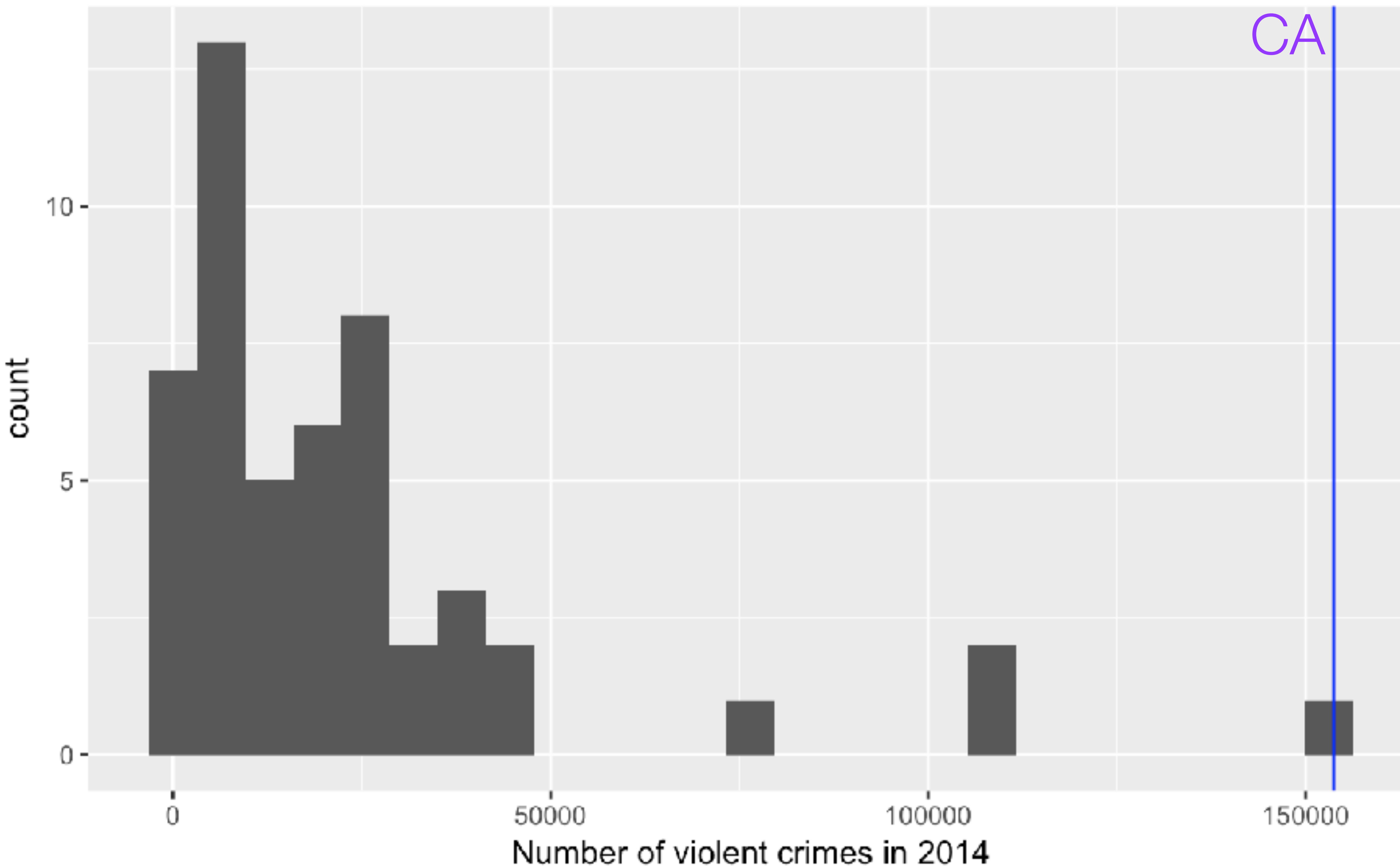
Session 11: Z-scores and Hypothesis testing review

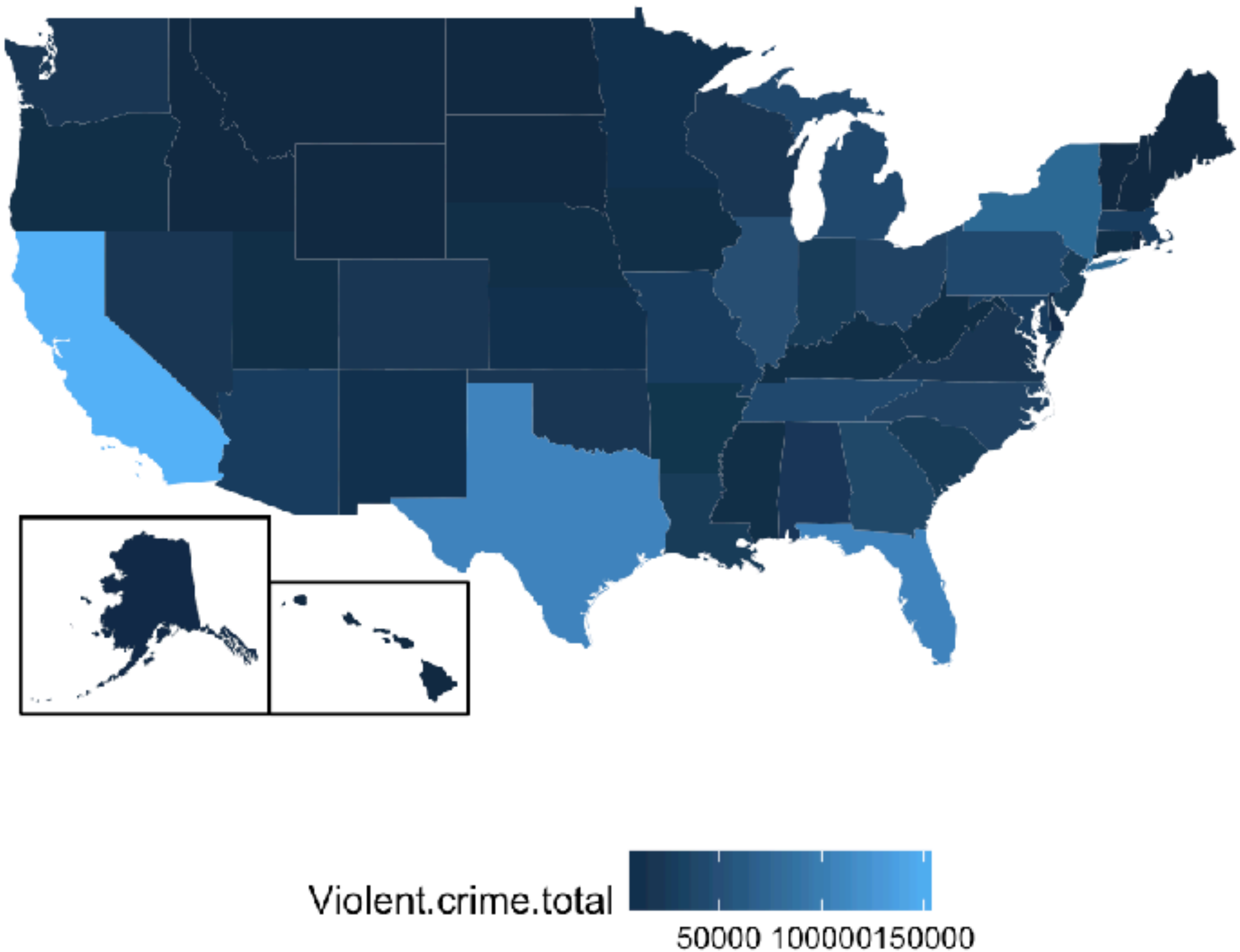
Stats 60/Psych 10
Ismael Lemhadri
Summer 2020

How safe is California compared to other states?

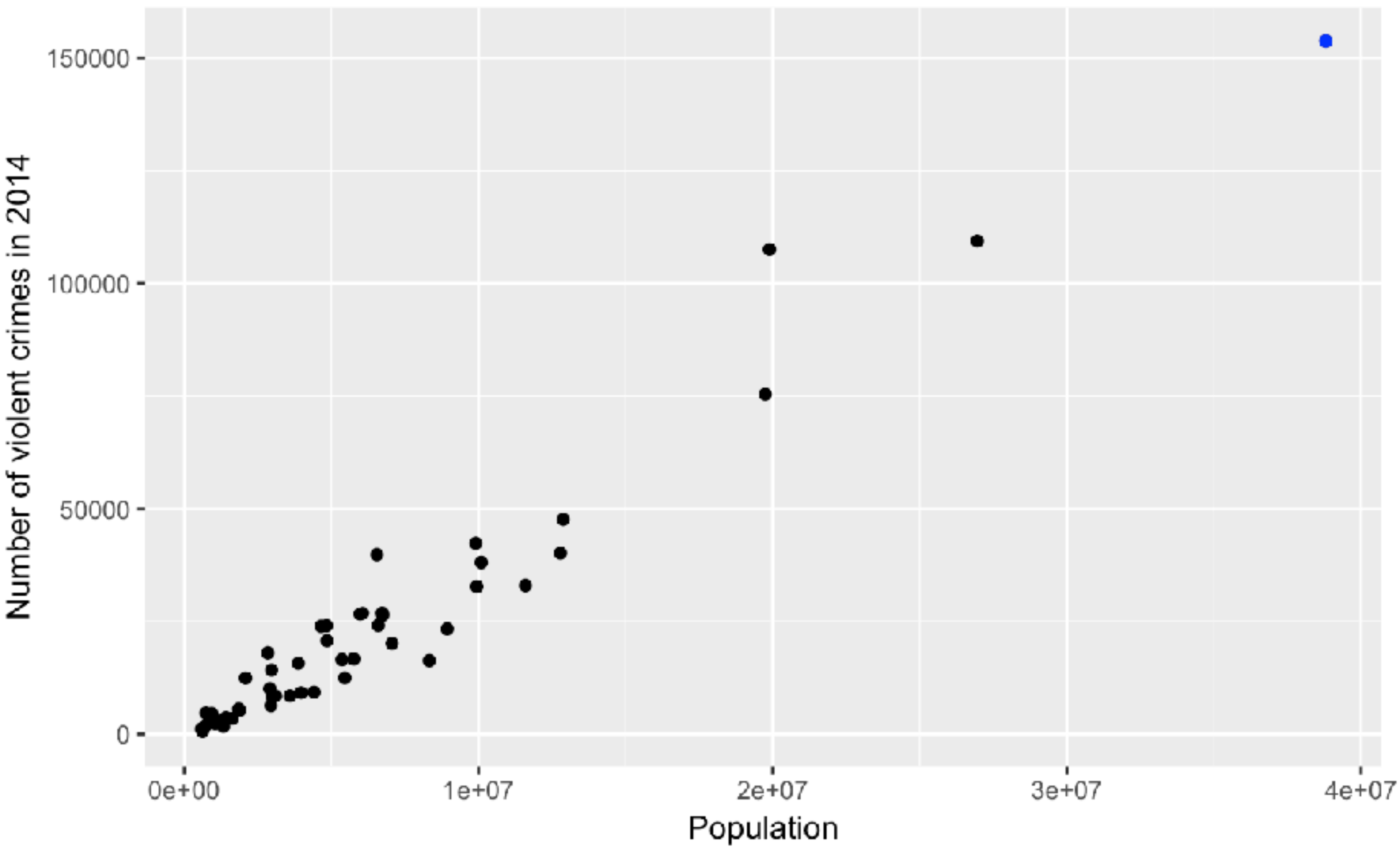
Whoa!


CA
count
10
5
0
0
50000
100000
150000
Number of violent crimes in 2014

Plotting geographical data in R



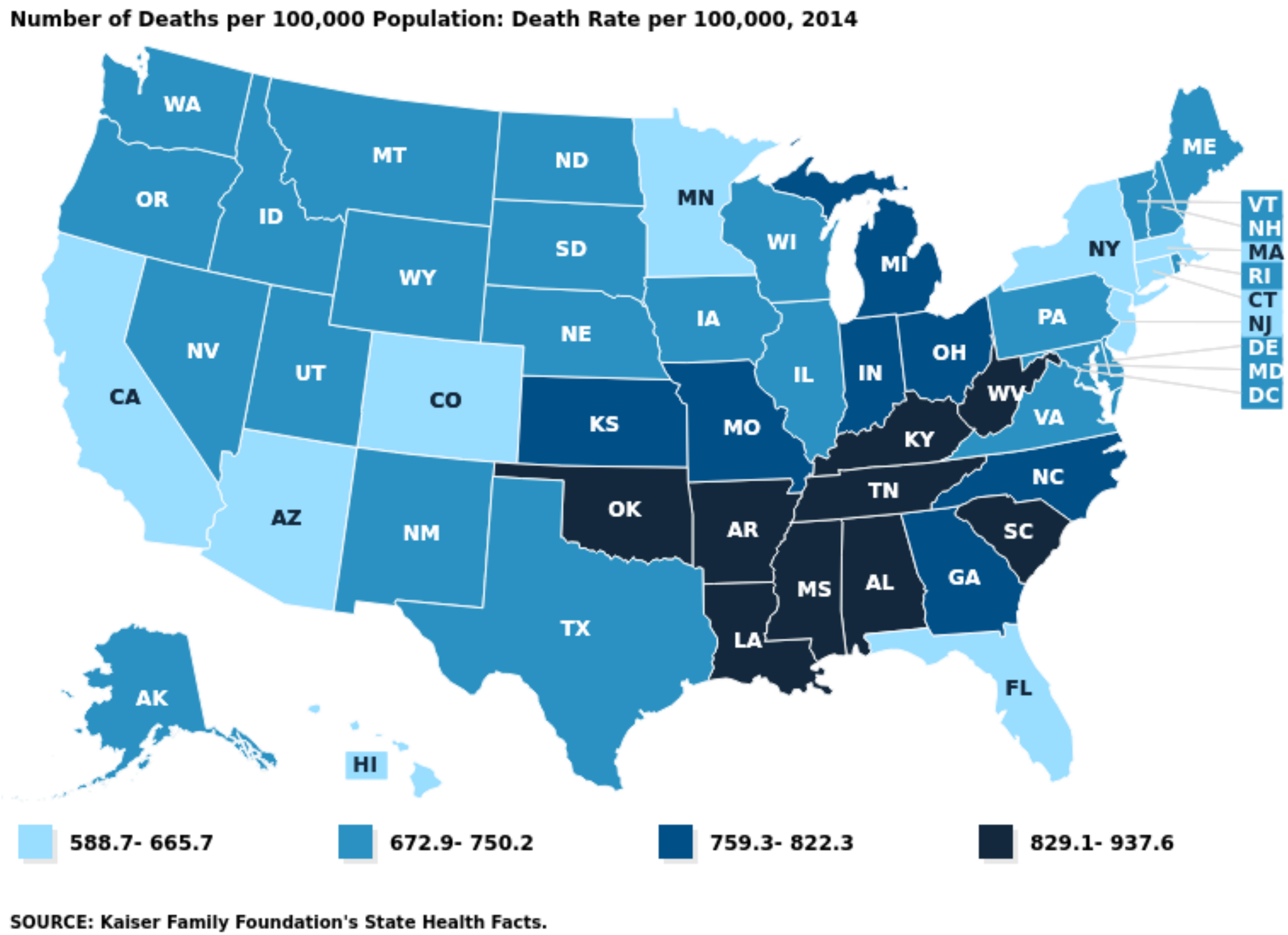
Number of crimes is closely related to population



Controlling for population differences

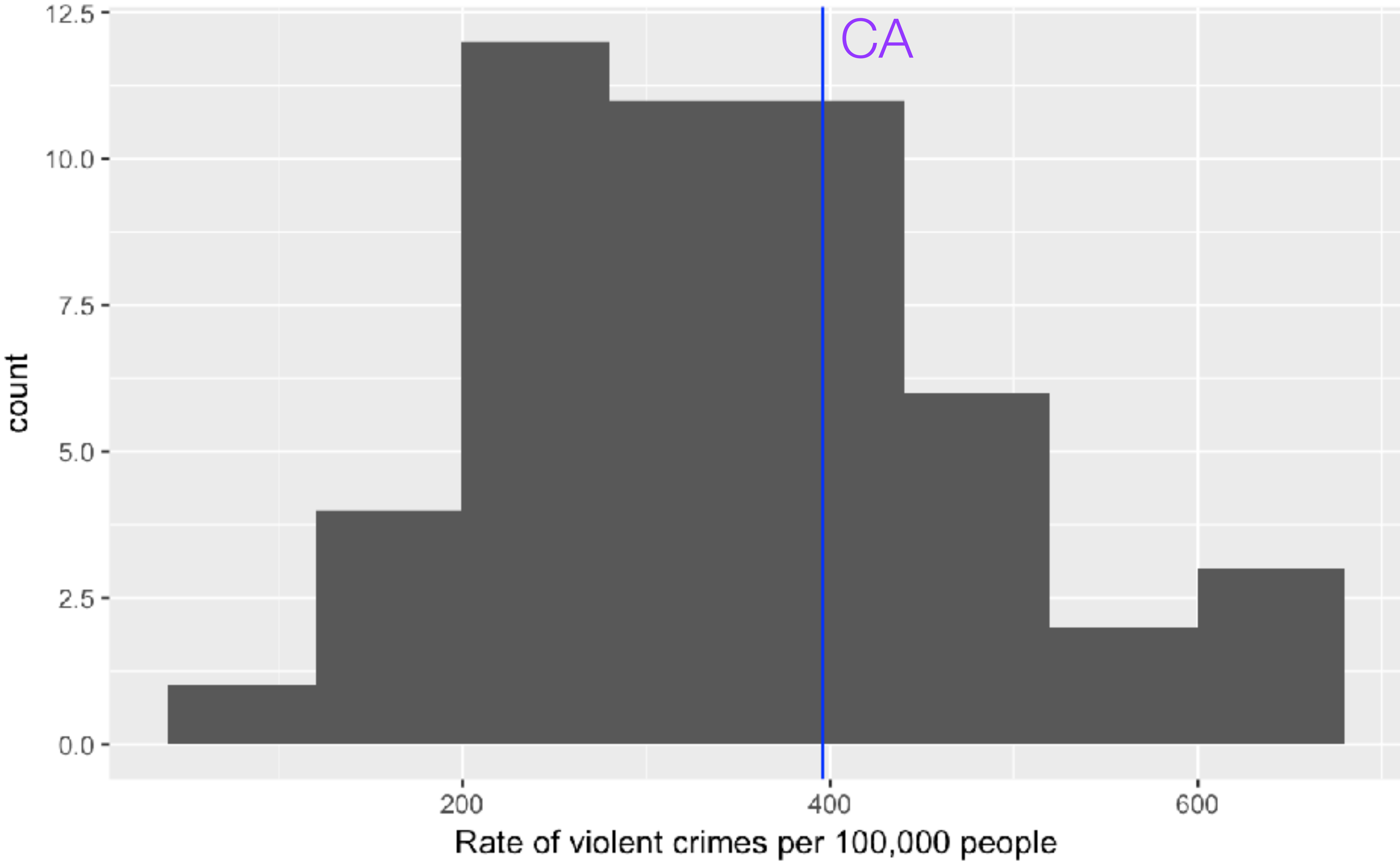
$$rate\, per\, 100,000 = 100,000 * \frac{number}{population}$$

Example: Deaths per 100,000 in 2014



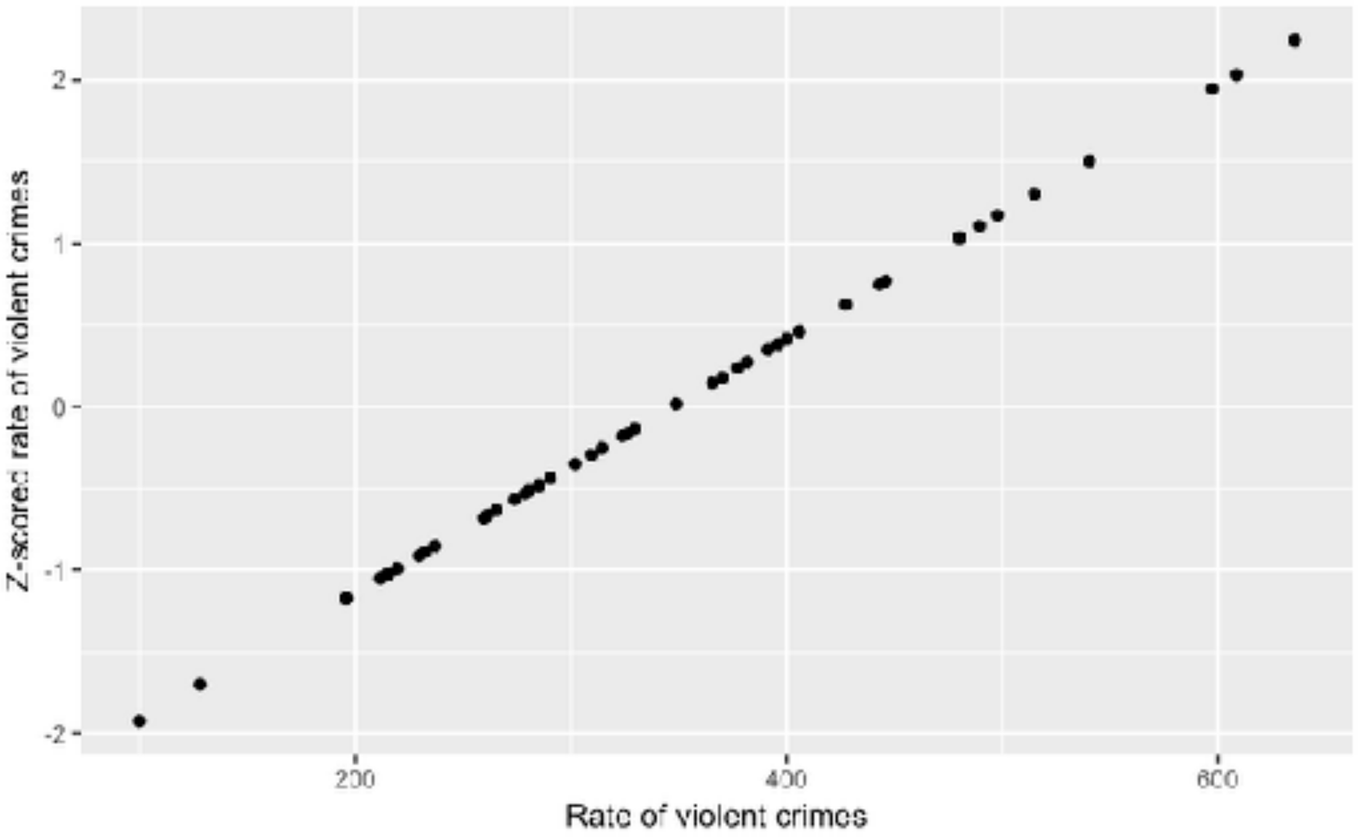
https://www.kff.org/statedata/

CA doesn't seem especially dangerous after all



Computing Z-scores

$$Z(X) = \frac{X - \mu}{\sigma}$$



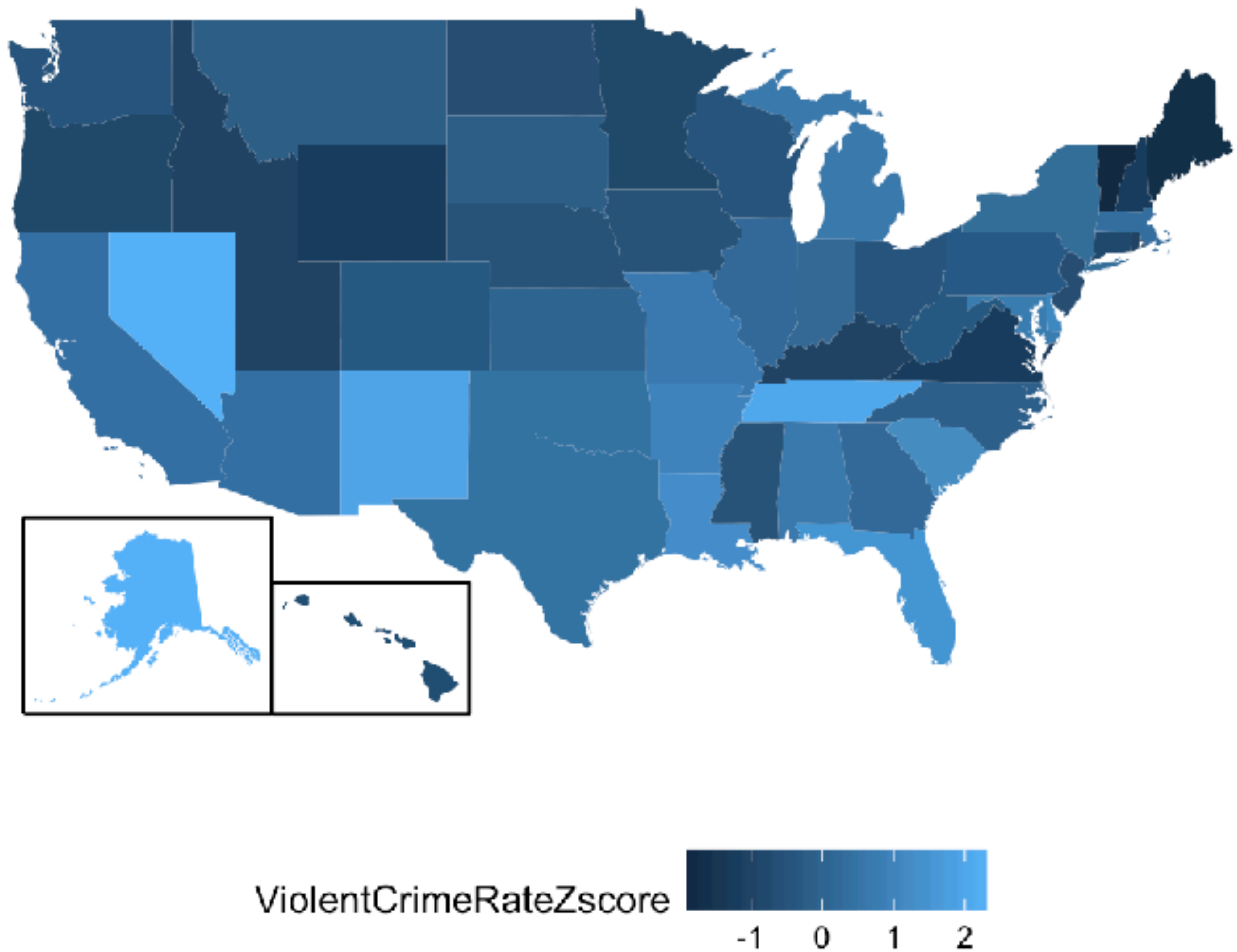
mean rate: 346.8
std of rate: 128.8
mean of Z-scored data: 1.46e-16 why isn't this zero?
std deviation of Z-scored data: 1

Numerical precision

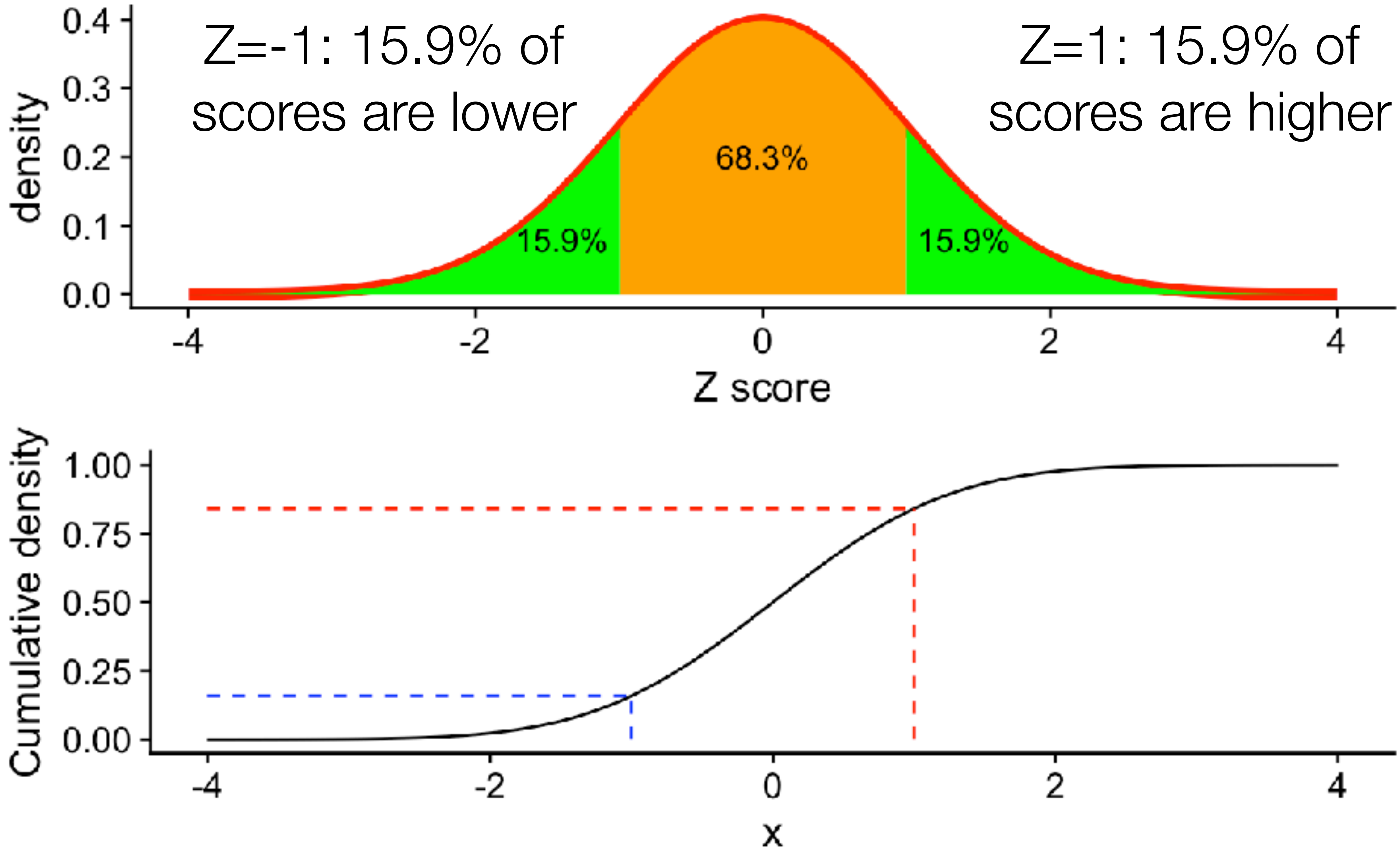
- Computers represent numbers with a given degree of precision
  - "Floating point"
    - The decimal point "floats"
    - 100 = 1X10^2 (usually abbreviated 1e2)
    - 200000 = 2e5
    - 0.01 = 1e-2
    - 0.000043 = 4.3e-5

```
## [1] "smallest number such that 1+x != 1 2.22044604925031e-16"

(1+.Machine$double.eps)==1
## [1] FALSE

(1+.Machine$double.eps/2)==1
## [1] TRUE


## [1] "largest number 1.79769313486232e+308"
```

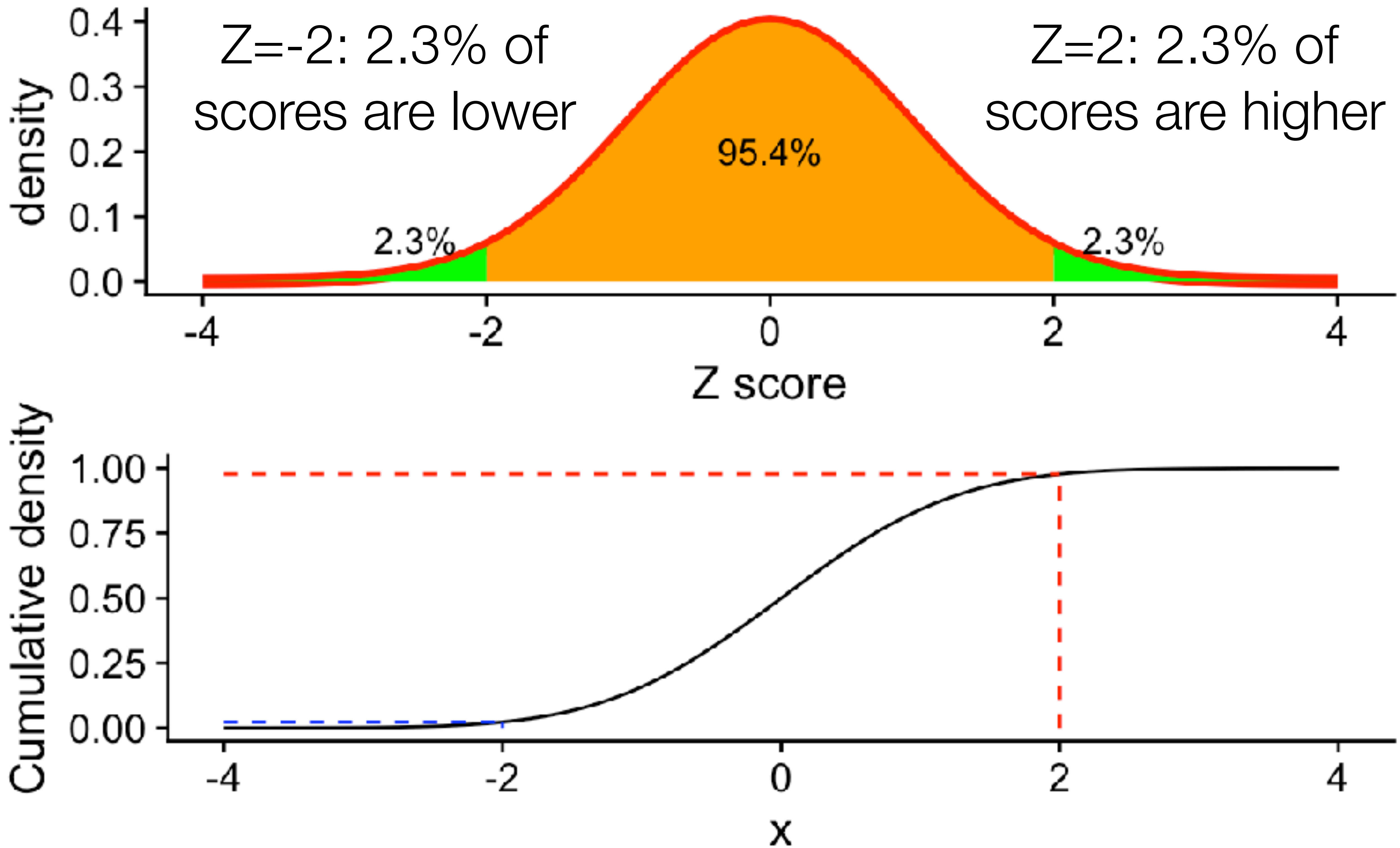
Map of violent crime rate Z-scores

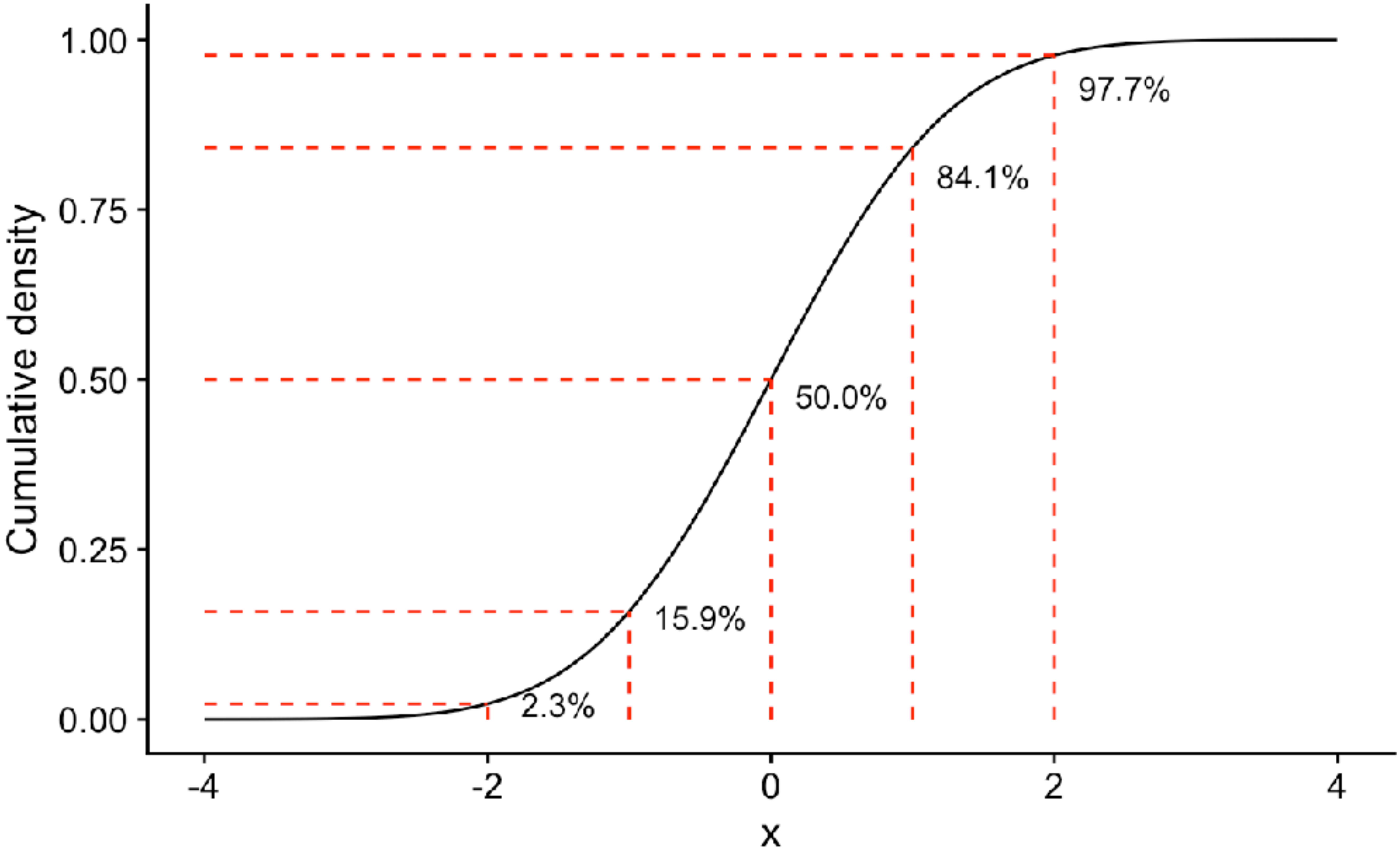


Interpreting Z-scores



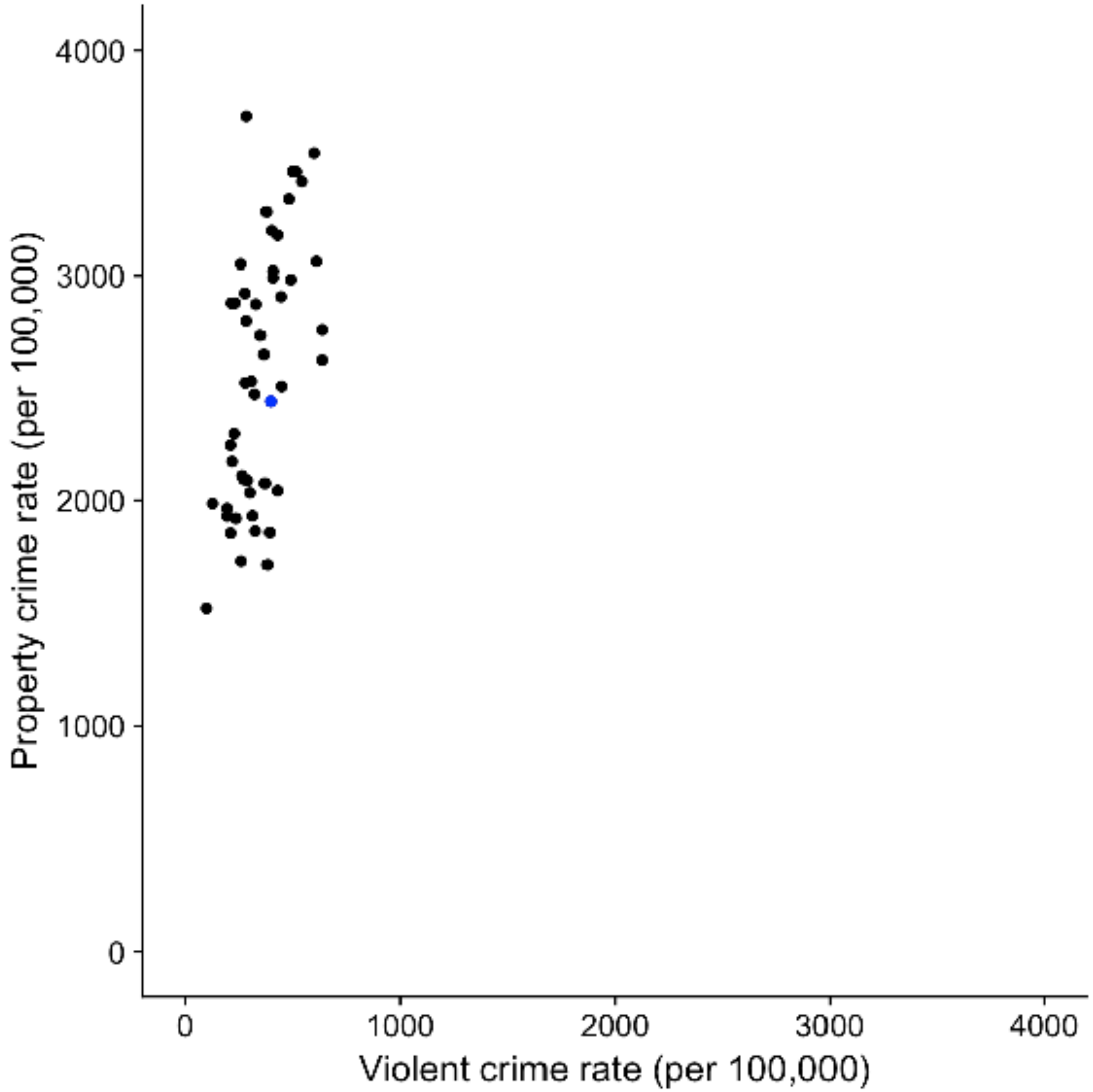
Interpreting Z-scores



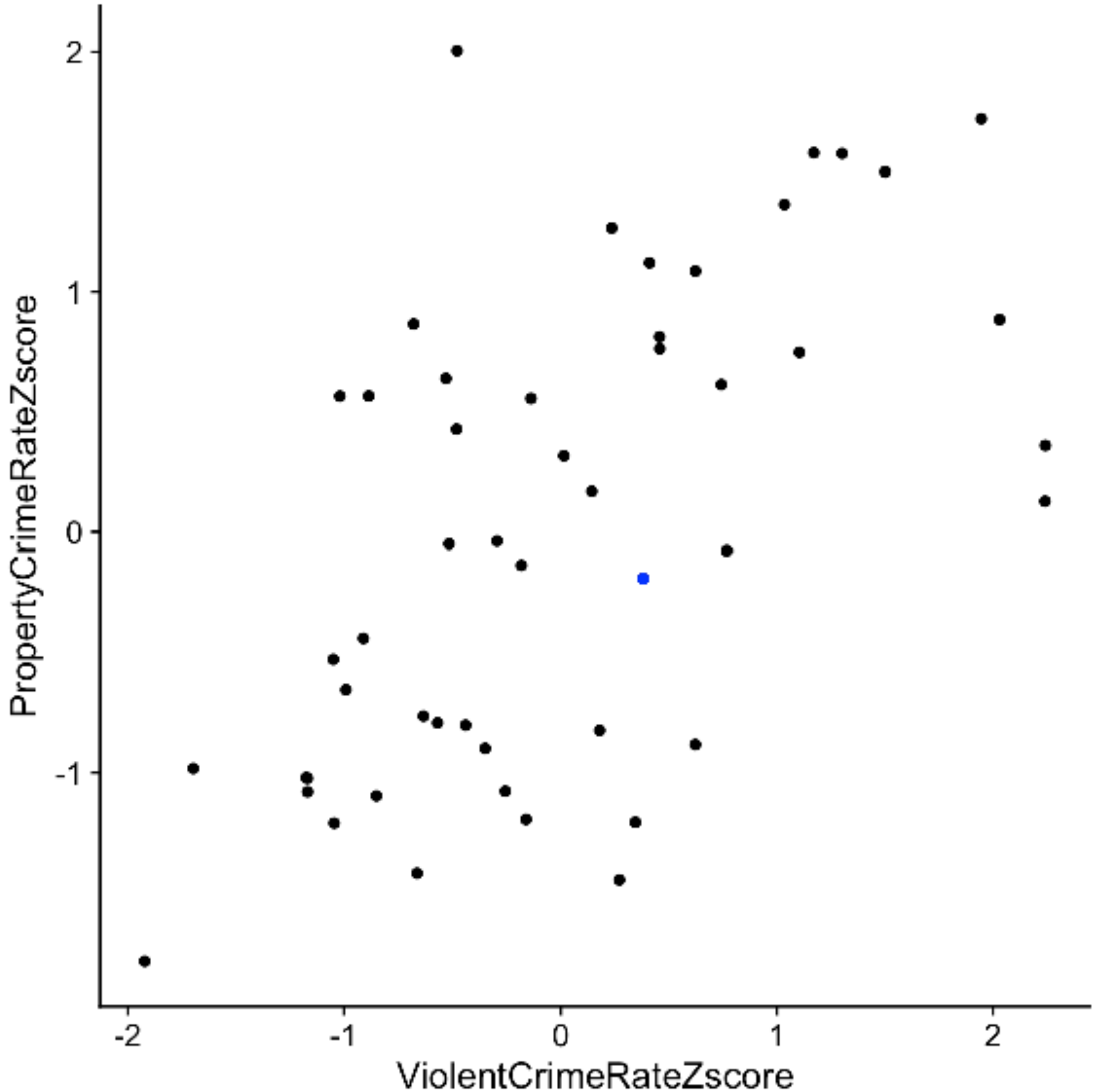


Cumulative density
1.00
0.75
0.50
0.25
0.00
97.7%
84.1%
50.0%
15.9%
2.3%
-4
-2
0
2
4
x

Comparing distributions

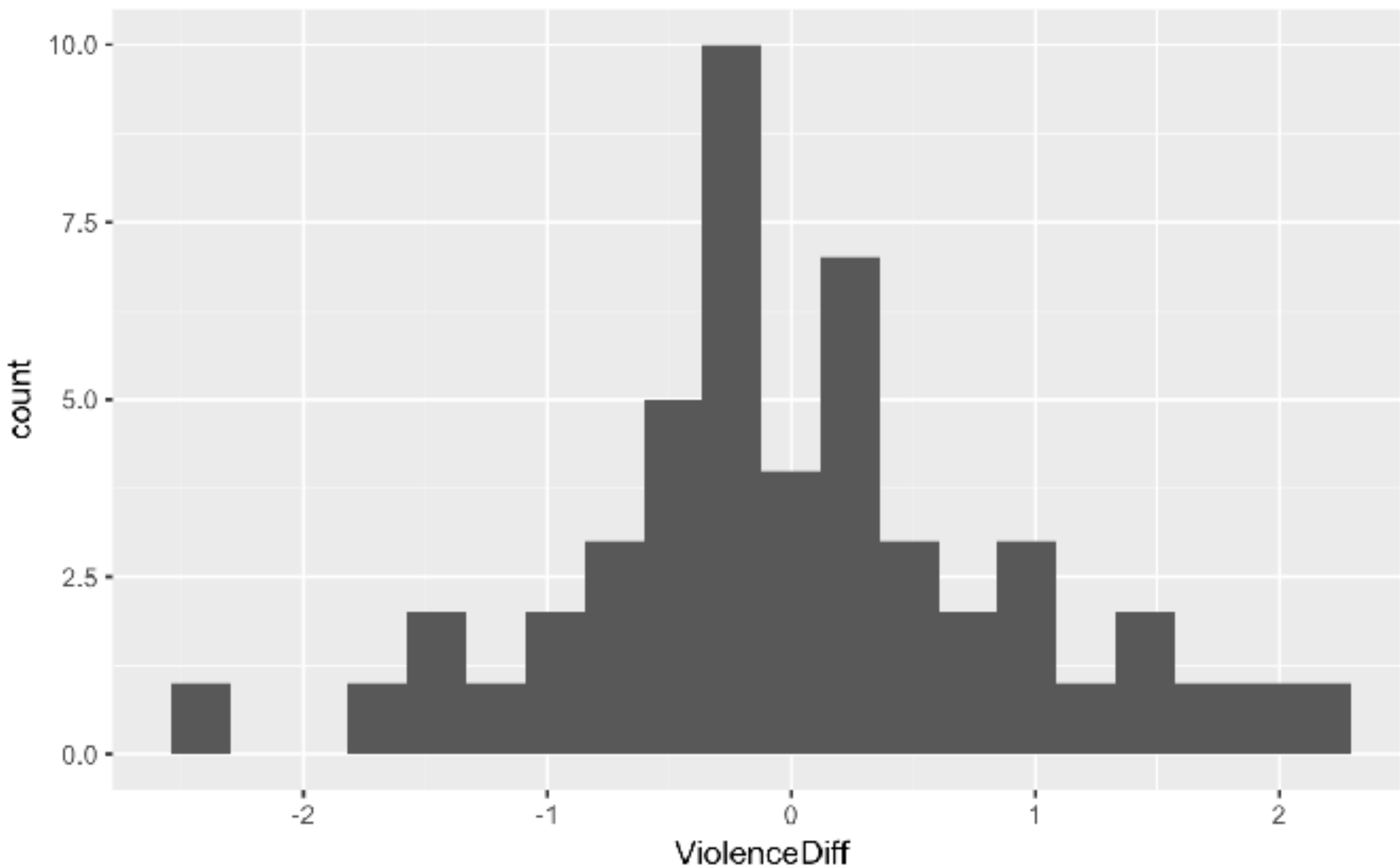


Z-scores allow direct comparison of different variables

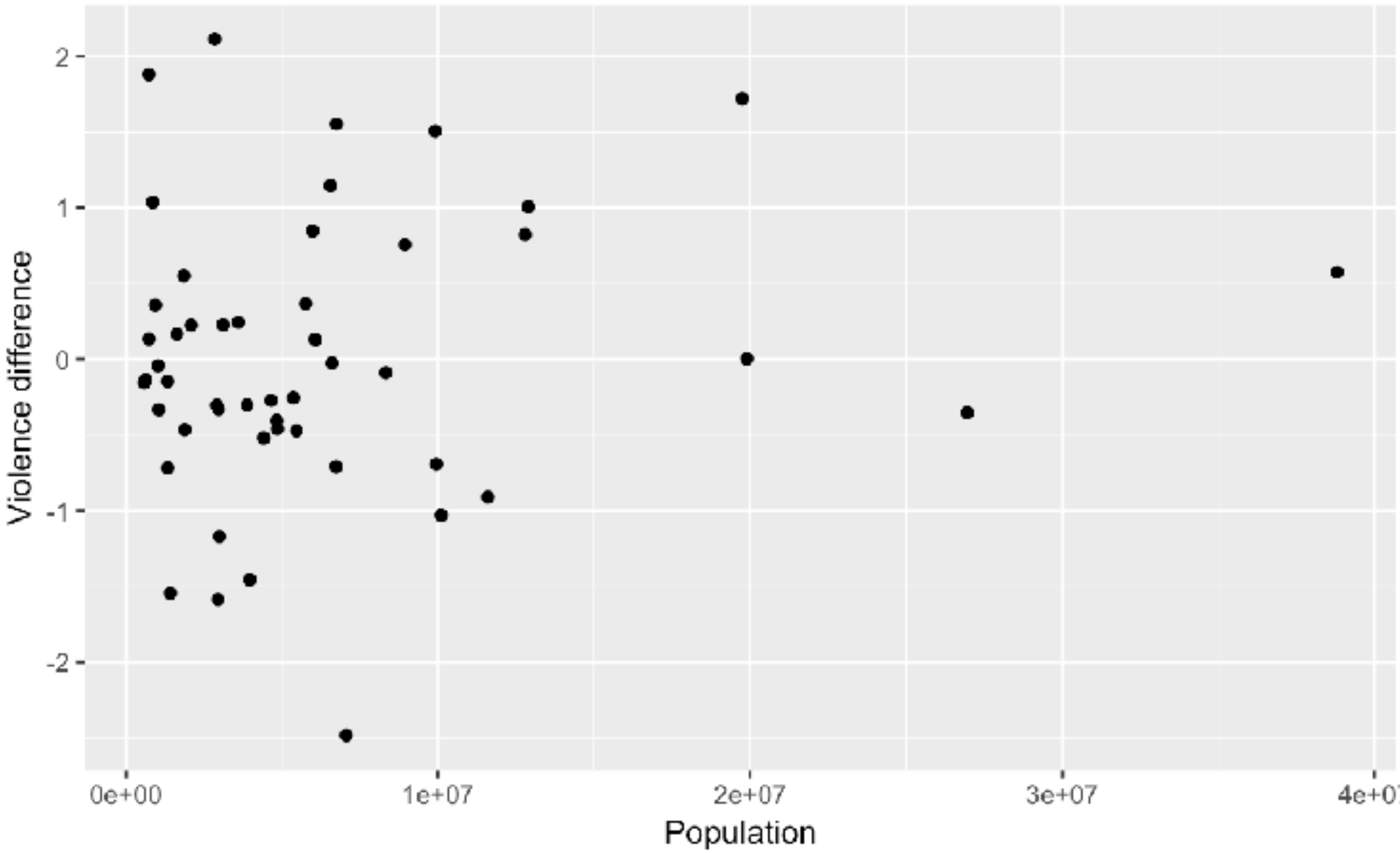


Comparing distributions directly using Z-scores

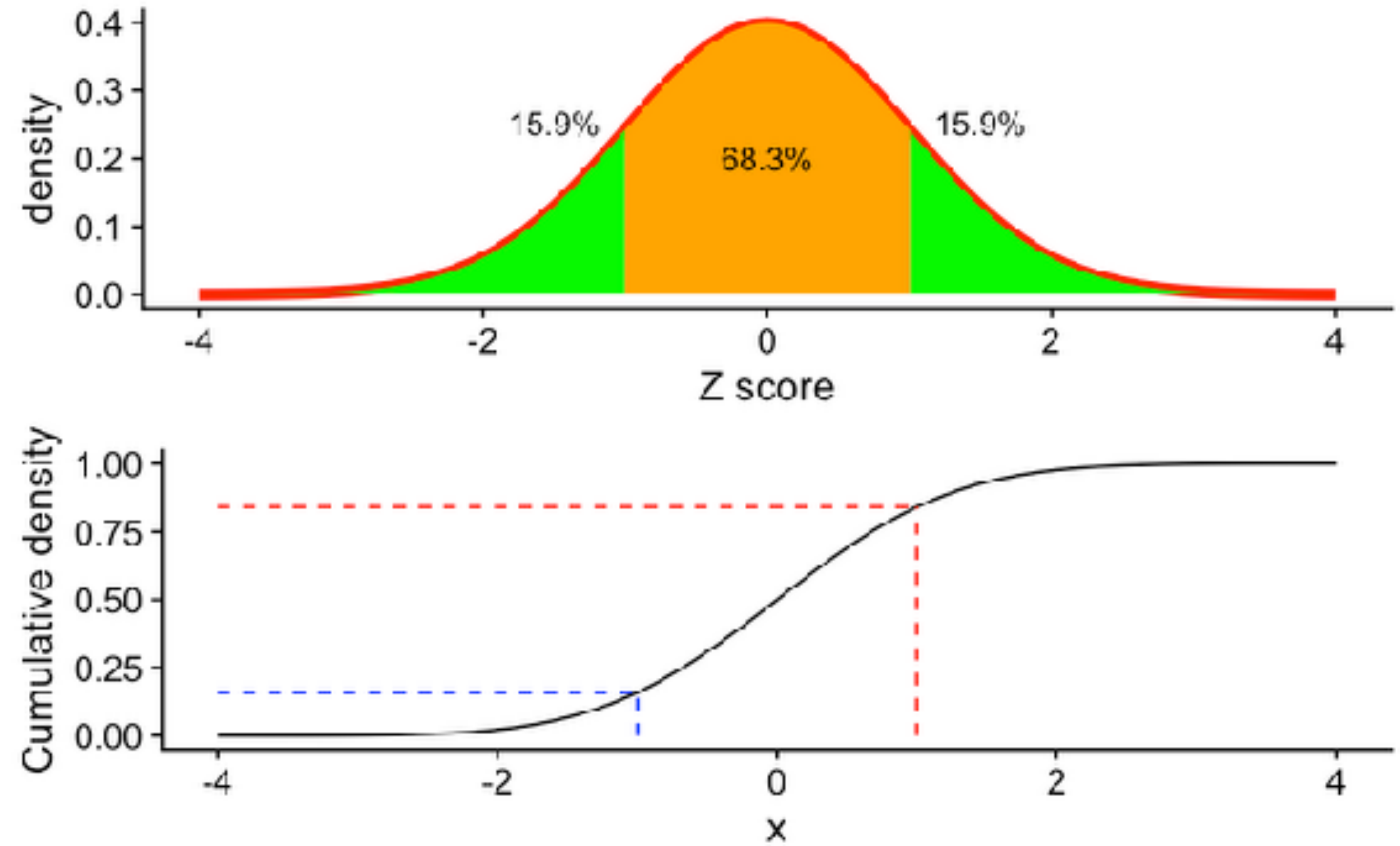
Violence difference = Z(violent crime) - Z(property crime)



- Why would smaller states have the most extreme positive *and* negative violence differences?

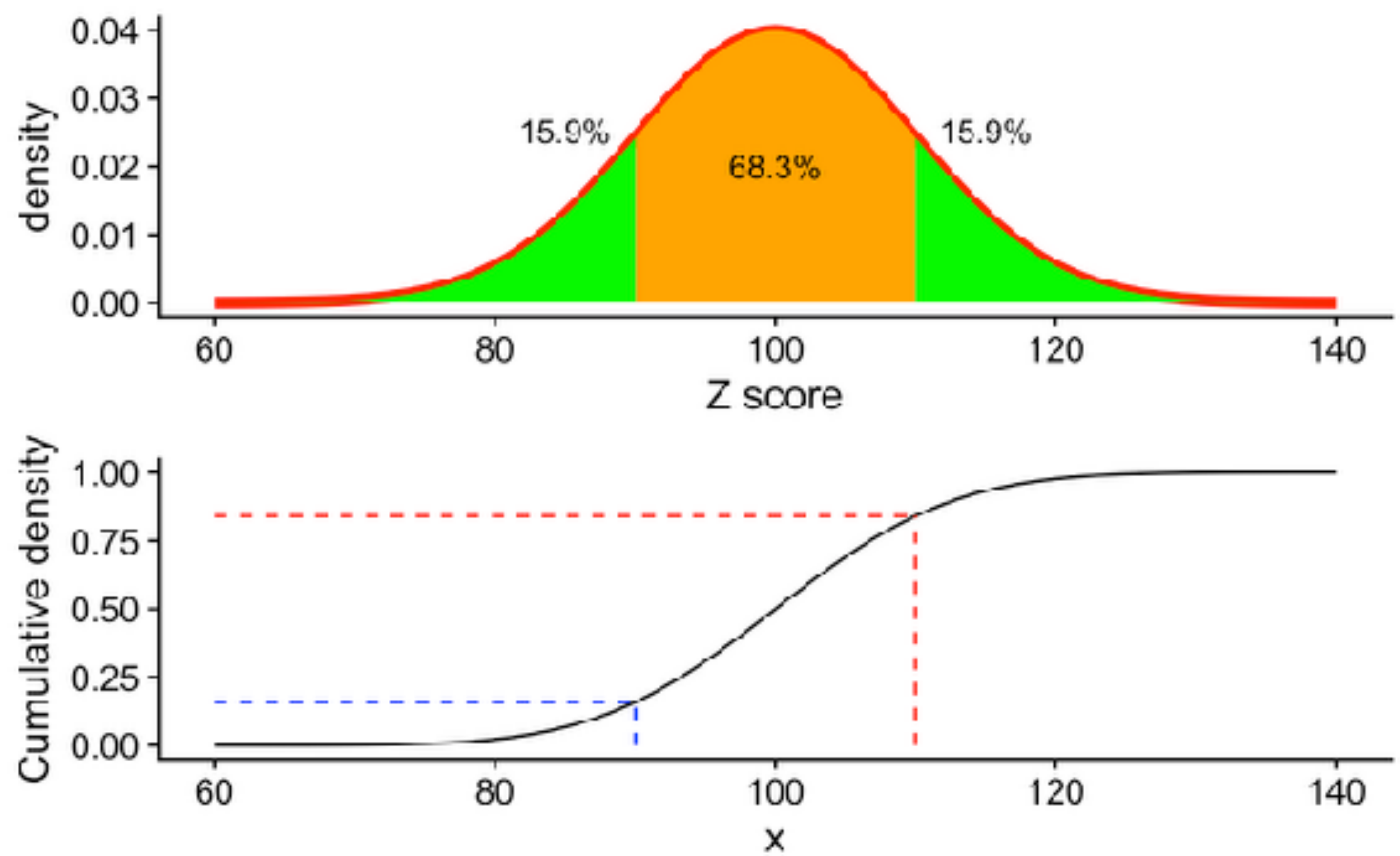


Standardized scores



$$StdScore(Z) =$$
$$Z * SD + mean$$

Example: IQ
mean=100
SD=10



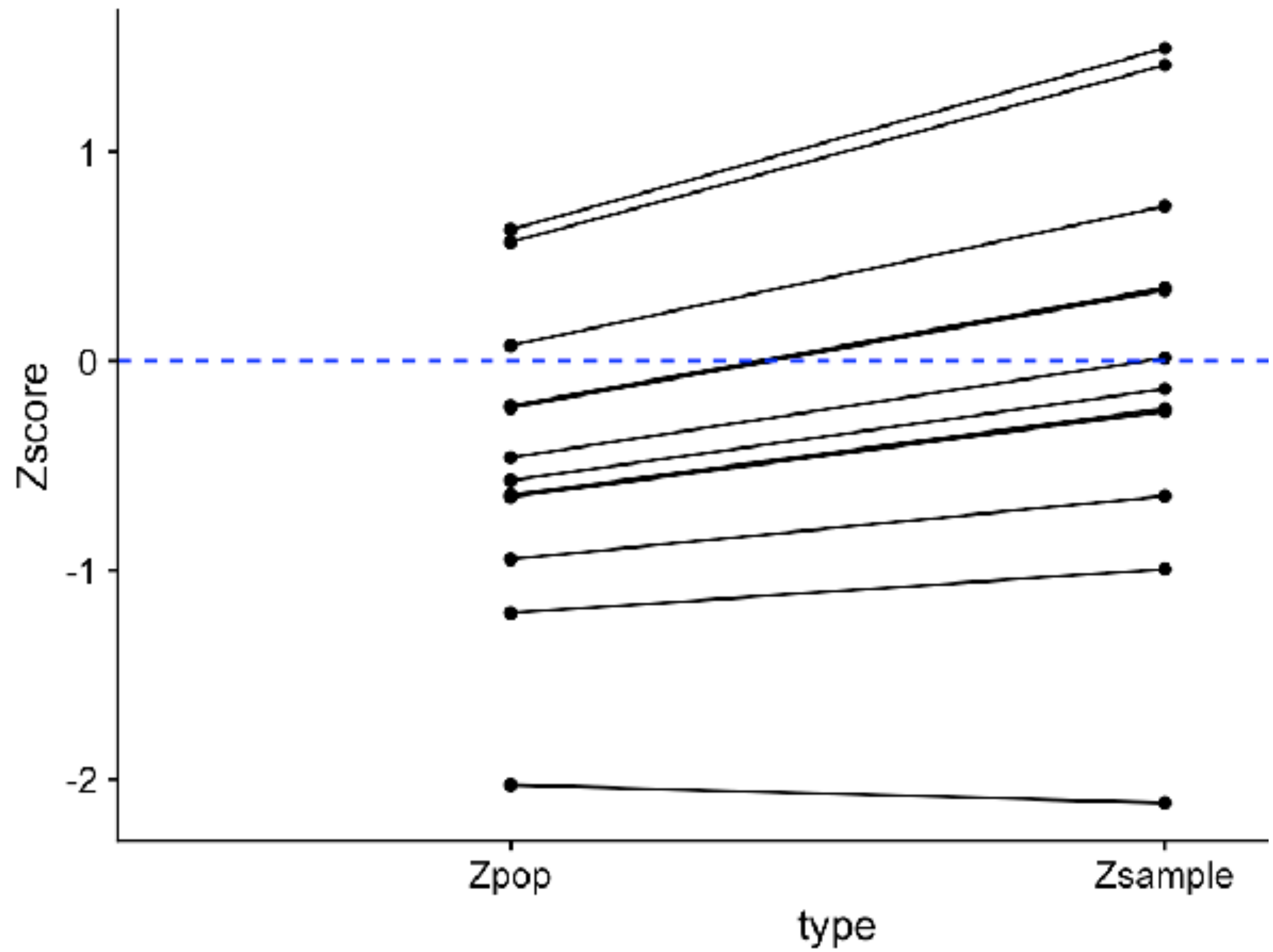
Population vs. sample Z-scores

- Sometimes we know the population parameters (mean and SD), so we can use those to create Z-scores
- But often we have to use the statistics estimated from a sample
- these can give different results!

Z-scores for NHANES Height

- Sampled 12 individuals from the NHANES population

	Mean	SD
Sample	164.09	7.44
Population	168.86	10.11



Using Z-scores in R

```
df <- tibble(raw=c(3,5,5,7,8,12,14,15))
df <- df %>%
 mutate(zscore=(raw - mean(raw))/sd(raw))
df
```

raw <dbl>	zscore <dbl>
3	-1.2494492
5	-0.8052006
5	-0.8052006
7	-0.3609520
8	-0.1388277
12	0.7496695
14	1.1939181
15	1.4160424

Recap: Z-scores

- Z-scores provide a way to standardize different types of data
- They don’t change the distribution
- They allow us to directly compare between different distributions

Hypothesis testing: A walkthrough

- We want to determine whether people in San Francisco are happier than people in Boston
  - We think that people in SF are happier.
- Let's say we have a biological test for a hormone that is related to happiness.
- We measure the levels of the hormone in 5 people from each city.

Data

	state	hormoneLevel
1	CA	13
2	CA	13
3	CA	15
4	CA	11
5	CA	11
6	MA	7
7	MA	10
8	MA	10
9	MA	7
10	MA	12

$$mean_{CA} = \frac{13+13+15+11+11}{5} = 12.6$$

$$std_{CA} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - mean_{CA})^2}{N-1}} = 1.67$$

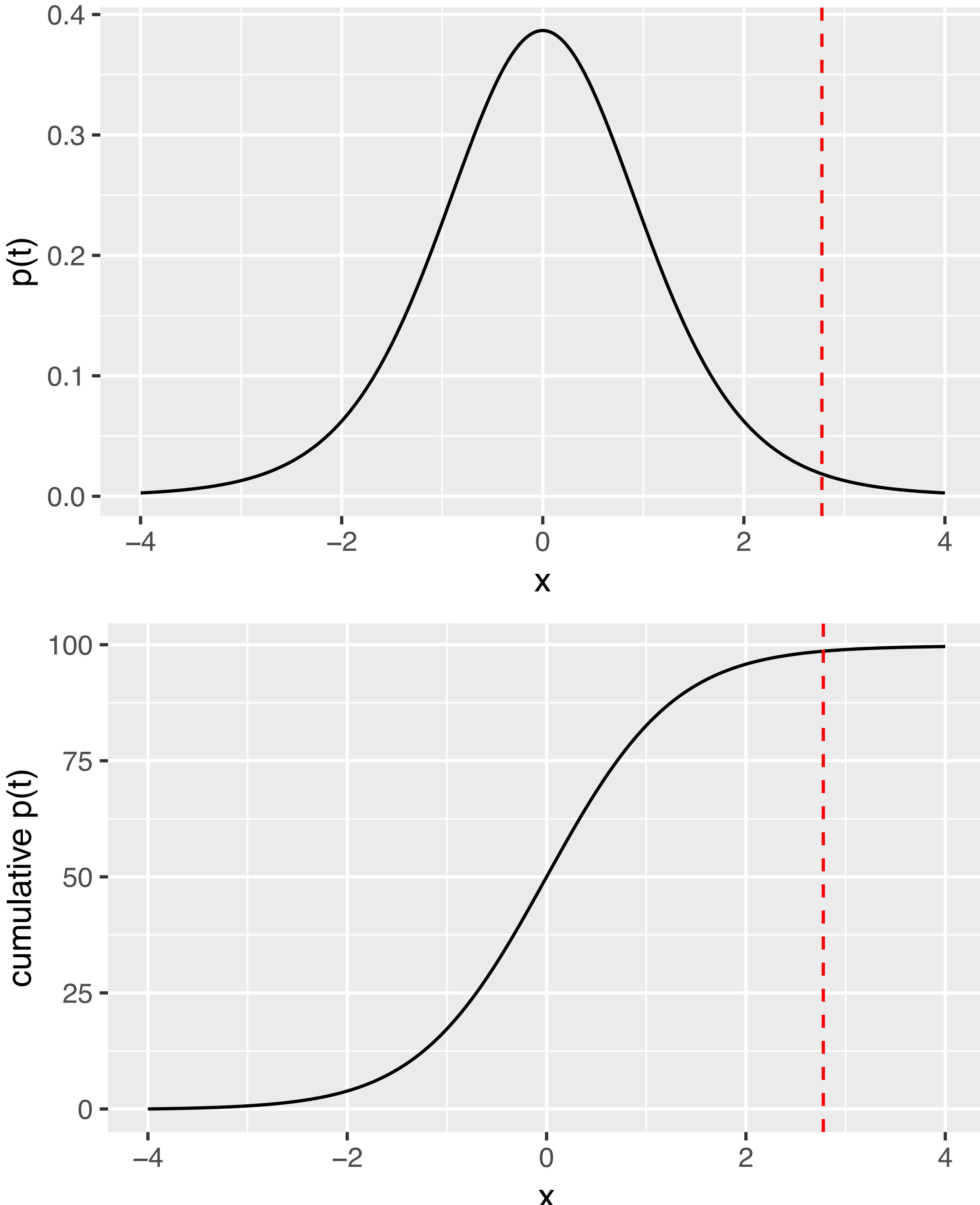
$$mean_{MA} = \frac{7+10+10+7+12}{5} = 9.2$$

$$std_{MA} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - mean_{MA})^2}{N-1}} = 2.17$$

$$t = \frac{mean_{CA} - meanMA}{\sqrt{\frac{std_{CA}^2}{5} + \frac{std_{MA}^2}{5}}} = 2.776$$

Data

	state	hormoneLevel
1	CA	13
2	CA	13
3	CA	15
4	CA	11
5	CA	11
6	MA	7
7	MA	10
8	MA	10
9	MA	7
10	MA	12



t = 2.776

Under null:

$$P(t_8 \geq 2.776 | H_O) = .012$$

BEYOND HYPOTHESIS TESTING

We would often like to know more than a yes/no decision about a hypothesis:

- How much uncertainty is there about the answer?
- Is the effect practically important in addition to be statistically significant?

CONFIDENCE INTERVALS

- An interval that will on average contain the true population parameter with a given probability
  - for example, the 95% confidence interval is an interval that will capture the true population parameter 95% of the time.
- this is *not* a statement about the population parameter; any particular confidence interval either does or does not contain the true parameter.

- As Jerzy Neyman, the inventor of the confidence interval, said: “The parameter is an unknown constant and no probability statement concerning its value may be made.”

ps Google Calendar -... SERA Axess PubMed Inbox - poldrack@... sklearn nipype Active Learning | ... Facebook FFIT Final Paper Final ...

Home Moments Notifi

22.1K Tweets

#FridayThoughts
23.3K Tweets

#FridayReads
1,626 Tweets

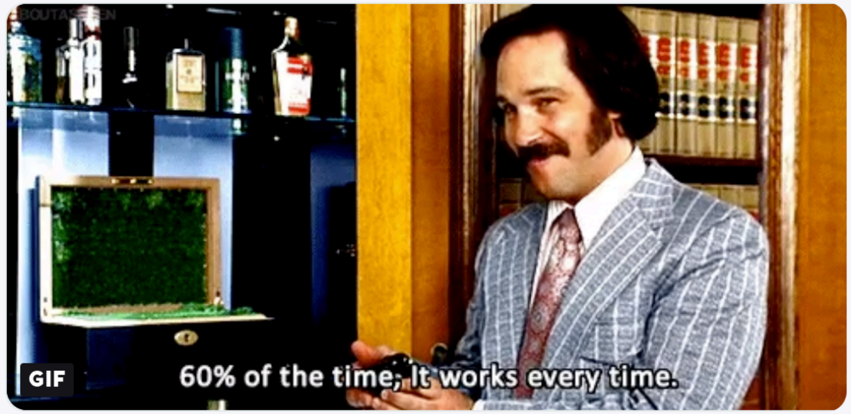
Forever Chemicals

Kraft Heinz
4,743 Tweets

Tweet

Will Gervais @wgervais · 2h

Does Anchorman have the pithiest description of how frequentist confidence intervals work of any modern film?



2 14

Will Gervais
@wgervais

Follow

🔥🔥🔥

d = .4, 95% CI [.25, .55]

- What does this mean?
- "Using a procedure that gets it right 95% of the time, we obtained this CI this time. We don't know if it's right or not, but we shall act as if it is."

EXAMPLE: WHAT IS THE MEAN WEIGHT OF ADULTS IN NHANES?

- Take a sample of 250 adults from NHANES

meanWeight	sdWeight	seWeight
82.77	22.27	1.408

COMPUTING CONFIDENCE INTERVALS

- The confidence interval for the mean is computed as:

$$CI = point\ estimate\ \pm critical\ value$$

- where the critical value is determined by the sampling distribution of the estimate.

CONFIDENCE INTERVALS USING THE NORMAL DISTRIBUTION

- If we know the population standard deviation, then we can use the normal distribution to compute a confidence interval.
  - We usually don't, but for our example of the NHANES dataset we do (it's 21.3 for weight).
- The critical value for 95% CI are the values of the standard normal distribution that capture 95% of the distribution
- these are simply the 2.5th percentile and the 97.5th percentile of the distribution, which we can compute using the `qnorm()` function in R, and come out to ± 1.96.

CONFIDENCE INTERVALS USING THE NORMAL DISTRIBUTION

The confidence interval for the mean ($\overline{X}$) is:

$$CI = \overline{X} \pm 1.96 * SE$$

where SE is the standard error:

$$SE = \frac{SD}{\sqrt{n}}$$

Using:

- the estimated mean from our sample (82.77)
- the known population standard deviation (21.3)

we can compute the confidence interval as:

$$CI = 82.76 \pm 1.96 * \frac{21.3}{\sqrt{250}}$$
$$= [80.13, 85.41]$$

CONFIDENCE INTERVALS USING THE T DISTRIBUTION

- In general we don't know the population SD
- the *t* distribution is more appropriate as a sampling distribution
- confidence intervals based on *t* will be slightly wider, due to extra uncertainty that arises for small samples.
- Instead of using the normal distribution to compute the percentiles, we use the distribution via `qt()` in R

$$CI = \overline{X} \pm t_{crit} * SE$$

- where t_{crit} is the critical t value.

How should the confidence interval using the t distribution relate to the one created using the normal distribution?

The CI using the t should be wider

The CI using the t should be narrow

It depends

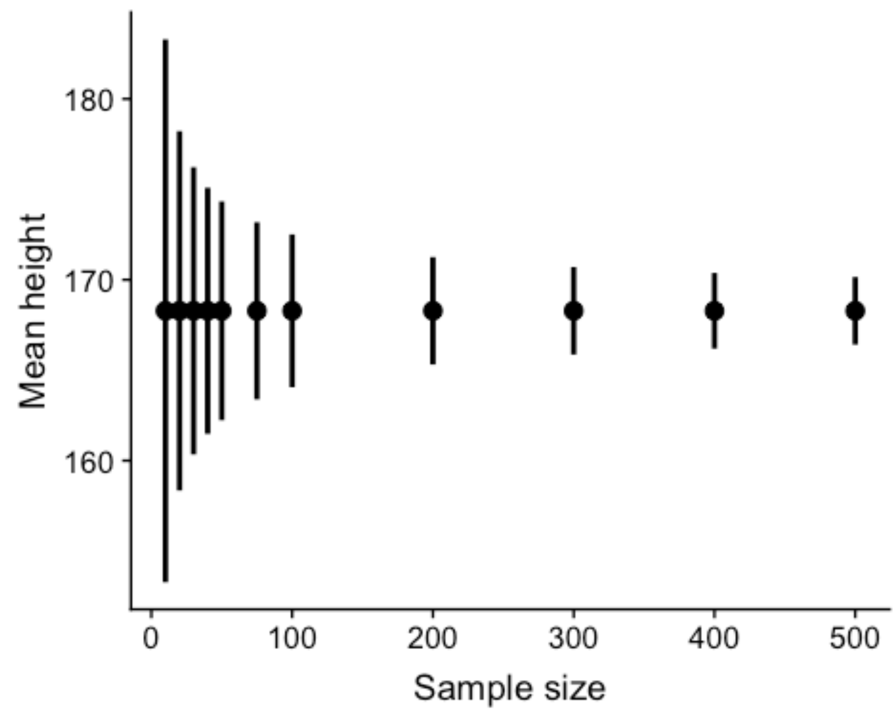
NHANES EXAMPLE

- As seen in the NHANES weight example (with sample size of 250) to the right, the confidence interval using the *t* distribution is slightly larger than the normal.
- Remember: the population mean is a fixed parameter (which we know is 79 because we have the entire population in this case)
  - In the long run, 95% of the confidence intervals will contain the true value

type	lower_cutoff	upper_cutoff
normal	80.12797	85.40963
t	79.99535	85.54225

CONFIDENCE INTERVALS AND SAMPLE SIZE

- Because the standard error decreases with sample size, the CI should get narrowers as the sample size increases
- The confidence interval becomes increasingly tighter as the sample size increases, but increasing samples provide diminishing returns
- the denominator of the confidence interval term is proportional to the square root of the sample size.



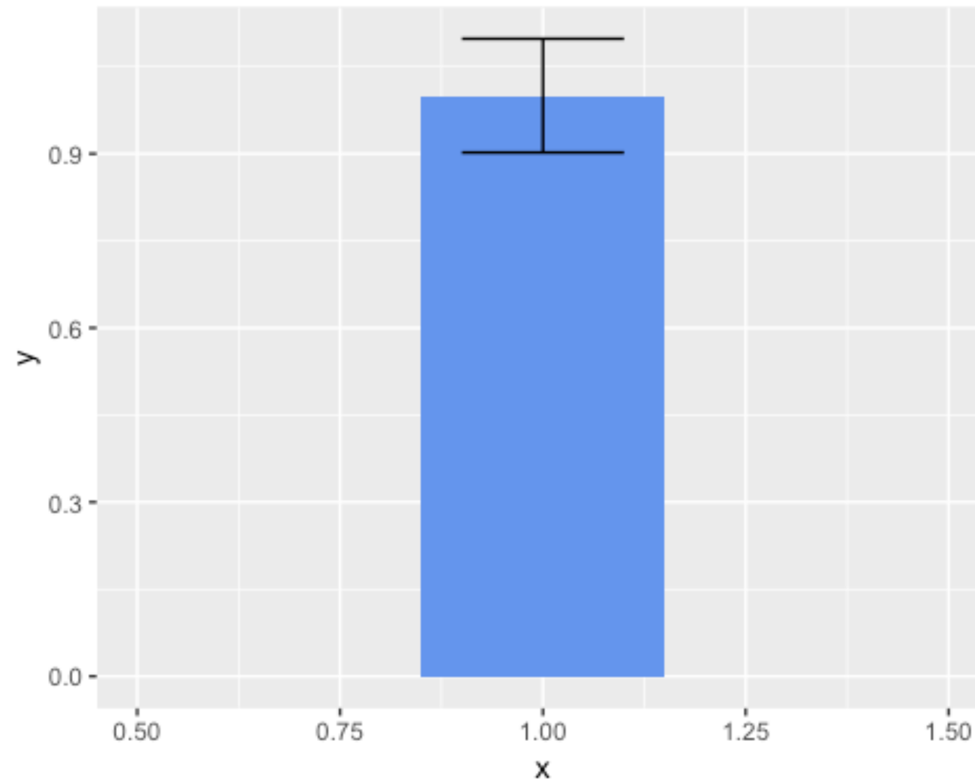
An example of the effect of sample size on the width of the confidence interval for the mean.

RELATION OF CONFIDENCE INTERVALS TO HYPOTHESIS TESTS

- There is a close relationship between confidence intervals and hypothesis tests.

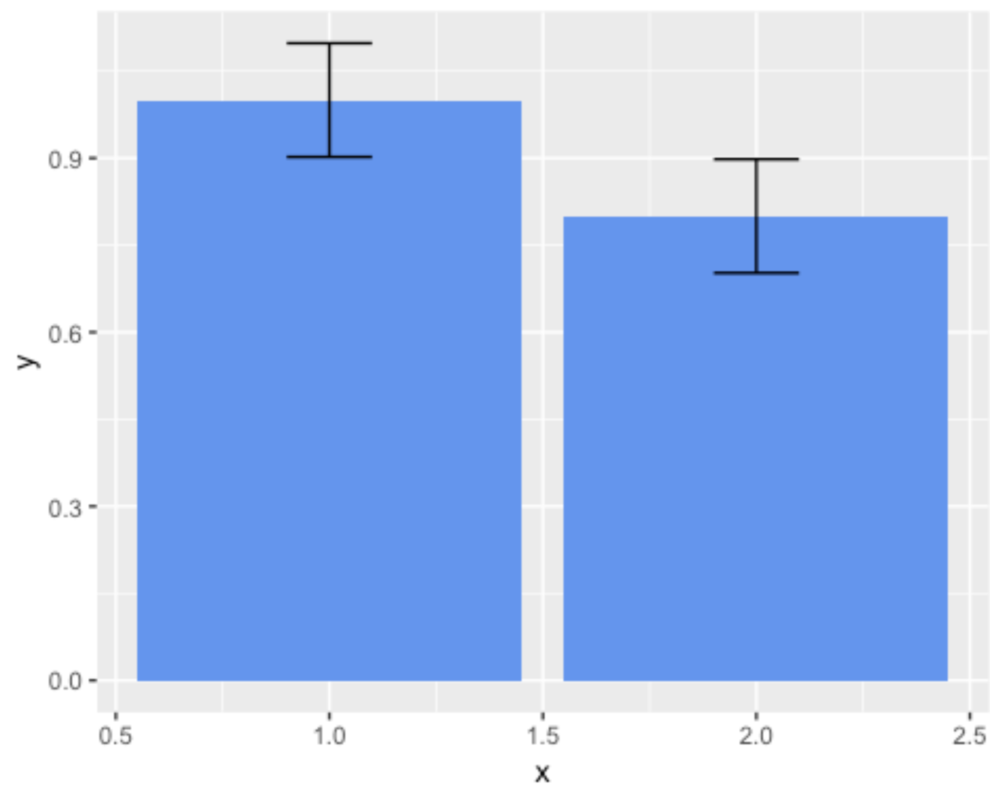
- If the confidence interval does not include the null hypothesis, then the associated statistical test would be statistically significant.

- In the plotted example, because the lower end of the 95 % CI is 0.9, a hypothesis test for the mean against any value below that would be significant.



RELATION OF CONFIDENCE INTERVALS TO HYPOTHESIS TESTS

- Things get trickier if we want to compare the means of two conditions
  - If each mean is contained within the confidence interval for the other mean, then there is certainly no significant difference at the chosen confidence level.
  - If there is no overlap between the confidence intervals, then there is certainly a significant difference at the chosen level.
  - Otherwise, it's complicated!
- In general we should avoid using the "visual test" for overlapping confidence intervals



Which of the following is the most appropriate interpretation of a 95% confidence interval?

We have 95% confidence that it contains the population mean

In the long run, it will contain the population mean 95% of the time

There is a 95% chance that the population mean is within the confidence interval

Any value outside the confidence interval has only a 5% chance of being the population mean