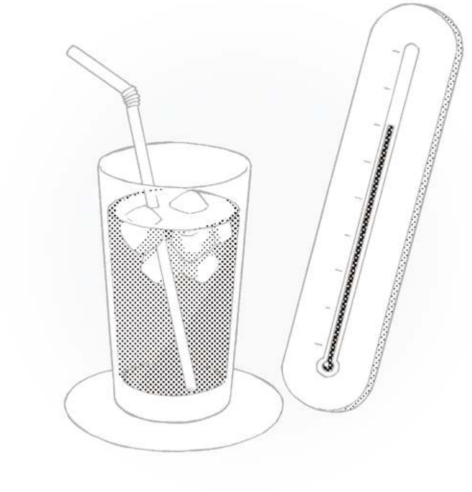
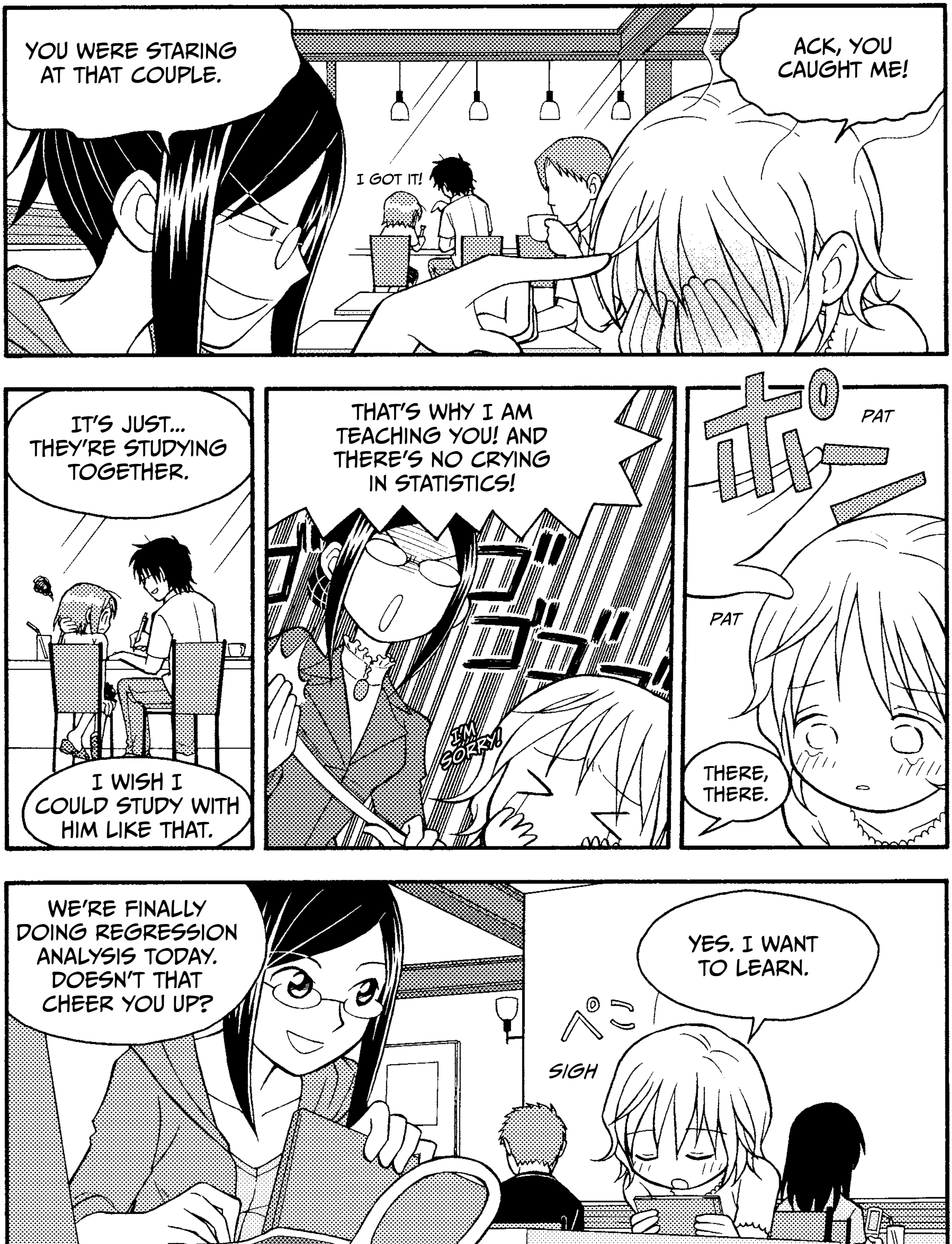
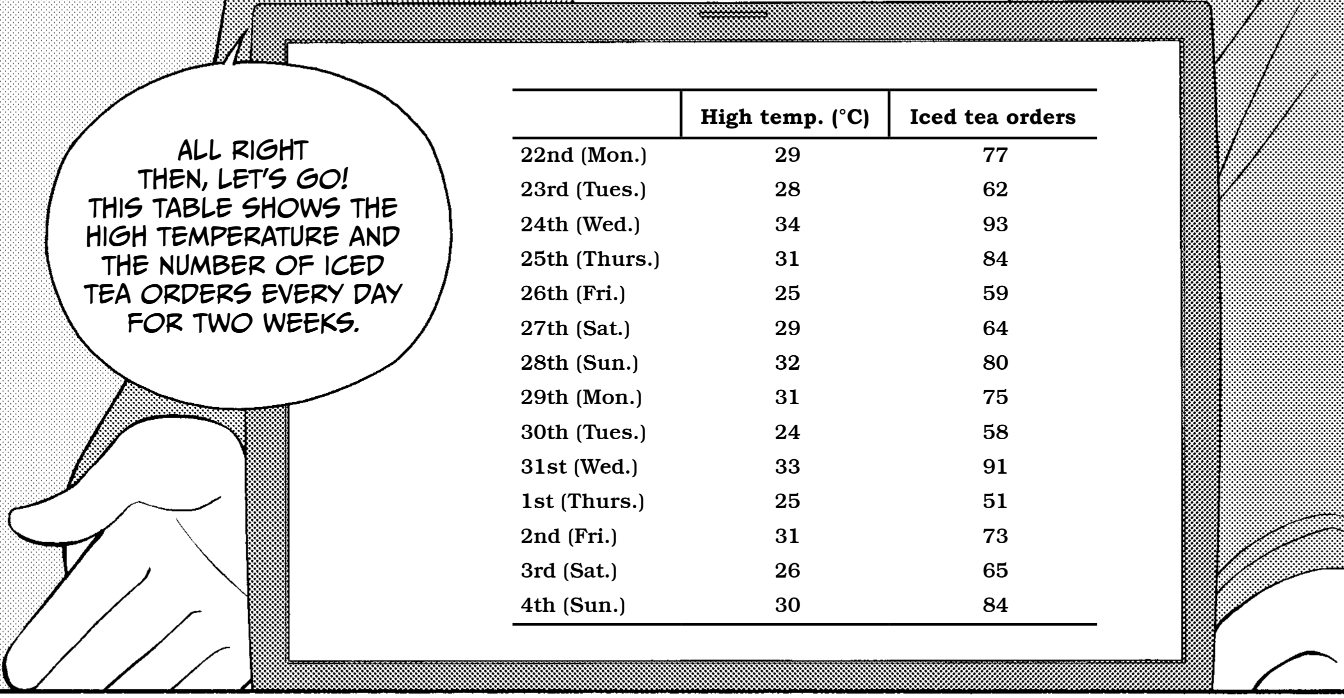
2

SIMPLE REGRESSION ANALYSIS



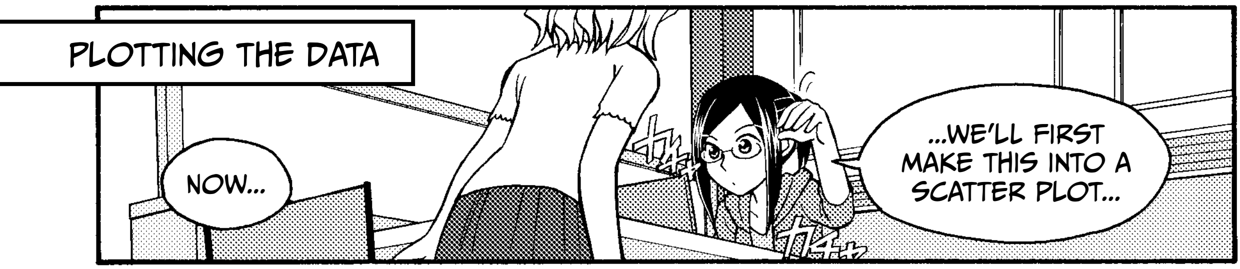

FIRST STEPS
THAT MEANS...
THERE IS A CONNECTION BETWEEN THE TWO, RIGHT?
EXACTLY!
WHERE DID YOU LEARN SO MUCH ABOUT REGRESSION ANALYSIS, MIU?
MIU!
ぼ
はっ
BLINK BLINK
EARTH TO MIU! ARE YOU THERE?

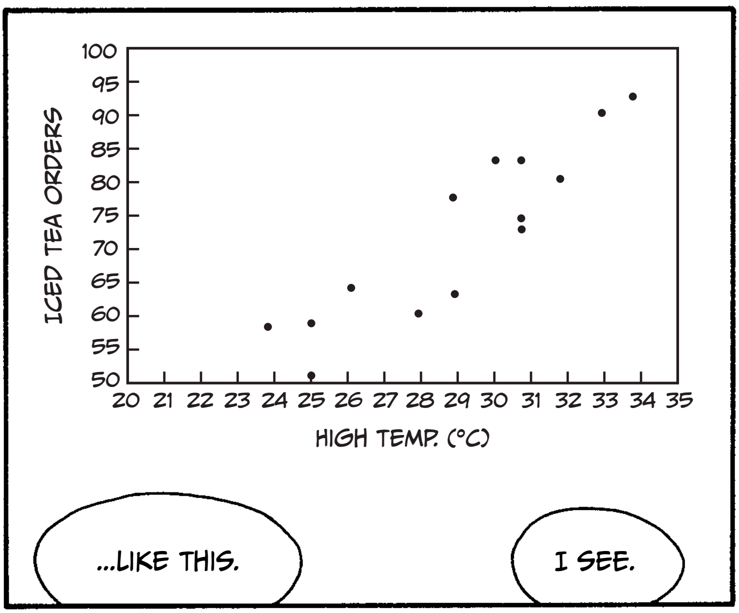

YOU WERE STARING AT THAT COUPLE.
I GOT IT!
ACK, YOU CAUGHT ME!
IT'S JUST... THEY'RE STUDYING TOGETHER.
I WISH I COULD STUDY WITH HIM LIKE THAT.
THAT'S WHY I AM TEACHING YOU! AND THERE'S NO CRYING IN STATISTICS!
I'M SORRY!
ポン
PAT
PAT
THERE, THERE.
WE'RE FINALLY DOING REGRESSION ANALYSIS TODAY. DOESN'T THAT CHEER YOU UP?
YES. I WANT TO LEARN.
ぺこ
SIGH

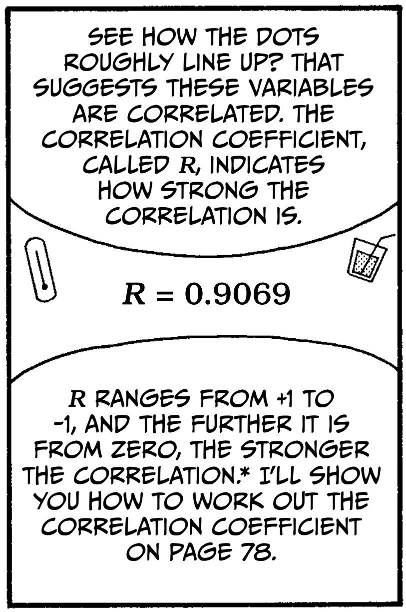


	High temp. (°C)	Iced tea orders
22nd (Mon.)	29	77
23rd (Tues.)	28	62
24th (Wed.)	34	93
25th (Thurs.)	31	84
26th (Fri.)	25	59
27th (Sat.)	29	64
28th (Sun.)	32	80
29th (Mon.)	31	75
30th (Tues.)	24	58
31st (Wed.)	33	91
1st (Thurs.)	25	51
2nd (Fri.)	31	73
3rd (Sat.)	26	65
4th (Sun.)	30	84

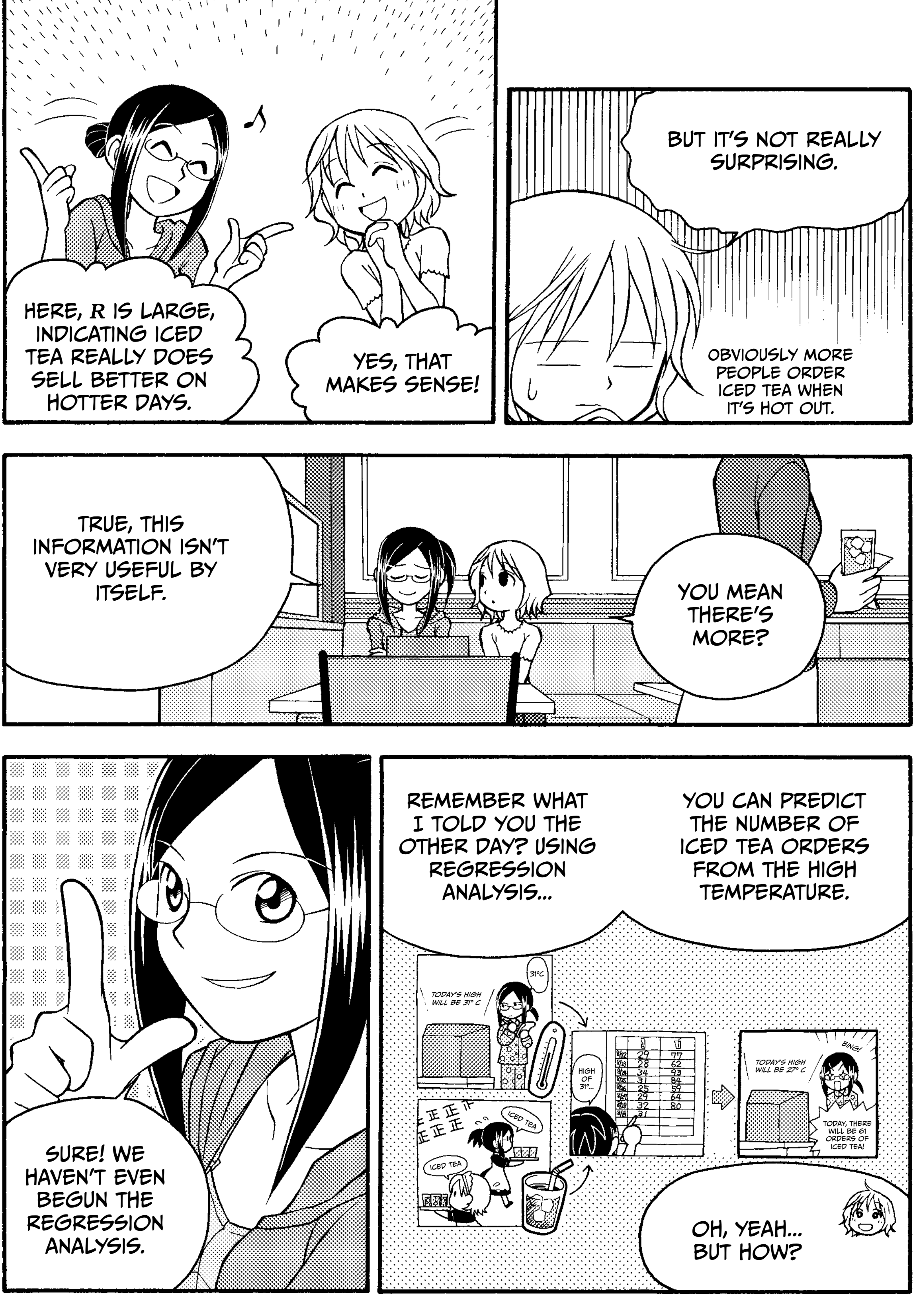
PLOTTING THE DATA



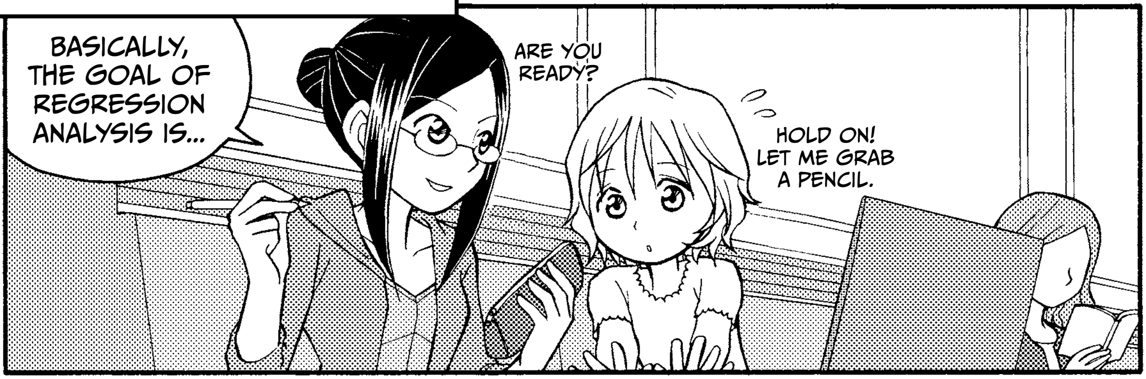


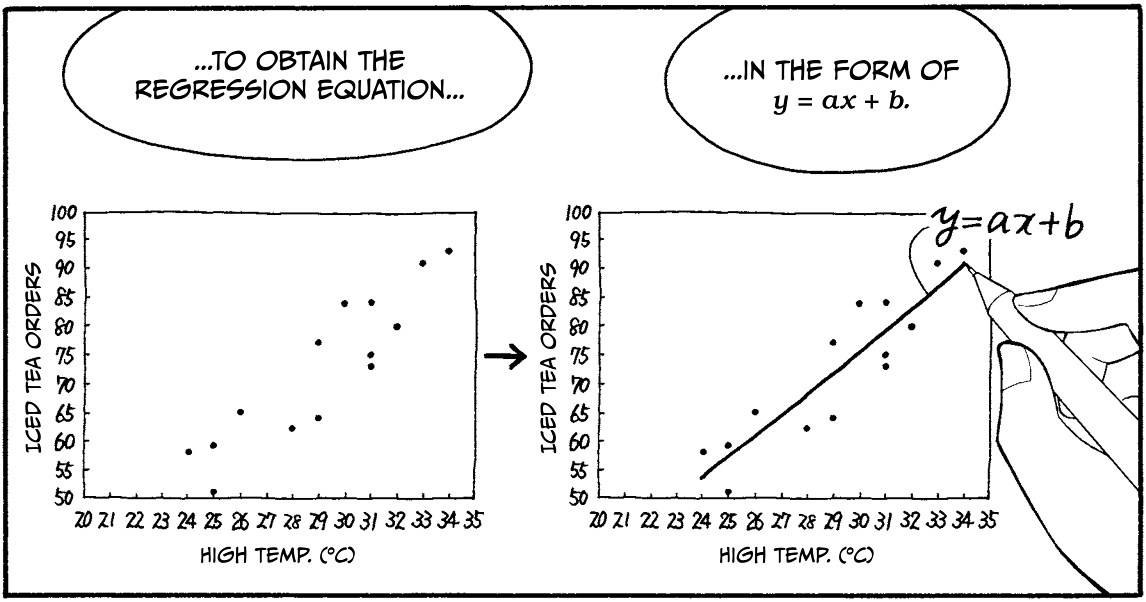


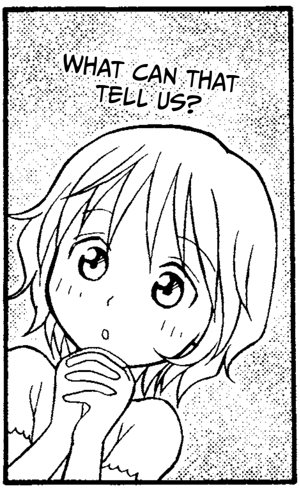
* A POSITIVE R VALUE INDICATES A POSITIVE RELATIONSHIP, MEANING AS x INCREASES, SO DOES y. A NEGATIVE R VALUE MEANS AS THE x VALUE INCREASES, THE y VALUE DECREASES.

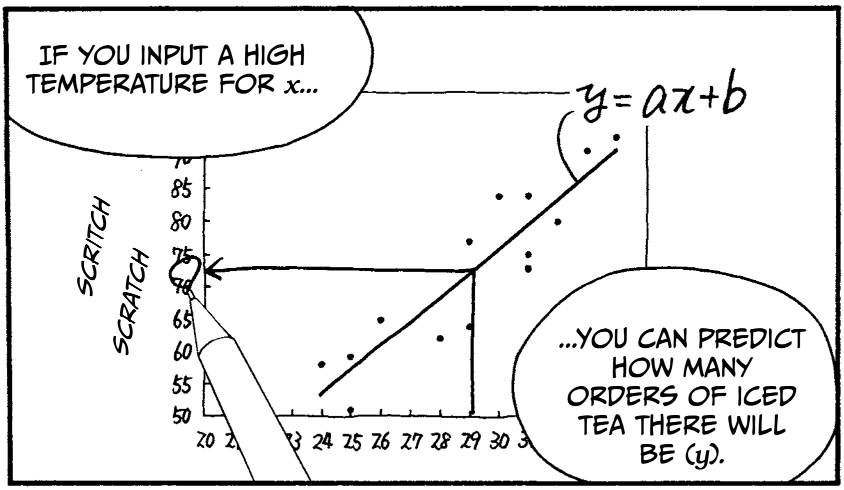

HERE, R IS LARGE, INDICATING ICED TEA REALLY DOES SELL BETTER ON HOTTER DAYS.
YES, THAT MAKES SENSE!
BUT IT'S NOT REALLY SURPRISING.
OBVIOUSLY MORE PEOPLE ORDER ICED TEA WHEN IT'S HOT OUT.
TRUE, THIS INFORMATION ISN'T VERY USEFUL BY ITSELF.
YOU MEAN THERE'S MORE?
SURE! WE HAVEN'T EVEN BEGUN THE REGRESSION ANALYSIS.
REMEMBER WHAT I TOLD YOU THE OTHER DAY? USING REGRESSION ANALYSIS...
YOU CAN PREDICT THE NUMBER OF ICED TEA ORDERS FROM THE HIGH TEMPERATURE.
TODAY'S HIGH WILL BE 31° C
31°C
ICED TEA
ICED TEA
HIGH OF 31°...
TODAY'S HIGH WILL BE 27° C
BING!
TODAY, THERE WILL BE 61 ORDERS OF ICED TEA!
OH, YEAH... BUT HOW?

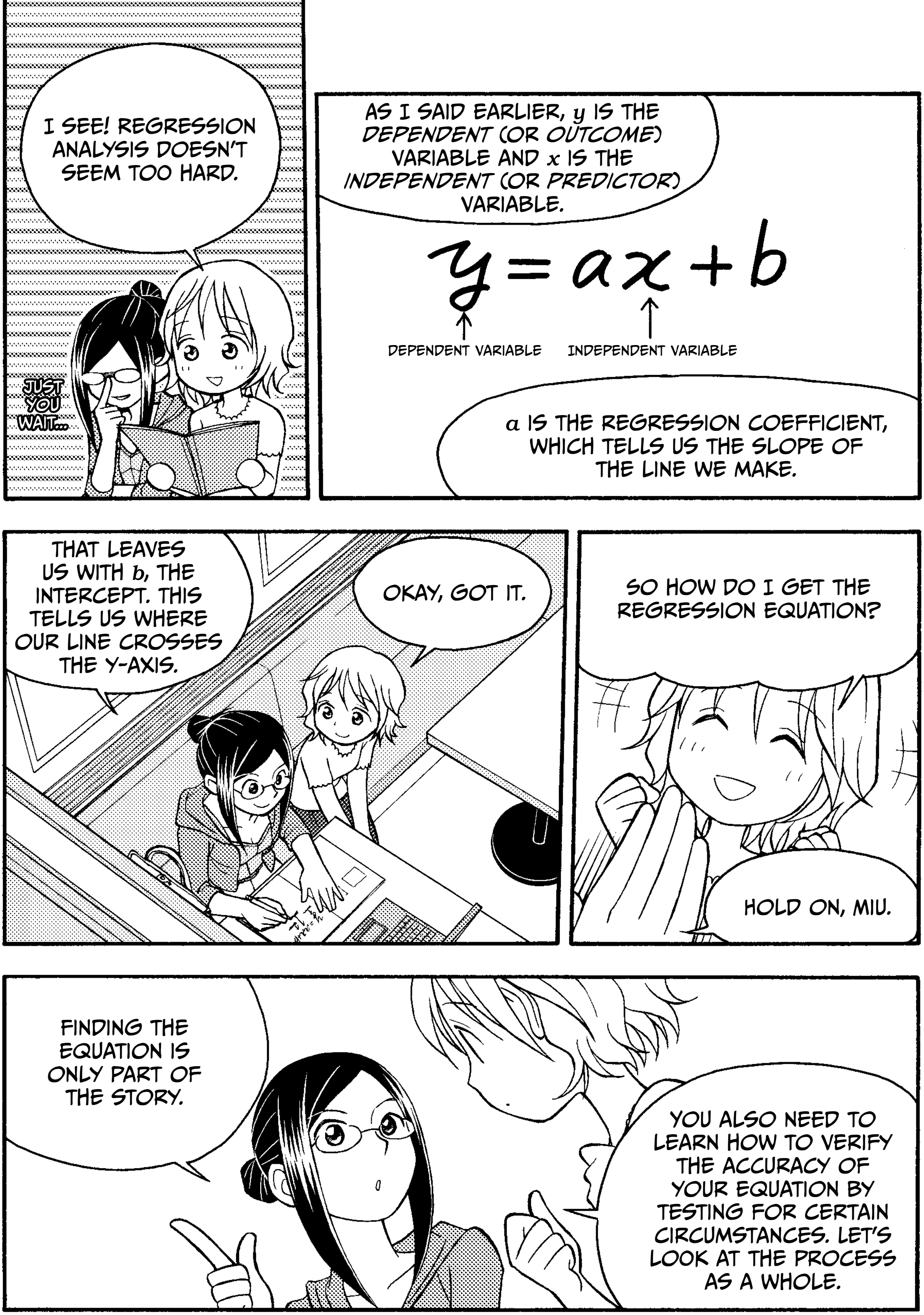
THE REGRESSION EQUATION



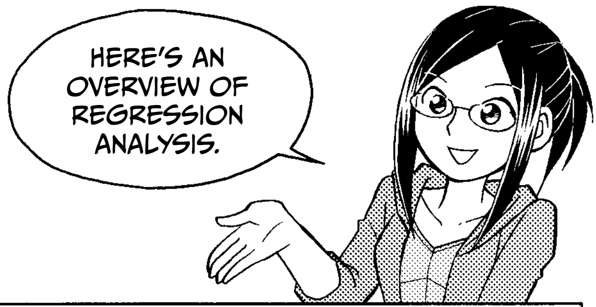






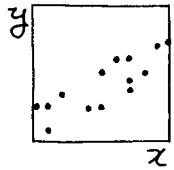

I SEE! REGRESSION ANALYSIS DOESN'T SEEM TOO HARD.
JUST YOU WAIT...
AS I SAID EARLIER, y IS THE *DEPENDENT* (OR *OUTCOME*) VARIABLE AND x IS THE *INDEPENDENT* (OR *PREDICTOR*) VARIABLE.
$y = ax + b$
DEPENDENT VARIABLE
INDEPENDENT VARIABLE
a IS THE REGRESSION COEFFICIENT, WHICH TELLS US THE SLOPE OF THE LINE WE MAKE.
THAT LEAVES US WITH b, THE INTERCEPT. THIS TELLS US WHERE OUR LINE CROSSES THE Y-AXIS.
OKAY, GOT IT.
SO HOW DO I GET THE REGRESSION EQUATION?
HOLD ON, MIU.
FINDING THE EQUATION IS ONLY PART OF THE STORY.
YOU ALSO NEED TO LEARN HOW TO VERIFY THE ACCURACY OF YOUR EQUATION BY TESTING FOR CERTAIN CIRCUMSTANCES. LET'S LOOK AT THE PROCESS AS A WHOLE.

GENERAL REGRESSION ANALYSIS PROCEDURE



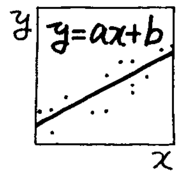
STEP 1

DRAW A SCATTER PLOT OF THE INDEPENDENT VARIABLE VERSUS THE DEPENDENT VARIABLE. IF THE DOTS LINE UP, THE VARIABLES MAY BE CORRELATED.



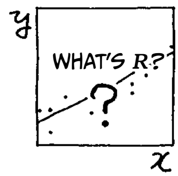
STEP 2

CALCULATE THE REGRESSION EQUATION.



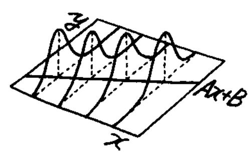
STEP 3

CALCULATE THE CORRELATION COEFFICIENT (*R*) AND ASSESS OUR POPULATION AND ASSUMPTIONS.



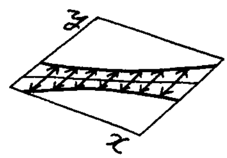
STEP 4

CONDUCT THE ANALYSIS OF VARIANCE.



STEP 5

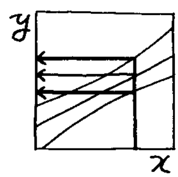
CALCULATE THE CONFIDENCE INTERVALS.

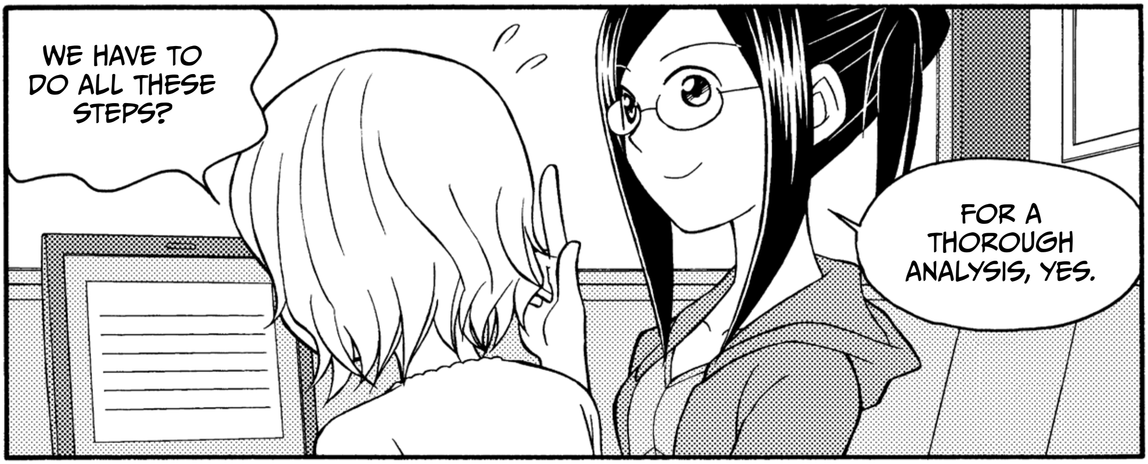


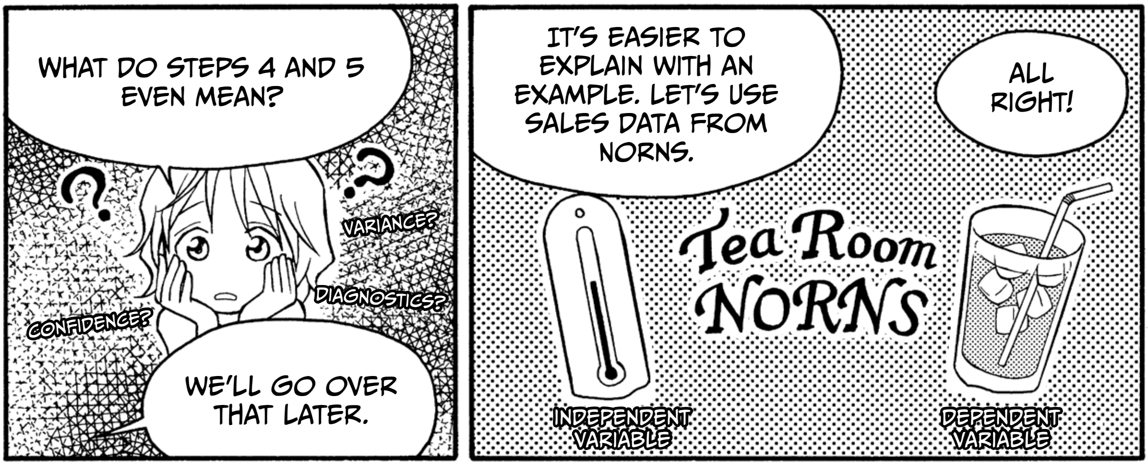
REGRESSION DIAGNOSTICS

STEP 6

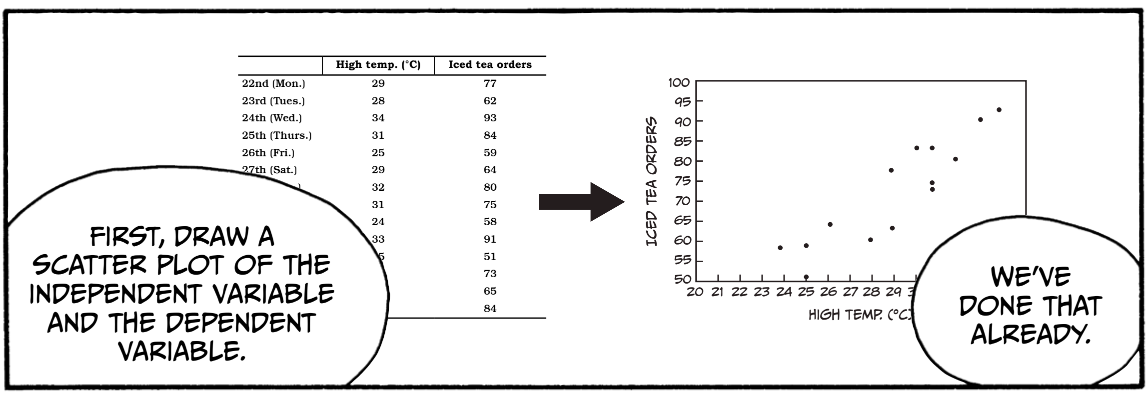
MAKE A PREDICTION!



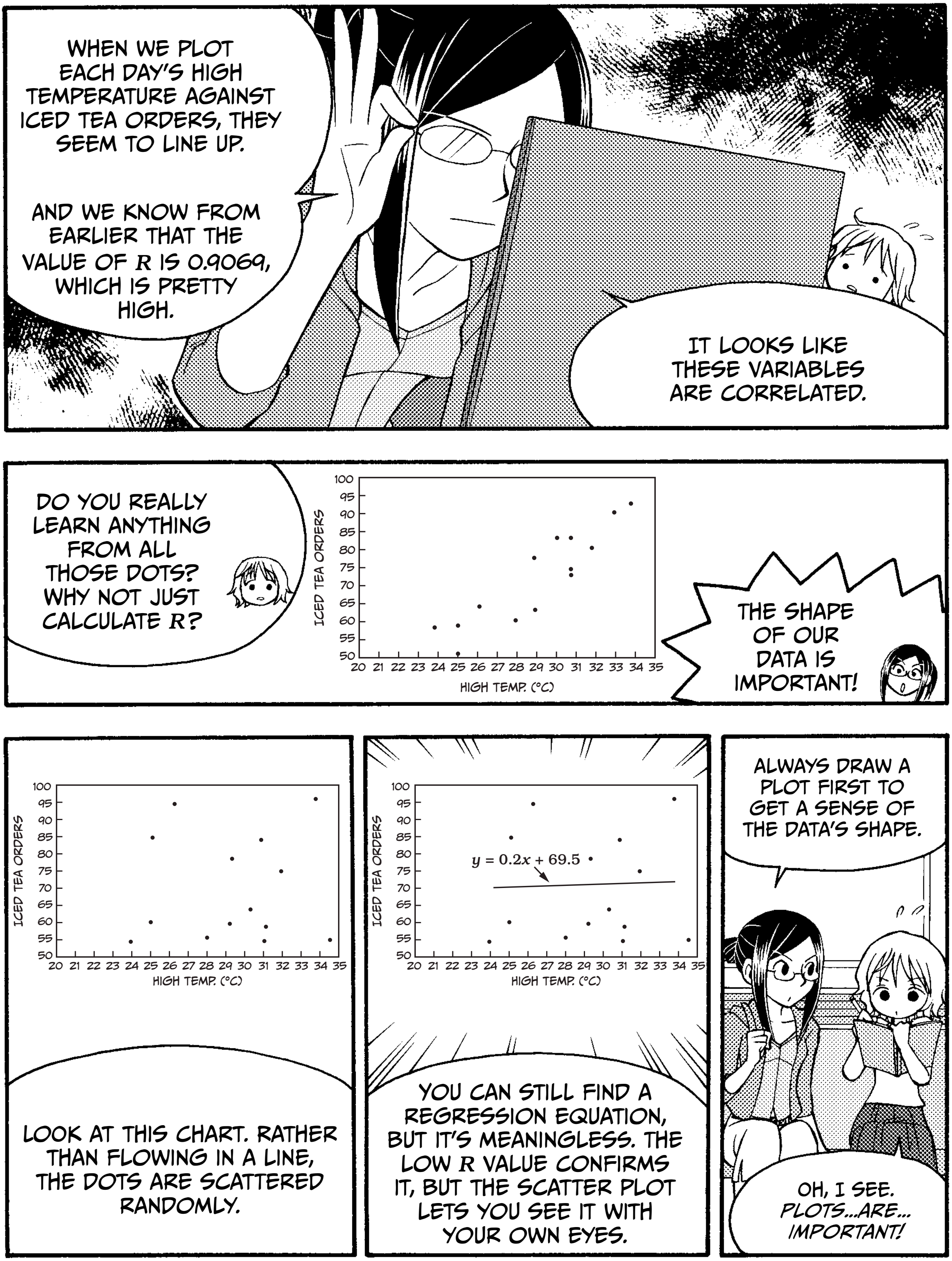




STEP 1: DRAW A SCATTER PLOT OF THE INDEPENDENT VARIABLE VERSUS THE DEPENDENT VARIABLE. IF THE DOTS LINE UP, THE VARIABLES MAY BE CORRELATED.

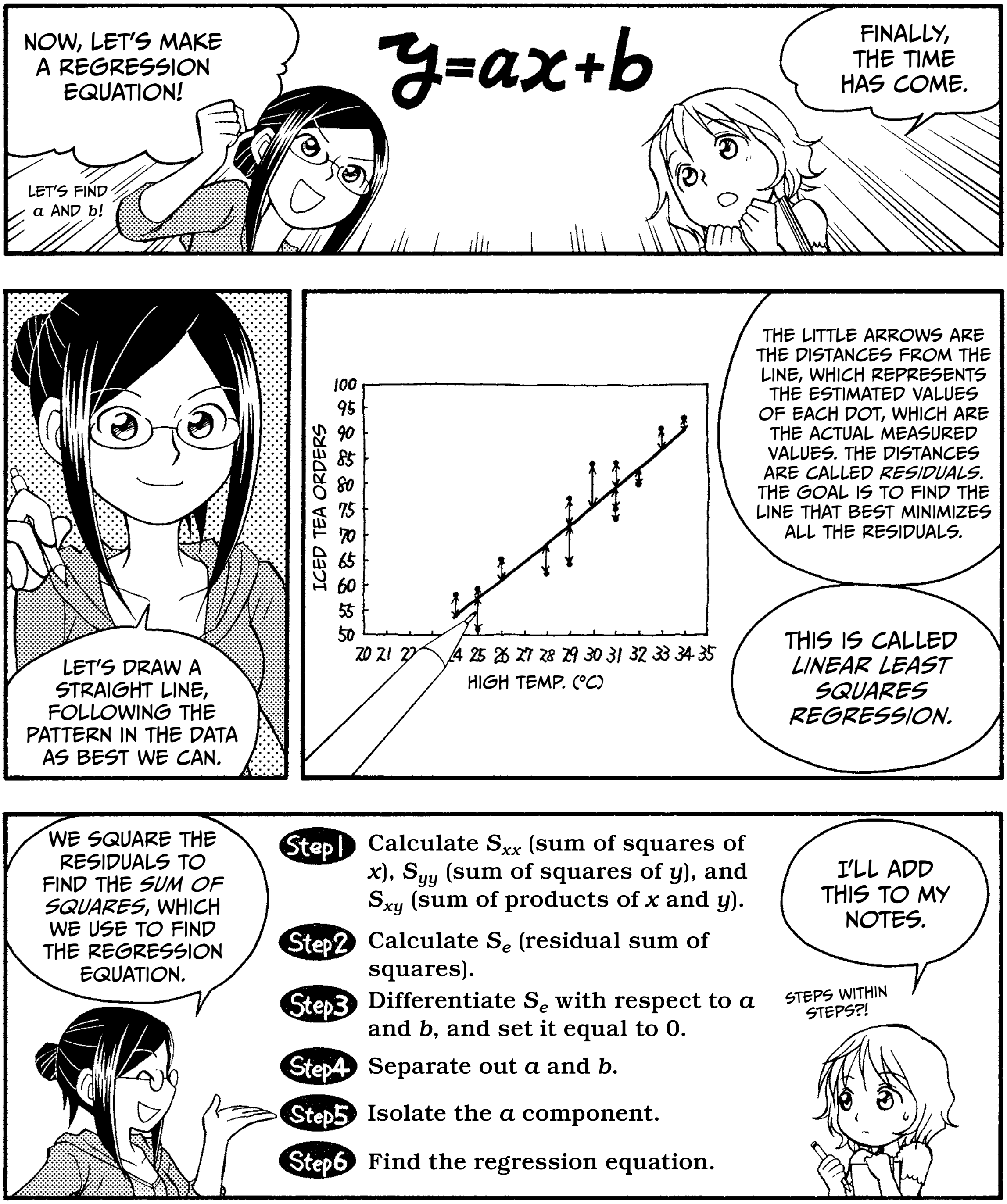


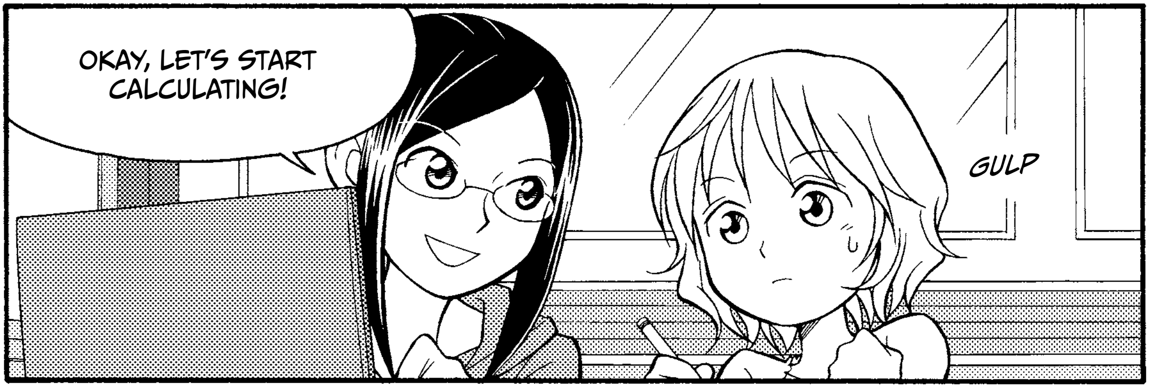
	High temp. (°C)	Iced tea orders
22nd (Mon.)	29	77
23rd (Tues.)	28	62
24th (Wed.)	34	93
25th (Thurs.)	31	84
26th (Fri.)	25	59
27th (Sat.)	29	64
	32	80
	31	75
	24	58
	33	91
		51
		73
		65
		84



WHEN WE PLOT EACH DAY'S HIGH TEMPERATURE AGAINST ICED TEA ORDERS, THEY SEEM TO LINE UP.
AND WE KNOW FROM EARLIER THAT THE VALUE OF R IS 0.9069, WHICH IS PRETTY HIGH.
IT LOOKS LIKE THESE VARIABLES ARE CORRELATED.
DO YOU REALLY LEARN ANYTHING FROM ALL THOSE DOTS? WHY NOT JUST CALCULATE R?
ICED TEA ORDERS
HIGH TEMP. (°C)
THE SHAPE OF OUR DATA IS IMPORTANT!
ICED TEA ORDERS
HIGH TEMP. (°C)
LOOK AT THIS CHART. RATHER THAN FLOWING IN A LINE, THE DOTS ARE SCATTERED RANDOMLY.
y = 0.2x + 69.5
ICED TEA ORDERS
HIGH TEMP. (°C)
YOU CAN STILL FIND A REGRESSION EQUATION, BUT IT'S MEANINGLESS. THE LOW R VALUE CONFIRMS IT, BUT THE SCATTER PLOT LETS YOU SEE IT WITH YOUR OWN EYES.
ALWAYS DRAW A PLOT FIRST TO GET A SENSE OF THE DATA'S SHAPE.
OH, I SEE. PLOTS...ARE... IMPORTANT!

STEP 2: CALCULATE THE REGRESSION EQUATION.





Step 1 Find

- The sum of squares of x, S_{xx}: $(x-\bar{x})^2$
- The sum of squares of y, S_{yy}: $(y-\bar{y})^2$
- The sum of products of x and y, S_{xy}: $(x-\bar{x})(y-\bar{y})$

Note: The bar over a variable (like $\bar{x}$) is a notation that means *average*. We can call this variable x-bar.

	High temp. in °C x	Iced tea orders y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
22nd (Mon.)	29	77	–0.1	4.4	0.0	19.6	–0.6
23rd (Tues.)	28	62	–1.1	–10.6	1.3	111.8	12.1
24th (Wed.)	34	93	4.9	20.4	23.6	417.3	99.2
25th (Thurs.)	31	84	1.9	11.4	3.4	130.6	21.2
26th (Fri.)	25	59	–4.1	–13.6	17.2	184.2	56.2
27th (Sat.)	29	64	–0.1	–8.6	0.0	73.5	1.2
28th (Sun.)	32	80	2.9	7.4	8.2	55.2	21.2
29th (Mon.)	31	75	1.9	2.4	3.4	5.9	4.5
30th (Tues.)	24	58	–5.1	–14.6	26.4	212.3	74.9
31st (Wed.)	33	91	3.9	18.4	14.9	339.6	71.1
1st (Thurs.)	25	51	–4.1	–21.6	17.2	465.3	89.4
2nd (Fri.)	31	73	1.9	0.4	3.4	0.2	0.8
3rd (Sat.)	26	65	–3.1	–7.6	9.9	57.8	23.8
4th (Sun.)	30	84	0.9	11.4	0.7	130.6	9.8
Sum	408	1016	0	0	129.7	2203.4	484.9
Average	29.1	72.6					
	↓ $\bar{x}$	↓ $\bar{y}$			↓ S_{xx}	↓ S_{yy}	↓ S_{xy}

* Some of the figures in this chapter are rounded for the sake of printing, but calculations are done using the full, unrounded values resulting from the raw data unless otherwise stated.

Step2 Find the residual sum of squares, S_e.

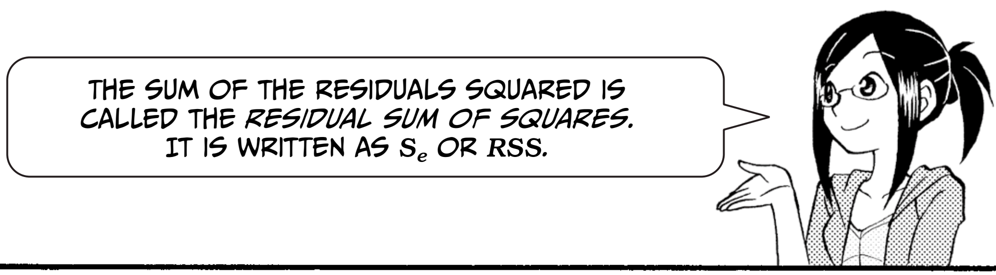
- y is the observed value.
- $\hat{y}$ is the the estimated value based on our regression equation.
- $y - \hat{y}$ is called the residual and is written as e.

Note: The caret in $\hat{y}$ is affectionately called a *hat*, so we call this parameter estimate y-hat.

	High temp. in °C x	Actual iced tea orders y	Predicted iced tea orders $\hat{y} = ax + b$	Residuals (e) $y - \hat{y}$	Squared residuals $(y - \hat{y})^2$
22nd (Mon.)	29	77	$a \times 29 + b$	$77 - (a \times 29 + b)$	$[77 - (a \times 29 + b)]^2$
23rd (Tues.)	28	62	$a \times 28 + b$	$62 - (a \times 28 + b)$	$[62 - (a \times 28 + b)]^2$
24th (Wed.)	34	93	$a \times 34 + b$	$93 - (a \times 34 + b)$	$[93 - (a \times 34 + b)]^2$
25th (Thurs.)	31	84	$a \times 31 + b$	$84 - (a \times 31 + b)$	$[84 - (a \times 31 + b)]^2$
26th (Fri.)	25	59	$a \times 25 + b$	$59 - (a \times 25 + b)$	$[59 - (a \times 25 + b)]^2$
27th (Sat.)	29	64	$a \times 29 + b$	$64 - (a \times 29 + b)$	$[64 - (a \times 29 + b)]^2$
28th (Sun.)	32	80	$a \times 32 + b$	$80 - (a \times 32 + b)$	$[80 - (a \times 32 + b)]^2$
29th (Mon.)	31	75	$a \times 31 + b$	$75 - (a \times 31 + b)$	$[75 - (a \times 31 + b)]^2$
30th (Tues.)	24	58	$a \times 24 + b$	$58 - (a \times 24 + b)$	$[58 - (a \times 24 + b)]^2$
31st (Wed.)	33	91	$a \times 33 + b$	$91 - (a \times 33 + b)$	$[91 - (a \times 33 + b)]^2$
1st (Thurs.)	25	51	$a \times 25 + b$	$51 - (a \times 25 + b)$	$[51 - (a \times 25 + b)]^2$
2nd (Fri.)	31	73	$a \times 31 + b$	$73 - (a \times 31 + b)$	$[73 - (a \times 31 + b)]^2$
3rd (Sat.)	26	65	$a \times 26 + b$	$65 - (a \times 26 + b)$	$[65 - (a \times 26 + b)]^2$
4th (Sun.)	30	84	$a \times 30 + b$	$84 - (a \times 30 + b)$	$[84 - (a \times 30 + b)]^2$
Sum	408	1016	$408a + 14b$	$1016 - (408a + 14b)$	S_e
Average	29.1	72.6	$29.1a + b$ $= \bar{x}a + b$	$72.6 - (29.1a + b)$ $= \bar{y} - (\bar{x}a + b)$	$= \dfrac{S_e}{14}$

$\bar{x}$ $\bar{y}$

$$S_e = \left[77 - (a \times 29 + b)\right]^2 + \cdots + \left[84 - (a \times 30 + b)\right]^2$$



Step3 Differentiate S_e with respect to a and b, and set it equal to 0. When differentiating $y=(ax+b)^{n-1}$ with respect to x, the result is

$$\frac{dy}{dx}=n(ax+b)^{n-1}\times a.$$

- Differentiate with respect to a.

$$\frac{dS_e}{da}=2\left[77-(29a+b)\right]\times(-29)+\cdots+2\left[84-(30a+b)\right]\times(-30)=0 \quad ❶$$

- Differentiate with respect to b.

$$\frac{dS_e}{db}=2\left[77-(29a+b)\right]\times(-1)+\cdots+2\left[84-(30a+b)\right]\times(-1)=0 \quad ❷$$

Step4 Rearrange ❶ and ❷ from the previous step.

Rearrange ❶.

$$2\left[77-(29a+b)\right]\times(-29)+\cdots+2\left[84-(30a+b)\right]\times(-30)=0$$

$$\left[77-(29a+b)\right]\times(-29)+\cdots+\left[84-(30a+b)\right]\times(-30)=0$$

DIVIDE BOTH SIDES BY 2.

$$29\left[(29a+b)-77\right]+\cdots+30\left[(30a+b)-84\right]=0$$

MULTIPLY BY -1.

$$(29\times29a+29\times b-29\times77)+\cdots+(30\times30a+30\times b-30\times84)=0$$

MULTIPLY.

$$(29^2+\cdots+30^2)a+(29+\cdots+30)b-(29\times77+\cdots+30\times84)=0$$

SEPARATE OUT a AND b.

❸

Rearrange ❷.

$$2\left[77-(29a+b)\right]\times(-1)+\cdots+2\left[84-(30a+b)\right]\times(-1)=0$$

$$\left[77-(29a+b)\right]\times(-1)+\cdots+\left[84-(30a+b)\right]\times(-1)=0$$

DIVIDE BOTH SIDES BY 2.

$$\left[(29a+b)-77\right]+\cdots+\left[(30a+b)-84\right]=0$$

MULTIPLY BY -1.

$$(29+\cdots+30)a+\underbrace{b+\cdots+b}_{14}-(77+\cdots+84)=0$$

SEPARATE OUT a AND b.

$$(29+\cdots+30)a+14b-(77+\cdots+84)=0$$

$$14b=(77+\cdots+84)-(29+\cdots+30)a$$

$$b=\frac{77+\cdots+84}{14}-\frac{29+\cdots+30}{14}a$$

SUBTRACT $14b$ FROM BOTH SIDES AND MULTIPLY BY -1.

❹ $b=\bar{y}-\bar{x}a$

ISOLATE b ON THE LEFT SIDE OF THE EQUATION.

❺

THE COMPONENTS IN ❹ ARE THE AVERAGES OF y AND x.

Step5 **Plug the value of b found in ❹ into line ❸ (❸ and ❹ are the results from Step 4).**

$$❸\quad (29^2+\cdots+30^2)a+(29+\cdots+30)\left(\frac{77+\cdots+84}{14}-\underbrace{\frac{29+\cdots+30}{14}a}_{❹}\right)-(29\times 77+\cdots+30\times 84)=0$$

NOW a IS THE ONLY VARIABLE.

$$(29^2+\cdots+30^2)a+\frac{(29+\cdots+30)(77+\cdots+84)}{14}-\frac{(29+\cdots+30)^2}{14}a-(29\times 77+\cdots+30\times 84)=0$$

$$\left[(29^2+\cdots+30^2)-\frac{(29+\cdots+30)^2}{14}\right]a+\frac{(29+\cdots+30)(77+\cdots+84)}{14}-(29\times 77+\cdots+30\times 84)=0$$

COMBINE THE a TERMS.

$$\left[(29^2+\cdots+30^2)-\frac{(29+\cdots+30)^2}{14}\right]a=(29\times 77+\cdots+30\times 84)-\frac{(29+\cdots+30)(77+\cdots+84)}{14}$$

TRANSPOSE.

Rearrange the left side of the equation.

$$(29^2+\cdots+30^2)-\frac{(29+\cdots+30)^2}{14}$$

$$=(29^2+\cdots+30^2)-2\times\frac{(29+\cdots+30)^2}{14}+\frac{(29+\cdots+30)^2}{14}$$

WE ADD AND SUBTRACT $\frac{(29+\cdots+30)^2}{14}$.

$$=(29^2+\cdots+30^2)-2\times(29+\cdots+30)\times\frac{29+\cdots+30}{14}+\left(\frac{29+\cdots+30}{14}\right)^2\times 14$$

THE LAST TERM IS MULTIPLIED BY $\frac{14}{14}$.

$$=(29^2+\cdots+30^2)-2\times(29+\cdots+30)\times\bar{x}+(\bar{x})^2\times 14$$

$\bar{x}=\frac{29+\cdots+30}{14}$

$$=(29^2+\cdots+30^2)-2\times(29+\cdots+30)\times\bar{x}+\underbrace{(\bar{x})^2+\cdots+(\bar{x})^2}_{14}$$

$$=\left[29^2-2\times 29\times\bar{x}+(\bar{x})^2\right]+\cdots+\left[30^2-2\times 30\times\bar{x}+(\bar{x})^2\right]$$

$$=(29-\bar{x})^2+\cdots+(30-\bar{x})^2$$

$$=S_{xx}$$

Rearrange the right side of the equation.

$$(29\times 77+\cdots+30\times 84)-\frac{(29+\cdots+30)(77+\cdots+84)}{14}$$

$$=(29\times 77+\cdots+30\times 84)-\frac{29+\cdots+30}{14}\times\frac{77+\cdots+84}{14}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-\bar{x}\times\bar{y}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-\bar{x}\times\bar{y}\times 14-\bar{x}\times\bar{y}\times 14+\bar{x}\times\bar{y}\times 14$$

WE ADD AND SUBTRACT $\bar{x}\times\bar{y}\times 14$.

$$=(29\times 77+\cdots+30\times 84)-\frac{29+\cdots+30}{14}\times\bar{y}\times 14-\bar{x}\times\frac{77+\cdots+84}{14}\times 14+\bar{x}\times\bar{y}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-(29+\cdots+30)\bar{y}-\bar{x}(77+\cdots+84)+\bar{x}\times\bar{y}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-(29+\cdots+30)\bar{y}-(77+\cdots+84)\bar{x}+\underbrace{\bar{x}\times\bar{y}+\cdots+\bar{x}\times\bar{y}}_{14}$$

$$=(29-\bar{x})(77-\bar{y})+\cdots+(30-\bar{x})(84-\bar{y})$$

$$=S_{xy}$$

$$S_{xx}a=S_{xy}$$

$$❻\quad a=\frac{S_{xy}}{S_{xx}}$$

ISOLATE a ON THE LEFT SIDE OF THE EQUATION.

Step 6 Calculate the regression equation.

From ❻ in Step 5, $a = \dfrac{S_{xy}}{S_{xx}}$. From ❺ in Step 4, $b = \bar{y} - \bar{x}a$.

If we plug in the values we calculated in Step 1,

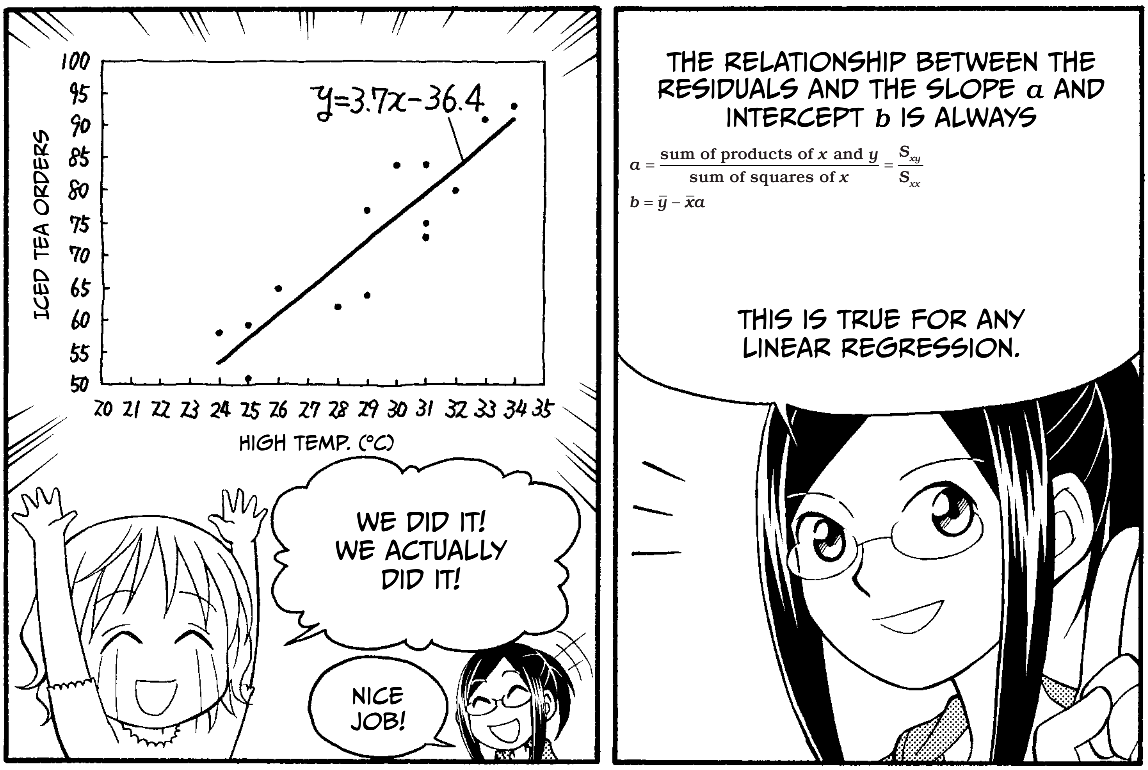
$$\begin{cases} a = \dfrac{S_{xx}}{S_{xy}} = \dfrac{484.9}{129.7} = 3.7 \\ b = \bar{y} - \bar{x}a = 72.6 - 29.1 \times 3.7 = -36.4 \end{cases}$$

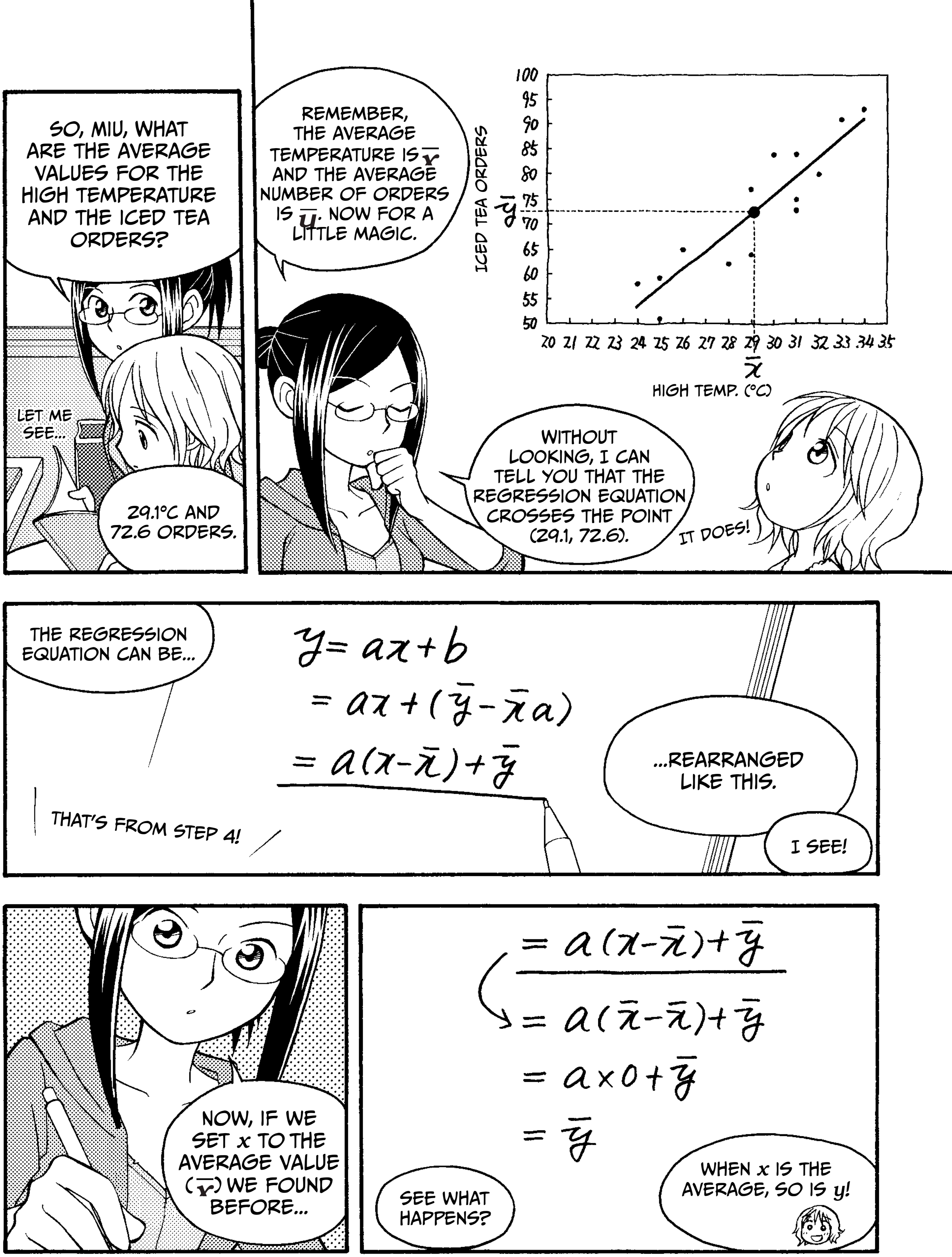
then the regression equation is

$$y = 3.7x - 36.4.$$

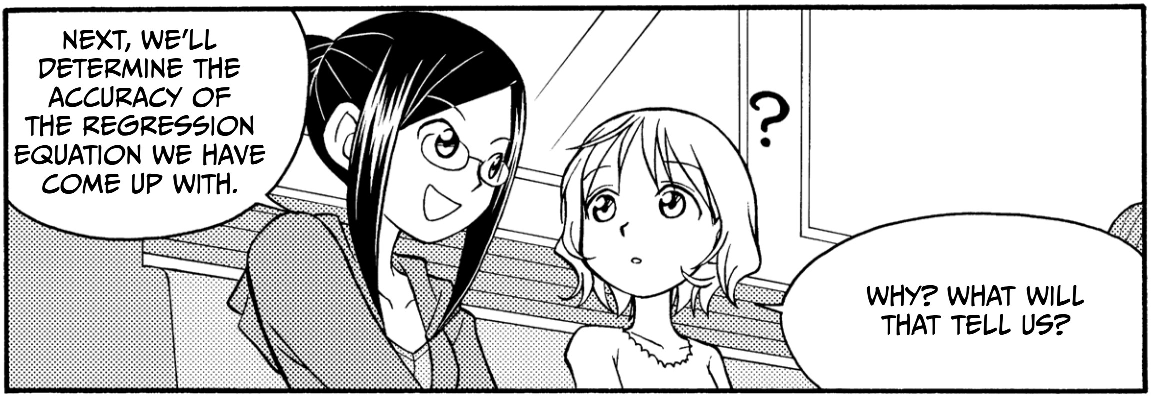
It's that simple!

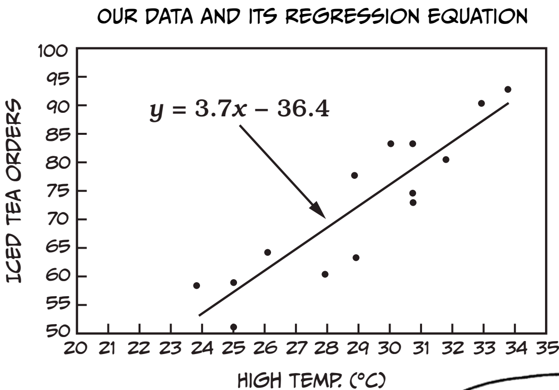
Note: The values shown are rounded for the sake of printing, but the result (36.4) was calculated using the full, unrounded values.

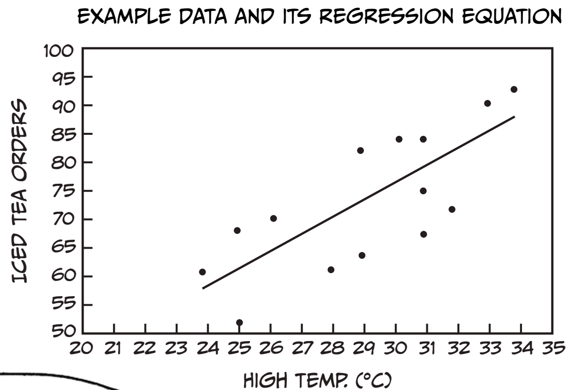


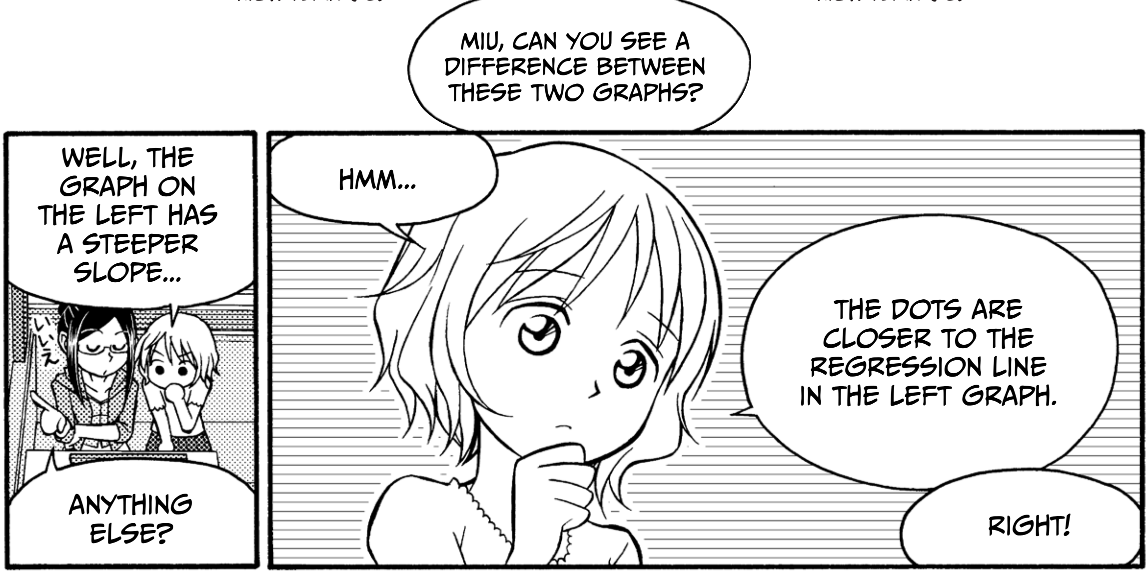

SO, MIU, WHAT ARE THE AVERAGE VALUES FOR THE HIGH TEMPERATURE AND THE ICED TEA ORDERS?
LET ME SEE...
29.1°C AND 72.6 ORDERS.
REMEMBER, THE AVERAGE TEMPERATURE IS $\bar{x}$ AND THE AVERAGE NUMBER OF ORDERS IS $\bar{y}$. NOW FOR A LITTLE MAGIC.
ICED TEA ORDERS
$\bar{y}$
100 95 90 85 80 75 70 65 60 55 50
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
$\bar{x}$
HIGH TEMP. (°C)
WITHOUT LOOKING, I CAN TELL YOU THAT THE REGRESSION EQUATION CROSSES THE POINT (29.1, 72.6).
IT DOES!
THE REGRESSION EQUATION CAN BE...
$y = ax + b$
$= ax + (\bar{y} - \bar{x}a)$
$= a(x - \bar{x}) + \bar{y}$
THAT'S FROM STEP 4!
...REARRANGED LIKE THIS.
I SEE!
NOW, IF WE SET x TO THE AVERAGE VALUE ($\bar{x}$) WE FOUND BEFORE...
$= a(x - \bar{x}) + \bar{y}$
$= a(\bar{x} - \bar{x}) + \bar{y}$
$= a \times 0 + \bar{y}$
$= \bar{y}$
SEE WHAT HAPPENS?
WHEN x IS THE AVERAGE, SO IS y!

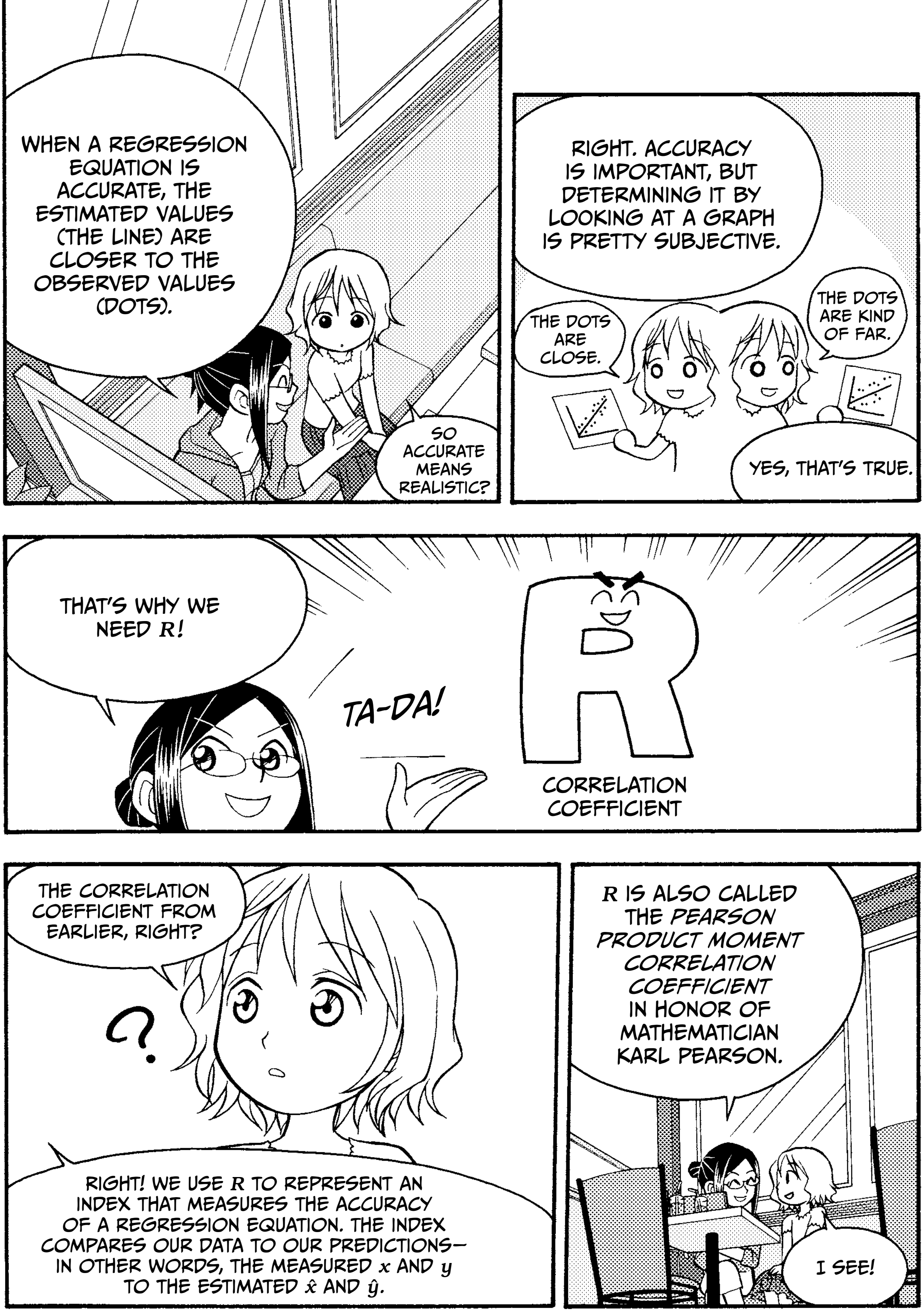
Step 3: Calculate the correlation coefficient (*R*) and assess our population and assumptions.

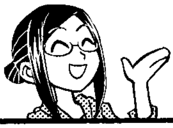






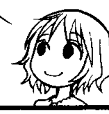


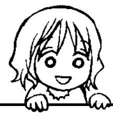

WHEN A REGRESSION EQUATION IS ACCURATE, THE ESTIMATED VALUES (THE LINE) ARE CLOSER TO THE OBSERVED VALUES (DOTS).
SO ACCURATE MEANS REALISTIC?
RIGHT. ACCURACY IS IMPORTANT, BUT DETERMINING IT BY LOOKING AT A GRAPH IS PRETTY SUBJECTIVE.
THE DOTS ARE CLOSE.
THE DOTS ARE KIND OF FAR.
YES, THAT'S TRUE.
THAT'S WHY WE NEED R!
TA-DA!
R
CORRELATION COEFFICIENT
THE CORRELATION COEFFICIENT FROM EARLIER, RIGHT?
?
RIGHT! WE USE R TO REPRESENT AN INDEX THAT MEASURES THE ACCURACY OF A REGRESSION EQUATION. THE INDEX COMPARES OUR DATA TO OUR PREDICTIONS—IN OTHER WORDS, THE MEASURED x AND y TO THE ESTIMATED $\hat{x}$ AND $\hat{y}$.
R IS ALSO CALLED THE *PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT* IN HONOR OF MATHEMATICIAN KARL PEARSON.
I SEE!



$$R = \frac{\text{sum of products } y \text{ and } \hat{y}}{\sqrt{\text{sum of squares of } y \times \text{sum of squares of } \hat{y}}} = \frac{S_{y\hat{y}}}{\sqrt{S_{yy} \times S_{\hat{y}\hat{y}}}}$$

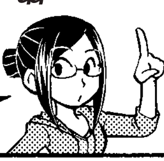
$$= \frac{1812.3}{\sqrt{2203.4 \times 1812.3}} = 0.9069$$

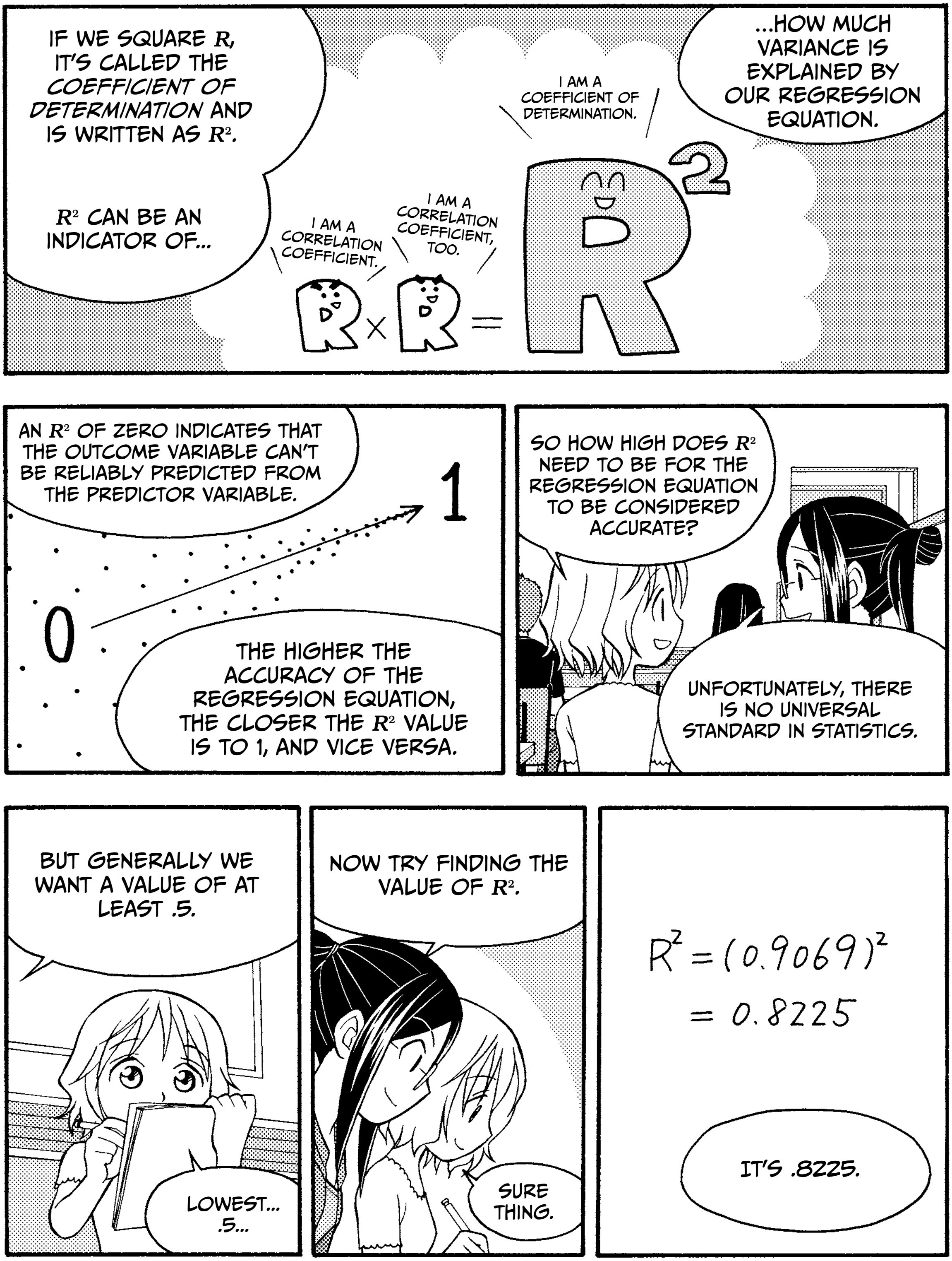


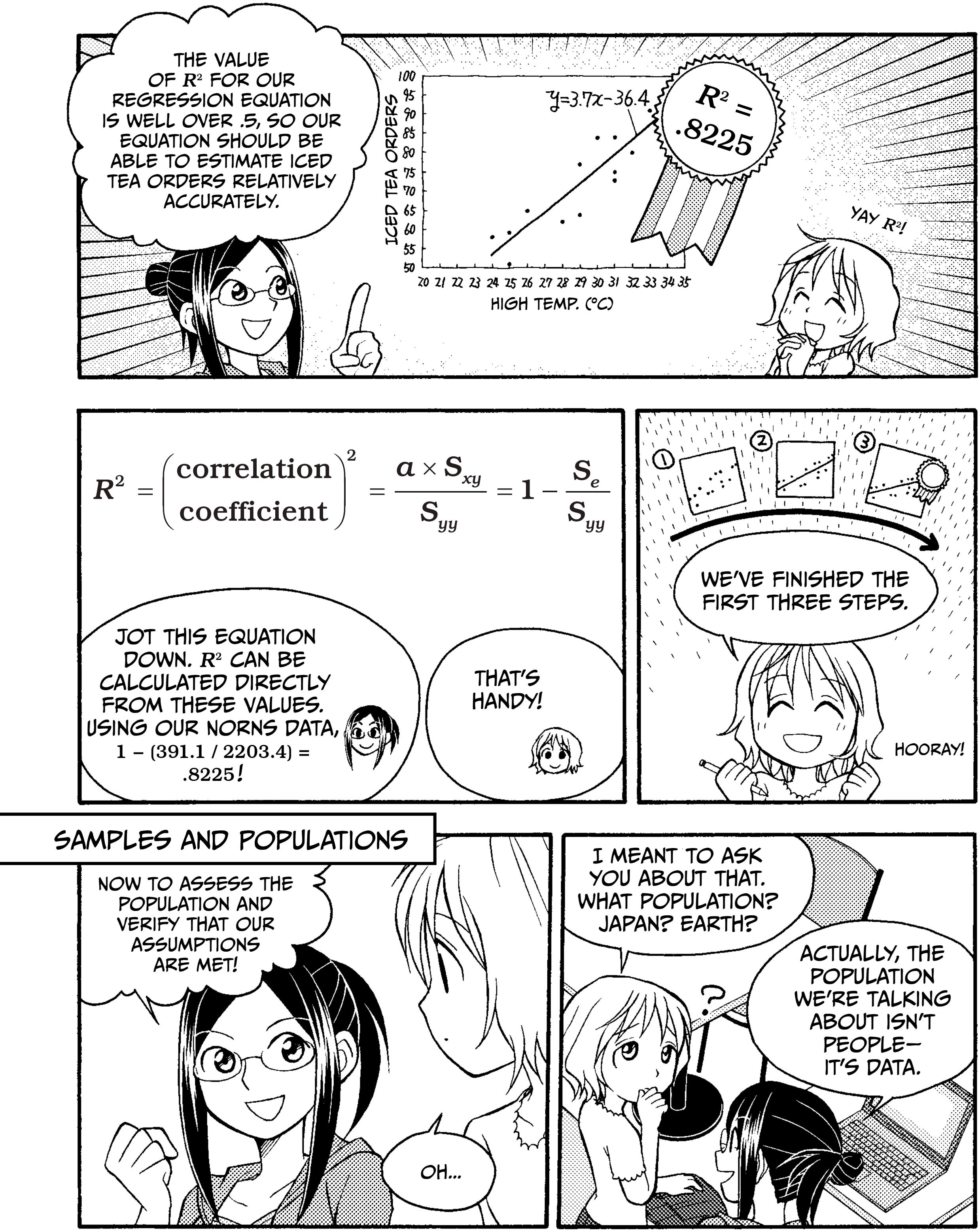


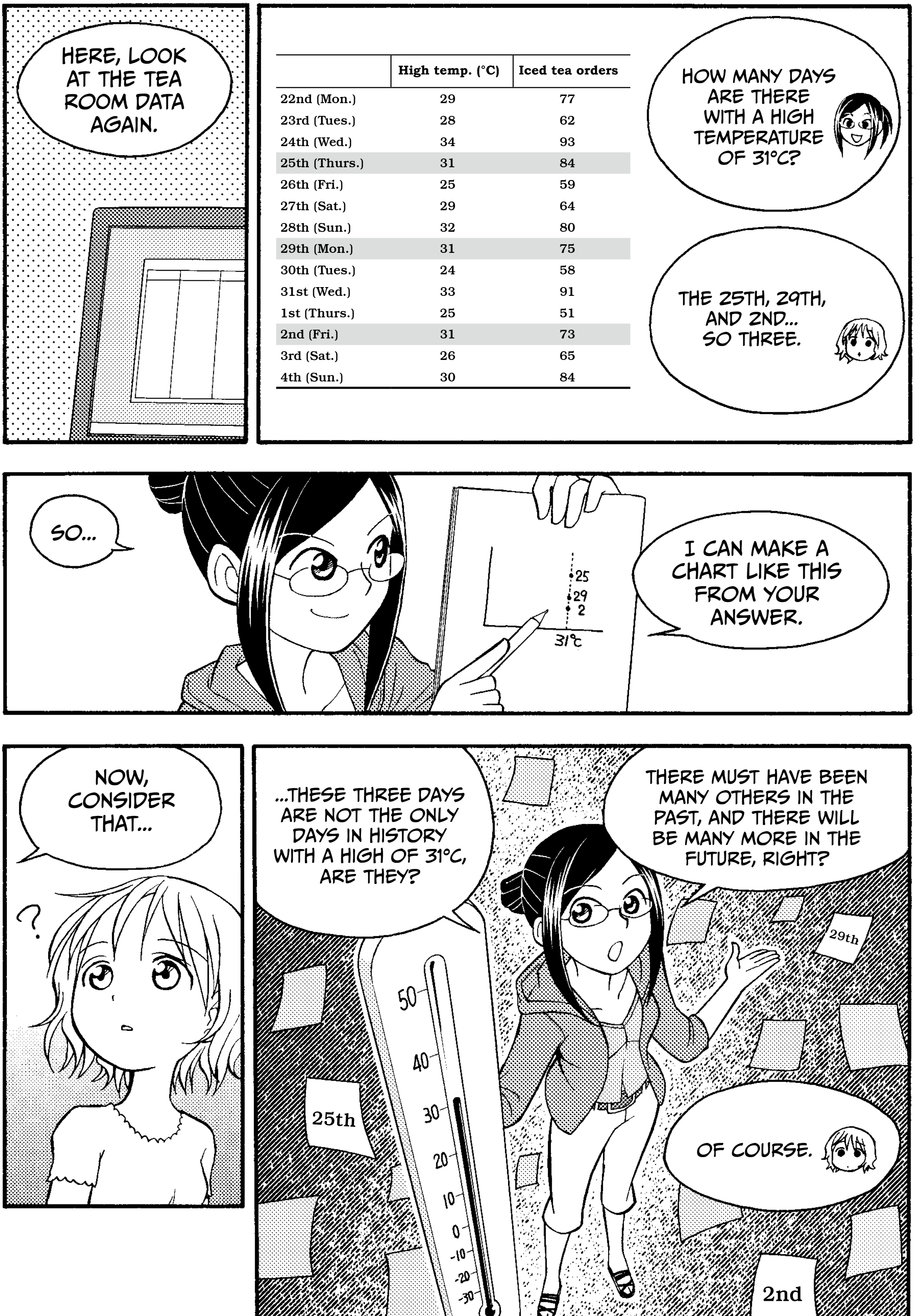
REGRESSION FUNCTION!

	Actual values y	Estimated values $\hat{y} = 3.7x - 36.4$	$y - \bar{y}$	$\hat{y} - \bar{\hat{y}}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{\hat{y}})^2$	$(y - \bar{y})(\hat{y} - \bar{\hat{y}})$	$(y - \hat{y})^2$
22nd (Mon.)	77	72.0	4.4	–0.5	19.6	0.3	–2.4	24.6
23rd (Tues.)	62	68.3	–10.6	–4.3	111.8	18.2	45.2	39.7
24th (Wed.)	93	90.7	20.4	18.2	417.3	329.6	370.9	5.2
25th (Thurs.)	84	79.5	11.4	6.9	130.6	48.2	79.3	20.1
26th (Fri.)	59	57.1	–13.6	–15.5	184.2	239.8	210.2	3.7
27th (Sat.)	64	72.0	–8.6	–0.5	73.5	0.3	4.6	64.6
28th (Sun.)	80	83.3	7.4	10.7	55.2	114.1	79.3	10.6
29th (Mon.)	75	79.5	2.4	6.9	5.9	48.2	16.9	20.4
30th (Tues.)	58	53.3	–14.6	–19.2	212.3	369.5	280.1	21.6
31st (Wed.)	91	87.0	18.4	14.4	339.6	207.9	265.7	16.1
1st (Thurs.)	51	57.1	–21.6	–15.5	465.3	239.8	334.0	37.0
2nd (Fri.)	73	79.5	0.4	6.9	0.2	48.2	3.0	42.4
3rd (Sat.)	65	60.8	–7.6	–11.7	57.3	138.0	88.9	17.4
4th (Sun.)	84	75.8	11.4	3.2	130.6	10.3	36.6	67.6
Sum	1016	1016	0	0	2203.4	1812.3	1812.3	391.1
Average	72.6	72.6						
	$\bar{y}$	$\bar{\hat{y}}$			S_{yy}	$S_{\hat{y}\hat{y}}$	$S_{y\hat{y}}$	S_e

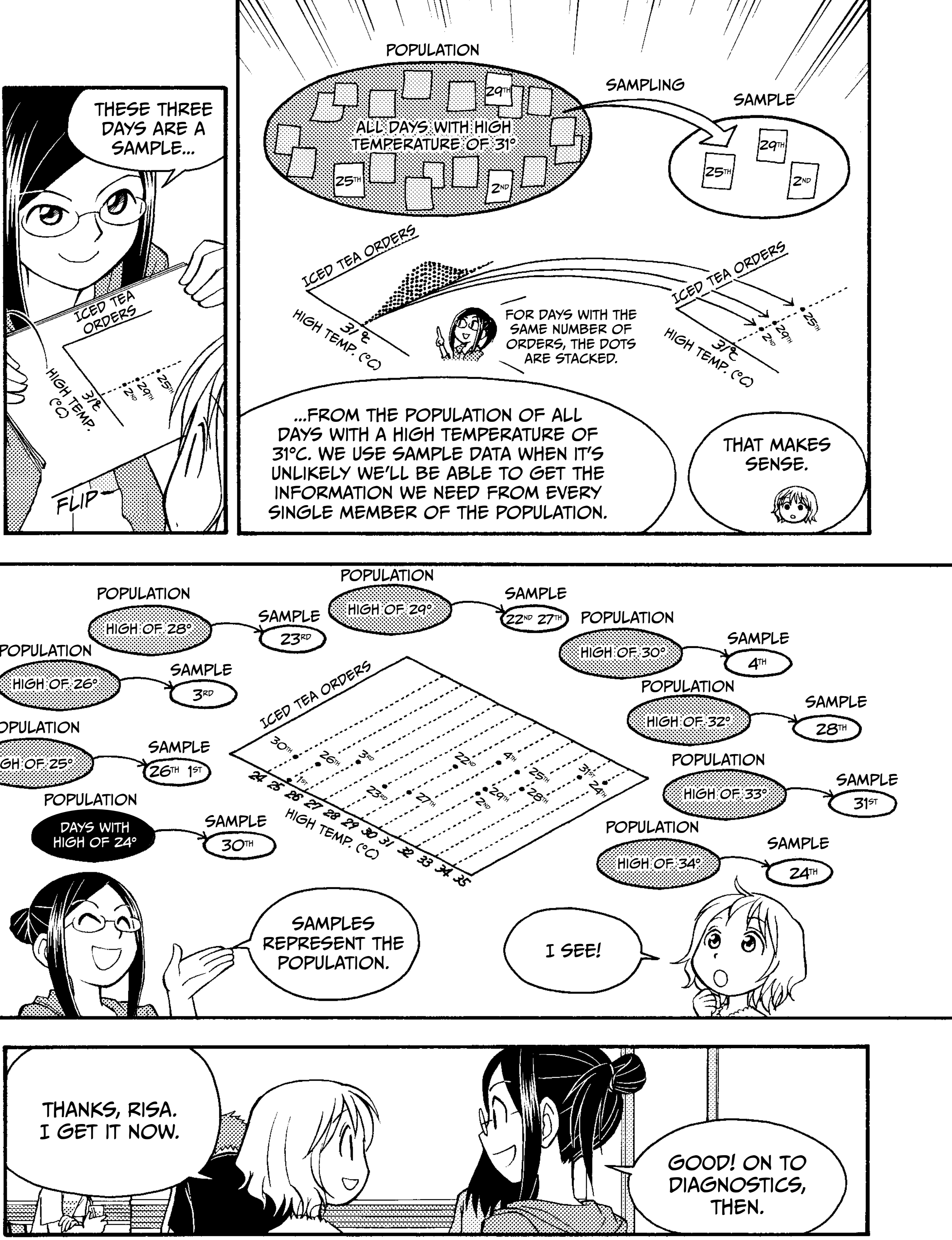


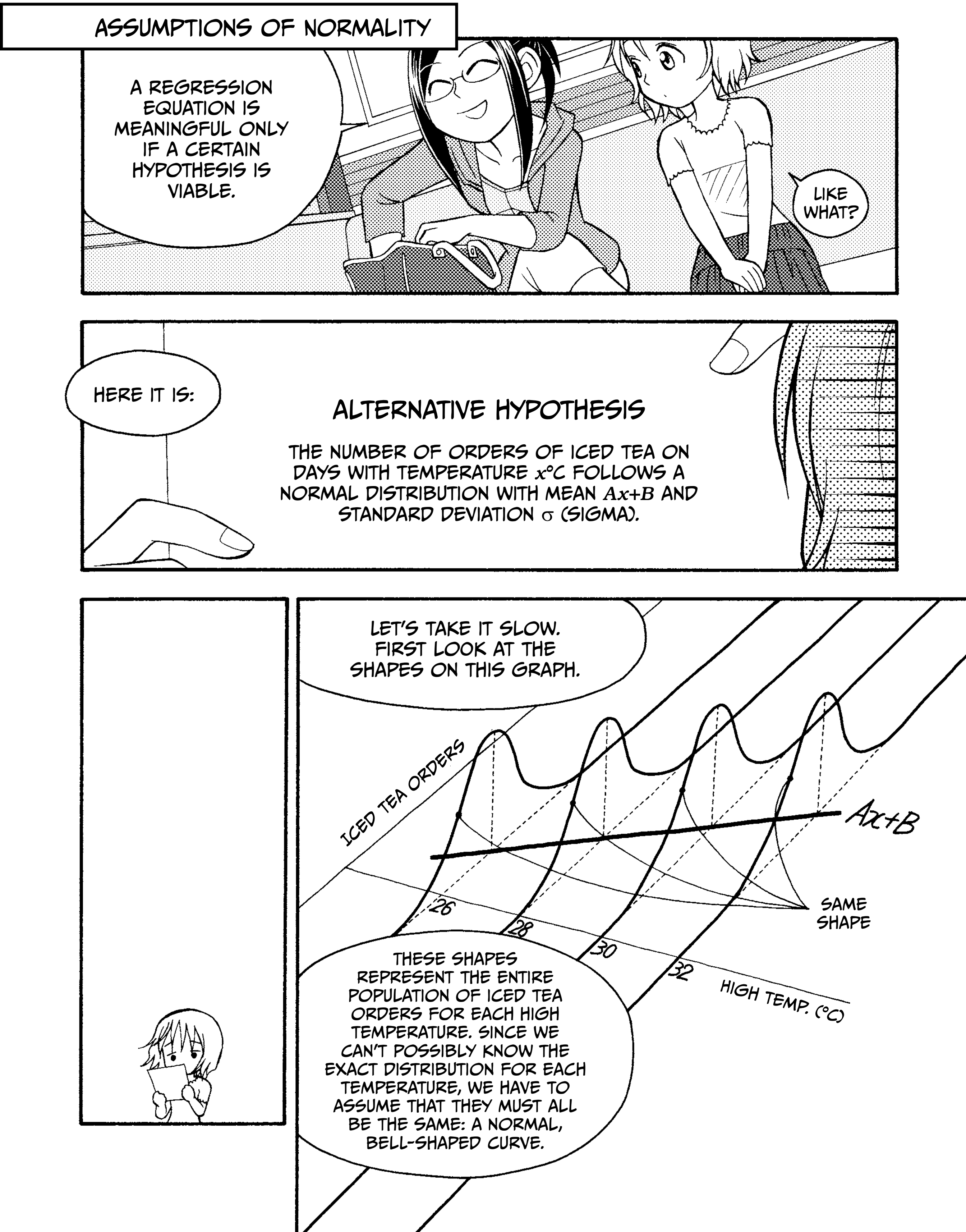

IF WE SQUARE *R*, IT'S CALLED THE *COEFFICIENT OF DETERMINATION* AND IS WRITTEN AS R^2.
R^2 CAN BE AN INDICATOR OF...
I AM A CORRELATION COEFFICIENT.
I AM A CORRELATION COEFFICIENT, TOO.
I AM A COEFFICIENT OF DETERMINATION.
$R \times R = R^2$
...HOW MUCH VARIANCE IS EXPLAINED BY OUR REGRESSION EQUATION.
AN R^2 OF ZERO INDICATES THAT THE OUTCOME VARIABLE CAN'T BE RELIABLY PREDICTED FROM THE PREDICTOR VARIABLE.
0
1
THE HIGHER THE ACCURACY OF THE REGRESSION EQUATION, THE CLOSER THE R^2 VALUE IS TO 1, AND VICE VERSA.
SO HOW HIGH DOES R^2 NEED TO BE FOR THE REGRESSION EQUATION TO BE CONSIDERED ACCURATE?
UNFORTUNATELY, THERE IS NO UNIVERSAL STANDARD IN STATISTICS.
BUT GENERALLY WE WANT A VALUE OF AT LEAST .5.
LOWEST... .5...
NOW TRY FINDING THE VALUE OF R^2.
SURE THING.
$R^2 = (0.9069)^2$
$= 0.8225$
IT'S .8225.

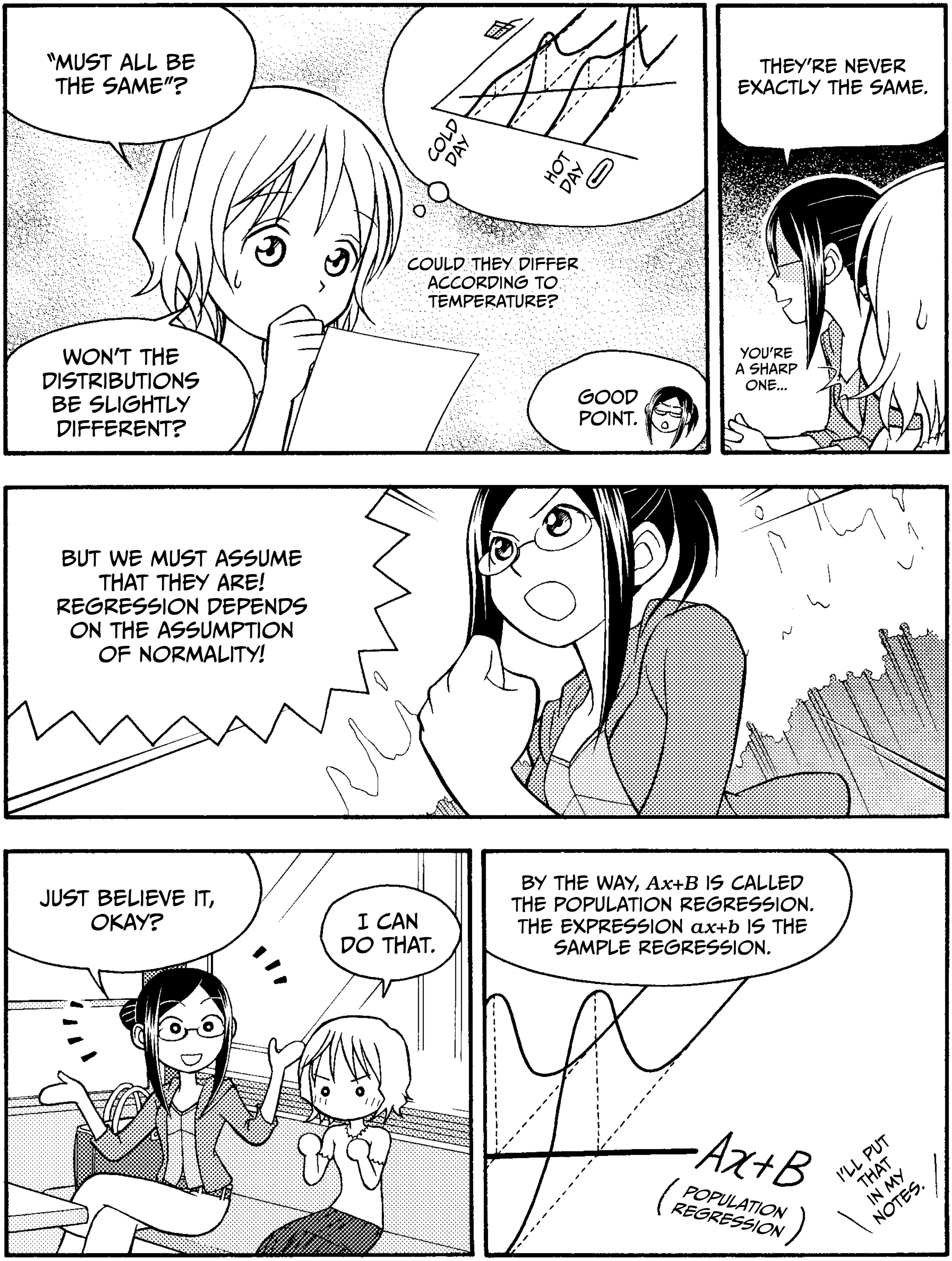

THE VALUE OF R^2 FOR OUR REGRESSION EQUATION IS WELL OVER .5, SO OUR EQUATION SHOULD BE ABLE TO ESTIMATE ICED TEA ORDERS RELATIVELY ACCURATELY.
ICED TEA ORDERS
HIGH TEMP. (°C)
y=3.7x−36.4
R^2 = .8225
YAY R^2!
$$R^2 = \left(\begin{matrix}\textbf{correlation}\\ \textbf{coefficient}\end{matrix}\right)^2 = \frac{a \times S_{xy}}{S_{yy}} = 1 - \frac{S_e}{S_{yy}}$$
JOT THIS EQUATION DOWN. R^2 CAN BE CALCULATED DIRECTLY FROM THESE VALUES. USING OUR NORNS DATA, 1 − (391.1 / 2203.4) = .8225!
THAT'S HANDY!
WE'VE FINISHED THE FIRST THREE STEPS.
HOORAY!
SAMPLES AND POPULATIONS
NOW TO ASSESS THE POPULATION AND VERIFY THAT OUR ASSUMPTIONS ARE MET!
OH...
I MEANT TO ASK YOU ABOUT THAT. WHAT POPULATION? JAPAN? EARTH?
ACTUALLY, THE POPULATION WE'RE TALKING ABOUT ISN'T PEOPLE—IT'S DATA.

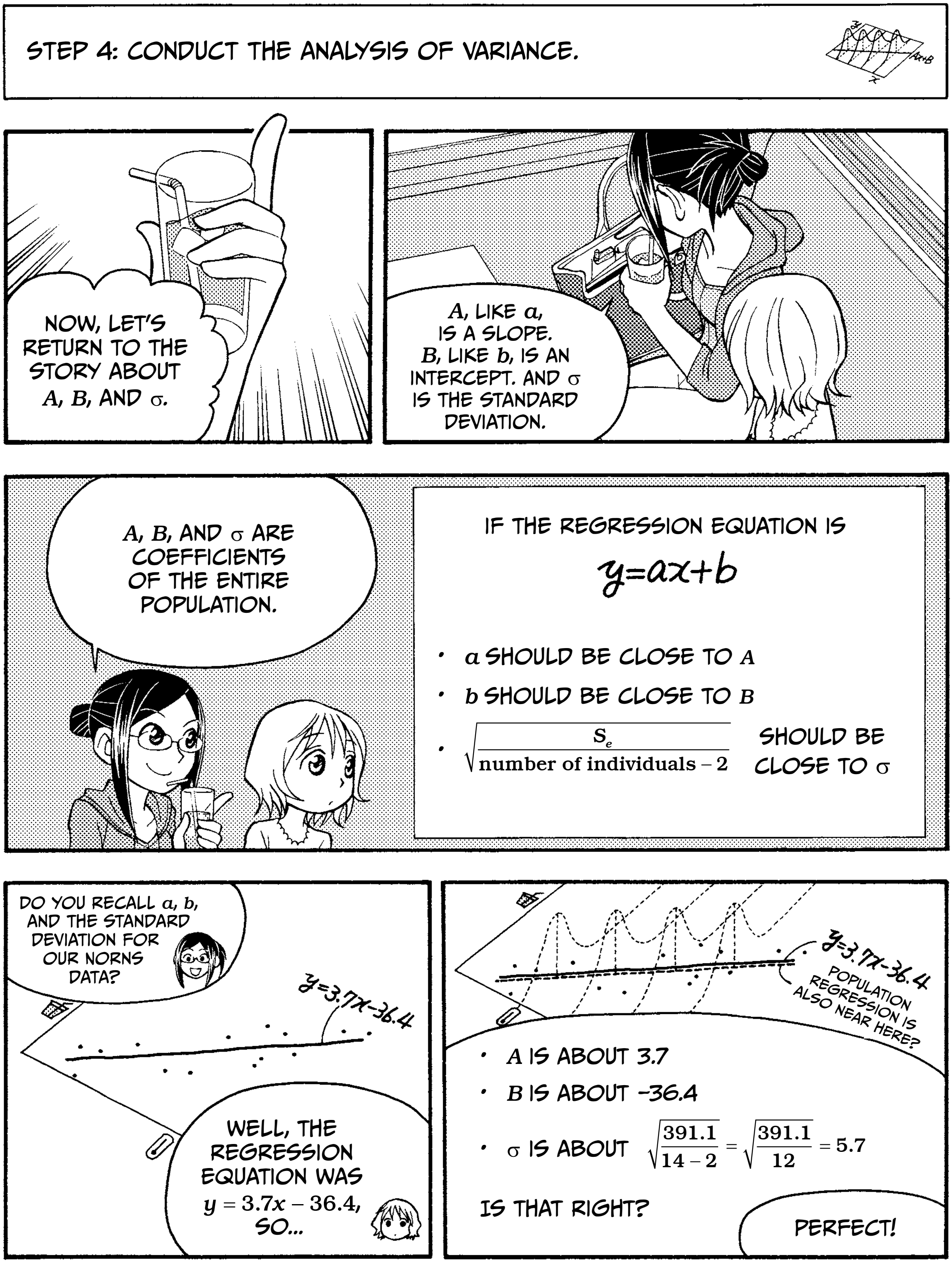


	High temp. (°C)	Iced tea orders
22nd (Mon.)	29	77
23rd (Tues.)	28	62
24th (Wed.)	34	93
25th (Thurs.)	31	84
26th (Fri.)	25	59
27th (Sat.)	29	64
28th (Sun.)	32	80
29th (Mon.)	31	75
30th (Tues.)	24	58
31st (Wed.)	33	91
1st (Thurs.)	25	51
2nd (Fri.)	31	73
3rd (Sat.)	26	65
4th (Sun.)	30	84

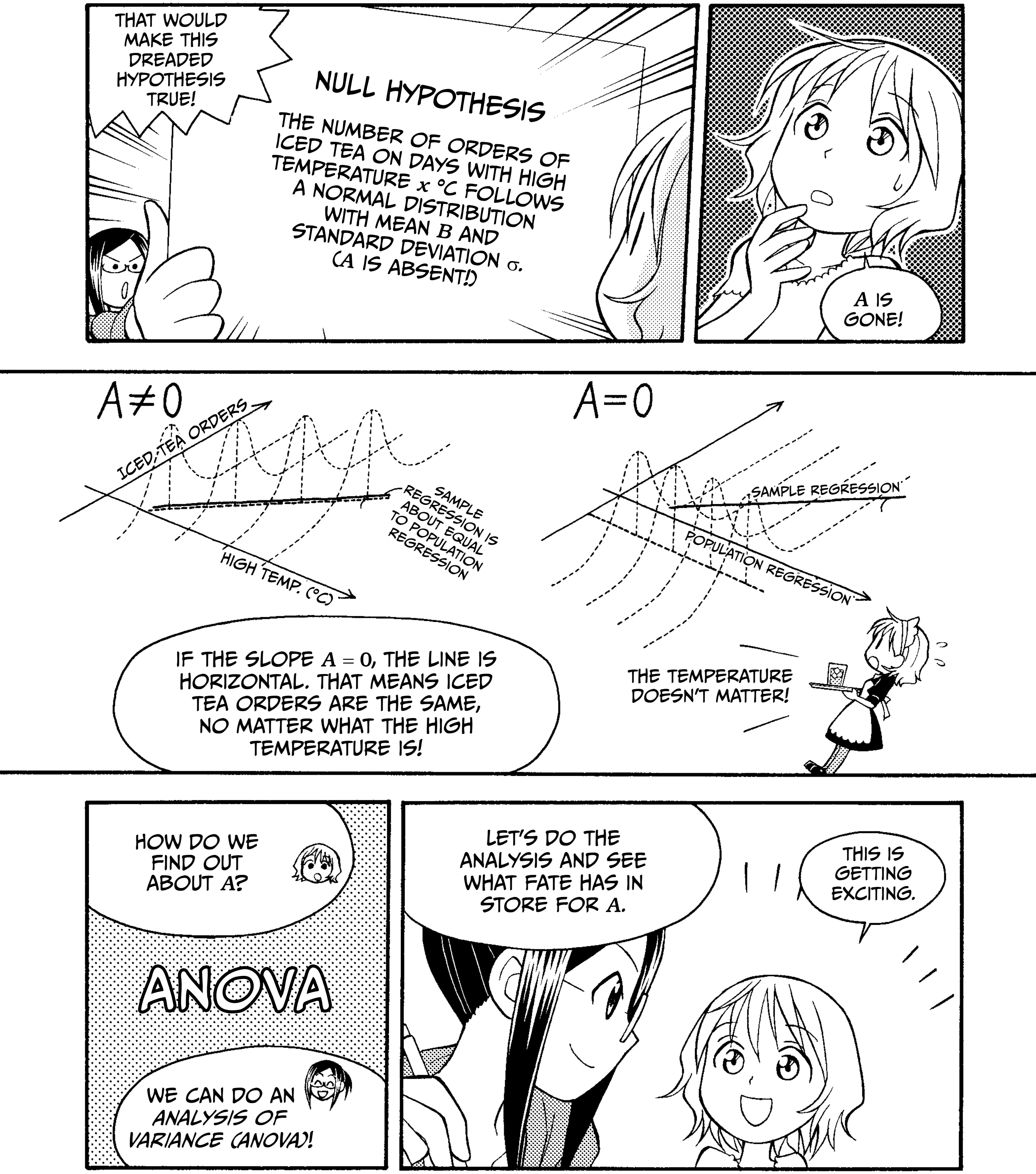

THESE THREE DAYS ARE A SAMPLE...
ICED TEA ORDERS
HIGH TEMP. (°C)
31°C
2ND 29TH 25TH
FLIP
POPULATION
ALL DAYS WITH HIGH TEMPERATURE OF 31°
29TH 25TH 2ND
SAMPLING
SAMPLE
29TH 25TH 2ND
ICED TEA ORDERS
HIGH TEMP. (°C)
31°C
FOR DAYS WITH THE SAME NUMBER OF ORDERS, THE DOTS ARE STACKED.
ICED TEA ORDERS
HIGH TEMP. (°C)
31°C
2ND 29TH 25TH
...FROM THE POPULATION OF ALL DAYS WITH A HIGH TEMPERATURE OF 31°C. WE USE SAMPLE DATA WHEN IT'S UNLIKELY WE'LL BE ABLE TO GET THE INFORMATION WE NEED FROM EVERY SINGLE MEMBER OF THE POPULATION.
THAT MAKES SENSE.
POPULATION DAYS WITH HIGH OF 24° SAMPLE 30TH
POPULATION HIGH OF 25° SAMPLE 26TH 1ST
POPULATION HIGH OF 26° SAMPLE 3RD
POPULATION HIGH OF 28° SAMPLE 23RD
POPULATION HIGH OF 29° SAMPLE 22ND 27TH
POPULATION HIGH OF 30° SAMPLE 4TH
POPULATION HIGH OF 32° SAMPLE 28TH
POPULATION HIGH OF 33° SAMPLE 31ST
POPULATION HIGH OF 34° SAMPLE 24TH
ICED TEA ORDERS
HIGH TEMP. (°C)
24 25 26 27 28 29 30 31 32 33 34 35
30TH 1ST 26TH 3RD 23RD 27TH 22ND 4TH 2ND 29TH 28TH 25TH 31ST 24TH
SAMPLES REPRESENT THE POPULATION.
I SEE!
THANKS, RISA. I GET IT NOW.
GOOD! ON TO DIAGNOSTICS, THEN.


ASSUMPTIONS OF NORMALITY
A REGRESSION EQUATION IS MEANINGFUL ONLY IF A CERTAIN HYPOTHESIS IS VIABLE.
LIKE WHAT?
HERE IT IS:
ALTERNATIVE HYPOTHESIS
THE NUMBER OF ORDERS OF ICED TEA ON DAYS WITH TEMPERATURE x°C FOLLOWS A NORMAL DISTRIBUTION WITH MEAN $Ax+B$ AND STANDARD DEVIATION σ (SIGMA).
LET'S TAKE IT SLOW. FIRST LOOK AT THE SHAPES ON THIS GRAPH.
ICED TEA ORDERS
Ax+B
SAME SHAPE
26
28
30
32
HIGH TEMP. (°C)
THESE SHAPES REPRESENT THE ENTIRE POPULATION OF ICED TEA ORDERS FOR EACH HIGH TEMPERATURE. SINCE WE CAN'T POSSIBLY KNOW THE EXACT DISTRIBUTION FOR EACH TEMPERATURE, WE HAVE TO ASSUME THAT THEY MUST ALL BE THE SAME: A NORMAL, BELL-SHAPED CURVE.


"MUST ALL BE THE SAME"?
COLD DAY
HOT DAY
COULD THEY DIFFER ACCORDING TO TEMPERATURE?
WON'T THE DISTRIBUTIONS BE SLIGHTLY DIFFERENT?
GOOD POINT.
THEY'RE NEVER EXACTLY THE SAME.
YOU'RE A SHARP ONE...
BUT WE MUST ASSUME THAT THEY ARE! REGRESSION DEPENDS ON THE ASSUMPTION OF NORMALITY!
JUST BELIEVE IT, OKAY?
I CAN DO THAT.
BY THE WAY, Ax+B IS CALLED THE POPULATION REGRESSION. THE EXPRESSION ax+b IS THE SAMPLE REGRESSION.
Ax+B
(POPULATION REGRESSION)
I'LL PUT THAT IN MY NOTES.

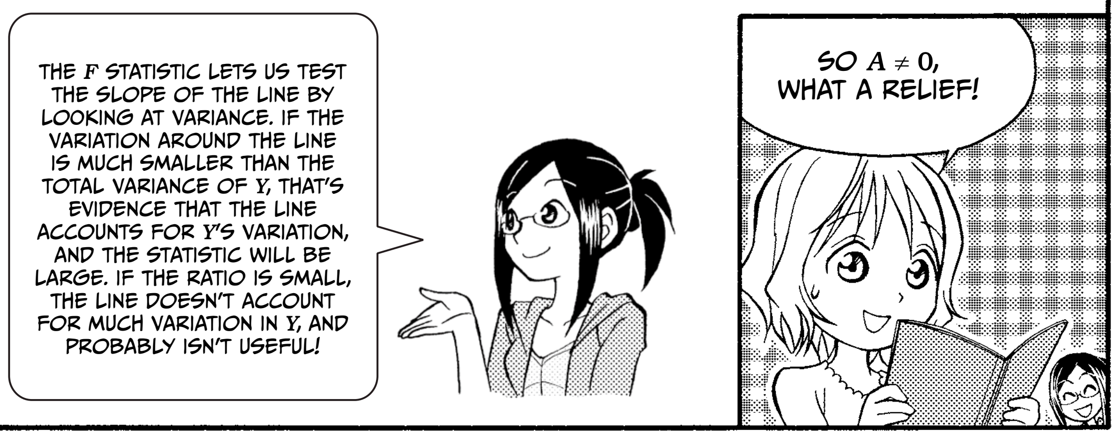

STEP 4: CONDUCT THE ANALYSIS OF VARIANCE.
NOW, LET'S RETURN TO THE STORY ABOUT A, B, AND σ.
A, LIKE a, IS A SLOPE. B, LIKE b, IS AN INTERCEPT. AND σ IS THE STANDARD DEVIATION.
A, B, AND σ ARE COEFFICIENTS OF THE ENTIRE POPULATION.
IF THE REGRESSION EQUATION IS
$y=ax+b$
• a SHOULD BE CLOSE TO A
• b SHOULD BE CLOSE TO B
• $\sqrt{\dfrac{S_e}{\text{number of individuals} - 2}}$ SHOULD BE CLOSE TO σ
DO YOU RECALL a, b, AND THE STANDARD DEVIATION FOR OUR NORNS DATA?
$y=3.7x-36.4$
WELL, THE REGRESSION EQUATION WAS $y = 3.7x - 36.4$, SO...
$y=3.7x-36.4$
POPULATION REGRESSION IS ALSO NEAR HERE?
• A IS ABOUT 3.7
• B IS ABOUT −36.4
• σ IS ABOUT $\sqrt{\dfrac{391.1}{14-2}} = \sqrt{\dfrac{391.1}{12}} = 5.7$
IS THAT RIGHT?
PERFECT!


"CLOSE TO" SEEMS SO VAGUE. CAN'T WE FIND A, B, AND σ WITH MORE CERTAINTY?
SINCE A, B, AND σ ARE COEFFICIENTS OF THE *POPULATION*, WE'D NEED TO USE ALL THE NORNS ICED TEA AND HIGH TEMPERATURE DATA THROUGHOUT HISTORY! WE COULD NEVER GET IT ALL.
HOWEVER...
...WE CAN DETERMINE ONCE AND FOR ALL WHETHER $A = 0$!
TAH-RUH!
...
YOU SHOULD LOOK MORE EXCITED! THIS IS IMPORTANT!
IMAGINE IF A WERE ZERO...

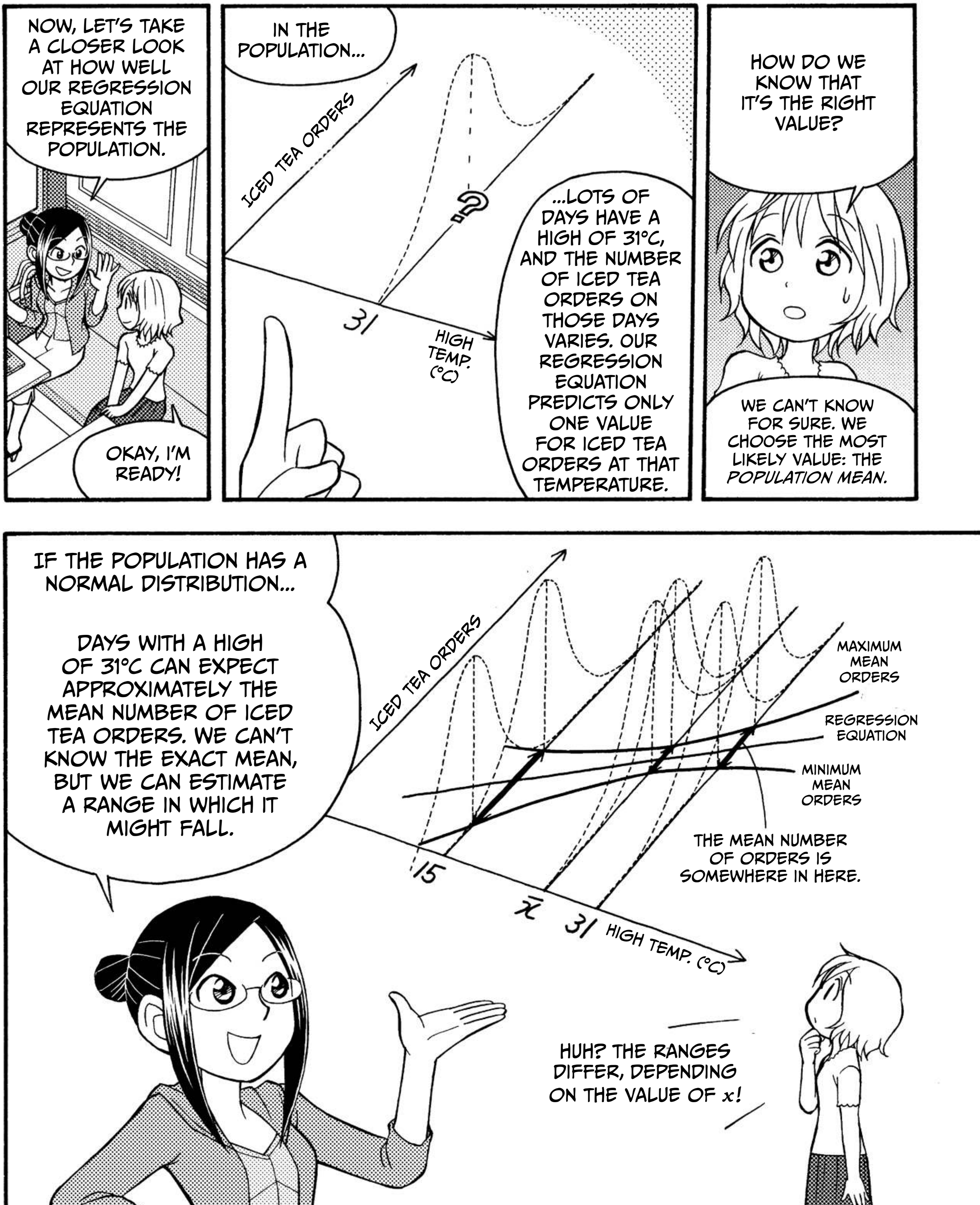

THAT WOULD MAKE THIS DREADED HYPOTHESIS TRUE!
NULL HYPOTHESIS
THE NUMBER OF ORDERS OF ICED TEA ON DAYS WITH HIGH TEMPERATURE x °C FOLLOWS A NORMAL DISTRIBUTION WITH MEAN B AND STANDARD DEVIATION σ. (A IS ABSENT!)
A IS GONE!
$A \neq 0$
ICED TEA ORDERS
SAMPLE REGRESSION IS ABOUT EQUAL TO POPULATION REGRESSION
HIGH TEMP. (°C)
$A = 0$
SAMPLE REGRESSION
POPULATION REGRESSION
IF THE SLOPE $A = 0$, THE LINE IS HORIZONTAL. THAT MEANS ICED TEA ORDERS ARE THE SAME, NO MATTER WHAT THE HIGH TEMPERATURE IS!
THE TEMPERATURE DOESN'T MATTER!
HOW DO WE FIND OUT ABOUT A?
ANOVA
WE CAN DO AN *ANALYSIS OF VARIANCE* (ANOVA)!
LET'S DO THE ANALYSIS AND SEE WHAT FATE HAS IN STORE FOR A.
THIS IS GETTING EXCITING.

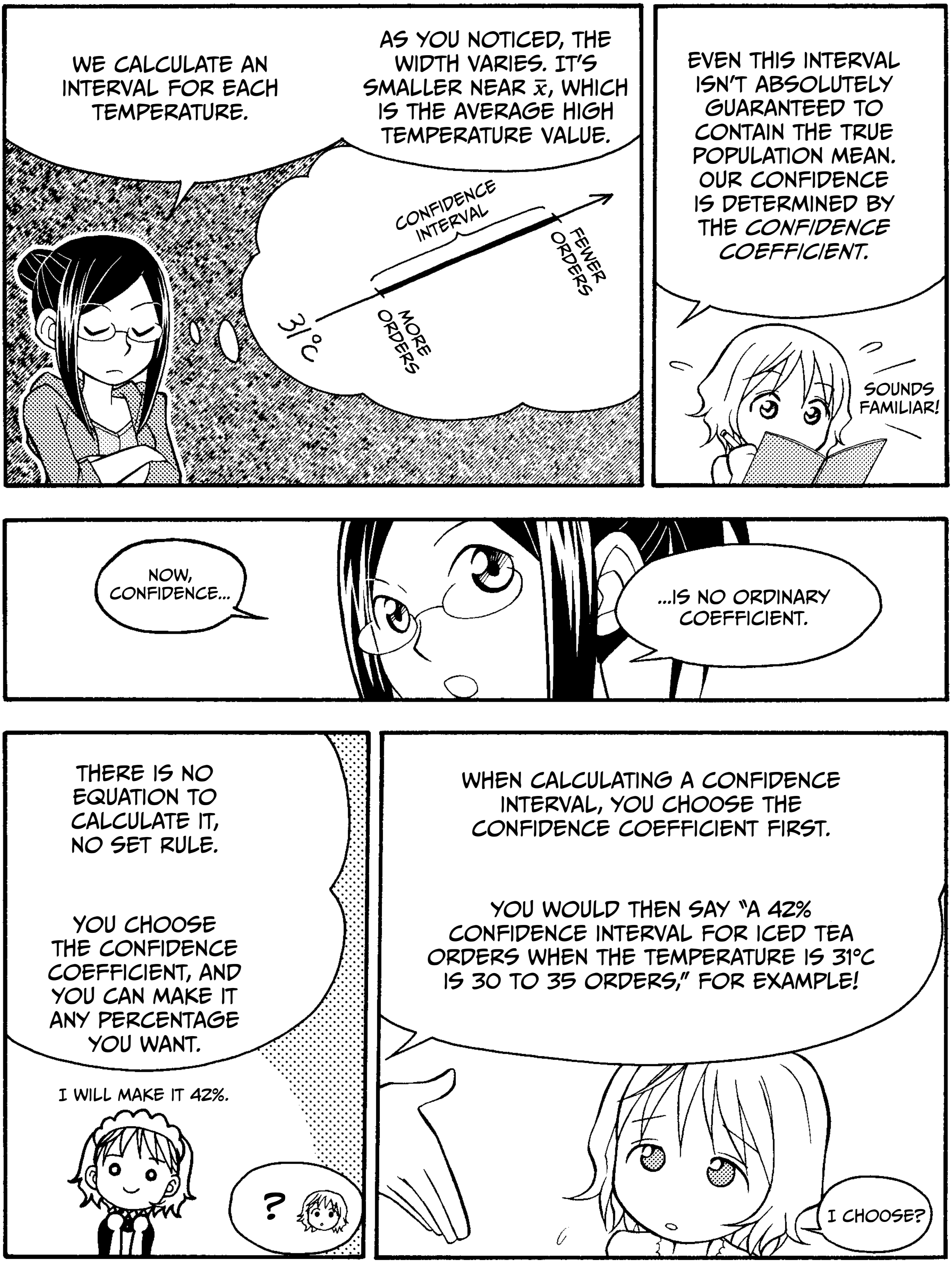
THE STEPS OF ANOVA

Step 1	Define the population.	The population is "days with a high temperature of x degrees."
Step 2	Set up a null hypothesis and an alternative hypothesis.	Null hypothesis is $A = 0$. Alternative hypothesis is $A \neq 0$.
Step 3	Select which hypothesis test to conduct.	We'll use analysis of one-way variance.
Step 4	Choose the significance level.	We'll use a significance level of .05.
Step 5	Calculate the test statistic from the sample data.	The test statistic is: $\dfrac{a^2}{\left(\dfrac{1}{S_{xx}}\right)} \div \dfrac{S_e}{\text{number of individuals} - 2}$ Plug in the values from our sample regression equation: $\dfrac{3.7^2}{\left(\dfrac{1}{129.7}\right)} \div \dfrac{391.1}{14 - 2} = 55.6$ The test statistic will follow an F distribution with first degree of freedom 1 and second degree of freedom 12 (number of individuals minus 2), if the null hypothesis is true.
Step 6	Determine whether the p-value for the test statistic obtained in Step 5 is smaller than the significance level.	At significance level .05, with d_1 being 1 and d_2 being 12, the critical value is 4.7472. Our test statistic is 55.6.
Step 7	Decide whether you can reject the null hypothesis.	Since our test statistic is greater than the critical value, we reject the null hypothesis.

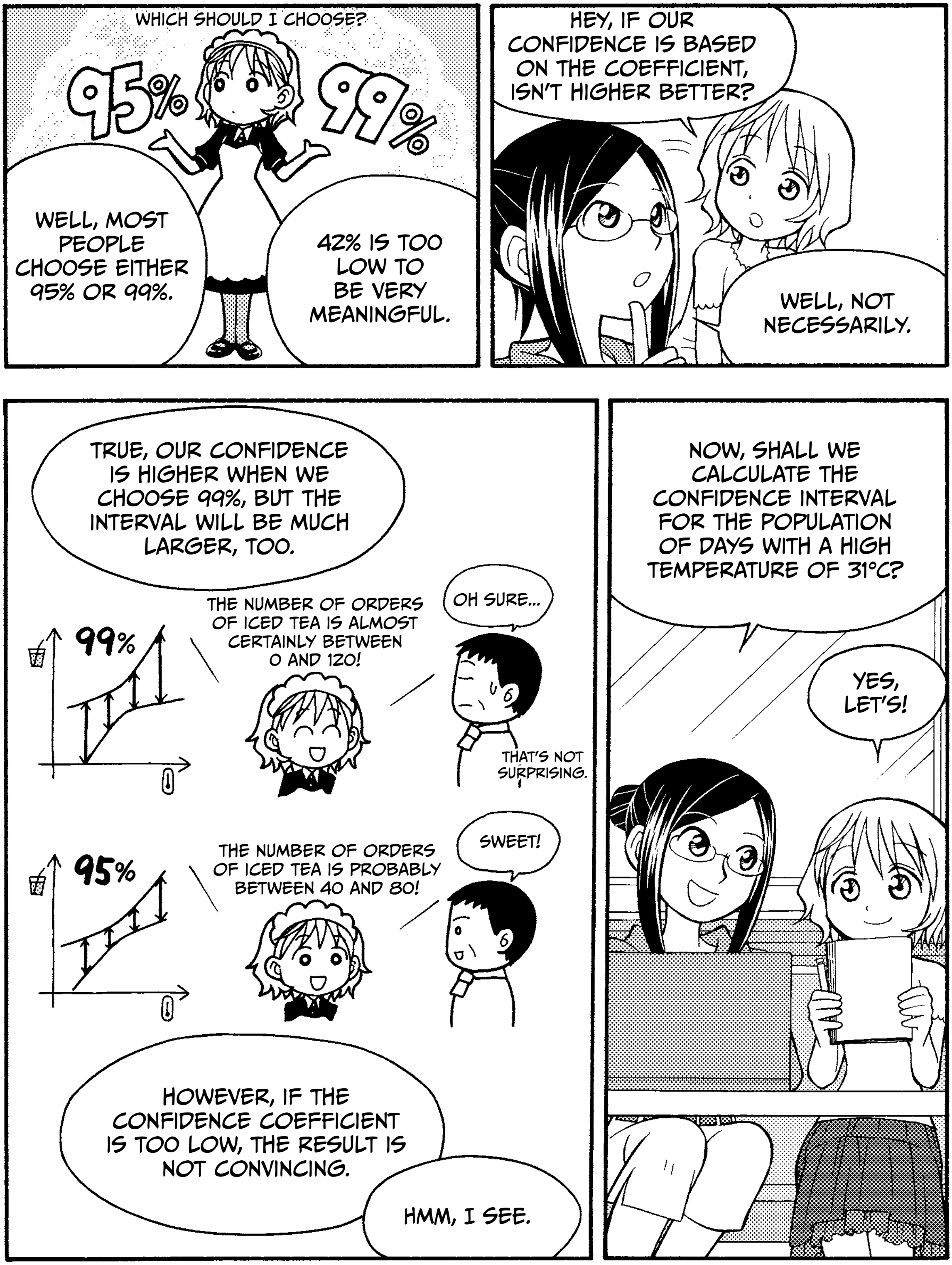


STEP 5: CALCULATE THE CONFIDENCE INTERVALS.





WE CALCULATE AN INTERVAL FOR EACH TEMPERATURE.
AS YOU NOTICED, THE WIDTH VARIES. IT'S SMALLER NEAR $\bar{x}$, WHICH IS THE AVERAGE HIGH TEMPERATURE VALUE.
CONFIDENCE INTERVAL
31°C
MORE ORDERS
FEWER ORDERS
EVEN THIS INTERVAL ISN'T ABSOLUTELY GUARANTEED TO CONTAIN THE TRUE POPULATION MEAN. OUR CONFIDENCE IS DETERMINED BY THE *CONFIDENCE COEFFICIENT.*
SOUNDS FAMILIAR!
NOW, CONFIDENCE...
...IS NO ORDINARY COEFFICIENT.
THERE IS NO EQUATION TO CALCULATE IT, NO SET RULE.
YOU CHOOSE THE CONFIDENCE COEFFICIENT, AND YOU CAN MAKE IT ANY PERCENTAGE YOU WANT.
I WILL MAKE IT 42%.
?
WHEN CALCULATING A CONFIDENCE INTERVAL, YOU CHOOSE THE CONFIDENCE COEFFICIENT FIRST.
YOU WOULD THEN SAY "A 42% CONFIDENCE INTERVAL FOR ICED TEA ORDERS WHEN THE TEMPERATURE IS 31°C IS 30 TO 35 ORDERS," FOR EXAMPLE!
I CHOOSE?


WHICH SHOULD I CHOOSE?
95%
99%
WELL, MOST PEOPLE CHOOSE EITHER 95% OR 99%.
42% IS TOO LOW TO BE VERY MEANINGFUL.
HEY, IF OUR CONFIDENCE IS BASED ON THE COEFFICIENT, ISN'T HIGHER BETTER?
WELL, NOT NECESSARILY.
TRUE, OUR CONFIDENCE IS HIGHER WHEN WE CHOOSE 99%, BUT THE INTERVAL WILL BE MUCH LARGER, TOO.
99%
THE NUMBER OF ORDERS OF ICED TEA IS ALMOST CERTAINLY BETWEEN 0 AND 120!
OH SURE...
THAT'S NOT SURPRISING.
95%
THE NUMBER OF ORDERS OF ICED TEA IS PROBABLY BETWEEN 40 AND 80!
SWEET!
HOWEVER, IF THE CONFIDENCE COEFFICIENT IS TOO LOW, THE RESULT IS NOT CONVINCING.
HMM, I SEE.
NOW, SHALL WE CALCULATE THE CONFIDENCE INTERVAL FOR THE POPULATION OF DAYS WITH A HIGH TEMPERATURE OF 31°C?
YES, LET'S!

HERE'S HOW TO CALCULATE A 95% CONFIDENCE INTERVAL FOR ICED TEA ORDERS ON DAYS WITH A HIGH OF 31°C.

$79.5^* - 3.9 = 75.6$

$$31 \times a + b = 31 \times 3.7 - 36.4 = 79.5$$

$79.5 + 3.9 = 83.4$

Distance from the estimated mean is

$$\sqrt{F(1, n-2; .05) \times \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \times \frac{S_e}{n-2}}$$

$$= \sqrt{F(1, 14-2; .05) \times \left(\frac{1}{14} + \frac{(31 - 29.1)^2}{129.7}\right) \times \frac{391.1}{14-2}}$$

$$= 3.9$$

where n is the number of data points in our sample and F is a ratio of two chi-squared distributions, as described on page 57.

TO CALCULATE A 99% CONFIDENCE INTERVAL, JUST CHANGE

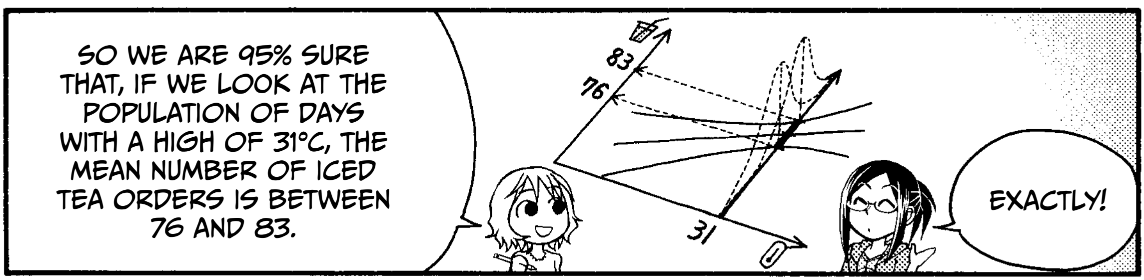
$F(1,14 - 2;.05) = 4.7$

TO

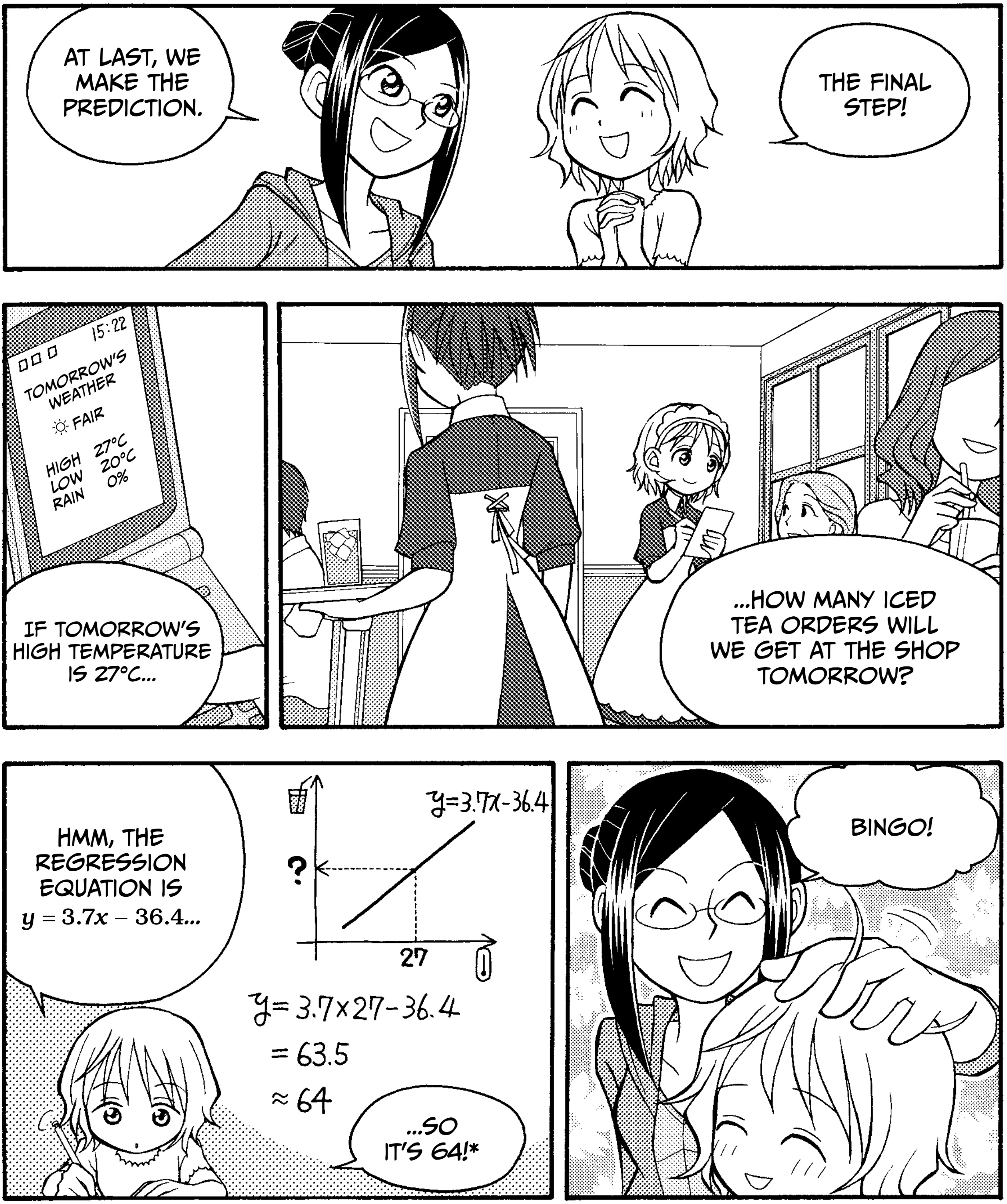
$F(1,14 - 2;.01) = 9.3$

(REFER TO PAGE 58 FOR AN EXPLANATION OF $F(1,n-2;.05) = 4.7$, AND SO ON.)

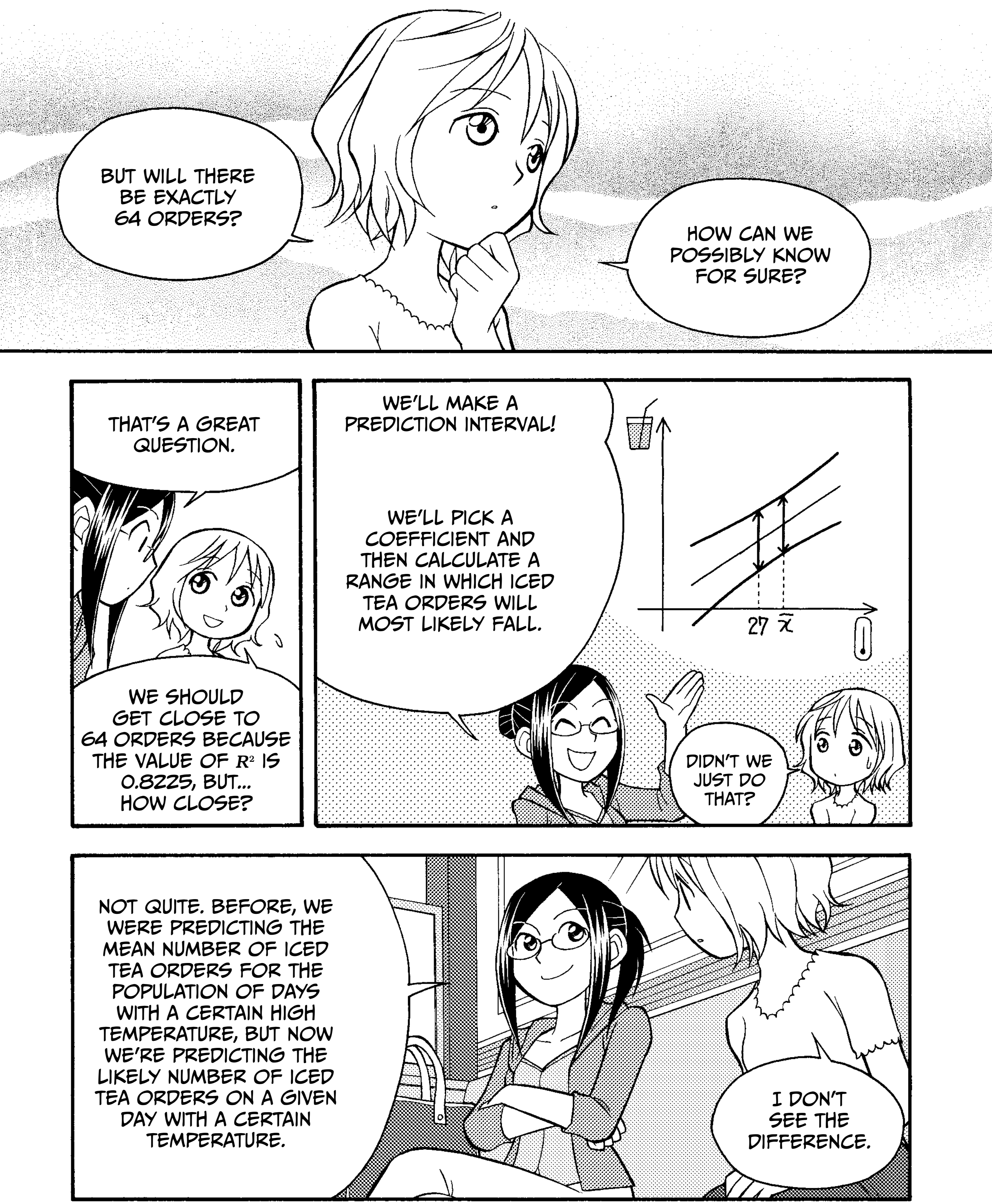
* THE VALUE 79.5 WAS CALCULATED USING UNROUNDED NUMBERS.

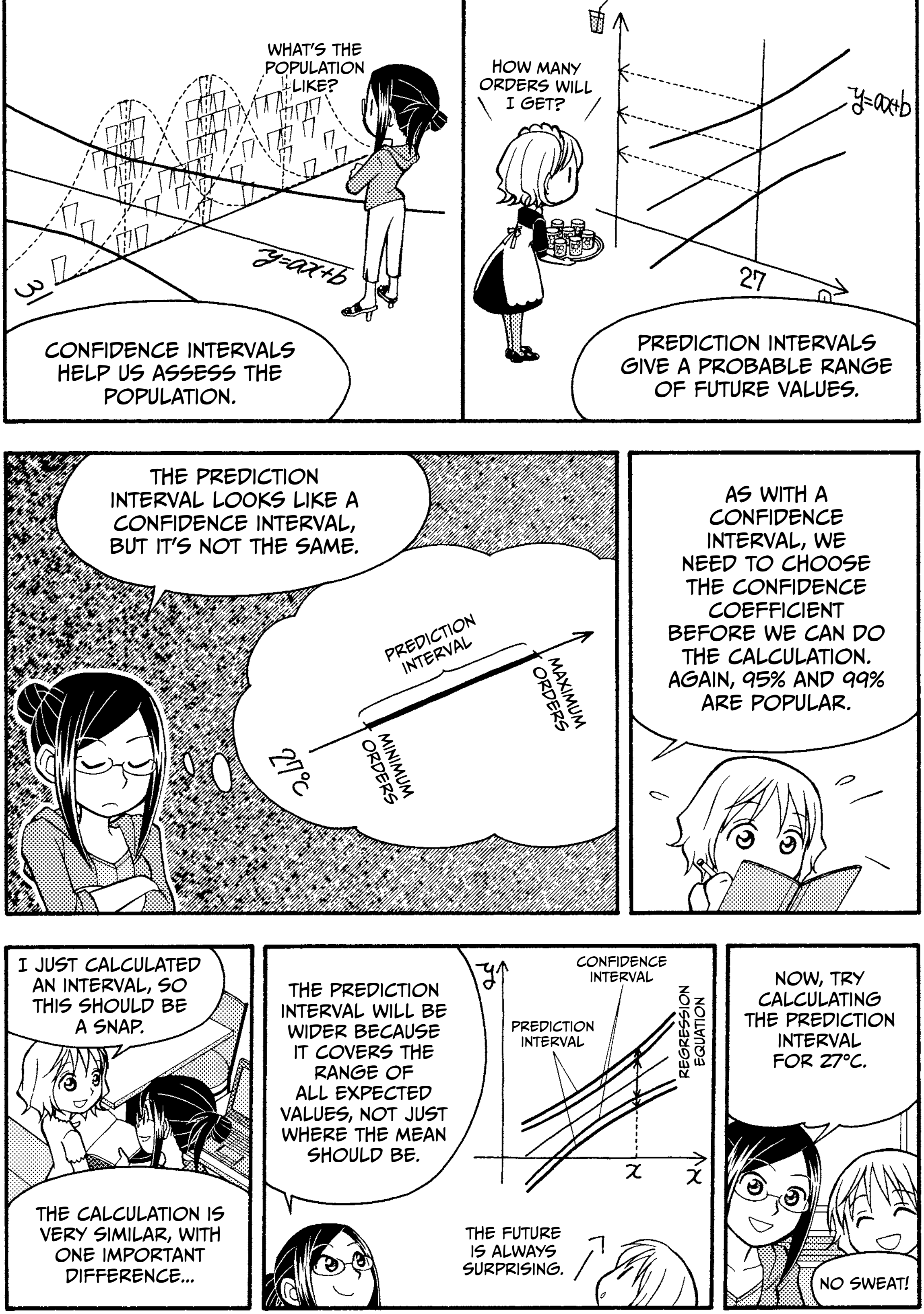


STEP 6: MAKE A PREDICTION!



* THIS CALCULATION WAS PERFORMED USING ROUNDED FIGURES. IF YOU'RE DOING THE CALCULATION WITH THE FULL, UNROUNDED FIGURES, YOU SHOULD GET 64.6.


BUT WILL THERE BE EXACTLY 64 ORDERS?
HOW CAN WE POSSIBLY KNOW FOR SURE?
THAT'S A GREAT QUESTION.
WE SHOULD GET CLOSE TO 64 ORDERS BECAUSE THE VALUE OF R^2 IS 0.8225, BUT... HOW CLOSE?
WE'LL MAKE A PREDICTION INTERVAL!
WE'LL PICK A COEFFICIENT AND THEN CALCULATE A RANGE IN WHICH ICED TEA ORDERS WILL MOST LIKELY FALL.
27
$\bar{x}$
DIDN'T WE JUST DO THAT?
NOT QUITE. BEFORE, WE WERE PREDICTING THE MEAN NUMBER OF ICED TEA ORDERS FOR THE POPULATION OF DAYS WITH A CERTAIN HIGH TEMPERATURE, BUT NOW WE'RE PREDICTING THE LIKELY NUMBER OF ICED TEA ORDERS ON A GIVEN DAY WITH A CERTAIN TEMPERATURE.
I DON'T SEE THE DIFFERENCE.


WHAT'S THE POPULATION LIKE?
y=ax+b
3
CONFIDENCE INTERVALS HELP US ASSESS THE POPULATION.
HOW MANY ORDERS WILL I GET?
y=ax+b
27
PREDICTION INTERVALS GIVE A PROBABLE RANGE OF FUTURE VALUES.
THE PREDICTION INTERVAL LOOKS LIKE A CONFIDENCE INTERVAL, BUT IT'S NOT THE SAME.
PREDICTION INTERVAL
MAXIMUM ORDERS
MINIMUM ORDERS
27°C
AS WITH A CONFIDENCE INTERVAL, WE NEED TO CHOOSE THE CONFIDENCE COEFFICIENT BEFORE WE CAN DO THE CALCULATION. AGAIN, 95% AND 99% ARE POPULAR.
I JUST CALCULATED AN INTERVAL, SO THIS SHOULD BE A SNAP.
THE CALCULATION IS VERY SIMILAR, WITH ONE IMPORTANT DIFFERENCE...
THE PREDICTION INTERVAL WILL BE WIDER BECAUSE IT COVERS THE RANGE OF ALL EXPECTED VALUES, NOT JUST WHERE THE MEAN SHOULD BE.
y
CONFIDENCE INTERVAL
PREDICTION INTERVAL
REGRESSION EQUATION
x
x
THE FUTURE IS ALWAYS SURPRISING.
NOW, TRY CALCULATING THE PREDICTION INTERVAL FOR 27°C.
NO SWEAT!

HERE'S HOW WE CALCULATE A 95% PREDICTION INTERVAL FOR TOMORROW'S ICED TEA SALES.

$64.6 - 13.1 = 51.5$

$$\begin{aligned} &27 \times a + b \\ &= 27 \times 3.7 - 36.4 \\ &= 64.6 \end{aligned}$$

$64.6 + 13.1 = 77.7$*

Distance from the estimated value is

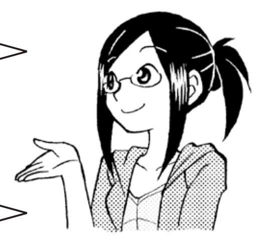
$$\sqrt{F(1, n-2; .05) \times \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \times \frac{S_e}{n-2}}$$

$$= \sqrt{F(1, 14-2; .05) \times \left(1 + \frac{1}{14} + \frac{(27 - 29.1)^2}{129.7}\right) \times \frac{391.1}{14-2}}$$

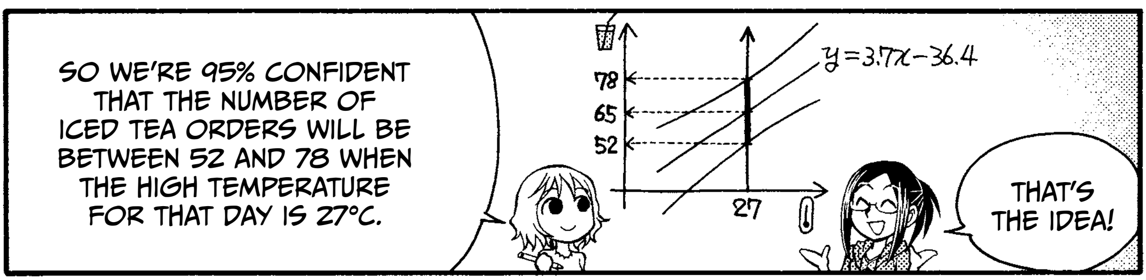
$$= 13.1$$

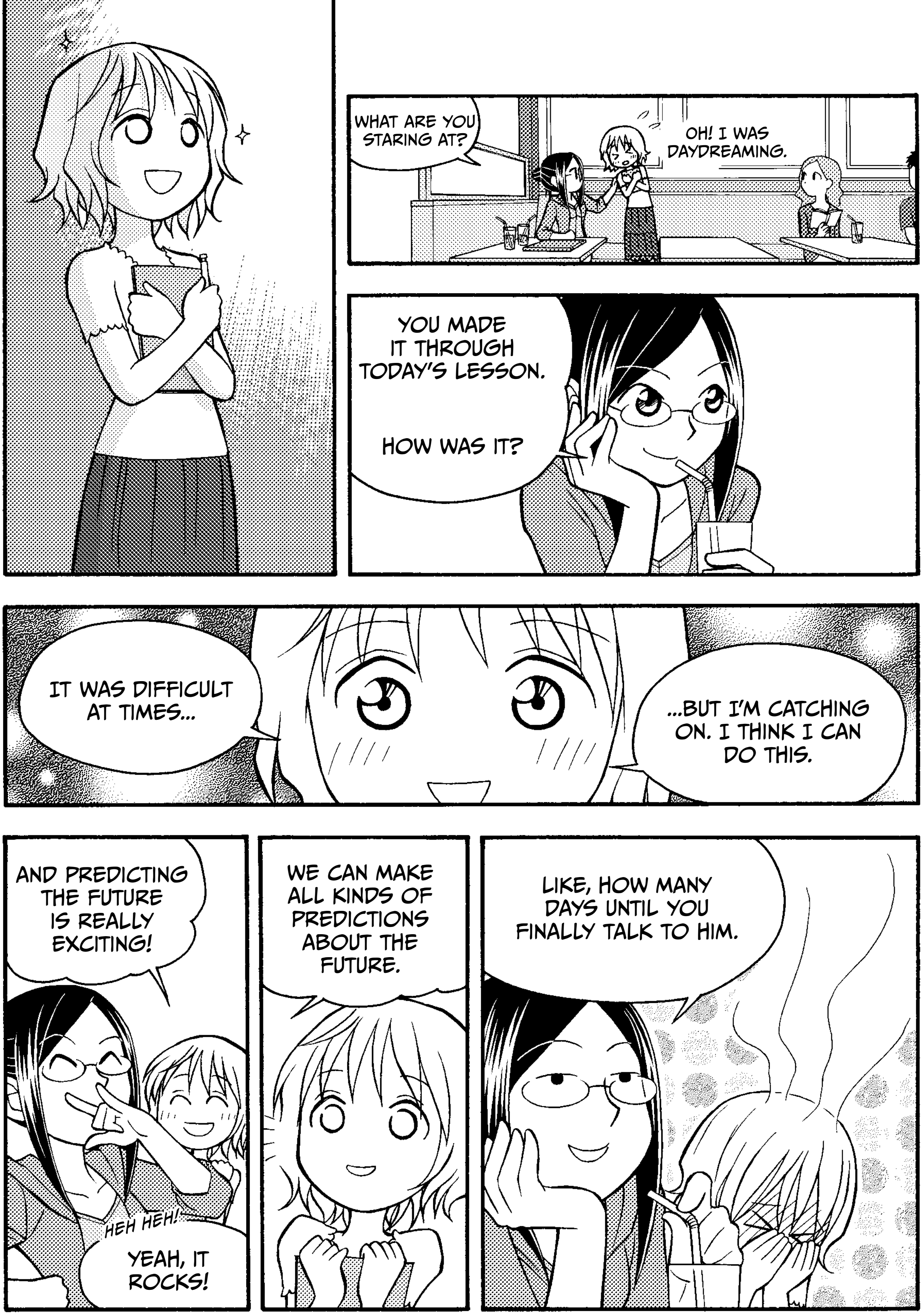
THE ESTIMATED NUMBER OF TEA ORDERS WE CALCULATED EARLIER (ON PAGE 95) WAS ROUNDED, BUT WE'VE USED THE NUMBER OF TEA ORDERS ESTIMATED USING UNROUNDED NUMBERS, 64.6, HERE.

HERE WE USED THE *F* DISTRIBUTION TO FIND THE PREDICTION INTERVAL AND POPULATION REGRESSION. TYPICALLY, STATISTICIANS USE THE *T* DISTRIBUTION TO GET THE SAME RESULTS.



* THIS CALCULATION WAS PERFORMED USING THE ROUNDED NUMBERS SHOWN HERE. THE FULL, UNROUNDED CALCULATION RESULTS IN 77.6.




WHAT ARE YOU STARING AT?
OH! I WAS DAYDREAMING.
YOU MADE IT THROUGH TODAY'S LESSON.
HOW WAS IT?
IT WAS DIFFICULT AT TIMES...
...BUT I'M CATCHING ON. I THINK I CAN DO THIS.
AND PREDICTING THE FUTURE IS REALLY EXCITING!
HEH HEH!
YEAH, IT ROCKS!
WE CAN MAKE ALL KINDS OF PREDICTIONS ABOUT THE FUTURE.
LIKE, HOW MANY DAYS UNTIL YOU FINALLY TALK TO HIM.

WHICH STEPS ARE NECESSARY?

Remember the regression analysis procedure introduced on page 68?

1. Draw a scatter plot of the independent variable versus the dependent variable. If the dots line up, the variables may be correlated.
2. Calculate the regression equation.
3. Calculate the correlation coefficient (R) and assess our population and assumptions.
4. Conduct the analysis of variance.
5. Calculate the confidence intervals.
6. Make a prediction!

In this chapter, we walked through each of the six steps, but it isn't always necessary to do every step. Recall the example of Miu's age and height on page 25.

- Fact: There is only one Miu in this world.
- Fact: Miu's height when she was 10 years old was 137.5 cm.

Given these two facts, it makes no sense to say that "Miu's height when she was 10 years old follows a normal distribution with mean $Ax + B$ and standard deviation σ." In other words, it's nonsense to analyze the population of Miu's heights at 10 years old. She was just one height, and we know what her height was.

In regression analysis, we either analyze the entire population or, much more commonly, analyze a sample of the larger population. When you analyze a sample, you should perform all the steps. However, since Steps 4 and 5 assess how well the sample represents the population, you can skip them if you're using data from an entire population instead of just a sample.

NOTE *We use the term* statistic *to describe a measurement of a characteristic from a sample, like a sample mean, and* parameter *to describe a measurement that comes from a population, like a population mean or coefficient.*

STANDARDIZED RESIDUAL

Remember that a *residual* is the difference between the *measured* value and the value *estimated* with the regression equation. The *standardized residual* is the residual divided by its estimated standard deviation. We use the standardized residual to assess whether a particular measurement deviates significantly from the trend. For example, say a group of thirsty joggers stopped by Norns on the 4th, meaning that though iced tea orders were expected to be about 76 based on that day's high temperature, customers actually placed 84 orders for iced tea. Such an event would result in a large standardized residual.

Standardized residuals are calculated by dividing each residual by an estimate of its standard deviation, which is calculated using the residual sum of squares. The calculation is a little complicated, and most statistics software does it automatically, so we won't go into the details of the calculation here.

Table 2-1 shows the standardized residual for the Norns data used in this chapter.

TABLE 2-1: CALCULATING THE STANDARDIZED RESIDUAL

	High temperature x	Measured number of orders of iced tea y	Estimated number of orders of iced tea $\hat{y} = 3.7x - 36.4$	Residual $y - \hat{y}$	Standardized residual
22nd (Mon.)	29	77	72.0	5.0	0.9
23rd (Tues.)	28	62	68.3	–6.3	–1.2
24th (Wed.)	34	93	90.7	2.3	0.5
25th (Thurs.)	31	84	79.5	4.5	0.8
26th (Fri.)	25	59	57.1	1.9	0.4
27th (Sat.)	29	64	72.0	–8.0	–1.5
28th (Sun.)	32	80	83.3	–3.3	–0.6
29th (Mon.)	31	75	79.5	–4.5	–0.8
30th (Tues.)	24	58	53.3	4.7	1.0
31st (Wed.)	33	91	87.0	4.0	0.8
1st (Thurs.)	25	51	57.1	–6.1	–1.2
2nd (Fri.)	31	73	79.5	–6.5	–1.2
3rd (Sat.)	26	65	60.8	4.2	0.8
4th (Sun.)	30	84	75.8	8.2	1.5

As you can see, the standardized residual on the 4th is 1.5. If iced tea orders had been 76, as expected, the standardized residual would have been 0.

Sometimes a measured value can deviate so much from the trend that it adversely affects the analysis. If the standardized residual is greater than 3 or less than –3, the measurement is considered an *outlier*. There are a number of ways to handle outliers, including removing them, changing them to a set value, or just keeping them in the analysis as is. To determine which approach is most appropriate, investigate the underlying cause of the outliers.

INTERPOLATION AND EXTRAPOLATION

If you look at the x values (high temperature) on page 64, you can see that the highest value is 34°C and the lowest value is 24°C. Using regression analysis, you can *interpolate* the number of iced tea orders on days with a high temperature between 24°C and 34°C and *extrapolate* the number of iced tea orders on days with a high below 24°C or above 34°C. In other words, extrapolation is the estimation of values that fall outside the range of your observed data.

Since we've only observed the trend between 24°C and 34°C, we don't know whether iced tea sales follow the same trend when the weather is extremely cold or extremely hot. Extrapolation is therefore less reliable than interpolation, and some statisticians avoid it entirely.

For everyday use, it's fine to extrapolate—as long as you're aware that your result isn't completely trustworthy. However, avoid using extrapolation in academic research or to estimate a value that's far beyond the scope of the measured data.

AUTOCORRELATION

The independent variable used in this chapter was high temperature; this is used to predict iced tea sales. In most places, it's unlikely that the high temperature will be 20°C one day and then shoot up to 30°C the next day. Normally, the temperature rises or drops gradually over a period of several days, so if the two variables are related, the number of iced tea orders should rise or drop gradually as well. Our assumption, however, has been that the deviation (error) values are random. Therefore, our predicted values do not change from day to day as smoothly as they might in real life.

When analyzing variables that may be affected by the passage of time, it's a good idea to check for autocorrelation. Autocorrelation occurs when the error is correlated over time, and it can indicate that you need to use a different type of regression model.

There's an index to describe autocorrelation—the *Durbin-Watson statistic*, which is calculated as follows:

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

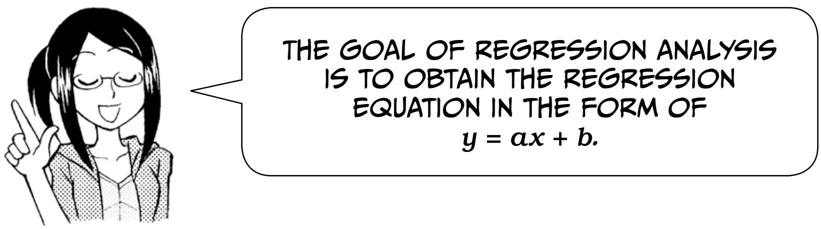
The equation can be read as "the sum of the square of each residual minus the previous residual, divided by the sum of each residual squared." You can calculate the value of the Durbin-Watson statistic for the example in this chapter:

$$\frac{(-6.3-5.0)^2+(2.3-(-6.3))^2+\cdots+(8.2-4.2)^2}{5.0^2+(-6.3)^2+\cdots+8.2^2}=1.8$$

The exact critical value of the Durbin-Watson test differs for each analysis, and you can use a table to find it, but generally we use 1 as a cutoff: a result less than 1 may indicate the presence of autocorrelation. This result is close to 2, so we can conclude that there is no autocorrelation in our example.

Nonlinear Regression

On page 66, Risa said:



This equation is linear, but regression equations don't have to be linear. For example, these equations may also be used as regression equations:

- $y=\frac{a}{x}+b$
- $y=a\sqrt{x}+b$
- $y=ax^2+bx+c$
- $y=a\times\log x+b$

The regression equation for Miu's age and height introduced on page 26 is actually in the form of $y=\frac{a}{x}+b$ rather than $y = ax + b$.

Of course, this raises the question of which type of equation you should choose when performing regression analysis on your own data. Below are some steps that can help you decide.

1. Draw a scatter plot of the data points, with the dependent variable values on the x-axis and the independent variable values on the y-axis. Examine the relationship between the variables suggested by the spread of the dots: Are they in roughly a straight line? Do they fall along a curve? If the latter, what is the shape of the curve?
2. Try the regression equation suggested by the shape in the variables plotted in Step 1. Plot the residuals (or standardized residuals) on the y-axis and the independent variable on the x-axis. The residuals should appear to be random, so if there is an obvious pattern in the residuals, like a curved shape, this suggests that the regression equation doesn't match the shape of the relationship.
3. If the residuals plot from Step 2 shows a pattern in the residuals, try a different regression equation and repeat Step 2. Try the shapes of several regression equations and pick one that appears to most closely match the data. It's usually best to pick the simplest equation that fits the data well.

TRANSFORMING NONLINEAR EQUATIONS INTO LINEAR EQUATIONS

There's another way to deal with nonlinear equations: simply turn them into linear equations. For an example, look at the equation for Miu's age and height (from page 26):

$$y = -\frac{326.6}{x} + 173.3$$

You can turn this into a linear equation. Remember:

$$\text{If } \frac{1}{x} = X, \text{ then } \frac{1}{X} = x.$$

So we'll define a new variable X, set it equal to $\frac{1}{x}$, and use X in the normal $y = aX + b$ regression equation. As shown on page 76, the value of a and b in the regression equation $y = aX + b$ can be calculated as follows:

$$\begin{cases} a = \dfrac{S_{Xy}}{S_{XX}} \\ b = \bar{y} - \bar{X}a \end{cases}$$

We continue with the analysis as usual. See Table 2-2.

TALBE 2-2: CALCULATING THE REGRESSION EQUATION

	Age x	$\frac{1}{\text{age}}$ $\frac{1}{x} = X$	Height y	$(X - \bar{X})$	$y - \bar{y}$	$(X - \bar{X})^2$	$(y - \bar{y})^2$	$(X - \bar{X})(y - \bar{y})$
	4	0.2500	100.1	0.1428	–38.1625	0.0204	1456.3764	–5.4515
	5	0.2000	107.2	0.0928	–31.0625	0.0086	964.8789	–2.8841
	6	0.1667	114.1	0.0595	–24.1625	0.0035	583.8264	–1.4381
	7	0.1429	121.7	0.0357	–16.5625	0.0013	274.3164	–0.5914
	8	0.1250	126.8	0.0178	–11.4625	0.0003	131.3889	–0.2046
	9	0.1111	130.9	0.0040	–7.3625	0.0000	54.2064	–0.0292
	10	0.1000	137.5	–0.0072	–0.7625	0.0001	0.5814	–0.0055
	11	0.0909	143.2	–0.0162	4.9375	0.0003	24.3789	–0.0802
	12	0.0833	149.4	–0.0238	11.1375	0.0006	124.0439	–0.2653
	13	0.0769	151.6	–0.0302	13.3375	0.0009	177.889	–0.4032
	14	0.0714	154.0	–0.0357	15.7375	0.0013	247.6689	–0.5622
	15	0.0667	154.6	–0.0405	16.3375	0.0016	266.9139	–0.6614
	16	0.0625	155.0	–0.0447	16.7375	0.0020	280.1439	–0.7473
	17	0.0588	155.1	–0.0483	16.8375	0.0023	283.5014	–0.8137
	18	0.0556	155.3	–0.0516	17.0375	0.0027	290.2764	–0.8790
	19	0.0526	155.7	–0.0545	17.4375	0.0030	304.0664	–0.9507
Sum	184	1.7144	2212.2	0.0000	0.0000	0.0489	5464.4575	–15.9563
Average	11.5	0.1072	138.3					

According to the table:

$$\begin{cases} a = \dfrac{S_{Xy}}{S_{XX}} = \dfrac{-15.9563}{0.0489} = -326.6^{*} \\ b = \bar{y} - \bar{X}a = 138.2625 - 0.1072 \times (-326.6) = 173.3 \end{cases}$$

So the regression equation is this:

$$\underset{\text{height}}{\underset{\uparrow}{y}} = -326.6\underset{\frac{1}{\text{age}}}{\underset{\uparrow}{X}} + 173.3$$

* If your result is slightly different from 326.6, the difference might be due to rounding. If so, it should be very small.

which is the same as this:

$$\underset{\substack{\uparrow \\ \text{height}}}{y} = -\frac{326.6}{\underset{\substack{\uparrow \\ \text{age}}}{x}} + 173.3$$

We've transformed our original, nonlinear equation into a linear one!