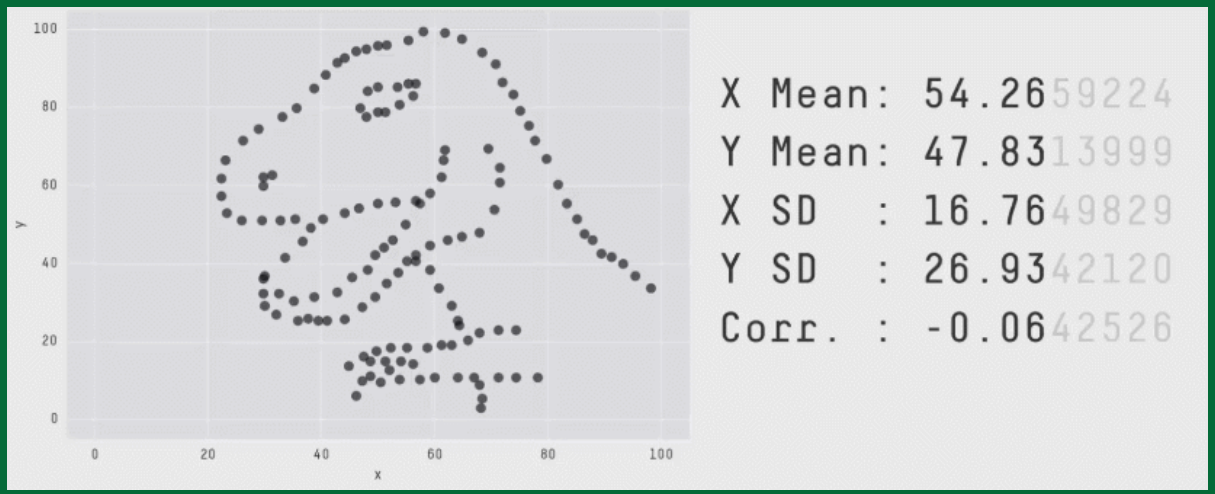
DSBA 5122: Visual Analytics

Class 5: Distributions and Uncertainty

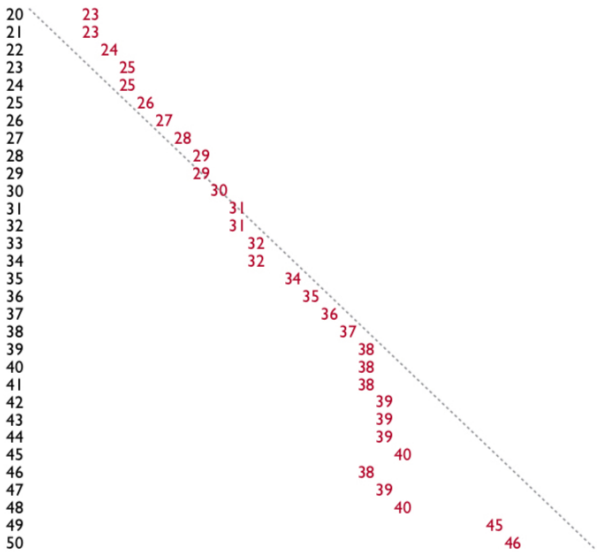
Ryan Wesslen

February 18, 2019

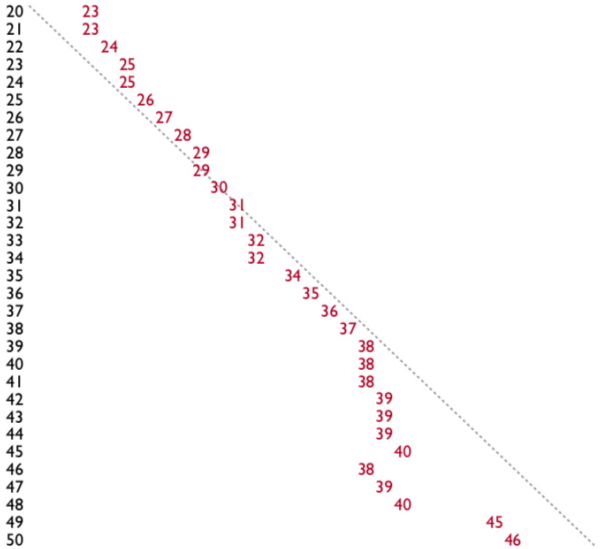
Why view Distributions? Cairo Ch. 7 & Wilke Ch. 7 - 9



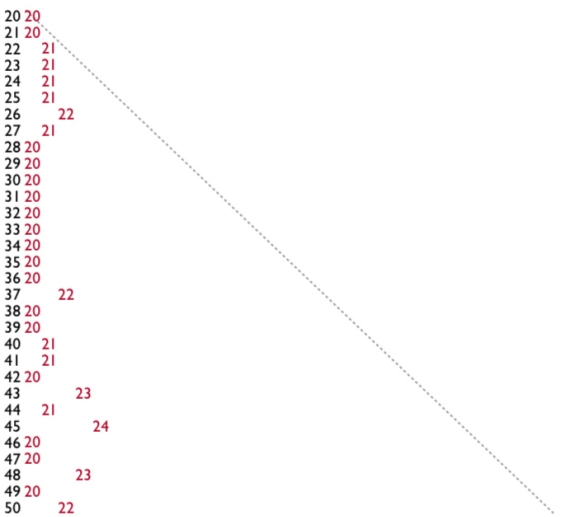
a woman's age vs. the age of the men who look best to her



a woman's age *vs. the age of the men who look best to her*



a man's age *vs. the age of the women who look best to him*



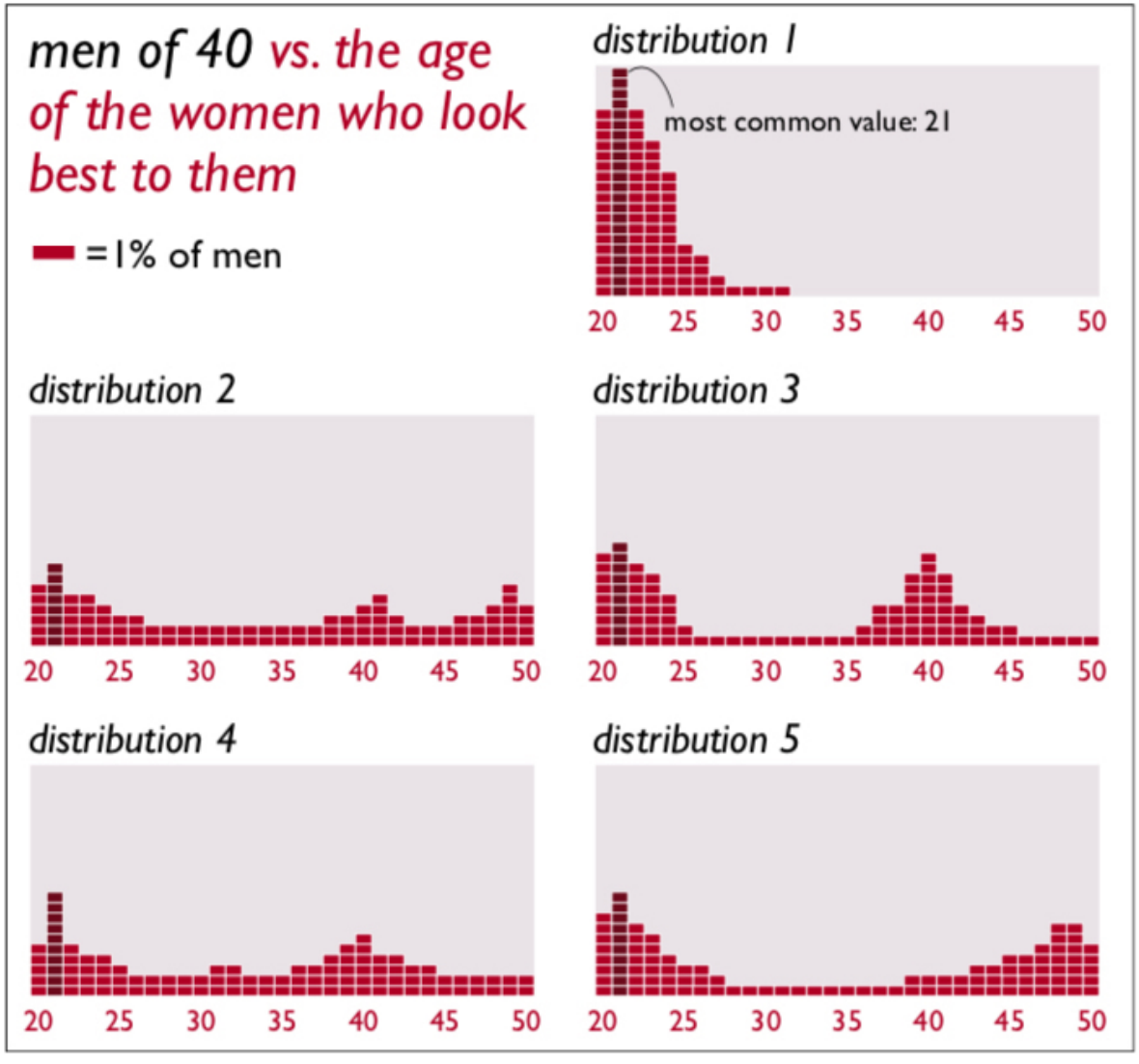
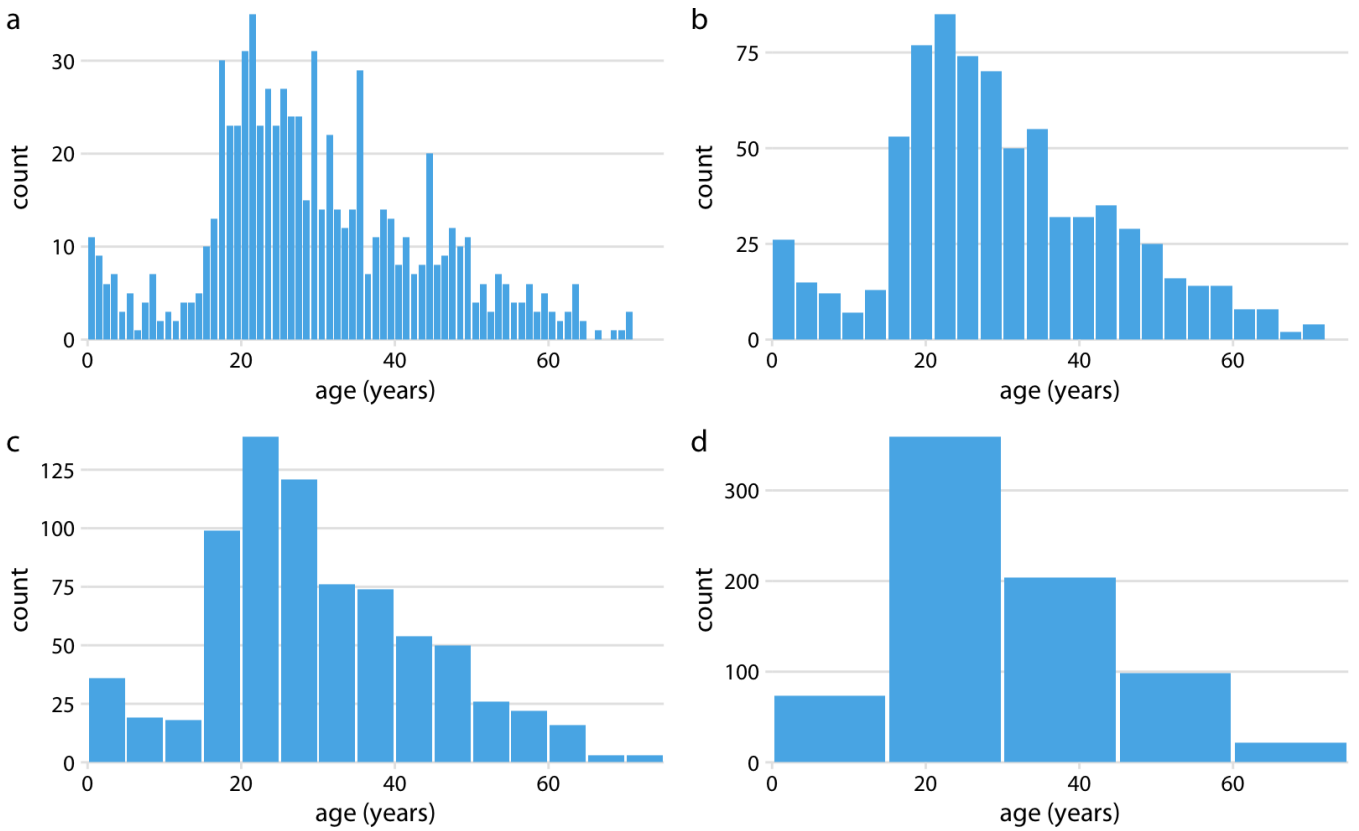
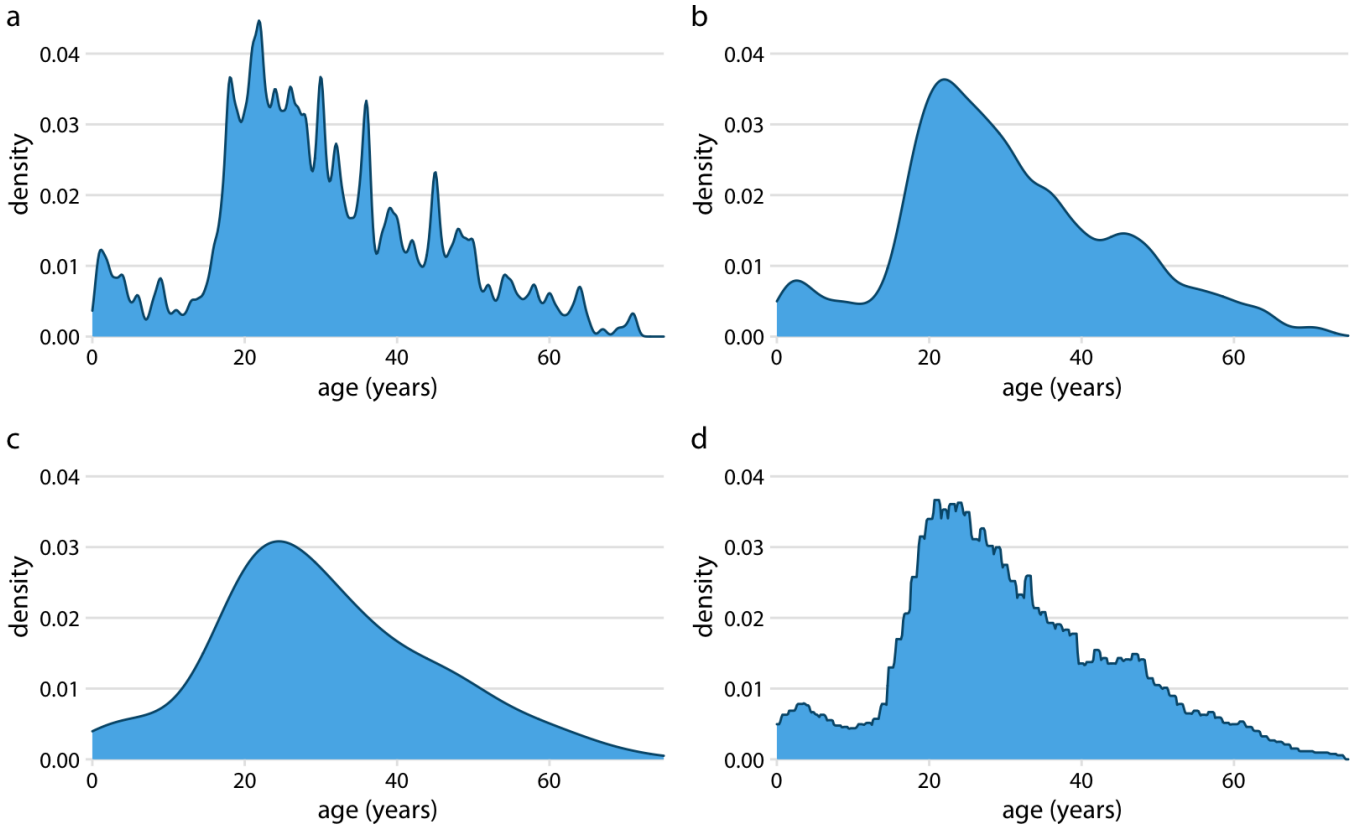
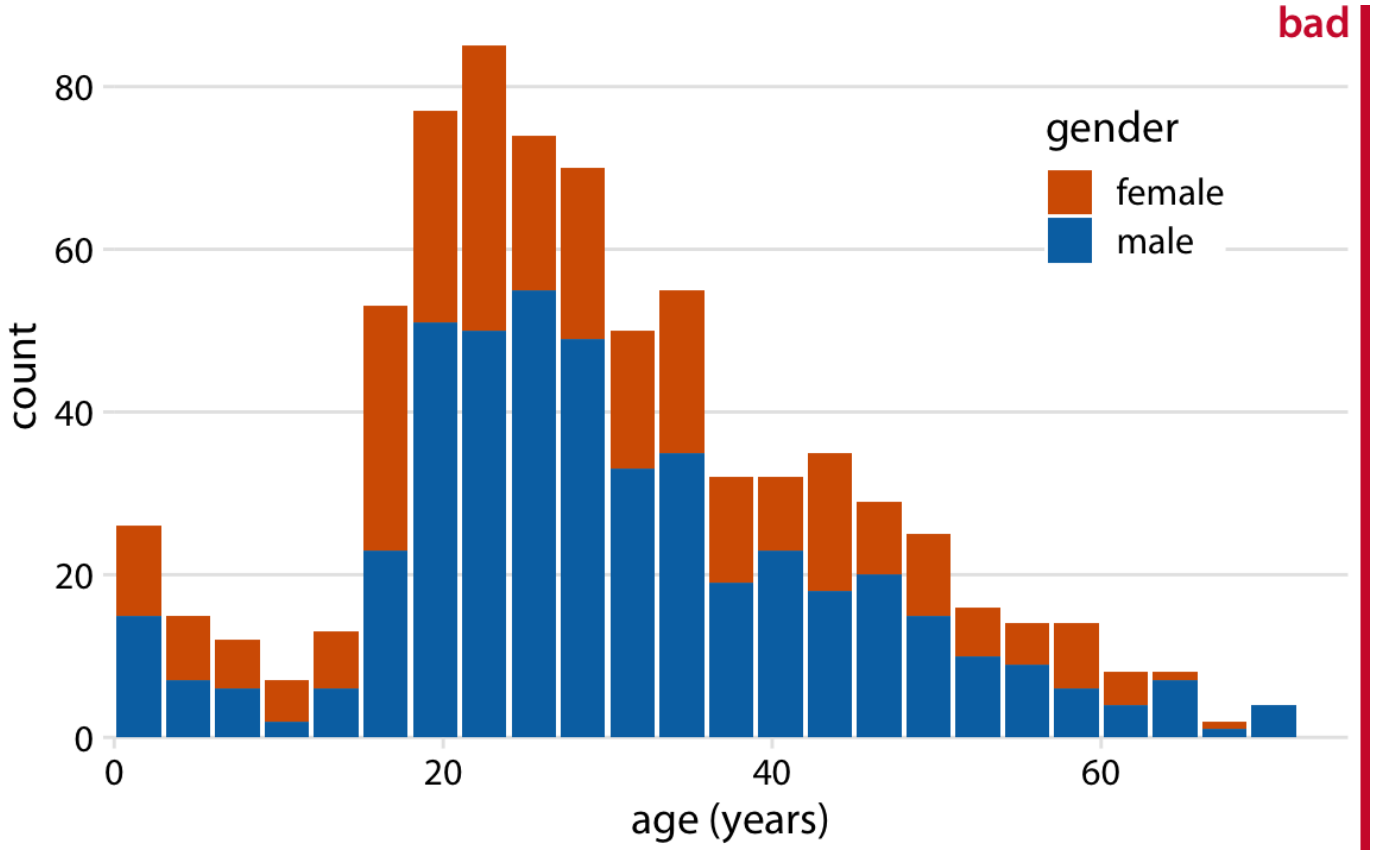
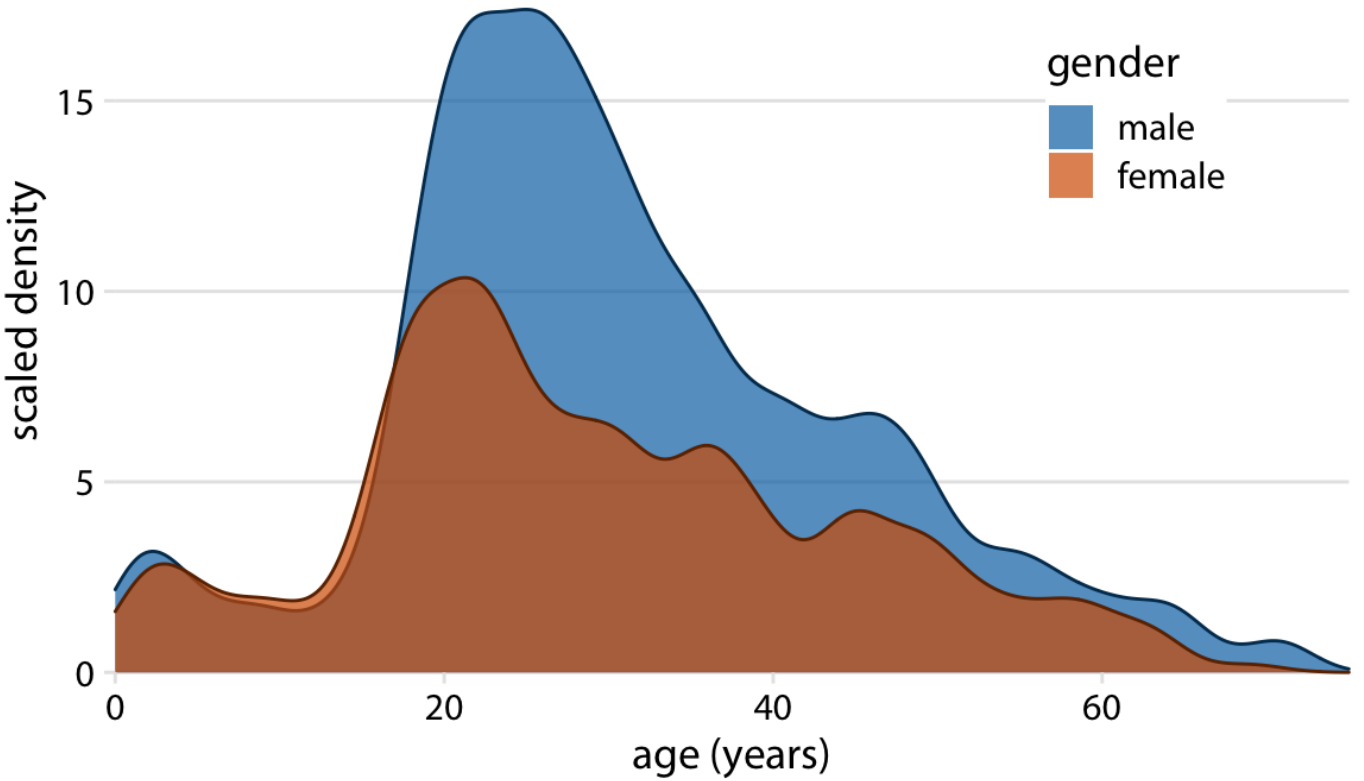


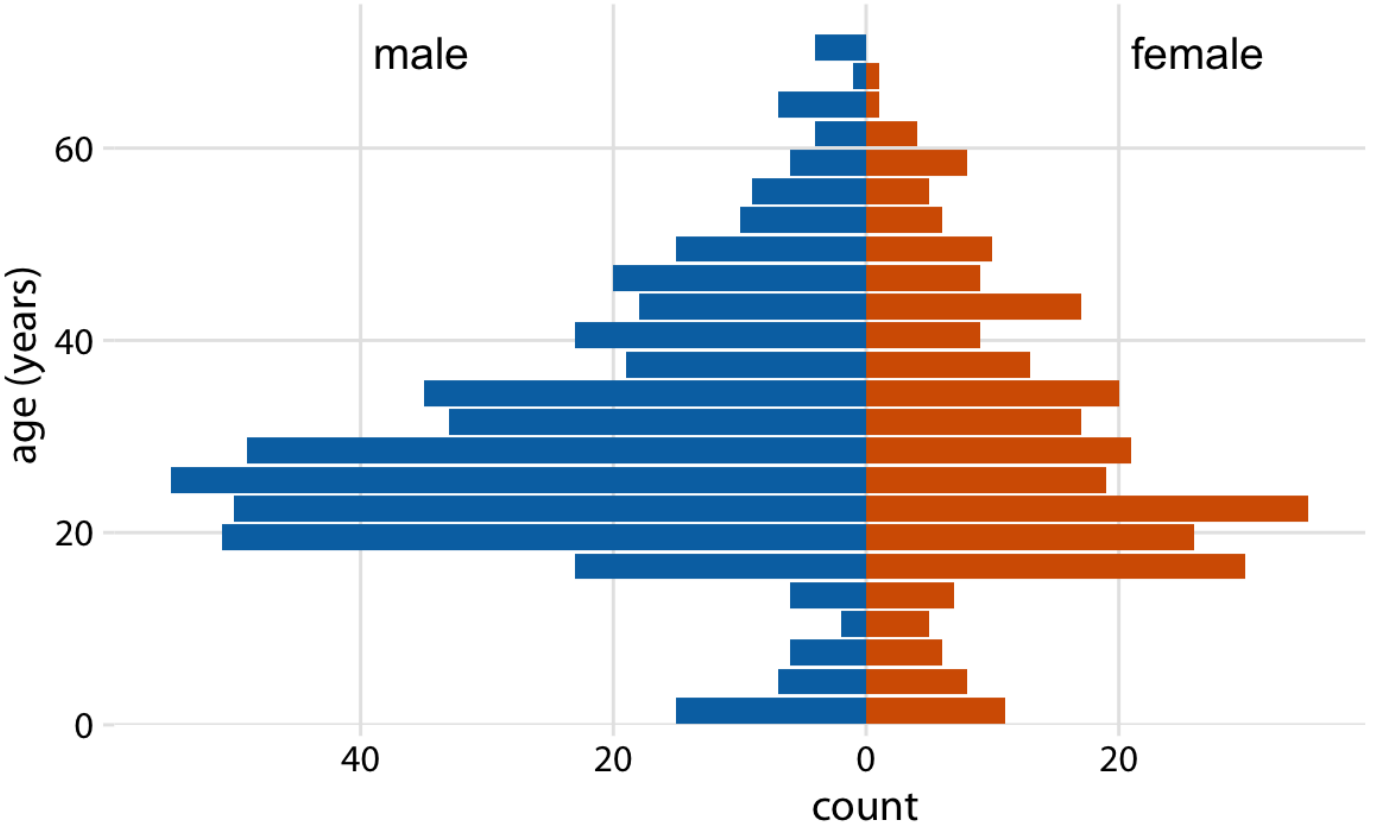
Figure 7.2 Five possible (and fictitious) distributions for the data corresponding to the preferences of men of 40. All of them have the same mode: 21.

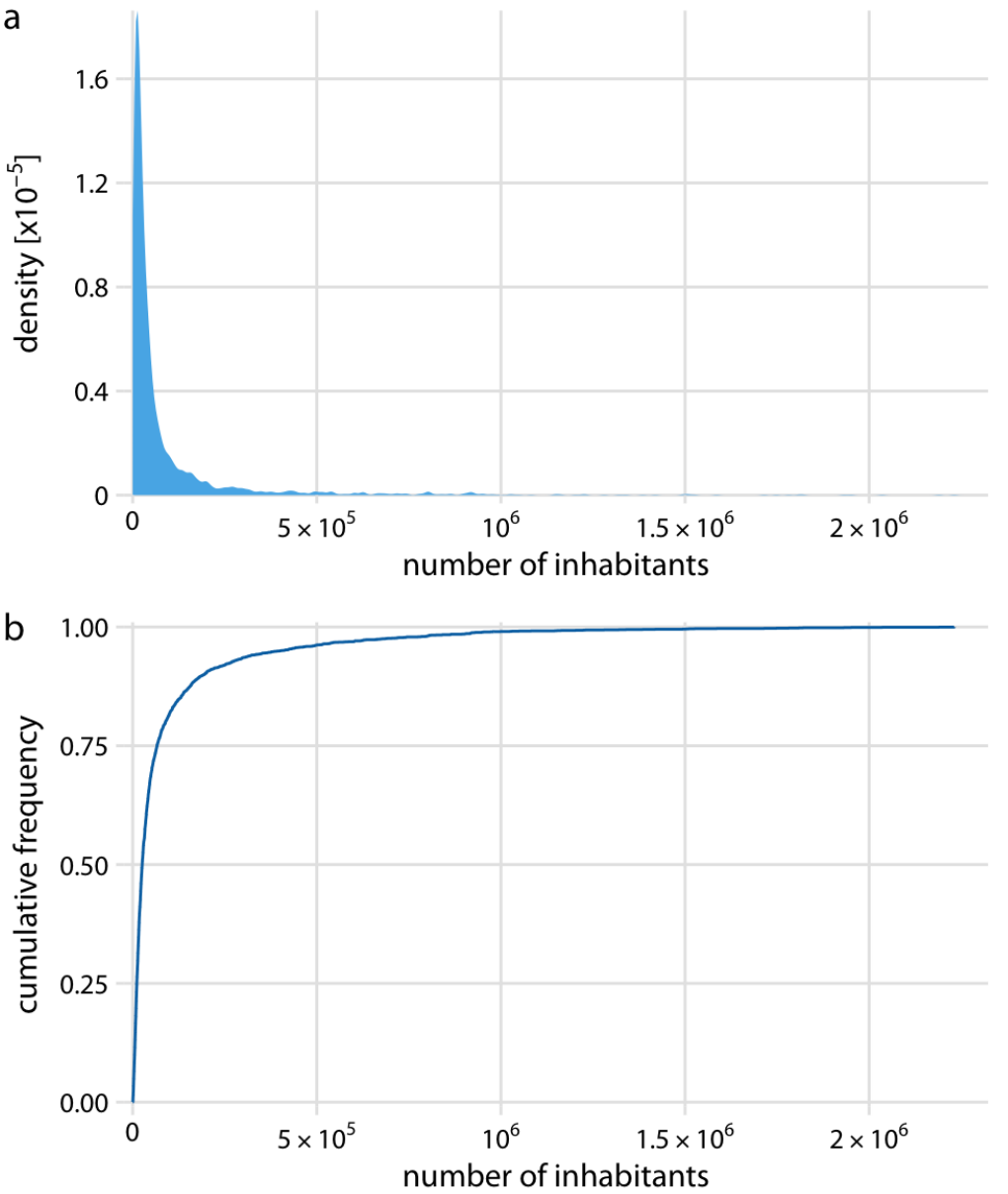

a
count
30
20
10
0
0
20
40
60
age (years)
b
count
75
50
25
0
0
20
40
60
age (years)
c
count
125
100
75
50
25
0
0
20
40
60
age (years)
d
count
300
200
100
0
0
20
40
60
age (years)

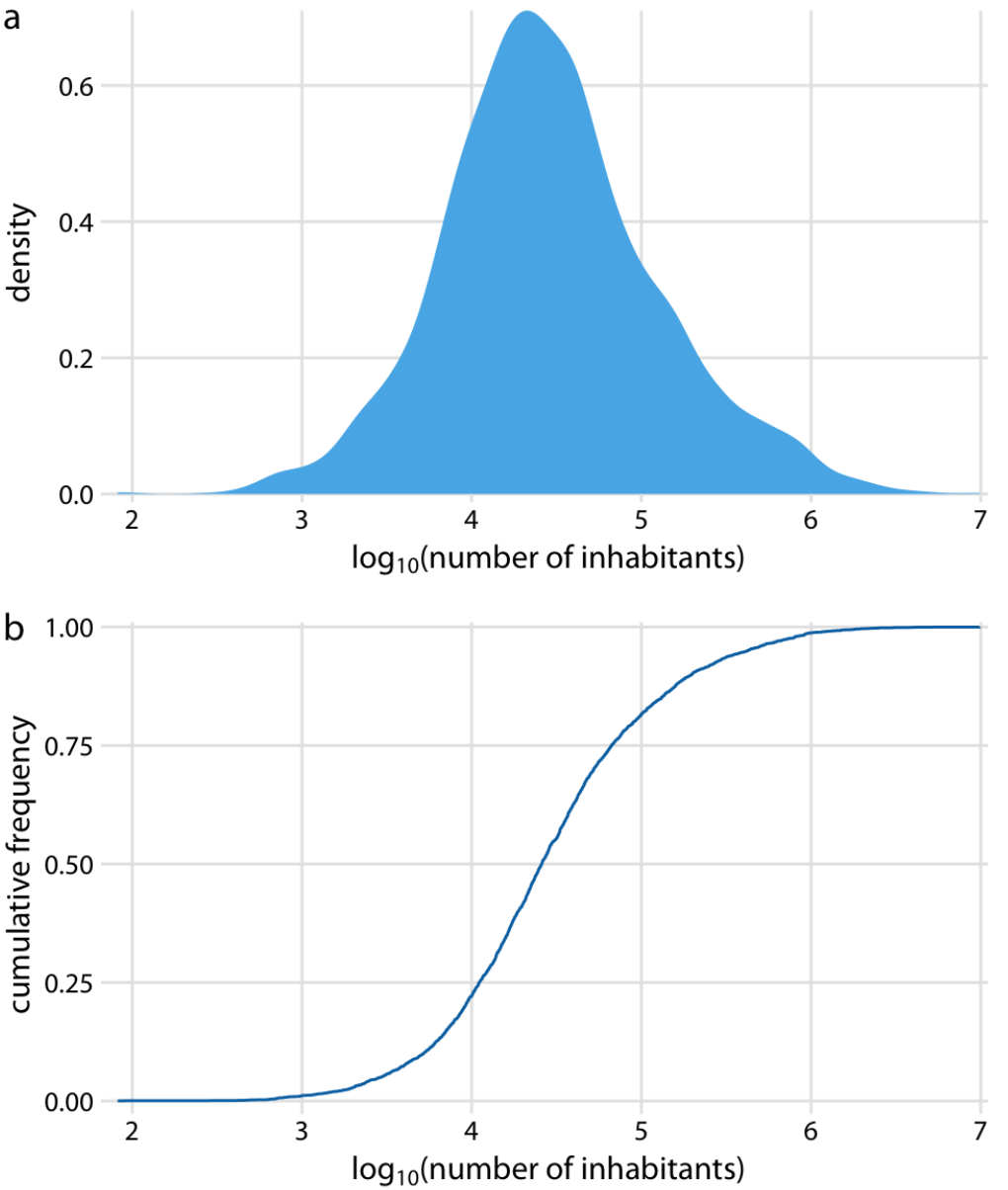

a
b
c
d
density
0.00
0.01
0.02
0.03
0.04
0
20
40
60
age (years)

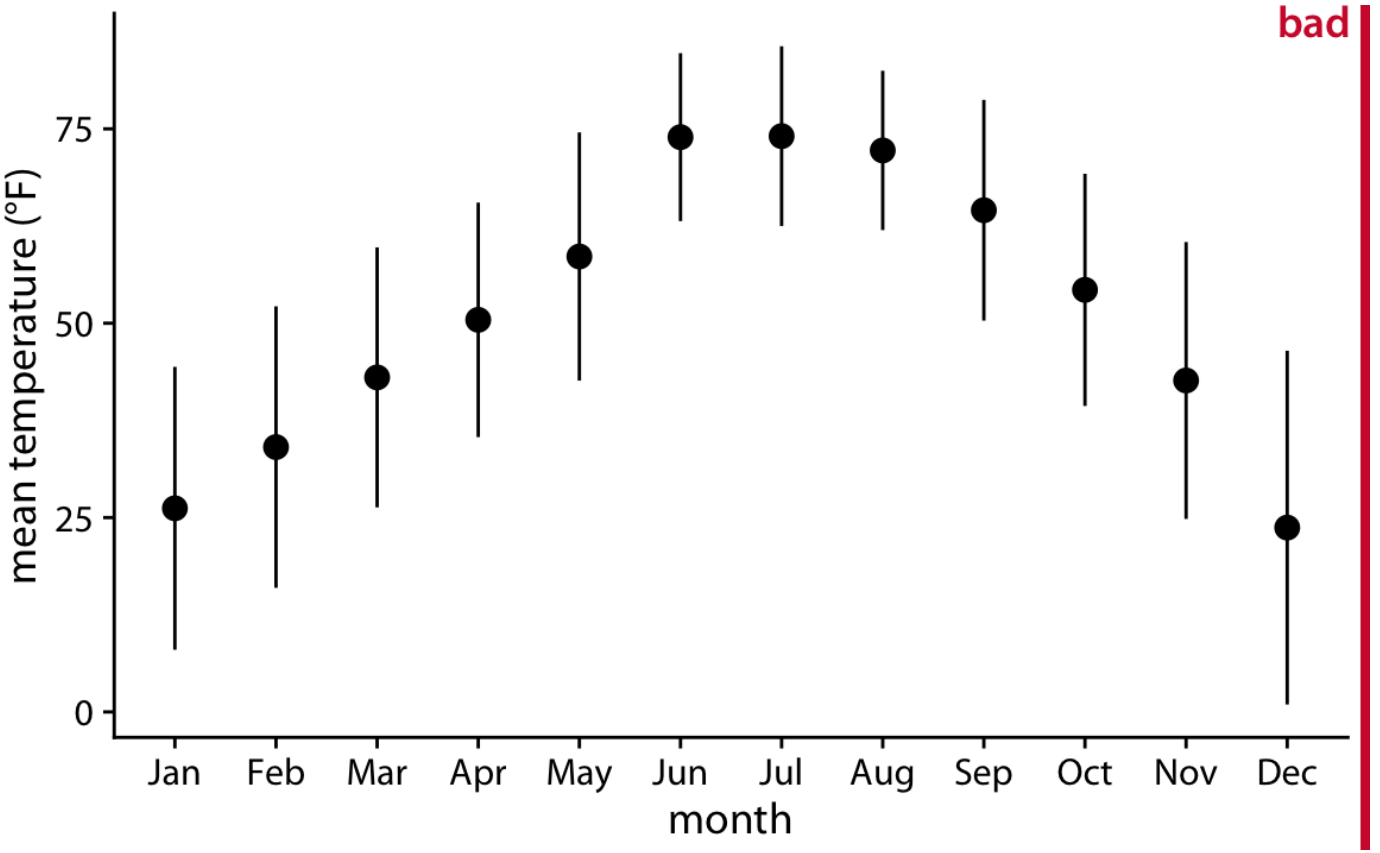

bad
gender
female
male
count
80
60
40
20
0
0
20
40
60
age (years)


gender
male
female
scaled density
15
10
5
0
0
20
40
60
age (years)

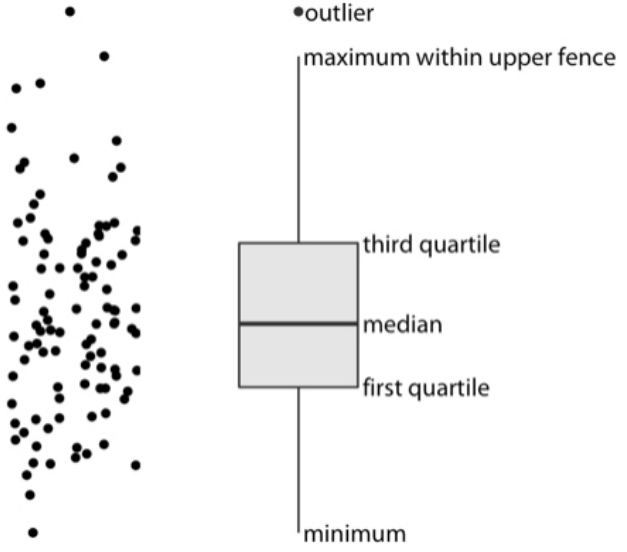

male
female
60
40
20
0
age (years)
40
20
0
20
count


a
density [x10^{-5}]
1.6
1.2
0.8
0.4
0
0
5×10^5
10^6
1.5×10^6
2×10^6
number of inhabitants
b
cumulative frequency
1.00
0.75
0.50
0.25
0.00
0
5×10^5
10^6
1.5×10^6
2×10^6
number of inhabitants

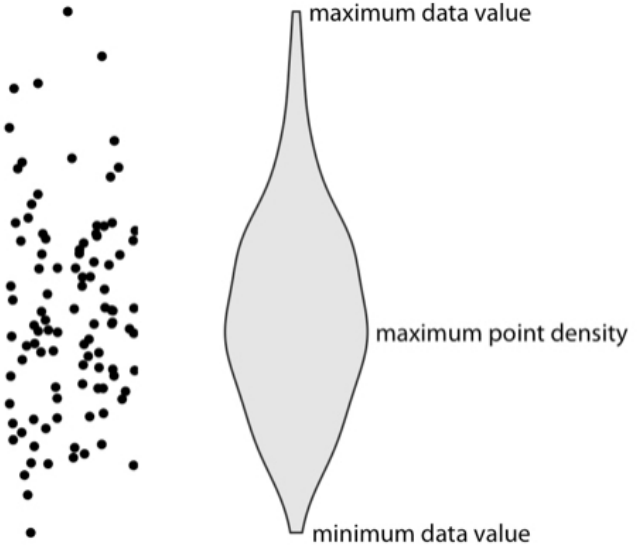

a
density
0.6
0.4
0.2
0.0
2
3
4
5
6
7
log10(number of inhabitants)
b
cumulative frequency
1.00
0.75
0.50
0.25
0.00
2
3
4
5
6
7
log10(number of inhabitants)

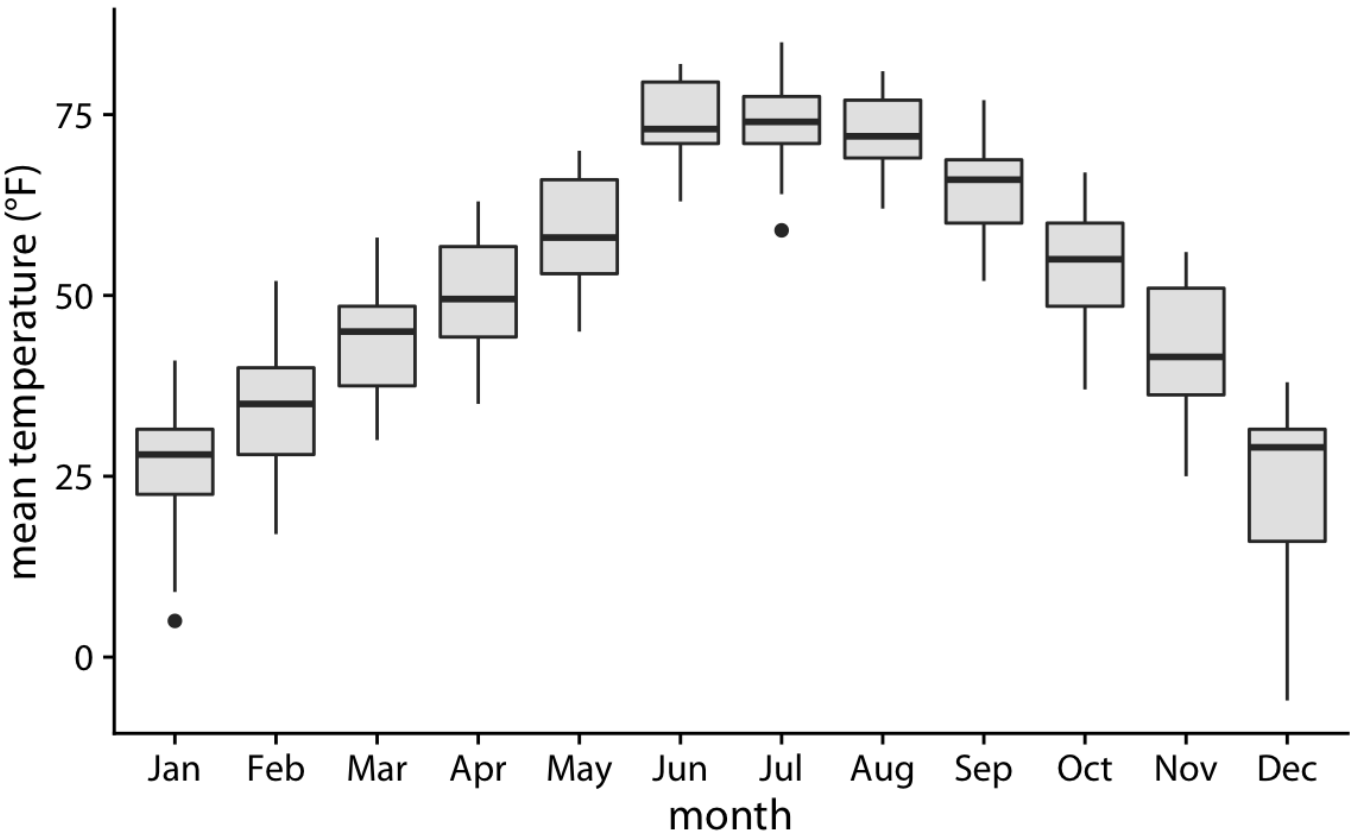

bad
mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

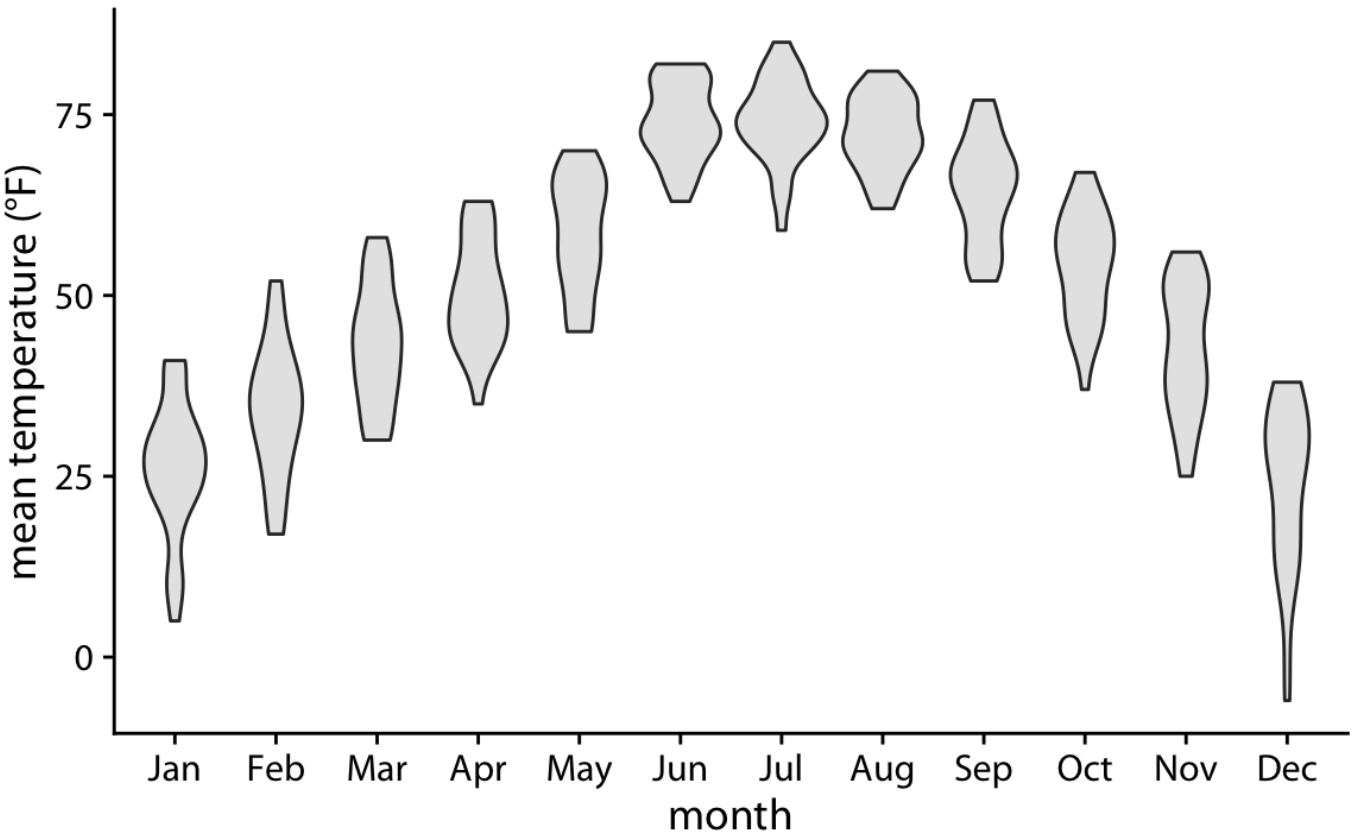
```
ggplot(df, aes(x, y)) + geom_boxplot()
```

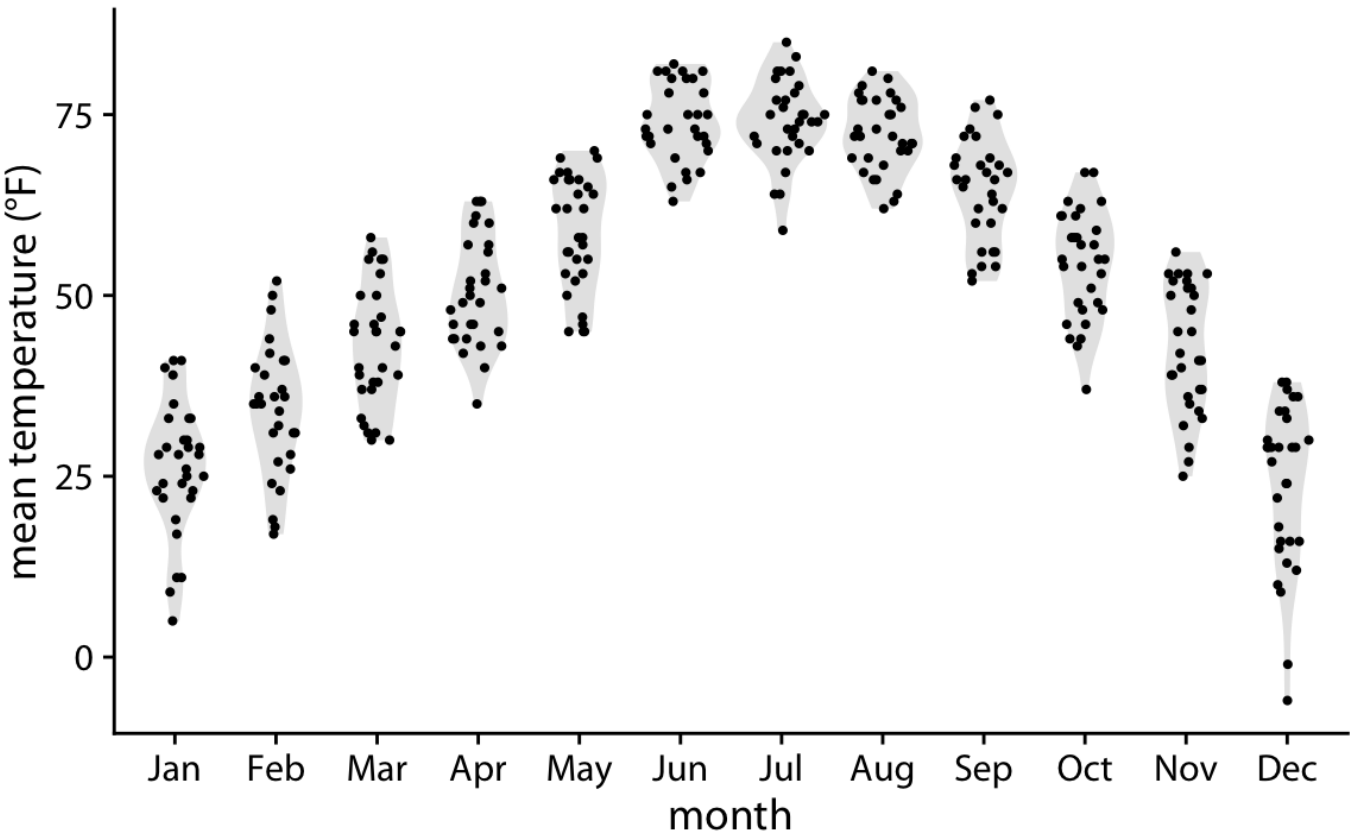


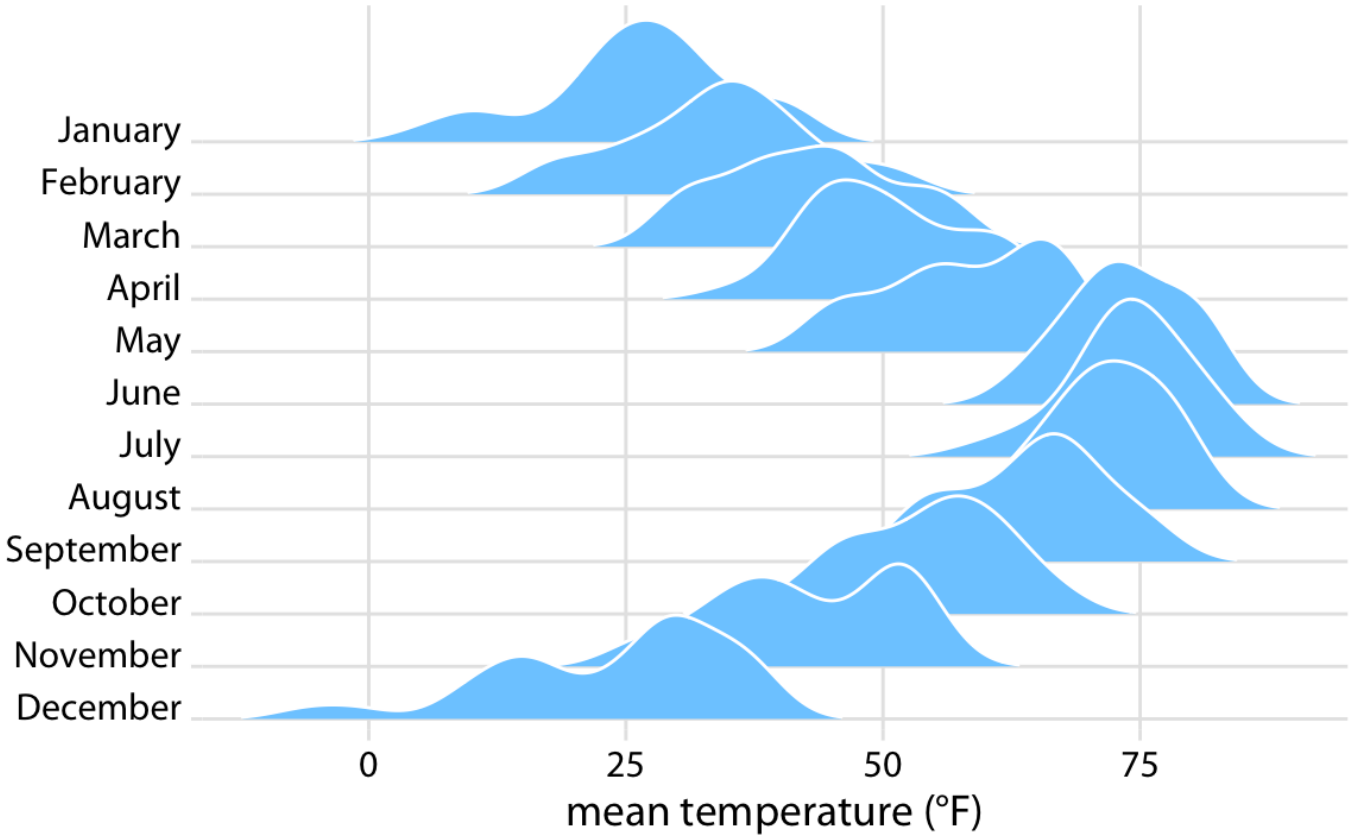

```
ggplot(df, aes(x, y)) + geom_violin()
```

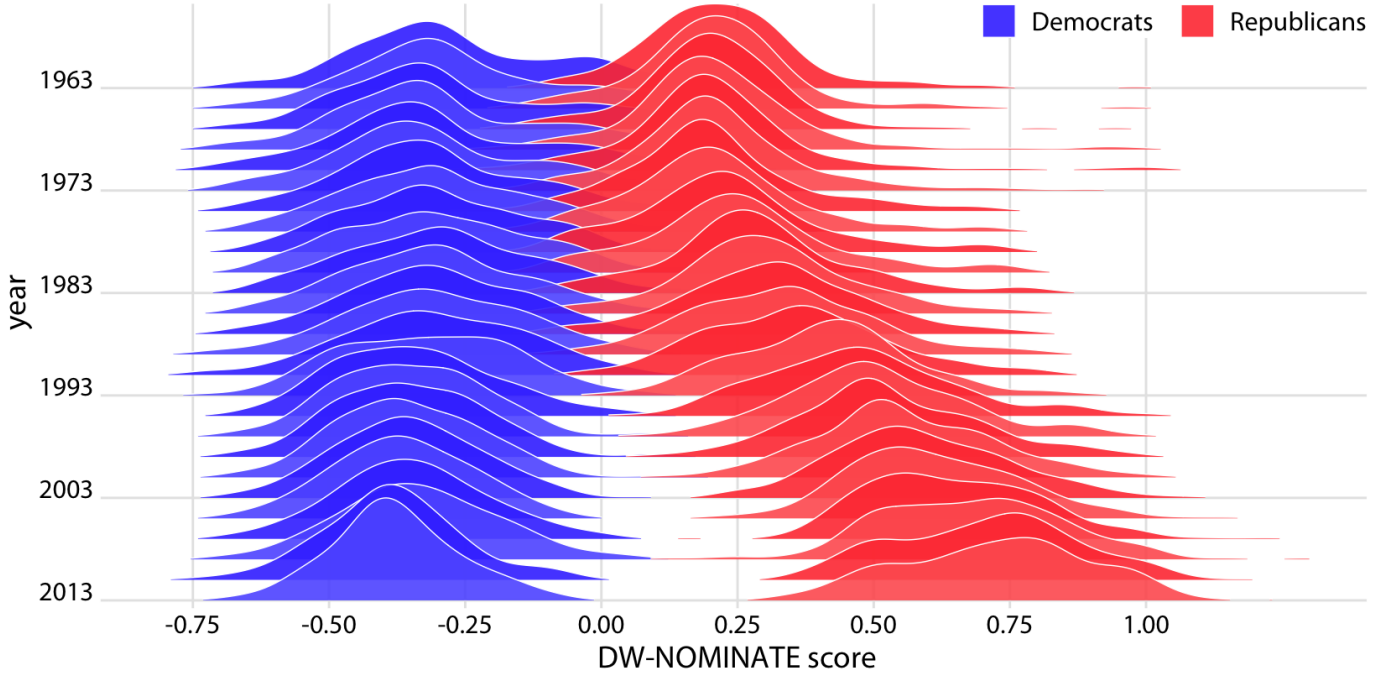



mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

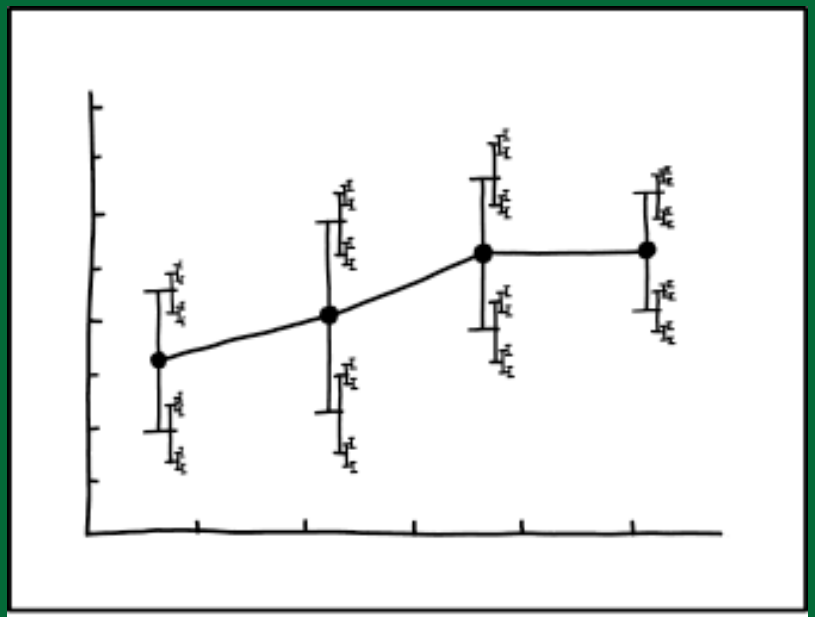

mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month


January
February
March
April
May
June
July
August
September
October
November
December
0
25
50
75
mean temperature (°F)


Democrats
Republicans
year
1963
1973
1983
1993
2003
2013
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
DW-NOMINATE score

Uncertainty: Cairo Ch. 10 & Wilke Ch. 16



xkcd

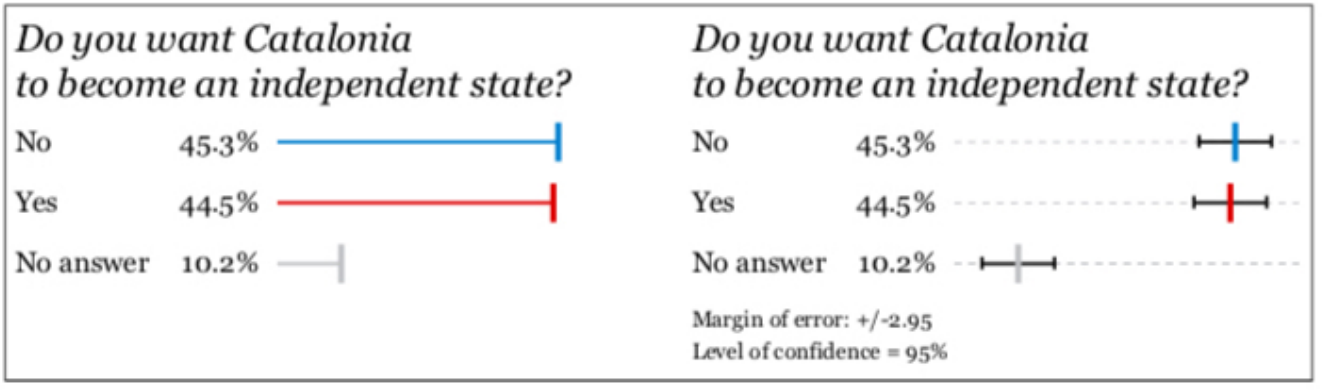


Figure 11.1 Displaying the margin of error can change your view of the data.

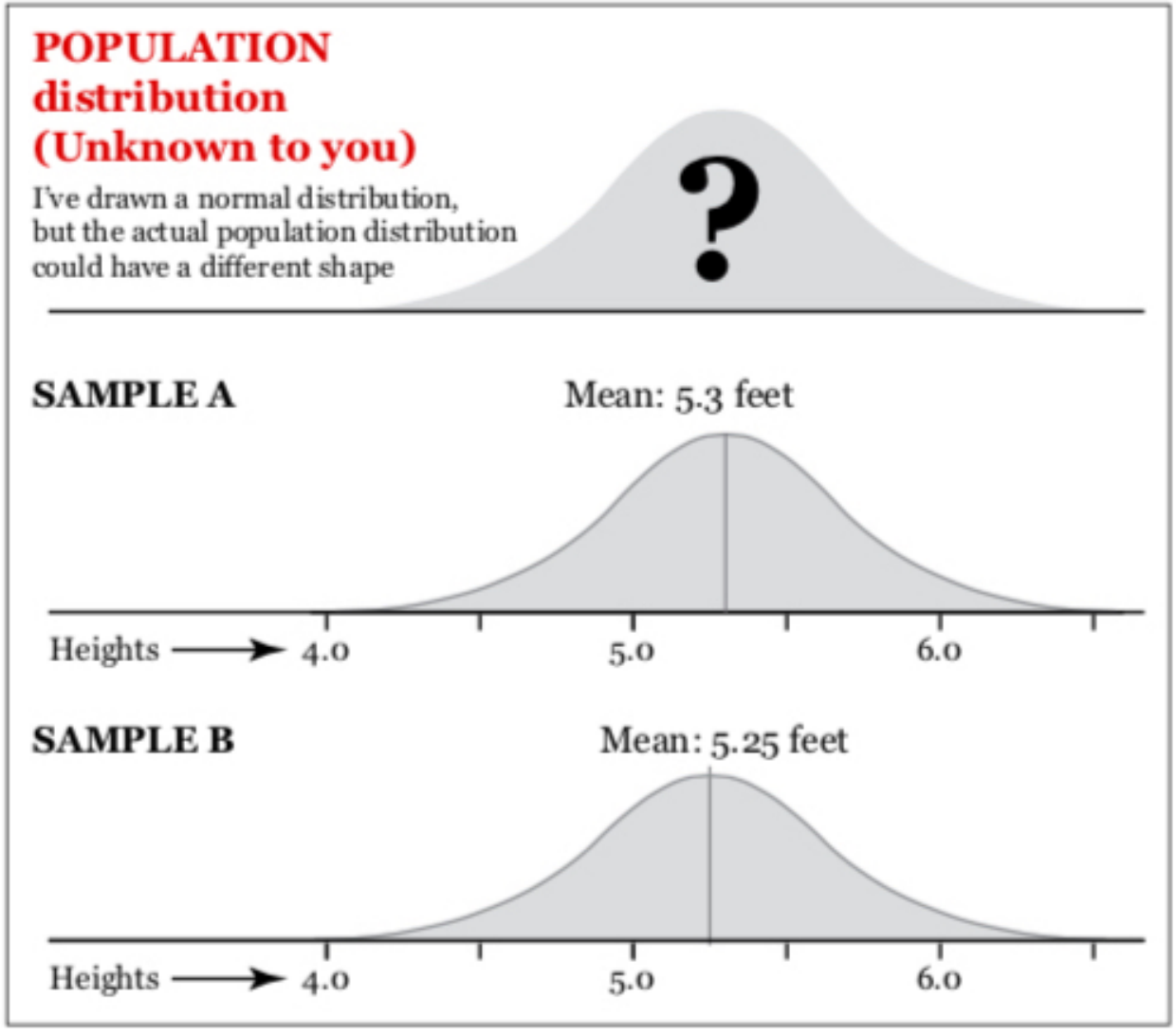


Figure 11.2 Population and samples.

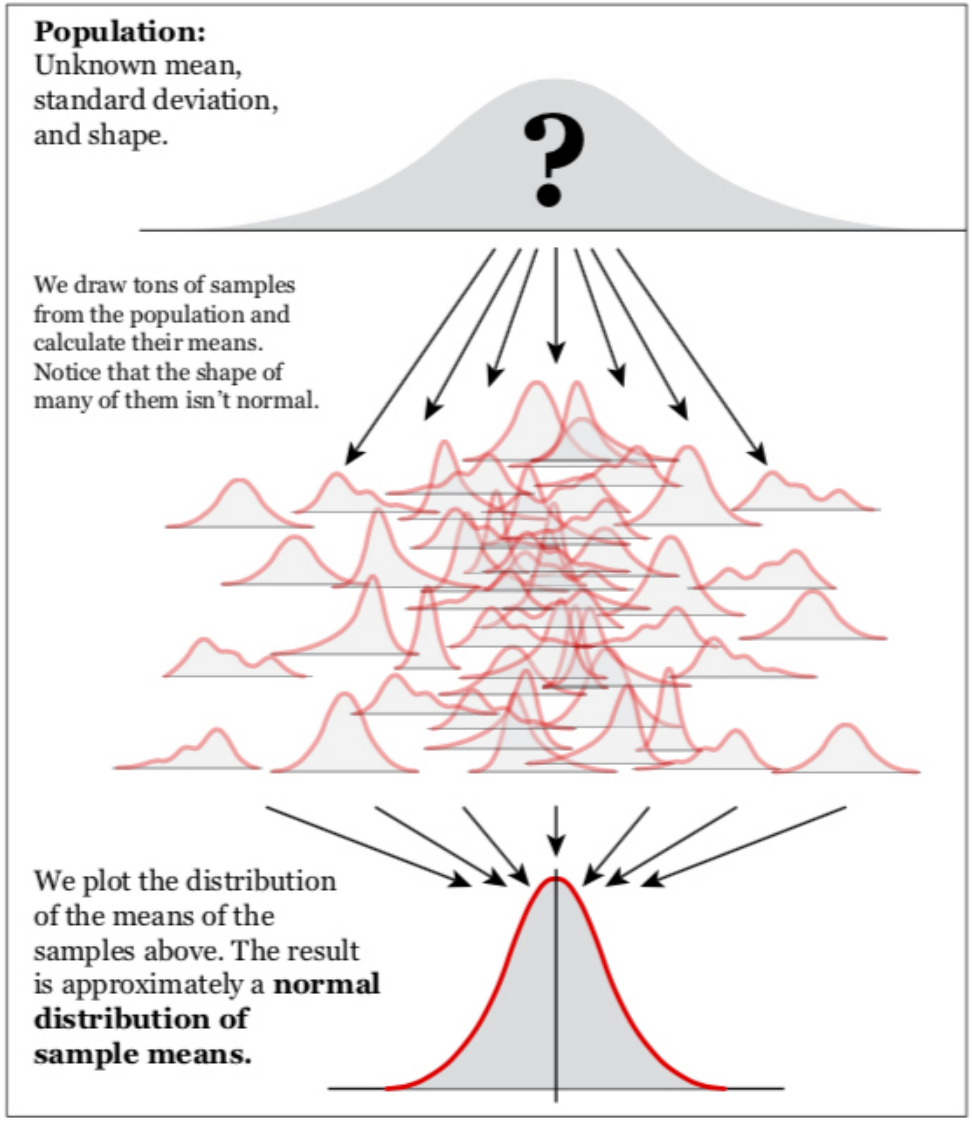
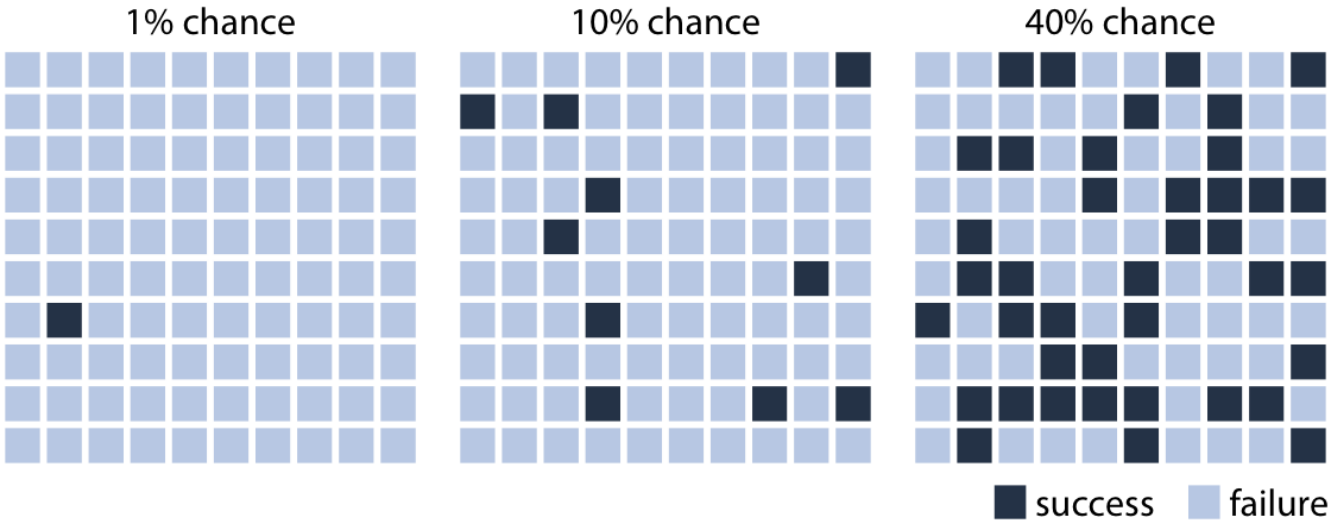
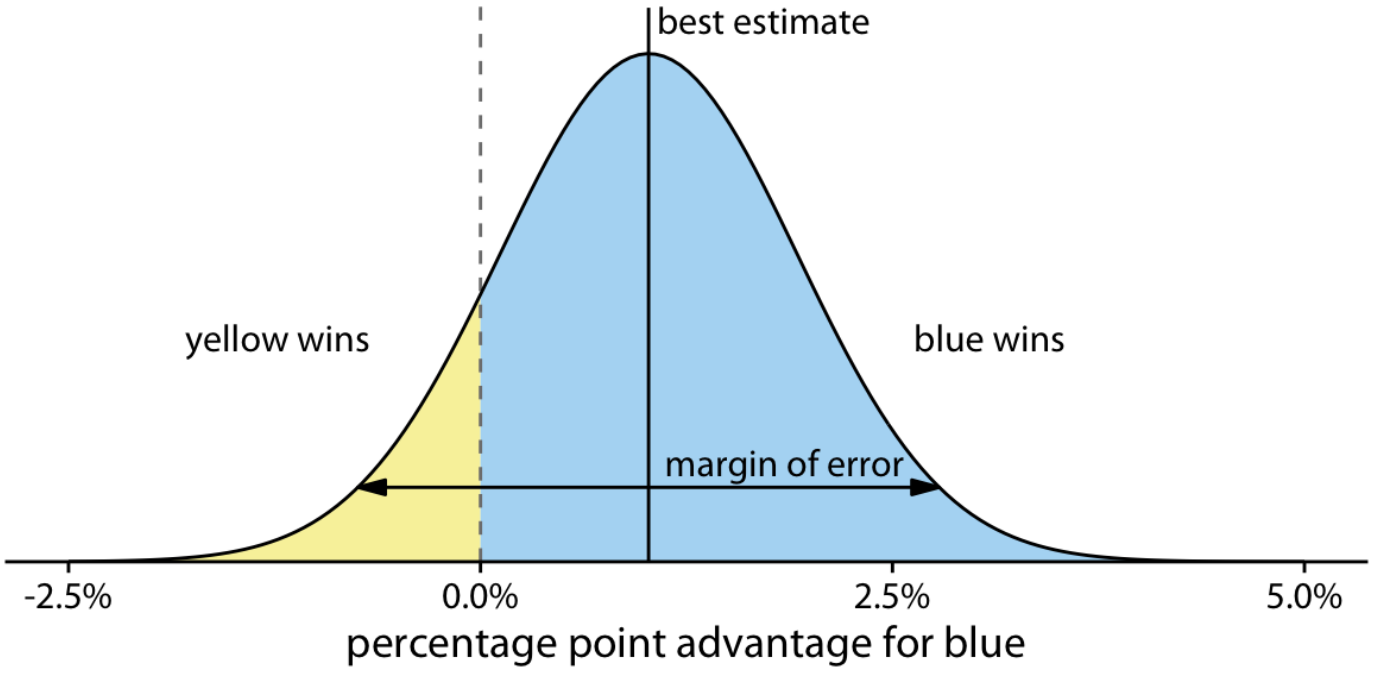


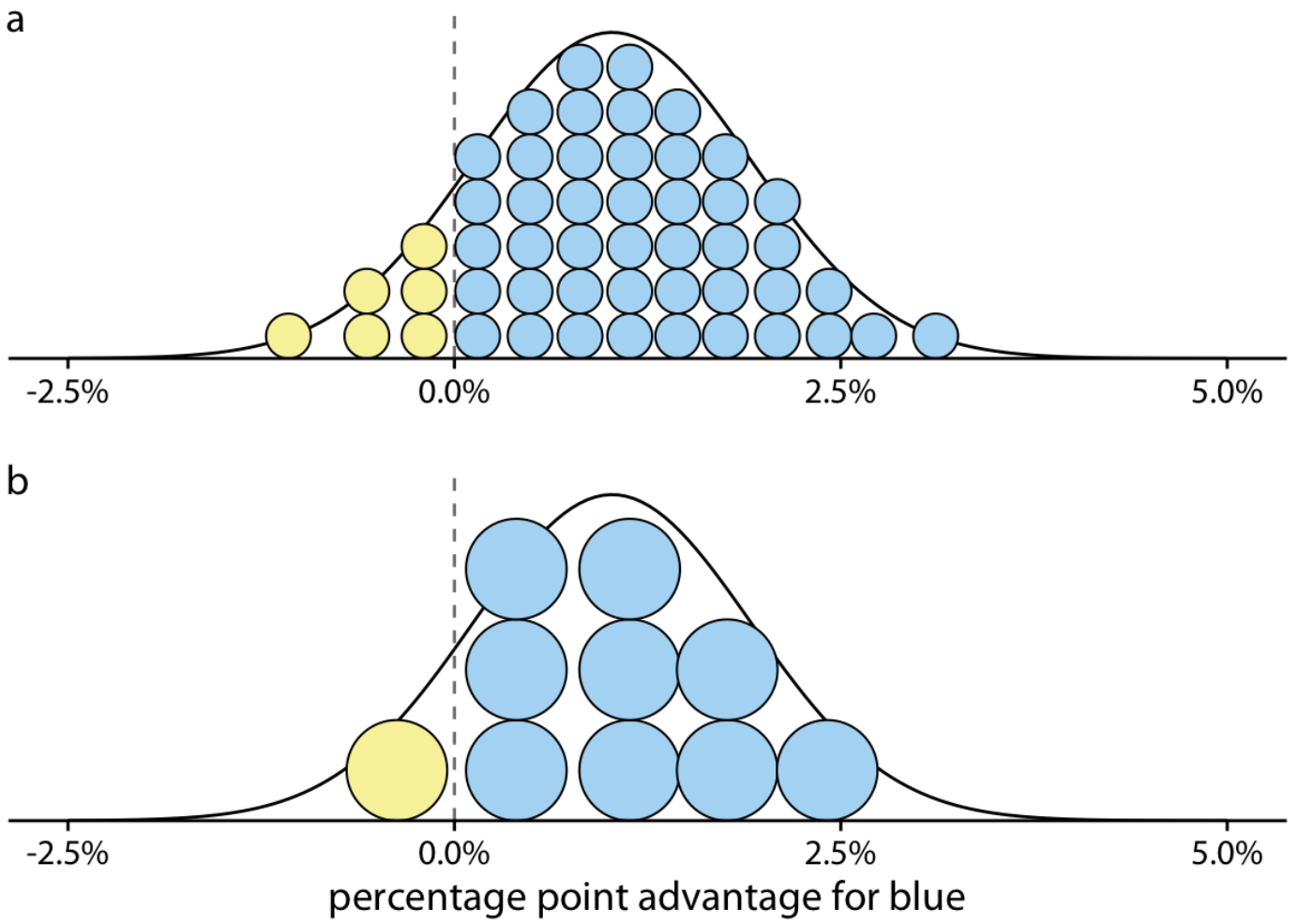
Figure 11.5 The distribution of (imaginary) sample means.

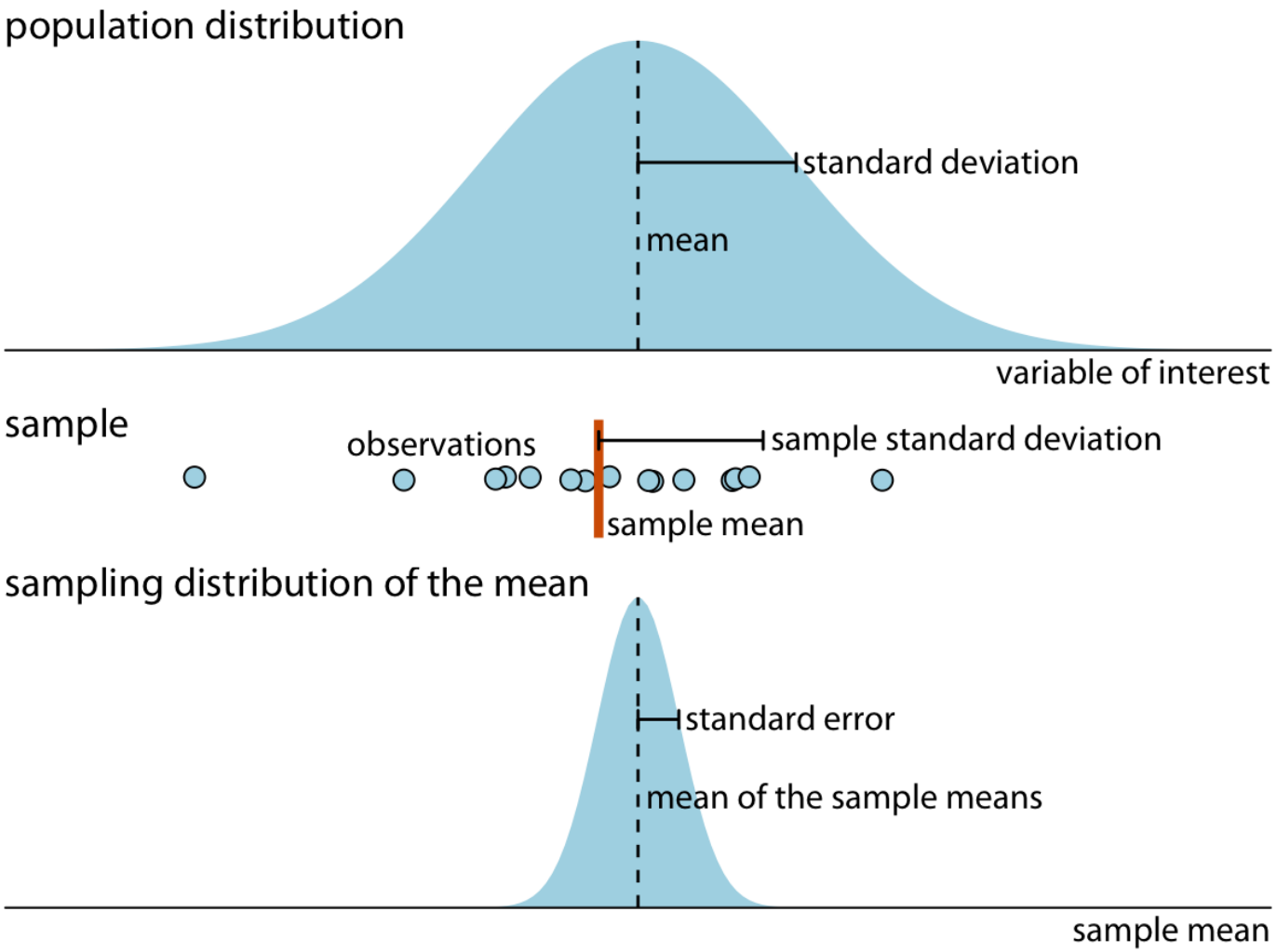
Why is visualizing uncertainty hard?

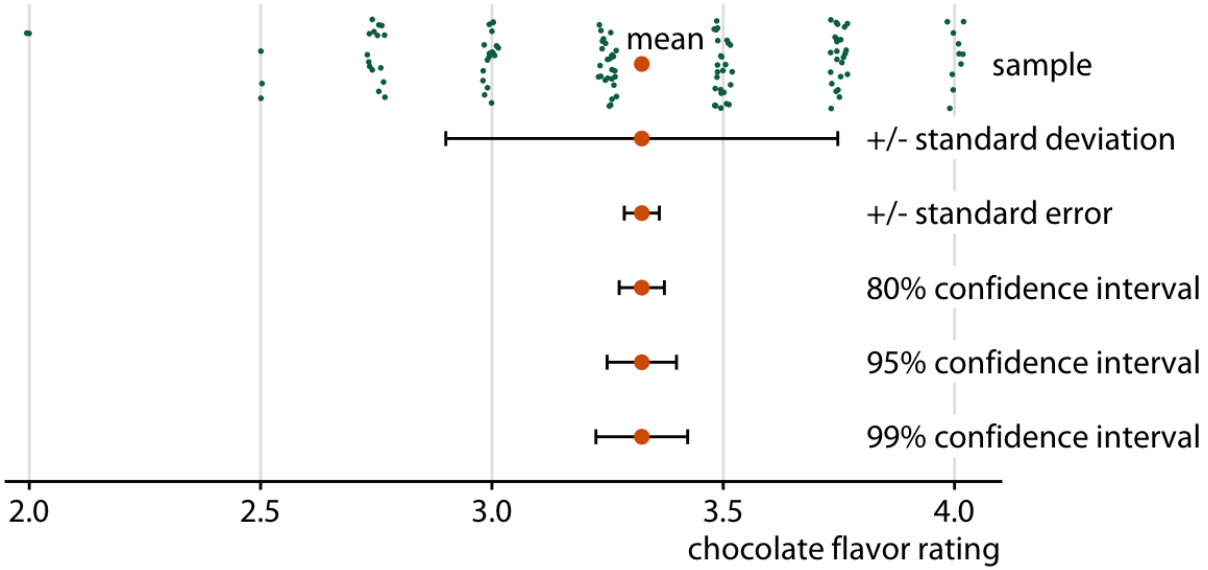
- Efficient encodings for uncertainty can be hard to find.
- Make sure people understand encodings (what does the plot mean?).
- Perceptual models of probability (e.g., quantile dot plot, HOP).
- Decisions under uncertainty (e.g., Gigerenzer et al or Monty Hall problem).
- Findings may not apply in all contexts.
- Plus, you still have to actually build it!

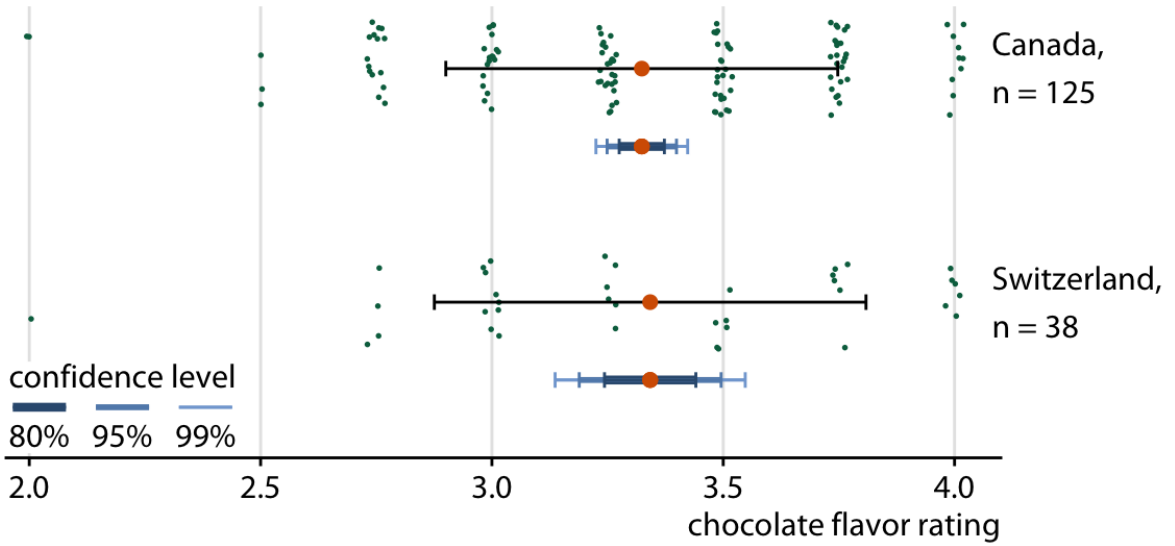

1% chance
10% chance
40% chance
success
failure


best estimate
yellow wins
blue wins
margin of error
-2.5%
0.0%
2.5%
5.0%
percentage point advantage for blue


a
-2.5%
0.0%
2.5%
5.0%
b
-2.5%
0.0%
2.5%
5.0%
percentage point advantage for blue


population distribution
standard deviation
mean
variable of interest
sample
observations
sample standard deviation
sample mean
sampling distribution of the mean
standard error
mean of the sample means
sample mean


mean
sample
+/- standard deviation
+/- standard error
80% confidence interval
95% confidence interval
99% confidence interval
2.0
2.5
3.0
3.5
4.0
chocolate flavor rating


Canada,
n = 125
Switzerland,
n = 38
confidence level
80%
95%
99%
2.0
2.5
3.0
3.5
4.0
chocolate flavor rating