Clinical statistics for non-statisticians: Day three

Steve Simon

Re-introduce yourself

Here’s one more interesting number about myself

- 20: I have a 20 year old son.

Tell us one more interesting number about yourself.

Speaker notes

Speaker notes

Back in 2004, my wife traveled to Russia to adopt a two year old child. Every year is full of excitement and challenges. The one thing I will warn you about is that the older they get, the more expensive their toys become. When he was little, I could spend a few dollars for a Hot Wheels car. Now, he has a 2008 Hummer, and it cost quite a bit more!

Outline of the three day course

- Day one: Numerical summaries and data visualization
- Day two: Hypothesis testing and sampling
- Day three: Statistical tests to compare treatment to a control and regression models

My goal: help you to become a better consumer of statistics

Speaker notes

Speaker notes

We are in the third day of the class. The topics you will see today include comparisons of a treatment to a control. This is a common setting in medical research. You will also get an introduction to regression models.

Day three topics

- Statistical tests to compare a treatment to a control
    - What tests should you use for categorical outcomes?
    - What tests should you use for continuous outcomes?
    - When should you use nonparametric tests?

Speaker notes

Speaker notes

The you use when comparing a treatment to a control depends on whether the outcome is continuous or categorical. You’ll also get a feel for when to use nonparametric tests instead of the traditional tests.

Day three topics (continued)

- Regression models
  - How does a regression model quantify trends
  - How does logistic regression differ from linear regression
  - What is a confounding variable
  - How should you control for or adjust for confounding

Speaker notes

Speaker notes

You’ll also get an introduction to regression models. I want to emphasize interpretation, both for linear regression and logistic regression. Regression models excel at identifying and controlling for confounding variables.

But first, a joke that relates to the topic of regression.

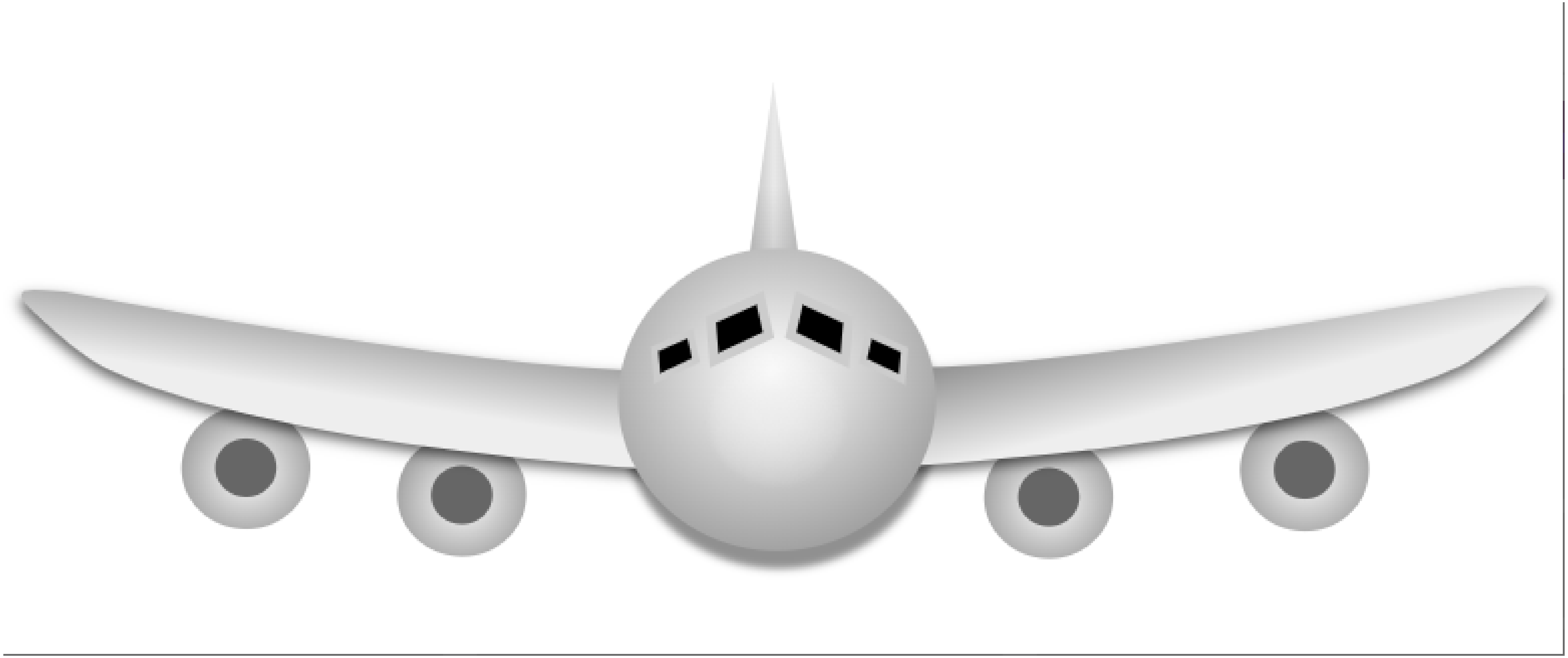


Figure 1: Image of a passenger jet with four engines

Speaker notes

Speaker notes

This image was downloaded from publicdomainvectors.org.

Two statisticians are on a plane flying from Amsterdam to Zurich. The flight proceeds normally, until 15 minutes go by, when they hear …

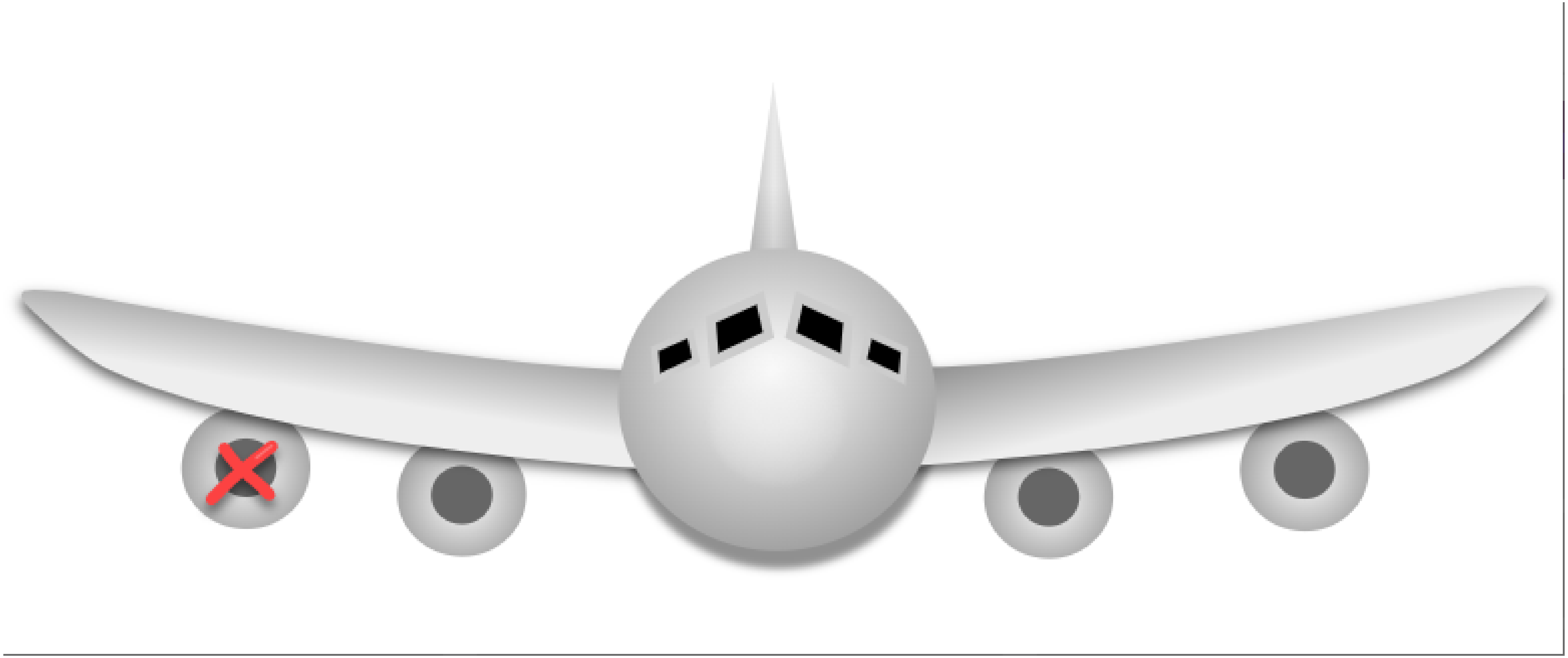


Figure 2: Image of a passenger jet with one bad engine

Speaker notes

Speaker notes

… a loud BANG! The pilot comes on the intercom and says “Ladies and gentlemen, one of the four engines on this plan just exploded. You’re all fine, but the flight that was supposed to take four hours will now take six hours. I apologize for any inconvenience.” So the two statisticians shrug their shoulders and start talking again about the geeky things that statisticians like to talk about. Fifteen minutes later, they hear …

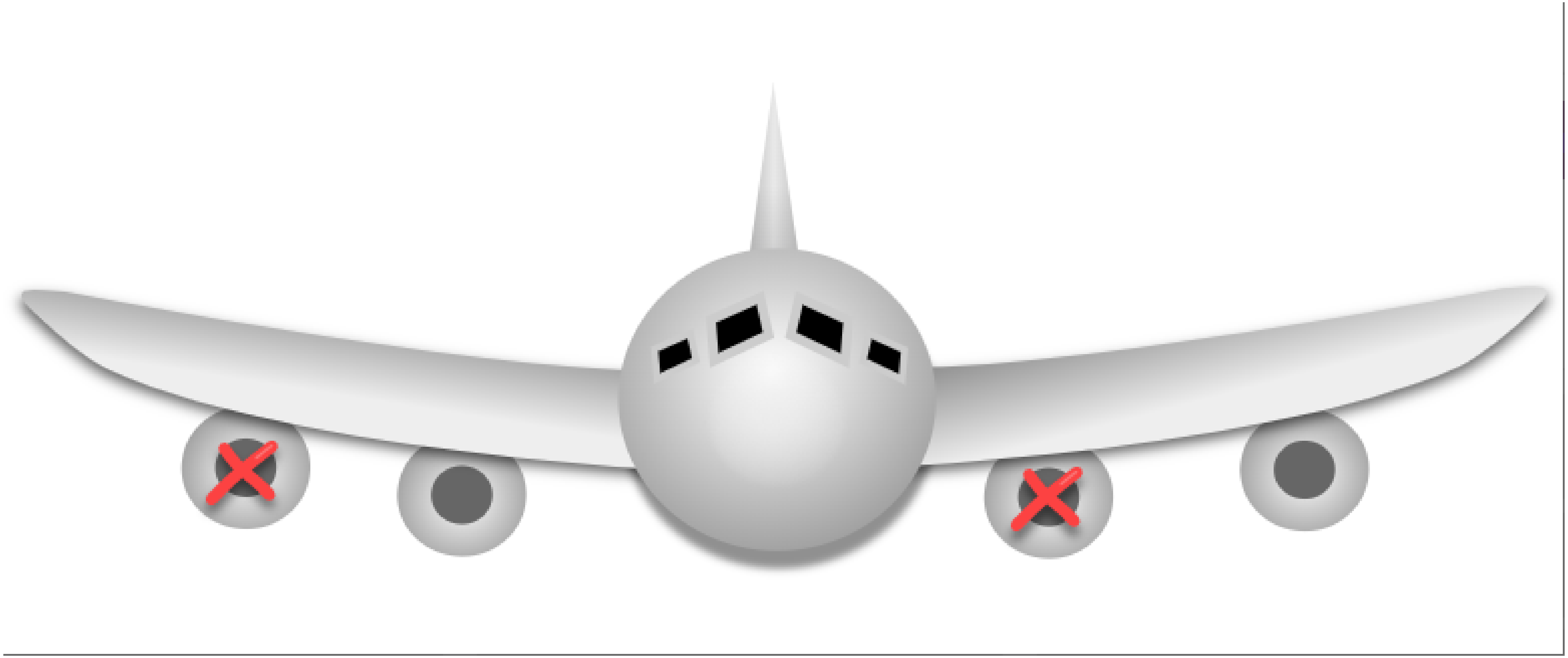


Figure 3: Image of a passenger jet with two bad engines

Speaker notes

… another loud BANG! The pilot comes on the intercom and says “Ladies and gentlemen, a second engines just exploded. You’re all fine, but the flight is now going to take eight hours. I greatly apologize for any inconvenience.” So the two statisticians shrug their shoulders and start talking again. You know what happens next. In fifteen minutes later, they hear …

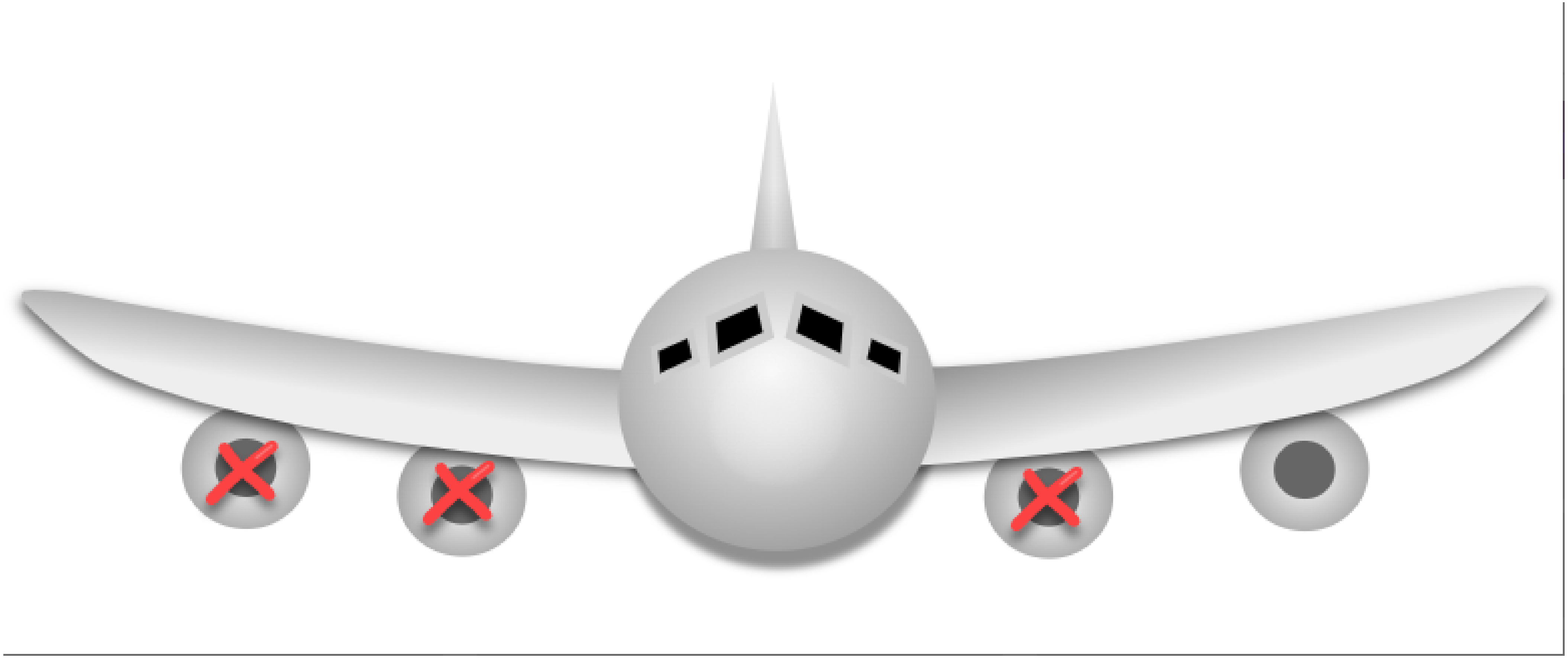


Figure 4: Image of a passenger jet with three bad engines

Speaker notes

… a third loud BANG! The pilot comes on the itnercom and says “Ladies and gentlemen, we just lost a third engine. Now every engine on this aircraft is very powerful so you all will arrive safely in Zurich. But the flight time is now ten hours.” At this point, one statistician turns to the other and says, “I hope this last engine doesn’t explode…”

“…OR WE’LL BE UP HERE FOREVER!”

This story illustrates a dangerous extrapolation. It offers an important point in the use of regression models, so I will return to it later in this talk.

Comparison of treatment and control

- Treatment, something new to help a patient
  - Active intervention
  - Randomized trial
- Exposure, something that a patient endures
  - Passive observation
  - Epidemiology study
- Control
  - Placebo, or
  - Usual standard of care

Speaker notes

Speaker notes

I am going to use the terms treatment and control a lot today, so you need to understand how I define these terms.

A treatment is something new that you as a researcher do to your patients. It could be a new drug, a medical device, an exercise regimen, or something else. It is intended to make things better. It might not make things better. In fact, it may at times make things worse. But the intent is important. A treatment is something that you do that you hope will make things better.

A treatment requires your active intervention. You do something different to the patient or ask them to do something different than what they might normally do. Treatments are usually evaluated in the context of a randomized trial.

In contrast, an exposure is something that a patient endures. It is thought that it might be harmful, but you're not sure. It could be a chemical exposure, such as a pesticide. It could be a dietary choice like black pudding. It could be a lifestyle choice like playing rugby. You, as a researcher, are a passive observer. You don't do something intentional to your patients or ask them to do something if you think it might be harmful. Exposures are usually examined in the context of an epidemiology study.

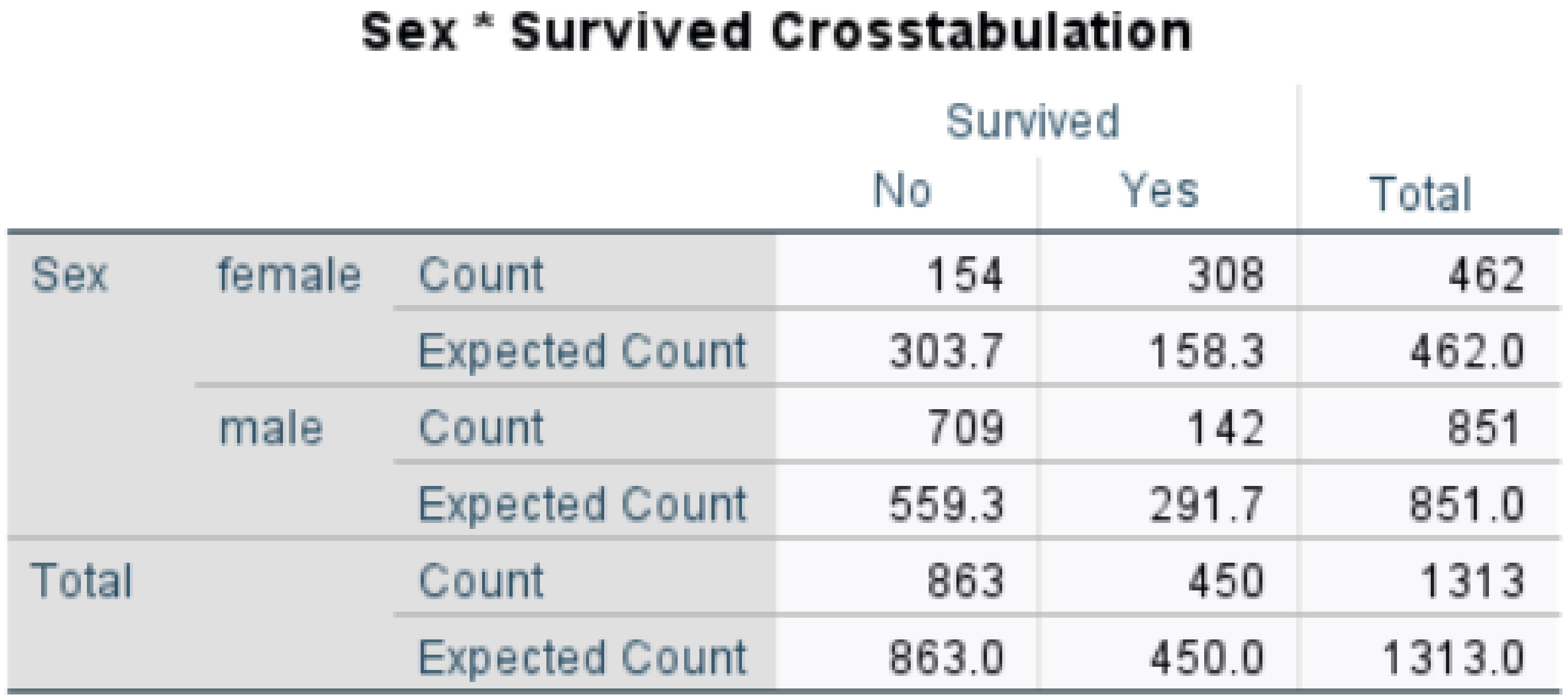
The control is sometimes a placebo or sometimes the usual standard of care which you want to compare to a treatment. In an exposure study, the control is the lack of an exposure.

Now there are not always easy distinctions between treatments and exposures. Is the use of CBD oil a treatment or an exposure?

Don't worry too much about the distinctions for now. The choice of what statistic to use is not influenced by whether you are studying a treatment or an exposure. The research design changes a lot, perhaps, but the same t-test that you use to compare a treatment mean to a control mean is the one you would use if you have an exposure mean instead.

Comparison of a binary outcome

$$X^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$



Sex * Survived Crosstabulation

			Survived		
			No	Yes	Total
Sex	female	Count	154	308	462
		Expected Count	303.7	158.3	462.0
	male	Count	709	142	851
		Expected Count	559.3	291.7	851.0
Total		Count	863	450	1313
		Expected Count	863.0	450.0	1313.0

Figure 5: Counts of dead and survived by sex with expected counts

Speaker notes

Speaker notes

The classic test for comparing a binary outcome in a treatment versus a control is the chi-squared test. Lay out the results in a two by two table. Here is the example of mortality versus sex that I discussed earlier. The chi-squared test compares the observed count (Oij) to the expected count (Eij). The expected count is the number dying or surviving, assuming that there is no difference between the treatment and control.

The overall survival rate on the Titanic across all passengers is 450 / 1,313 or 34%. If that rate applied equally to men and women, you would expect to see 158.3 survivors among the women. We observed a lot more survivors, 308. If the rate applied equally to men, you'd expect to see 291.7 survivors. This is a lot more than the 142 that actually did survive. You can do similar calculations for the expected number of deaths in each group.

In all four cells there is a large discrepancy between the observed counts and what you'd expect if the survival and mortality rates were the same for both men and women. This leads to a large value of the test statistic and you would reject the hypothesis that men and women were equally likely to survive.

Alternative approach, the odds ratio

```
        Died  Survived
Females  154       308
Males    709       142

Survival odds for Females 2 to 1 in favor (308 / 154).
Survival odds for Males 5 to 1 against (142/ 709).

Odds ratio = (2/1) / (1/5) = 10

95% CI (7.7, 13)
```

Speaker notes

The odds ratio is a common way to describe changes in risk. The odds of survival for women was 2 to 1 in favor. The ratio of survivors (308) to deaths (154) is exactly 2, so you can cite the odds as 2 to 1 in favor of survival. In men, the ratio of survivors (142) to deaths (709) is about 1/5. I am rounding a bit here for simplicity.

The odds ratio compares the 2 to 1 odds in favor to the 5 to 1 odds against to get 10. Women fared 10 times better than men because their odds of survival were 10 times better.

Alternative approach, relative risk

```
        Died  Survived
Females  154       308
Males    709       142

Survival probability for 66.7%.
Survival probability for Males 16.7%.

Relative risk = 0.667 / 0.167 = 4

95% CI (3.4, 4.7)
```

Which is the better measure?

- Two schools of thought
    - Relative risk is better
        - More natural interpretation
    - Odds ratio is better - Symmetric with respect to outcome
- Cannot use relative risk for certain datasets

Both are inferior to absolute risk reduction

```
        Died  Survived
Females  154       308
Males    709       142

Survival probability for 66.7%.
Survival probability for Males 16.7%.

Absolute risk reduction = 0.667 - 0.167 = 0.5

95% CI (0.45, 0.55)
```

Comparison of multinomial outcome

- Multinomial = 3 or more categories
- Beyond the scope of this class
    - Multinomial logistic regression
    - Ordinal logistic regression

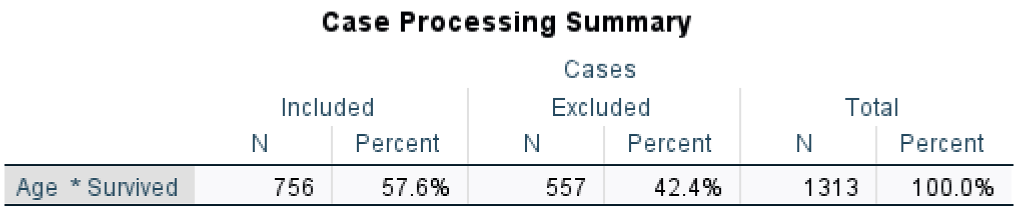
Comparison of a continuous outcome

- Two cases
  - Independent (unpaired) samples
  - Paired samples

Two sample test

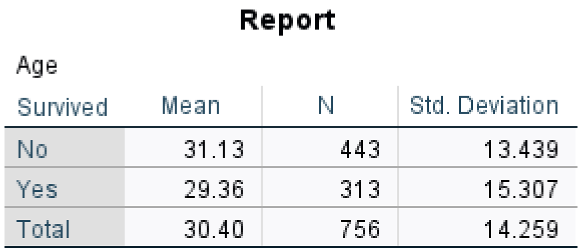
- Is $(\bar{X}_1 - \bar{X}_2)$ close to zero?
- How much sampling error?
  - $S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Comparison of ages of deaths/survivors



Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Age * Survived	756	57.6%	557	42.4%	1313	100.0%



Report

Age

Survived	Mean	N	Std. Deviation
No	31.13	443	13.439
Yes	29.36	313	15.307
Total	30.40	756	14.259

95% CI (-0.3, 3.8)

Paired samples

```
Room Before  After
 121   11.8   10.1
 125    7.1    3.8
 163    8.2    7.2
 218   10.1   10.5
 233   10.8    8.3
 264   14     12
 324   14.6   12.1
 325   14     13.7
```

Average change

```
Room Before  After Change
 121   11.8   10.1   -1.7
 125    7.1    3.8   -3.3
 163    8.2    7.2   -1.0
 218   10.1   10.5    0.4
 233   10.8    8.3   -2.5
 264   14     12     -2.0
 324   14.6   12.1   -2.5
 325   14     13.7   -0.3
```

$$\bar{D} = -1.61,\ S_D = 1.24$$

95% CI (-2.65, -0.58)

Assumptions for t-tests

- t-tests require two or more assumptions
  - Patients are independent
  - Outcome is normally distributed
  - For two sample t-test, equal variation

Nonparametric test

- Uses ranks of the data
- Does not rely on normality assumption
- Does not rely on Central Limit Theorem

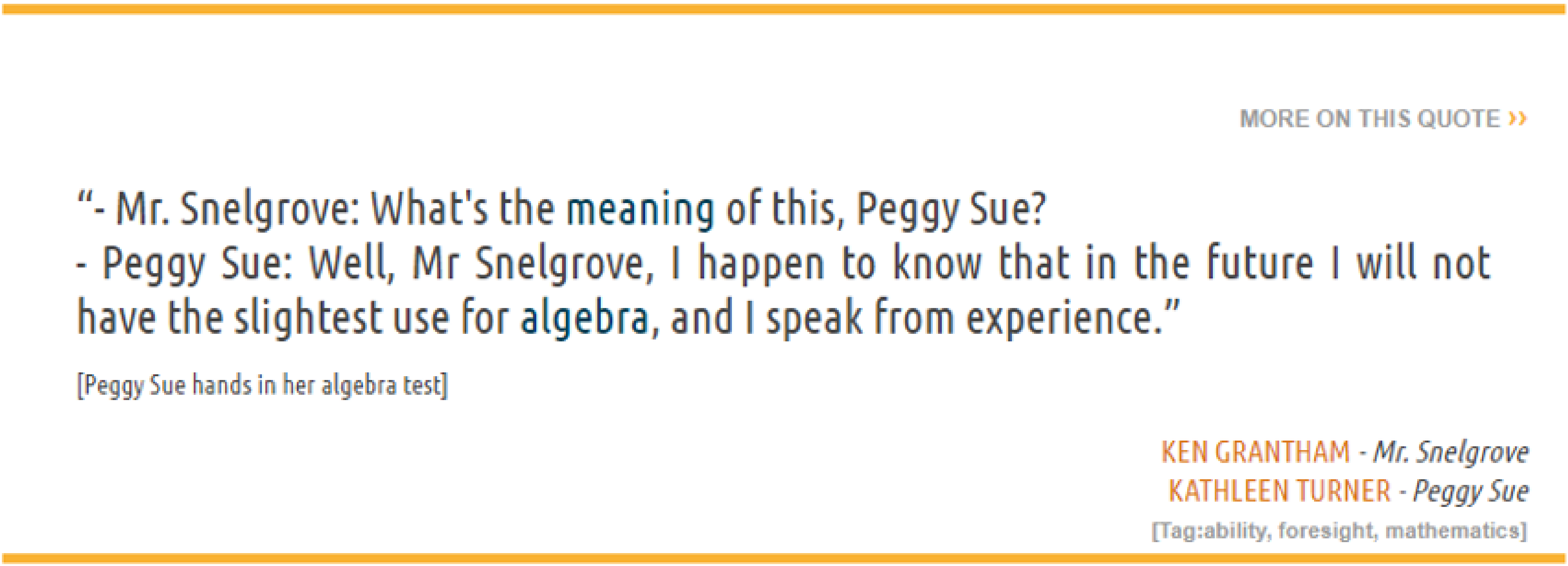
Wilcoxon signed rank test

				Absolute	
Room	Before	After	Change	Change	Rank
121	11.8	10.1	-1.7	1.7	4
125	7.1	3.8	-3.3	3.3	8
163	8.2	7.2	-1.0	1.0	3
218	10.1	10.5	0.4	0.4	2
233	10.8	8.3	-2.5	2.5	6/7
264	14	12	-2.0	2.0	5
324	14.6	12.1	-2.5	2.5	6/7
325	14	13.7	-0.3	0.3	1

p = 0.023

Criticisms of nonparametric tests

- Not easy to get confidence intervals
- Difficult to do risk adjustments



MORE ON THIS QUOTE ››

“- Mr. Snelgrove: What's the meaning of this, Peggy Sue?
- Peggy Sue: Well, Mr Snelgrove, I happen to know that in the future I will not have the slightest use for algebra, and I speak from experience.”

[Peggy Sue hands in her algebra test]

KEN GRANTHAM - *Mr. Snelgrove*
KATHLEEN TURNER - *Peggy Sue*
[Tag:ability, foresight, mathematics]

Figure 6: Quote from “Peggy Sue Got Married

Speaker notes

I want to get a quick feel for your background and interests. Here's a quote from a romantic comedy starring Kathleen Turner from 1986. A forty year old woman, played by Kathleen Turner, travels back in time to her high school senior year, 1960. She has an amusing interchange with her high school math teacher.

"I happen to know that in the future, I will not have the slightest use for algebra, and I speak from experience."

Think back to your high school algebra class.

Do you remember any important formulas from that class?

Did you hate, hate, hate high school algebra?

Did you love high school algebra?

Big question: Will you use high school algebra in your future?

Source: https://www.moviequotes.com/s-movie/peggy-sue-got-married/

Pop quiz

- From high school algebra.
  - Pythagorean theorem
    - ?
  - Quadratic formula
    - ?
  - Equation for a straight line
    - ?

Pop quiz answers

- From high school algebra.
  - Pythagorean theorem
    - $a^2 + b^2 = c^2$
  - Quadratic formula
    - $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$
  - Formula for a straight line
    - $y = mx + b$

Equation of a straight line

- $y = mx + b$
  - $m =$ slope $= \triangle y / \triangle x$
  - $b =$ y-intercept

In linear regression

- y: dependent variable
- x: independent variable
- Slope: estimated average change in y when x increases by one unit.
- Intercept: estimated average value of y when x equals zero.

Example: does mother’s age affect duration of breast feeding?

- Study of breast feeding with pre-term infants
  - Difficulty: mother leaves hospital first

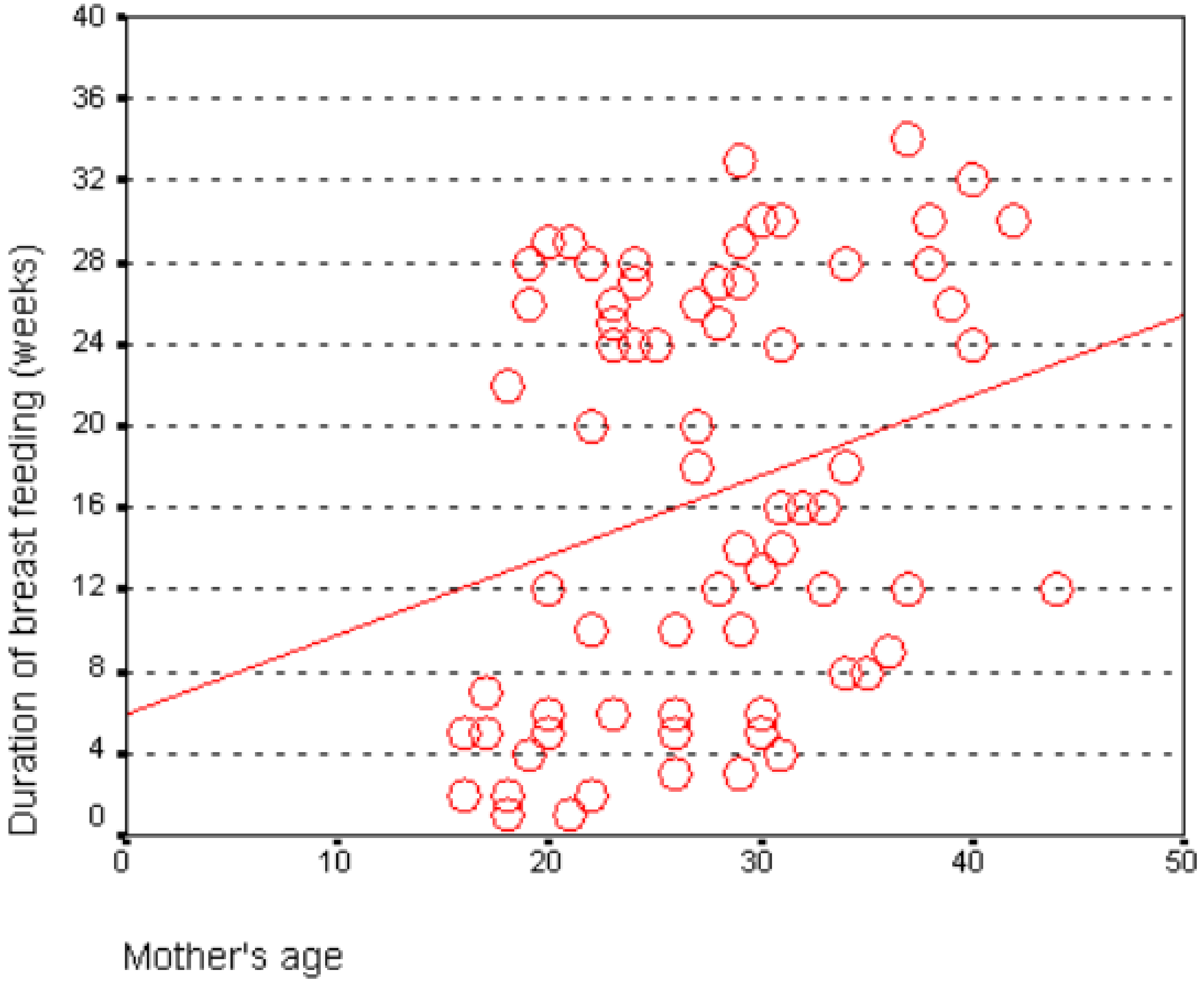
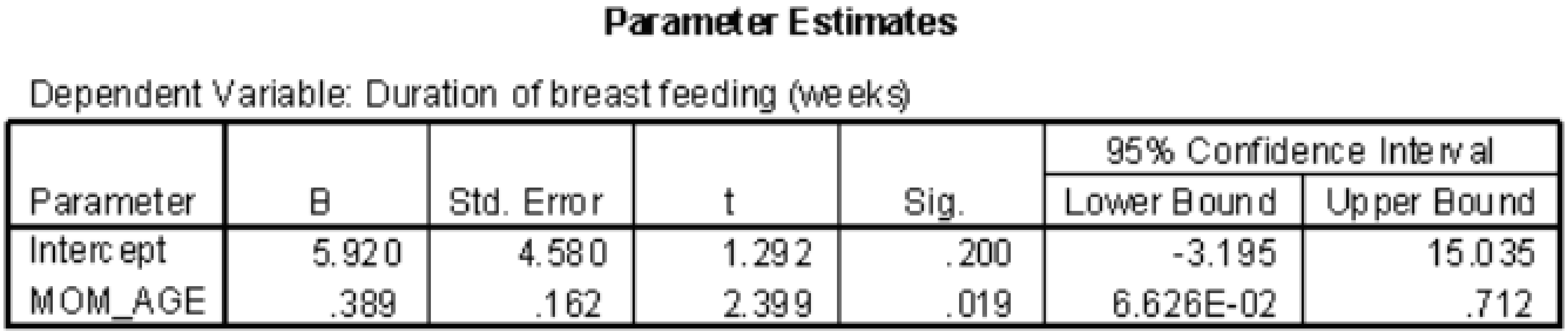


Figure 7: Scatterplot with regression line for age

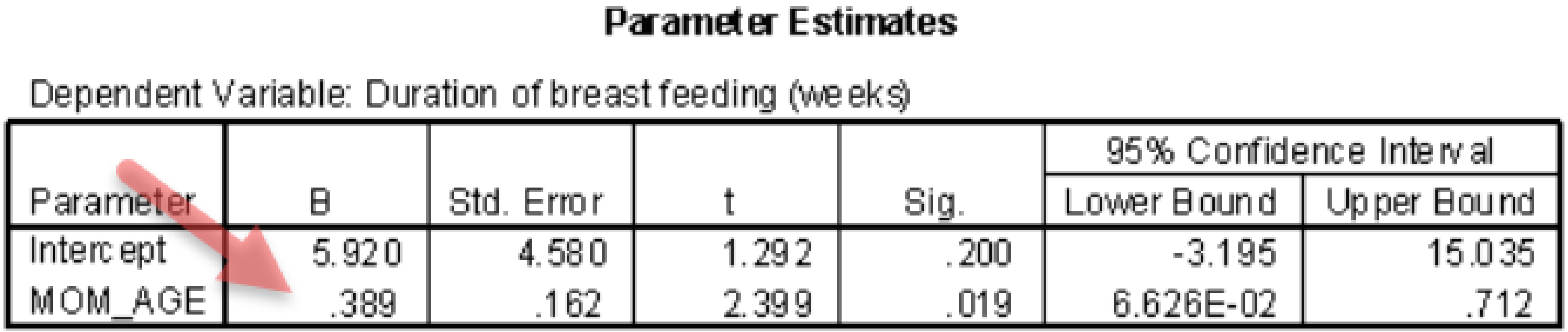


Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Figure 8: Linear regression output



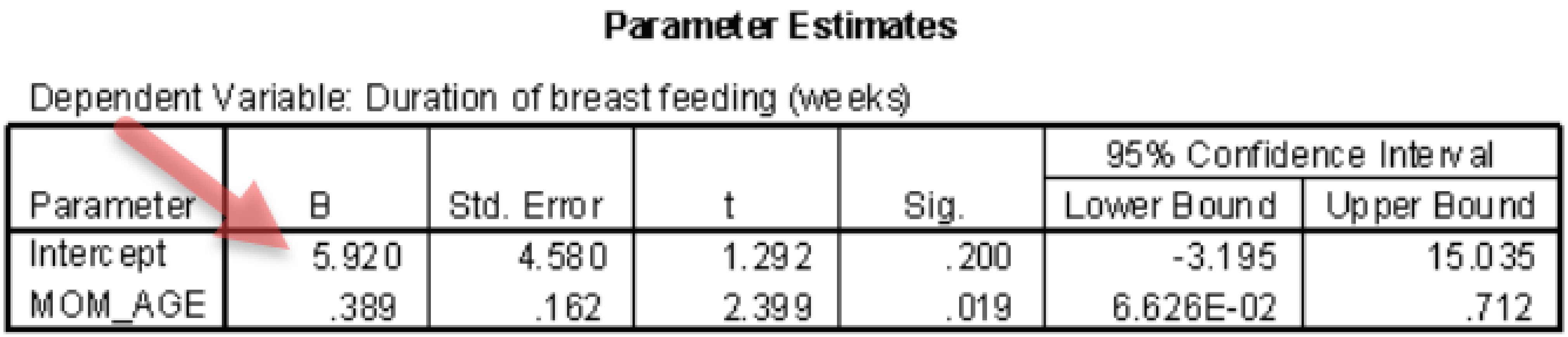
Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Figure 9: Linear regression output, slope

- Slope = 0.4
  - The estimated average duration of breast feeding increases by 0.4 weeks for every increase of one year in the mother’s age.



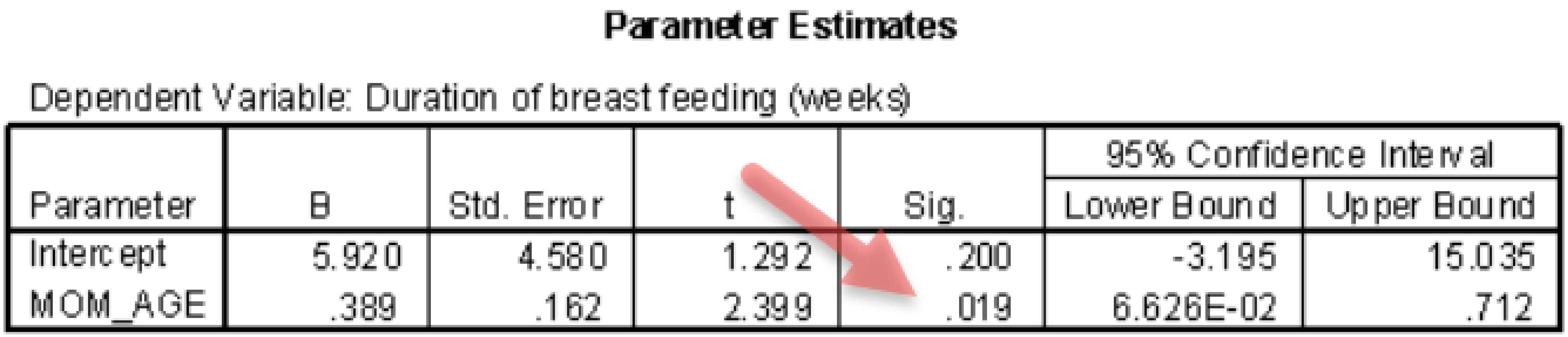
Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Figure 10: Linear regression output, intercept

- Intercept = 5.9
  - The estimated average duration of breast feeding is 5.9 weeks for a mother with age = 0.
  - Clearly an inappropriate extrapolation



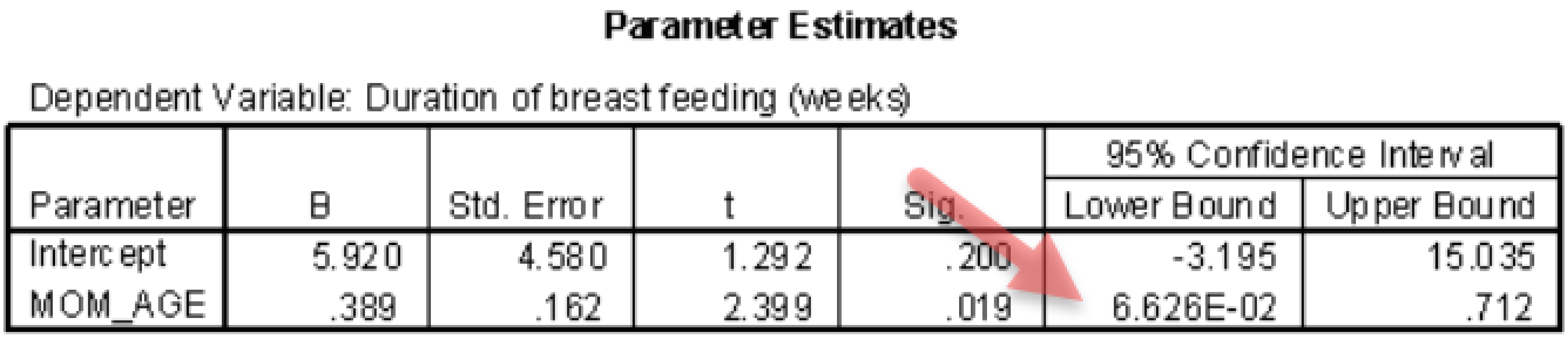
Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Figure 11: Linear regression output, p-value

- p-value=0.019
  - Reject the null hypothesis and conclude that there is a positive relationship between mother’s age and duration of breast feeding.



Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5.920	4.580	1.292	.200	-3.195	15.035
MOM_AGE	.389	.162	2.399	.019	6.626E-02	.712

Figure 12: Linear regression output, confidence interval

- 95% Confidence interval (0.066 to 0.71)
  - Note 6.626E-02 $= 6.626 \times 10^{-2}$
  - You are 95% confident that the true regression slope is positive.

Example: does treatment affect duration of breast feeding?

- Both groups: encourage breast feeding when mom is in hospital
  - Intervention: feed infants through ng tube when mom is away
  - Control: Feeding using bottles when mom is away

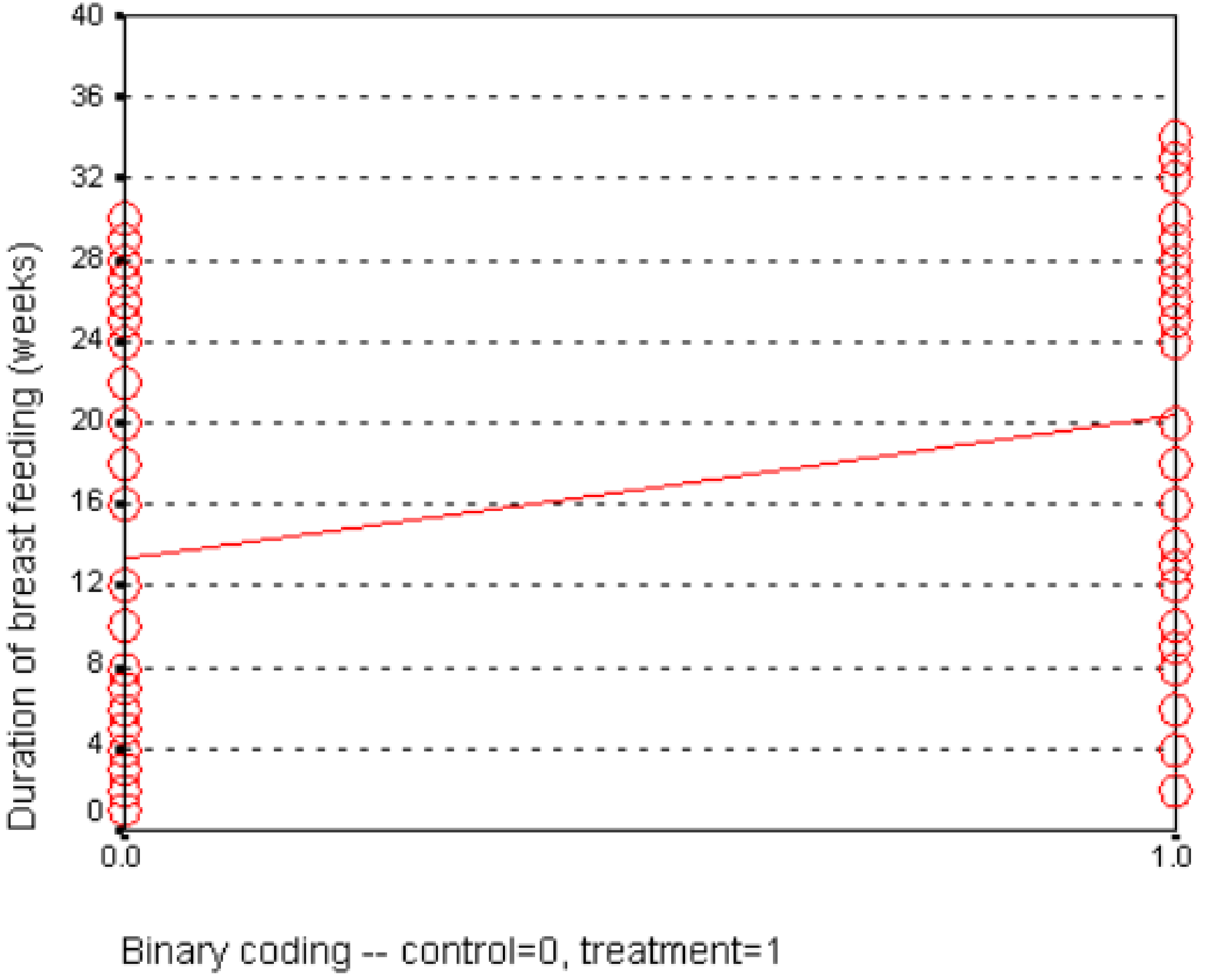
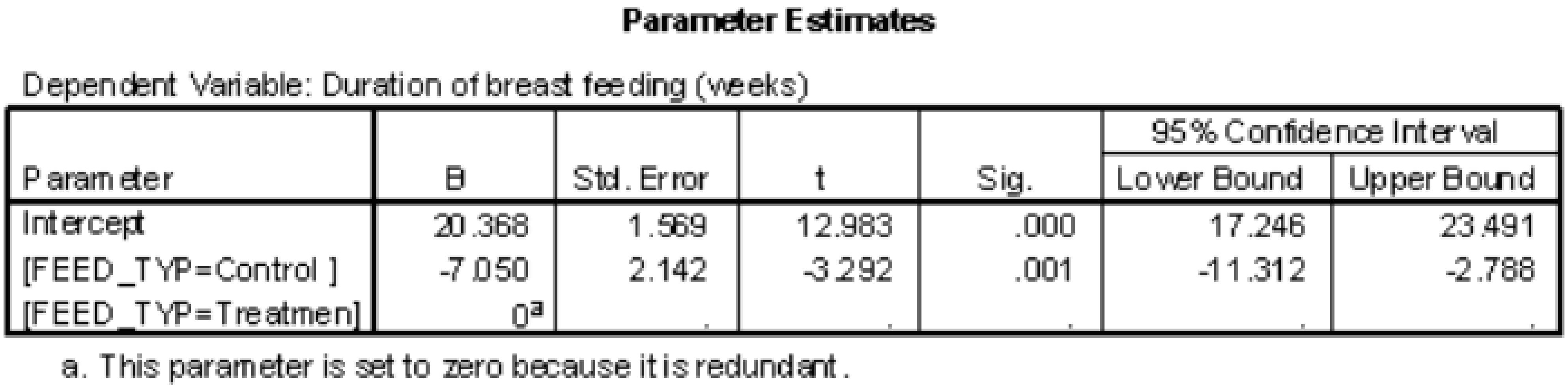


Figure 13: Scatterplot with regression line for treatment=1



Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

					95% Confidence Interval	
Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
Intercept	20.368	1.569	12.983	.000	17.246	23.491
[FEED_TYP=Control]	-7.050	2.142	-3.292	.001	-11.312	-2.788
[FEED_TYP=Treatmen]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 14: Linear regression output, treatment

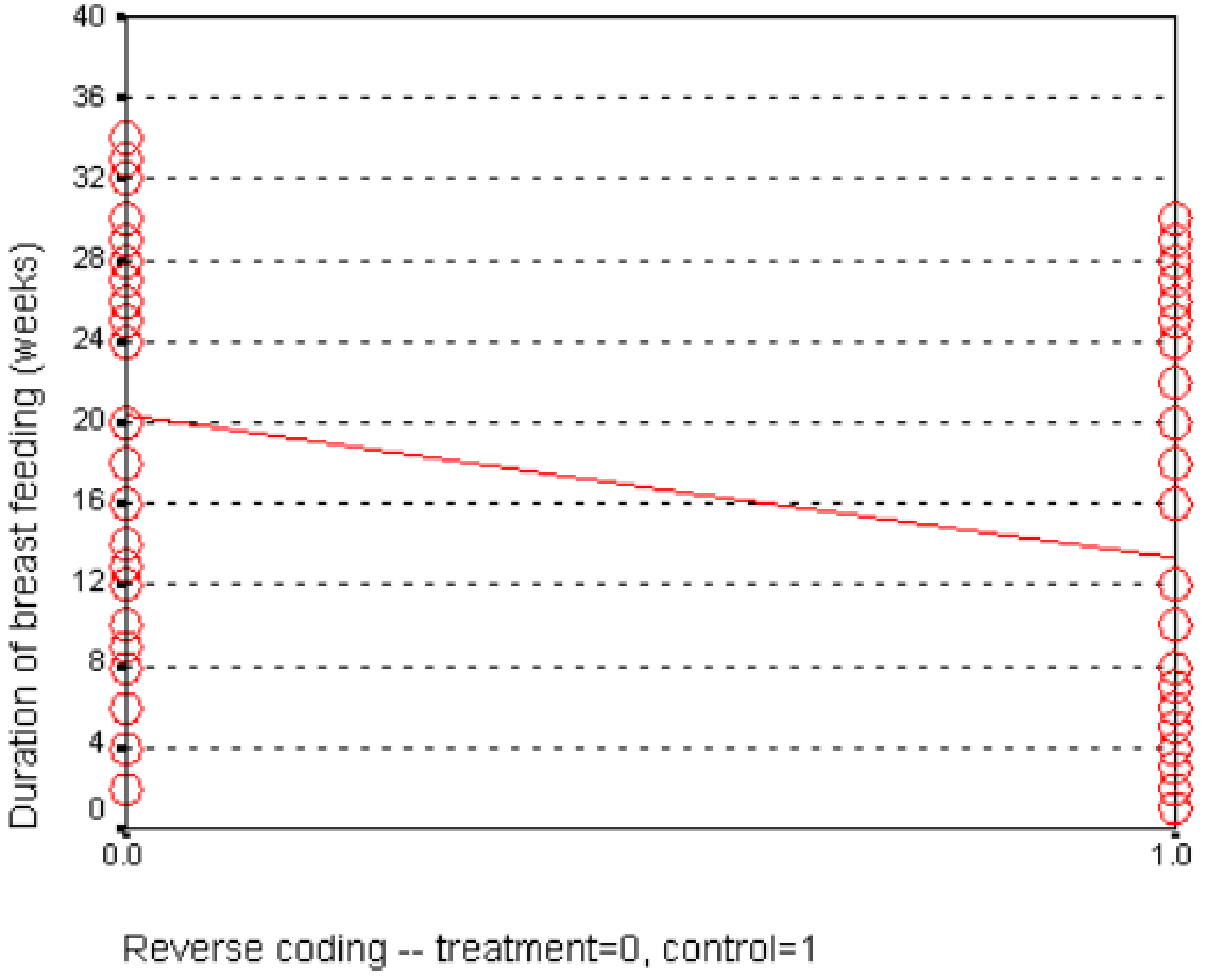


Figure 15: Scatterplot with regression line for control=1

Linear regression with two independent variables

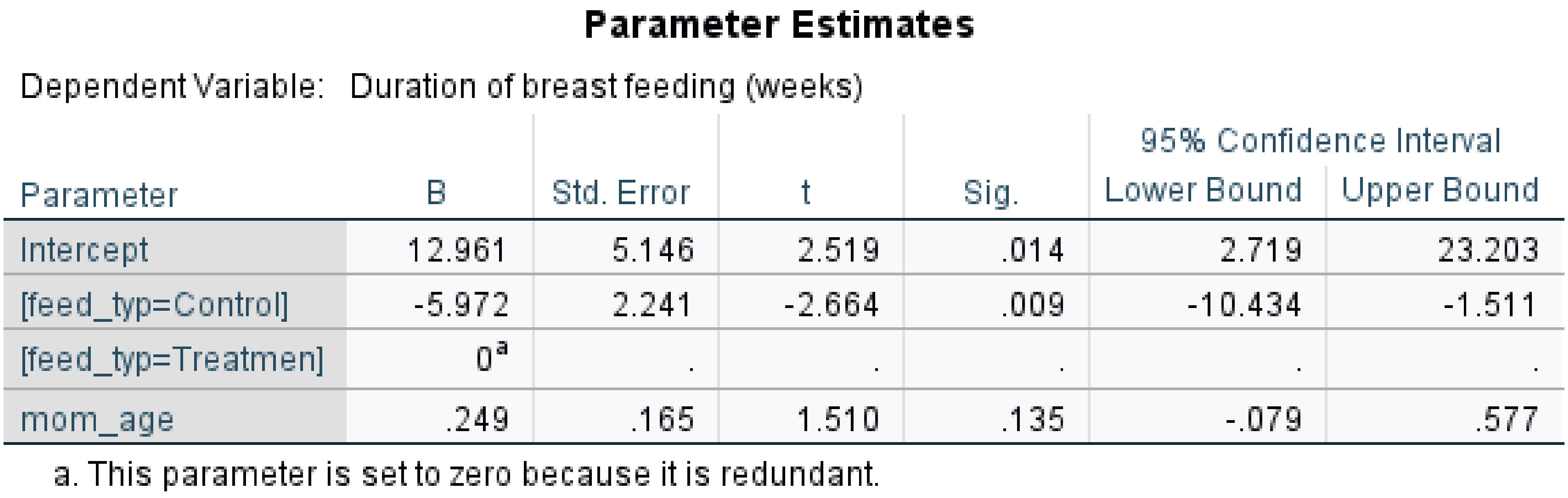
- Intercept
- Slope for first independent variable
- Slope for second independent variable

Interpretation of intercept and slopes

- Intercept: estimated average value of y when x1 and x2 both equal zero.
- Slope for x1: estimated average change in y when x increases by one unit **and x2 is held constant**.
- Slope for x2: similar interpretation

Adjusting for covariate imbalance

- Covariate: variable not of direct interest in the research
  - but has to be accounted for to draw valid conclusions
- Covariate imbalance: a difference in average levels of the covariate between treatment and control
  - Threat to the validity of the research
- Example: average age of mothers
  - 25 in control group, 29 in treatment group
- Covariate imbalance not quite same as confounding



Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	12.961	5.146	2.519	.014	2.719	23.203
[feed_typ=Control]	-5.972	2.241	-2.664	.009	-10.434	-1.511
[feed_typ=Treatmen]	0[a]	.	.	.	.	.
mom_age	.249	.165	1.510	.135	-.079	.577

[a] This parameter is set to zero because it is redundant.

Figure 16: Multiple linear regression output

Adjusted means

- Unadjusted
  - Treatment: $12.961 + 0.249 \times 29 = 20.2$
  - Control: $12.961 - 5.972 + 0.249 \times 25 = 13.2$
- Adjusted
  - Treatment: $12.961 + 0.249 \times 27 = 19.7$
  - Control: $12.961 - 5.972 + 0.249 \times 27 = 13.7$

Small group exercises

- Group 1: Effect of sex and height on fev
- Group 2: Effect of smoking and age on fev Examples in the medical literature

Joke about prediction models

- Risks during surgery
  - P[death] = 0.6
- If risk doubles
  - P[death] = 1.2

Speaker notes

A doctor is advising her patient about the risks of an upcoming surgery. She warned that the probability that the patient would die during surgery was 60%. Then she looked up an said, no wait, the risk is twice as big in your demographic group. The chances that you will die during surgery is actually 120%. The patient seemed a bit confused. I know what a 100% risk of mortality would be—I'm a goner. But what would a 120% risk of mortality be? The doctor replied, that is a fate worse than death.

Logistic regression

- Binary outcome
- Linear on a log odds scale

Speaker notes

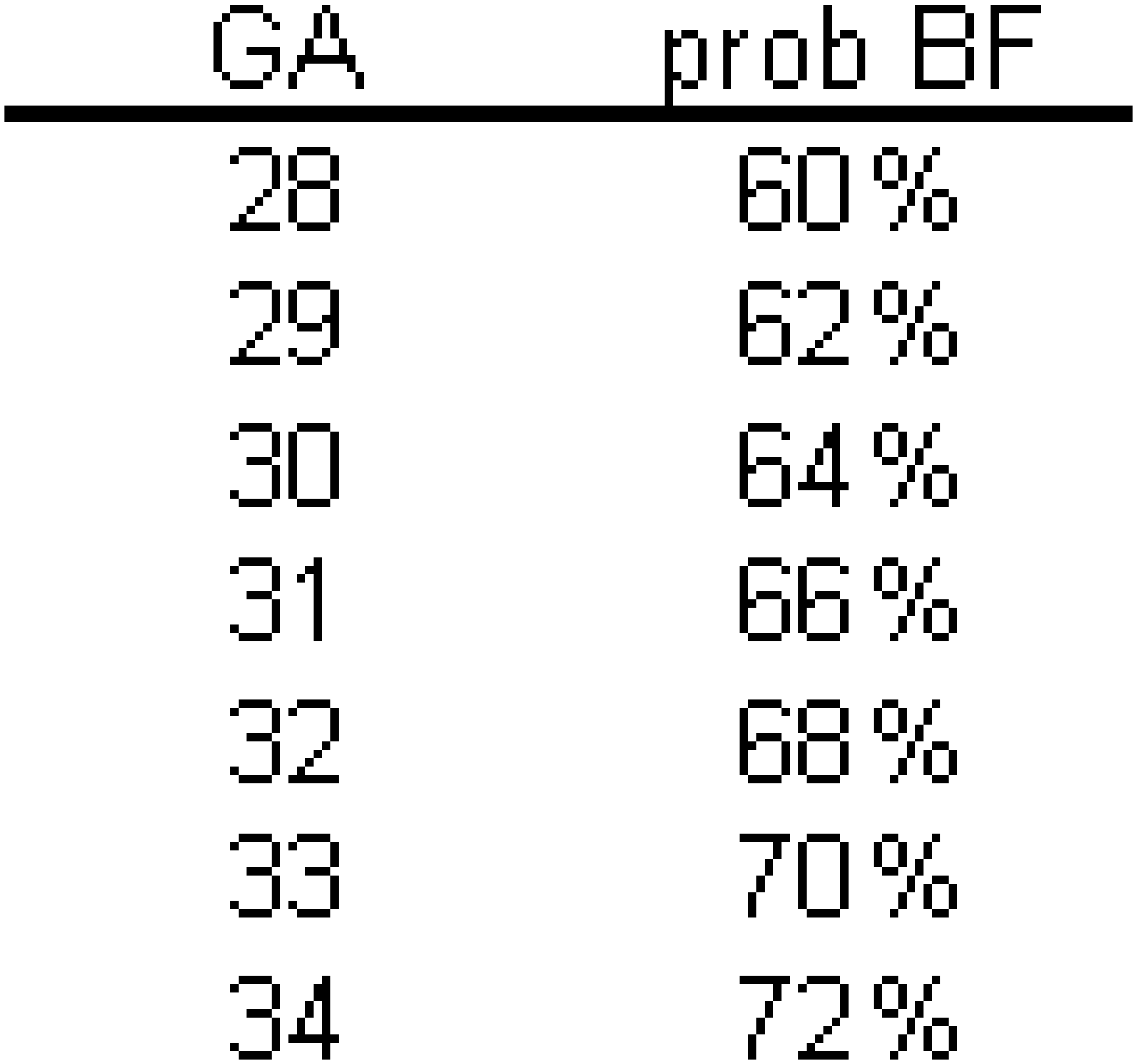
The logistic regression model is a model that uses a binary (two possible values) outcome variable. Examples of a binary variable are mortality (live/dead), and morbidity (healthy/diseased). Sometimes you might take a continuous outcome and convert it into a binary outcome. For example, you might be interested in the length of stay in the hospital for mothers during an unremarkable delivery. A binary outcome might compare mothers who were discharged within 48 hours versus mothers discharged more than 48 hours.

The covariates in a logistic regression model represent variables that might be associated with the outcome variable. Covariates can be either continuous or categorical variables.

For binary outcomes, you might find it helpful to code the variable using indicator variables. An indicator variable equals either zero or one. Use the value of one to represent the presence of a condition and zero to represent absence of that condition. As an example, let 1=diseased, 0=healthy.

Indicator variables have many nice mathematical properties. One simple property is that the average of an indicator variable equals the observed probability in your data of the specific condition for that variable.

A logistic regression model examines the relationship between one or more independent variable and the log odds of your binary outcome variable. Log odds seem like a complex way to describe your data, but when you are dealing with probabilities, this approach leads to the simplest description of your data that is consistent with the rules of probability.



GA	prob BF
28	60%
29	62%
30	64%
31	66%
32	68%
33	70%
34	72%

Figure 17: A linear trend in probability

prob BF = 4 + 2*GA

Speaker notes

Let’s consider an artificial data example where we collect data on the gestational age of infants (GA), which is a continuous variable, and the probability that these infants will be breast feeding at discharge from the hospital (BF), which is a binary variable. We expect an increasing trend in the probability of BF as GA increases. Premature infants are usually sicker and they have to stay in the hospital longer. Both of these present obstacles to BF.

A linear model would presume that the probability of BF increases as a linear function of GA. You can represent a linear function algebraically as

prob BF = a + b*GA

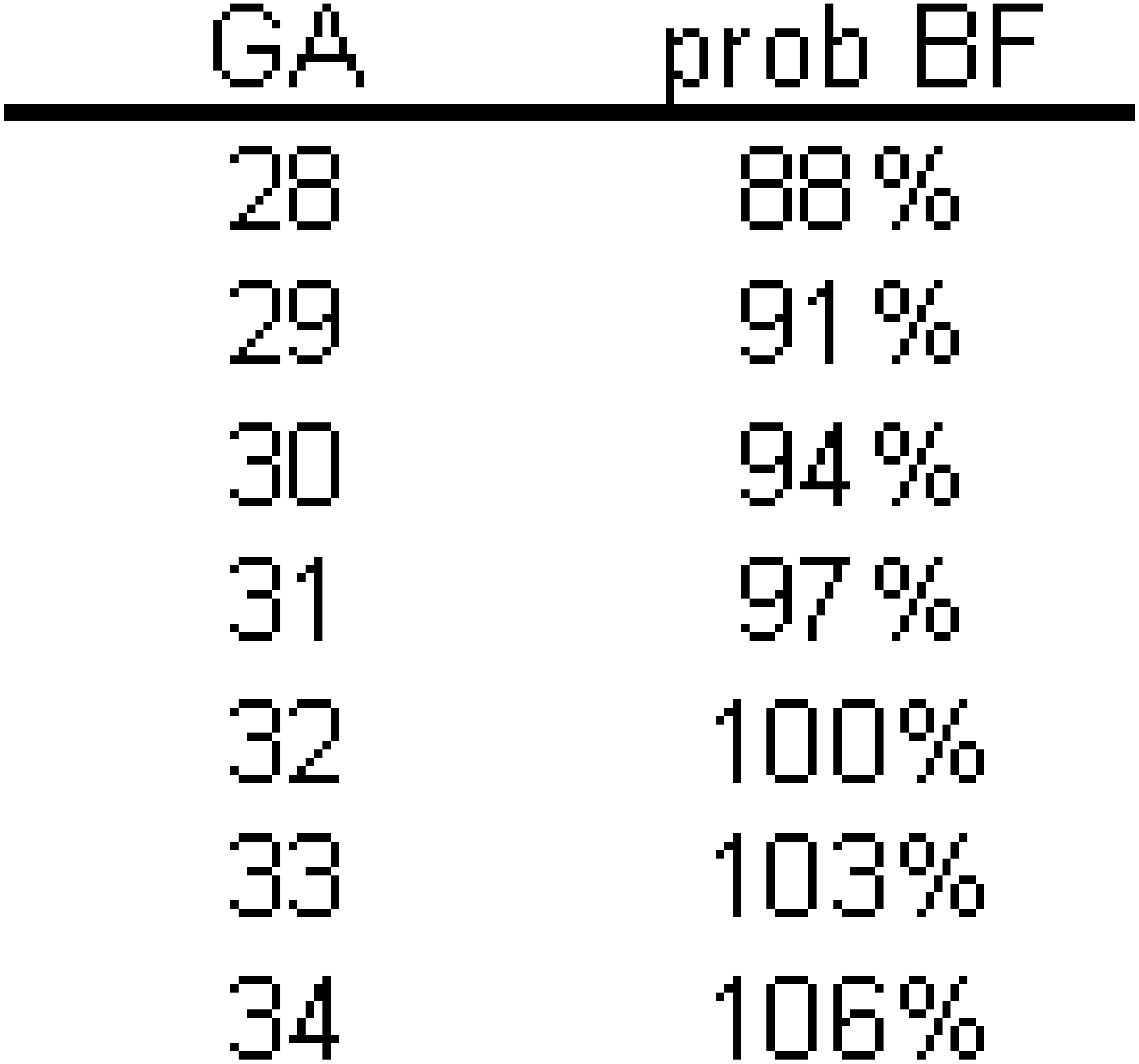
This means that each unit increase in GA would add b percentage points to the probability of BF. The table shown below gives an example of a linear function.

This table represents the linear function

prob BF = 4 + 2*GA

which means that you can get the probability of BF by doubling GA and adding 4. So an infant with a gestational age of 30 would have a probability of 4+2*30 = 64.

A simple interpretation of this model is that each additional week of GA adds an extra 2% to the probability of BF. We could call this an additive probability model.



GA	prob BF
28	88 %
29	91 %
30	94 %
31	97 %
32	100%
33	103%
34	106%

Figure 18: A bad linear trend in probability

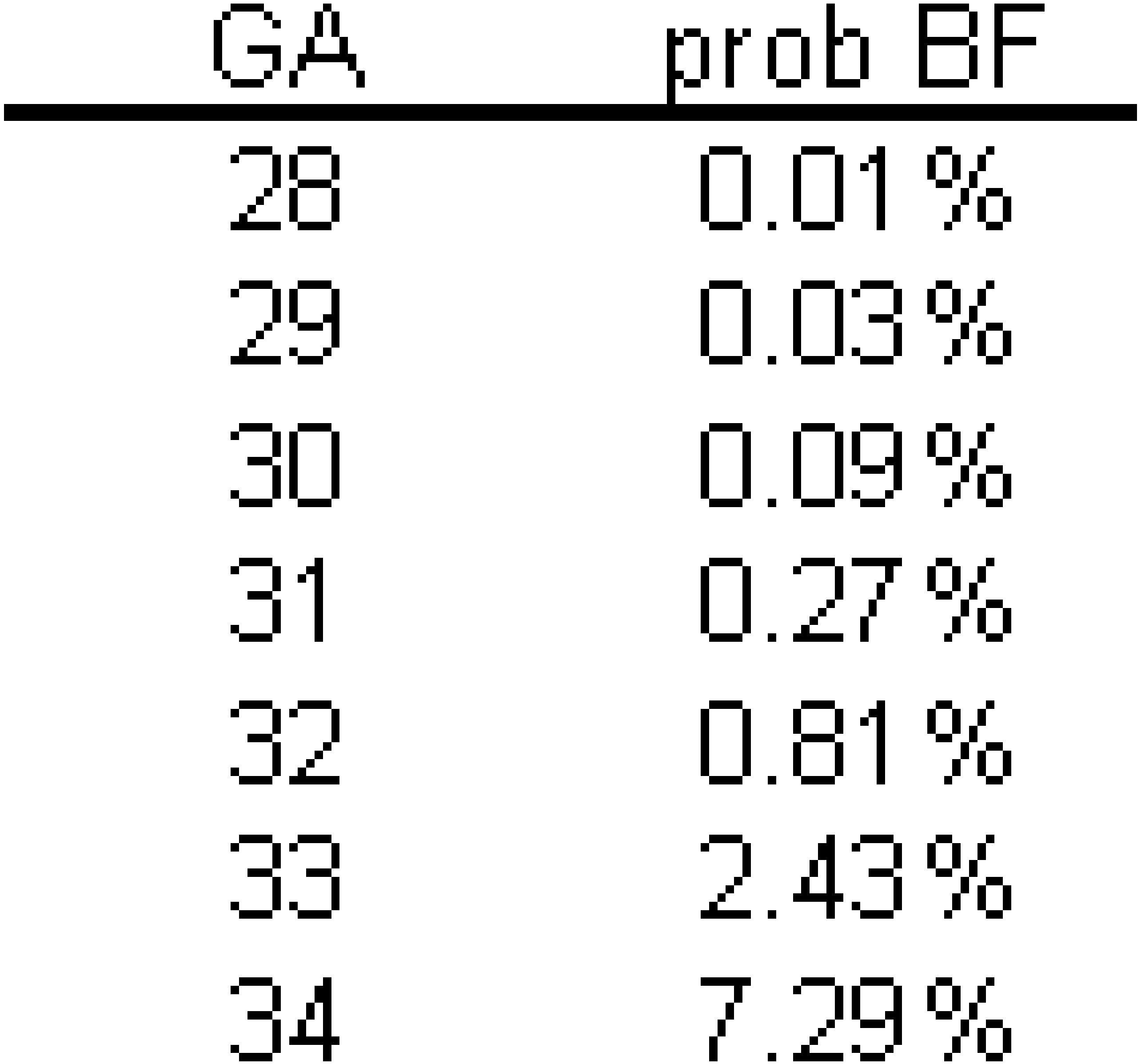
Speaker notes

I’m not an expert on BF; what little experience I’ve had with the topic occurred over 65 years ago. But I do know that an additive probability model tends to have problems when you get probabilities close to 0% or 100%. Let’s change the linear model slightly to the following:

prob BF = 4 + 3*GA

This model would produce the following table of probabilities.

You may find it difficult to explain what a probability of 106% means. This is a reason to avoid using a additive model for estimating probabilities. In particular, try to avoid using an additive model unless you have good reason to expect that all of your estimated probabilities will be between 20% and 80%.



GA	prob BF
28	0.01 %
29	0.03 %
30	0.09 %
31	0.27 %
32	0.81 %
33	2.43 %
34	7.29 %

Figure 19: Multiplicative trend in probabilities

Speaker notes

It’s worthwhile to consider a different model here, a multiplicative model for probability, even though it suffers from the same problems as the additive model.

In a multiplicative model, you change the probabilities by multiplying rather than adding. Here’s a simple example.

In this example, each extra week of GA produces a tripling in the probability of BF. Contrast this to the linear models shown above, where each extra week of GA adds 2% or 3% to the probability of BF.

A multiplicative model can’t produce any probabilities less than 0%, but it’s pretty easy to get a probability bigger than 100%. A multiplicative model for probability is actually quite attractive, as long as you have good reason to expect that all of the probabilities are small, say less than 20%.

Using odds

- Three to one in favor of victory
    - Expect three wins for every loss
- Four to one odds against victory
    - Expect four losses for every win
- Odds = Prob / (1- Prob)
- Prob = Odds / (Odds + 1)

Speaker notes

Speaker notes

The relationship between odds and probability Another approach is to try to model the odds rather than the probability of BF. You see odds mentioned quite frequently in gambling contexts. If the odds are three to one in favor of your favorite football team, that means you would expect a win to occur about three times as often as a loss. If the odds are four to one against your team, you would expect a loss to occur about four times as often as a win.

You need to be careful with odds. Sometimes the odds represent the odds in favor of winning and sometimes they represent the odds against winning. Usually it is pretty clear from the context. When you are told that your odds of winning the lottery are a million to one, you know that this means that you would expect to having a losing ticket about a million times more often than you would expect to hit the jackpot.

It’s easy to convert odds into probabilities and vice versa. With odds of three to one in favor, you would expect to see roughly three wins and only one loss out of every four attempts. In other words, your probability for winning is 0.75.

If you expect the probability of winning to be 20%, you would expect to see roughly one win and four losses out of every five attempts. In other words, your odds are 4 to 1 against.

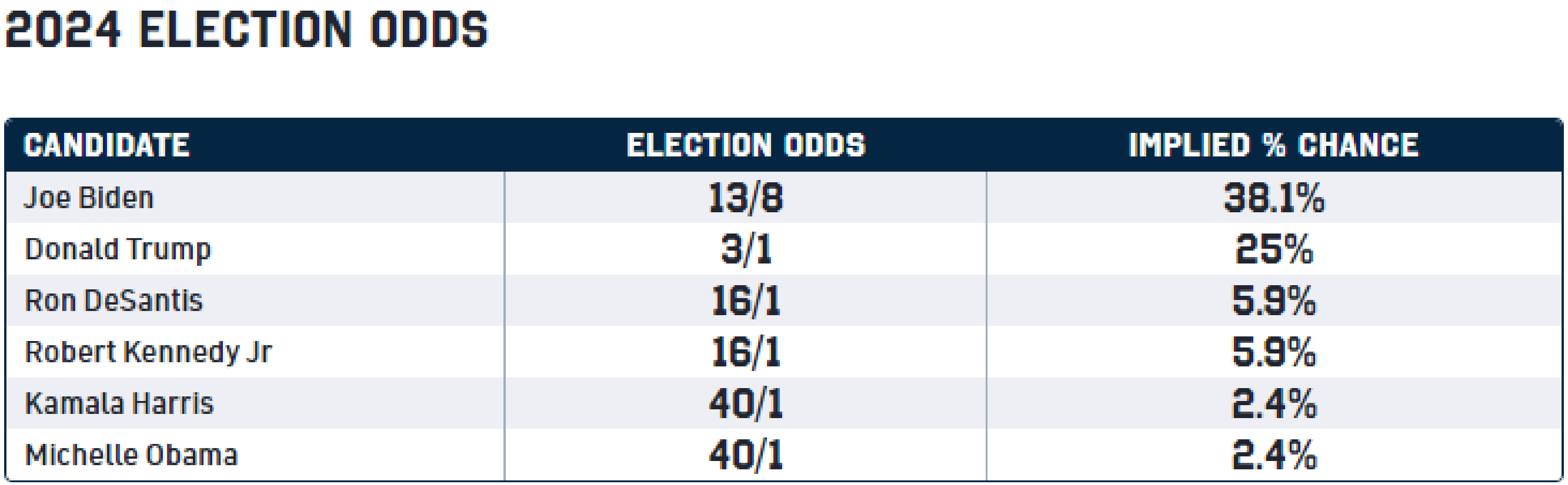
The formulas for conversion are

odds = prob / (1-prob)

and

prob = odds / (1+odds).

In medicine and epidemiology, when an event is less likely to happen and more likely not to happen, we represent the odds as a value less than one. So odds of four to one against an event would be represented by the fraction 1/5 or 0.2. When an event is more likely to happen than not, we represent the odds as a value greater than one. So odds of three to one in favor of an event would be represented simply as an odds of 3. With this convention, odds are bounded below by zero, but have no upper bound.



2024 ELECTION ODDS

CANDIDATE	ELECTION ODDS	IMPLIED % CHANCE
Joe Biden	13/8	38.1%
Donald Trump	3/1	25%
Ron DeSantis	16/1	5.9%
Robert Kennedy Jr	16/1	5.9%
Kamala Harris	40/1	2.4%
Michelle Obama	40/1	2.4%

Odds for winning election to U.S. president in 2024

- Biden: $\frac{8/13}{1+8/13} = \frac{8}{21} = 0.381$
- Trump: $\frac{1/3}{1+1/3} = \frac{1}{4} = 0.25$
- DeSantis: $\frac{1/16}{1+1/16} = \frac{1}{17} = 0.059$

Speaker notes

Speaker notes

To convert from odds to probability, use the formula odds/(1+odds). You have to flip these around because 40 to 1 odds does not mean that Michelle Obama has 40 chances to win for every one chance of a loss.

Table downloaded from oddschecker.com

Probability of winning 2022 World Cup

```
Brazil: 30.8%
Argentina: 18.2%
France: 16.7%
Spain: 13.3%
England: `10%
Portugal: 7.7%
Netherlands: 5.3%
Croatia: 2.8%
```

```
Switzerland: 1.5%
Japan: 1.5%
Morocco: 1.2%
USA: 1.1%
Senegal: 1%
South Korea: 0.67%
Poland: 0.55%
Australia: 0.5%
```

Argentina:

$$\frac{0.182}{1-0.182} = 0.2225 \approx 2/9$$

France:

$$\frac{0.167}{1-0.167} = 0.2004 \approx 1/5$$

Speaker notes

These probabilities were computed from a table of odds posted at the beginning of the round of 16 for the football world cup. Convert these back to odds.

These odds were taken from a December 2, 2022 blog post on the DraftKings website.

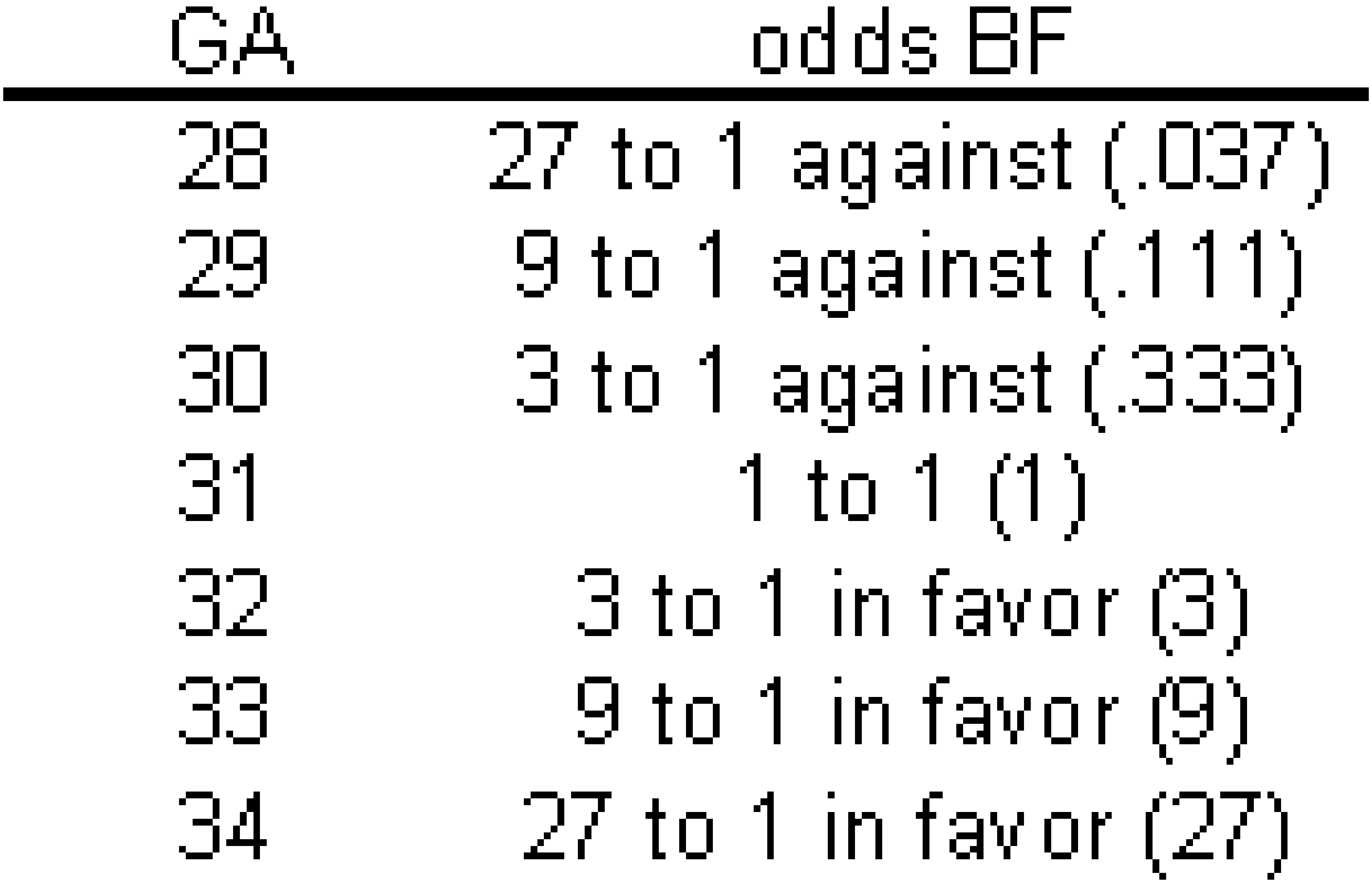
Odds against winning 2022 football World Cup

```
Brazil: 9 to 4
Argentina: 9 to 2
France: 5 to 1
Spain: 13 to 2
England: 9 to 1
Portugal: 12 to 1
Netherlands: 18 to 1
Croatia: 35 to 1
```

```
Switzerland: 65 to 1
Japan: 65 to 1
Morocco: 80 to 1
USA: 90 to 1
Senegal: 100 to 1
South Korea: 150 to 1
Poland: 180 to 1
Australia: 200 to 1
```

Speaker notes

Here are all the odds. Notice that the United States was rightfully given almost no chance of winning. But wait until the women’s football World Cup.



GA	odds BF
28	27 to 1 against (.037)
29	9 to 1 against (.111)
30	3 to 1 against (.333)
31	1 to 1 (1)
32	3 to 1 in favor (3)
33	9 to 1 in favor (9)
34	27 to 1 in favor (27)

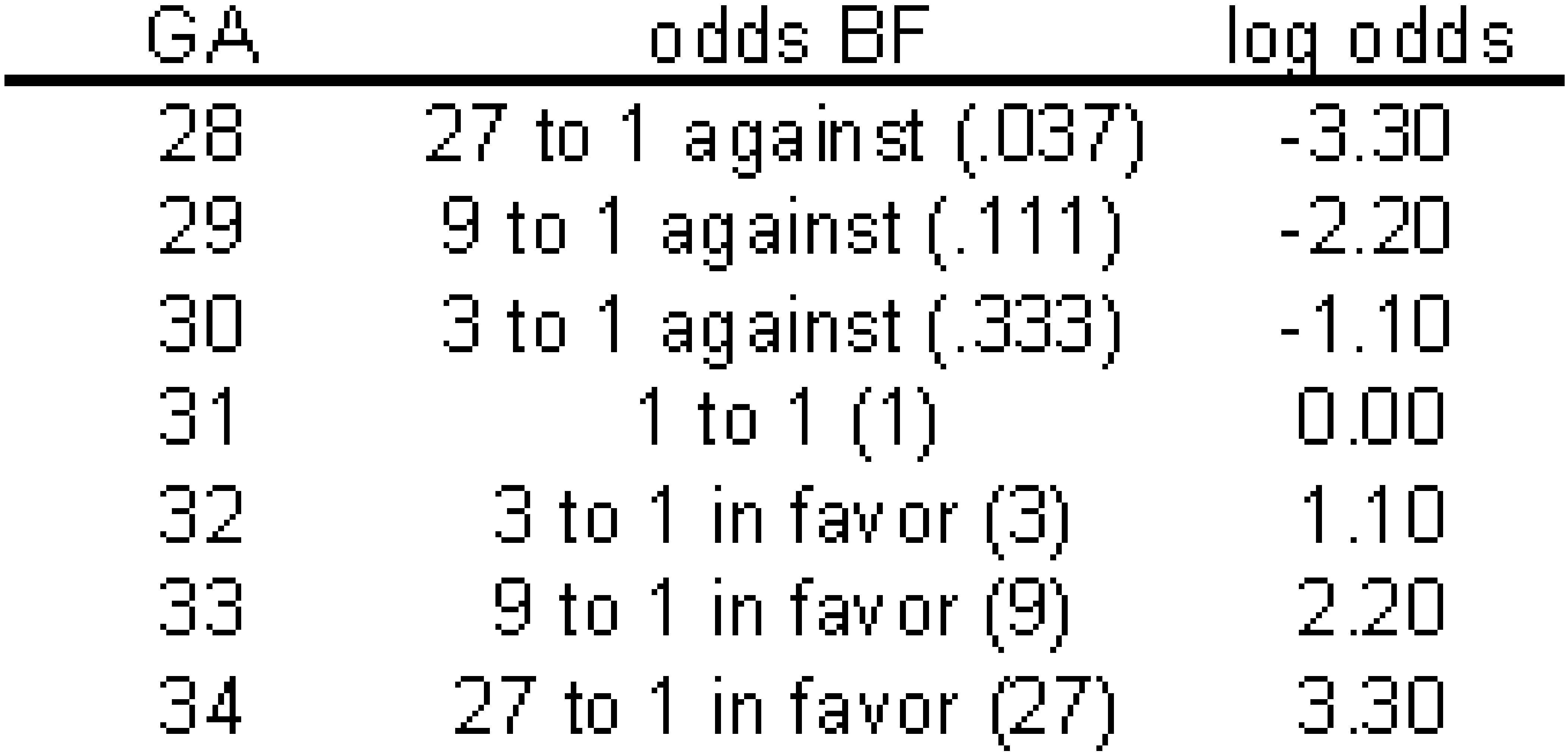
Figure 20: Multiplicative trend for odds

Speaker notes

Speaker notes

Let’s consider a multiplicative model for the odds (not the probability) of BF.

This model implies that each additional week of GA triples the odds of BF. A multiplicative model for odds is nice because it can’t produce any meaningless estimates.



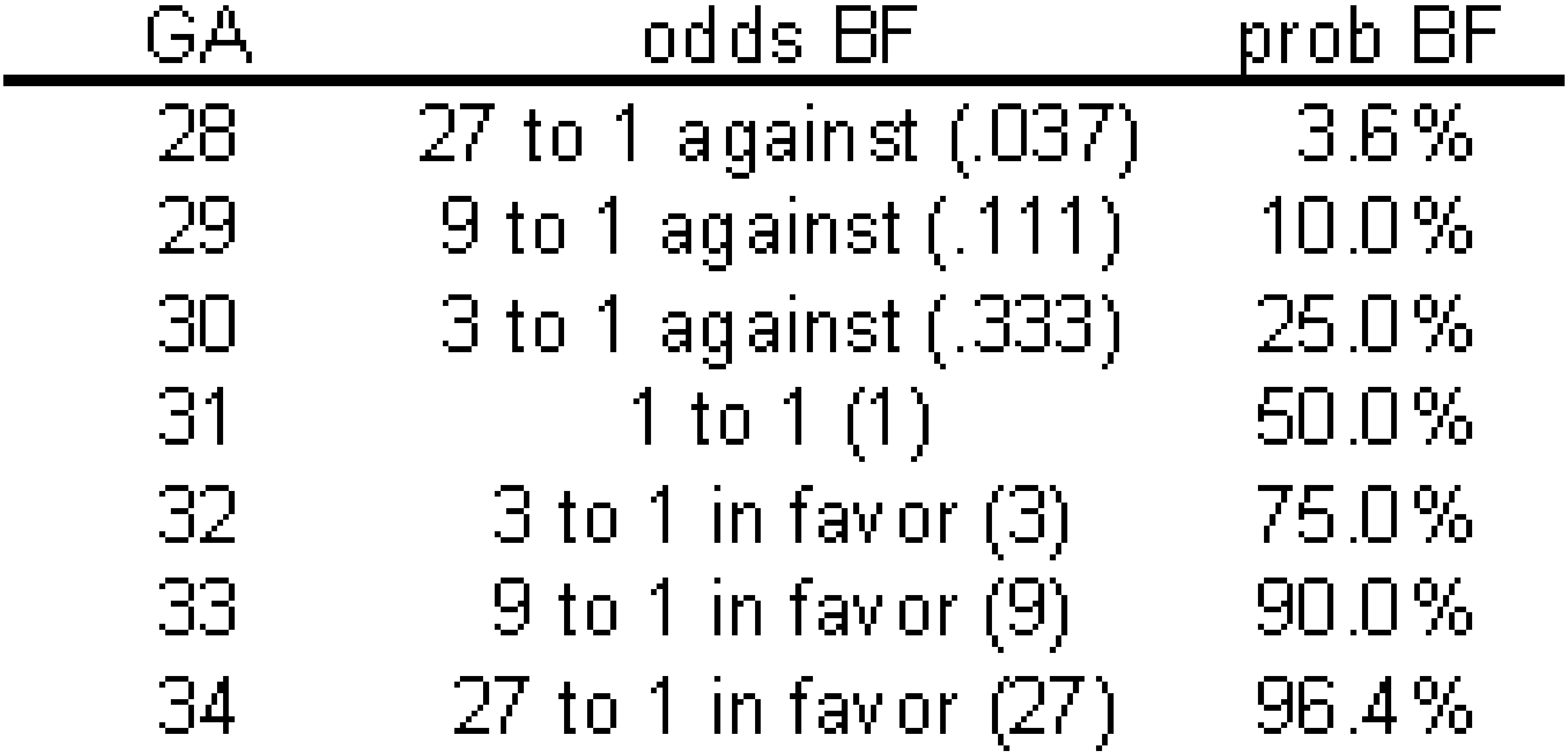
GA	odds BF	log odds
28	27 to 1 against (.037)	-3.30
29	9 to 1 against (.111)	-2.20
30	3 to 1 against (.333)	-1.10
31	1 to 1 (1)	0.00
32	3 to 1 in favor (3)	1.10
33	9 to 1 in favor (9)	2.20
34	27 to 1 in favor (27)	3.30

Additive trend in log odds

Speaker notes

It’s interesting to look at how the logarithm of the odds behave.

Notice that an extra week of GA adds 1.1 units to the log odds. So you can describe this model as linear (additive) in the log odds. When you run a logistic regression model in SPSS or other statistical software, it uses a model just like this, a model that is linear on the log odds scale. This may not seem too important now, but when you look at the output, you need to remember that SPSS presents all of the results in terms of log odds. If you want to see results in terms of probabilities instead of logs, you have to transform your results.



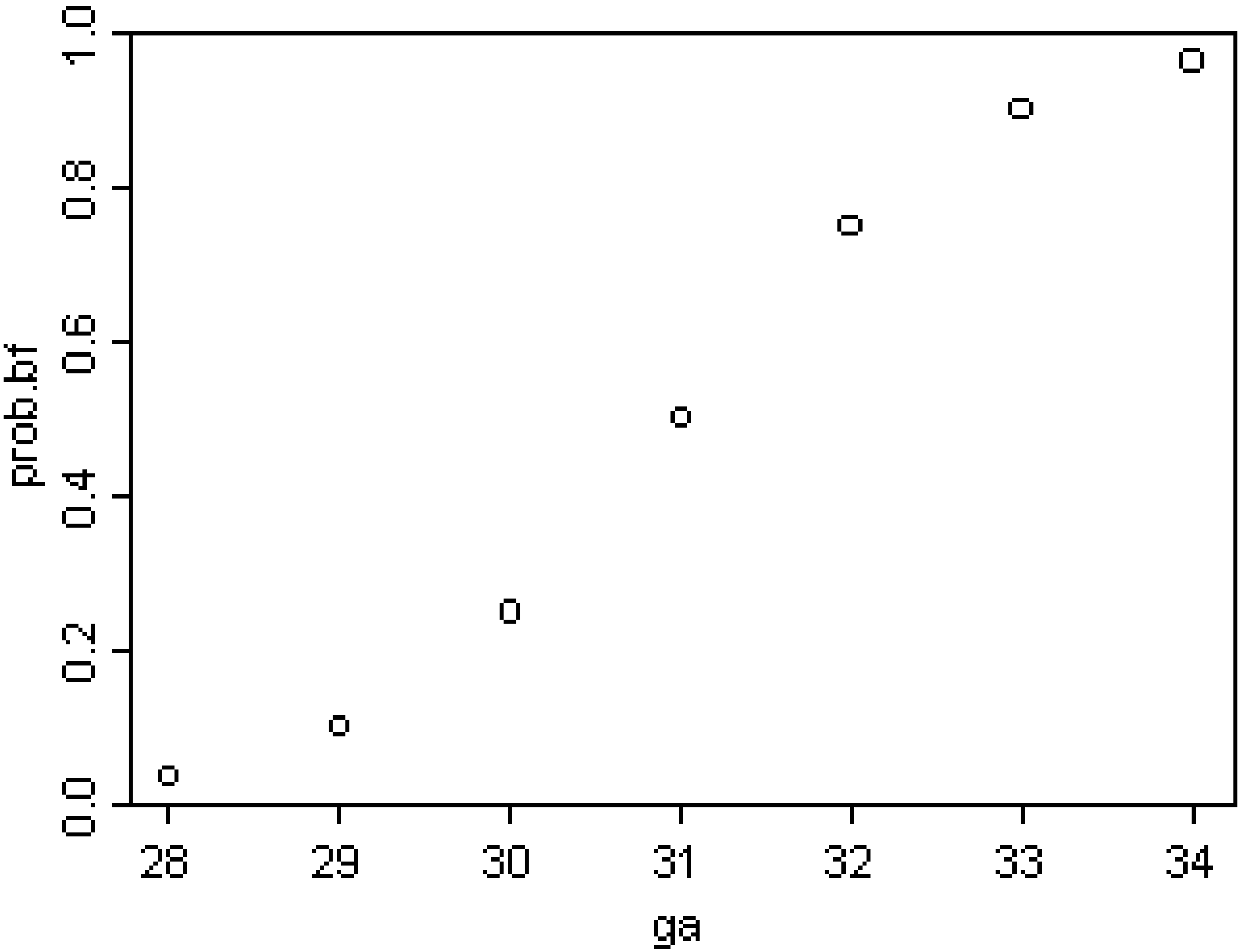
GA	odds BF	prob BF
28	27 to 1 against (.037)	3.6%
29	9 to 1 against (.111)	10.0%
30	3 to 1 against (.333)	25.0%
31	1 to 1 (1)	50.0%
32	3 to 1 in favor (3)	75.0%
33	9 to 1 in favor (9)	90.0%
34	27 to 1 in favor (27)	96.4%

Figure 21: Odds converted into probabilities

Speaker notes

Let’s look at how the probabilities behave in this model.

Notice that even when the odds get as large as 27 to 1, the probability still stays below 100%. Also notice that the probabilities change in neither an additive nor a multiplicative fashion.



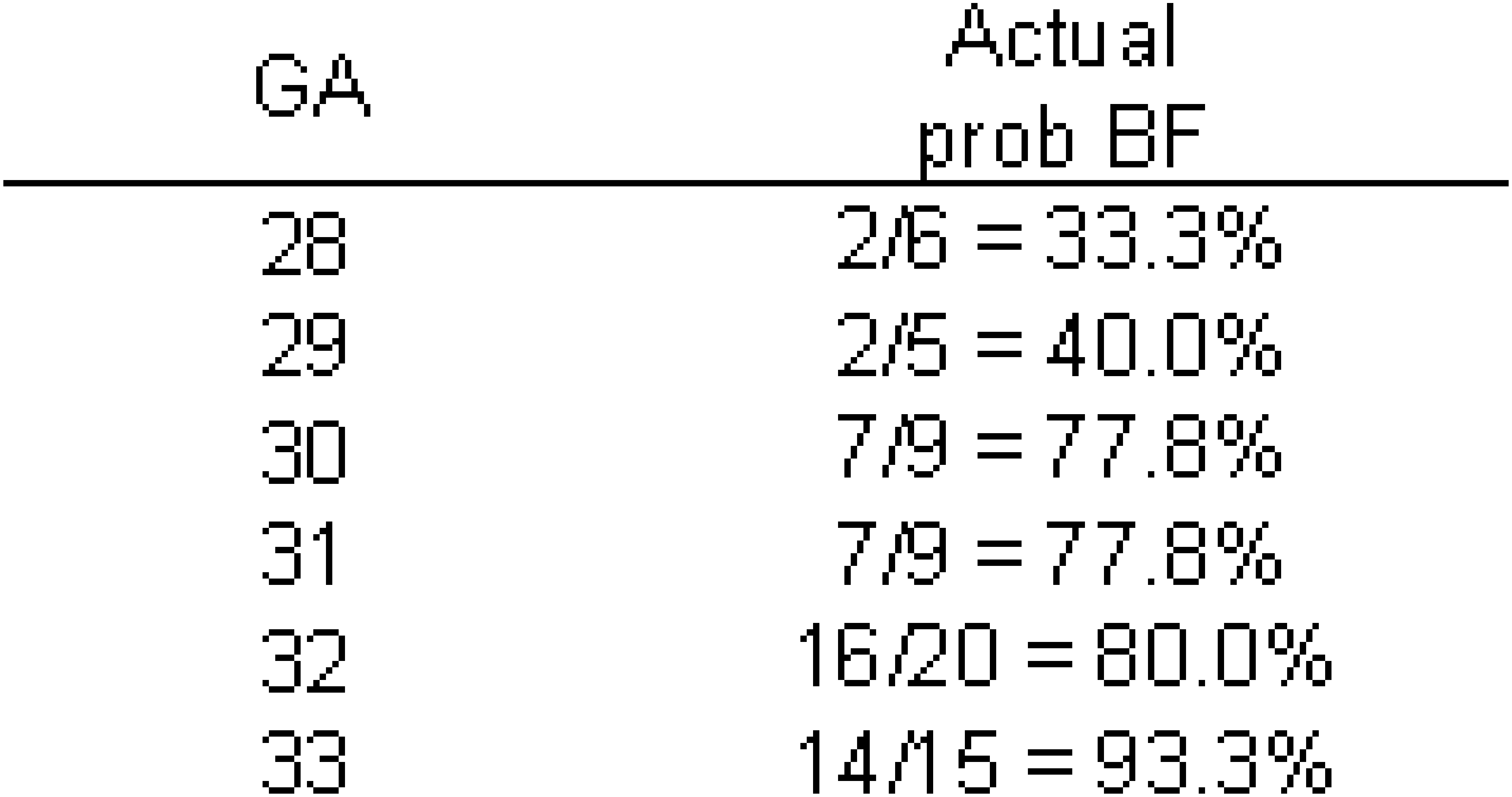
S-shaped curve

Speaker notes

Speaker notes

A graph shows what is happening.

The probabilities follow an S-shaped curve that is characteristic of all logistic regression models. The curve levels off at zero on one side and at one on the other side. This curve ensures that the estimated probabilities are always between 0% and 100%.



GA	Actual prob BF
28	2/6 = 33.3%
29	2/5 = 40.0%
30	7/9 = 77.8%
31	7/9 = 77.8%
32	16/20 = 80.0%
33	14/15 = 93.3%

Figure 22: Actual data on gestational age

$$log\ odds = -16.72 + 0.577 \times ga$$

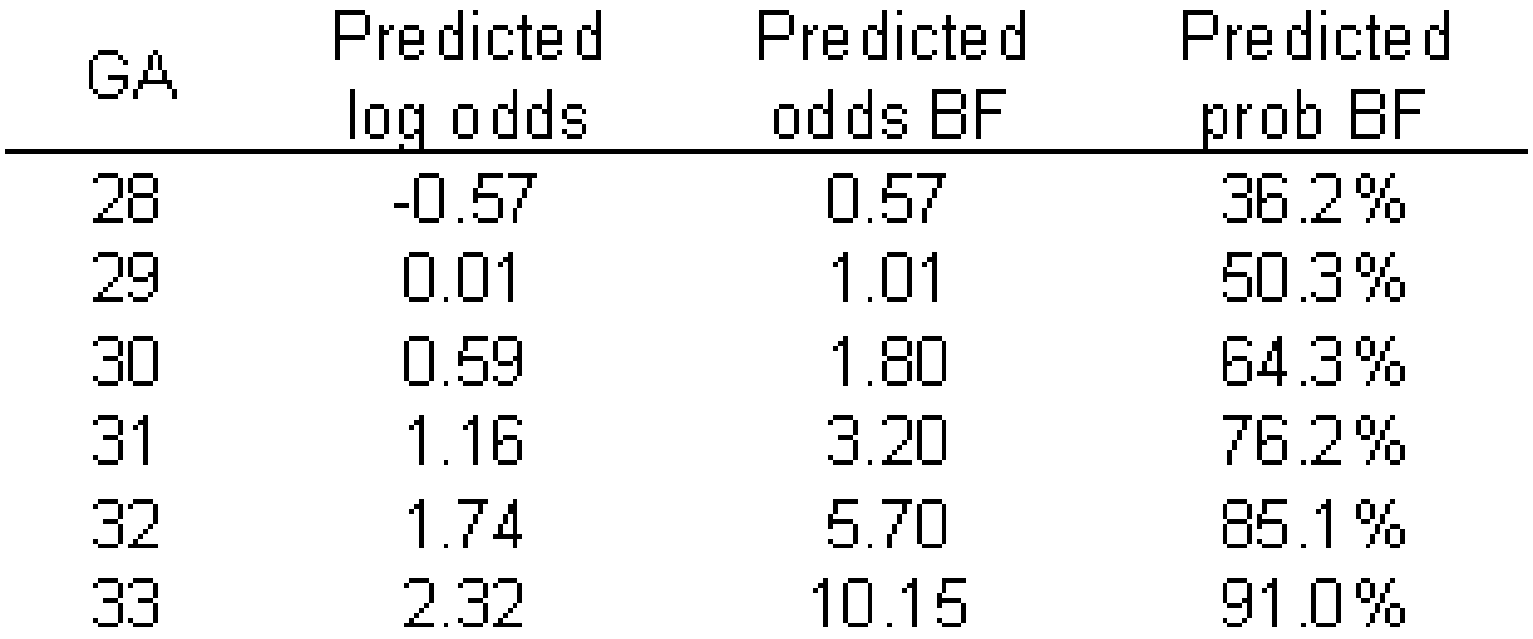
Speaker notes

An example of a log odds model with real data

There are other approaches that also work well for this type of data, such as a probit model, that I won’t discuss here. But I did want to show you what the data relating GA and BF really looks like.

I’ve simplified this data set by removing some of the extreme gestational ages. The estimated logistic regression model is

log odds = -16.72 + 0.577*GA



GA	Predicted log odds	Predicted odds BF	Predicted prob BF
28	-0.57	0.57	36.2%
29	0.01	1.01	50.3%
30	0.59	1.80	64.3%
31	1.16	3.20	76.2%
32	1.74	5.70	85.1%
33	2.32	10.15	91.0%

Figure 23: Predicted log odds

- Let’s examine these calculations for GA = 30.
  - log odds = -16.72 + 0.577*30 = 0.59
  - odds = exp(0.59) = 1.80
  - prob = 1.80 / (1+1.80) = 0.643

Speaker notes

Speaker notes

The table below shows the predicted log odds, and the calculations needed to transform this estimate back into predicted probabilities.

Let’s examine these calculations for GA = 30. The predicted log odds would be

log odds = -16.72 + 0.577*30 = 0.59

Convert from log odds to odds by exponentiating.

And finally, convert from odds back into probability.

The predicted probability of 64.3% is reasonably close to the true probability (77.8%).

Ratio of successive odds

1.01/0.57 = 1.78

1.80/1.01 = 1.78

3.20/1.80 = 1.78

5.70/3.20 = 1.78

Speaker notes

You might also want to take note of the predicted odds. Notice that the ratio of any odds to the odds in the next row is 1.78. For example,

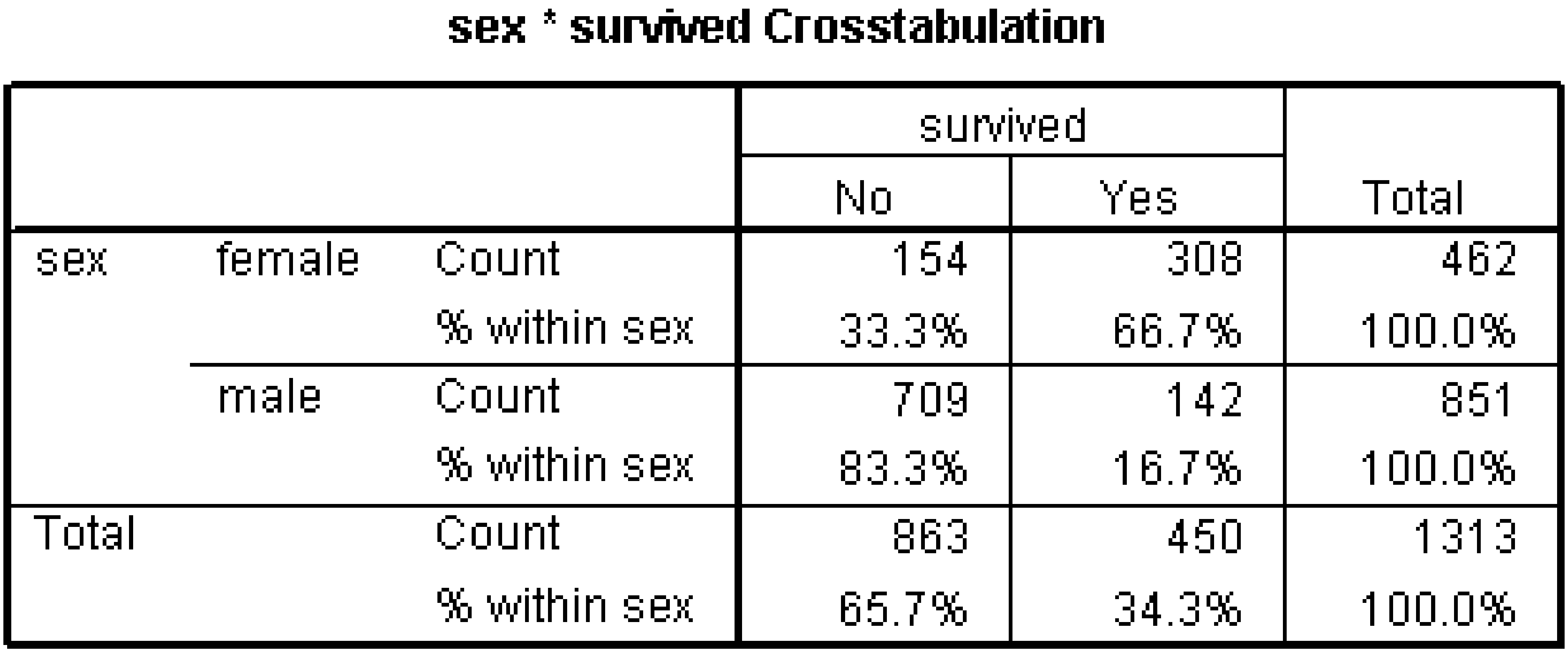
3.20/1.80 = 1.78

5.70/3.20 = 1.78

It’s not a coincidence that you get the same value when you exponentiate the slope term in the log odds equation.

exp(0.59) = 1.78

This is a general property of the logistic model. The slope term in a logistic regression model represents the log of the odds ratio. This represents the increase (decrease) in risk as the independent variable increases by one unit.



sex * survived Crosstabulation

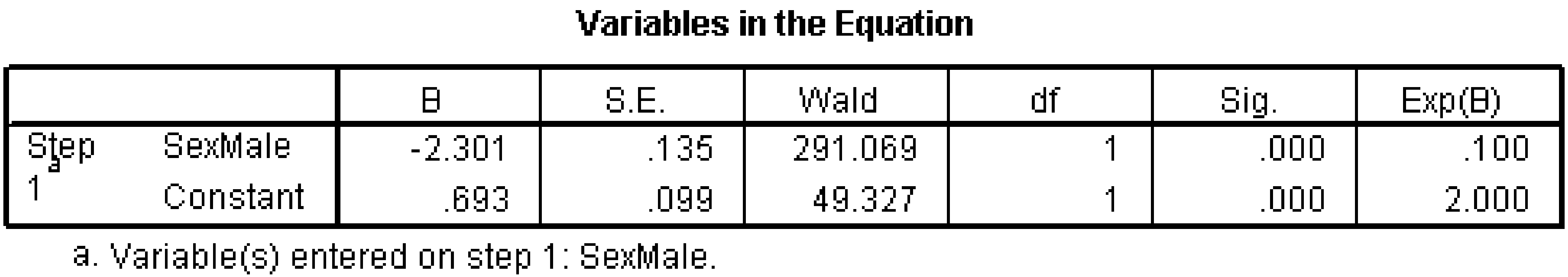
			survived		Total
			No	Yes	
sex	female	Count	154	308	462
		% within sex	33.3%	66.7%	100.0%
	male	Count	709	142	851
		% within sex	83.3%	16.7%	100.0%
Total		Count	863	450	1313
		% within sex	65.7%	34.3%	100.0%

Figure 24: Titanic probabilities for death and survival

Speaker notes

Speaker notes

Here is the table of survival versus sex on the Titanic.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SexMale	-2.301	.135	291.069	1	.000	.100
	Constant	.693	.099	49.327	1	.000	2.000

a. Variable(s) entered on step 1: SexMale.

Figure 25: Logistic regression for Titanic data

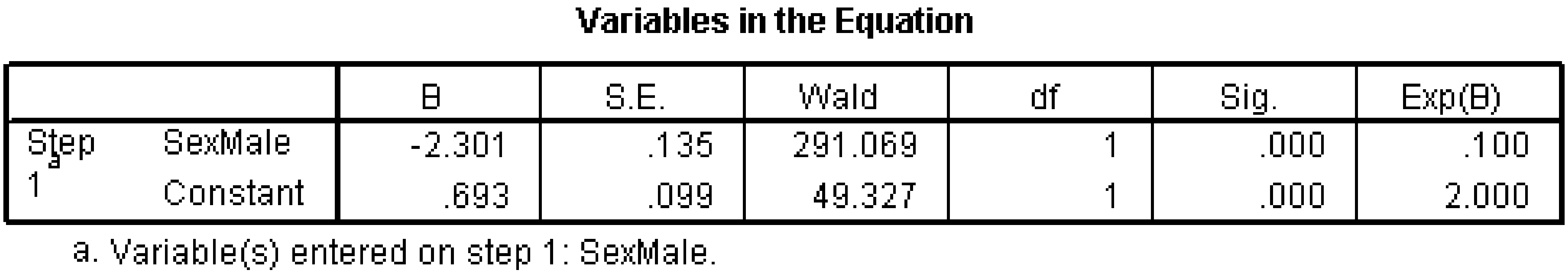
- Female
  - log odds = 0.693
  - odds = 2
  - prob = 0.667

Speaker notes

Speaker notes

Let’s start with the CONSTANT row of the data. This has an interpretation similar to the intercept in the linear regression model. the B column represents the estimated log odds when SexMale=0. Above, you saw that the odds for dying were 2 to 1 against for females, and the natural logarithm of 2 is 0.693. The last column, EXP(B) represents the odds, or 2.000. You need to be careful with this interpretation, because sometimes SPSS will report the odds in favor of an event and sometimes it will report the odds against an event. You have to look at the crosstabulation to be sure which it is.

Probability calculations for males



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SexMale	-2.301	.135	291.069	1	.000	.100
	Constant	.693	.099	49.327	1	.000	2.000

a. Variable(s) entered on step 1: SexMale.

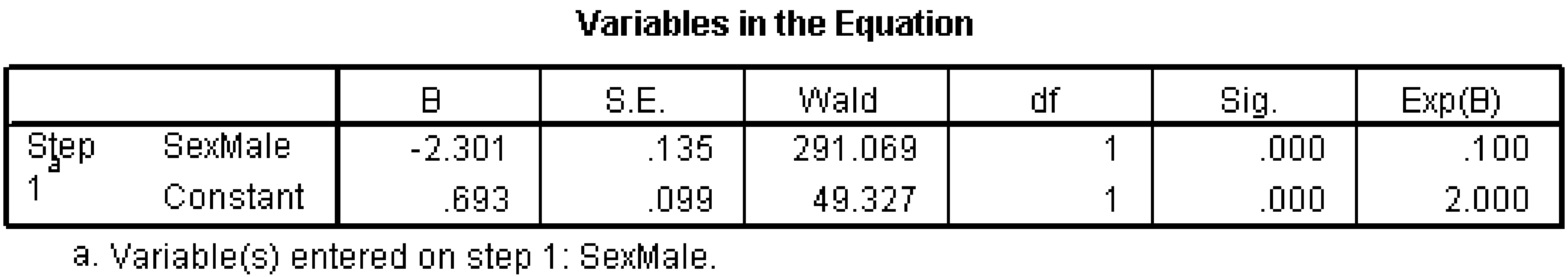
- Male
  - log odds = 0.693 - 2.301 = -1.608
  - odds = 0.2003
  - prob = 0.167

Speaker notes

Speaker notes

Let’s start with the CONSTANT row of the data. This has an interpretation similar to the intercept in the linear regression model. the B column represents the estimated log odds when SexMale=0. Above, you saw that the odds for dying were 2 to 1 against for females, and the natural logarithm of 2 is 0.693. The last column, EXP(B) represents the odds, or 2.000. You need to be careful with this interpretation, because sometimes SPSS will report the odds in favor of an event and sometimes it will report the odds against an event. You have to look at the crosstabulation to be sure which it is.

The SexMale row has an interpretation similar to the slope term in a linear regression model. The B column represents the estimated change in the log odds when SexMale increases by one unit. This is effectively the log odds ratio. We computed the odds ratio above, and -2.301 is the natural logarithm of 0.1. The last column, EXP(B) provides you with the odds ratio (0.100).



Variables in the Equation

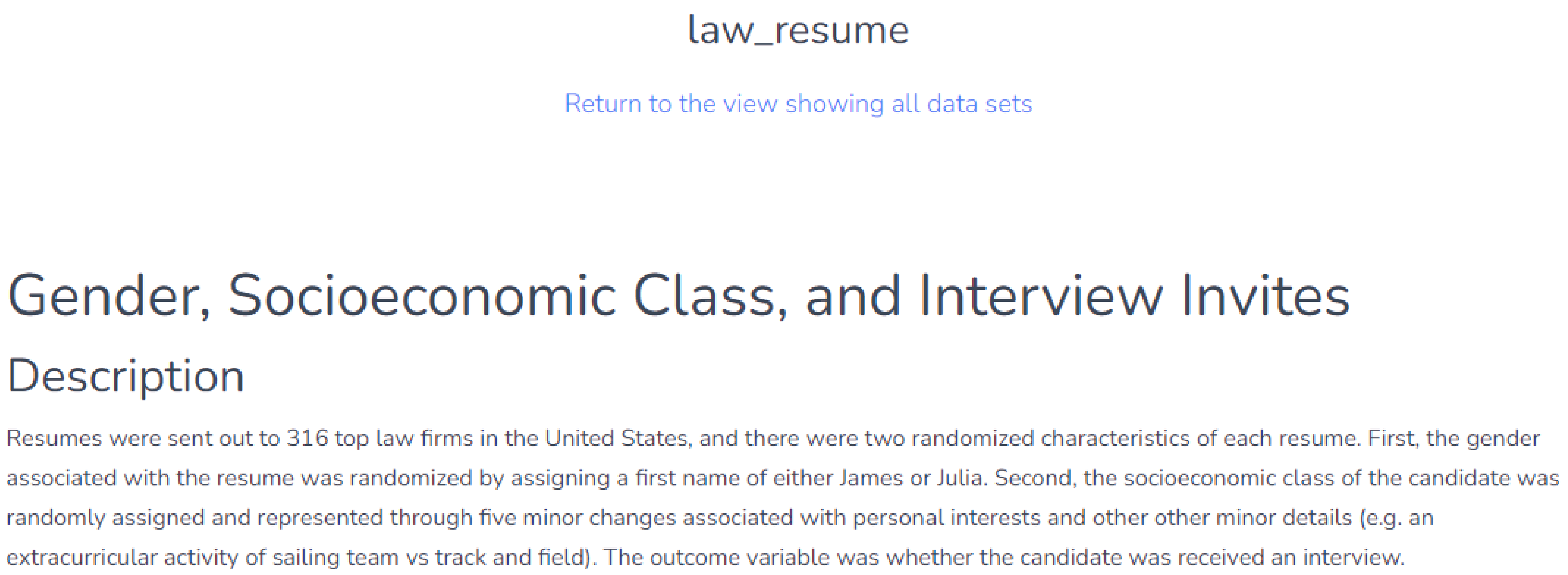
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SexMale	-2.301	.135	291.069	1	.000	.100
	Constant	.693	.099	49.327	1	.000	2.000

a. Variable(s) entered on step 1: SexMale.

- log odds ratio = -2.301
  - odds ratio = 0.1

Speaker notes

The SexMale row has an interpretation similar to the slope term in a linear regression model. The B column represents the estimated change in the log odds when SexMale increases by one unit. This is effectively the log odds ratio. We computed the odds ratio above, and -2.301 is the natural logarithm of 0.1. The last column, EXP(B) provides you with the odds ratio (0.100).



law_resume

Return to the view showing all data sets

Gender, Socioeconomic Class, and Interview Invites

Description

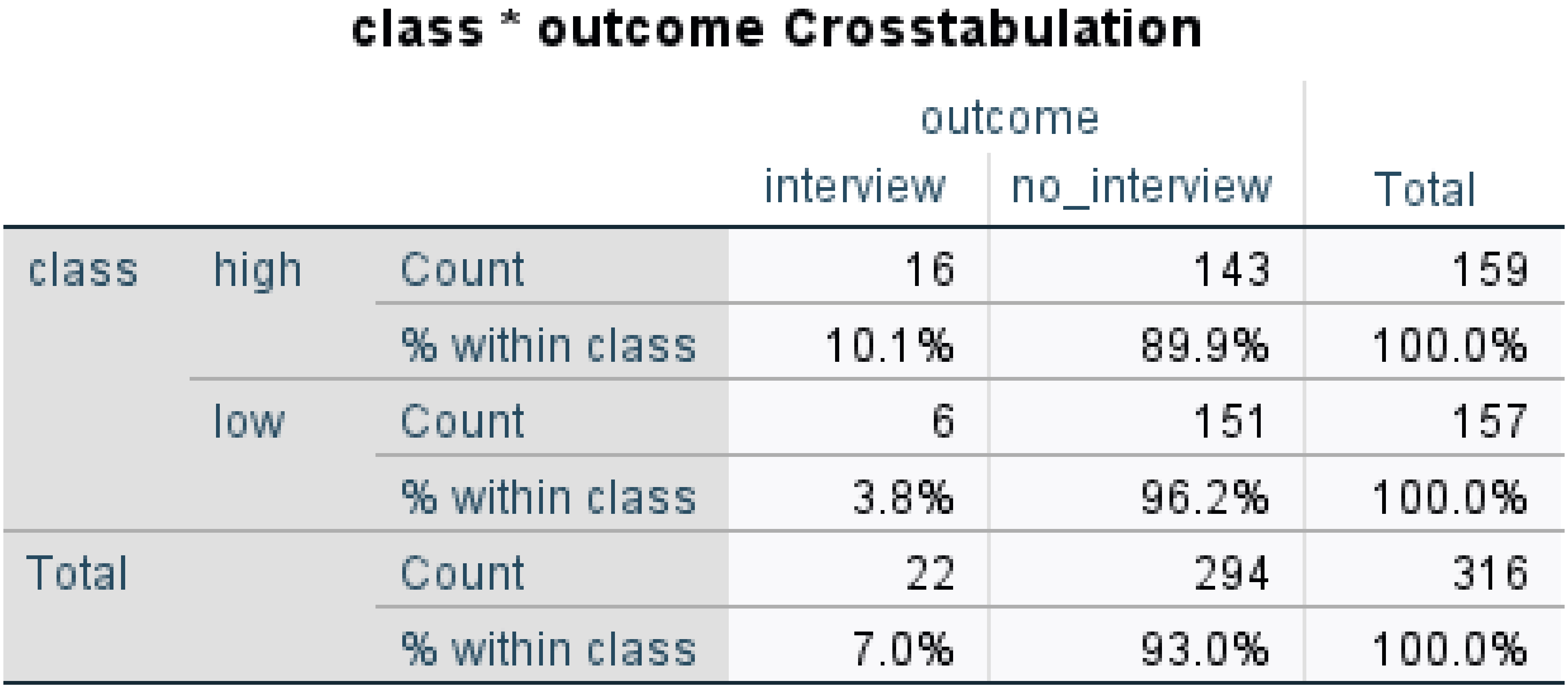
Resumes were sent out to 316 top law firms in the United States, and there were two randomized characteristics of each resume. First, the gender associated with the resume was randomized by assigning a first name of either James or Julia. Second, the socioeconomic class of the candidate was randomly assigned and represented through five minor changes associated with personal interests and other other minor details (e.g. an extracurricular activity of sailing team vs track and field). The outcome variable was whether the candidate was received an interview.

Figure 26: Description of the interview invite dataset

Speaker notes

Speaker notes

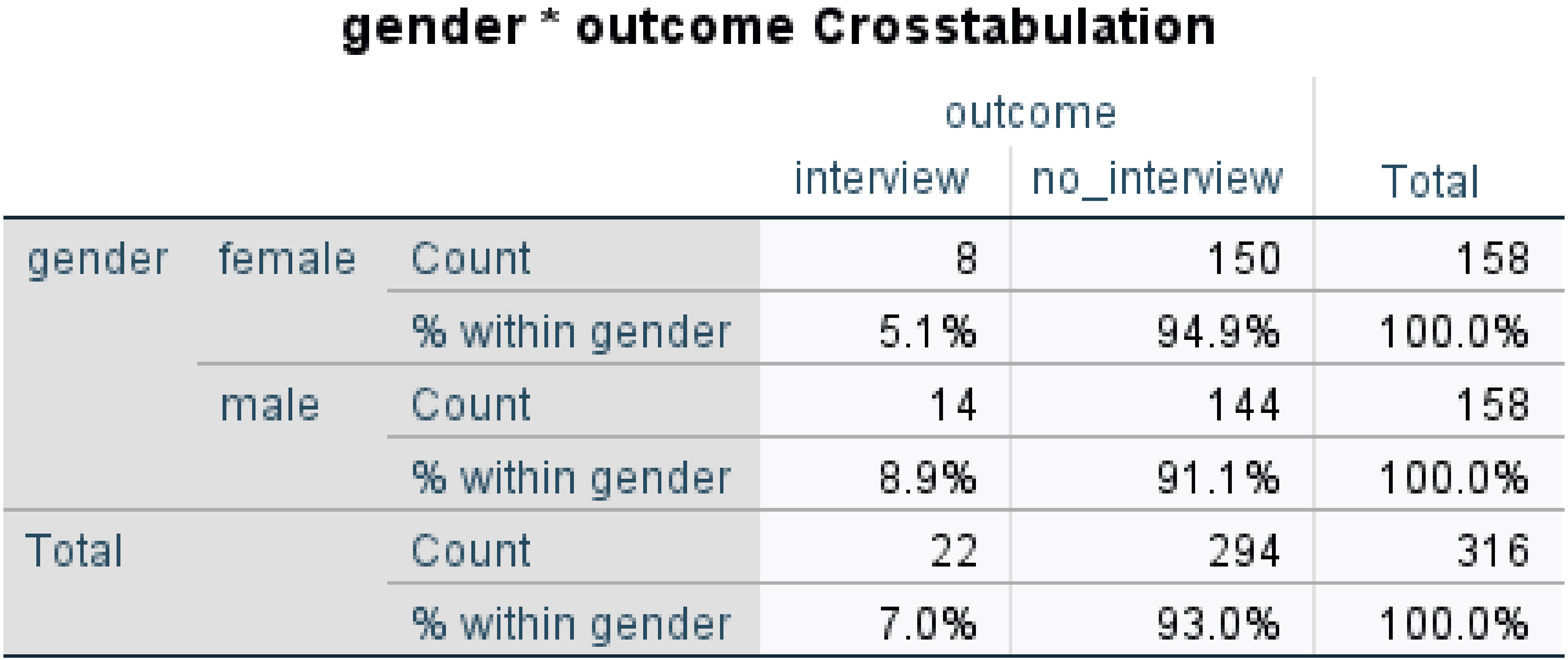
“Resumes were sent out to 316 top law firms in the United States, and there were two randomized characteristics of each resume. First, the gender associated with the resume was randomized by assigning a first name of either James or Julia. Second, the socioeconomic class of the candidate was randomly assigned and represented through five minor changes associated with personal interests and other other minor details (e.g. an extracurricular activity of sailing team vs track and field). The outcome variable was whether the candidate was received an interview.”



class * outcome Crosstabulation

			outcome		
			interview	no_interview	Total
class	high	Count	16	143	159
		% within class	10.1%	89.9%	100.0%
	low	Count	6	151	157
		% within class	3.8%	96.2%	100.0%
Total		Count	22	294	316
		% within class	7.0%	93.0%	100.0%

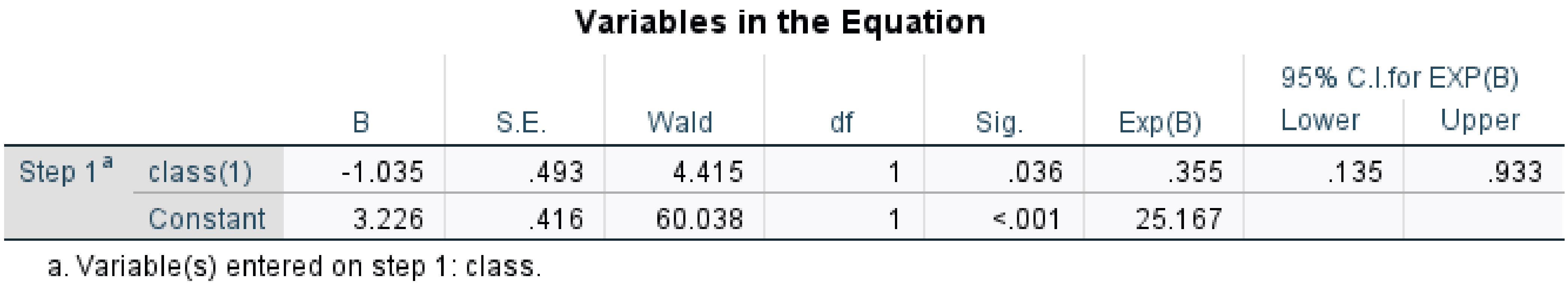
Figure 27: Crosstabulation of class and interview



gender * outcome Crosstabulation

			outcome		
			interview	no_interview	Total
gender	female	Count	8	150	158
		% within gender	5.1%	94.9%	100.0%
	male	Count	14	144	158
		% within gender	8.9%	91.1%	100.0%
Total		Count	22	294	316
		% within gender	7.0%	93.0%	100.0%

Figure 28: Crosstabulation of gender and interview

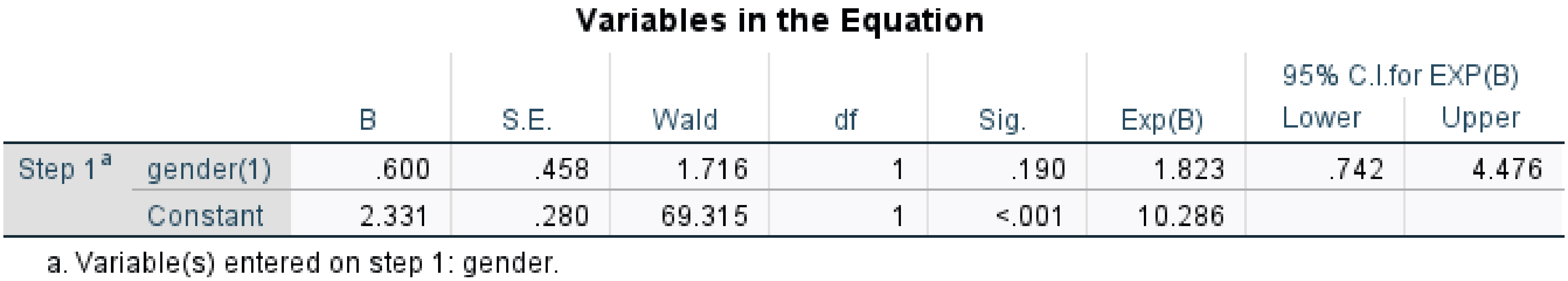


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	class(1)	-1.035	.493	4.415	1	.036	.355	.135	.933
	Constant	3.226	.416	60.038	1	<.001	25.167		

a. Variable(s) entered on step 1: class.

Figure 29: Logistic regression model for class

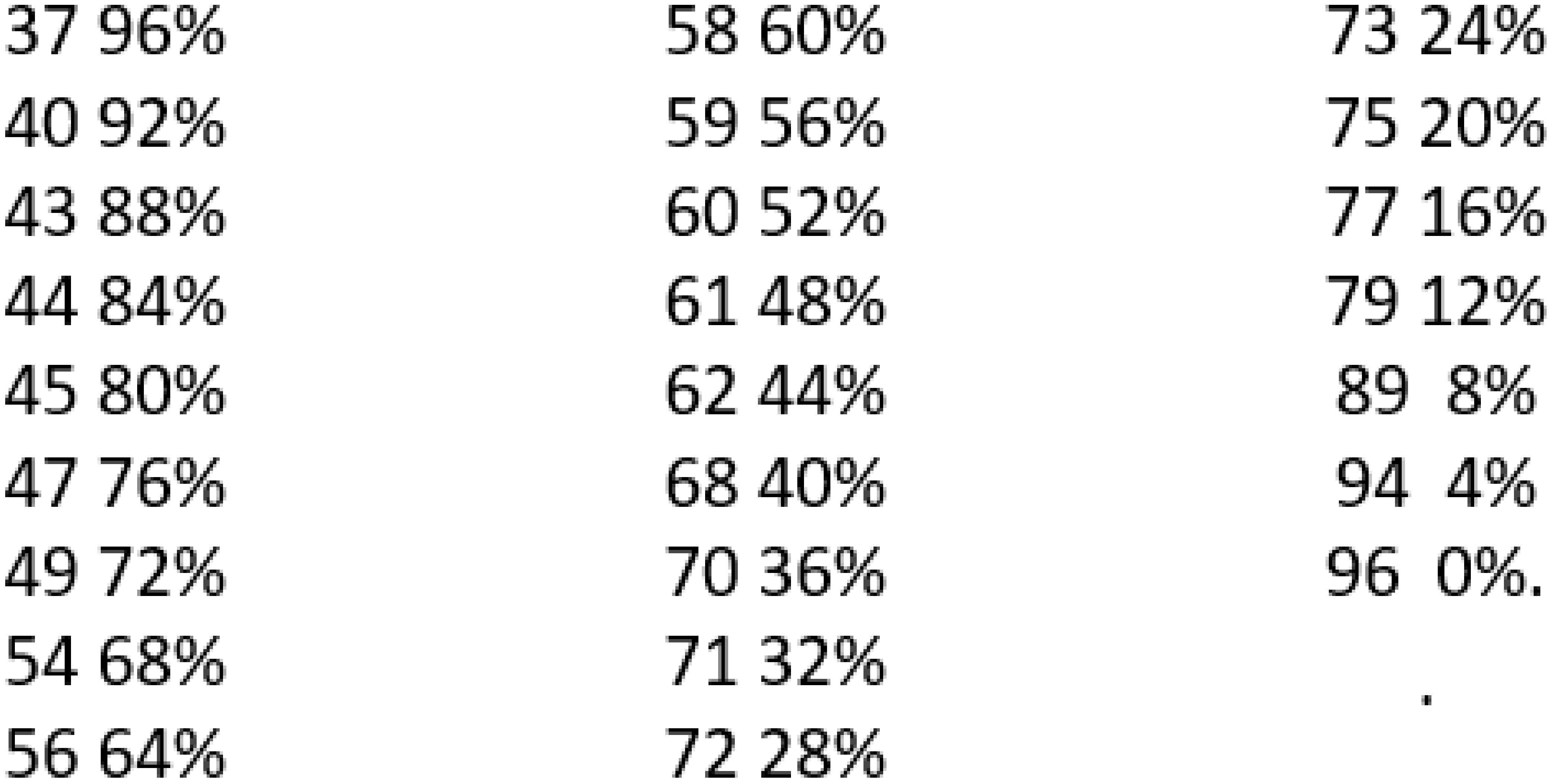


Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	gender(1)	.600	.458	1.716	1	.190	1.823	.742	4.476
	Constant	2.331	.280	69.315	1	<.001	10.286		

a. Variable(s) entered on step 1: gender.

Figure 30: Logistic regression model for gender



37	96%
40	92%
43	88%
44	84%
45	80%
47	76%
49	72%
54	68%
56	64%
58	60%
59	56%
60	52%
61	48%
62	44%
68	40%
70	36%
71	32%
72	28%
73	24%
75	20%
77	16%
79	12%
89	8%
94	4%
96	0%.

Figure 31: Fruit fly data, round 1

Speaker notes

Speaker notes

This data represents survival time for a group of fruit flies and is a subset of a larger data set found at the Data and Story Library (DASL). The data set has been slightly modified to simplify some of these explanations.

There are 25 flies in the sample, with the first fly dying on day 37 and the last fly dying on day 96.

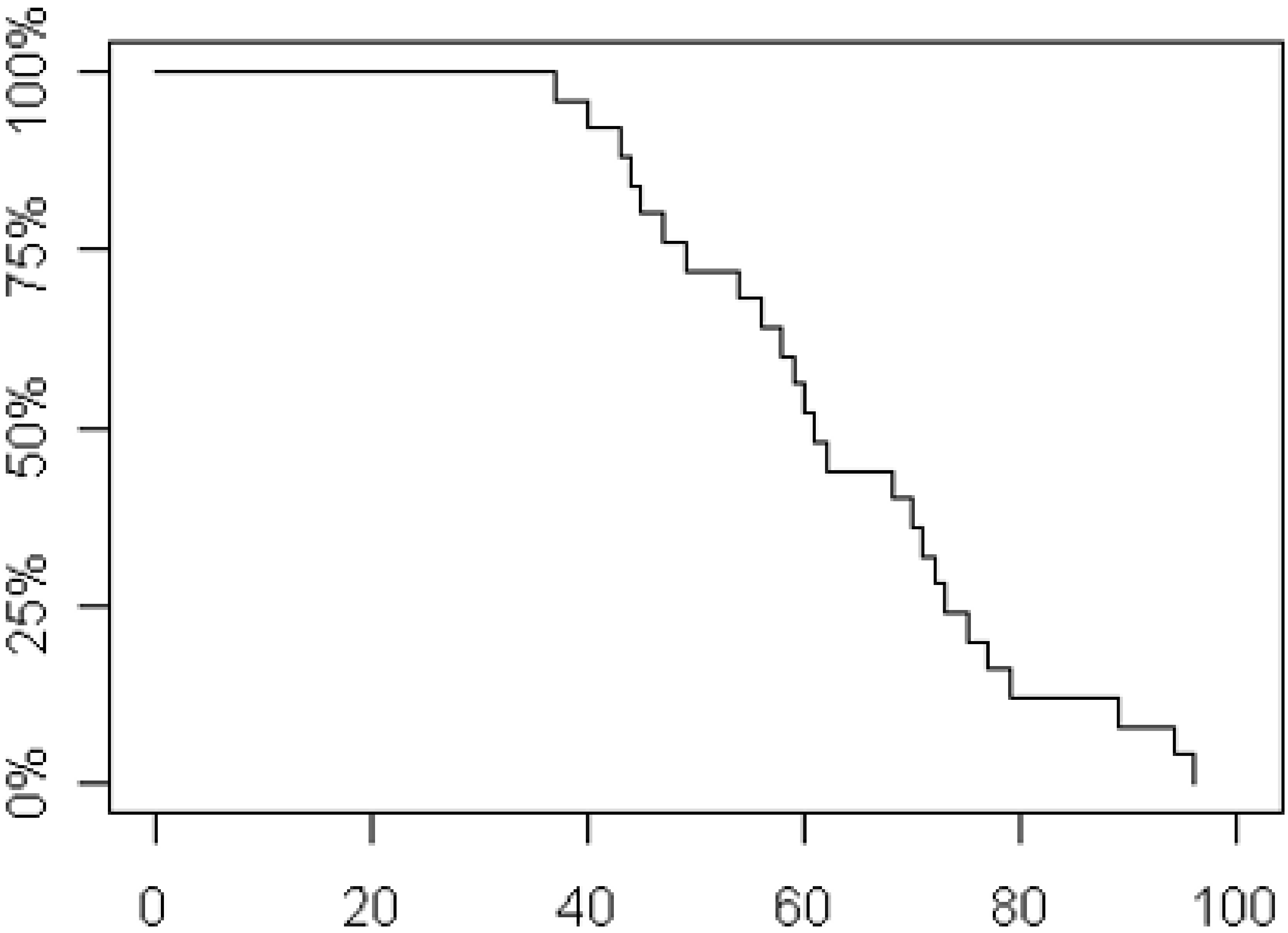


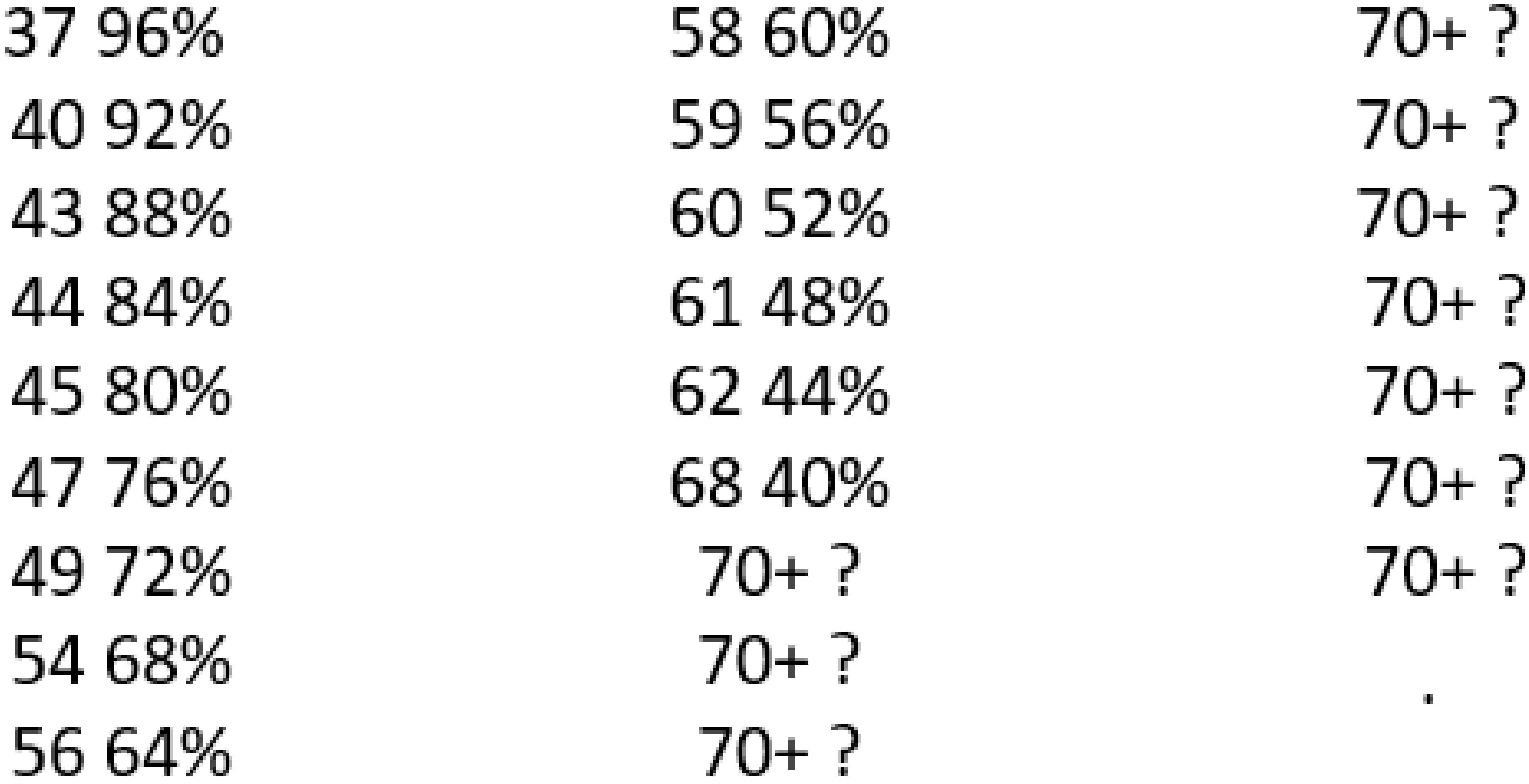
Figure 32: Fruit fly graph, round 1

Speaker notes

Speaker notes

If you wanted to estimate the survival probability for this data, you would draw a curve that decreases by 4% (1/25) every time a fly dies.

Fruit fly data (round 2)



37 96%
40 92%
43 88%
44 84%
45 80%
47 76%
49 72%
54 68%
56 64%

58 60%
59 56%
60 52%
61 48%
62 44%
68 40%
70+ ?
70+ ?
70+ ?

70+ ?
70+ ?
70+ ?
70+ ?
70+ ?
70+ ?
70+ ?
.

Figure 33: Fruit fly data, round 2

Speaker notes

Speaker notes

Now let’s alter the experiment. Suppose that totally by accident, a technician leaves the screen cover open on day 70 and all the flies escape. You’re probably worried that the whole experiment has been ruined. But don’t be so pessimistic. You still have complete information on survival of the fruit flies up to their 70th day of life. Here’s how you would present the data and estimate the survival probabilities.

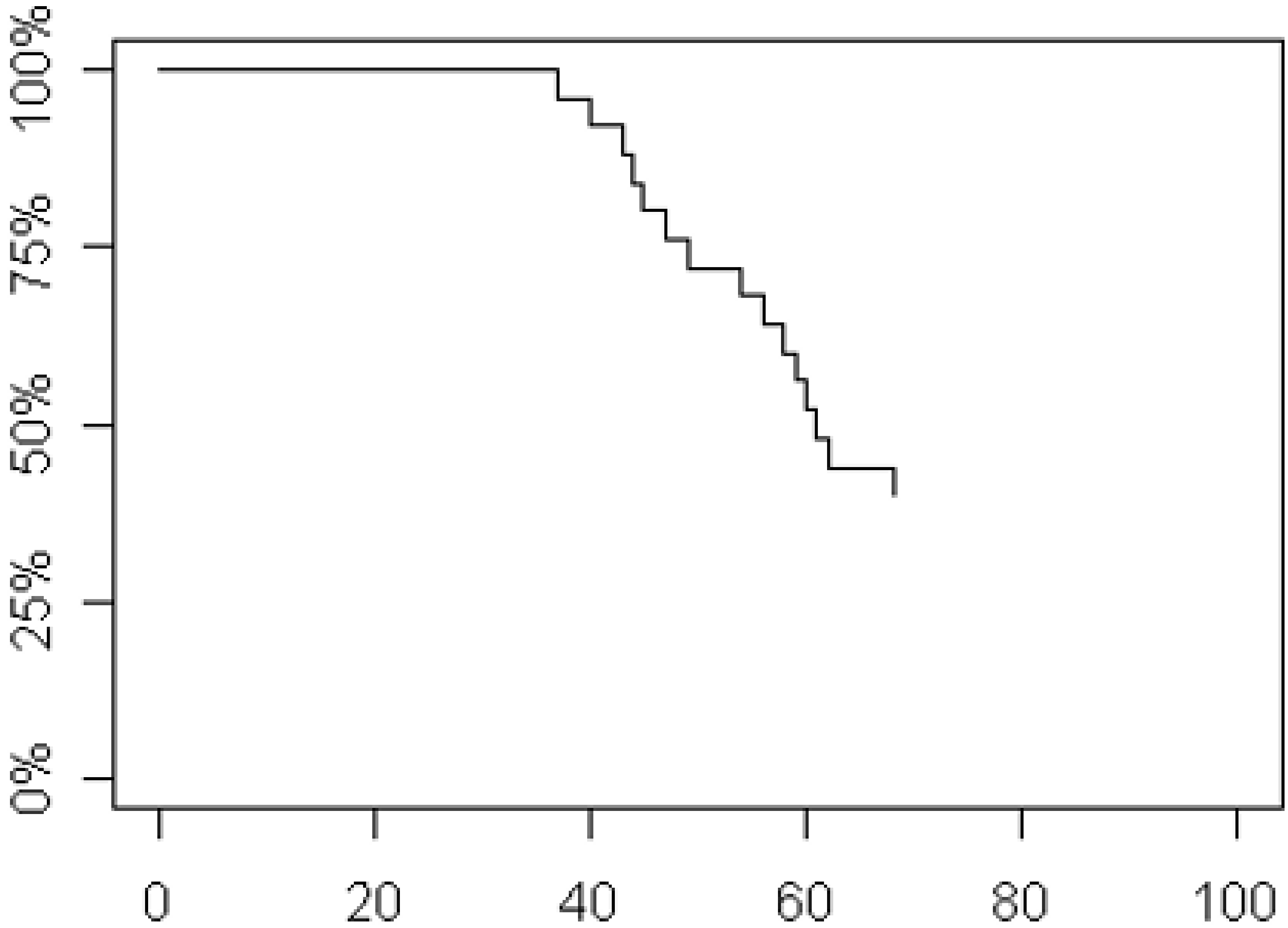
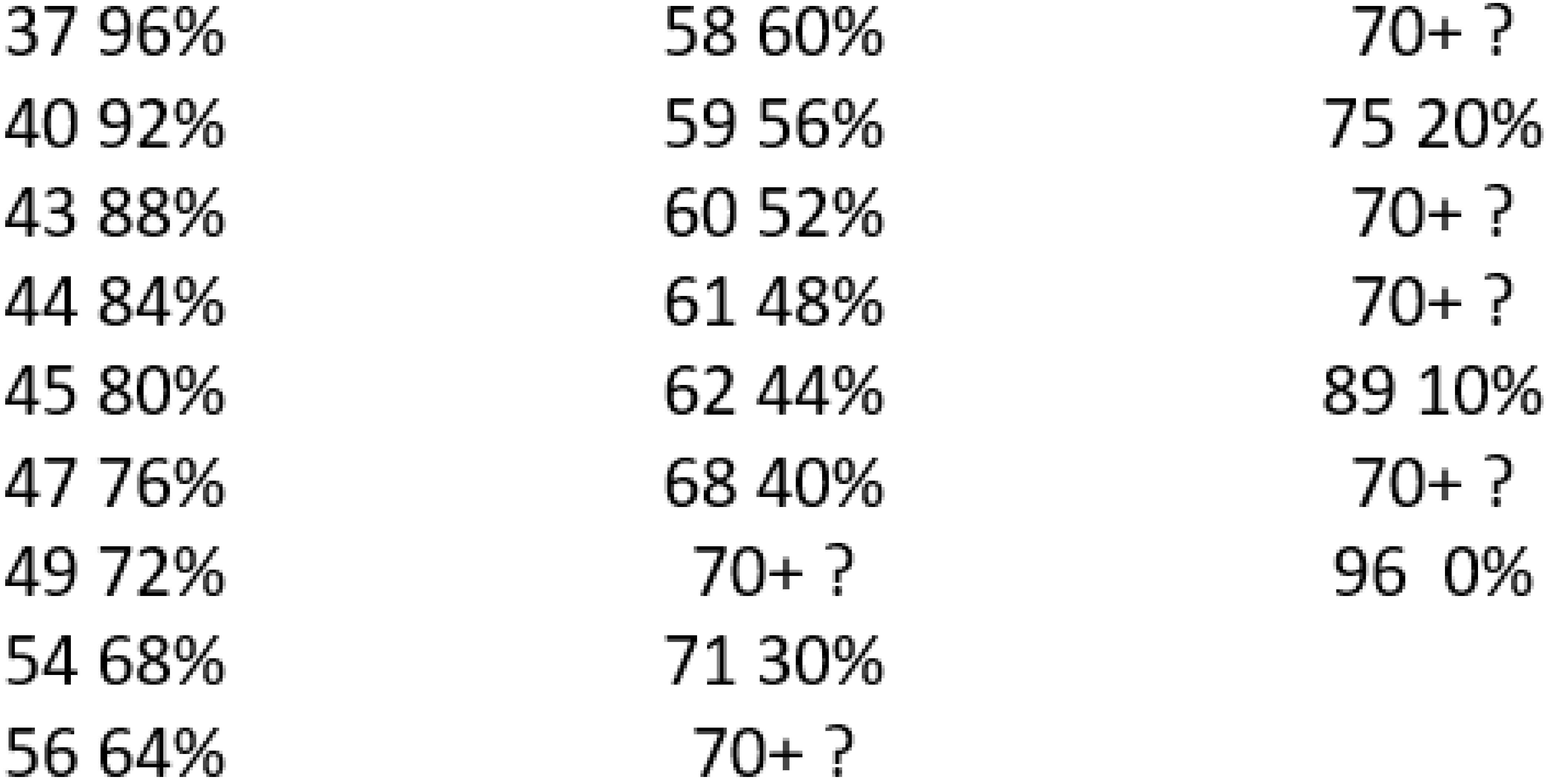


Figure 34: Fruit fly graph, round 2

Speaker notes

We clearly have enough data to make several important statements about survival probability. For example, the median survival time is 61 days because roughly half of the flies had died before this day.

By the way, you might be tempted to ignore the ten flies who escaped. But that would seriously bias your results. The median survival time, for example, of the 15 flies who did not escape, for example, is only 54 days which is much smaller than the actual median.



37 96%

40 92%

43 88%

44 84%

45 80%

47 76%

49 72%

54 68%

56 64%

58 60%

59 56%

60 52%

61 48%

62 44%

68 40%

70+ ?

71 30%

70+ ?

70+ ?

75 20%

70+ ?

70+ ?

89 10%

70+ ?

96 0%

Figure 35: Fruit fly data, round 3

Speaker notes

Speaker notes

Let’s look at a third experiment, where the screen cover is left open and all but four of the remaining flies escape. It turns out that those four remaining flies who didn’t bug out will allow us to still get reasonable estimates of survival probabilities beyond 70 days. Here is the data and the survival probabilities.

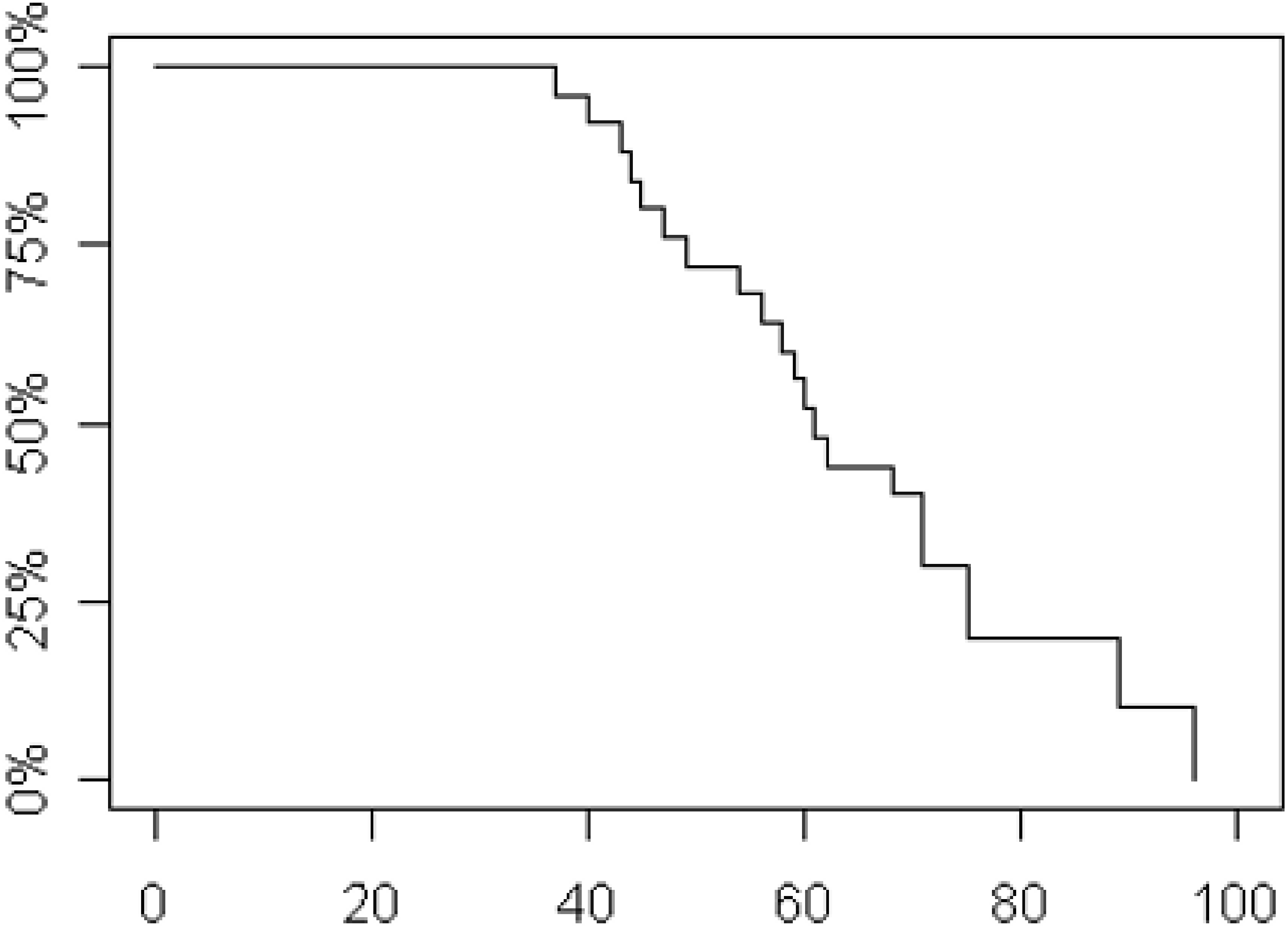


Figure 36: Fruit fly graph, round 3

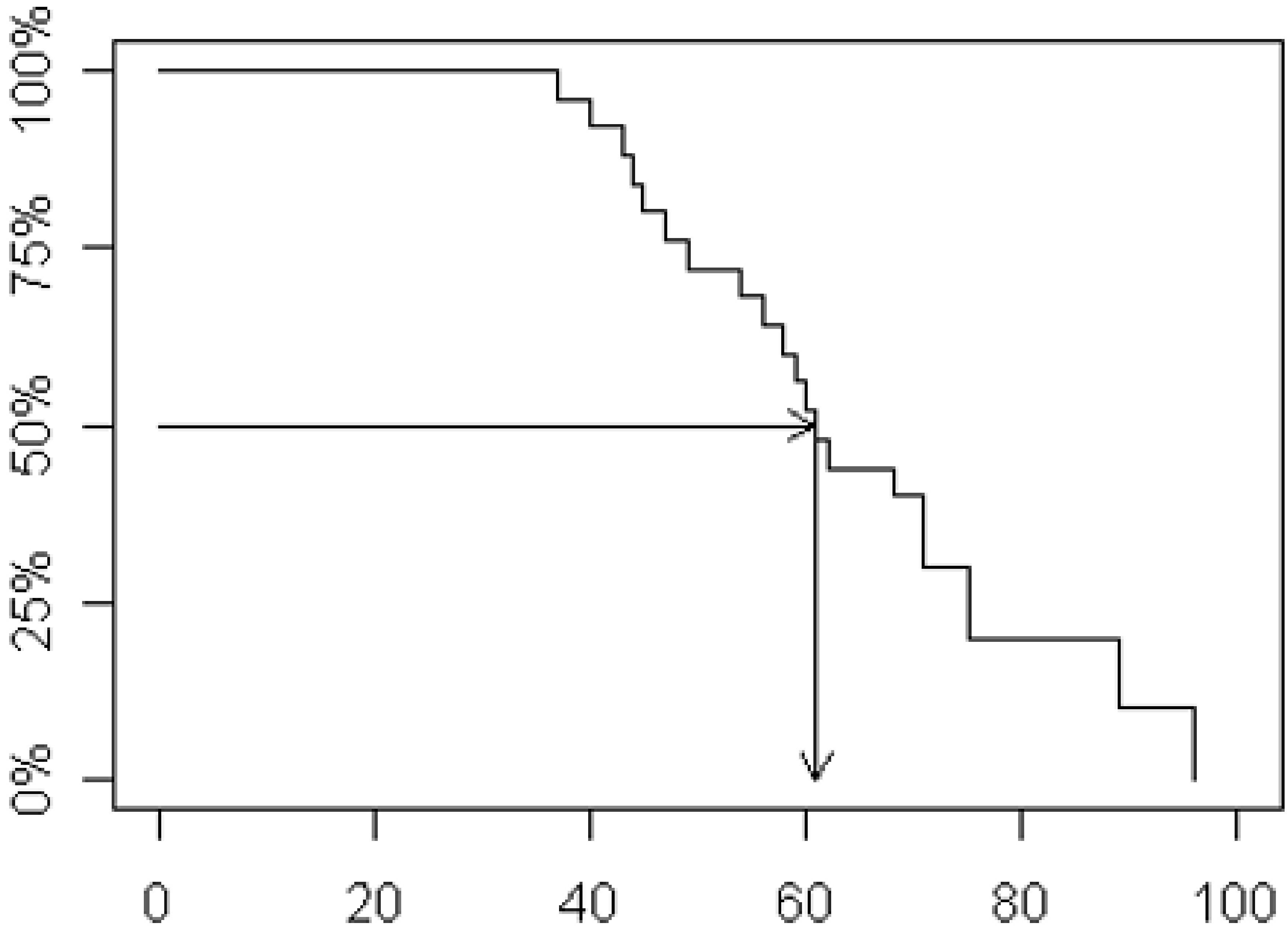


Figure 37: Fruit fly graph, estimating the median

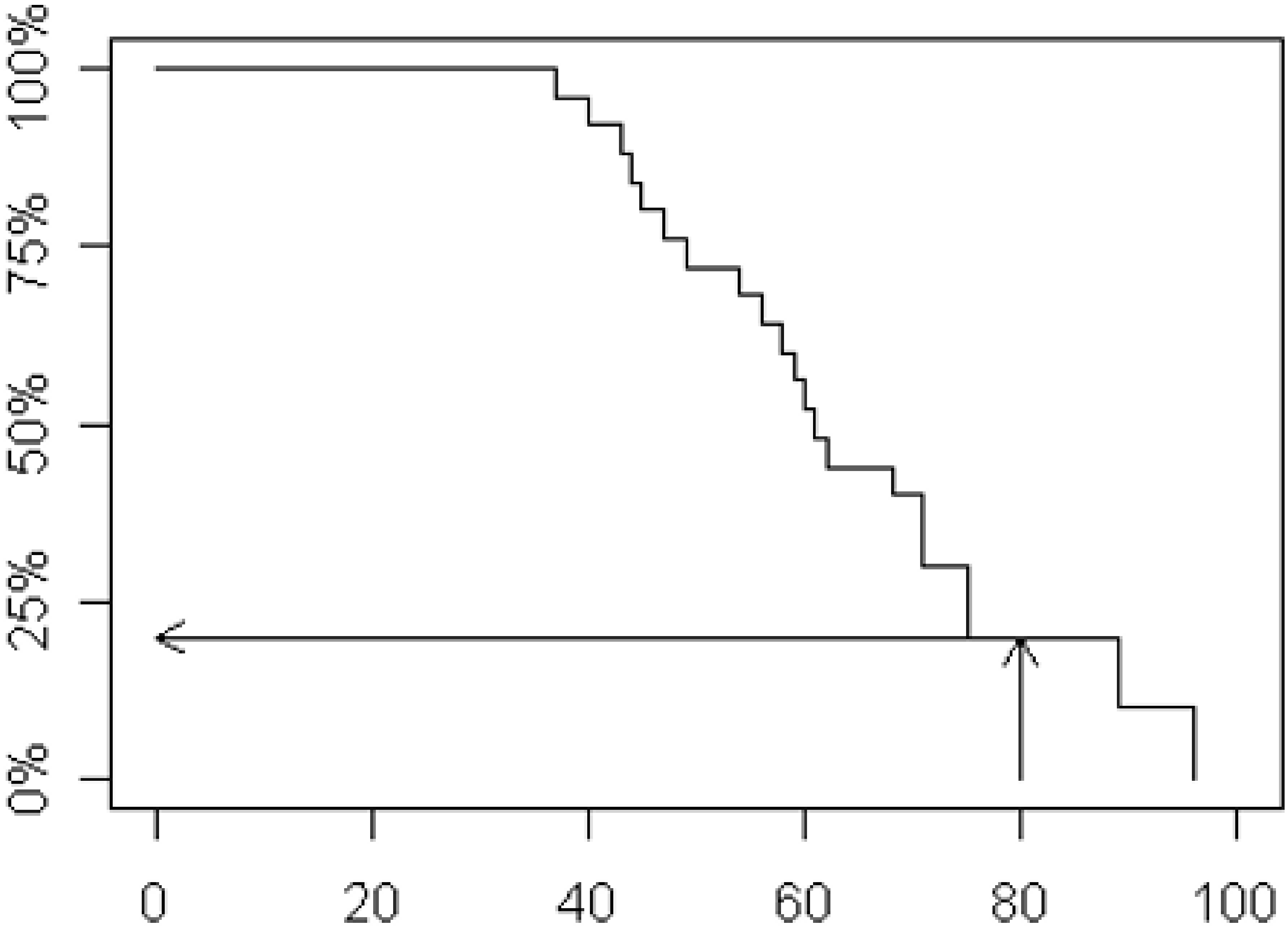


Figure 38: Fruit fly graph, estimating survival probability

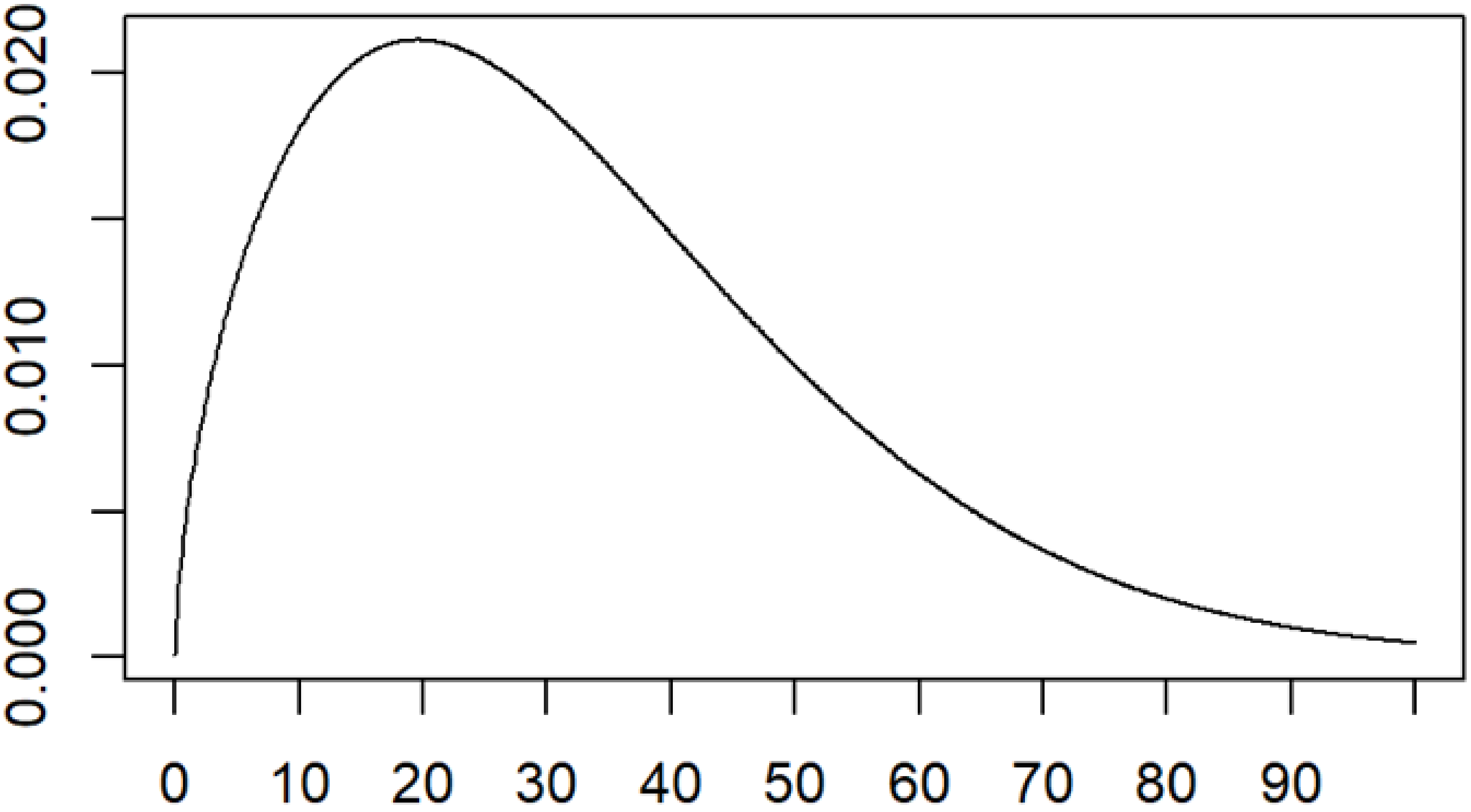


Figure 39: Hypothetical survival distribution

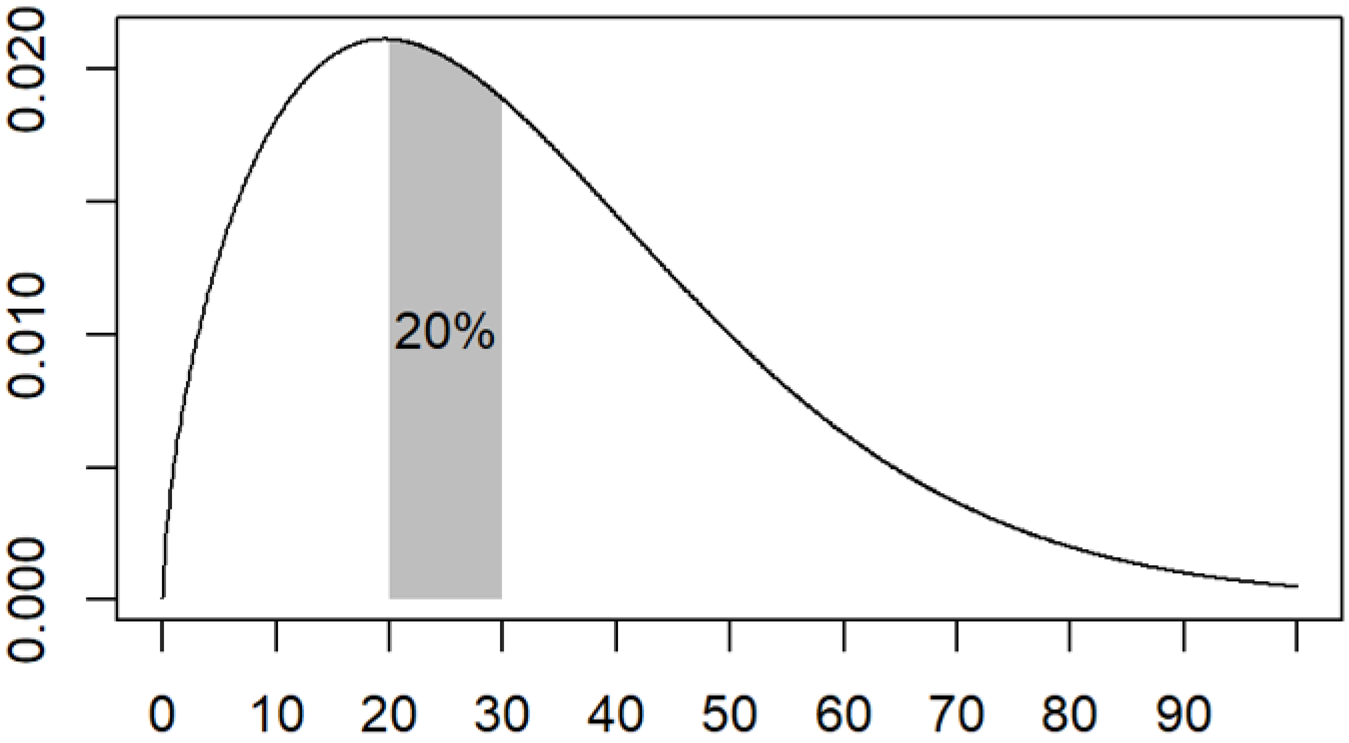


Figure 40: Hypothetical survival distribution, probability for 20-30 years

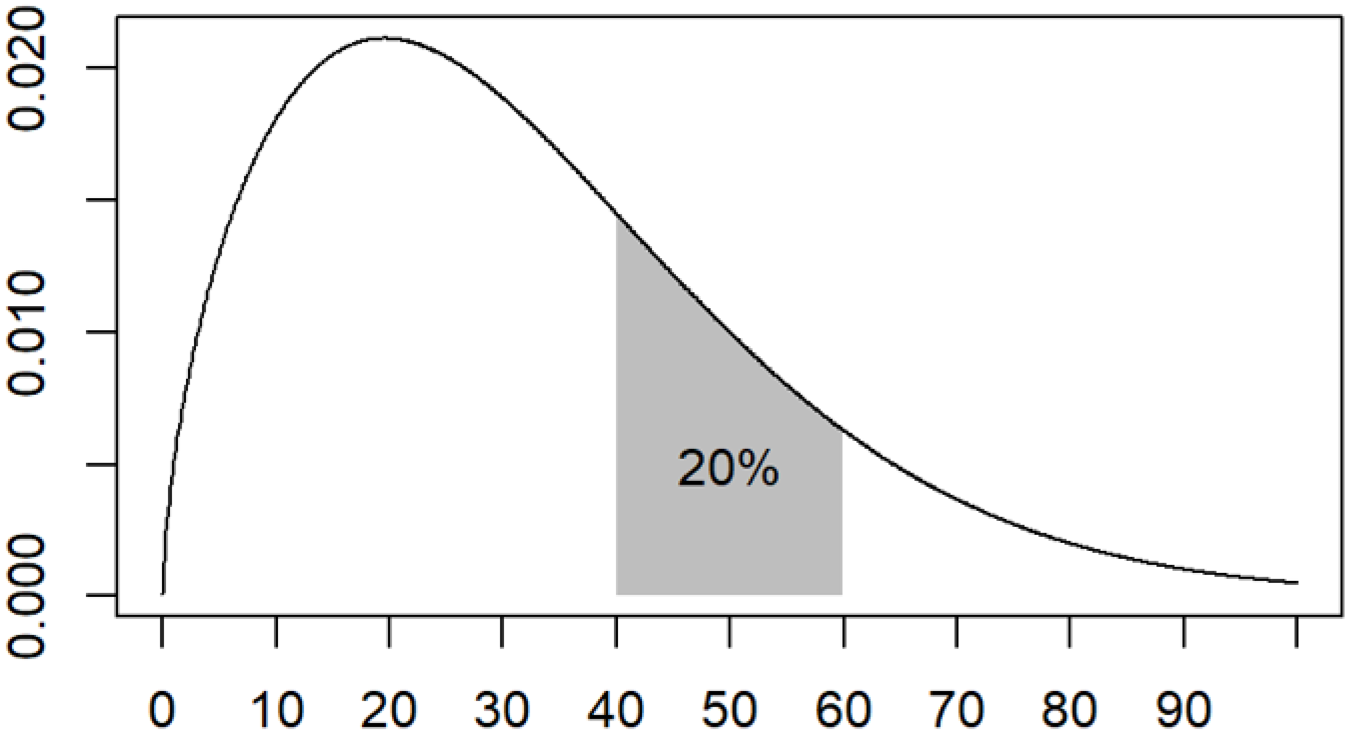


Figure 41: Hypothetical survival distribution, probability for 40-60 years

Defining the hazard function (1/2)

- To make a fair comparison
  - Adjust by the probability of surviving up to age 20 or age 40.
  - Calculate a death rate by dividing by the time range.
  - Calculate over a narrow time interval, Δt.

Defining the hazard function (2/2)

- The hazard function is defined as
  - $h(t) = (P[t \leq T \leq t+\Delta t] / \Delta t) / P[T \geq t]$
- Key points
  - adjusted for the number surviving to that time ($P[T \geq t]$),
  - calculated as a rate
    - ($P[t \leq T \leq T+\Delta t] / \Delta t$) is not a probability, and
  - computed over a narrow time interval.

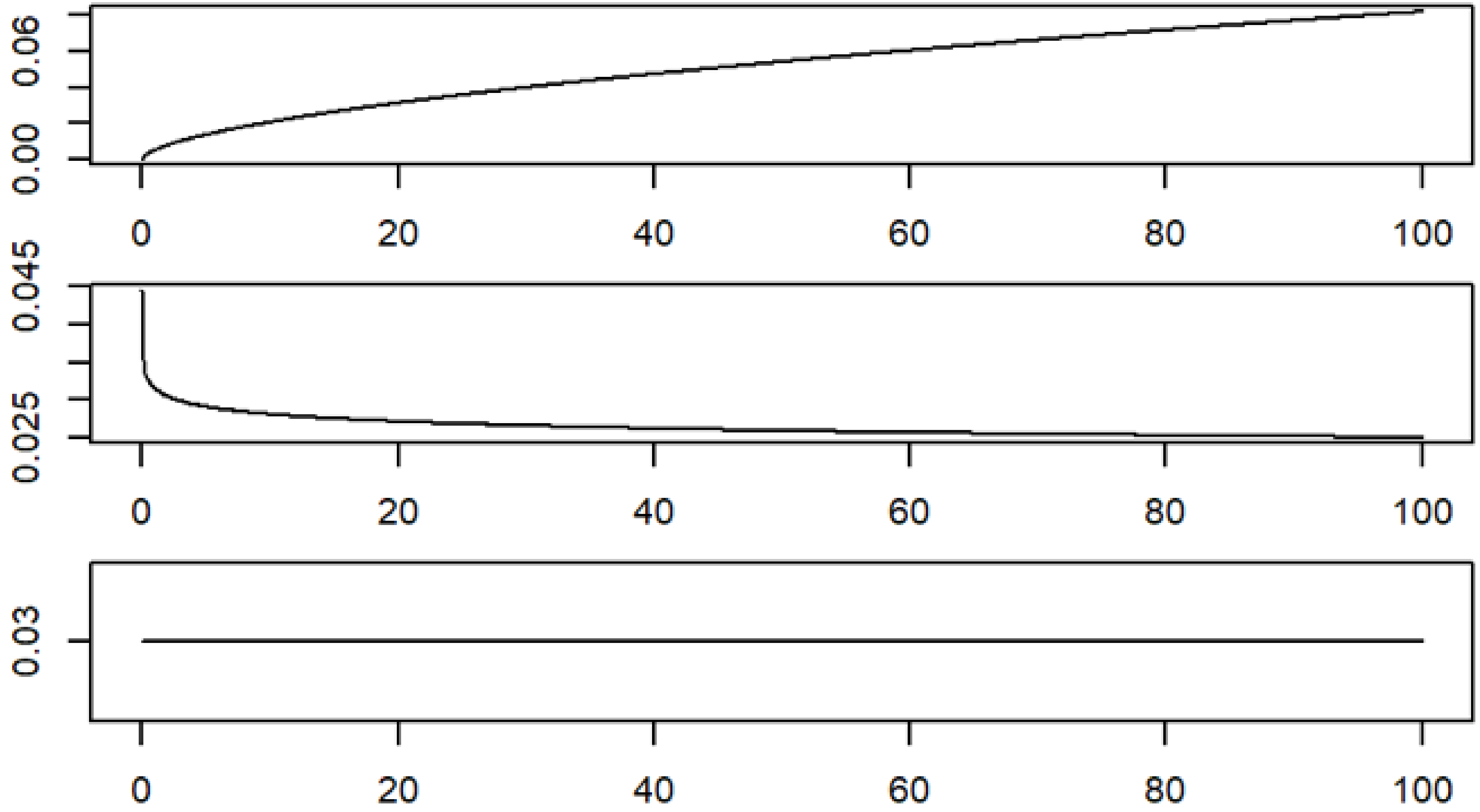


Figure 42: Increasing, decreasing, and constant hazard functions

Summary

- Day one: Numerical summaries and data visualization
- Day two: Hypothesis testing and sampling
- Day three: Statistical tests to compare treatment to a control and regression models

My goal: help you to become a better consumer of statistics

Email (mail@pmean.com) or visit …

- Github site github.com/pmean
  - talks: clinical-statistics, sample-size-justification, power
  - classes: clinical-research-methodology, survival-models
  - papers-and-presentations: illustrating-linear-regression, power
- websites
  - www.pmean.com
    - original website, blog, new website
  - Try googling *topic* site:pmean.com

Speaker notes

Speaker notes

I am a part-time independent consultant, but I offer one hour of advice on any topic for free. So email me. I love quick questions and often take questions that I am asked by email and turn them into web pages.

Anything I do, I put up on the web and make it available for free to anyone. I have a github site that is intended for programming, but I also use a progamming language, R Markdown, to create all my talks.

My website is currently undergoing re-organization. I am try to update a lot of old broken pages and consolidate my old website and blog into a new site that will be easier to maintain and update. The new website also uses R Markdown.

Half of the stuff is projects I started but never got around to finishing. So I apologize if you see something and it is only half-baked. Send me an encouraging email if you want me to update something that you find important in your work.