Econ 270 Lecture 2

Sam Gifford

2024-01-17

Types of Data (1.2.2)

- Quantitative
  - Interval: differences are meaningful
  - Ratio: nonnegative with a "true zero"
  - Discrete or Continuous
    - Discrete use counting numbers
- Qualitative (categorical)
  - Ordinal: order matters
  - Nominal: No order

Quantitative Examples

- Money (ratio)
- Weight (ratio)
- Time (interval)
- Temperature (F or C) (interval)
  - Kelvin is ratio

Quantitative Data

- I don't care about the distinction between ratio and interval
- Important thing is that we can interpret values in context
  - When comparing income levels of earners, should we compare differences or ratios?
- Data classifications aren't necessarily standardized

Qualitative Examples

- Name (nominal)
- State (nominal)
- Educational Attainment (ordinal)
- Student Evaluation (ordinal)
  - Labeled 1-5, but not a true quantitative variable

Examples

	Student ID	SAT Score	age	Grade
1:	1	1537	18	A
2:	2	1618	19	B
3:	3	1516	18	C
4:	4	1760	19	A
5:	5	1633	18	B

A Quantitative B Ordinal C Nominal

Summarizing Data

- Some data analysis techniques only work for numeric data
- We'll start with some more holistic measures, then move to single-number summary statistics

Quantiles and Box Plots (2.1.5)

- Suppose you have data that consistent of students exam scores
- You're interested in understanding the general distribution of these scores

83 74 87 58 71 59 83 90 80 75 77 87 92 66 63 76 93 98 79 72

Order Statistics

- One easy thing we can do is sort the data
- What are the largest and smallest values?

```
45 50 52 58 59 59 61 61 63 63 65 66 67 67 71 72 72 73 74 75
```

Quantiles and Box Plots (2.1.5)

- We define the median as the 'middle' value: half of all exam scores are below this, and half are above
  - If there's an even number a common convention is to take the average
- What is the median of the following data?

```
13 2 10 8 5
```

Quantiles and Box Plots (2.1.5)

- Median for the exams?

45 50 52 58 59 59 61 61 63 63 65 66 67 67 71 72 72 73 74 75

Quantiles and Box Plots (2.1.5)

- The median is a measure of central tendency
- Often we're interested in more than just the average student
- One more general way is to divide the class into 4 groups instead of 2

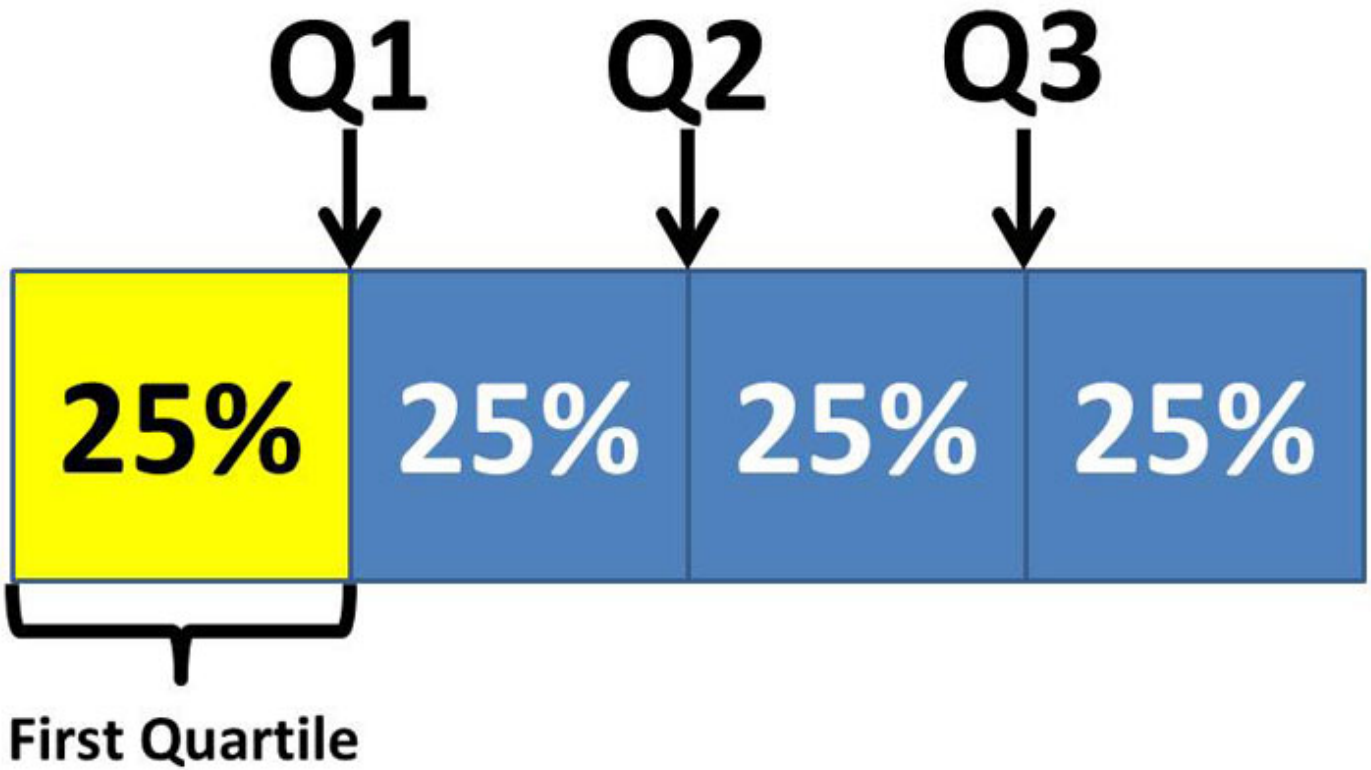


Figure 1: from sciencenewsforstudents.org

Quantiles and Box Plots (2.1.5)

What is the third quartile (Q3) of exam scores?

45 50 52 58 59 59 61 61 63 63 65 66 67 67 71 72 72 73 74 75

General quantiles

- In general we define as percentile as the number such that p% of the data lies below this point
- The probability notation for this is rather confusing: $x : P(X \leq x) \geq p$
- The 1st, 2nd, and 3rd quartiles are the 25th, 50th, and 75th percentiles
  - The 2nd quartile is also the median
  - The top 1% is the 99th percentile

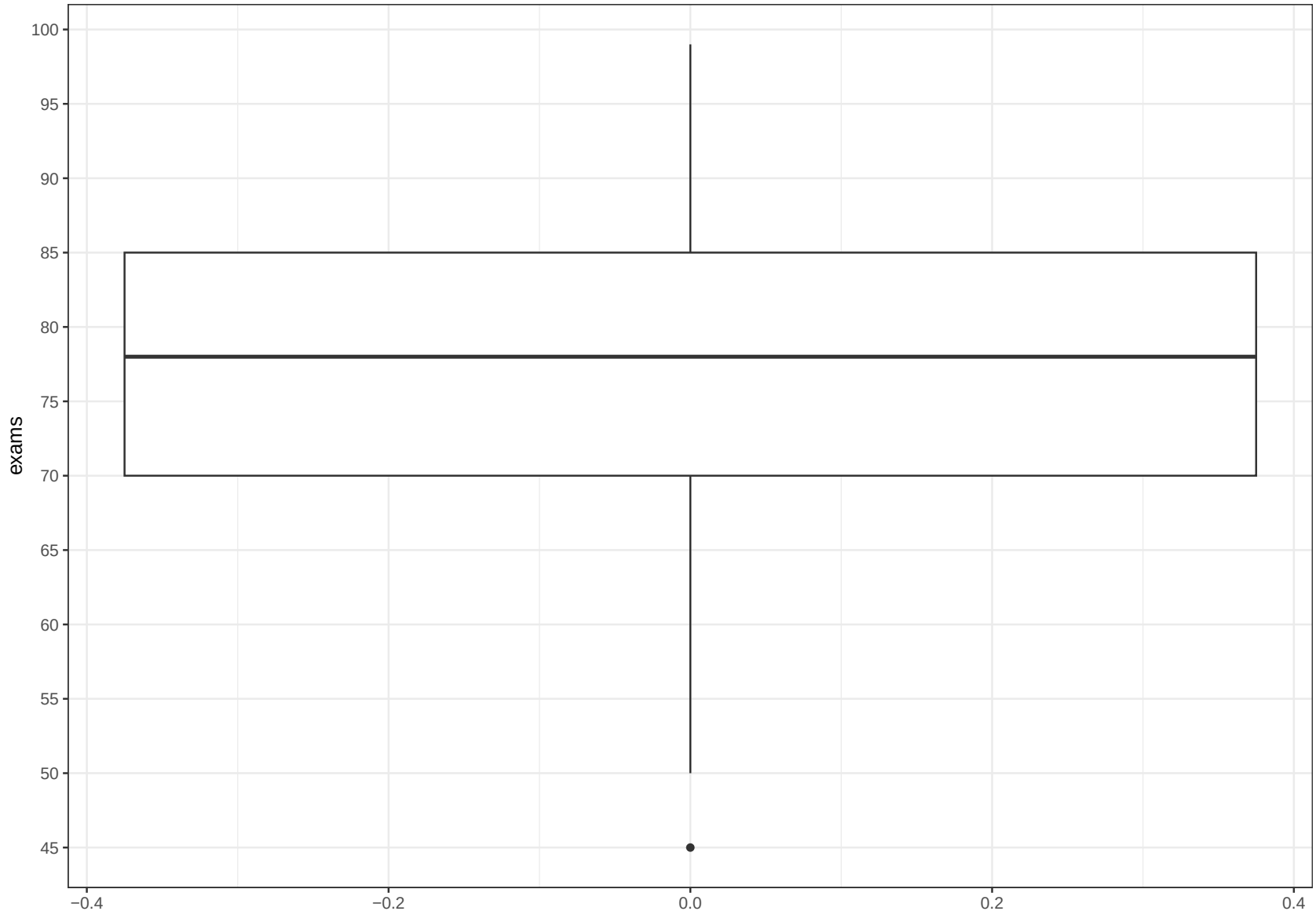
Interquartile Range and Outliers

- The difference between the first and third quartiles is called the interquartile range (IQR)
  - IQR = Q3-Q1
  - It gives a measure of spread, or dispersion, of the distribution
- An outlier is an extreme observation that is atypical
  - No consistent definition, but sometimes defined to be more than 1.5 times the IQR from the median

Box Plots

- Using the quartiles and the IQR we can give a plot that is a good summary of our data
- The Q1, the median, and Q3 are all plotted as a line, creating a box
- Whiskers extend from the top and bottom of the box until it either reaches the max/min of the data, or until it reaches 1.5 times the IQR, whichever comes first
- Any values outside this range (the outliers) are individually plotted

Box Plots

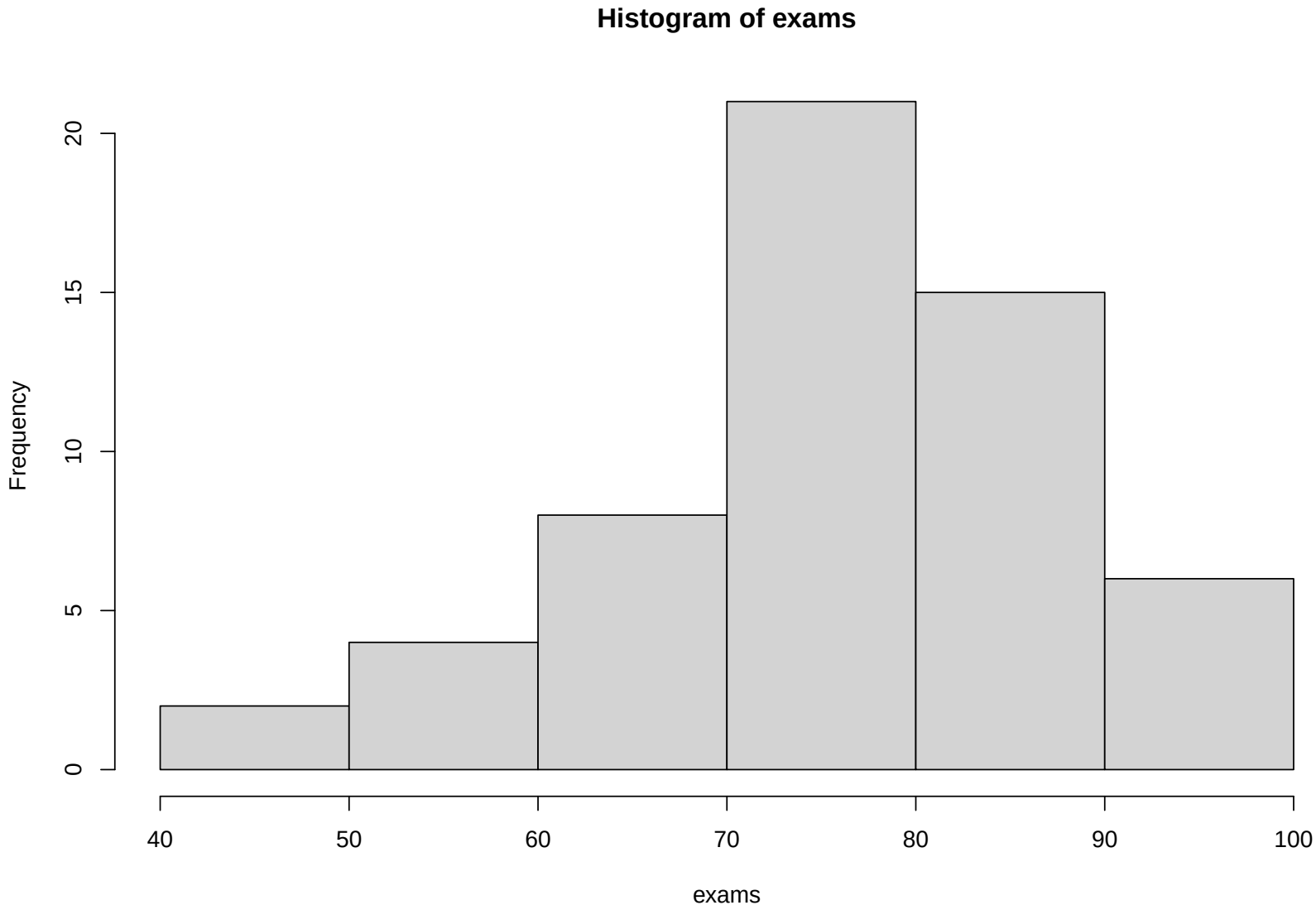


Histogram (2.1.3)

- Boxplots provide a good summary, but leave out information
- We can instead bin our exams into ranges and count how many students fall in each
  - e.g. 18 students scored in the 70s, and 15 in the 80s
- If the bins are of equal size this plot is called a histogram

Histogram

- This is a ‘left-skewed’ exam



Skewness

- If the tail tapers off to the right, a distribution is skewed right
  - Generally means that the mean (average) is greater than than the median
- A distribution with no skew is symmetric

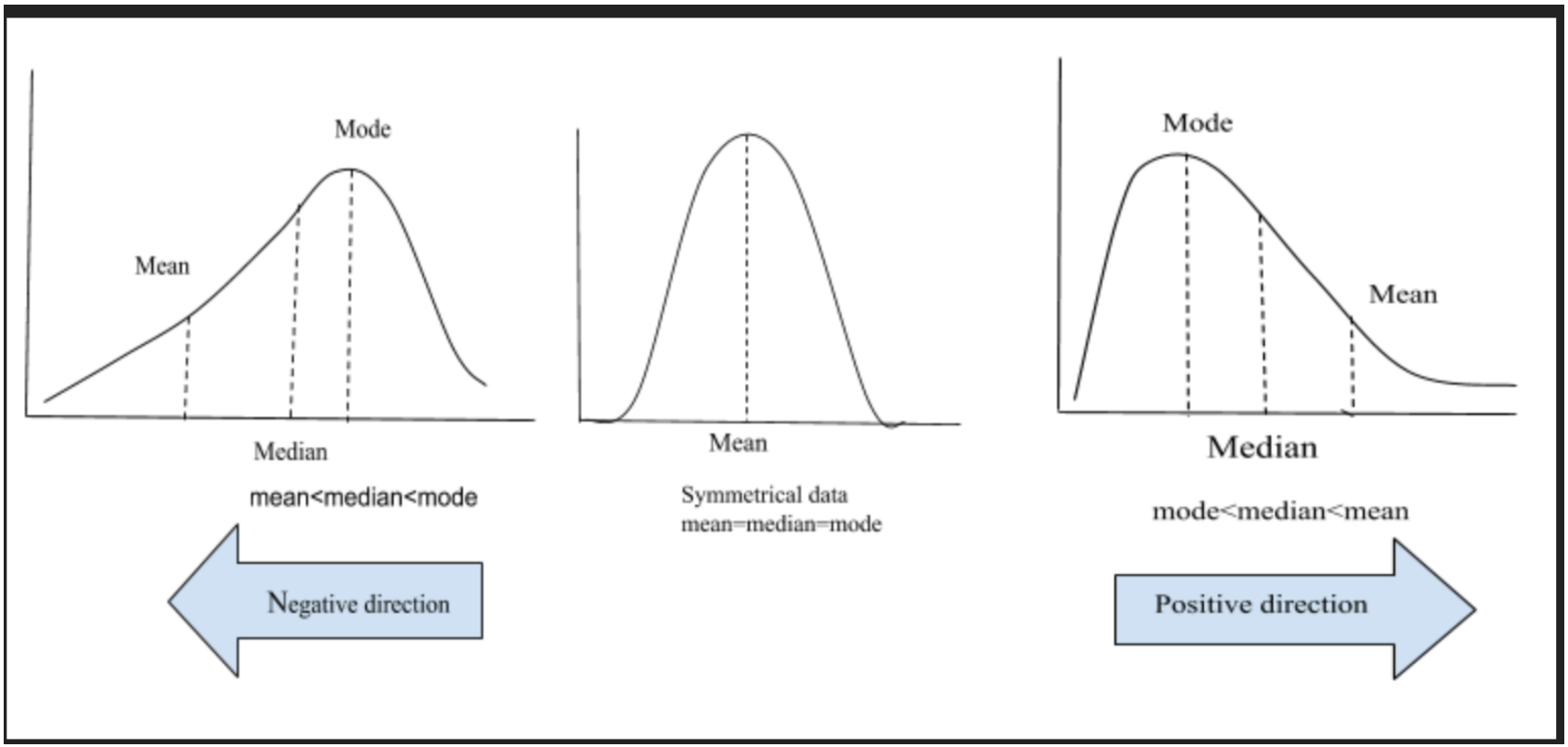
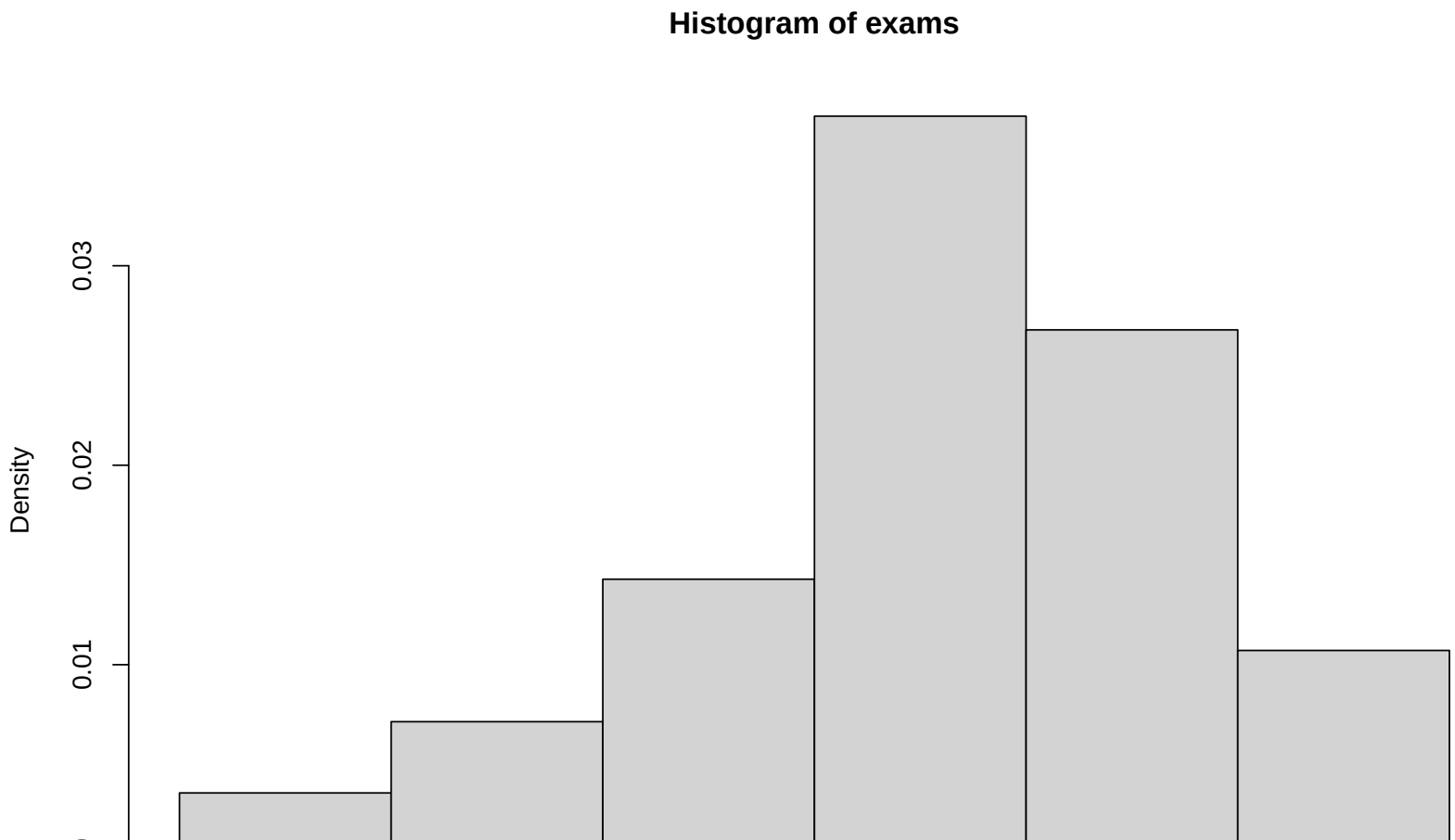


Figure 2: from alevelsmath

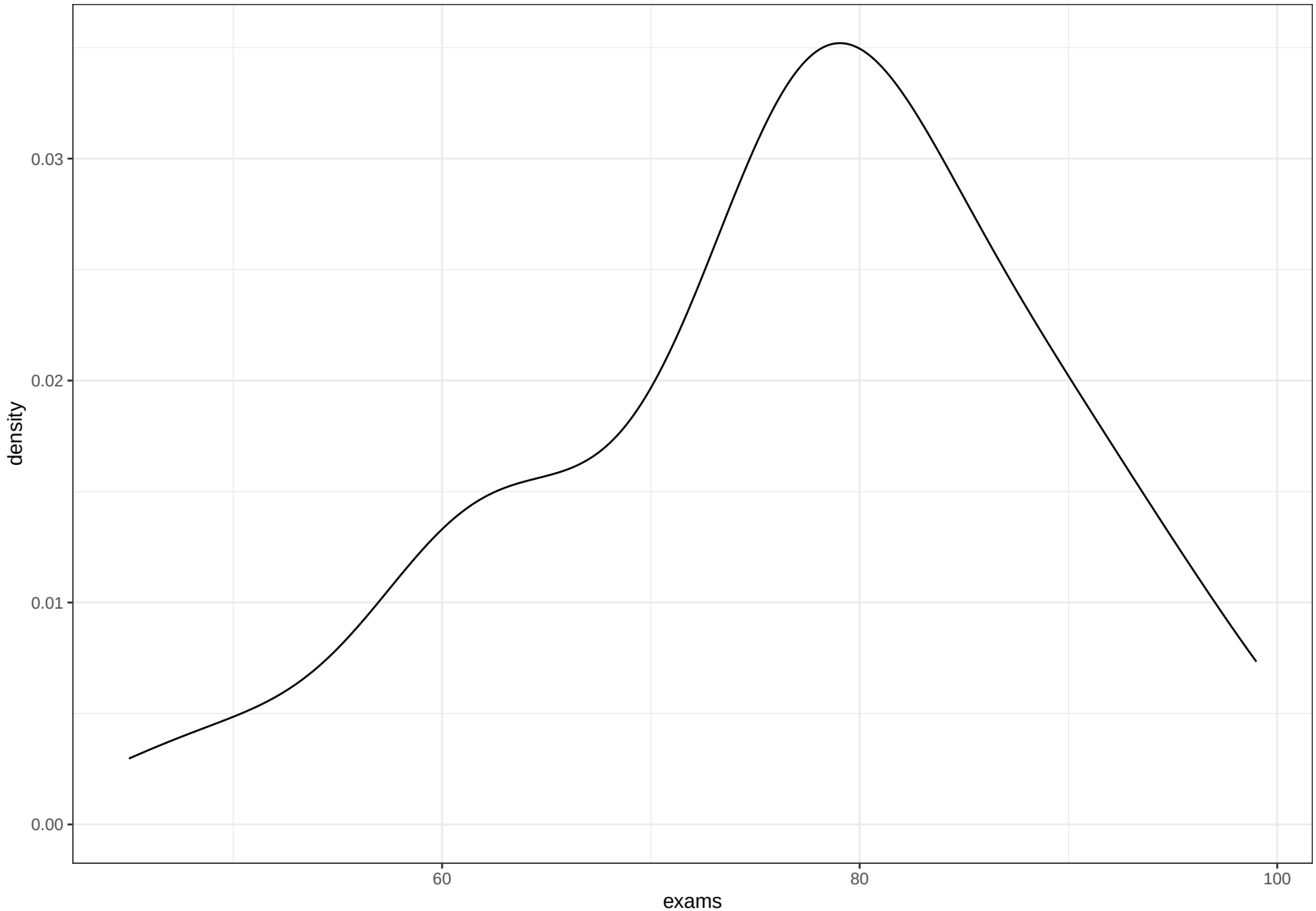
Absolute vs Relative

- We can plot histograms with either the counts or the percentages. Percentages tend to be more useful when comparing distributions
- This plot uses ‘density’ - multiply by the bin width to get the percentage



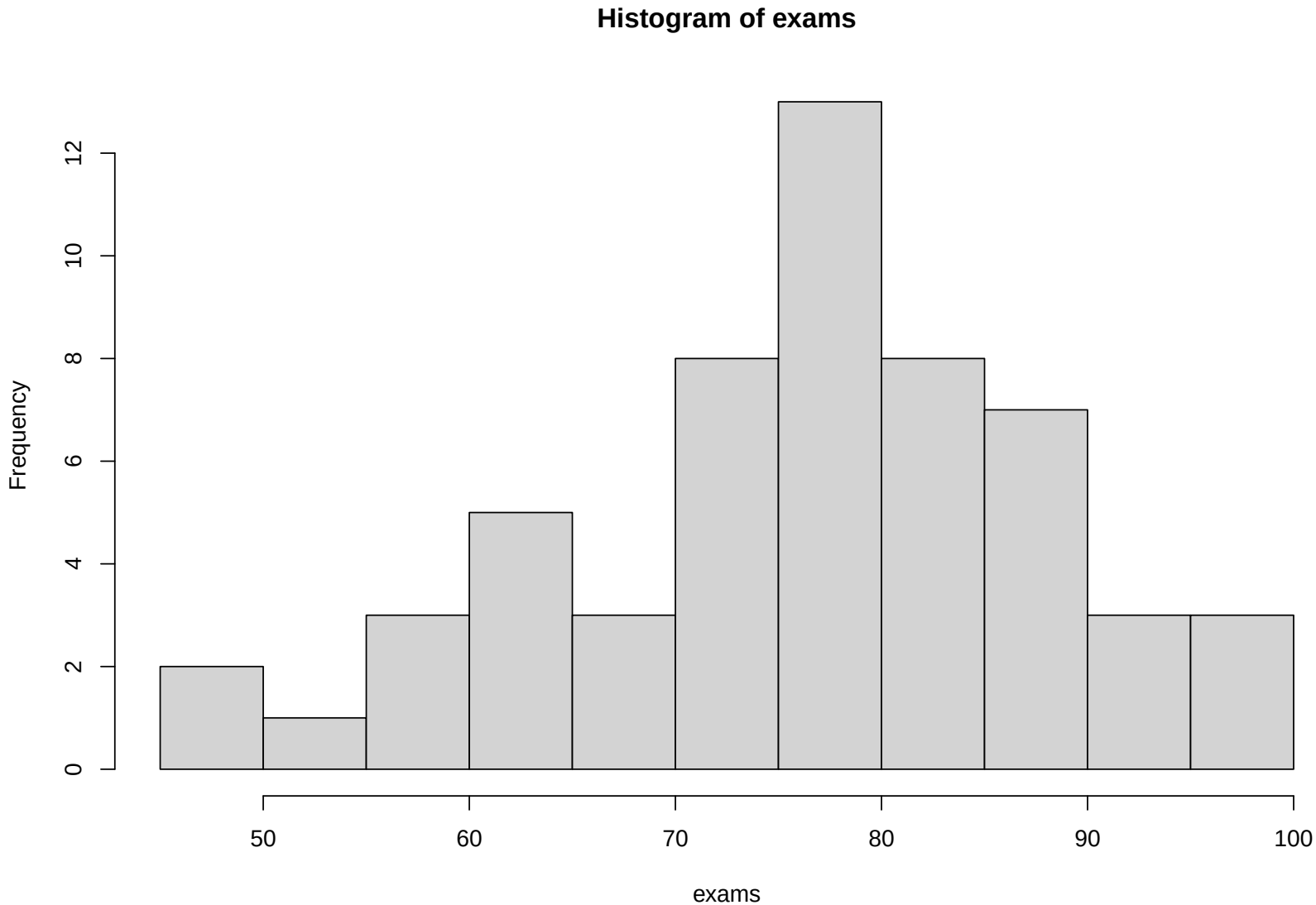
Kernel Density

- An alternate way to draw this is with a kernel density plot
- Good for overlaying distributions



Bin width

- More bins will reveal more details, but be less smooth
- Optimal bin width is more of an art than a science



Bimodality

- The mode is the most frequent observation in data
- If the histogram has two peaks, it is called bimodal
- The exams are bimodal - there is a peak round 65 and around 75 in the data

View Tradeoffs

- Note that in the first histogram the skewness is easier to see, while in the second the bimodality is easier to see
- We often want multiple views of data to get a sense of the distribution

Qualitative Data

- We can still construct something akin to a histogram for ordinal data
- For nominal data, we can construct a regular bar graph to display count information

Summary statistics (2.1.2)

- We often want to view the whole distribution, but sometimes we just need a simple summary
- Summary statistics can give a brief overview of the data as a single number
- The quartiles and median are all examples of such summary statistics

Central Tendency

- The center of a distribution is usually important - it represents the ‘average’ data point
- The median is one measure of central tendency
- The more common one is the mean (or average)

Mean

- The mean, $\bar{x}$, is just the average of a set of data
- $\bar{x} = \frac{x_1+x_2+...+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - We will see some alternative notation when discussing probability distributions
- This is the 'balancing point' of a distribution - equal weight is on both sides

Mean vs Median

- Consider the following set of numbers. What is the mean?
- {1,1,1,2,2,2,3,3,3,22}

Robustness

- For the prior sample, removing 22 yields a mean of 2.
- By comparison, the median would be 2 in both cases
- The median is said to be robust: outliers in the data do not have a large effect
- The mean is not robust

Why the mean

- If the median is more typical of the average, why use the mean at all?
- Consider calculating average insurance claims: the mean is what affects the profitability of the insurance company, not the median!
- Usually we (should) care about both

Weighted Means

- Sometimes observations are weighted. In this case we can just multiply by the weight before averaging
- Consider if our observations were aggregated beforehand. What is the average exam score?

```
##    Score N
## 1:    40 1
## 2:    50 2
## 3:    60 8
```

Weighted Mean Calculation

- We can 'expand' the data and calculate the mean: {40,50,50,60,60,60,60,60,60,60,60}
- Or we can directly weight it: 40/11 + 50*2/11 + 60*8/11
- Both answers are identical

Dispersion (spread)

- In addition to the center of the distribution, we usually care about the dispersion, or spread
- The interquartile range is one example of dispersion
  - Like the median, it is robust to outliers

Variance (2.1.4)

- The more common way of measuring spread is with the variance
- The average squared distance of each point to the center
  - $var(x) = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$
- Why squared distance?

Practice

- What is the variance of the set $\{1,2,3,4,5\}$?

Standard Deviation

- Squared units are hard to interpret. We usually take the square root of the variance for this reason
- This is called the standard deviation, usually labeled as σ (sigma)
  - $\sigma = \sqrt{var(x)}$
  - $var(x)$ is sometimes labeled σ^2

A note on sample vs population

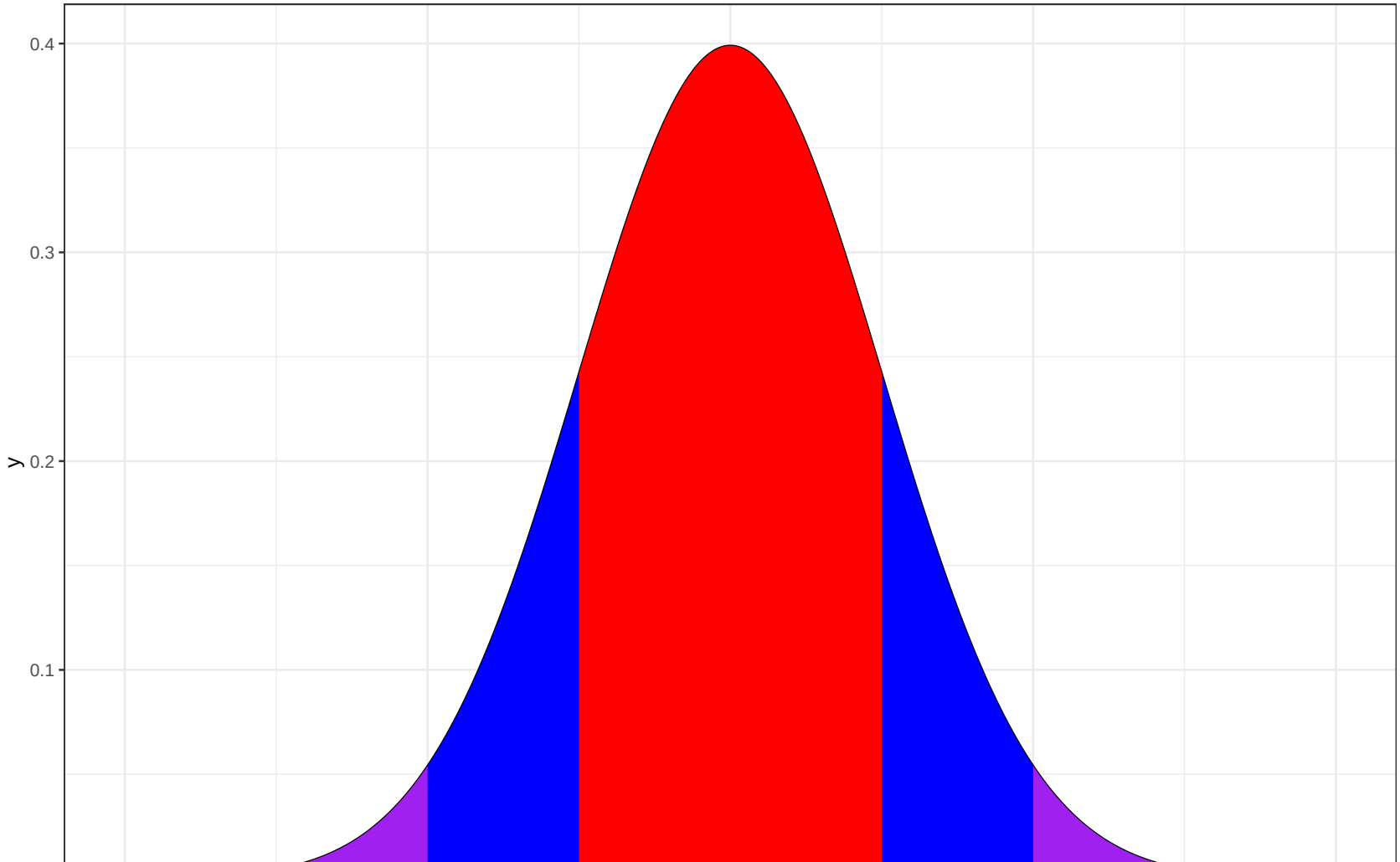
- We're holding off on discussing populations vs samples, but the formulas for variance and standard deviation will be different
  - We'll prove this when we get to discussing sampling

Interpreting Standard Deviation

- Standard deviation is 'almost' the average distance from the mean of the distribution
- Chebyshev's inequality tells us a bit more: $1 - \frac{1}{k^2}$ of the data is within k standard deviations of the mean
- This means that 75% of the data is always within 2 standard deviations of the mean
- About 95% of the data is within 4.5 standard deviations of the mean

Interpreting Standard Deviation

- For a special distribution called a normal distribution, 68% of the data falls within 1 standard deviation of the mean
- 95% falls within 2 standard deviation, and 99.7% within 3 standard deviations



Z-scores

- It is common to 'normalize' data by subtracting the mean and dividing by the standard deviation
- Gives a new distribution with mean 0 and standard deviation 1
- The resulting value is the z-score. It's the number of standard deviations above or below the mean
- Example: Exams had a mean of 76 and standard deviation of 12. A score of 82 would have a z-score of 0.5

Absolute Deviance

- Why not just use the average of the absolute distance from the mean, instead of squared distance?
  - Historically easier to calculate (absolute value is not differentiable)
  - Has synergy with the normal distribution
  - Absolute moments are seeing more usage in modern machine learning
    - Still a tradeoff. General L^p spaces in topology are informative

Other moments

- We can also use summary statistics to categorize the skewness and tail weight
  - Skewness relies on cubed deviations from the mean
  - Tail weight (kurtosis) uses deviations to the fourth power
- Deviations raised to an integer power like this are called central moments. We won't use them much in this class, but are sometimes useful
  - They're frequently used in formal proofs of statistical properties