THE MANGA GUIDE TO

STATISTICS

SHIN TAKAHASHI
TREND-PRO, CO., LTD.

Ohmsha
no starch press

PRAISE FOR THE *MANGA GUIDE* SERIES

"A fun and fairly painless lesson on what many consider to be a less-than-thrilling subject."
—SCHOOL LIBRARY JOURNAL

"This is really what a good math text should be like. . . . It presents statistics as something *fun*, and something enlightening."
—GOOD MATH, BAD MATH

"If you want to introduce a subject that kids wouldn't normally be very interested in, give it an amusing storyline and wrap it in cartoons."
—MAKE

"Definitely way better than trying to comprehend a bland statistics book."
—DR. DOBB'S CODETALK

"The most enjoyable tech book I've ever read."
—LINUX PRO MAGAZINE

"This is a solid book, and I wish there were more like it in the IT world."
—SLASHDOT

"Especially if you have a genuine enthusiasm for the material, the *Manga Guides* are effective learning tools."
—SKEPTICBLOG

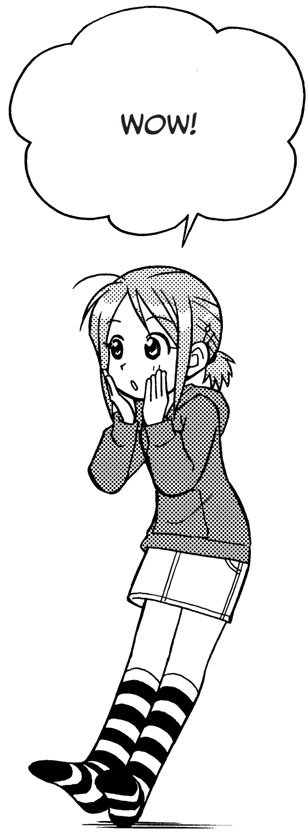
"The art is fantastic, and the teaching method is both fun and educational."
—ACTIVE ANIME

"Great for anyone wanting an introduction or a refresher on statistics."
—PARENTING SQUAD

"A light, impressively non-oppressive read, especially considering the technical nature of its subject."
—AIN'T IT COOL NEWS

"Makes accessible a very intimidating subject, letting the reader have fun while still delivering the goods."
—GEEKDAD, WIRED.COM

"An awfully fun, highly educational read."
—FRAZZLEDDAD



THE MANGA GUIDE TO STATISTICS

THE MANGA GUIDE™ TO
STATISTICS

SHIN TAKAHASHI
TREND-PRO CO., LTD.

no starch press

This English edition is co-published by No Starch Press and Ohmsha, Ltd.

11 10 09 2 3 4 5 6 7 8 9

ISBN-10: 1-59327-189-1
ISBN-13: 978-1-59327-189-3

Publisher: William Pollock
Author: Shin Takahashi
Illustrator: Iroha Inoue
Scenario Writer: re_akino
Producer: TREND-PRO Co., Ltd.
Production Editor: Megan Dunchak
Developmental Editor: Tyler Ortman
Technical Reviewers: Tom Bowler and Whitney Ortman
Compositor: Riley Hoffman
Proofreader: Christine Lias
Indexer: Karin Arrigioni

For information on book distributors or translations, please contact No Starch Press, Inc. directly:
No Starch Press, Inc.
555 De Haro Street, Suite 250, San Francisco, CA 94107
phone: 415.863.9900; fax: 415.863.9950; info@nostarch.com; http://www.nostarch.com/

Library of Congress Cataloging-in-Publication Data

```
Takahashi, Shin.
  The manga guide to statistics / Shin Takahashi and Trend-pro Co. -- 1st ed.
      p. cm.
  Includes index.
  ISBN-13: 978-1-59327-189-3
  ISBN-10: 1-59327-189-1
 1.  Mathematical statistics--Comic books, strips, etc. 2.  Mathematical statistics--Caricatures and cartoons.
I. Trend-pro Co. II. Title.
  QA276.T228 2009
  519.5--dc22
                                                                      2008042157
```

TABLE OF CONTENTS

PREFACE

This is an introductory book on statistics. The intended readers are:

- Those who need to conduct data analysis for research or business
- Those who do not necessarily need to conduct data analysis now but are interested in getting an idea of what the world of statistics is like
- Those who have already acquired general knowledge of statistics and want to learn more

Statistics is one of the areas of mathematics most closely related to everyday life and business. Familiarizing yourself with statistics may come in handy in situations like:

- Estimating how many servings of fried noodles you can sell at a food stand you are planning to set up during a school festival
- Estimating whether you will be able to pass a certification exam
- Comparing the probability that a sick person will get better between a case in which medicine *X* is used and a case in which it is not used

This book consists of seven chapters. Basically, each chapter is organized in the following sections:

- Cartoon
- Text explanation to supplement the cartoon
- Exercise and answer
- Summary

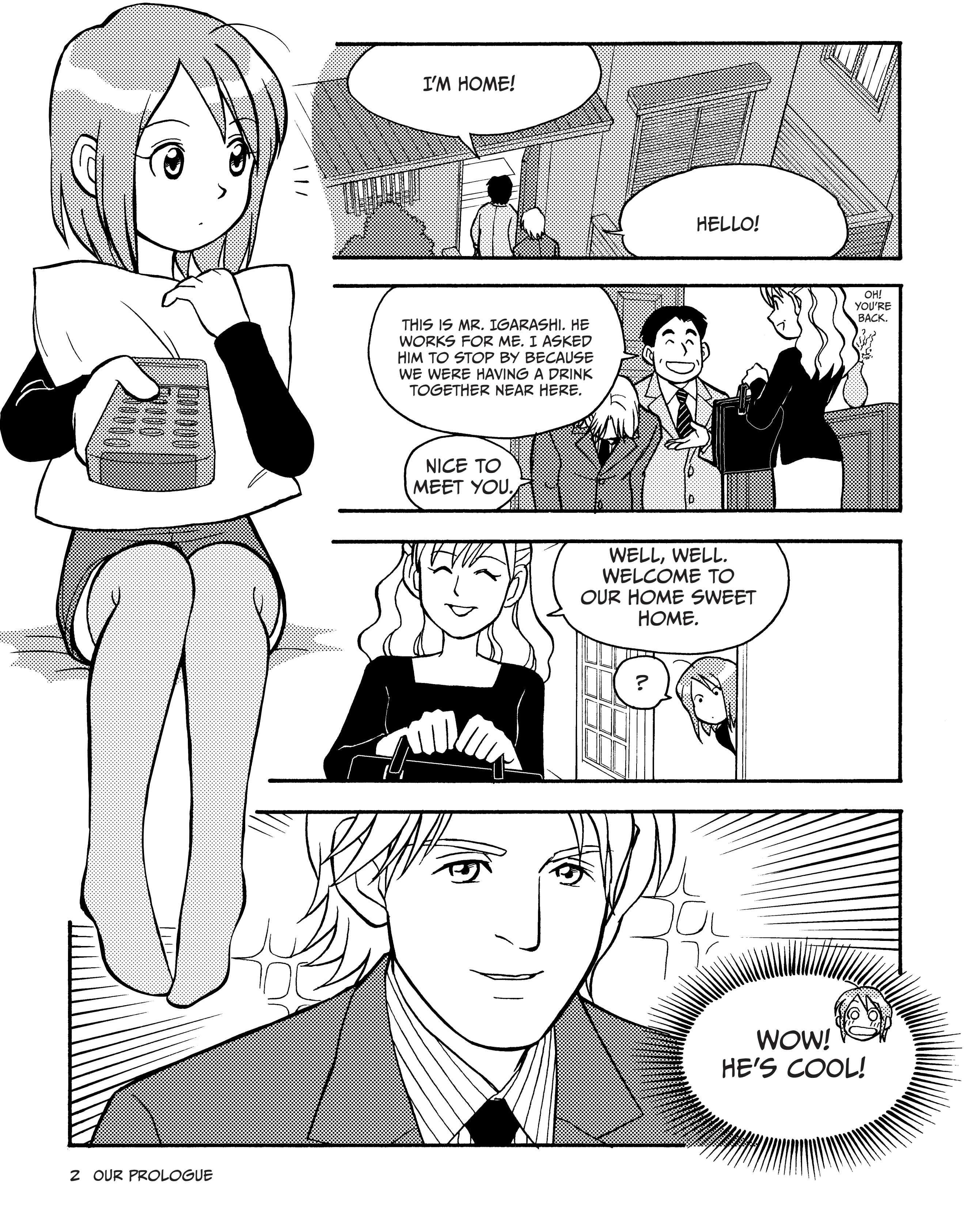
You can learn a lot by just reading the cartoon section, but deeper understanding and knowledge will be acquired if you read the other sections as well.

I would be very pleased if you start feeling that statistics is fun and useful after reading this book.

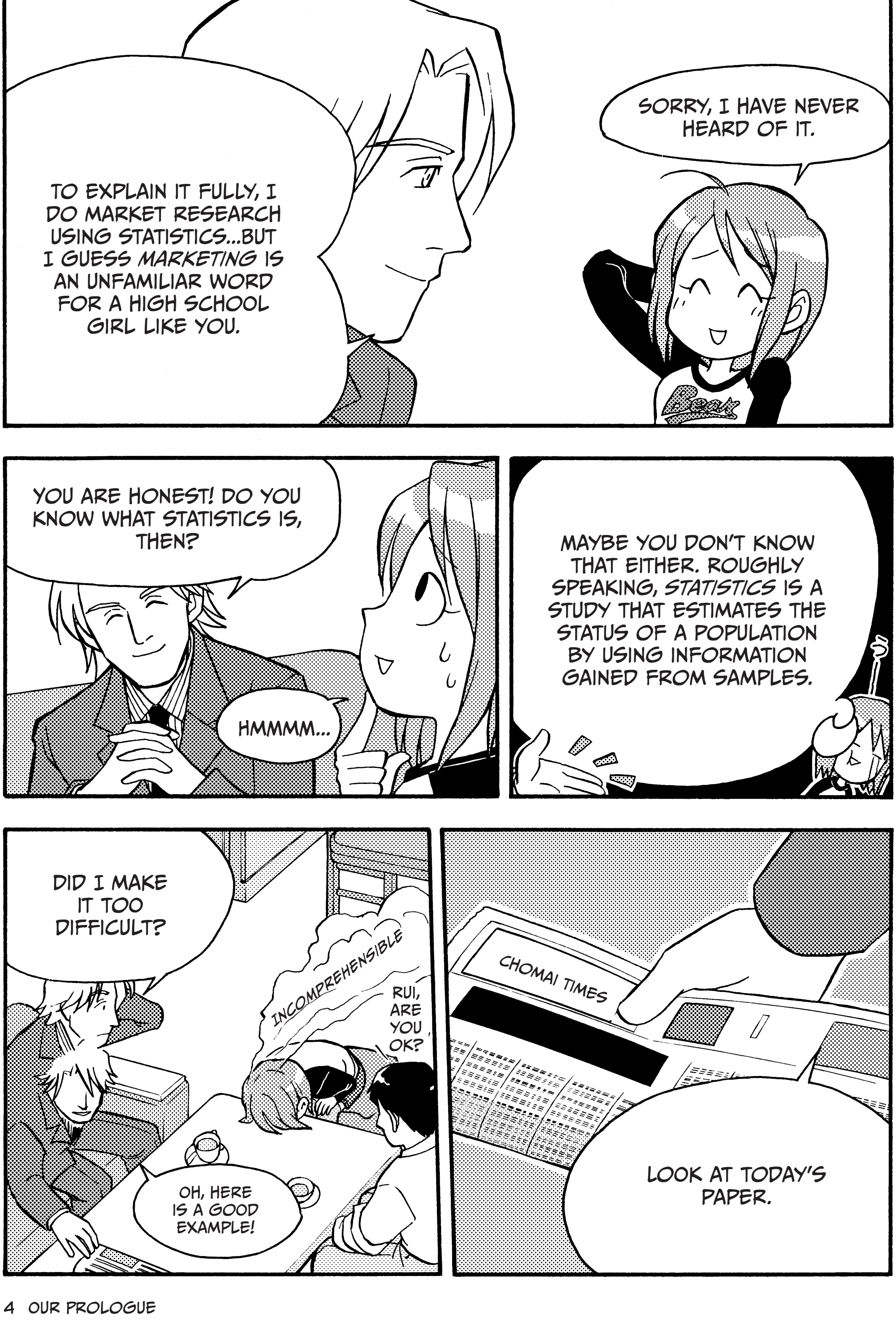
I would like to thank the staff in the development department of Ohmsha, Ltd., who offered me the opportunity to write this book. I would also like to thank TREND-PRO, Co., Ltd. for making my manuscript into a cartoon, the scenario writer, re_akino, and the illustrator, Iroha Inoue. Last but not least, I would like to thank Dr. Sakaori Fumitake of the College of Social Relations at Rikkyo University. He provided me with invaluable advice while I was preparing the manuscript for this book.

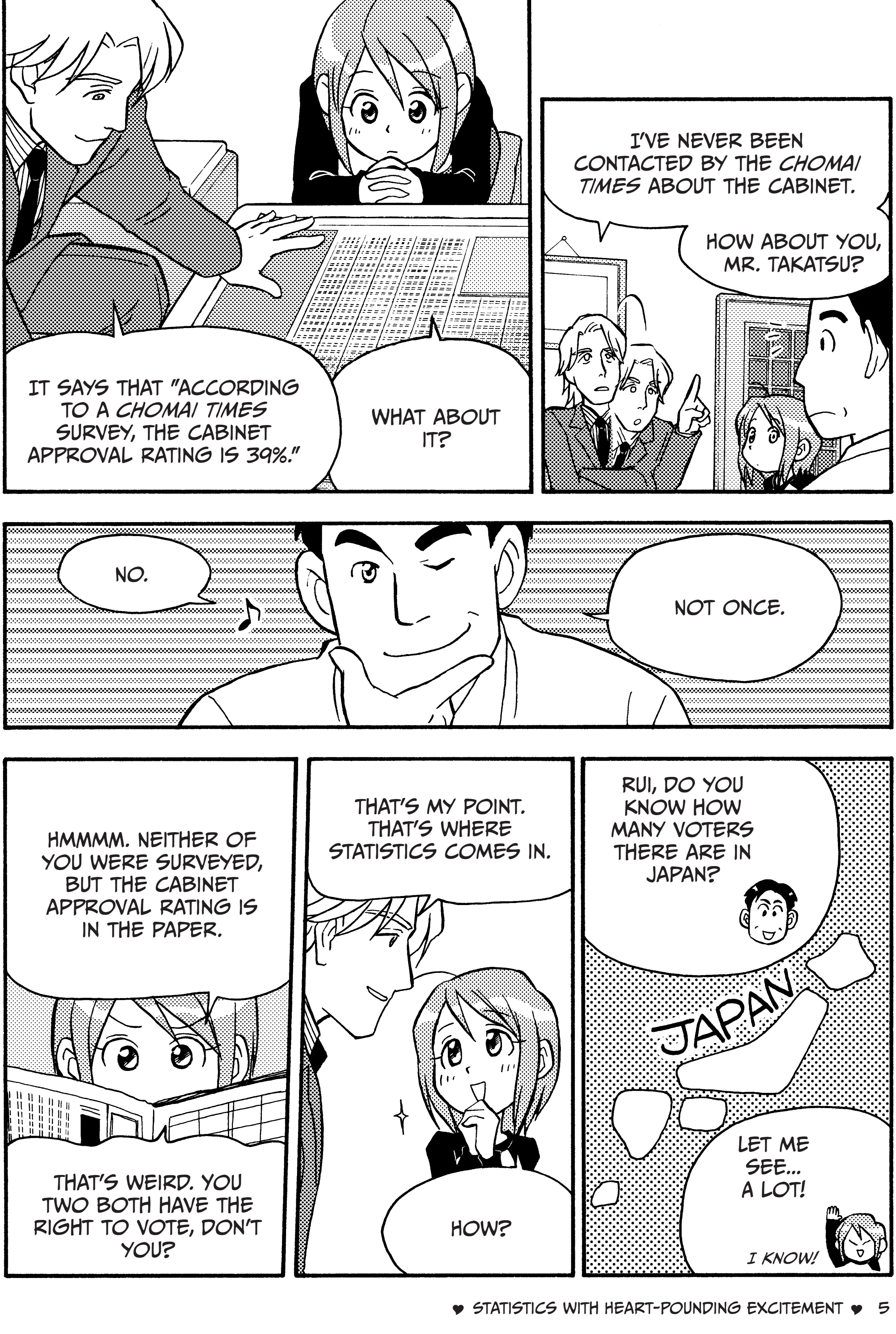
SHIN TAKAHASHI

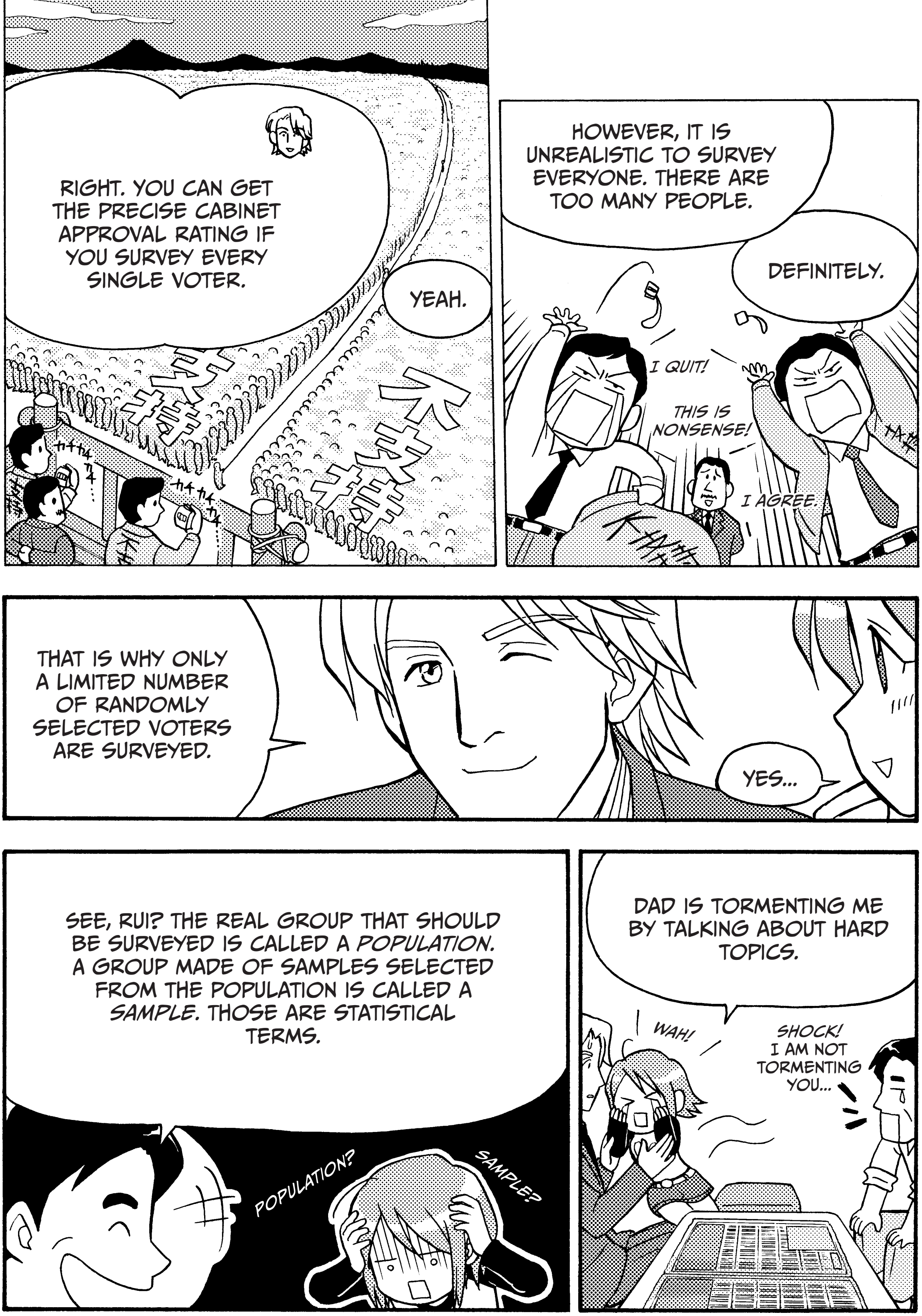
Our Prologue: Statistics with ♥ Heart-Pounding Excitement ♥

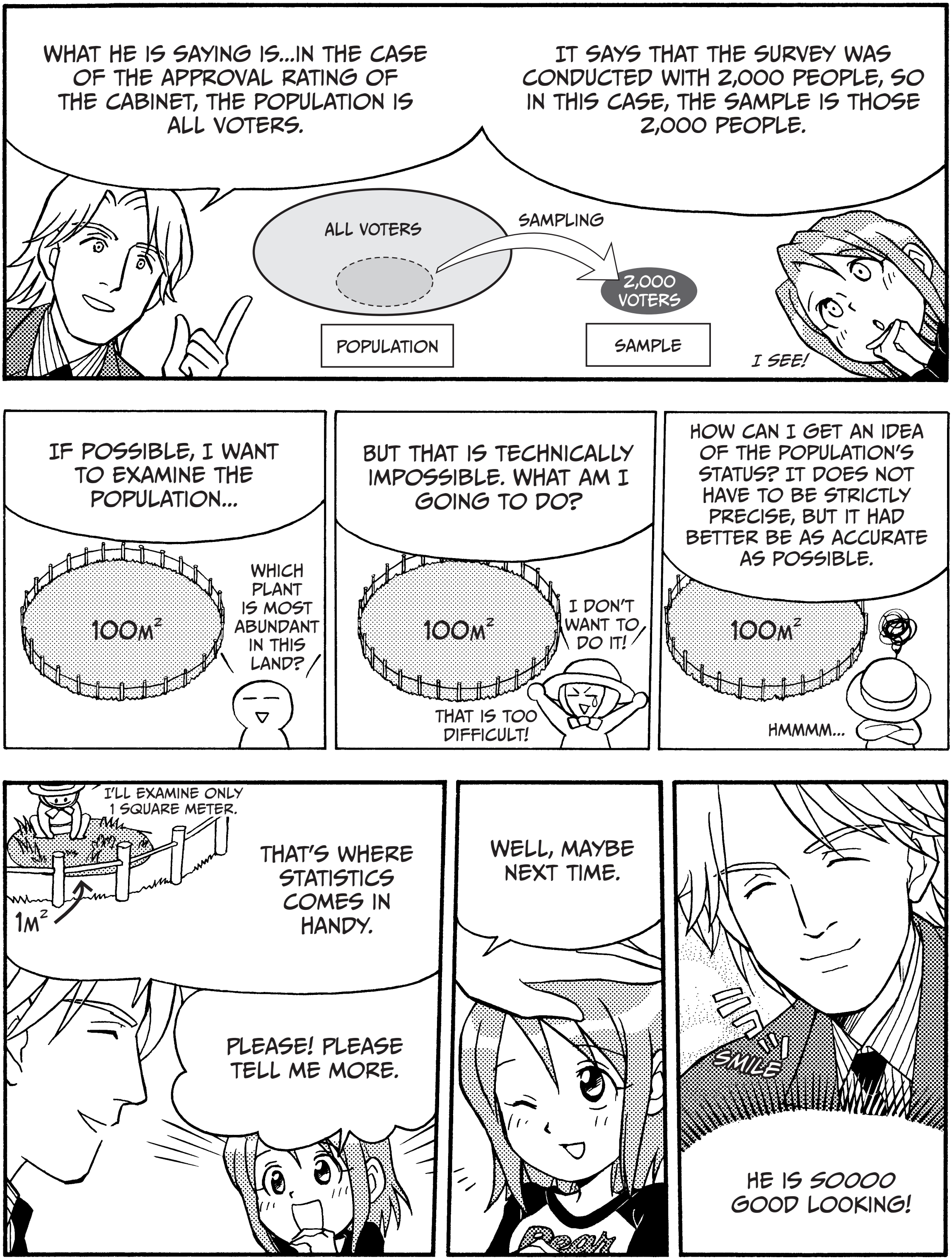

I'M HOME!
HELLO!
THIS IS MR. IGARASHI. HE WORKS FOR ME. I ASKED HIM TO STOP BY BECAUSE WE WERE HAVING A DRINK TOGETHER NEAR HERE.
OH! YOU'RE BACK.
NICE TO MEET YOU.
WELL, WELL. WELCOME TO OUR HOME SWEET HOME.
?
WOW! HE'S COOL!

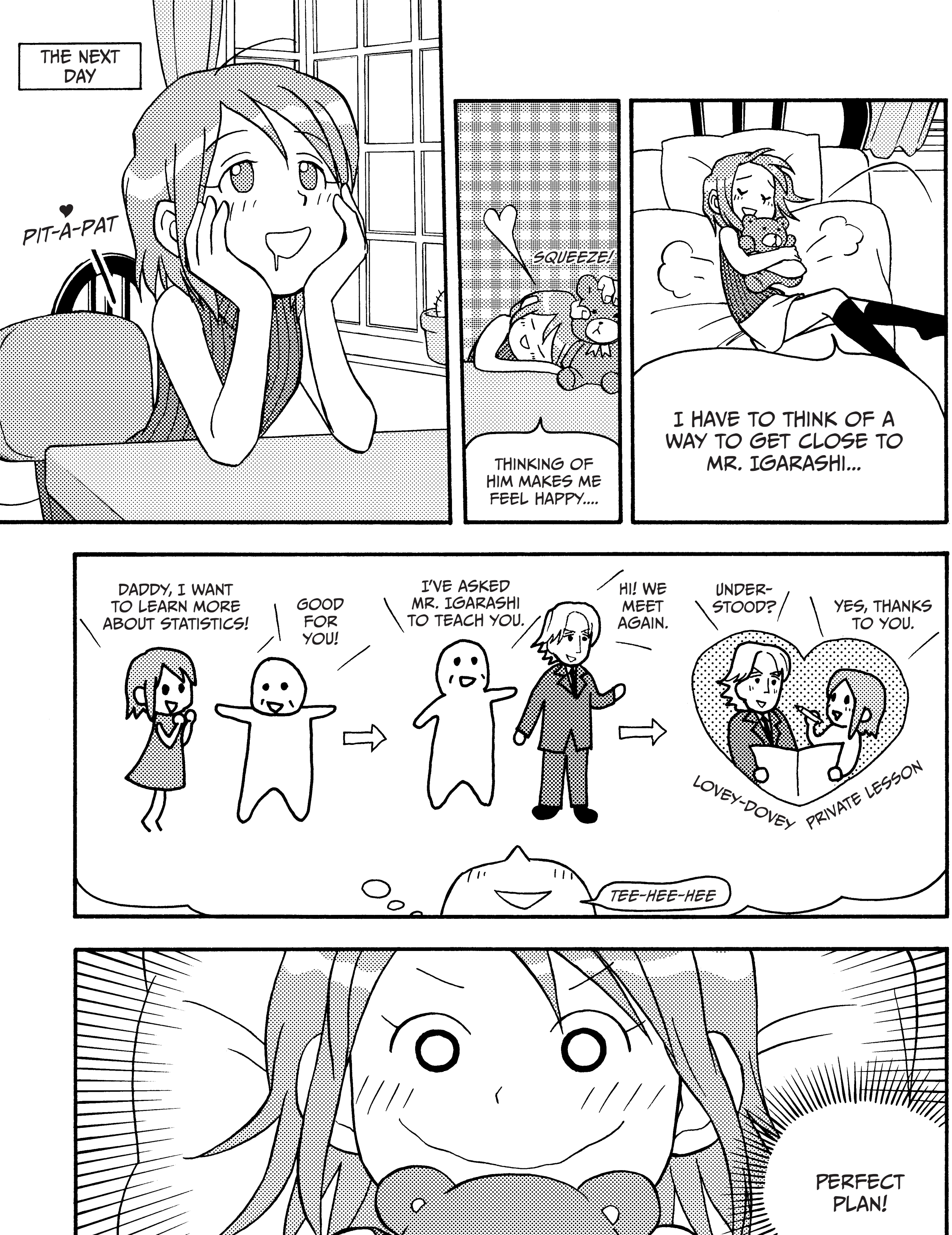

I'M HOME, RUI! SAY HELLO TO MR. IGARASHI. HE WORKS FOR ME.
GOOD EVENING.
PLEASE TAKE A SEAT.
TH... THANK YOU.
PLEASE USE THE CUSHION.
WOULD YOU LIKE A CUP OF COFFEE?
YOU HAVE A PRETTY DAUGHTER.
OH!
YOU FLATTER ME!
BUT I DO NOT DENY IT!
MR. IGARASHI, WHAT KIND OF WORK DO YOU DO?
WELL, AS I WORK FOR THE SAME COMPANY AS YOUR FATHER...
IN SHORT, I DO MARKETING.
MARKETING?
マーケティング?


TO EXPLAIN IT FULLY, I DO MARKET RESEARCH USING STATISTICS...BUT I GUESS *MARKETING* IS AN UNFAMILIAR WORD FOR A HIGH SCHOOL GIRL LIKE YOU.
SORRY, I HAVE NEVER HEARD OF IT.
YOU ARE HONEST! DO YOU KNOW WHAT STATISTICS IS, THEN?
HMMMM...
MAYBE YOU DON'T KNOW THAT EITHER. ROUGHLY SPEAKING, *STATISTICS* IS A STUDY THAT ESTIMATES THE STATUS OF A POPULATION BY USING INFORMATION GAINED FROM SAMPLES.
DID I MAKE IT TOO DIFFICULT?
INCOMPREHENSIBLE
RUI, ARE YOU OK?
OH, HERE IS A GOOD EXAMPLE!
CHOMAI TIMES
LOOK AT TODAY'S PAPER.

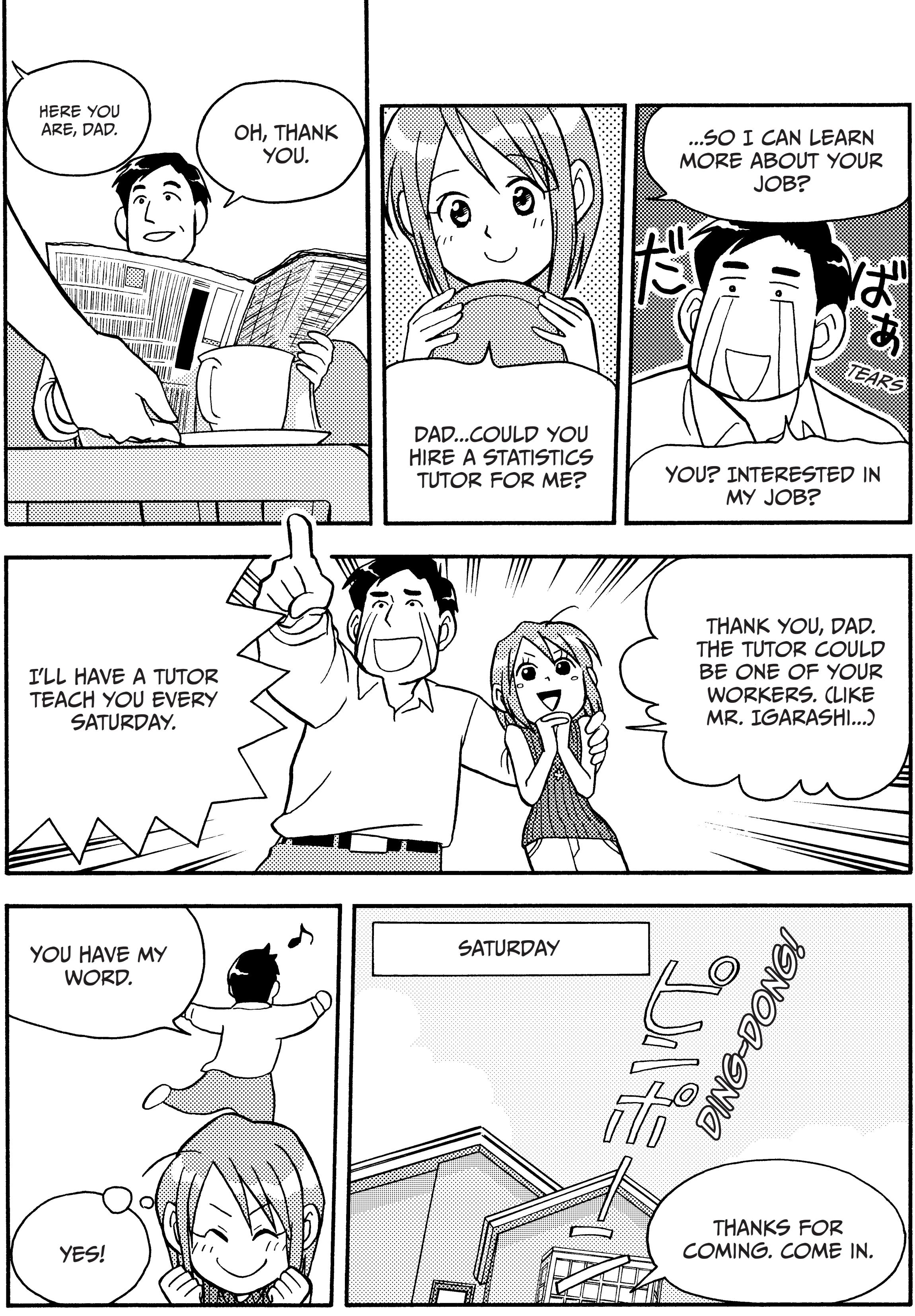

IT SAYS THAT "ACCORDING TO A *CHOMAI TIMES* SURVEY, THE CABINET APPROVAL RATING IS 39%."
WHAT ABOUT IT?
I'VE NEVER BEEN CONTACTED BY THE *CHOMAI TIMES* ABOUT THE CABINET.
HOW ABOUT YOU, MR. TAKATSU?
NO.
NOT ONCE.
HMMMM. NEITHER OF YOU WERE SURVEYED, BUT THE CABINET APPROVAL RATING IS IN THE PAPER.
THAT'S WEIRD. YOU TWO BOTH HAVE THE RIGHT TO VOTE, DON'T YOU?
THAT'S MY POINT. THAT'S WHERE STATISTICS COMES IN.
HOW?
RUI, DO YOU KNOW HOW MANY VOTERS THERE ARE IN JAPAN?
JAPAN
LET ME SEE... A LOT!
I KNOW!


RIGHT. YOU CAN GET THE PRECISE CABINET APPROVAL RATING IF YOU SURVEY EVERY SINGLE VOTER.
YEAH.
支持
不支持
HOWEVER, IT IS UNREALISTIC TO SURVEY EVERYONE. THERE ARE TOO MANY PEOPLE.
DEFINITELY.
I QUIT!
THIS IS NONSENSE!
I AGREE.
THAT IS WHY ONLY A LIMITED NUMBER OF RANDOMLY SELECTED VOTERS ARE SURVEYED.
YES...
SEE, RUI? THE REAL GROUP THAT SHOULD BE SURVEYED IS CALLED A *POPULATION*. A GROUP MADE OF SAMPLES SELECTED FROM THE POPULATION IS CALLED A *SAMPLE*. THOSE ARE STATISTICAL TERMS.
POPULATION?
SAMPLE?
DAD IS TORMENTING ME BY TALKING ABOUT HARD TOPICS.
WAH!
SHOCK! I AM NOT TORMENTING YOU...


WHAT HE IS SAYING IS...IN THE CASE OF THE APPROVAL RATING OF THE CABINET, THE POPULATION IS ALL VOTERS.
IT SAYS THAT THE SURVEY WAS CONDUCTED WITH 2,000 PEOPLE, SO IN THIS CASE, THE SAMPLE IS THOSE 2,000 PEOPLE.
ALL VOTERS
SAMPLING
2,000 VOTERS
POPULATION
SAMPLE
I SEE!
IF POSSIBLE, I WANT TO EXAMINE THE POPULATION...
100m²
WHICH PLANT IS MOST ABUNDANT IN THIS LAND?
BUT THAT IS TECHNICALLY IMPOSSIBLE. WHAT AM I GOING TO DO?
100m²
I DON'T WANT TO DO IT!
THAT IS TOO DIFFICULT!
HOW CAN I GET AN IDEA OF THE POPULATION'S STATUS? IT DOES NOT HAVE TO BE STRICTLY PRECISE, BUT IT HAD BETTER BE AS ACCURATE AS POSSIBLE.
100m²
HMMMM...
I'LL EXAMINE ONLY 1 SQUARE METER.
1m²
THAT'S WHERE STATISTICS COMES IN HANDY.
PLEASE! PLEASE TELL ME MORE.
WELL, MAYBE NEXT TIME.
SMILE
HE IS SOOOO GOOD LOOKING!



THE NEXT DAY
PIT-A-PAT
SQUEEZE!
THINKING OF HIM MAKES ME FEEL HAPPY....
I HAVE TO THINK OF A WAY TO GET CLOSE TO MR. IGARASHI...
DADDY, I WANT TO LEARN MORE ABOUT STATISTICS!
GOOD FOR YOU!
I'VE ASKED MR. IGARASHI TO TEACH YOU.
HI! WE MEET AGAIN.
UNDER-STOOD?
YES, THANKS TO YOU.
LOVEY-DOVEY PRIVATE LESSON
TEE-HEE-HEE
PERFECT PLAN!


HERE YOU ARE, DAD.
OH, THANK YOU.
DAD...COULD YOU HIRE A STATISTICS TUTOR FOR ME?
...SO I CAN LEARN MORE ABOUT YOUR JOB?
だばぁ
TEARS
YOU? INTERESTED IN MY JOB?
I'LL HAVE A TUTOR TEACH YOU EVERY SATURDAY.
THANK YOU, DAD. THE TUTOR COULD BE ONE OF YOUR WORKERS. (LIKE MR. IGARASHI...)
YOU HAVE MY WORD.
YES!
SATURDAY
ピンポーン
DING-DONG!
THANKS FOR COMING. COME IN.


HE'S HERE!
RUI! YOUR TEACHER IS HERE.
COMING!
STOMP STOMP STOMP
HI...

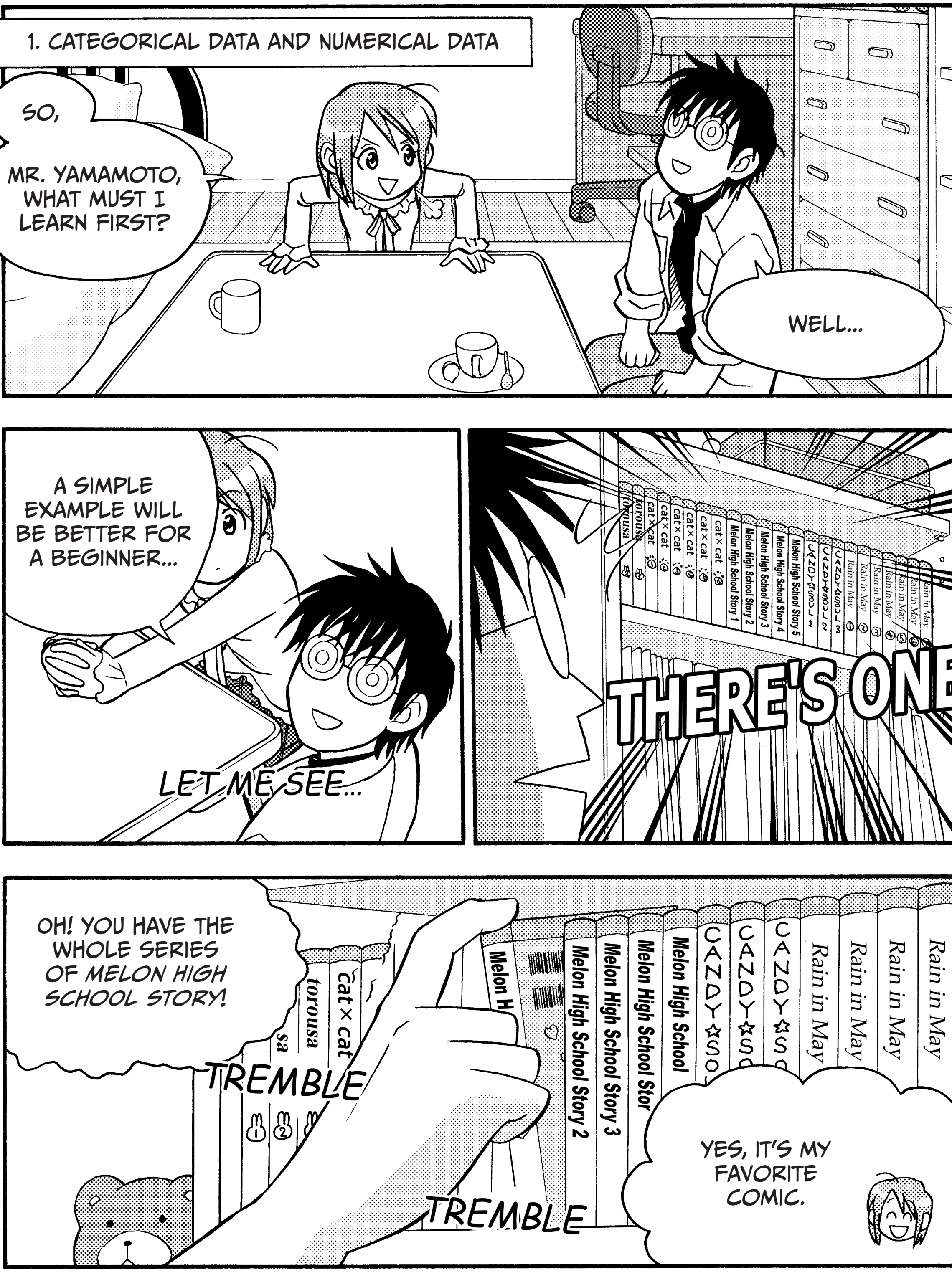

ボロッ…
RUI, THIS IS MY EMPLOYEE, MAMORU YAMAMOTO.
HOW DO YOU DO.
WHO IS THIS GUY?!
ガ
DAD...ISN'T MR. IGARASHI COMING?
IGARASHI? MAMORU LIVES CLOSER TO OUR HOUSE. HE IS GOOD AT TEACHING, TOO.
STUDY HARD!
HO HO HO!

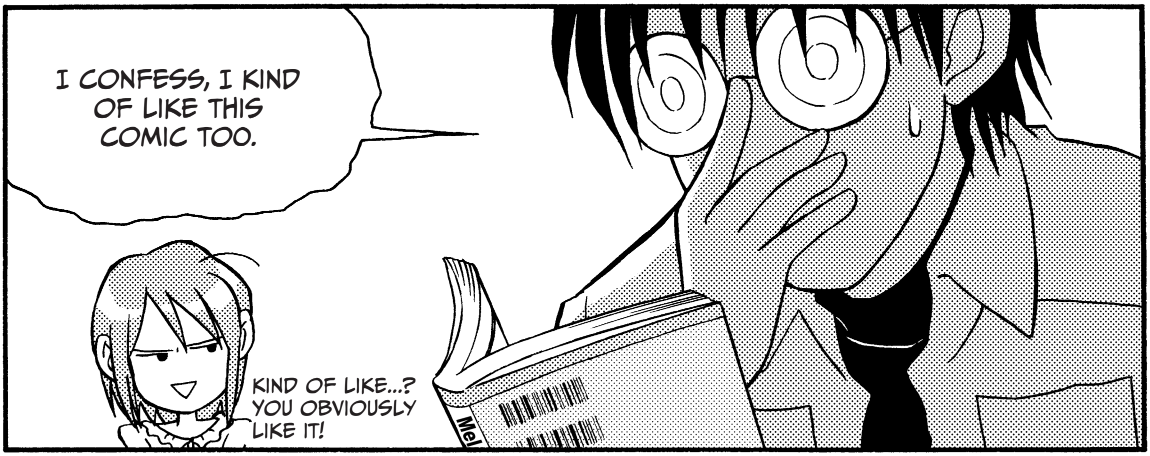

THIS IS A NIGHTMARE.
SHALL WE BEGIN, RUI?
UNNNGH...
MR. IGARASHI, I WORKED HARD TO MASTER STATISTICS!
GREAT! WHY DON'T YOU WORK WITH ME.
UM... RUI?
THIS IS NOT WHAT I HAD IN MIND...
I HAVE AN IDEA!
?
I'LL USE HIM TO STUDY STATISTICS, AND THEN APPROACH MR. IGARASHI!
NEVER GIVE UP!
LET'S START RIGHT NOW!
OH... OK.
THUS THE LESSON BEGAN.

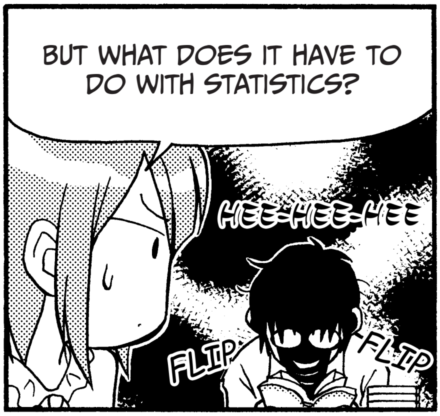
1

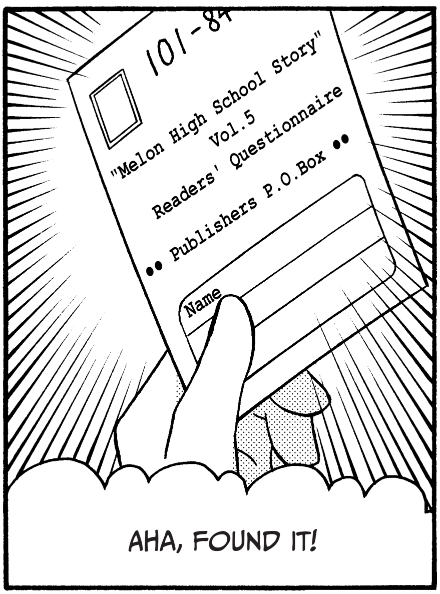
DETERMINING DATA TYPES

1. CATEGORICAL DATA AND NUMERICAL DATA









Melon High School Story Vol. 5 Reader Questionnaire

Q1. What is your impression of *Melon High School Story* Vol. 5?

1. Very fun
2. Rather fun
3. Average
4. Rather boring
5. Very boring

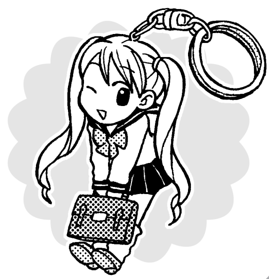
Q2. Sex

1. Female
2. Male

Q3. Age ________ years old

Q4. How many comics do you purchase per month? ________ titles

A Rina keychain will be given away to 30 lucky winners among those who send back this questionnaire!

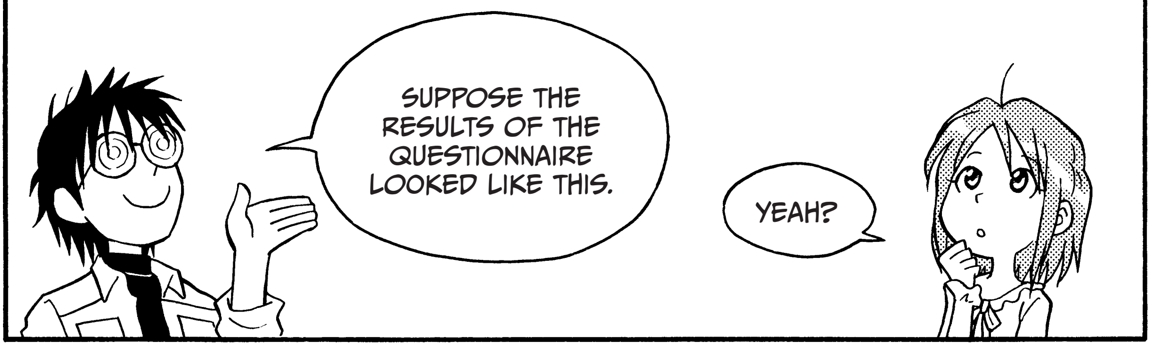


THANK YOU FOR YOUR CONTRIBUTION. YOUR VALUABLE OPINION WILL BE REFLECTED IN OUR FUTURE PUBLICATIONS.

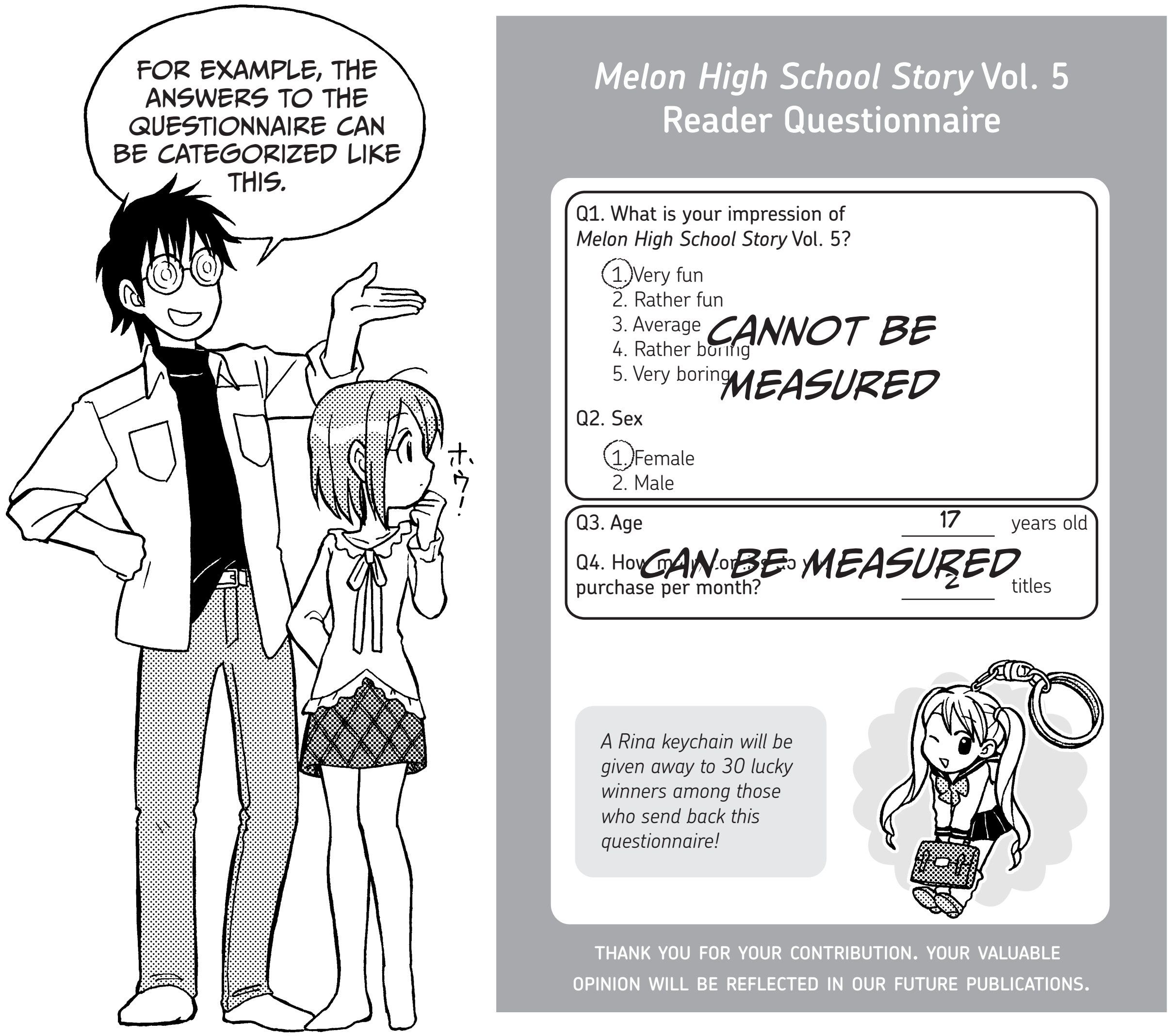

WHAT ABOUT THIS QUESTIONNAIRE?
IF YOU ANSWER THAT QUESTIONNAIRE...
FLASH!
101-8460
"Melon High School Story"
I BELIEVED YOU!
I'LL DO MY BEST!
I MADE THIS.
BUT I LIKE YOU!
YOU MIGHT BE ABLE TO GET A RINA KEYCHAIN! (IF YOU ARE LUCKY ENOUGH...)
GOD, I WISH I COULD WIN ONE...
YOU DORK!
NOW, BACK TO STUDYING...
SO SUDDENLY?
JUST ANSWER THAT QUESTIONNAIRE.
ALL RIGHT.
DONE!
SCRIBBLE SCRIBBLE
GOOD.

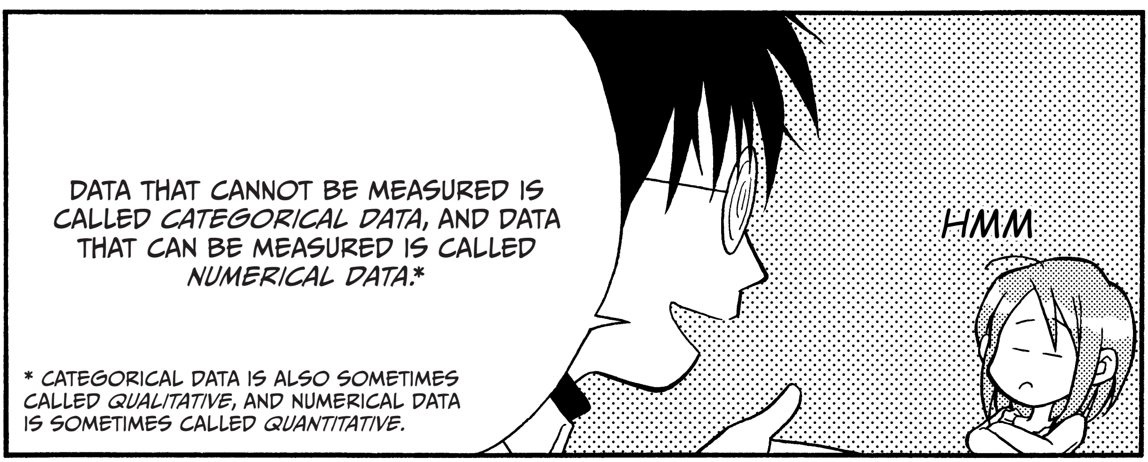
QUESTIONNAIRE RESULTS

RESPONDENT	Q1 YOUR IMPRESSION OF *MELON HIGH SCHOOL STORY*	Q2 SEX	Q3 AGE	Q4 COMIC BOOK PURCHASES PER MONTH
RUI	VERY FUN	FEMALE	17	2
A	RATHER FUN	FEMALE	17	1
B	AVERAGE	MALE	18	5
C	RATHER BORING	MALE	22	7
D	RATHER FUN	FEMALE	25	4
E	VERY BORING	MALE	20	3
F	VERY FUN	FEMALE	16	1
G	RATHER FUN	FEMALE	17	2
H	AVERAGE	MALE	18	0
I	AVERAGE	FEMALE	21	3
...	...	...	...	...



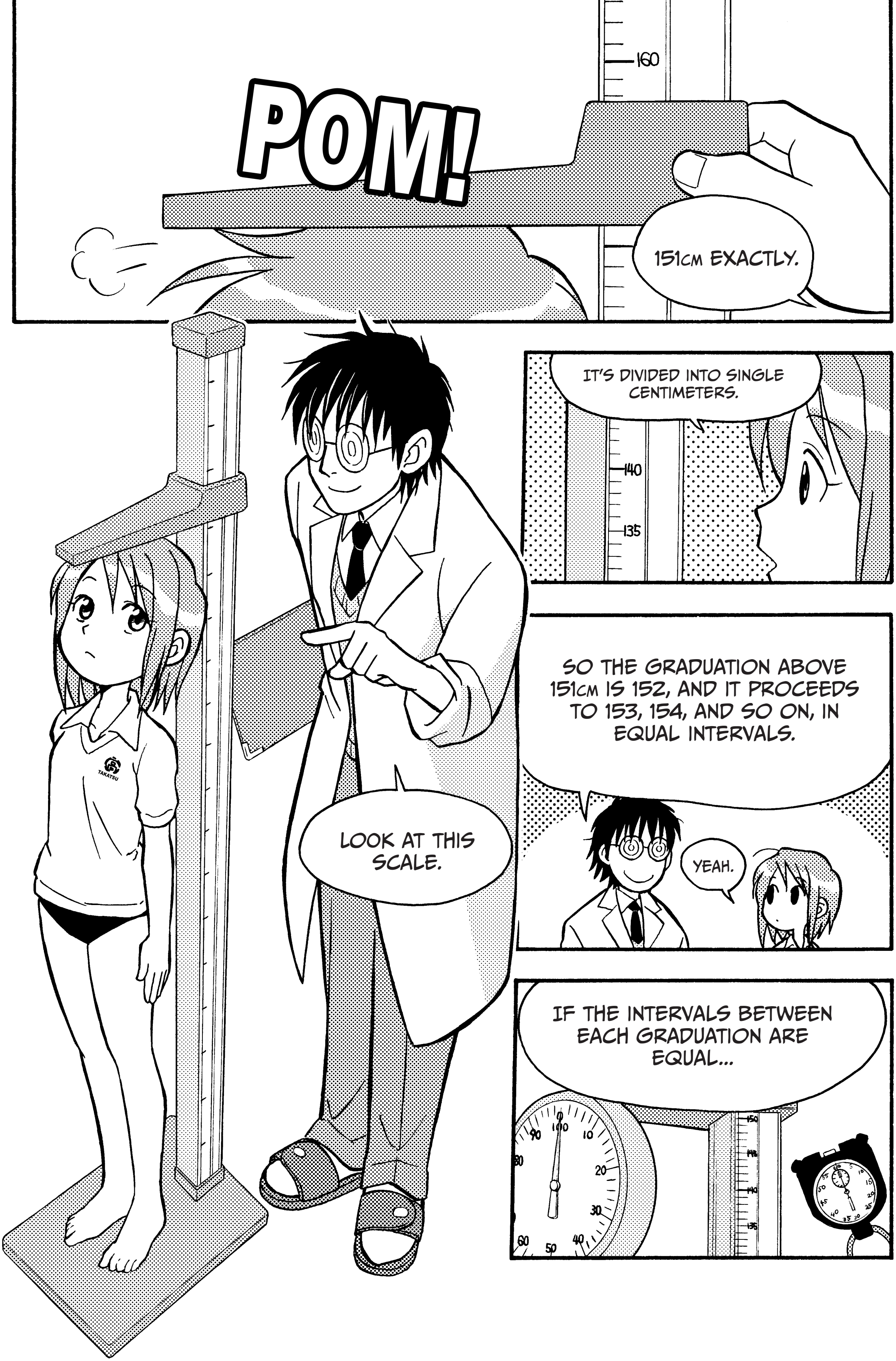

PERSONALLY, I AM EXTREMELY ANXIOUS TO KNOW IF *MELON HIGH SCHOOL STORY* WAS WELL RECEIVED BY THE READERS!
YOU "MELON HIGH SCHOOL" FREAK!
BUT LET'S STAY FOCUSED. THE POINT OF TODAY'S LESSON IS TO CATEGORIZE DATA.
THUMP
WHAT DO YOU MEAN?
WHEN YOU LOOK AT DATA, THERE ARE TWO TYPES:
DATA THAT YOU CAN MEASURE AND DATA THAT YOU CANNOT MEASURE.
CAN BE MEASURED
CANNOT BE MEASURED

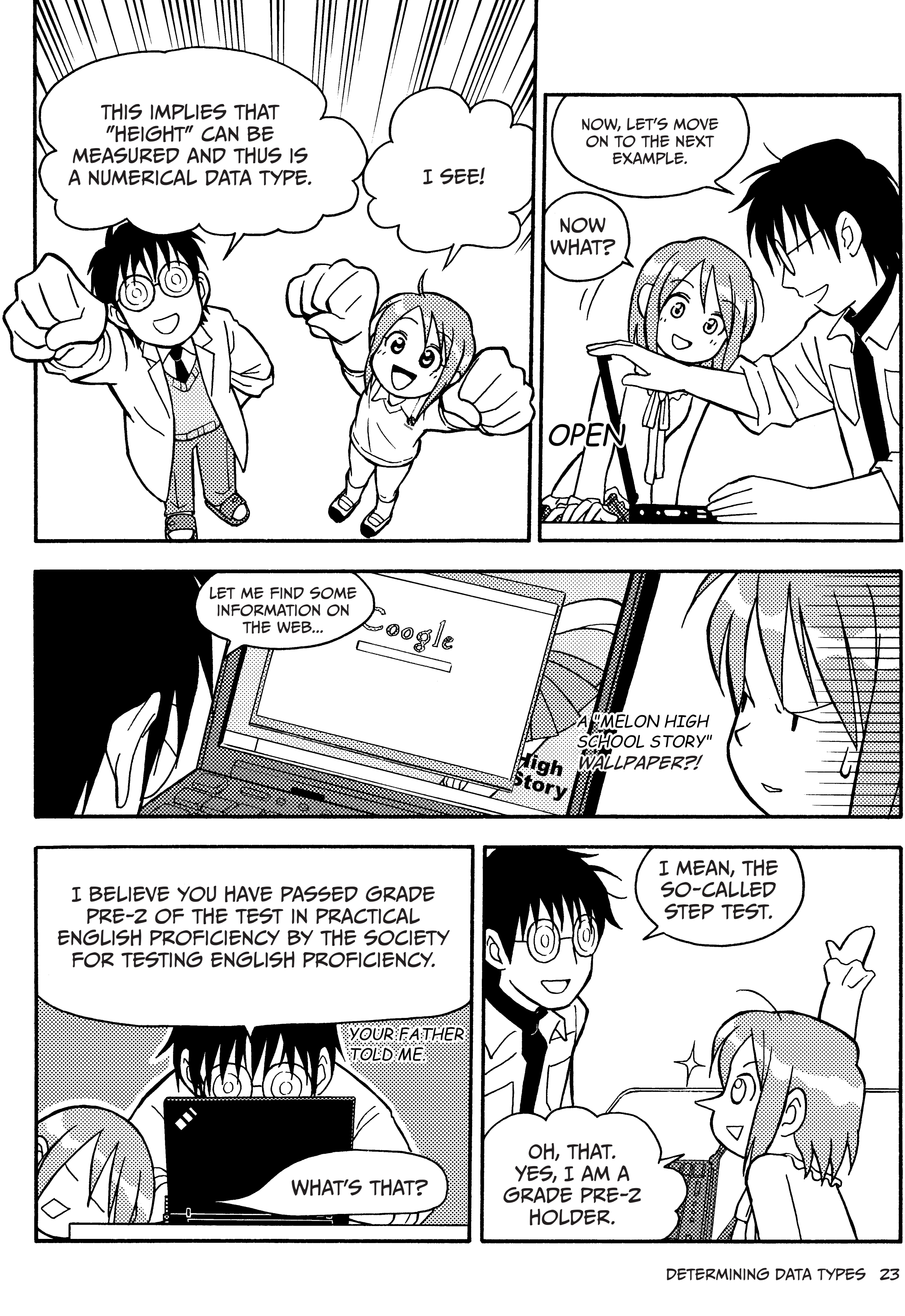

FOR EXAMPLE, THE ANSWERS TO THE QUESTIONNAIRE CAN BE CATEGORIZED LIKE THIS.
ホウ!
Melon High School Story Vol. 5
Reader Questionnaire
Q1. What is your impression of *Melon High School Story* Vol. 5?
1. Very fun
2. Rather fun
3. Average
4. Rather boring
5. Very boring
CANNOT BE MEASURED
Q2. Sex
1. Female
2. Male
Q3. Age 17 years old
Q4. How
purchase per month? 2 titles
CAN BE MEASURED
A Rina keychain will be given away to 30 lucky winners among those who send back this questionnaire!
THANK YOU FOR YOUR CONTRIBUTION. YOUR VALUABLE OPINION WILL BE REFLECTED IN OUR FUTURE PUBLICATIONS.

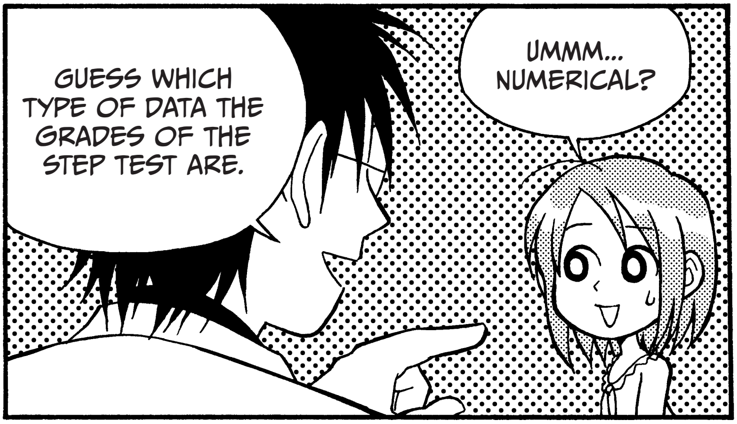

DATA THAT CANNOT BE MEASURED IS CALLED *CATEGORICAL DATA*, AND DATA THAT CAN BE MEASURED IS CALLED *NUMERICAL DATA*.*
HMM
* CATEGORICAL DATA IS ALSO SOMETIMES CALLED *QUALITATIVE*, AND NUMERICAL DATA IS SOMETIMES CALLED *QUANTITATIVE*.

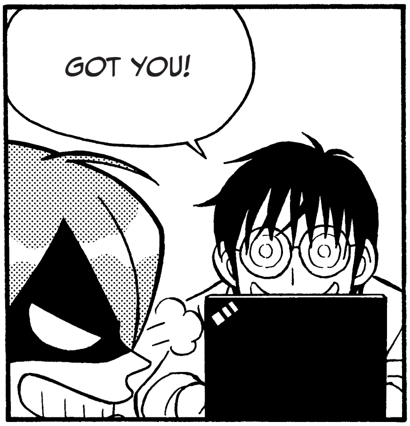
2. AN EXAMPLE OF TRICKY CATEGORICAL DATA
BUT TO ME...
THE FIRST QUESTION DOES NOT LOOK LIKE CATEGORICAL DATA...
Q1. What is your impr
Melon High School Sto
1. Very fun
2. Rather fun
Average
Rather boring
Very boring
I UNDERSTAND WHY YOU FEEL THAT WAY.
TELL ME WHY IT IS!
THOUGH IT DOES NOT LOOK LIKE CATEGORICAL DATA, THIS IS INDEED DATA THAT CANNOT BE MEASURED.
HOW COME?
BECAUSE THE DATA CANNOT BE EQUALLY SEPARATED.
I STILL DON'T GET IT.
HERE'S A GOOD EXAMPLE.


NOW, TELL ME YOUR...UMM...
WHAT?
WHAT?
WEIGHT.
THAT IS SOMETHING YOU SHOULD NEVER ASK A LADY!
I'M SORRY...
Melon High School Story
THEN, COULD YOU TELL ME YOUR HEIGHT?
151cm.
WHY DIDN'T YOU ASK ME THAT IN THE FIRST PLACE?
I'M GLAD YOU'RE NOT UPSET BY THIS QUESTION.
WHEN YOU MEASURE YOUR HEIGHT, YOU USE THE SCALE, DON'T YOU?
OF COURSE!


POM!
151cm EXACTLY.
LOOK AT THIS SCALE.
IT'S DIVIDED INTO SINGLE CENTIMETERS.
SO THE GRADUATION ABOVE 151cm IS 152, AND IT PROCEEDS TO 153, 154, AND SO ON, IN EQUAL INTERVALS.
YEAH.
IF THE INTERVALS BETWEEN EACH GRADUATION ARE EQUAL...


THIS IMPLIES THAT "HEIGHT" CAN BE MEASURED AND THUS IS A NUMERICAL DATA TYPE.
I SEE!
NOW, LET'S MOVE ON TO THE NEXT EXAMPLE.
NOW WHAT?
OPEN
LET ME FIND SOME INFORMATION ON THE WEB...
Coogle
A "MELON HIGH SCHOOL STORY" WALLPAPER?!
I BELIEVE YOU HAVE PASSED GRADE PRE-2 OF THE TEST IN PRACTICAL ENGLISH PROFICIENCY BY THE SOCIETY FOR TESTING ENGLISH PROFICIENCY.
YOUR FATHER TOLD ME.
WHAT'S THAT?
I MEAN, THE SO-CALLED STEP TEST.
OH, THAT. YES, I AM A GRADE PRE-2 HOLDER.

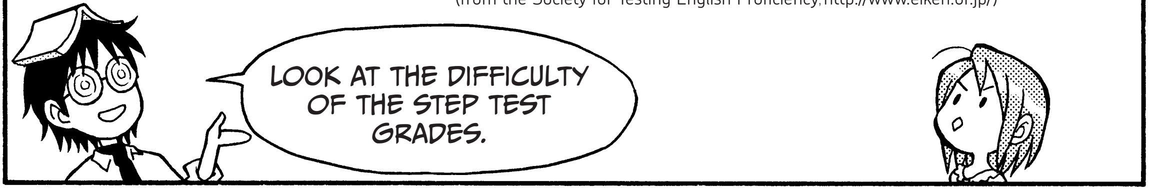


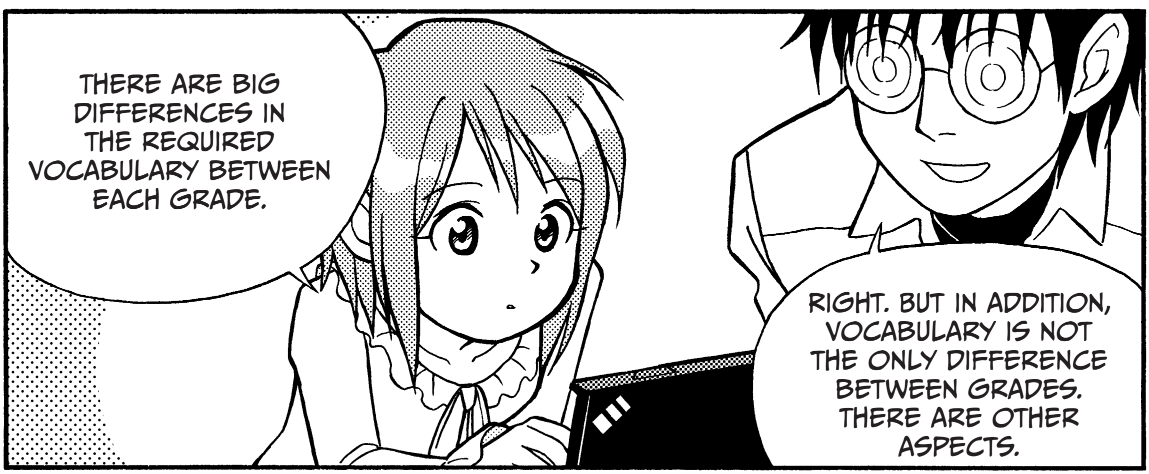


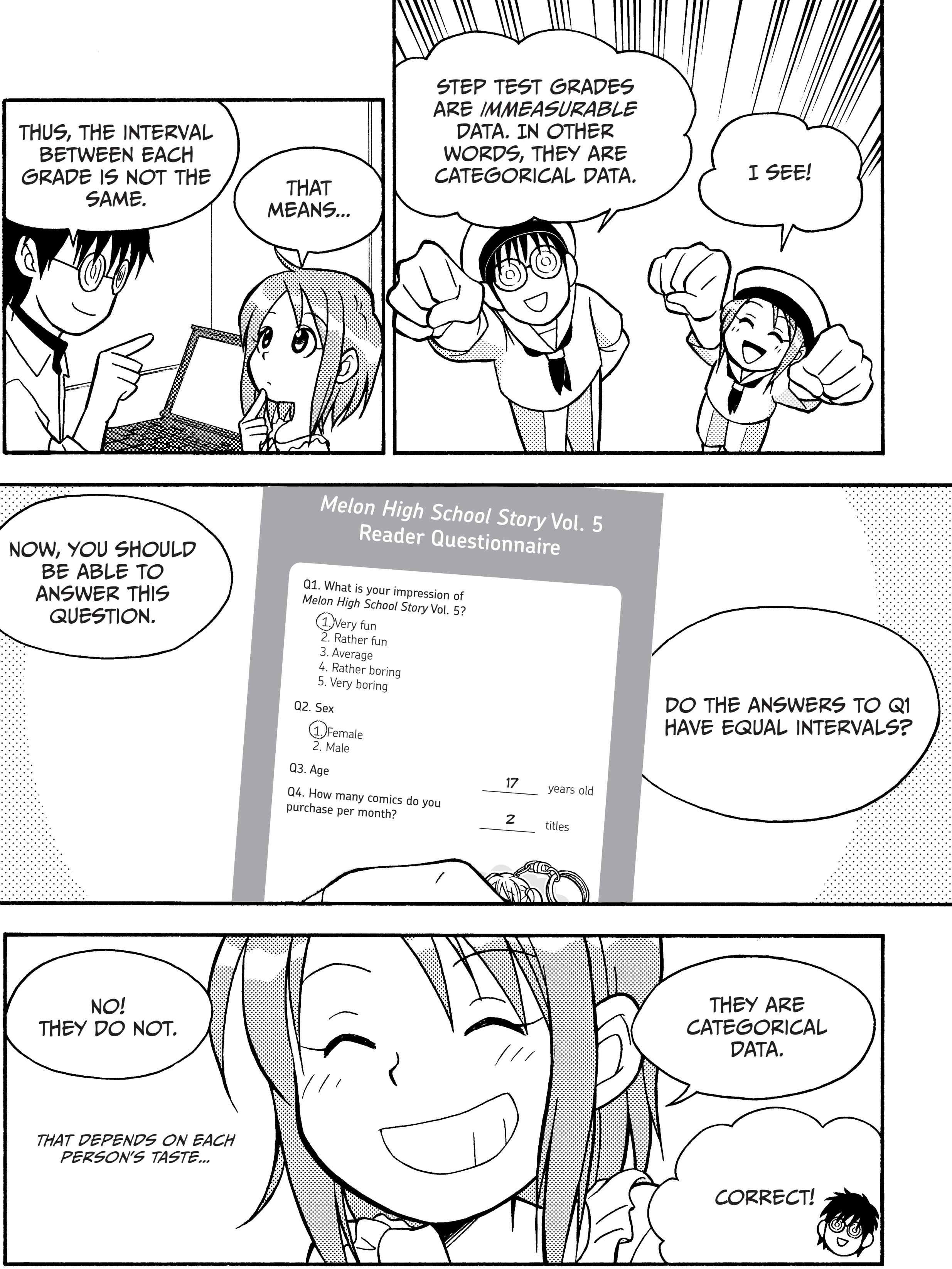
THE STEP TEST GRADES

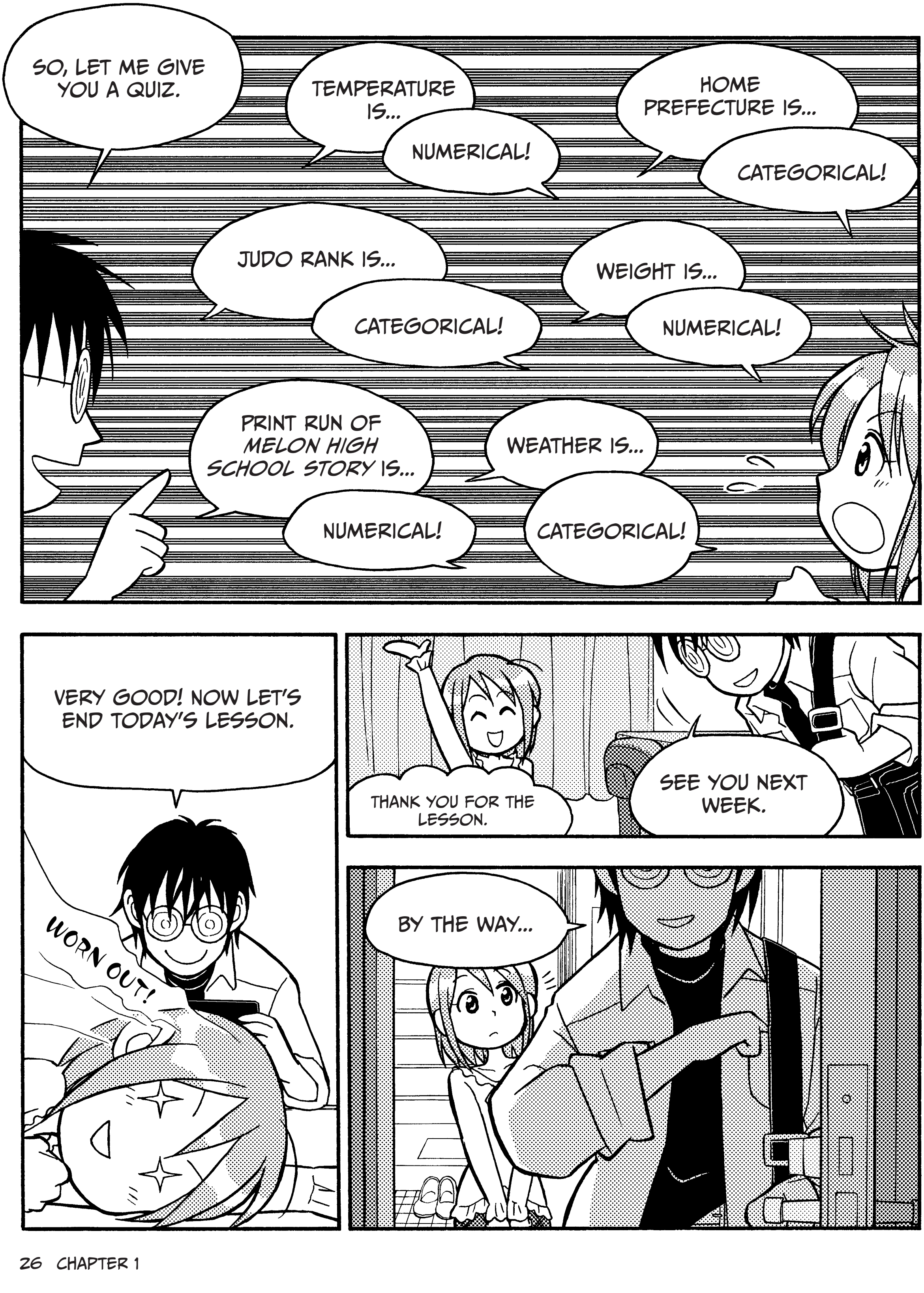
Grade	Requirements
Grade 1	Advanced university graduate level, vocabulary 10,000–15,000 words
Grade 2	High school graduate level, vocabulary 5,100 words
Grade 3	Junior high school graduate level, vocabulary 2,100 words
Grade 4	Intermediate junior high school level, vocabulary 1,300 words
Grade 5	Beginner junior high school level, vocabulary 600 words

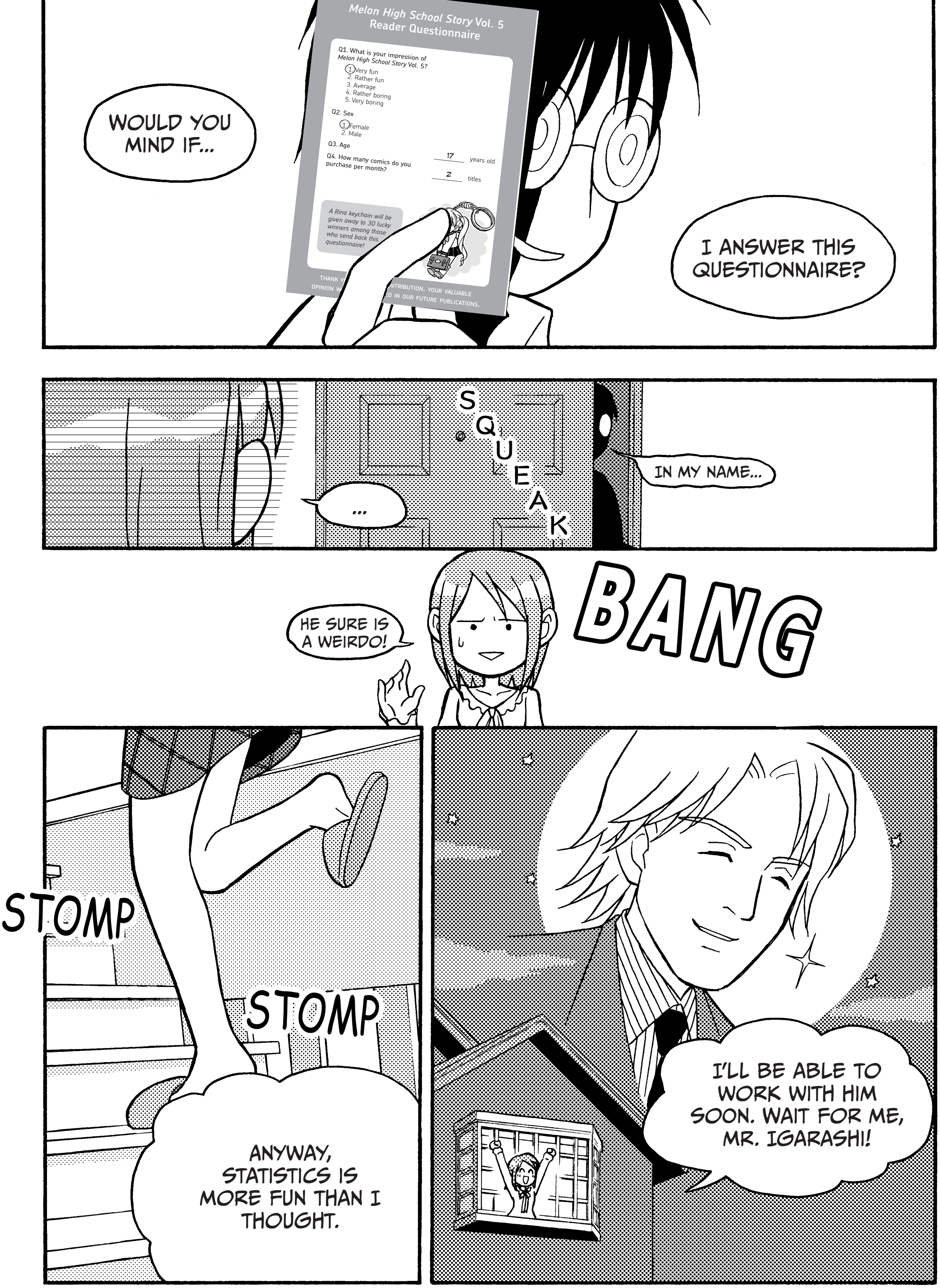
(from the Society for Testing English Proficiency, http://www.eiken.or.jp/)






THUS, THE INTERVAL BETWEEN EACH GRADE IS NOT THE SAME.
THAT MEANS...
STEP TEST GRADES ARE *IMMEASURABLE* DATA. IN OTHER WORDS, THEY ARE CATEGORICAL DATA.
I SEE!
NOW, YOU SHOULD BE ABLE TO ANSWER THIS QUESTION.
Melon High School Story Vol. 5
Reader Questionnaire
Q1. What is your impression of *Melon High School Story* Vol. 5?
1. Very fun
2. Rather fun
3. Average
4. Rather boring
5. Very boring
Q2. Sex
1. Female
2. Male
Q3. Age
17 years old
Q4. How many comics do you purchase per month?
2 titles
DO THE ANSWERS TO Q1 HAVE EQUAL INTERVALS?
NO! THEY DO NOT.
THEY ARE CATEGORICAL DATA.
THAT DEPENDS ON EACH PERSON'S TASTE...
CORRECT!


SO, LET ME GIVE YOU A QUIZ.
TEMPERATURE IS...
NUMERICAL!
HOME PREFECTURE IS...
CATEGORICAL!
JUDO RANK IS...
CATEGORICAL!
WEIGHT IS...
NUMERICAL!
PRINT RUN OF *MELON HIGH SCHOOL STORY* IS...
NUMERICAL!
WEATHER IS...
CATEGORICAL!
VERY GOOD! NOW LET'S END TODAY'S LESSON.
WORN OUT!
THANK YOU FOR THE LESSON.
SEE YOU NEXT WEEK.
BY THE WAY...


Melon High School Story Vol. 5
Reader Questionnaire
Q1. What is your impression of Melon High School Story Vol. 5?
1. Very fun
2. Rather fun
3. Average
4. Rather boring
5. Very boring
Q2. Sex
1. Female
2. Male
Q3. Age
17 years old
Q4. How many comics do you purchase per month?
2 titles
A Rina keychain will be given away to 30 lucky winners among those who send back this questionnaire!
WOULD YOU MIND IF...
I ANSWER THIS QUESTIONNAIRE?
SQUEAK
IN MY NAME...
...
BANG
HE SURE IS A WEIRDO!
STOMP
STOMP
ANYWAY, STATISTICS IS MORE FUN THAN I THOUGHT.
I'LL BE ABLE TO WORK WITH HIM SOON. WAIT FOR ME, MR. IGARASHI!

3. How Multiple-Choice Answers Are Handled in Practice

As mentioned on page 25, the multiple-choice answers for the first question of the readers' questionnaire are categorical data. However, in practice, it is possible to handle such data as numerical data when processing consumer questionnaires and so on. Some examples are below.

Very fun	⇨	5 points
Rather fun	⇨	4 points
Average	⇨	3 points
Rather boring	⇨	2 points
Very boring	⇨	1 point

Very fun	⇨	2 points
Rather fun	⇨	1 point
Average	⇨	0 points
Rather boring	⇨	–1 points
Very boring	⇨	–2 points

The same data is handled differently in theory and in practice. Keep in mind that data may be categorized differently in different situations.

Exercise and Answer

Exercise

Determine whether the data in the following table is categorical data or numerical data.

Respondent	Blood type	Opinion on sports drink *X*	Comfortable air conditioning temperature (°C)	100m track race record (seconds)
Mr./Ms. A	B	Not good	25	14.1
Mr./Ms. B	A	Good	24	12.2
Mr./Ms. C	AB	Good	25	17.0
Mr./Ms. D	O	Average	27	15.6
Mr./Ms. E	A	Not good	24	18.4
...	...	...	...	...

Answer

Blood type and opinion on sports drink *X* are examples of categorical data. Comfortable air conditioning temperature and 100m track race record are examples of numerical data.

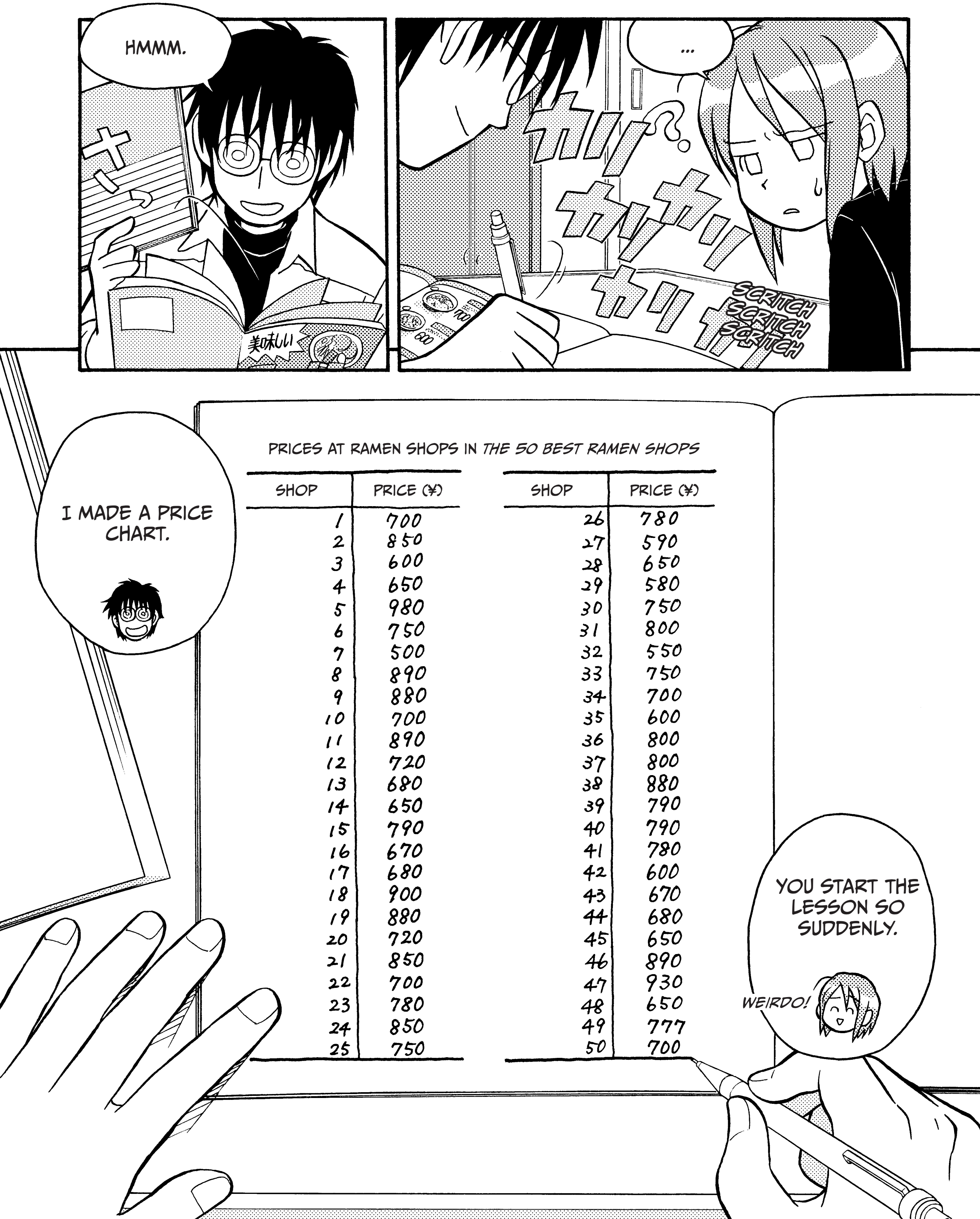
Summary

- Data is classified as categorical data or numerical data.
- Some data, such as "very fun" or "very boring," is theoretically categorical data. However, in practice, it is possible to treat it as numerical data.

2

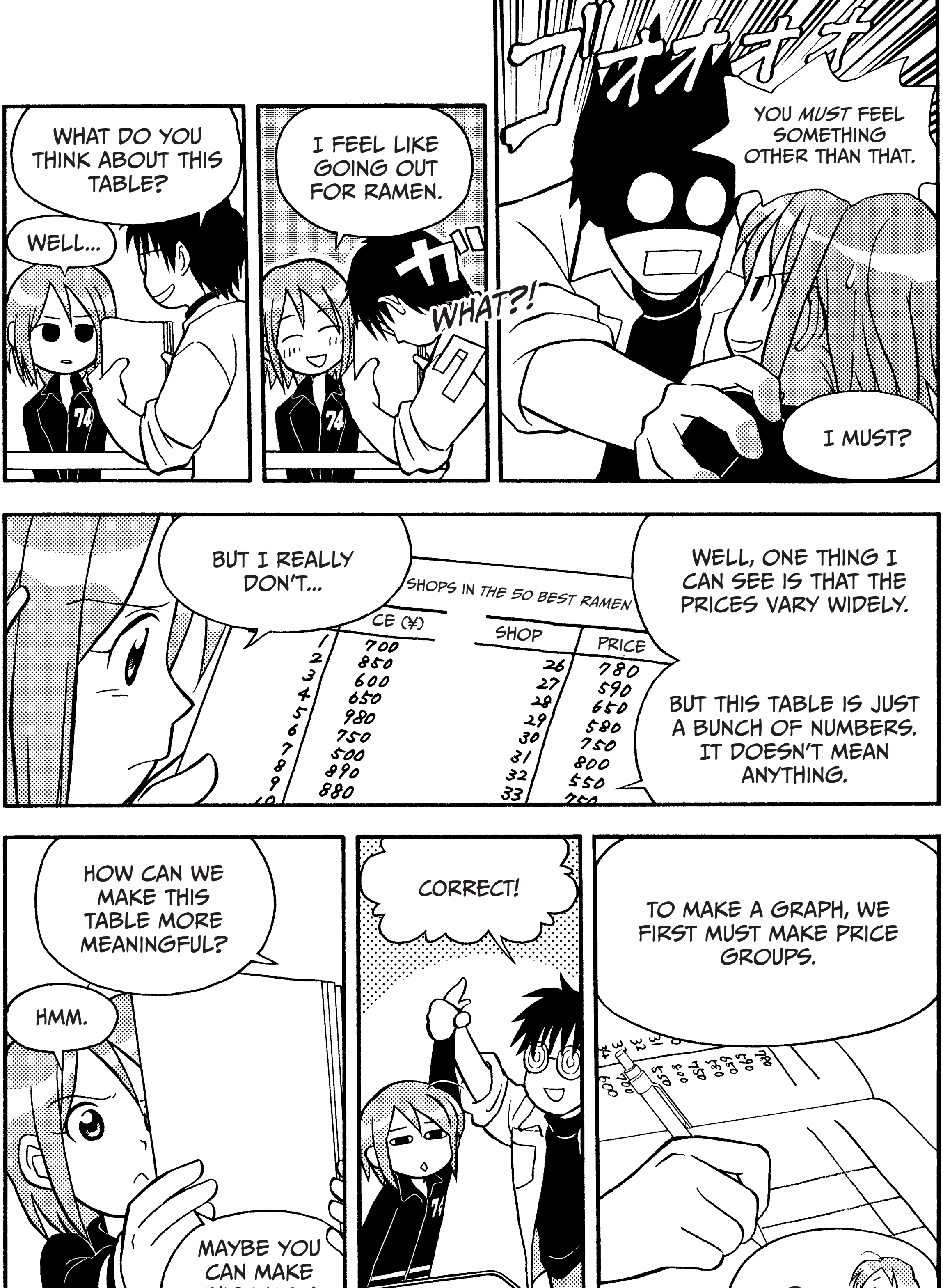
GETTING THE BIG PICTURE: UNDERSTANDING NUMERICAL DATA


1. FREQUENCY DISTRIBUTION TABLES AND HISTOGRAMS
真剣
SQUEAK
ギイ…
HELLO, RUI.
HELLO, MR. YAMAMOTO!
WHAT ARE YOU READING? YOU LIKE RAMEN?
美味しい ラーメン* best 50
*THE 50 BEST RAMEN SHOPS
I WAS LOOKING AT THIS MAGAZINE TO CHOOSE WHICH RESTAURANT TO EAT AT.
OH!
THEY ALL LOOK SO GOOD, DON'T THEY?
渋谷
500
890
880
700
890



PRICES AT RAMEN SHOPS IN *THE 50 BEST RAMEN SHOPS*

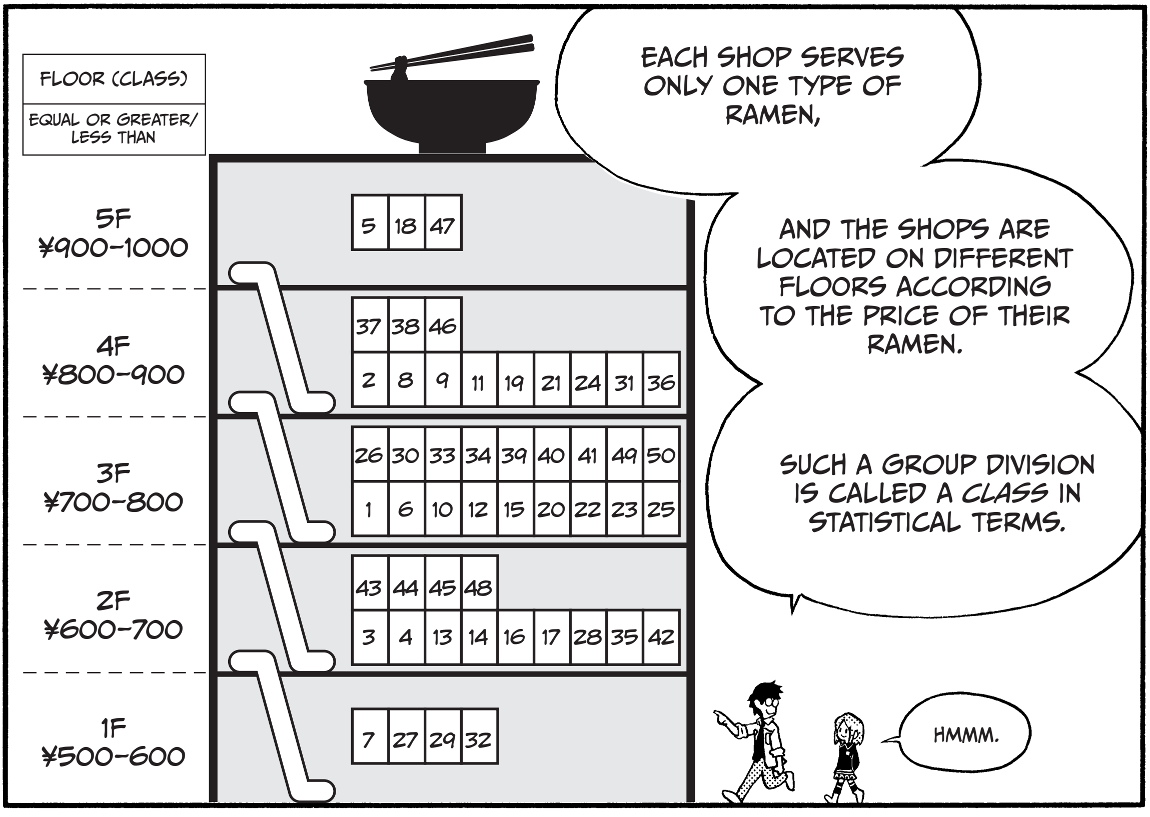
SHOP	PRICE (¥)	SHOP	PRICE (¥)
1	700	26	780
2	850	27	590
3	600	28	650
4	650	29	580
5	980	30	750
6	750	31	800
7	500	32	550
8	890	33	750
9	880	34	700
10	700	35	600
11	890	36	800
12	720	37	800
13	680	38	880
14	650	39	790
15	790	40	790
16	670	41	780
17	680	42	600
18	900	43	670
19	880	44	680
20	720	45	650
21	850	46	890
22	700	47	930
23	780	48	650
24	850	49	777
25	750	50	700



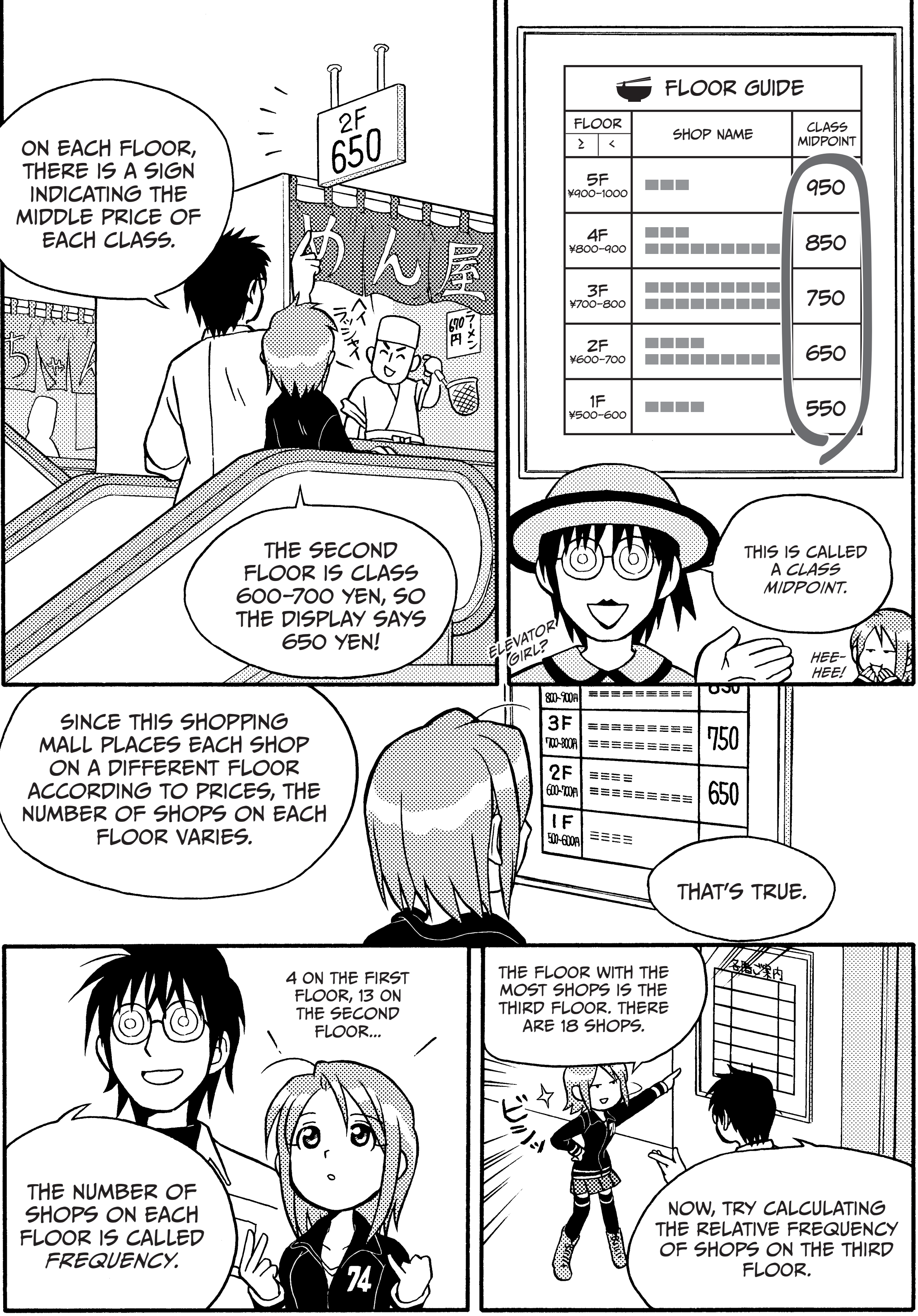
WHAT DO YOU THINK ABOUT THIS TABLE?
WELL...
I FEEL LIKE GOING OUT FOR RAMEN.
WHAT?!
YOU *MUST* FEEL SOMETHING OTHER THAN THAT.
I MUST?
BUT I REALLY DON'T...
SHOPS IN THE 50 BEST RAMEN
SHOP
PRICE
WELL, ONE THING I CAN SEE IS THAT THE PRICES VARY WIDELY.
BUT THIS TABLE IS JUST A BUNCH OF NUMBERS. IT DOESN'T MEAN ANYTHING.
HOW CAN WE MAKE THIS TABLE MORE MEANINGFUL?
HMM.
MAYBE YOU CAN MAKE THIS INTO A GRAPH?
CORRECT!
TO MAKE A GRAPH, WE FIRST MUST MAKE PRICE GROUPS.
?

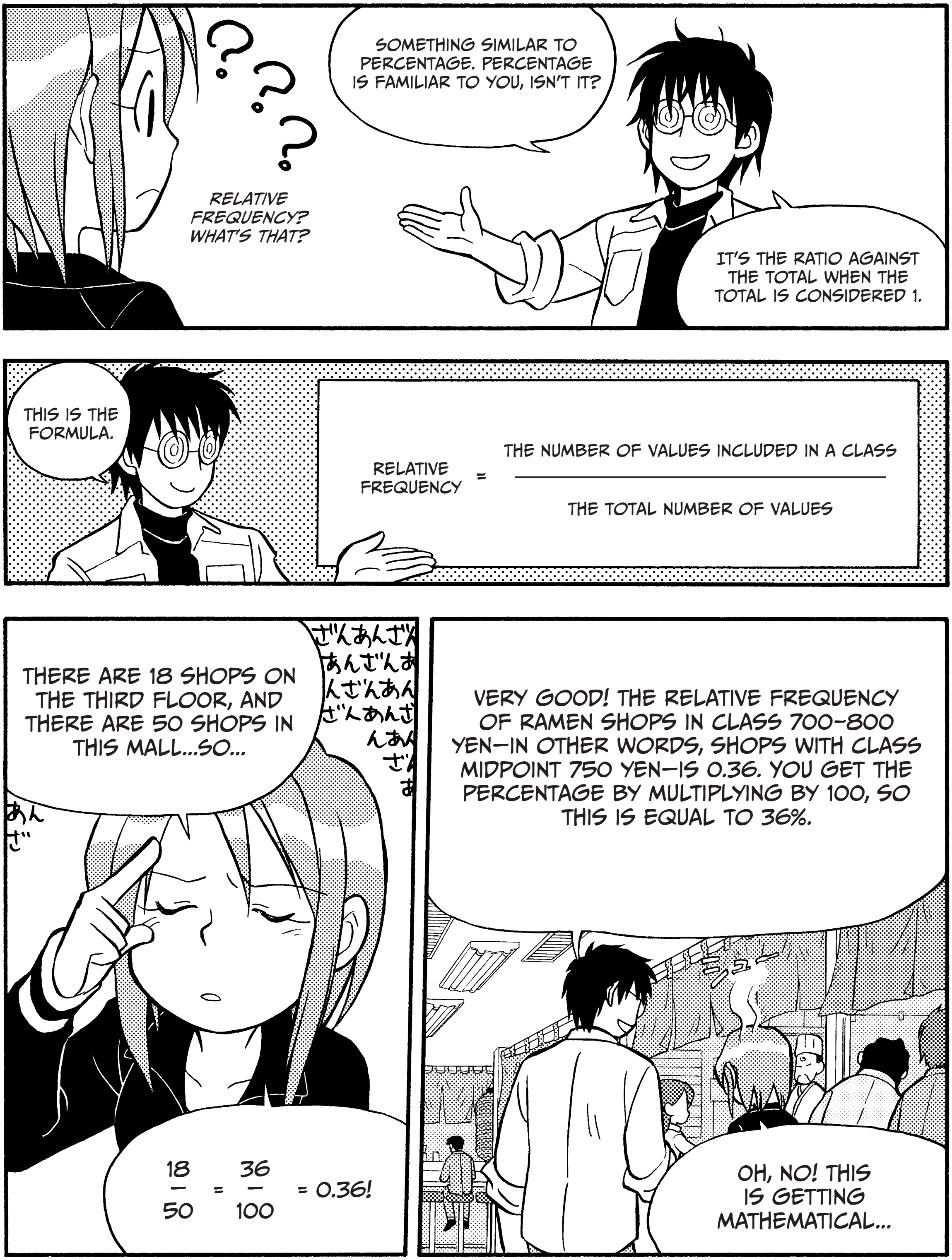
SUPPOSE...
*PUSH

THERE IS A HUGE SHOPPING MALL CONSISTING OF 50 RAMEN SHOPS...AND *ONLY* RAMEN SHOPS.
ラーメン万歳
SOMEHOW RUI HAS TURNED INTO AN ELEVATOR GIRL?
I WOULD GO EVERY DAY!



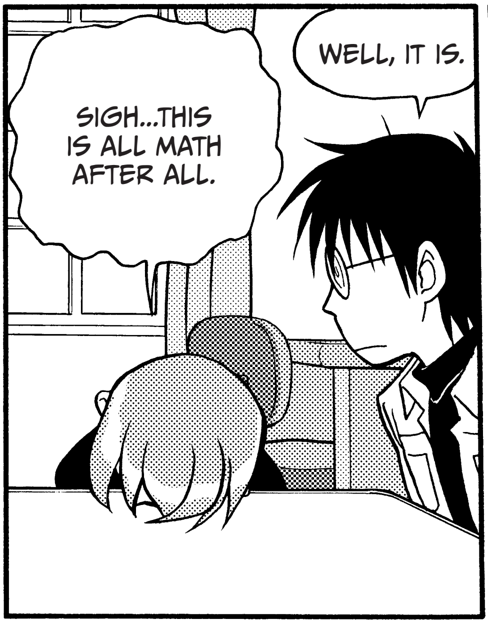
FLOOR (CLASS)
EQUAL OR GREATER/ LESS THAN
5F ¥900–1000
5 18 47
4F ¥800–900
37 38 46
2 8 9 11 19 21 24 31 36
3F ¥700–800
26 30 33 34 39 40 41 49 50
1 6 10 12 15 20 22 23 25
2F ¥600–700
43 44 45 48
3 4 13 14 16 17 28 35 42
1F ¥500–600
7 27 29 32
EACH SHOP SERVES ONLY ONE TYPE OF RAMEN,
AND THE SHOPS ARE LOCATED ON DIFFERENT FLOORS ACCORDING TO THE PRICE OF THEIR RAMEN.
SUCH A GROUP DIVISION IS CALLED A *CLASS* IN STATISTICAL TERMS.
HMMM.

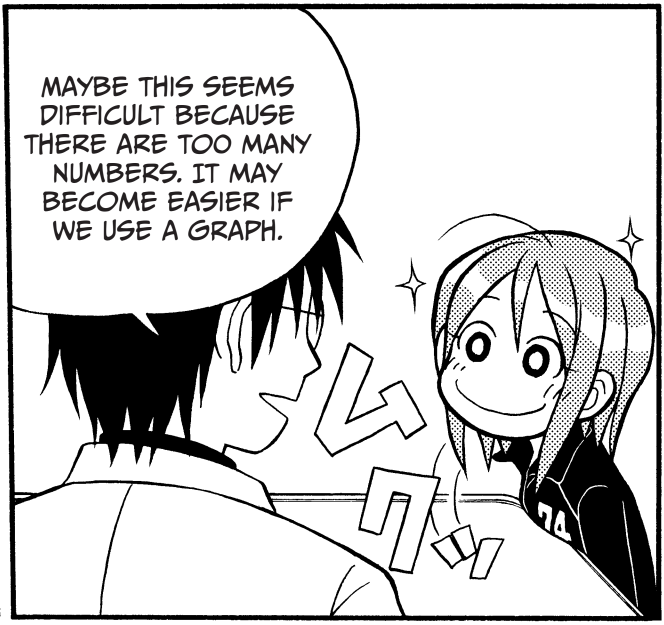

2F 650
ON EACH FLOOR, THERE IS A SIGN INDICATING THE MIDDLE PRICE OF EACH CLASS.
THE SECOND FLOOR IS CLASS 600–700 YEN, SO THE DISPLAY SAYS 650 YEN!
FLOOR GUIDE
FLOOR ≥ <
SHOP NAME
CLASS MIDPOINT
5F ¥900–1000 950
4F ¥800–900 850
3F ¥700–800 750
2F ¥600–700 650
1F ¥500–600 550
THIS IS CALLED A CLASS MIDPOINT.
ELEVATOR GIRL?
HEE-HEE!
SINCE THIS SHOPPING MALL PLACES EACH SHOP ON A DIFFERENT FLOOR ACCORDING TO PRICES, THE NUMBER OF SHOPS ON EACH FLOOR VARIES.
THAT'S TRUE.
4 ON THE FIRST FLOOR, 13 ON THE SECOND FLOOR...
THE NUMBER OF SHOPS ON EACH FLOOR IS CALLED *FREQUENCY*.
THE FLOOR WITH THE MOST SHOPS IS THE THIRD FLOOR. THERE ARE 18 SHOPS.
NOW, TRY CALCULATING THE RELATIVE FREQUENCY OF SHOPS ON THE THIRD FLOOR.

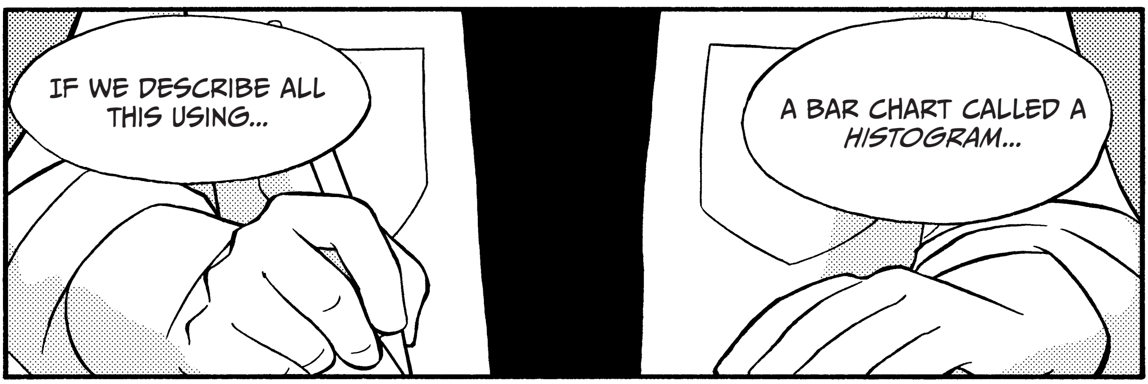

RELATIVE FREQUENCY? WHAT'S THAT?
SOMETHING SIMILAR TO PERCENTAGE. PERCENTAGE IS FAMILIAR TO YOU, ISN'T IT?
IT'S THE RATIO AGAINST THE TOTAL WHEN THE TOTAL IS CONSIDERED 1.
THIS IS THE FORMULA.
RELATIVE FREQUENCY = THE NUMBER OF VALUES INCLUDED IN A CLASS / THE TOTAL NUMBER OF VALUES
THERE ARE 18 SHOPS ON THE THIRD FLOOR, AND THERE ARE 50 SHOPS IN THIS MALL...SO...
$\frac{18}{50} = \frac{36}{100} = 0.36!$
VERY GOOD! THE RELATIVE FREQUENCY OF RAMEN SHOPS IN CLASS 700–800 YEN—IN OTHER WORDS, SHOPS WITH CLASS MIDPOINT 750 YEN—IS 0.36. YOU GET THE PERCENTAGE BY MULTIPLYING BY 100, SO THIS IS EQUAL TO 36%.
OH, NO! THIS IS GETTING MATHEMATICAL...

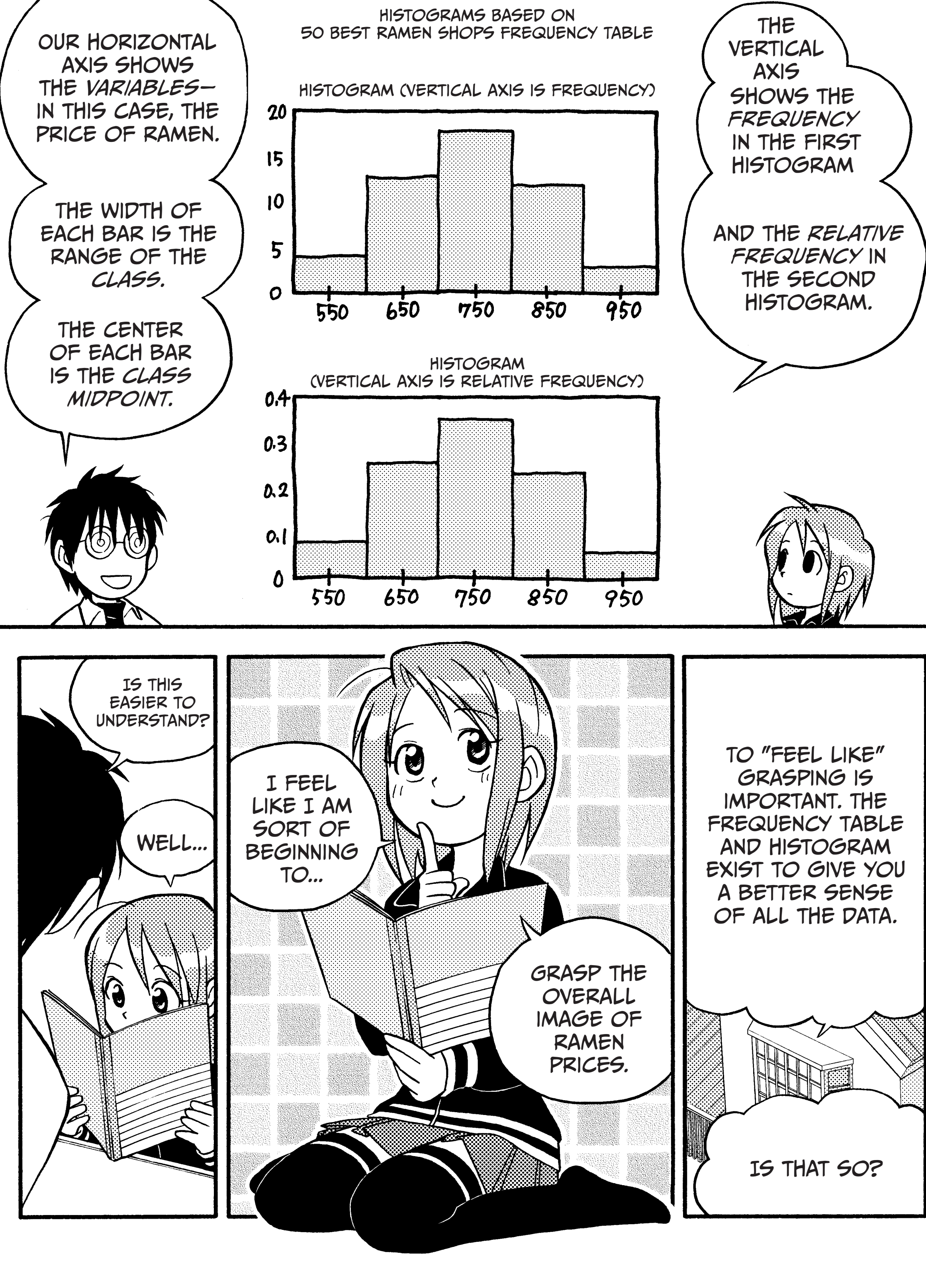
50 BEST RAMEN SHOPS FREQUENCY TABLE

CLASS (EQUAL OR GREATER/ LESS THAN)	CLASS MIDPOINT	FREQUENCY	RELATIVE FREQUENCY
500–600	550	4	0.08
600–700	650	13	0.26
700–800	750	18	0.36
800–900	850	12	0.24
900–1000	950	3	0.06
SUM		50	1.00

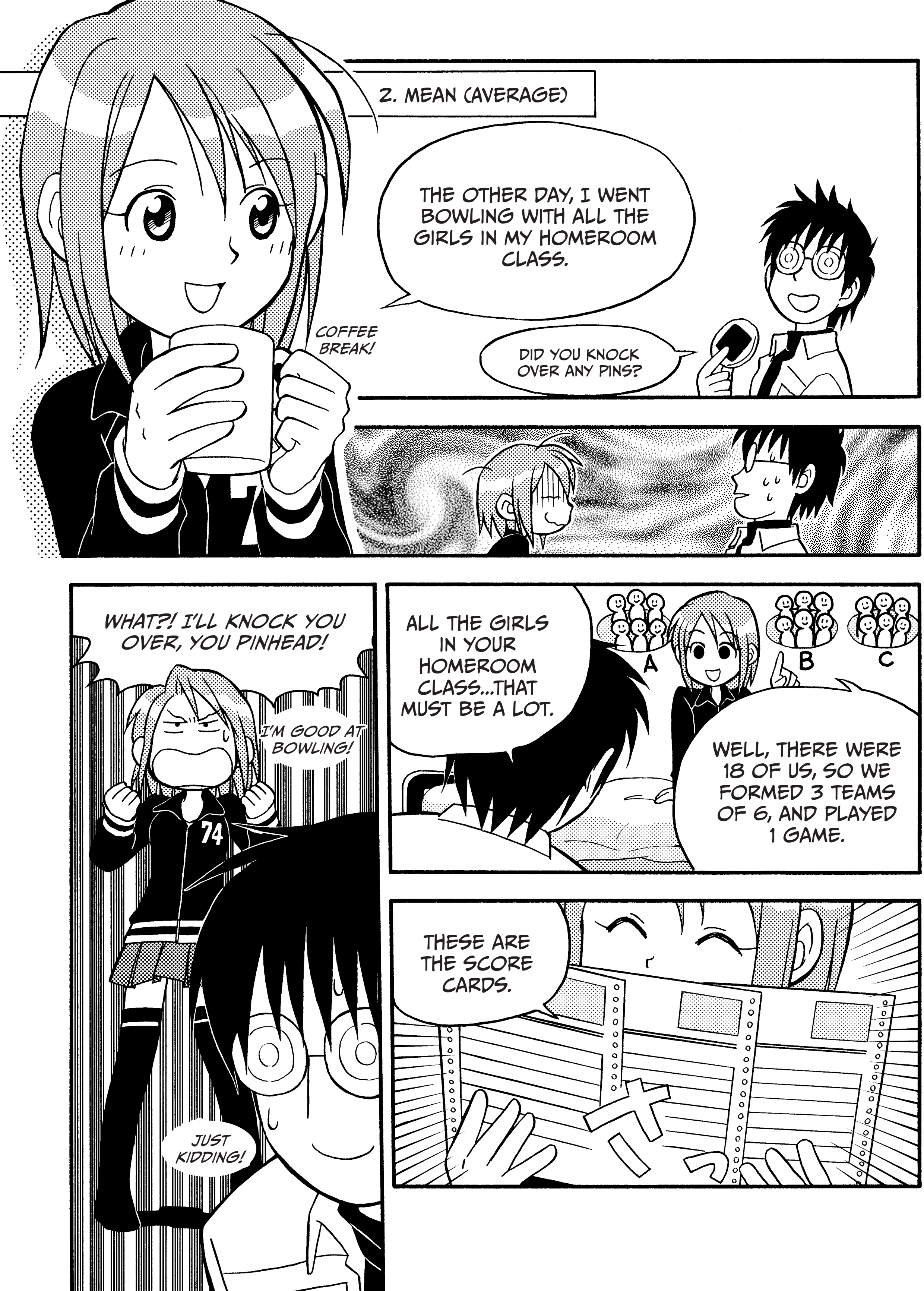


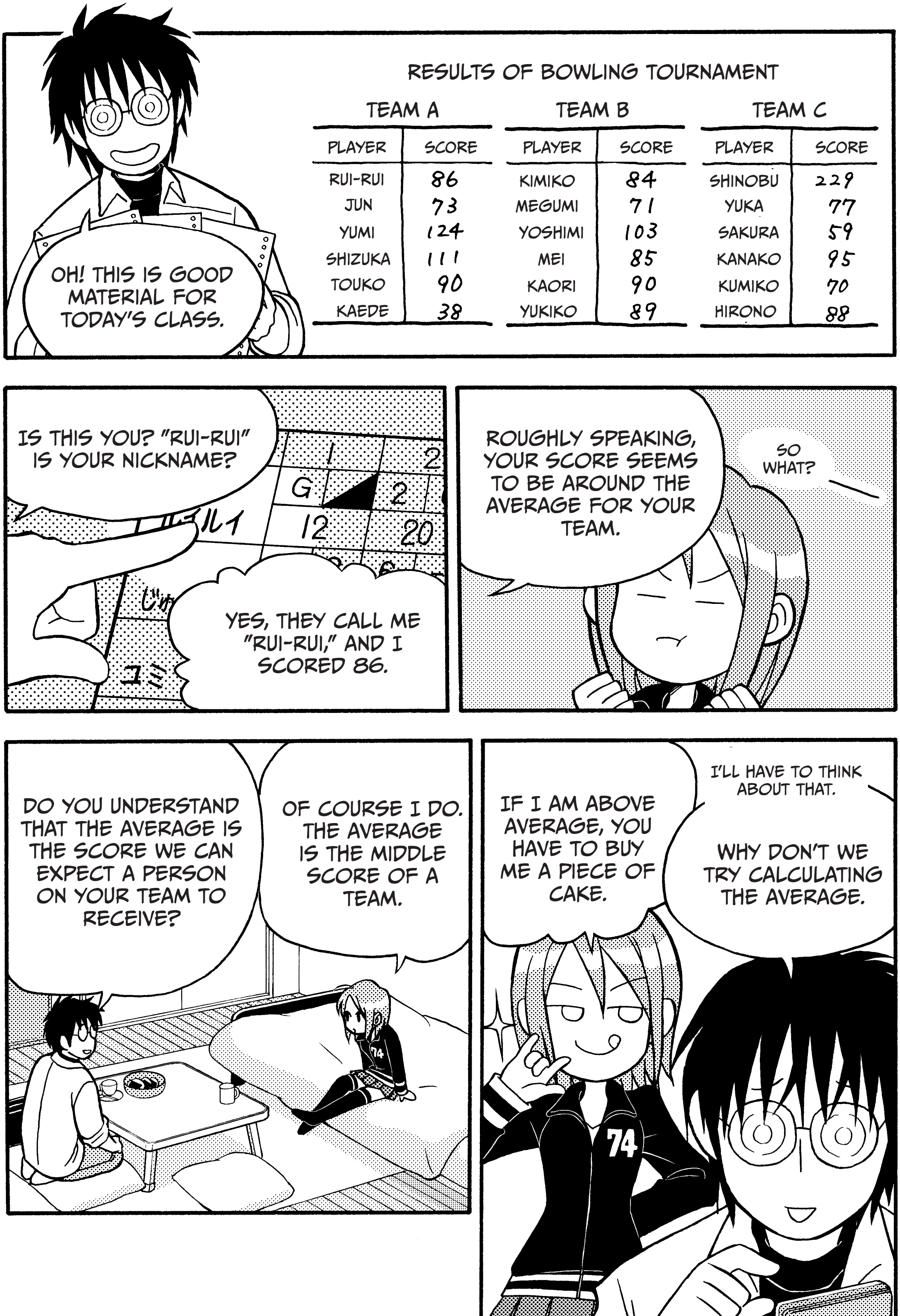




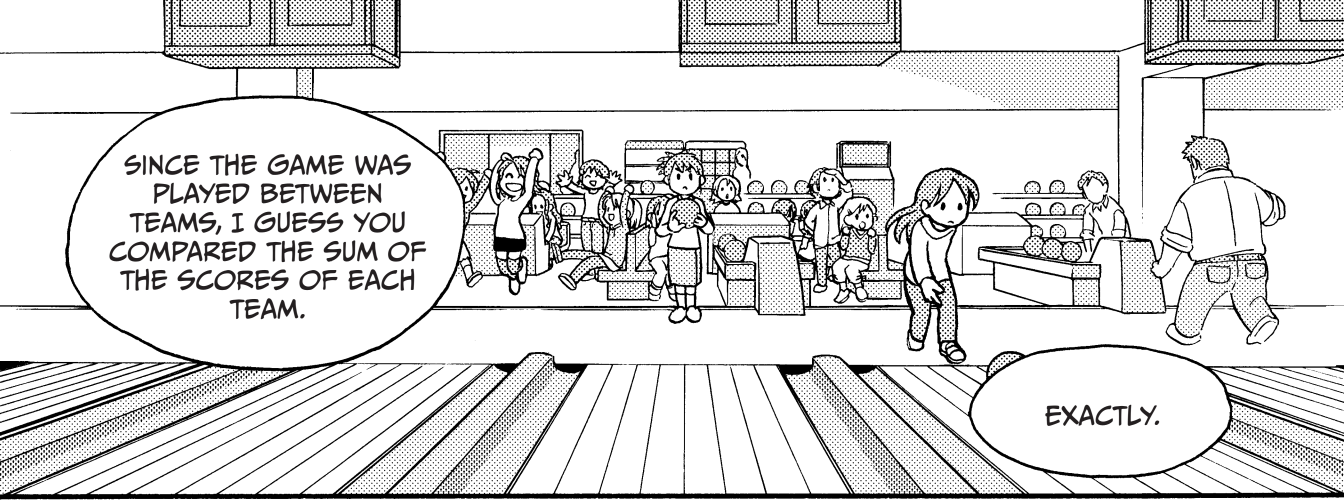


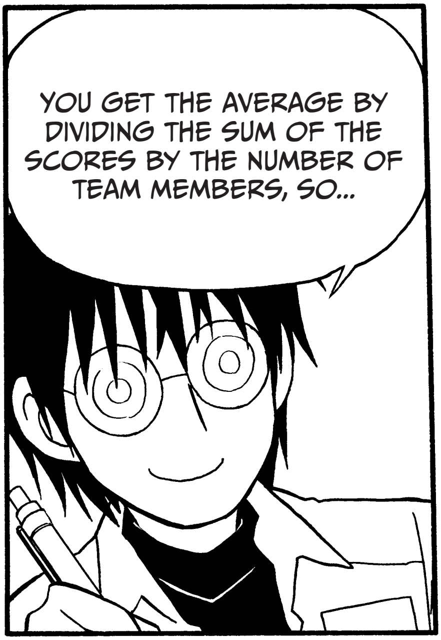
OUR HORIZONTAL AXIS SHOWS THE *VARIABLES*— IN THIS CASE, THE PRICE OF RAMEN.
THE WIDTH OF EACH BAR IS THE RANGE OF THE *CLASS*.
THE CENTER OF EACH BAR IS THE *CLASS MIDPOINT*.
HISTOGRAMS BASED ON 50 BEST RAMEN SHOPS FREQUENCY TABLE
HISTOGRAM (VERTICAL AXIS IS FREQUENCY)
20
15
10
5
0
550
650
750
850
950
HISTOGRAM (VERTICAL AXIS IS RELATIVE FREQUENCY)
0.4
0.3
0.2
0.1
0
550
650
750
850
950
THE VERTICAL AXIS SHOWS THE *FREQUENCY* IN THE FIRST HISTOGRAM
AND THE *RELATIVE FREQUENCY* IN THE SECOND HISTOGRAM.
IS THIS EASIER TO UNDERSTAND?
WELL...
I FEEL LIKE I AM SORT OF BEGINNING TO...
GRASP THE OVERALL IMAGE OF RAMEN PRICES.
TO "FEEL LIKE" GRASPING IS IMPORTANT. THE FREQUENCY TABLE AND HISTOGRAM EXIST TO GIVE YOU A BETTER SENSE OF ALL THE DATA.
IS THAT SO?


2. MEAN (AVERAGE)
THE OTHER DAY, I WENT BOWLING WITH ALL THE GIRLS IN MY HOMEROOM CLASS.
COFFEE BREAK!
DID YOU KNOCK OVER ANY PINS?
WHAT?! I'LL KNOCK YOU OVER, YOU PINHEAD!
I'M GOOD AT BOWLING!
JUST KIDDING!
ALL THE GIRLS IN YOUR HOMEROOM CLASS...THAT MUST BE A LOT.
A
B
C
WELL, THERE WERE 18 OF US, SO WE FORMED 3 TEAMS OF 6, AND PLAYED 1 GAME.
THESE ARE THE SCORE CARDS.


RESULTS OF BOWLING TOURNAMENT

TEAM A PLAYER	TEAM A SCORE	TEAM B PLAYER	TEAM B SCORE	TEAM C PLAYER	TEAM C SCORE
RUI-RUI	86	KIMIKO	84	SHINOBU	229
JUN	73	MEGUMI	71	YUKA	77
YUMI	124	YOSHIMI	103	SAKURA	59
SHIZUKA	111	MEI	85	KANAKO	95
TOUKO	90	KAORI	90	KUMIKO	70
KAEDE	38	YUKIKO	89	HIRONO	88





TEAM A

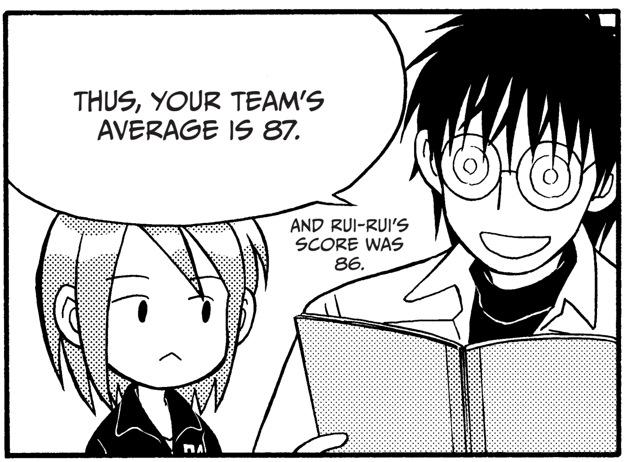
$$\frac{86+73+124+111+90+38}{6}=\frac{522}{6}=87$$

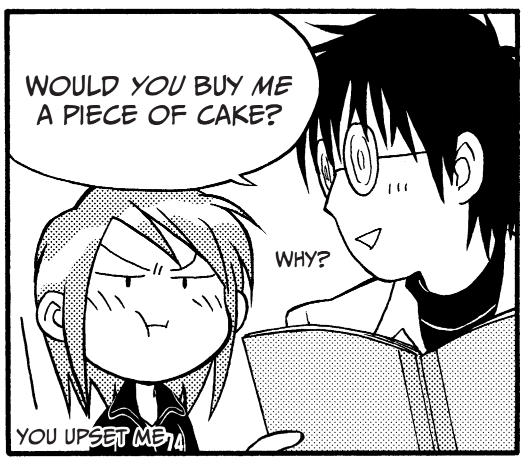
TEAM B

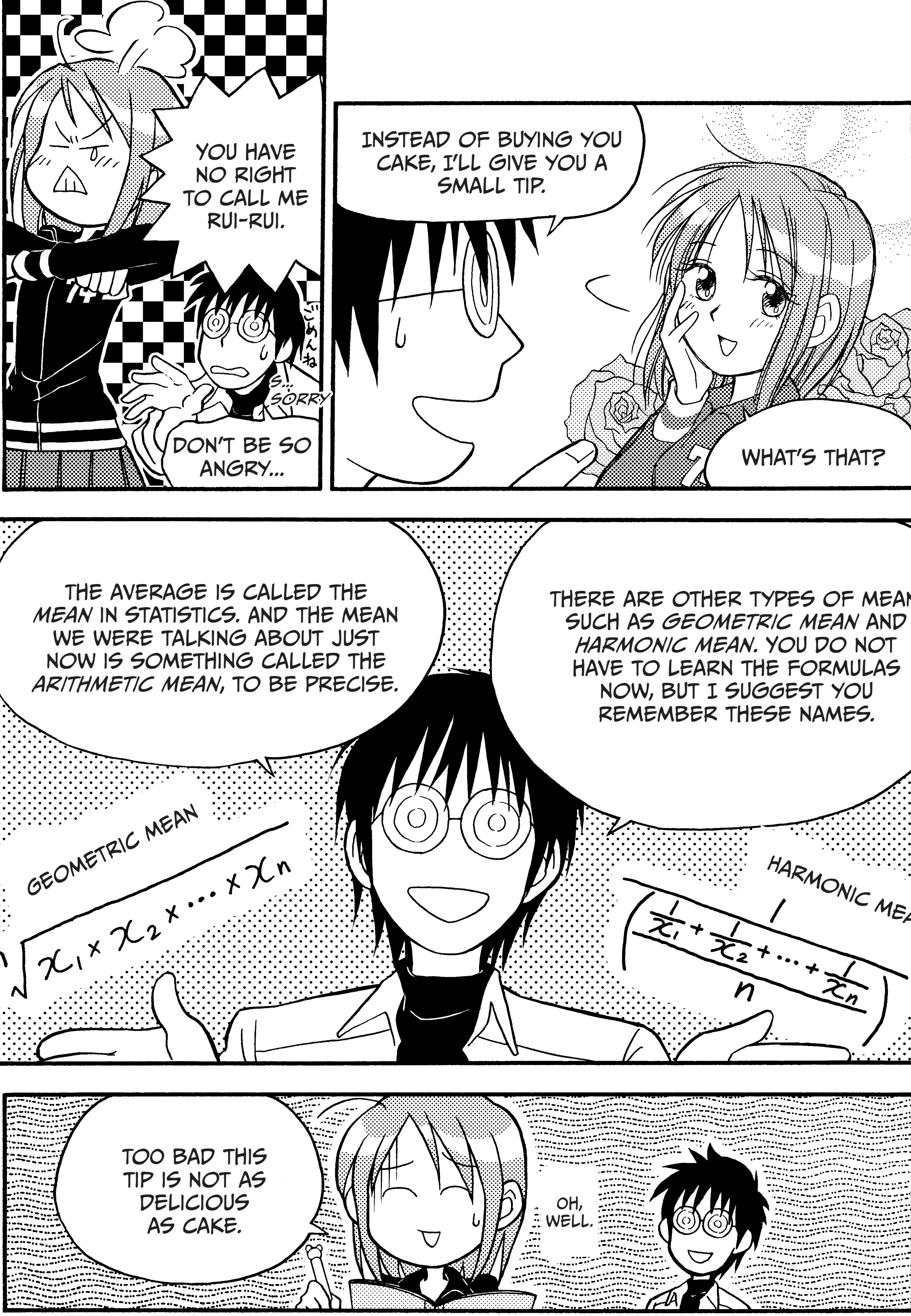
$$\frac{84+71+103+85+90+89}{6}=\frac{522}{6}=87$$

TEAM C

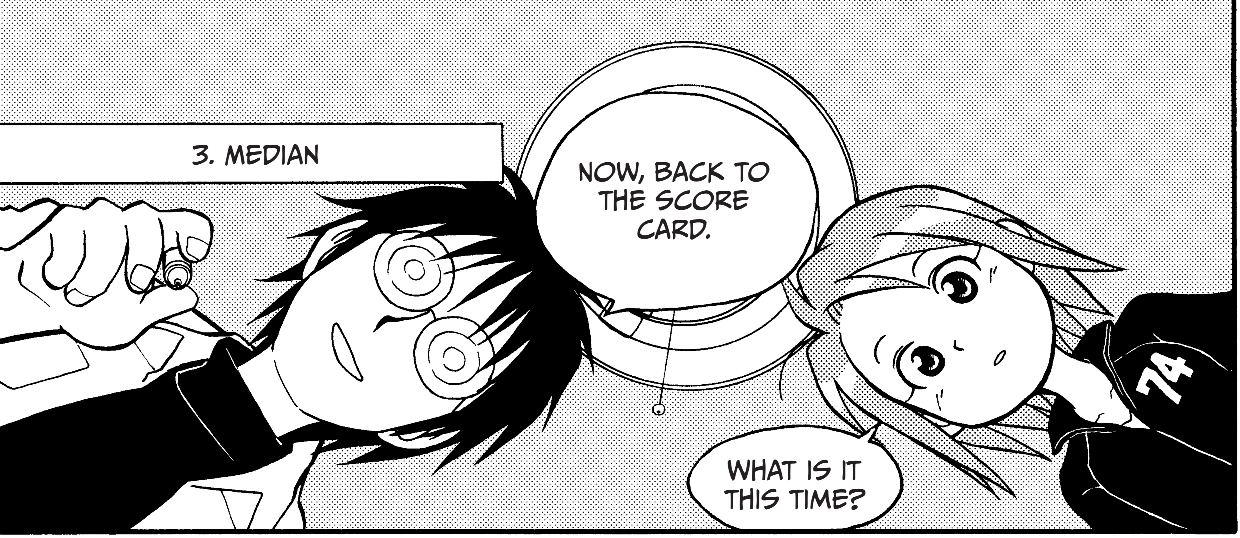
$$\frac{229+77+59+95+70+88}{6}=\frac{618}{6}=103$$






YOU HAVE NO RIGHT TO CALL ME RUI-RUI.
ごめんね
S... SORRY
DON'T BE SO ANGRY...
INSTEAD OF BUYING YOU CAKE, I'LL GIVE YOU A SMALL TIP.
WHAT'S THAT?
THE AVERAGE IS CALLED THE *MEAN* IN STATISTICS. AND THE MEAN WE WERE TALKING ABOUT JUST NOW IS SOMETHING CALLED THE *ARITHMETIC MEAN*, TO BE PRECISE.
THERE ARE OTHER TYPES OF MEAN, SUCH AS *GEOMETRIC MEAN* AND *HARMONIC MEAN*. YOU DO NOT HAVE TO LEARN THE FORMULAS NOW, BUT I SUGGEST YOU REMEMBER THESE NAMES.
GEOMETRIC MEAN
$\sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$
HARMONIC MEAN
$\left(\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}{n} \right)^{-1}$
TOO BAD THIS TIP IS NOT AS DELICIOUS AS CAKE.
OH, WELL.

3. MEDIAN

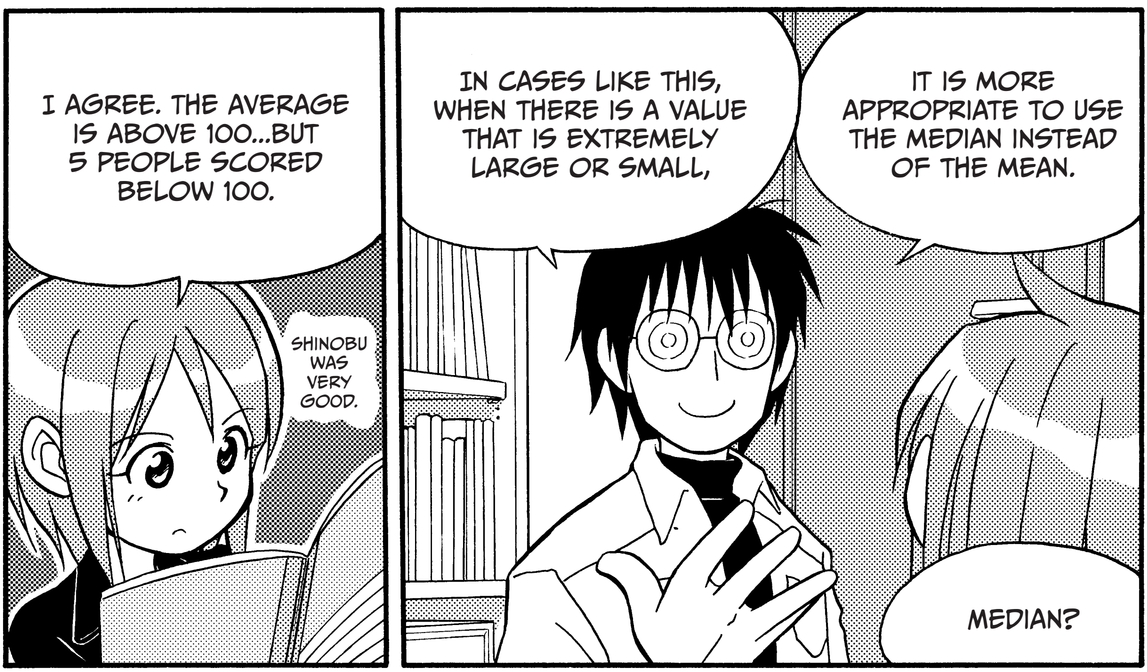


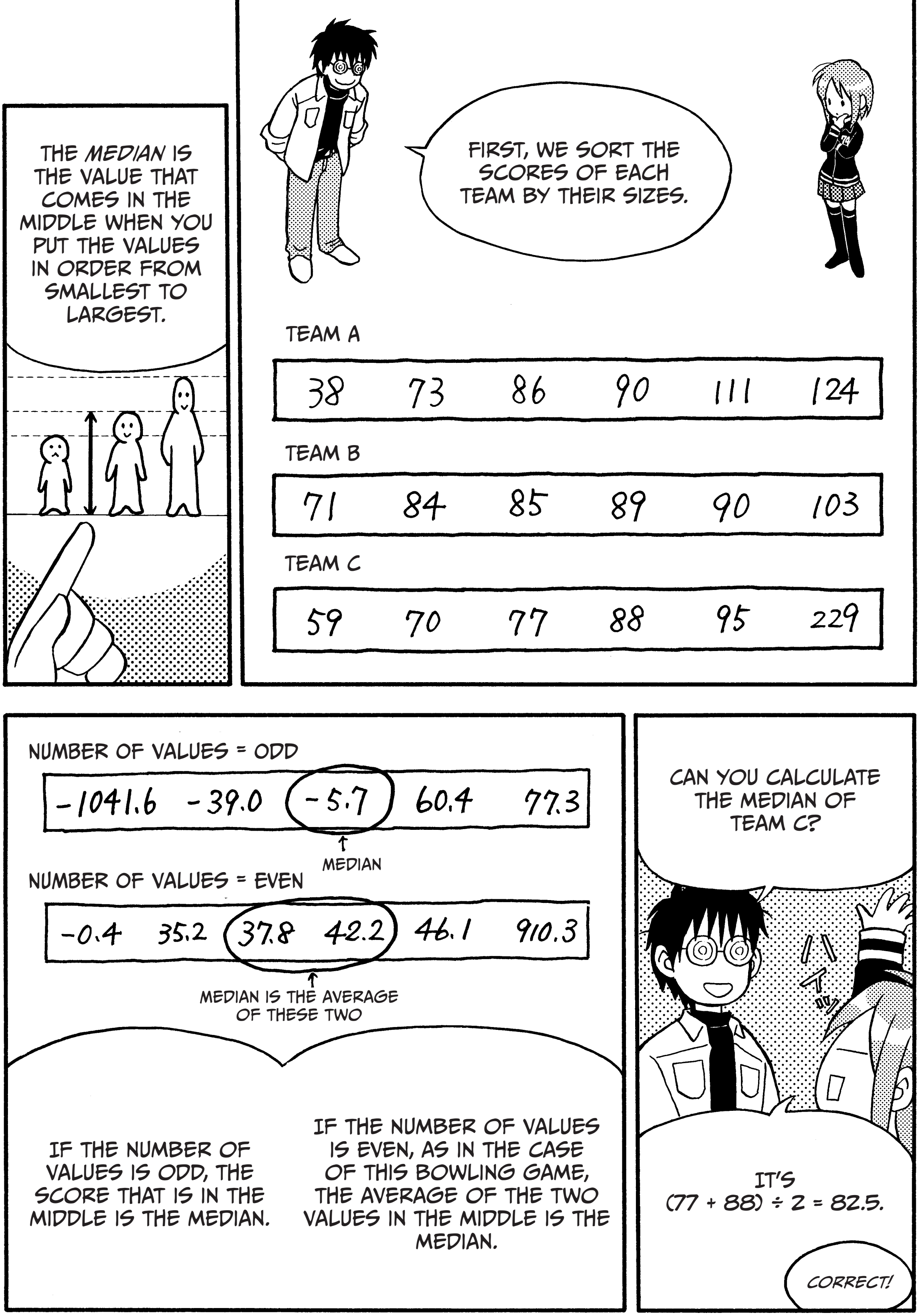
LET'S IGNORE TEAMS A AND B FOR NOW, AND LOOK AT TEAM C...

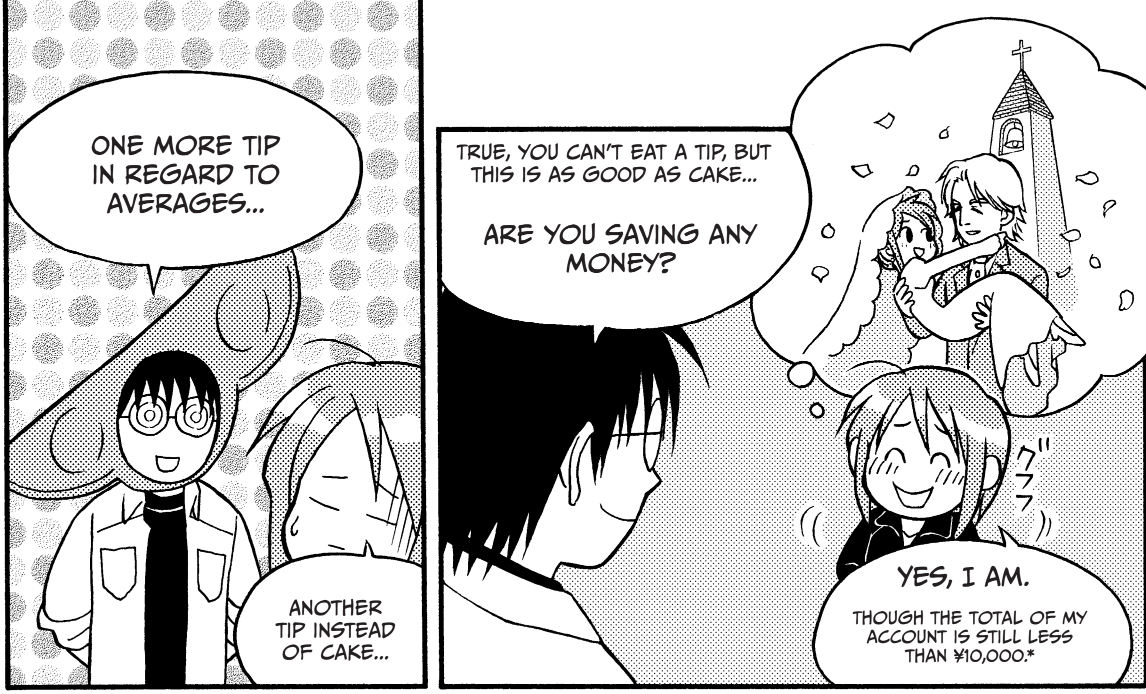
RESULTS OF BOWLING TOURNAMENT

TEAM A		TEAM B		TEAM C	
PLAYER	SCORE	PLAYER	SCORE	PLAYER	SCORE
RUI-RUI	86	KIMIKO	84	SHINOBU	229
JUN	73	MEGUMI	71	YUKA	77
YUMI	124	YOSHIMI	103	SAKURA	59
SHIZUKA	111	MEI	85	KANAKO	95
TOUKO	90	KAORI	90	KUMIKO	70
KAEDE	38	YUKIKO	89	HIRONO	88

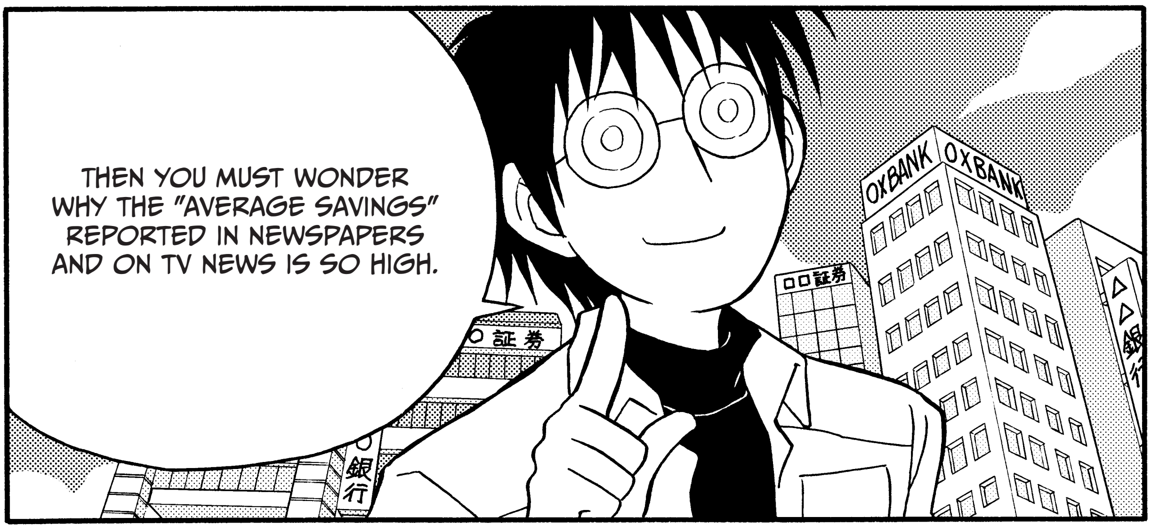
HERE, I DON'T THINK YOU CAN REALLY SAY THAT THE AVERAGE IS "ROUGHLY THE SCORE OF EACH PERSON."

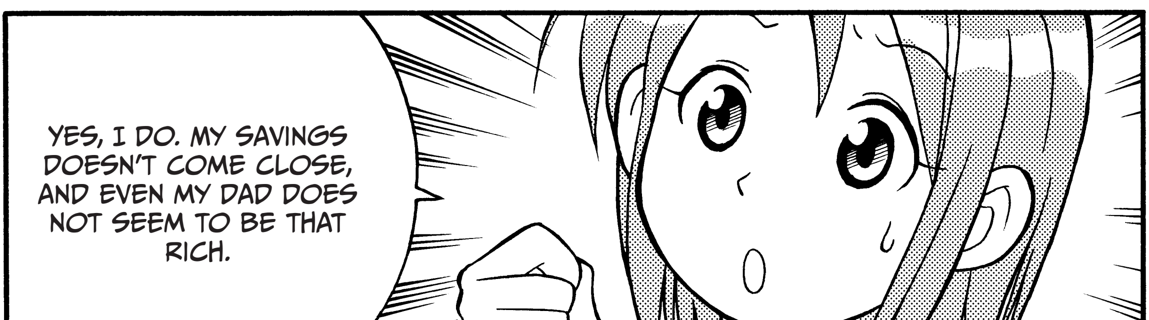


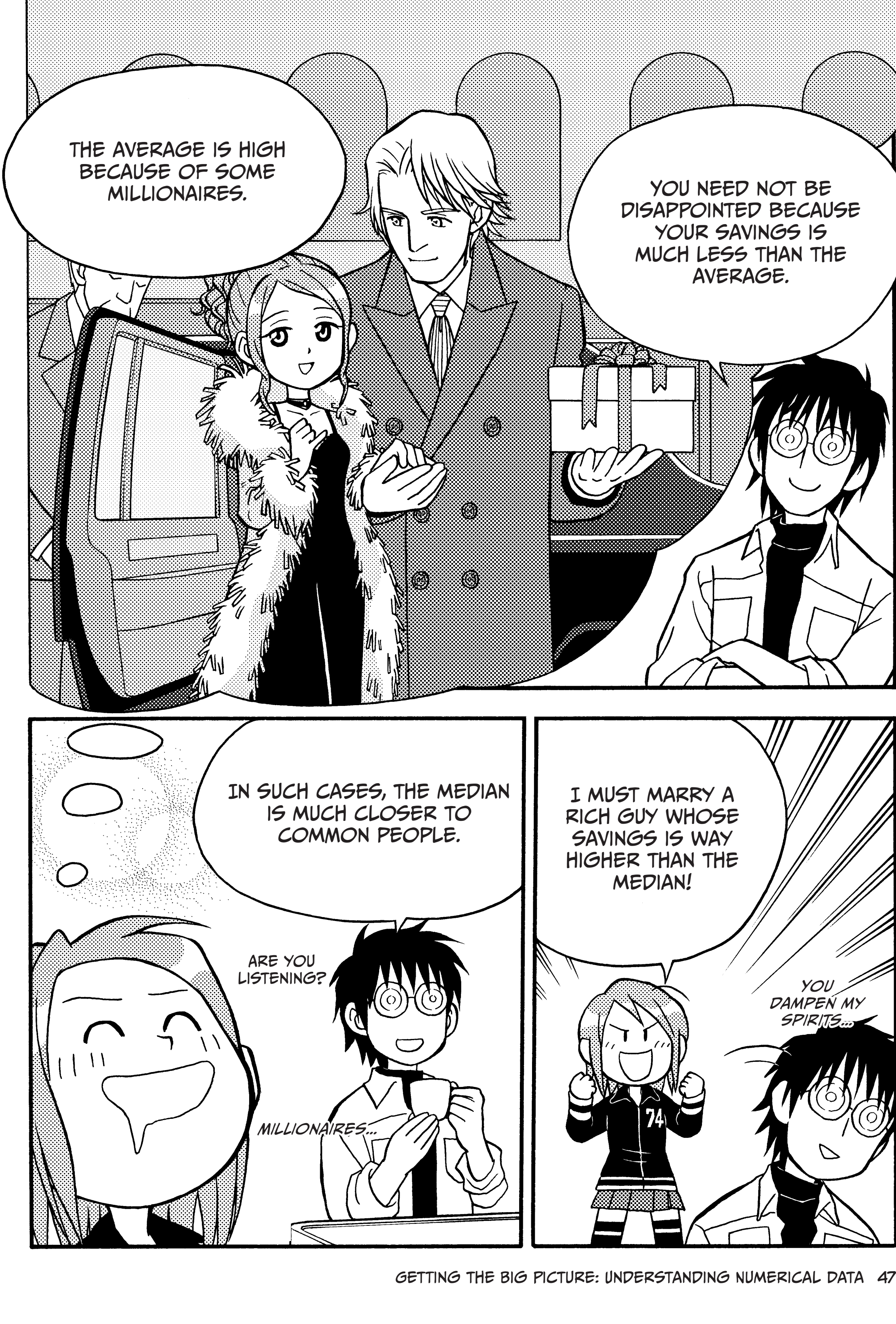

THE *MEDIAN* IS THE VALUE THAT COMES IN THE MIDDLE WHEN YOU PUT THE VALUES IN ORDER FROM SMALLEST TO LARGEST.
FIRST, WE SORT THE SCORES OF EACH TEAM BY THEIR SIZES.
TEAM A
38 73 86 90 111 124
TEAM B
71 84 85 89 90 103
TEAM C
59 70 77 88 95 229
NUMBER OF VALUES = ODD
-1041.6 -39.0 -5.7 60.4 77.3
MEDIAN
NUMBER OF VALUES = EVEN
-0.4 35.2 37.8 42.2 46.1 910.3
MEDIAN IS THE AVERAGE OF THESE TWO
IF THE NUMBER OF VALUES IS ODD, THE SCORE THAT IS IN THE MIDDLE IS THE MEDIAN.
IF THE NUMBER OF VALUES IS EVEN, AS IN THE CASE OF THIS BOWLING GAME, THE AVERAGE OF THE TWO VALUES IN THE MIDDLE IS THE MEDIAN.
CAN YOU CALCULATE THE MEDIAN OF TEAM C?
ハイッ
IT'S (77 + 88) ÷ 2 = 82.5.
CORRECT!

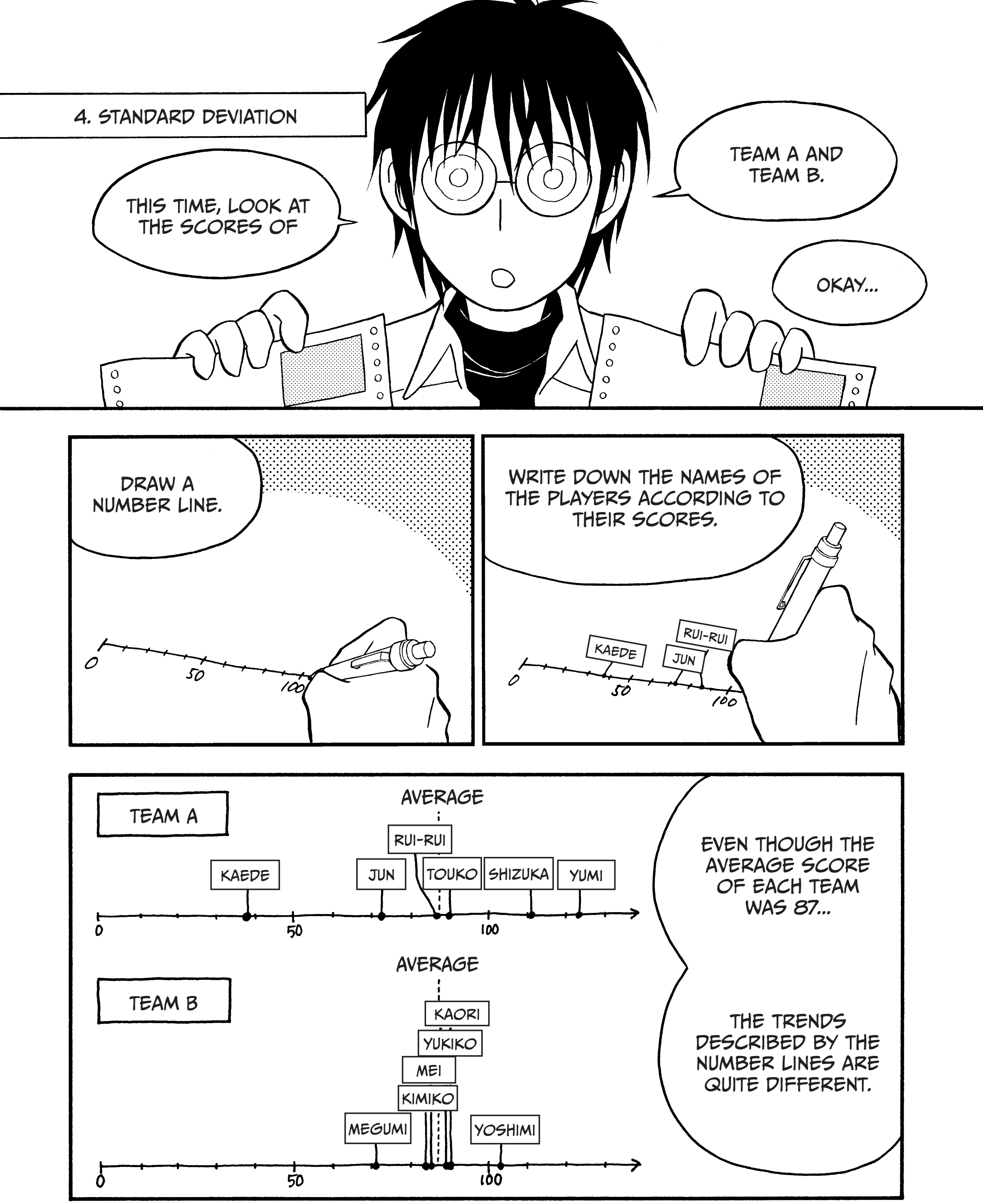


*LESS THAN $100

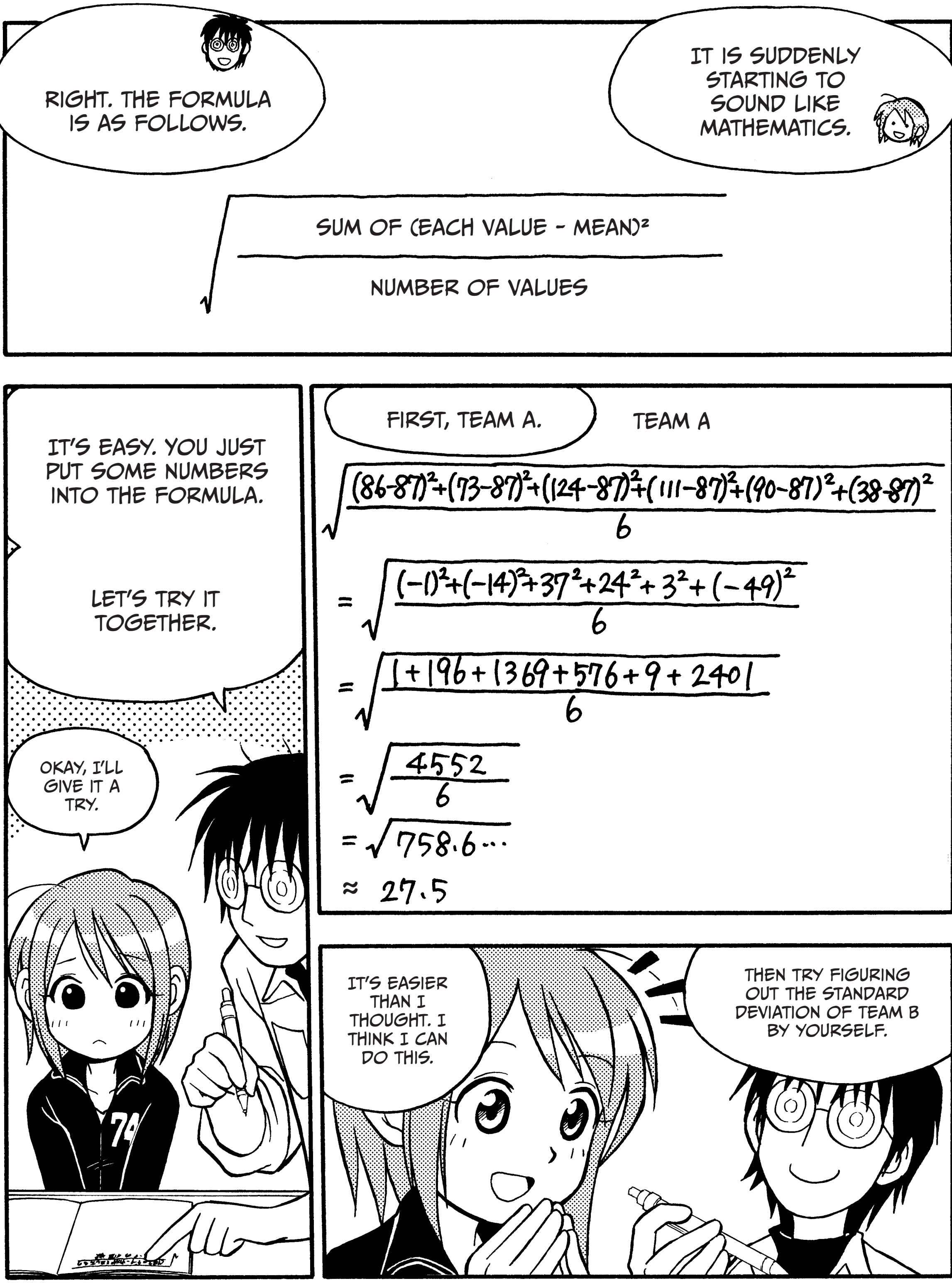


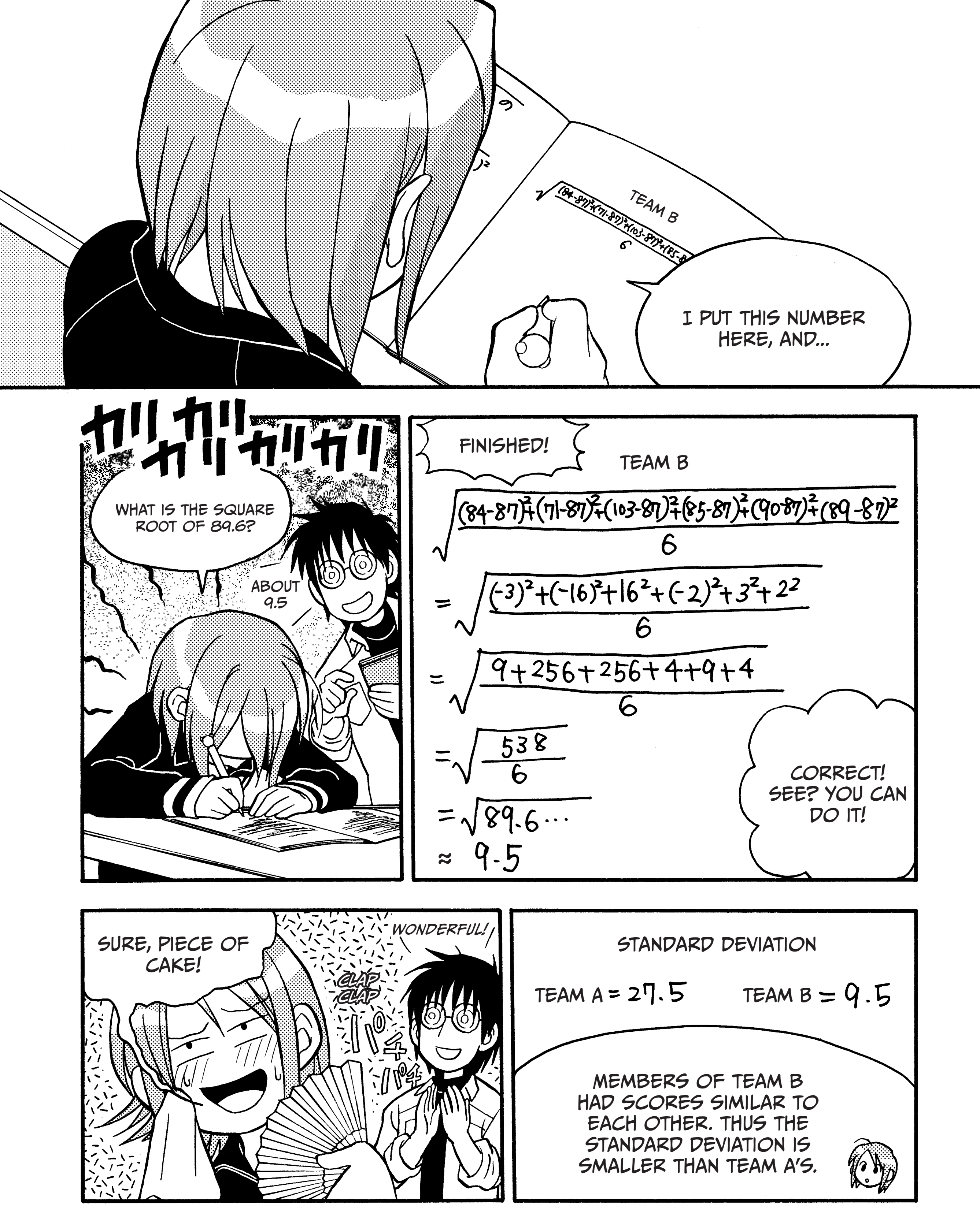


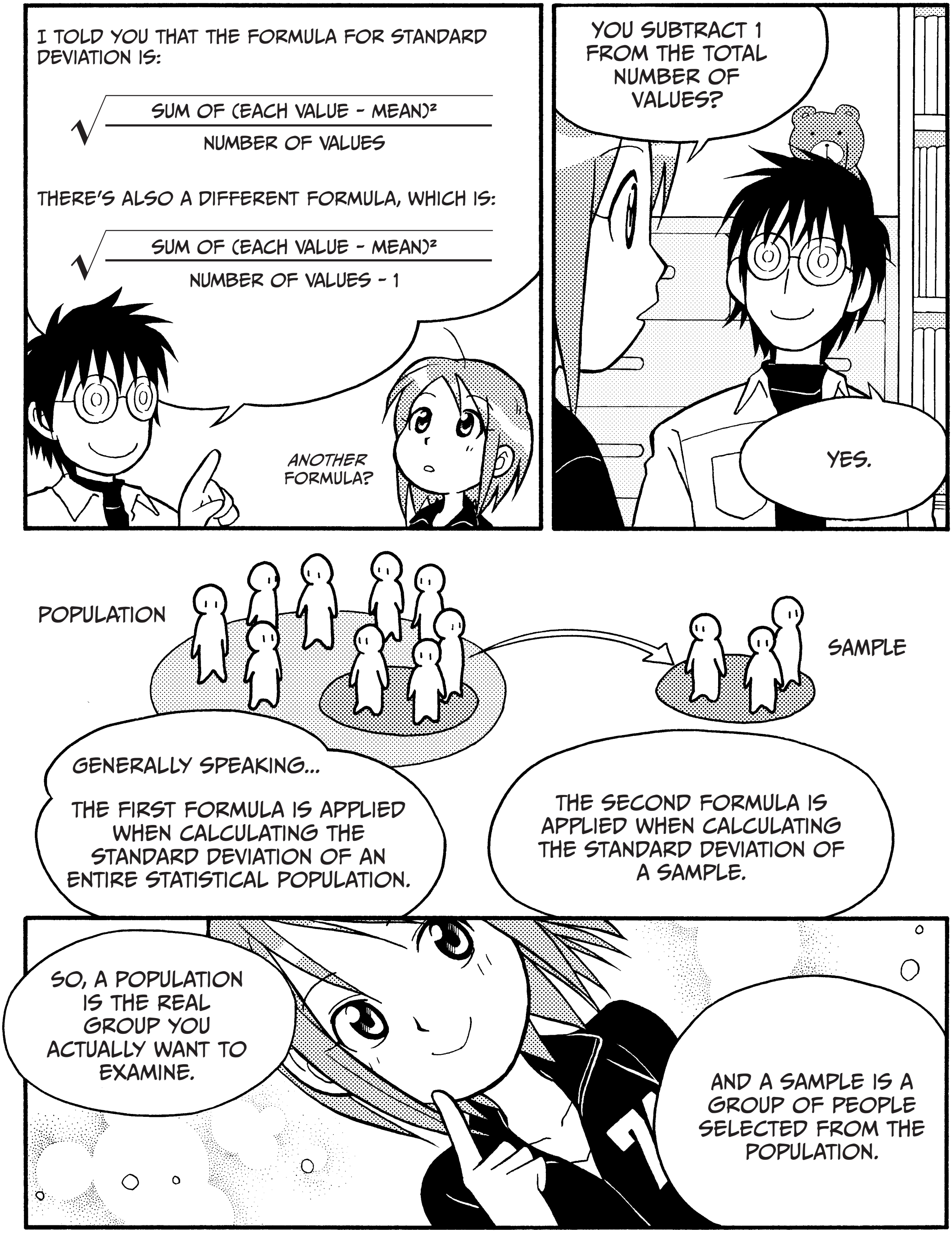

THE AVERAGE IS HIGH BECAUSE OF SOME MILLIONAIRES.
YOU NEED NOT BE DISAPPOINTED BECAUSE YOUR SAVINGS IS MUCH LESS THAN THE AVERAGE.
IN SUCH CASES, THE MEDIAN IS MUCH CLOSER TO COMMON PEOPLE.
ARE YOU LISTENING?
MILLIONAIRES...
I MUST MARRY A RICH GUY WHOSE SAVINGS IS WAY HIGHER THAN THE MEDIAN!
YOU DAMPEN MY SPIRITS...
74


4. STANDARD DEVIATION
THIS TIME, LOOK AT THE SCORES OF
TEAM A AND TEAM B.
OKAY...
DRAW A NUMBER LINE.
0
50
100
WRITE DOWN THE NAMES OF THE PLAYERS ACCORDING TO THEIR SCORES.
KAEDE
JUN
RUI-RUI
0
50
100
TEAM A
AVERAGE
RUI-RUI
KAEDE
JUN
TOUKO
SHIZUKA
YUMI
0
50
100
TEAM B
AVERAGE
KAORI
YUKIKO
MEI
KIMIKO
MEGUMI
YOSHIMI
0
50
100
EVEN THOUGH THE AVERAGE SCORE OF EACH TEAM WAS 87...
THE TRENDS DESCRIBED BY THE NUMBER LINES ARE QUITE DIFFERENT.

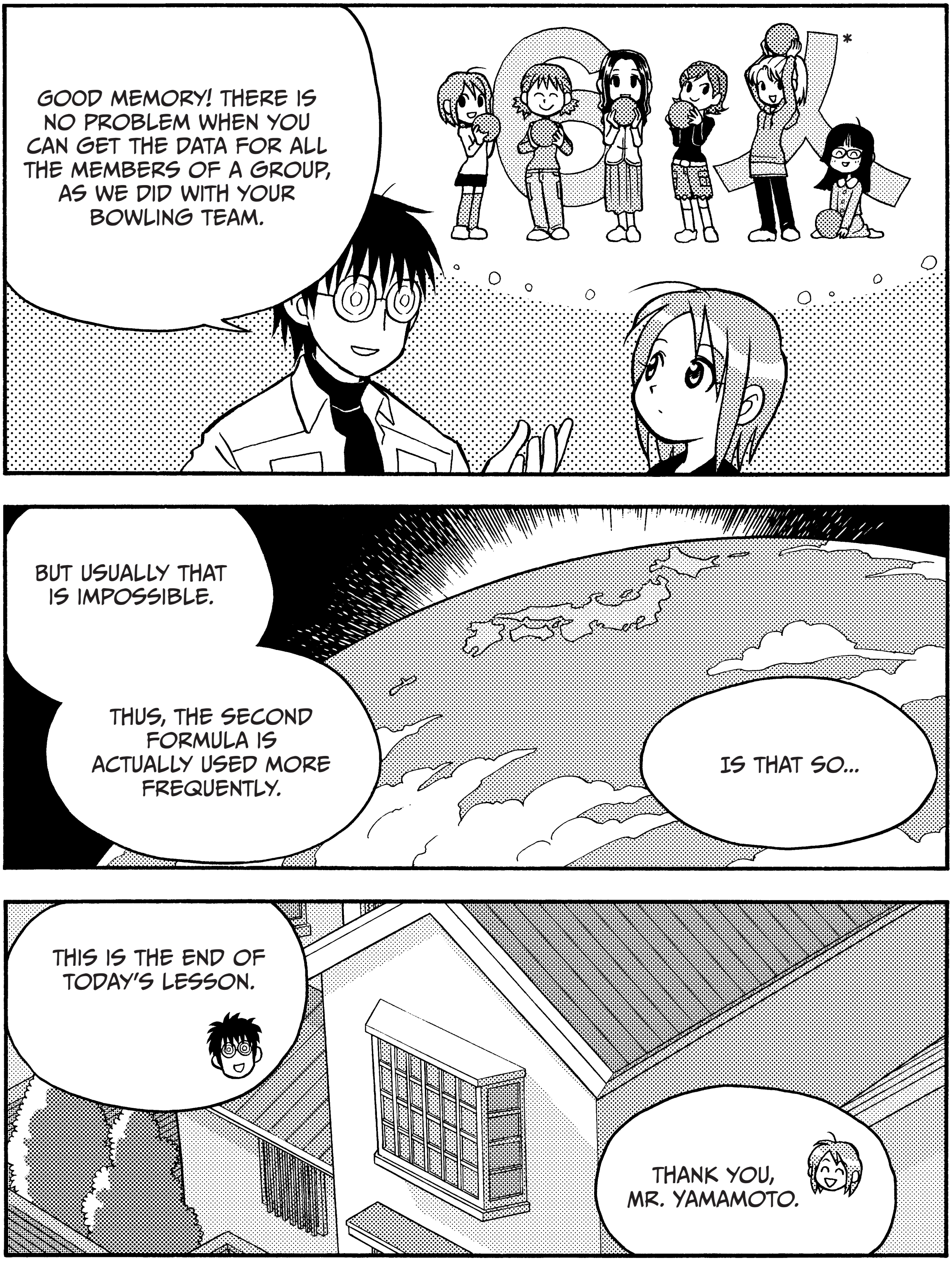

THEY SURE ARE. TEAM A'S SCORES VARY FROM LOW TO HIGH, BUT TEAM B'S SCORES ARE MORE SIMILAR.
STANDARD DEVIATION IS USED TO DESCRIBE SUCH SCATTERING OF DATA.
ANOTHER NEW TERM?!
IN SHORT, STANDARD DEVIATION IS *AN INDICATOR TO SHOW THE DIFFERENCE FROM THE MEAN OF* EACH VALUE IN THAT SET.
INDICATOR TO SHOW THE DIFFERENCE...?
THE MINIMUM STANDARD DEVIATION IS ZERO, AND AS THE "SCATTERING OF DATA" INCREASES, SO DOES THE STANDARD DEVIATION.
0 (MINIMUM)
NO SCATTERING (ALL VALUES ARE THE SAME)
SCATTERED
GUESS WHICH STANDARD DEVIATION IS LARGER, TEAM A'S OR TEAM B'S?
HMMM...
MAYBE TEAM A'S?


RIGHT. THE FORMULA IS AS FOLLOWS.
IT IS SUDDENLY STARTING TO SOUND LIKE MATHEMATICS.
$\sqrt{\dfrac{\text{SUM OF (EACH VALUE - MEAN)}^2}{\text{NUMBER OF VALUES}}}$
IT'S EASY. YOU JUST PUT SOME NUMBERS INTO THE FORMULA.
LET'S TRY IT TOGETHER.
OKAY, I'LL GIVE IT A TRY.
FIRST, TEAM A.
TEAM A
$\sqrt{\dfrac{(86-87)^2+(73-87)^2+(124-87)^2+(111-87)^2+(90-87)^2+(38-87)^2}{6}}$
$= \sqrt{\dfrac{(-1)^2+(-14)^2+37^2+24^2+3^2+(-49)^2}{6}}$
$= \sqrt{\dfrac{1+196+1369+576+9+2401}{6}}$
$= \sqrt{\dfrac{4552}{6}}$
$= \sqrt{758.6\cdots}$
≈ 27.5
IT'S EASIER THAN I THOUGHT. I THINK I CAN DO THIS.
THEN TRY FIGURING OUT THE STANDARD DEVIATION OF TEAM B BY YOURSELF.


TEAM B
I PUT THIS NUMBER HERE, AND...
カリカリ
カリカリカリ
WHAT IS THE SQUARE ROOT OF 89.6?
ABOUT 9.5
FINISHED!
TEAM B
$\sqrt{\frac{(84-87)^2+(71-87)^2+(103-87)^2+(85-87)^2+(90-87)^2+(89-87)^2}{6}}$
$=\sqrt{\frac{(-3)^2+(-16)^2+16^2+(-2)^2+3^2+2^2}{6}}$
$=\sqrt{\frac{9+256+256+4+9+4}{6}}$
$=\sqrt{\frac{538}{6}}$
$=\sqrt{89.6\cdots}$
≈ 9.5
CORRECT! SEE? YOU CAN DO IT!
SURE, PIECE OF CAKE!
WONDERFUL!
CLAP CLAP
パチ
パチ
STANDARD DEVIATION
TEAM A = 27.5
TEAM B = 9.5
MEMBERS OF TEAM B HAD SCORES SIMILAR TO EACH OTHER. THUS THE STANDARD DEVIATION IS SMALLER THAN TEAM A'S.


I TOLD YOU THAT THE FORMULA FOR STANDARD DEVIATION IS:
$\sqrt{\dfrac{\text{SUM OF (EACH VALUE - MEAN)}^2}{\text{NUMBER OF VALUES}}}$
THERE'S ALSO A DIFFERENT FORMULA, WHICH IS:
$\sqrt{\dfrac{\text{SUM OF (EACH VALUE - MEAN)}^2}{\text{NUMBER OF VALUES - 1}}}$
ANOTHER FORMULA?
YOU SUBTRACT 1 FROM THE TOTAL NUMBER OF VALUES?
YES.
POPULATION
SAMPLE
GENERALLY SPEAKING...
THE FIRST FORMULA IS APPLIED WHEN CALCULATING THE STANDARD DEVIATION OF AN ENTIRE STATISTICAL POPULATION.
THE SECOND FORMULA IS APPLIED WHEN CALCULATING THE STANDARD DEVIATION OF A SAMPLE.
SO, A POPULATION IS THE REAL GROUP YOU ACTUALLY WANT TO EXAMINE.
AND A SAMPLE IS A GROUP OF PEOPLE SELECTED FROM THE POPULATION.

*6 MEMBERS


GOOD MEMORY! THERE IS NO PROBLEM WHEN YOU CAN GET THE DATA FOR ALL THE MEMBERS OF A GROUP, AS WE DID WITH YOUR BOWLING TEAM.
BUT USUALLY THAT IS IMPOSSIBLE.
THUS, THE SECOND FORMULA IS ACTUALLY USED MORE FREQUENTLY.
IS THAT SO...
THIS IS THE END OF TODAY'S LESSON.
THANK YOU, MR. YAMAMOTO.

5. THE RANGE OF CLASS OF A FREQUENCY TABLE

If you felt that something was unclear in "Frequency Distribution Tables and Histograms" on page 32, take another look here at the table introduced on page 38.

TABLE 2-1: 50 BEST RAMEN SHOPS FREQUENCY TABLE

Class (equal or greater/less than)	Class midpoint	Frequency	Relative frequency
500–600	550	4	0.08
600–700	650	13	0.26
700–800	750	18	0.36
800–900	850	12	0.24
900–1000	950	3	0.06
Sum		50	1.00

As you can see, the range of class in this table is 100. The range was not determined according to any kind of mathematical standard—I set the range subjectively. Determining the range of class is up to the person who is analyzing the data.

But shouldn't there be a way to set the range of class mathematically? A frequency table may seem invalid if its range is determined subjectively.

There is a way to figure out the range of class mathematically. This is explained on the following pages. You'll also find a sample calculation using the data in Table 2-1.

Step 1

Calculate the number of classes using the Sturges' Rule below:

$$1 + \frac{\log_{10}(\text{number of values})}{\log_{10}2}$$

$$1 + \frac{\log_{10}50}{\log_{10}2} = 1 + 5.6438... = 6.6438... \approx 7$$

Step 2

Calculate the range of class using the formula below:

$$\frac{(\text{the maximum value}) - (\text{the minimum value})}{\text{the number of classes calculated from the Sturges' Rule}}$$

$$\frac{980 - 500}{7} = \frac{480}{7} = 68.5714... \approx 69$$

Below is a frequency chart organized according to the range of class as calculated by the formula in step 2.

Table 2-2: 50 Best Ramen Shops Frequency Table (Range of Class Determined Mathematically)

Class (equal or greater/less than)	Class midpoint	Frequency	Relative frequency
500–569	534.5	2	0.04
569–638	603.5	5	0.10
638–707	672.5	15	0.30
707–776	741.5	6	0.12
776–845	810.5	10	0.20
845–914	879.5	10	0.20
914–983	948.5	2	0.04
Sum		50	1.00

What do you think of this? Does this table seem even less convincing compared to Table 2-1? And why is the interval 69 yen?

If you try to explain to people that “this was calculated by a formula called the Sturges’ Rule,” they will only get mad and say, “Who cares about Stur . . . whatever! Why did you set the interval to a weird amount like 69 yen?”

To summarize, some people may hesitate to set the range of class subjectively. However, as the table above indicates, determining the range of class with the Sturges’ Rule does not necessarily provide a convincing table. A frequency table is, after all, a tool to help you visualize data. The analyst should set the range of class to any amount he or she thinks is appropriate.

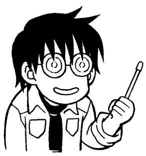
6. Estimation Theory and Descriptive Statistics

In the prologue, we explain that statistics can make an estimate about the situation of the population based on information collected from samples. To tell the truth, this explanation is not necessarily correct.

Statistics can be roughly classified into two categories: estimation theory and descriptive statistics. The one introduced in the prologue is the former. What, then, is descriptive statistics? It is a kind of a statistics that aims to describe the status of a group simply and clearly by organizing data. Descriptive statistics regards the group as the population.

Perhaps this explanation of descriptive statistics is abstract and difficult to understand. Here is an example to help clarify things. Remember when I figured out the mean and standard deviation of Rui's bowling team? This was not because I was trying to estimate the status of a population from the information collected from Rui's team. I calculated the mean and standard deviation purely because I wanted to describe the status of Rui's team simply. That kind of statistics is descriptive statistics.

Exercise and Answer



Exercise

The table below is a record of a high school girls' 100m track race.

Runner	**100m track race (seconds)**
Ms. A	16.3
Ms. B	22.4
Ms. C	18.5
Ms. D	18.7
Ms. E	20.1

1. What is the average?
2. What is the median?
3. What is the standard deviation?

ANSWER

1. The arithmetic mean is $\dfrac{16.3 + 22.4 + 18.5 + 18.7 + 20.1}{5} = \dfrac{96}{5} = 19.2$

2. The median is 18.7. 16.3 18.5 18.7 20.1 22.4

3. The standard deviation is

$$\sqrt{\frac{(16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{5}}$$

$$= \sqrt{\frac{(-2.9)^2 + 3.2^2 + (-0.7)^2 + (-0.5)^2 + 0.9^2}{5}}$$

$$= \sqrt{\frac{8.41 + 10.24 + 0.49 + 0.25 + 0.81}{5}}$$

$$= \sqrt{\frac{20.2}{5}}$$

$$= \sqrt{4.04}$$

$$\approx 2.01$$

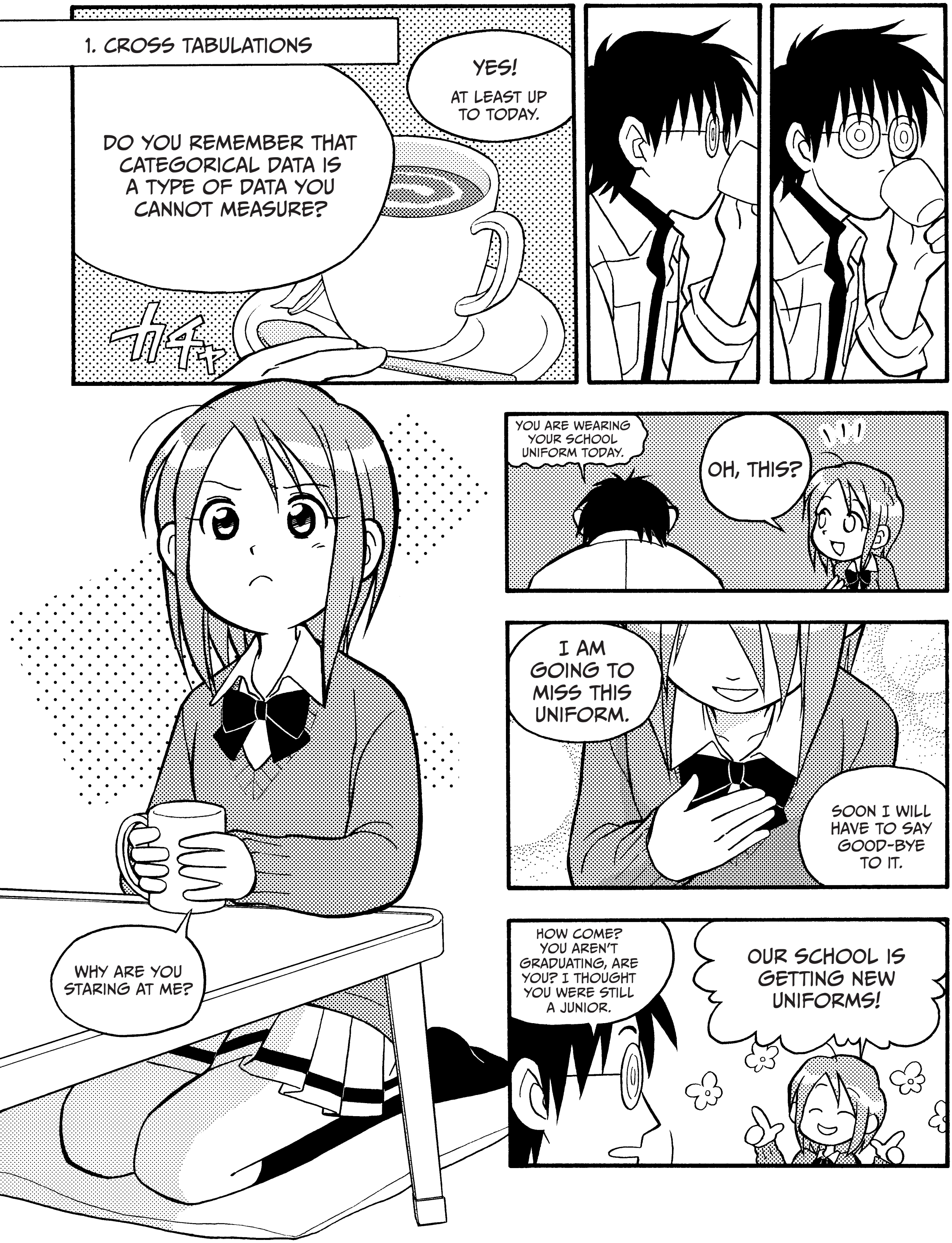
SUMMARY

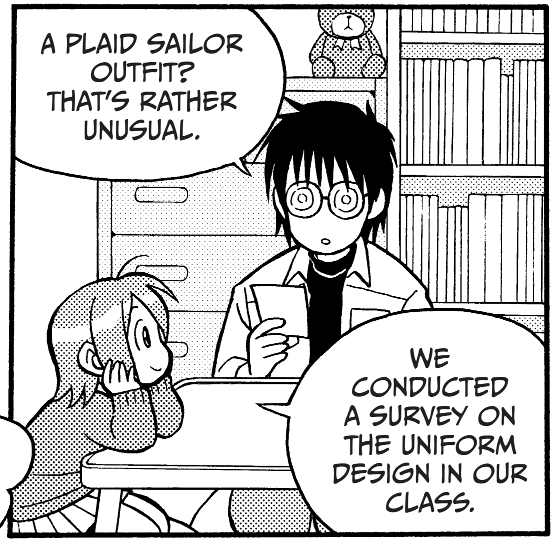
- To visualize the big picture of the data intuitively, create a frequency table or draw a histogram.
- When making a frequency table, the range of class may be determined by the Sturges' Rule.
- To visualize the data mathematically, calculate the arithmetic mean, median, and standard deviation.
- When there is an extremely large or small value in the data set, it is more appropriate to use the median than the arithmetic mean.
- *Standard deviation* is an index to describe "the size of scattering" of the data.

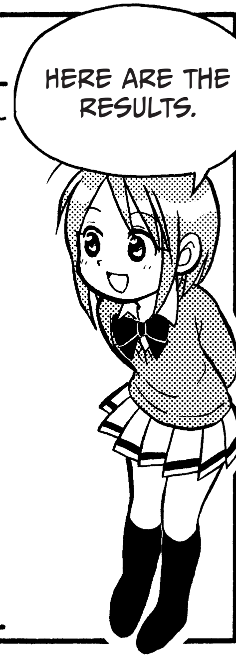
3

Getting the Big Picture: Understanding Categorical Data

1. CROSS TABULATIONS

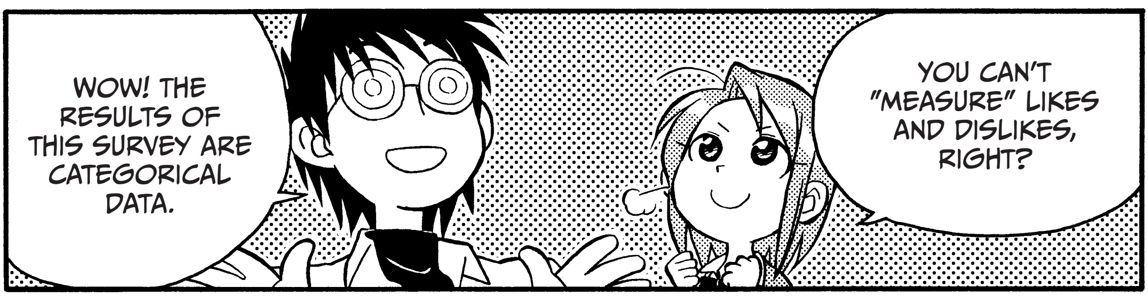


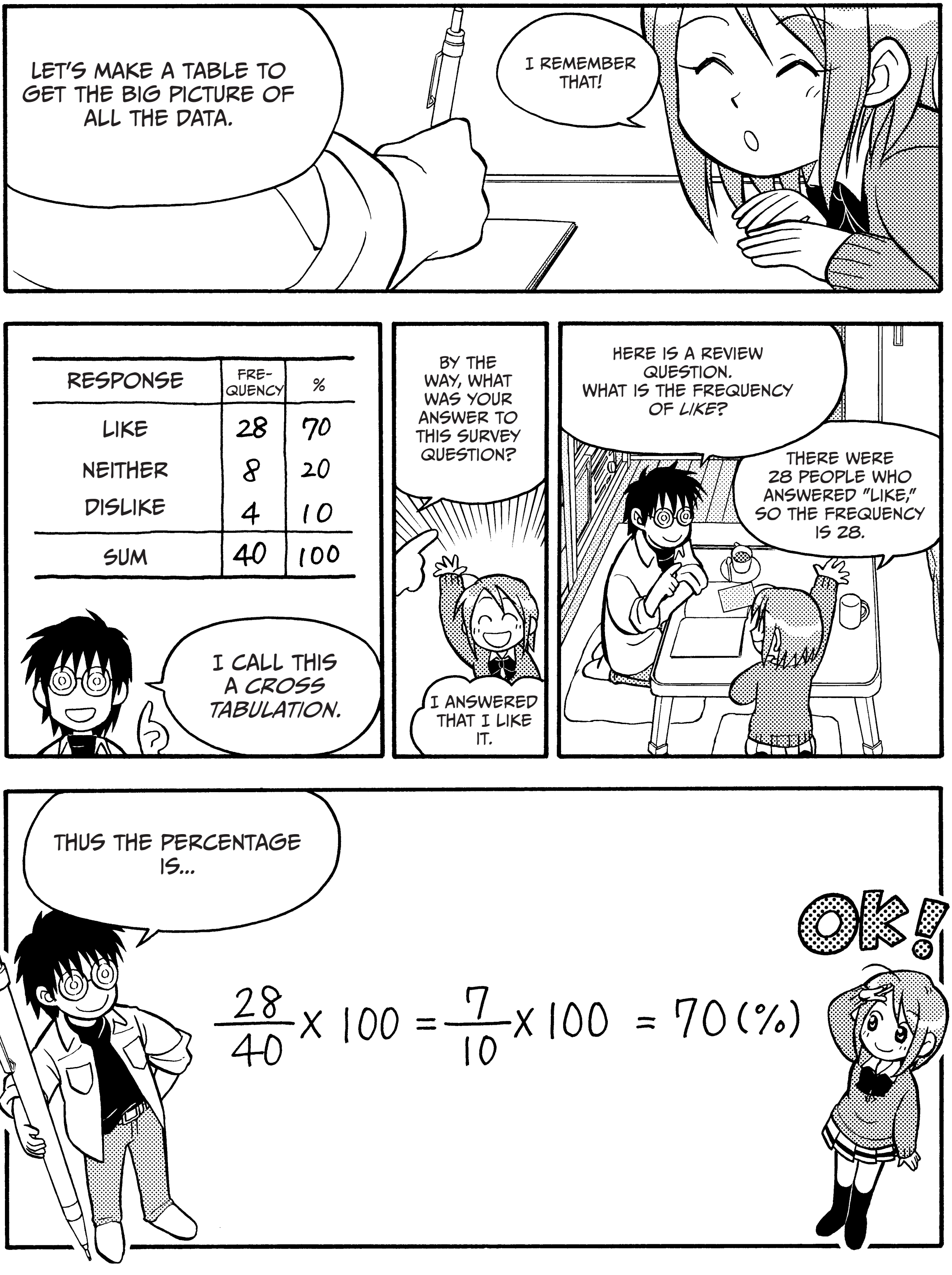




DO YOU LIKE OR DISLIKE THE NEW UNIFORM DESIGN?

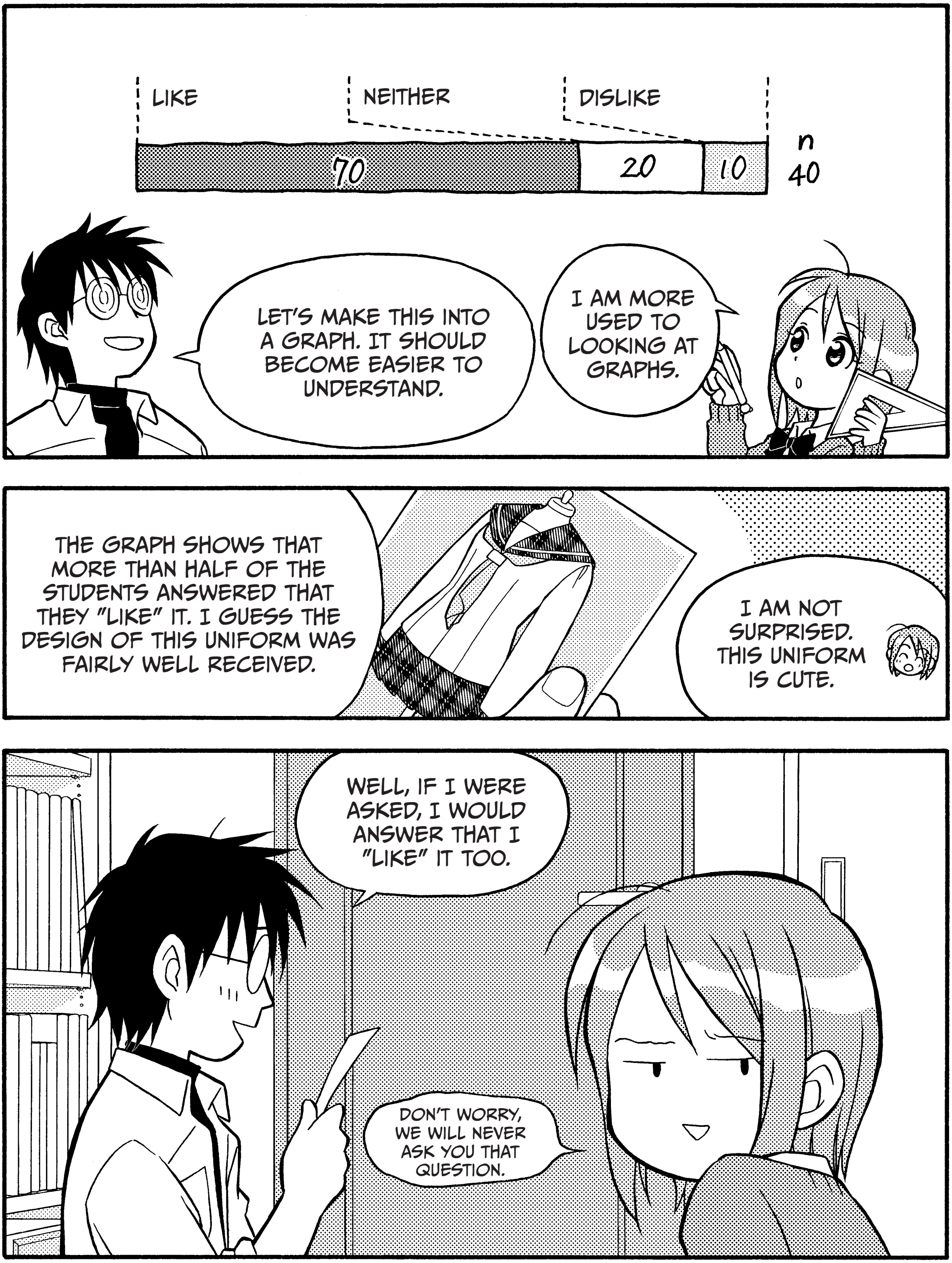
	RESPONSE		RESPONSE		RESPONSE
1	LIKE	16	NEITHER	31	NEITHER
2	NEITHER	17	LIKE	32	NEITHER
3	LIKE	18	LIKE	33	LIKE
4	NEITHER	19	LIKE	34	DISLIKE
5	DISLIKE	20	LIKE	35	LIKE
6	LIKE	21	LIKE	36	LIKE
7	LIKE	22	LIKE	37	LIKE
8	LIKE	23	DISLIKE	38	LIKE
9	LIKE	24	NEITHER	39	NEITHER
10	LIKE	25	LIKE	40	LIKE
11	LIKE	26	LIKE		
12	LIKE	27	DISLIKE		
13	NEITHER	28	LIKE		
14	LIKE	29	LIKE		
15	LIKE	30	LIKE		



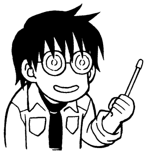


RESPONSE	FRE-QUENCY	%
LIKE	28	70
NEITHER	8	20
DISLIKE	4	10
SUM	40	100

$$\frac{28}{40} \times 100 = \frac{7}{10} \times 100 = 70(\%)$$


LIKE
NEITHER
DISLIKE
70
20
10
n
40
LET'S MAKE THIS INTO A GRAPH. IT SHOULD BECOME EASIER TO UNDERSTAND.
I AM MORE USED TO LOOKING AT GRAPHS.
THE GRAPH SHOWS THAT MORE THAN HALF OF THE STUDENTS ANSWERED THAT THEY "LIKE" IT. I GUESS THE DESIGN OF THIS UNIFORM WAS FAIRLY WELL RECEIVED.
I AM NOT SURPRISED. THIS UNIFORM IS CUTE.
WELL, IF I WERE ASKED, I WOULD ANSWER THAT I "LIKE" IT TOO.
DON'T WORRY, WE WILL NEVER ASK YOU THAT QUESTION.

Exercise and Answer



Exercise

A newspaper took a survey on political party A, which hopes to win the next election. The results are below.

Respondent	Do you expect party A to win or lose against party B?
1	Lose
2	Lose
3	Lose
4	I don't know
5	Win
6	Lose
7	Win
8	I don't know
9	Lose
10	Lose

Make a cross tabulation from these survey results.

Answer

Below is the cross tabulation.

Response	Frequency	%
Win	2	20
I don't know	2	20
Lose	6	60
Sum	10	100

Summary

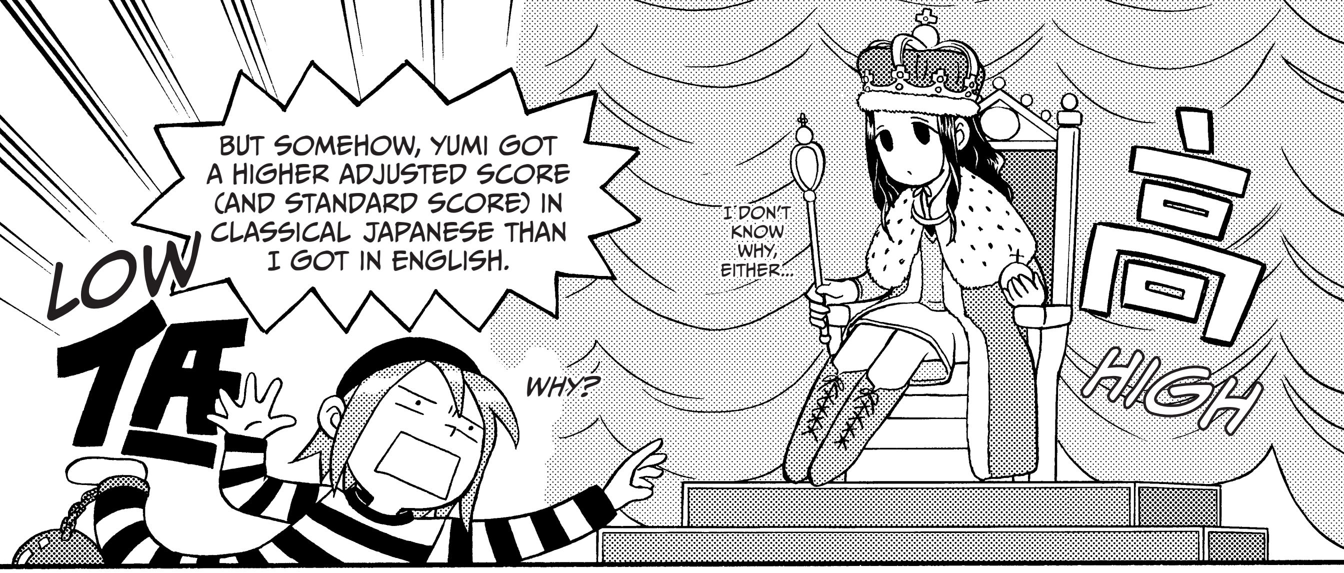
- One way to see the big picture of all the data is to make a cross tabulation.

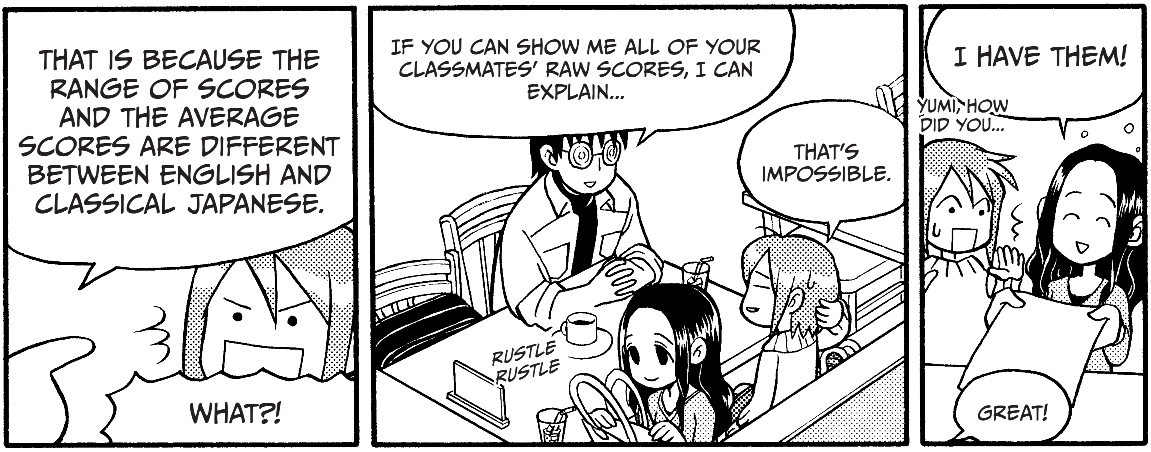
4

STANDARD SCORE AND DEVIATION SCORE

1. NORMALIZATION AND STANDARD SCORE

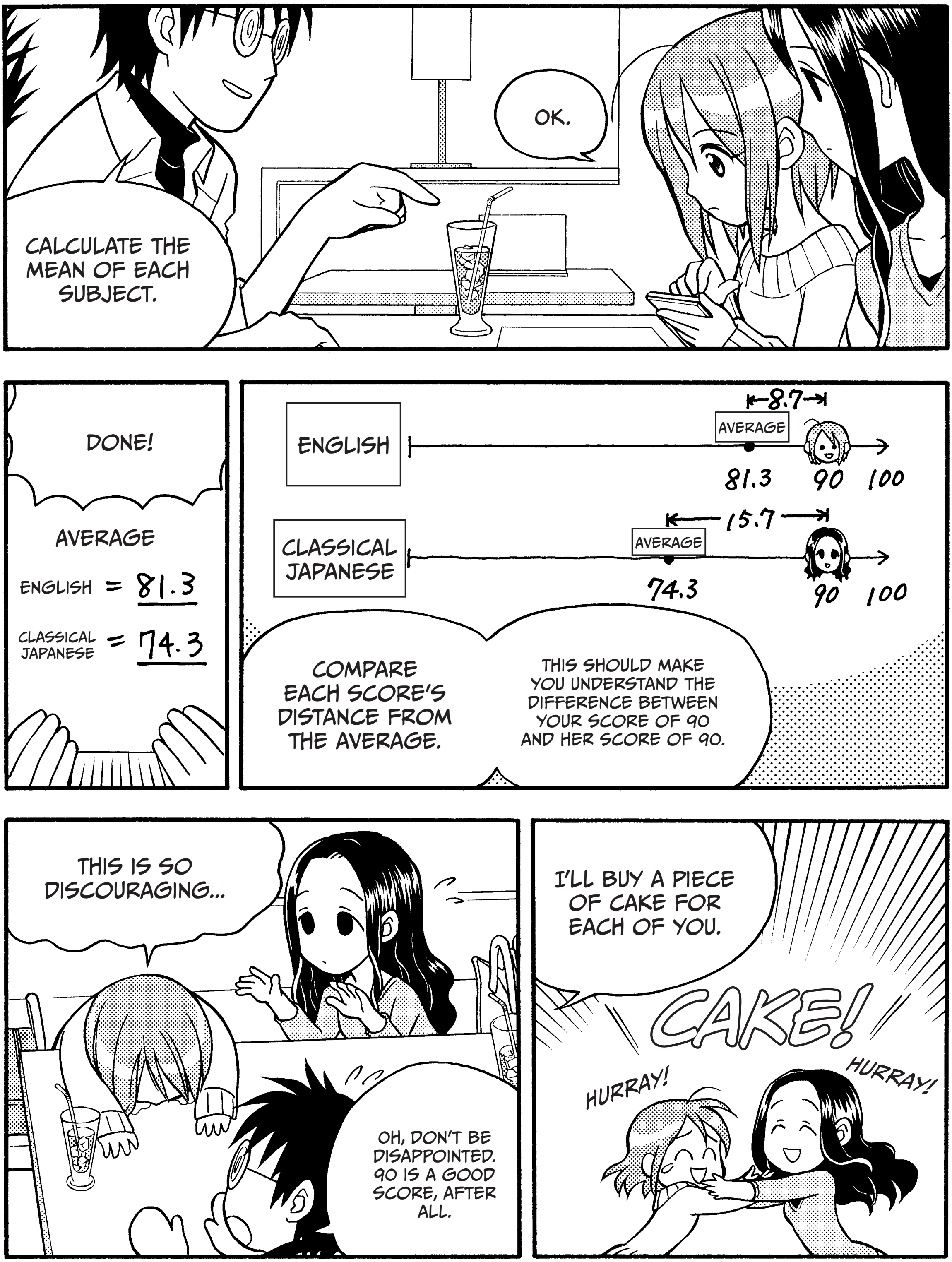
* ADJUSTING TEST RESULTS BASED ON STANDARD SCORE IS COMMONLY KNOWN AS *GRADING ON A CURVE.*

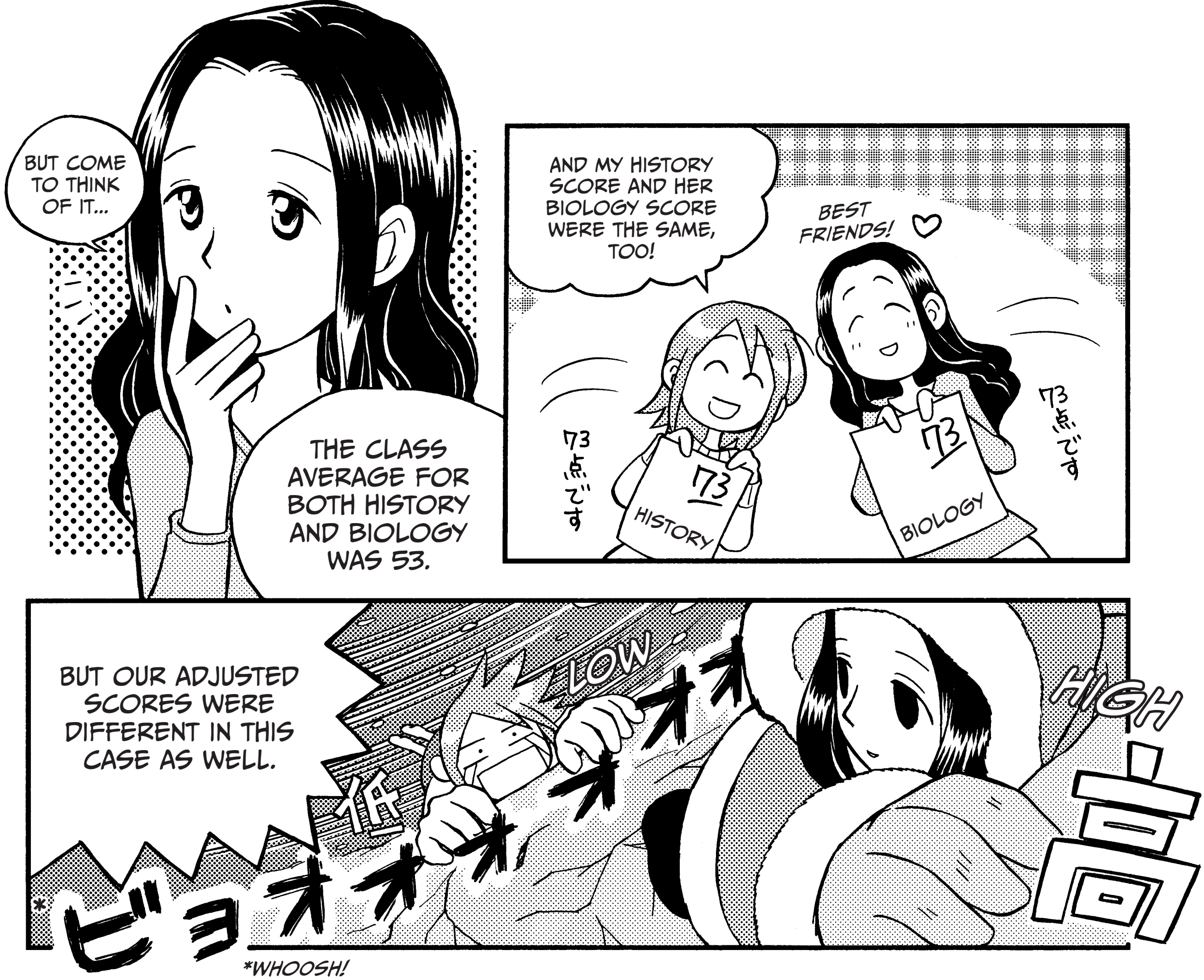




RAW TEST SCORES (OUT OF 100)

STUDENT	ENGLISH	CLASSICAL JAPANESE	STUDENT	ENGLISH	CLASSICAL JAPANESE
RUI	90	71	H	67	85
YUMI	81	90	I	87	93
A	73	79	J	78	89
B	97	70	K	85	78
C	85	67	L	96	74
D	60	66	M	77	65
E	74	60	N	100	78
F	64	83	O	92	53
G	72	57	P	86	80

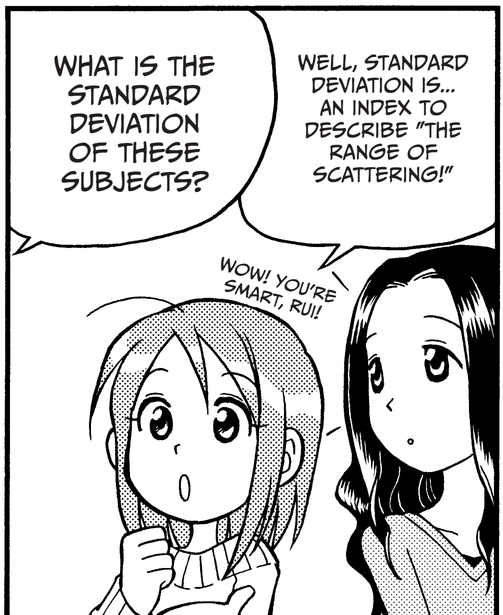

CALCULATE THE MEAN OF EACH SUBJECT.
OK.
DONE!
AVERAGE
ENGLISH = 81.3
CLASSICAL JAPANESE = 74.3
ENGLISH
AVERAGE
8.7
81.3
90
100
CLASSICAL JAPANESE
AVERAGE
15.7
74.3
90
100
COMPARE EACH SCORE'S DISTANCE FROM THE AVERAGE.
THIS SHOULD MAKE YOU UNDERSTAND THE DIFFERENCE BETWEEN YOUR SCORE OF 90 AND HER SCORE OF 90.
THIS IS SO DISCOURAGING...
OH, DON'T BE DISAPPOINTED. 90 IS A GOOD SCORE, AFTER ALL.
I'LL BUY A PIECE OF CAKE FOR EACH OF YOU.
CAKE!
HURRAY!
HURRAY!

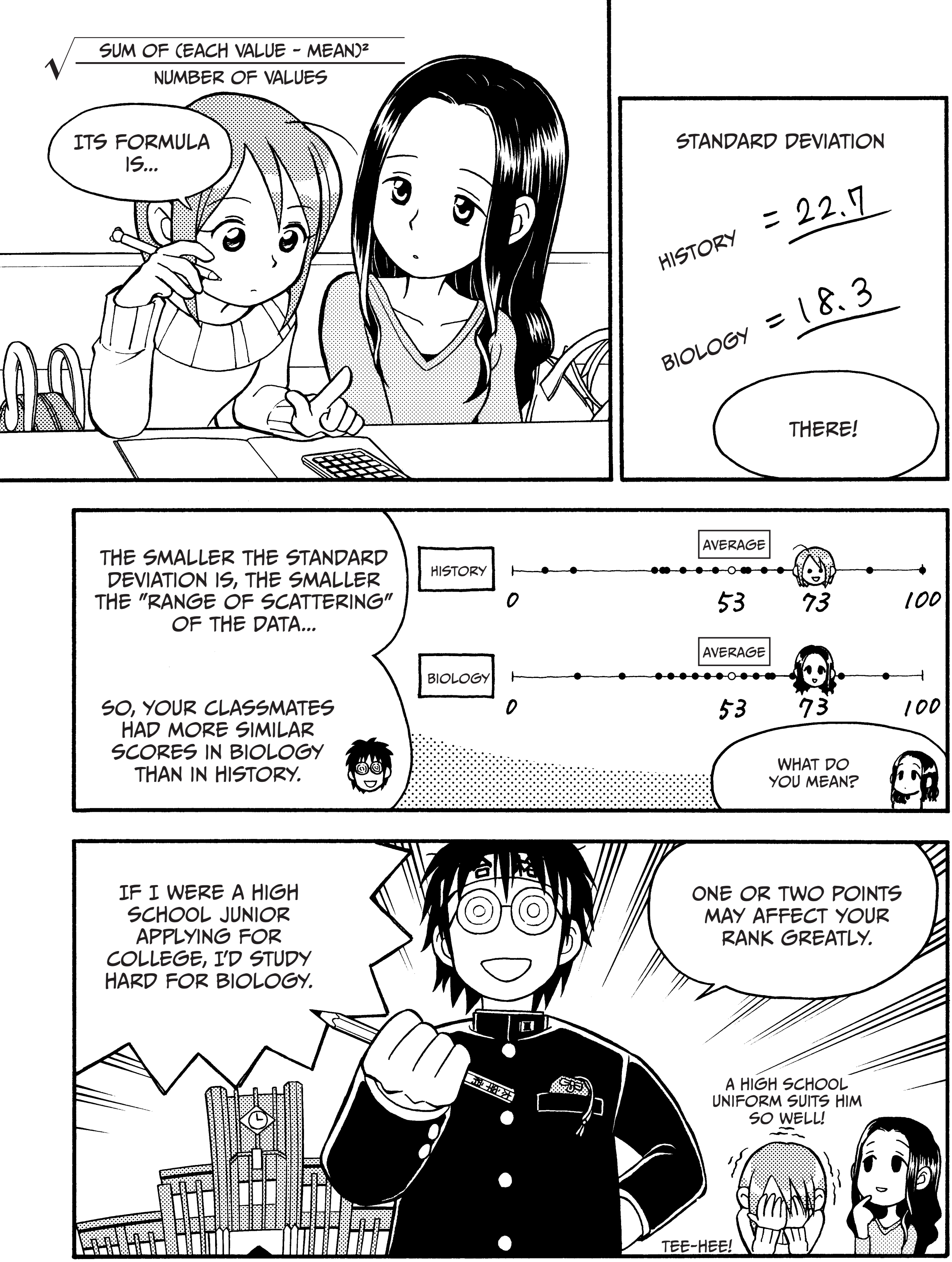


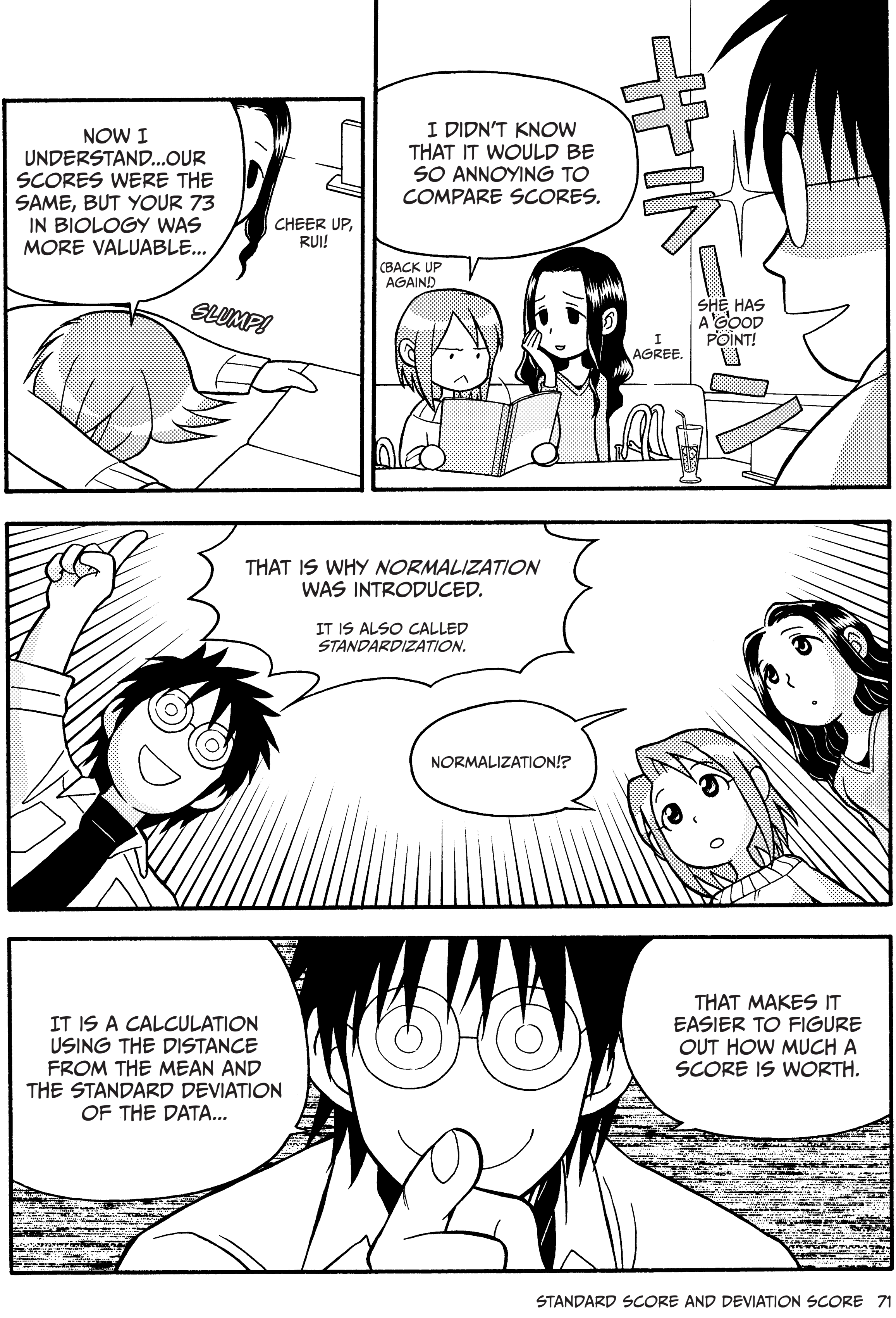
EVEN THOUGH THE DIFFERENCES BETWEEN OUR SCORES AND THE AVERAGES WERE THE SAME!

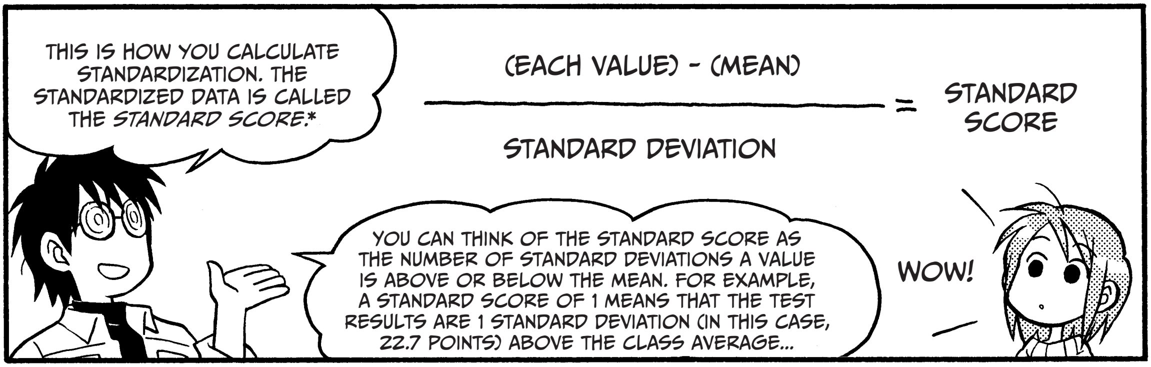
HMMM...

STUDENT	HISTORY	BIOLOGY	STUDENT	HISTORY	BIOLOGY
RUI	73	59	H	7	50
YUMI	61	73	I	53	41
A	14	47	J	100	62
B	41	38	K	57	44
C	49	63	L	45	26
D	87	56	M	56	91
E	69	15	N	34	35
F	65	53	O	37	53
G	36	80	P	70	68
			AVERAGE	53	53

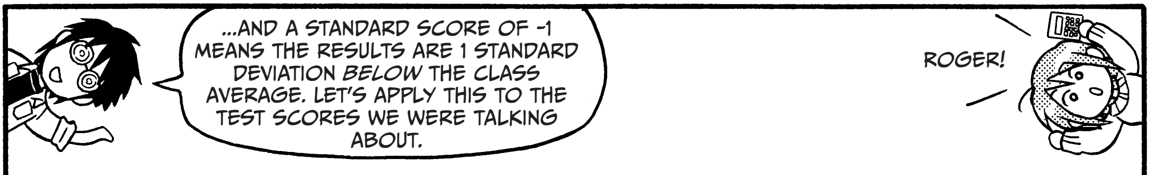



$\sqrt{\dfrac{\text{SUM OF (EACH VALUE - MEAN)}^2}{\text{NUMBER OF VALUES}}}$
ITS FORMULA IS...
STANDARD DEVIATION
HISTORY = 22.7
BIOLOGY = 18.3
THERE!
THE SMALLER THE STANDARD DEVIATION IS, THE SMALLER THE "RANGE OF SCATTERING" OF THE DATA...
HISTORY
AVERAGE
0
53
73
100
BIOLOGY
AVERAGE
0
53
73
100
SO, YOUR CLASSMATES HAD MORE SIMILAR SCORES IN BIOLOGY THAN IN HISTORY.
WHAT DO YOU MEAN?
IF I WERE A HIGH SCHOOL JUNIOR APPLYING FOR COLLEGE, I'D STUDY HARD FOR BIOLOGY.
ONE OR TWO POINTS MAY AFFECT YOUR RANK GREATLY.
A HIGH SCHOOL UNIFORM SUITS HIM SO WELL!
TEE-HEE!


NOW I UNDERSTAND...OUR SCORES WERE THE SAME, BUT YOUR 73 IN BIOLOGY WAS MORE VALUABLE...
CHEER UP, RUI!
SLUMP!
I DIDN'T KNOW THAT IT WOULD BE SO ANNOYING TO COMPARE SCORES.
(BACK UP AGAIN!)
I AGREE.
SHE HAS A GOOD POINT!
THAT IS WHY *NORMALIZATION* WAS INTRODUCED.
IT IS ALSO CALLED *STANDARDIZATION*.
NORMALIZATION!?
IT IS A CALCULATION USING THE DISTANCE FROM THE MEAN AND THE STANDARD DEVIATION OF THE DATA...
THAT MAKES IT EASIER TO FIGURE OUT HOW MUCH A SCORE IS WORTH.



* STANDARD SCORE IS ALSO CALLED Z-SCORE.

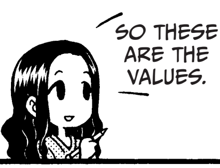


RESULTS AND STANDARD SCORES OF HISTORY AND BIOLOGY TESTS

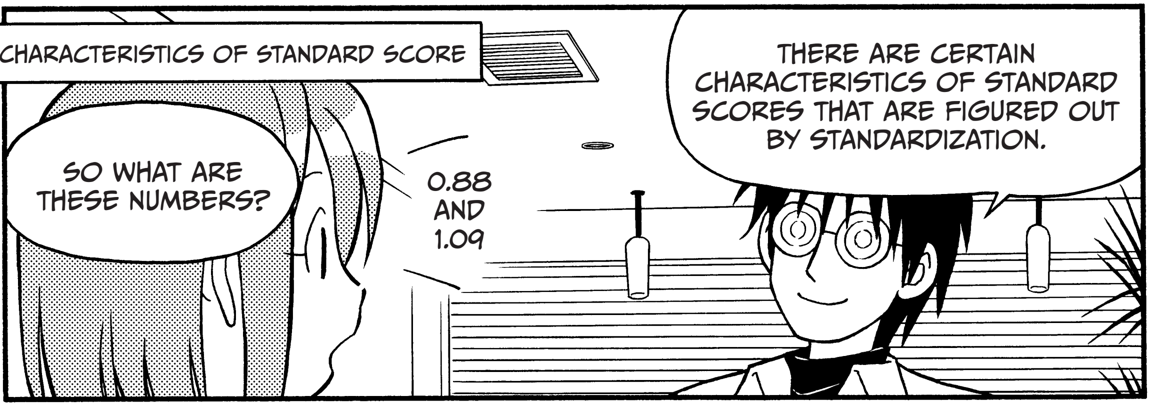
STUDENT	HISTORY	BIOLOGY	STANDARD SCORE OF HISTORY	STANDARD SCORE OF BIOLOGY
RUI	73	59	0.88	0.33
YUMI	61	73	0.35	1.09
A	14	47	-1.71	-0.33
B	41	38	-0.53	-0.82
C	49	63	-0.18	0.55
D	87	56	1.49	0.16
E	69	15	0.70	-2.08
F	65	53	0.53	0
G	36	80	-0.75	1.48
H	7	50	-2.02	-0.16
I	53	41	0	-0.66
J	100	62	2.07	0.49
K	57	44	0.18	-0.49
L	45	26	-0.35	-1.48
M	56	91	0.13	2.08
N	34	35	-0.84	-0.98
O	37	53	-0.70	0
P	70	68	0.75	0.82
AVERAGE	53	53	0	0
STANDARD DEVIATION	22.7	18.3	1	1

STANDARD SCORE OF RUI'S HISTORY TEST $\frac{73-53}{22.7} = \frac{20}{22.7} = 0.88$

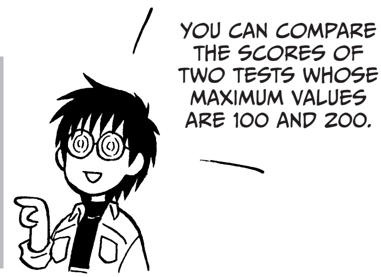
STANDARD SCORE OF YUMI'S BIOLOGY TEST $\frac{73-53}{18.3} = \frac{20}{18.3} = 1.09$



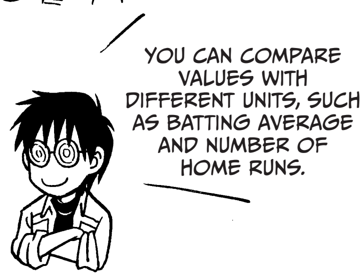
2. CHARACTERISTICS OF STANDARD SCORE

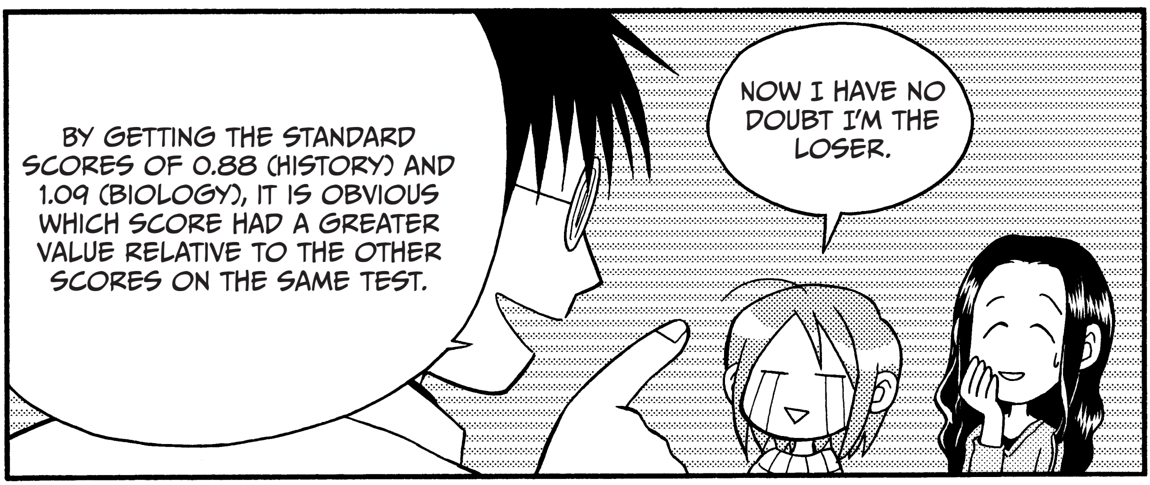


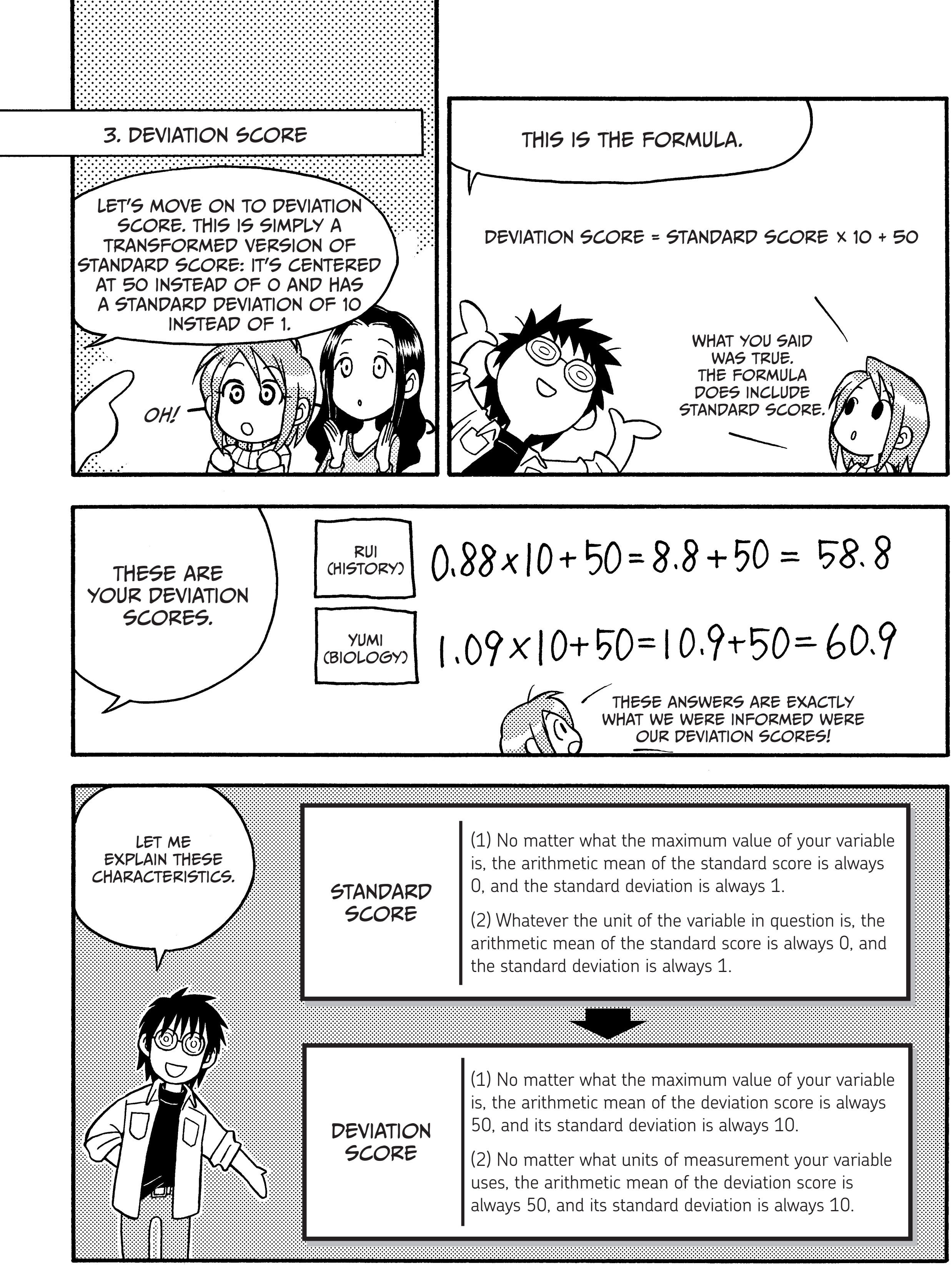
(1) No matter what the maximum value of your variable is, the arithmetic mean of the standard score is always 0, and the standard deviation is always 1.

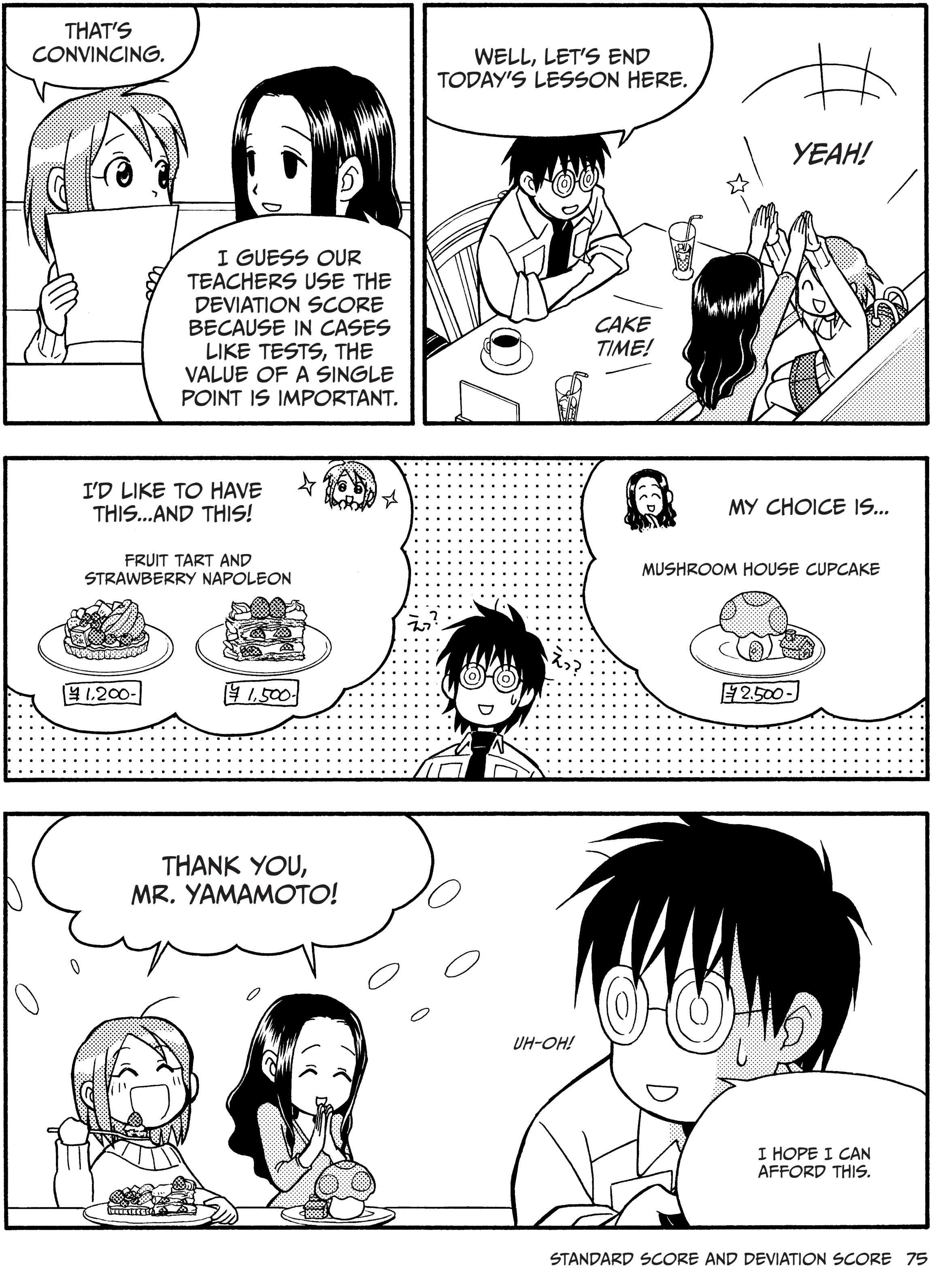


(2) Whatever the unit of the variable in question is, the arithmetic mean of the standard score is always 0, and the standard deviation is always 1.






3. DEVIATION SCORE
LET'S MOVE ON TO DEVIATION SCORE. THIS IS SIMPLY A TRANSFORMED VERSION OF STANDARD SCORE: IT'S CENTERED AT 50 INSTEAD OF 0 AND HAS A STANDARD DEVIATION OF 10 INSTEAD OF 1.
OH!
THIS IS THE FORMULA.
DEVIATION SCORE = STANDARD SCORE × 10 + 50
WHAT YOU SAID WAS TRUE. THE FORMULA DOES INCLUDE STANDARD SCORE.
THESE ARE YOUR DEVIATION SCORES.
RUI (HISTORY)
0.88 × 10 + 50 = 8.8 + 50 = 58.8
YUMI (BIOLOGY)
1.09 × 10 + 50 = 10.9 + 50 = 60.9
THESE ANSWERS ARE EXACTLY WHAT WE WERE INFORMED WERE OUR DEVIATION SCORES!
LET ME EXPLAIN THESE CHARACTERISTICS.
STANDARD SCORE
(1) No matter what the maximum value of your variable is, the arithmetic mean of the standard score is always 0, and the standard deviation is always 1.
(2) Whatever the unit of the variable in question is, the arithmetic mean of the standard score is always 0, and the standard deviation is always 1.
DEVIATION SCORE
(1) No matter what the maximum value of your variable is, the arithmetic mean of the deviation score is always 50, and its standard deviation is always 10.
(2) No matter what units of measurement your variable uses, the arithmetic mean of the deviation score is always 50, and its standard deviation is always 10.



THAT'S CONVINCING.
I GUESS OUR TEACHERS USE THE DEVIATION SCORE BECAUSE IN CASES LIKE TESTS, THE VALUE OF A SINGLE POINT IS IMPORTANT.
WELL, LET'S END TODAY'S LESSON HERE.
YEAH!
CAKE TIME!
I'D LIKE TO HAVE THIS...AND THIS!
FRUIT TART AND STRAWBERRY NAPOLEON
¥1,200-
¥1,500-
えっ?
えっ?
MY CHOICE IS...
MUSHROOM HOUSE CUPCAKE
¥2,500-
THANK YOU, MR. YAMAMOTO!
UH-OH!
I HOPE I CAN AFFORD THIS.

4. Interpretation of Deviation Score

Special caution is necessary when interpreting deviation scores. As explained on page 74, the definition of deviation score is:

$$\text{deviation score} = \text{standard score} \times 10 + 50 = \frac{(\text{each value} - \text{mean})}{\text{standard deviation}} \times 10 + 50$$

As mentioned on page 62, Rui's class has a total of 40 students, and as mentioned on page 40, there are 18 girls in the class. The example of deviation score on page 69 is not for the whole class, but is for the girls only. If the story were about the whole class, the mean and standard deviation would have been different from those for the girls only. Naturally, the deviation scores for Rui and Yumi would have been different as well. In fact, when everybody in the class is taken into consideration, Rui has the higher deviation score. Table 4-1 shows the test results for the whole class. Try calculating the deviation score.

To tell you the answer in advance, the deviation score for Rui's history test is 59.1, and that of Yumi's biology test is 56.7.

Suppose the same test is given to students in classes 1 and 2. The mean and standard deviation of class 1 are calculated individually, and deviation scores are obtained according to those amounts. Similarly, mean, standard deviation, and deviation scores for class 2 are obtained. Student A in class 1 has a deviation score of 57. Student B in class 2 has the same deviation score of 57. Outwardly, students A and B seem to have the same ability. However, the mean and standard deviation used to calculate these two deviation scores differ, because they come from two different classes. Unless the mean and standard deviation of the two classes are equal, you cannot compare the deviation scores of the two students.

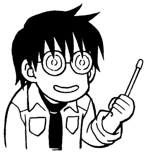
Here is another example. Suppose student A takes an entrance exam at a prep school in April and gets a deviation score of 54. After studying hard at a special summer course, student A takes an entrance exam at a different prep school in September. The deviation score is 62. It may seem that student A's proficiency has increased. However, the exam and the students taking it in April are different from the exam and the students taking it in September. Therefore, you cannot compare the deviation scores for these two exams, because the data used to calculate the mean and standard deviation of the April and September exams is different. In exam situations, you can only compare deviation scores for a group of students who all take the same exam. Keep these facts in mind when you interpret deviation scores.

TABLE 4-1: TEST RESULTS OF HISTORY AND BIOLOGY (ALL MEMBERS OF RUI'S CLASS)

Girls	History	Biology	Boys	History	Biology
Rui	73	59	a	54	2
Yumi	61	73	b	93	7
A	14	47	c	91	98
B	41	38	d	37	85
C	49	63	e	44	100
D	87	56	f	16	29
E	69	15	g	12	57
F	65	53	h	44	37
G	36	80	i	4	95
H	7	50	j	17	39
I	53	41	k	66	70
J	100	62	l	53	14
K	57	44	m	14	97
L	45	26	n	73	39
M	56	91	o	6	75
N	34	35	p	22	80
O	37	53	q	69	77
P	70	68	r	95	14
			s	16	24
			t	37	91
			u	14	36
			v	88	76

	History	Biology
Average of the whole class	48.0	54.9
Standard deviation of the whole class	27.5	26.9

EXERCISE AND ANSWER



EXERCISE

Below are the results of a high school girls' 100m track race.

Runner	100m track race (seconds)
Ms. A	16.3
Ms. B	22.4
Ms. C	18.5
Ms. D	18.7
Ms. E	20.1
Mean	19.2
Standard deviation	2.01

1. Demonstrate that the mean of the standard scores of the 100m track race is 0.
2. Demonstrate that the standard deviation of the standard score of the 100m track race is 1.

ANSWER

1. Mean of the standard score of the 100m track race

$$= \frac{\left(\frac{16.3 - 19.2}{2.01}\right) + \left(\frac{22.4 - 19.2}{2.01}\right) + \left(\frac{18.5 - 19.2}{2.01}\right) + \left(\frac{18.7 - 19.2}{2.01}\right) + \left(\frac{20.1 - 19.2}{2.01}\right)}{5}$$

$$= \frac{\left\{\frac{(16.3 - 19.2) + (22.4 - 19.2) + (18.5 - 19.2) + (18.7 - 19.2) + (20.1 - 19.2)}{2.01}\right\}}{5}$$

The numerator has been clarified.

$$= \frac{\left\{\frac{16.3 + 22.4 + 18.5 + 18.7 + 20.1 - 19.2 - 19.2 - 19.2 - 19.2 - 19.2}{2.01}\right\}}{5}$$

The numerator has been reorganized so that each value and (-19.2) are separate.

$$= \frac{\left\{\frac{96 - 19.2 \times 5}{2.01}\right\}}{5}$$

$$= \frac{\left\{\frac{96 - 96}{2.01}\right\}}{5}$$

$$= \frac{0}{5}$$

$$= 0$$

2. Standard deviation of the standard score of the 100m track race

$$= \sqrt{\frac{\left(\frac{16.3 - 19.2}{2.01} - 0\right)^2 + \left(\frac{22.4 - 19.2}{2.01} - 0\right)^2 + \left(\frac{18.5 - 19.2}{2.01} - 0\right)^2 + \left(\frac{18.7 - 19.2}{2.01} - 0\right)^2 + \left(\frac{20.1 - 19.2}{2.01} - 0\right)^2}{5}}$$

$$= \sqrt{\frac{\left(\frac{16.3 - 19.2}{2.01}\right)^2 + \left(\frac{22.4 - 19.2}{2.01}\right)^2 + \left(\frac{18.5 - 19.2}{2.01}\right)^2 + \left(\frac{18.7 - 19.2}{2.01}\right)^2 + \left(\frac{20.1 - 19.2}{2.01}\right)^2}{5}}$$

$$= \sqrt{\frac{\left\{\frac{(16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{2.01^2}\right\}}{5}}$$

The numerator has been clarified.

$$= \sqrt{\frac{1}{2.01^2} \times \frac{(16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{5}}$$

The numerator has been clarified.

$$= \frac{1}{2.01} \times \sqrt{\frac{(16.3 - 19.2)^2 + (22.4 - 19.2)^2 + (18.5 - 19.2)^2 + (18.7 - 19.2)^2 + (20.1 - 19.2)^2}{5}}$$

$$= \frac{1}{\text{standard deviation of the 100m track race}} \times \text{standard deviation of the 100m track race}$$

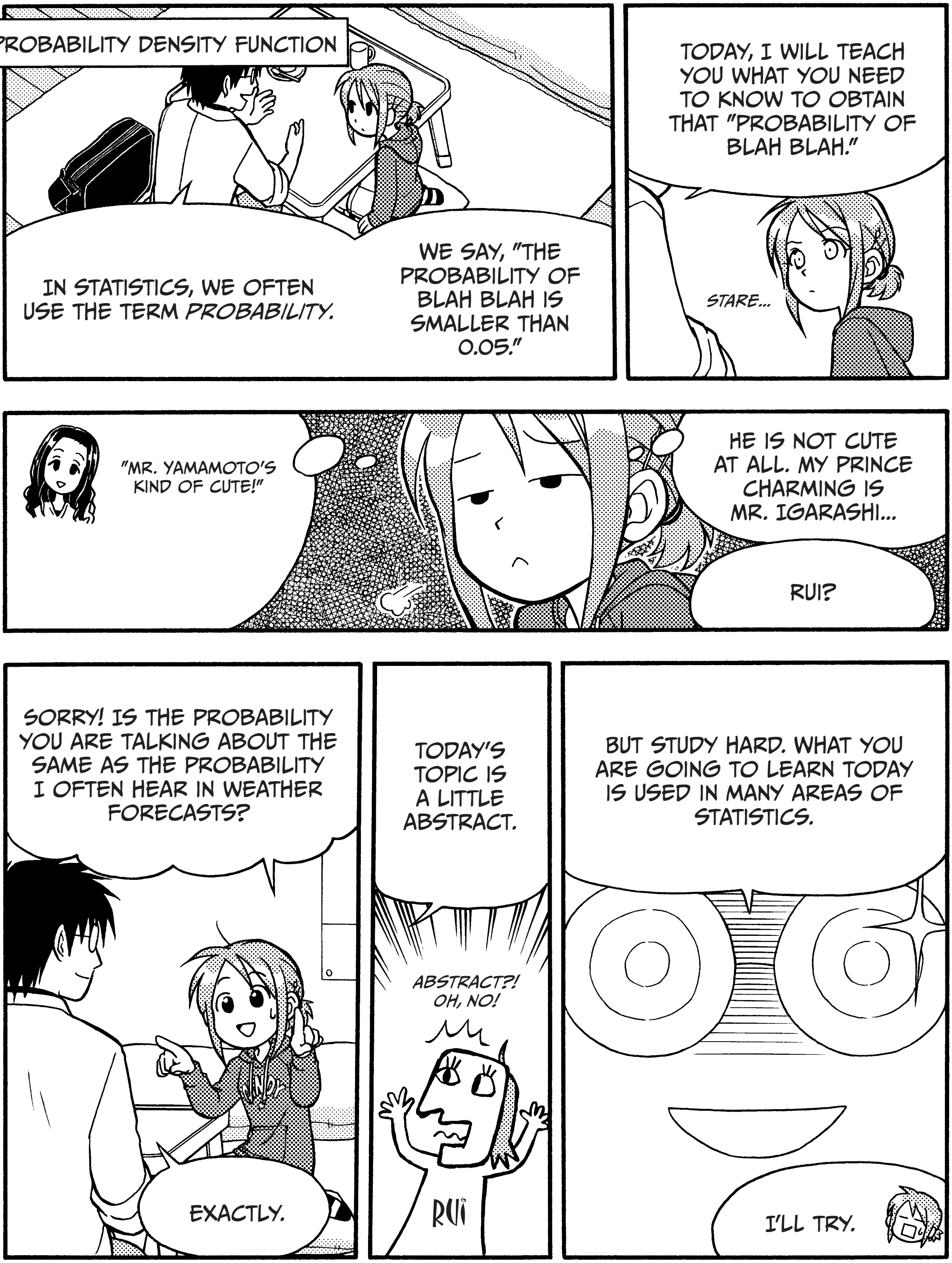
Carefully look at the table on page 78.

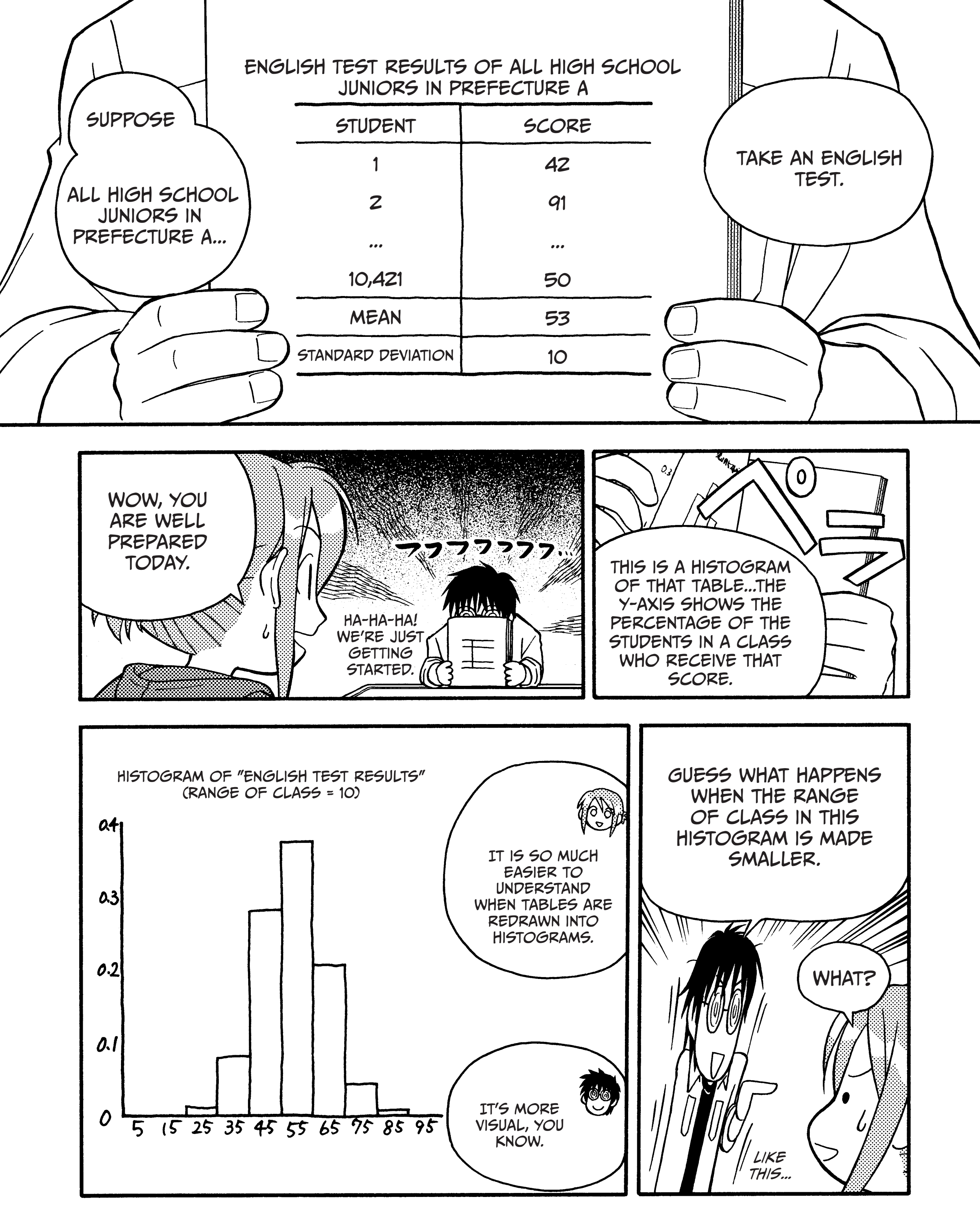
$$= 1$$

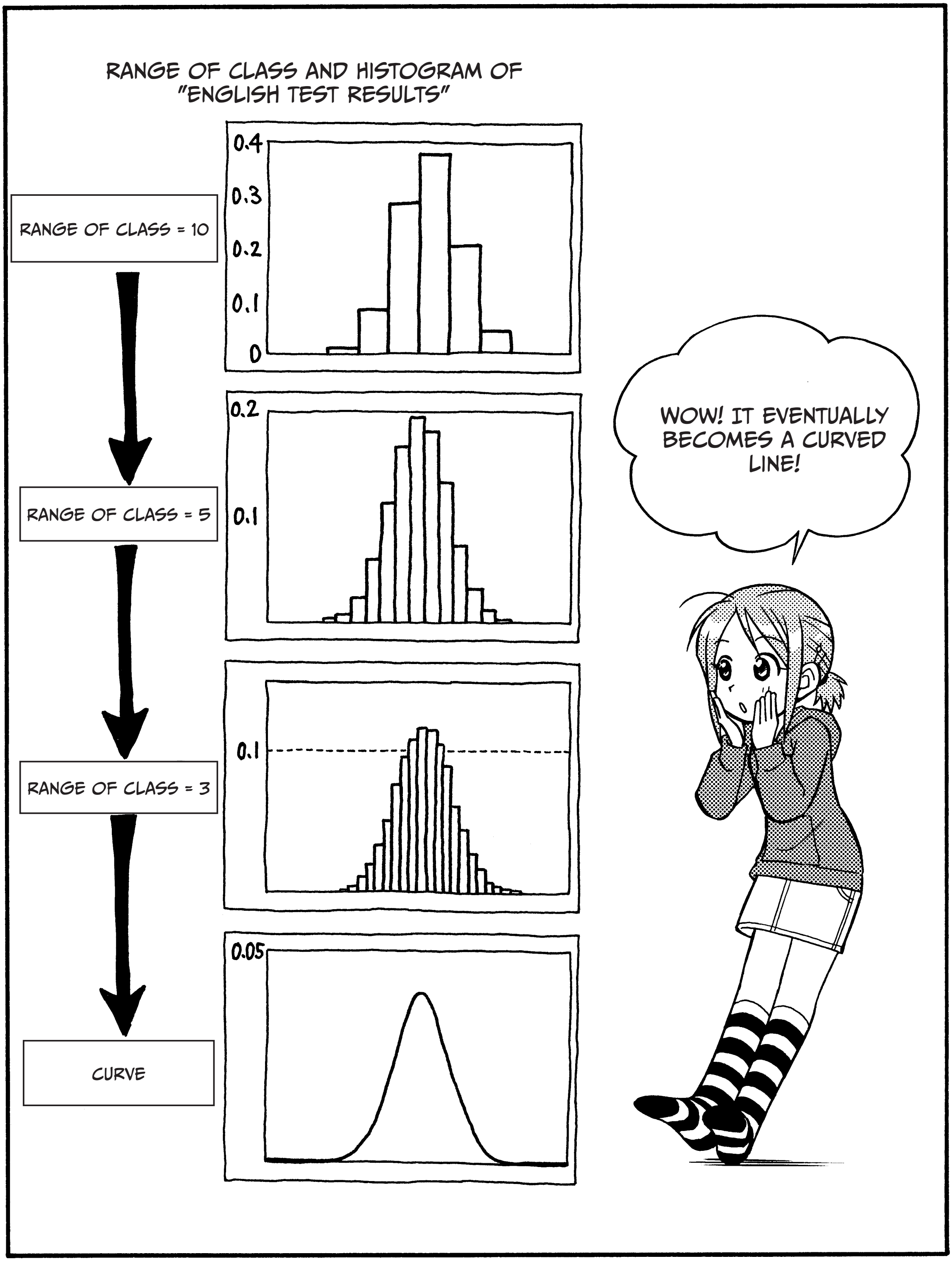
SUMMARY

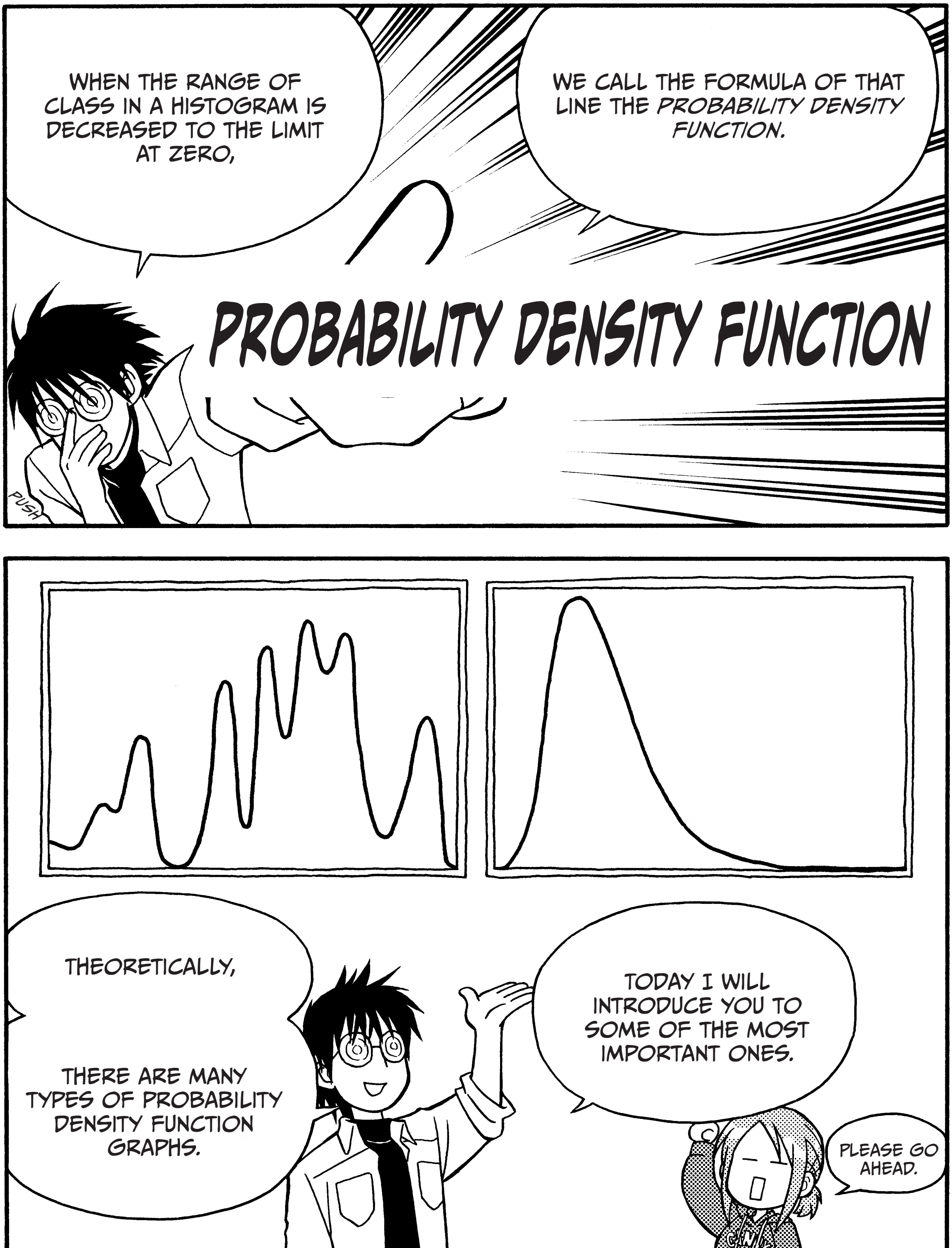
- *Standardization* helps you examine the value of a data point relative to the rest of your data by using its distance from the mean and "the size of scattering" of the data.
- Use standardization to:
    - Compare variables with different ranges
    - Compare variables that use different units of measurements
- A data point that has been standardized is called the *standard score* for that observation. Deviation score is an application of standard score.

5
LET'S OBTAIN THE PROBABILITY

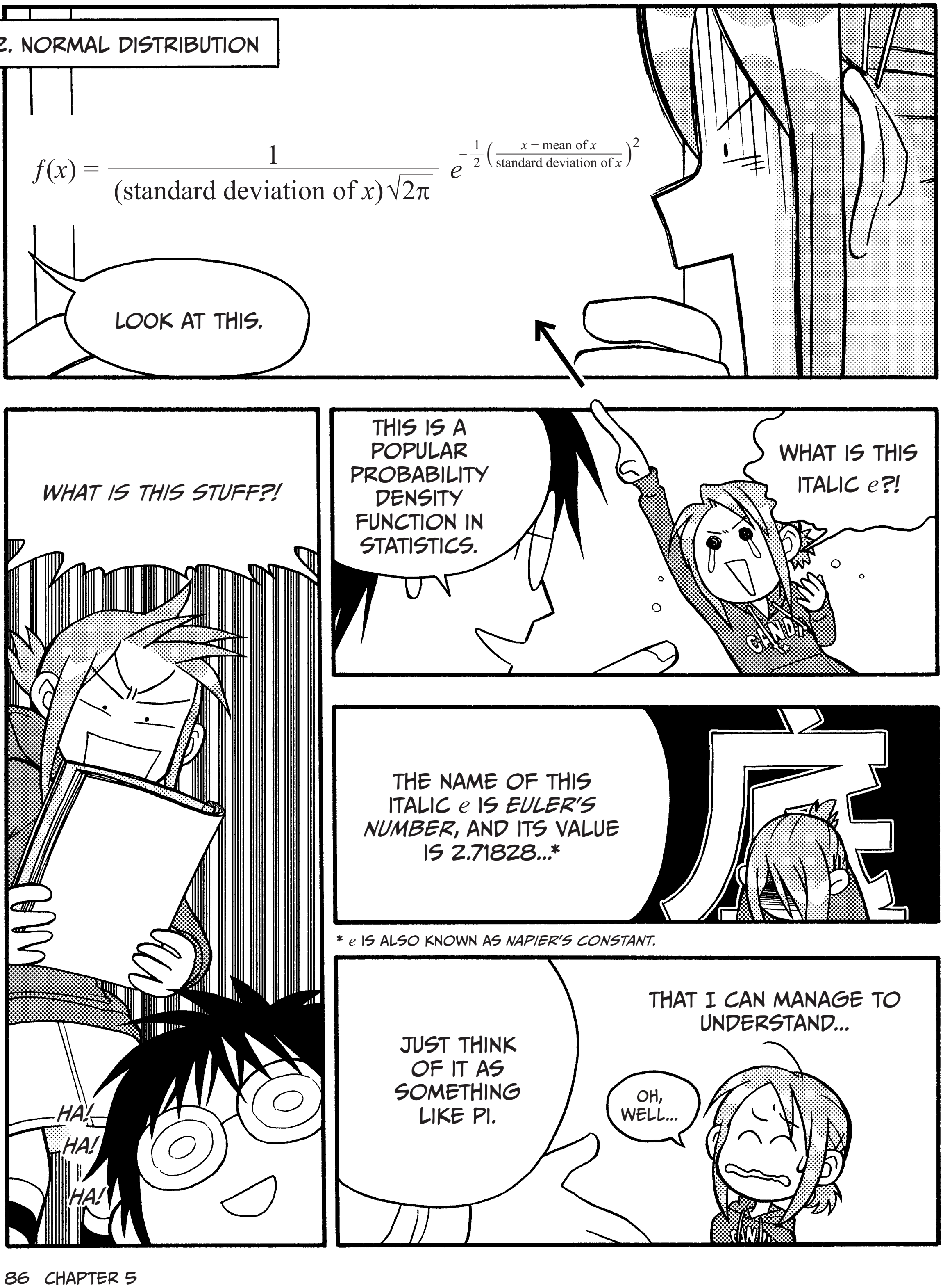

1. PROBABILITY DENSITY FUNCTION
IN STATISTICS, WE OFTEN USE THE TERM *PROBABILITY*.
WE SAY, "THE PROBABILITY OF BLAH BLAH IS SMALLER THAN 0.05."
TODAY, I WILL TEACH YOU WHAT YOU NEED TO KNOW TO OBTAIN THAT "PROBABILITY OF BLAH BLAH."
STARE...
"MR. YAMAMOTO'S KIND OF CUTE!"
HE IS NOT CUTE AT ALL. MY PRINCE CHARMING IS MR. IGARASHI...
RUI?
SORRY! IS THE PROBABILITY YOU ARE TALKING ABOUT THE SAME AS THE PROBABILITY I OFTEN HEAR IN WEATHER FORECASTS?
EXACTLY.
TODAY'S TOPIC IS A LITTLE ABSTRACT.
ABSTRACT?! OH, NO!
RUI
BUT STUDY HARD. WHAT YOU ARE GOING TO LEARN TODAY IS USED IN MANY AREAS OF STATISTICS.
I'LL TRY.


ENGLISH TEST RESULTS OF ALL HIGH SCHOOL JUNIORS IN PREFECTURE A
STUDENT | SCORE
1 | 42
2 | 91
... | ...
10,421 | 50
MEAN | 53
STANDARD DEVIATION | 10
SUPPOSE
ALL HIGH SCHOOL JUNIORS IN PREFECTURE A...
TAKE AN ENGLISH TEST.
WOW, YOU ARE WELL PREPARED TODAY.
HA-HA-HA! WE'RE JUST GETTING STARTED.
THIS IS A HISTOGRAM OF THAT TABLE...THE Y-AXIS SHOWS THE PERCENTAGE OF THE STUDENTS IN A CLASS WHO RECEIVE THAT SCORE.
HISTOGRAM OF "ENGLISH TEST RESULTS" (RANGE OF CLASS = 10)
0.4
0.3
0.2
0.1
0
5 15 25 35 45 55 65 75 85 95
IT IS SO MUCH EASIER TO UNDERSTAND WHEN TABLES ARE REDRAWN INTO HISTOGRAMS.
IT'S MORE VISUAL, YOU KNOW.
GUESS WHAT HAPPENS WHEN THE RANGE OF CLASS IN THIS HISTOGRAM IS MADE SMALLER.
WHAT?
LIKE THIS...

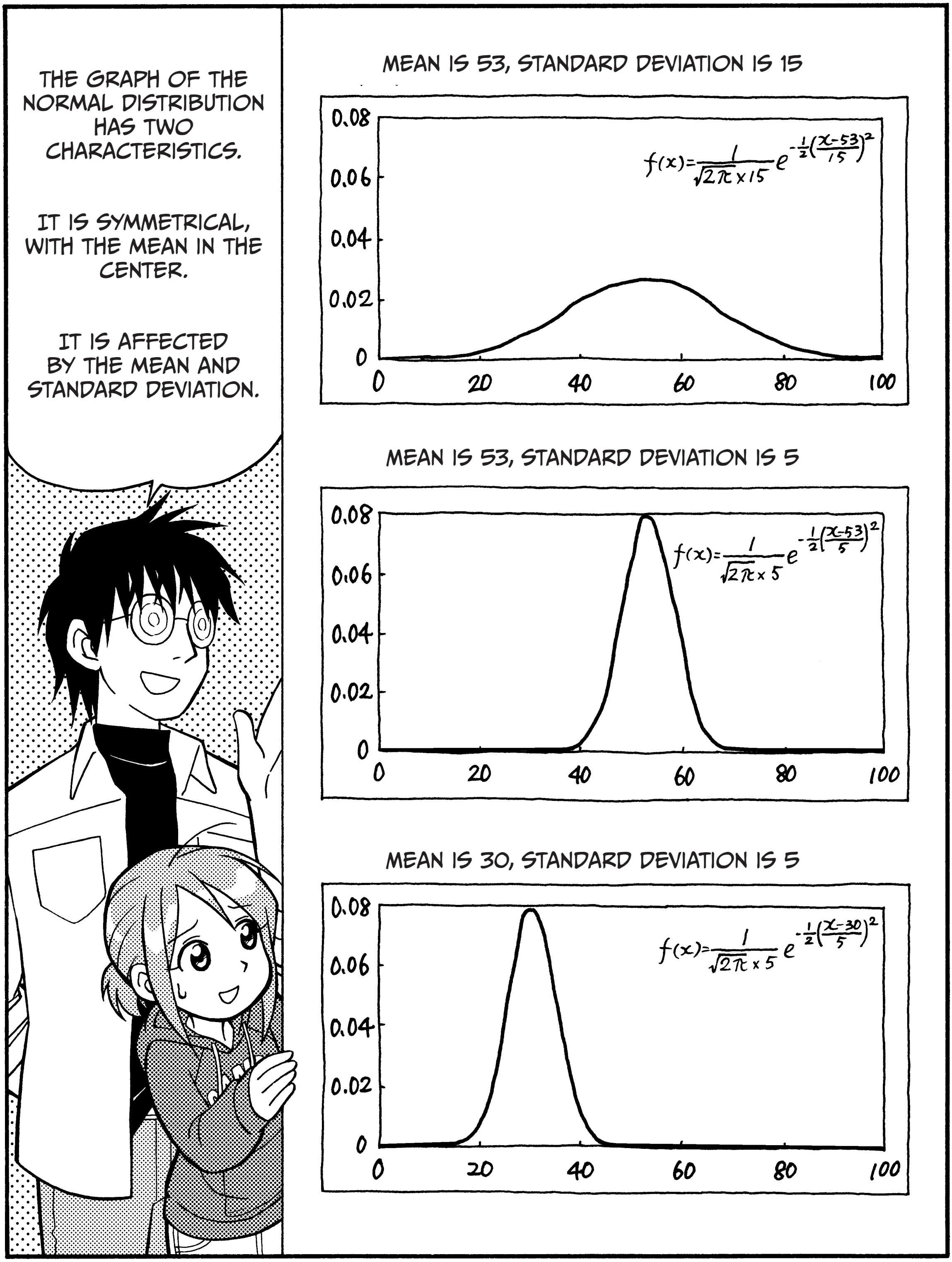

RANGE OF CLASS AND HISTOGRAM OF
"ENGLISH TEST RESULTS"
RANGE OF CLASS = 10
0.4
0.3
0.2
0.1
0
RANGE OF CLASS = 5
0.2
0.1
RANGE OF CLASS = 3
0.1
CURVE
0.05
WOW! IT EVENTUALLY BECOMES A CURVED LINE!

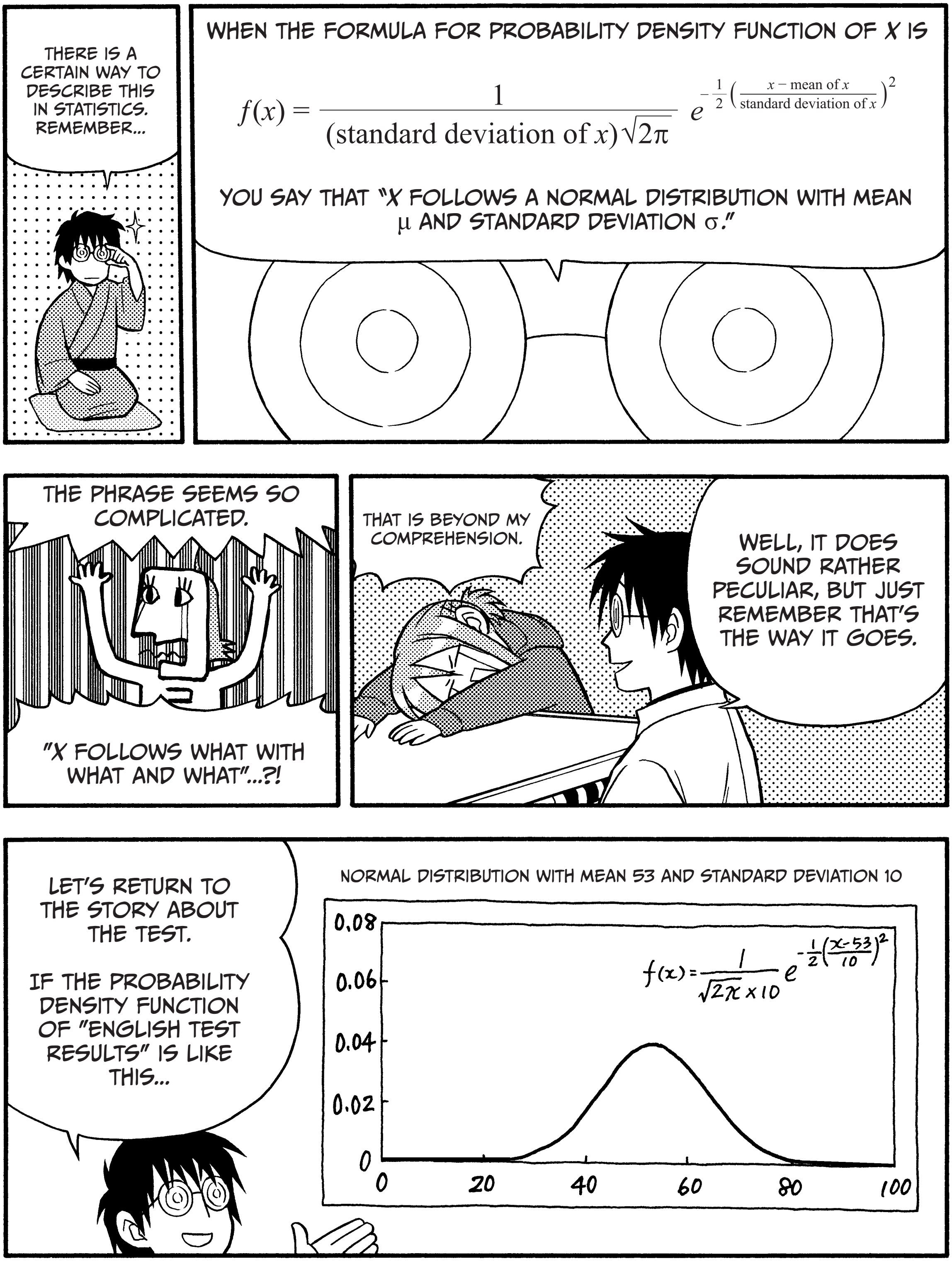

WHEN THE RANGE OF CLASS IN A HISTOGRAM IS DECREASED TO THE LIMIT AT ZERO,
WE CALL THE FORMULA OF THAT LINE THE *PROBABILITY DENSITY FUNCTION.*
PROBABILITY DENSITY FUNCTION
PUSH
THEORETICALLY,
THERE ARE MANY TYPES OF PROBABILITY DENSITY FUNCTION GRAPHS.
TODAY I WILL INTRODUCE YOU TO SOME OF THE MOST IMPORTANT ONES.
PLEASE GO AHEAD.

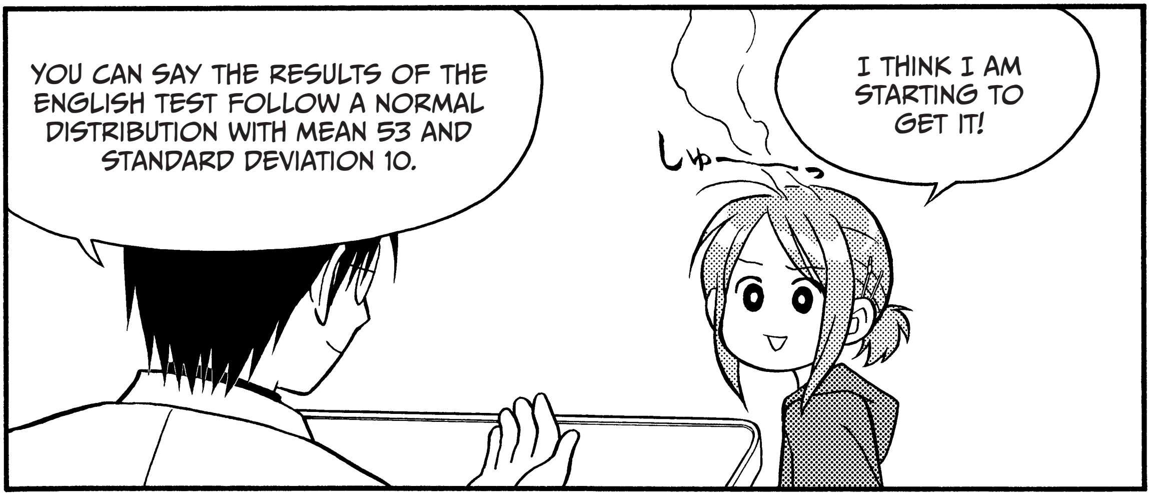
2. NORMAL DISTRIBUTION



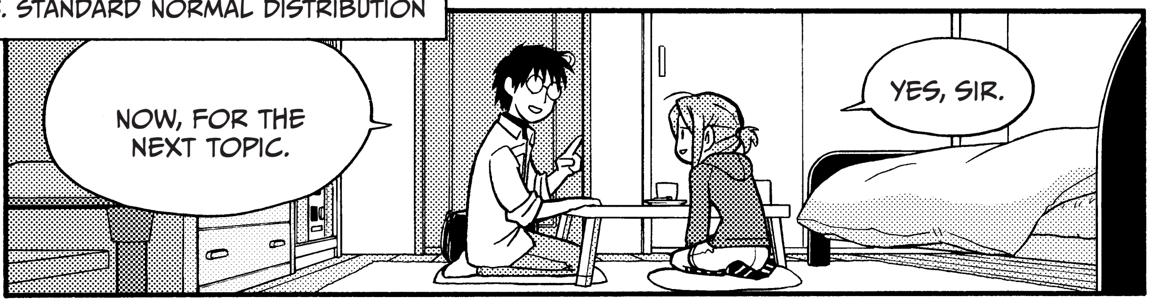
* e IS ALSO KNOWN AS *NAPIER'S CONSTANT*.

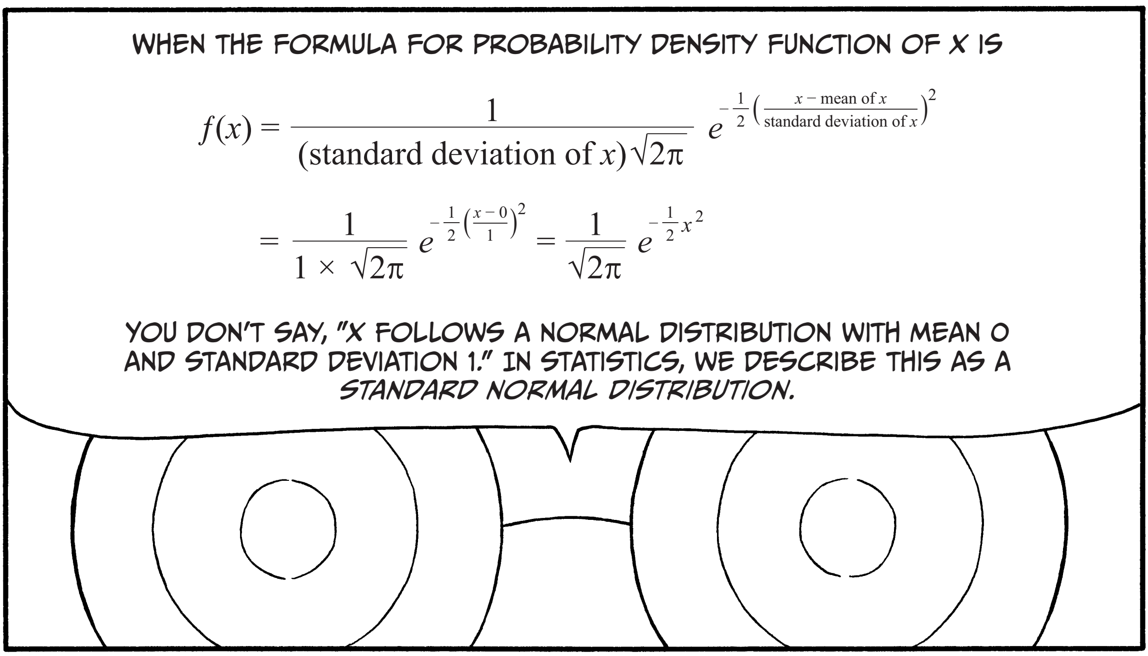

THE GRAPH OF THE NORMAL DISTRIBUTION HAS TWO CHARACTERISTICS.
IT IS SYMMETRICAL, WITH THE MEAN IN THE CENTER.
IT IS AFFECTED BY THE MEAN AND STANDARD DEVIATION.
MEAN IS 53, STANDARD DEVIATION IS 15
$f(x)=\frac{1}{\sqrt{2\pi}\times 15}e^{-\frac{1}{2}\left(\frac{x-53}{15}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
MEAN IS 53, STANDARD DEVIATION IS 5
$f(x)=\frac{1}{\sqrt{2\pi}\times 5}e^{-\frac{1}{2}\left(\frac{x-53}{5}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
MEAN IS 30, STANDARD DEVIATION IS 5
$f(x)=\frac{1}{\sqrt{2\pi}\times 5}e^{-\frac{1}{2}\left(\frac{x-30}{5}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100

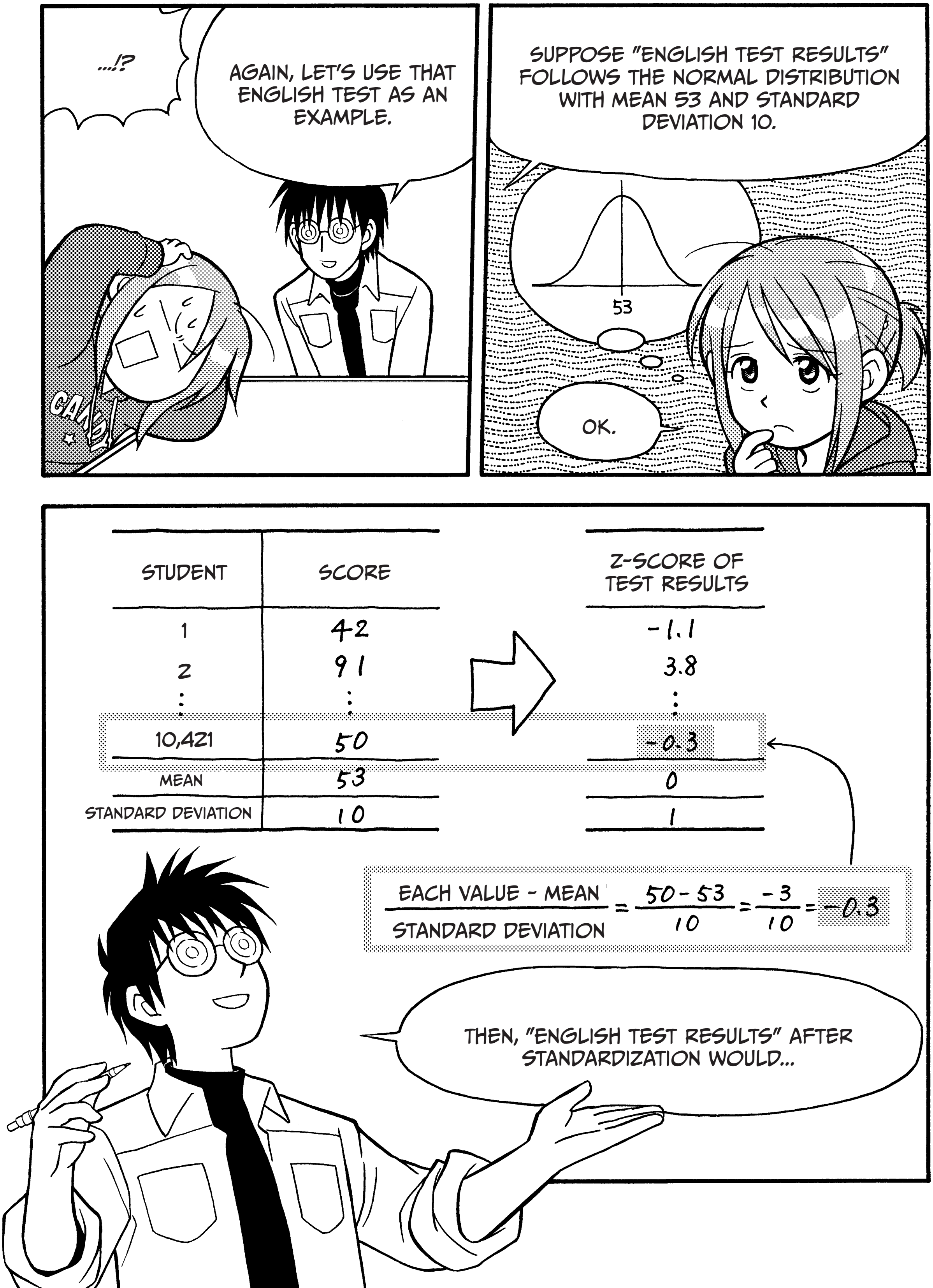

THERE IS A CERTAIN WAY TO DESCRIBE THIS IN STATISTICS. REMEMBER...
WHEN THE FORMULA FOR PROBABILITY DENSITY FUNCTION OF X IS
$$f(x) = \frac{1}{(\text{standard deviation of } x)\sqrt{2\pi}} \, e^{-\frac{1}{2}\left(\frac{x - \text{mean of } x}{\text{standard deviation of } x}\right)^2}$$
YOU SAY THAT "X FOLLOWS A NORMAL DISTRIBUTION WITH MEAN μ AND STANDARD DEVIATION σ."
THE PHRASE SEEMS SO COMPLICATED.
"X FOLLOWS WHAT WITH WHAT AND WHAT"...?!
THAT IS BEYOND MY COMPREHENSION.
WELL, IT DOES SOUND RATHER PECULIAR, BUT JUST REMEMBER THAT'S THE WAY IT GOES.
LET'S RETURN TO THE STORY ABOUT THE TEST.
IF THE PROBABILITY DENSITY FUNCTION OF "ENGLISH TEST RESULTS" IS LIKE THIS...
NORMAL DISTRIBUTION WITH MEAN 53 AND STANDARD DEVIATION 10
$f(x) = \frac{1}{\sqrt{2\pi} \times 10} e^{-\frac{1}{2}\left(\frac{x-53}{10}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100

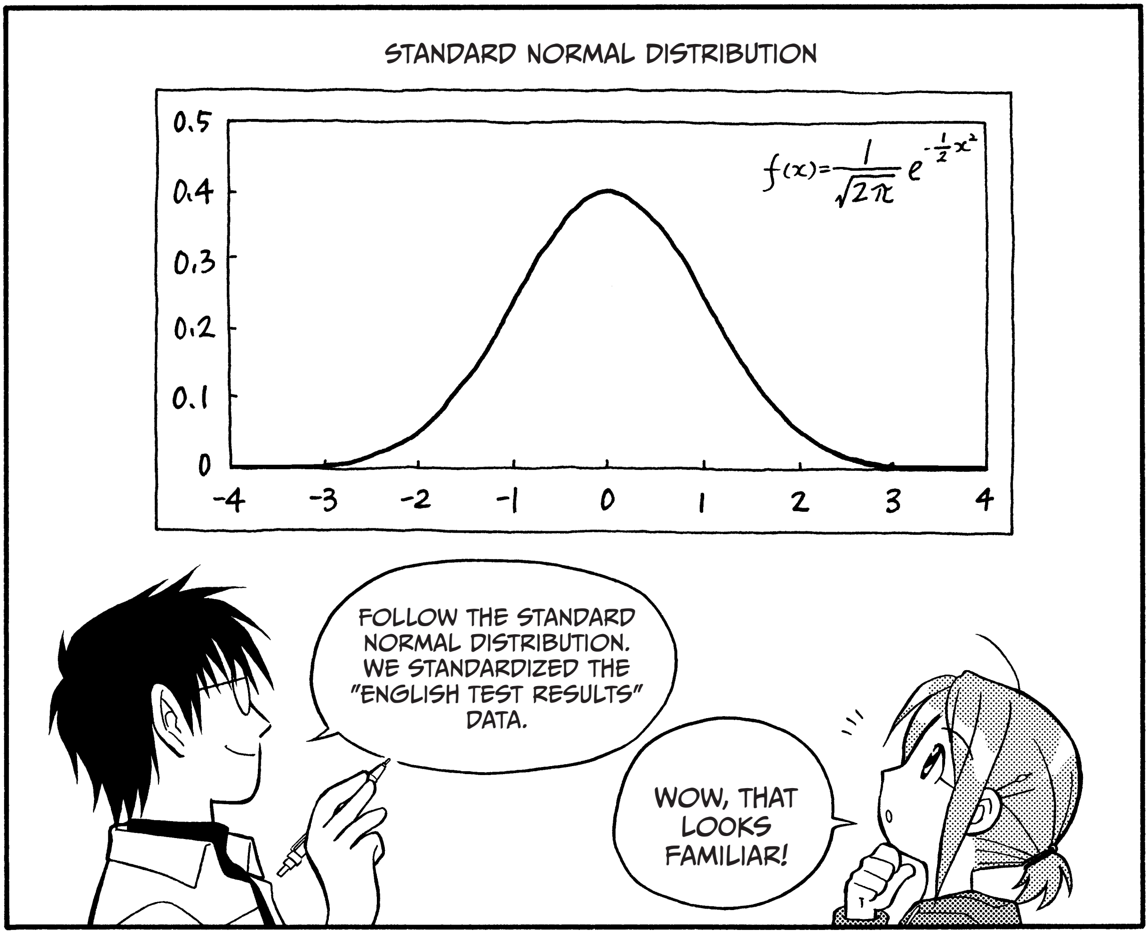


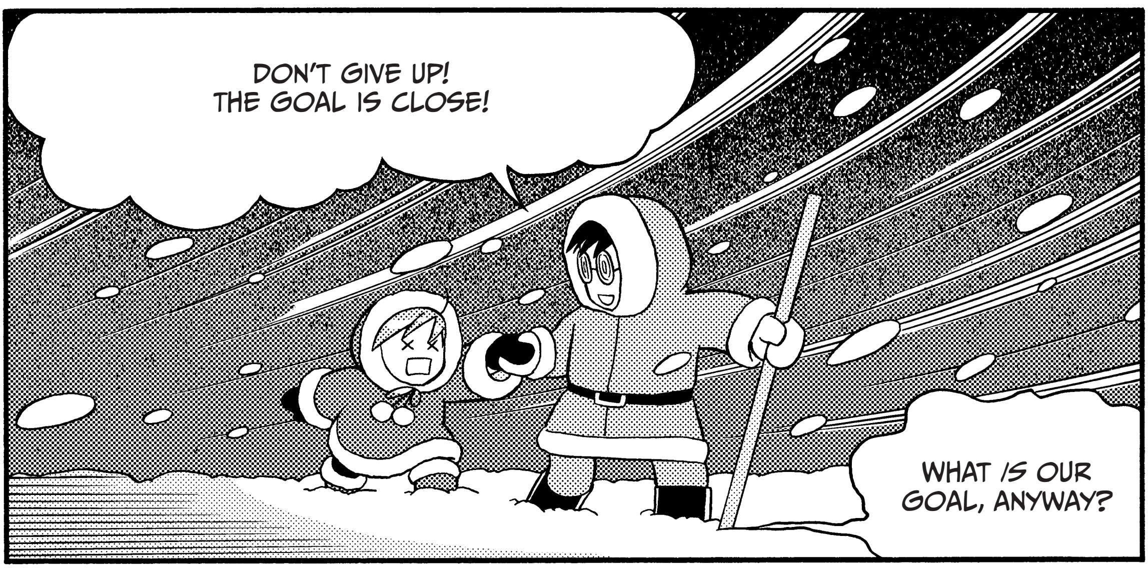
3. STANDARD NORMAL DISTRIBUTION






...!?
AGAIN, LET'S USE THAT ENGLISH TEST AS AN EXAMPLE.
SUPPOSE "ENGLISH TEST RESULTS" FOLLOWS THE NORMAL DISTRIBUTION WITH MEAN 53 AND STANDARD DEVIATION 10.
53
OK.
STUDENT | SCORE
1 | 42
2 | 91
10,421 | 50
MEAN | 53
STANDARD DEVIATION | 10
Z-SCORE OF TEST RESULTS
-1.1
3.8
-0.3
0
1
EACH VALUE - MEAN / STANDARD DEVIATION = 50 - 53 / 10 = -3 / 10 = -0.3
THEN, "ENGLISH TEST RESULTS" AFTER STANDARDIZATION WOULD...


STANDARD NORMAL DISTRIBUTION
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$
FOLLOW THE STANDARD NORMAL DISTRIBUTION. WE STANDARDIZED THE "ENGLISH TEST RESULTS" DATA.
WOW, THAT LOOKS FAMILIAR!


DON'T GIVE UP! THE GOAL IS CLOSE!
WHAT *IS* OUR GOAL, ANYWAY?

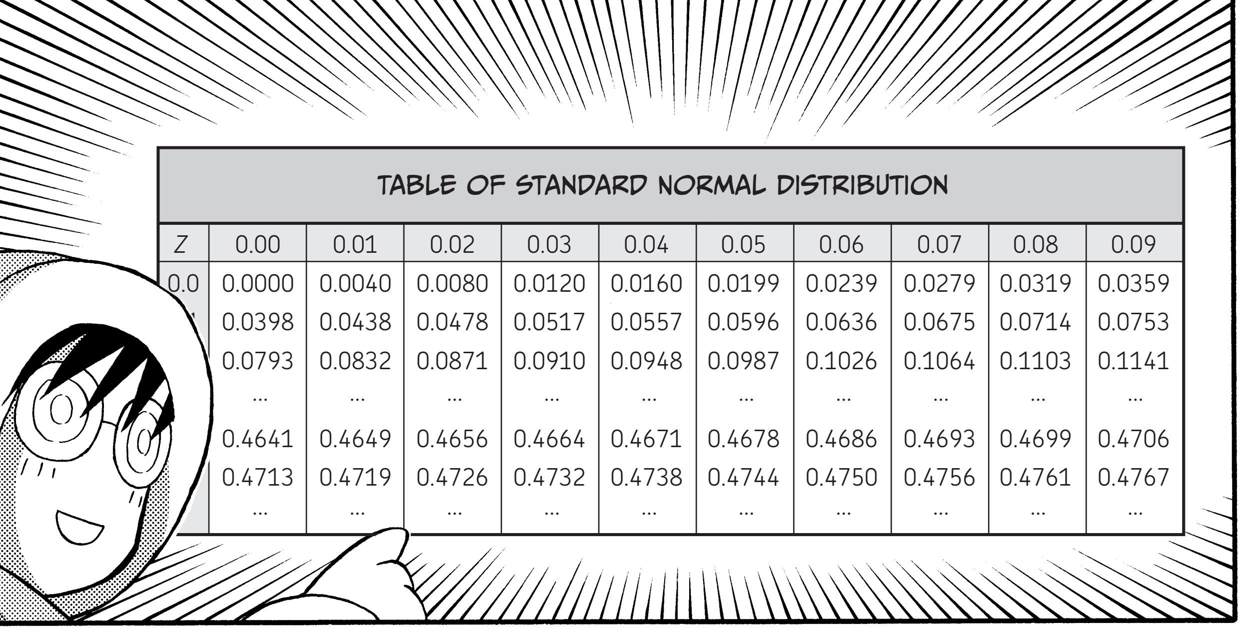
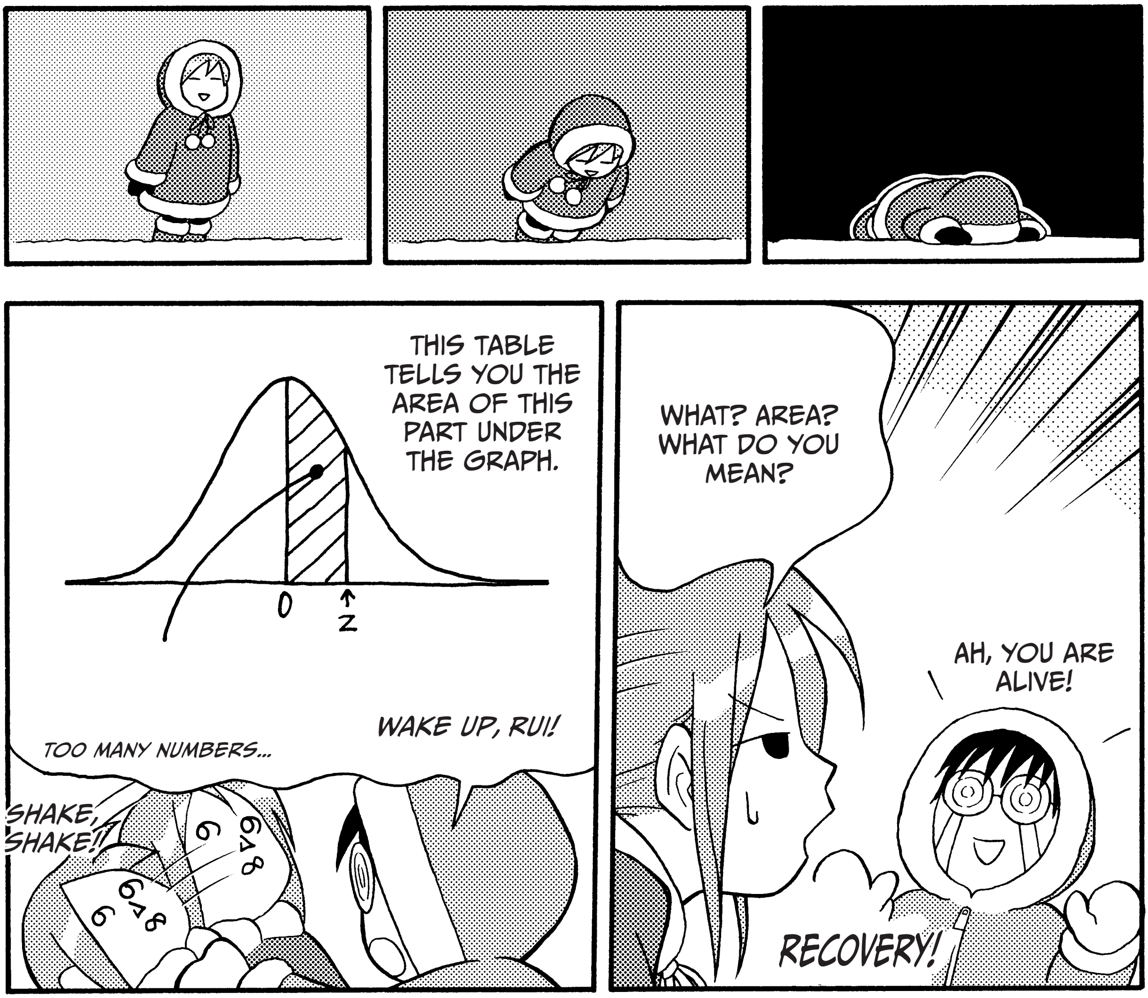
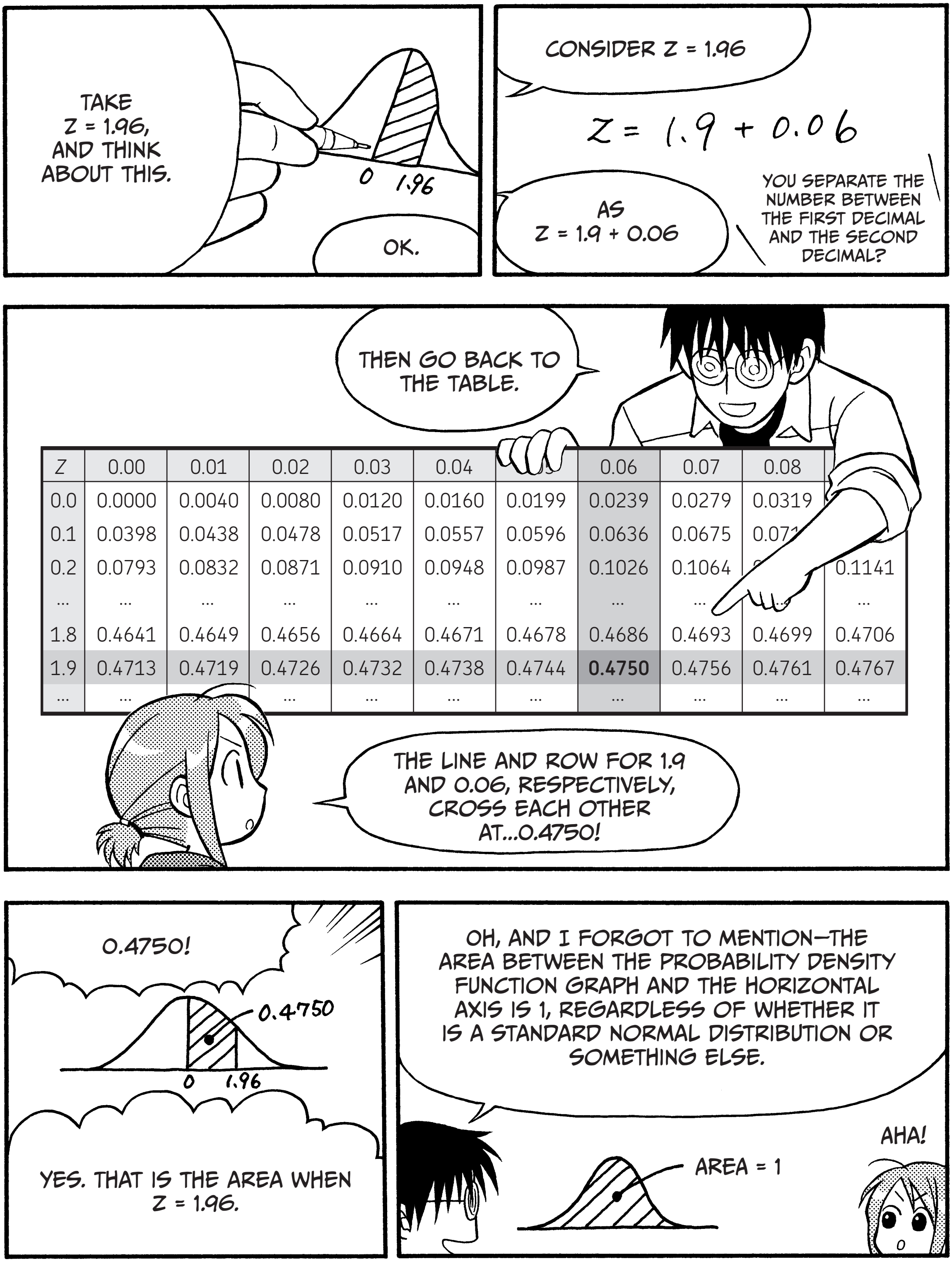


Table of Standard Normal Distribution										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
	...	...	...	...	...	...	...	...	...	...
	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
	...	...	...	...	...	...	...	...	...	...

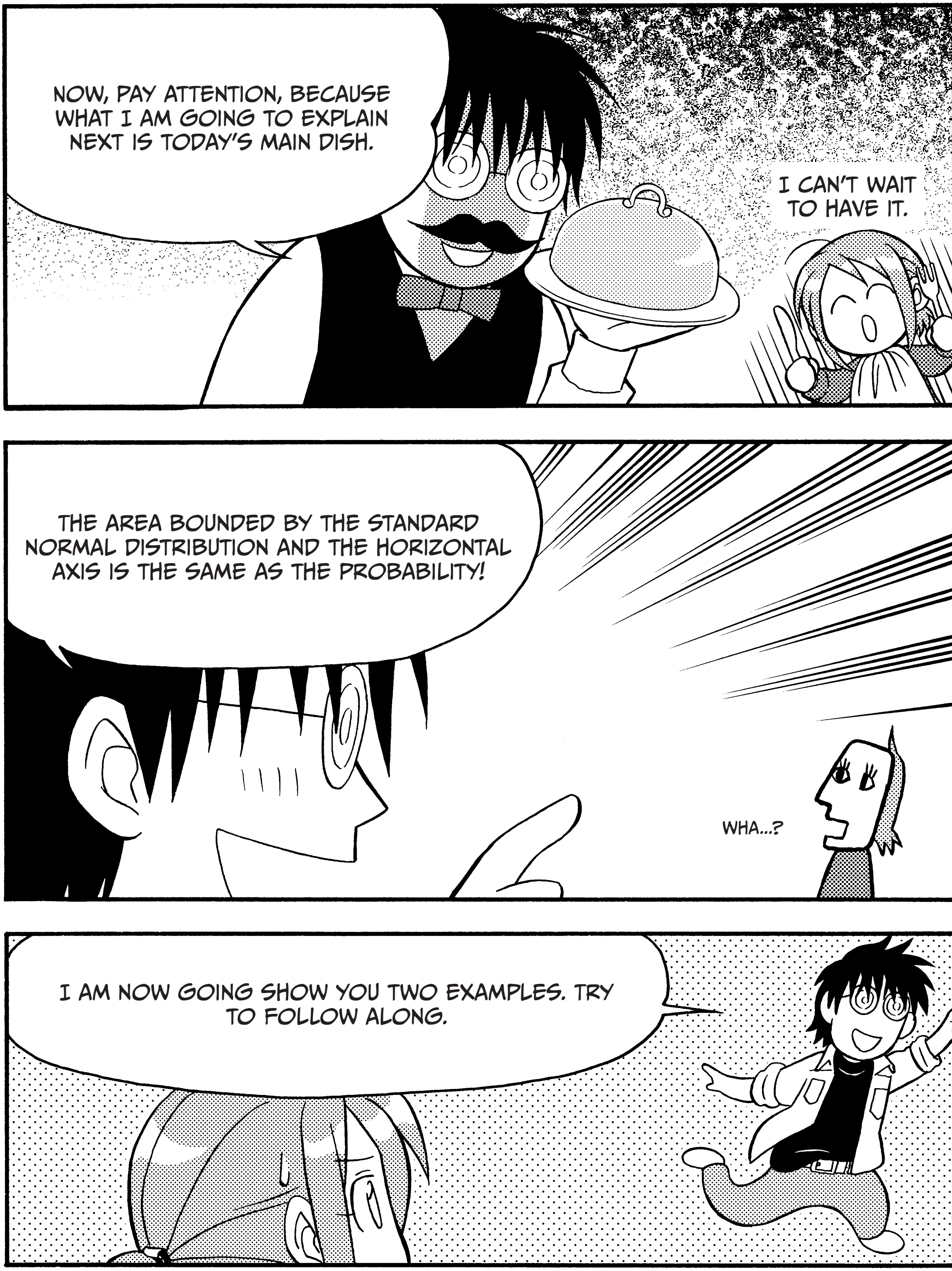




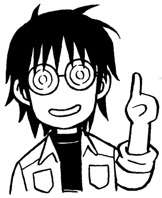
TAKE Z = 1.96, AND THINK ABOUT THIS.
0
1.96
OK.
CONSIDER Z = 1.96
Z = 1.9 + 0.06
AS Z = 1.9 + 0.06
YOU SEPARATE THE NUMBER BETWEEN THE FIRST DECIMAL AND THE SECOND DECIMAL?
THEN GO BACK TO THE TABLE.

Z	0.00	0.01	0.02	0.03	0.04		0.06	0.07	0.08	
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.071	
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064		0.1141
...	...	...	...	...	...	...	...	...	...	...
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	**0.4750**	0.4756	0.4761	0.4767
...	...		...	...	...	...	...	...	...	...

THE LINE AND ROW FOR 1.9 AND 0.06, RESPECTIVELY, CROSS EACH OTHER AT...0.4750!
0.4750!
0.4750
0
1.96
YES. THAT IS THE AREA WHEN Z = 1.96.
OH, AND I FORGOT TO MENTION—THE AREA BETWEEN THE PROBABILITY DENSITY FUNCTION GRAPH AND THE HORIZONTAL AXIS IS 1, REGARDLESS OF WHETHER IT IS A STANDARD NORMAL DISTRIBUTION OR SOMETHING ELSE.
AREA = 1
AHA!

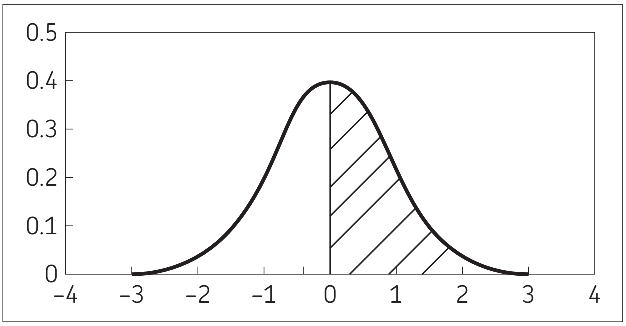

NOW, PAY ATTENTION, BECAUSE WHAT I AM GOING TO EXPLAIN NEXT IS TODAY'S MAIN DISH.
I CAN'T WAIT TO HAVE IT.
THE AREA BOUNDED BY THE STANDARD NORMAL DISTRIBUTION AND THE HORIZONTAL AXIS IS THE SAME AS THE PROBABILITY!
WHA...?
I AM NOW GOING SHOW YOU TWO EXAMPLES. TRY TO FOLLOW ALONG.

EXAMPLE 1

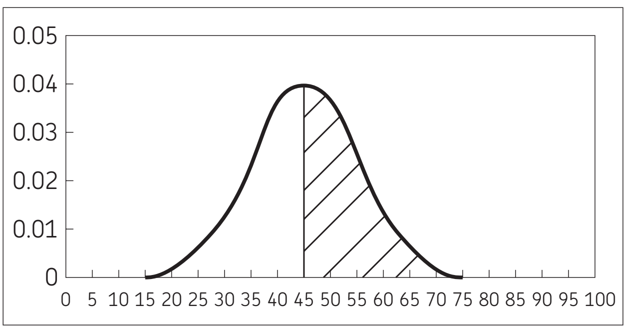


All high school freshmen in prefecture B took a math test. After the tests were marked, the test results turned out to follow a normal distribution with a mean of 45 and a standard deviation of 10. Now, think carefully. The five sentences below all have the same meaning.

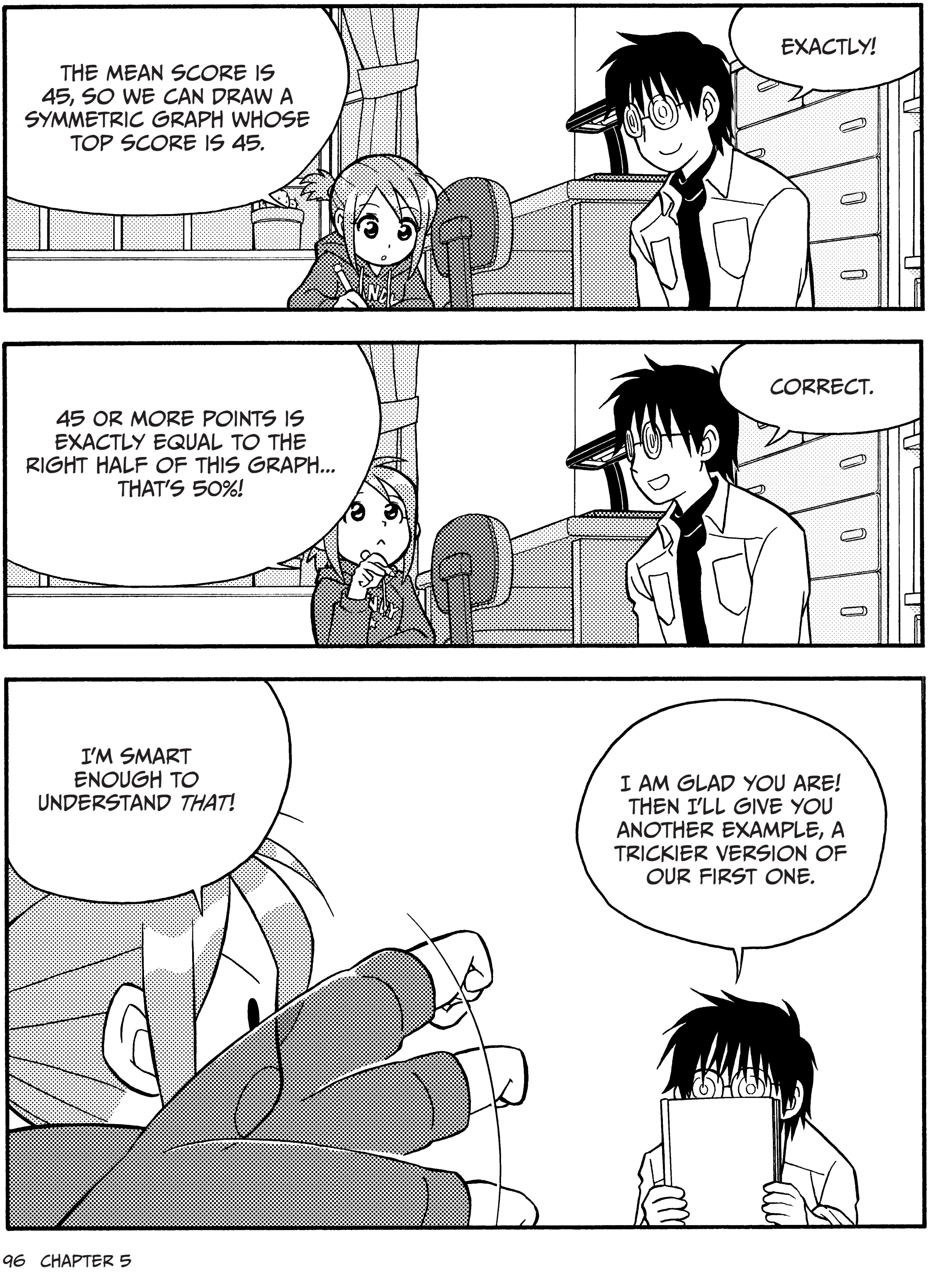
1. In a normal distribution with an average of 45 and a standard deviation of 10, the shaded area in the chart below is 0.5.



2. The ratio of students who scored 45 points or more is 0.5 (50% of all students tested).
3. When one student is randomly chosen from all students tested, the probability that the student's score is 45 or more is 0.5 (50%).
4. In a normal distribution of standardized "math test results," the ratio of students with a standard score of 0 or more is 0.5 (50% of all students tested).

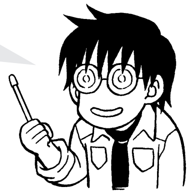


5. When one student's results are randomly chosen from all of those tested in a normal distribution of standardized "math test results," the probability that the selected student's standard score is 0 or more is 0.5 (50%).

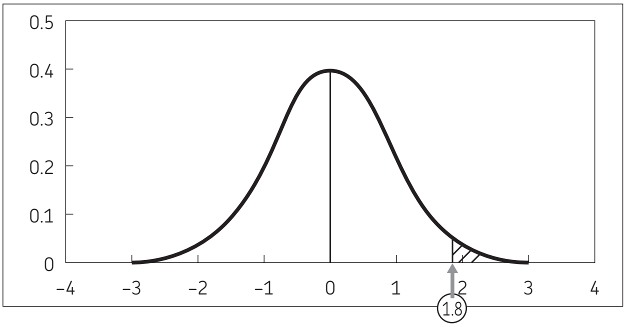

THE MEAN SCORE IS 45, SO WE CAN DRAW A SYMMETRIC GRAPH WHOSE TOP SCORE IS 45.
EXACTLY!
45 OR MORE POINTS IS EXACTLY EQUAL TO THE RIGHT HALF OF THIS GRAPH... THAT'S 50%!
CORRECT.
I'M SMART ENOUGH TO UNDERSTAND *THAT*!
I AM GLAD YOU ARE! THEN I'LL GIVE YOU ANOTHER EXAMPLE, A TRICKIER VERSION OF OUR FIRST ONE.

EXAMPLE II

All high school freshmen in prefecture B took a math test. Now, think carefully. The five sentences below all have the same meaning.



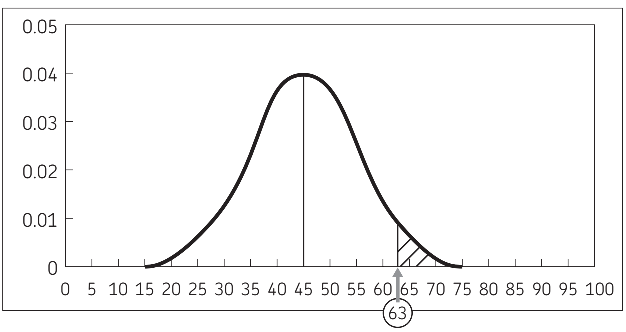
1. In a normal distribution with a mean of 45 and a standard deviation of 10, the shaded area in the chart below is 0.5 - 0.4641 = 0.0359.



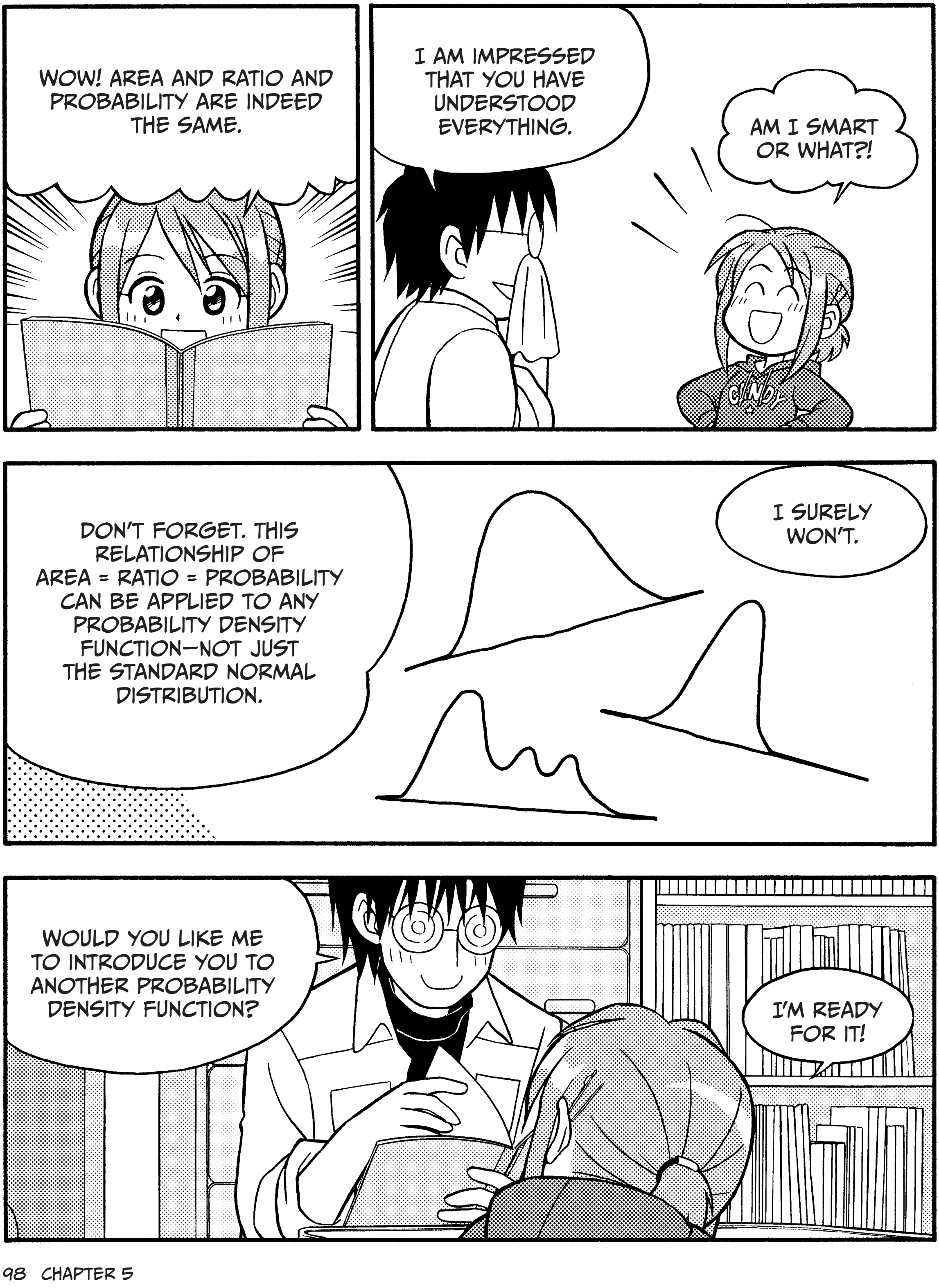
2. The ratio of students who scored 63 points or more is 0.5 – 0.4641 = 0.0359 (3.59% of all students tested).

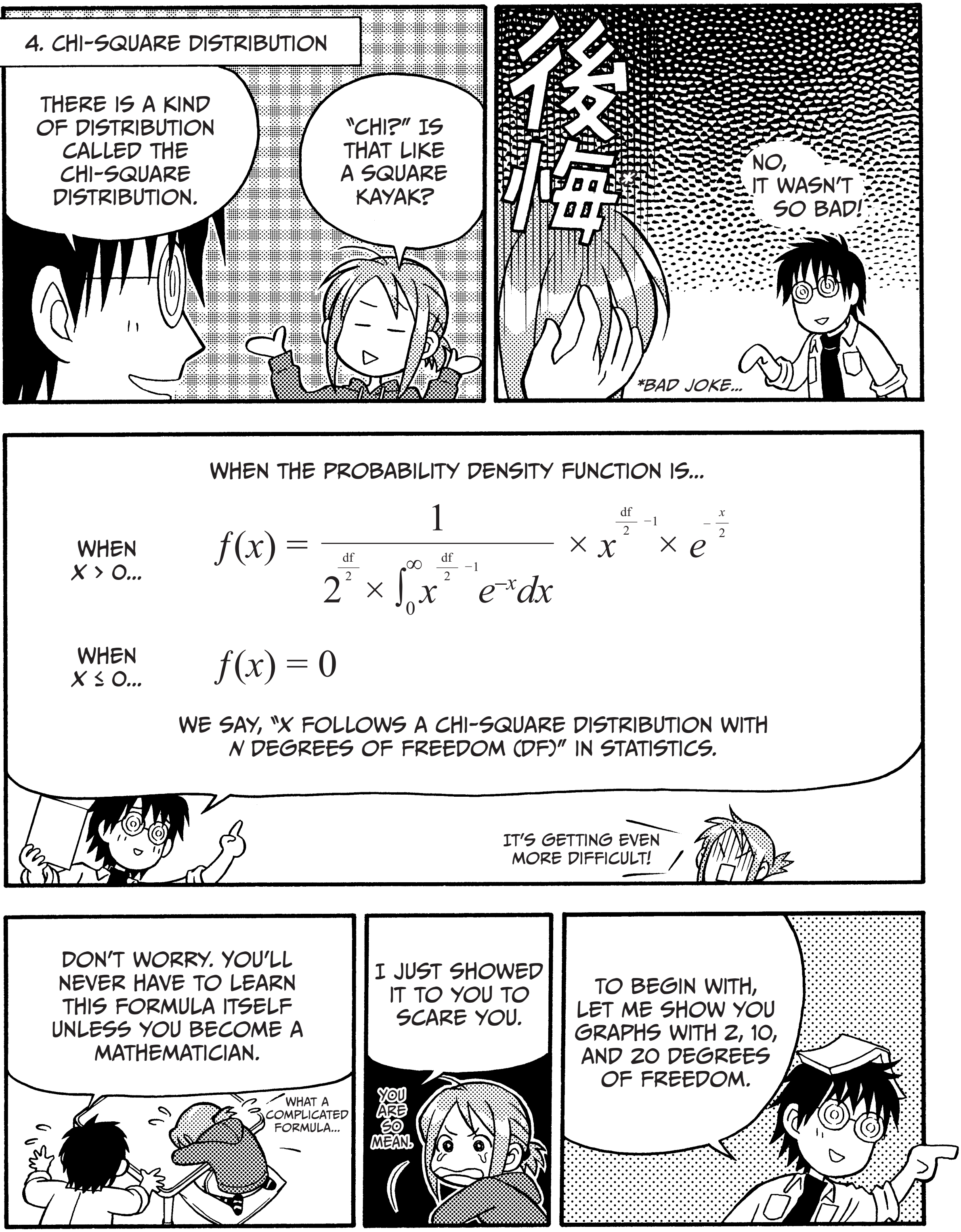
3. When one student is randomly chosen from all those tested, the probability that the student's score is 63 or more is 0.5 – 0.4641 = 0.0359 (3.59%).

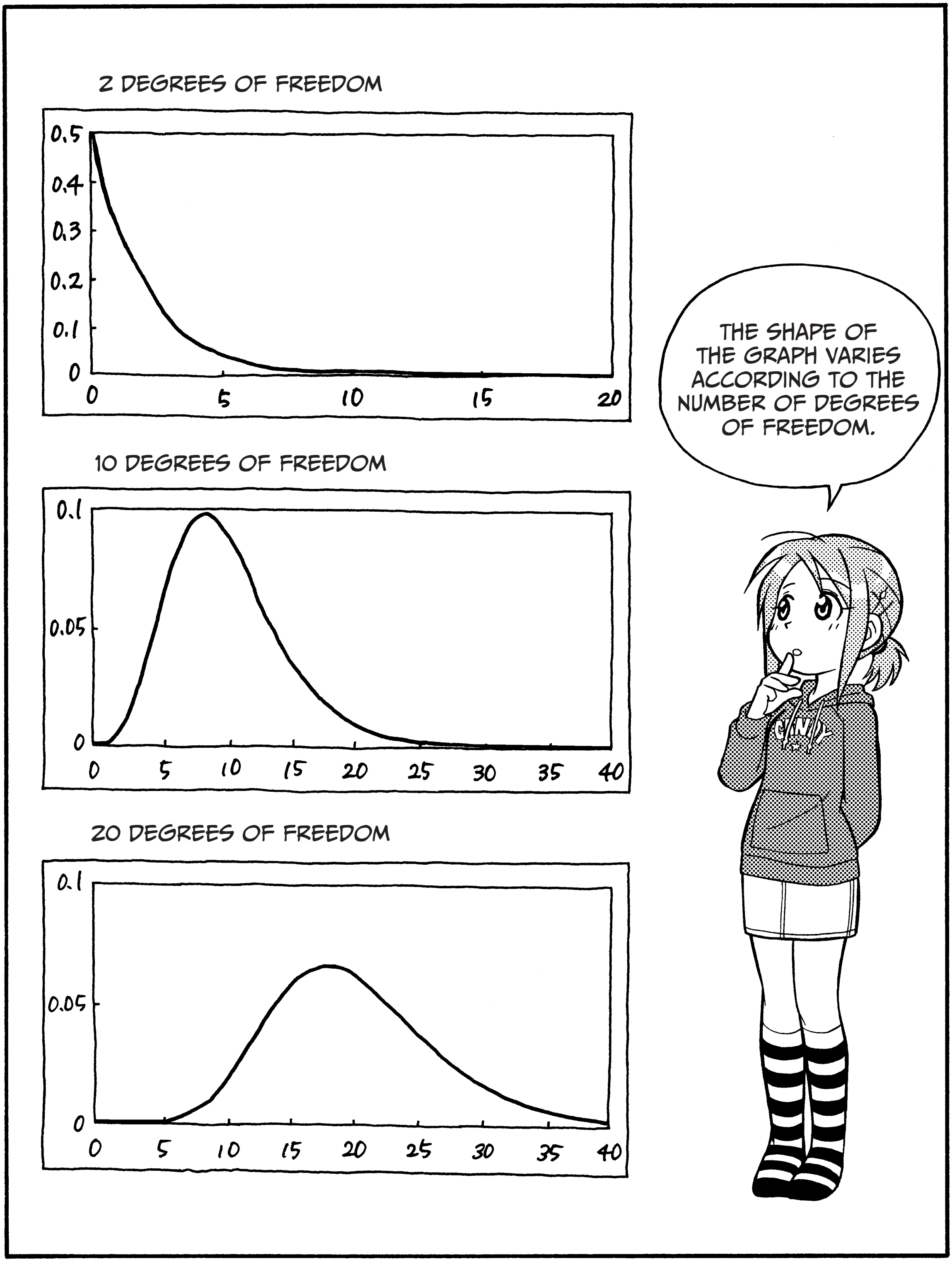
4. In a normal distribution of standardized test results, the ratio of students with standard scores (or z-scores) of 1.8 or more [(each value – average) ÷ standard deviation = (63 – 45) ÷ 10 = 18 ÷ 10 = 1.8] is 3.59% (0.5 – 0.4641 = 0.0359). You can also obtain this value from a table of standard normal distribution.

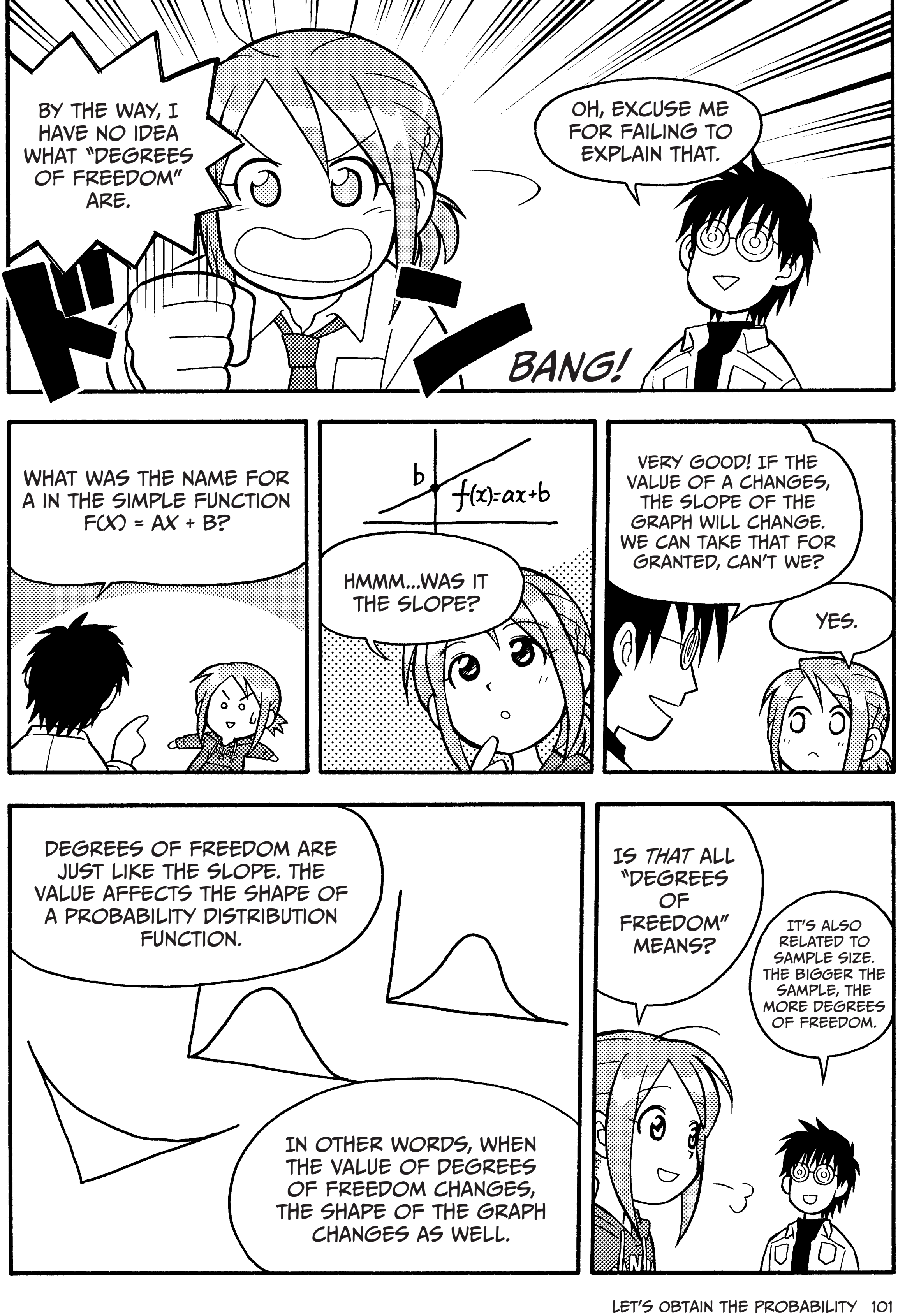


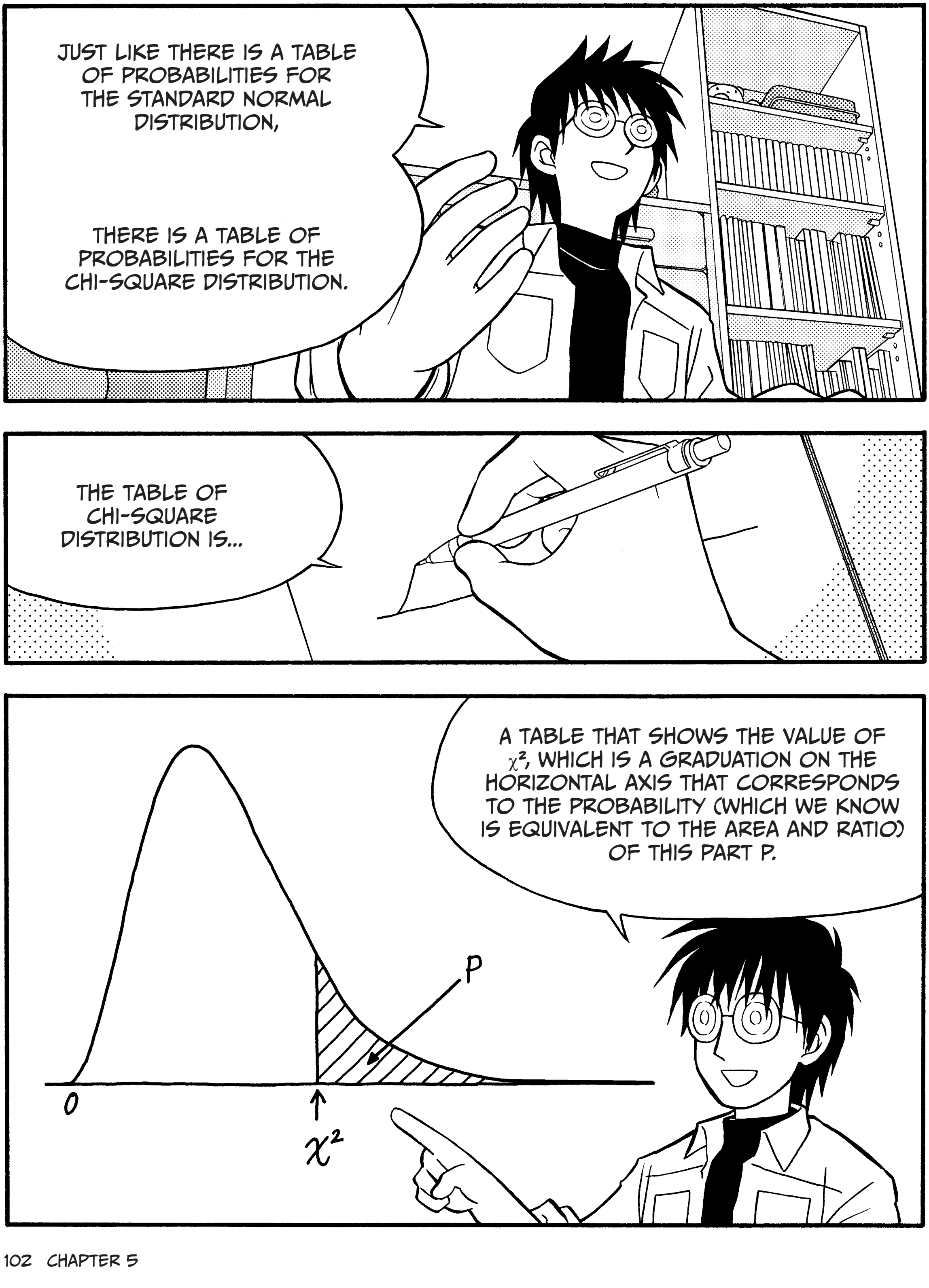
5. When one student is randomly chosen from all those tested in a normal distribution of standardized "math test results," the probability that the student's standard score is 1.8 or more is 0.5 – 0.4641 = 0.0359 (3.59%).

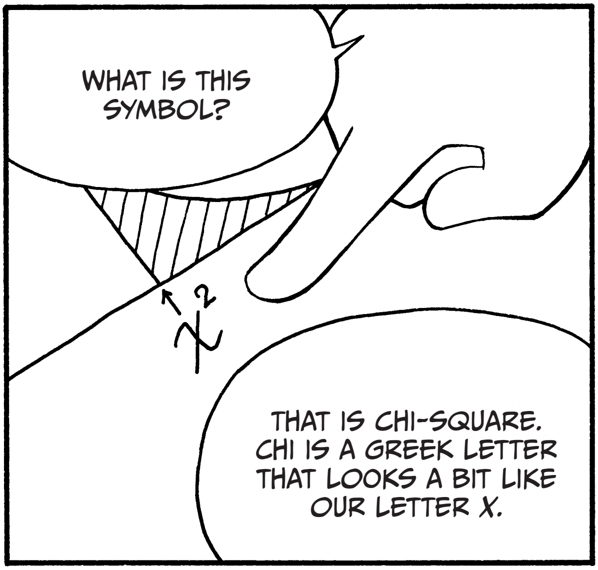

WOW! AREA AND RATIO AND PROBABILITY ARE INDEED THE SAME.
I AM IMPRESSED THAT YOU HAVE UNDERSTOOD EVERYTHING.
AM I SMART OR WHAT?!
CANDY
DON'T FORGET. THIS RELATIONSHIP OF AREA = RATIO = PROBABILITY CAN BE APPLIED TO ANY PROBABILITY DENSITY FUNCTION—NOT JUST THE STANDARD NORMAL DISTRIBUTION.
I SURELY WON'T.
WOULD YOU LIKE ME TO INTRODUCE YOU TO ANOTHER PROBABILITY DENSITY FUNCTION?
I'M READY FOR IT!

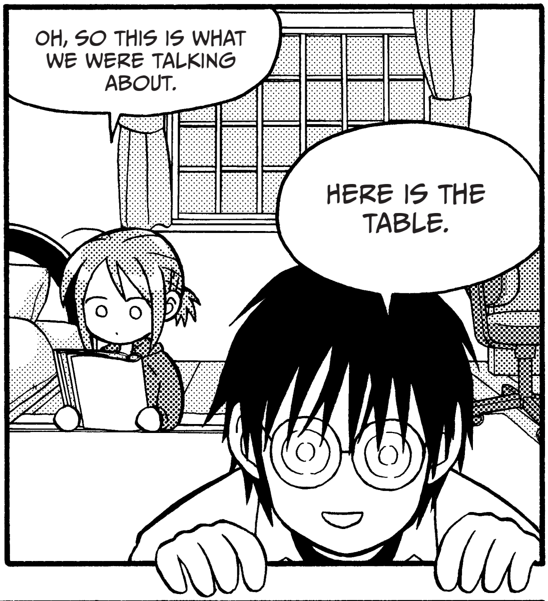

4. CHI-SQUARE DISTRIBUTION
THERE IS A KIND OF DISTRIBUTION CALLED THE CHI-SQUARE DISTRIBUTION.
"CHI?" IS THAT LIKE A SQUARE KAYAK?
後悔*
NO, IT WASN'T SO BAD!
*BAD JOKE...
WHEN THE PROBABILITY DENSITY FUNCTION IS...
WHEN $x > 0$...
$$f(x) = \frac{1}{2^{\frac{\mathrm{df}}{2}} \times \int_0^\infty x^{\frac{\mathrm{df}}{2}-1} e^{-x} dx} \times x^{\frac{\mathrm{df}}{2}-1} \times e^{-\frac{x}{2}}$$
WHEN $x \leq 0$...
$$f(x) = 0$$
WE SAY, "X FOLLOWS A CHI-SQUARE DISTRIBUTION WITH *N* DEGREES OF FREEDOM (DF)" IN STATISTICS.
IT'S GETTING EVEN MORE DIFFICULT!
DON'T WORRY. YOU'LL NEVER HAVE TO LEARN THIS FORMULA ITSELF UNLESS YOU BECOME A MATHEMATICIAN.
WHAT A COMPLICATED FORMULA...
I JUST SHOWED IT TO YOU TO SCARE YOU.
YOU ARE SO MEAN.
TO BEGIN WITH, LET ME SHOW YOU GRAPHS WITH 2, 10, AND 20 DEGREES OF FREEDOM.


2 DEGREES OF FREEDOM
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
10 DEGREES OF FREEDOM
0.1
0.05
0
0
5
10
15
20
25
30
35
40
20 DEGREES OF FREEDOM
0.1
0.05
0
0
5
10
15
20
25
30
35
40
THE SHAPE OF THE GRAPH VARIES ACCORDING TO THE NUMBER OF DEGREES OF FREEDOM.


BY THE WAY, I HAVE NO IDEA WHAT "DEGREES OF FREEDOM" ARE.
OH, EXCUSE ME FOR FAILING TO EXPLAIN THAT.
BANG!
WHAT WAS THE NAME FOR A IN THE SIMPLE FUNCTION F(X) = AX + B?
b
f(x)=ax+b
HMMM...WAS IT THE SLOPE?
VERY GOOD! IF THE VALUE OF A CHANGES, THE SLOPE OF THE GRAPH WILL CHANGE. WE CAN TAKE THAT FOR GRANTED, CAN'T WE?
YES.
DEGREES OF FREEDOM ARE JUST LIKE THE SLOPE. THE VALUE AFFECTS THE SHAPE OF A PROBABILITY DISTRIBUTION FUNCTION.
IN OTHER WORDS, WHEN THE VALUE OF DEGREES OF FREEDOM CHANGES, THE SHAPE OF THE GRAPH CHANGES AS WELL.
IS *THAT* ALL "DEGREES OF FREEDOM" MEANS?
IT'S ALSO RELATED TO SAMPLE SIZE. THE BIGGER THE SAMPLE, THE MORE DEGREES OF FREEDOM.


JUST LIKE THERE IS A TABLE OF PROBABILITIES FOR THE STANDARD NORMAL DISTRIBUTION,
THERE IS A TABLE OF PROBABILITIES FOR THE CHI-SQUARE DISTRIBUTION.
THE TABLE OF CHI-SQUARE DISTRIBUTION IS...
A TABLE THAT SHOWS THE VALUE OF χ^2, WHICH IS A GRADUATION ON THE HORIZONTAL AXIS THAT CORRESPONDS TO THE PROBABILITY (WHICH WE KNOW IS EQUIVALENT TO THE AREA AND RATIO) OF THIS PART P.
P
0
χ^2

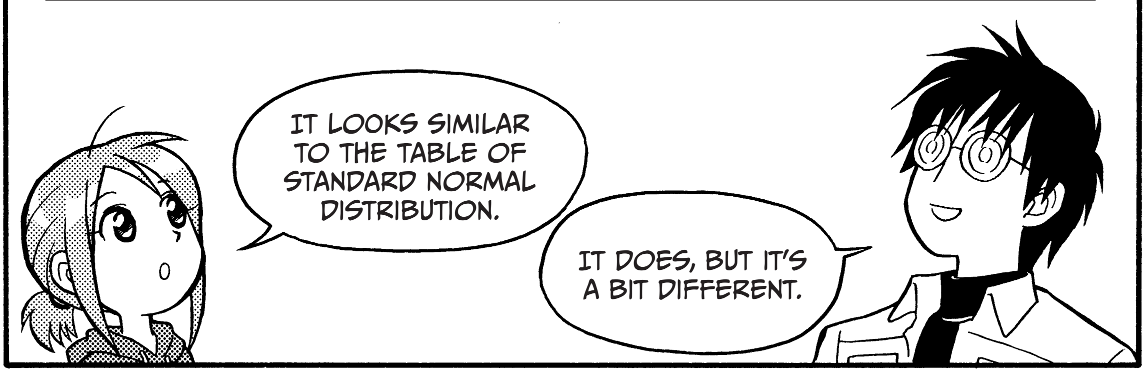


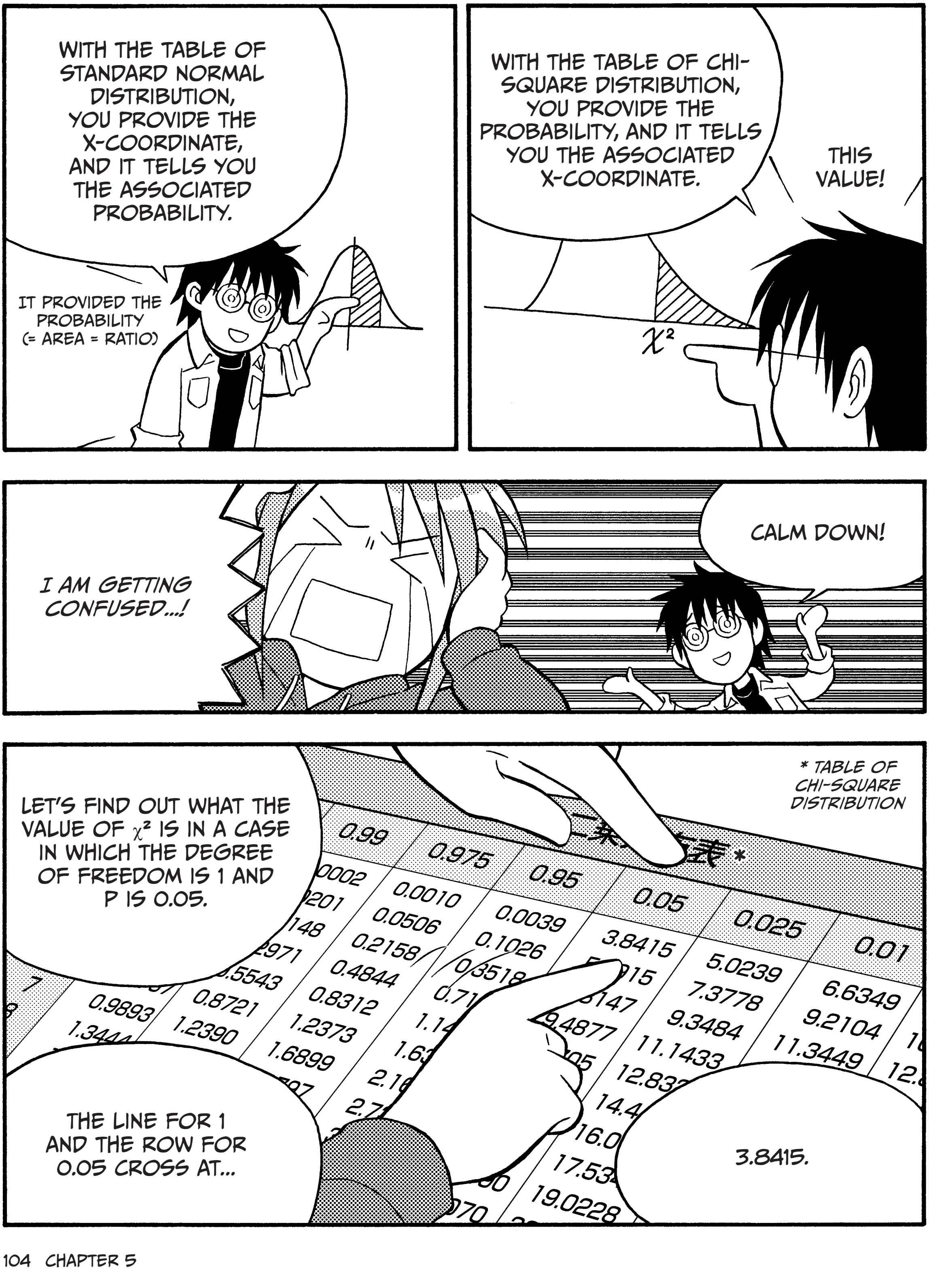


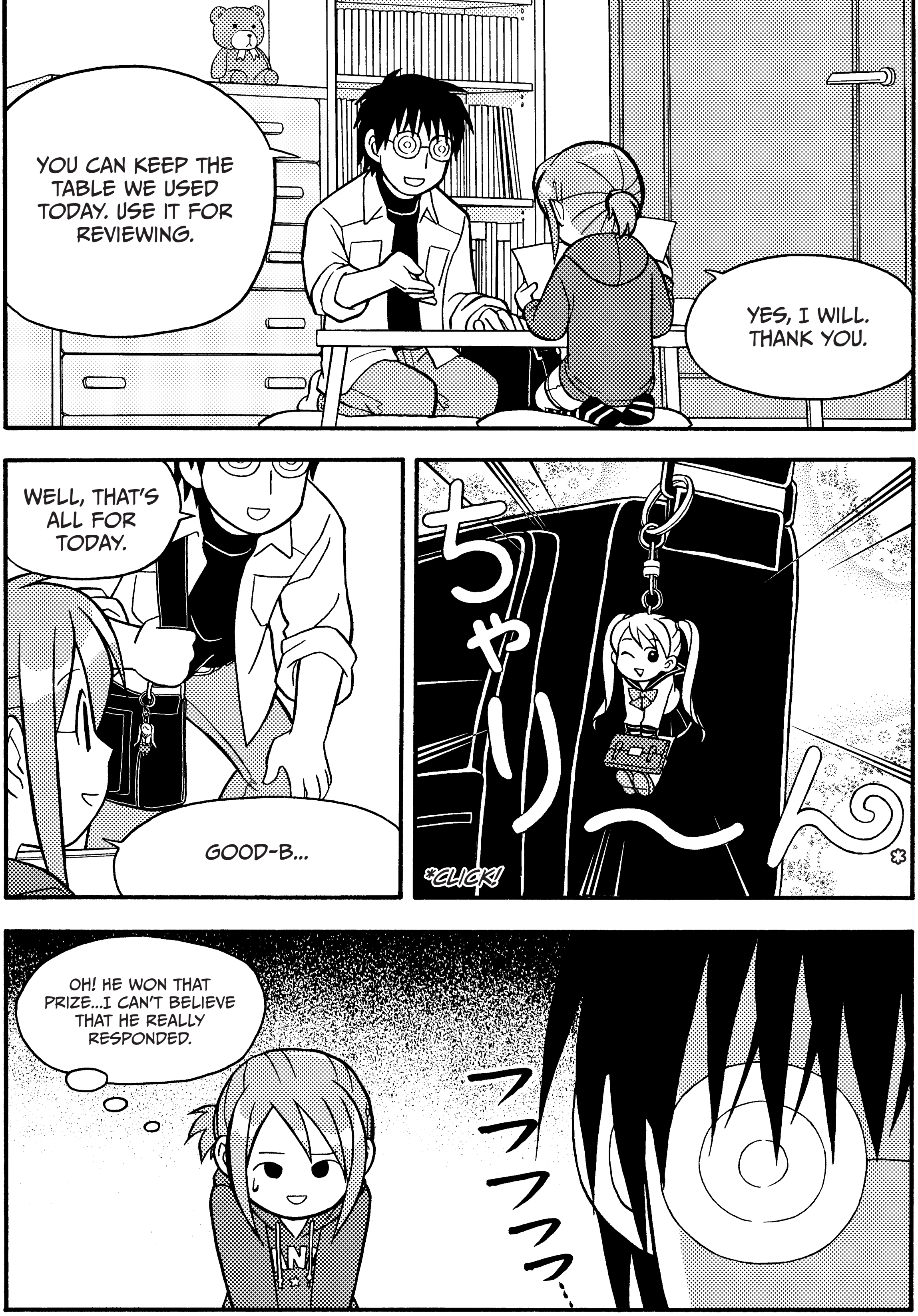
DEGREES OF FREEDOM

TABLE OF CHI-SQUARE DISTRIBUTION

P	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000039	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5893
10	2.1558	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1881
...	...	...	...	...	...	...	...	...




WITH THE TABLE OF STANDARD NORMAL DISTRIBUTION, YOU PROVIDE THE X-COORDINATE, AND IT TELLS YOU THE ASSOCIATED PROBABILITY.
IT PROVIDED THE PROBABILITY (= AREA = RATIO)
WITH THE TABLE OF CHI-SQUARE DISTRIBUTION, YOU PROVIDE THE PROBABILITY, AND IT TELLS YOU THE ASSOCIATED X-COORDINATE.
THIS VALUE!
χ^2
I AM GETTING CONFUSED...!
CALM DOWN!
LET'S FIND OUT WHAT THE VALUE OF χ^2 IS IN A CASE IN WHICH THE DEGREE OF FREEDOM IS 1 AND P IS 0.05.
* TABLE OF CHI-SQUARE DISTRIBUTION
0.99
0.975
0.95
0.05
0.025
0.01
THE LINE FOR 1 AND THE ROW FOR 0.05 CROSS AT...
3.8415.


YOU CAN KEEP THE TABLE WE USED TODAY. USE IT FOR REVIEWING.
YES, I WILL. THANK YOU.
WELL, THAT'S ALL FOR TODAY.
GOOD-B...
ちゃり〜ん
*CLICK!
OH! HE WON THAT PRIZE...I CAN'T BELIEVE THAT HE REALLY RESPONDED.
フフフフフ...

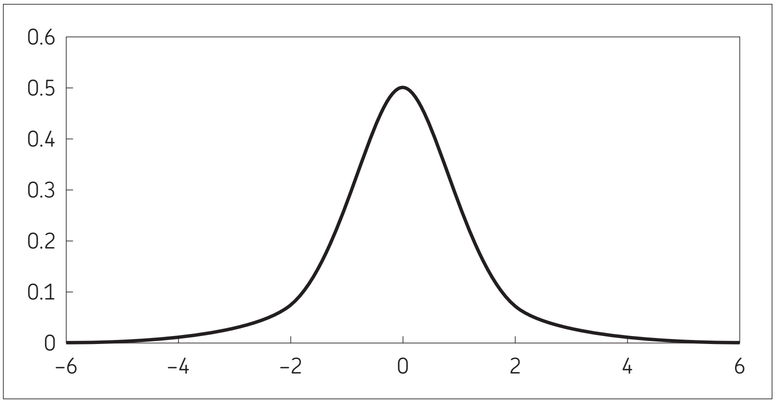
5. T Distribution

The probability density function below is a popular topic in statistics.

$$f(x) = \frac{\int_0^\infty x^{\frac{\text{df}+1}{2}-1} e^{-x}dx}{\sqrt{\text{df} \times \pi} \times \int_0^\infty x^{\frac{\text{df}}{2}-1} e^{-x}dx} \times \left(1 + \frac{x^2}{\text{df}}\right)^{-\frac{\text{df}+1}{2}}$$

When the probability density function for *x* looks like this, we say, "*x* follows a t distribution with *n* degrees of freedom."

Here is a case with 5 degrees of freedom:



6. F Distribution

The probability density function below is a popular topic in statistics.

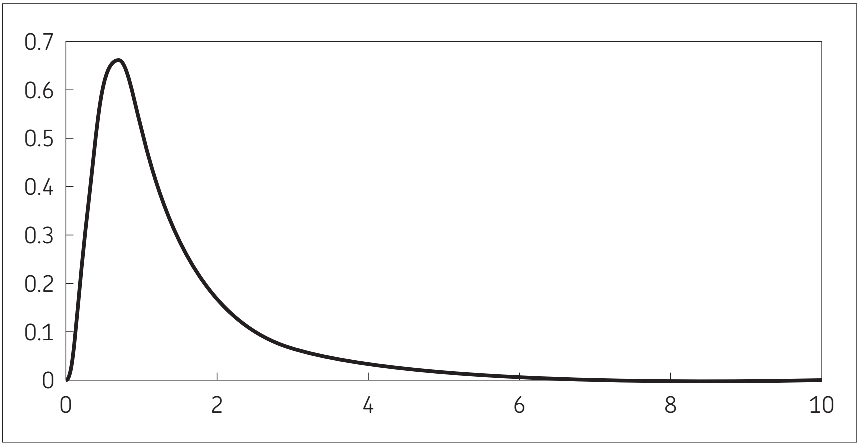
when $x > 0$:

$$f(x) = \frac{\left(\int_0^\infty x^{\frac{\left(\text{first df} + \text{second df}\right)}{2}-1} e^{-x}dx\right) \times \left(\text{first df}\right)^{\frac{\text{first df}}{2}} \times \left(\text{second df}\right)^{\frac{\text{second df}}{2}}}{\left(\int_0^\infty x^{\frac{\text{first df}}{2}-1} e^{-x}dx\right) \times \left(\int_0^\infty x^{\frac{\text{second df}}{2}-1} e^{-x}dx\right)} \times \frac{x^{\frac{\text{first df}}{2}-1}}{\left(\text{first df} \times x + \text{second df}\right)^{\frac{\left(\text{first df} + \text{second df}\right)}{2}}}$$

when $x \leq 0$: $f(x) = 0$

When the probability density function for *x* looks like this, we say, "*x* follows an F distribution with the first degree of freedom *m* and the second degree of freedom *n*."

Here is a case in which the first degree of freedom is 10 and the second degree of freedom is 5:



7. Distributions and Excel

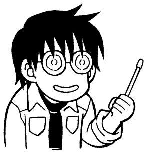
Until the rise of personal computers (roughly speaking, around the beginning of the 1990s), it was difficult for an individual to calculate the probability without tables of standard normal distribution or chi-square distribution. However, these tables of distribution are not used much anymore—you can use Excel functions to find the same values as the ones provided by the tables. This enables individuals to calculate even more types of values than the ones found in the tables of distribution. Table 5-1 summarizes Excel functions related to various distributions. (Refer to the appendix on page 191 for more information on making calculations with Excel.)

Table 5-1: Excel Functions Related to Various Distributions

Distribution	Functions	Feature of the function
normal*	NORMDIST	Calculates the probability that corresponds to a point on the horizontal axis.
normal	NORMINV	Calculates a point on the horizontal axis that corresponds to the probability.
standard normal	NORMSDIST	Calculates the probability that corresponds to a point on the horizontal axis.
standard normal	NORMSINV	Calculates a point on the horizontal axis that corresponds to the probability.
chi-square	CHIDIST	Calculates the probability that corresponds to a point on the horizontal axis.
chi-square	CHIINV	Calculates a point on the horizontal axis that corresponds to the probability.
t	TDIST	Calculates the probability that corresponds to a point on the horizontal axis.
t	TINV	Calculates a point on the horizontal axis that corresponds to the probability.
F	FDIST	Calculates the probability that corresponds to a point on the horizontal axis.
F	FINV	Calculates a point on the horizontal axis that corresponds to the probability.

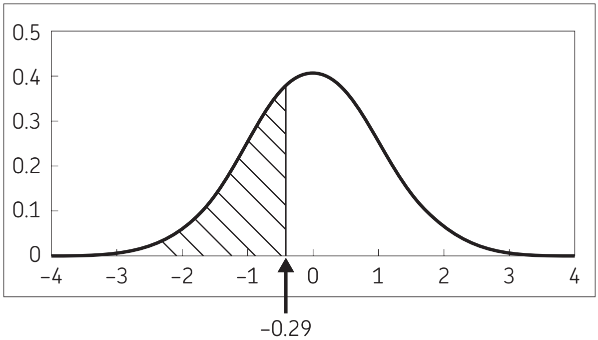
* The probability density function for normal distribution is affected by the mean and standard deviation. Thus, it is impossible to make a "table of normal distribution," and no such thing exists in this world. However, by using Excel, you can conveniently calculate the values and make a table relevant to the normal distribution.

EXERCISE AND ANSWER



EXERCISE

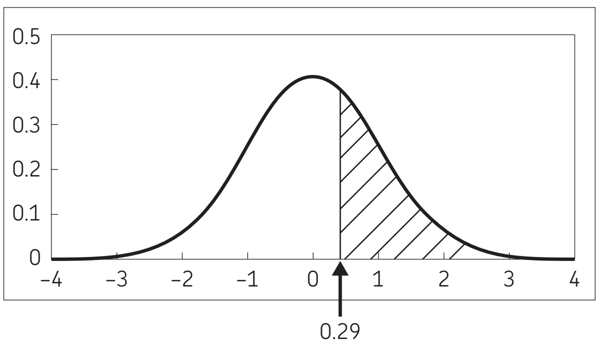
1. Calculate the probability (the shaded area in the graph below) using the table of standard normal distribution on page 93.



2. Calculate the value of χ^2 when there are 2 degrees of freedom and P is 0.05 using the table of chi-square distribution on page 103.

ANSWER

1. Because the standard normal distribution is symmetrical, the probability in question is equal to the probability shown in the graph below.



The probability when z = 0.29 = 0.2 + 0.09 is 0.1141 according to the table of standard normal distribution. Therefore, the probability to be obtained is 0.5 - 0.1141 = 0.3859.

2. The value of χ^2 to be obtained is 5.9915 according to the table of chi-square distribution.

SUMMARY

- Some of the most common probability density functions are:
  - Normal distribution
  - Standard normal distribution
  - Chi-square distribution
  - t distribution
  - F distribution
- The area between the probability density function and the horizontal axis is 1. This area is equivalent to a ratio or a probability.
- By using an Excel function or a table of probabilities for the appropriate distribution, you can calculate:
  - The probability that corresponds to a point on the horizontal axis
  - The point on the horizontal axis that corresponds to the probability