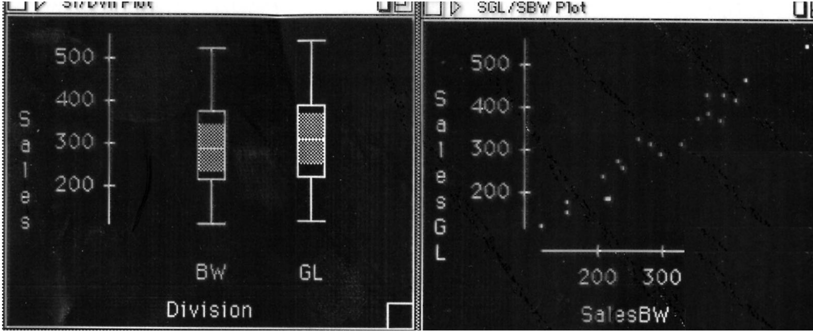
Math 1710 Class 36

Chi Square
Dr. Back

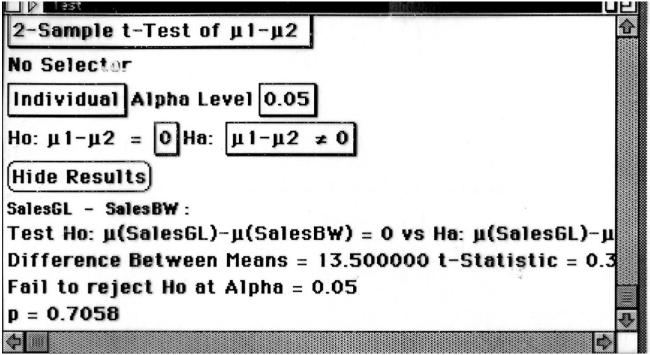
Nov. 20, 2009

MP as more Sensitive

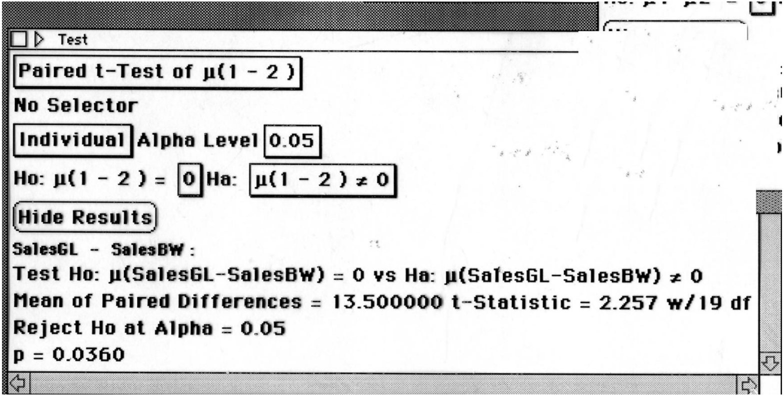
Sales figures for two car brands by district.



MP as more Sensitive



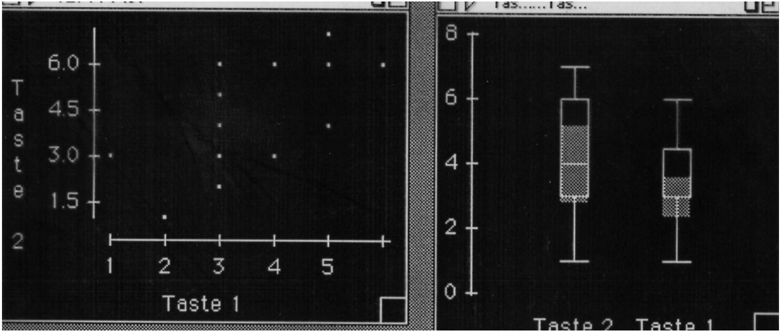
MP as more Sensitive



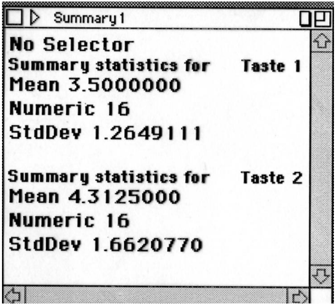
MP as more Sensitive

Taste Test Results on 2 teas:
Each of 16 people tasted both teas and gave them a rating form 1 to 7, 1 being best.

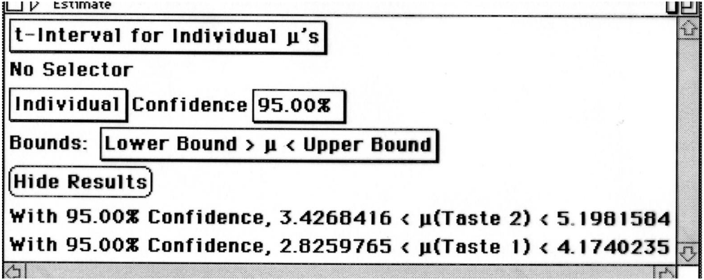
MP as more Sensitive



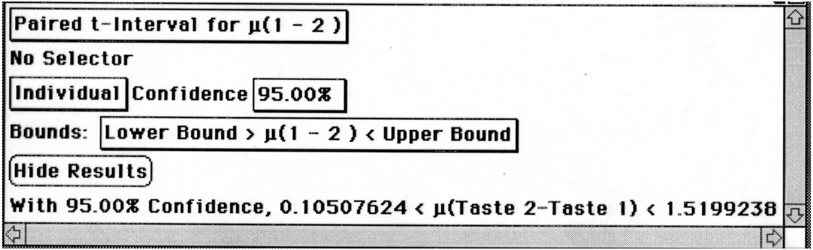
MP as more Sensitive



MP as more Sensitive



MP as more Sensitive



Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

$n < 15$	unimodal and symmetric
$15 <= n <= 40$	avoid strong skewness and outliers (unimodal and sym best)
$n > 40$	pretty much ok

Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

$n < 15$ unimodal and symmetric

$15 <= n <= 40$ avoid strong skewness and outliers (unimodal and sym best)

$n > 40$ pretty much ok

What Happens if Not Satisfied:

random sampling -
could be critical;
might be ok if "representative"
representative hard/impossible to define

Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

$n < 15$ unimodal and symmetric

$15 <= n <= 40$ avoid strong skewness and outliers (unimodal and sym best)

$n > 40$ pretty much ok

What Happens if Not Satisfied:

plausible independence -
could be critical
sometimes just a working hypothesis

Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

$n < 15$ unimodal and symmetric
$15 <= n <= 40$ avoid strong skewness and outliers (unimodal and sym best)
$n > 40$ pretty much ok

What Happens if Not Satisfied:

10% condition -
results in overestimation of samp. dist. st dev
gradual breakdown in formulas, not method

Conditions for t Tests

- plausible independence
- random sampling
- 10% condition
- nearly normal condition

$n < 15$ unimodal and symmetric
$15 <= n <= 40$ avoid strong skewness and outliers
(unimodal and sym best)
$n > 40$ pretty much ok

What Happens if Not Satisfied:

nearly normal -
no guarantee
progressive reduction of accuracy

Conditions for t Tests

For 2-sample inference, we add the

independence groups assumption.

The chance of an individual in one of the groups assuming a certain value should be independent of the values assumed by any of the individuals in the other group.

Robustness

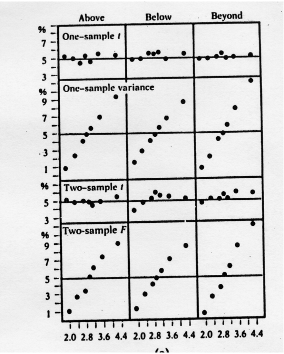
Formulas for t-inference and regression inference are based on assumptions of normality of the data.
Yet most distributions are not normal. (Although the CLT makes averages of lots normal.)
So it is perhaps remarkable that t-inference methods for moderate size data sets without outliers are typically pretty good. Why is this?

Robustness

Formulas for t-inference and regression inference are based on assumptions of normality of the data.
Yet most distributions are not normal. (Although the CLT makes averages of lots normal.)
So it is perhaps remarkable that t-inference methods for moderate size data sets without outliers are typically pretty good. Why is this?
One answer is robustness against non-normality.

Robustness

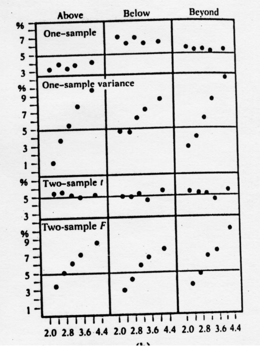
Actual confidence level vs kurtosis



for some symmetric $n = 25$ test distributions in a 1975 Biometrika paper of Pearson and Please. (kurtosis=3 in the normal case.)

Robustness

Actual confidence level vs kurtosis



for some skewed $n = 25$ test distributions in a 1975 Biometrika paper of Pearson and Please.

Chi Square Test of Independence

Flu Shots (categorical) vs Getting the Flu

	0-shot	1-shot	2-shot
flu	24	9	13
no flu	289	100	565

Chi Square Test of Independence

Hypotheses:

- H_0: The categorical variables FluShot? and GetTheFlu? are independent.
- H_a: The categorical variables are not independent. *(i.e. there is an association.)*

Chi Square Test of Independence

With Marginals

	0-shot	1-shot	2-shot	
flu	24	9	13	46
no flu	289	100	565	954
	313	109	578	1000

Chi Square Test of Independence

With Marginals

	0-shot	1-shot	2-shot	
flu	24	9	13	46
no flu	289	100	565	954
	313	109	578	1000

If independent, the expected number of 0-shot people getting the flu would be

$$\text{overall fraction getting flu} \times \text{number 0-shot people}$$

$$= \frac{46}{1000} \times 313 = 14.4$$

Chi Square Test of Independence

With Marginals

	0-shot	1-shot	2-shot	
flu	24	9	13	46
no flu	289	100	565	954
	313	109	578	1000

If independent, the expected number of 0-shot people getting the flu would be

$$\text{overall fraction getting flu} \times \text{number 0-shot people}$$

$$= \frac{46}{1000} \times 313 = 14.4$$

Chi Square Test of Independence

If independent, the expected number of 0-shot people getting the flu would also be

$$\text{overall fraction getting 0-shots} \times \text{number getting flu}$$

$$= \frac{313}{1000} \times 46 = 14.4$$

Chi Square Test of Independence

If independent, the expected number of 0-shot people getting the flu would be

$$\text{overall fraction getting flu} \times \text{number 0-shot people} = \frac{46}{1000} \times 313 = 14.4$$

This comes down to

$$\text{expected value} = \frac{\text{row sum} \times \text{col sum}}{total}$$

Chi Square Test of Independence

With Expected Values

	0-shot	1-shot	2-shot	
flu	24 (14.4)	9 (5.0)	13 (26.6)	46
no flu	289 (298.6)	100 (104)	565 (551.4)	954
	313	109	578	1000

Chi Square Test of Independence

With Expected Values

	0-shot	1-shot	2-shot	
flu	24 (14.4)	9 (5.0)	13 (26.6)	46
no flu	289 (298.6)	100 (104)	565 (551.4)	954
	313	109	578	1000

$$\chi^2 = \Sigma \frac{(Obs - Exp)^2}{Exp}$$

Chi Square Test of Independence

With Expected Values

	0-shot	1-shot	2-shot	
flu	24 (14.4)	9 (5.0)	13 (26.6)	46
no flu	289 (298.6)	100 (104)	565 (551.4)	954
	313	109	578	1000

$$\chi^2 = \Sigma \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots (5 \text{ more terms})$$

Chi Square Test of Independence

With Expected Values

	0-shot	1-shot	2-shot	
flu	24 (14.4)	9 (5.0)	13 (26.6)	46
no flu	289 (298.6)	100 (104)	565 (551.4)	954
	313	109	578	1000

$$\chi^2 = \Sigma \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots = 17.35$$

Chi Square Test of Independence

With Expected Values

	0-shot	1-shot	2-shot	
flu	24 (14.4)	9 (5.0)	13 (26.6)	46
no flu	289 (298.6)	100 (104)	565 (551.4)	954
	313	109	578	1000

$$\chi^2 = \Sigma \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots = 17.35$$

χ square tests also have a degrees of freedom.
For a test of independence, the number of rows (nr) and the number of columns (nc) in the original contingency table are the numbers of categories of the resp. cat. vars.

$$df = (nr - 1)(nc - 1)$$

Chi Square Test of Independence

With Expected Values

	0-shot	1-shot	2-shot	
flu	24 (14.4)	9 (5.0)	13 (26.6)	46
no flu	289 (298.6)	100 (104)	565 (551.4)	954
	313	109	578	1000

$$\chi^2 = \Sigma \frac{(Obs - Exp)^2}{Exp} = \frac{(24 - 14.4)^2}{14.4} + \ldots = 17.35$$

$$df = (nr - 1)(nc - 1)$$

Here

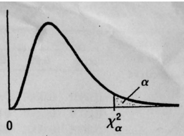
$$df = (2 - 1)(3 - 1) = 2.$$

Using Table χ

Right tail probability

Table X

Values of χ^2_α



df	0.10	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
...					
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.143	32.852	36.191	38.582

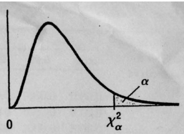
These are all critical values χ^*.

Using Table χ

Right tail probability

Table X

Values of χ^2_α

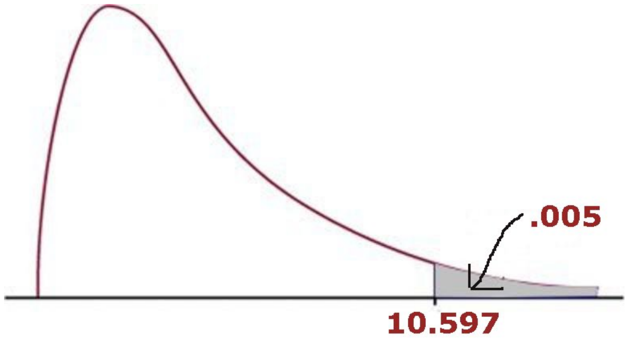


df	0.10	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
...					
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.143	32.852	36.191	38.582

These are all critical values χ^*.
For example $P(\chi > 5.991) = .05$ for the χ square distribution with 2 df.

Using Table χ

Our χ square-statistic of 17.35 is more extreme than any on the df=2 row of the table.
The picture



shows what the critical value $\chi^* = 10.597$ for a tail prob. of .005 means.

Using Table χ

So our tail probability and P-value are both less than .005 and we reject the null. Flu shots are making a difference.

Using Table χ

What we just did is actually a hypothesis test.

Using Table χ

Hypotheses:

- H_0: The categorical variables FluShot? and GetTheFlu? are independent.
- H_a: The categorical variables are not independent. *(i.e. there is an association.)*

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.

The **Chi Square distribution with** d **degrees of freedom** is given by:

$$\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2$$

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.

The **Chi Square distribution with d degrees of freedom** is given by:

$$\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2$$

Chi Square Distribution Formula

$$f(x) = \frac{1}{\Gamma(\frac{d}{2})2^{\frac{d}{2}}} x^{\frac{d}{2}-1} e^{\frac{-x}{2}}$$

where d is the number of degrees of freedom.

Chi Square Distribution as an RV

Let $X_1, \ldots X_d$ be d independent standard normal random variables.
The **Chi Square distribution with** d **degrees of freedom** is given by:

$$\chi^2 = X_1^2 + X_2^2 + \ldots X_d^2$$

Chi Square Distribution Formula

$$f(x) = \frac{1}{\Gamma(\frac{d}{2})2^{\frac{d}{2}}} x^{\frac{d}{2}-1} e^{\frac{-x}{2}}$$

where d is the number of degrees of freedom.
The mean is d and the variance is $2d$.

Chi Square Hypotheses

Test of Independence: Two Categorical Variables.

- H_0: The two categorical variables are independent.
- H_a: There is an association between the variables.

Chi Square Hypotheses

Test of Homogeneity: Multiple samples/populations. and Another Categorical Variable.

- H_0: The categorical variable has the same distribution within each population.
- H_a: The distributions differ among some of the populations.

(Which Population? could be viewed as a cat. var.)

Chi Square Hypotheses

Same Chi Square statistic,
Degrees of Freedom = (numrows-1)(numcols-1) for both indep. and homog.

Chi Square Hypotheses

Same Chi Square statistic,
Degrees of Freedom = (numrows-1)(numcols-1) for both indep. and homog.
The concept of "what is a population" does not have a clear answer, so we will not require you to distinguish between indep. and homog. on exams or homework.
(You need to know lots about the design to truly tell these apart.)

Chi Square Hypotheses

The concept of "what is a population" does not have a clear answer, so we will not require you to distinguish between indep. and homog. on exams or homework.
(You need to know lots about the design to truly tell these apart.)
In fact indep. and homog. have different exact mmathematical models, but both are approximated by the same χ^2.

Chi Square Hypotheses

Goodness of Fit: Frequencies of One Cat Var
AND a hypothesized distribution.

- H_0: The cat. var. follows the hypothesized distribution.
- H_a: The cat. var. doesn't.

Degrees of Freedom =number of cells -1.

Goodness of Fit Example

Suppose a maze has 3 exits. 90 rats run the maze and choose the following exits:

	door 1	door 2	door 3
frequency	23	36	31

Goodness of Fit Example

Suppose a maze has 3 exits. 90 rats run the maze and choose the following exits:

	door 1	door 2	door 3
frequency	23	36	31

The null hypothesis is that all doors are equally likely. Is this data strong evidence that not all doors are equally likely?

Goodness of Fit Example

	door 1	door 2	door 3
frequency	23	36	31

In this test, "equally likely" could be replaced by any hypothesized distribution of interest; e.g. there might be a theory to be evaluated that 38% exit door 1, 42% exit door 2, and 20% exit door 3.

Goodness of Fit Example

	door 1	door 2	door 3
frequency	23	36	31

In this test, expected values come from the hypothesized distribution, NOT $\frac{rowsum \cdot colsum}{total}$.

Goodness of Fit Example

(The row sum 90 is the total number of rats.)
The expected values are all $\frac{1}{3} \cdot 90$.

	door 1	door 2	door 3	
frequency	23	36	31	90
expected values	(30)	(30)	(30)	

Goodness of Fit Example

(The row sum 90 is the total number of rats.)
The expected values are all $\frac{1}{3} \cdot 90$.

	door 1	door 2	door 3	
frequency	23	36	31	90
expected values	(30)	(30)	(30)	

The Chi Square statistic is again

$$\chi^2 = \Sigma \frac{(Obs - Exp)^2}{Exp}$$

but the degrees of freedom are *numcells* $-$ 1.

Goodness of Fit Example

(The row sum 90 is the total number of rats.)
The expected values are all $\frac{1}{3} \cdot 90$.

	door 1	door 2	door 3	
frequency	23	36	31	90
expected values	(30)	(30)	(30)	

Here $df = 3 - 1 = 2$ and

$$\chi^2 = \frac{(23-30)^2}{30} + \frac{(36-30)^2}{30} + \frac{(31-30)^2}{30} = \frac{86}{30} = 2.87.$$

Goodness of Fit Example

(The row sum 90 is the total number of rats.)
The expected values are all $\frac{1}{3} \cdot 90$.

	door 1	door 2	door 3	
frequency	23	36	31	90
expected values	(30)	(30)	(30)	

Our P-value is greater than .1. We retain the null. All exits might be equally likely.

Why $\frac{(Obs - Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment.
Each cell corresponds to a particular value of each of the two categorical variables.

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment.

Notation:

- n The total number of observed subjects.
- $\hat{p}$ The observed proportion in the cell of interest.
- p The true proportion in the cell of interest.

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment. Then

Observed $n\hat{p}$.

Expected np.

$$\frac{(Obs-Exp)^2}{Exp} = \frac{(n\hat{p}-np)^2}{np} = \left(\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}\right)^2 q$$

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment. Then

Observed $n\hat{p}$.

Expected np.

$$\frac{(Obs - Exp)^2}{Exp} = \frac{(n\hat{p} - np)^2}{np} = \left(\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}\right)^2 q$$

Thus except for the extra q (often near 1), we have the square of a Z-statistic!

Why $\frac{(Obs-Exp)^2}{Exp}$?

Focus on 1 cell in our contingency table for a moment.
Then

Observed $n\hat{p}$.

Expected np.

$$\frac{(Obs-Exp)^2}{Exp} = \frac{(n\hat{p}-np)^2}{np} = \left(\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}\right)^2 q$$

Thus except for the extra q (often near 1), we have the square of a Z-statistic!

The degrees of freedom being less than the number of cells may be viewed as the result of careful analysis of all those extra q factors.

2×2 Tables

Degrees of Freedom

Why df=(numrows-1)(numcols-1) for test of independence? (a 3 by 4 example: df=2*3=6)

Degrees of Freedom

Why df=(numrows-1)(numcols-1) for test of independence? (a 3 by 4 example: df=2*3=6)

Imagine these marginals . . .

known	known	known		300
known	known	known		300
				300
100	200	300	400	1000

Degrees of Freedom

Idea: Consider all 3 by 4 contingency tables with the same marginal totals. These determine the expected values. How well the observed values match the expected tests the null hypothesis of independence.

Degrees of Freedom

The idea is that if we specify the observed values in the 6 known cells, then the rest of the cells are determined by the marginal totals. In other words df=6 corresponds to 6 cells can vary freely, the rest are determined.

Degrees of Freedom

Suppose we know the values in the 2×3 upper left:

40	80	100		300
30	110	150		300
				300
100	200	300	400	1000

Degrees of Freedom

Then the remaining values are determined by subtraction

40	80	100	80	300
30	110	150	10	300
30	10	50	310	400
100	200	300	400	1000

C

Regression Inference Questions

Basic Setup:

1) Data (x_i, y_i), $1 \leq i \leq n$ leads to line of regression

$$\hat{y} = b_0 + b_1 x$$

2) Assume an ideal line

$$\hat{y} = \beta_0 + \beta_1 x$$

3) Together with an error process ϵ_i folowing an $N(0, \sigma)$ law (independent for each i.)

4) So that individual observations come from

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Regression Inference Questions

Natural Questions in Regression:

1) Estimate β_0 and β_1.
2) Estimate the accuracy of b_0 and b_1 as estimators of β_0 and β_1.
3) Estimate σ, the standard deviation of the error process.

For a given value x^* of x:

4) How accurately does the regression estimate $b_0 + b_1 x^*$ approximate an actual y observation when $x = x^*$.
5) How accurately does the regression estimate $b_0 + b_1 x^*$ approximate the average of a lot of y observations when $x = x^*$.