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.

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

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?

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

Figure 1: Image of a passenger jet with four engines

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

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

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

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

Comparison of a binary outcome

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

Sex * Survived Crosstabulation

			No	Yes	Total
			Survived		
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

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)

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

Room	Before	After	Change	Absolute 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

"- 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

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

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

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
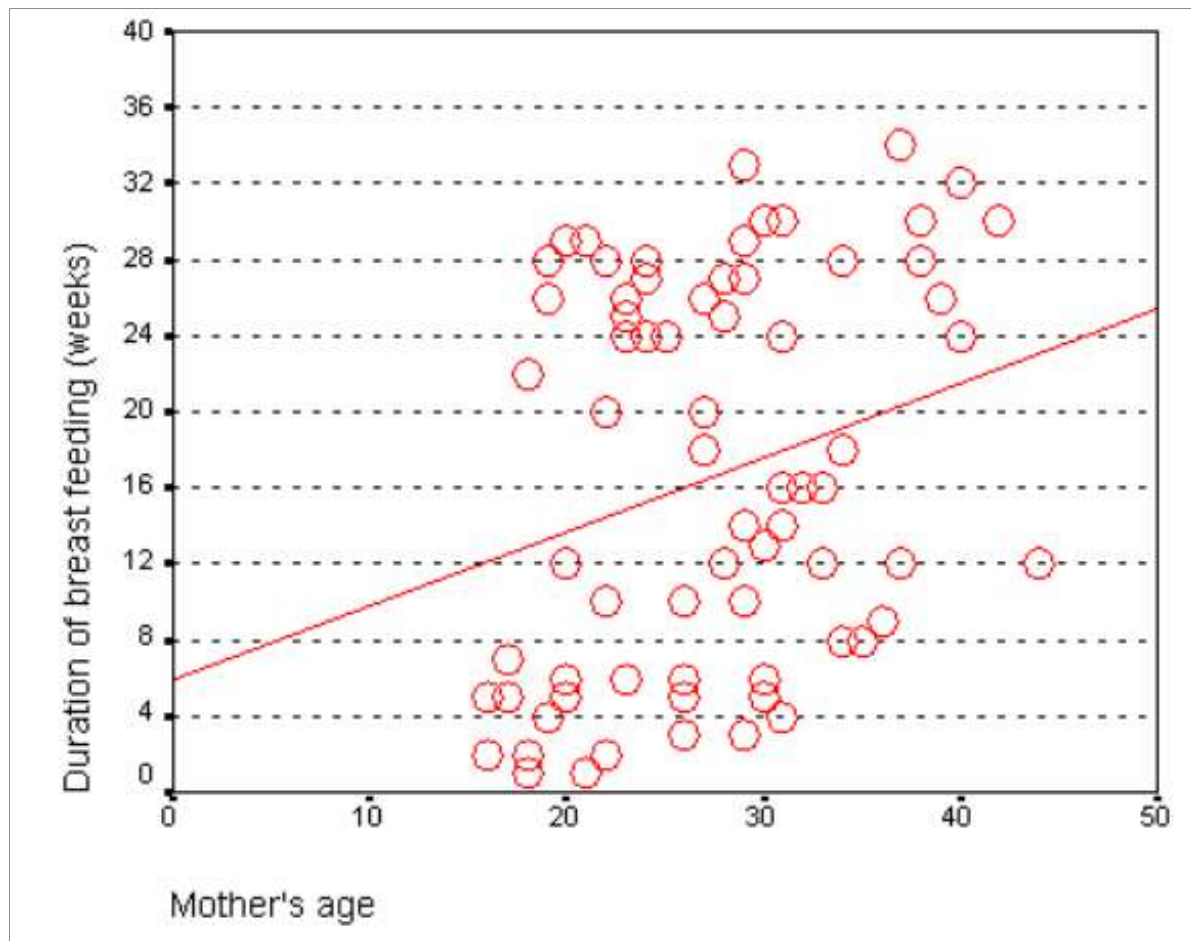
Figure 7: Scatterplot with regression line for age

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	95% Confidence Interval 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
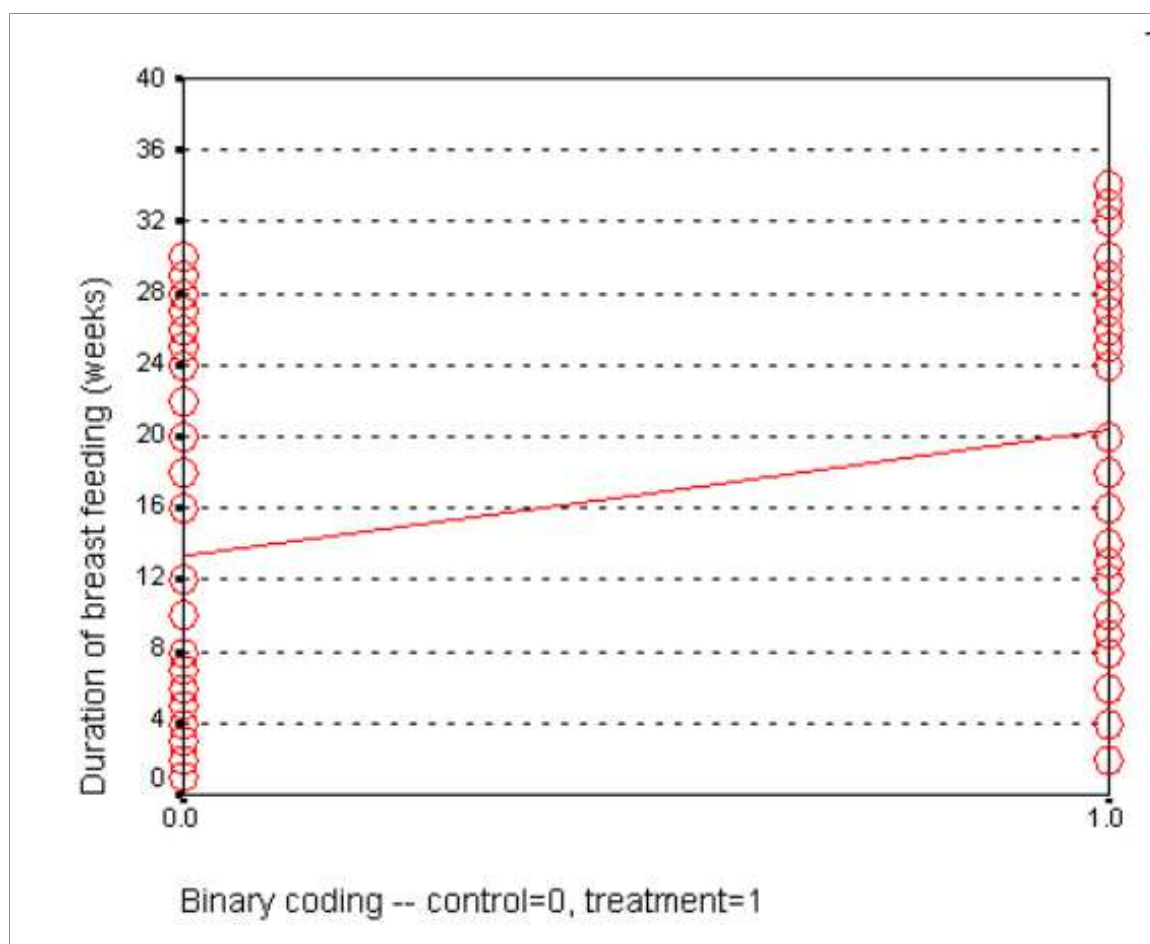
Figure 13: Scatterplot with regression line for treatment=1

Parameter Estimates

Dependent Variable: Duration of breast feeding (weeks)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					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]	0ᵃ	.	.	.	.	.

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

Figure 14: Linear regression output, treatment
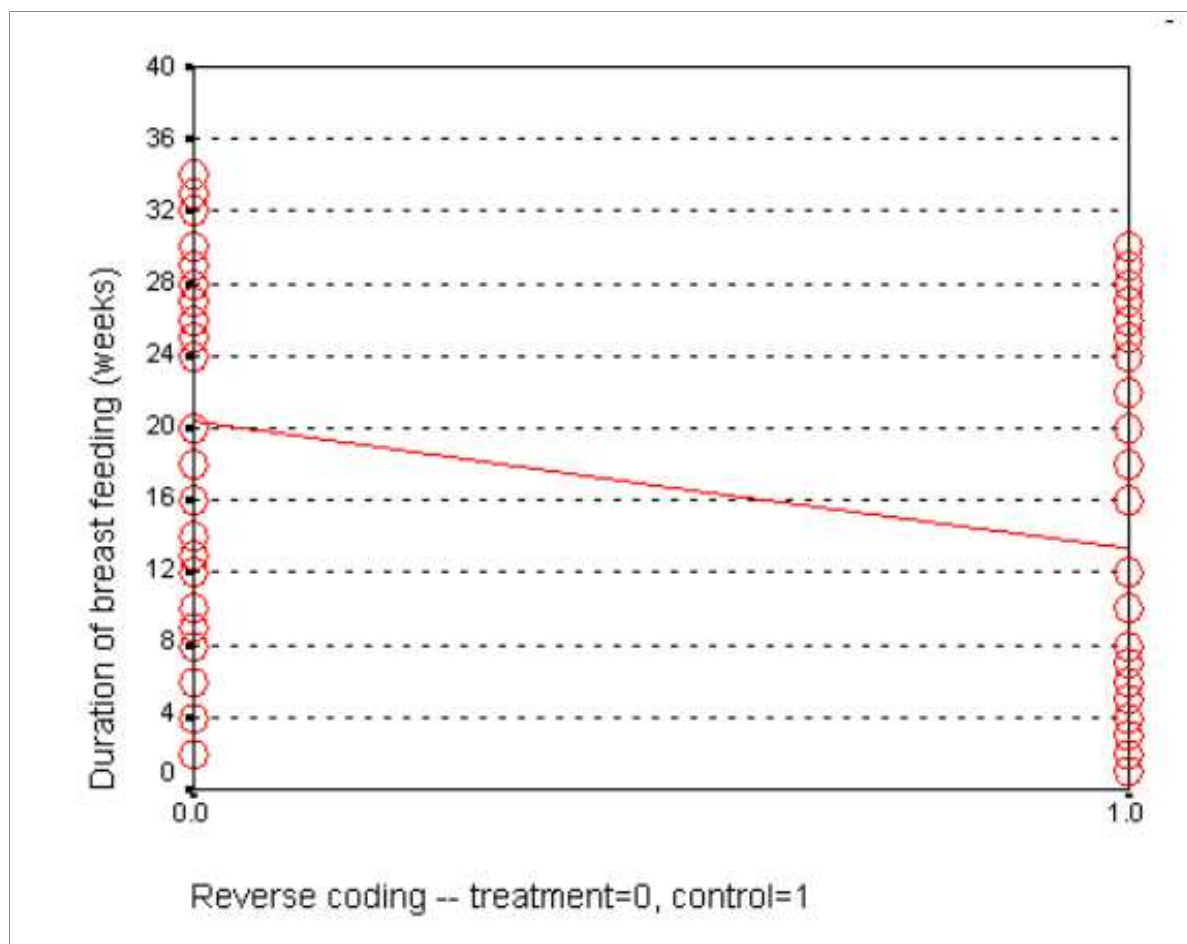
Figure 15: Scatterplot with regression line for control=1

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 x1 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

Logistic regression

- Binary outcome
- Linear on a log odds scale

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

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

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

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)

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: $\frac{1/16}{1+1/16} = \frac{1}{17} = 0.059$

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%
```

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

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

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

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
```

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

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

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
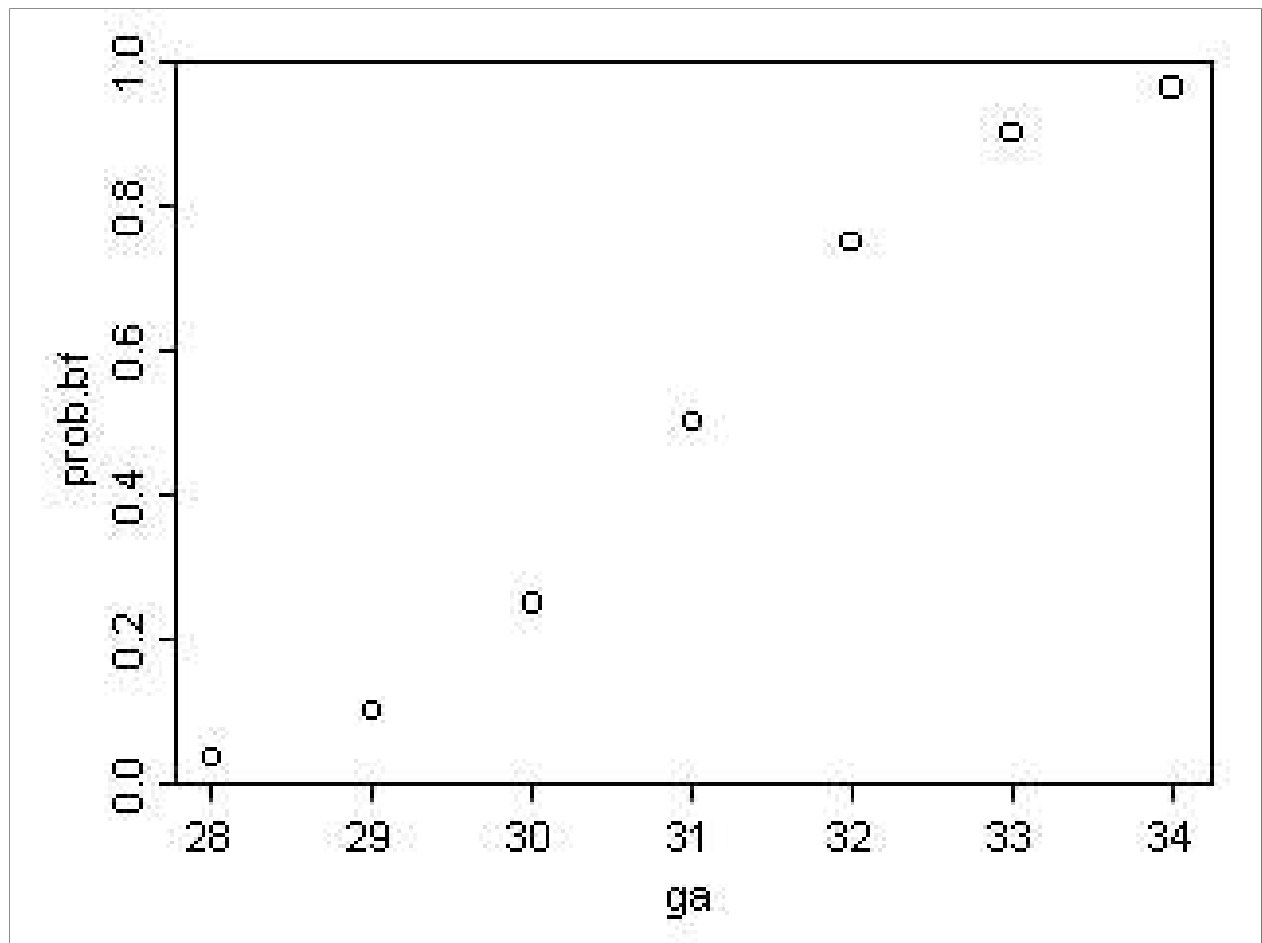
S-shaped curve

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

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

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

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

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

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

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

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	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

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

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

Figure 31: Fruit fly data, round 1
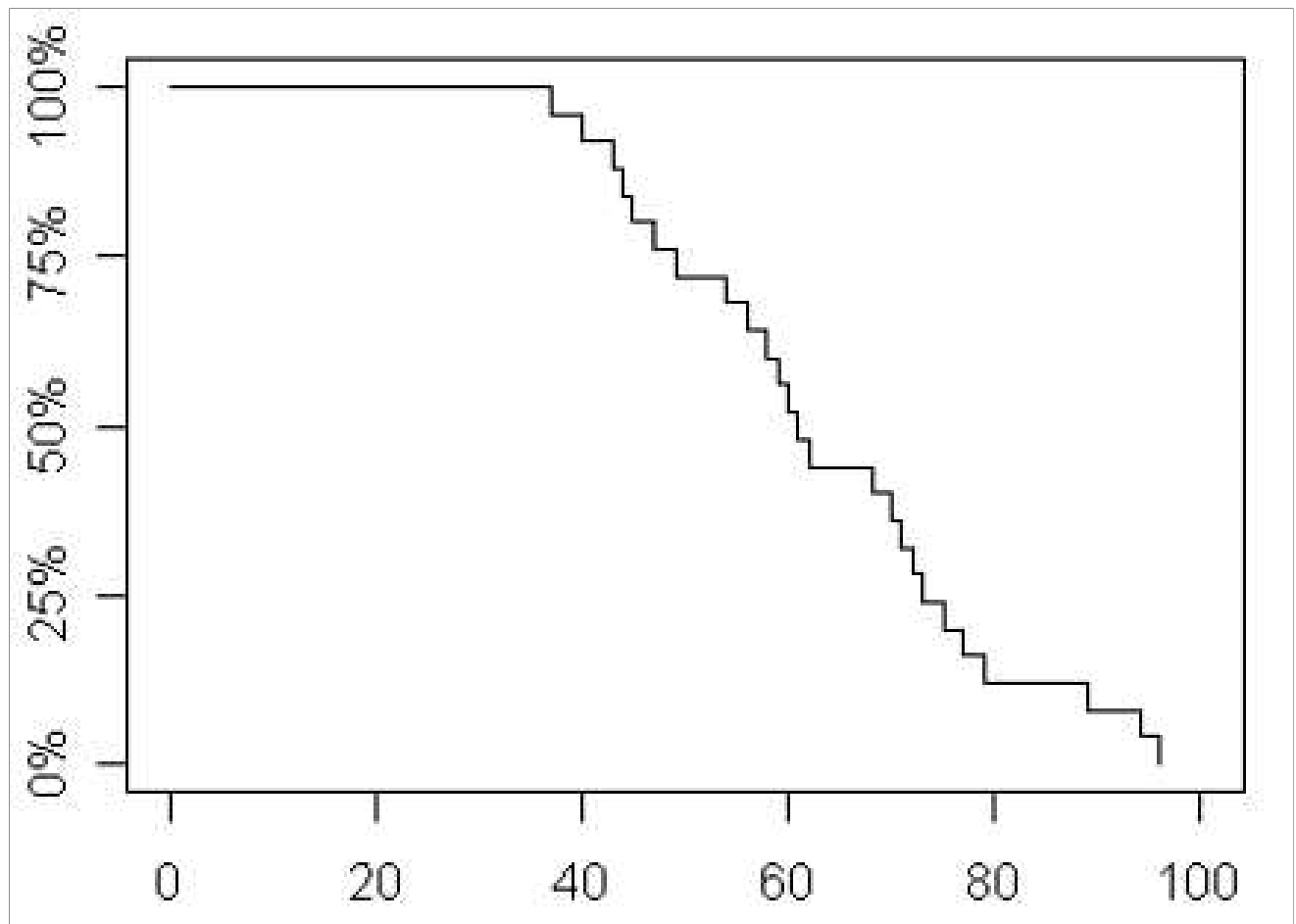
Figure 32: Fruit fly graph, round 1

Fruit fly data (round 2)

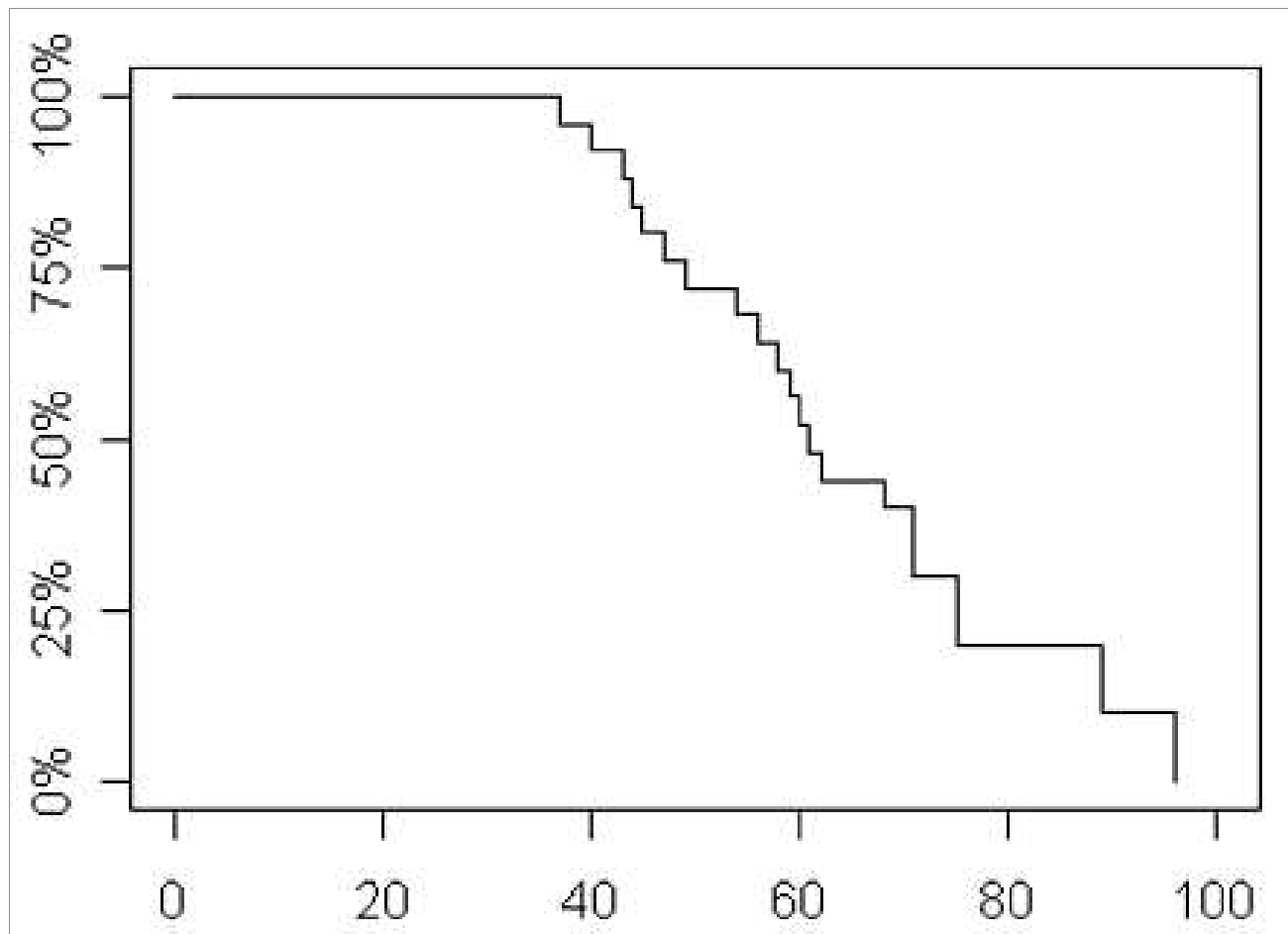|		
37 96%	58 60%	70+
40 92%	59 56%	70+
43 88%	60 52%	70+
44 84%	61 48%	70+
45 80%	62 44%	70+
47 76%	68 40%	70+
49 72%	70+ ?	70+
54 68%	70+ ?	
56 64%	70+ ?	

Figure 33: Fruit fly data, round 2
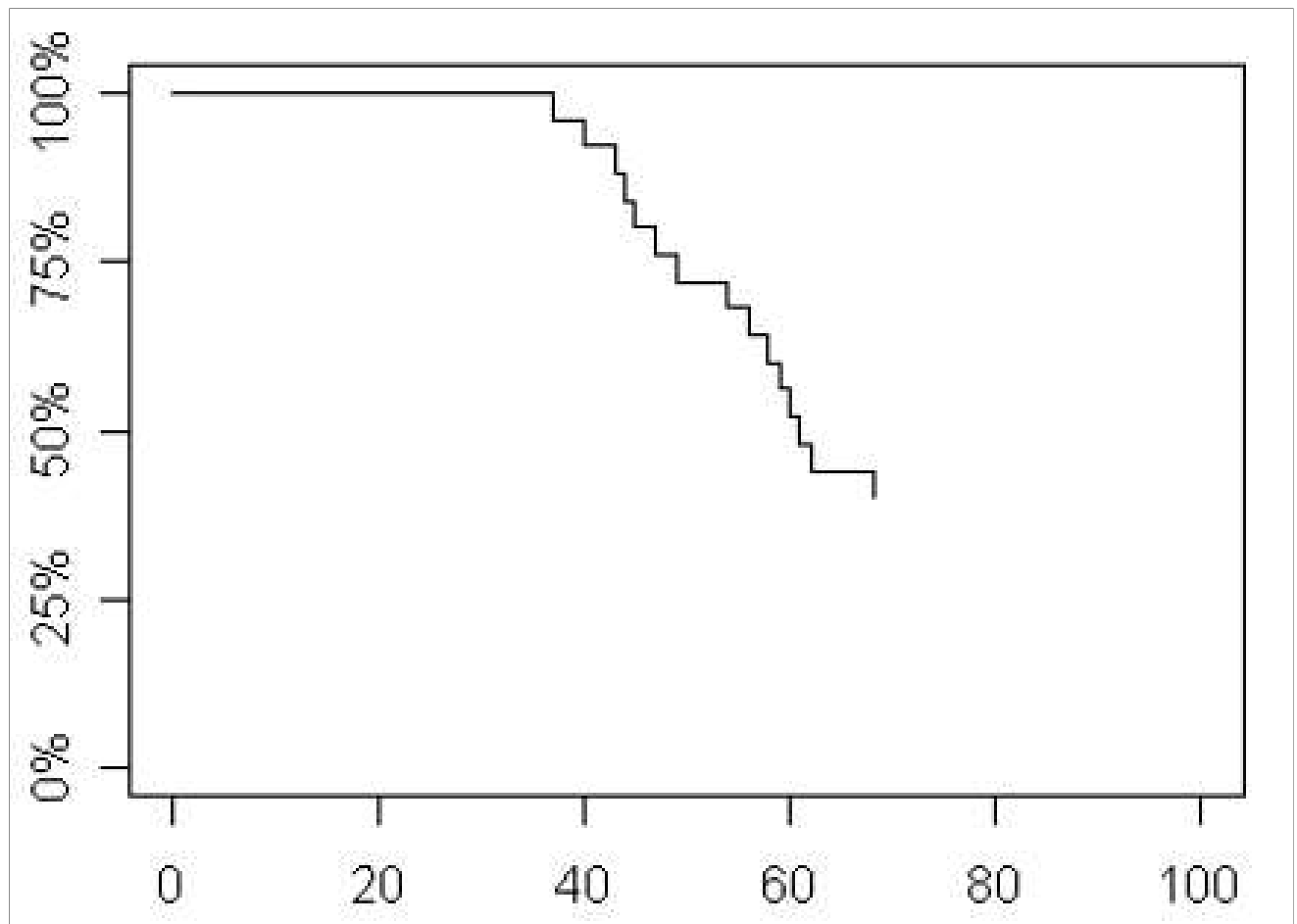
Figure 34: Fruit fly graph, round 2

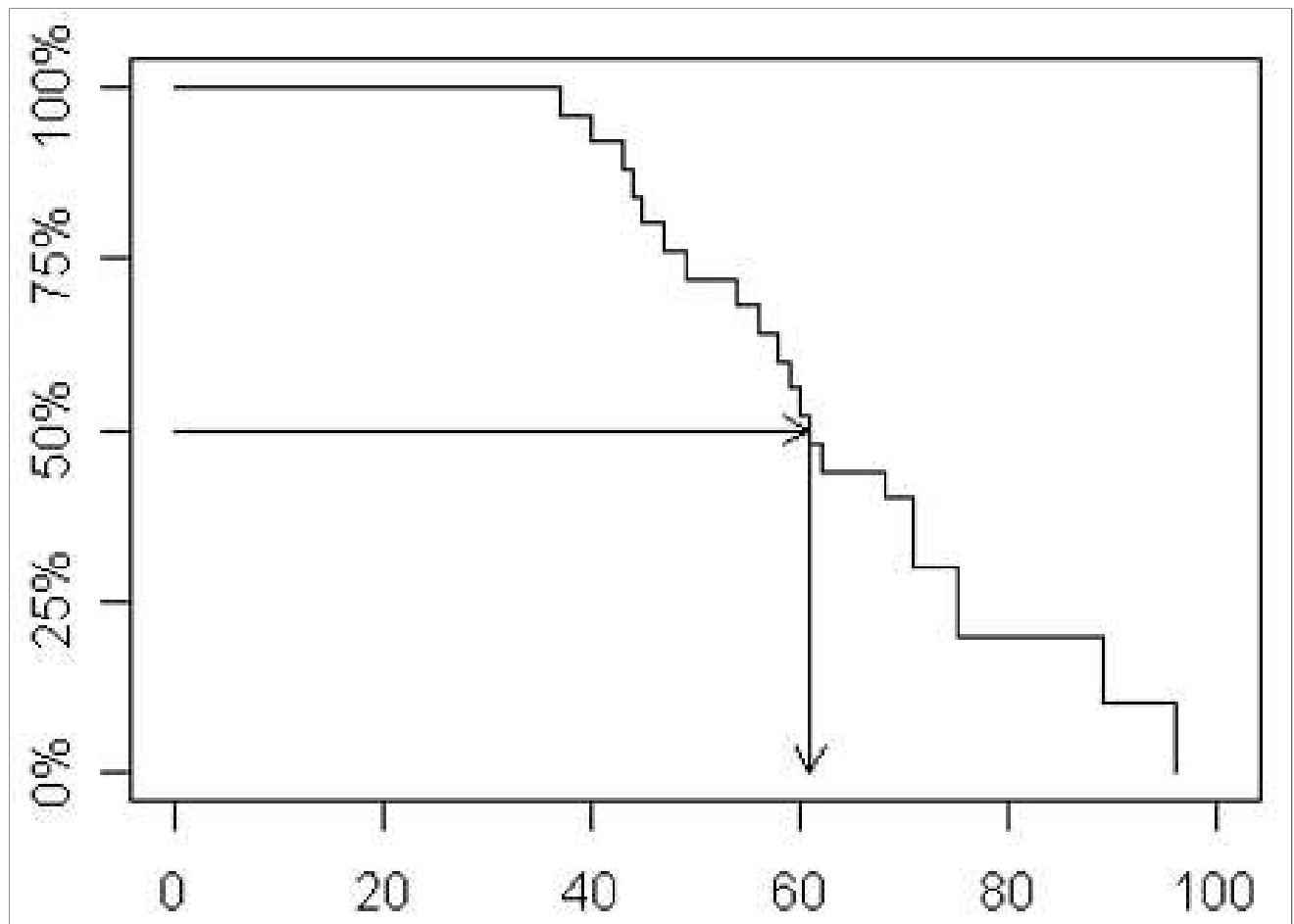|				
37	96%	58	60%	70+ ?
40	92%	59	56%	75 20%
43	88%	60	52%	70+ ?
44	84%	61	48%	70+ ?
45	80%	62	44%	89 10%
47	76%	68	40%	70+ ?
49	72%		70+ ?	96 0%
54	68%	71	30%	
56	64%		70+ ?	

Figure 35: Fruit fly data, round 3

Figure 36: Fruit fly graph, round 3
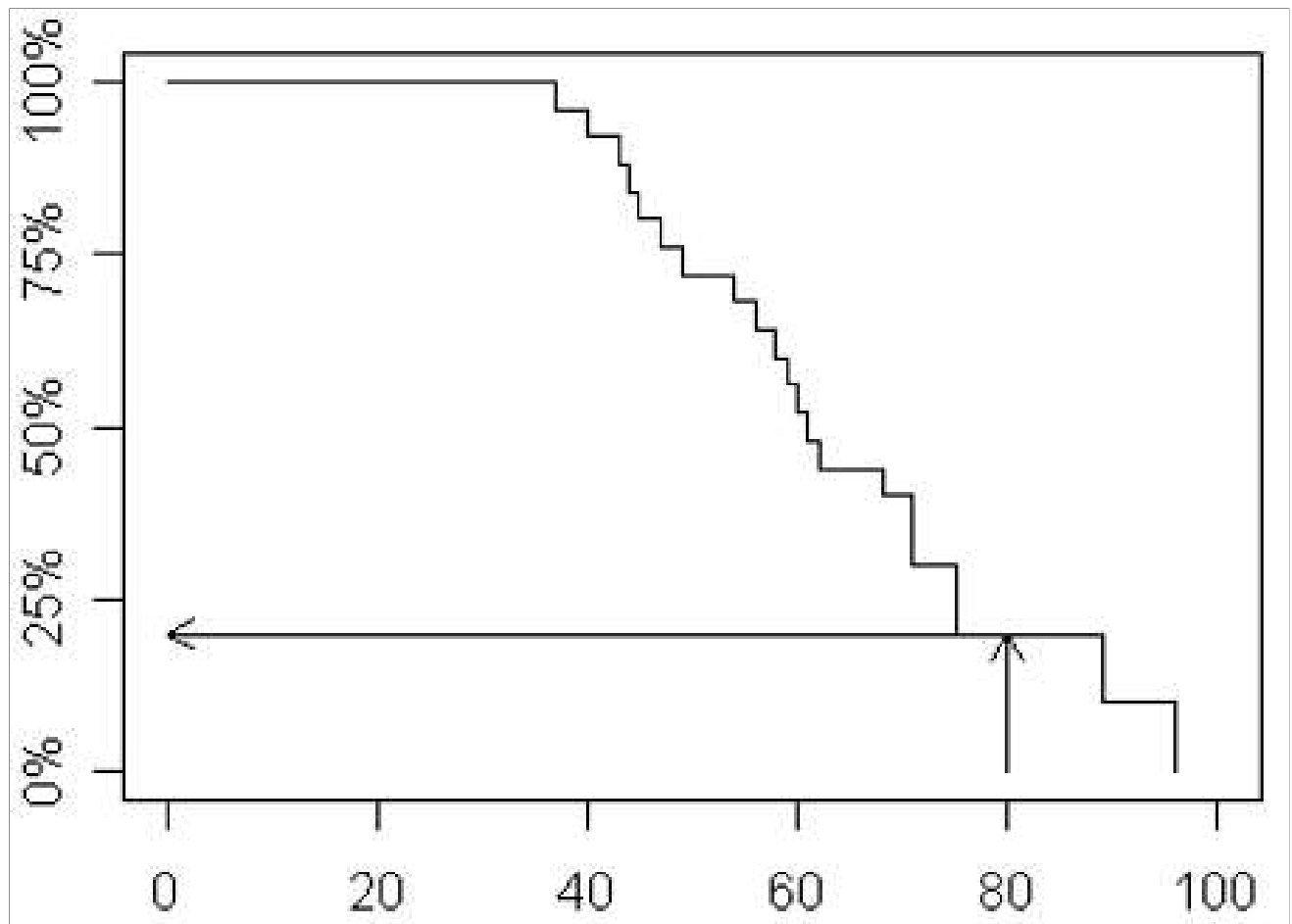
Figure 37: Fruit fly graph, estimating the median

Figure 38: Fruit fly graph, estimating survival probability
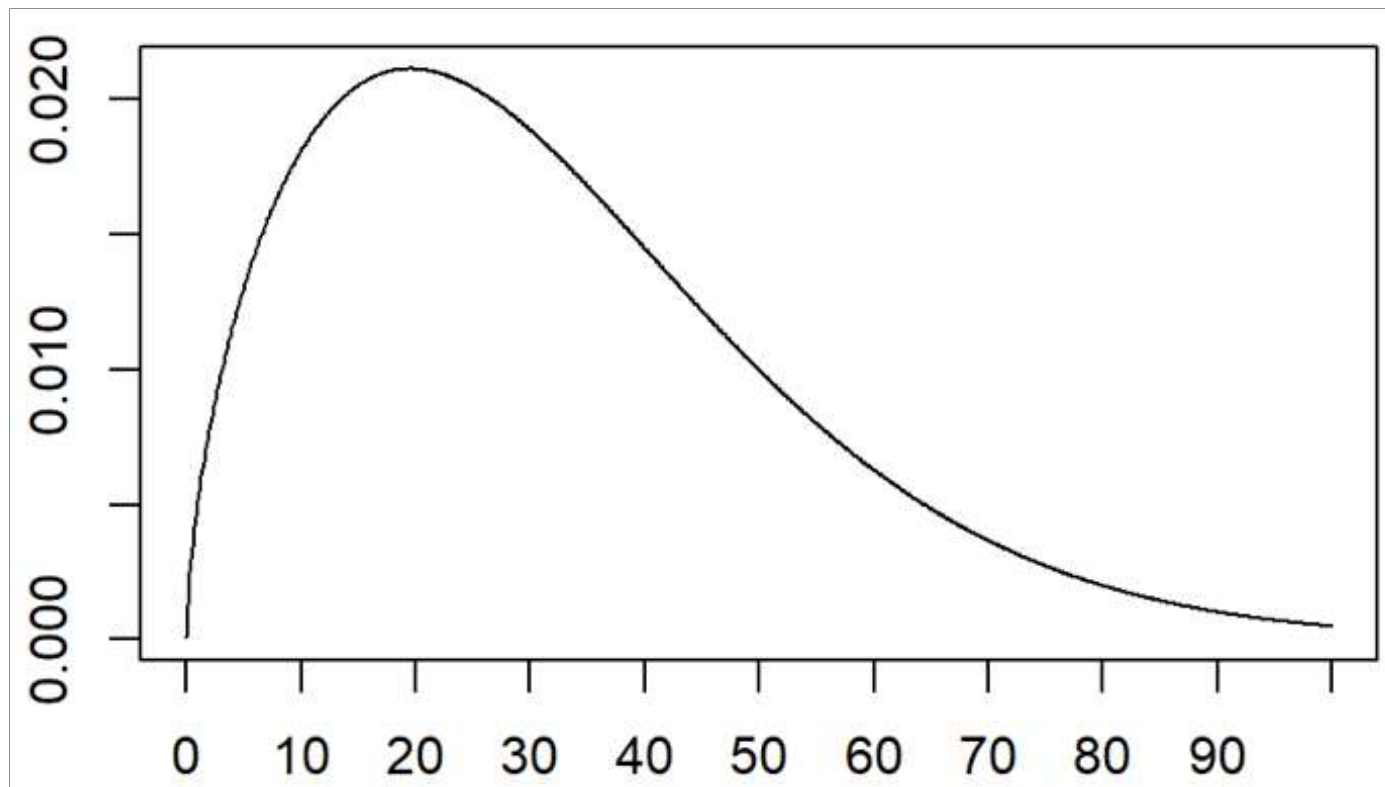
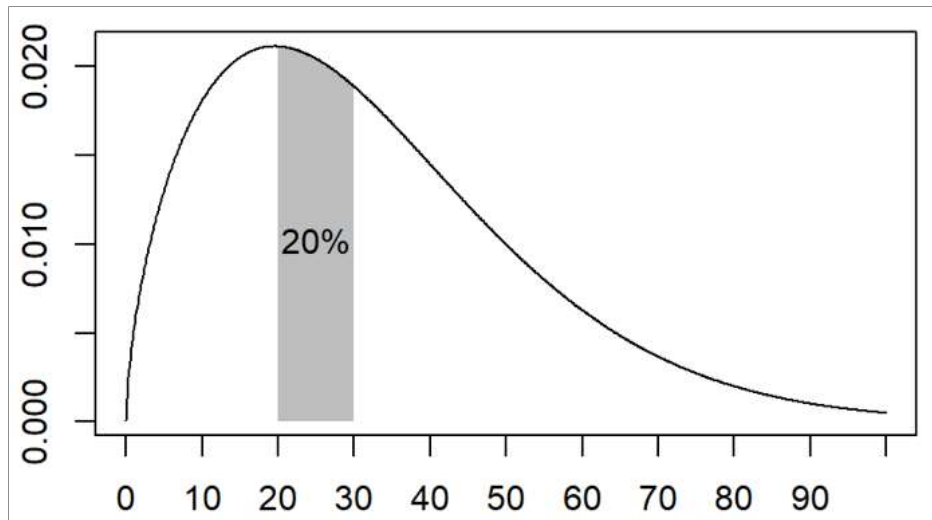Figure 39: Hypothetical survival distribution

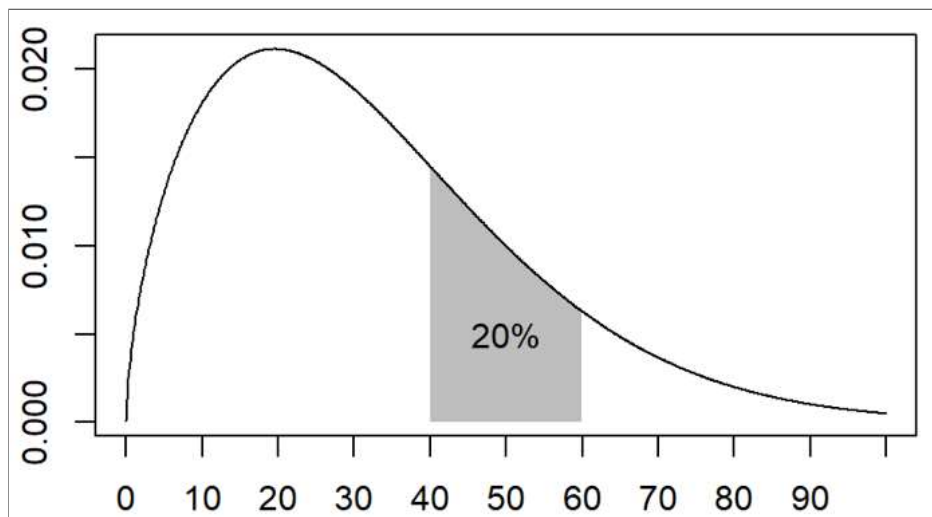Figure 40: Hypothetical survival distribution, probability for 20-30 years

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 ≤ T ≤ t+Δt] / Δt) / P[T ≥ t]
- Key points
  - adjusted for the number surviving to that time (P[T ≥ t]),
  - calculated as a rate
    - (P[t ≤ T ≤ T+Δt] / Δt) is not a probability, and
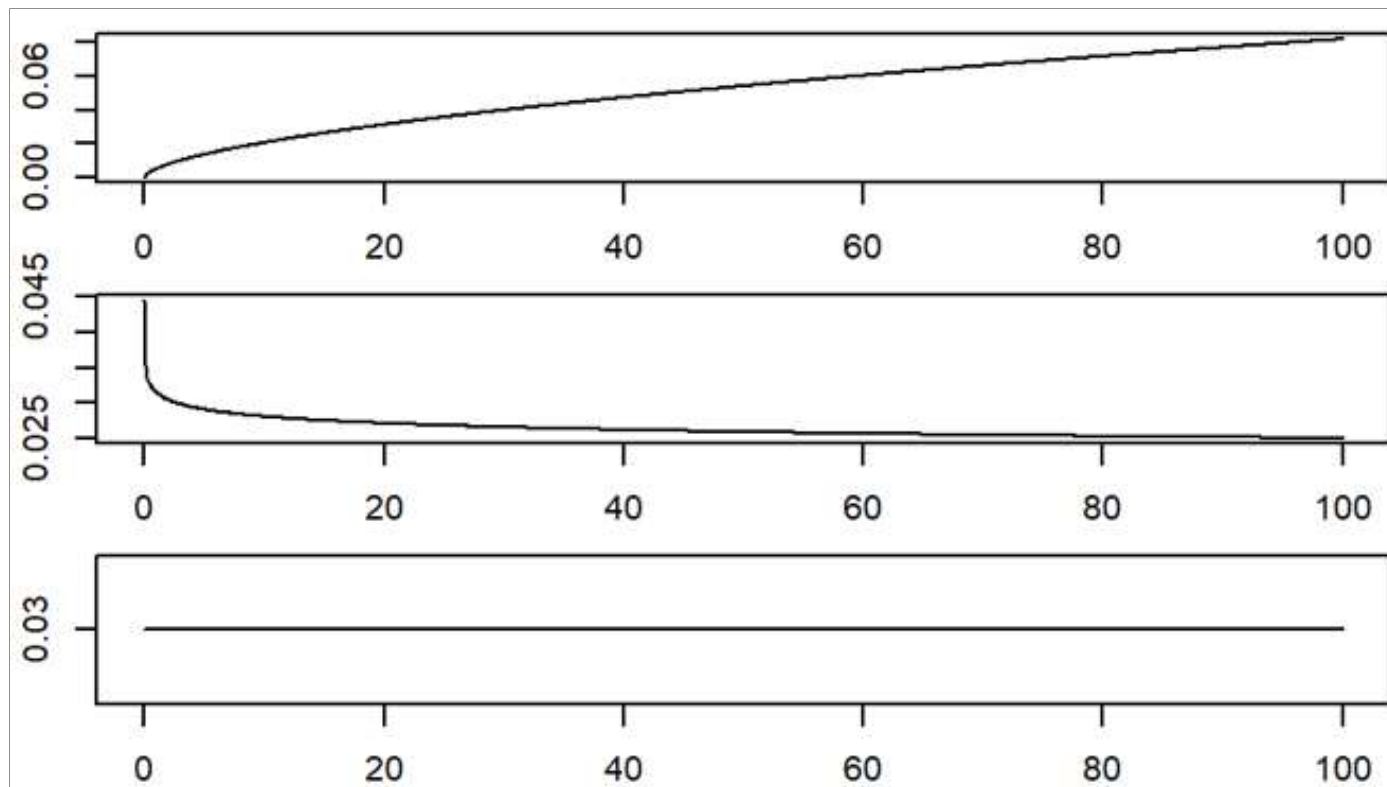 - computed over a narrow time interval.

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

Error   ×