How to display data badly

Karl W Broman

Department of Biostatistics & Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman
github.com/kbroman
@kwbroman

Using Microsoft Excel to obscure your data and annoy your readers

Karl W Broman

Department of Biostatistics & Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman
github.com/kbroman
@kwbroman

Inspiration

This lecture was inspired by

> H Wainer (1984) How to display data badly. American Statistician 38(2): 137–147

Dr. Wainer was the first to elucidate the principles of the bad display of data.

The now widespread use of Microsoft Excel has resulted in remarkable advances in the field.

General principles

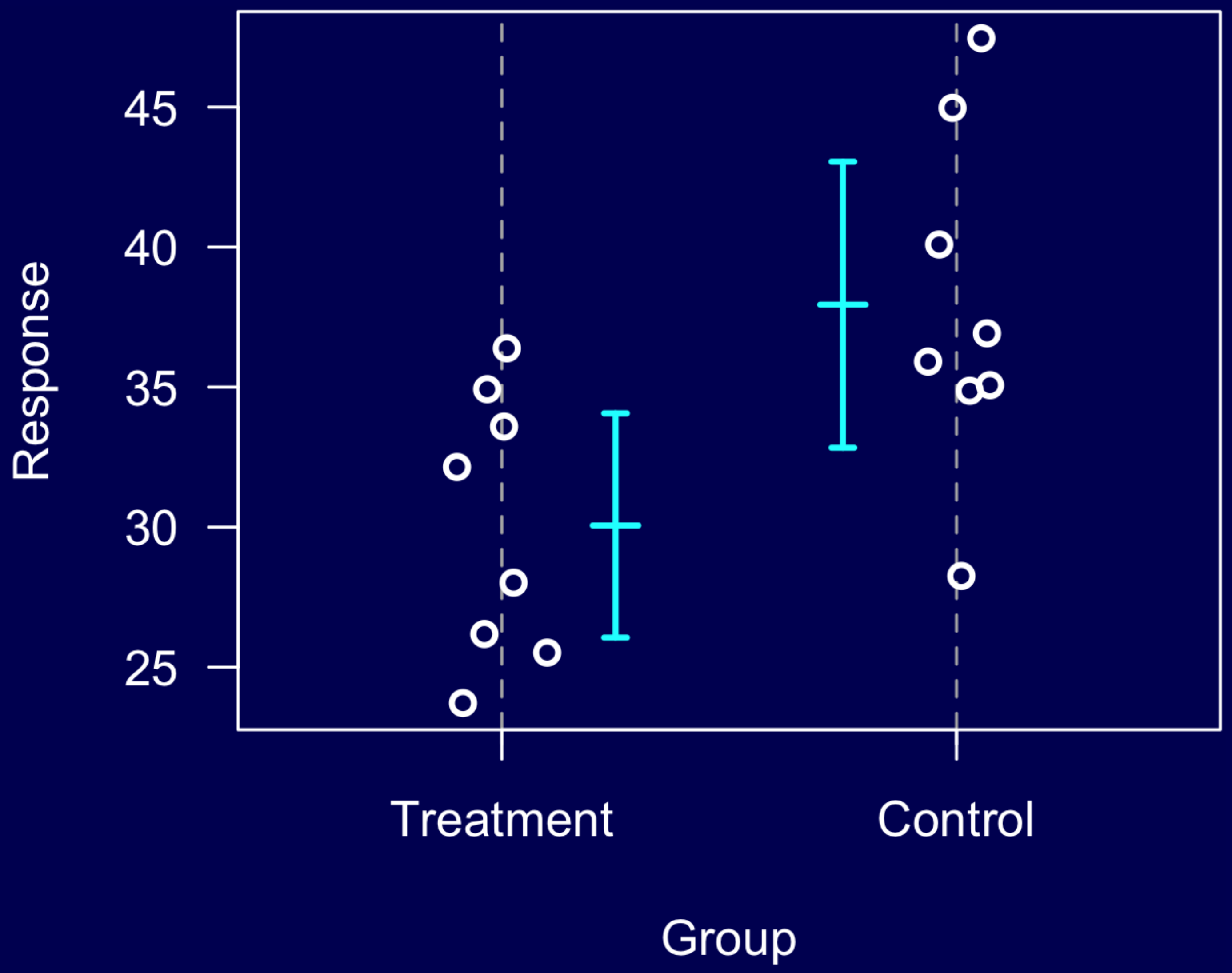
The aim of good data graphics:

Display data accurately and clearly.

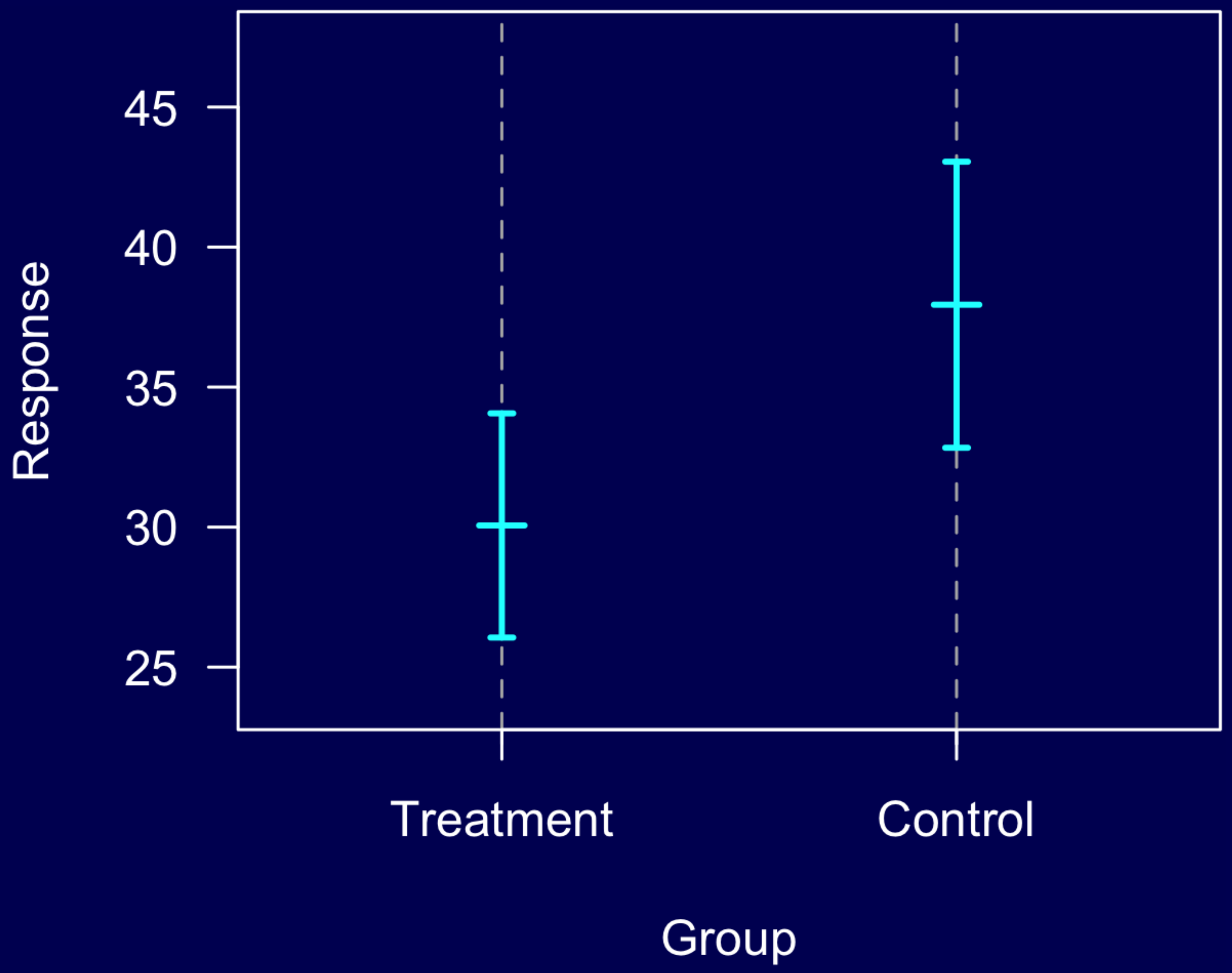
Some rules for displaying data badly:

- Display as little information as possible.
- Obscure what you do show (with chart junk).
- Use pseudo-3d and color gratuitously.
- Make a pie chart (preferably in color and 3d).
- Use a poorly chosen scale.
- Ignore sig figs.

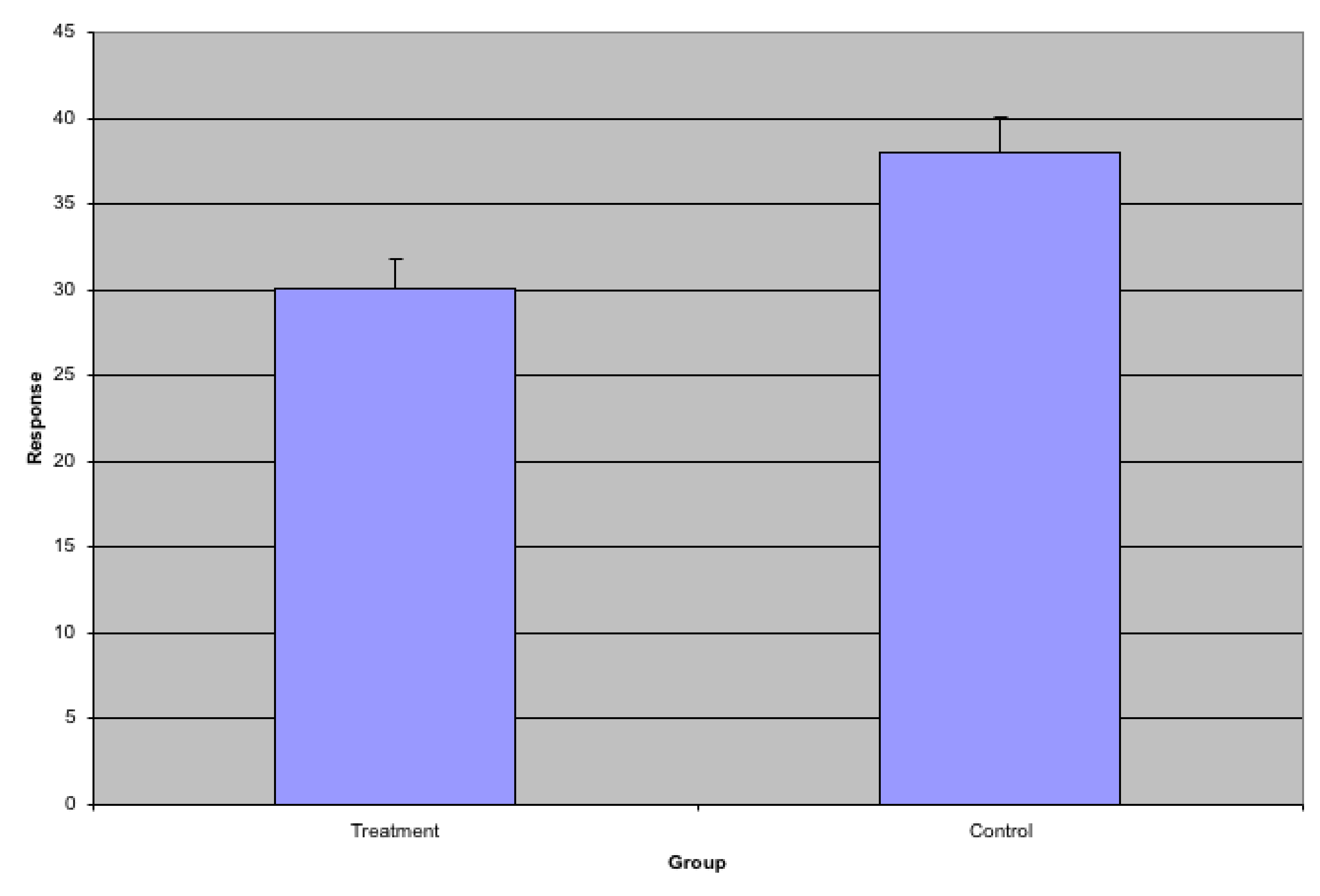
Example 1



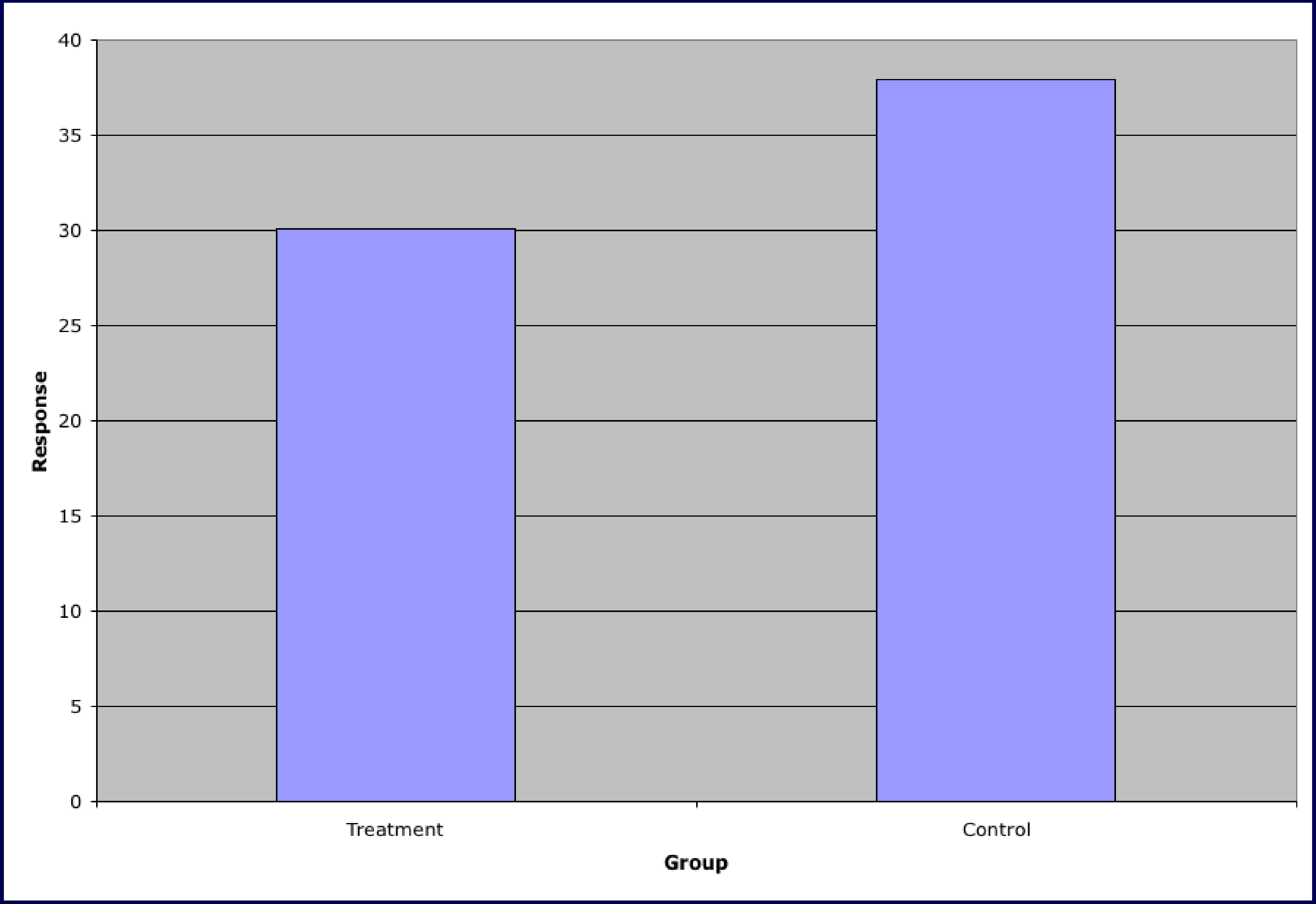
Example 1



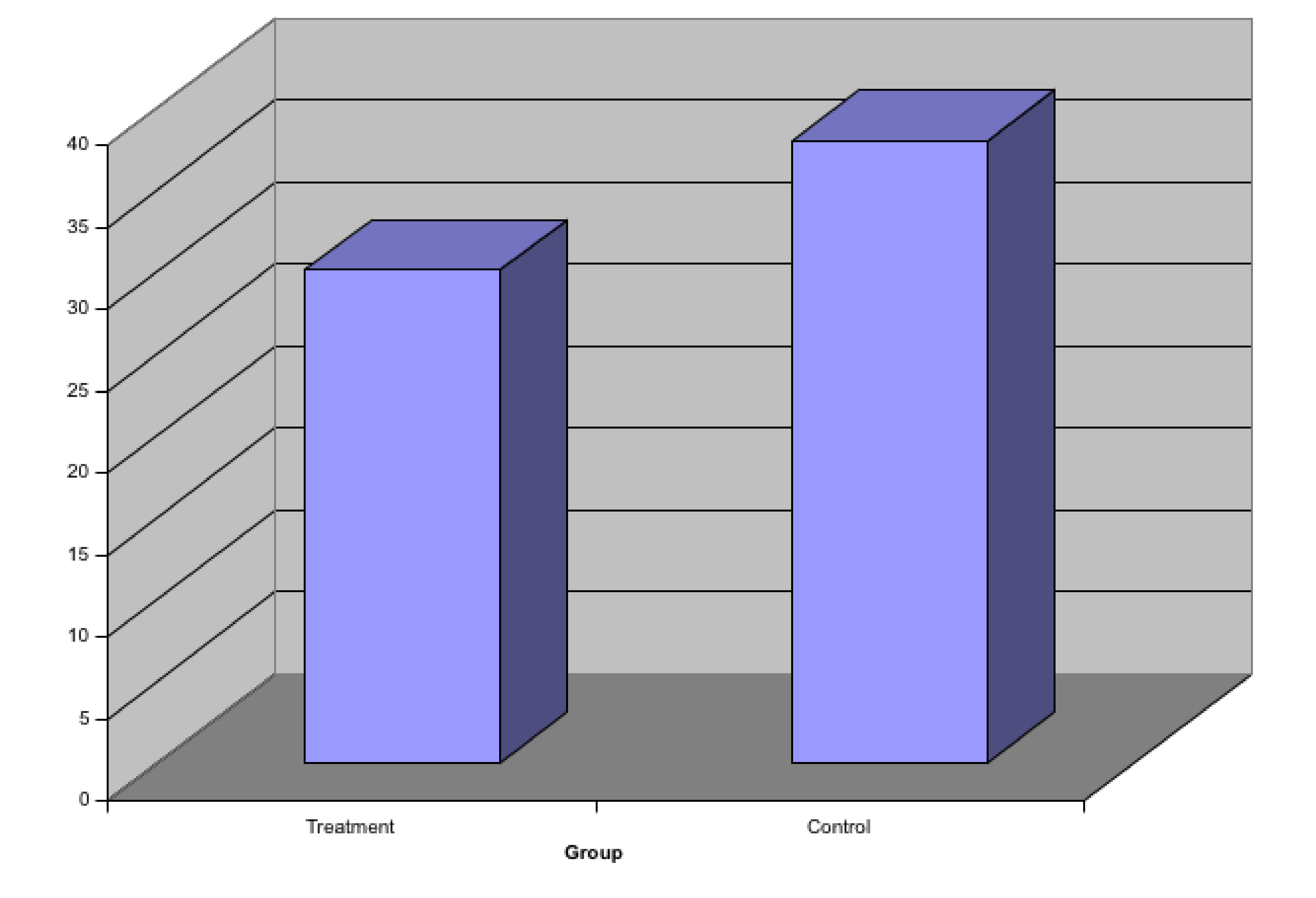
Example 1



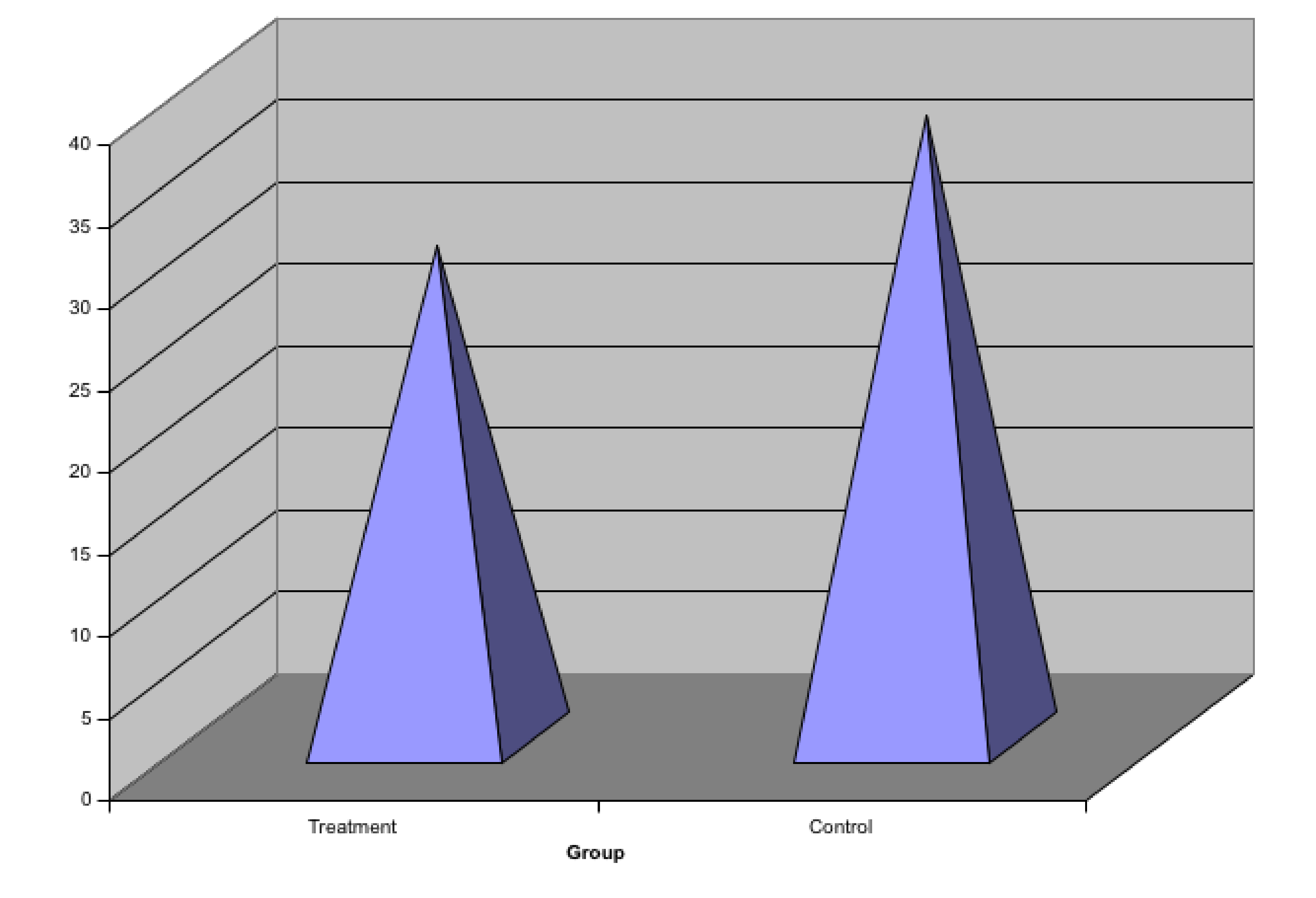
Example 1



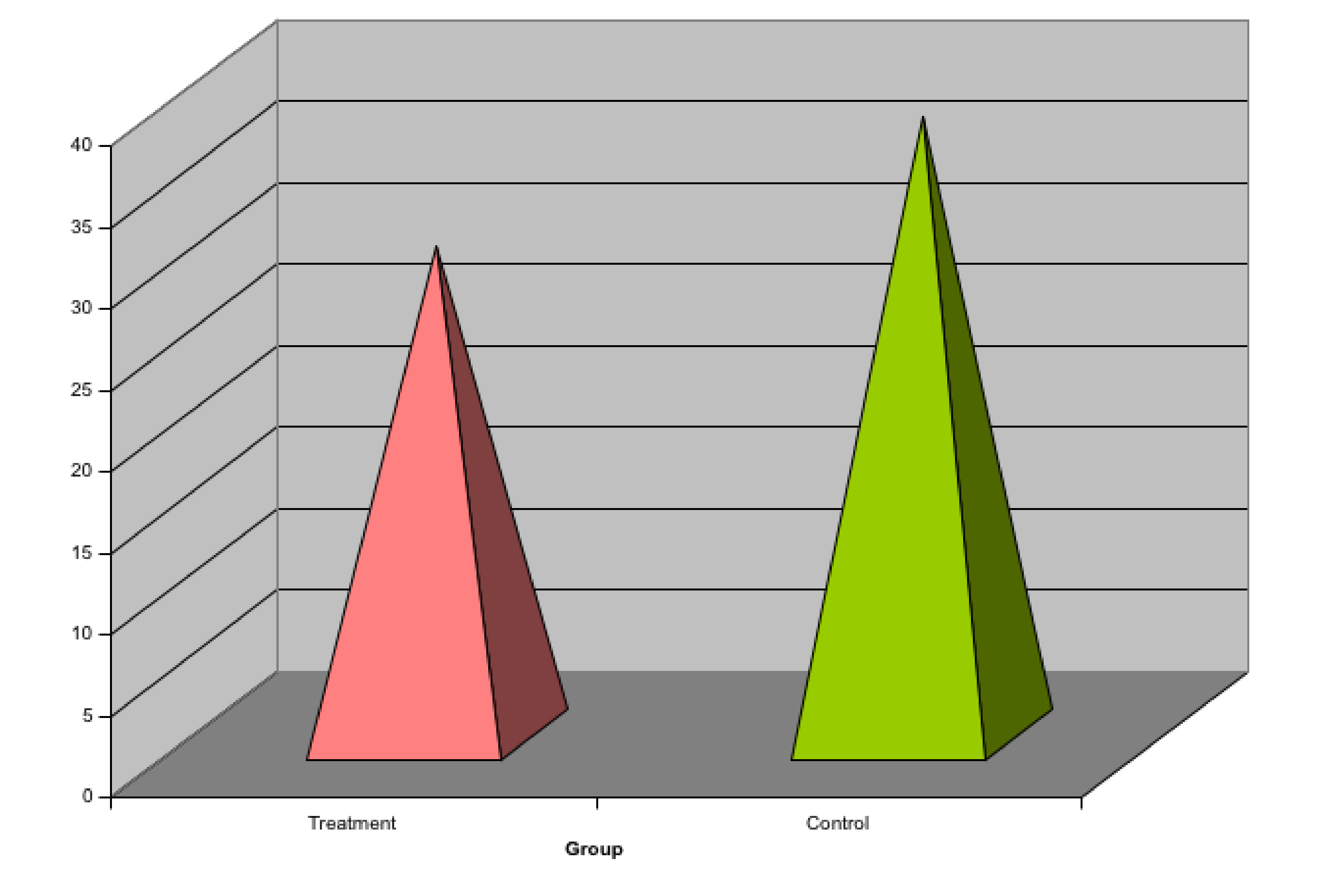
Example 1



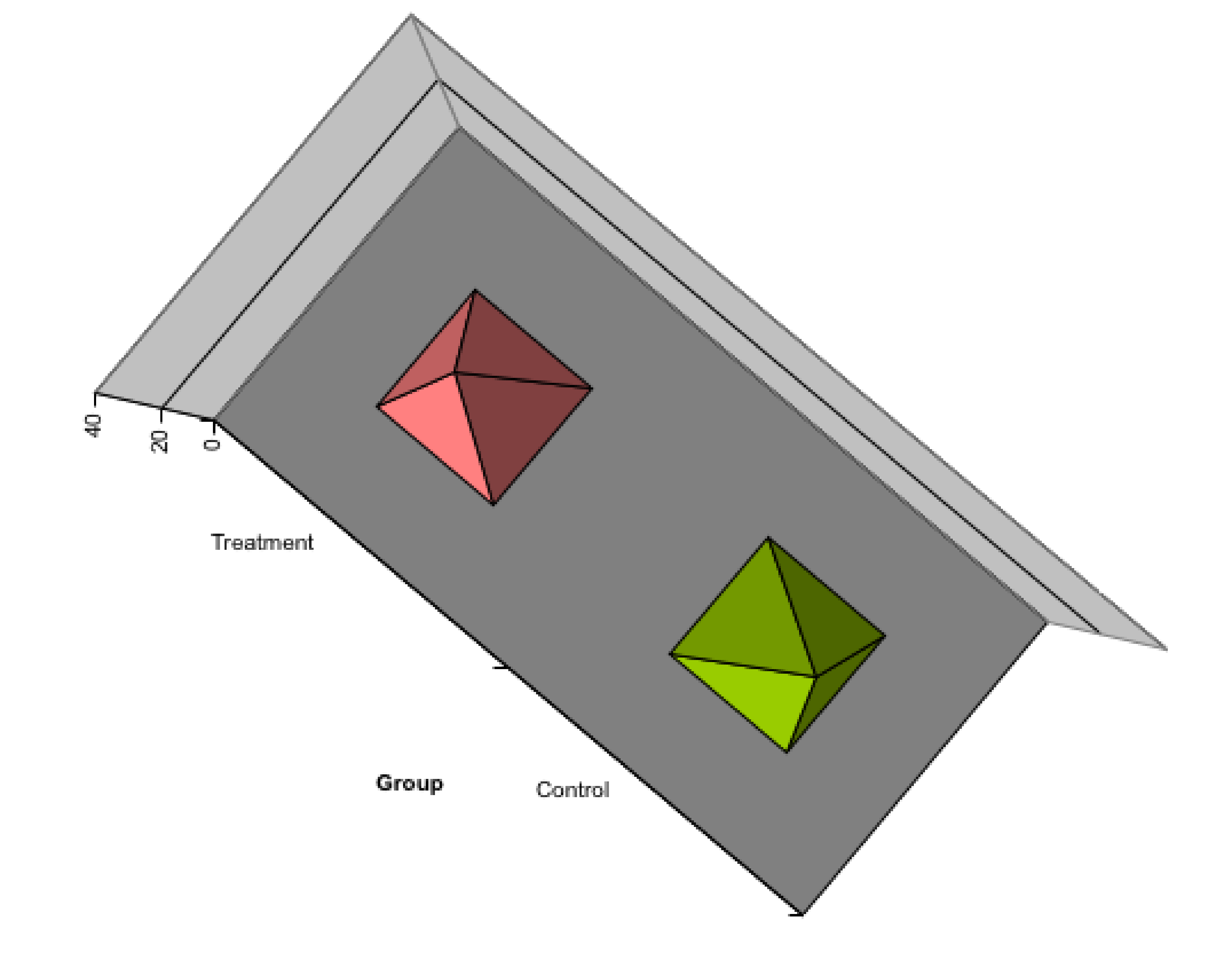
Example 1



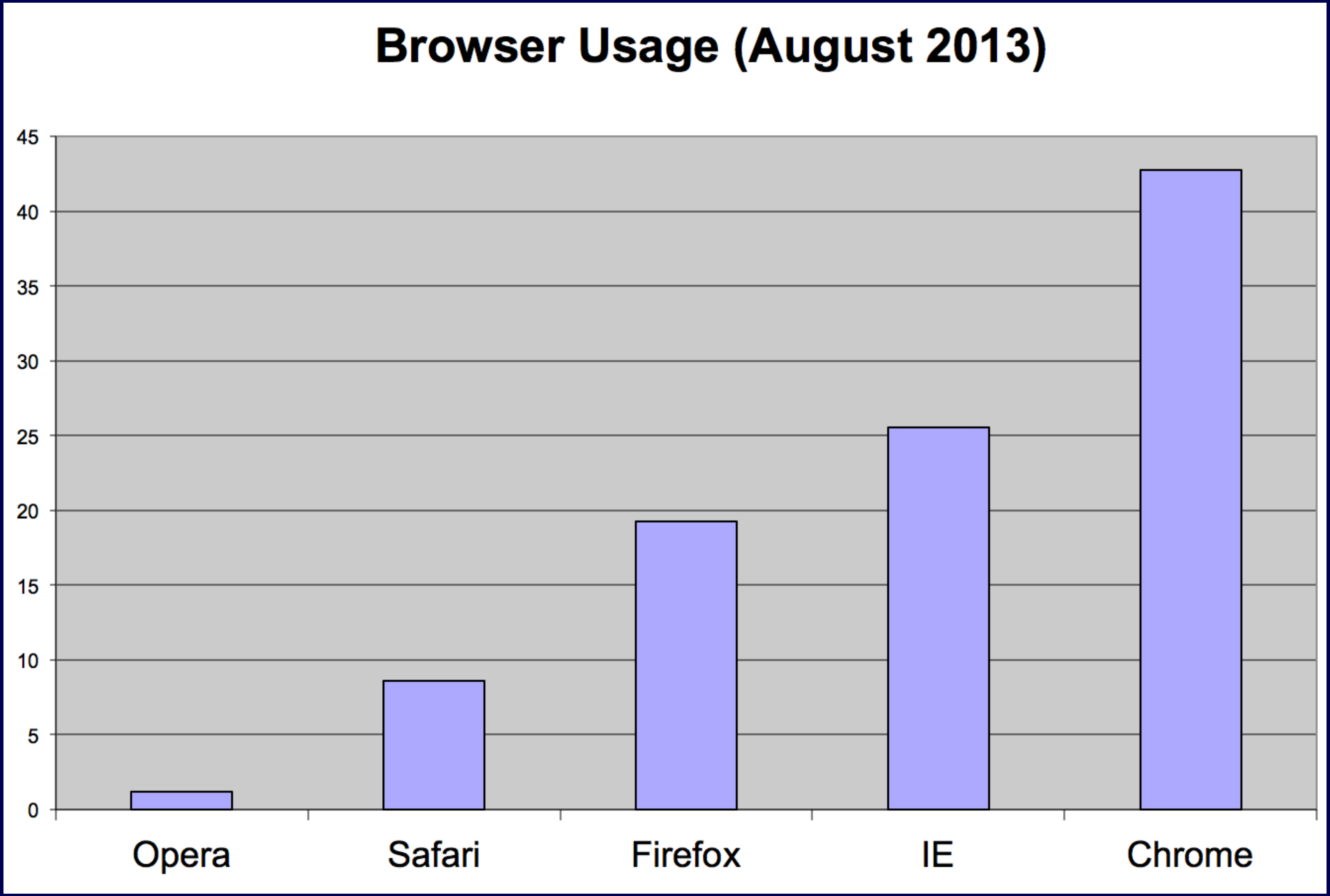
Example 1



Example 1

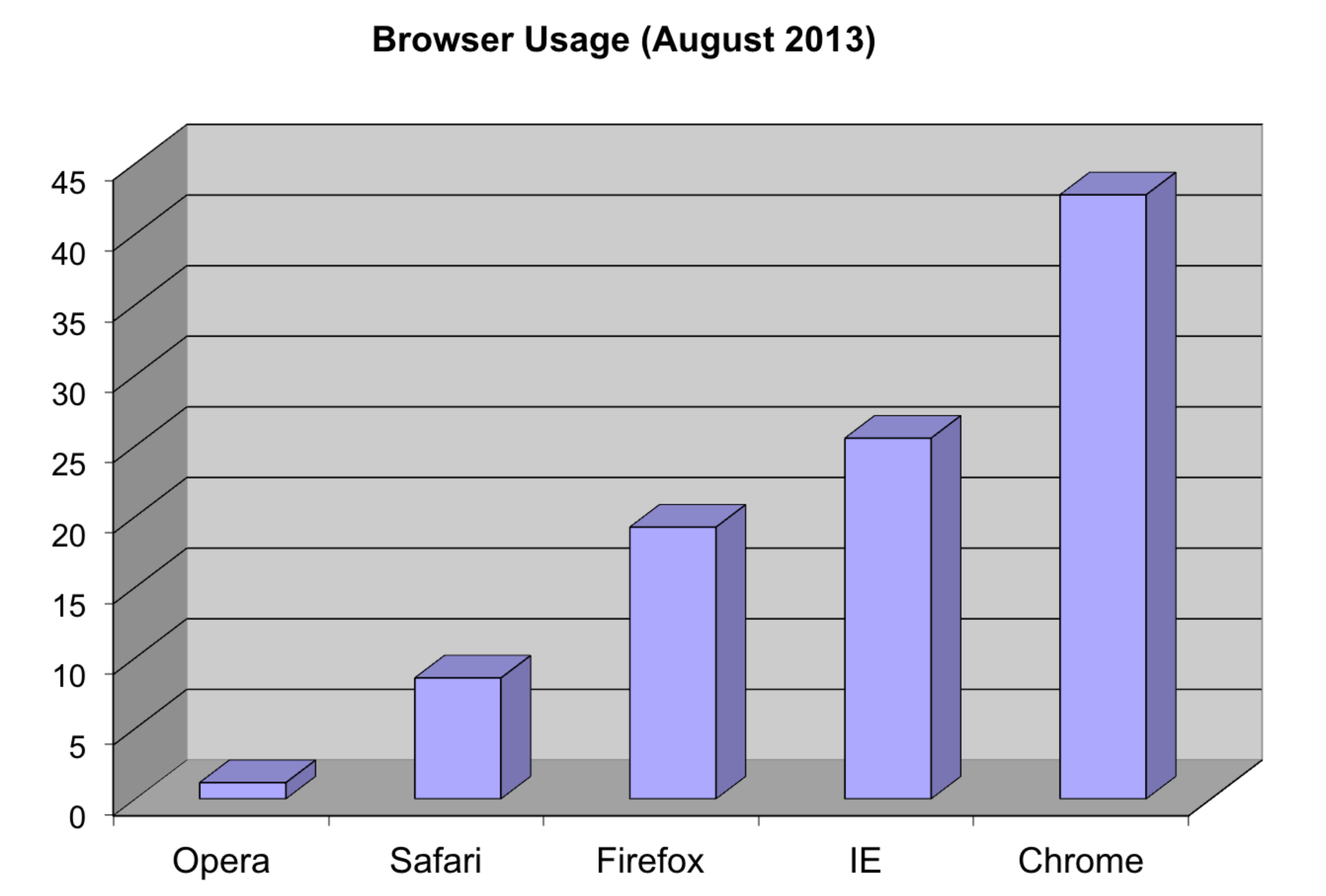


Example 2

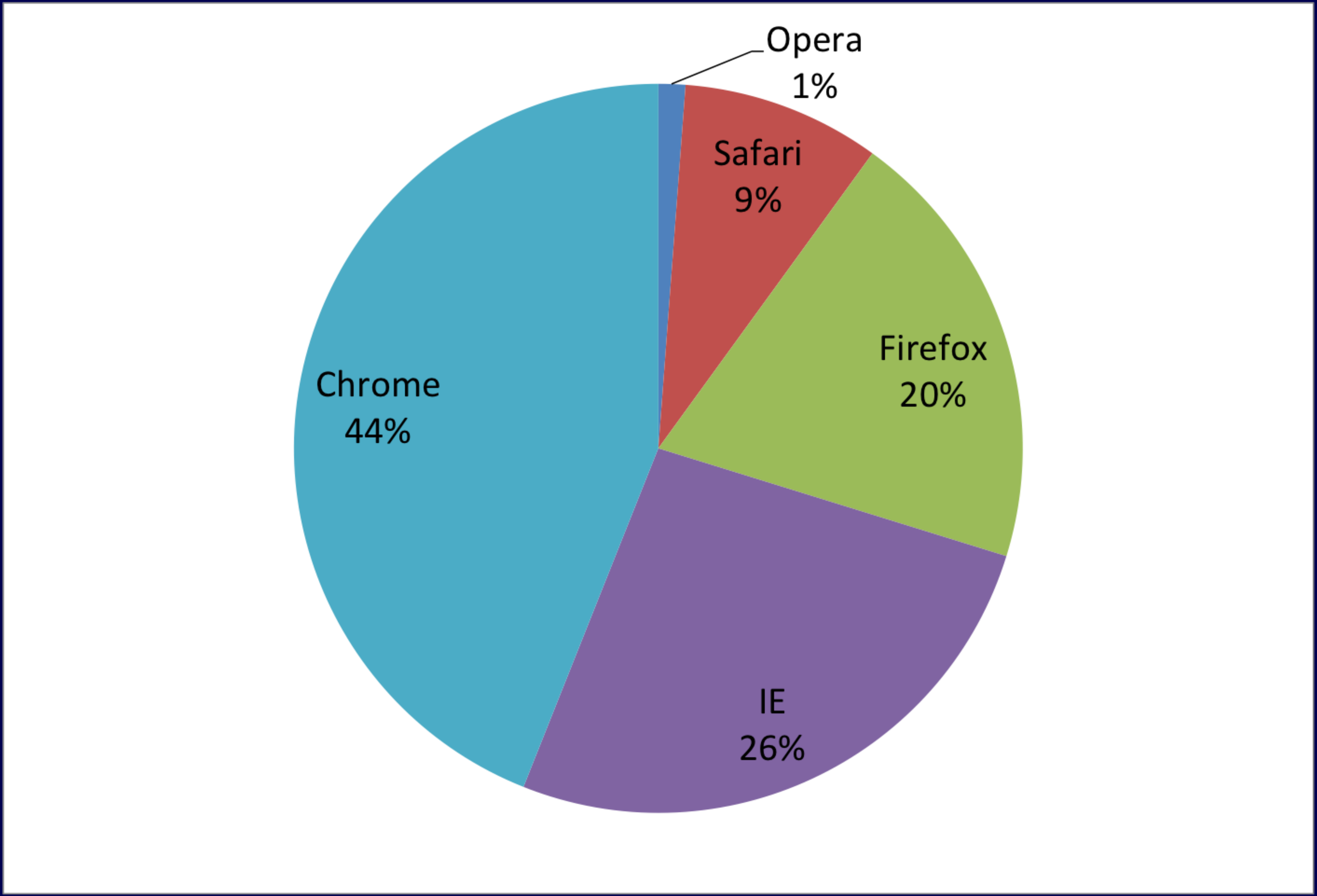


Browser Usage (August 2013)
45
40
35
30
25
20
15
10
5
0
Opera
Safari
Firefox
IE
Chrome

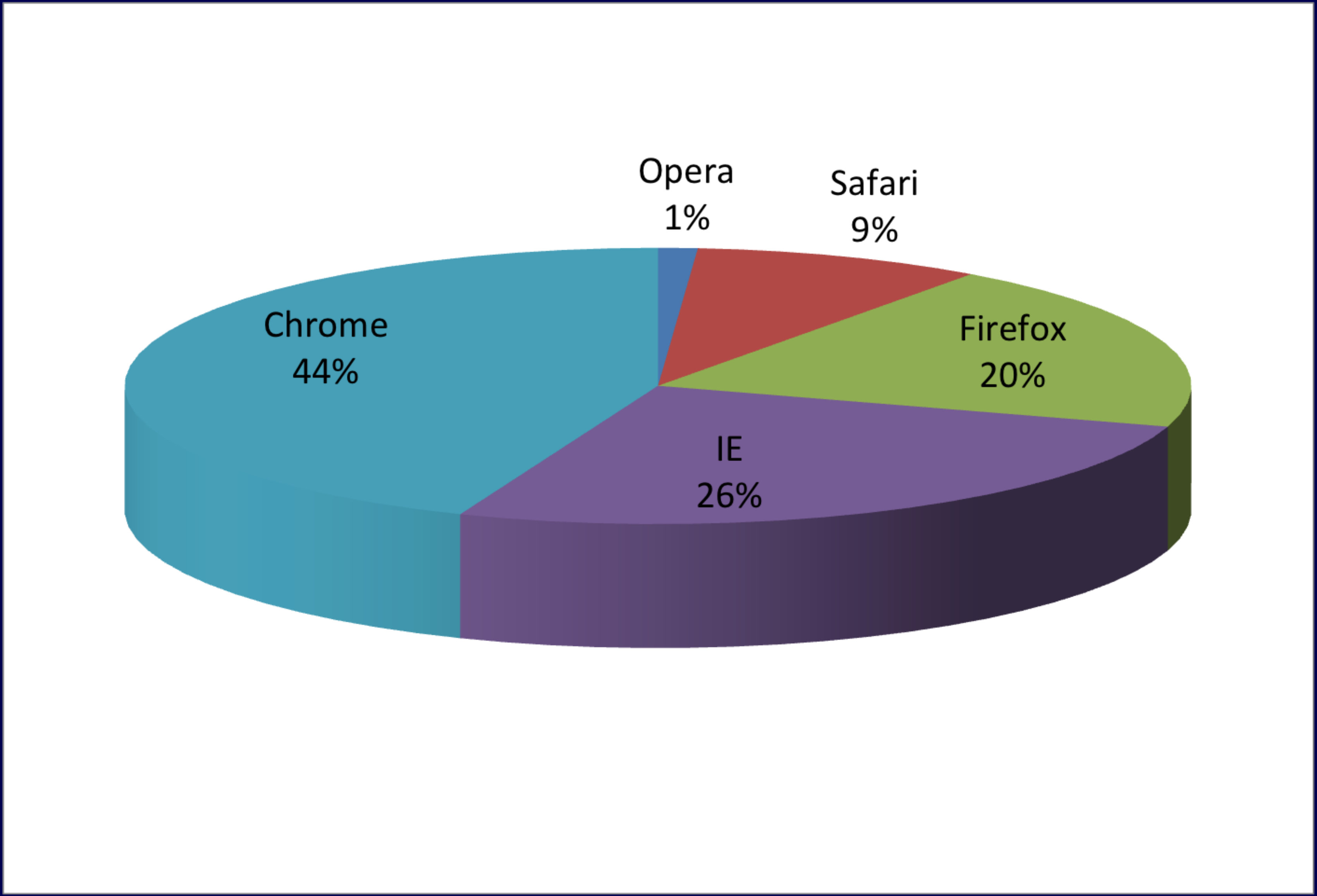
Example 2


Browser Usage (August 2013)
45
40
35
30
25
20
15
10
5
0
Opera
Safari
Firefox
IE
Chrome

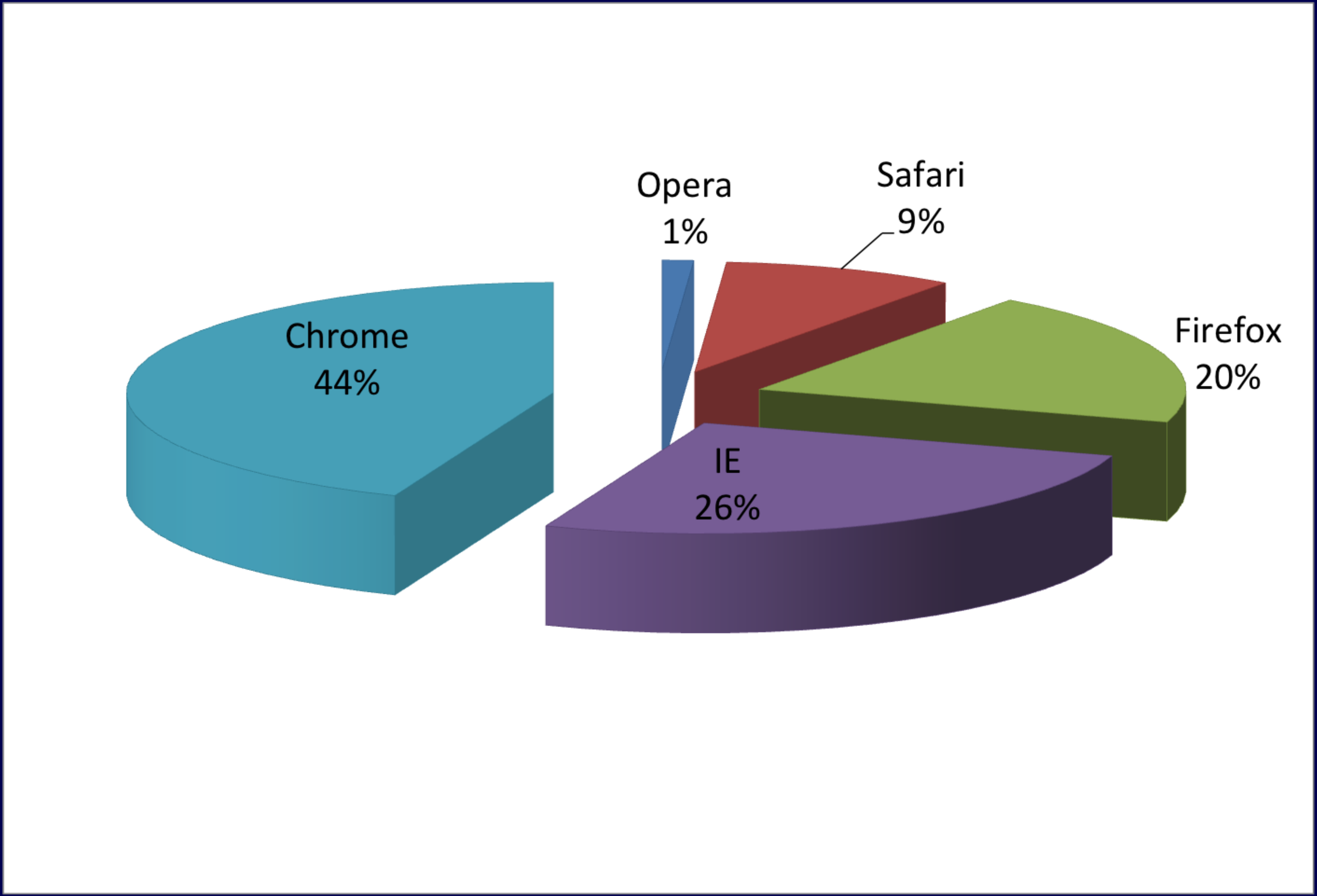
Example 2



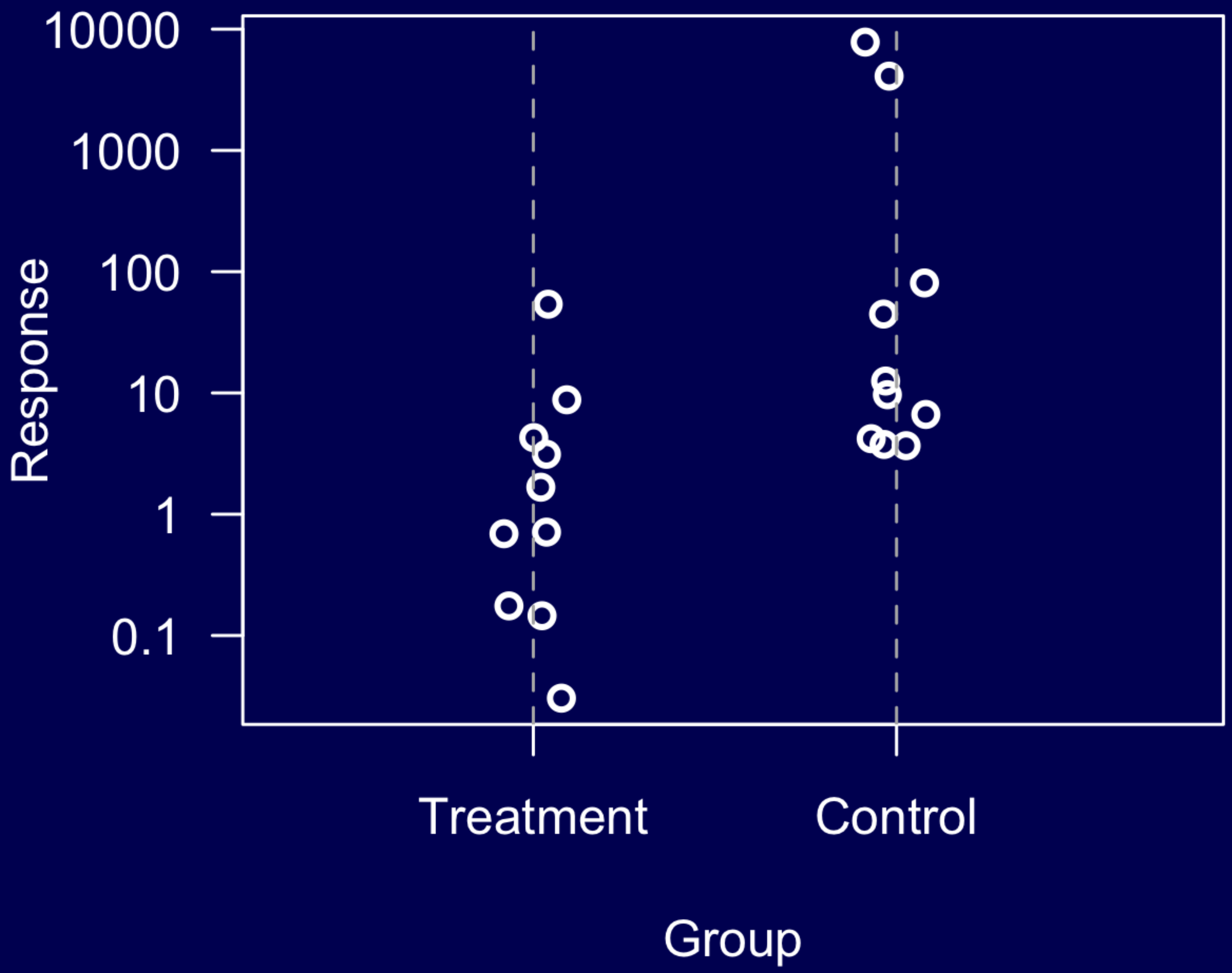
Example 2



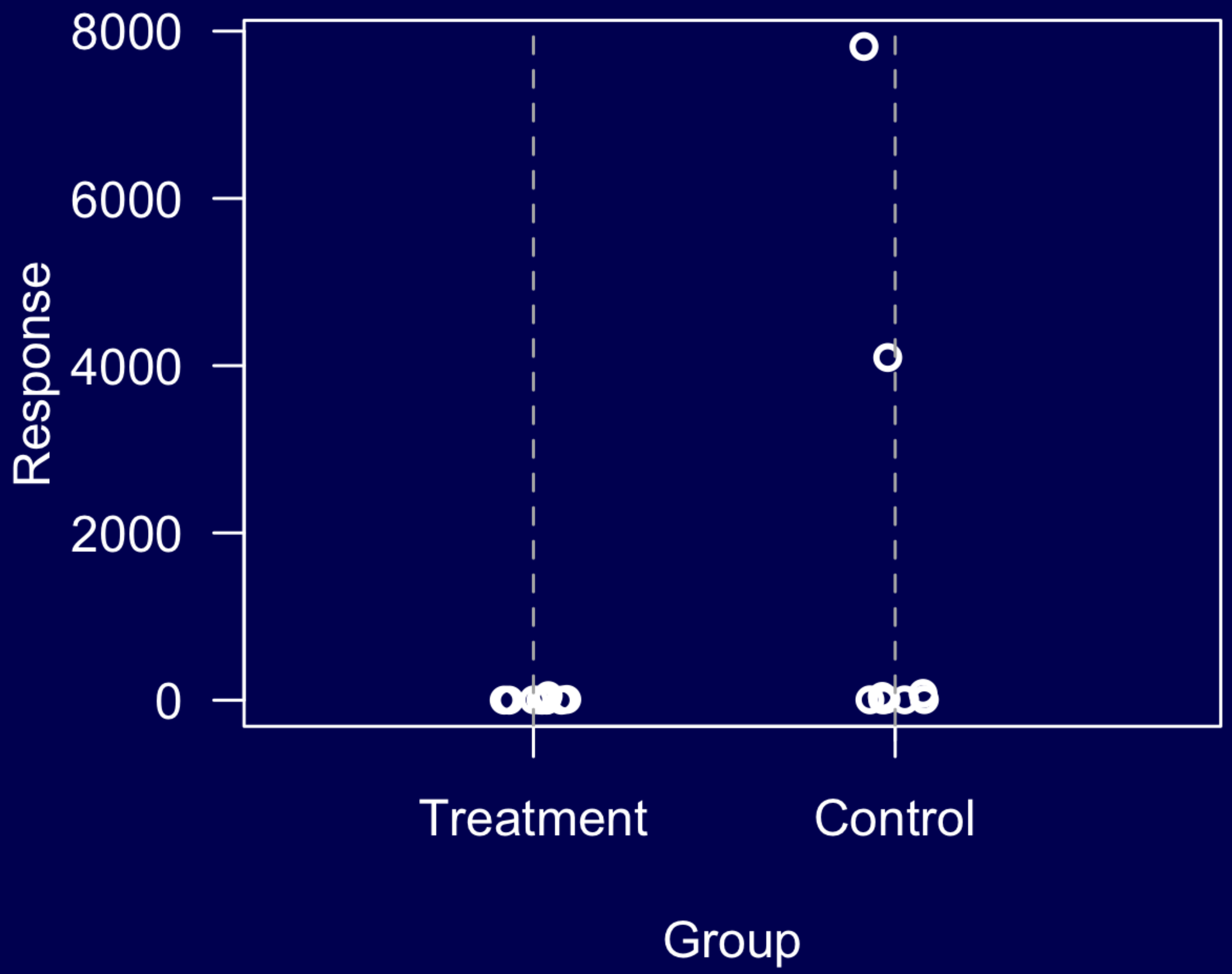
Example 2



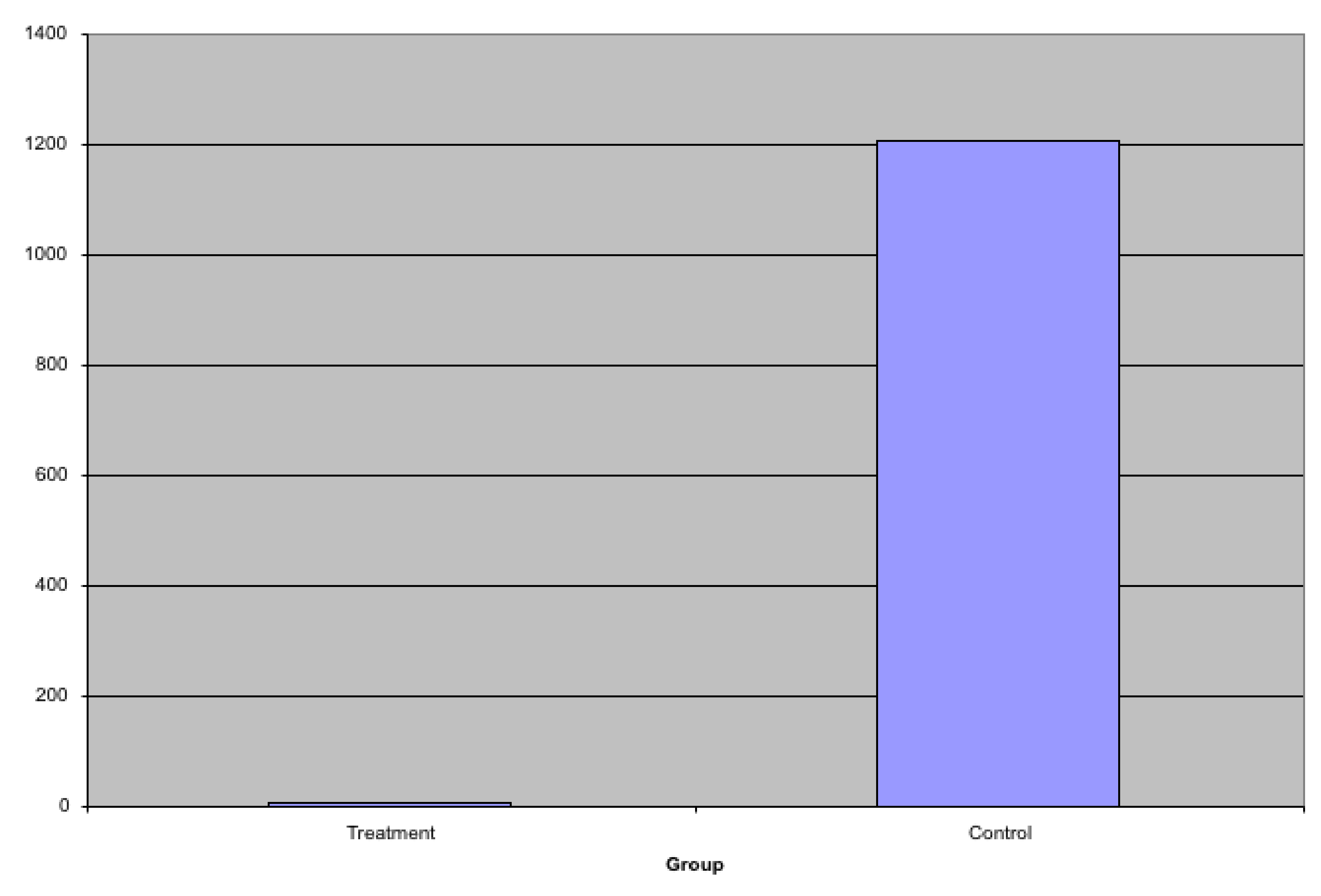
Example 3



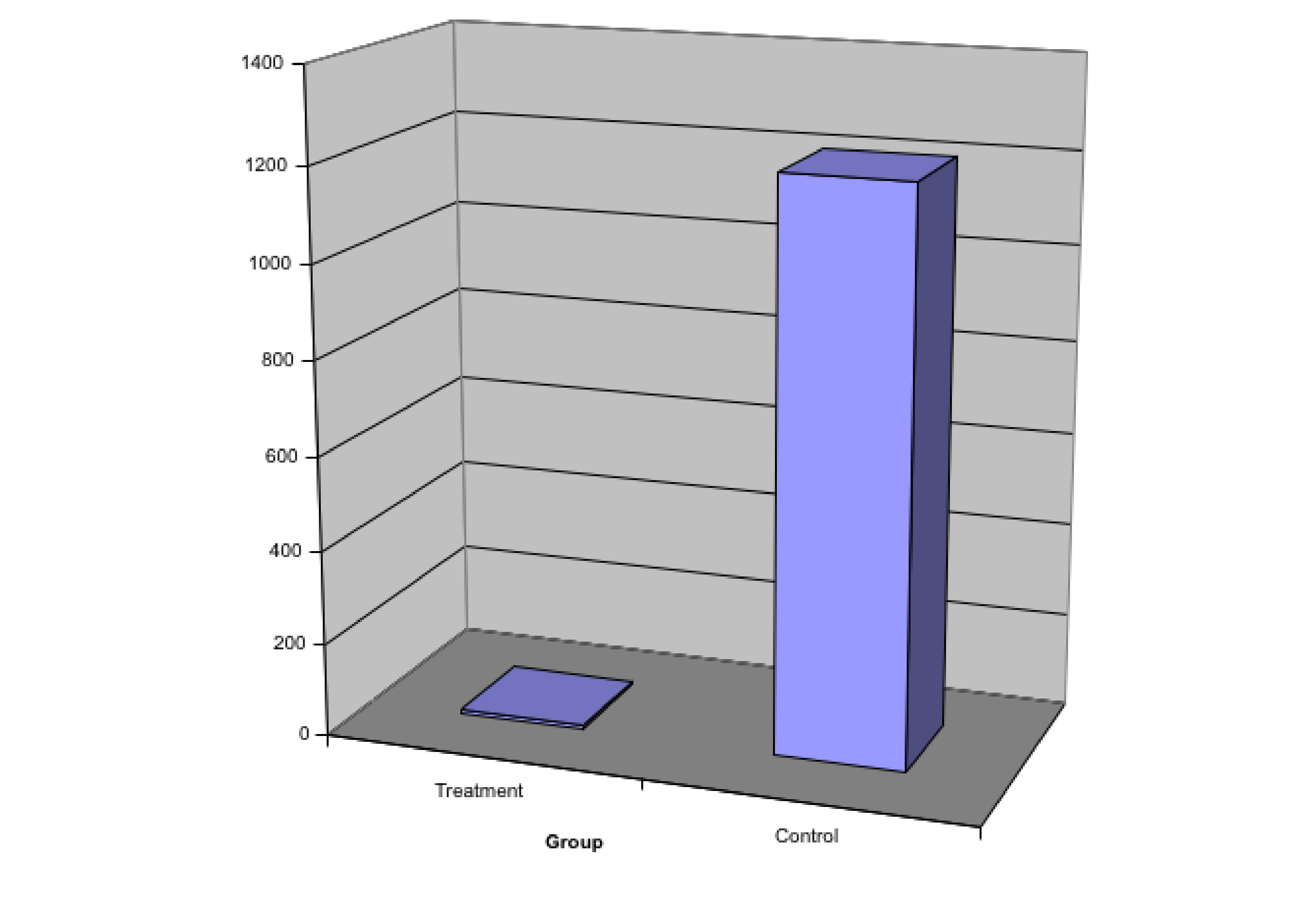
Example 3



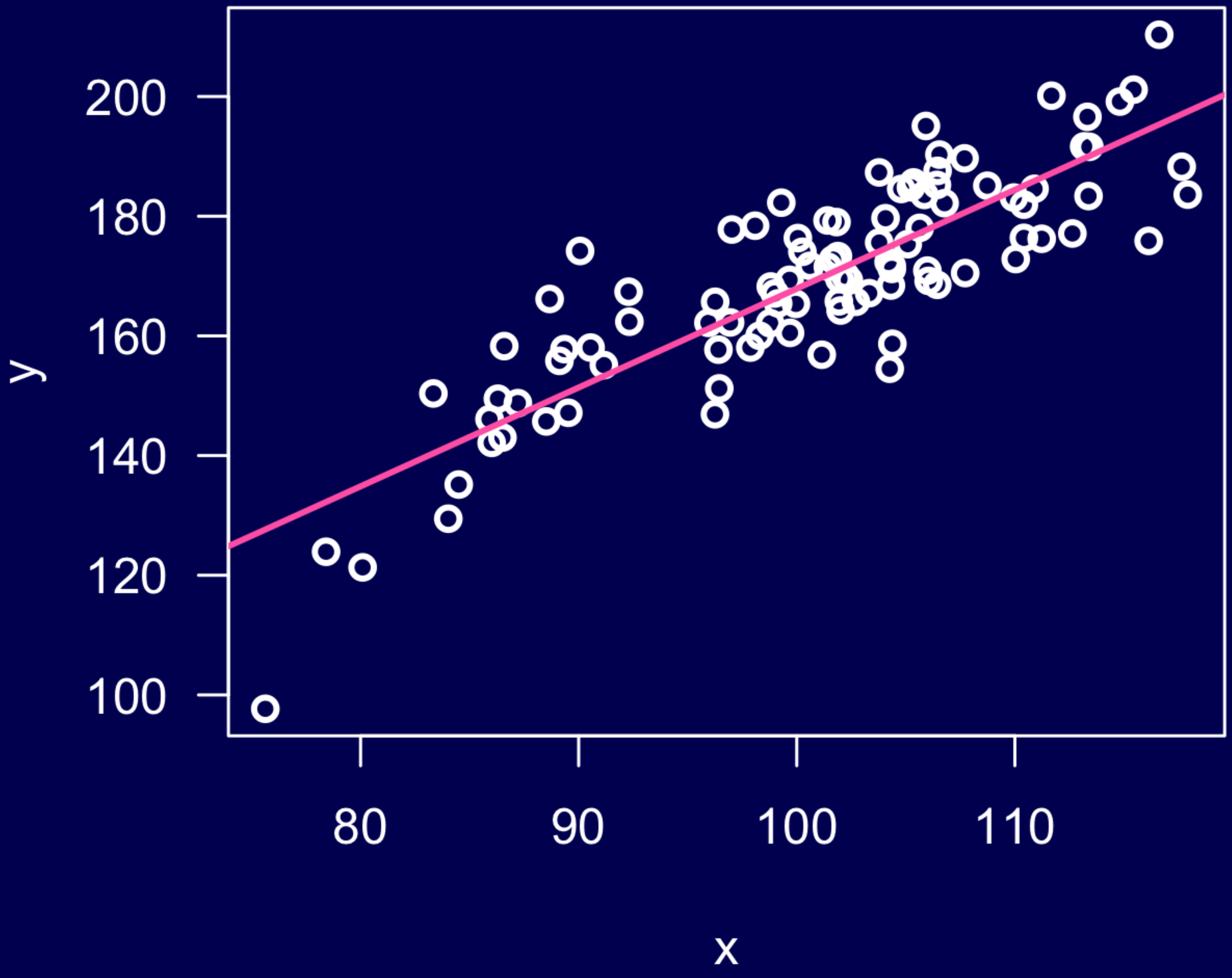
Example 3



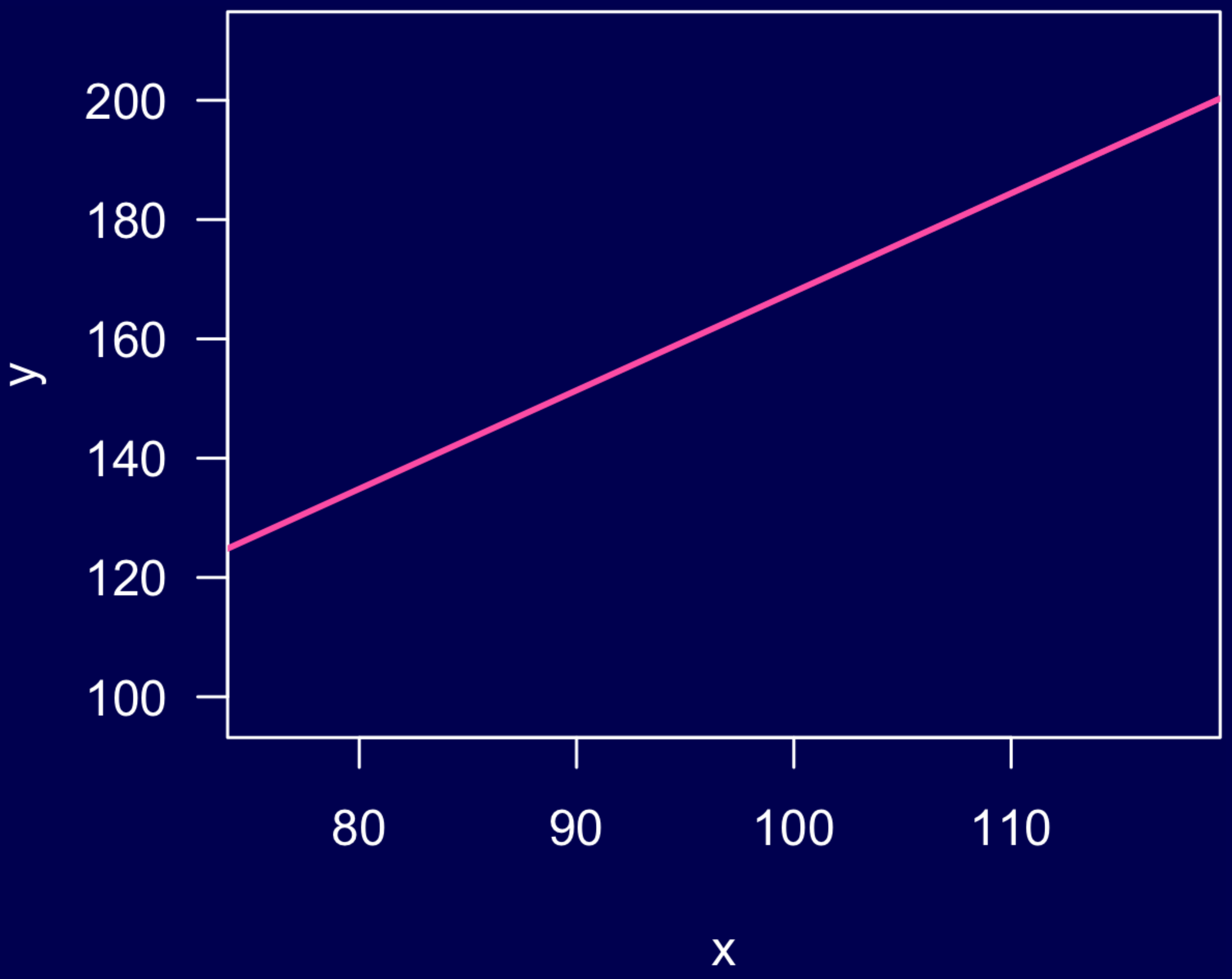
Example 3



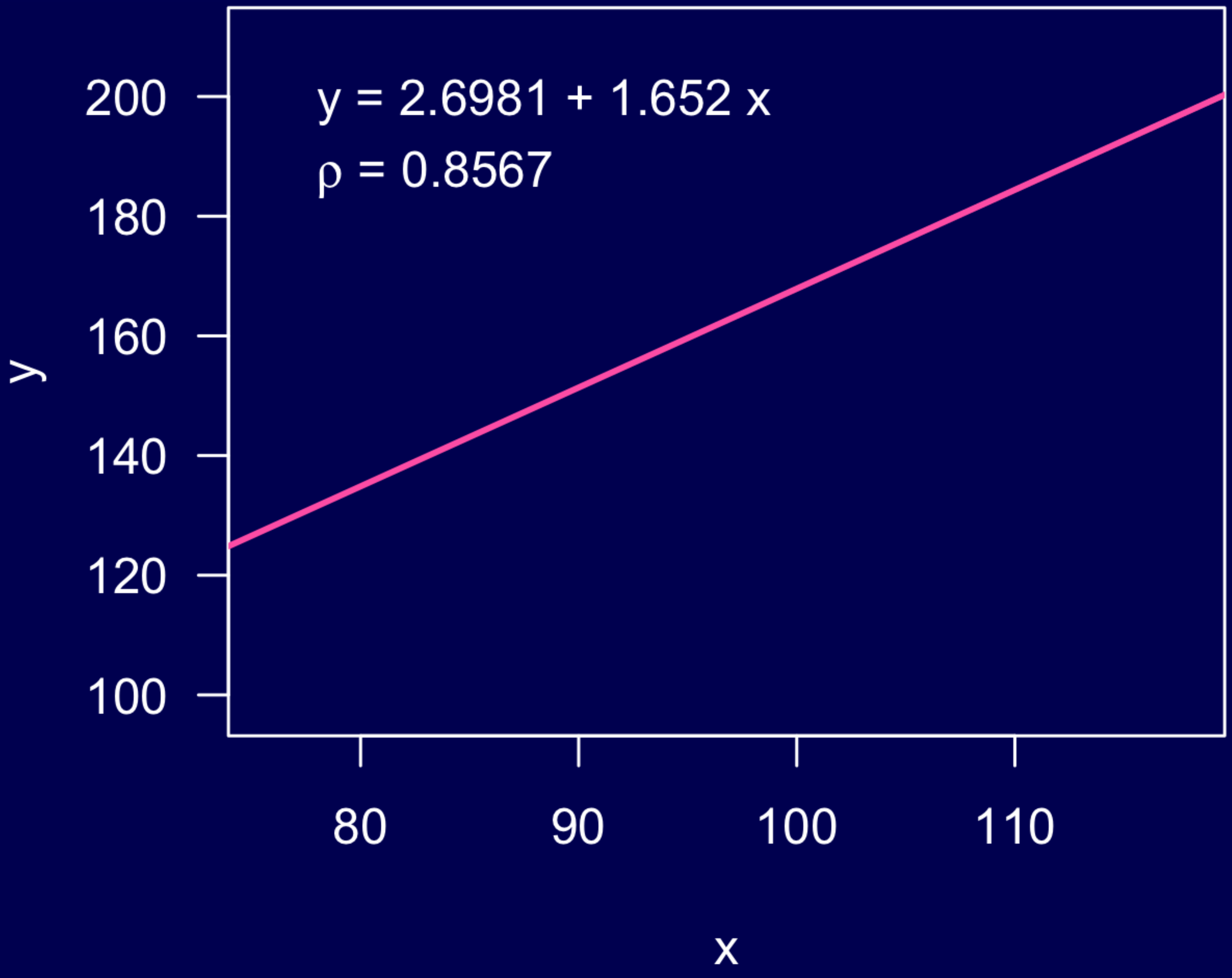
Example 4



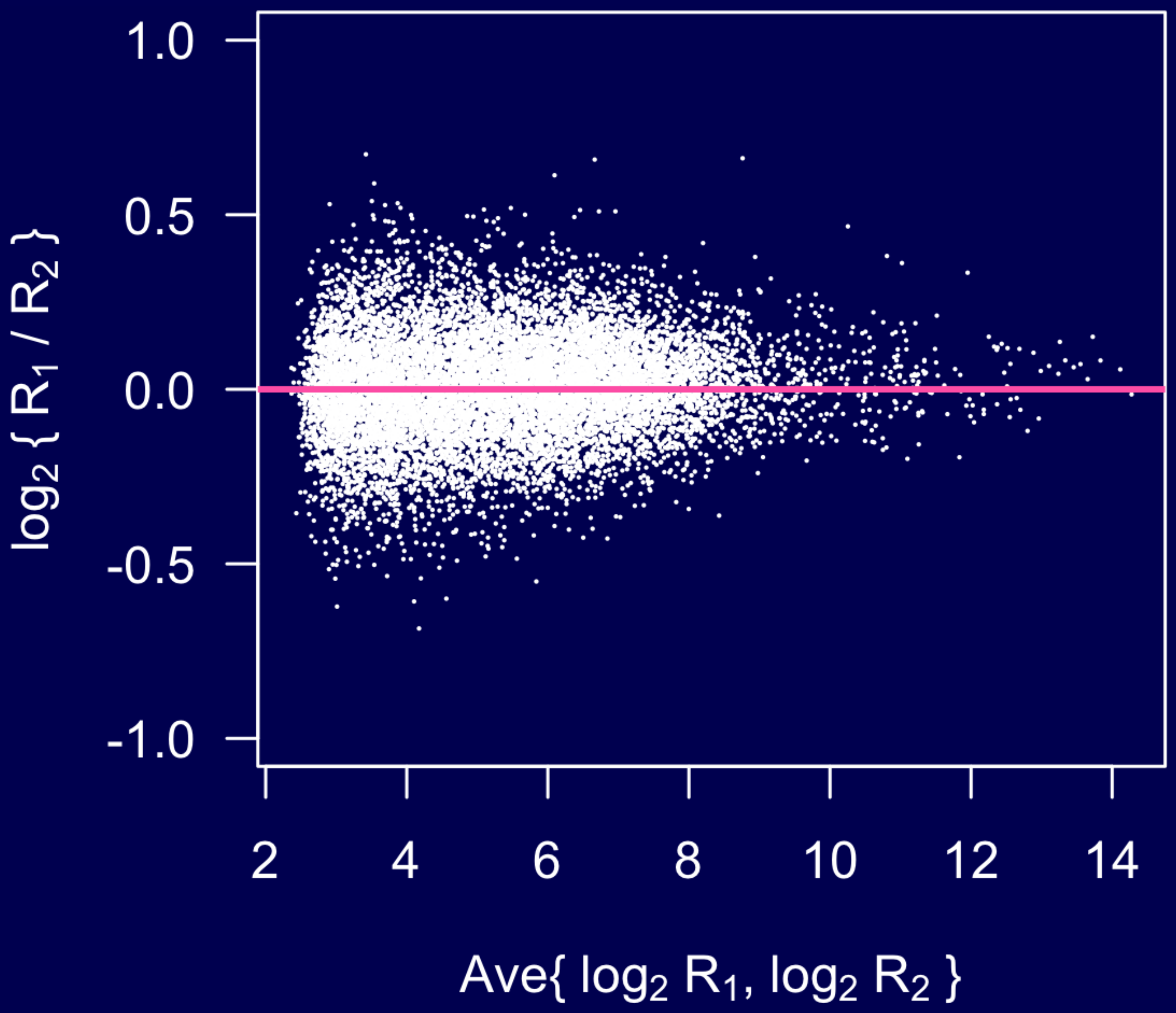
Example 4



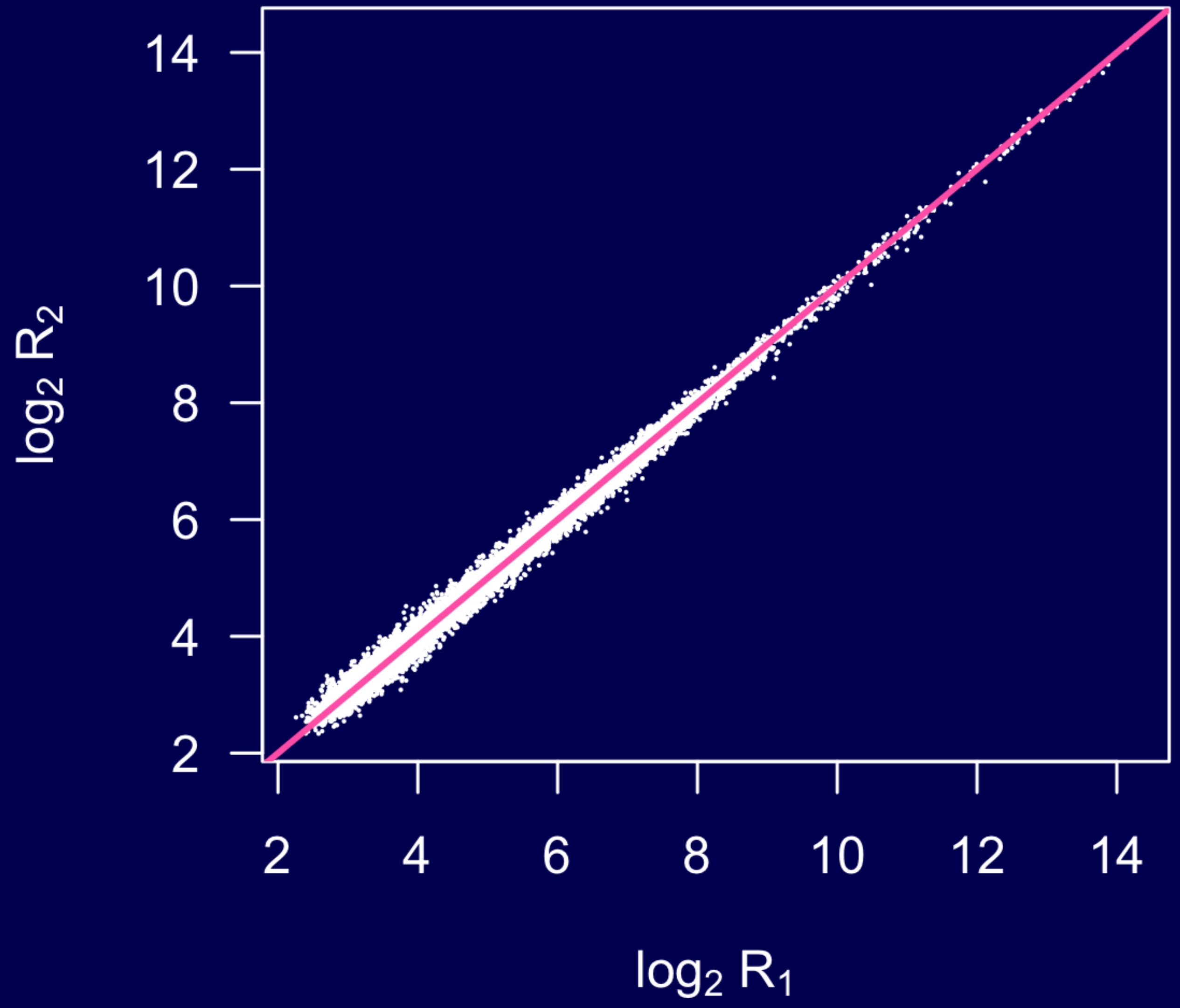
Example 4



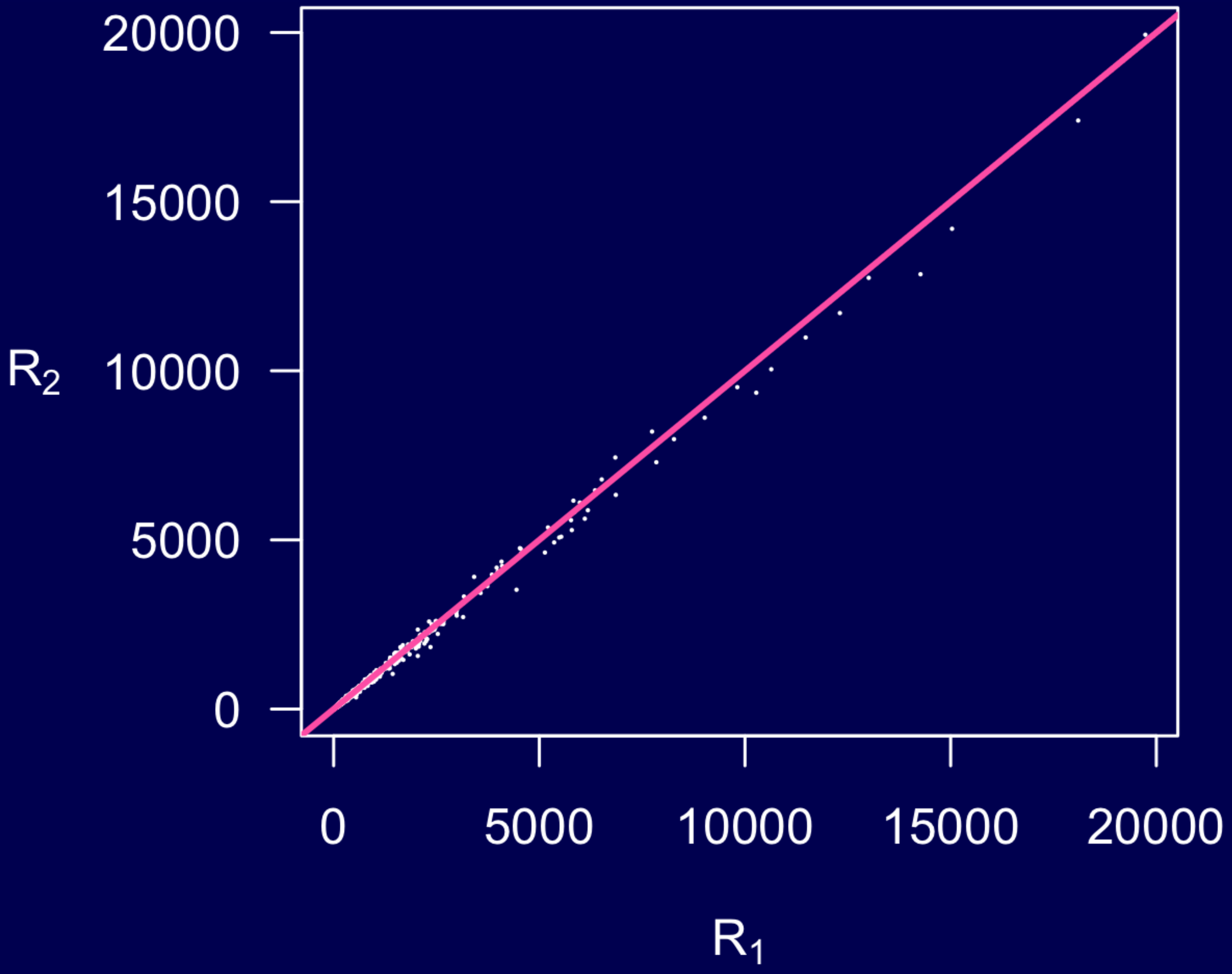
Example 5



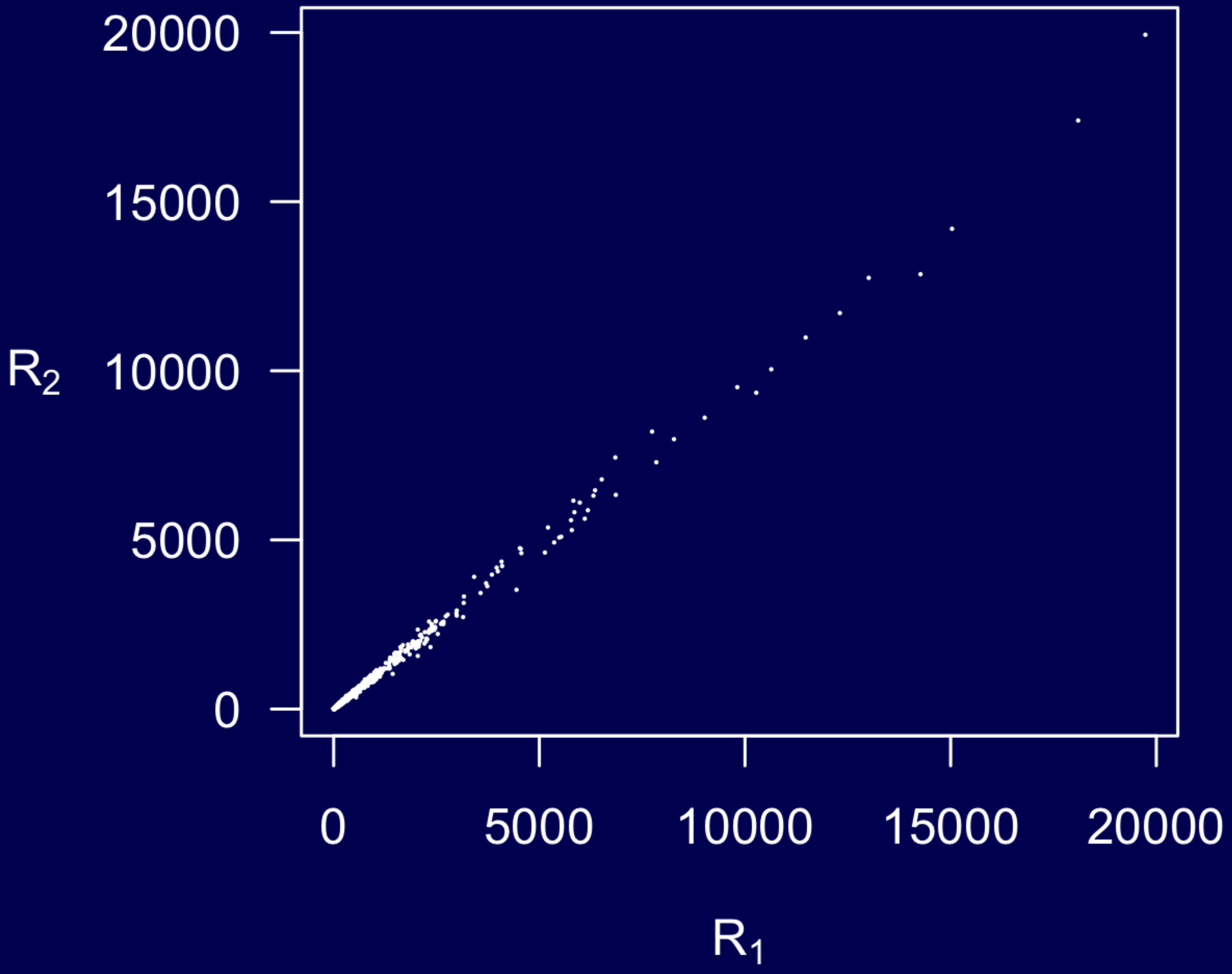
Example 5



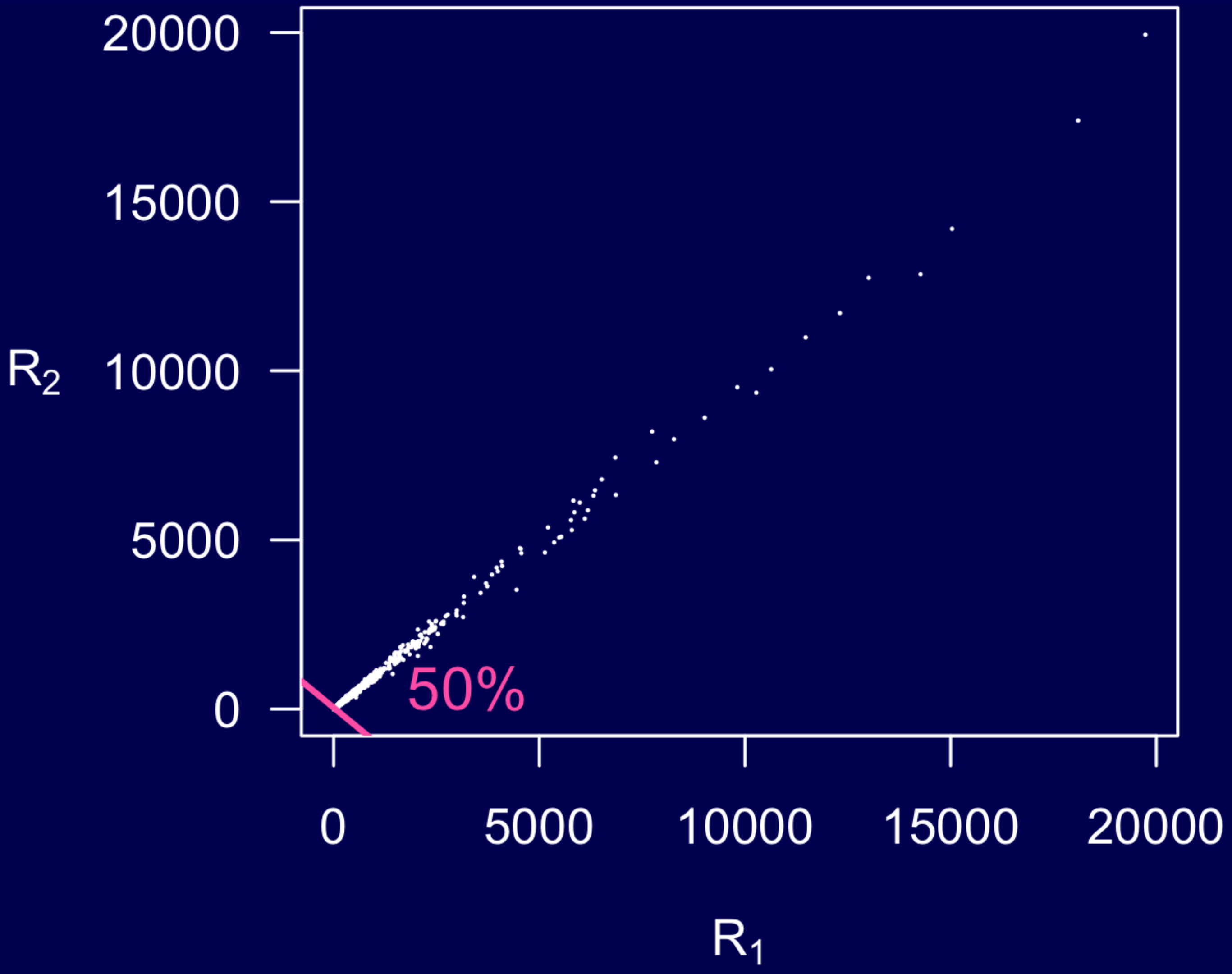
Example 5



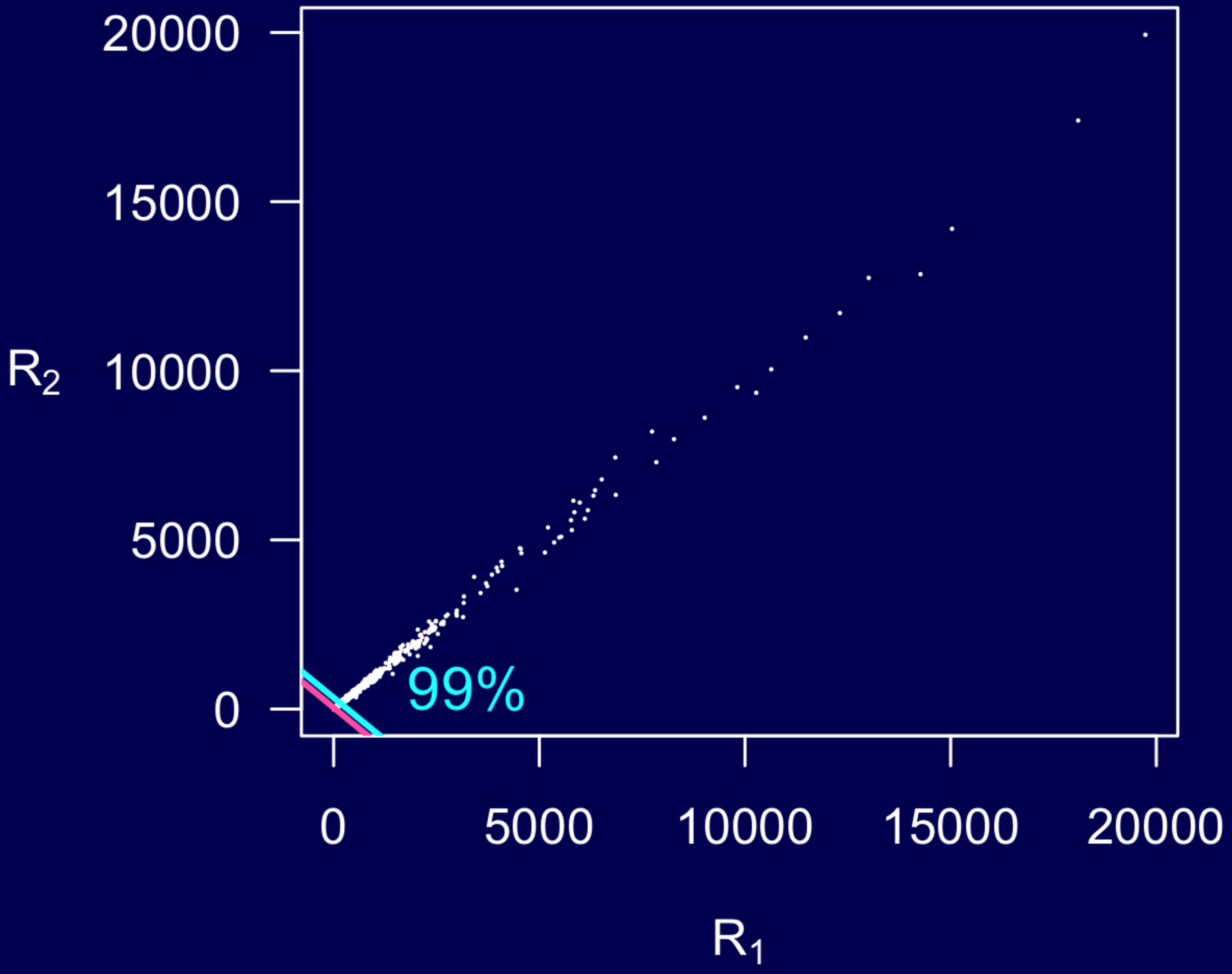
Example 5



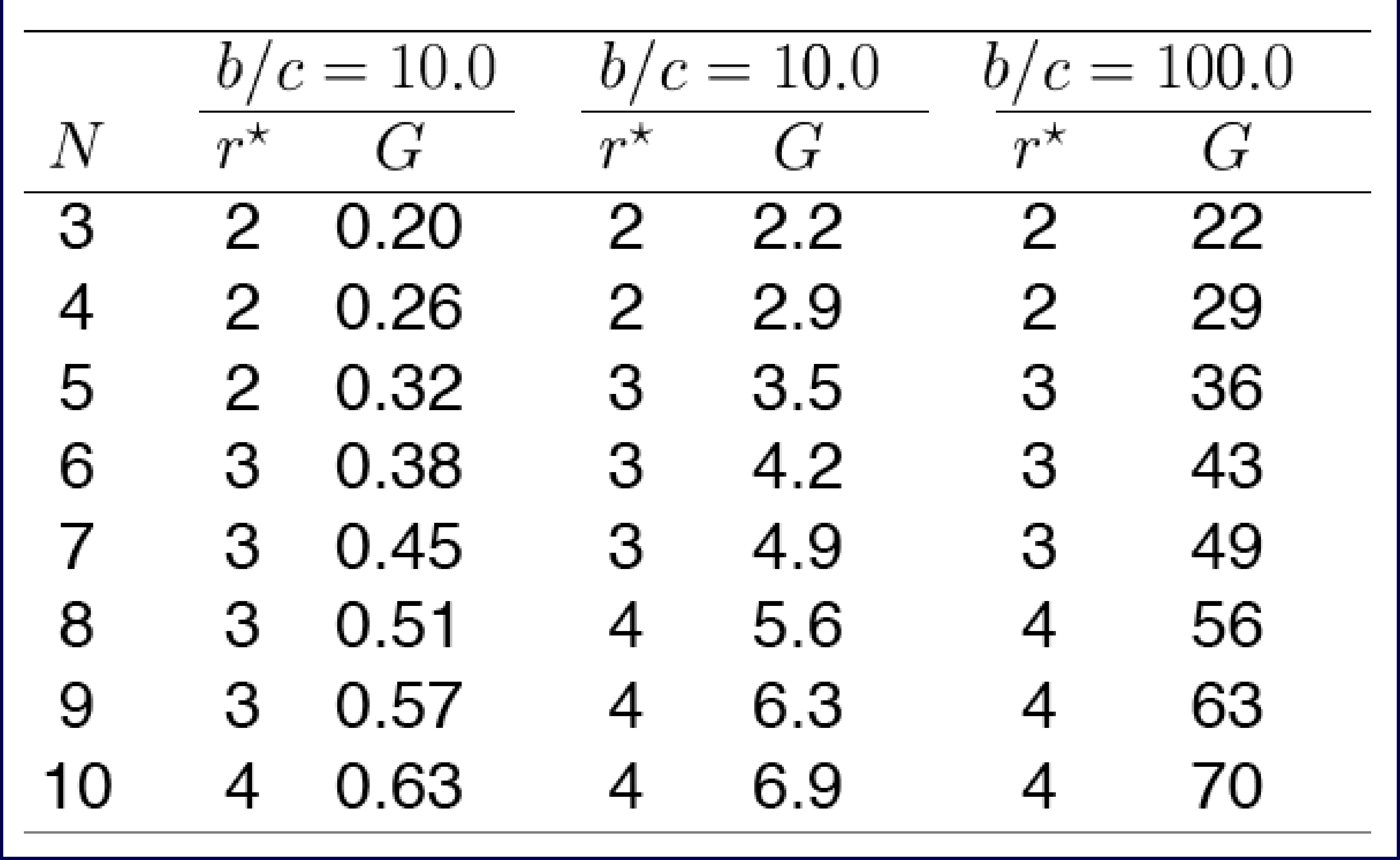
Example 5



Example 5

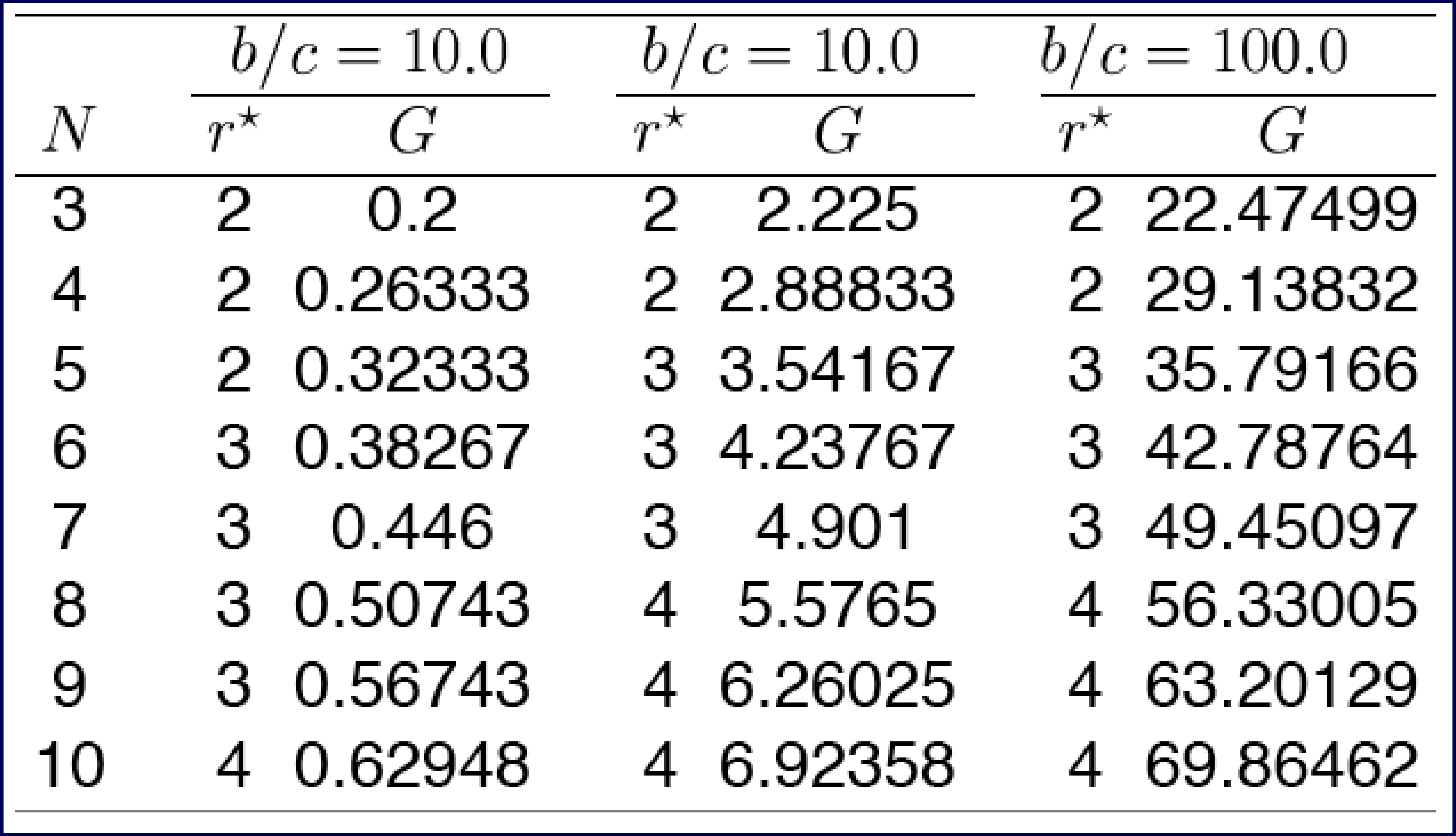


Example 6



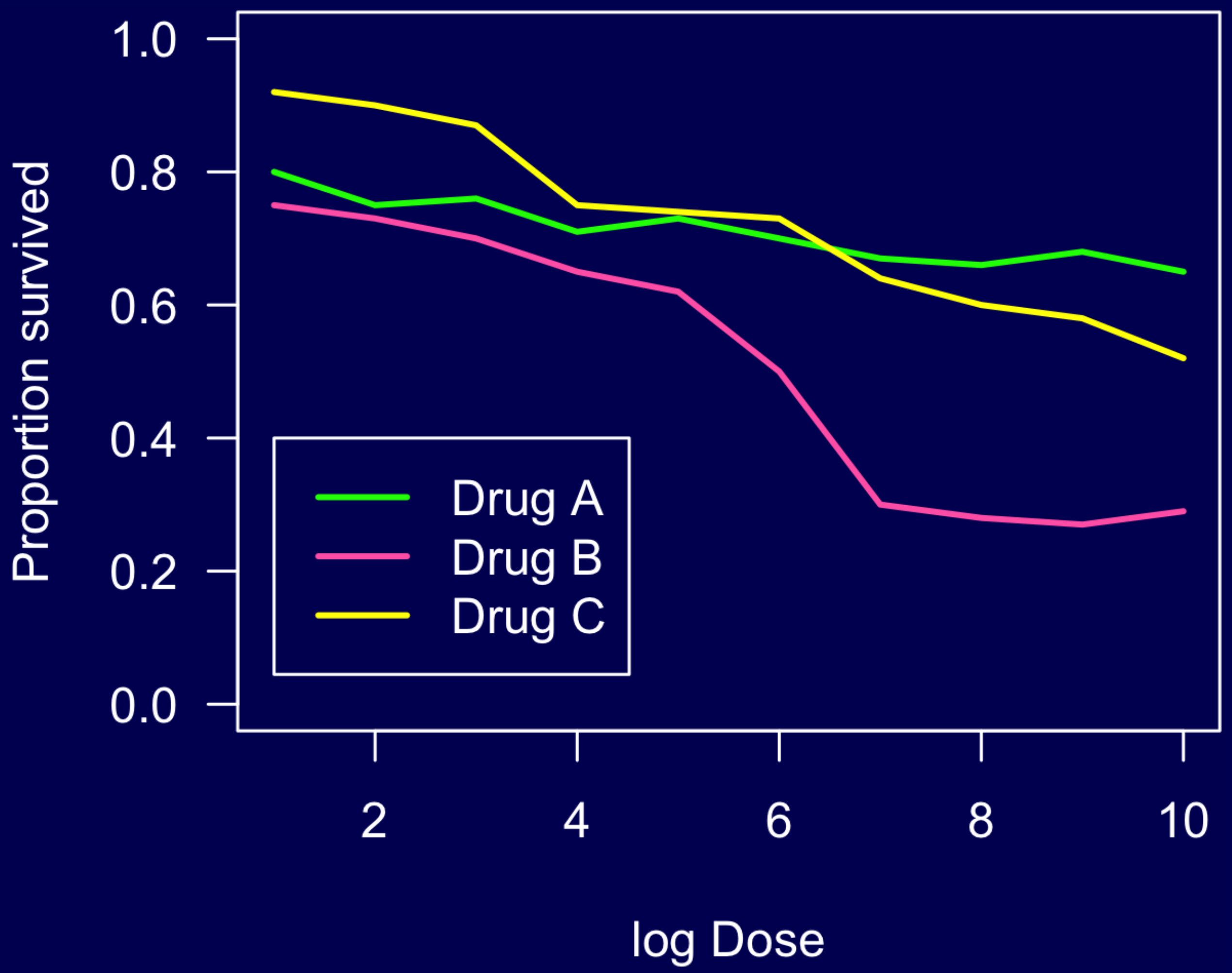
	$b/c = 10.0$		$b/c = 10.0$		$b/c = 100.0$	
N	$r^\star$	G	$r^\star$	G	$r^\star$	G
3	2	0.20	2	2.2	2	22
4	2	0.26	2	2.9	2	29
5	2	0.32	3	3.5	3	36
6	3	0.38	3	4.2	3	43
7	3	0.45	3	4.9	3	49
8	3	0.51	4	5.6	4	56
9	3	0.57	4	6.3	4	63
10	4	0.63	4	6.9	4	70

Example 6

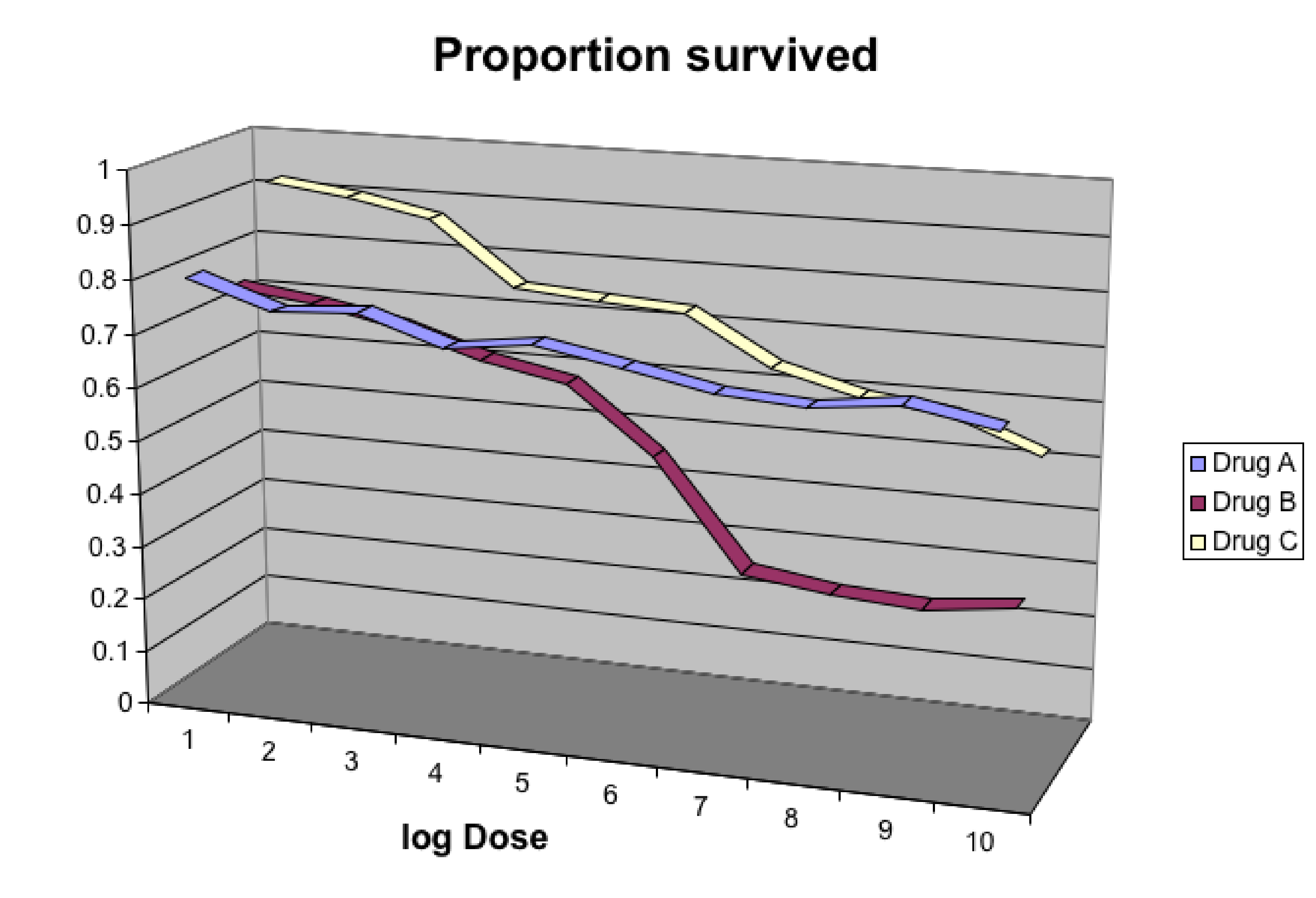


	$b/c = 10.0$		$b/c = 10.0$		$b/c = 100.0$	
N	$r^\star$	G	$r^\star$	G	$r^\star$	G
3	2	0.2	2	2.225	2	22.47499
4	2	0.26333	2	2.88833	2	29.13832
5	2	0.32333	3	3.54167	3	35.79166
6	3	0.38267	3	4.23767	3	42.78764
7	3	0.446	3	4.901	3	49.45097
8	3	0.50743	4	5.5765	4	56.33005
9	3	0.56743	4	6.26025	4	63.20129
10	4	0.62948	4	6.92358	4	69.86462

Example 7


1.0
0.8
0.6
0.4
0.2
0.0
Proportion survived
2
4
6
8
10
log Dose
Drug A
Drug B
Drug C

Example 7


Proportion survived
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
log Dose
Drug A
Drug B
Drug C

Displaying data well

- Be accurate and clear.
- Let the data speak.
  - Show as much information as possible, taking care not to obscure the message.
- Science not sales.
  - Avoid unnecessary frills (esp. gratuitous 3d).
- In tables, every digit should be meaningful. Don't drop ending 0's.

Further reading

- ER Tufte (1983) The visual display of quantitative information. Graphics Press.
- ER Tufte (1990) Envisioning information. Graphics Press.
- ER Tufte (1997) Visual explanations. Graphics Press.

- WS Cleveland (1993) Visualizing data. Hobart Press.
- WS Cleveland (1994) The elements of graphing data. CRC Press.

- A Gelman, C Pasarica, R Dodhia (2002) Let's practice what we preach: Turning tables into graphs. The American Statistician 56:121-130

- NB Robbins (2004) Creating more effective graphs. Wiley

- Nature Methods columns: `http://bang.clearscience.info/?p=546`

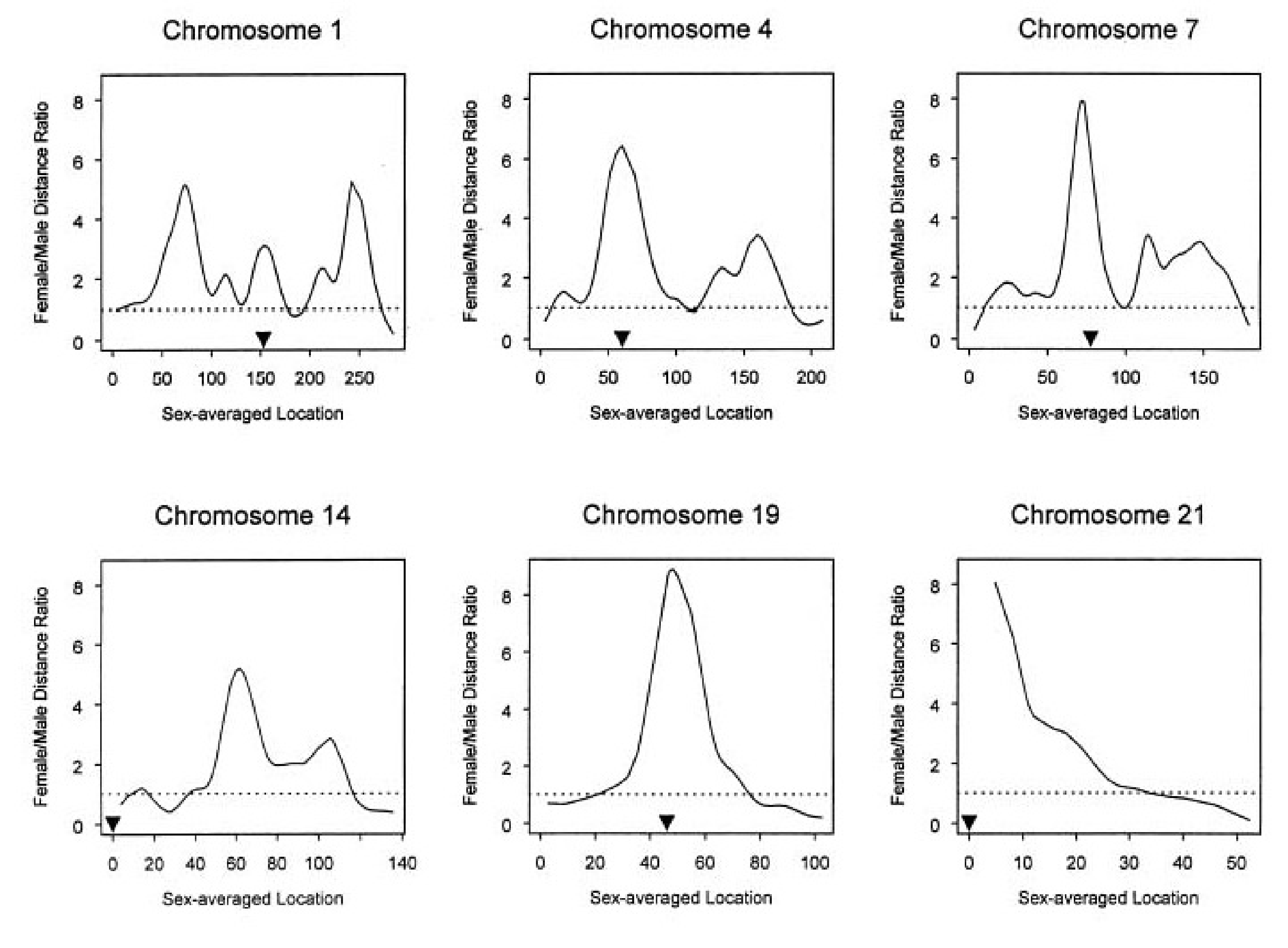
The top ten worst graphs

With apologizes to the authors, we provide the following list of the top ten worst graphs in the scientific literature.

As these examples indicate, good scientists can make mistakes.

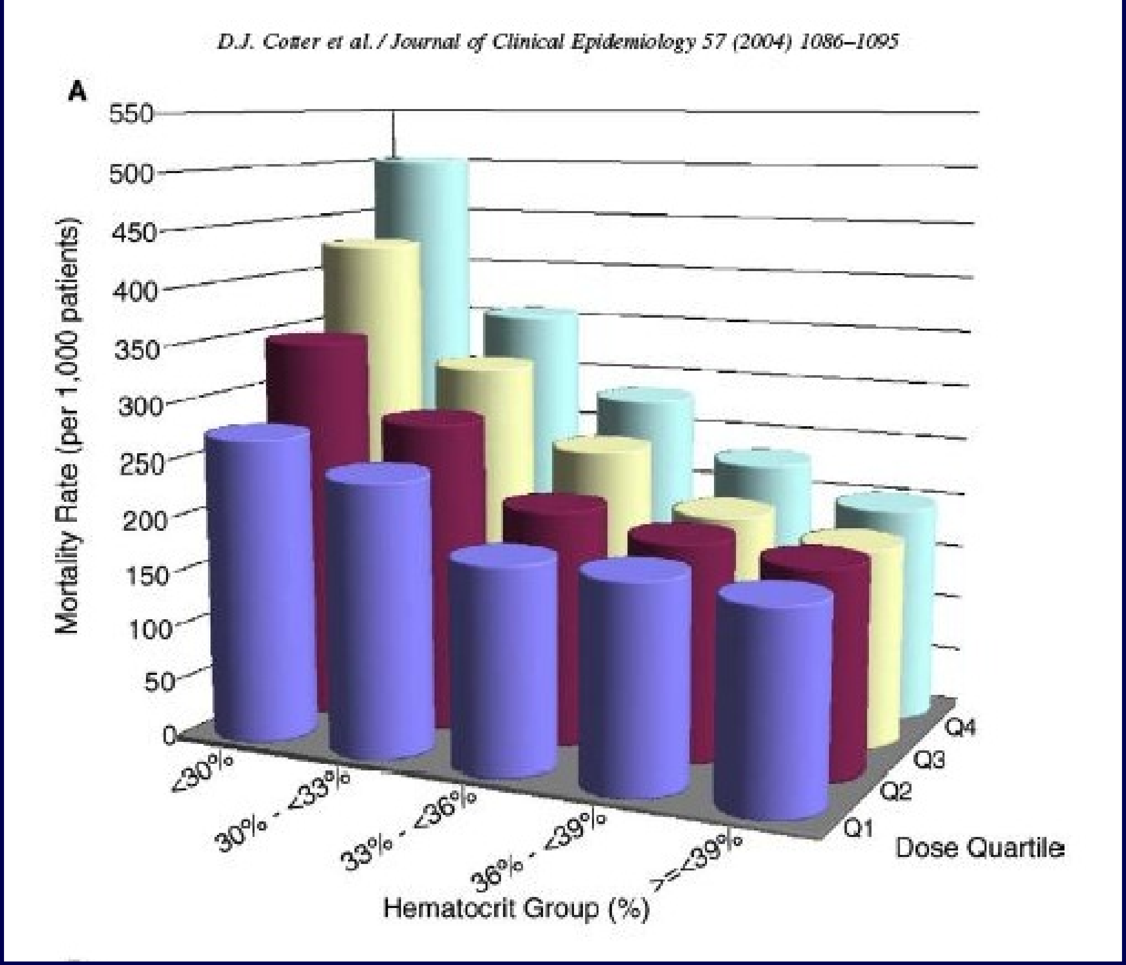
http://www.biostat.wisc.edu/~kbroman/topten_worstgraphs

10

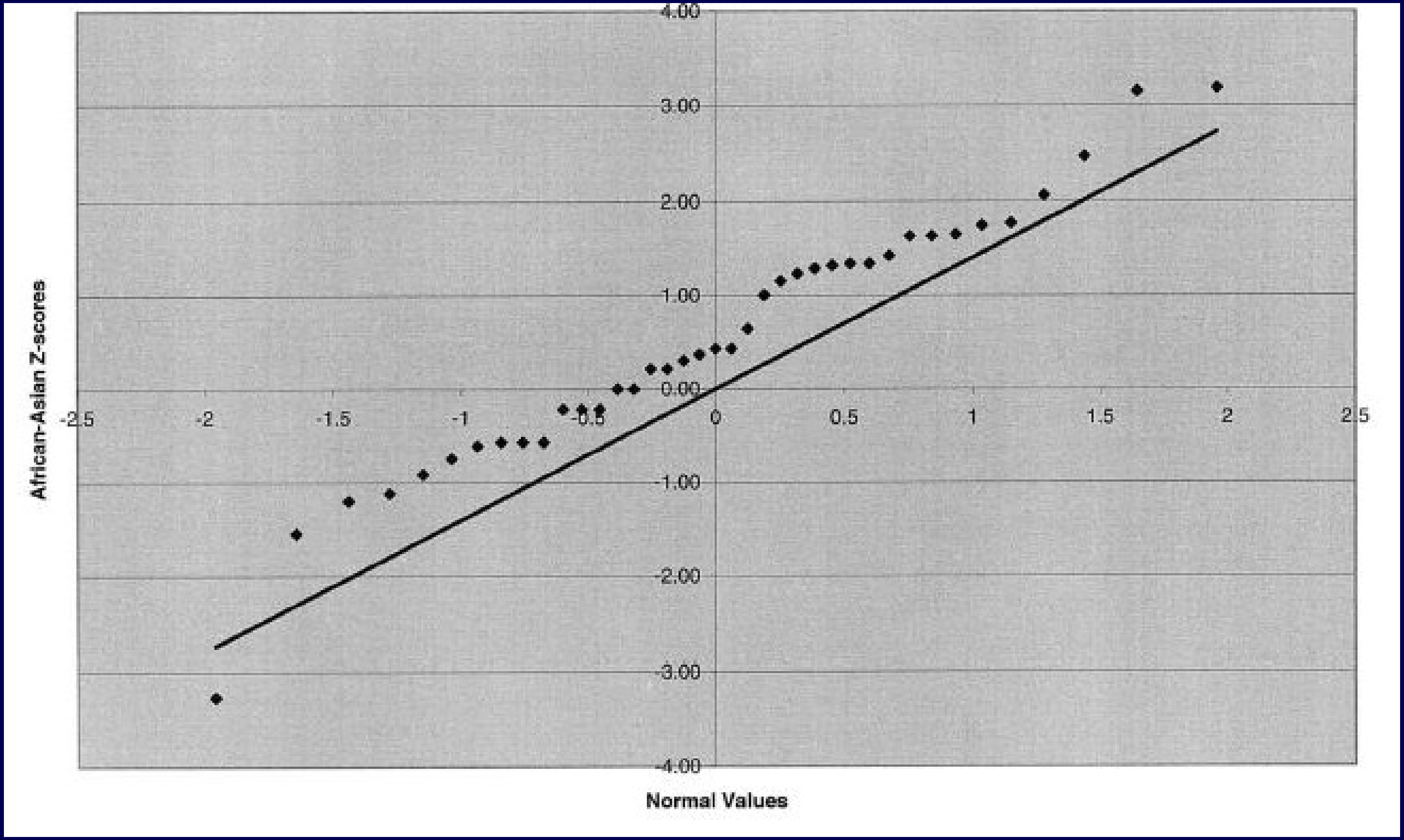


Broman et al., Am J Hum Genet 63:861-869, 1998, Fig. 1

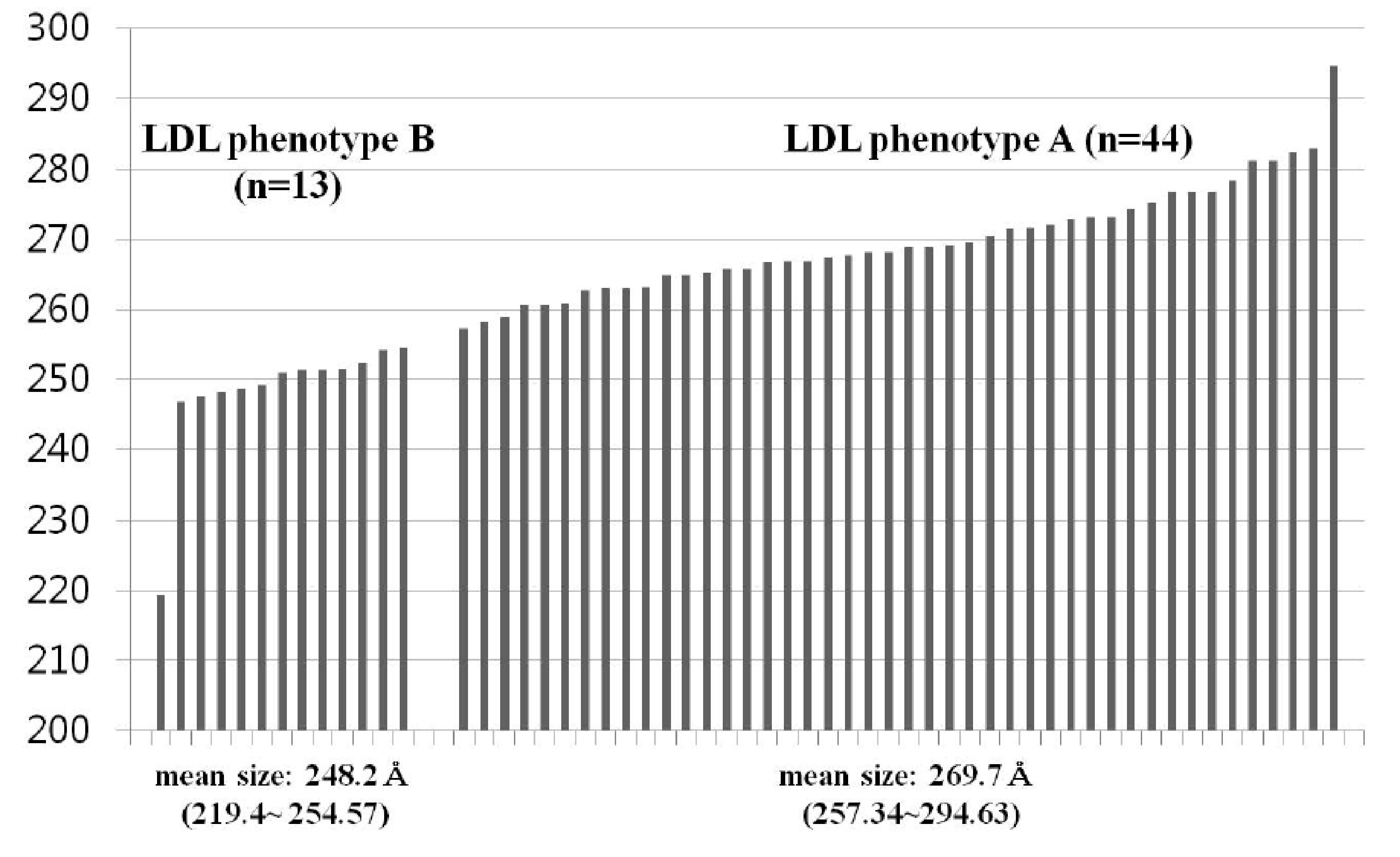
9



Cotter et al., J Clin Epidemiol 57:1086-1095, 2004, Fig 2

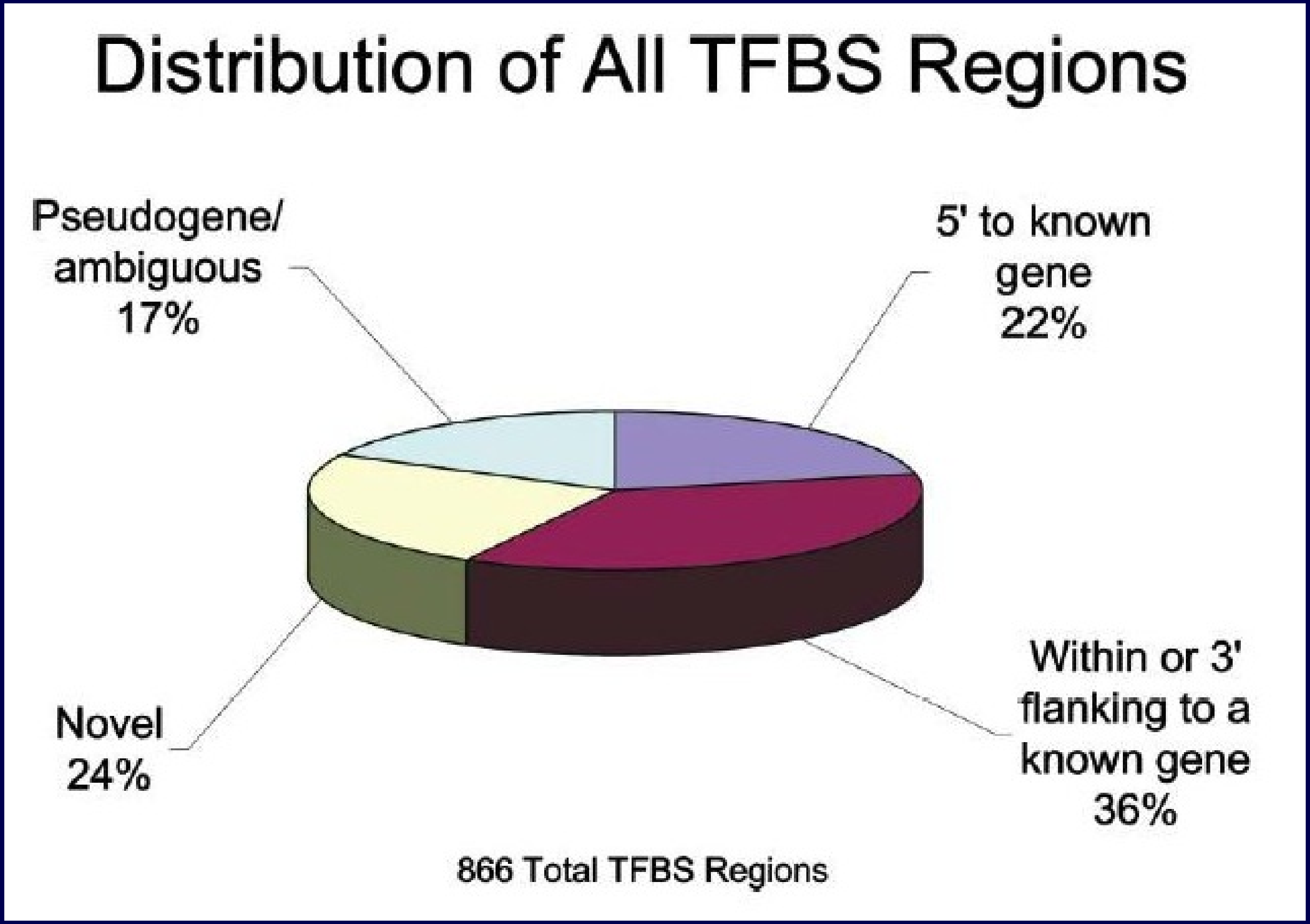


Jorgenson et al., Am J Hum Genet 76:276-290, 2005, Fig 2



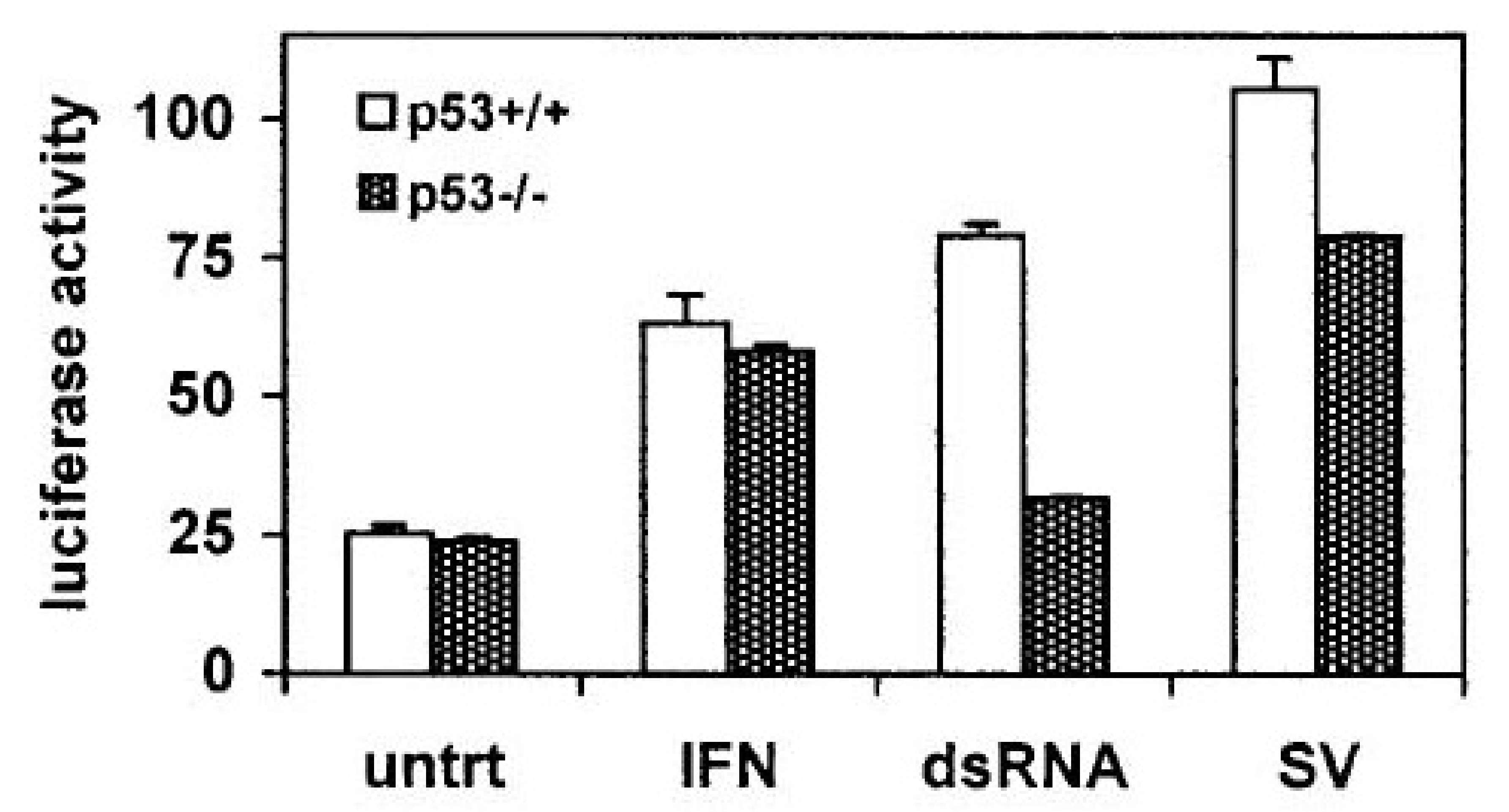
Kim et al., Nutr Res Prac 6:120-125, 2012 Fig 1

6

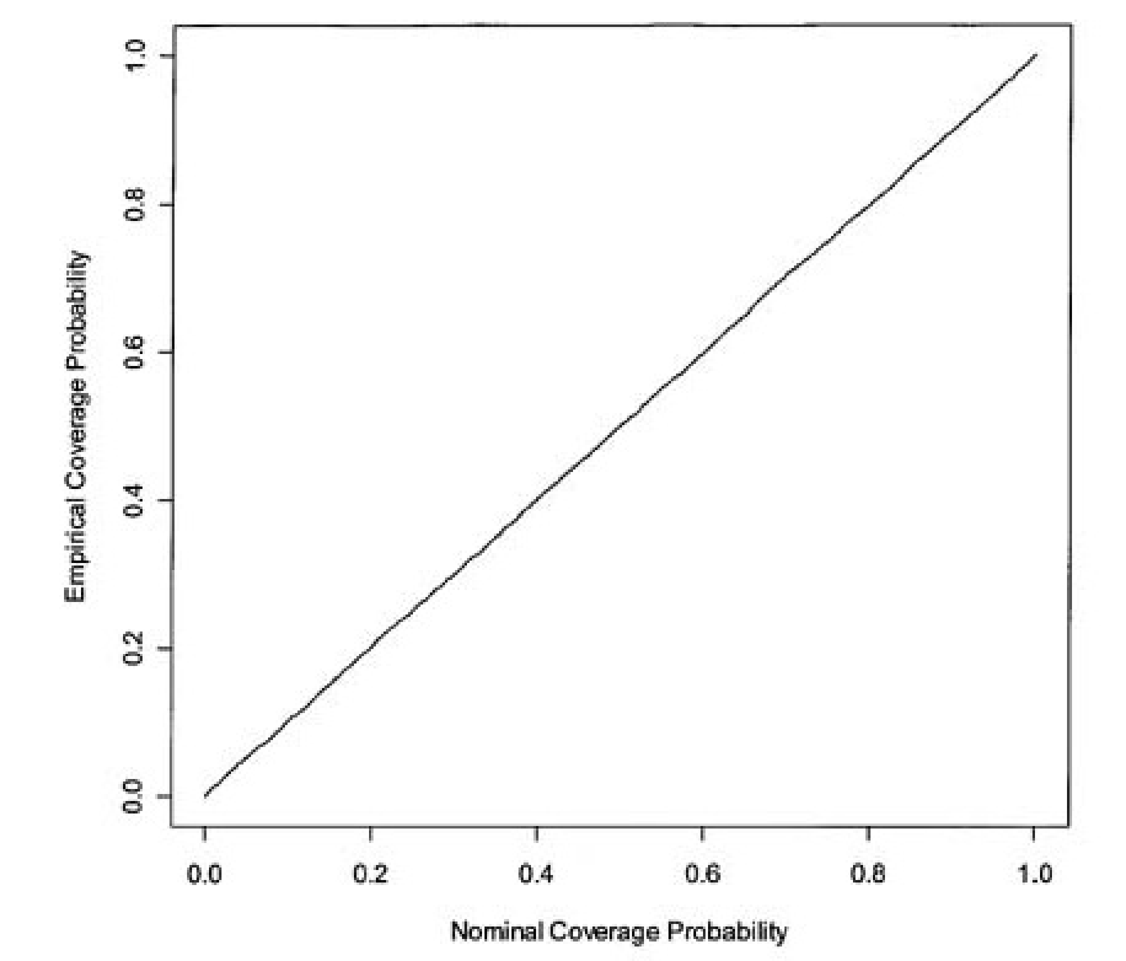


Cawley et al., Cell 116:499-509, 2004, Fig 1

5

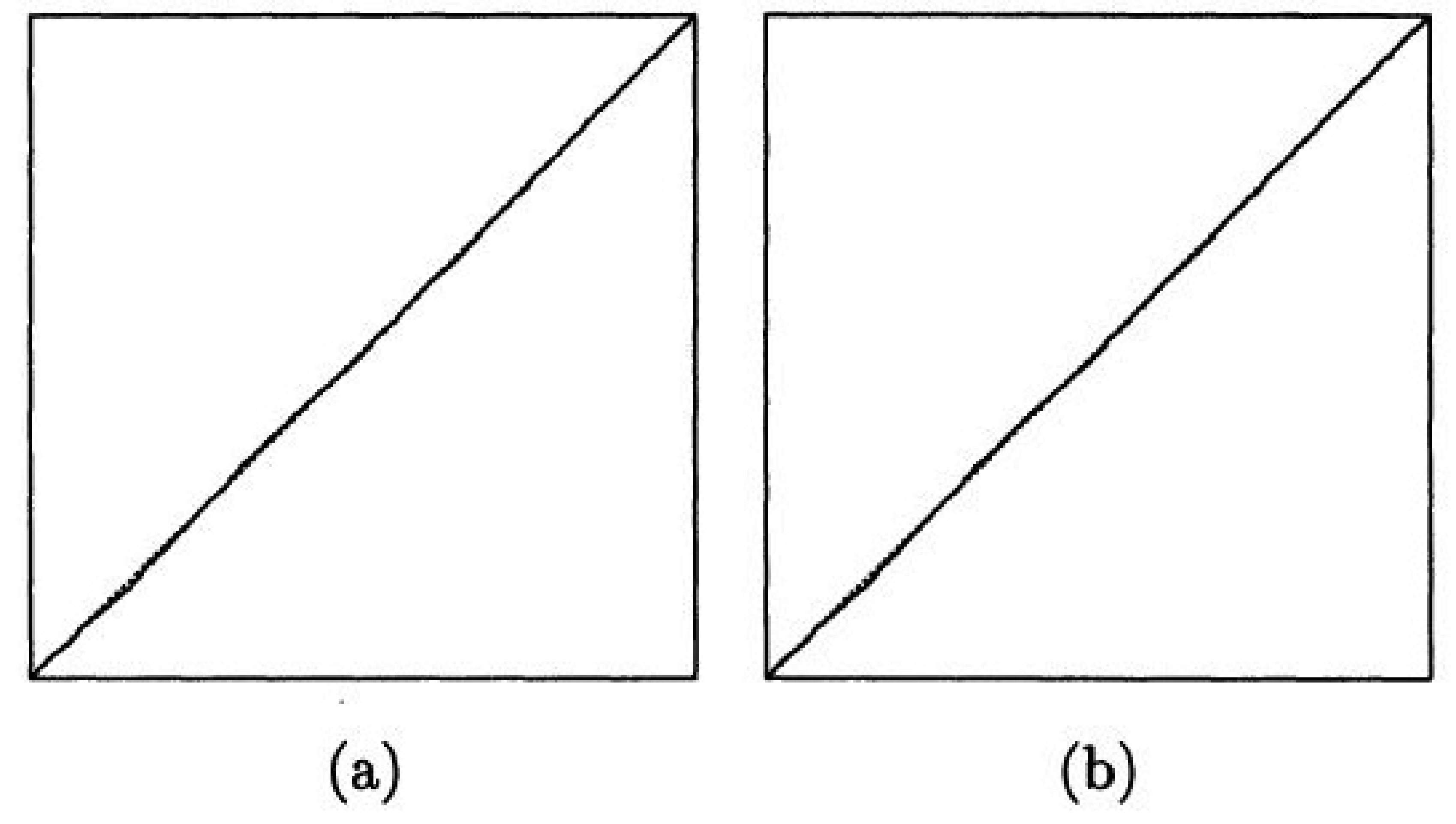


Hummer et al., J Virol 75:7774-7777, 2001, Fig 4



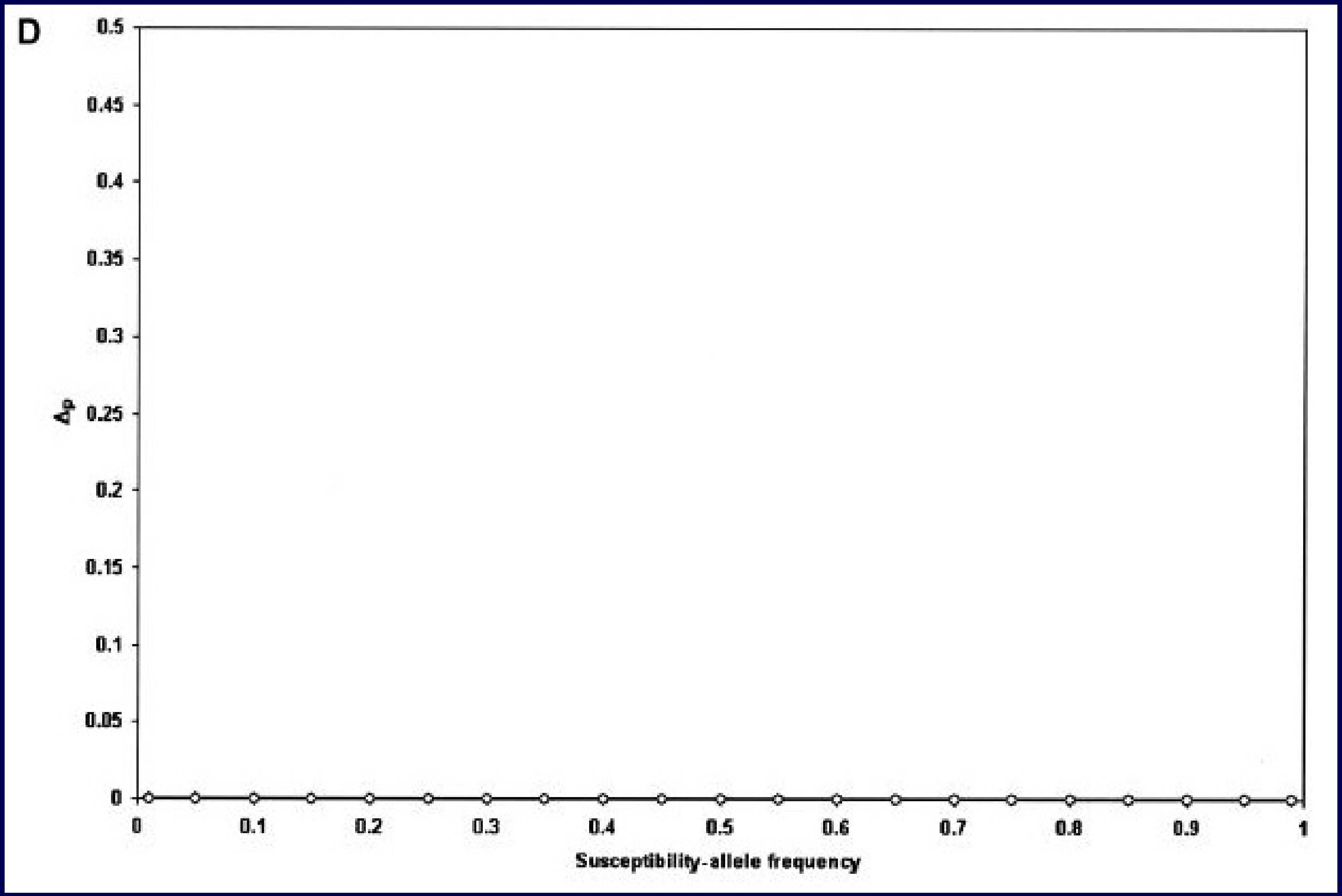
Epstein and Satten, Am J Hum Genet 73:1316-1329, 2003, Fig 1

3



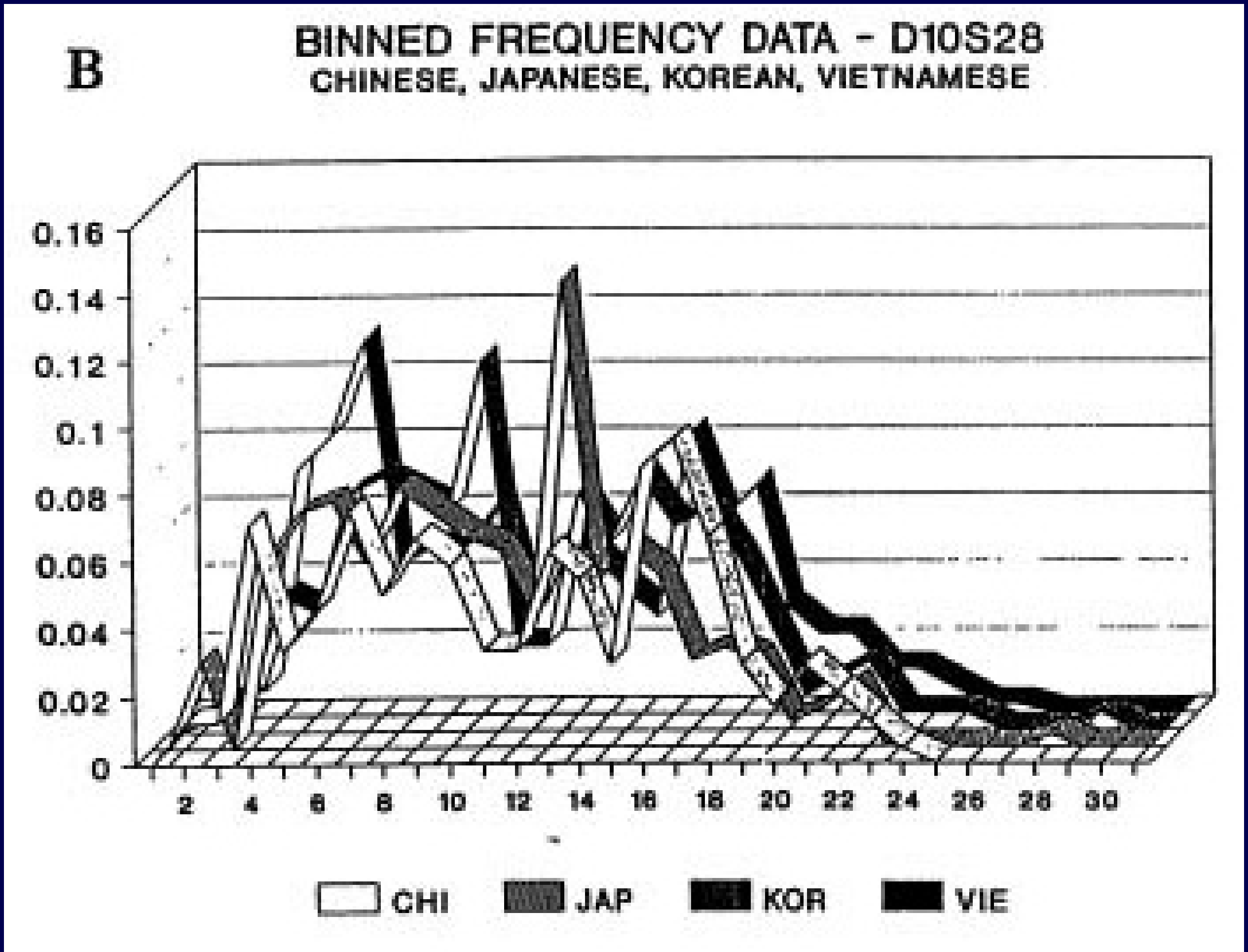
Mykland et al., J Am Stat Asso 90:233-241, 1995, Fig 1

2



Wittke-Thompson et al., Am J Hum Genet 76:967-986, Fig 1

1



Roeder, Stat Sci 9:222-278, 1994, Fig 4

More on data visualization

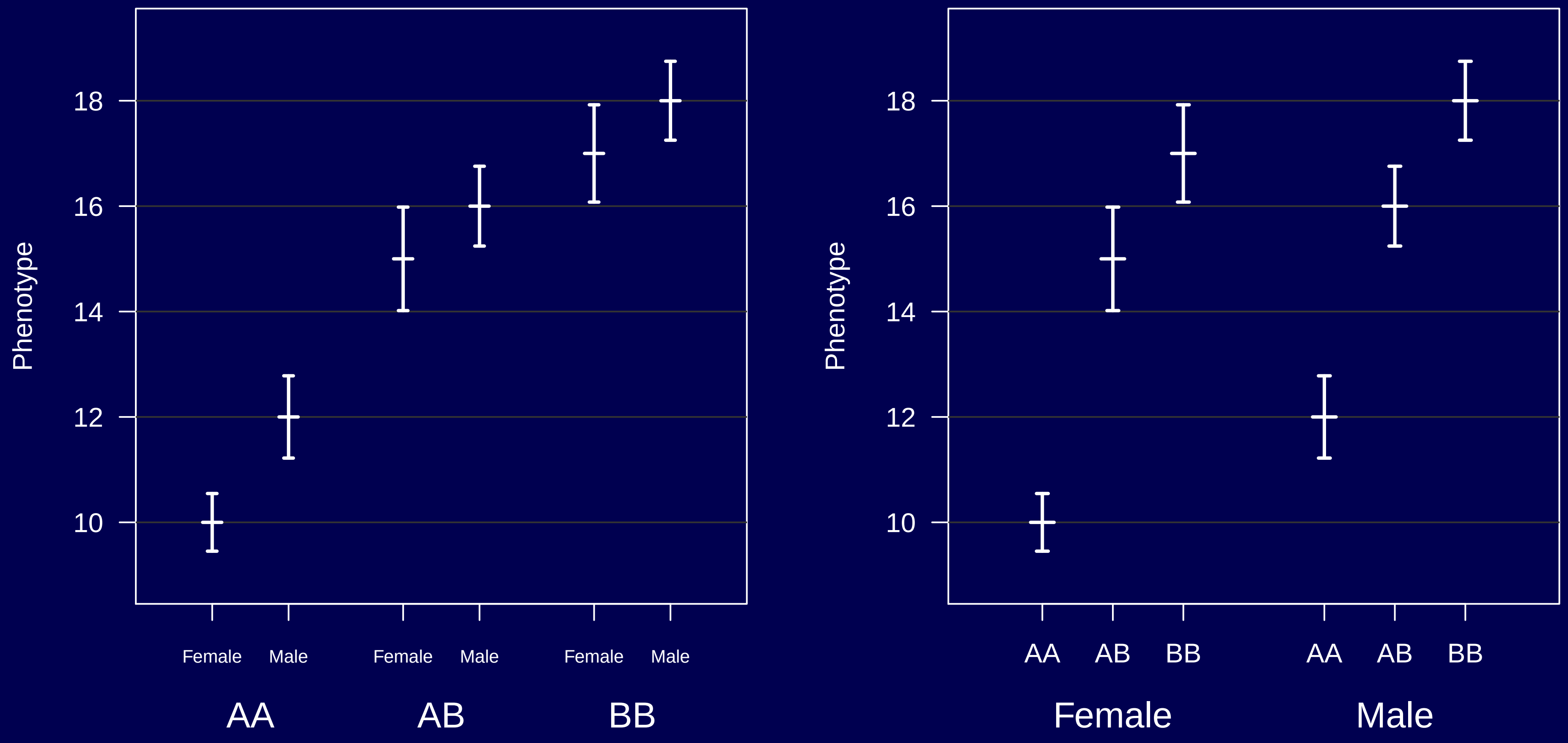
Karl W Broman

Department of Biostatistics & Medical Informatics
University of Wisconsin – Madison

`www.biostat.wisc.edu/~kbroman`

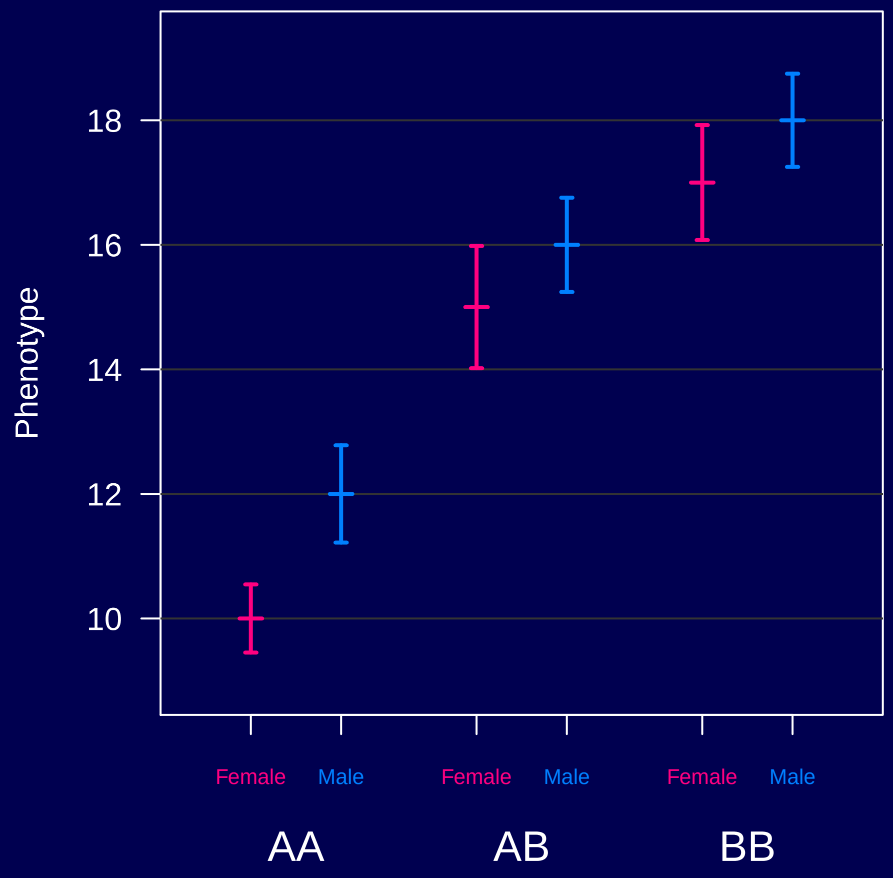
Ease comparisons

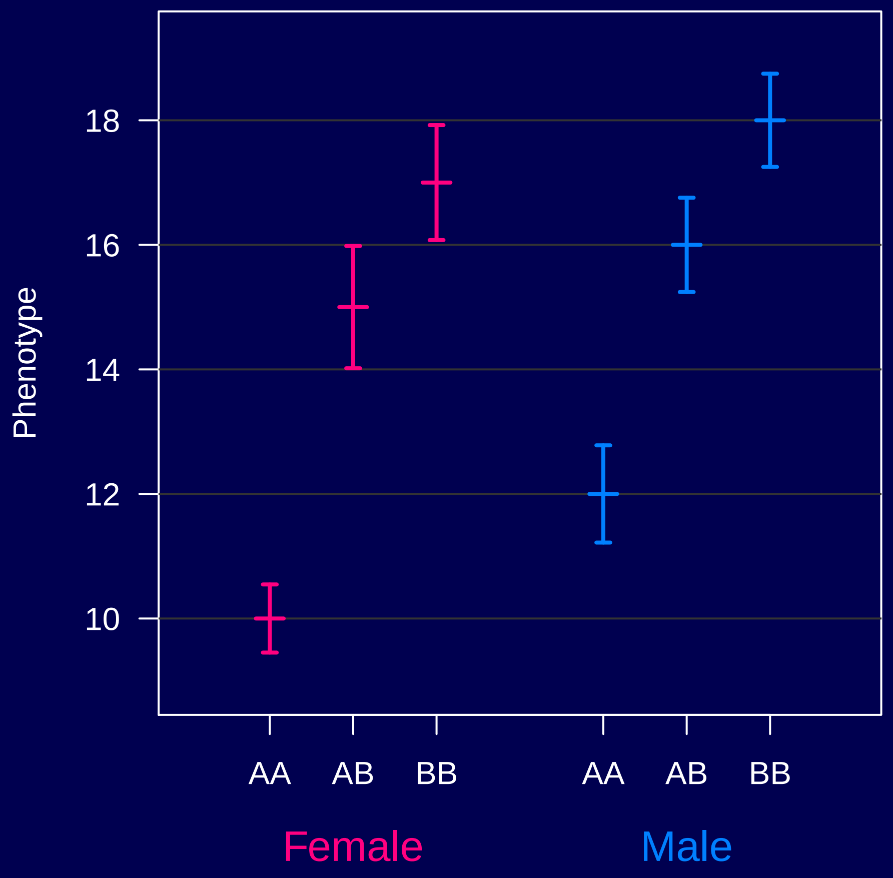
(things to be compared should be adjacent)



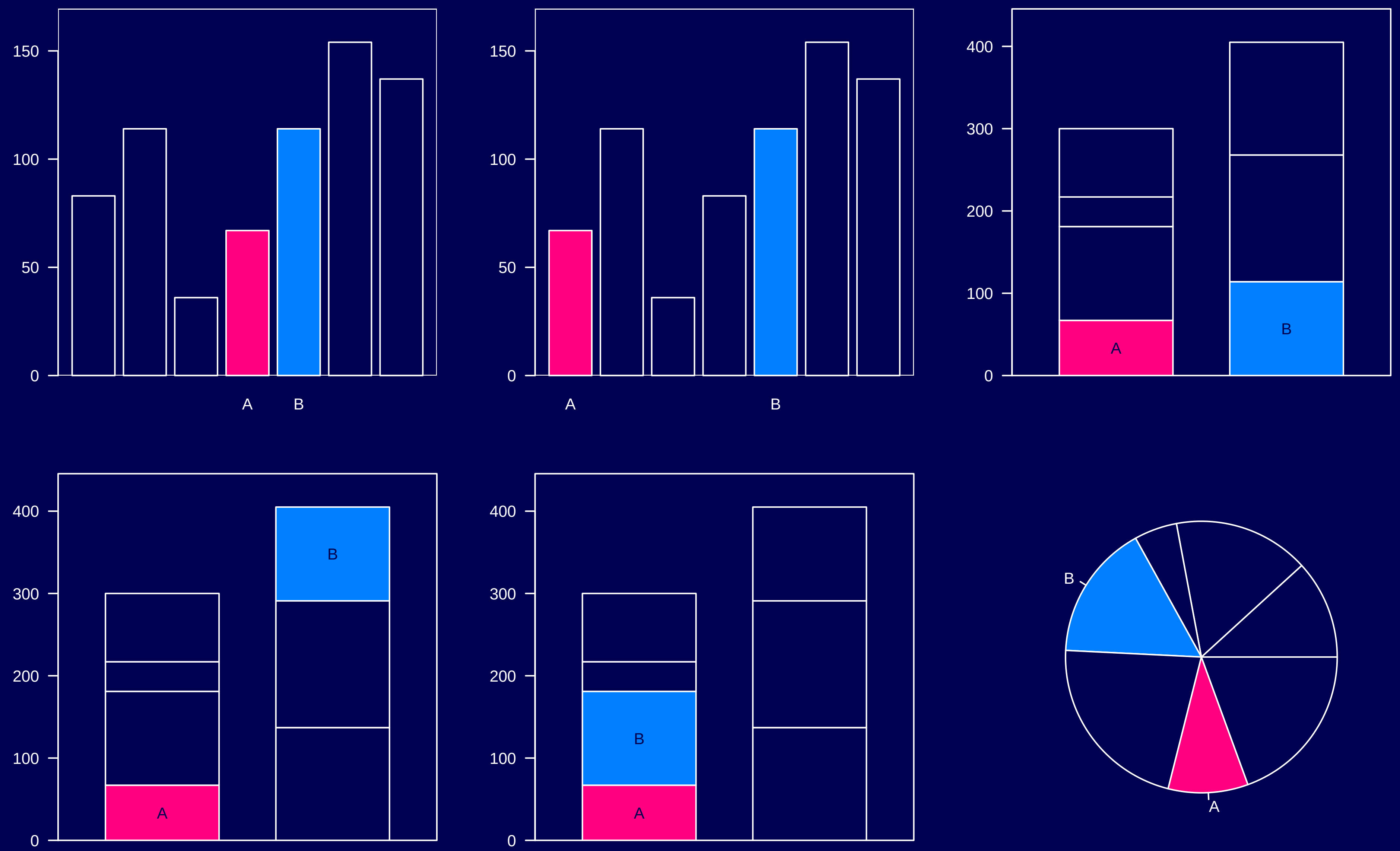
Ease comparisons

(add a bit of color)



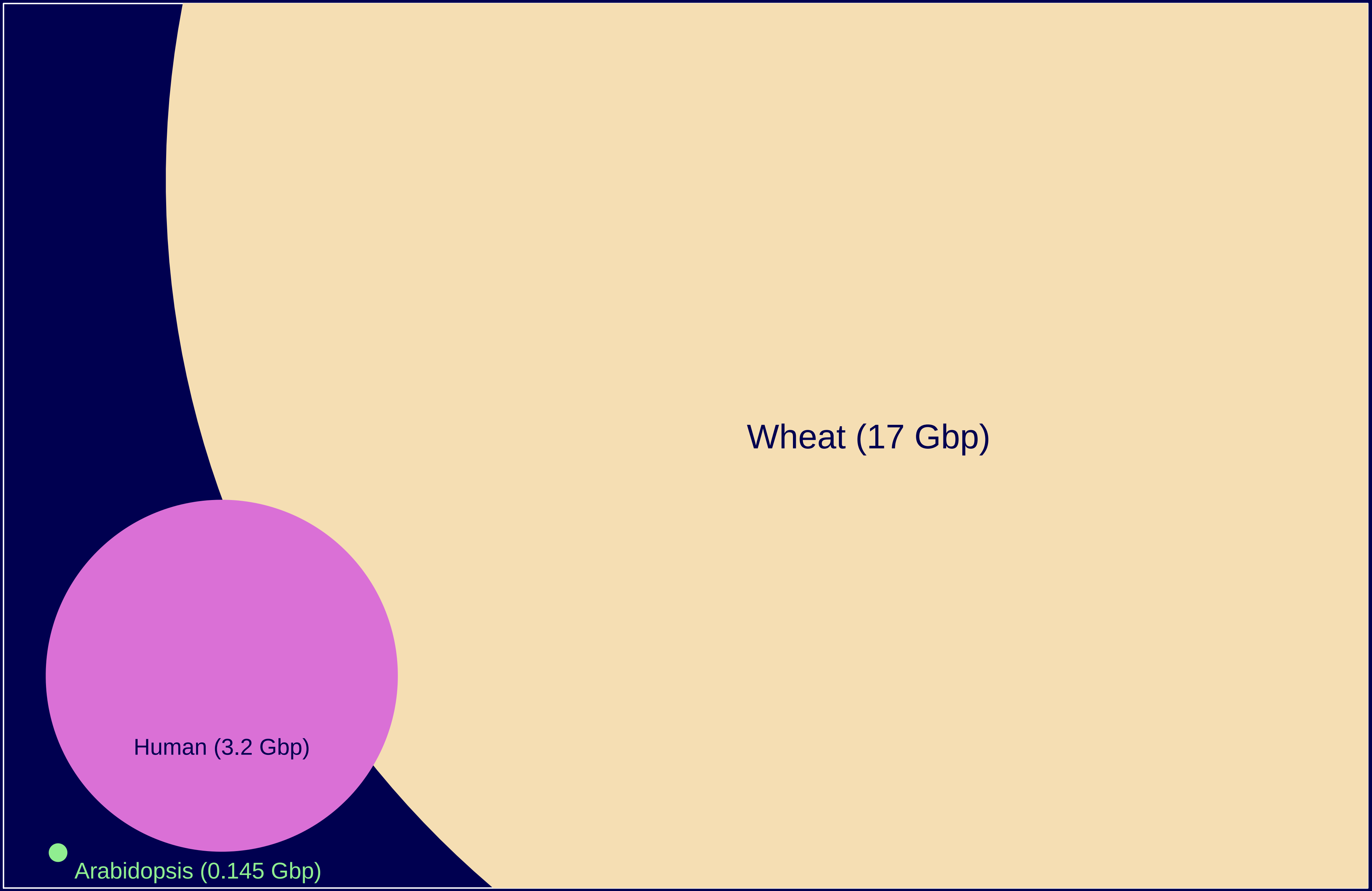


Which comparison is easiest?



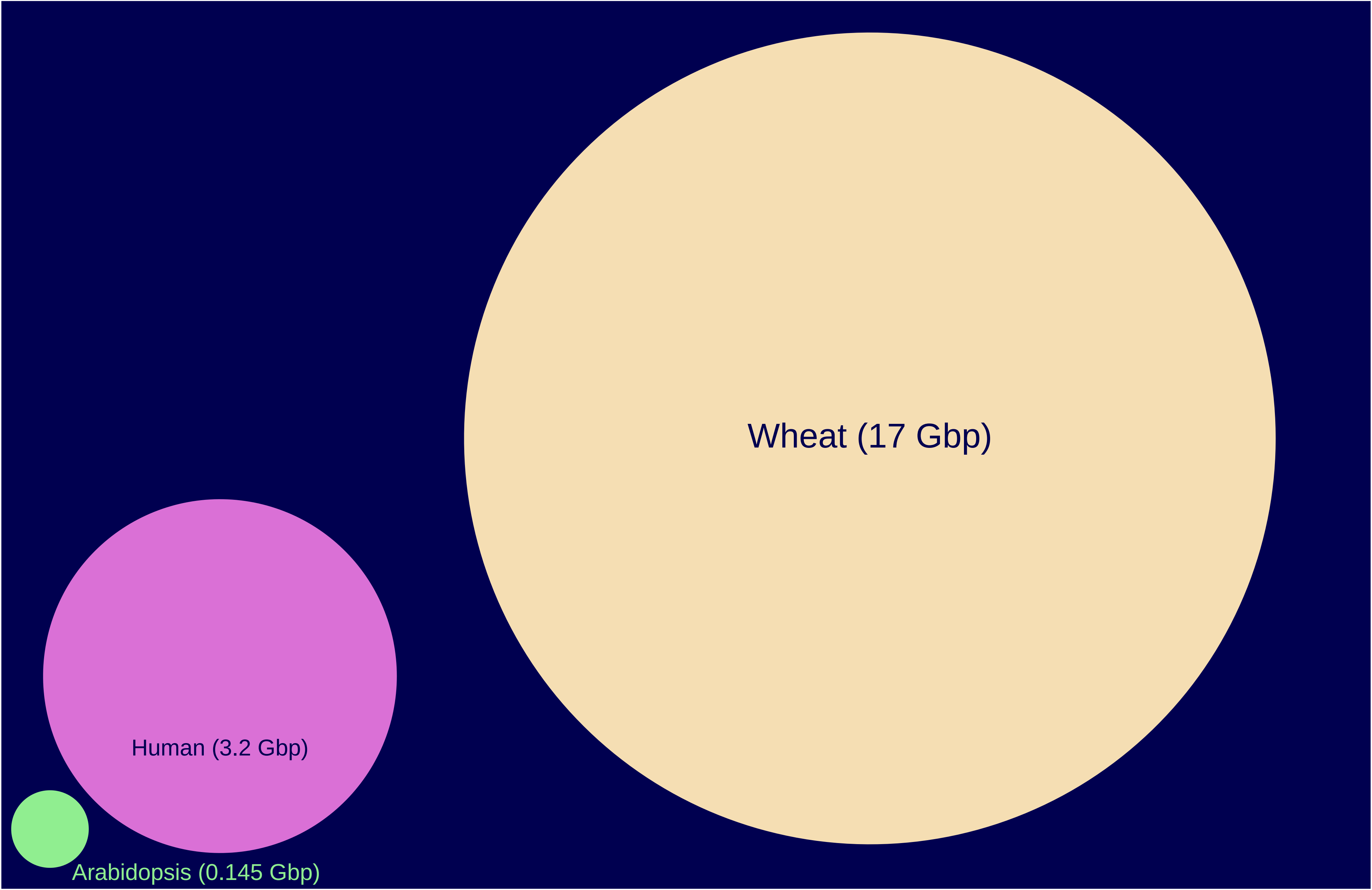
Don’t distort the quantities

(value $\propto$ radius)



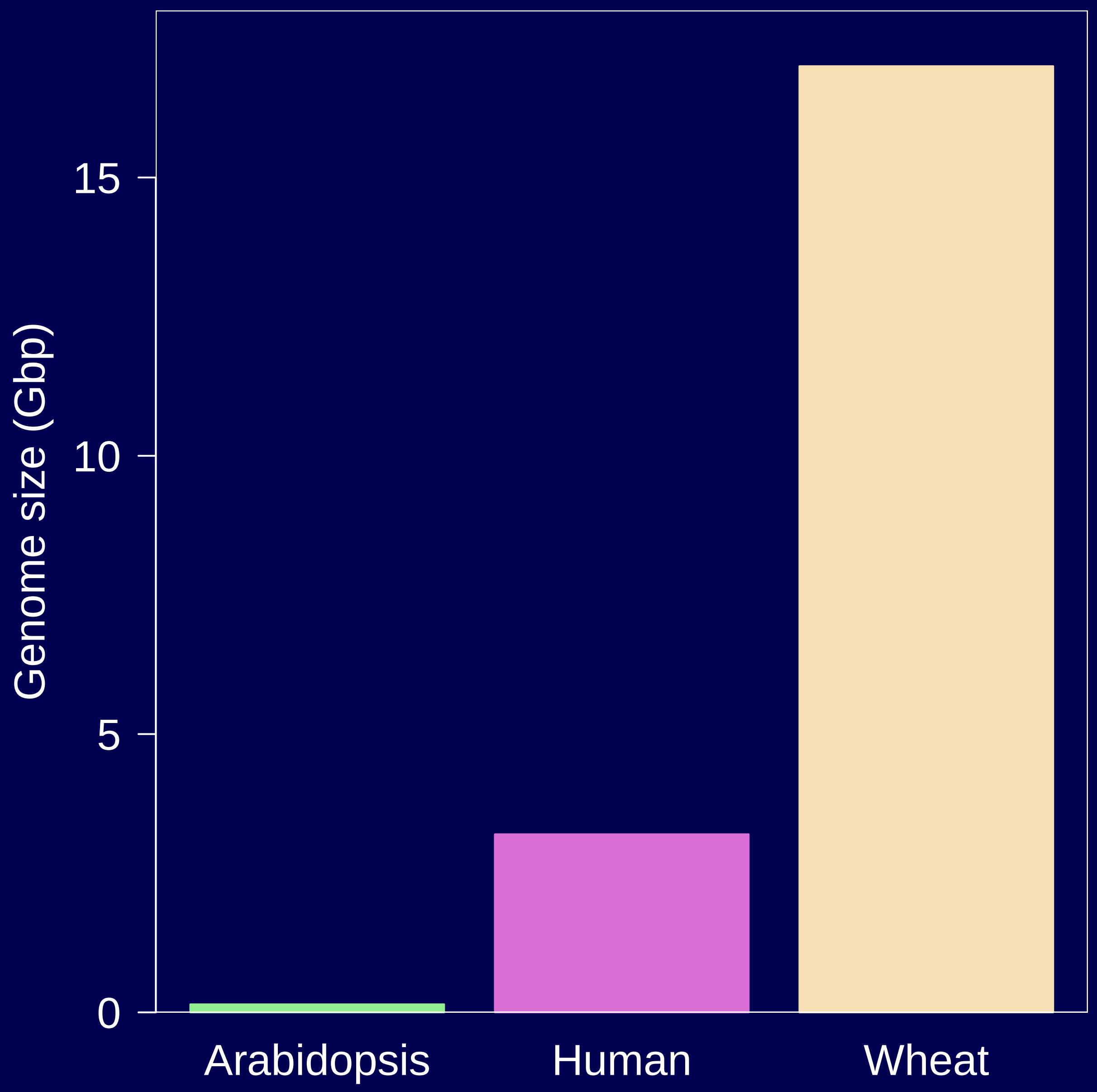
Don’t distort the quantities

(value $\propto$ area)



Don’t use areas at all

(value $\propto$ length)



Encoding data

Quantities

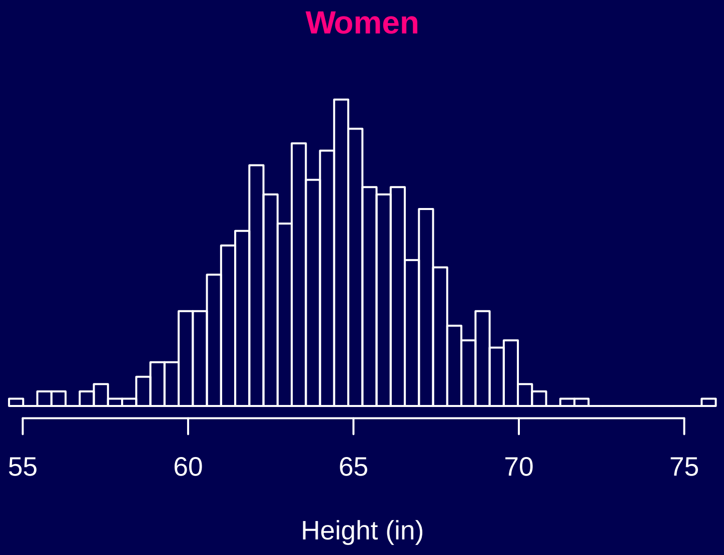
- Position
- Length
- Angle
- Area
- Luminance (light/dark)
- Chroma (amount of color)

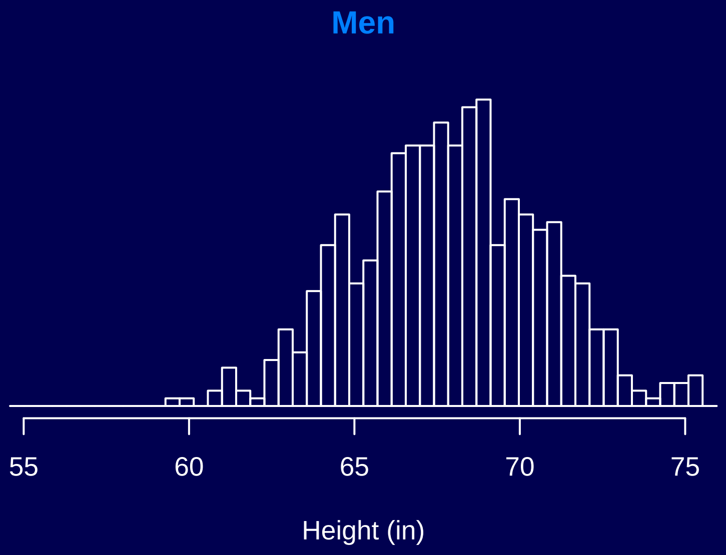
Categories

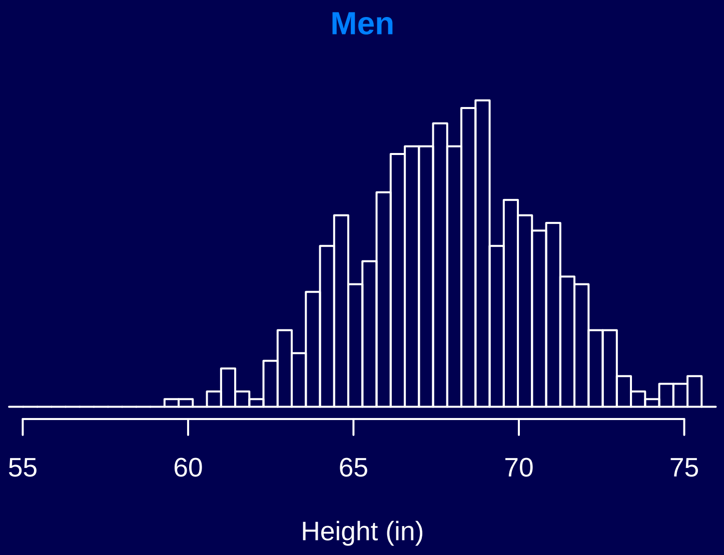
- Shape
- Hue (which color)
- Texture
- Width

Ease comparisons

(align things vertically)

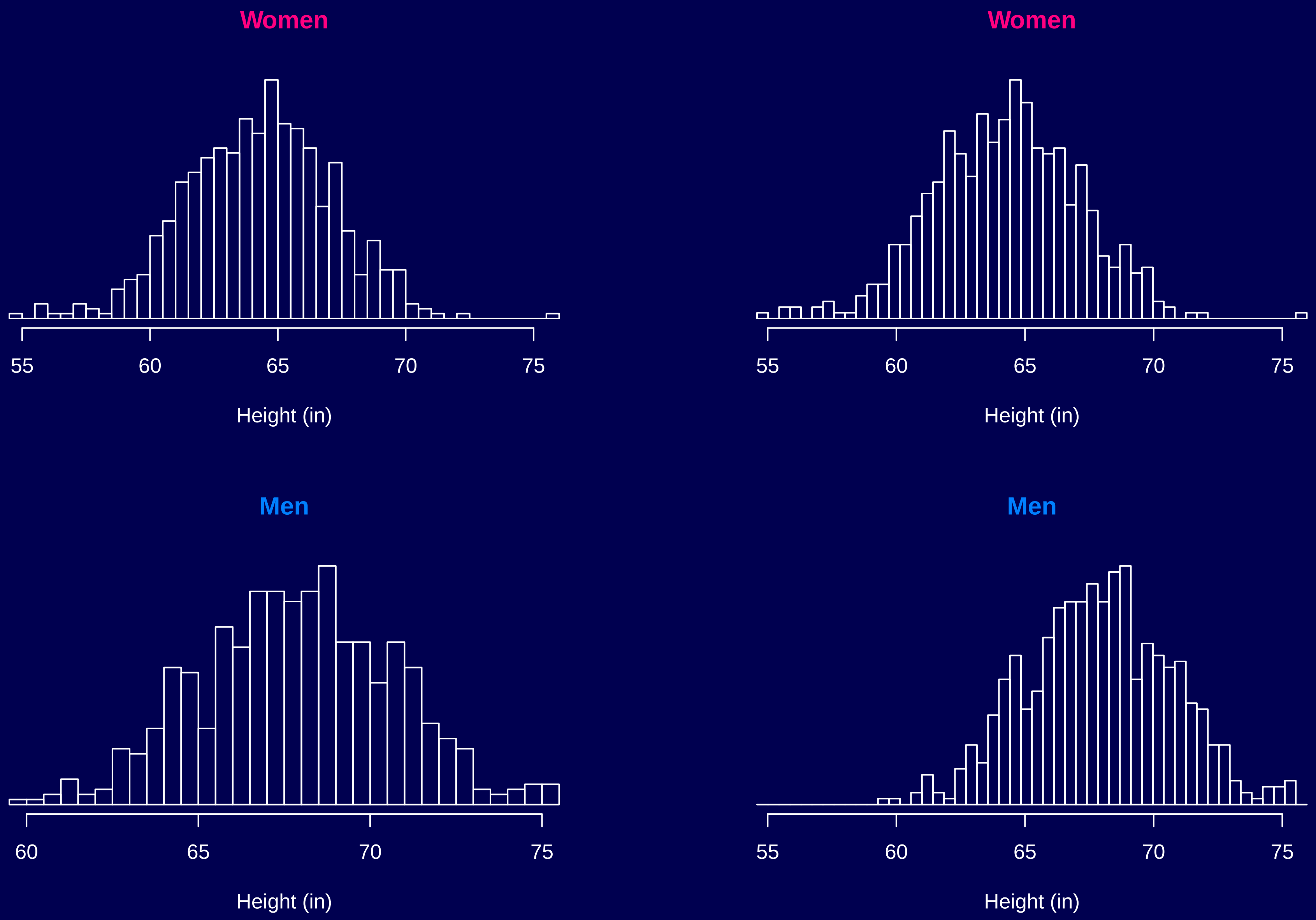




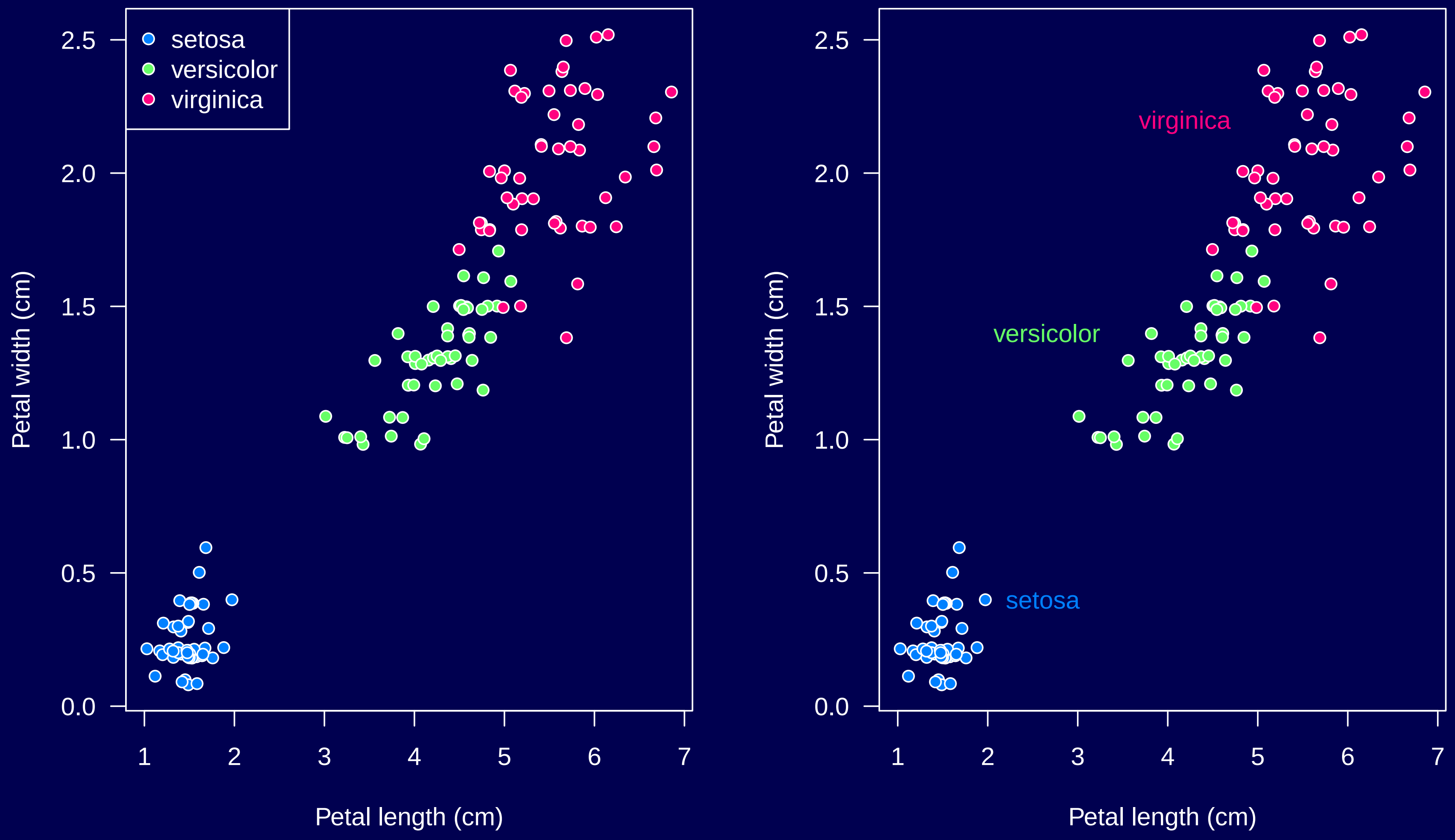


Ease comparisons

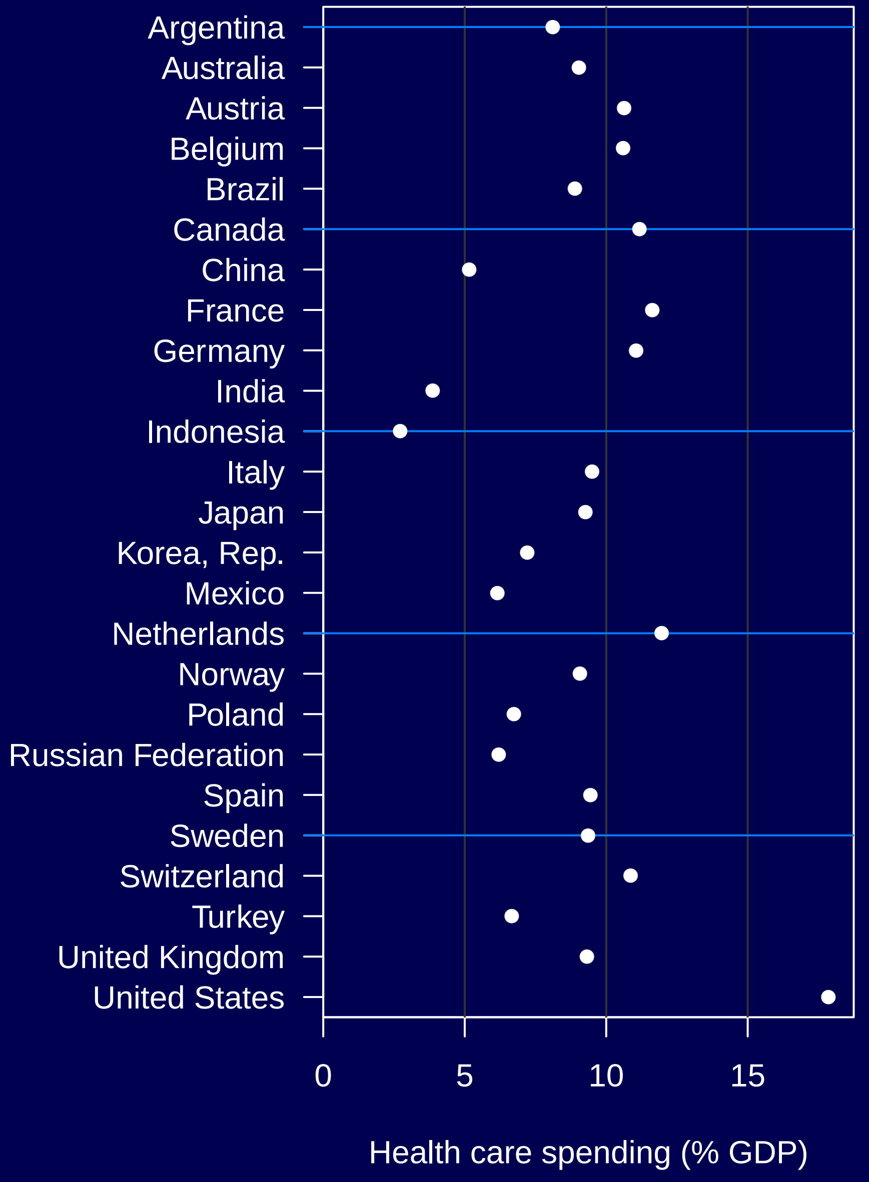
(use common axes)

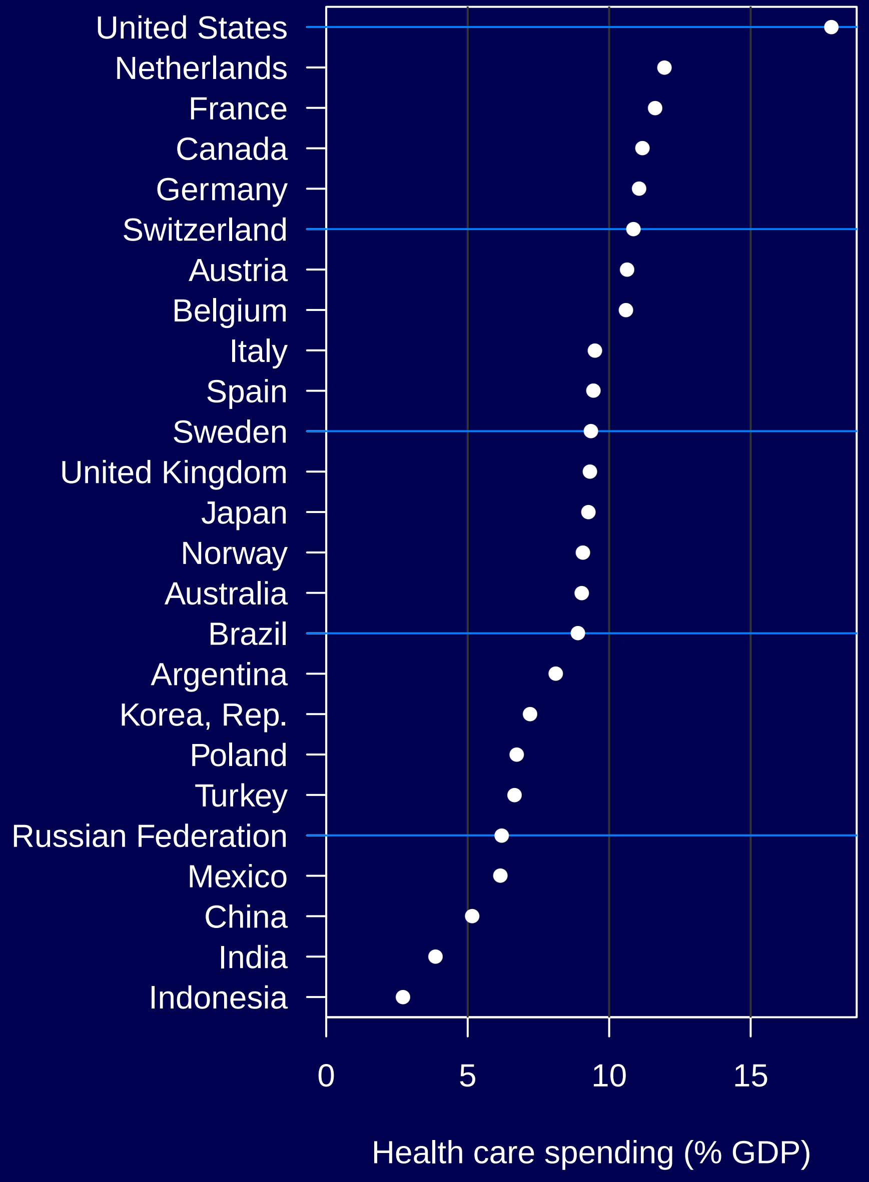


Use labels not legends

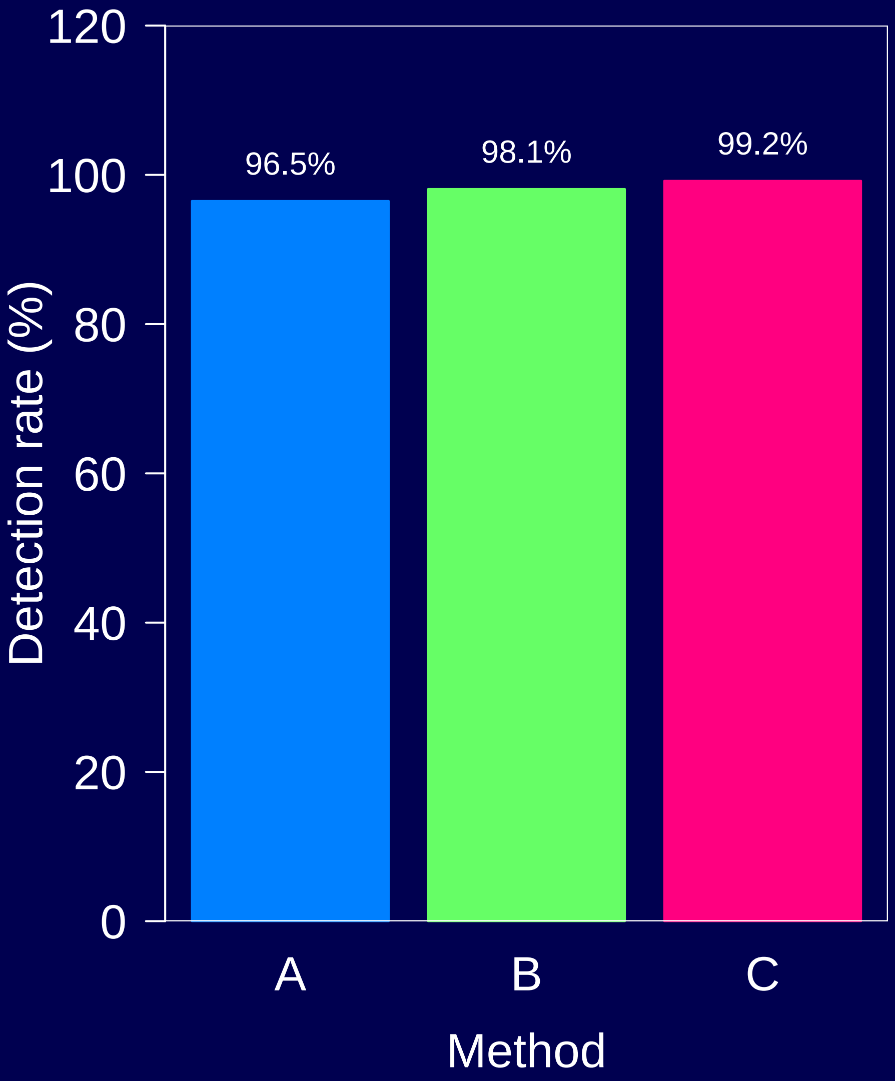


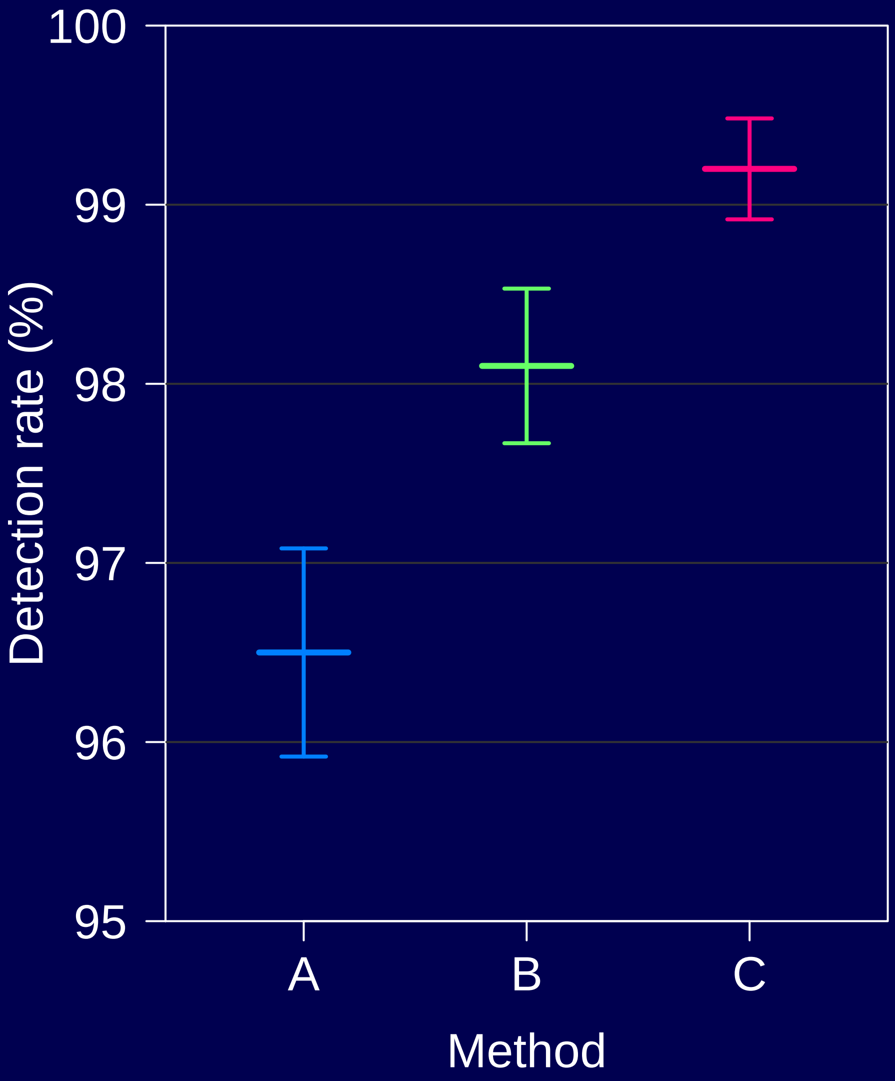
Don’t sort alphabetically





Must you include 0?





Summary

- Put the things to be compared next to each other
- Use color to set things apart, but consider color blind folks
- Use position rather than angle or area to represent quantities
- Align things vertically to ease comparisons
- Use common axis limits to ease comparisons
- Use labels rather than legends
- Sort on meaningful variables (not alphabetically)
- Must 0 be included in the axis limits?
- Consider taking logs and/or differences

Inspirations

- Hadley Wickham (slides at `http://courses.had.co.nz`)
- Naomi Robbins (*Creating more effective graphs*)
- Howard Wainer
- Andrew Gelman
- Edward Tufte