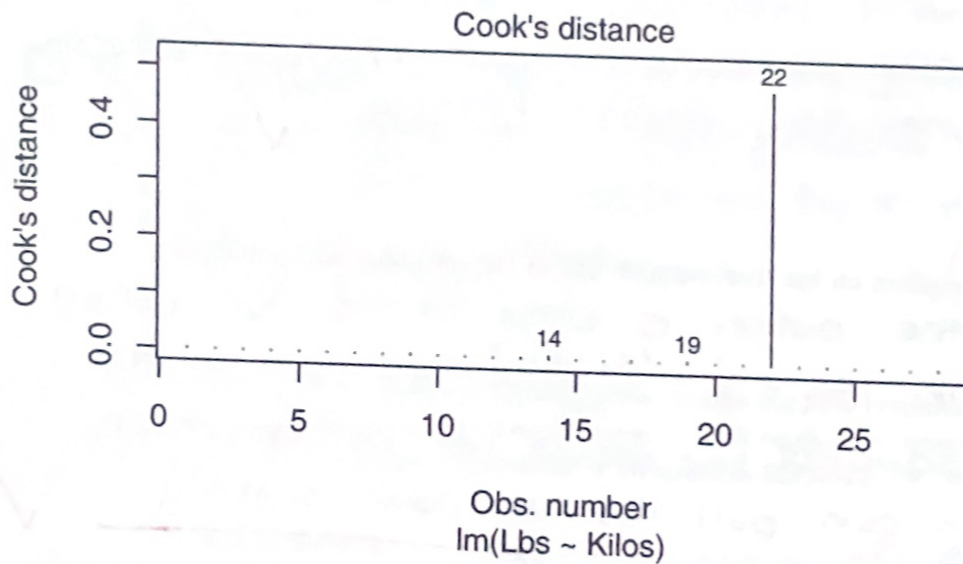
Multiple R-squared: 0.876, Adjusted R-squared: 0.871

F-statistic: 183 on 1 and 26 DF, p-value: 2.78e-13

```
plot(Lbs ~ Kilos, data = dogs)
abline(lm_out)
```



```
plot(lm_out, which = 4)
```



- (a) If one did not know how many pounds were in a kilogram, how many pounds per kilogram would one estimate based on the fitted model?

Based on the data, one would estimate that there are 2.179 lbs in a kg ($\hat{\beta}_1 = 2.179$)

- (b) Give an explanation of what appears in the Cook's distance plot.

Based on the Cook's distance plot there seems to be 1 significant outlier, with the possibility of 2 more less significant outliers.

- (c) Give the name of the dog whose pound and kilogram weights are outlying.

Dog 22, Kvindfolket with weights 37 lbs and 26 kgs

- (d) What do you think happened here to create the outlier?

There was probably a mistake in recording the weight, with maybe mistyping 37 when they meant 57 lbs.

- (e) Describe the effect on the least-squares line of the outlying data point.

Since the outlier is close to the $\bar{x}_n$, based on the $lev_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{S_{xx}}$, the pull on the least-squares line should not be too much, but it can pull the line down slightly.

- (f) State whether the outlier has a high leverage value and explain your answer.

The outlier will not have a high leverage value as the x_i is close to $\bar{x}_n$, making the numerator of the second part of the equation small. So, the leverage will be low.

- (g) From the R output above, report the value of $\hat{\sigma}$. Give a careful assessment of the quality of this estimate (is it likely too large, too small, or neither? Why?)

$$\hat{\sigma} = 3.88$$

This value is the standard deviation of the error term. This value is likely too large given the significant outlier in the data, making the estimated deviation from the least squares line greater. ✓

- (h) Give the value of the coefficient of determination for the fitted model.

$$R^2 = .876$$

- (i) After removing the outlier from the data, the coefficient of determination for predicting the pound weights from the kilogram weights becomes $R^2 = 1$. Give an interpretation of this value.

R^2 explains the proportion of the variation in Y explained by X . $R^2 = 1$ means 100% of the variation of Y is explained by X making this model perfectly linear. ✓

- (j) After removing the outlier from the data, one obtains the 99% confidence intervals shown below for β_0 and β_1 . Give an interpretation of both confidence intervals in light of the fact that the true number of pounds in a kilogram is 2.204623. $\rightarrow$ true β_1

	0.5 %	99.5 %
(Intercept)	-0.12	0.916
Kilos	2.16	2.203

The true value of β_1 does not fall in the confidence interval $[2.16, 2.203]$ since the actual conversion factor is 2.204623. This suggests there is something off with the data. β_0 does fall in the interval of $[-.12, .916]$ since the real β_0 should have a value of 0 (0lbs = 0 kgs). The β_0 interval seems to be heavy to the right with β_1 being too heavy to the left. ✓



2

The data set of which a few rows are printed below contains the length, diameter, number of rings, shucked weight, and sex (“M” for male, “F” for female, or “I” for infant) on a number of abalones in variables with the names `len`, `diam`, `rings`, `wt`, and `sex`, respectively.

```
head(ab)
```

	len	diam	rings	wt	sex
1	0.350	0.265	7	0.0995	M
2	0.530	0.420	9	0.2565	F
3	0.440	0.365	10	0.2155	M
4	0.330	0.255	7	0.0895	I
5	0.425	0.300	8	0.1410	I
6	0.530	0.415	20	0.2370	F

The R code below fits and displays some output for three models.

Model 1:

```
model1 <- lm(log(wt) ~ sex + log(len) + log(rings) + sex:log(len), data = ab)
summary(model1)
```

Call:

```
lm(formula = log(wt) ~ sex + log(len) + log(rings) + sex:log(len),
    data = ab)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.7169	-0.1199	-0.0076	0.1070	1.2264

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2265	0.0372	32.94	< 2e-16 ***
sexI	-0.0455	0.0242	-1.88	0.06 .
sexM	0.1207	0.0228	5.29	1.3e-07 ***
log(len)	3.1126	0.0319	97.62	< 2e-16 ***
log(rings)	-0.1732	0.0123	-14.03	< 2e-16 ***

```
sexI:log(len)  0.0347    0.0359    0.97    0.33
sexM:log(len)  0.1714    0.0381    4.50  7.1e-06 ***
---
```

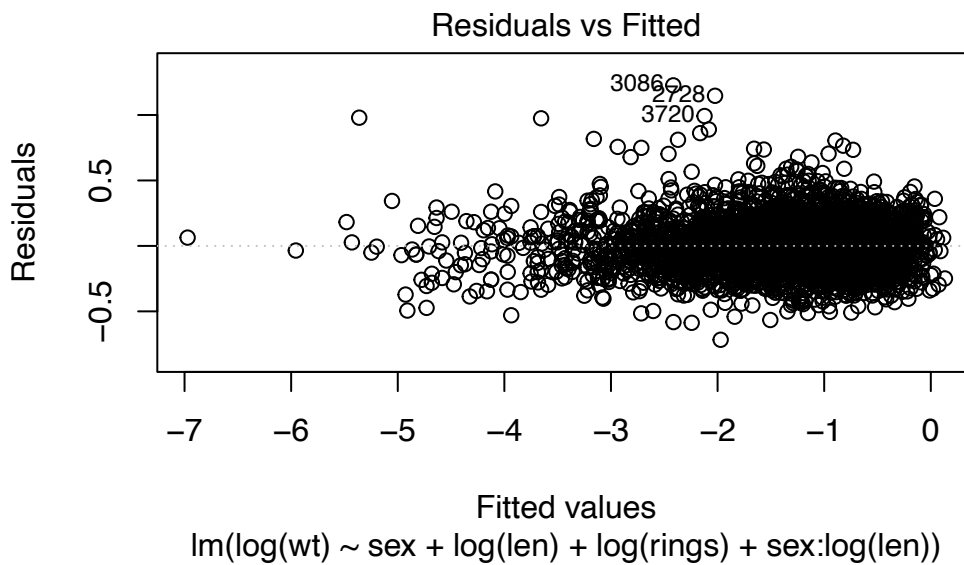
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.181 on 4166 degrees of freedom

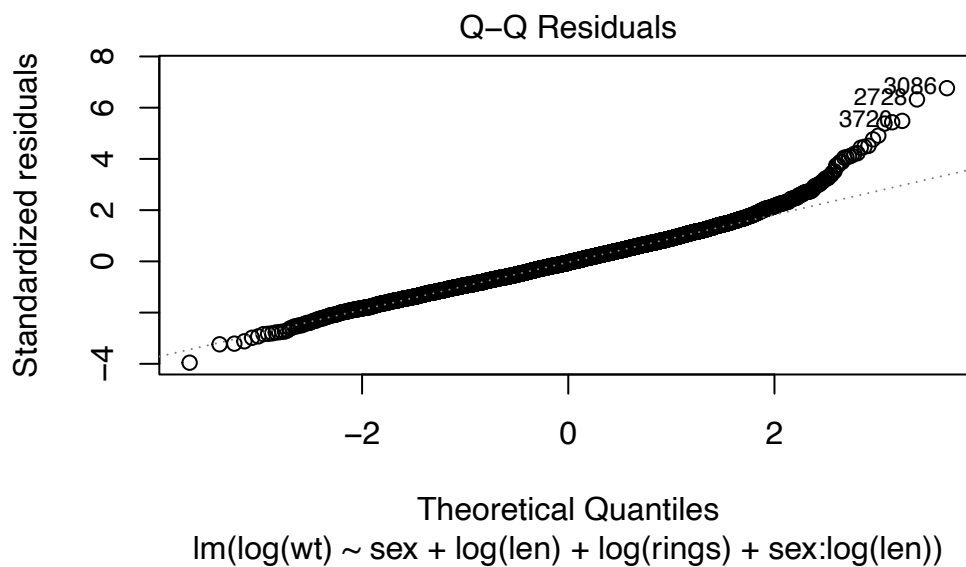
Multiple R-squared: 0.956, Adjusted R-squared: 0.956

F-statistic: 1.5e+04 on 6 and 4166 DF, p-value: <2e-16

```
plot(model1,which=1,add.smooth=F)
```



```
plot(model1,which=2)
```



Model 2:

```
model2 <- lm(log(wt) ~ log(len) + log(rings), data = ab)
summary(model2)
```

Call:

```
lm(formula = log(wt) ~ log(len) + log(rings), data = ab)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.711	-0.120	-0.007	0.108	1.209

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2439	0.0347	35.9	<2e-16 ***
log(len)	3.2271	0.0143	225.7	<2e-16 ***
log(rings)	-0.1543	0.0121	-12.7	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.184 on 4170 degrees of freedom

Multiple R-squared: 0.954, Adjusted R-squared: 0.954

F-statistic: 4.36e+04 on 2 and 4170 DF, p-value: <2e-16

Model 3:

```
model3 <- lm(log(rings) ~ sex + log(len) + sex:log(len), data = ab)
summary(model3)
```

Call:

```
lm(formula = log(rings) ~ sex + log(len) + sex:log(len), data = ab)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.6628	-0.1539	-0.0365	0.1148	0.9381

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.6290	0.0229	114.94	< 2e-16 ***
sexI	0.1057	0.0303	3.49	0.00049 ***
sexM	0.0601	0.0286	2.10	0.03562 *
log(len)	0.4538	0.0394	11.52	< 2e-16 ***
sexI:log(len)	0.3530	0.0447	7.89	3.9e-15 ***
sexM:log(len)	0.1407	0.0478	2.95	0.00325 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.228 on 4167 degrees of freedom
Multiple R-squared: 0.492, Adjusted R-squared: 0.492
F-statistic: 808 on 5 and 4167 DF, p-value: <2e-16

(a) How many abalones are in the data set?

4173

(b) In Model 1, what is the role of the interaction term entered as `sex:log(len)` in the `lm()` function?

The role of the interaction term is to determine how the slope of length is impacted by the sex of the abalones

- (c) Write a formula based on Model 1 for predicting the shucked weight of a ~~female~~ abalone based on its length and number of rings.

$$wt = e^{1.2265 \cdot len^{3.1126} \cdot rings^{-0.1732}}$$

$$\log wt = 1.2265 + 3.1126 \log(len) - 0.1732 \log(rings) \quad (!)$$

- (d) Based on Model 1, what increase in the log weight is associated with an increase of one in log length for infant abalones?

$$\text{slope} = 3.1126 + 0.0347 = 3.1473$$

$$\begin{array}{r} 3.1126 \\ 0.0347 \\ \hline 3.1473 \end{array}$$

- (e) Based on the Model 1 output, is there strong evidence to suggest that the effect of length on weight differs between male and female abalones? Explain your answer.

since $p < 0.05$ this suggests that there is strong evidence that the effect of length on weight differs between male and female. $\text{sex M: } \log(len) \rightarrow 0.1714$ $p \text{ value} = 7.1 \times 10^{-6}$ ✓

- (f) For Model 1 one obtains $SS_{\text{Error}} = 137.145$, while for Model 2 one obtains $SS_{\text{Error}} = 141.116$. Use these values in an expression for a test statistic for testing the significance of the sex variables and their interactions with the log length. You do not need to evaluate your expression.

$$F = \frac{(141.116 - 137.145) / (4170 - 4166)}{137.145 / 4166}$$

- (g) For the test statistic constructed in the previous part one obtains a p-value approximately equal to zero. What do we learn from this result? Explain carefully.

Reject $H_0 \rightarrow$ sex's interaction w length are significant predictors of weight.

- (h) Using the Model 3 output, give the variance inflation factor for the $\log(\text{rings})$ variable as a predictor in Model 1.

$$\frac{1}{1 - 0.492}$$

- (i) Explain the reason for looking at the residuals versus fitted values plot and give an interpretation of what you see in this plot for Model 1.

To see if there is linearity in the data. In Model 1, this is true because there are no curved trends. Also to check if there is constant variance which is also true for the same reason.

- (j) Explain the reason for looking at the normal quantile-quantile plot of the residuals and give an interpretation of what you see in the plot for Model 1.

Analyzing the normality of errors. Seems to be normal in Model 1 based on the linear slope of the line with some possible irregularities on right side.

3

The numbers read into R below are the weights in pounds of several wooden pallets delivered to a farm by a produce distributor.

```
wt <- c(41,34,40,44,33,42,52,38,32,31,31,35,  
        39,44,42,42,35,33,40,48,51,32,41,35,  
        38,48,37,35,42,41,40,47,40,46,33,38,  
        51,39,40)
```

The sample mean is $\bar{X}_n = 39.744$ and the sample standard deviation is $S_n = 5.72$. Suppose you want to know if the mean weight of all pallets belonging to the distributor is less than 40 pounds.

- (a) Write down the null and alternate hypotheses of interest.

$$H_0: \mu \geq \mu_0$$

$$H_1: \mu < \mu_0$$

- (b) Give an expression for the test statistic for testing these hypotheses, using the mean, standard deviation, and size of the sample above. You do not have to evaluate your expression.

$$\frac{39.744 - 40}{5.72 / \sqrt{39}}$$



- (c) The value of the test statistic turns out to be -0.28 . State whether you should reject the null hypothesis on the basis of this value. Explain your answer.

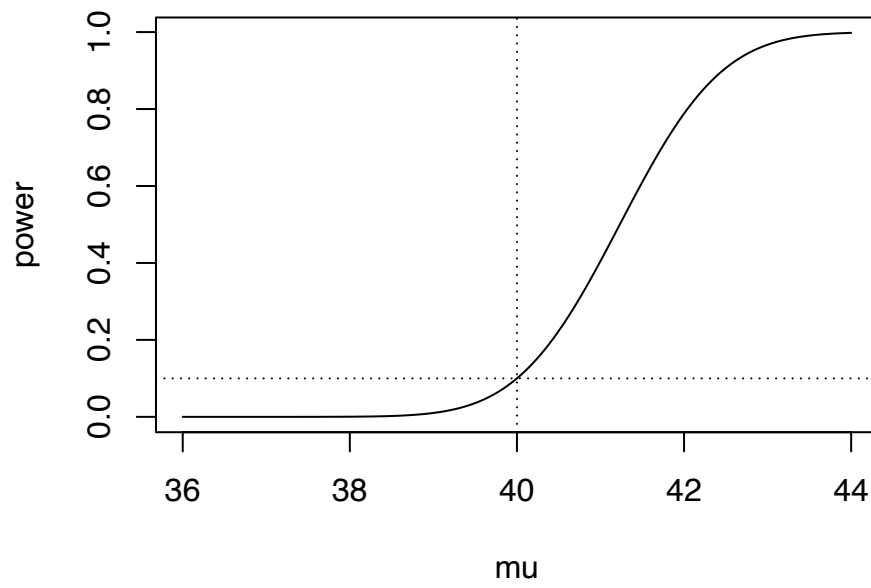
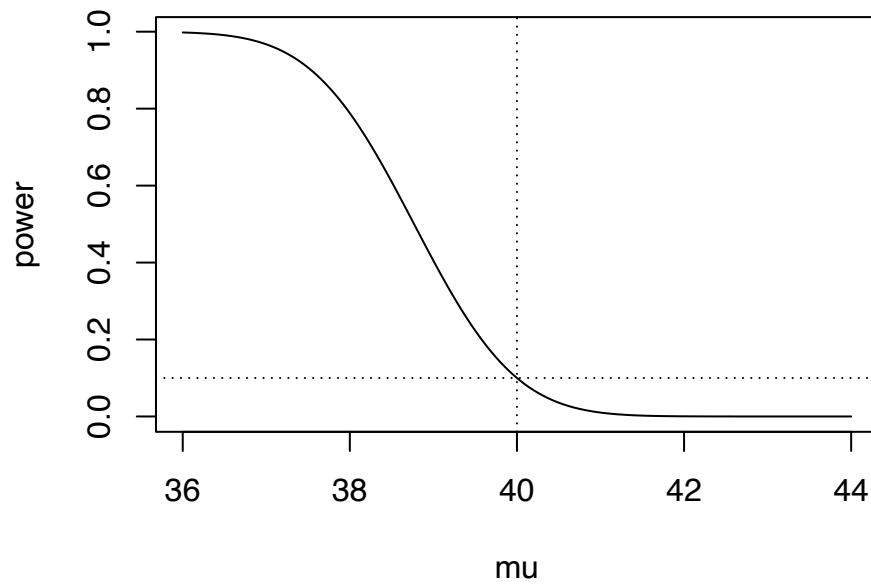
I would fail to reject my null hypothesis due to the small test value. While I can't say what the critical value is I know from experience it is further away than -0.28

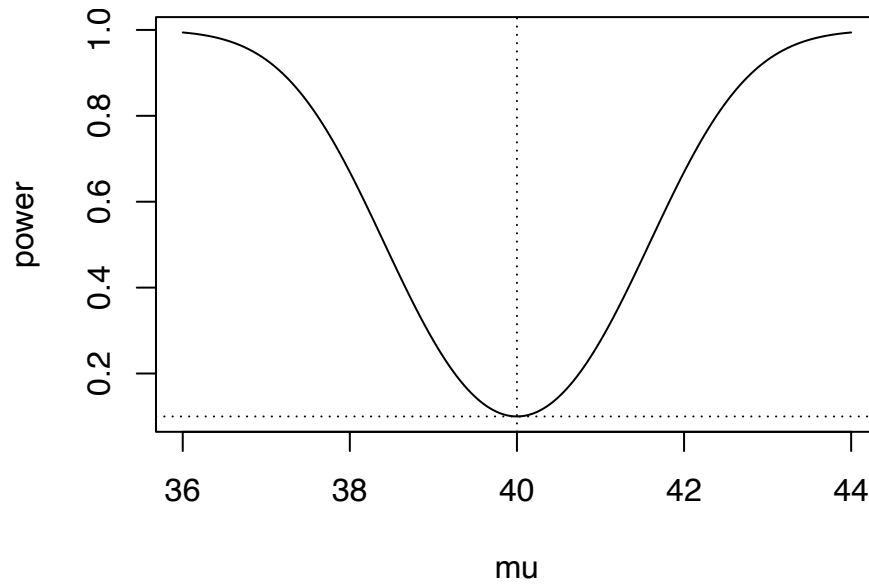
- (d) Of the three power curves shown below (on the following pages), only one has the correct shape to be the power curve for a test of the hypotheses from part a. According to this curve, report the approximate probability of rejecting the null hypothesis if the mean weight of all pallets owned by the distributor is equal to 39 pounds.

40%

- (e) According to the same curve, give the approximate probability of a Type II error if the mean weight of all the pallets owned by the distributor is equal to 38 pounds.

20%





Bibliography

Nansen, Fridtjof, and Otto Neumann Sverdrup. 1898. *Farthest North: Being the Record of a Voyage of Exploration of the Ship "Fram" 1893-96, and of a Fifteen Months' Sleigh Journey by Dr. Nansen and Lieut. Johansen*. Harper & Brothers Publishers.