Central Tendency and Dispersion

Paul Johnson[1]

[1]Political Science & Psychology, KU

2020

Outline

Here's what I hope you will learn

- Definition of "variable" and notation for writing about variables
- Ways to Describe Numeric Variables
- Central Tendency: Mean, Median, Mode
- Dispersion: Variance, Standard Deviation, etc.
- Rescalings

Outline

Variable

Variable a collection of scores that represent observations.

Example:

$$height = \{6.0, 5.1, 4.2, 5.8, 5.4\} \tag{1}$$

Subscript $height_i$: $height_1$ is observation 1, $height_2$ is observation 2, and so forth

Clarification: Social scientists refer to this as "a sample" with 5 observations, but I notice engineers and machine learning people refer to it as 5 samples in a collection.

Common Notation

More abstractly

$$x = \{x_1, x_2, x_3, x_4, \ldots, x_N\} \tag{2}$$

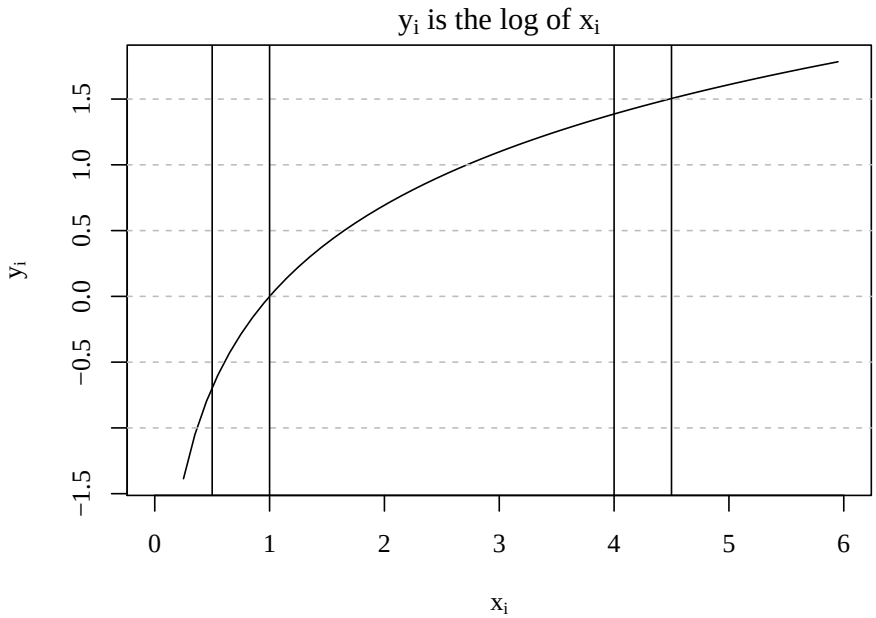
Or perhaps more succinctly

$$x_i, for\, i \in \{1, ..., N\} \tag{3}$$

1. N: capital N refers to "sample size" or "number of observations" (in most social sciences).
2. Usually, when I talk about x_i, I mean to refer to any of the individual observations in x.
3. Set notation
   1. $\in$ means "element of," as in $i \in N$ or $x_2 \in X = \{x_1, x_2, \ldots, x_N\}$.
   2. $\forall$ abbreviation of "for all", so "$for\, i \in N$" might be $\forall i \in N$.

Numeric Variables

- NUMERIC variables: accept mathematical transformations
- The range from $\{minimum, maximum\}$ is (subjectively) meaningful
- From x_i to $2 \times x_i$: there is twice as much of it
- Analysis may be altered (improved or damaged) by transformations



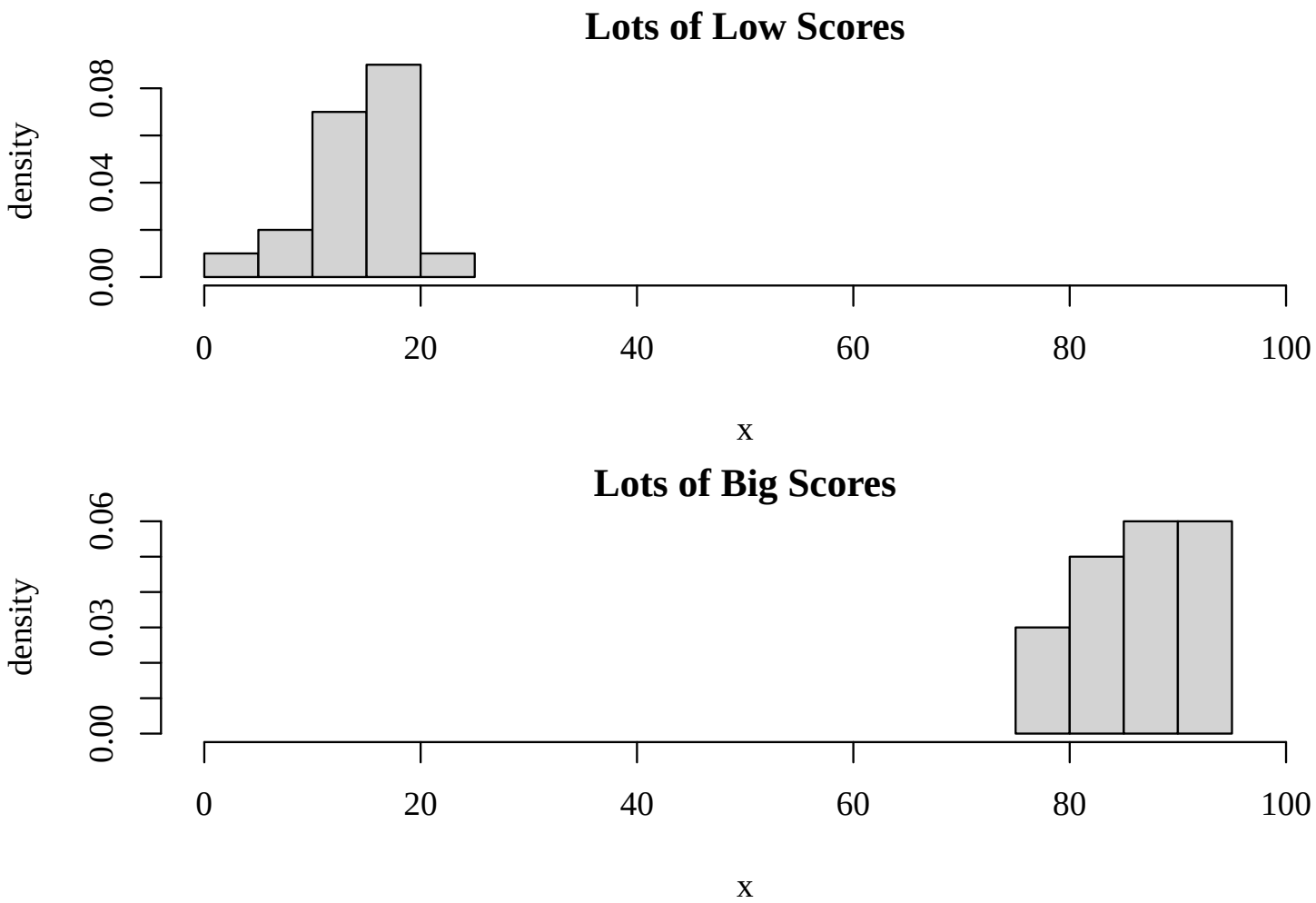
- $\log()$ magnifies the importance of a step from 0.5 to 1 and shrinks the importance of a step from 4 to 4.5.

Terminology to Describe Variables

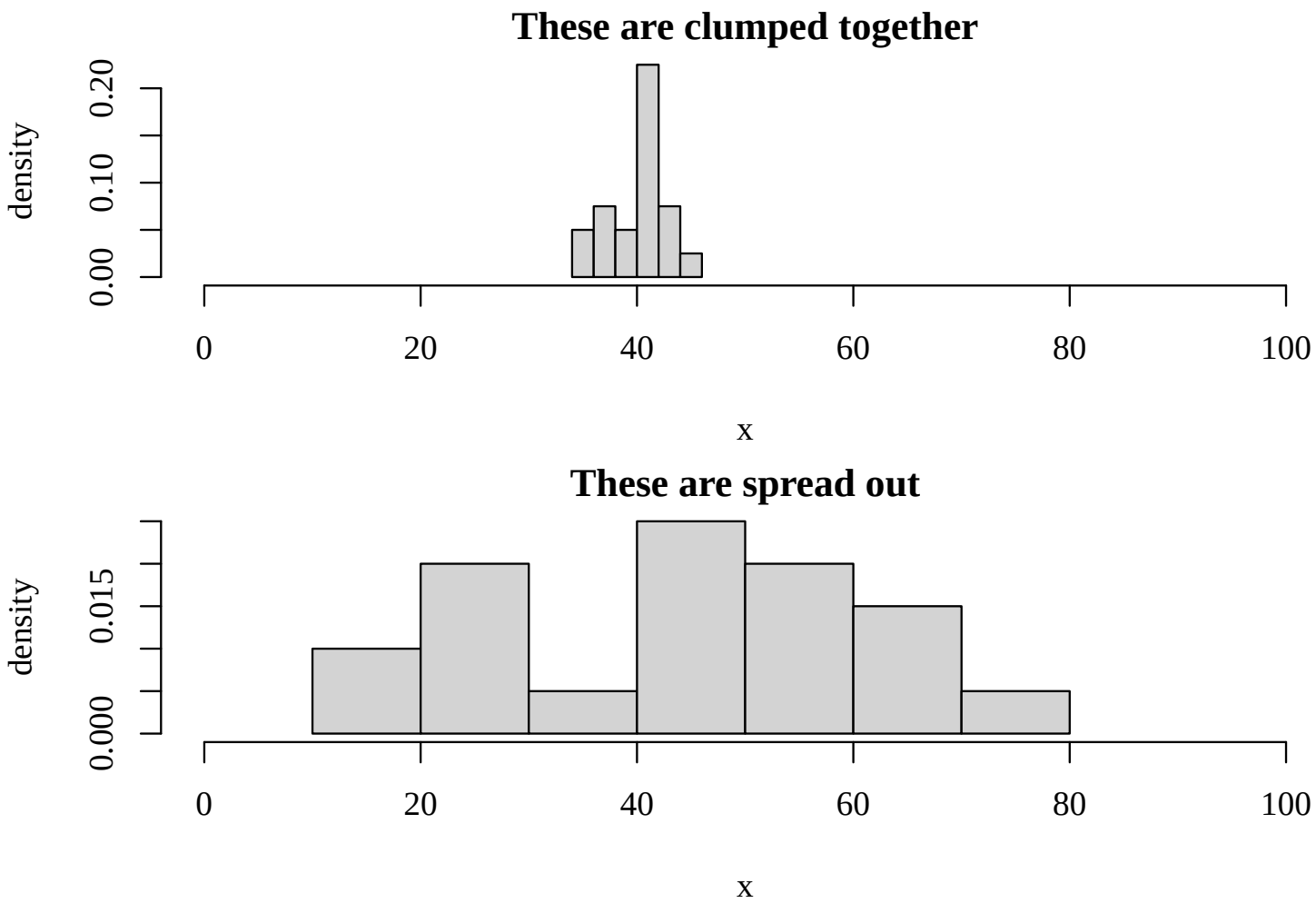
- Central Tendency: Where, "generally" are the scores? Is there a "meaningful" (subjective) characterization of where "most" scores are situated
- Dispersion: How "spread out" are the scores? Is it not meaningful to talk about a "typical" observation?
- Shape of Distributions: Do the observations appear to be
  - Unimodal (one most-likely score, others less likely)
  - Symmetric or Skewed

Outline

Histograms: Compare Two Variables



Histograms: Compare Two Variables

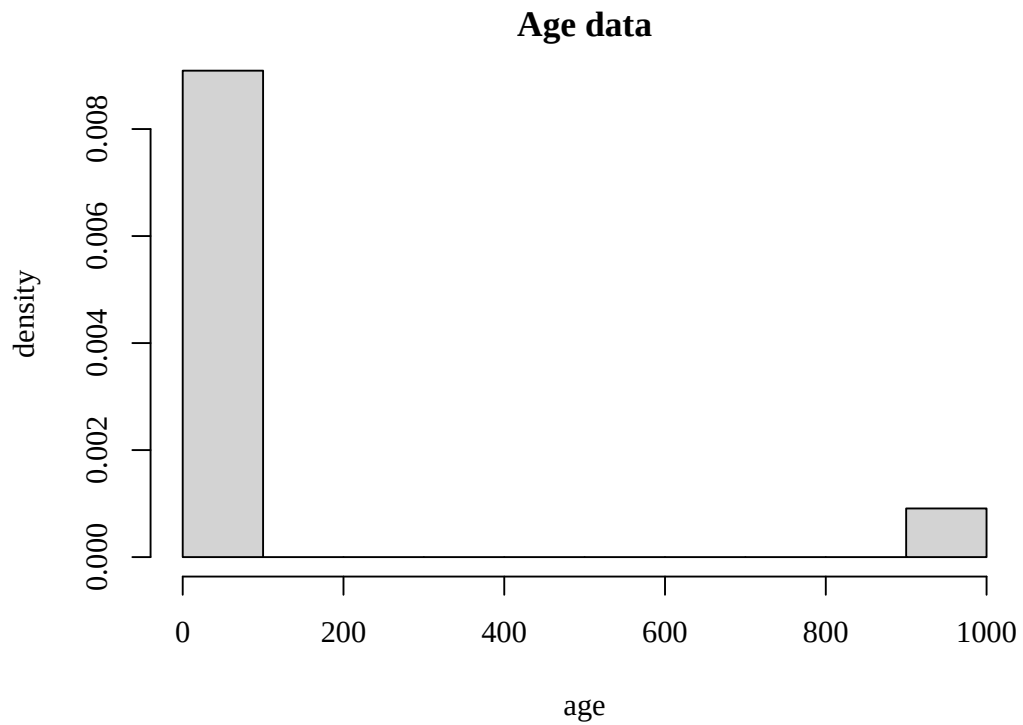


Define "Histogram", Please

- Group observations into "bins" of similar scores
- Draw bars to represent the proportion of all scores that fall into each bin
- The areas of the bars should sum to 1.0
- The hist function can produce the bins and counts, without drawing a plot (see the argument plot=FALSE)

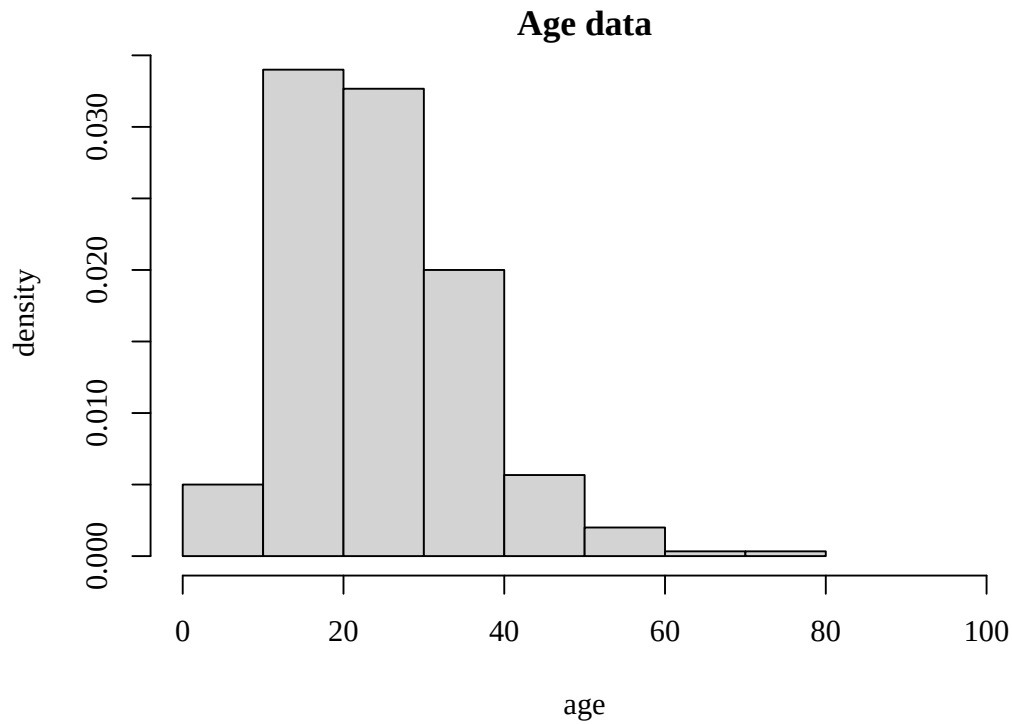
Histograms: Check for Data Errors

- Suppose your data is supposed to be human age
- But somebody put in 999 for "missing" data points



Histograms: Check for Data Errors

- If you ignore (remove) the cases that are equal to 999 (or set them to NA)
- Generally, whenever you get new data from anybody/anywhere, a histogram is a good "first check" on it.



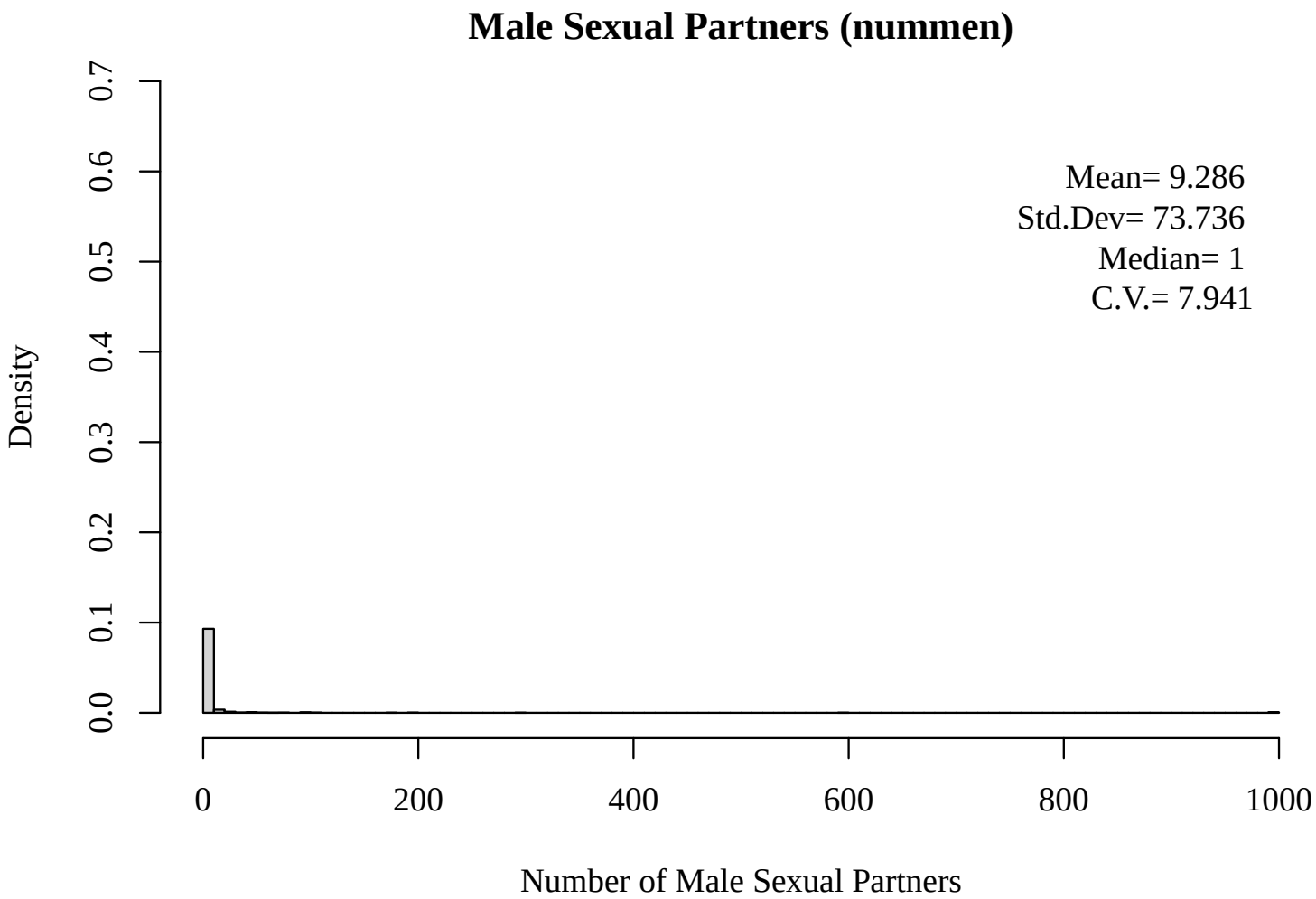
Various "transformations" might be applied

- I'm cautious about fiddling with data
- Some transformations are not "harmless"
- Goal: Be honest with self & others about changes applied to data, including
  - omission of missing or extreme observations
  - multiplicative re-scaling
  - nonlinear transformations (log, Box-Cox, etc.)

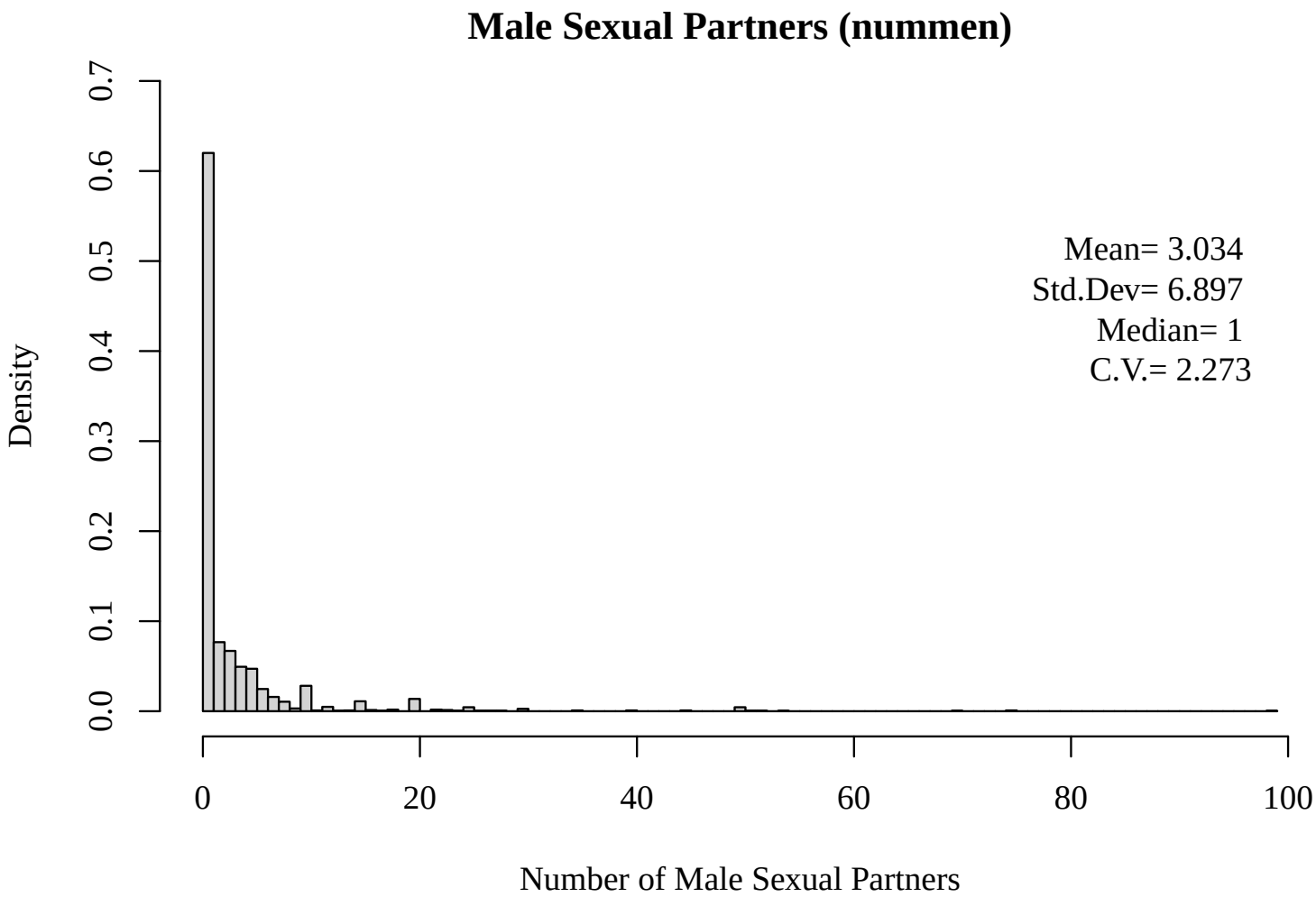
Some Examples from the General Social Survey

/stat/DataSets/GSS/gss-subset2.Rda

Histogram: Spot Typos/Unusual Scores



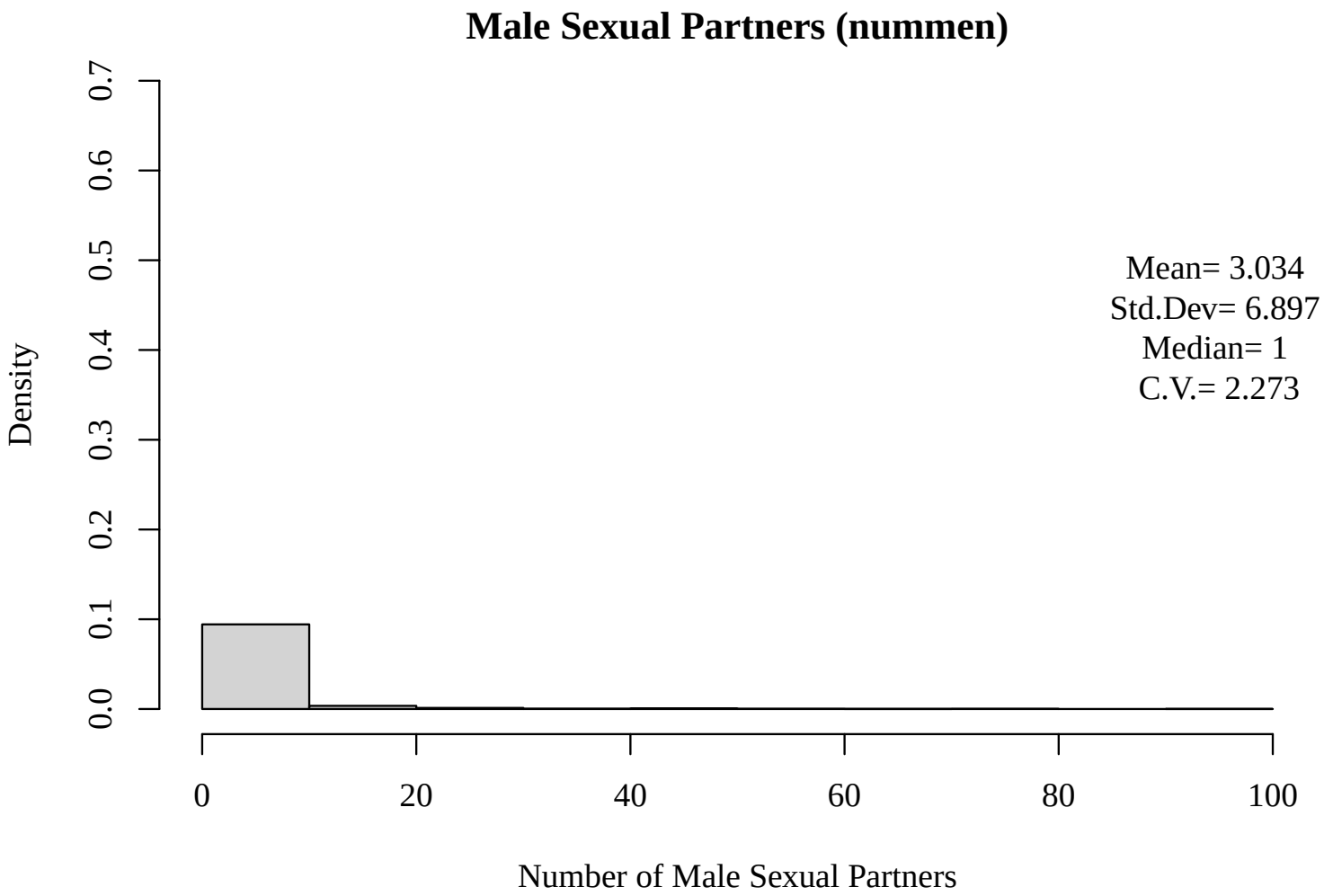
Histogram: Eliminate values greater than 99



The Size of the Bins Can Make a Difference

- Narrow bars have more detail, possibly less generalizability (harder to see patterns)
- Wide bars smooth out too many bumps, hide details
- Many algorithms proposed to choose bin width to automate production of "good" histograms.

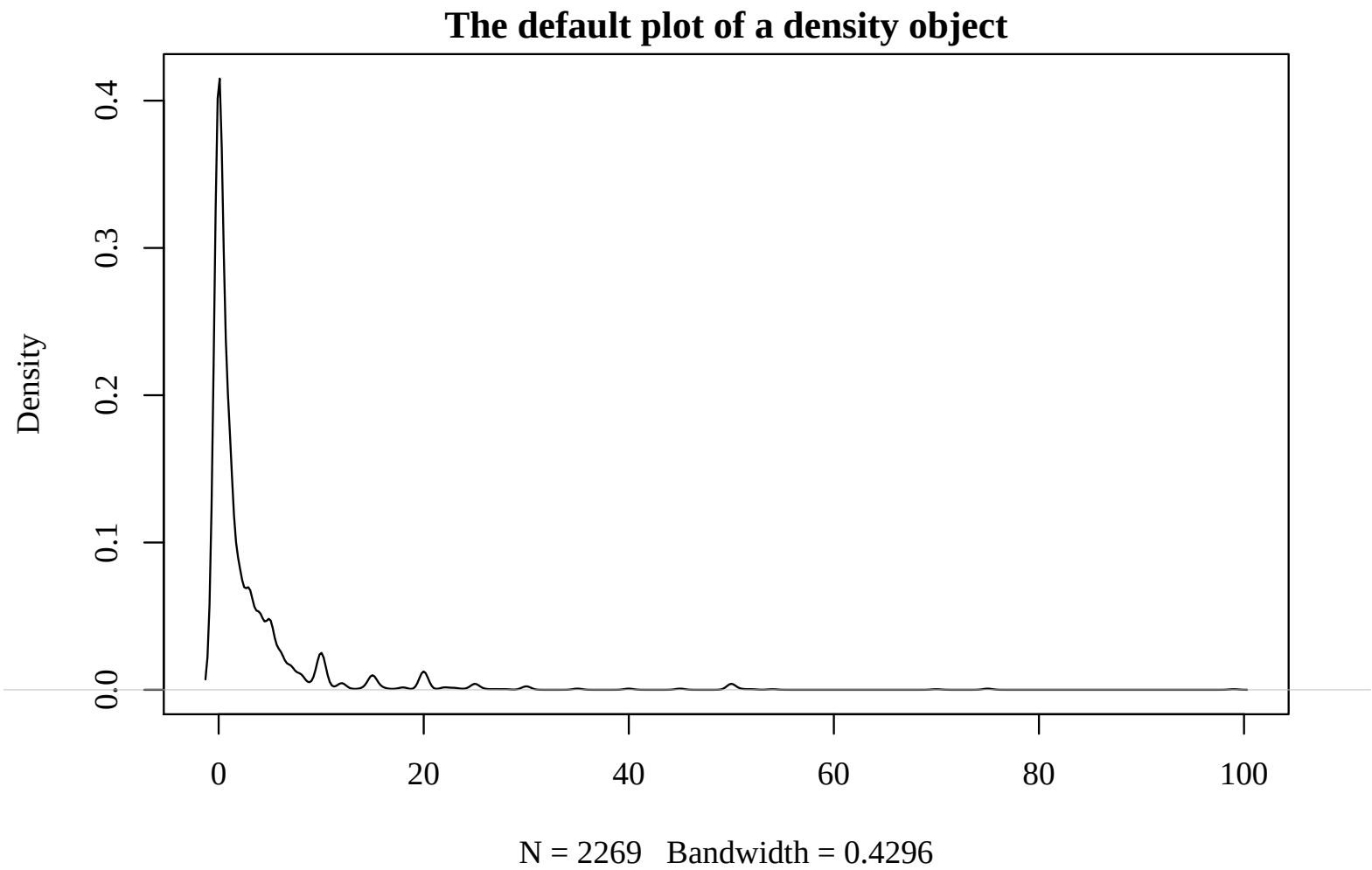
Histogram: Fatter Bars!



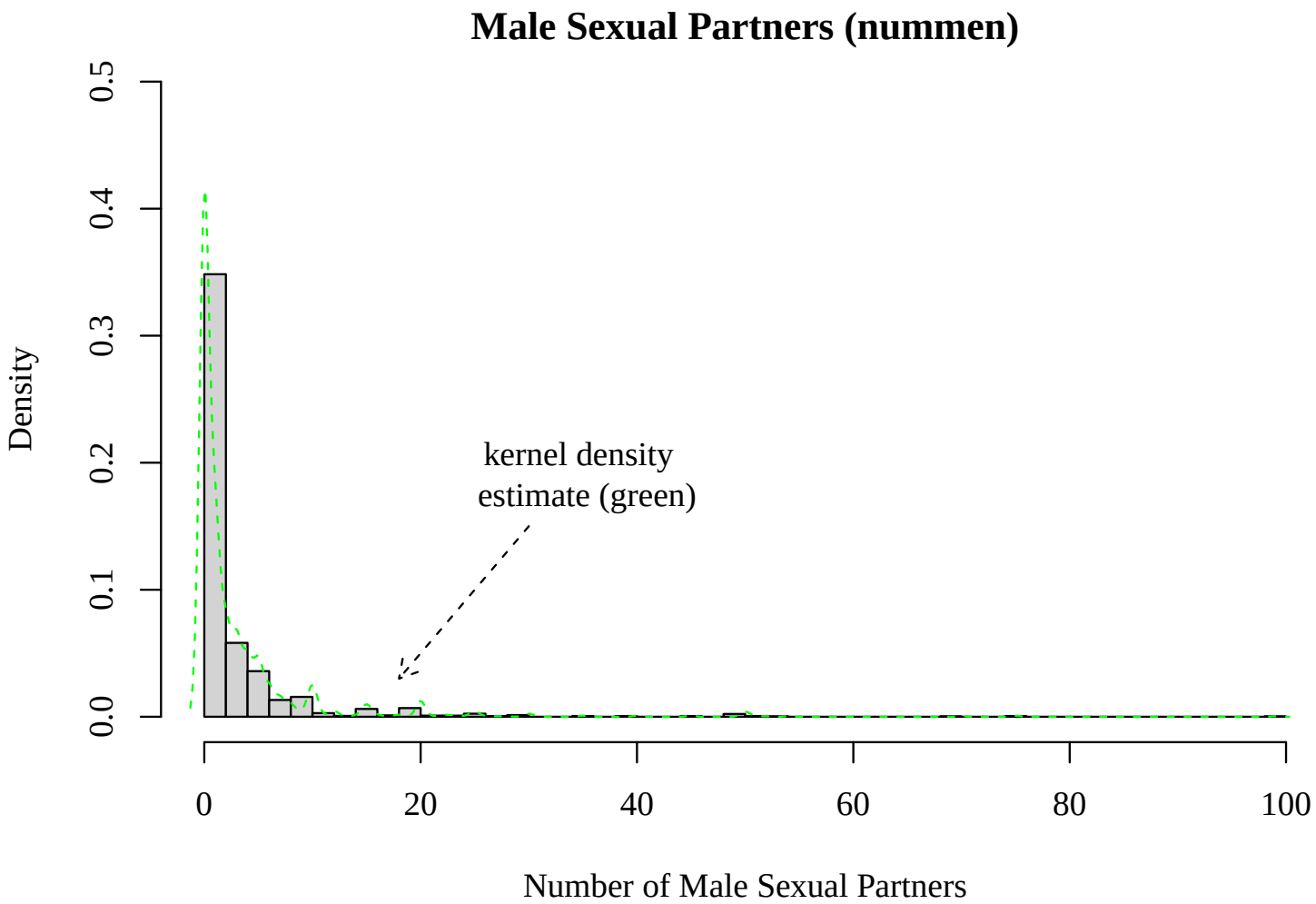
A Smoothing Curve: Kernel Density Estimate (KDE)

- Because of the (subjective) "bin width" problem, other density estimation methods have been developed
- The kernel density estimate is a "smoothing" method that estimates the density at each value, putting more weight on nearby observations than far away ones.
- Some propose to replace histograms with KDE

The Density Estimates



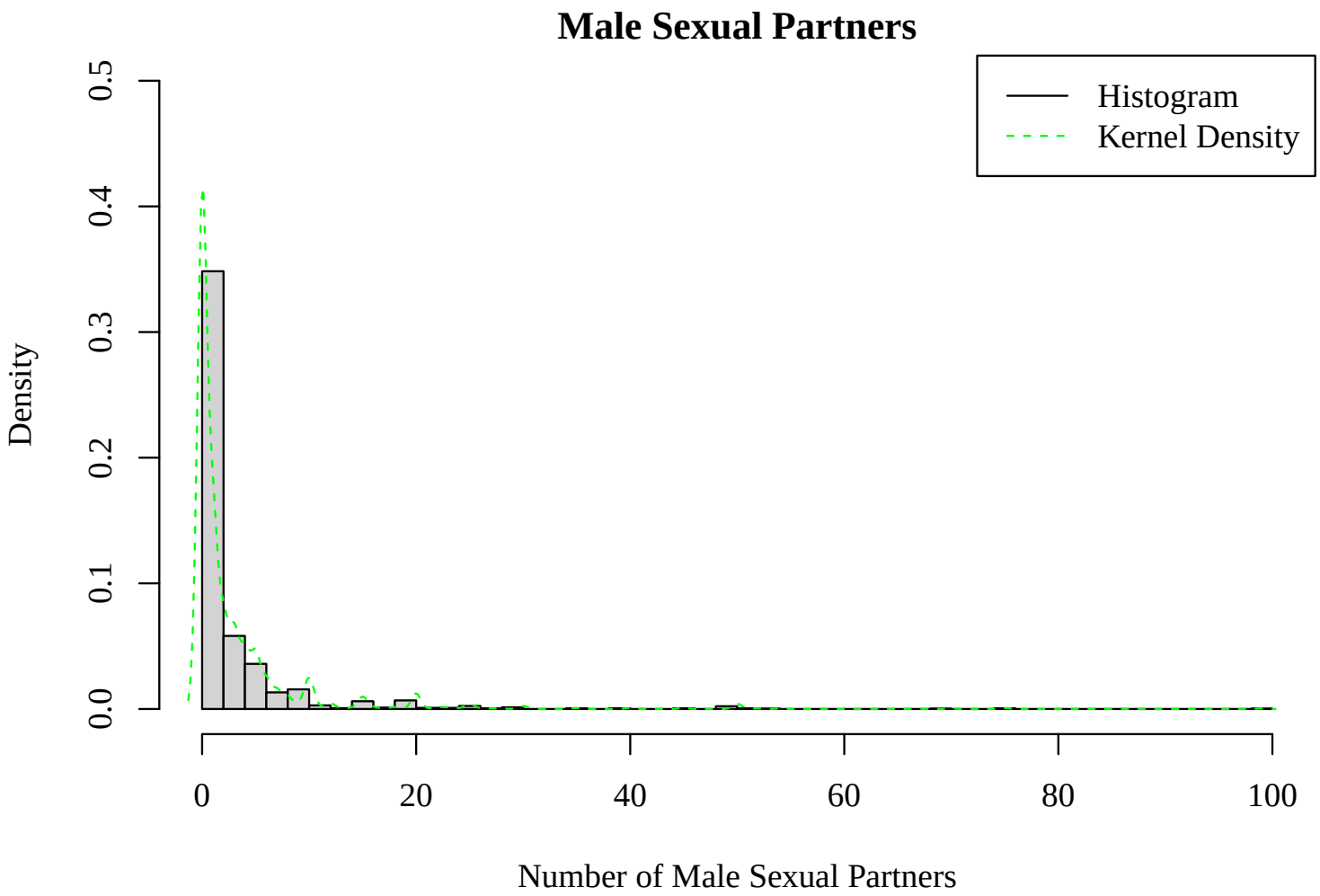
Histogram with Density Super-imposed



Histogram: More on Customizing Histograms

- My lectures in guides/Rcourse (plot-1, plot-2) have plenty of additional detail on beautifying plots.

Histogram: with a "legend"



Outline

Convey Same Info Without a Graph?

- What if your publisher will not allow you the space for a histogram?
- Convey same information without a picture?
- Need to develop terminology to describe and compare what we see.

Mean = Average, Common index of "central tendency"

- "central tendency" is, vaguely speaking, the "middle" of a symmetric distribution

  Mean: (AKA "average").

$$\bar{x} = \frac{Sum\,Of\,All\,Scores}{Number\,Of\,Scores} \tag{4}$$

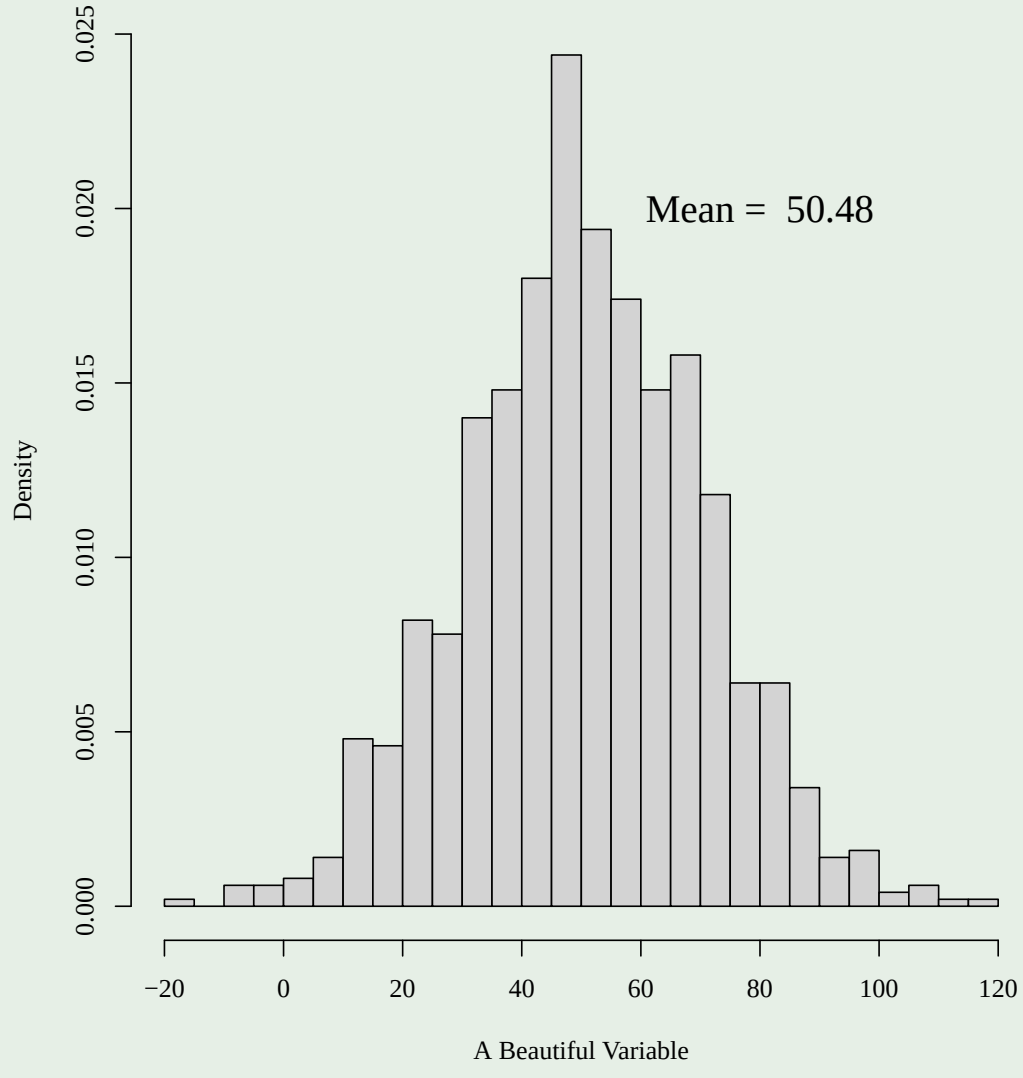
  Given x, add up the scores, then divide by N.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{5}$$

- About Notation:
  - Some math books use "m" for estimated mean, because the true (unknown) value is μ (Greek "mu").
  - $\bar{x}$ is common notation for average, but I don't know why.
  - Sometiems I write μ for the true value and $\hat{\mu}$ for the estimate.
  - Sometimes I write $\widehat{M(x)}$ for the estimate

- I manufactured a sample of pleasantly symmetrical random data
- The sample mean is 50.485
- Appears (to me)
  - unimodal (one peak)
  - symmetric (more or less)

A Histogram with 30 Bins



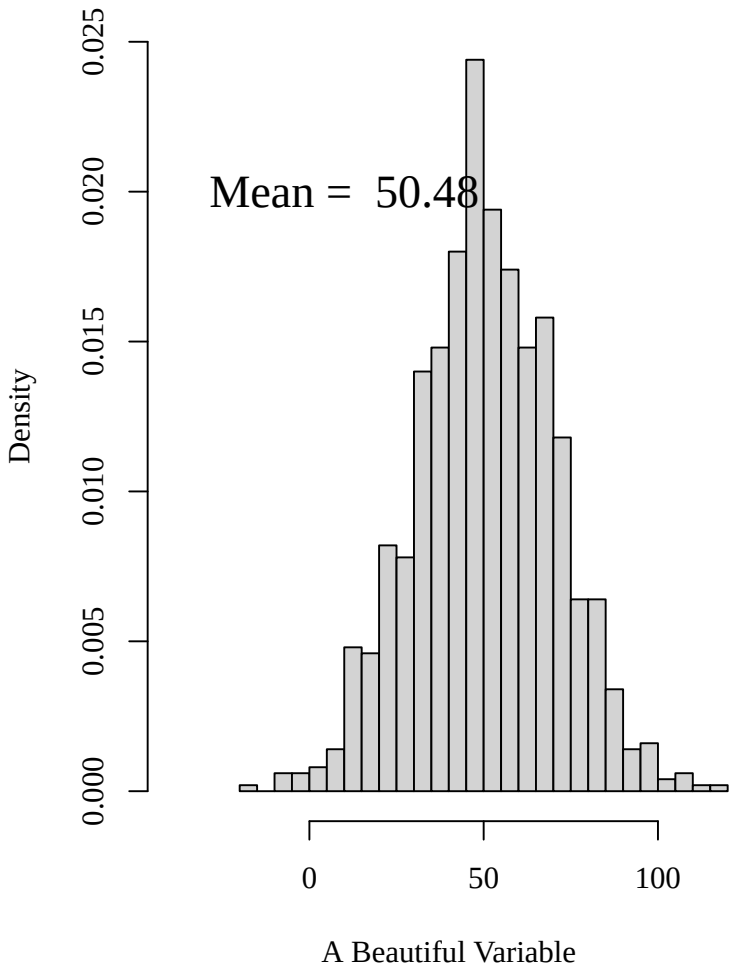
You too can manufacture Normal samples

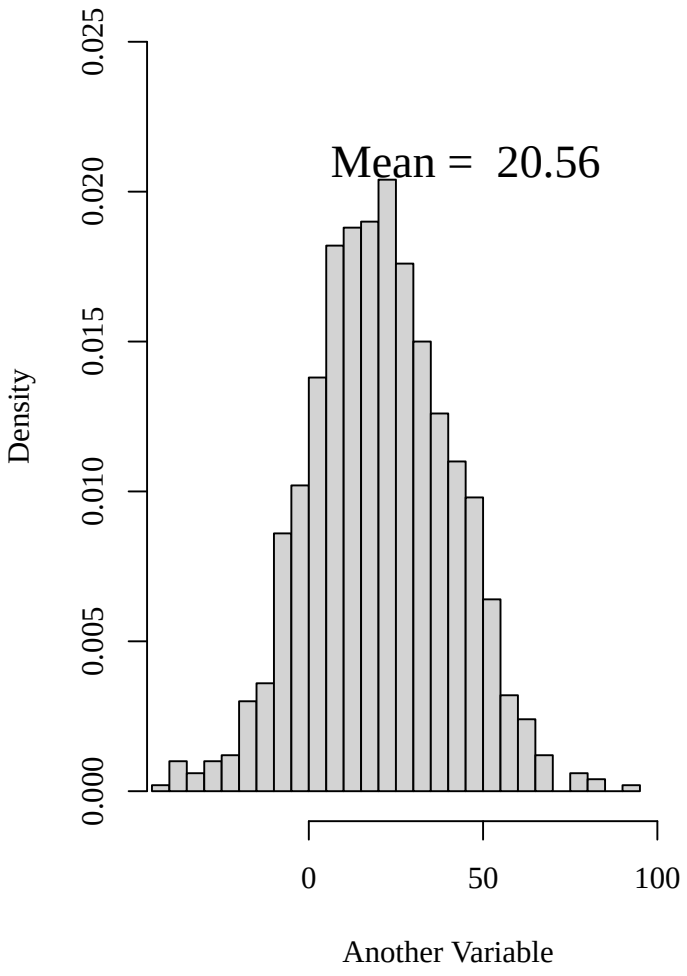
- I used R's rnorm function to draw some example observations

```
set.seed(1234321)
myx <- rnorm(1000, mean=50, sd=20)
```

- That creates 1000 observations from the Normal distribution, $N(50, 20^2)$
- We specify 2 parameters
  - 50 is the parameter mu (μ), the "true mean"
  - 20 is the parameter sigma (σ), which controls the "dispersion" of the scores.
- "Gaussian distribution" another name for the Normal.
- In case you wondered, the sample standard deviation is 19.977

Compare 2 variables





Outline

Variance

Variance: the average of squared deviations about the mean (AKA mean squared error)

1. Calculate the difference between the i'th case and the mean:

$$x_i - \bar{x} \tag{6}$$

2. Square that:

$$(x_i - \bar{x})^2 \tag{7}$$

3. Do the same for all and add them up:

$$\sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{8}$$

Variance ...

4. Then divide by N.

$$Var(x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \tag{9}$$

- In some contexts, it is preferred to divide by $N - 1$. That is needed to
  1. create an "unbiased estimate" of the true mean squared error of a data-generating process
  2. to use the variance as a component in further calculations, such as a T-test

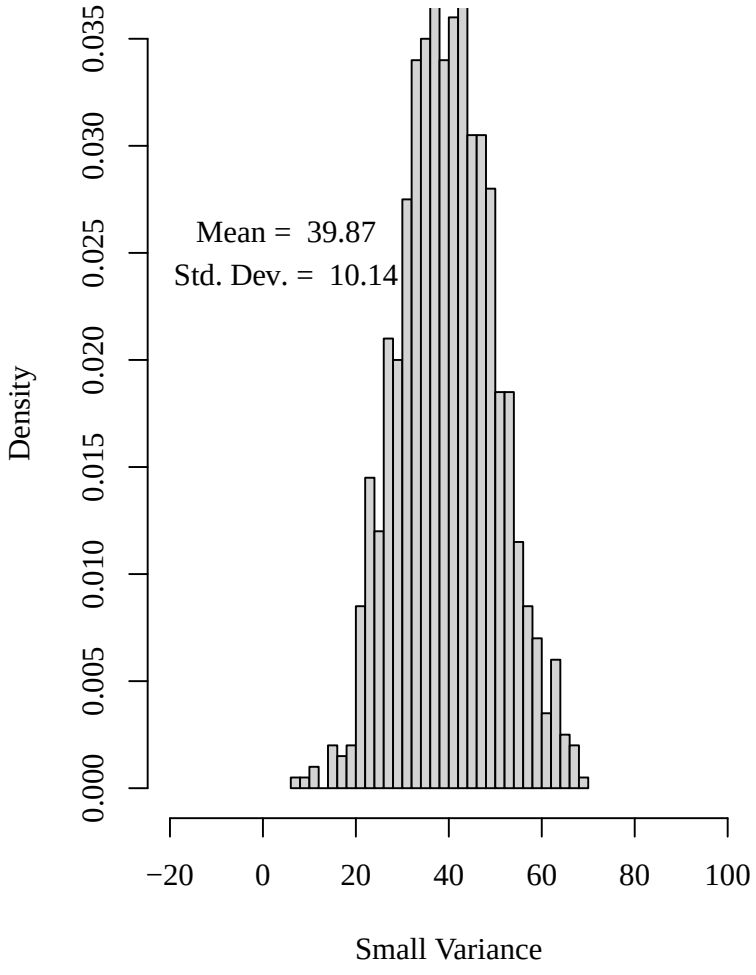
Standard Deviation

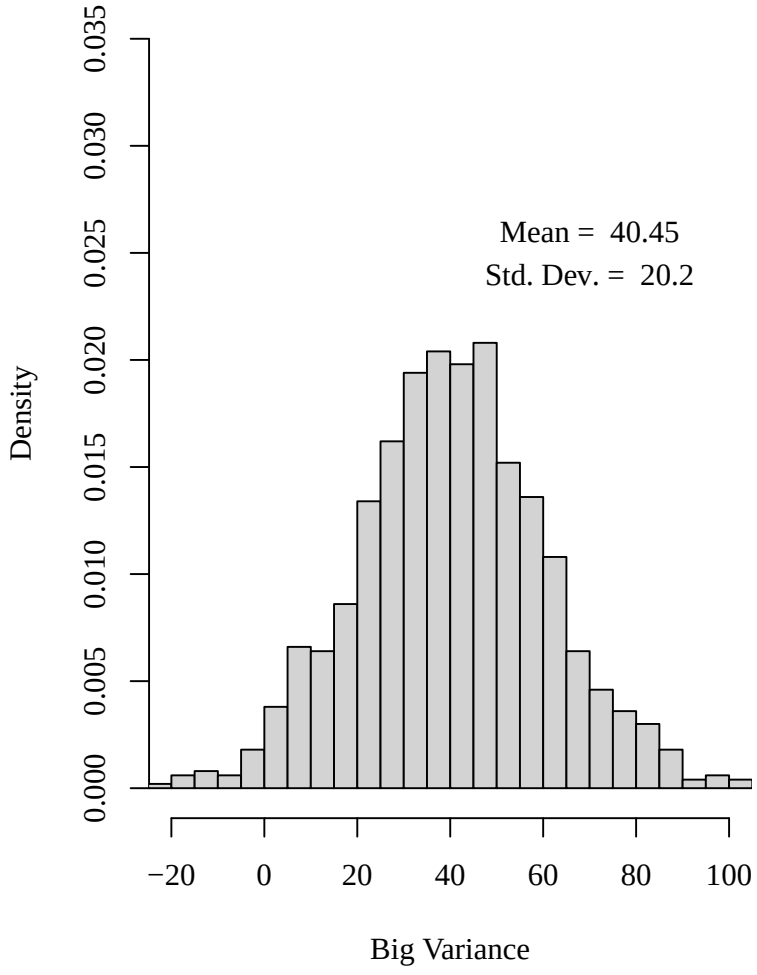
Standard Deviation: the square root of the variance.

$$Std.Dev.(x) = \sqrt{Var(x)} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{10}$$

- Var and Std.Dev. serve same purpose.
- Std.Dev. has an advantage: it is measured (roughly speaking) on the same scale as the mean. (see below on "scaling")

Compare 2 variables





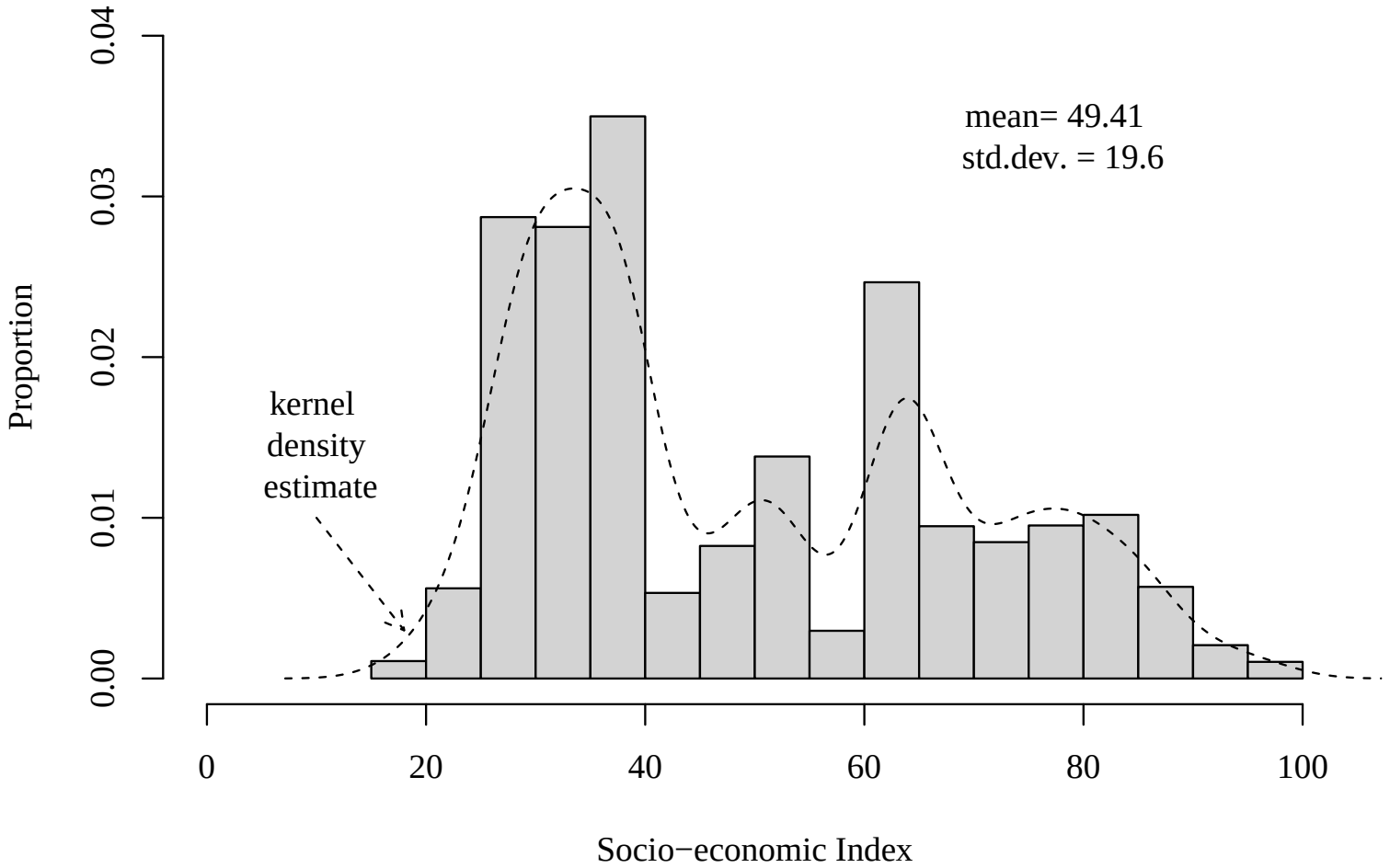
About Notation

Call the observed variance what you want!

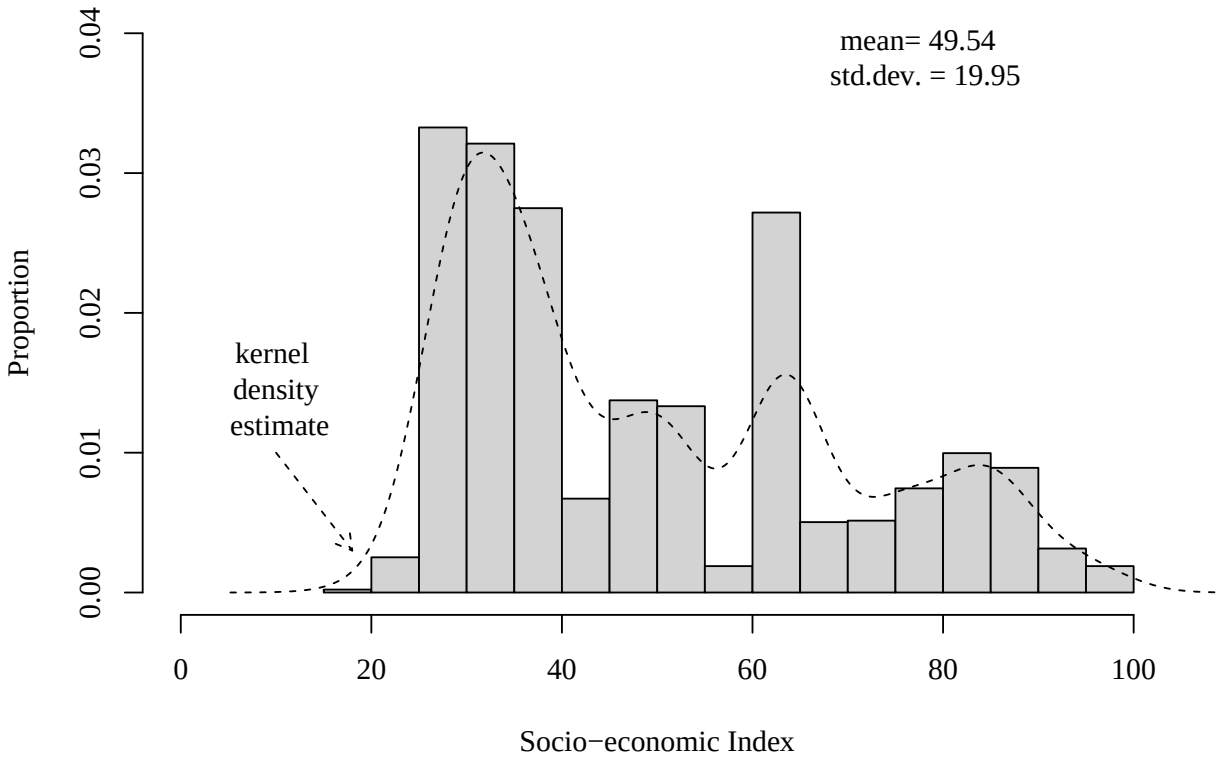
1. Older stats books
   1. σ^2 is the "true (unknown) variance
   2. s^2 is the estimate. "s" short for "sigma"
2. I like the hat notation, estimated variance $\widehat{\sigma^2}$.
3. I also like notation with the "true" variance $\mathrm{Var}(x)$ while an estimate is $\widehat{\mathrm{Var}(x)}$.

You are allowed to use any symbol you prefer, as long as you are clear to define the term when you use it.

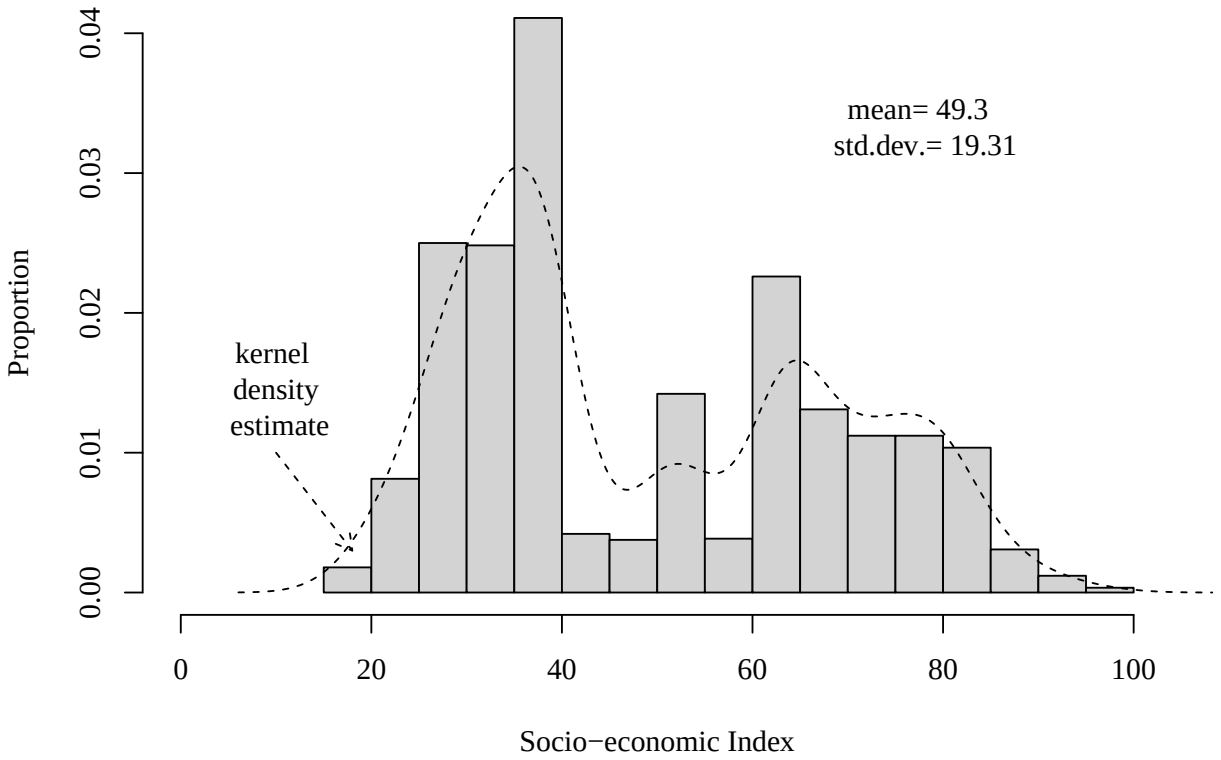
Socio Economic Status



Socio Economic Status: Only Men



Socio Economic Status: Women



Other Diversity Indicators

- Inter-Quartile range: group data by ordered quarters, and then think of the range between 25 percentile and 75 percentile as a diversity indicator.
- Many possible diversity indicators, including
  - gini index (often used for income inequality)
  - the mean of absolute valued differences

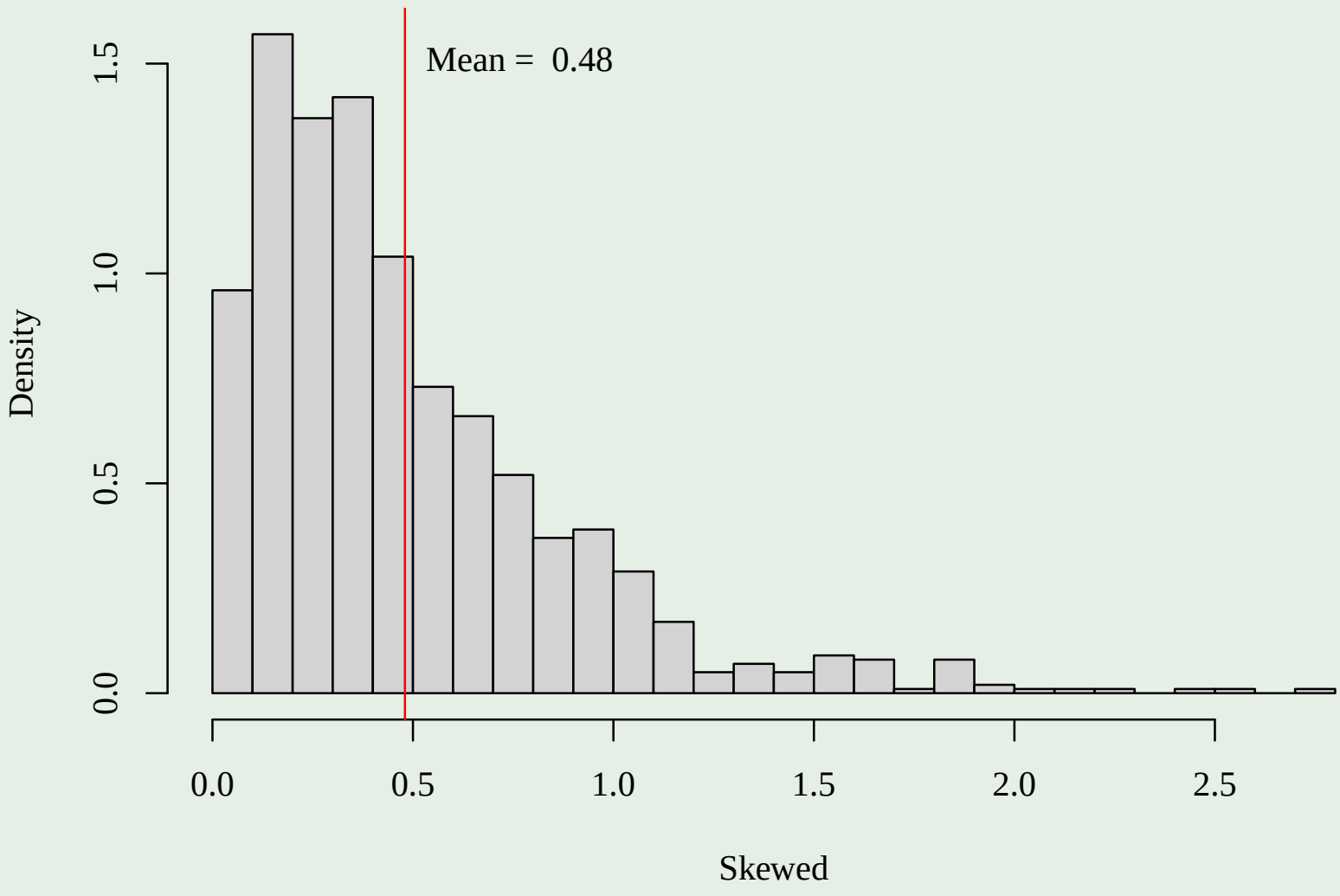
$$Mean\ Absolute\ Deviation = \frac{\sum_{i=1}^{N} |x - \bar{x}|}{N} \tag{11}$$

Outline

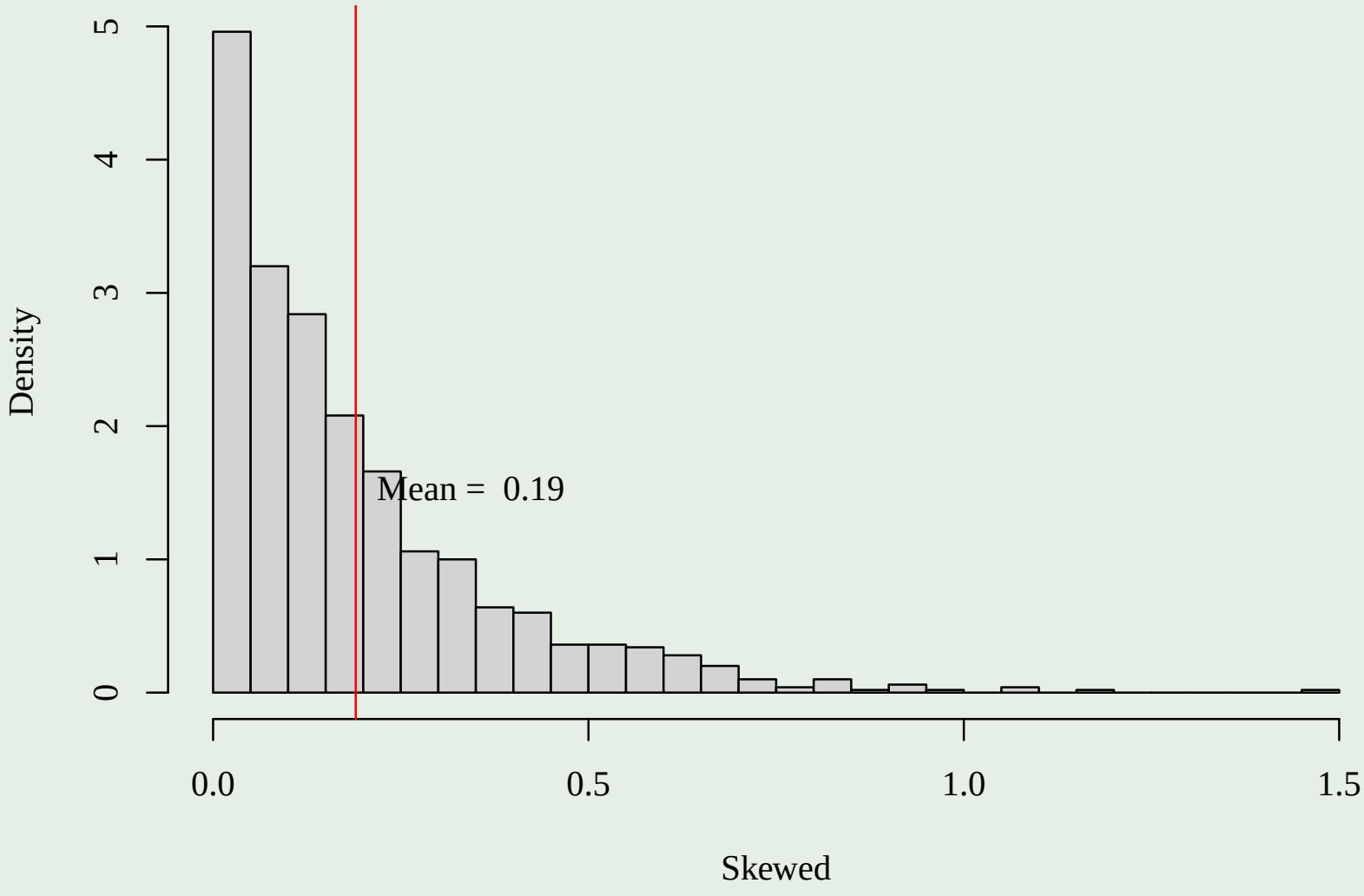
Symmetry Definition

- A distribution is symmetric if the chance of observing a score $\bar{x} - c$ is the same as observing $\bar{x} + c$.
- If a distribution is symmetric, then we have no trouble conveying the idea of its 'location'.
- The mean is in the middle!

A Nonsymmetric Distribution



Another Nonsymmetric Distribution



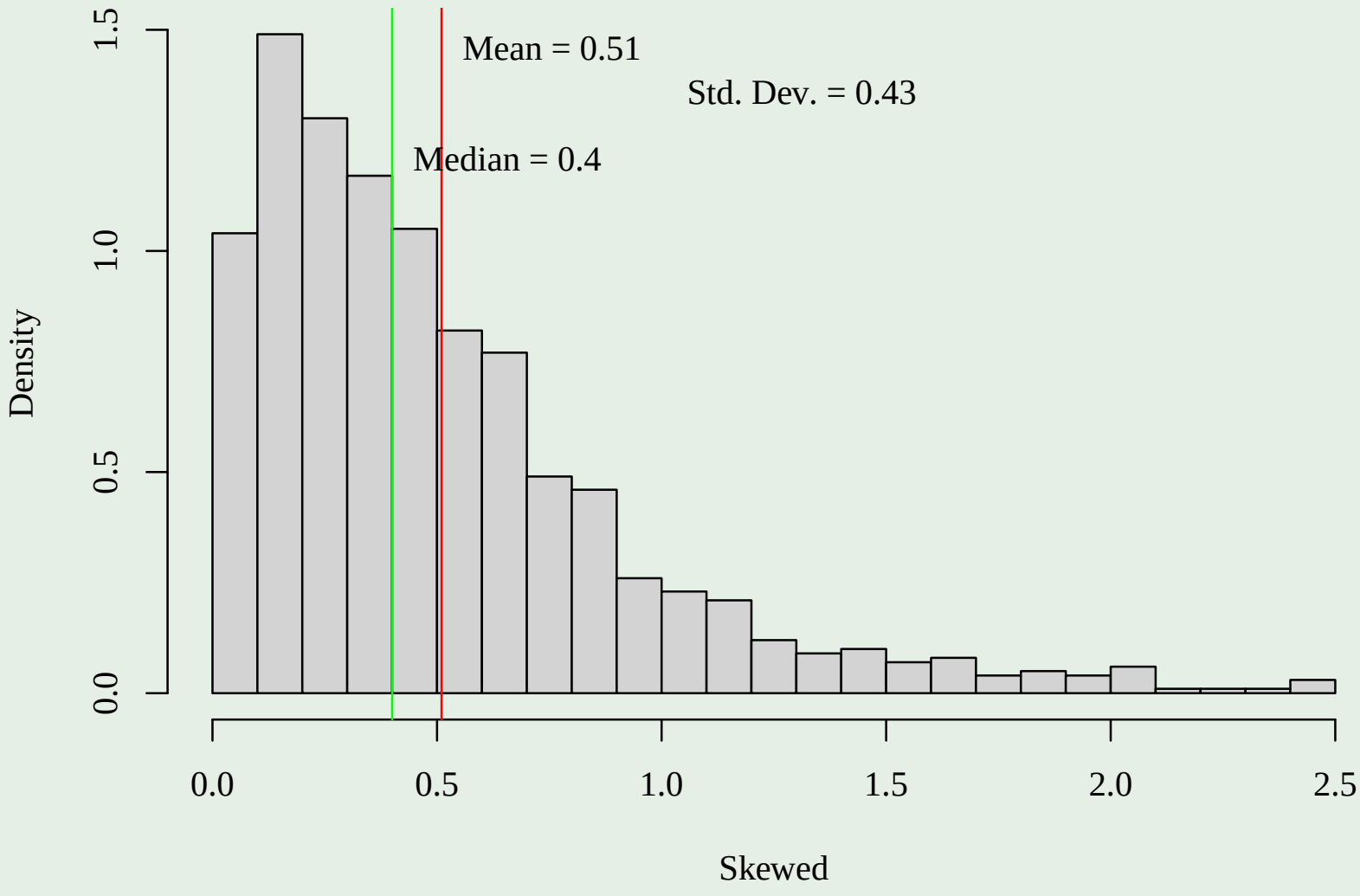
Median: Center Case

Median: The "center observation," the number of observations that are larger equals the number that is smaller.

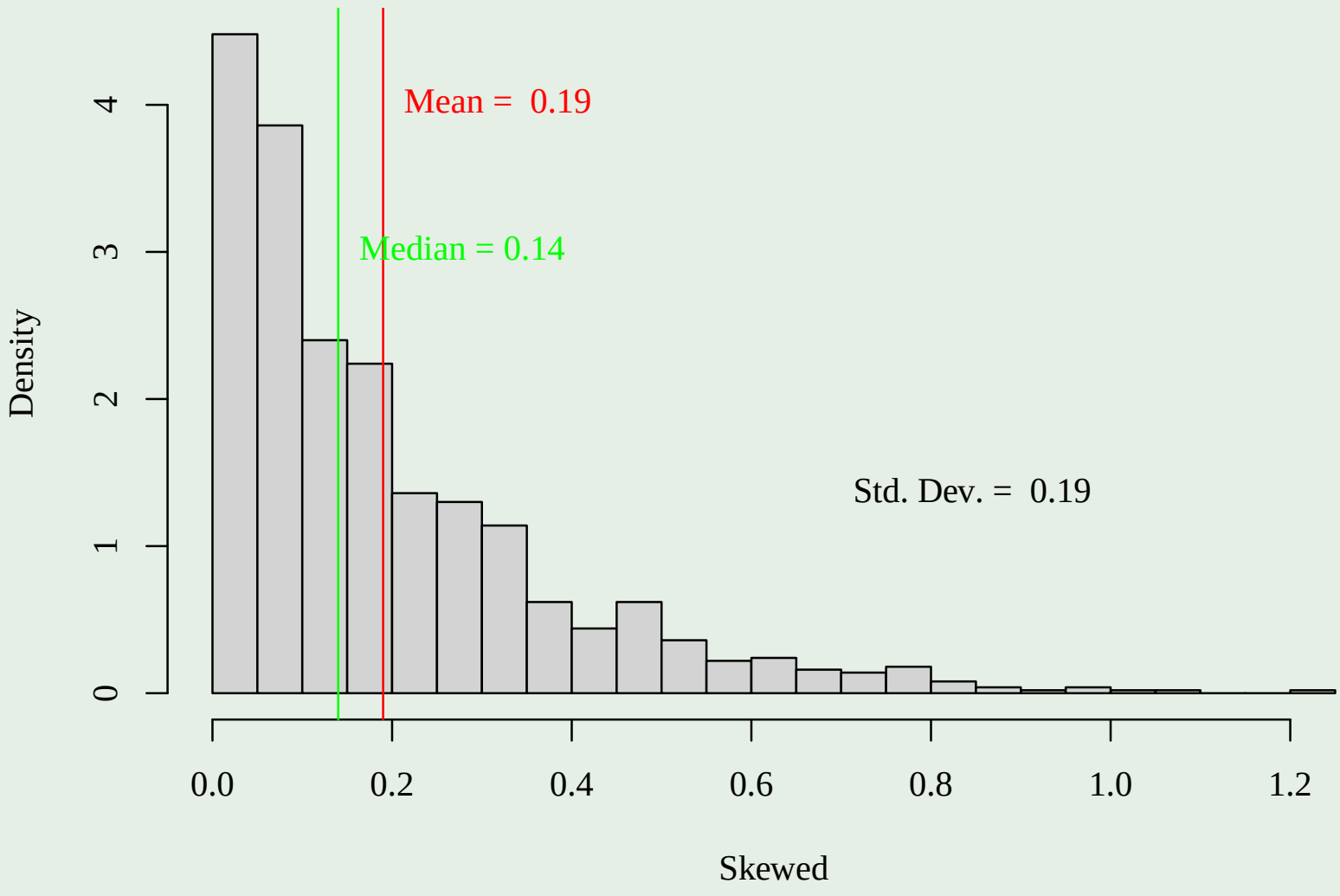
Questions:

1. When do you think the mean and median are likely to be the same?
2. Can you think of a situation in which the median may be more meaningful than the mean?

Add the median. Helpful?



Another Nonsymmetric Distribution



When To Emphasize The Mode

If lots of observations are clumped up at one point, it is worth noting!

Suppose I collected data like this:

$$X = \{1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 50\}$$

If almost all of the scores are "2", we should tell the reader. Note about Level of Measurement

- Mean only useful if we have numerical data (silly to average "low", "medium", "high")
- Median requires ordered data, either numerical or ordered categorical
- Problem with the mean: it is distorted by a change in one value on either side (change one 50 to 5,000,000 and note the mean changes)
- Median is a more "robust" estimate (jargon: high 'breakdown point')

Outline

Should the Scale Matter?

- The temperature in Celsius is 10. The temperature in Farenheit is 50 (32+9/5*10).
- My income in dollars is 68,000. My income in Euros is 43,000 and in Pesos it is 1,126,123.
- Sometimes, we receive data in one format, but convert to another
- Simple scale conversions SHOULD NOT substantively change the conclusions we will draw.
- If simple scale conversions seem to matter, be VERY cautious.

The Mean Scales With The Data

- Take variable $X = \{x_1, x_2, \ldots, x_N\}$, and multiply each value by 10 to create $newx$

$$newx = \{10x_1, 10x_2, \ldots, 10x_N\} \tag{12}$$

- The mean of $newX$ is obviously 10 the mean of old x. See?

$$Mean(newX) = \frac{10x_1 + 10x_2 + \ldots + 10x_N}{N} = 10\frac{\sum_{i=1}^{N} x_i}{N}$$

$$Mean(newX) = \overline{newX} = 10 \times \bar{x}$$

- Generally (meaning always), the mean of $(k \times X)$ is equal to k times the mean of X.

My First Big Fact

- State that as a theorem. k_1 and k_2 are any non-zero constants. X is any variable. Create a new variable $newX = k_1 + k_2 X$

The Mean scales proportionally. Given constants k_1, k_2

$$Mean(k_1 + k_2 X) = k_1 + k_2 \times Mean(X) \tag{13}$$

- The point: The Mean changes in a completely predictable way when the data is re-scaled by addition and multiplication. Just apply same same re-scaling to the old mean.

The Variance Doesn't Scale Proportionally

- Suppose variance of X is $var(X)$
- Create $newX$ by multiplying by 10, $newX = 10 \cdot X$
- The variance of $newX$ is $10^2 Var(X)$

General Result for Variance of Re-scaled Variables

Calculate the Variance of a re-scaled Variable, X. Given k_1 , k_2

$$Var(k_1 + k_2 \cdot X) = k_2^2 \cdot Var(X) \tag{14}$$

- Adding k_1 does not change the dispersion at all, it just shifts the scores.
- The variance of $newX = k_1 + k_2X$ is $k_2^2 \times Var(x)$

Implication: Don't re-calculate mean and variance if is proportionally re-scaled.

- Celsius temperature data, x. Suppose the mean is, 100.
- Rescale that data to Fahrenheit

$$xF_i = 32 + \frac{9}{5}x_i \tag{15}$$

- Some students want to re-run xF_i through the mean function, but they don't need to.
- The mean of xF is $32 + (9/5)Mean(x) = 32 + (9/5)100 = 212$.

But the Standard Deviation Scales Proportionally!

- The variance of xF is $(9/5)^2 \times Var(x)$, which is NOT linear
- However, recall standard deviation is $\sqrt{Var(x)}$, so the standard deviation would be

$$Std.Dev.(xF) = \sqrt{(9/5)^2 \times Var(x)} = (9/5) \times Std.Dev.(x) \quad (16)$$

- Like the mean, the standard deviation scales proportionally.

Standard Deviation of kX is $k \times Std.Dev.(X)$

$$Std.Dev(k \cdot X) = k \cdot Std.Dev(X) \quad (17)$$

The ratio $mean/std.dev.$ is Also Scale Invariant

- Recall $Mean(k \cdot x) = kMean(x)$
- And $Std.Dev.(k \cdot x) = kStd.Dev.(x)$
- Then the ratio of the mean to the standard deviation is not affected by k

$$\frac{Mean(k \cdot x)}{Std.Dev.(k \cdot x)} = \frac{k \cdot Mean(x)}{k \cdot Std.Dev.(x)} = \frac{Mean(x)}{Std.Dev.(x)} \tag{18}$$

- And the converse is also true

$$\frac{k \cdot Std.Dev(x)}{k \cdot Mean(x)} = \frac{Std.Dev(x)}{Mean(x)} \tag{19}$$

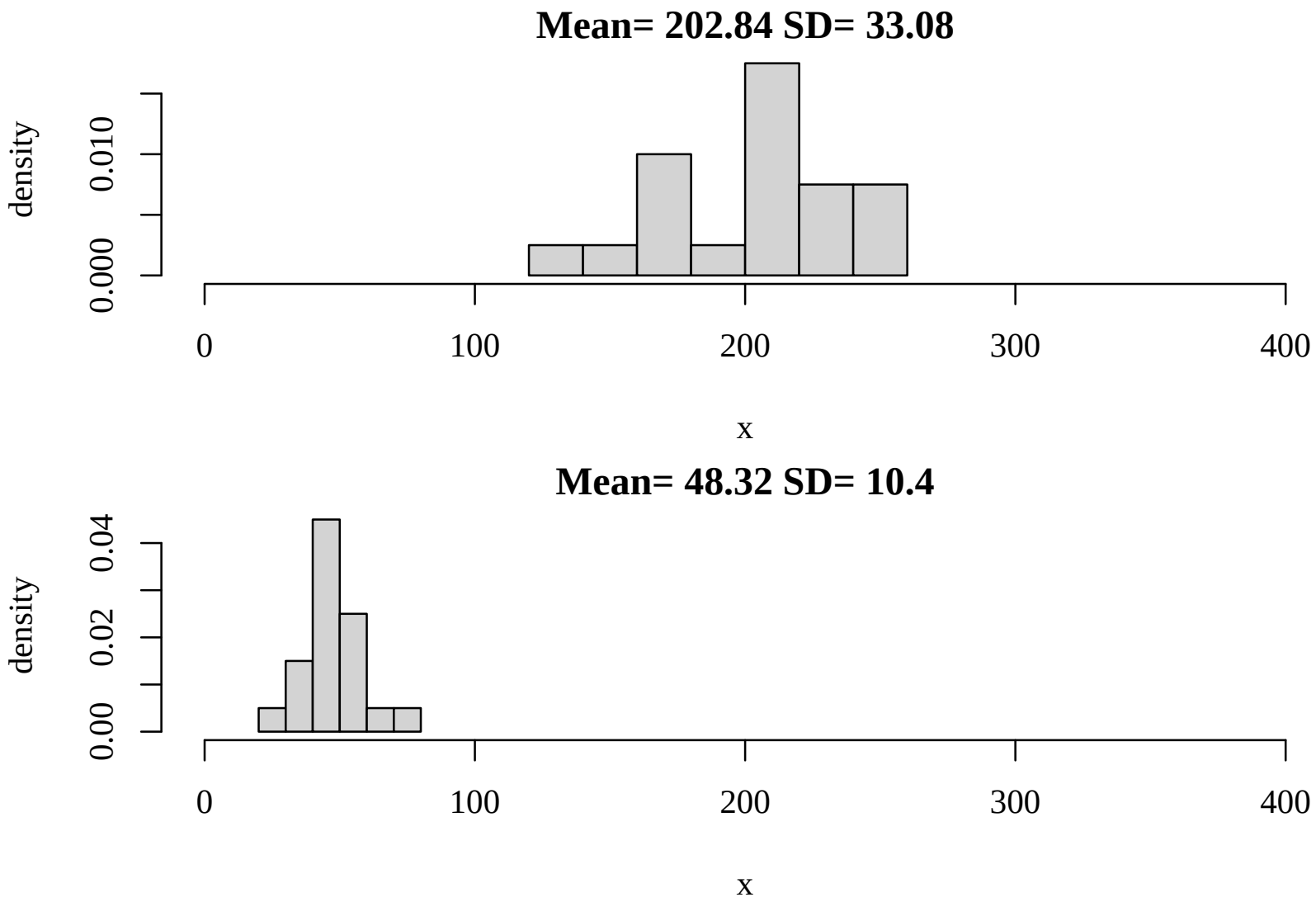
Coefficient of Variation is Std.Dev(x)/M(x)

Coefficient of variation, CV.

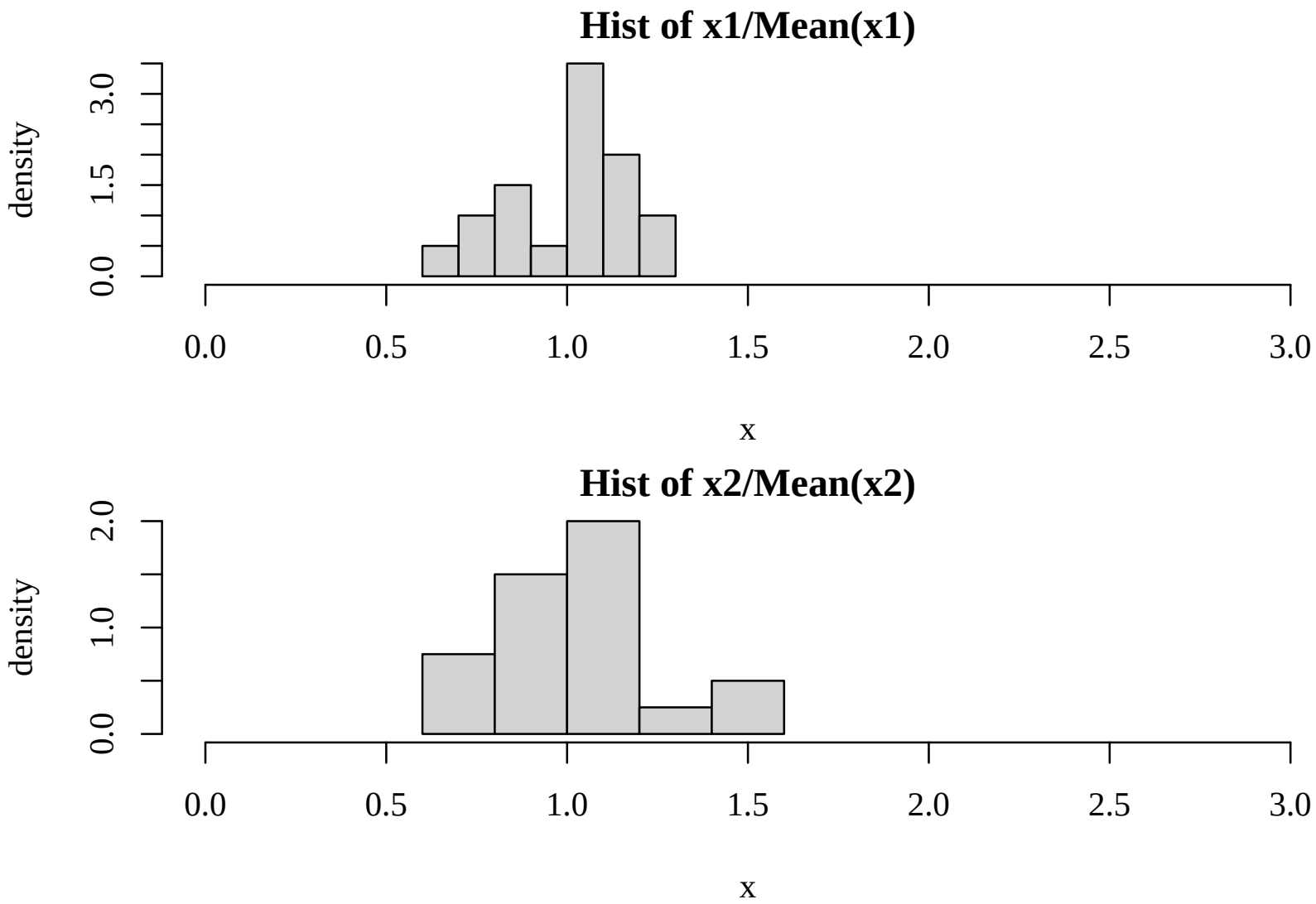
Question: is "this distribution" more "spread out" than "that one"?

- This is a difficult, possibly silly question when distributions are fundamentally different
- But, if they have roughly the same "shape", then the re-scaling might make them comparable.

Compare dispersion of 2 disparate variables



Compare 2: plot x/Mean(x)



Summarize those 2 variables

```
                      top      bottom
Min.           129.8727392 29.6653058
1st Qu.        174.5457051 43.6833138
Median         213.9866159 48.9267030
Mean           202.8352460 48.3206232
3rd Qu.        225.8233332 50.6007836
Max.           246.5122581 73.1726698
mean           202.8352460 48.3206232
sd              33.0755854 10.3983793
sd.over.mean     0.1630663  0.2151955
```

Outline

Mean-center x_i

- Mean centered data (aka "data in deviations form")

$$MeanCentered(x_i) = x_i - Mean(x_i) \tag{20}$$

- Do we need abbreviation for that? x_i^{MC} or $\widetilde{x}_i$ or ?
- The mean of a centered variable is always 0
- The variance and standard deviation are unchanged by centering
- Sometimes mean-centered data may faciliate interpretation of results.

Standardized Variables

Standardized Variables.

- Standardize means "divide $MeanCentered(x_i)$ by standard deviation".

$$\frac{x_i - \bar{x}}{\sigma_x} \tag{21}$$

- Since $M(x)/Std.Dev(x)$ is scale invariant, it makes $MeanCentered(x_i)/Std.Dev(x)$ will also be unaffected by re-scaling of the observations.
- The letter "Z" is often used to refer to standardized variables.

Standardized implies Mean 0, Std.Dev 1

- Mean

$$Mean\,of\,Z = \bar{Z} = M(Z) = \mu_Z = 0$$

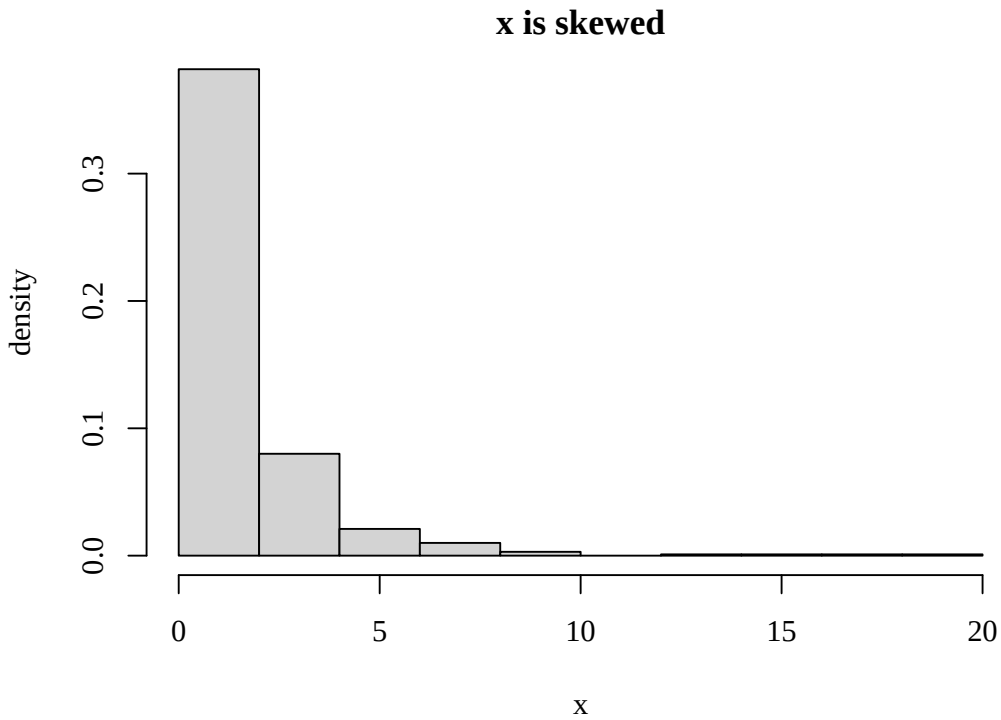
- Standard deviation

$$Std.Dev.(Z) = SD(Z) = \sigma_Z = 1$$

- Standardization helps with some "machine learning" procedures, may help psychologists compare variables

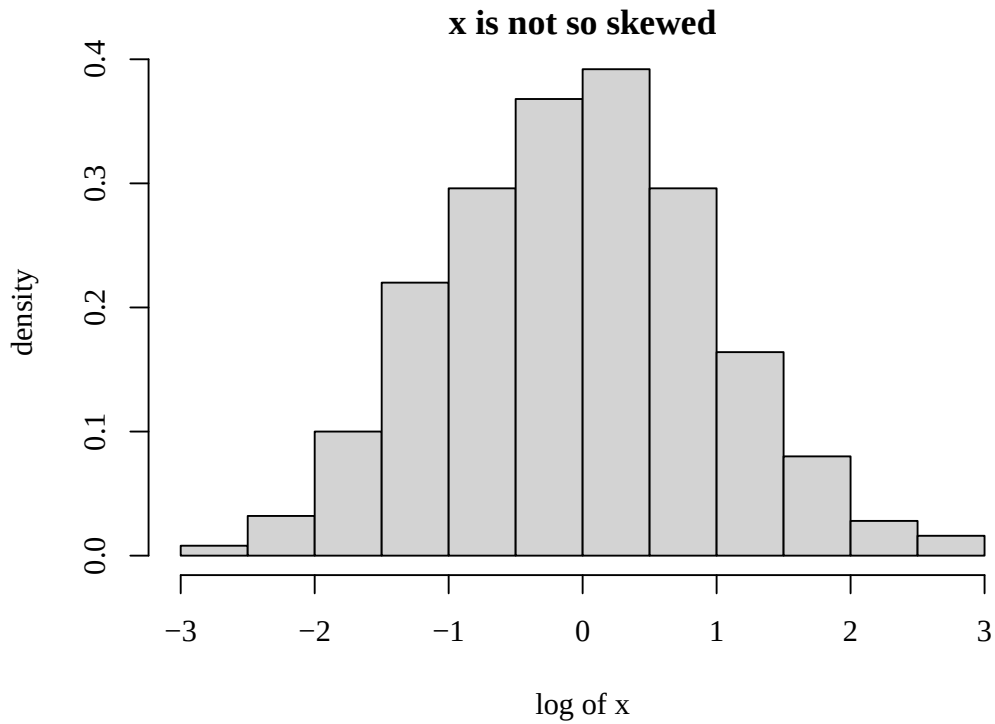
The log is the most commonly applied nonlinear transformation

- We often gather data that is "clumped" on the left
- Examples, income, education



The log is the most commonly applied nonlinear transformation

- The distribution of log(x) appears more symmetric



Difficult to say for sure if logging a variable is good or bad

- Some methods books will recommend logging all variables, claiming that it almost always makes analysis "work better" in some sense.
- Please just remember it is a possibilty

R functions to remember

x is a variable

- mean(x, na.rm = TRUE)
- sd(x, na.rm = TRUE)
- var(x, na.rm = TRUE)
- median(x, na.rm = TRUE)
- range(x, na.rm = TRUE)
- quantile(x, na.rm = TRUE)
- summary(x)
- rockchalk::summarize(x)
- hist(x, prob = TRUE)
- xdens <- density(x)
- lines(xdens)
- plot(xdens)

References

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing.

Session

Output from R (R Core Team (2020))

```
sessionInfo()
```

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

```

Session ...

```
other attached packages:
[1] stationery_0.98.30

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.4.6     digest_0.6.25    plyr_1.8.6       xtable_1.8-4
 [5] evaluate_0.14    zip_2.0.4        rlang_0.4.7      stringi_1.4.6
 [9] openxlsx_4.1.5   rmarkdown_2.3    tools_4.0.2      foreign_0.8-79
[13] kutils_1.70      xfun_0.15        compiler_4.0.2   htmltools_0.5.0
[17] knitr_1.29
```