Problem 1: Statistics - Advanced Statistics

Suppose that the random variable $X \sim N\left(\mu,\sigma^2\right)$

(a) Find the moment-generating function of X.

(b) Using this moment generating function verify that the mean of the random variable is μ and the variance is σ^2.

(c) Suppose that $Y = \dfrac{\left(X-\mu\right)^2}{\sigma^2}$. Verify that Y has a χ^2 distribution with 1 degree of freedom.

Solution: (a) By definition, we have

$$M_X\left(t\right) = E\left(e^{tX}\right) = \int_{-\infty}^{\infty} \frac{e^{tx}}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}+tx} dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2-2x\mu+\mu^2+2\sigma^2 tx}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2-2x\left(\mu+\sigma^2 t\right)+\mu^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2-2x\left(\mu+\sigma^2 t\right)+\left(\mu+\sigma^2 t\right)^2+\mu^2-\left(\mu+\sigma^2 t\right)^2}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(x-\mu-\sigma^2 t\right)+2\sigma^2\mu t+\sigma^4 t^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(x-\mu-\sigma^2 t\right)}{2\sigma^2}+\mu t+\frac{\sigma^2 t^2}{2}} dx$$

$$= e^{\mu t+\frac{\sigma^2 t}{2}}\left(\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(x-\mu-\sigma^2 t\right)}{2\sigma^2}}dx\right)=e^{\mu t+\frac{\sigma^2 t^2}{2}}\cdot 1=e^{\mu t+\frac{\sigma^2 t^2}{2}}$$

(b) Now, differentiating we get:

$$M_X^{'}\left(t\right)=\exp\left(\mu t+\frac{\sigma^2 t^2}{2}\right)\left(\mu+\sigma^2 t\right)$$

and hence,

$$\mu = M_X^{'}\left(0\right)=\exp\left(0\right)\left(\mu+0\right)=\mu$$

On the other hand,

$$M_X^{"}\left(t\right)=\exp\left(\mu t+\frac{\sigma^2 t^2}{2}\right)\left(\mu+\sigma^2 t\right)^2+\exp\left(\mu t+\frac{\sigma^2 t^2}{2}\right)\left(\sigma^2\right)$$

which means that

$$E\left(X^2\right)=M_X^{"}\left(t\right)=\exp\left(0\right)\left(\mu\right)^2+\exp\left(0\right)\left(\sigma^2\right)=\mu^2+\sigma^2$$

which implies that $\sigma^2 = E\left(X^2\right)-\mu^2=\operatorname{var}\left(X\right)$.

(c) In fact, since

$$\frac{\left(X-\mu\right)}{\sigma}\sim N\left(0,1\right)$$

then by definition, $Y = \frac{(X-\mu)^2}{\sigma^2} = \left(\frac{X-\mu}{\sigma}\right)^2$ has a χ^2 distribution with 1 degree of freedom.

Problem 2: Statistics - Analysis of Variance

A company sells 3 items: swimming pools, spas, saunas. Owner decides to see whether age of the sales rep and the type of item affect monthly sales. At a=0.05, analyze the data shown, using a two-way ANOVA. Sales are given in hundreds of dollars for a randomly selected month, and five salespeople were selected for each group.

Age of Salesperson	Product Pool	Spa	Sauna
Over 30	56, 23, 52, 28, 35	43, 25, 16, 27, 32	47, 43, 52, 61, 74
30 or under	16, 14, 18, 27, 31	58, 62, 68, 72, 83	15, 14, 22, 16, 27

Solution: We need to test for

$$H_0: \text{ the means are the same for all the products}$$

and for

$$H_0: \text{ the means are the same for all ages of the sales persons}$$

and

$$H_0: \text{ there's no interaction bewteen Product and Age of Sales Person}$$

The ANOVA summary table is shown below:

ANOVA summary table for #15.

Source	**SS**	**d.f.**	**MS**	***F***
Age	168.033	1	168.033	$\frac{168.033}{107.25} = 1.5667$
Product	1762.067	2	$\frac{1762.067}{2} = 881.0335$	$\frac{881.0335}{107.5} = 8.214765$
Interaction	7955.267	2	$\frac{7955.267}{2} = 3977.6335$	$\frac{3977.6335}{107.5} = 37.08749$
Within	2574.000	24	$\frac{2574}{24} = 107.25$	
Total	12,459.367	29		

The critical value for (1, 24) degrees of freedom and a 0.05 significance level is 4.2597.
The critical value for (2, 24) degrees of freedom and a 0.05 significance level is 3.4028.
From this we conclude that the both Product and the interaction between Product and Age affect significantly the means, at the 0.05 significance level.

Problem 3: Statistics - Analysis of Variance

A researcher at an accounting firm wants to out whether the current ratio for three industries is about the same. Random samples of eight firms in industry A, six firms in industry B, and six firms in industry C are available. The ratios are as follows:

Industry A: 1.38, 1.55, 1.90, 2.00, 1.22, 2.11, 1.98, 1.61

Industry B: 2.33, 2.50, 2.79, 3.01, 1.99, 2.45

Industry C: 1.06, 1.37, 1.09, 1.65, 1.44, 1.11

Conduct the test at α = 0.05, and state your conclusion.

Solution: We are comparing the population mean ratio for three different industries. Our hypotheses are. Let's call μ_A, μ_B, μ_C, to those mean ratios, our hypotheses are

$$H_0: \text{the current ratio for three industries is not signicantly different}$$
$$H_1: \text{the current ratio for three industries is signicantly different}$$

We use ANOVA (analysis of variance) to test this hypothesis. One of the assumptions for ANOVA is that the population variances are the same. We'll use an F-test to compare all thee pairs of variances, obtaining that

F-Test Two-Sample for Variances

	A	*B*
Mean	1.71875	2.511667
Variance	0.105298	0.126977
Observations	8	6
df	7	5
F	0.829272	
P(F<=f) one-tail	0.395638	
F Critical one-tail	0.251792	

F-Test Two-Sample for Variances

	B	*C*
Mean	2.511666667	1.286667
Variance	0.126976667	0.056747
Observations	6	6
df	5	5
F	2.237605733	

P(F<=f) one-tail	0.198705266
F Critical one-tail	5.050338814

F-Test Two-Sample for Variances

	A	*C*
Mean	1.71875	1.286667
Variance	0.105298	0.056747
Observations	8	6
df	7	5
F	1.855584	
P(F<=f) one-tail	0.2571	
F Critical one-tail	4.875858	

All the p-values are greater than 0.05, the level of significance, and therefore none of them is significant, which means that *we can consider all the population variances to be equal*. Now we can safely apply the ANOVA procedure. Using EXCEL we get

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F critical*
Between Groups	4.658116	2	2.329057917	23.913683	1.14517E-05	3.591538
Within Groups	1.655704	17	0.097394363			
Total	6.31382	19				

Here we observe that the p-value is significant (because it's less than 0.05) which means that we reject the null hypothesis, and we accept the alternative. Therefore, *the current ratio for three industries cannot be considered the same, for the given level of significance*.

Problem 4: Statistics - Confidence Intervals

Find the margin of error, E, given that the college students annual earnings has a mean, $p = \$3967$, and sample size n = 74 and a standard deviation, a = \$874 and the degree of confidence is 95%.

Solution: We know that

$$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{1.96 \times 874}{\sqrt{74}} = 199.13686$$

Problem 5: Statistics - Confidence Intervals

The mean time before a certain headache remedy starts to work is 12 minutes with a standard deviation of 3. A new coating is used to help swallow the pill. A sample of 38 people using the pills with the new coating showed the mean time before it started to work was 14 minutes. Is there a difference with the new coating?

Test at the .01 level. State the hypotheses and identify the claim, find the critical value(s), compute the test value, make the decision, and summarize the results.

Solution: We have to test the following hypotheses:

$$H_0 : \mu = 12$$
$$H_A : \mu \neq 12 \ \text{(claim)}$$

We use a two tailed test z-test (the population variance is known). The critical values for this test are

$$z_{\alpha/2} = \pm 2.575$$

The rejection region is given by

$$R = \{z \in \mathbb{R} : \ |z| > 2.575\}$$

We compute now the z-statistics:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{14 - 12}{3 / \sqrt{38}} = 4.109609$$

This means that we reject the null hypothesis. *In other words, we have enough evidence to support the claim that there's a difference with the new coating*, at the 0.01 significance level.

Problem 6: Statistics - Confidence Intervals

An employment counselor found that in a sample of 100 unemployed workers, 65% were not interested in returning to work. Find the 95% confidence interval of the true proportion of workers who do not wish to return to work.

Solution: For the true proportion, the 95% confidence interval is given by

$$CI = \left(\hat{p} - 1.96 \times \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \ \hat{p} + 1.96 \times \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right) = (0.556514, \ 0.743486)$$

Problem 7: Statistics - Confidence Intervals

Sales of a new line of athletic footwear are crucial to the success of a newly formed company. The company wishes to estimate the avg. weekly sales of the new footwear to within \$150 with 98% reliability. The initial sales indicate the standard deviation of the weekly sales figures are approx. \$1525. How many weeks of data must be sampled for the company to get the information it desires?

Solution: The 98% Margin of error is equal to

$$MOE = 2.33 \times \frac{1525}{\sqrt{n}}$$

We need to estimate the avg. weekly sales of the new footwear to within $150 with 98%, which means

$$MOE = 2.33 \times \frac{1525}{\sqrt{n}} \le 150 \;\Leftrightarrow\; \left(\frac{2.33 \times 1525}{150}\right)^2 \le n \;\Leftrightarrow\; n \ge 561.1371$$

Problem 8: Statistics - Confidence Intervals

A company wants to market a new product to their present customers. They survey 100 potential customers and find that 21 would buy the new product. If you presently have 38,232 customers, determine the actual number of customers that would purchase the new product with a 95% level of confidence.

Solution: Let X be the number of customers. We know that a 95% confidence interval is given by

$$CI = \left(\mu - z_{\alpha/2} \times \sigma,\; \mu + z_{\alpha/2} \times \sigma\right)$$

where $\mu = Np = 38232 \times 0.21 = 8028.72$,

$$\sigma = \sqrt{Np(1-p)} = \sqrt{38232 \times 0.21 \times 0.79} = 79.641$$

and $z_{\alpha/2} = 1.96$. That means that the confidence interval is

$$CI = \left(\mu - z_{\alpha/2} \times \sigma,\; \mu + z_{\alpha/2} \times \sigma\right) = (7862.624,\; 8184.816)$$

Problem 9: Statistics - Control Charts

The Highway Patrol has a target of 28 traffic tickets per week with standard deviation 5 tickets per week for a stretch of mountain highway. The number of tickets issued for 15 consecutive weeks is given below. t = number of week, x = number of tickets

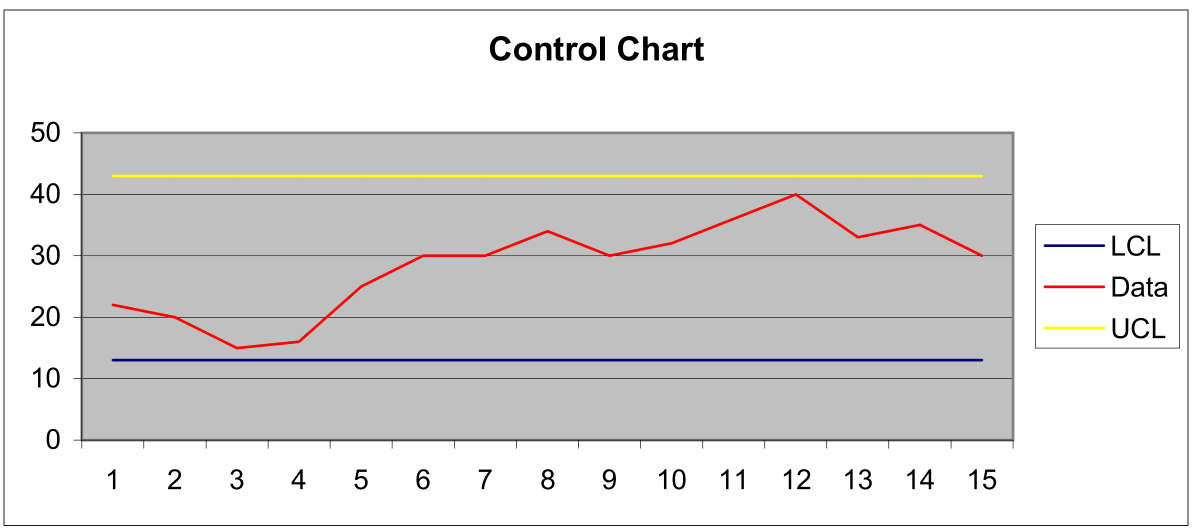
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x	22	20	15	16	25	30	30	34	30	32	36	40	33	35	30

a) Make a control chart for the above data.
b) Determine whether the process is in statistical control. If it is not, specify which out-of-control signals are present.

Solution:

a) The lower and upper limits are given by $\mu \pm 3\sigma = 28 \pm 15$. Then

$$LCL = 13,\ \text{UCL=43}$$



b) The process is in statistical control because the data don't go beyond the lower and upper limit.

Problem 10: Statistics - Correlation and Regression

If, after the formulation of a least squares regression model, the 95% confidence interval for β_1 is [-1.43; 2.62], what should be true about the overall F-test for a regression relationship? Explain.

Solution: The overall F-test is used to test the hypothesis

$$H_0 : \beta_1 = 0$$

Since [-1.43; 2.62] contains β_1 with 95% confidence, that means that we cannot reject the null hypothesis, which means that the F-statistic has to be less than the critical value.

Problem 11: Statistics - Correlation and Regression

A breeder of thoroughbred horses wishes to model the relationship between the gestation period and the length of life of a horse. The breeder believes that the two variables may follow a linear trend. The information in the table was supplied to the breeder form various thoroughbred stables across the state

Horse	Gestation period x (days)	Life Length y (years)	Horse	Gestation period x (days)	Life Length y (years)
1	416	24	5	356	22
2	279	25.5	6	403	23.5
3	298	20	7	265	21
4	307	21.5			

Summary statistics yield SS_{xx} = 21,752, SS_{xy} = 236.5, SS_{yy} = 22, x-bar = 332, and y-bar =22.5. Calculate SSE, s^2, and s.

Solution: We first compute $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{236.5}{21,752} = 0.010873$$

We now compute SSE :

$$SSE = SS_T - \hat{\beta}_1 SS_{XY} = \sum y_i^2 - n\overline{y}^2 - 0.010873 \times 236.5 = 3565.75 - 7 \times 22.5^2 - 0.010873 \times 236.5$$

$$= 19.42864$$

Finally, we compute

$$s^2 = \frac{SSE}{n-2} = \frac{19.42864}{5} = 3.885728$$

which means that

$$s = 1.971225.$$

Problem 12: Statistics - Correlation and Regression

The owner of a Ford dealer wants to study the relationship between the age of a car and its selling price. Listed below, there is a random sample of 12 used cars sold last year:

Car	Age	Price
1	9	8.1
2	7	6
3	11	3.6
4	12	4
5	8	5
6	7	10
7	8	7.6
8	11	8
9	10	8

10	12	6
11	6	8.6
12	6	8

(a) If we want to estimate the selling price on the basis of the age of a car, which variable is the dependent variable and which is the independent variable?

(b) Draw a scatter diagram

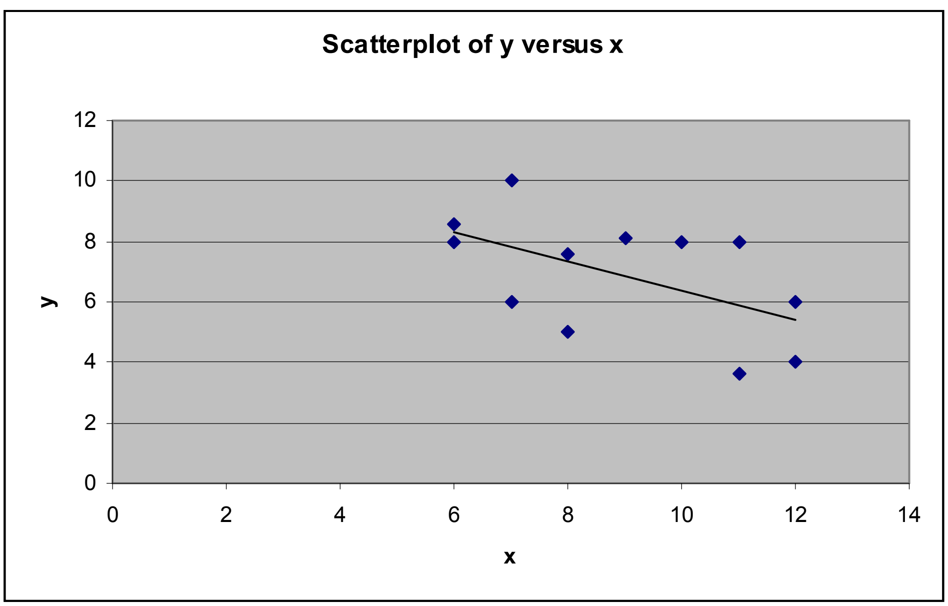
(c) Compute the coefficient of correlation

(d) Compute the coefficient of determination

(e) Interpret.

Solution: (a) The dependent variable is selling price, and the independent variable is age.

(b) First, we have the following scatterplot of the data:



(c) The following table is obtained:

	x	y	x · y	x²	y²
	9	8.1	72.9	81	65.61
	7	6	42	49	36
	11	3.6	39.6	121	12.96
	12	4	48	144	16
	8	5	40	64	25
	7	10	70	49	100
	8	7.6	60.8	64	57.76
	11	8	88	121	64
	10	8	80	100	64
	12	6	72	144	36
	6	8.6	51.6	36	73.96
	6	8	48	36	64
Sum	**107**	**82.9**	**712.9**	**1009**	**615.29**

In order to run a regression analysis, we need first to compute the correlation coefficient, in order to determine whether or not it makes sense to use a linear model.

Based on the results from the table above, we get the following results for the correlation coefficient:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}} = \frac{12\times 712.9 - 107\times 82.9}{\sqrt{12\times 1009 - 107^2}\sqrt{12\times 615.29 - 82.9^2}}$$

$$= -0.5436$$

(d) The coefficient of determination is

$$r^2 = -0.5436^2 = 0.2956$$

which means that 29.56% of the variation in y is explained by x.

(e) It is not surprising, because by common sense we know that, on average, the older a car is, the cheaper the price.

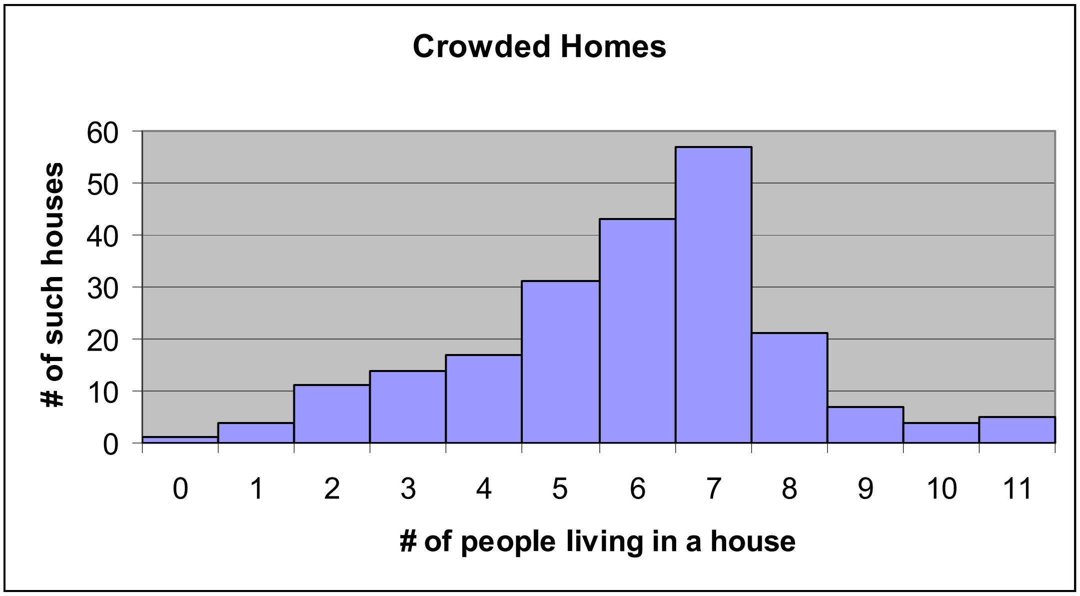
Problem 13: Statistics - Descriptive Statistics

What is the purpose of a survey? What are three examples of a survey and what potential purpose could each be used for?

Solution: Generally speaking, a survey corresponds to a way to gather information about a certain population based on the information provided by a reduced amount of individuals from that population. Surveys can be use to forecast the winner of the presidential election 2 weeks before the actual election, or to predict the success of a new product based on a marketing strategy, or determine what percentage of people is watching the TV show "Deal or not deal", based on a phone survey.

Problem 14: Statistics - Descriptive Statistics

The following display shows the number of people officially living in 215 randomly selected housing units in a densely populated area.



Use the above picture to answer the following questions. You may give approximate answers.

a) In this sample, what percentage of the houses, have less than 5 people living in it?

b) In this sample, what percentage of the houses, have more than 7 people living in it?

Solution: (a) The percentage is

$$p = \frac{(17+13+11+4+2)}{215} \times 100 = 21.9\%$$

(b) The percentage is

$$p = \frac{(21+7+5+6)}{215} \times 100 = 18.1\%$$

Problem 15: Statistics - Descriptive Statistics

Using the following values for X: $X = 9, 1, 1, 1$

Find the value of the following expressions:

a. ΣX^2
b. $(\Sigma X)^2$
c. $\Sigma(X - 3)$
d. $\Sigma(X-3)^3$

Solution: We have the following table:

X	X^2	X-3	$(X-3)^3$
9	81	6	216
1	1	-2	-8
1	1	-2	-8
1	1	-2	-8
12	84	0	192

which means that $\sum X^2 = 84$, $\left(\sum X\right)^2 = 12^2 = 144$, $\sum(X-3) = 0$ and $\sum(X-3)^3 = 192$

Problem 16: Statistics - Hypothesis Testing

2. Judy Povich is a fashion designer artist who designes the display windows in front of a large clothing store in New York City. Electronic counters at the entrances total the number of people entering the store each business day. Before Judy was hired by the store, the mean number of people entering the store each day was 3218. However, since Judy has started working, it is thought that this

number increased. A random sample of 42 business days after Judy began work gave an average $\overline{X}$ = 3392 people entering the store each day. The sample standard deviation was s = 287 people. Does this indicate that the average number of people entering the store each day has increased? Use a 1% level of significance.

(a) What is the null hypothesis? What is the alternate hypothesis? Will we use a left-tailed, right-tailed, or two-tailed test? What is the level of significance?

(b) What sampling distribution will we use? What is the critical value zo (or critical values $\pm z$o)

(c) Sketch the critical region and show the critical value (or critical values).

(d) Calculate the z value corresponding to the sample statistic $\overline{X}$ and show it's location on the sketch of part (c)

(e) Based on your answers for parts (a) to (d), shall we reject or fail to reject (i.e., "accept") the null hypothesis at the given level of significance α ? Explain your conclusion in the context of the problem.

(f) Are the data statistically significant?

Solution: (a) The hypotheses are

$$H_0 : \mu = 3218$$
$$H_A : \mu > 3218$$

This corresponds to a right-tailed rejection region. The level of significance is $\alpha = 0.01$.

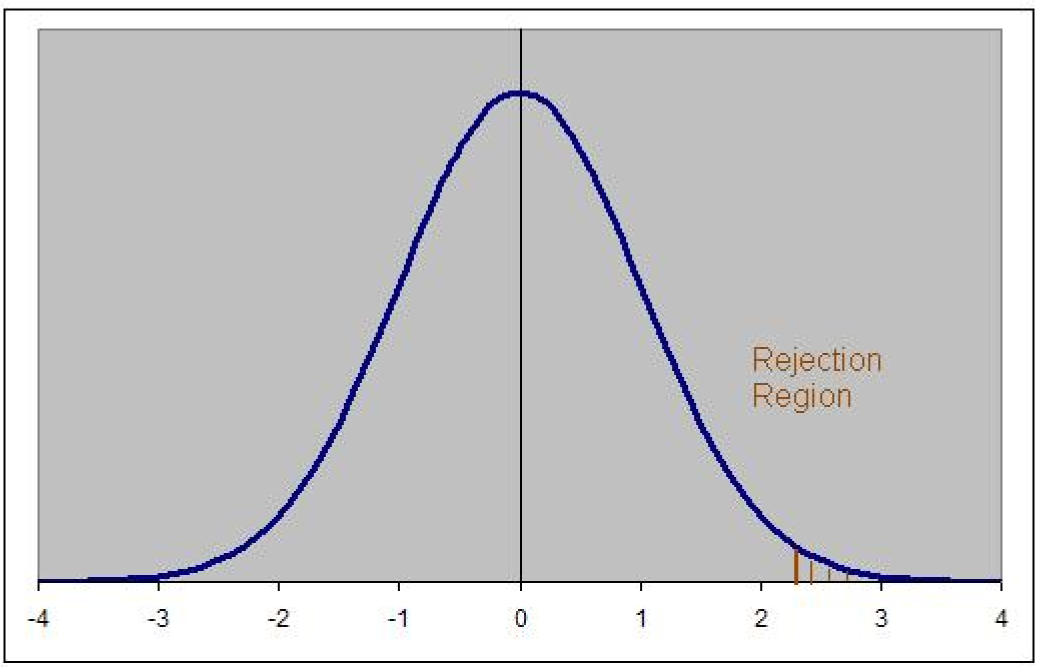
(b) We have that

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

has approximately a standard normal distribution. The critical value corresponds to

$$z_0 = 2.32635$$

(c) We have the following graph

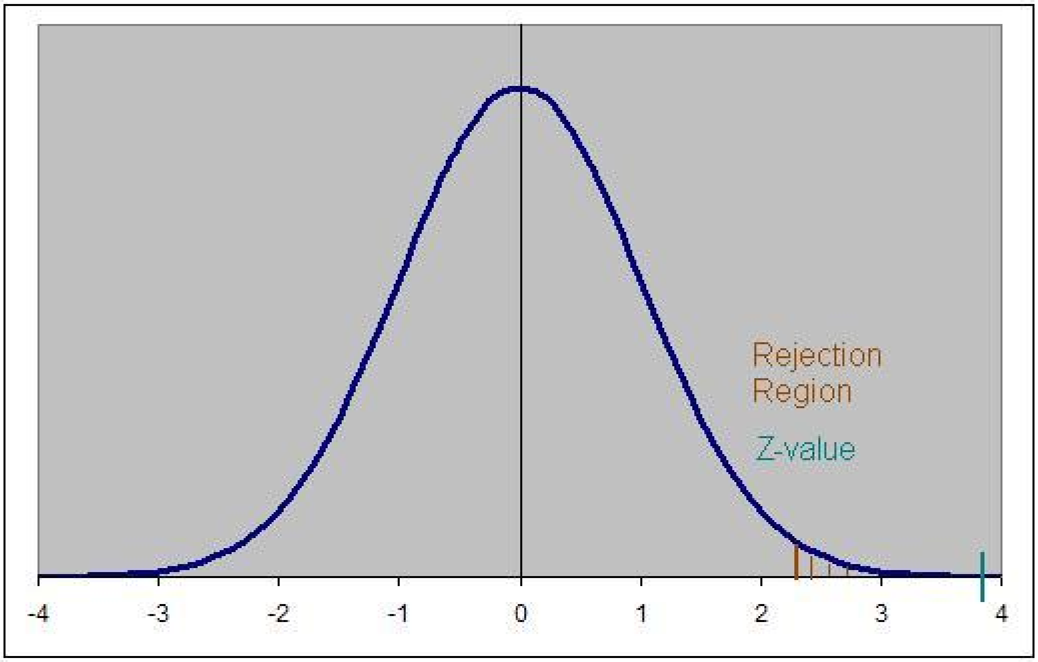


(d) The Z-value is computed as

$$Z = \frac{3392 - 3218}{287 / \sqrt{42}} = 3.92909$$

Free Statistics Help

StatisticsHelp.org



(e) We reject the null hypothesis because the Z-score is greater than the critical value z_0. This means that we have enough evidence to claim that the new designer draws more people to the store, with a significance of $\alpha = 0.01$

(f) Yes.

Problem 17: Statistics - Hypothesis Testing

The management of White Industries is considering a new method of assembling its golf cart. The present method requires 42.3 minutes, on the average, to assemble a cart. The mean assembly time for a random sample of 24 carts, using the new method, was 40.6 minutes, and the standard deviation of the sample was 2.7 minutes. Using the .10 level of significance, can we conclude that the assembly time using the new method is faster?

Solution: We need to test the following hypotheses

$$H_0 : \mu = 42.3$$
$$H_A : \mu < 42.3$$

Since the population variance is unknown, and the sample size is not big enough, we use a left-tailed t-test. The t-statistics is computed as

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}} = \frac{40.6 - 42.3}{2.7 / \sqrt{24}} = -3.08454$$

The critical t-value for this left tailed test for 23 degrees of freedom and $\alpha = 0.10$ is -1.31946. Since the t-statistics is less than the t-critical, we reject the null hypothesis. *In other words, we have enough evidence to claim that the new method is faster.*

The p-value is computed as

$$p = \Pr\left(t < -3.8454\right) = 0.002618$$