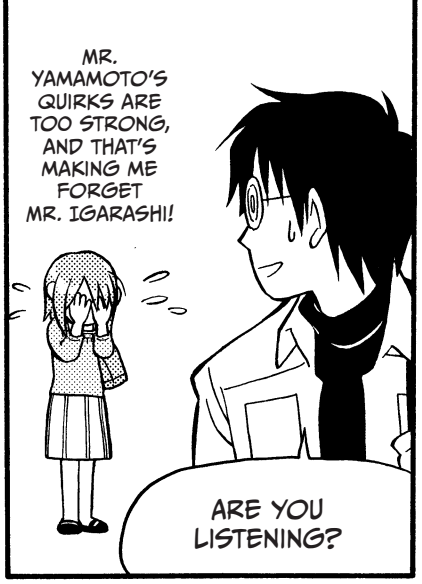
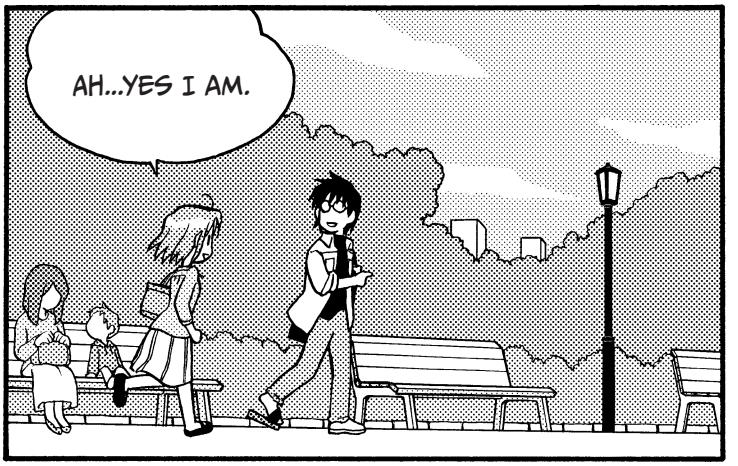
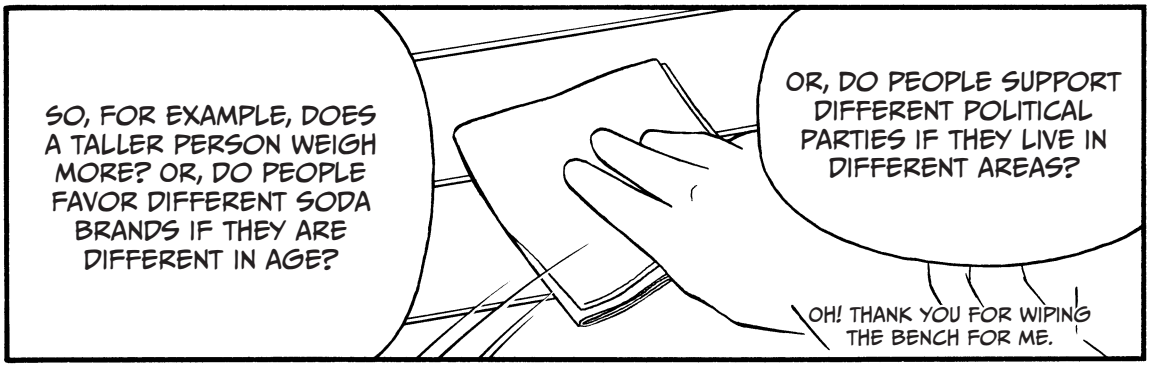
6

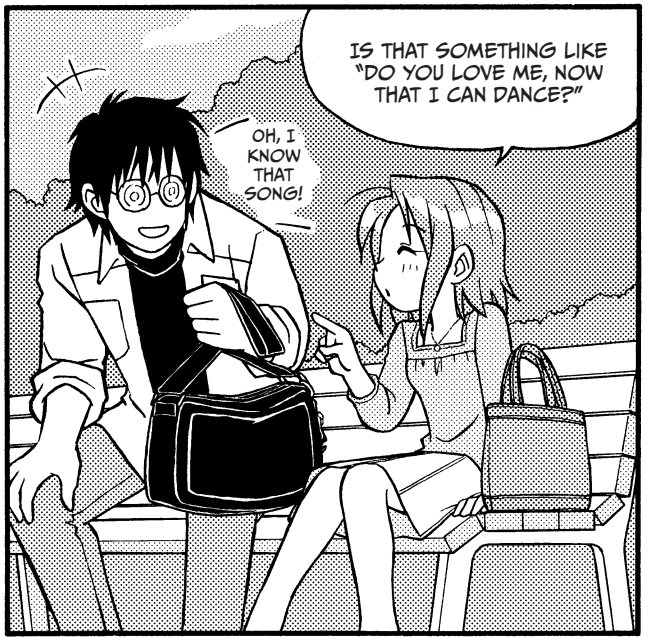
LET'S LOOK AT THE RELATIONSHIP BETWEEN TWO VARIABLES

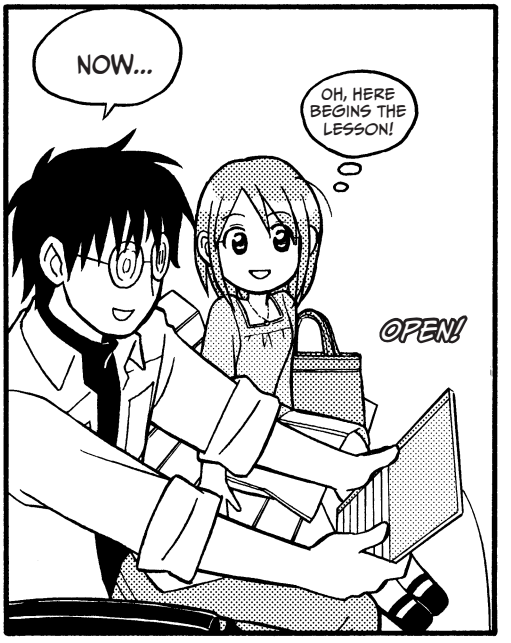

ずるぺた
NICE WEATHER.
ずるぺた*
*SCRITCH, CLUMP
WHY ARE YOU WEARING THAT SANDAL?! YOU'RE WEARING TWO DIFFERENT SHOES!
OOPS! YOU'RE RIGHT.
NOT A BAD DAY TO STUDY OUTDOORS.
IF I WERE HERE WITH MR. IGARASHI... I'D BE SO HAPPY.
MR. IGARASHI... MR. IGARASHI...?
OH, NO! I AM BEGINNING TO LOSE MY MEMORY OF HIM.
TODAY, I AM GOING TO TALK ABOUT THE RELATIONSHIP BETWEEN TWO VARIABLES.
ガーン
ずるぺた*
*SCRITCH, CLUMP
ずるぺた


MR. YAMAMOTO'S QUIRKS ARE TOO STRONG, AND THAT'S MAKING ME FORGET MR. IGARASHI!
ARE YOU LISTENING?

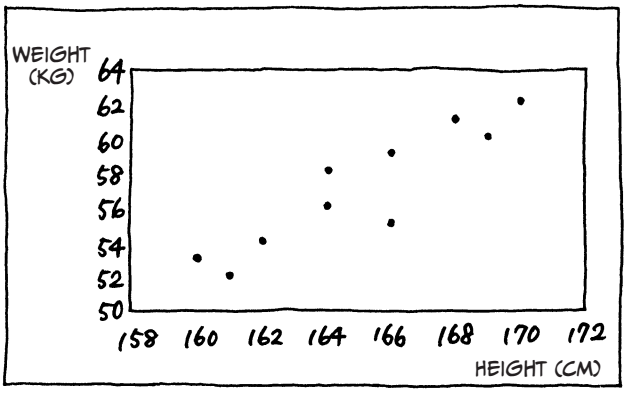

AH...YES I AM.


SO, FOR EXAMPLE, DOES A TALLER PERSON WEIGH MORE? OR, DO PEOPLE FAVOR DIFFERENT SODA BRANDS IF THEY ARE DIFFERENT IN AGE?
OR, DO PEOPLE SUPPORT DIFFERENT POLITICAL PARTIES IF THEY LIVE IN DIFFERENT AREAS?
OH! THANK YOU FOR WIPING THE BENCH FOR ME.


IS THAT SOMETHING LIKE "DO YOU LOVE ME, NOW THAT I CAN DANCE?"
OH, I KNOW THAT SONG!

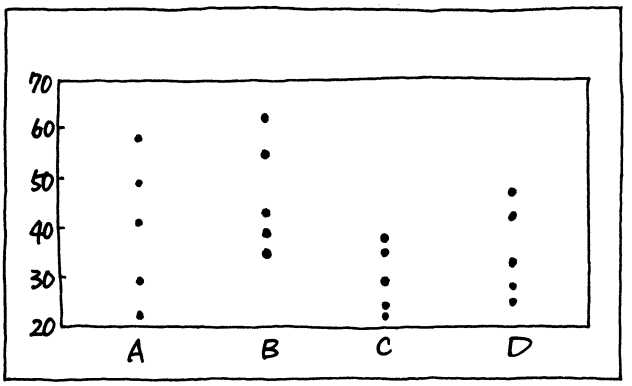

NOW...
OH, HERE BEGINS THE LESSON!
OPEN!

SCATTER CHART OF HEIGHT AND WEIGHT



NUMERICAL AND NUMERICAL

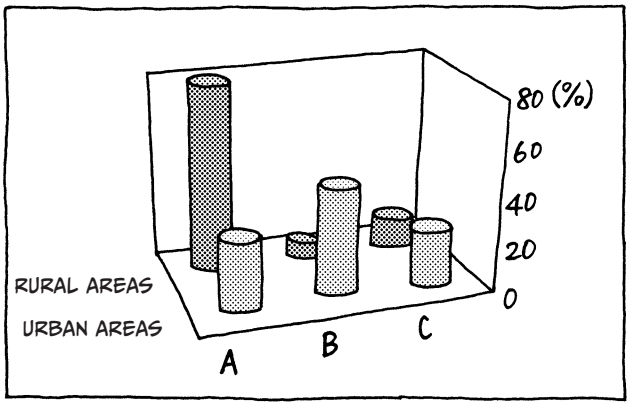
SCATTER CHART OF FAVORITE SODA BRAND AND AGE



NUMERICAL AND CATEGORICAL

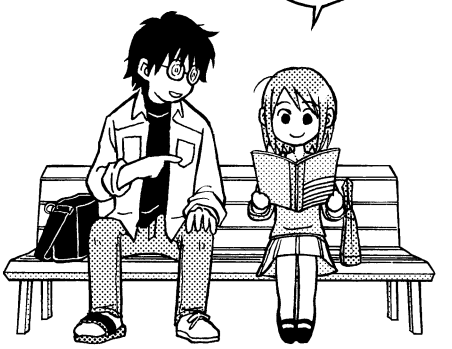
YOU CAN SEE WHETHER OR NOT TWO VARIABLES ARE RELATED TO EACH OTHER BY DRAWING A CHART.

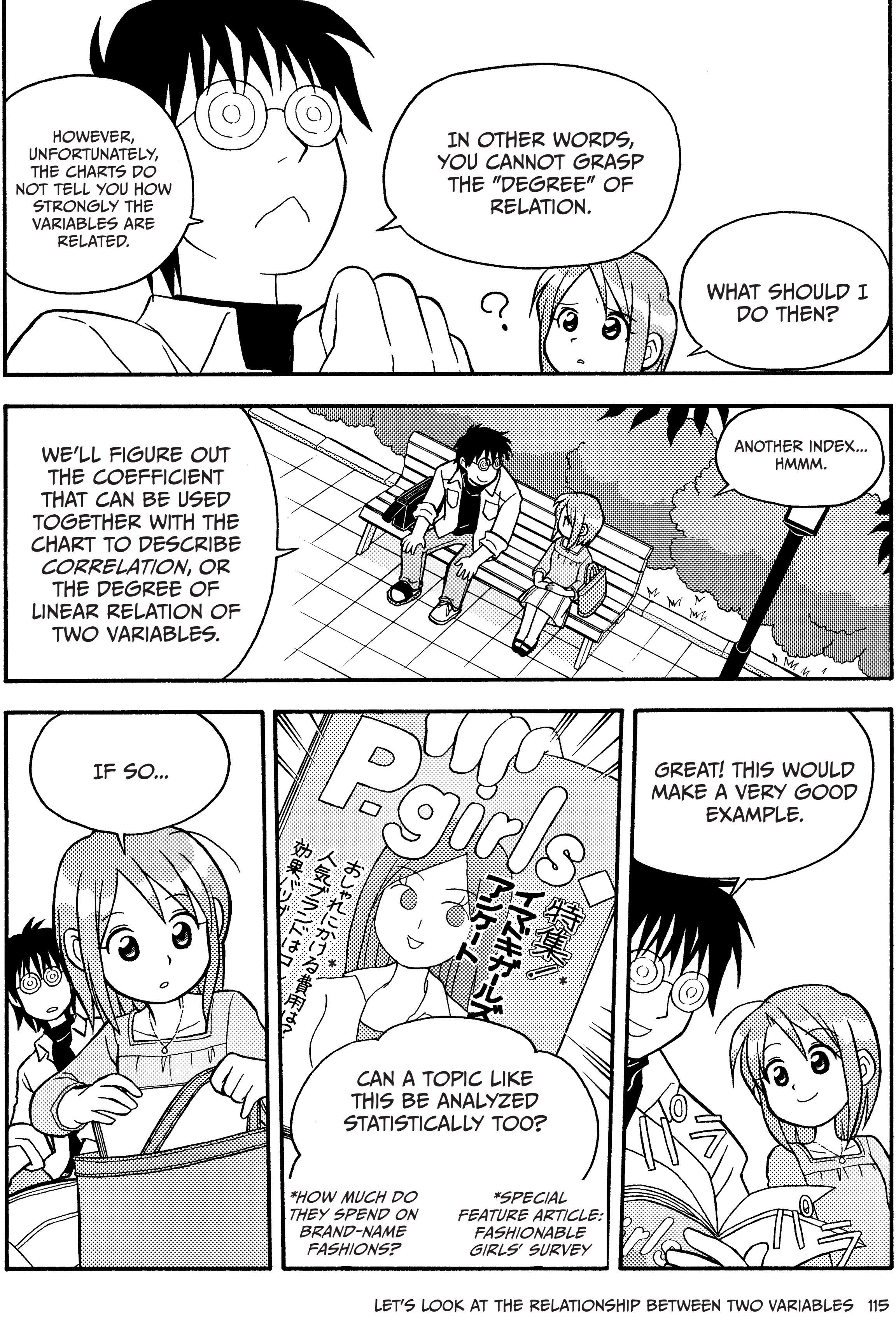
CYLINDER CHART OF PLACE OF RESIDENCE AND SUPPORT OF POLITICAL PARTY X



CATEGORICAL AND CATEGORICAL

AHA!




HOWEVER, UNFORTUNATELY, THE CHARTS DO NOT TELL YOU HOW STRONGLY THE VARIABLES ARE RELATED.
IN OTHER WORDS, YOU CANNOT GRASP THE "DEGREE" OF RELATION.
WHAT SHOULD I DO THEN?
WE'LL FIGURE OUT THE COEFFICIENT THAT CAN BE USED TOGETHER WITH THE CHART TO DESCRIBE *CORRELATION*, OR THE DEGREE OF LINEAR RELATION OF TWO VARIABLES.
ANOTHER INDEX... HMMM.
IF SO...
P.girls
特集！イマドキガールズアンケート*
CAN A TOPIC LIKE THIS BE ANALYZED STATISTICALLY TOO?
*HOW MUCH DO THEY SPEND ON BRAND-NAME FASHIONS?
*SPECIAL FEATURE ARTICLE: FASHIONABLE GIRLS' SURVEY
GREAT! THIS WOULD MAKE A VERY GOOD EXAMPLE.

1. CORRELATION COEFFICIENT

OH, HERE IS A SURVEY ON MAKEUP EXPENDITURES AND CLOTHES EXPENDITURES.

BOTH VARIABLES ARE NUMERICAL!

Street survey!

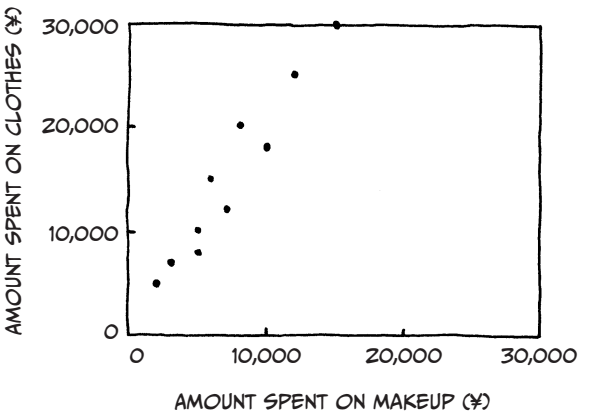
Ten ladies in their 20s answered
Monthly Expenditures on Makeup and Clothes

Respondent	Amount spent on makeup (¥)	Amount spent on clothes (¥)
Ms. A	3,000	7,000
Ms. B	5,000	8,000
Ms. C	12,000	25,000
Ms. D	2,000	5,000
Ms. E	7,000	12,000
Ms. F	15,000	30,000
Ms. G	5,000	10,000
Ms. H	6,000	15,000
Ms. I	8,000	20,000
Ms. J	10,000	18,000

WHY DON'T YOU MAKE A CHART FIRST.

YES, SIR!

SCATTER CHART OF MONTHLY EXPENDITURES ON MAKEUP AND CLOTHES

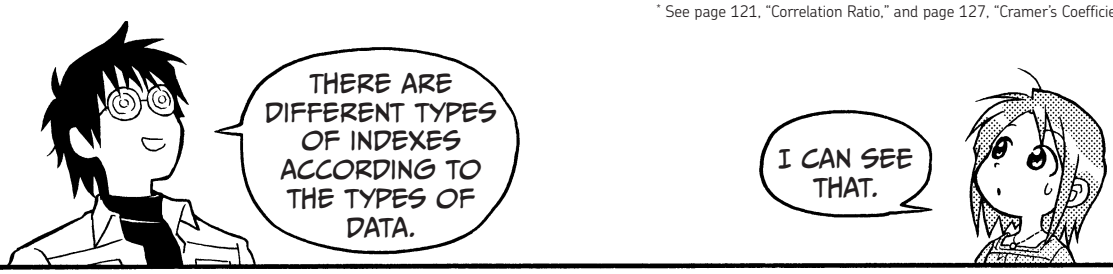


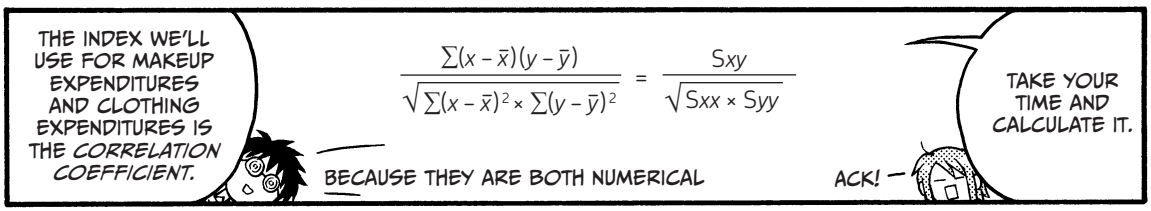
OBVIOUSLY, PEOPLE WHO SPEND MORE ON MAKEUP SPEND MORE ON THEIR CLOTHES AS WELL.

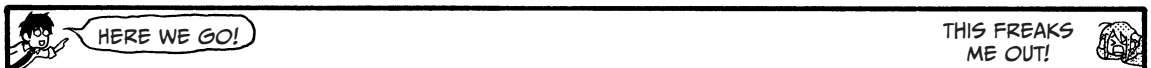
THEN, WHY DON'T WE TRY FIGURING OUT THE DEGREE OF RELATIONSHIP?

Data types	Index	Value range	Formula
Numerical and numerical	Correlation coefficient	–1 – 1	$\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \times \sum(y-\bar{y})^2}} = \frac{Sxy}{\sqrt{Sxx \times Syy}}$
Numerical and categorical	Correlation ratio*	0 – 1	$\frac{\text{interclass variance}}{\text{intraclass variance + interclass variance}}$
Categorical and categorical	Cramer's coefficient*	0 – 1	$\sqrt{\frac{\chi_0^2}{\text{the total number of values} \times (\min\{\text{the number of lines in the cross tabulation, the number of rows in the cross tabulation}\} - 1)}}$

* See page 121, "Correlation Ratio," and page 127, "Cramer's Coefficient."



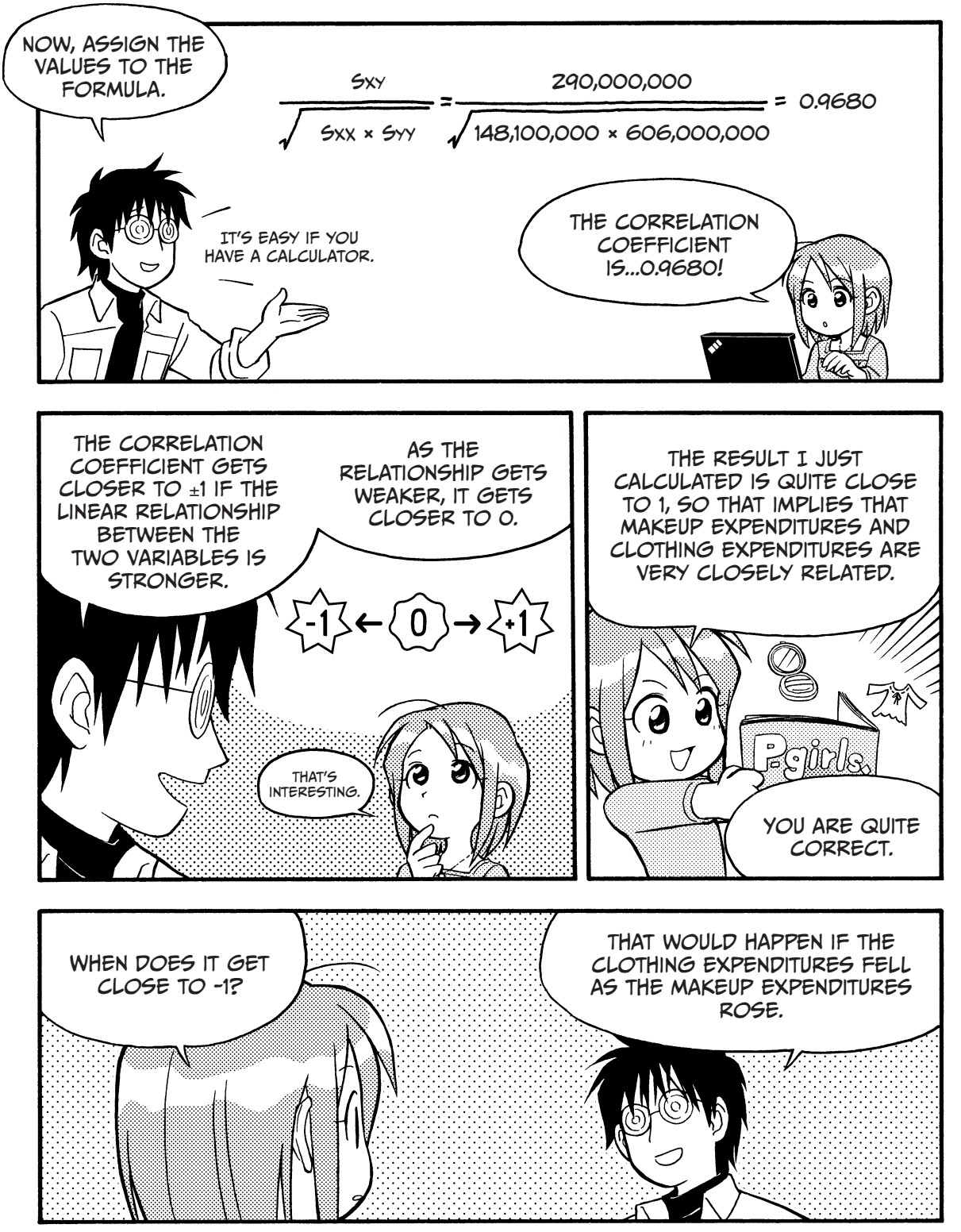


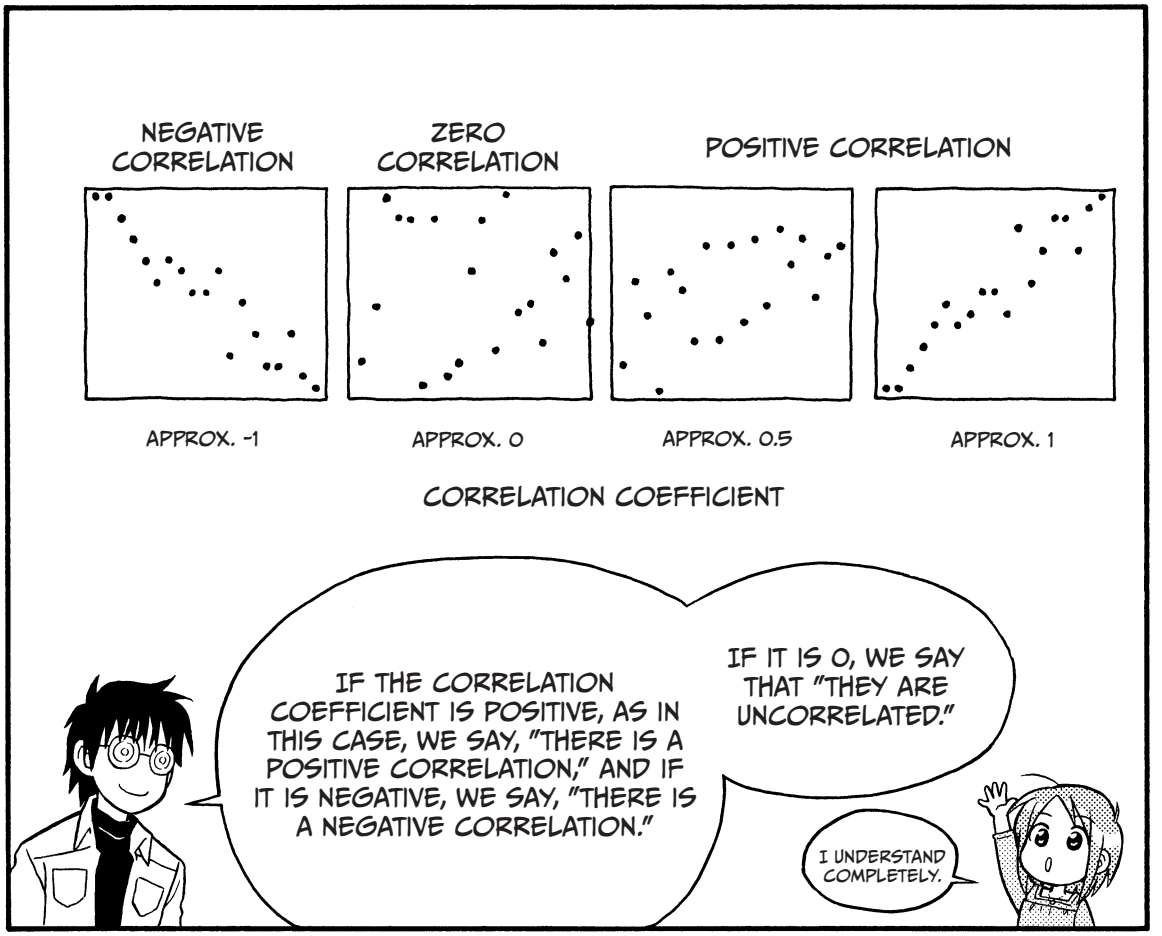


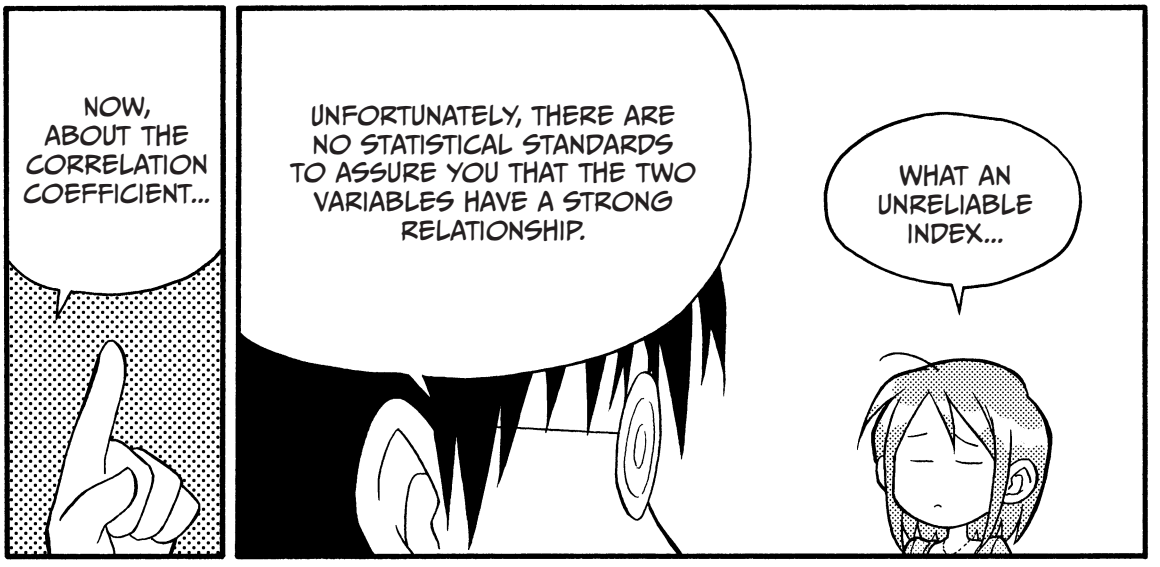
THE PROCESS FOR CALCULATING THE CORRELATION COEFFICIENT FOR MONTHLY EXPENDITURES ON MAKEUP AND CLOTHES

	Amount spent on makeup (¥)	Amount spent on clothes (¥)					
	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
Ms. A	3,000	7,000	-4,300	-8,000	18,490,000	64,000,000	34,400,000
Ms. B	5,000	8,000	-2,300	-7,000	5,290,000	49,000,000	16,100,000
Ms. C	12,000	25,000	4,700	10,000	22,090,000	100,000,000	47,000,000
Ms. D	2,000	5,000	-5,300	-10,000	28,090,000	100,000,000	53,000,000
Ms. E	7,000	12,000	-300	-3,000	90,000	9,000,000	900,000
Ms. F	15,000	30,000	7,700	15,000	59,290,000	225,000,000	115,500,000
Ms. G	5,000	10,000	-2,300	-5,000	5,290,000	25,000,000	11,500,000
Ms. H	6,000	15,000	-1,300	0	1,690,000	0	0
Ms. I	8,000	20,000	700	5,000	490,000	25,000,000	3,500,000
Ms. J	10,000	18,000	2,700	3,000	7,290,000	9,000,000	8,100,000
Sum	73,000	150,000	0	0	148,100,000	606,000,000	290,000,000
Mean	7,300	15,000					

$\bar{x}$ $\bar{y}$ Sxx Syy Sxy

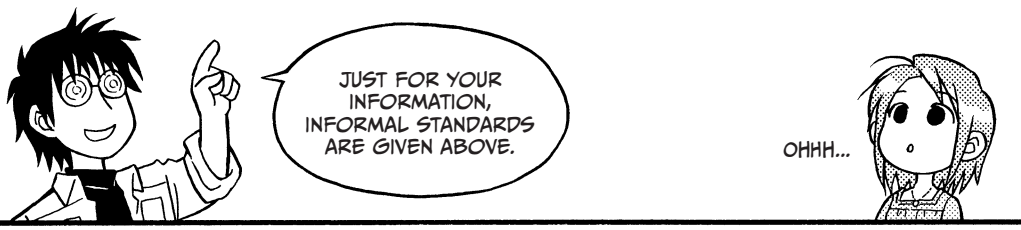

NOW, ASSIGN THE VALUES TO THE FORMULA.
$\frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} = \frac{290{,}000{,}000}{\sqrt{148{,}100{,}000 \times 606{,}000{,}000}} = 0.9680$
IT'S EASY IF YOU HAVE A CALCULATOR.
THE CORRELATION COEFFICIENT IS...0.9680!
THE CORRELATION COEFFICIENT GETS CLOSER TO ±1 IF THE LINEAR RELATIONSHIP BETWEEN THE TWO VARIABLES IS STRONGER.
AS THE RELATIONSHIP GETS WEAKER, IT GETS CLOSER TO 0.
-1 ← 0 → +1
THAT'S INTERESTING.
THE RESULT I JUST CALCULATED IS QUITE CLOSE TO 1, SO THAT IMPLIES THAT MAKEUP EXPENDITURES AND CLOTHING EXPENDITURES ARE VERY CLOSELY RELATED.
P-girls.
YOU ARE QUITE CORRECT.
WHEN DOES IT GET CLOSE TO -1?
THAT WOULD HAPPEN IF THE CLOTHING EXPENDITURES FELL AS THE MAKEUP EXPENDITURES ROSE.


NEGATIVE CORRELATION
ZERO CORRELATION
POSITIVE CORRELATION
APPROX. -1
APPROX. 0
APPROX. 0.5
APPROX. 1
CORRELATION COEFFICIENT
IF THE CORRELATION COEFFICIENT IS POSITIVE, AS IN THIS CASE, WE SAY, "THERE IS A POSITIVE CORRELATION," AND IF IT IS NEGATIVE, WE SAY, "THERE IS A NEGATIVE CORRELATION."
IF IT IS 0, WE SAY THAT "THEY ARE UNCORRELATED."
I UNDERSTAND COMPLETELY.


NOW, ABOUT THE CORRELATION COEFFICIENT...
UNFORTUNATELY, THERE ARE NO STATISTICAL STANDARDS TO ASSURE YOU THAT THE TWO VARIABLES HAVE A STRONG RELATIONSHIP.
WHAT AN UNRELIABLE INDEX...

INFORMAL STANDARDS OF THE CORRELATION COEFFICIENT

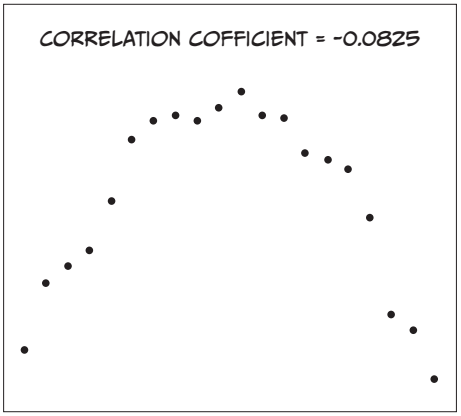
Absolute value of the correlation coefficient		Detailed description	Rough description
1.0–0.9	⇨	Very strongly related	
0.9–0.7	⇨	Fairly strongly related	Related
0.7–0.5	⇨	Fairly weakly related	
Below 0.5	⇨	Very weakly related	Not related



WARNING

I mentioned earlier that the correlation coefficient is an index that shows the degree of *linear* relation between two numerical variables.

SAMPLE OF DATA UNSUITABLE FOR CORRELATION COEFFICIENT

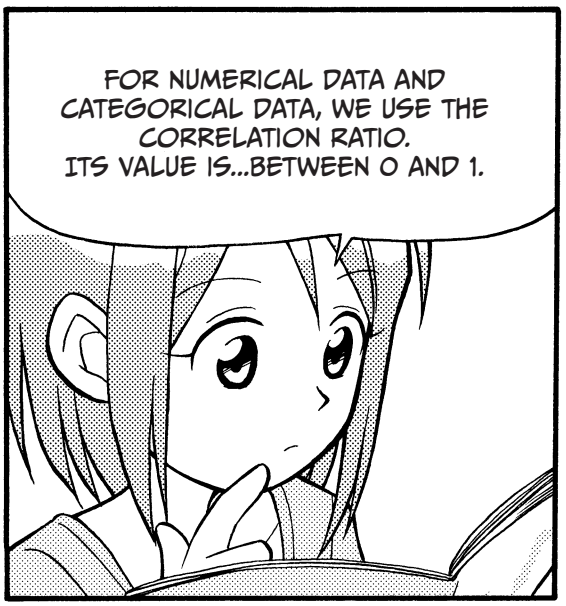


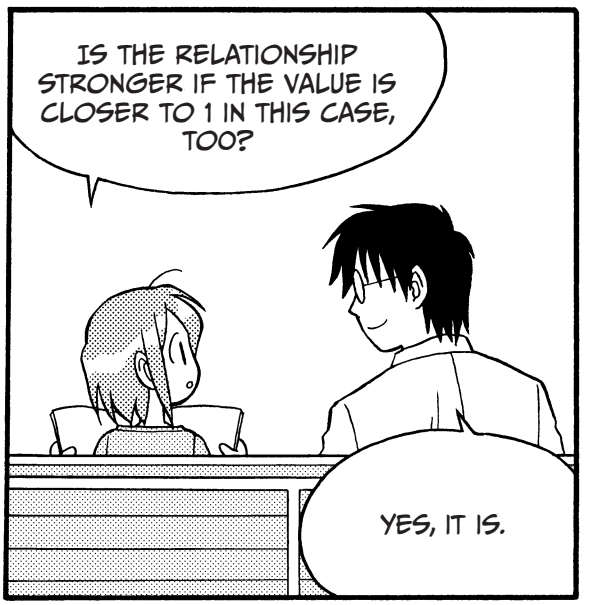
For example, the two variables are obviously related in this chart. However, the correlation coefficient is almost 0 because the relationship is *non-linear*.

Street Survey in Everyhills

Age and Favorite Fashion Brand

Respondent	Age	Brand
Ms. A	27	Theremes
Ms. B	33	Channelior
Ms. C	16	Bureperry
Ms. D	29	Bureperry
Ms. E	32	Channelior
Ms. F	23	Theremes
Ms. G	25	Channelior
Ms. H	28	Theremes
Ms. I	22	Bureperry
Ms. J	18	Bureperry
Ms. K	26	Channelior
Ms. L	26	Theremes
Ms. M	15	Bureperry
Ms. N	29	Channelior
Ms. O	26	Bureperry





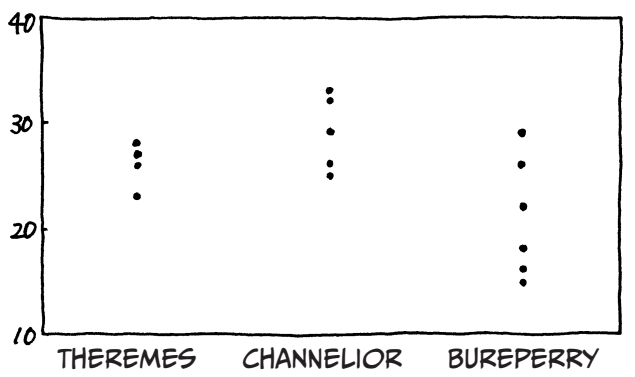
I WILL RE-ORGANIZE THE TABLE.

HMMM.

FAVORITE FASHION BRAND AND AGE

THEREMES	CHANNELIOR	BUREPERRY	
23	25	15	
26	26	16	
27	29	18	
28	32	22	
	33	26	
		29	
SUM	104	145	126	375
MEAN	26	29	21	25

SCATTER CHART OF FAVORITE FASHION BRAND AND AGE



NEXT, LET'S MAKE A CHART.

WOW! IT SEEMS LIKE THERE ARE CERTAIN CORRELATIONS.

AT THIS POINT, LET'S ACTUALLY CALCULATE THE VALUE OF THE CORRELATION RATIO.

ALL RIGHT, LET'S DO IT!

The value of the correlation ratio can be calculated by following steps 1 through 4 below.

Step 1

Do the calculations in the table below.

		Sum
(Theremes – average for Theremes)²	$(23 - 26)^2 = (-3)^2 = 9$ $(26 - 26)^2 = 0^2 = 0$ $(27 - 26)^2 = 1^2 = 1$ $(28 - 26)^2 = 2^2 = 4$	14 S_{TT}
(Channelior – average for Channelior)²	$(25 - 29)^2 = (-4)^2 = 16$ $(26 - 29)^2 = (-3)^2 = 9$ $(29 - 29)^2 = 0^2 = 0$ $(32 - 29)^2 = 3^2 = 9$ $(33 - 29)^2 = 4^2 = 16$	50 S_{CC}
(Bureperry – average for Bureperry)²	$(15 - 21)^2 = (-6)^2 = 36$ $(16 - 21)^2 = (-5)^2 = 25$ $(18 - 21)^2 = (-3)^2 = 9$ $(22 - 21)^2 = 1^2 = 1$ $(26 - 21)^2 = 5^2 = 25$ $(29 - 21)^2 = 8^2 = 64$	160 S_{BB}

Step 2

Calculate the intraclass variance ($S_{TT} + S_{CC} + S_{BB}$ = how much the data within each category varies).

$$S_{TT} + S_{CC} + S_{BB} = 14 + 50 + 160 = 224$$

Step 3

Calculate the interclass variance, or how different the categories are from each other.

(number of votes for Theremes) × (average for Theremes – average for all data)^2
+ (number of votes for Channelior) × (average for Channelior – average for all data)^2
+ (number of votes for Bureperry) × (average for Bureperry – average for all data)^2

$4 \times (26 - 25)^2 + 5 \times (29 - 25)^2 + 6 \times (21 - 25)^2$

$= 4 \times 1 + 5 \times 16 + 6 \times 16$

$= 4 + 80 + 96$

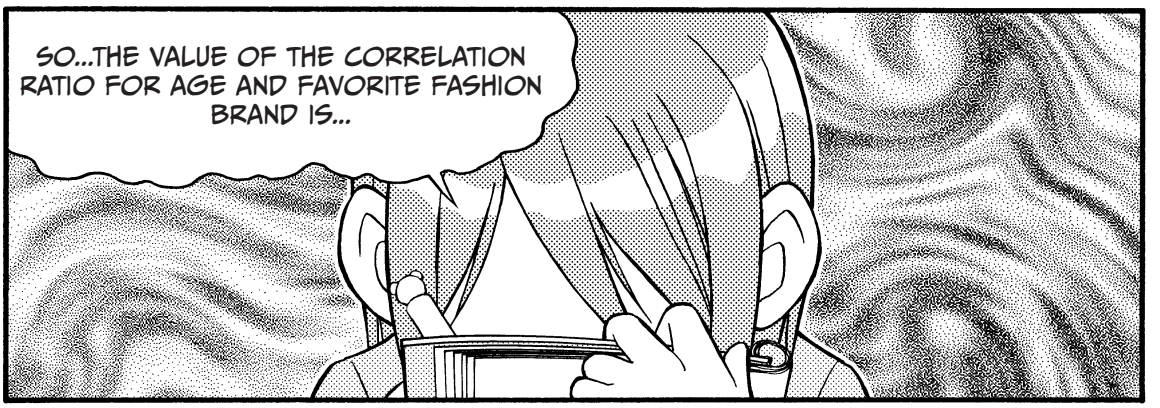
$= 180$

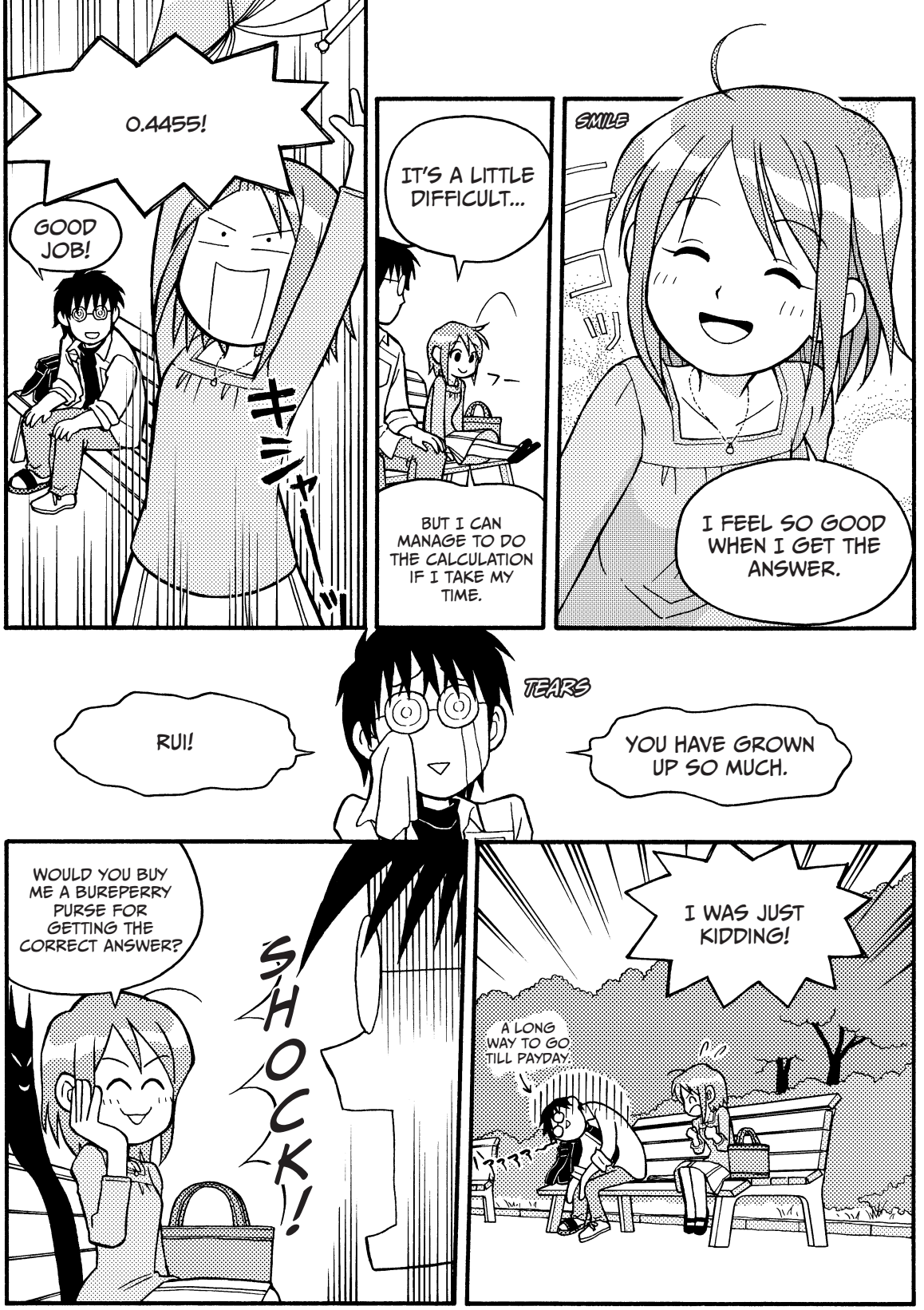
Step 4

Calculate the value of the correlation ratio.

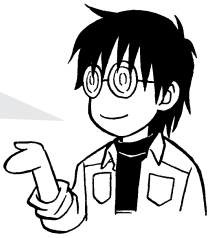
$$\frac{\text{interclass variance}}{\text{intraclass variance} + \text{interclass variance}}$$

$$\frac{180}{224 + 180} = \frac{180}{404} = 0.4455$$

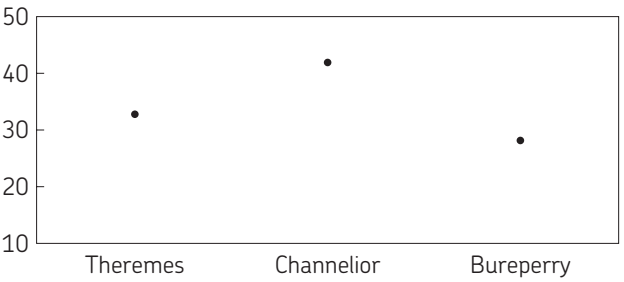



0.4455!
GOOD JOB!
キラーン
IT'S A LITTLE DIFFICULT...
フー
BUT I CAN MANAGE TO DO THE CALCULATION IF I TAKE MY TIME.
SMILE
I FEEL SO GOOD WHEN I GET THE ANSWER.
TEARS
RUI!
YOU HAVE GROWN UP SO MUCH.
WOULD YOU BUY ME A BUREPERRY PURSE FOR GETTING THE CORRECT ANSWER?
SHOCK!
I WAS JUST KIDDING!
A LONG WAY TO GO TILL PAYDAY.
ァァァァ

As explained earlier, the value of the correlation ratio is between 0 and 1. The stronger the correlation is between the two variables, the closer the value is to 1, and the weaker the correlation is between two variables, the closer the value is to 0. Refer to the charts below for more details.

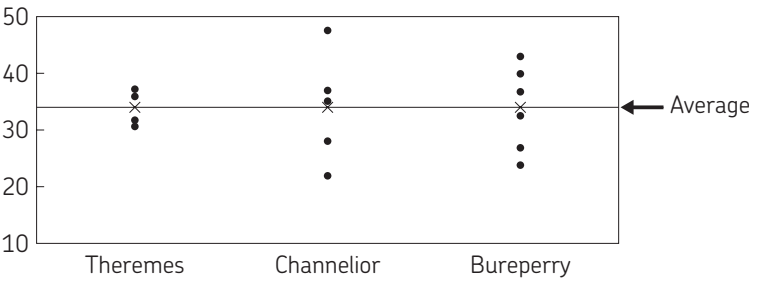


Here is a scatter chart of favorite fashion brand and age (when the correlation ratio is 1).



correlation ratio is 1 ⇔ data included in each group is the same ⇔ intraclass variance is 0

Here is a scatter chart of favorite fashion brand and age (when the correlation ratio is 0).



correlation ratio is 0 ⇔ average for each group is the same ⇔ intraclass variance is 0

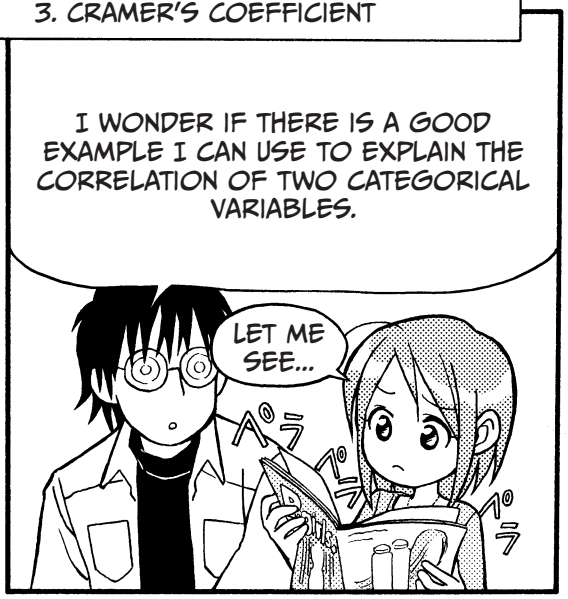
Unfortunately, there are no statistical standards such as "the two variables have a strong correlation if the correlation ratio is above a certain benchmark." However, informal standards are given below.

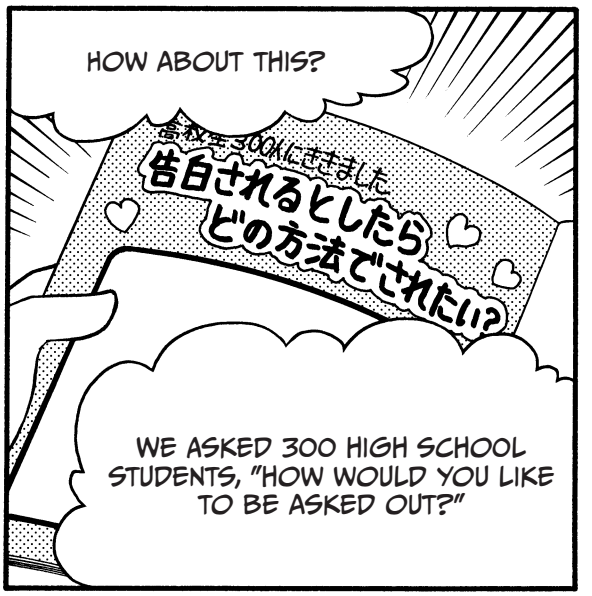
INFORMAL STANDARDS OF THE CORRELATION RATIO

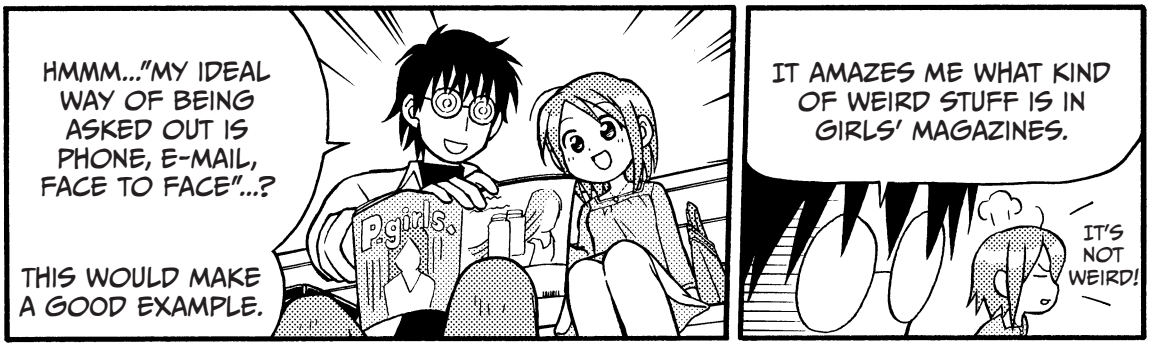
Correlation ratio		Detailed description	Rough description
1.0–0.8	⇨	Very strongly related	
0.8–0.5	⇨	Fairly strongly related	Related
0.5–0.25	⇨	Fairly weakly related	
Below 0.25	⇨	Very weakly related	Not related

The result of the calculation for the case in question was 0.4455, so the variables are fairly weakly related!

3. CRAMER'S COEFFICIENT







CROSS TABULATION OF SEX AND DESIRED WAY OF BEING ASKED OUT

		DESIRED WAY OF BEING ASKED OUT			SUM
		PHONE	E-MAIL	FACE TO FACE	
SEX	FEMALE	34	61	53	148
	MALE	38	40	74	152
SUM		72	101	127	300

This indicates that 74 out of 152 males answered that they'd like to be asked out directly.

CROSS TABULATION OF SEX AND DESIRED WAY OF BEING ASKED OUT (HORIZONTAL PERCENTAGE TABLE)

		DESIRED WAY OF BEING ASKED OUT			SUM
		PHONE	E-MAIL	FACE TO FACE	
SEX	FEMALE	23%	41%	36%	100%
	MALE	25%	26%	49%	100%
SUM		24%	34%	42%	100%

This shows that 49% ($\frac{74}{152} \times 100$) of the 152 males would like to be asked out directly.


IT INDEED SEEMS THAT THERE IS A DIFFERENCE IN THE DESIRED WAY OF BEING ASKED OUT BETWEEN GIRLS AND BOYS.
IN OTHER WORDS, THERE IS A CORRELATION BETWEEN SEX AND DESIRED WAY OF BEING ASKED OUT.
I SHOULD CONTACT MR. IGARASHI DIRECTLY IF I AM GOING TO ASK HIM OUT...
RUI? ARE YOU LISTENING?
SURE, I AM.
WHAT WAS THE INDEX TO EXPRESS THE DEGREE OF CORRELATION BETWEEN TWO PIECES OF CATEGORICAL DATA?
THE CRAMER'S COEFFICIENT!
THE CRAMER'S COEFFICIENT IS ALSO CALLED THE *CRAMER'S V* OR AN *INDEPENDENT COEFFICIENT.*
I CANNOT CRAM SO MUCH NEW VOCABULARY INTO MY HEAD!
WELL, THAT WAS KIND OF FUNNY.
YOU DON'T HAVE TO BE NICE TO ME.
* BAD JOKE AGAIN...

The Cramer's coefficient can be calculated by following steps 1 through 5 below.

Step 1

Prepare a cross tabulation. The values surrounded by the bold frame are called *actual measurement frequencies*.

		Desired way of being asked out			Sum
		Phone	E-mail	Face to face	
Sex	Female	34	61	53	148
	Male	38	40	74	152
Sum		72	101	127	300

Step 2

Do the calculations in the table below. The values surrounded by the bold frame are called *expected frequencies*.

		Desired way of being asked out			Sum
		Phone	E-mail	Face to face	
Sex	Female	$\frac{148 \times 72}{300}$	$\frac{148 \times 101}{300}$	$\frac{148 \times 127}{300}$	148
	Male	$\frac{152 \times 72}{300}$	$\frac{152 \times 101}{300}$	$\frac{152 \times 127}{300}$	152
Sum		72	101	127	300

$$\frac{\text{sum of male} \times \text{sum of face to face}}{\text{total number of values}}$$

Formula A

If sex and desired way of being asked out have no relationship, the ratio between phone, e-mail, and face to face should be

$$72 : 101 : 127 = \frac{72}{72+101+127} : \frac{101}{72+101+127} : \frac{127}{72+101+127}$$

$$= \frac{72}{300} : \frac{101}{300} : \frac{127}{300}$$

for both males and females, according to the sum column in the table in step 2. Thus, our expected frequency (Formula A) shows the predicted number of males who wish to be asked out directly when there is no relationship between sex and desired way of being asked out is $152 \times (127 \div 300) = (152 \times 127) \div 300$, or

$$152 \times \frac{127}{300} = \frac{152 \times 127}{300} = 64.3$$

Step 3

Calculate $\dfrac{(\text{actual frequency} - \text{expected frequency})^2}{\text{expected frequency}}$ for each square.

		Desired way of being asked out			Sum
		Phone	E-mail	Face to face	
Sex	Female	$\dfrac{\left(34 - \frac{148 \times 72}{300}\right)^2}{\frac{148 \times 72}{300}}$	$\dfrac{\left(61 - \frac{148 \times 101}{300}\right)^2}{\frac{148 \times 101}{300}}$	$\dfrac{\left(53 - \frac{148 \times 127}{300}\right)^2}{\frac{148 \times 127}{300}}$	148
	Male	$\dfrac{\left(38 - \frac{152 \times 72}{300}\right)^2}{\frac{152 \times 72}{300}}$	$\dfrac{\left(40 - \frac{152 \times 101}{300}\right)^2}{\frac{152 \times 101}{300}}$	$\dfrac{\left(74 - \frac{152 \times 127}{300}\right)^2}{\frac{152 \times 127}{300}}$	152
Sum		72	101	127	300

The bigger the gap between the actual frequencies and the expected frequencies, the larger the values in each square become.

Step 4

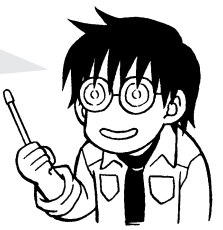
Calculate the sum of the value inside the bold frame in the table of step 3. This value is called *Pearson's chi-square test statistic*. It will be written as χ_0^2 from now on.

$$\chi_0^2 = \frac{\left(34 - \frac{148 \times 72}{300}\right)^2}{\frac{148 \times 72}{300}} + \frac{\left(61 - \frac{148 \times 101}{300}\right)^2}{\frac{148 \times 101}{300}} + \frac{\left(53 - \frac{148 \times 127}{300}\right)^2}{\frac{148 \times 127}{300}}$$

$$+ \frac{\left(38 - \frac{152 \times 72}{300}\right)^2}{\frac{152 \times 72}{300}} + \frac{\left(40 - \frac{152 \times 101}{300}\right)^2}{\frac{152 \times 101}{300}} + \frac{\left(74 - \frac{152 \times 127}{300}\right)^2}{\frac{152 \times 127}{300}}$$

$$= 8.0091$$

As can be understood from step 3, the more the actual measurements diverge from their expected frequencies, or the greater the correlation between sex and desired way of being asked out, the larger Pearson's chi-square test statistic (χ_0^2) becomes.



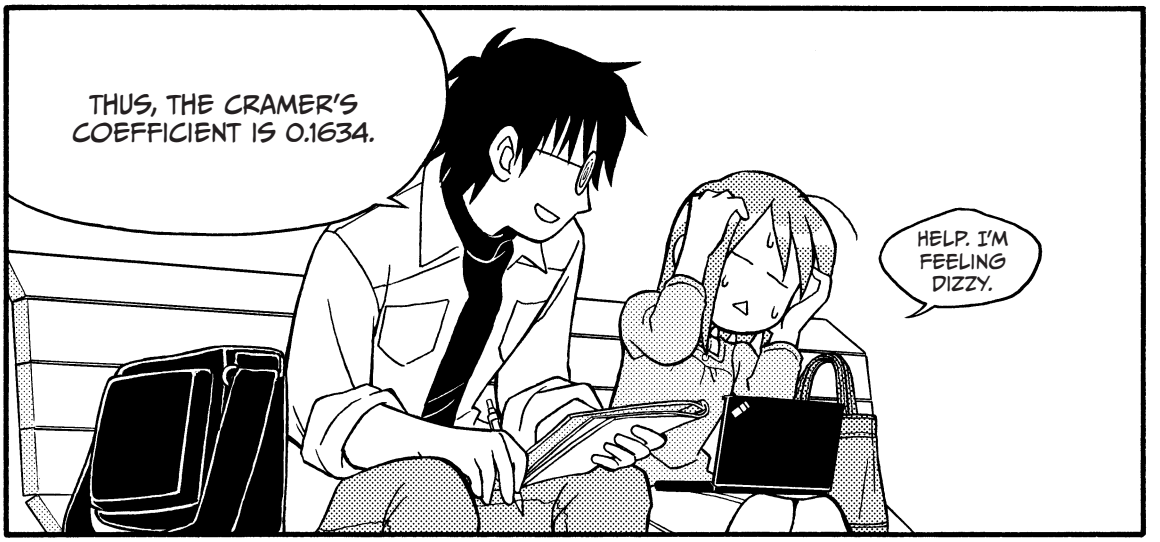
Step 5

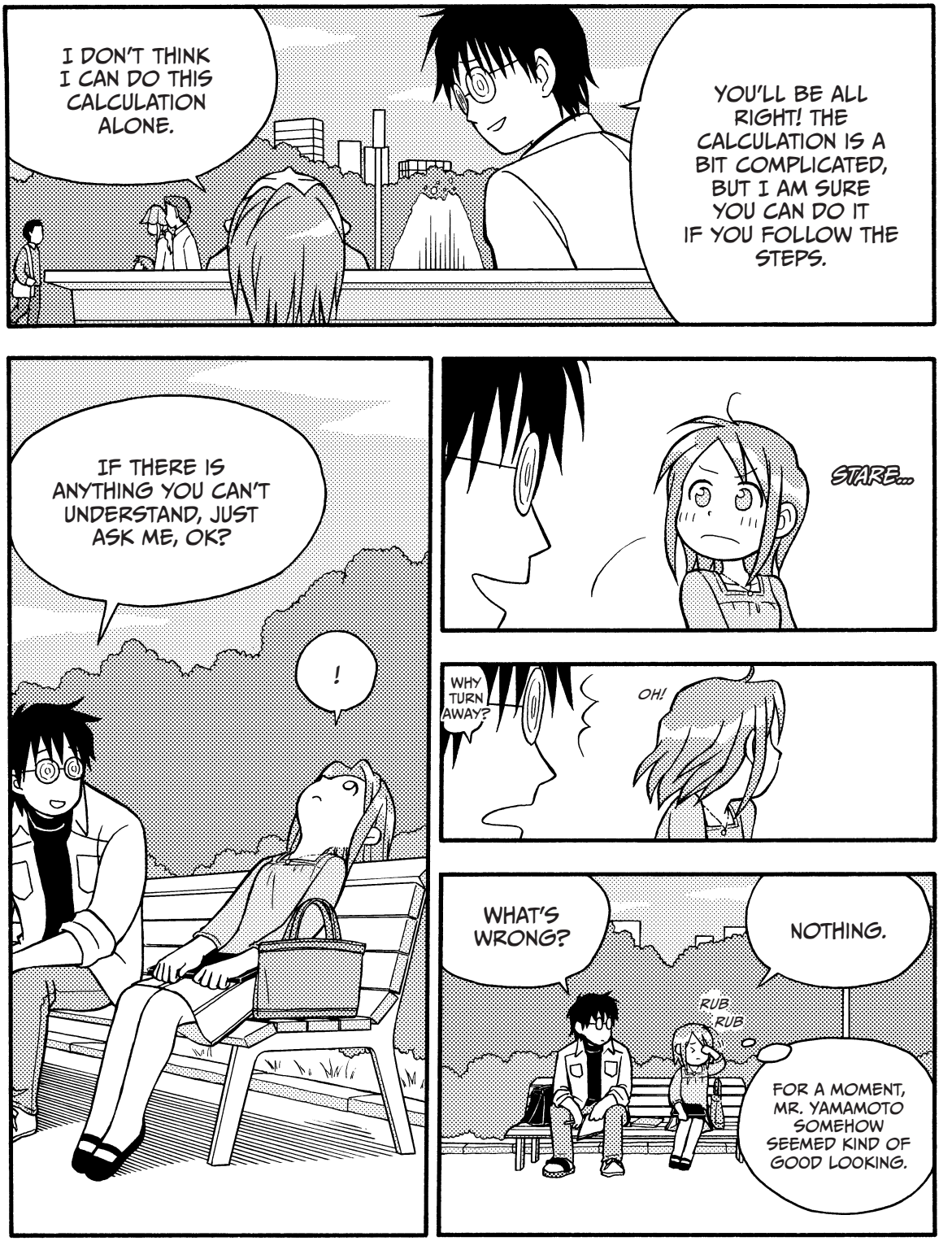
Calculate the Cramer's coefficient.

$$\sqrt{\frac{\chi_0^2}{\text{the total number of values} \times (\min\{\text{the number of lines in the cross tabulation, the number of rows in the cross tabulation}\} - 1)}}$$

min{*a*,*b*} means "whichever is smaller, *a* or *b*."

$$\sqrt{\frac{8.0091}{300 \times \min\{2,3\} - 1}} = \sqrt{\frac{8.0091}{300 \times (2 - 1)}} = \sqrt{\frac{8.0091}{300}} = 0.1634$$




I DON'T THINK I CAN DO THIS CALCULATION ALONE.
YOU'LL BE ALL RIGHT! THE CALCULATION IS A BIT COMPLICATED, BUT I AM SURE YOU CAN DO IT IF YOU FOLLOW THE STEPS.
IF THERE IS ANYTHING YOU CAN'T UNDERSTAND, JUST ASK ME, OK?
!
STARE...
WHY TURN AWAY?
OH!
WHAT'S WRONG?
NOTHING.
RUB RUB
FOR A MOMENT, MR. YAMAMOTO SOMEHOW SEEMED KIND OF GOOD LOOKING.

As explained earlier, the Cramer's coefficient is between 0 and 1. The stronger the correlation between two variables, the closer the coefficient gets to 1, and the weaker the correlation, the closer the coefficient gets to 0. See the cross tabulation (horizontal percentage table) below for more details.

Here is the cross tabulation of sex and desired way of being asked out (horizontal percentage table) when the value of the Cramer's coefficient is 1.

		Desired way of being asked out			Sum
		Phone	E-mail	Face to face	
Sex	Female	17%	83%	0%	100%
	Male	0%	0%	100%	100%

Cramer's coefficient is 1 ⇔ the preferences of female and male are completely different

Here is the cross tabulation of sex and desired way of being asked out (horizontal percentage table) when the value of the Cramer's coefficient is 0.

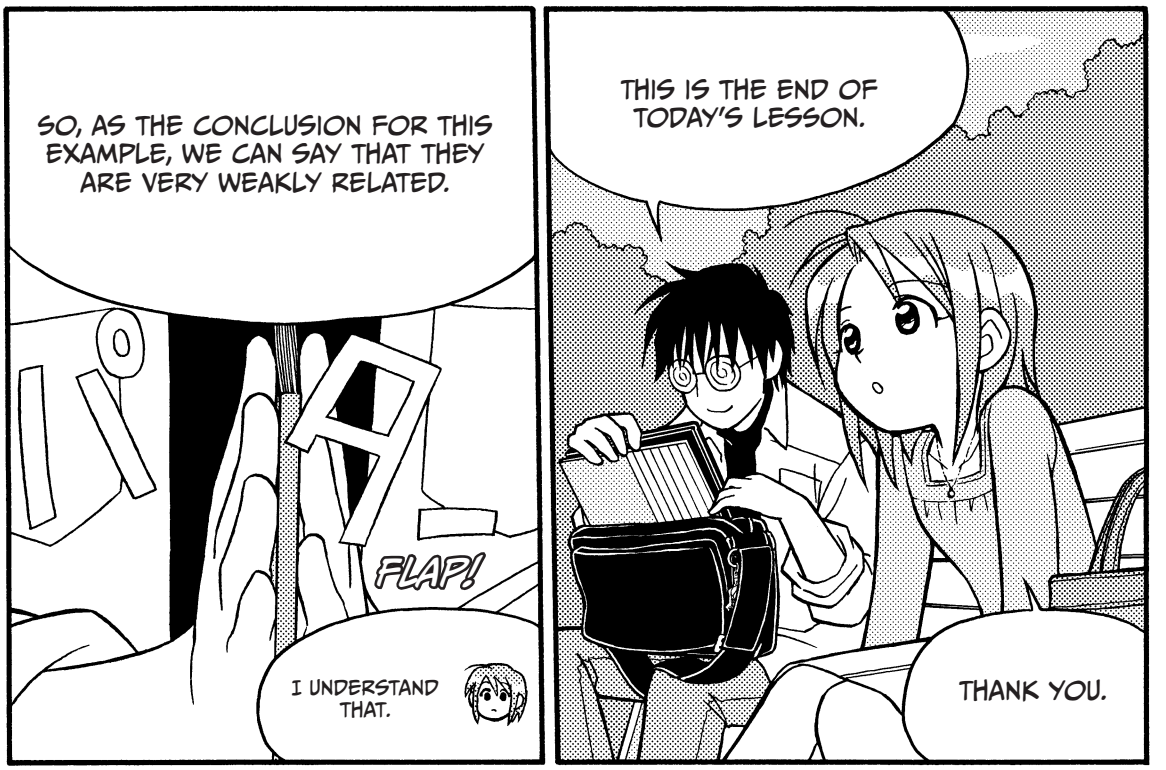
		Desired way of being asked out			Sum
		Phone	E-mail	Face to face	
Sex	Female	17%	48%	35%	100%
	Male	17%	48%	35%	100%

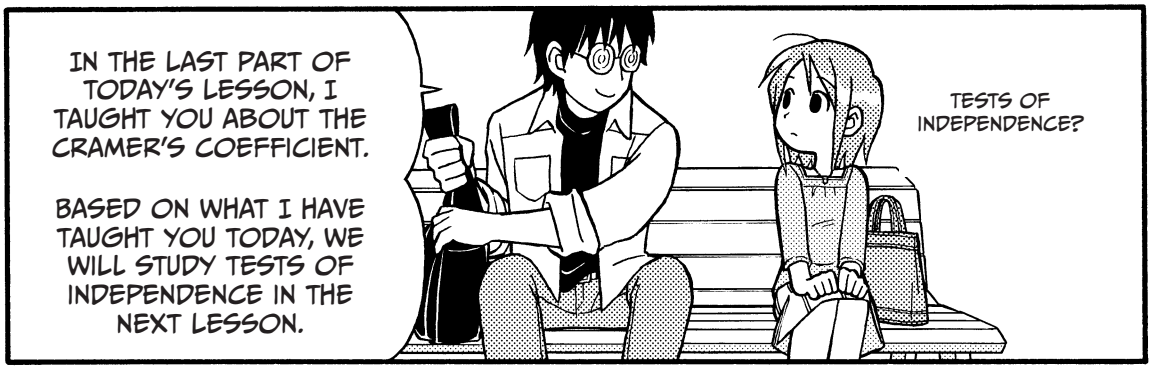
Cramer's coefficient is 0 ⇔ the preferences of female and male are the same

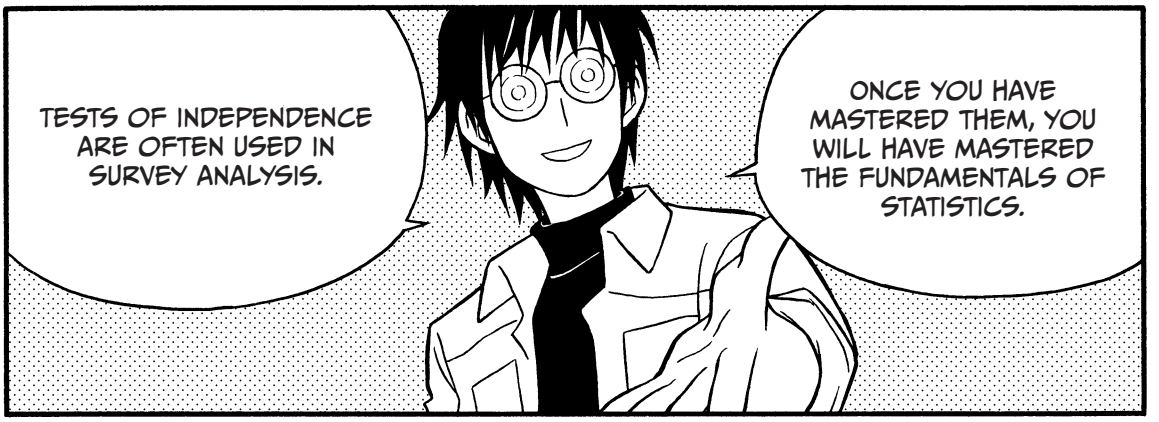
Unfortunately, there are no statistical standards such as "the two variables have a strong correlation if the Cramer's coefficient is above a certain benchmark." However, informal standards are given below.

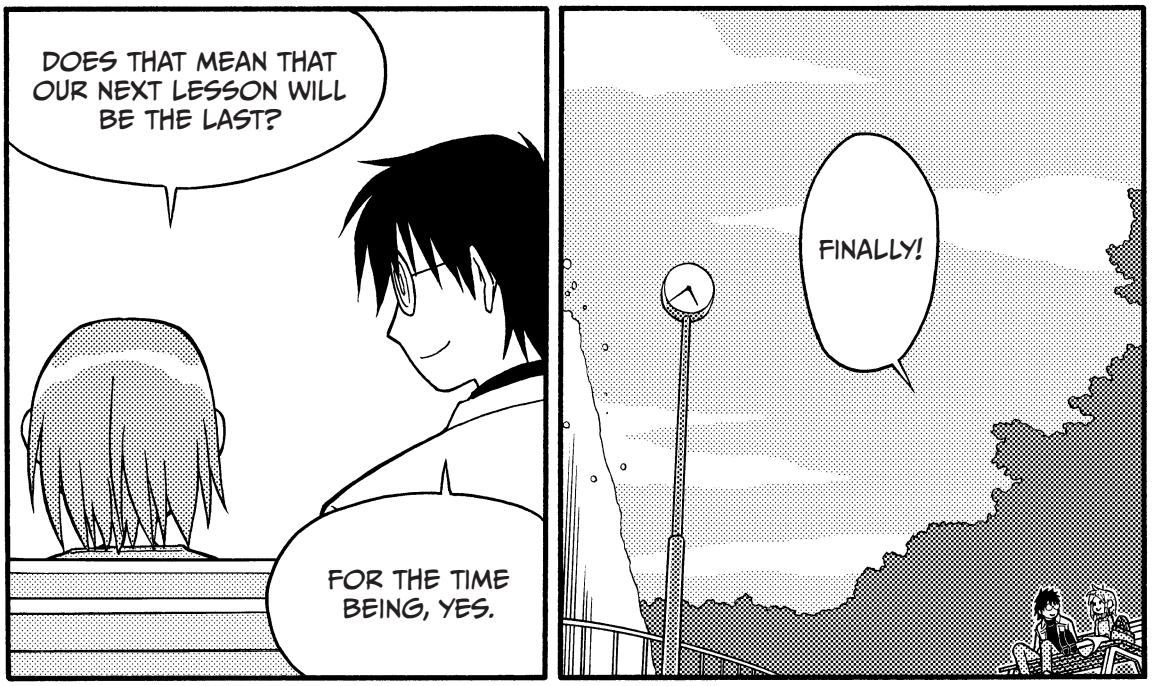
INFORMAL STANDARDS OF THE CRAMER'S COEFFICIENT

Cramer's coefficient		Detailed description	Rough description
1.0–0.8	⇨	Very strongly related	
0.8–0.5	⇨	Fairly strongly related	Related
0.5–0.25	⇨	Fairly weakly related	
Below 0.25	⇨	Very weakly related	Not related




IN THE LAST PART OF TODAY'S LESSON, I TAUGHT YOU ABOUT THE CRAMER'S COEFFICIENT.
BASED ON WHAT I HAVE TAUGHT YOU TODAY, WE WILL STUDY TESTS OF INDEPENDENCE IN THE NEXT LESSON.
TESTS OF INDEPENDENCE?


TESTS OF INDEPENDENCE ARE OFTEN USED IN SURVEY ANALYSIS.
ONCE YOU HAVE MASTERED THEM, YOU WILL HAVE MASTERED THE FUNDAMENTALS OF STATISTICS.


DOES THAT MEAN THAT OUR NEXT LESSON WILL BE THE LAST?
FOR THE TIME BEING, YES.
FINALLY!

EXERCISE

Company *X* runs a casual dining restaurant. Its financial status was declining recently. Thus, Company *X* decided to study its customers' needs and conducted a survey of randomly chosen people, age 20 or older, residing in Japan. The table below shows the results of this survey.

Respondent	**What food do you often have in a casual dining restaurant?**	**If a free drink is to be served after a meal, which would you prefer? Coffee or tea?**
1	Chinese	Coffee
2	European	Coffee
...	...	...
250	Japanese	Tea

Below is a cross tabulation made using the table above.

		Preference for coffee or tea		Sum
		Coffee	Tea	
Type of food often ordered	Japanese	43	33	76
	European	51	53	104
	Chinese	29	41	70
Sum		123	127	250

Calculate the Cramer's coefficient for the food often ordered in casual dining restaurants and the preferred free drink of either coffee or tea.

ANSWER

Step 1

Prepare a cross tabulation.

		Preference for coffee or tea		Sum
		Coffee	Tea	
Type of food often ordered	Japanese	43	33	76
	European	51	53	104
	Chinese	29	41	70
Sum		123	127	250

Step 2

Calculate the expected frequency.

		Preference for coffee or tea		Sum
		Coffee	Tea	
Type of food often ordered	Japanese	$\frac{76 \times 123}{250}$	$\frac{76 \times 127}{250}$	76
	European	$\frac{104 \times 123}{250}$	$\frac{104 \times 127}{250}$	104
	Chinese	$\frac{70 \times 123}{250}$	$\frac{70 \times 127}{250}$	70
Sum		123	127	250

Step 3

Calculate

$$\frac{(\text{actual measurement frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

for each square.

		Preference for coffee or tea		Sum
		Coffee	Tea	
Type of food often ordered	Japanese	$\dfrac{\left(43 - \dfrac{76 \times 123}{250}\right)^2}{\dfrac{76 \times 123}{250}}$	$\dfrac{\left(33 - \dfrac{76 \times 127}{250}\right)^2}{\dfrac{76 \times 127}{250}}$	76
	European	$\dfrac{\left(51 - \dfrac{104 \times 123}{250}\right)^2}{\dfrac{104 \times 123}{250}}$	$\dfrac{\left(53 - \dfrac{104 \times 127}{250}\right)^2}{\dfrac{104 \times 127}{250}}$	104
	Chinese	$\dfrac{\left(29 - \dfrac{70 \times 123}{250}\right)^2}{\dfrac{70 \times 123}{250}}$	$\dfrac{\left(41 - \dfrac{70 \times 127}{250}\right)^2}{\dfrac{70 \times 127}{250}}$	70
Sum		123	127	250

Step 4

Calculate the sum of the value inside the bold frame in the table in step 3, which is the value of Pearson's chi-square test statistic (χ_0^2).

$$\chi_0^2 = \frac{\left(43 - \frac{76 \times 123}{250}\right)^2}{\frac{76 \times 123}{250}} + \frac{\left(33 - \frac{76 \times 127}{250}\right)^2}{\frac{76 \times 127}{250}} + \frac{\left(51 - \frac{104 \times 123}{250}\right)^2}{\frac{104 \times 123}{250}} + \frac{\left(53 - \frac{104 \times 127}{250}\right)^2}{\frac{104 \times 127}{250}} + \frac{\left(29 - \frac{70 \times 123}{250}\right)^2}{\frac{70 \times 123}{250}} + \frac{\left(41 - \frac{70 \times 127}{250}\right)^2}{\frac{70 \times 127}{250}}$$

$$= 3.3483$$

Step 5

Calculate the Cramer's coefficient.

$$\sqrt{\frac{\chi_0^2}{\text{the total number of values} \times \left(\min\left\{\begin{array}{l}\text{the number of lines} \\ \text{in the cross tabulation}\end{array}, \begin{array}{l}\text{the number of rows} \\ \text{in the cross tabulation}\end{array}\right\} - 1\right)}}$$

$$\sqrt{\frac{3.3483}{250 \times (\min\{3,2\} - 1)}} = \sqrt{\frac{3.3483}{250 \times (2 - 1)}} = \sqrt{\frac{3.3483}{250}} = 0.1157$$

SUMMARY

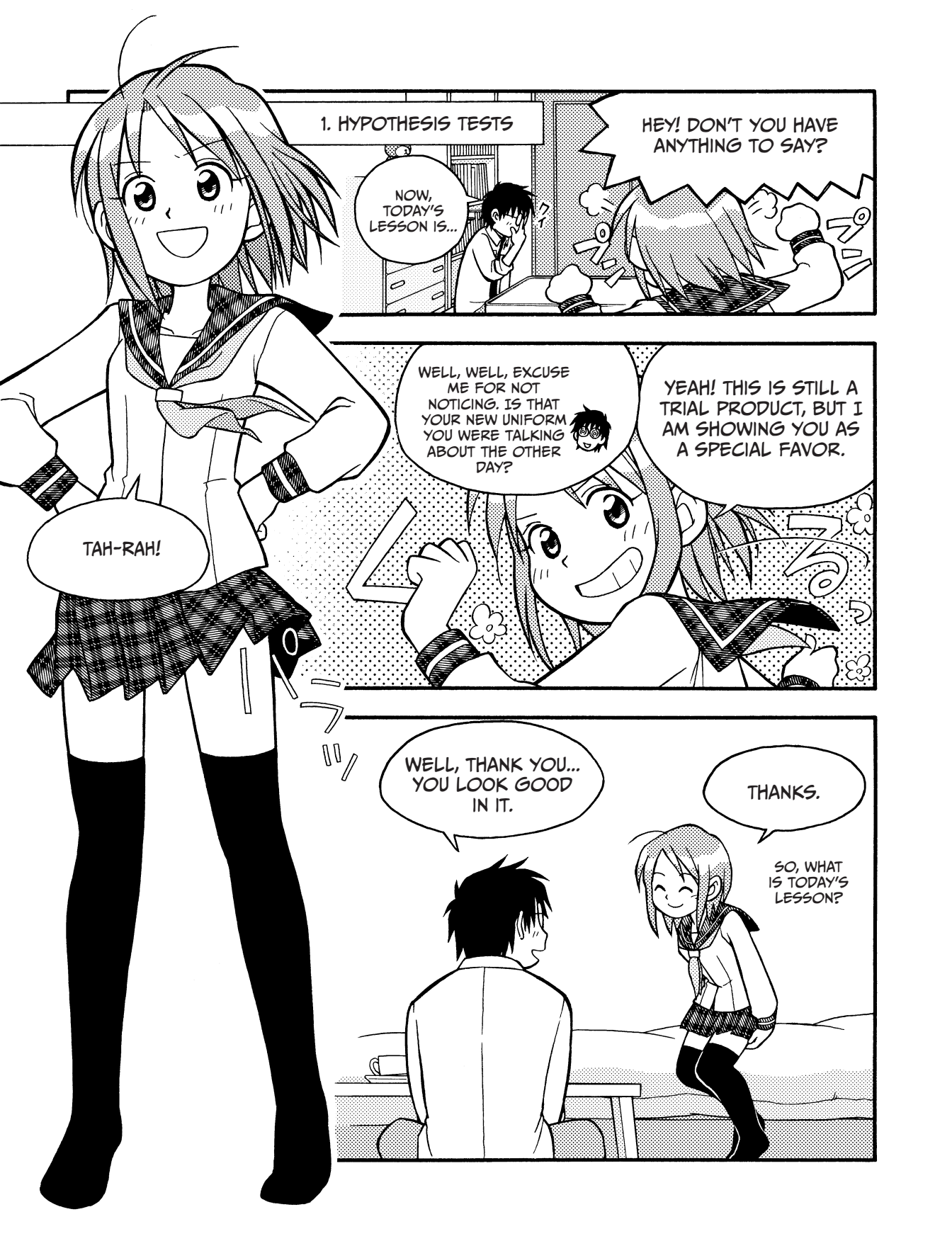
- The index used to describe the degree of correlation between numerical data and numerical data is the *correlation coefficient*.
- The index used to describe the degree of correlation between numerical data and categorical data is the *correlation ratio*.
- The index used to describe the degree of correlation between categorical data and categorical data is the *Cramer's coefficient* (sometimes called the *Cramer's V* or an *independent coefficient*).
- The characteristics of the correlation coefficient, correlation ratio, and Cramer's coefficient are shown in the table below.

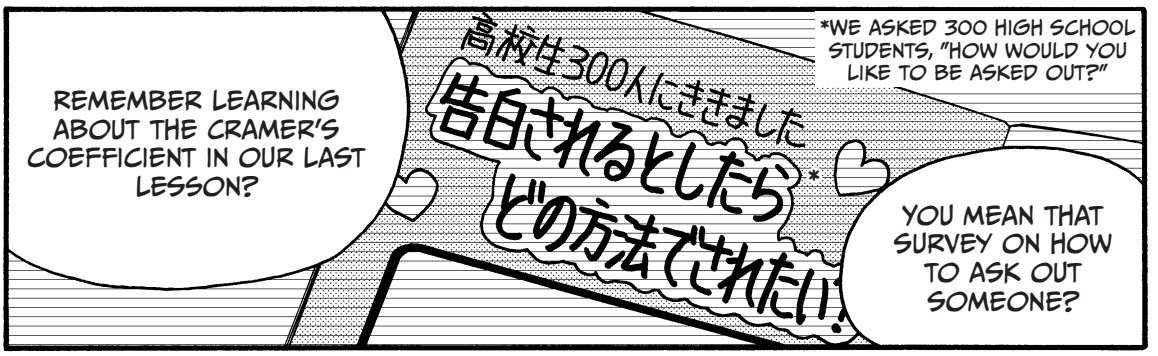
	Minimum	Maximum	The value when the two variables are not correlated at all	The value when the two variables are most strongly correlated
Correlation coefficient	–1	1	0	–1 or 1
Correlation ratio	0	1	0	1
Cramer's coefficient	0	1	0	1

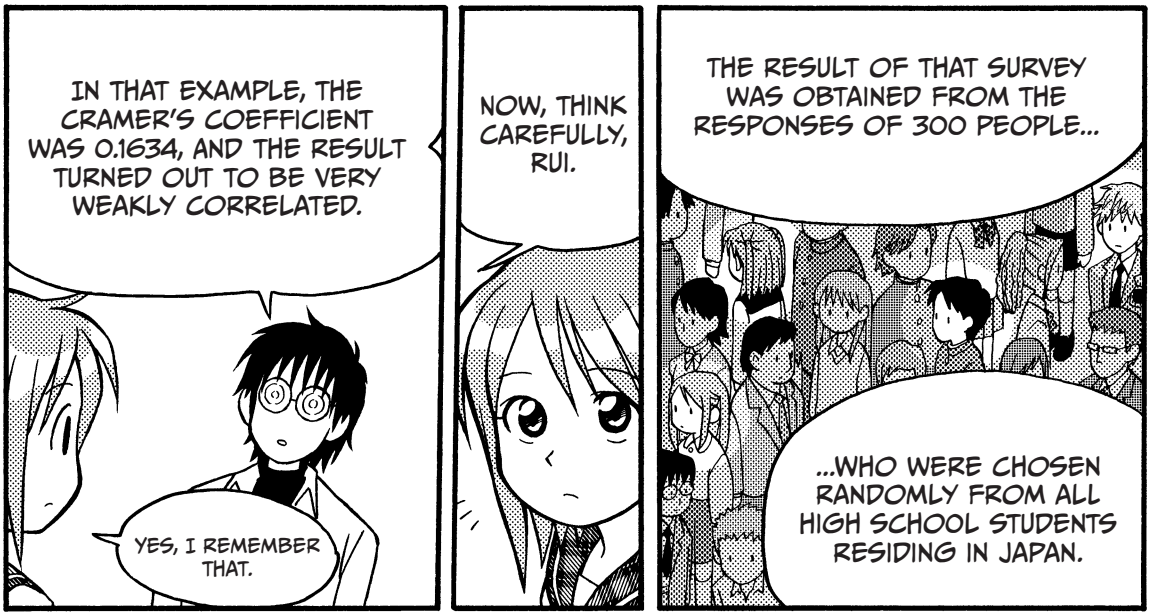
- There are no statistical standards for the correlation coefficient, correlation ratio, and Cramer's coefficient, such as "the two variables have a strong correlation if the value is above a certain benchmark."

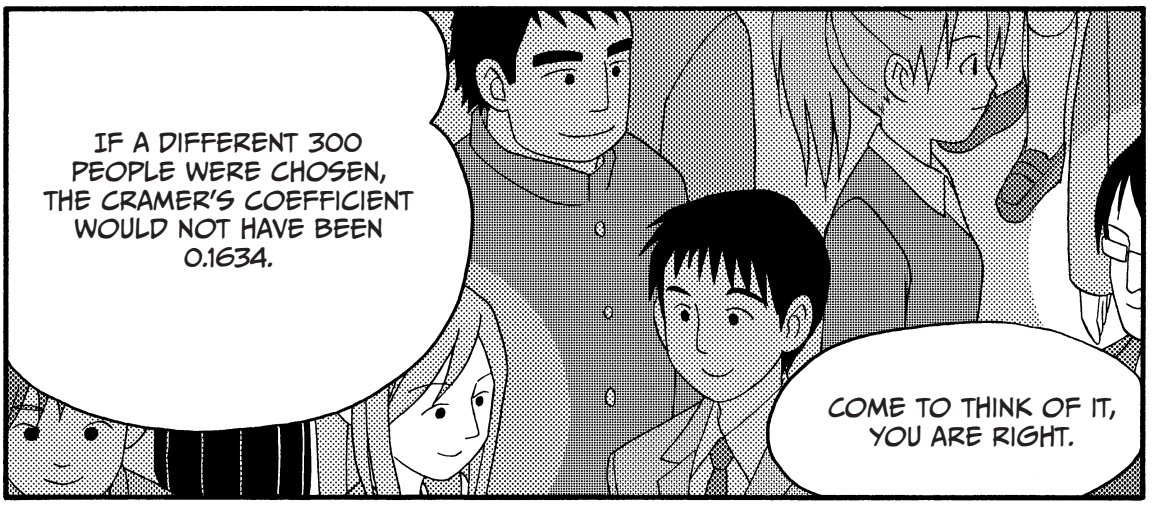
7

LET'S EXPLORE THE HYPOTHESIS TESTS

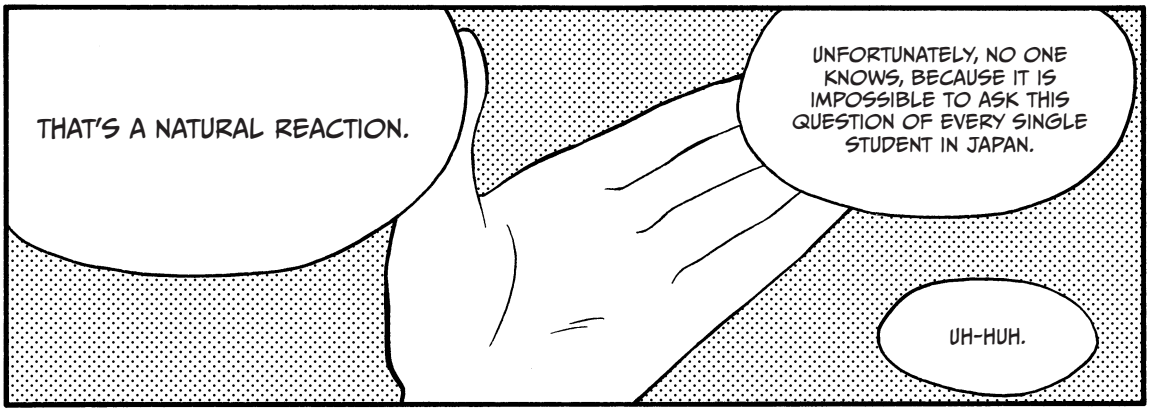

1. HYPOTHESIS TESTS
NOW, TODAY'S LESSON IS...
HEY! DON'T YOU HAVE ANYTHING TO SAY?
TAH-RAH!
WELL, WELL, EXCUSE ME FOR NOT NOTICING. IS THAT YOUR NEW UNIFORM YOU WERE TALKING ABOUT THE OTHER DAY?
YEAH! THIS IS STILL A TRIAL PRODUCT, BUT I AM SHOWING YOU AS A SPECIAL FAVOR.
WELL, THANK YOU... YOU LOOK GOOD IN IT.
THANKS.
SO, WHAT IS TODAY'S LESSON?

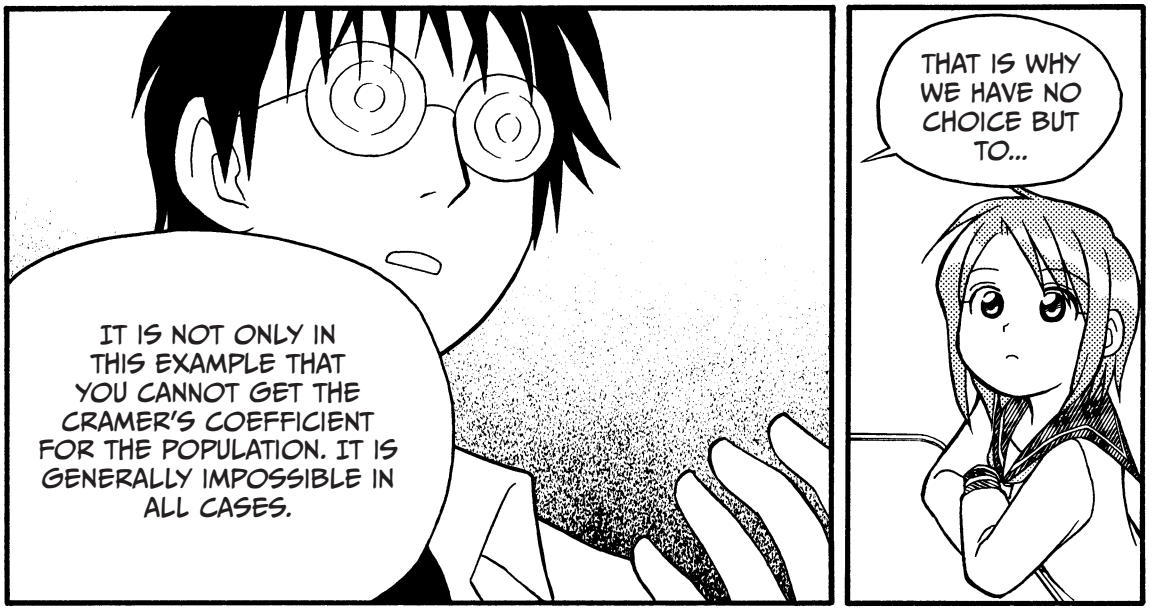

REMEMBER LEARNING ABOUT THE CRAMER'S COEFFICIENT IN OUR LAST LESSON?
高校生300人にききました
告白されるとしたらどの方法でされたい?
*WE ASKED 300 HIGH SCHOOL STUDENTS, "HOW WOULD YOU LIKE TO BE ASKED OUT?"
YOU MEAN THAT SURVEY ON HOW TO ASK OUT SOMEONE?

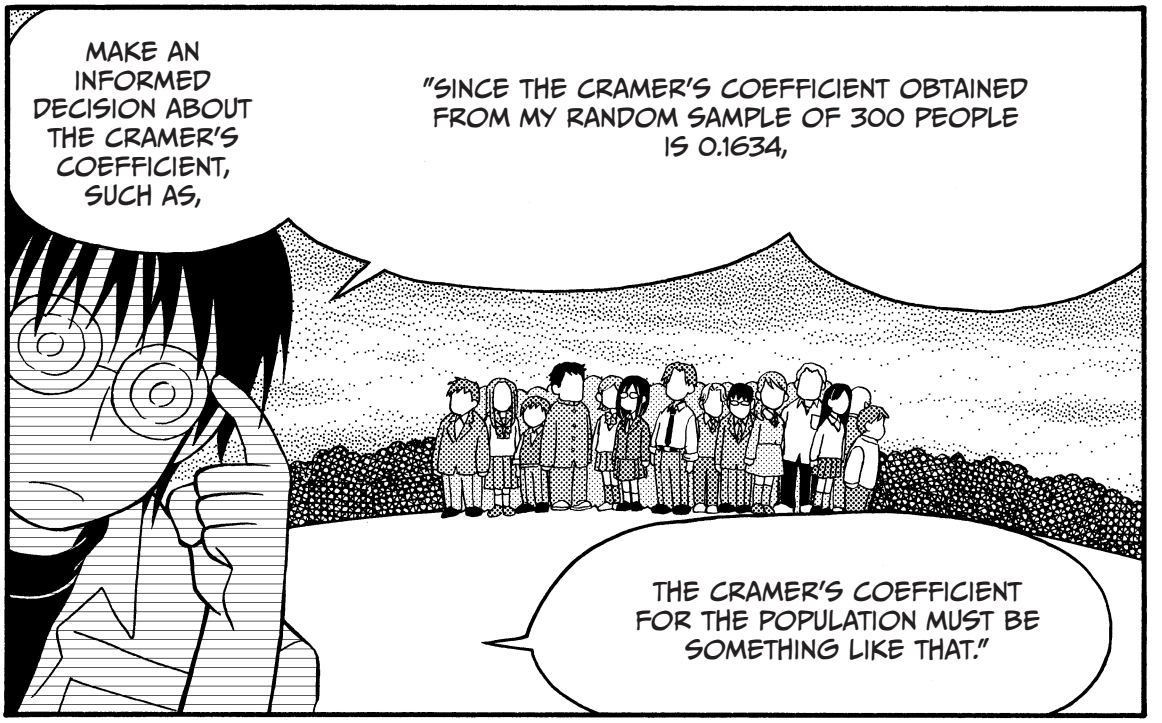

IN THAT EXAMPLE, THE CRAMER'S COEFFICIENT WAS 0.1634, AND THE RESULT TURNED OUT TO BE VERY WEAKLY CORRELATED.
YES, I REMEMBER THAT.
NOW, THINK CAREFULLY, RUI.
THE RESULT OF THAT SURVEY WAS OBTAINED FROM THE RESPONSES OF 300 PEOPLE...
...WHO WERE CHOSEN RANDOMLY FROM ALL HIGH SCHOOL STUDENTS RESIDING IN JAPAN.


IF A DIFFERENT 300 PEOPLE WERE CHOSEN, THE CRAMER'S COEFFICIENT WOULD NOT HAVE BEEN 0.1634.
COME TO THINK OF IT, YOU ARE RIGHT.


DO YOU HAVE ANY IDEA WHAT THE VALUE OF THE CRAMER'S COEFFICIENT FOR THE POPULATION OF THIS EXAMPLE, ALL HIGH SCHOOL STUDENTS RESIDING IN JAPAN, WOULD BE IN THE FIRST PLACE?
HOW AM I SUPPOSED TO KNOW THAT?

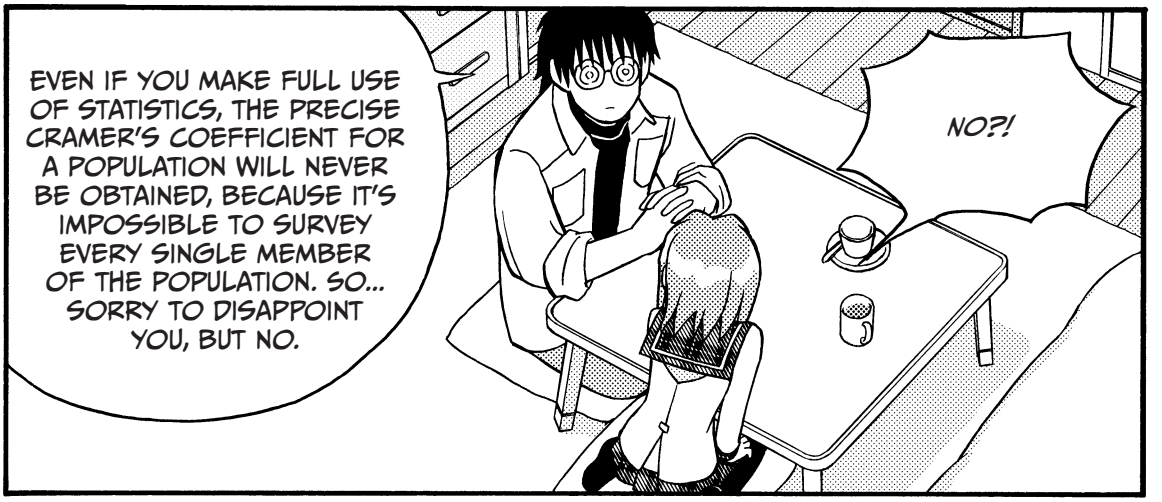

THAT'S A NATURAL REACTION.
UNFORTUNATELY, NO ONE KNOWS, BECAUSE IT IS IMPOSSIBLE TO ASK THIS QUESTION OF EVERY SINGLE STUDENT IN JAPAN.
UH-HUH.

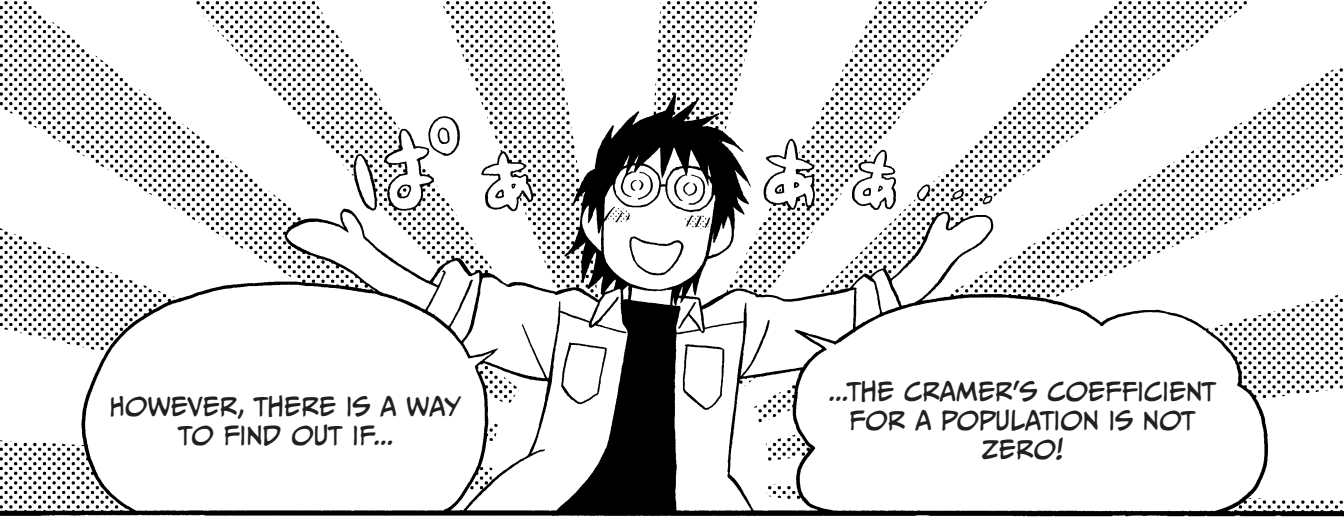

IT IS NOT ONLY IN THIS EXAMPLE THAT YOU CANNOT GET THE CRAMER'S COEFFICIENT FOR THE POPULATION. IT IS GENERALLY IMPOSSIBLE IN ALL CASES.
THAT IS WHY WE HAVE NO CHOICE BUT TO...

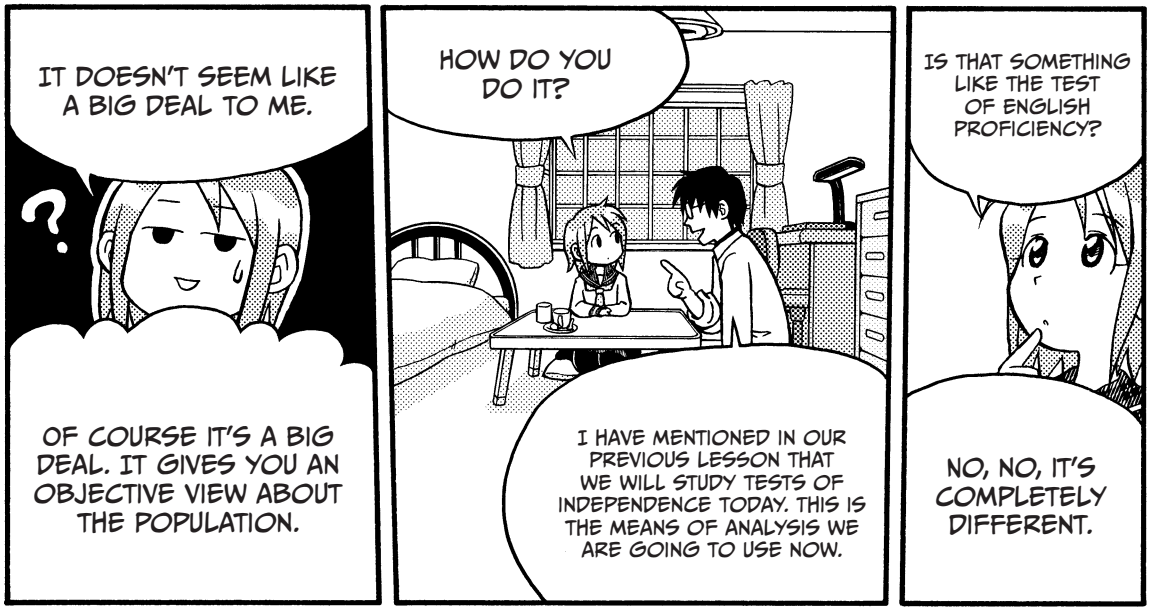

MAKE AN INFORMED DECISION ABOUT THE CRAMER'S COEFFICIENT, SUCH AS,
"SINCE THE CRAMER'S COEFFICIENT OBTAINED FROM MY RANDOM SAMPLE OF 300 PEOPLE IS 0.1634,
THE CRAMER'S COEFFICIENT FOR THE POPULATION MUST BE SOMETHING LIKE THAT."

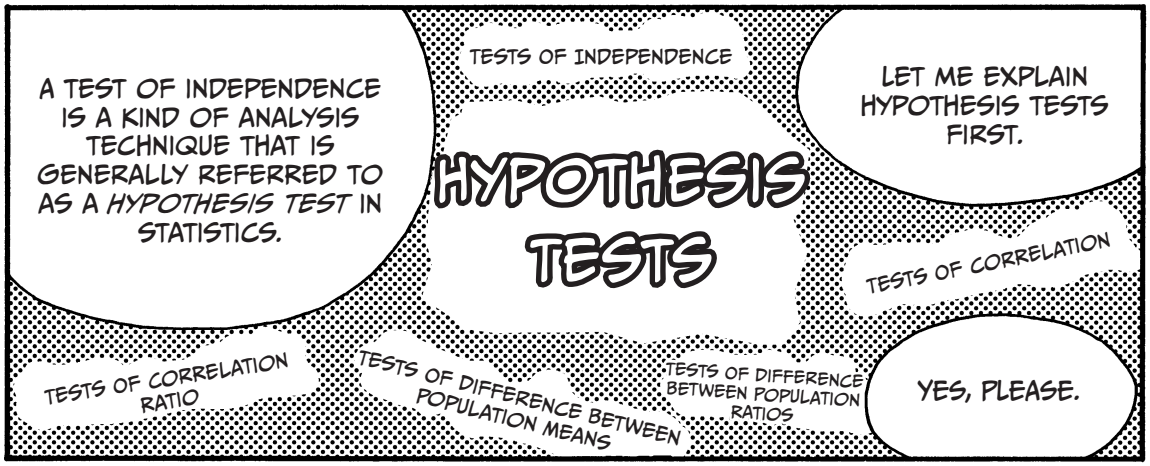

THAT'S NOT AS CONVINCING AS I THOUGHT.

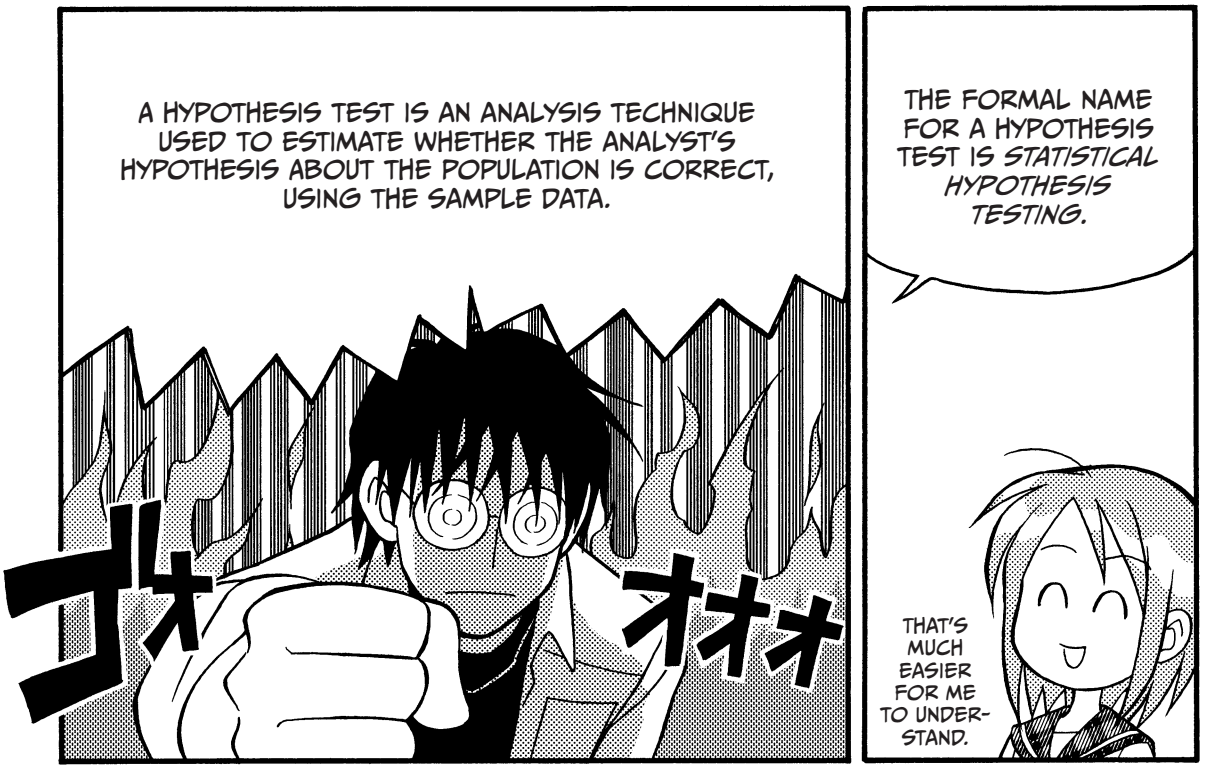

I BET THERE IS A WAY TO DO SOMETHING ABOUT IT USING STATISTICS AS IN OTHER CASES.
ビッ

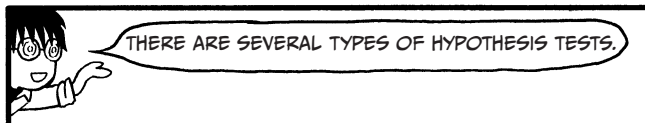

EVEN IF YOU MAKE FULL USE OF STATISTICS, THE PRECISE CRAMER'S COEFFICIENT FOR A POPULATION WILL NEVER BE OBTAINED, BECAUSE IT'S IMPOSSIBLE TO SURVEY EVERY SINGLE MEMBER OF THE POPULATION. SO... SORRY TO DISAPPOINT YOU, BUT NO.
NO?!


HOWEVER, THERE IS A WAY TO FIND OUT IF...
...THE CRAMER'S COEFFICIENT FOR A POPULATION IS NOT ZERO!


IT DOESN'T SEEM LIKE A BIG DEAL TO ME.
?
OF COURSE IT'S A BIG DEAL. IT GIVES YOU AN OBJECTIVE VIEW ABOUT THE POPULATION.
HOW DO YOU DO IT?
I HAVE MENTIONED IN OUR PREVIOUS LESSON THAT WE WILL STUDY TESTS OF INDEPENDENCE TODAY. THIS IS THE MEANS OF ANALYSIS WE ARE GOING TO USE NOW.
IS THAT SOMETHING LIKE THE TEST OF ENGLISH PROFICIENCY?
NO, NO, IT'S COMPLETELY DIFFERENT.


A TEST OF INDEPENDENCE IS A KIND OF ANALYSIS TECHNIQUE THAT IS GENERALLY REFERRED TO AS A *HYPOTHESIS TEST* IN STATISTICS.
TESTS OF INDEPENDENCE
HYPOTHESIS TESTS
LET ME EXPLAIN HYPOTHESIS TESTS FIRST.
TESTS OF CORRELATION
TESTS OF CORRELATION RATIO
TESTS OF DIFFERENCE BETWEEN POPULATION MEANS
TESTS OF DIFFERENCE BETWEEN POPULATION RATIOS
YES, PLEASE.

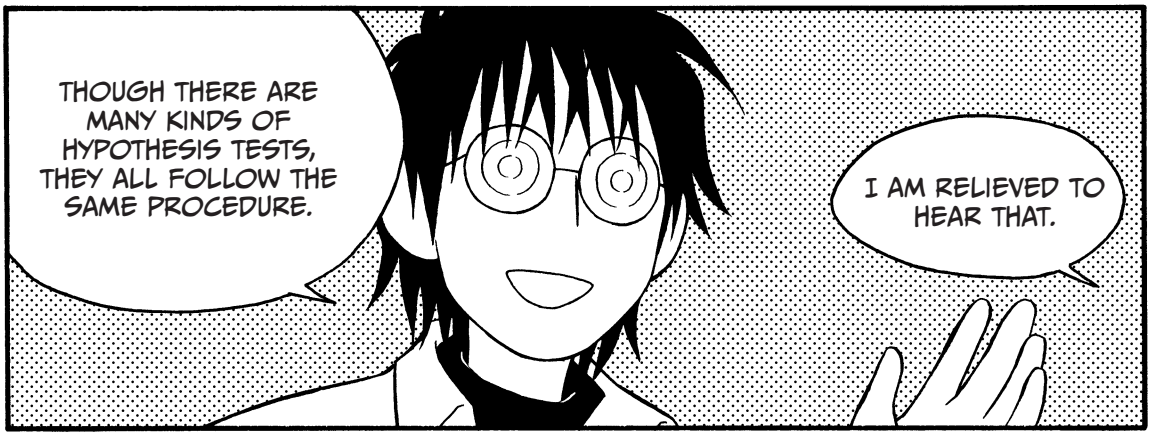




EXAMPLES OF HYPOTHESIS TESTS

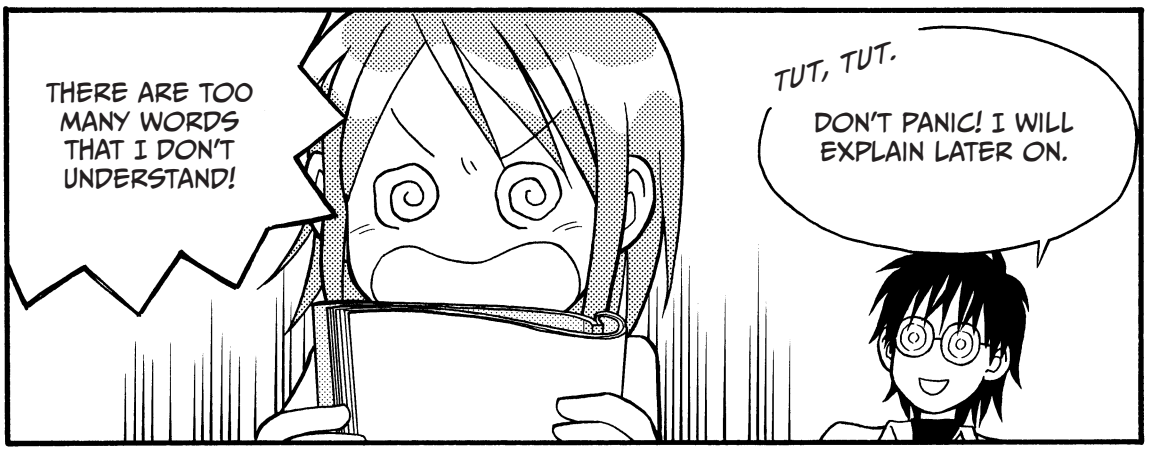
Name	Example of use
Tests of independence	Estimates whether the value of the Cramer's coefficient for sex and desired way of being asked out is zero for a population
Tests of correlation ratio	Estimates whether the value of the correlation ratio for favorite fashion brand and age is zero for a population
Tests of correlation	Estimates whether the correlation coefficient for amount spent on makeup and amount spent on clothes is zero for a population
Tests of difference between population means	Estimates whether allowances are different between high school girls in Tokyo and Osaka*
Tests of difference between population ratios	Estimates whether the approval rating of cabinet *X* is different between voters residing in urban areas and rural areas*

* Note that two populations are being considered.



PROCEDURE FOR A HYPOTHESIS TEST

Step 1	Define the population.
Step 2	Set up a null hypothesis and an alternative hypothesis.
Step 3	Select which hypothesis test to conduct.
Step 4	Determine the significance level.
Step 5	Obtain the test statistic from the sample data.
Step 6	Determine whether the test statistic obtained in step 5 is in the critical region.
Step 7	If the test statistic is in the critical region, you must reject the null hypothesis. If not, you fail to reject the null hypothesis.



NOW I WILL EXPLAIN TODAY'S MAIN TOPIC, A TEST OF INDEPENDENCE.

I TOLD YOU THAT A TEST OF INDEPENDENCE IS AN ANALYSIS TECHNIQUE USED TO ESTIMATE WHETHER THE CRAMER'S COEFFICIENT FOR A POPULATION IS ZERO.

YES, YOU DID.

TO PUT IT DIFFERENTLY, IT IS AN ANALYSIS TECHNIQUE USED TO ESTIMATE WHETHER TWO VARIABLES IN A CROSS TABULATION ARE CORRELATED.

I SEE! THAT IS WHY IT IS APPLIED TO SURVEY ANALYSES.

		DESIRED WAY OF BEING ASKED OUT			
		PHONE	E-MAIL	FACE TO FACE	SUM
SEX	F	34	61	53	148
	M	38	40	74	152
	SUM	72	101	127	

THIS TEST OF INDEPENDENCE IS ALSO CALLED A *CHI-SQUARE TEST.*

NOT THAT CHI-WHATEVER AGAIN!

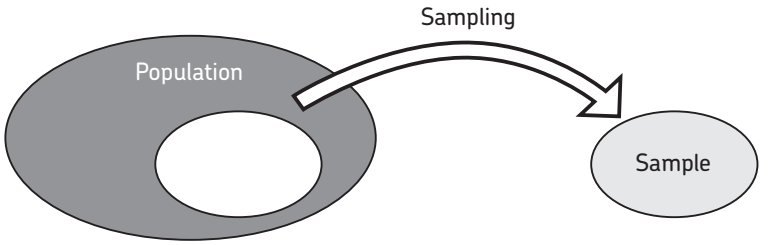
EXPLANATION

Pearson's chi-square test statistic (χ_0^2) and chi-square distribution

Before giving an actual example of a test of independence, I would like to explain an important fact that is fundamental to tests of independence. Though it is impossible to do this in reality, suppose the below experiment is conducted.

Step 1

Take a random sample of 300 students from the population "all high school students residing in Japan."



Step 2

Conduct the survey on page 127 with the 300 people chosen in step 1 to obtain the chi-square statistic (χ_0^2).

Step 3

Put the 300 people back into the population.

Step 4

Repeat steps 1 through 3 over and over.

In this experiment, if the value of the Cramer's coefficient for the population "all high school students residing in Japan" is 0, the graph of Pearson's chi-square test statistic (χ_0^2) turns out to be a chi-square distribution with 2 degrees of freedom. In other words, if the value of the Cramer's coefficient for the population "all high school students residing in Japan" is 0, then Pearson's chi-square test statistic (χ_0^2) follows a chi-square distribution with 2 degrees of freedom.

- See pages 130–133 for information on how to obtain Pearson's chi-square test statistic (χ_0^2).
- See page 100 for information on a chi-square distribution with 2 degrees of freedom.

We have actually conducted this experiment. In carrying out the experiment, we set the restrictions below.

- As it is impossible to experiment with the actual population of "all high school students residing in Japan," the group of 10,000 people in Table 7-1 will be regarded as "all high school students residing in Japan" instead.
- We assume that the Cramer's coefficient for "all high school students residing in Japan" is 0. This means that the ratio of those who prefer being asked out by phone to those who prefer being asked out by e-mail to those who prefer being asked out directly is equal for girls and boys (see page 135). The cross tabulation for Table 7-1 is Table 7-2.
- Since it is otherwise endless, we will stop repeating steps 1 through 3 after 10,000 times.

TABLE 7-1: DESIRED WAY OF BEING ASKED OUT
(ALL HIGH SCHOOL STUDENTS RESIDING IN JAPAN)

Respondent	Sex	Desired way of being asked out
1	Female	Face to face
2	Female	Phone
...	...	...
10,000	Male	E-mail

TABLE 7-2: CROSS TABULATION OF SEX AND DESIRED WAY OF BEING ASKED OUT

		Desired way of being asked out			Sum
		Phone	E-mail	Face to face	
Sex	Female	400	1,600	2,000	4,000
	Male	600	2,400	3,000	6,000
Sum		1,000	4,000	5,000	10,000

Table 7-3 shows the result of the experiment. Figure 7-1 is a histogram made according to Table 7-3.

TABLE 7-3: RESULT OF EXPERIMENT

Experiment	Pearson's chi-square test statistic (χ_0^2)
1	0.8598
2	0.7557
...	...
10,000	2.7953

FIGURE 7-1: A HISTOGRAM BASED ON TABLE 7-3 (RANGE OF CLASS = 1)

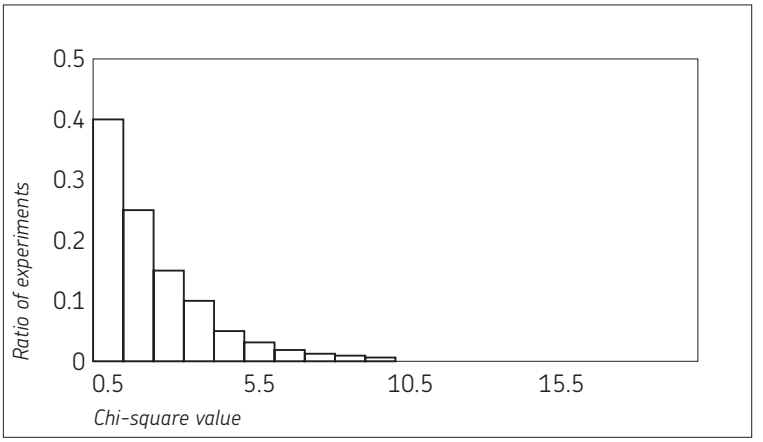
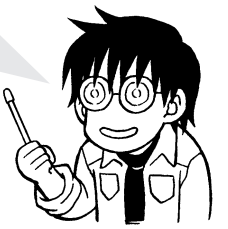


Figure 7-1 indeed looks similar to the graph on page 100, "2 Degrees of Freedom." It seems to be correct that Pearson's chi-square test statistic (χ_0^2) follows a chi-square distribution with 2 degrees of freedom. Though this has nothing to do with the experiment itself, here is one point to note. Two degrees of freedom comes from:

$$(2 - 1) \times (3 - 1) = 1 \times 2 = 2$$

↑ 2 patterns: female and male

↑ 3 patterns: phone, e-mail, and face to face



I will not go into why such a strange calculation is applied, as it is a topic much too advanced for the level of this book. But don't worry—even if you don't fully understand the calculation, you won't be at any disadvantage.

I SEE THAT THE VALUE OF THE CRAMER'S COEFFICIENT FOR "ALL HIGH SCHOOL STUDENTS RESIDING IN JAPAN" IS ZERO...WHICH MEANS THERE WAS NO RELATIONSHIP BETWEEN SEX AND DESIRED WAY OF BEING ASKED OUT.

I SUPPOSE...

THE RATIO OF PREFERENCE IS THE SAME FOR FEMALES AND MALES!

THEN, I SURVEY 300 PEOPLE SELECTED FROM "ALL HIGH SCHOOL STUDENTS RESIDING IN JAPAN."

REPEAT THAT OVER AND OVER AND OVER AND OVER!

Surveys

Surveys

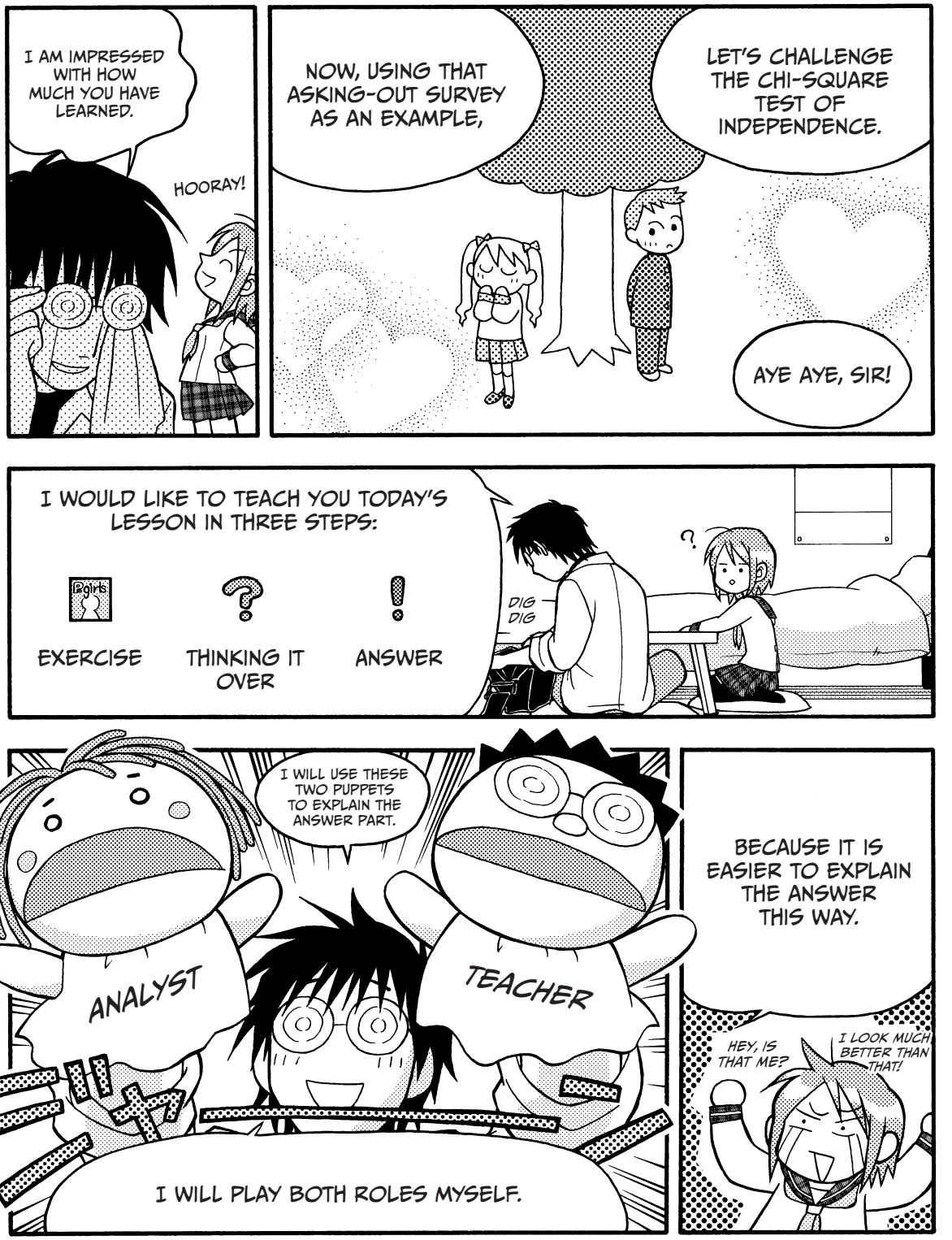
AFTER THAT, I CALCULATE PEARSON'S CHI-SQUARE TEST STATISTIC.

ADD $\left(\frac{\text{ACTUAL MEASUREMENT FREQUENCY} - \text{EXPECTED FREQUENCY}}{\text{EXPECTED FREQUENCY}}\right)^2$ IN EACH SQUARE

THE GRAPH OBTAINED FROM THE RESULT IS A CHI-SQUARE DISTRIBUTION WITH 2 DEGREES OF FREEDOM!

I FINALLY...

HAVE THE ANSWER!


I AM IMPRESSED WITH HOW MUCH YOU HAVE LEARNED.
HOORAY!
NOW, USING THAT ASKING-OUT SURVEY AS AN EXAMPLE,
LET'S CHALLENGE THE CHI-SQUARE TEST OF INDEPENDENCE.
AYE AYE, SIR!
I WOULD LIKE TO TEACH YOU TODAY'S LESSON IN THREE STEPS:
Boys
Girls
EXERCISE
THINKING IT OVER
ANSWER
DIG DIG
I WILL USE THESE TWO PUPPETS TO EXPLAIN THE ANSWER PART.
ANALYST
TEACHER
I WILL PLAY BOTH ROLES MYSELF.
BECAUSE IT IS EASIER TO EXPLAIN THE ANSWER THIS WAY.
HEY, IS THAT ME?
I LOOK MUCH BETTER THAN THAT!

EXERCISE

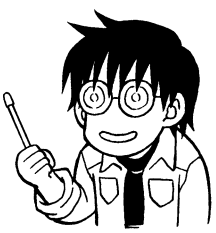
P-Girls Magazine decided to publish an article titled "We Asked 300 High School Students, 'How Would You Like to Be Asked Out?'" In order to prepare the article, a journalist randomly chose 300 people from all the high school students residing in Japan and took a survey. The table below is the result of this survey.

Respondent	Desired way of being asked out	Age	Sex
1	Face to face	17	Female
2	Phone	15	Female
...	...	...	...
300	E-mail	18	Male

The table below is the cross tabulation of sex and desired way of being asked out.

		Desired way of being asked out			Sum
		Phone	E-mail	Face to face	
Sex	Female	34	61	53	148
	Male	38	40	74	152
Sum		72	101	127	300

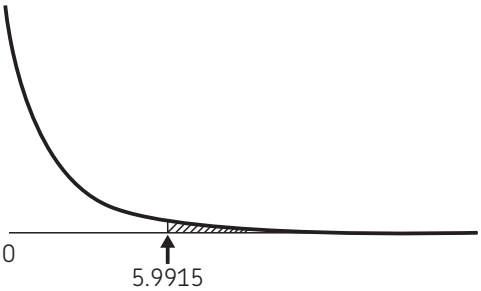
Using the chi-square test of independence, estimate if the Cramer's coefficient for sex and desired way of being asked out in the population "all high school students residing in Japan" is greater than 0. This is the same as estimating with a test of independence whether sex and desired way of being asked out are correlated. Remember that the significance level (explained later) is 0.05.



THINKING IT OVER

As explained on pages 152–154, Pearson's chi-square test statistic (χ_0^2) follows a chi-square distribution with 2 degrees of freedom if the null hypothesis states that the value of the Cramer's coefficient for the population "all high school students residing in Japan" is 0. If that's true, then the probability that χ_0^2 obtained from the 300 people who have been chosen randomly is 5.9915 or more is 0.05.

FIGURE 7-2: PROBABILITY THAT χ_0^2 IS 5.9915 OR MORE



This is clear from the table of chi-square distribution on page 103.

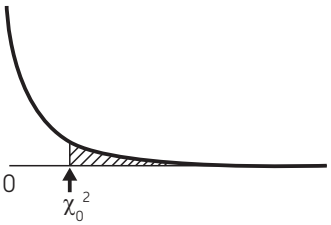
χ_0^2 for this exercise has already been calculated on page 132. It is 8.0091. True, this figure has been calculated based on data from 300 randomly chosen people, but doesn't this seem too large? Taking into consideration the comment on page 132, isn't it natural to assume that the Cramer's coefficient for the population "all high school students residing in Japan" is greater than 0?

Remember that the process for a chi-square test of independence (not limited to this exercise) goes like this:

1. Assume a null hypothesis that "the Cramer's coefficient for the population is 0" for the time being.
2. Calculate χ_0^2 from the sample data.
3. If χ_0^2 is too large, reject the null hypothesis and conclude that "the Cramer's coefficient for the population is greater than 0."

As χ_0^2 becomes larger, the probability shown as the shaded area in Figure 7-3 naturally becomes smaller.

FIGURE 7-3: PROBABILITY IN CORRESPONDENCE TO χ_0^2

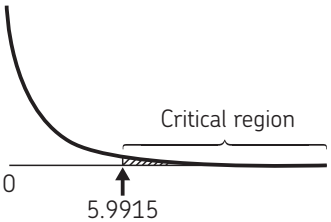


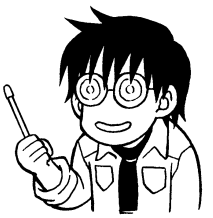
In chi-square tests of independence, if the probability shown as the shaded area in Figure 7-3 is less than or equal to the value called the *significance level*, reject the null hypothesis and conclude that "the Cramer's coefficient for the population is greater than 0." In general, the significance level (also called the *alpha value* and expressed by the symbol α) is considered to be 0.05 or 0.01.

It is up to the analyst which significance level to use. Suppose we decide to use 0.05 as the significance level in this case. The significance level is in fact the probability expressed as the shaded area in Figure 7-3.

The shaded area in Figure 7-4 is called the *critical region*.

FIGURE 7-4: CRITICAL REGION
(WHEN SIGNIFICANCE LEVEL IS 0.05)





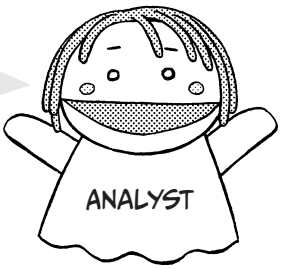
ANSWER

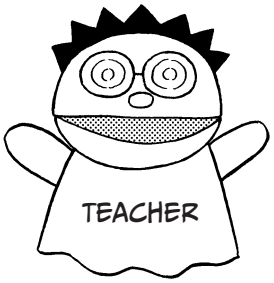
Step 1

Define the population.

The population is:

ALL HIGH SCHOOL STUDENTS RESIDING IN JAPAN





In this exercise, the population was defined as "all high school students residing in Japan." Thus, in this particular exercise, step 1 is unnecessary.

However, for "Tests of difference between population ratios" in the table on page 149, the populations in question are "voters residing in urban areas" and "voters residing in rural areas." Where are the urban areas exactly? Are they Tokyo and Osaka? Or are they the capitals of the prefectures? This must be specified by the analyst.

I repeat: When you are actually doing a hypothesis test, you must determine the population. No matter which hypothesis test you are trying to carry out, you must not fail to properly define the population.

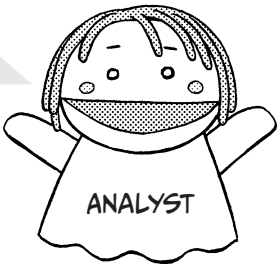
Otherwise, you might fall into a situation in which you are lost, wondering, "What was I trying to estimate?" Lots of statisticians fall into traps like this. Take great care about this point.

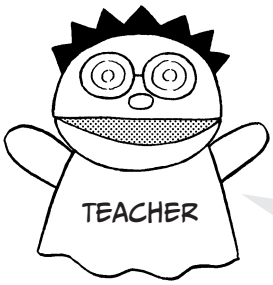
Step 2

Set up a null hypothesis and an alternative hypothesis.

The null hypothesis is: "The Cramer's coefficient for the population is 0. In other words, sex and desired way of being asked out are not correlated."

The alternative hypothesis is: "The Cramer's coefficient for the population is greater than 0. In other words, sex and desired way of being asked out are correlated."



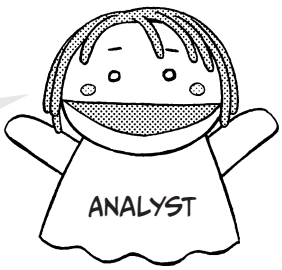


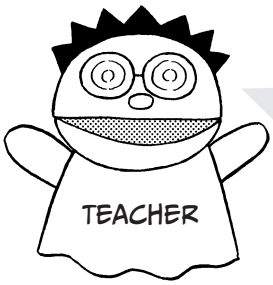
An explanation of null hypotheses and alternative hypotheses is given on page 170.

Step 3

Choose which hypothesis test to do.

I am going to do a chi-square test of independence.





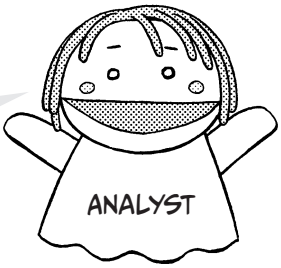
This exercise asks you to do a chi-square test of independence. So in this particular exercise, step 3 is unnecessary.

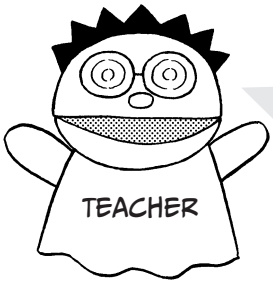
(When you are actually doing a hypothesis test and not an exercise, you must select the hypothesis test suitable for the objective of analysis on your own.)

Step 4

Determine the significance level.

I will use 0.05 as the significance level.



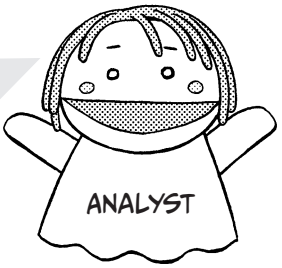


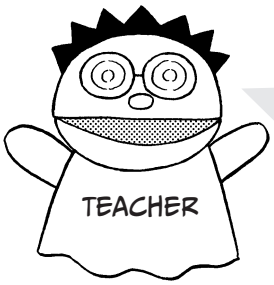
The exercise assigns 0.05 as the significance level, so in this particular exercise, step 4 is unnecessary. When you are actually doing a hypothesis test and not an exercise, you must determine the significance level. As mentioned earlier, normally either 0.05 or 0.01 is used. The smaller the P-value computed from the sample data, the stronger the evidence is against the null hypothesis. In general, the symbol α is used to express the significance level (alpha value).

Step 5

Calculate the test statistic from the sample data.

I am trying to do a chi-square test of independence. Thus the test statistic is Pearson's chi-square test statistic (χ_0^2). The value of χ_0^2 for this exercise has already been calculated on page 132: $\chi_0^2 = 8.0091$.





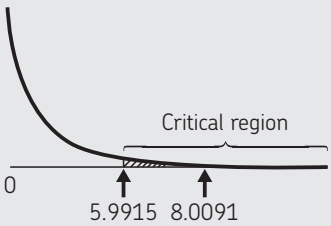
The test statistic is obtained from a function that calculates a single value from the sample data. Different kinds of hypothesis tests have different test statistics. As mentioned above, the value for a test of independence is χ_0^2, and in the case of tests of correlation (see page 149), the test statistic is as below.

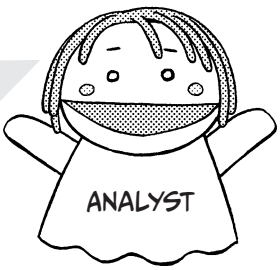
$$\frac{\text{correlation coefficient}^2 \times \sqrt{\text{number of values} - 2}}{1 - \sqrt{\text{correlation coefficient}^2}}$$

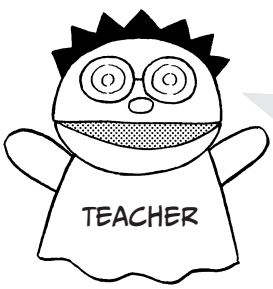
Step 6

Determine whether or not the test statistic from step 5 is in the critical region.

Pearson's chi-square test statistic (χ_0^2) is 8.0091. As the significance level (α) is 0.05, the critical region is 5.9915 or above, according to the table of chi-square distribution on page 103. As shown in the chart below, the test statistic is within the critical region.





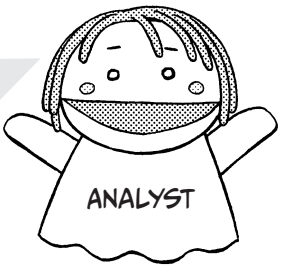


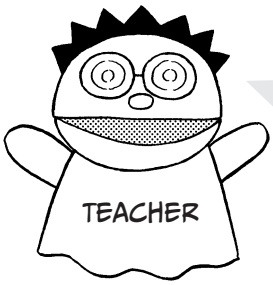
The critical region changes depending on the significance level (α). If α in this exercise was 0.01 instead of 0.05, the critical region would be 9.2104 or above, according to the table of chi-square distribution on page 103.

Step 7

If the test statistic is in the critical region in step 6, you reject the null hypothesis. If not, you fail to reject the null hypothesis. In this case, the test statistic was in the critical region.

Thus the alternative hypothesis, "the Cramer's coefficient for the population is greater than 0," is correct!

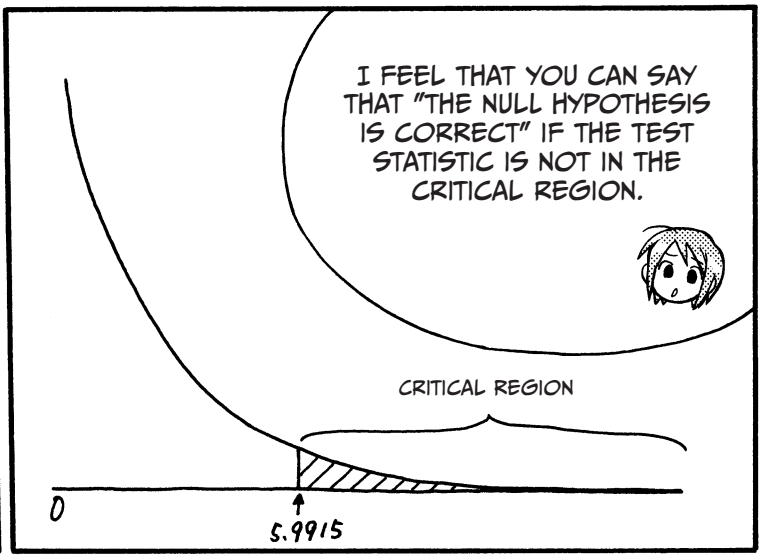


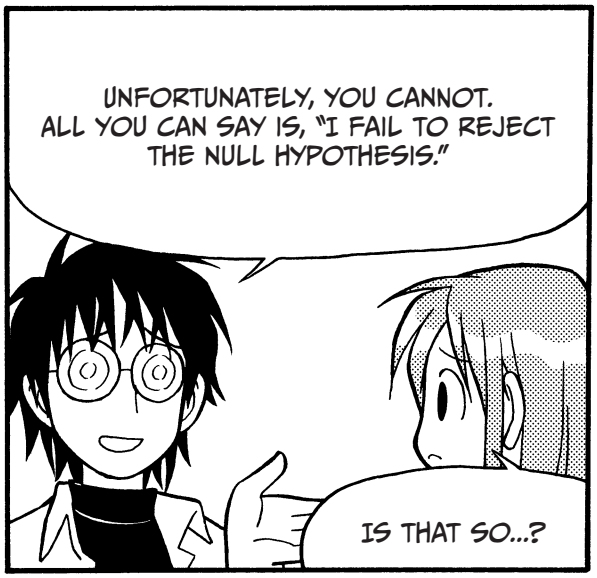


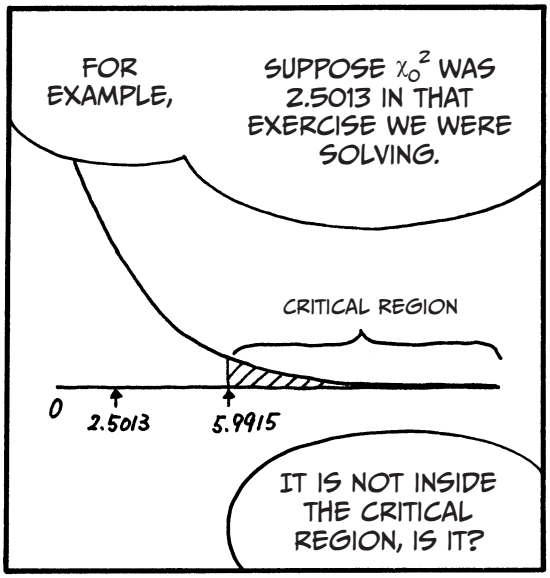
You cannot conclude that the alternative hypothesis is absolutely correct in a hypothesis test, even if the test statistic is within the critical region. The only conclusion you can make is, "I would like to say that the alternative hypothesis is 'absolutely' correct . . . but there is, at most, a ($\alpha \times 100$)% possibility that the null hypothesis is correct."


THAT'S ABOUT ALL.
I SEE...


BUT STILL, SOMETHING ABOUT STEP 7 SEEMS RATHER UNCONVINCING.
?

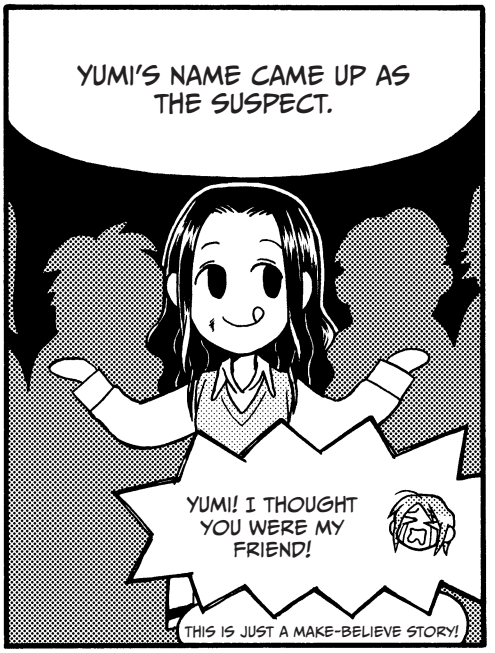

I FEEL THAT YOU CAN SAY THAT "THE NULL HYPOTHESIS IS CORRECT" IF THE TEST STATISTIC IS NOT IN THE CRITICAL REGION.
CRITICAL REGION
0
5.9915

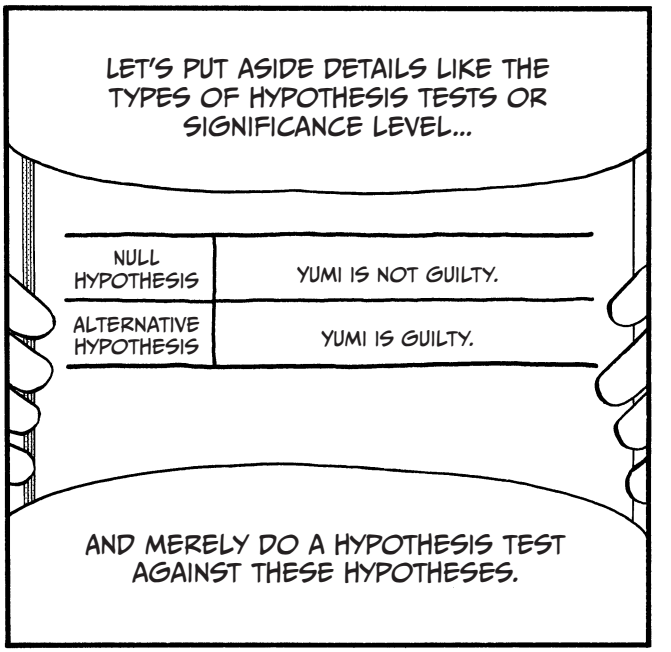

UNFORTUNATELY, YOU CANNOT. ALL YOU CAN SAY IS, "I FAIL TO REJECT THE NULL HYPOTHESIS."
IS THAT SO...?

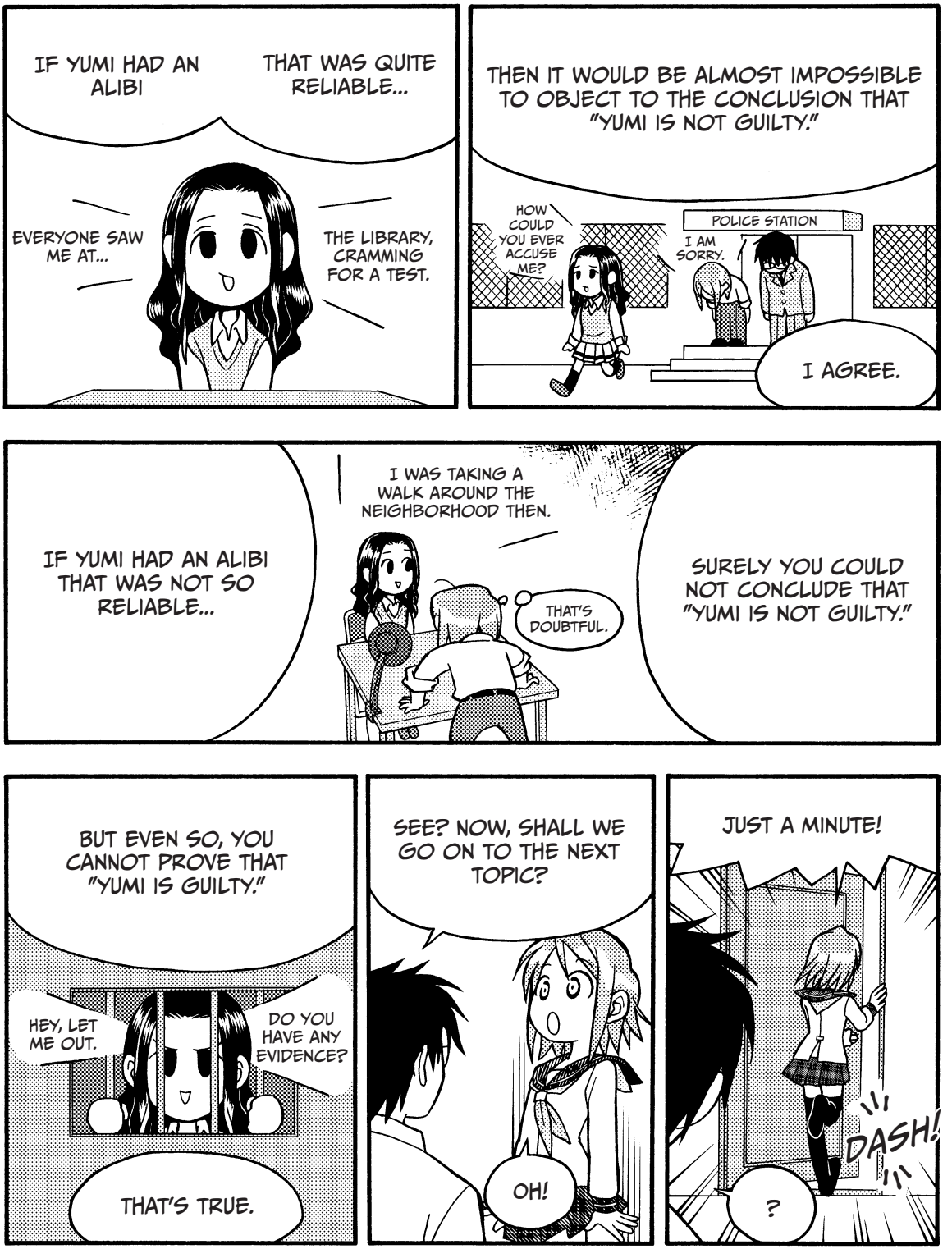

FOR EXAMPLE,
SUPPOSE χ_0^2 WAS 2.5013 IN THAT EXERCISE WE WERE SOLVING.
CRITICAL REGION
0
2.5013
5.9915
IT IS NOT INSIDE THE CRITICAL REGION, IS IT?

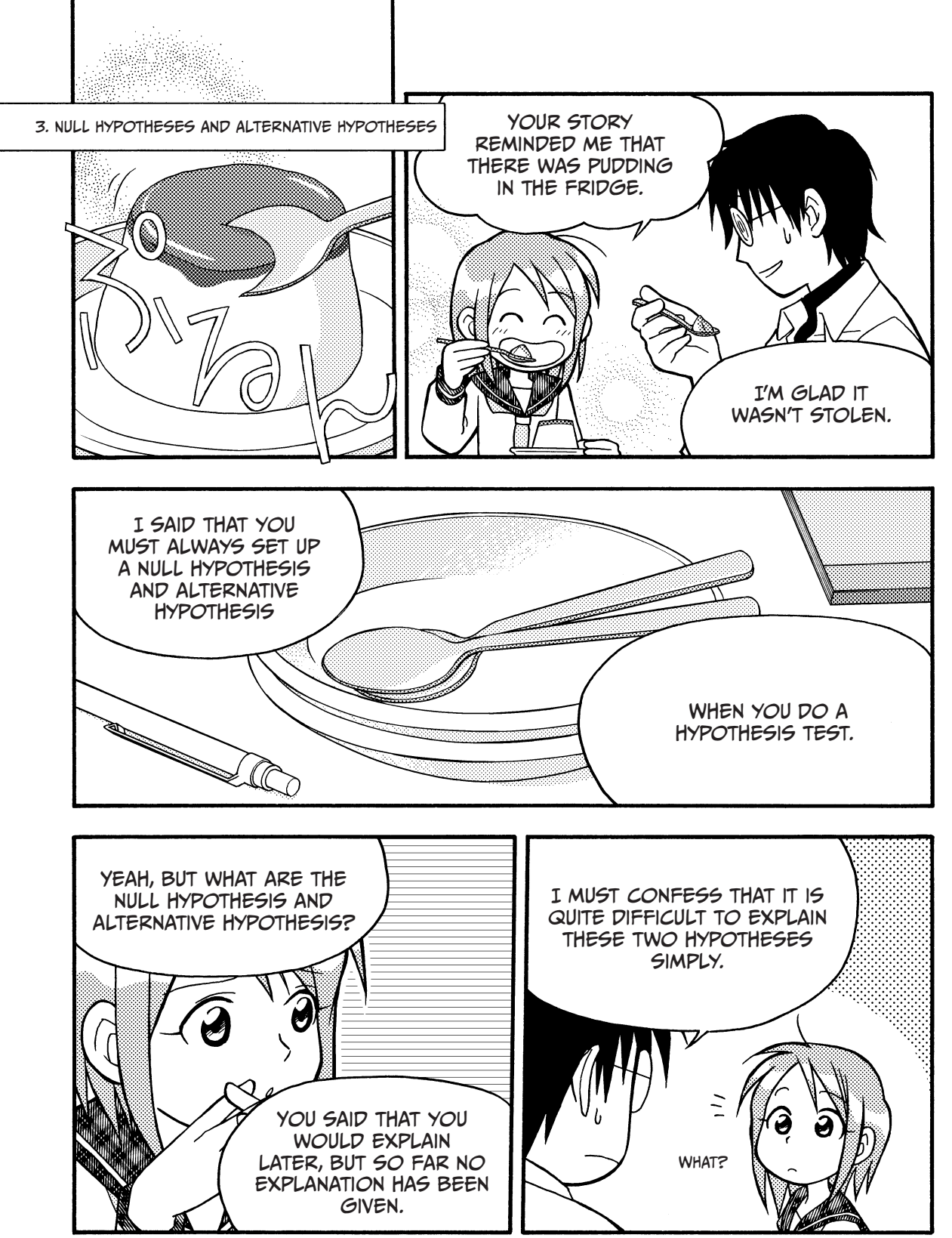

AS A MATTER OF COURSE, YOU CANNOT ACCEPT THE ALTERNATIVE HYPOTHESIS THAT "THE CRAMER'S COEFFICIENT FOR THE POPULATION IS GREATER THAN 0."
フムフム〜
HOWEVER, THERE IS NO WAY YOU CAN AFFIRM THE NULL HYPOTHESIS THAT "THE CRAMER'S COEFFICIENT FOR THE POPULATION IS 0."

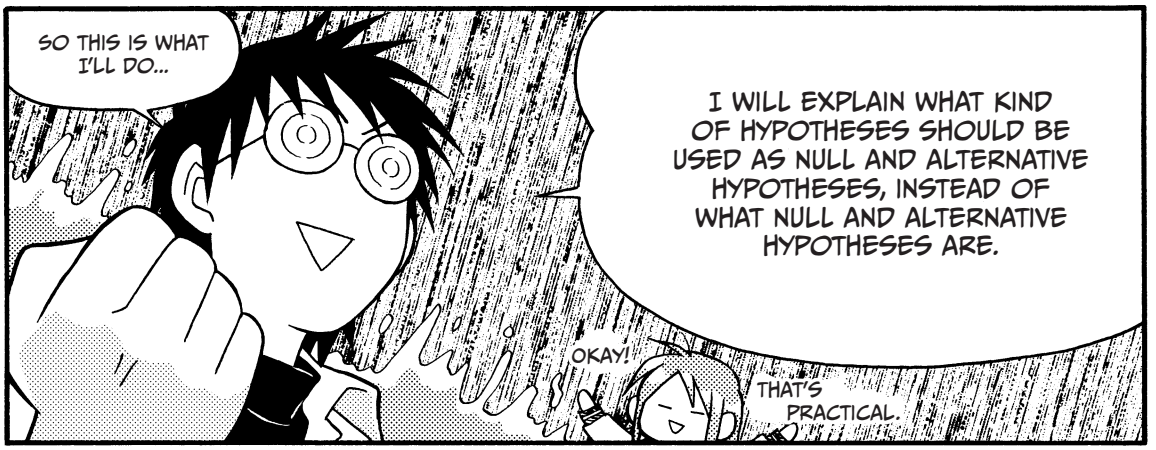

HERE'S A LITTLE STORY THAT MIGHT HELP YOU UNDERSTAND IT BETTER.
SUPPOSE SOMEONE STOLE AND ATE A PUDDING YOU WERE SAVING TO EAT LATER.
WHO STOLE MY PUDDING!?


YUMI'S NAME CAME UP AS THE SUSPECT.
YUMI! I THOUGHT YOU WERE MY FRIEND!
THIS IS JUST A MAKE-BELIEVE STORY!


LET'S PUT ASIDE DETAILS LIKE THE TYPES OF HYPOTHESIS TESTS OR SIGNIFICANCE LEVEL...
NULL HYPOTHESIS
YUMI IS NOT GUILTY.
ALTERNATIVE HYPOTHESIS
YUMI IS GUILTY.
AND MERELY DO A HYPOTHESIS TEST AGAINST THESE HYPOTHESES.


IF YUMI HAD AN ALIBI
THAT WAS QUITE RELIABLE...
EVERYONE SAW ME AT...
THE LIBRARY, CRAMMING FOR A TEST.
THEN IT WOULD BE ALMOST IMPOSSIBLE TO OBJECT TO THE CONCLUSION THAT "YUMI IS NOT GUILTY."
HOW COULD YOU EVER ACCUSE ME?
POLICE STATION
I AM SORRY.
I AGREE.
IF YUMI HAD AN ALIBI THAT WAS NOT SO RELIABLE...
I WAS TAKING A WALK AROUND THE NEIGHBORHOOD THEN.
THAT'S DOUBTFUL.
SURELY YOU COULD NOT CONCLUDE THAT "YUMI IS NOT GUILTY."
BUT EVEN SO, YOU CANNOT PROVE THAT "YUMI IS GUILTY."
HEY, LET ME OUT.
DO YOU HAVE ANY EVIDENCE?
THAT'S TRUE.
SEE? NOW, SHALL WE GO ON TO THE NEXT TOPIC?
OH!
JUST A MINUTE!
DASH!
?

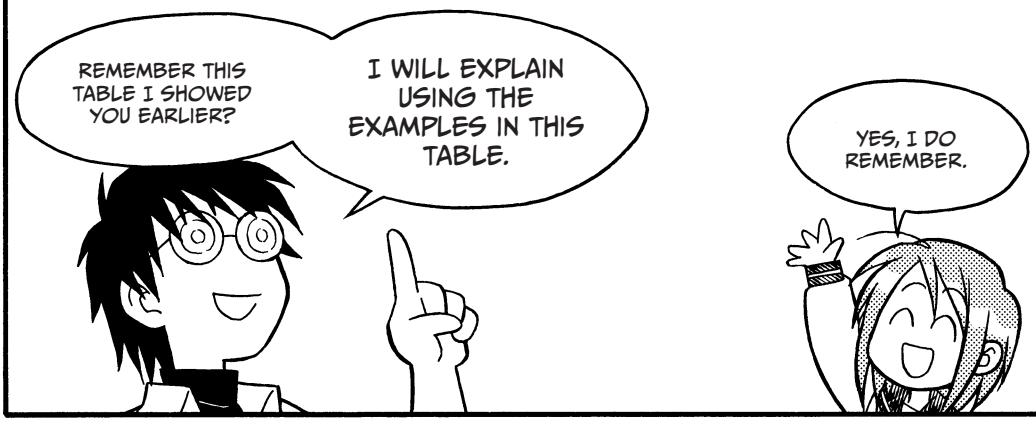

3. NULL HYPOTHESES AND ALTERNATIVE HYPOTHESES
YOUR STORY REMINDED ME THAT THERE WAS PUDDING IN THE FRIDGE.
I'M GLAD IT WASN'T STOLEN.
I SAID THAT YOU MUST ALWAYS SET UP A NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS
WHEN YOU DO A HYPOTHESIS TEST.
YEAH, BUT WHAT ARE THE NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS?
YOU SAID THAT YOU WOULD EXPLAIN LATER, BUT SO FAR NO EXPLANATION HAS BEEN GIVEN.
I MUST CONFESS THAT IT IS QUITE DIFFICULT TO EXPLAIN THESE TWO HYPOTHESES SIMPLY.
WHAT?



EXAMPLES OF HYPOTHESIS TESTS

Name	Example of use
Tests of independence	Estimates whether the value of the Cramer's coefficient for sex and desired way of being asked out is zero for a population
Tests of correlation ratio	Estimates whether the value of the correlation ratio for favorite fashion brand and age is zero for a population
Tests of correlation	Estimates whether the correlation coefficient for amount spent on makeup and amount spent on clothes is zero for a population
Tests of difference between population means	Estimates whether allowances are different between high school girls in Tokyo and Osaka*
Tests of difference between population ratios	Estimates whether the approval rating of cabinet *X* is different between voters residing in urban areas and rural areas*

* Note that two populations are being considered.



TESTS OF INDEPENDENCE

Null hypothesis	The Cramer's coefficient for sex and desired way of being asked out is 0 for a population.
Alternative hypothesis	The Cramer's coefficient for sex and desired way of being asked out is greater than 0 for a population.

TESTS OF CORRELATION RATIO

Null hypothesis	The correlation ratio for favorite fashion brand and age is 0 for a population.
Alternative hypothesis	The correlation ratio for favorite fashion brand and age is greater than 0 for a population.

TESTS OF CORRELATION

Null hypothesis	The correlation coefficient for amount spent on makeup and amount spent on clothes is 0 for a population.
Alternative hypothesis	The correlation coefficient for amount spent on makeup and amount spent on clothes is not 0 for a population. or The correlation coefficient for amount spent on makeup and amount spent on clothes is greater than 0 for a population. or The correlation coefficient for amount spent on makeup and amount spent on clothes is less than 0 for a population.

TESTS OF DIFFERENCE BETWEEN POPULATION MEANS

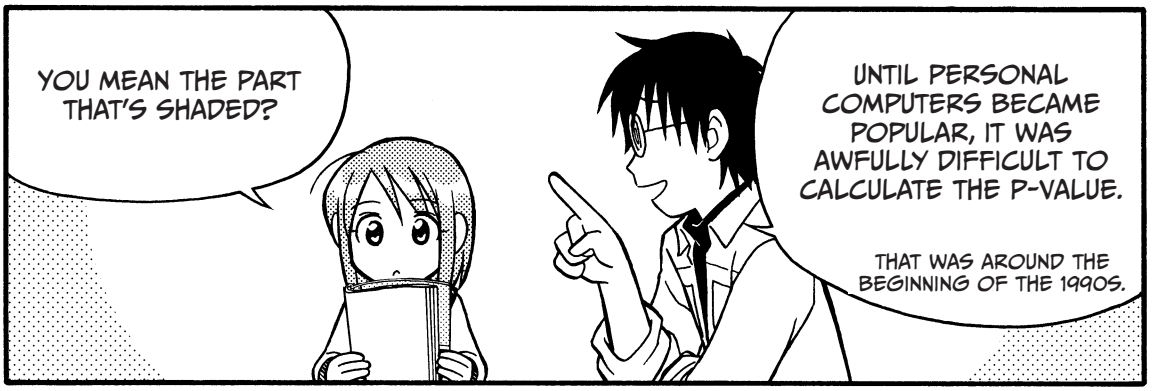
Null hypothesis	The allowances of high school girls in Tokyo and Osaka are the same.
Alternative hypothesis	The allowances of high school girls in Tokyo and Osaka are not the same. or The allowances of high school girls in Tokyo are larger than those of high school girls in Osaka. or The allowances of high school girls in Tokyo are smaller than those of high school girls in Osaka.

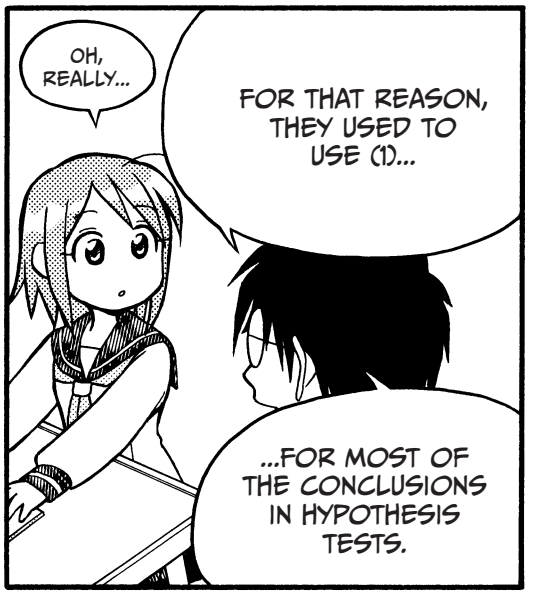
TESTS OF DIFFERENCE BETWEEN POPULATION RATIOS

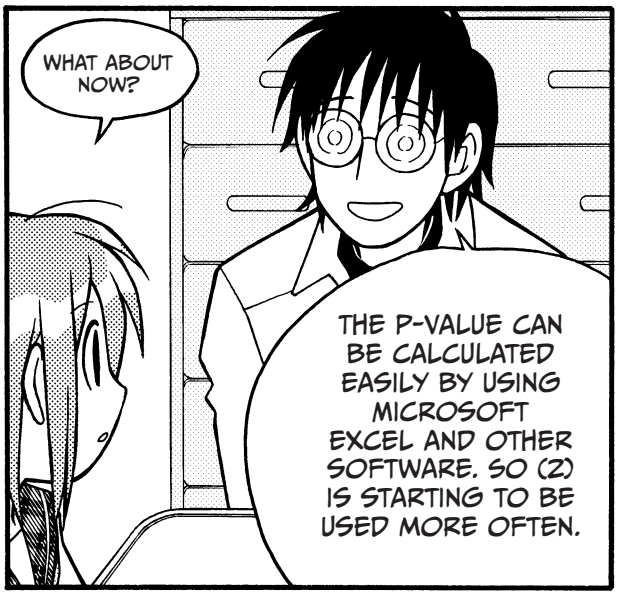
Null hypothesis	The approval ratings of cabinet *X* for voters residing in urban areas and rural areas are the same.
Alternative hypothesis	The approval ratings of cabinet *X* for voters residing in urban areas and rural areas are not the same. or The approval rating of cabinet *X* for voters residing in urban areas is higher than that of voters residing in rural areas. or The approval rating of cabinet *X* for voters residing in urban areas is lower than that of voters residing in rural areas.

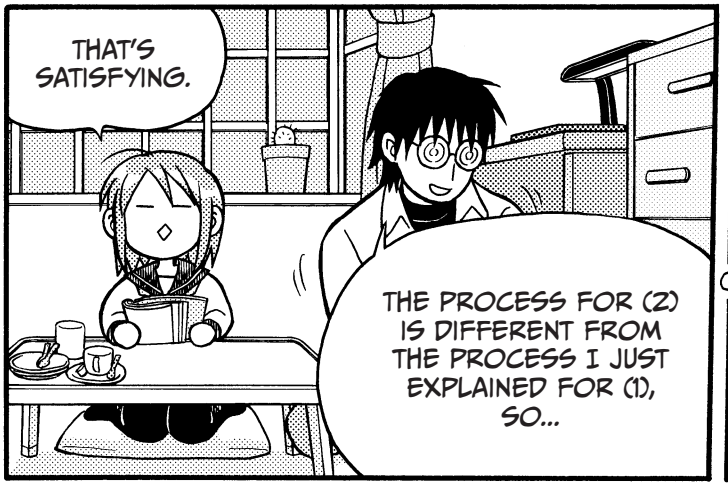

FOLLOWING ALONG SO FAR?
I HOPE YOU NOW UNDERSTAND WHY "THE CRAMER'S COEFFICIENT FOR THE POPULATION IS 'ALMOST' ZERO" IS NOT THE ONLY NULL HYPOTHESIS, AND THERE EXIST SOME NULL HYPOTHESES THAT SEEM DIFFICULT TO PROVE, SUCH AS "THE CRAMER'S COEFFICIENT FOR THE POPULATION IS ZERO."
YES I DO. SO MANY OF THEM ARE VERY EXTREME.
ALSO, DID YOU NOTICE THAT NULL HYPOTHESES ARE POSITIVE AND USE PHRASES LIKE "...IS..." OR "...ARE THE SAME"?
ALTERNATIVE HYPOTHESES ARE NEGATIVE, WITH PHRASES LIKE "...IS NOT..." OR "...ARE NOT THE SAME."
THAT'S TRUE.
THIS IS HOW YOU SHOULD THINK ABOUT THIS MATTER.
USE A HYPOTHESIS THAT SEEMS DIFFICULT TO PROVE AND IS POSITIVE FOR THE NULL HYPOTHESIS.
THAT'S CONVINCING.
THEN SET A HYPOTHESIS THAT IS OPPOSITE TO THE NULL HYPOTHESIS AS THE ALTERNATIVE HYPOTHESIS.

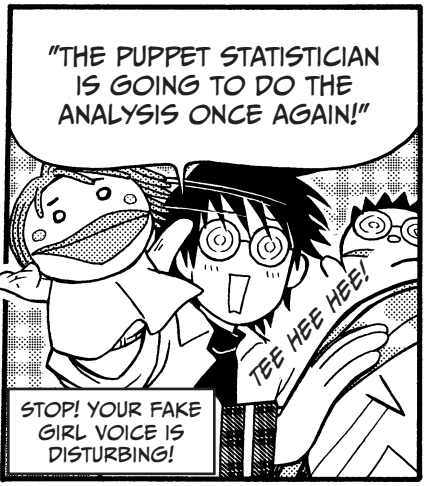

4. P-VALUE AND PROCEDURE FOR HYPOTHESIS TESTS
WHEN MAKING A CONCLUSION IN A HYPOTHESIS TEST...
(1) WHETHER THE TEST STATISTIC IS IN THE CRITICAL REGION
(2) WHETHER THE P-VALUE IS SMALLER THAN THE SIGNIFICANCE LEVEL
THERE ARE TWO WAYS TO MAKE A JUDGMENT.
YOU TOLD ME ABOUT THE FIRST ONE, BUT NOT THE SECOND.
WHAT'S THE P-VALUE?
THOUGH THERE ARE SOME DIFFERENCES DEPENDING ON WHICH HYPOTHESIS TEST YOU ARE DOING,
IN TESTS OF INDEPENDENCE, THE *P-VALUE*...
...IS A PROBABILITY THAT GIVES YOU A VALUE OF χ_0^2 THE SAME AS OR GREATER THAN WHAT HAS BEEN CALCULATED IN THE CASE IN QUESTION, WHEN THE NULL HYPOTHESIS IS TRUE.
IN THE PREVIOUS EXAMPLE...
0
$\chi_0^2 = 8.0091$
IT IS THE PROBABILITY SHOWN HERE.


YOU MEAN THE PART THAT'S SHADED?
UNTIL PERSONAL COMPUTERS BECAME POPULAR, IT WAS AWFULLY DIFFICULT TO CALCULATE THE P-VALUE.
THAT WAS AROUND THE BEGINNING OF THE 1990S.


OH, REALLY...
FOR THAT REASON, THEY USED TO USE (1)...
...FOR MOST OF THE CONCLUSIONS IN HYPOTHESIS TESTS.


WHAT ABOUT NOW?
THE P-VALUE CAN BE CALCULATED EASILY BY USING MICROSOFT EXCEL AND OTHER SOFTWARE. SO (2) IS STARTING TO BE USED MORE OFTEN.


THAT'S SATISFYING.
THE PROCESS FOR (2) IS DIFFERENT FROM THE PROCESS I JUST EXPLAINED FOR (1), SO...


"THE PUPPET STATISTICIAN IS GOING TO DO THE ANALYSIS ONCE AGAIN!"
TEE HEE HEE!
STOP! YOUR FAKE GIRL VOICE IS DISTURBING!

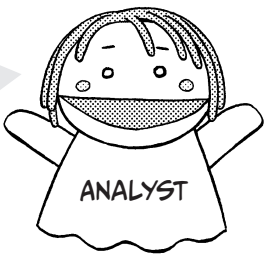
Step 6p

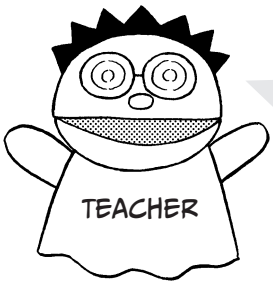
Determine whether or not the P-value corresponding to the test statistic obtained in step 5 is smaller than the significance level.

The significance level is 0.05. Since Pearson's chi-square test statistic (χ_0^2, which is the test statistic in this case) is 8.0091, the P-value is 0.0182.

$0.0182 < 0.05$

Thus the P-value is smaller.



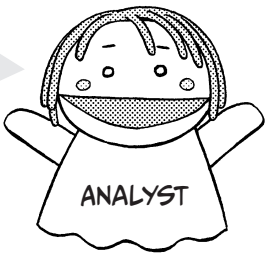


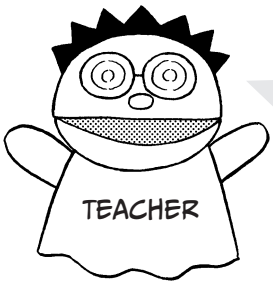
As mentioned before, you can calculate the P-value using Excel (though this depends on what type of hypothesis test you are doing). See page 208 for details.

Step 7p

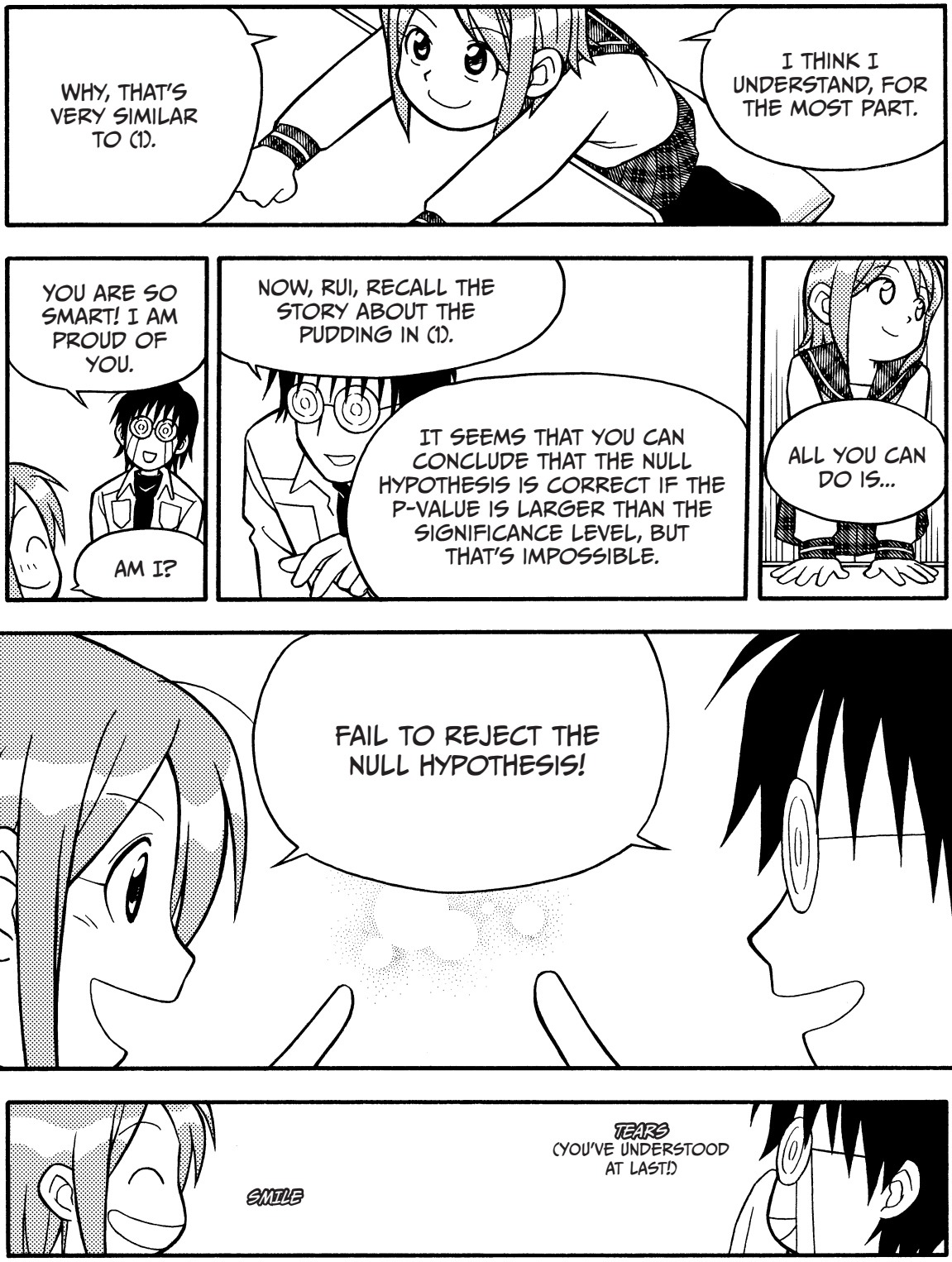
If the P-value is smaller than the significance level in step 6p, you reject the null hypothesis. If not, you fail to reject the null hypothesis.

The P-value was smaller than the significance level. Therefore, you conclude in favor of the alternative hypothesis, "the Cramer's coefficient for the population is greater than 0."





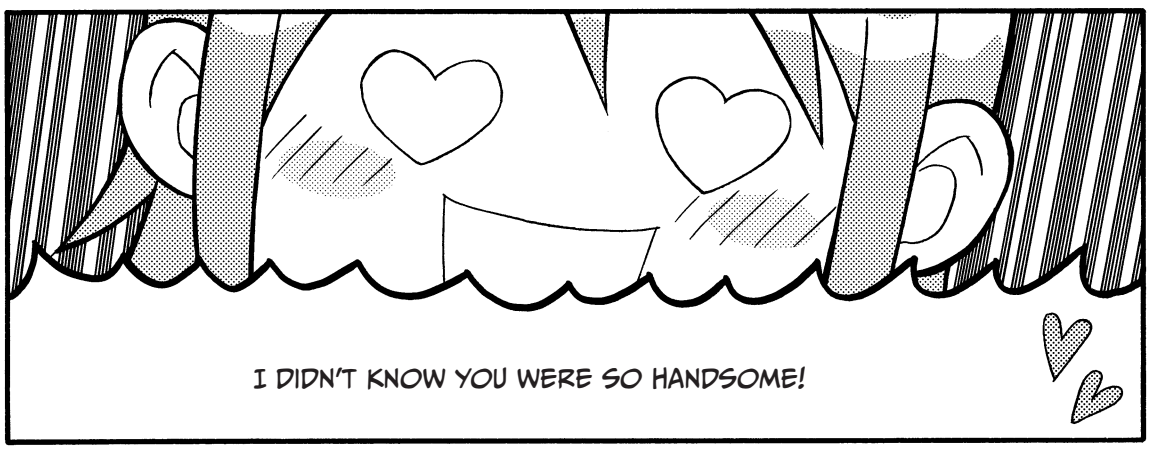
Even if the P-value was smaller than the significance level, you cannot really conclude that the alternative hypothesis is "absolutely" correct in a hypothesis test. The only conclusion you can make is: "I would like to say that the alternative hypothesis is 'absolutely' correct . . . but there is a ($\alpha \times 100$)% possibility that the null hypothesis is correct."


WHY, THAT'S VERY SIMILAR TO (1).
I THINK I UNDERSTAND, FOR THE MOST PART.
YOU ARE SO SMART! I AM PROUD OF YOU.
AM I?
NOW, RUI, RECALL THE STORY ABOUT THE PUDDING IN (1).
IT SEEMS THAT YOU CAN CONCLUDE THAT THE NULL HYPOTHESIS IS CORRECT IF THE P-VALUE IS LARGER THAN THE SIGNIFICANCE LEVEL, BUT THAT'S IMPOSSIBLE.
ALL YOU CAN DO IS...
FAIL TO REJECT THE NULL HYPOTHESIS!
SMILE
TEARS
(YOU'VE UNDERSTOOD AT LAST!)


YOU HAVE WORKED VERY HARD THROUGHOUT THESE LESSONS, RUI.
THANK YOU VERY MUCH, MR. YAMAMOTO.
ペコ
AT FIRST, STATISTICS SEEMED DIFFICULT TO ME, BUT I HAVE LEARNED SO MUCH ABOUT IT.
I AM BEGINNING TO FEEL THAT TABULATION OF SURVEYS AND STUFF LIKE THAT IS KIND OF FUN.
YOU ARE A STUDENT WORTH TEACHING.
I AM SO GLAD TO HEAR YOU SAY SO.
I WANT TO DO OTHER TYPES OF ANALYSIS AS WELL.
IF YOU ARE SO ENTHUSIASTIC, WHY DON'T YOU WORK IN THE FIELD MY COLLEAGUES AND I WORK IN?
OH! I ALMOST FORGOT. I STARTED STUDYING STATISTICS TO WORK WITH MR. IGARASHI...
I GUESS IT IS TIME FOR ME TO LEAVE.


I MUST GO SEE MR. IGARASHI.
WHAT? IGARASHI? HE'S GONE...
...ON VACATION. HE IS ON HIS HONEYMOON.
WHAT? HE'S MARRIED?!
THAT CAN'T BE TRUE... I HAVE BEEN STUDYING STATISTICS ALL THIS TIME IN VAIN...!?
ガーン
ガーン
RUI? I THOUGHT YOU WERE STUDYING FOR YOUR OWN SAKE..!
SHUT UP!
HEY, BE CAREFUL!
CRASH
SORRY...


ARE YOU ALL RIGHT?


I DIDN'T KNOW YOU WERE SO HANDSOME!


MR. YAMAMOTO, PLEASE DO KEEP TEACHING ME THIS AND THAT!
AND SO THEIR LESSONS WILL CONTINUE...
...MAYBE...MAYBE NOT.

5. TESTS OF INDEPENDENCE AND TESTS OF HOMOGENEITY

There is a hypothesis test very similar to a test of independence called a *test of homogeneity*. Below is an example of a test of homogeneity. As you read it, think about how it is different from a test of independence.

EXAMPLE

P-Girls Magazine published an article titled, "We Asked 300 High School Students, 'How Would You Like to Be Asked Out?'" The choices were phone, e-mail, or face to face.

> HYPOTHESIS: THE RATIO OF PHONE TO E-MAIL TO FACE-TO-FACE IS DIFFERENT BETWEEN HIGH SCHOOL GIRLS AND BOYS.

To find out if this hypothesis is true or not, a journalist actually conducted a survey by randomly choosing respondents from each of the two groups, "all high school girls residing in Japan" and "all high school boys residing in Japan." The table below is the result.

Respondent	Desired way of being asked out	Age	Sex
1	Face to face	17	Female
...	...	...	...
148	E-mail	16	Female
149	Phone	15	Male
...	...	...	...
300	E-mail	18	Male

The cross tabulation of sex and desired way of being asked out is the table below.

		Desired way of being asked out			
		Phone	E-mail	Face to face	Sum
Sex	**Female**	34	61	53	148
	Male	38	40	74	152
Sum		72	101	127	300

Estimate whether or not the hypothesis stated above is correct using a test of homogeneity. Use 0.05 as the significance level.

PROCEDURE

Step 1	Define the population.	The population in this case is "all high school girls residing in Japan" and "all high school boys residing in Japan."
Step 2	Set up a null hypothesis and an alternative hypothesis.	The null hypothesis is "the ratio of phone to e-mail to face to face is the same for high school girls and boys." The alternative hypothesis is "the ratio of phone to e-mail to face to face is different between high school girls and boys."
Step 3	Choose which hypothesis test to do.	A test of homogeneity will be applied.
Step 4	Determine the significance level.	The significance level is 0.05.
Step 5	Calculate the test statistic from the sample data.	A test of homogeneity is being used in this exercise. Therefore, the test statistic is Pearson's chi-square test statistic. The value of χ_0^2 in this exercise has already been calculated on page 132. $\chi_0^2 = 8.0091$ Pearson's chi-square test statistic (χ_0^2) in this exercise follows a chi-square distribution of degrees of freedom $(2 - 1) \times (3 - 1) = 1 \times 2 = 2$, if the null hypothesis is true.
Step 6	Determine whether the test statistic in step 5 is in the critical region.	The test statistic χ_0^2 is 8.0091. Since the significance level is 0.05, the critical region is 5.9915 or more, according to the table of chi-square distribution on page 103. The test statistic is within the critical region.
Step 7	If the test statistic is in the critical region in step 6, reject the null hypothesis and conclude in favor of the alternative. If not, fail to reject the null hypothesis.	The test statistic was within the critical region. Thus, you conclude in favor of the alternative hypothesis, "the ratio of phone to e-mail to face to face is different between high school girls and boys."

Don't you think that both the exercise and procedure are quite similar to those for a test of independence? Let's now look at the differences between tests of independence and tests of homogeneity. There are three things to note.

First, the population defined is different. There is only one population ("all high school students residing in Japan") in the former. In the latter, there are two populations: "all high school girls residing in Japan" and "all high school boys residing in Japan."

Also, the hypotheses are different. In the former,

Null hypothesis	The Cramer's coefficient for the population is 0. In other words, sex and desired way of being asked out are not correlated.
Alternative hypothesis	The Cramer's coefficient for the population is greater than 0. In other words, sex and desired way of being asked out are correlated.

In the latter,

Null hypothesis	The ratio of phone to e-mail to face to face is the same for high school girls and boys.
Alternative hypothesis	The ratio of phone to e-mail to face to face is different between high school girls and boys.

Finally, the order of procedure is different. In the former, the hypothesis is set after the data is collected, whereas the hypothesis is set before collecting the data in the latter.

As confirmed in the previous paragraph, tests of independence and tests of homogeneity have obvious differences. However, in practice, people tend to do tests of homogeneity when they are actually intending to do tests of independence, or vice versa. Be careful.

6. HYPOTHESIS TEST CONCLUSIONS

Up to this point, we have expressed the conclusion of a hypothesis test as follows:

> IF THE TEST STATISTIC IS IN THE CRITICAL REGION, YOU CAN CONCLUDE, "I REJECT THE NULL HYPOTHESIS." IF NOT, YOU CONCLUDE, "I FAIL TO REJECT THE NULL HYPOTHESIS."

But there are other ways to express the conclusions of hypothesis tests. They are summarized below.

TABLE 7-4: EXPRESSIONS OF HYPOTHESIS TEST CONCLUSIONS

When the test statistic is in the critical region	When the test statistic is not in the critical region
• Conclude in favor of the alternative hypothesis • Conclude that the result is statistically significant • Reject the null hypothesis	• Fail to reject the null hypothesis • Conclude that the result is not statistically significant • Accept the null hypothesis

The expressions "it is statistically significant" and "it is not statistically significant" seem to be popular in introductions to statistics. So why did we use an unpopular expression on purpose? I recognize that many beginners to hypothesis tests use the expression "it is significant" without actually understanding the meaning of the phrase. They seem to be merely confirming the test statistic or P-value. If you do not set a proper null and alternative hypothesis, the meaning of *significant* will be ambiguous. Beginners' definitions of their populations are frequently unclear as well.

I used to think I shouldn't be so strict with beginners. But it's impossible to make an accurate conclusion with uncertain null and alternative hypotheses. So in this book, I use the expressions "reject the null hypothesis" and "fail to reject the null hypothesis" so that you will get into the habit of thinking hard about your hypotheses.

EXERCISE AND ANSWER

EXERCISE

The table below is the same as the cross tabulation found on page 138.

		Preference for coffee or tea		Sum
		Coffee	Tea	
Type of food often ordered	Japanese	43	33	76
	European	51	53	104
	Chinese	29	41	70
Sum		123	127	250

Using a chi-square test of independence, estimate if the Cramer's coefficient for type of food often ordered and preference for coffee or tea in the population "people of age 20 or older residing in Japan" is greater than 0. This is the same as estimating whether there is a correlation between type of food often ordered and preference for coffee or tea. Use 0.01 as the significance level.

ANSWER

Step 1	Define the population.	The population in this case is "people of age 20 or older residing in Japan."
Step 2	Set up a null hypothesis and an alternative hypothesis.	The null hypothesis is "type of food often ordered and preference for coffee or tea are not correlated." The alternative hypothesis is "type of food often ordered and preference for coffee or tea are correlated."
Step 3	Choose which hypothesis test to do.	A chi-square test of independence will be applied.
Step 4	Determine the significance level.	The significance level is 0.01.
Step 5	Calculate the test statistic from the sample data.	A chi-square test of independence is being used in this exercise. Therefore, the test statistic is Pearson's chi-square test statistic (χ_0^2). The value of χ_0^2 in this exercise has already been calculated on page 141. $\chi_0^2 = 3.3483$
Step 6	Determine whether the test statistic obtained in step 5 is in the critical region.	The test statistic χ_0^2 is 3.3483. Because the significance level (α) is 0.01, the critical region is 9.2104 or above, according to the table of chi-square distribution on page 103. The test statistic is not within the critical region.
Step 7	If the test statistic is in the critical region in step 6, reject the null hypothesis. If not, fail to reject the null hypothesis.	The test statistic was not within the critical region. Thus, the null hypothesis "type of food often ordered and preference for coffee or tea are not correlated" cannot be rejected.

SUMMARY

- A *hypothesis test* is an analysis technique used to estimate whether the analyst's hypothesis about the population is correct using the sample data.
- The formal name for a hypothesis test is *statistical hypothesis testing*.
- Test statistics are obtained from a function that calculates a single value from the sample data.
- In general, 0.05 or 0.01 is used as the significance level.
- The *critical region* is an area that corresponds to the significance level (also called the *alpha value* and expressed by the symbol α).
- A *chi-square test of independence* is an analysis technique used to estimate whether the Cramer's coefficient for a population is 0. It can also be said that it is an analysis technique used to estimate whether the two variables in a cross tabulation are correlated.
- If the Cramer's coefficient for a population is 0, Pearson's chi-square test statistic follows a chi-square distribution.
- The *P-value* in a test of independence is a probability that gives a Pearson's chi-square test statistic equal to or greater than the value earned in the case when the null hypothesis is true.
- When making a conclusion in a hypothesis test, there are two bases of judgment:

  1. Whether the test statistic is in the critical region
  2. Whether the P-value is smaller than the significance level

- The process of analysis in any hypothesis test is the same as the process for the test of independence or any other kind of test. The actual procedure is:

Step 1	Define the population.
Step 2	Set up a null hypothesis and an alternative hypothesis.
Step 3	Choose which hypothesis test to do.
Step 4	Determine the significance level.
Step 5	Calculate the value of the test statistic from the sample data.
Step 6	Determine whether the test statistic obtained in step 5 is in the critical region.
Step 7	If the test statistic is in the critical region in step 6, reject the null hypothesis. If not, fail to reject the null hypothesis.
Step 6p	Determine whether the P-value corresponding to the test statistic obtained in step 5 is smaller than the significance level.
Step 7p	If the P-value is smaller than the significance level in step 6p, reject the null hypothesis. If not, fail to reject the null hypothesis.

LET'S CALCULATE USING EXCEL

This appendix contains instructions for calculating various statistics using Microsoft Excel. You'll learn how to do the following things:

1. Make a frequency table
2. Calculate arithmetic mean, median, and standard deviation
3. Make a cross tabulation
4. Calculate the standard score and the deviation score
5. Calculate the probability of the standard normal distribution
6. Calculate the point on the horizontal axis of the chi-square distribution
7. Calculate the correlation coefficient
8. Perform tests of independence

You can download these Excel files and follow along (get them at *http://www.nostarch.com/mg_statistics.htm*). Readers who are not familiar with Excel should try "Calculating Arithmetic Mean, Median, and Standard Deviation" on page 195 first.

1. MAKING A FREQUENCY TABLE

This exercise uses the ramen restaurant prices on page 33.

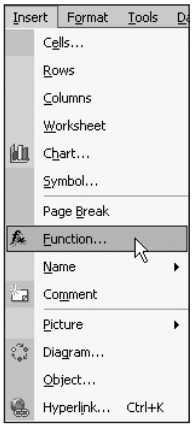
Step 1

Select cell J3.

	A	B	C	D	E	F	G	H	I	J
1		Price (yen)			Price (yen)					
2	Ramen shop 1	700		Ramen shop 26	780		Equal or greater	Less than	Equal or less	Frequency
3	Ramen shop 2	850		Ramen shop 27	590		500	600	599	
4	Ramen shop 3	600		Ramen shop 28	650		600	700	699	
5	Ramen shop 4	650		Ramen shop 29	580		700	800	799	
6	Ramen shop 5	980		Ramen shop 30	750		800	900	899	
7	Ramen shop 6	750		Ramen shop 31	800		900	1000	999	
8	Ramen shop 7	500		Ramen shop 32	550					
9	Ramen shop 8	890		Ramen shop 33	750					
10	Ramen shop 9	880		Ramen shop 34	700					
11	Ramen shop 10	700		Ramen shop 35	600					
12	Ramen shop 11	890		Ramen shop 36	800					
13	Ramen shop 12	720		Ramen shop 37	800					
14	Ramen shop 13	680		Ramen shop 38	880					
15	Ramen shop 14	650		Ramen shop 39	790					
16	Ramen shop 15	790		Ramen shop 40	790					
17	Ramen shop 16	670		Ramen shop 41	780					
18	Ramen shop 17	680		Ramen shop 42	600					
19	Ramen shop 18	900		Ramen shop 43	670					
20	Ramen shop 19	880		Ramen shop 44	680					
21	Ramen shop 20	720		Ramen shop 45	650					
22	Ramen shop 21	850		Ramen shop 46	890					
23	Ramen shop 22	700		Ramen shop 47	930					
24	Ramen shop 23	780		Ramen shop 48	650					
25	Ramen shop 24	850		Ramen shop 49	777					
26	Ramen shop 25	750		Ramen shop 50	700					

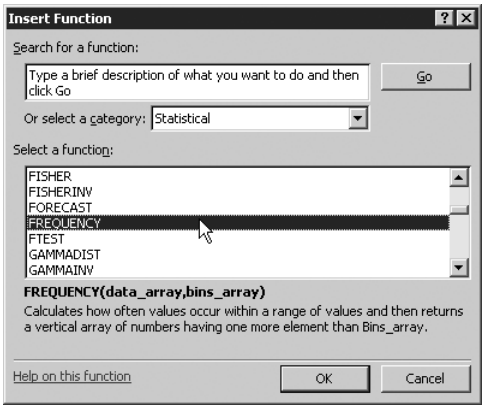
Step 2

Select **Insert ▸ Function**.



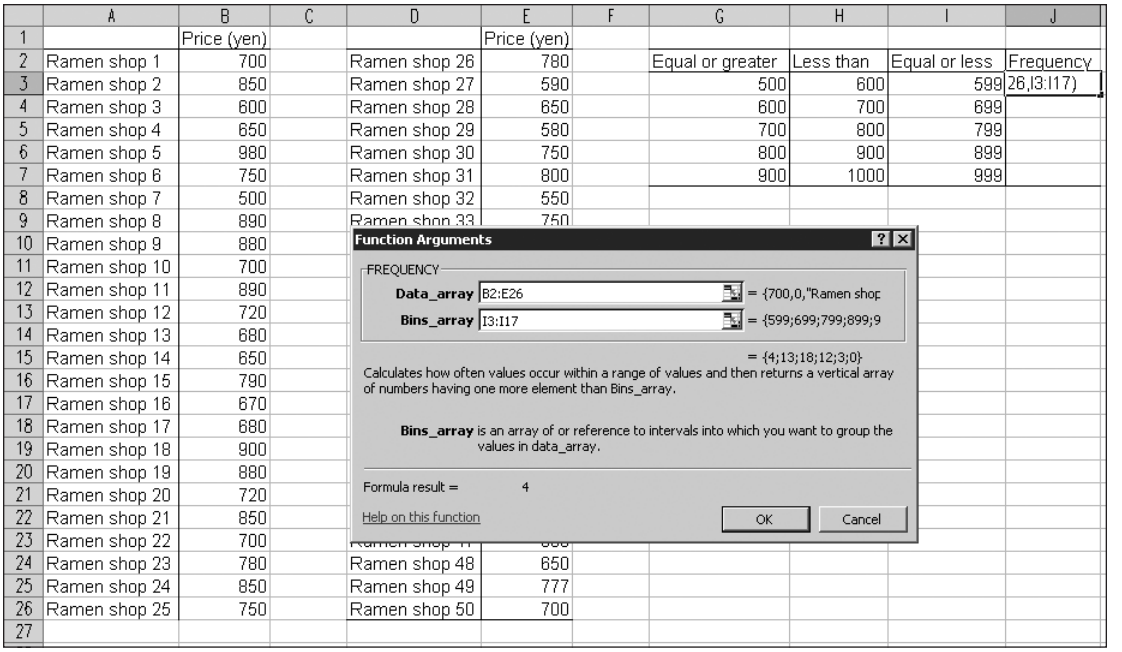
Step 3

Select **Statistical** from the category dropdown menu, and then select **FREQUENCY** as the name of the function.



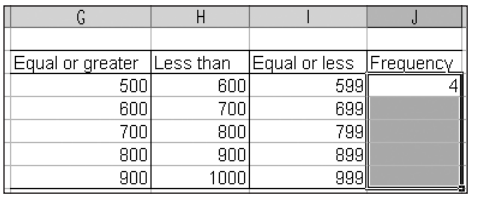
Step 4

Select the area shown in the figure below, and click **OK**.



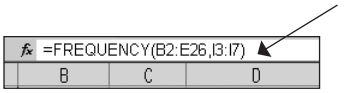
Step 5

Start with cell J3, and select the area from cell J3 to J7 as shown below.



Step 6

Click this part in the formula bar.



Step 7

Press ENTER while holding down the SHIFT key and CTRL key at the same time.

Step 8

Now you have the frequency of each class!

G	H	I	J
Equal or greater	Less than	Equal or less	Frequency
500	600	599	4
600	700	699	13
700	800	799	18
800	900	899	12
900	1000	999	3

2. CALCULATING ARITHMETIC MEAN, MEDIAN, AND STANDARD DEVIATION

This data comes from Rui's classmates' bowling scores on page 41.

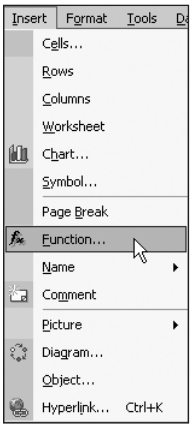
Step 1

Select cell B10.

	A	B
1		Team A
2	Rui-Rui	86
3	Jun	73
4	Yumi	124
5	Shizuka	111
6	Touko	90
7	Kaede	38
8		
9		
10	Average	
11	Median	
12	Standard Deviation	

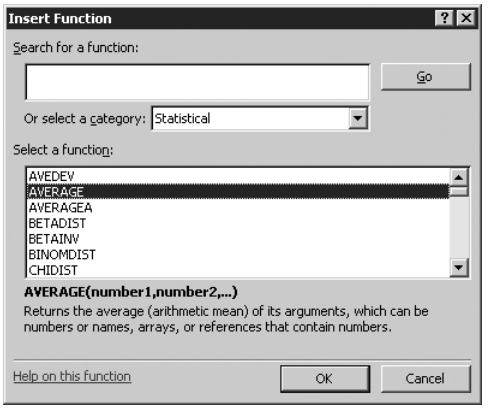
Step 2

Select **Insert ▸ Function**.



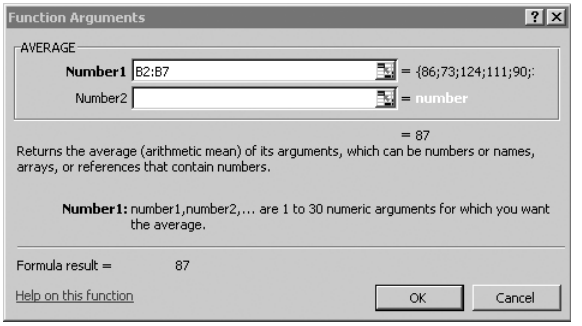
Step 3

Select **Statistical** in the category dropdown, and then select **AVERAGE**.



Step 4

Type the range shown in the figure below, and click **OK**.



Step 5

Now you have the average score for the team.

	A	B
1		Team A
2	Rui-Rui	86
3	Jun	73
4	Yumi	124
5	Shizuka	111
6	Touko	90
7	Kaede	38
8		
9		
10	Average	87
11	Median	
12	Standard Deviation	

You can calculate the median and standard deviation by following steps 1 through 5 and using the functions MEDIAN and STDEVP in step 2.

3. MAKING A CROSS TABULATION

The data for this table is Rui's classmates' responses to the new uniform design, found on page 61.

Step 1

Select cell F20, then select **Insert ▸ Function**.

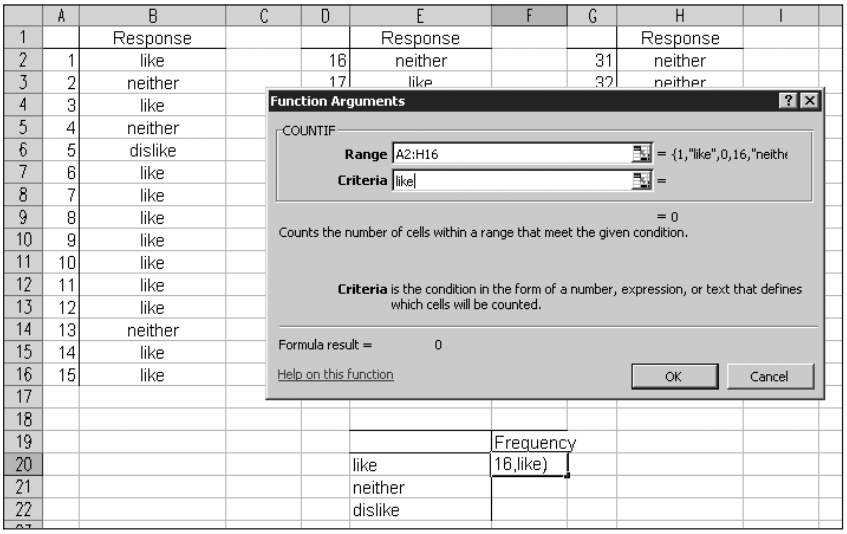
	A	B	C	D	E	F	G	H
1		Response			Response			Response
2	1	like		16	neither		31	neither
3	2	neither		17	like		32	neither
4	3	like		18	like		33	like
5	4	neither		19	like		34	dislike
6	5	dislike		20	like		35	like
7	6	like		21	like		36	like
8	7	like		22	like		37	like
9	8	like		23	dislike		38	like
10	9	like		24	neither		39	neither
11	10	like		25	like		40	like
12	11	like		26	like			
13	12	like		27	dislike			
14	13	neither		28	like			
15	14	like		29	like			
16	15	like		30	like			
17								
18								
19						Frequency		
20					like			
21					neither			
22					dislike			

Step 2

Select **Statistical** in the category dropdown, and then select **COUNTIF** as the name of the function.

Step 3

Select the area shown in the figure below, type *like* in the Criteria text box, and then click **OK**.



Step 4

Now you have the total number of Rui's classmates who like the new uniform.

	A	B	C	D	E	F	G	H
1		Response			Response			Response
2	1	like		16	neither		31	neither
3	2	neither		17	like		32	neither
4	3	like		18	like		33	like
5	4	neither		19	like		34	dislike
6	5	dislike		20	like		35	like
7	6	like		21	like		36	like
8	7	like		22	like		37	like
9	8	like		23	dislike		38	like
10	9	like		24	neither		39	neither
11	10	like		25	like		40	like
12	11	like		26	like			
13	12	like		27	dislike			
14	13	neither		28	like			
15	14	like		29	like			
16	15	like		30	like			
17								
18								
19						Frequency		
20					like	28		
21					neither			
22					dislike			

Step 5

You can obtain the frequency of *neither* and *dislike* by following steps 1 through 4 and typing those words instead of *like* in step 3.

4. CALCULATING THE STANDARD SCORE AND THE DEVIATION SCORE

This exercise uses the test data from page 72.

Steps 1 through 8 show the process for obtaining the standard score.

Steps 9 through 11 show the process for obtaining the deviation score. There is an Excel function for calculating standard score, but there is no function for calculating deviation score. However, the deviation score can be calculated fairly easily if we use the result of the standard score calculation.

Step 1

Select cell E2.

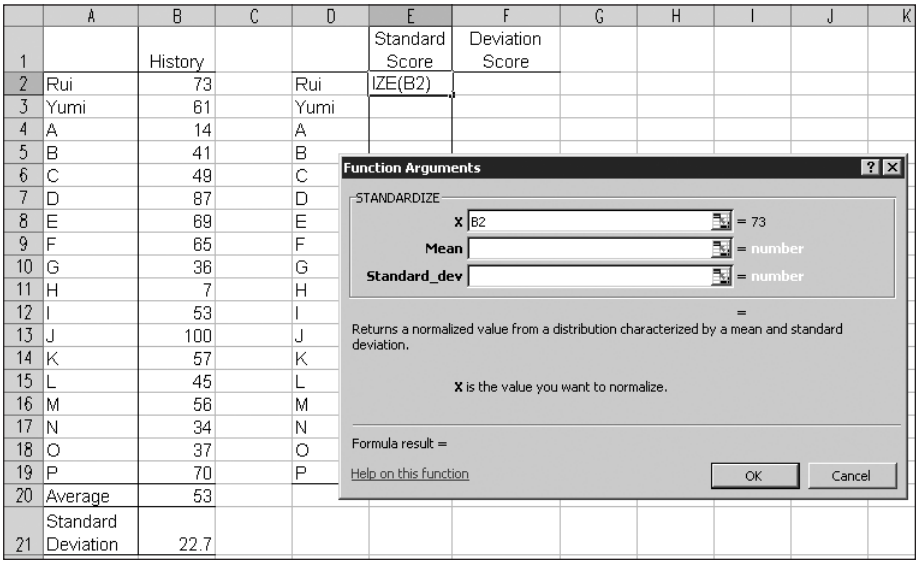
	A	B	C	D	E	F
1		History			Standard Score	Deviation Score
2	Rui	73		Rui		
3	Yumi	61		Yumi		
4	A	14		A		
5	B	41		B		
6	C	49		C		
7	D	87		D		
8	E	69		E		
9	F	65		F		
10	G	36		G		
11	H	7		H		
12	I	53		I		
13	J	100		J		
14	K	57		K		
15	L	45		L		
16	M	56		M		
17	N	34		N		
18	O	37		O		
19	P	70		P		
20	Average	53				
21	Standard Deviation	22.7				

Step 2

Select **Insert ▸ Function**. Then select **Statistical**, and then select **STANDARDIZE** as the name of the function.

Step 3

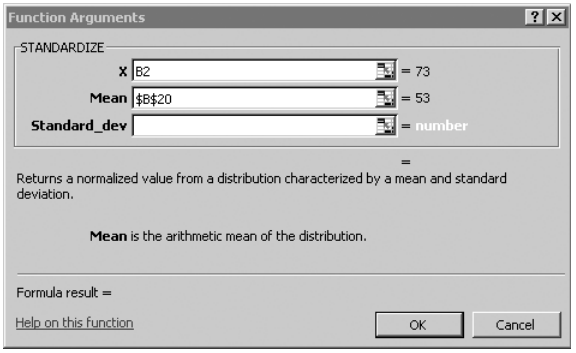
Select cell B2.



	A	B	C	D	E	F
1		History			Standard Score	Deviation Score
2	Rui	73		Rui	IZE(B2)	
3	Yumi	61		Yumi		
4	A	14		A		
5	B	41		B		
6	C	49		C		
7	D	87		D		
8	E	69		E		
9	F	65		F		
10	G	36		G		
11	H	7		H		
12	I	53		I		
13	J	100		J		
14	K	57		K		
15	L	45		L		
16	M	56		M		
17	N	34		N		
18	O	37		O		
19	P	70		P		
20	Average	53				
21	Standard Deviation	22.7				

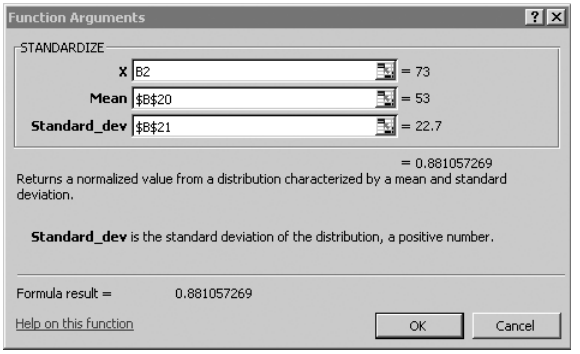
Step 4

Select B20 for Mean, press F4 once, and confirm that B20 has changed to B20.



Step 5

Select cell B21 for Standard_dev, press F4 once, and after confirming that B21 has changed to B21, click **OK**.



Step 6

Confirm that Rui's standard score has been calculated.

	A	B	C	D	E	F
1		History			Standard Score	Deviation Score
2	Rui	73		Rui	0.88	
3	Yumi	61		Yumi		
4	A	14		A		
5	B	41		B		
6	C	49		C		
7	D	87		D		
8	E	69		E		
9	F	65		F		
10	G	36		G		
11	H	7		H		
12	I	53		I		
13	J	100		J		
14	K	57		K		
15	L	45		L		
16	M	56		M		
17	N	34		N		
18	O	37		O		
19	P	70		P		
20	Average	53				
21	Standard Deviation	22.7				

Step 7

Put the point of the arrow near the bottom-right side of cell E2, confirm that the arrow has changed to a black cross, drag down to cell E19 by holding down the left button of the mouse, and let go of the button when you finish dragging.

	A	B	C	D	E	F
1		History			Standard Score	Deviation Score
2	Rui	73		Rui	0.88	
3	Yumi	61		Yumi		
4	A	14		A		
5	B	41		B		
6	C	49		C		
7	D	87		D		
8	E	69		E		
9	F	65		F		
10	G	36		G		
11	H	7		H		
12	I	53		I		
13	J	100		J		
14	K	57		K		
15	L	45		L		
16	M	56		M		
17	N	34		N		
18	O	37		O		
19	P	70		P		
20	Average	53				
21	Standard Deviation	22.7				

Step 8

Now you should have everyone's standard score!

	A	B	C	D	E	F
1		History			Standard Score	Deviation Score
2	Rui	73		Rui	0.88	
3	Yumi	61		Yumi	0.35	
4	A	14		A	-1.72	
5	B	41		B	-0.53	
6	C	49		C	-0.18	
7	D	87		D	1.50	
8	E	69		E	0.70	
9	F	65		F	0.53	
10	G	36		G	-0.75	
11	H	7		H	-2.03	
12	I	53		I	0.00	
13	J	100		J	2.07	
14	K	57		K	0.18	
15	L	45		L	-0.35	
16	M	56		M	0.13	
17	N	34		N	-0.84	
18	O	37		O	-0.70	
19	P	70		P	0.75	
20	Average	53				
21	Standard Deviation	22.7				

Step 9

Select cell F2 and type *=E2*10+50*, then press ENTER.

	A	B	C	D	E	F
1		History			Standard Score	Deviation Score
2	Rui	73		Rui	0.88	=E2*10+50
3	Yumi	61		Yumi	0.35	
4	A	14		A	-1.72	
5	B	41		B	-0.53	
6	C	49		C	-0.18	
7	D	87		D	1.50	
8	E	69		E	0.70	
9	F	65		F	0.53	
10	G	36		G	-0.75	
11	H	7		H	-2.03	
12	I	53		I	0.00	
13	J	100		J	2.07	
14	K	57		K	0.18	
15	L	45		L	-0.35	
16	M	56		M	0.13	
17	N	34		N	-0.84	
18	O	37		O	-0.70	
19	P	70		P	0.75	
20	Average	53				
21	Standard Deviation	22.7				

Step 10

Drag down to cell F19, as you did in step 7.

Step 11

Now you have the class's deviation score.

	A	B	C	D	E	F
1		History			Standard Score	Deviation Score
2	Rui	73		Rui	0.88	58.81
3	Yumi	61		Yumi	0.35	53.52
4	A	14		A	-1.72	32.82
5	B	41		B	-0.53	44.71
6	C	49		C	-0.18	48.24
7	D	87		D	1.50	64.98
8	E	69		E	0.70	57.05
9	F	65		F	0.53	55.29
10	G	36		G	-0.75	42.51
11	H	7		H	-2.03	29.74
12	I	53		I	0.00	50.00
13	J	100		J	2.07	70.70
14	K	57		K	0.18	51.76
15	L	45		L	-0.35	46.48
16	M	56		M	0.13	51.32
17	N	34		N	-0.84	41.63
18	O	37		O	-0.70	42.95
19	P	70		P	0.75	57.49
20	Average	53				
21	Standard Deviation	22.7				

5. CALCULATING THE PROBABILITY OF THE STANDARD NORMAL DISTRIBUTION

For this example, we'll use the data from page 93.

Step 1

Select cell B2.

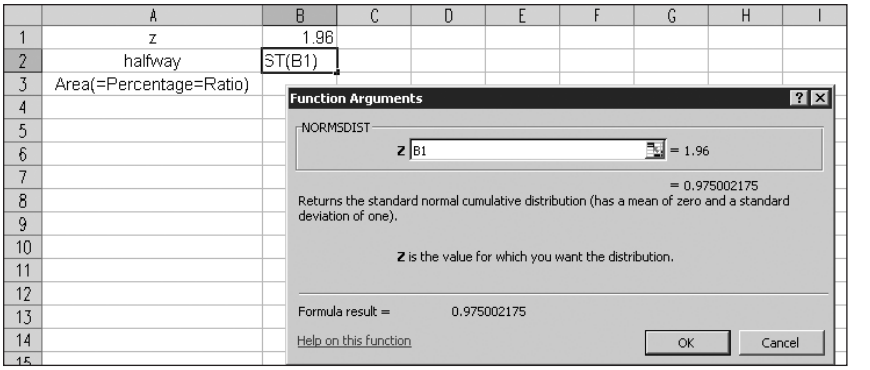
	A	B
1	z	1.96
2	halfway	
3	Area(=Percentage=Ratio)	

Step 2

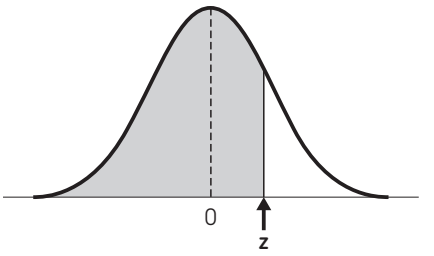
Select **Insert ▸ Function**, then select **Statistical**, and then select **NORMSDIST**.

Step 3

Select cell B1, and click **OK**.



In fact, NORMSDIST is a function to calculate the probability shown in the figure below.



Step 4

Type *=B2-0.5* in cell B3.

	A	B
1	z	1.96
2	halfway	0.975
3	Area(=Percentage=Ratio)	=B2-0.5

Step 5

Now you have the area.

	A	B
1	z	1.96
2	halfway	0.975
3	Area(=Percentage=Ratio)	0.475

6. CALCULATING THE POINT ON THE HORIZONTAL AXIS OF THE CHI-SQUARE DISTRIBUTION

The data for this exercise comes from page 104.

Step 1

Select cell B3.

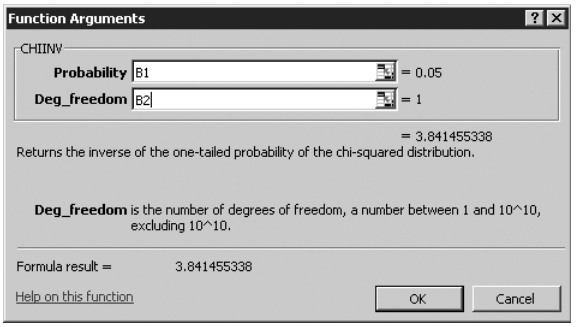
	A	B
1	P	0.05
2	Degrees of freedom	1
3	Chi-square	

Step 2

Select **Insert ▸ Function**, then select **Statistical**, and then select **CHIINV**.

Step 3

Select cells B1 and B2, and then click **OK**.



Step 4

Now you're done.

	A	B
1	P	0.05
2	Degrees of freedom	1
3	Chi-square	3.84146

7. CALCULATING THE CORRELATION COEFFICIENT

This data comes from the *P-Girls Magazine* survey found on page 116.

Step 1

Select cell B14.

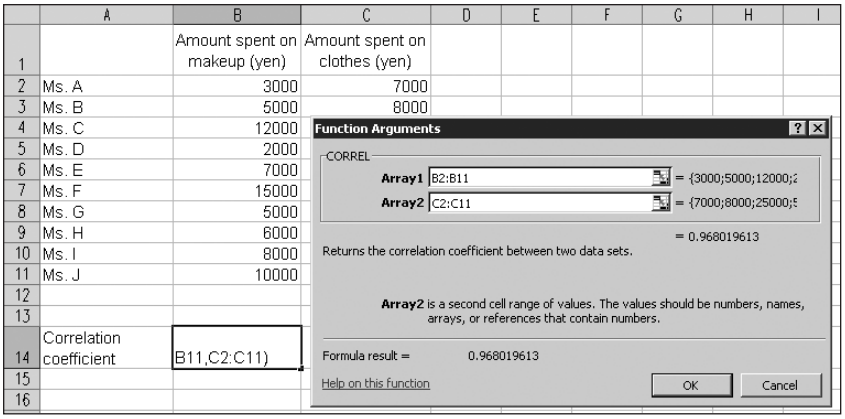
	A	B	C
1		Amount spent on makeup (yen)	Amount spent on clothes (yen)
2	Ms. A	3000	7000
3	Ms. B	5000	8000
4	Ms. C	12000	25000
5	Ms. D	2000	5000
6	Ms. E	7000	12000
7	Ms. F	15000	30000
8	Ms. G	5000	10000
9	Ms. H	6000	15000
10	Ms. I	8000	20000
11	Ms. J	10000	18000
12			
13			
14	Correlation coefficient		

Step 2

Select **Insert ▸ Function**, then select **Statistical**, and then select **CORREL**.

Step 3

Select the area shown in the figure below, and then click **OK**.



	A	B	C	D	E	F	G	H	I
1		Amount spent on makeup (yen)	Amount spent on clothes (yen)						
2	Ms. A	3000	7000						
3	Ms. B	5000	8000						
4	Ms. C	12000							
5	Ms. D	2000							
6	Ms. E	7000							
7	Ms. F	15000							
8	Ms. G	5000							
9	Ms. H	6000							
10	Ms. I	8000							
11	Ms. J	10000							
12									
13									
14	Correlation coefficient	B11,C2:C11)							
15									
16									

Step 4

Now you have the correlation coefficient.

	A	B	C
1		Amount spent on makeup (yen)	Amount spent on clothes (yen)
2	Ms. A	3000	7000
3	Ms. B	5000	8000
4	Ms. C	12000	25000
5	Ms. D	2000	5000
6	Ms. E	7000	12000
7	Ms. F	15000	30000
8	Ms. G	5000	10000
9	Ms. H	6000	15000
10	Ms. I	8000	20000
11	Ms. J	10000	18000
12			
13			
14	Correlation coefficient	0.968019613	

NOTE *Unfortunately, there are no Excel functions for calculating the correlation ratio or the Cramer's coefficient.*

8. PERFORMING TESTS OF INDEPENDENCE

This data is from the dating survey on page 157.

Step 1

Select cell B8.

	A	B	C	D	E
1		Phone	E-mail	Face to face	Sum
2	Female	34	61	53	148
3	Male	38	40	74	152
4	Sum	72	101	127	300
5					
6					
7		Phone	E-mail	Face to face	
8	Female				
9	Male				
10					
11					
12	P-value				

Step 2

Type *=E2*B4/E4* in cell B8. Do not press ENTER yet.

	A	B	C	D	E
1		Phone	E-mail	Face to face	Sum
2	Female	34	61	53	148
3	Male	38	40	74	152
4	Sum	72	101	127	300
5					
6					
7		Phone	E-mail	Face to face	
8	Female	=E2*B4/E4			
9	Male				
10					
11					
12	P-value				

Step 3

Select *E2* in the equation you just typed, press F4 three times, and confirm that *E2* has changed to *$E2*. Do not press ENTER yet.

	A	B	C	D	E
1		Phone	E-mail	Face to face	Sum
2	Female	34	61	53	148
3	Male	38	40	74	152
4	Sum	72	101	127	300
5					
6					
7		Phone	E-mail	Face to face	
8	Female	=$E2*B4/E4			
9	Male				
10					
11					
12	P-value				

Step 4

Select *B4* in the equation in cell B8, press F4 twice, and confirm that *B4* has changed to *B$4*. Select *E4* in the equation in cell B8, press F4 once, and confirm that *E4* has changed to *E4*. Then press ENTER.

	A	B	C	D	E
1		Phone	E-mail	Face to face	Sum
2	Female	34	61	53	148
3	Male	38	40	74	152
4	Sum	72	101	127	300
5					
6					
7		Phone	E-mail	Face to face	
8	Female	=$E2*B$4/E4			
9	Male				
10					
11					
12	P-value				

Step 5

Select cell B8, put the point of the arrow near the bottom right side of cell B8, confirm that the arrow has changed to a black cross, drag down to cell D8 by holding down the left button of the mouse, and let go of the button when you finish dragging.

	A	B	C	D	E
1		Phone	E-mail	Face to face	Sum
2	Female	34	61	53	148
3	Male	38	40	74	152
4	Sum	72	101	127	300
5					
6					
7		Phone	E-mail	Face to face	
8	Female	35.52			
9	Male				
10					
11					
12	P-value				

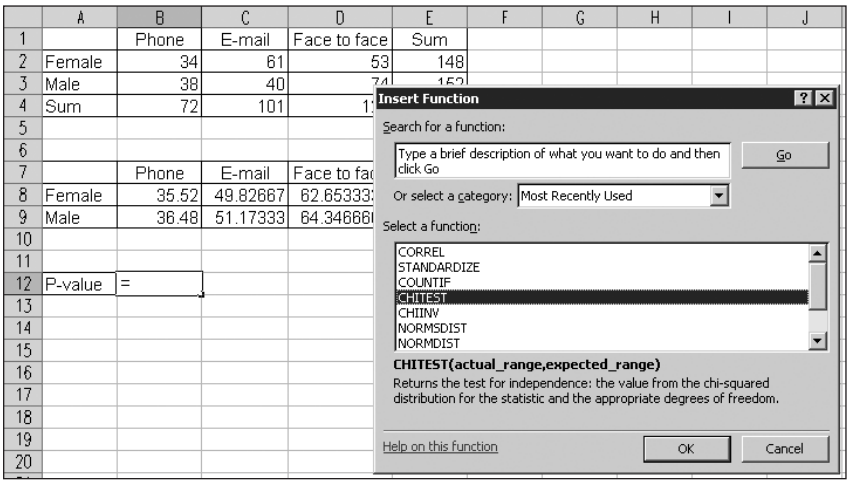
Step 6

Select the area from cell B8 to D8, put the point of the arrow near the bottom right side of cell D8, confirm that the arrow has changed to a black cross, drag down to cell D9 by holding down the left button of the mouse, and let go of the button when you finish dragging.

	A	B	C	D	E
1		Phone	E-mail	Face to face	Sum
2	Female	34	61	53	148
3	Male	38	40	74	152
4	Sum	72	101	127	300
5					
6					
7		Phone	E-mail	Face to face	
8	Female	35.52	49.82667	62.6533333	
9	Male				
10					
11					
12	P-value				

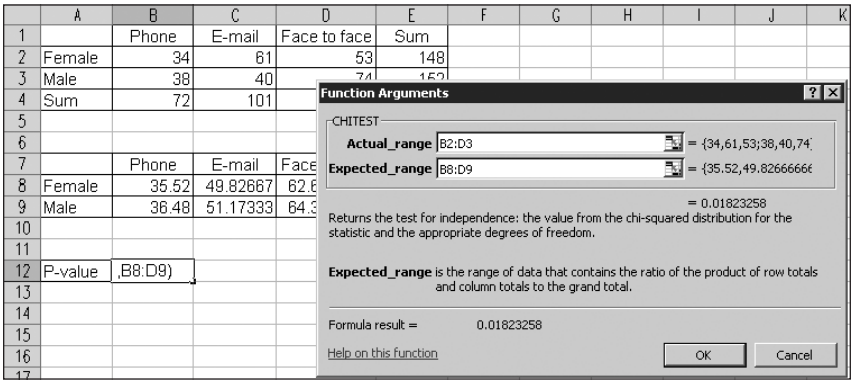
Step 7

Select cell B12, select **Insert ▸ Function**, then select **Statistical**, and then select **CHITEST**.



Step 8

Select the area shown in the figure below, and then click **OK**.



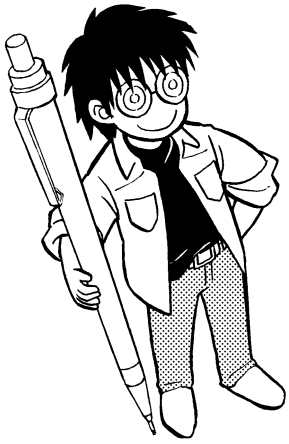
Step 9

Now you're done. You can confirm that the calculated value is equal to the P-value on page 177.

	A	B	C	D	E
1		Phone	E-mail	Face to face	Sum
2	Female	34	61	53	148
3	Male	38	40	74	152
4	Sum	72	101	127	300
5					
6					
7		Phone	E-mail	Face to face	
8	Female	35.52	49.82667	62.65333333	
9	Male	36.48	51.17333	64.34666667	
10					
11					
12	P-value	0.018233			

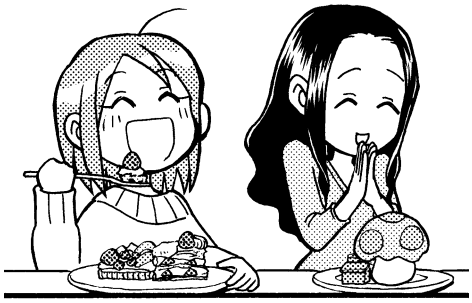
INDEX

NOTES



NOTES

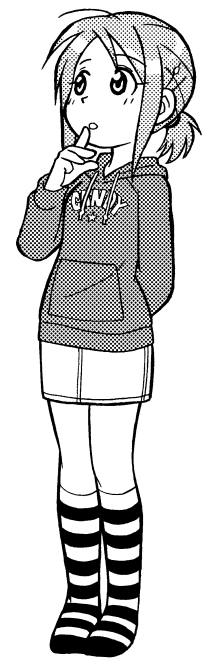
NOTES



NOTES

NOTES

NOTES



ABOUT THE AUTHOR

Shin Takahashi graduated from the Graduate School of Design of Kyushu University. Takahashi has worked as a lecturer and a data analyst and is currently active as a technical writer. He also wrote *Factor Analysis* and *Regression Analysis* in the Manga Guide series.

PRODUCTION TEAM FOR THE JAPANESE EDITION

Production: TREND-PRO Co., Ltd.

Founded in 1988, TREND-PRO produces newspaper and magazine advertisements incorporating manga for a wide range of clients from government agencies to major corporations and associations. Recently, TREND-PRO participated in advertisement and publishing projects using digital content. Some examples of past creations are publicly available at the company's website, *http://www.ad-manga.com/*.

Ikeden Bldg., 3F, 2-12-5 Shinbashi, Minato-ku, Tokyo, Japan

Telephone: 03-3519-6769; Fax: 03-3519-6110

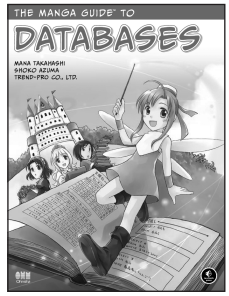
Scenario Writer: re_akino

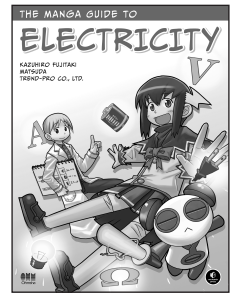
Artist: Iroha Inoue

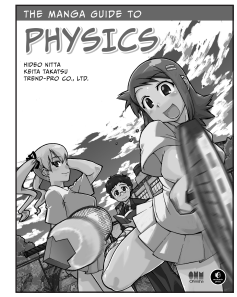
MORE MANGA GUIDES

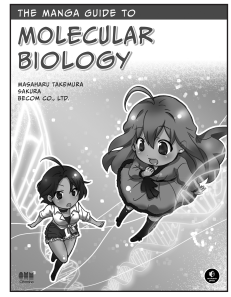
The *Manga Guide* series is a co-publication of No Starch Press and Ohmsha, Ltd. of Tokyo, Japan, one of Japan's oldest and most respected scientific and technical book publishers. Each title in the best-selling *Manga Guide* series is the product of the combined work of a manga illustrator, scenario writer, and expert scientist or mathematician. Once each title is translated into English, we rewrite and edit the translation as necessary and have an expert review each volume for technical accuracy. The result is the English version you hold in your hands.

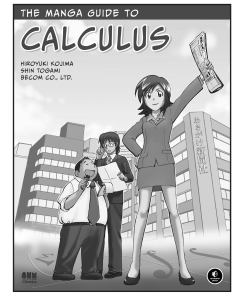
Find more *Manga Guides* at your favorite bookstore, and learn more about the series at *http://www.edumanga.me/*.







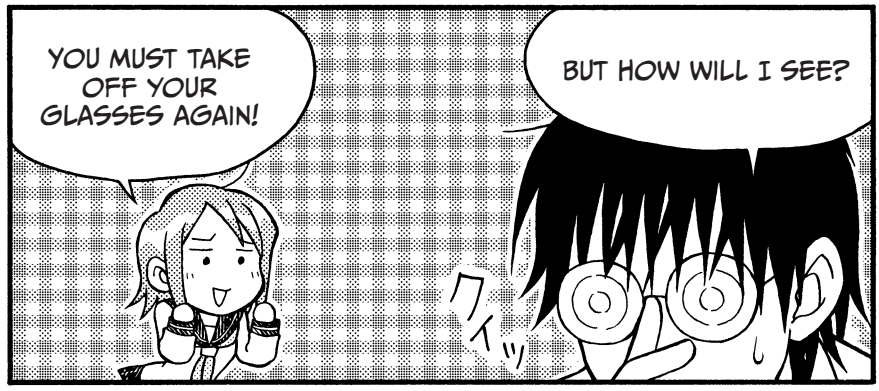




UPDATES

Visit *http://www.nostarch.com/mg_statistics.htm* for updates and errata, and to download the Excel files from the appendix.

The Manga Guide to Statistics is set in CCMeanwhile and Chevin. The book was printed and bound at Malloy Incorporated in Ann Arbor, Michigan. The paper is Glatfelter Spring Forge 60# Smooth Eggshell, which is certified by the Sustainable Forestry Initiative (SFI).



STATISTICS WITH HEART-POUNDING EXCITEMENT!

IF YOU'RE FEELING ANXIOUS ABOUT STATISTICS, OR YOU JUST NEED TO GET A HANDLE ON YOUR DATA, ***THE MANGA GUIDE TO STATISTICS*** WILL HELP YOU CONQUER THAT "I'M NO GOOD AT MATH" FEELING. THIS CARTOON GUIDE WILL HAVE YOU ON YOUR WAY TO STATISTICAL LITERACY IN NO TIME. AND BECAUSE NO MATHEMATICS BOOK IS COMPLETE WITHOUT THEM, IT HAS EXERCISES (AND ANSWERS), TOO, SO YOU CAN PRACTICE WHAT YOU LEARN.

FOLLOW ALONG AS THE EVER-PATIENT MR. YAMAMOTO TEACHES RUI HOW TO:

- CALCULATE THE MEAN, MEDIAN, AND STANDARD DEVIATION OF BOWLING SCORES
- GRAPH RAMEN NOODLE PRICES ON A HISTOGRAM
- DETERMINE THE PROBABILITY OF GETTING AN A ON A MATH TEST
- CALCULATE THE CRAMER'S COEFFICIENT TO DETERMINE HOW BOYS AND GIRLS PREFER TO BE ASKED OUT
- LEARN HOW STANDARD SCORE IS USED TO ADJUST TEST RESULTS WHEN TEACHERS "GRADE ON A CURVE"

THESE AND OTHER QUIRKY, REAL-WORLD EXAMPLES MAKE IT EASY FOR YOU TO LEARN WHAT MANY PEOPLE FIND DIFFICULT TO MASTER.

IF YOU WANT TO LEARN STATISTICS BUT YOU FEEL LIKE YOUR HEAD IS GOING TO EXPLODE, OR IF YOU JUST NEED A GREAT REFRESHER, LET MR. YAMAMOTO AND RUI BE YOUR GUIDES.

Ohmsha

no starch press

THE FINEST IN GEEK ENTERTAINMENT™
www.nostarch.com

$19.95 ($19.95 CDN)

SHELVE IN: *REFERENCE/MATHEMATICS*

ISBN: 978-1-59327-189-3

9 781593 271893 51995

6 89145 71891 1