Exploratory Data Analysis – Part 1

MSA 2023

First things first! Explore your data

- What kind of variables do you have?
- What do their distributions look like?
- Are there any anomalies?
- Do they have any interesting associations?

First things first! Explore your data

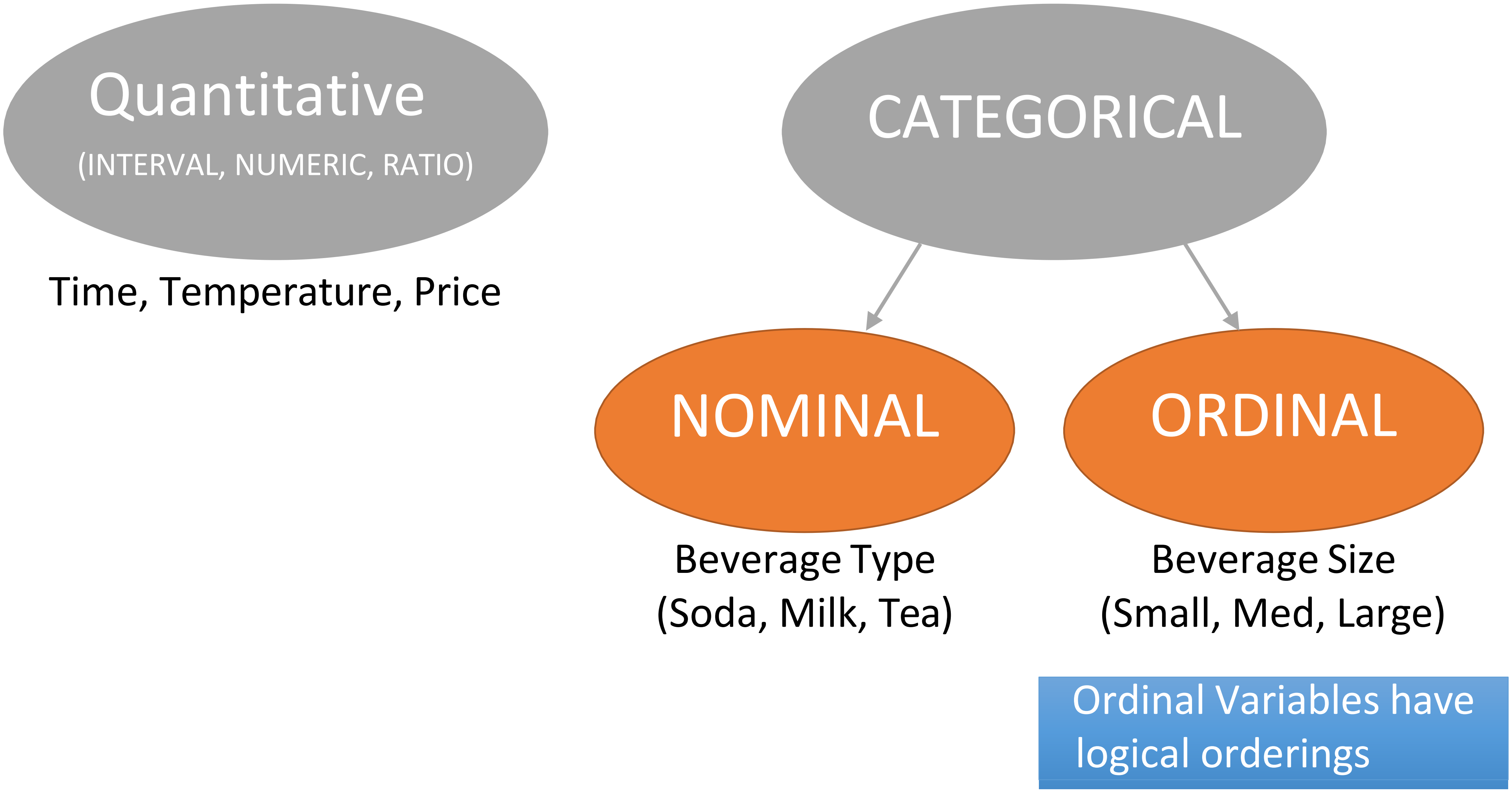
- **What kind of variables do you have?**
- What do their distributions look like?
- Are there any anomalies?
- Do they have any interesting associations?

Quantities or Qualities of Interest

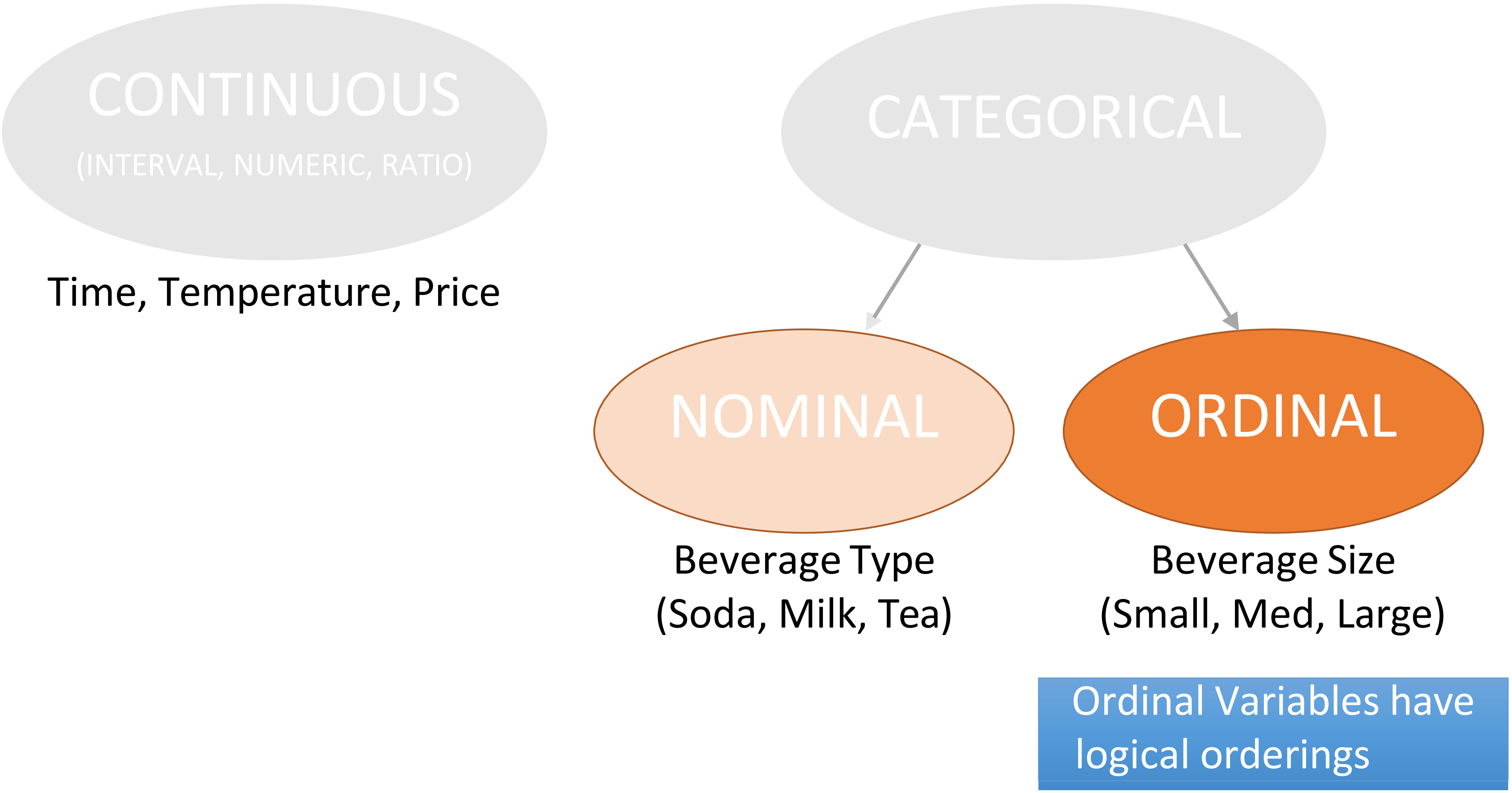
Columns of data equivalently called:

- Variables
- Attributes
- Features
- Predictors/Targets
- Factors
- Inputs/Outputs
- Covariates

Types of Attributes



Types of Attributes



Ordinal Variables

Ordinal variables are treated as *either* **categorical** *or* **quantitative.**

- The levels become dummy variables if treated **categorically**:

Size
S
M
L

Small	Med.	Large
1	0	0
0	1	0
0	0	1

- The levels are given values if treated **quantitatively**:

Size
S
M
L

$\longrightarrow$

Size
1
2
3

Ordinal Variables

Ordinal variables are treated as *either* **categorical** *or* **quantitative.**

- The levels become dummy variables if treated **categorically**:

Size
S
M
L

Small	Med.	Large
1	0	0
0	1	0
0	0	1

- The levels are given values if treated **quantitatively**:

Size
S
M
L

Size
1
2
3

BUT, these values do not have to be the integers 1,2,3... other techniques like *optimal scaling* are used to find the "optimal" values for each level based on linearity.

Optimal Scaling

Primary idea: Ordinal variables need not be equally spaced levels in terms of the target.

Example: Consider the effect of education level on salary. Estimate the salary difference you'd expect to find between individuals with different education levels, all else constant.

- "No HS Degree" vs. "GED"
- "Bachelor's Degree" vs. "Master's Degree"

Reasonable to expect a bigger salary increase stemming from Master's degree vs. a GED. Requires a careful definition of a "1-unit" change in education.

Optimal Scaling and Target Level Encoding

Example:

- "No HS Degree" vs. "GED"
- "Bachelor's Degree" vs. "Master's Degree"

Reasonable to expect a bigger salary increase stemming from Master's degree vs. a GED. Requires a careful definition of a "1-unit" change in education.

Education	Education
No HS degree	1
GED	2
HS diploma	3
Bachelors	10
Masters	16
PhD	20

Doesn't have to be arbitrary values.

Could use the actual expected increase in Salary using the training data to create this valuation! (Target level encoding)

The Ames Real Estate Data Set

```r
install.packages("AmesHousing")
library(AmesHousing)
```

```r
ames <- make_ordinal_ames()
```

Display structure of any R object

```r
str(ames)
```

```
$ MS_SubClass        : Factor w/ 16 levels "One_Story_1946_and_Newer_All_Styles",..: 1 1 1 1 6 6 12 12 12 6 ...
$ MS_Zoning          : Factor w/ 7 levels "Floating_Village_Residential",..: 3 2 3 3 3 3 3 3 3 3 ...
$ Lot_Frontage       : num [1:2930] 141 80 81 93 74 78 41 43 39 60 ...
$ Lot_Area           : int [1:2930] 31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
$ Street             : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
$ Alley              : Factor w/ 3 levels "Gravel","No_Alley_Access",..: 2 2 2 2 2 2 2 2 2 2 ...
$ Lot_Shape          : Ord.factor w/ 4 levels "Irregular"<"Moderately_Irregular"<..: 3 4 3 4 3 3 4 3 3 4 ...
$ Land_Contour       : Ord.factor w/ 4 levels "Low"<"HLS"<"Bnk"<..: 4 4 4 4 4 4 4 2 4 4 ...
$ Utilities          : Ord.factor w/ 4 levels "ELO"<"NoSeWa"<..: 4 4 4 4 4 4 4 4 4 4 ...
$ Lot_Config         : Factor w/ 5 levels "Corner","CulDSac",..: 1 5 1 1 5 5 5 5 5 5 ...
$ Land_Slope         : Ord.factor w/ 3 levels "Sev"<"Mod"<"Gtl": 3 3 3 3 3 3 3 3 3 3 ...
$ Neighborhood       : Factor w/ 29 levels "North_Ames","College_Creek",..: 1 1 1 1 7 7 17 17 17 7 ...
$ Condition_1        : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 3 3 3 ...
$ Condition_2        : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 3 ...
$ Bldg_Type          : Factor w/ 5 levels "OneFam","TwoFmCon",..: 1 1 1 1 1 1 5 5 5 1 ...
$ House_Style        : Factor w/ 8 levels "One_and_Half_Fin",..: 3 3 3 3 8 8 3 3 3 8 ...
$ Overall_Qual       : Ord.factor w/ 10 levels "Very_Poor"<"Poor"<..: 6 5 6 7 5 6 8 8 8 7 ...
$ Overall_Cond       : Ord.factor w/ 10 levels "Very_Poor"<"Poor"<..: 5 6 6 5 5 6 5 5 5 5 ...
```

Describing Distributions Part 1: Quantification

Statistics measuring location, spread, and shape

First things first! Explore your data

- What kind of variables do you have?
- **What do their distributions look like?**
- **Are there any anomalies?**
- Do they have any interesting associations?

Describing Distributions

- Center/Location
- Spread/Variation
- Shape
- Anomalous Observations

Describing Distributions

- **Center/Location**
- Spread/Variation
- Shape
- Anomalous Observations

Measures of Central Tendency

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

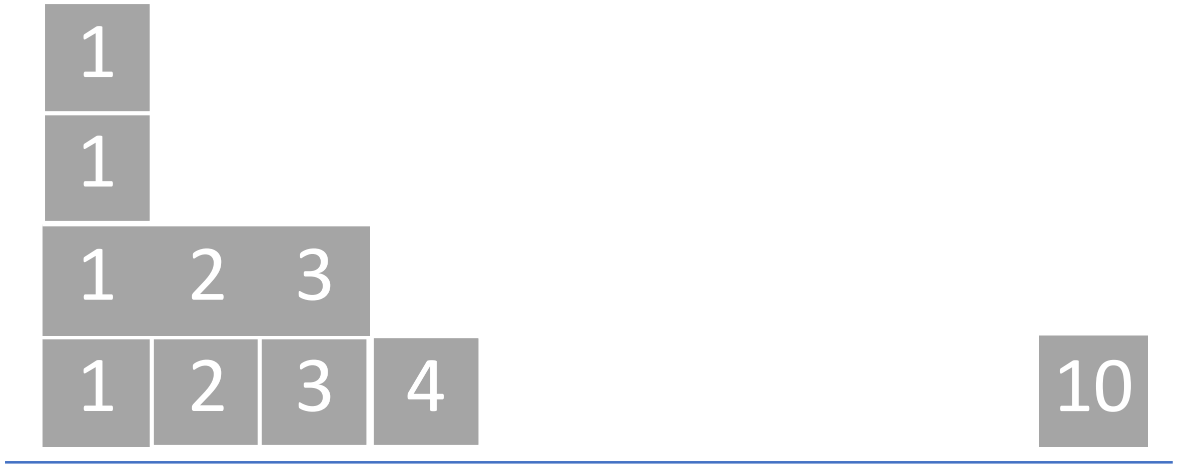
Average value.
Affected by outliers.

Median

Middle value.
50th Percentile.
Unaffected by outliers.

Mode

Most frequent value.
Typical for categorical data.



Measures of Central Tendency

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

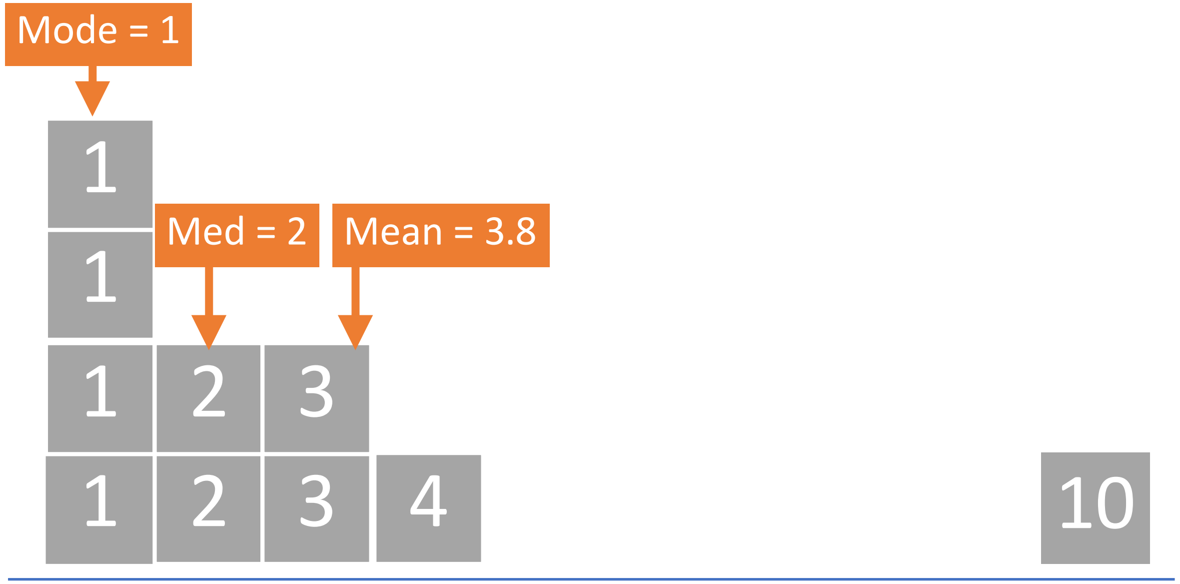
Average value.
Affected by outliers.

Median

Middle value.
50th Percentile.
Unaffected by outliers.

Mode

Most frequent value.
Typical for categorical data.



Measures of Location

Percentiles

A point, x_p, in your data (or on its range) for which p% of the data is $\leq x_p$

98		
95		
92	92	Third Quartile (75^{th} percentile)
90		
85		
81	83	Median (50^{th} percentile)
79		
70	70	First Quartile (25^{th} percentile)
70		
55		

Describing Distributions

- Center/Location
- **Spread/Variation**
- Shape
- Anomalous Observations

Measures of Spread/Dispersion

Range

Difference between
the minimum and maximum
data values

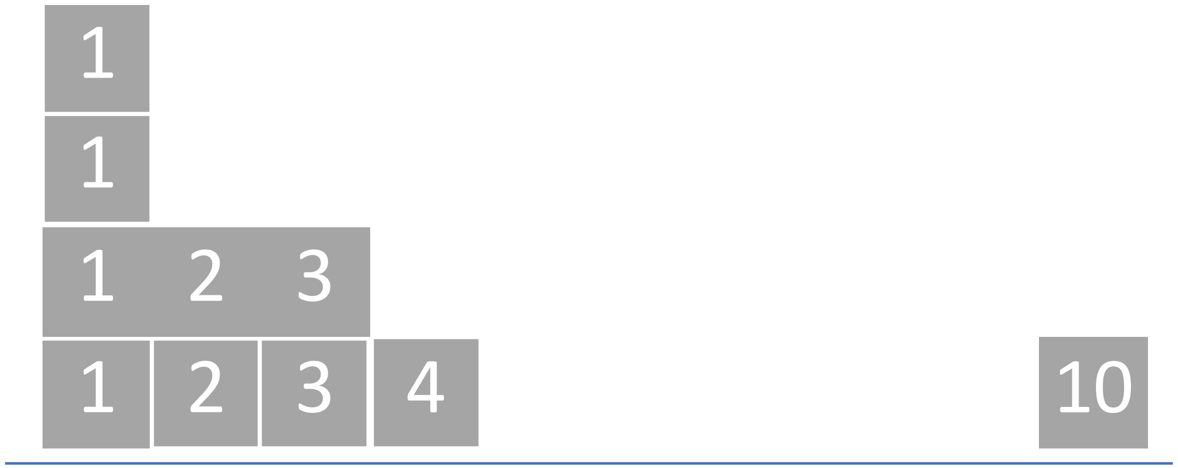
Interquartile Range (IQR)

Difference between
third and first quartile.

Variance (σ^2) and Standard Deviation ( )

Dispersion of the data
around the mean

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



Measures of Spread/Dispersion

Range

Difference between
the minimum and maximum
data values

Interquartile Range (IQR)

Difference between
third and first quartile.

Variance (σ^2) and Standard Deviation ( )

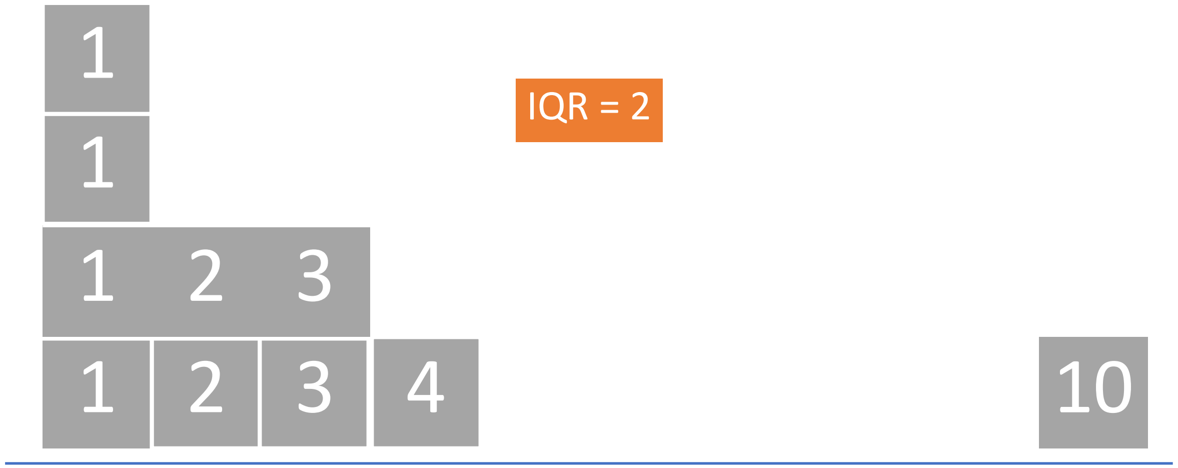
Dispersion of the data
around the mean

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Range = 9

IQR = 2

Variance = 7.51

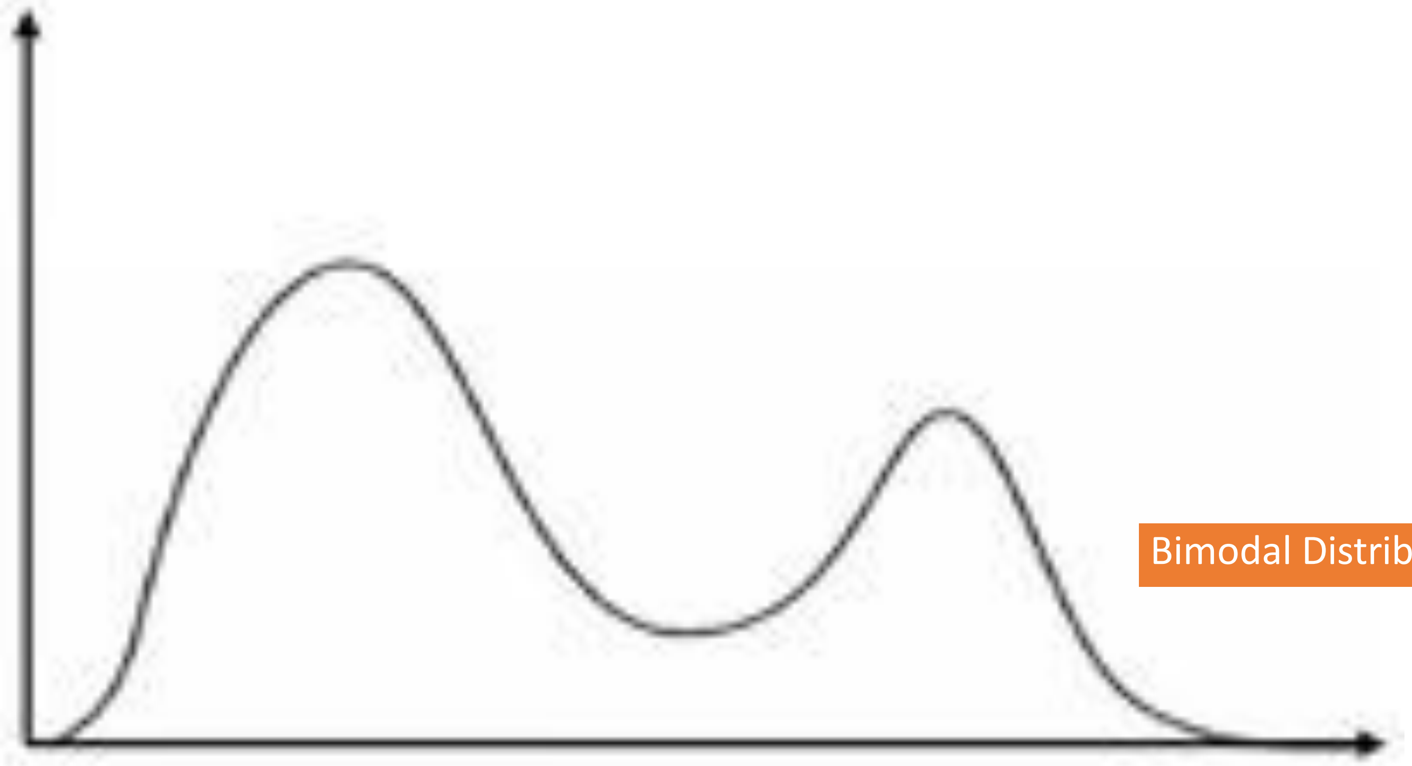


Describing Distributions

- Center/Location
- Spread/Variation
- **Shape**
- Anomalous Observations

Measures of Shape

Modality — Is the distribution unimodal? Bimodal? More Complicated?

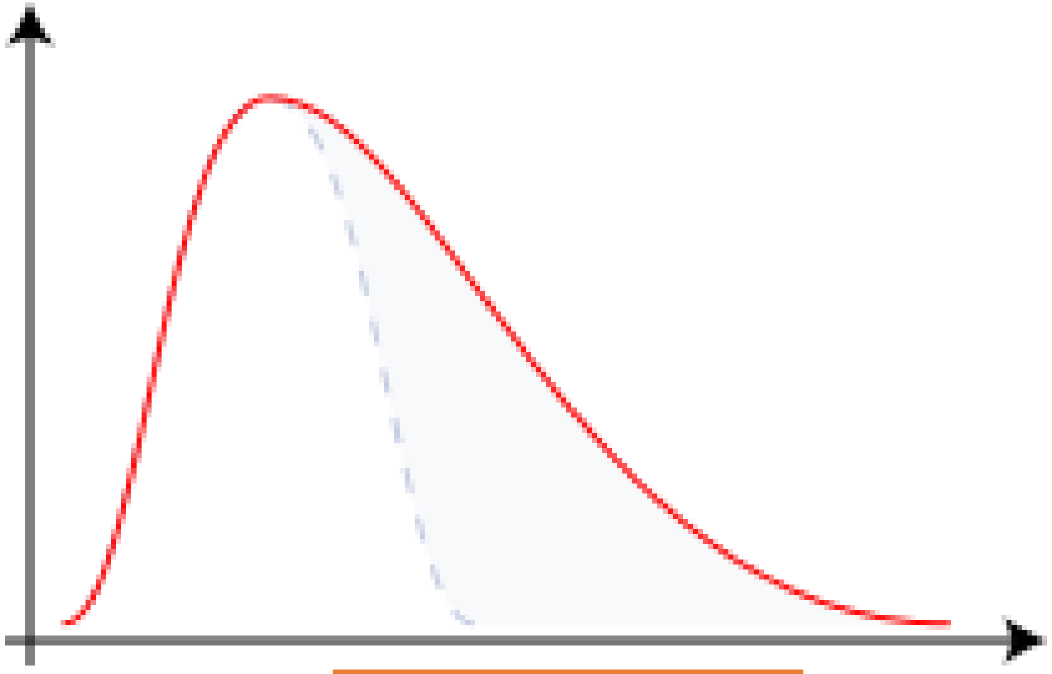


Bimodal Distribution

Measures of Shape

Skew Is the distribution symmetric? Or does it have a longer tail on one side?

Left-Skewed Distribution



Right-Skewed Distribution

Measures of Shape

Kurtosis Does the distribution have thicker/thinner tails than a normal distribution with same mean and variance?

Leptokurtic Distribution

More data in the tails than a normal distribution.

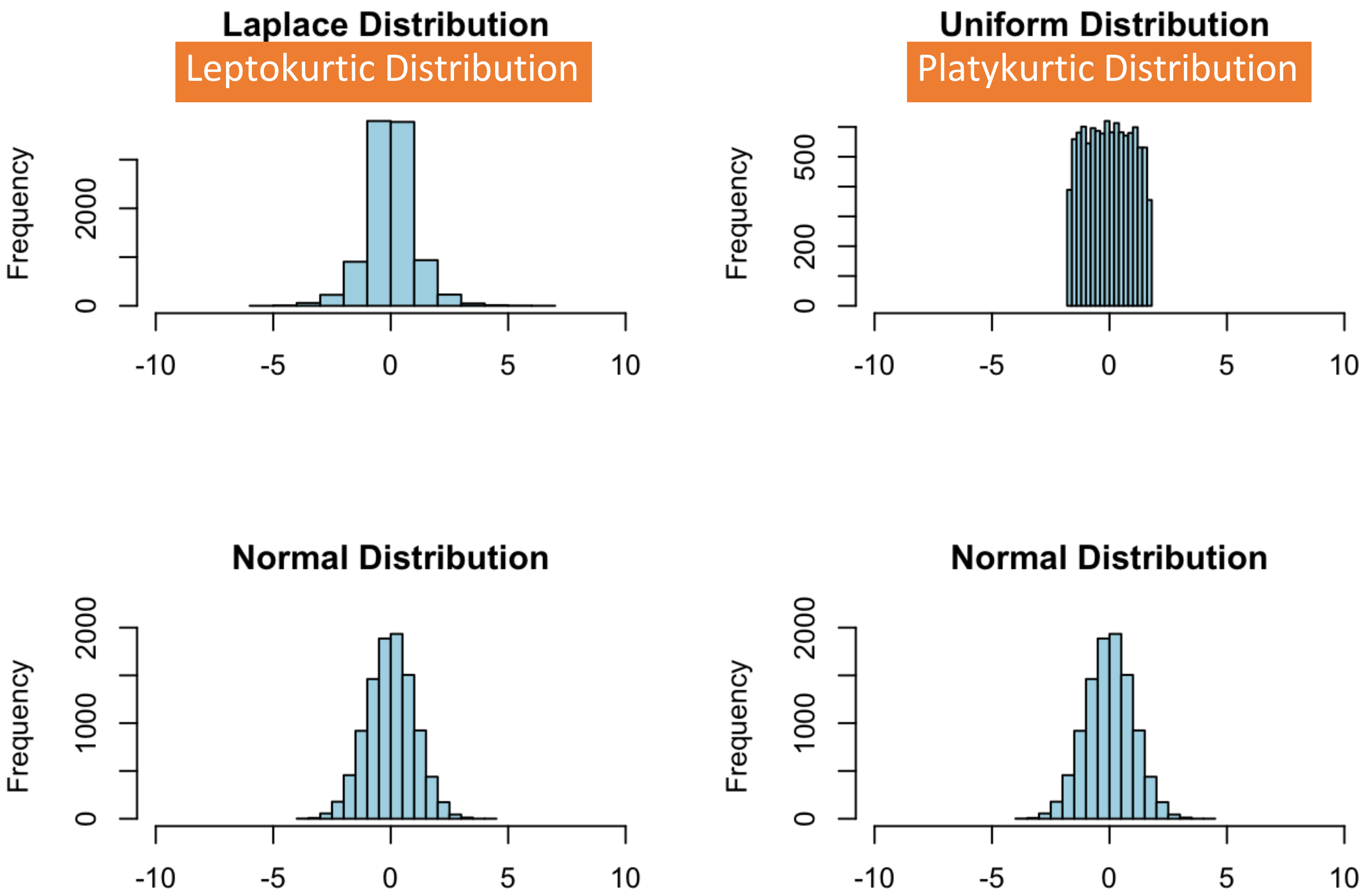
Platykurtic Distribution

Less data in the tails than a normal distribution.

Kurtosis has to do with the *asymptotic* behavior of a distribution - how quickly does the density tend toward zero as we approach +/- ?

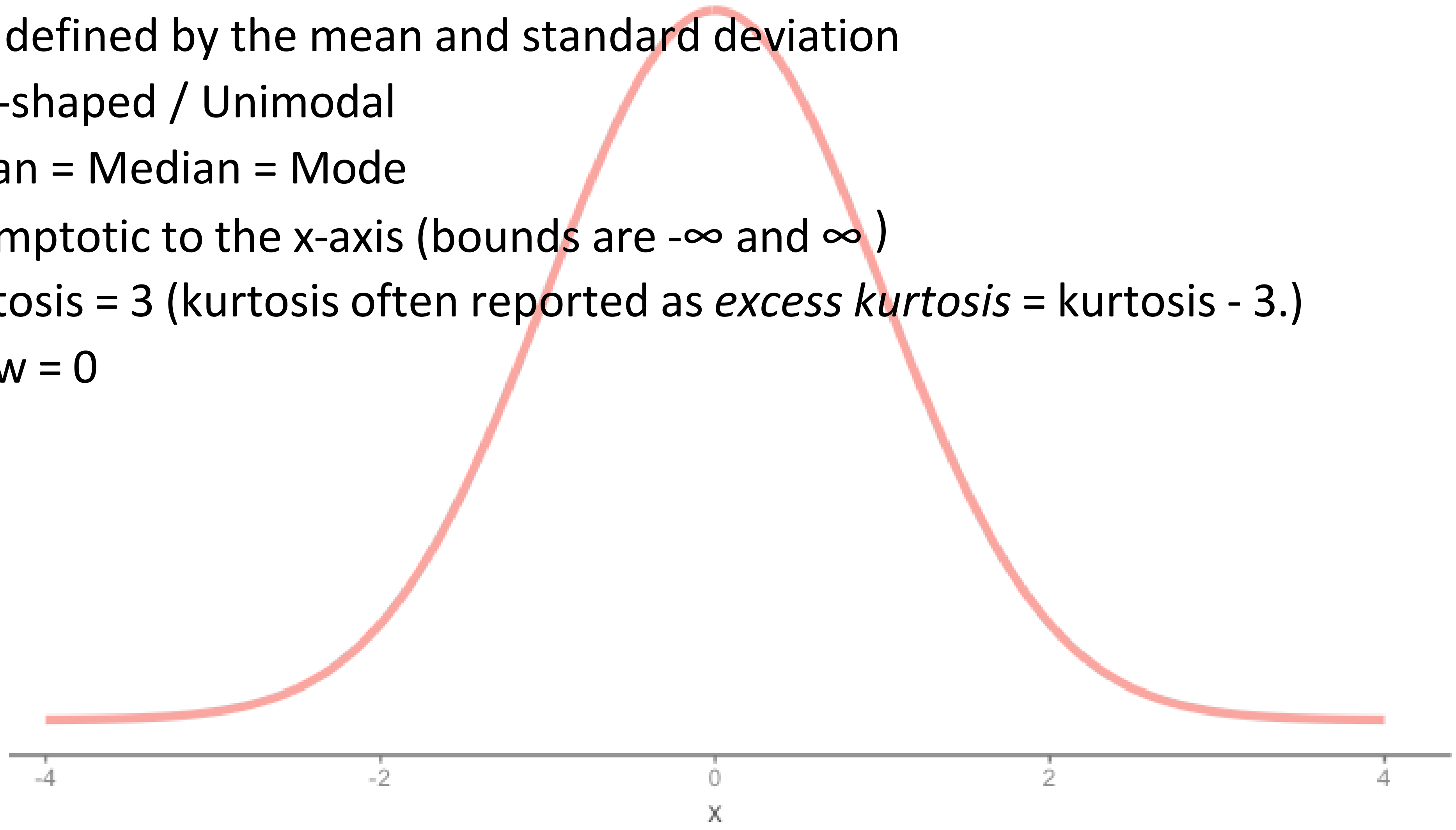
Measures of Shape

Kurtosis Does the distribution have thicker/thinner tails than a normal distribution with same mean and variance?

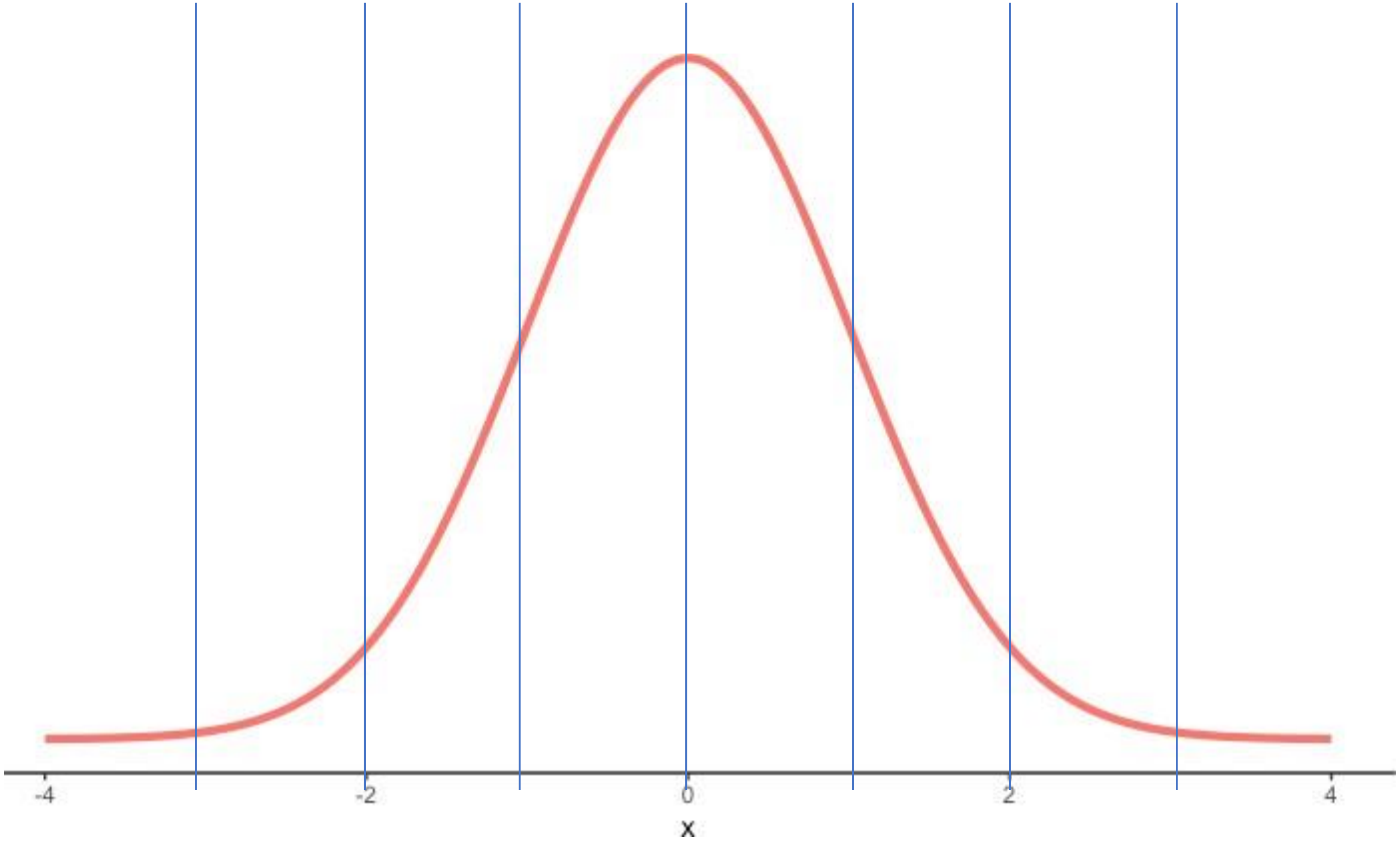


The Normal Distribution

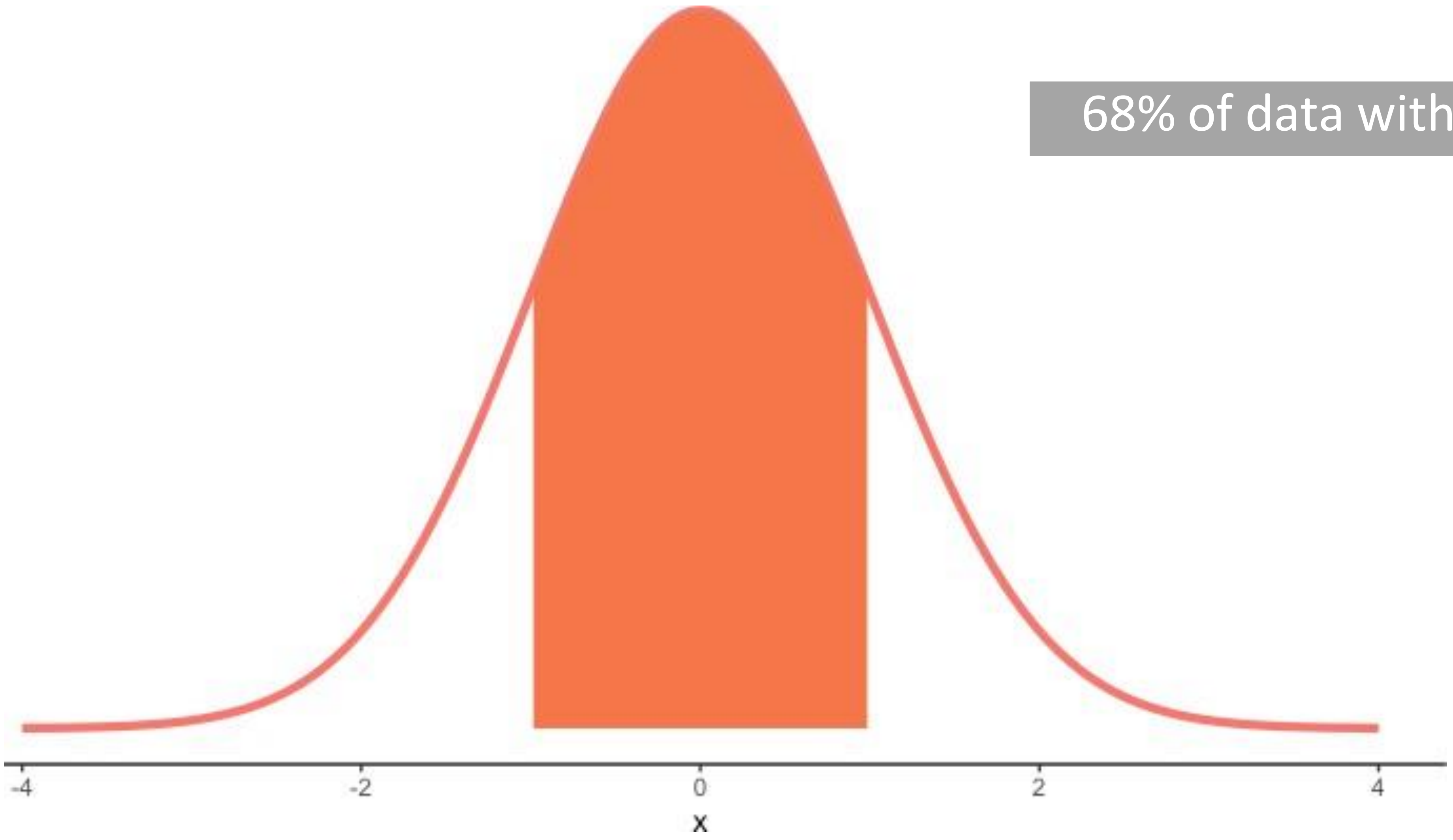
- Symmetric
- Full defined by the mean and standard deviation
- Bell-shaped / Unimodal
- Mean = Median = Mode
- Asymptotic to the x-axis (bounds are -∞ and ∞)
- Kurtosis = 3 (kurtosis often reported as *excess kurtosis* = kurtosis - 3.)
- Skew = 0



The Normal Distribution and the Empirical Rule

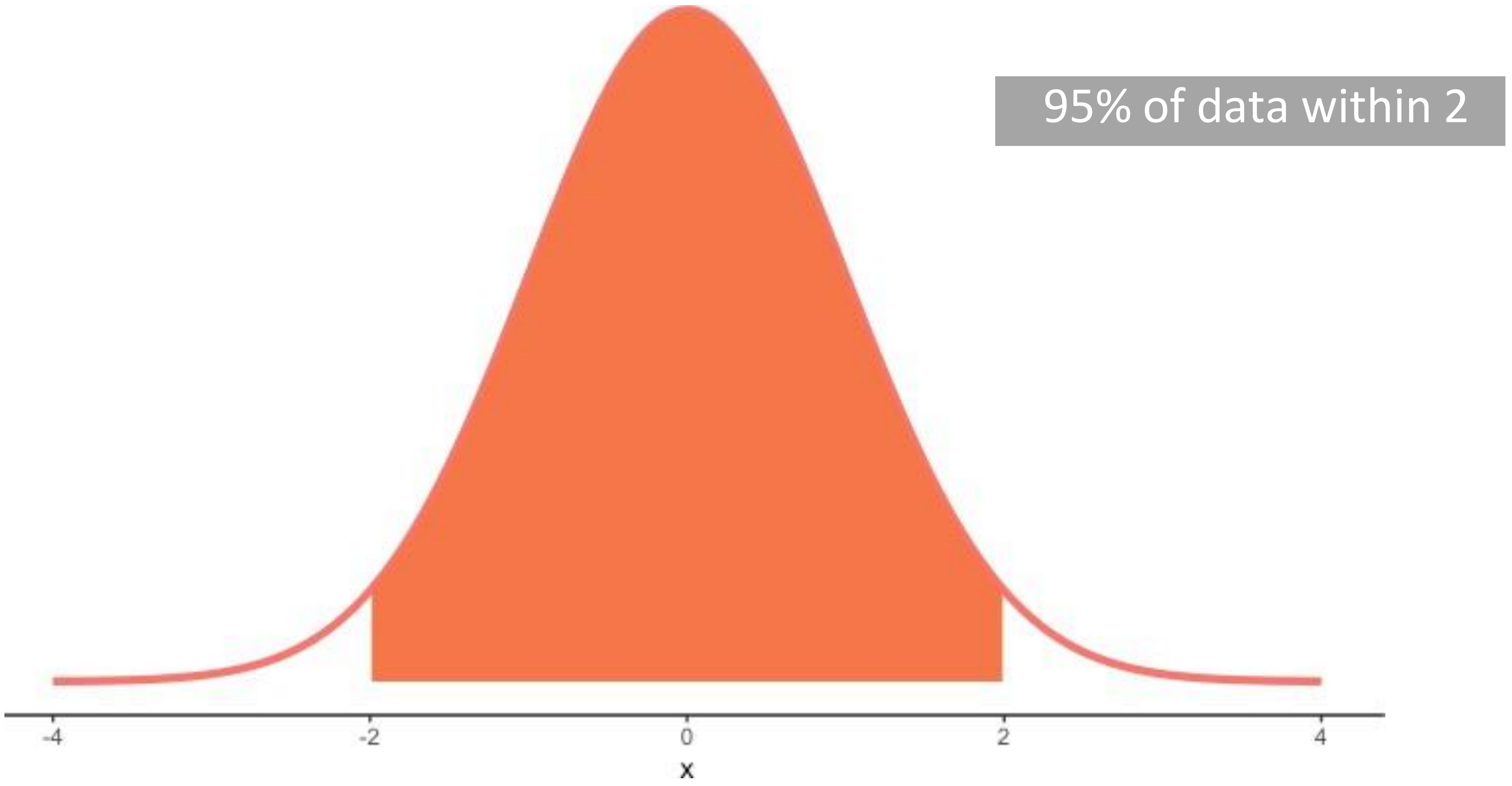


The Normal Distribution and the Empirical Rule

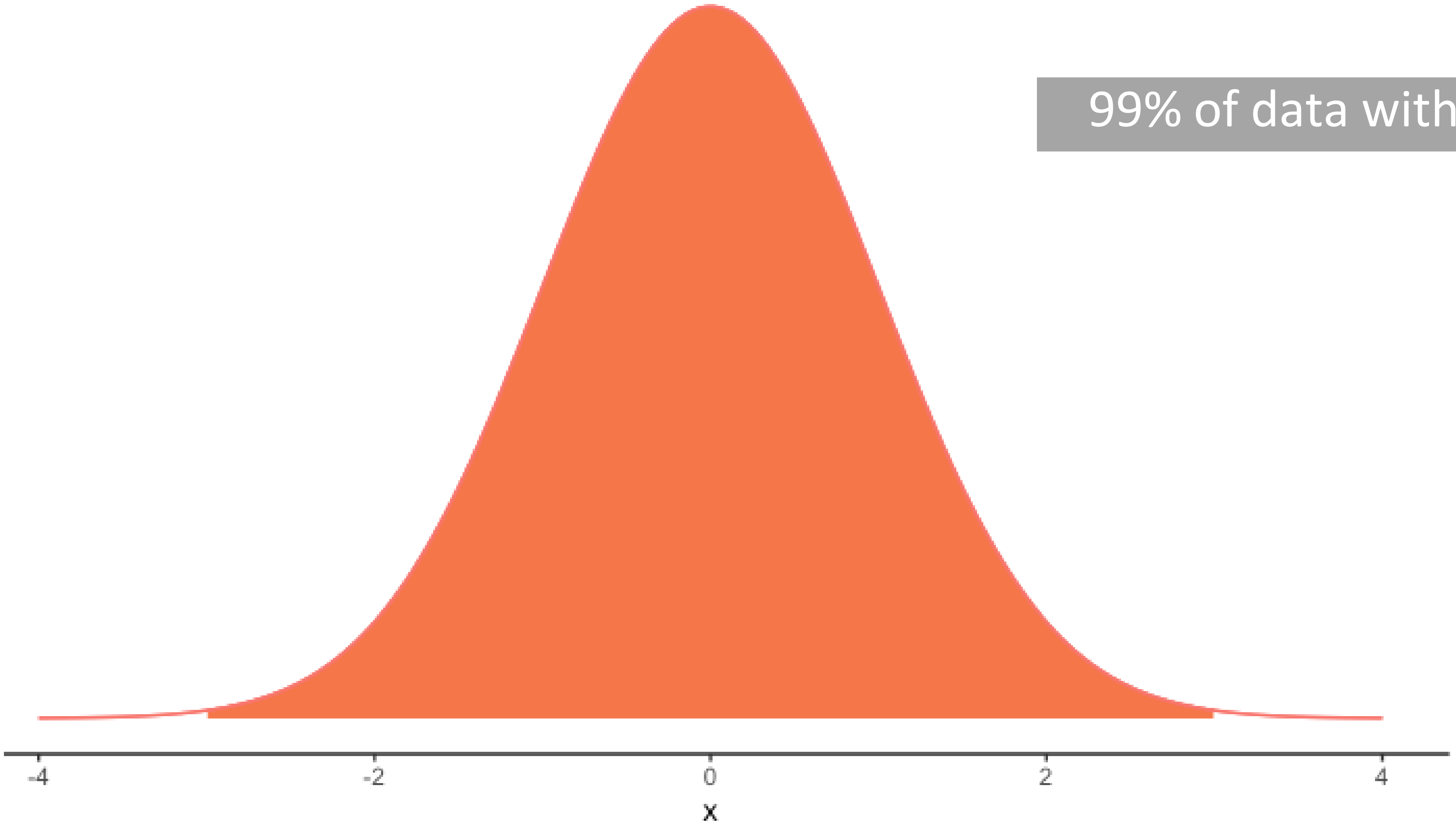


68% of data within 1

The Normal Distribution and the Empirical Rule



The Normal Distribution and the Empirical Rule



99% of data within 3 ’

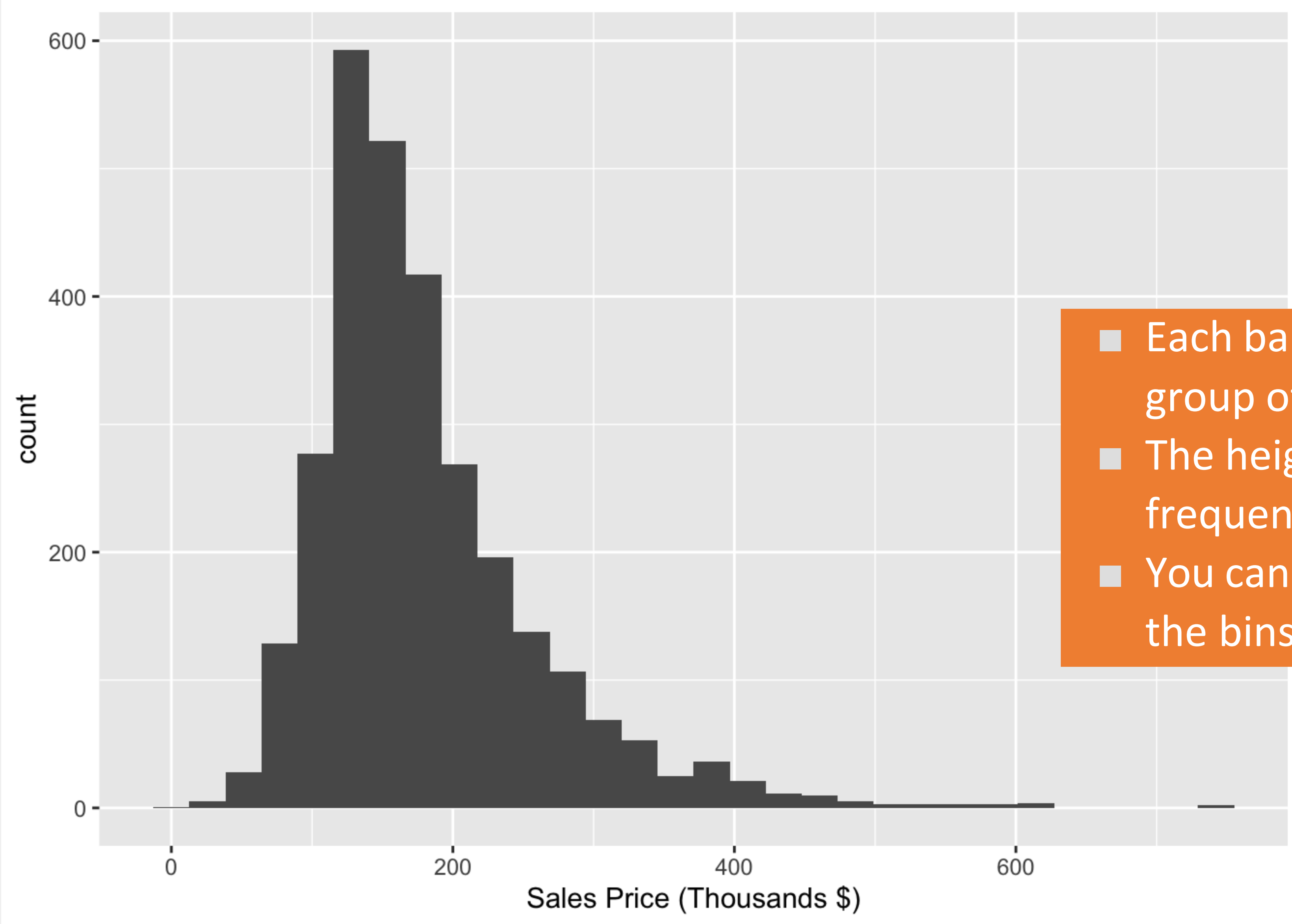
Describing Distributions Part 2: Visualization

Histograms, Density Plots, QQ-Plots and Box-Plots

Graphical Displays of Distributions

- Histograms
- Normal Probability Plots (QQ-Plots)
- Box Plots

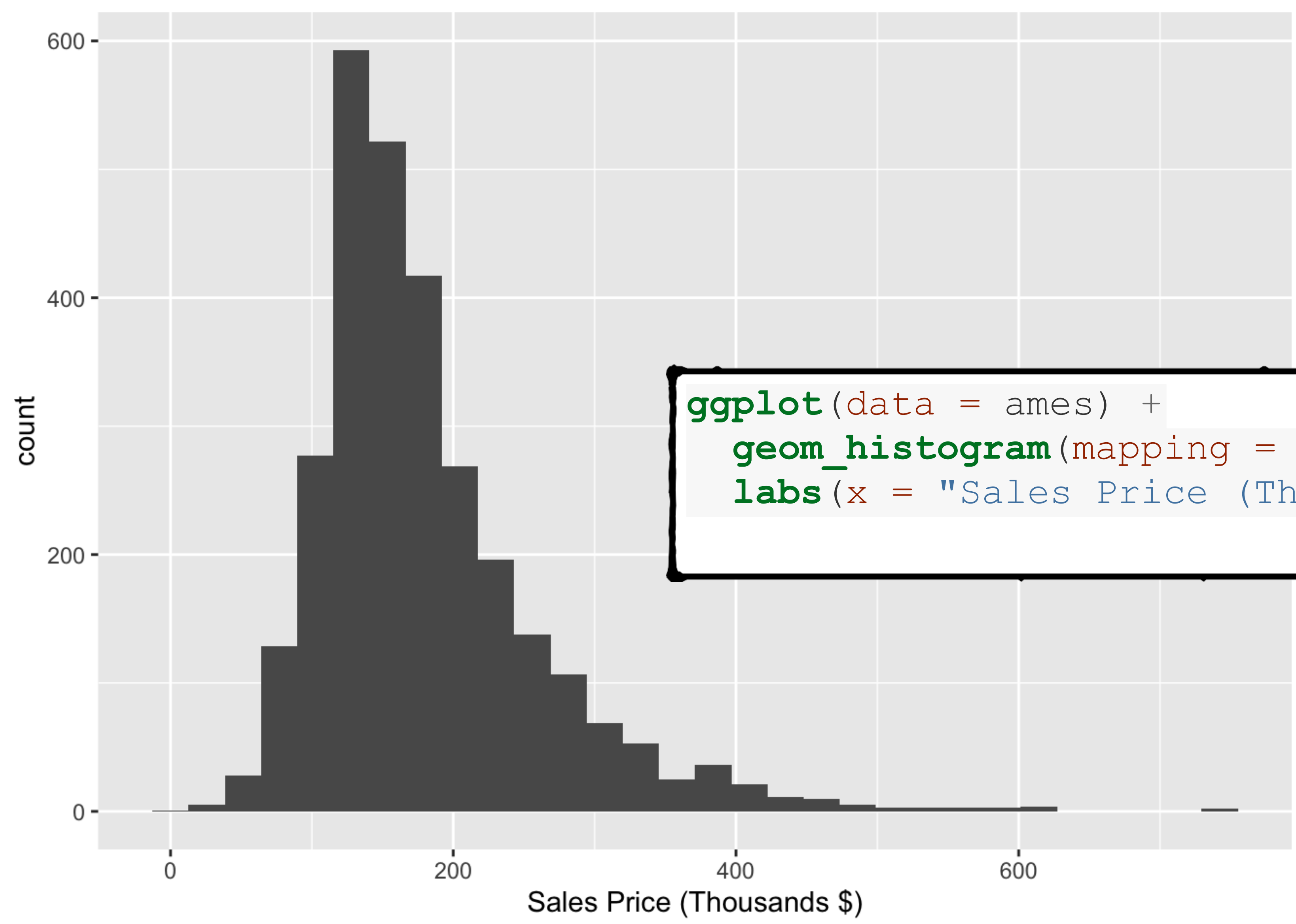
Graphical Displays of Distributions



- **Histograms**
- Normal Probability Plots (QQ-Plots)
- Box Plots

- Each bar in the histogram represents a group of values (a *bin*)
- The height of the bar represents the frequency or percent of values in the bin
- You can specify the number or width of the bins as desired

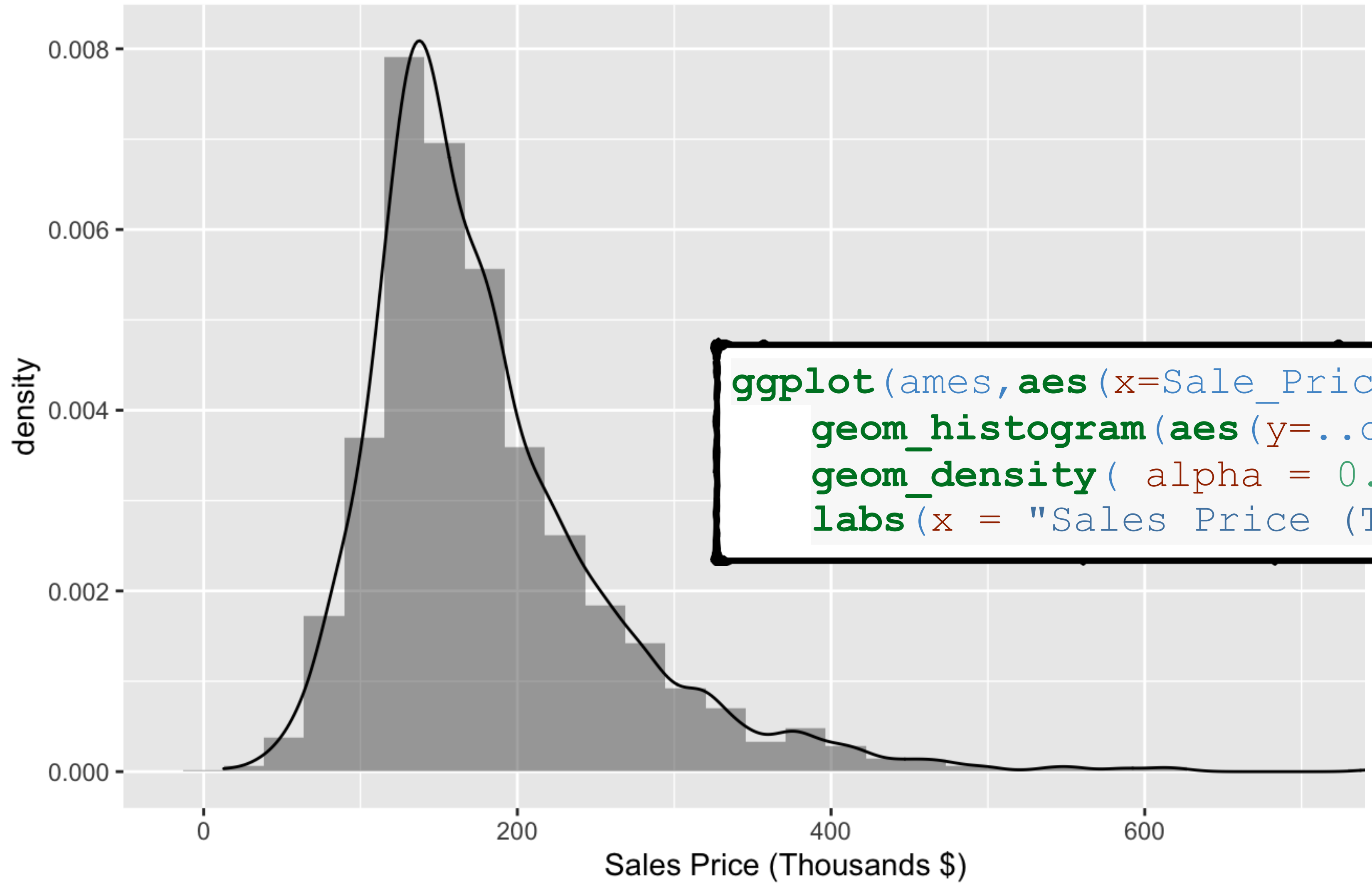
Graphical Displays of Distributions



- **Histograms**
- Normal Probability Plots (QQ-Plots)
- Box Plots

```
ggplot(data = ames) +
  geom_histogram(mapping = aes(x = Sale_Price/1000)) +
  labs(x = "Sales Price (Thousands $)")
```

Graphical Displays of Distributions

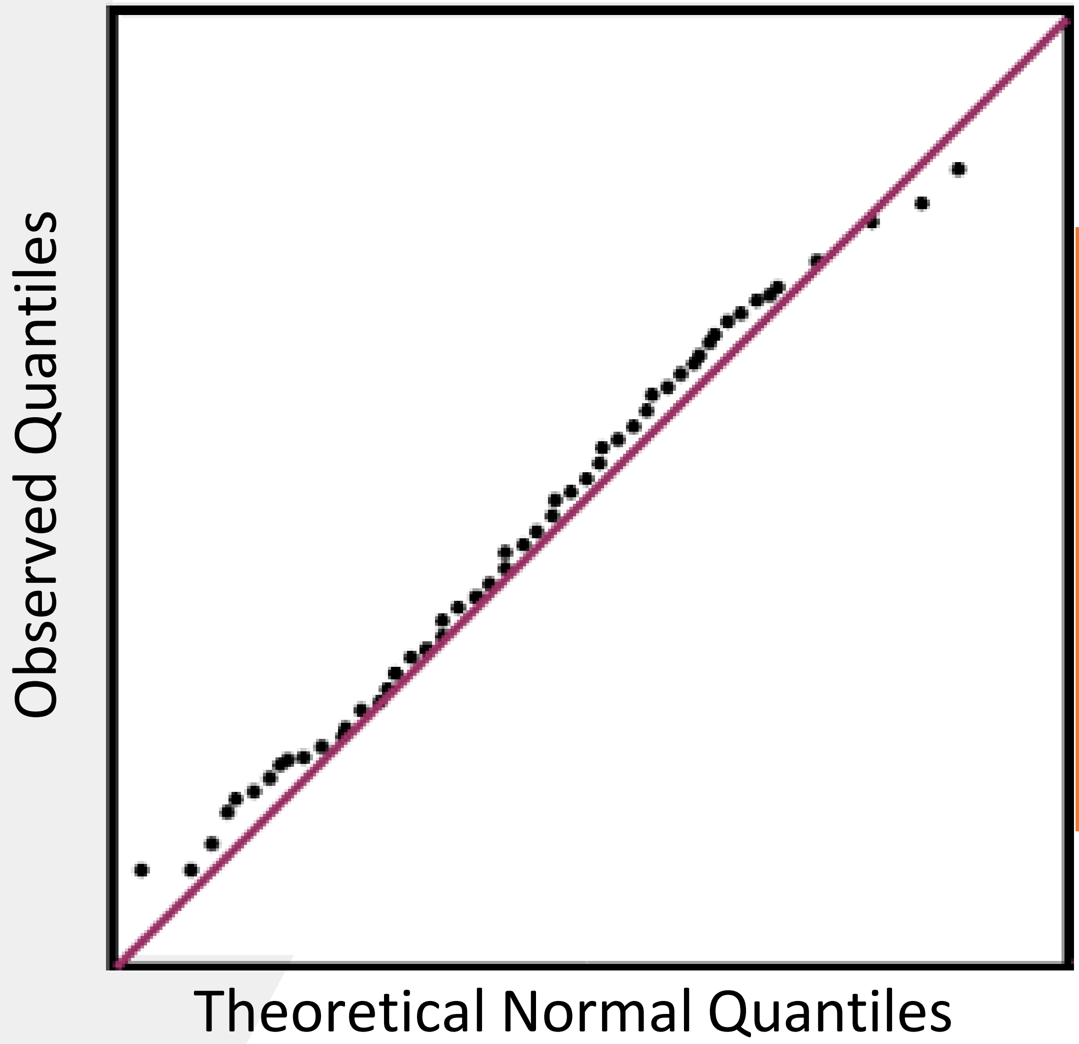



```
ggplot(ames,aes(x=Sale_Price/1000)) +
    geom_histogram(aes(y=..density..), alpha=0.5) +
    geom_density( alpha = 0.2) +
    labs(x = "Sales Price (Thousands $)")
```

- **Histograms**
- Normal Probability Plots (QQ-Plots)
- Box Plots

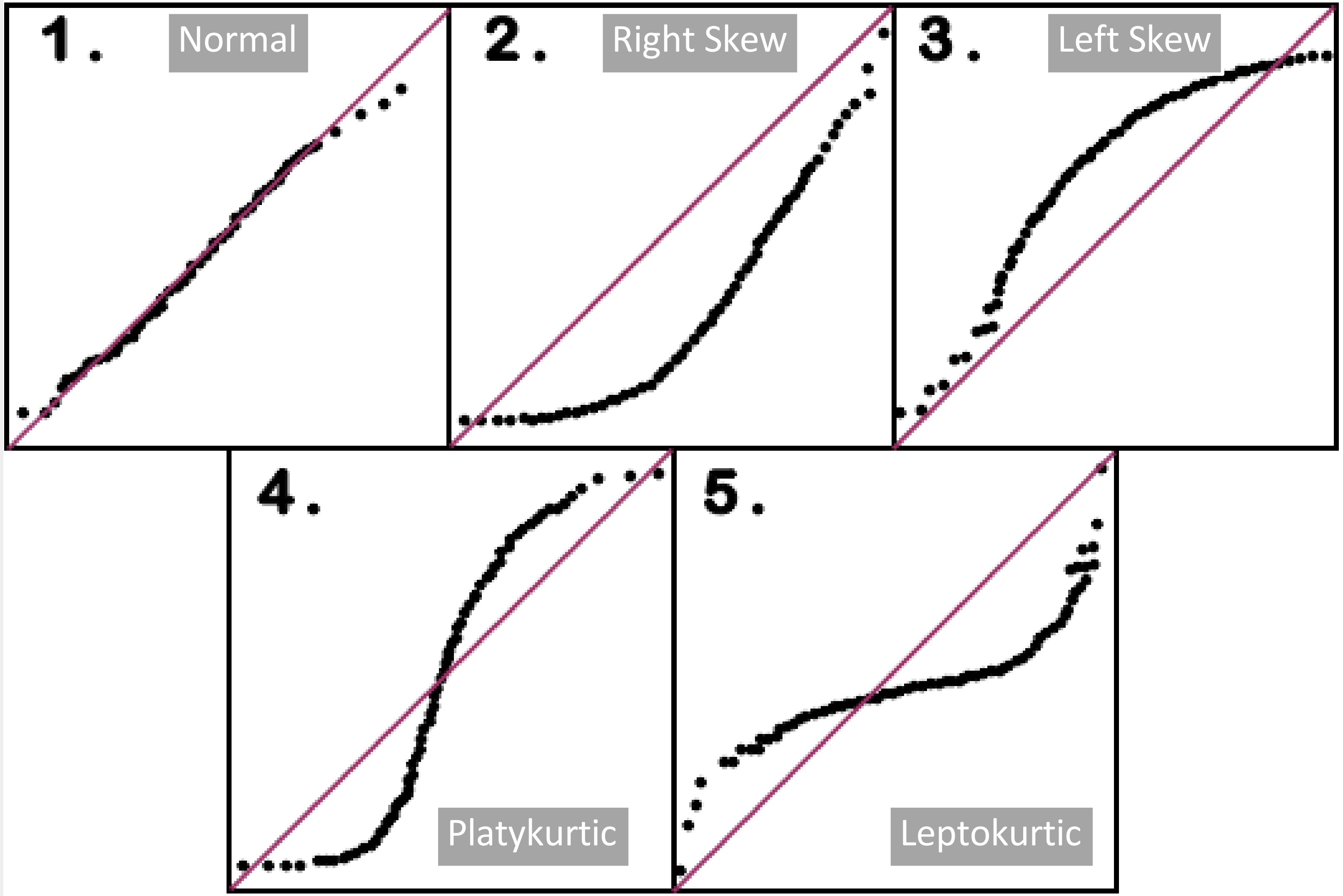
Graphical Displays of Distributions

- Histograms
- **Normal Probability Plots (QQ-Plots)**
- Box Plots



- Used to compare two distributions, typically to verify that a variable is approximately normal.
- Compare observed quantiles to theoretical quantiles of a normal distribution with the same mean and variance
- If the points follow the line diagonal line, the distribution is normal.

Graphical Displays of Distributions

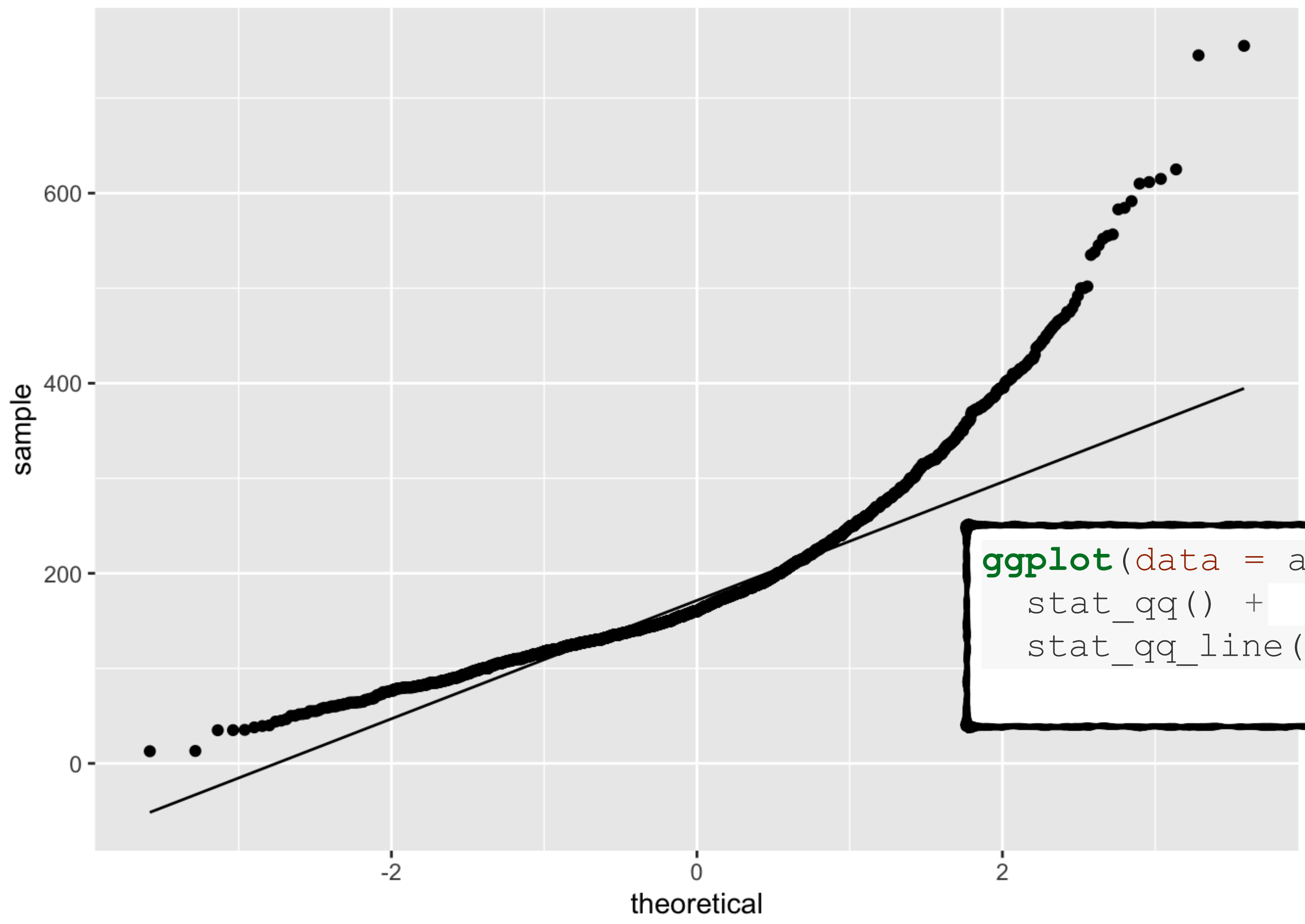


- Histograms
- **Normal Probability Plots (QQ-Plots)**
- Box Plots

- Quadratic patterns indicate problems with skew
- Cubic patterns indicate problems with kurtosis

Graphical Displays of Distributions

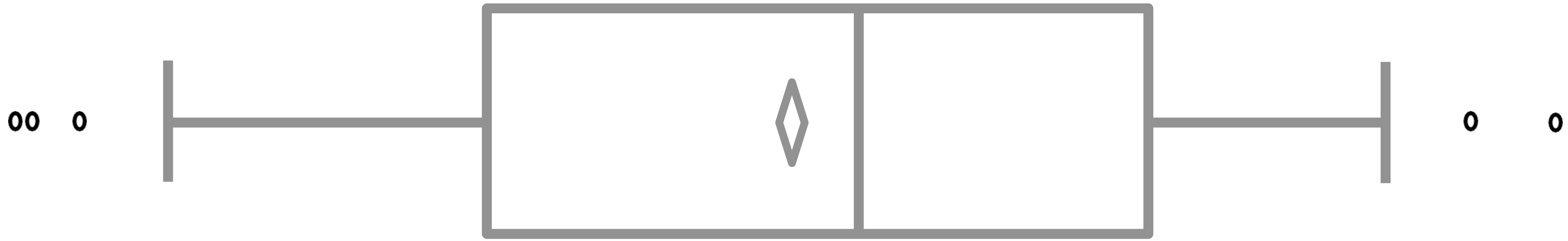
- Histograms
- **Normal Probability Plots (QQ-Plots)**
- Box Plots




```
ggplot(data = ames, aes(sample = Sale_Price/1000)) +
  stat_qq() +
  stat_qq_line()
```

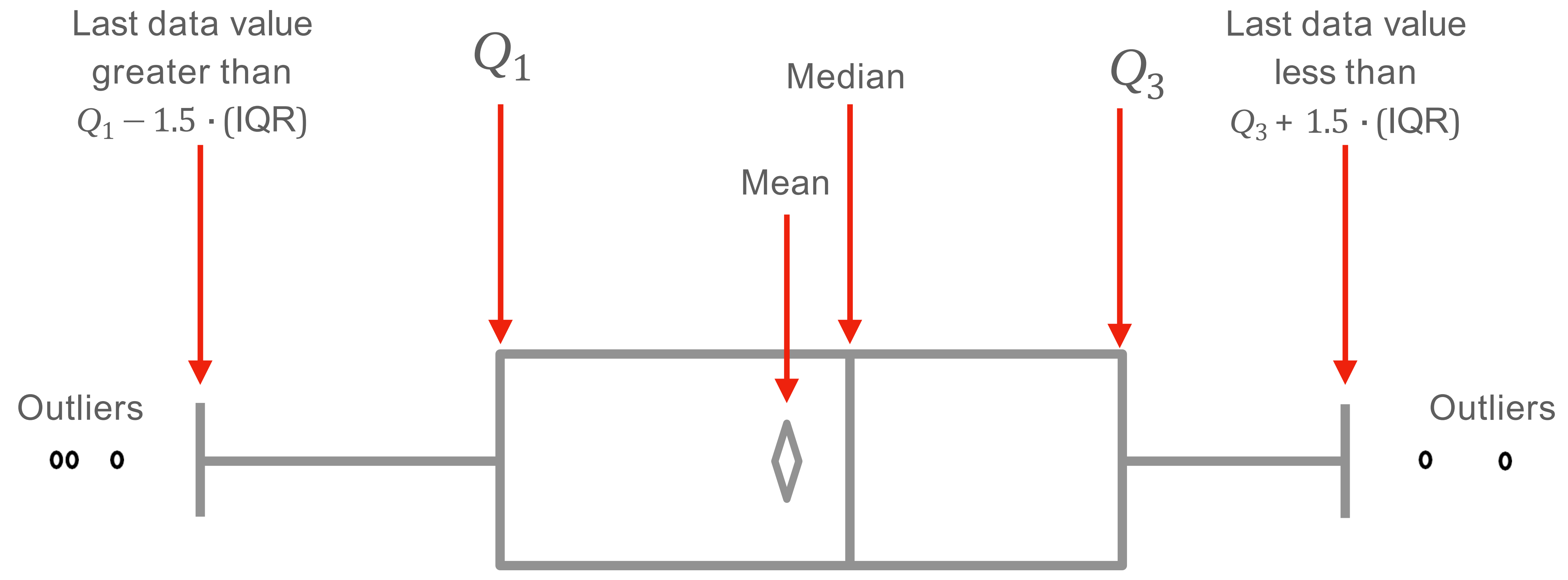
Graphical Displays of Distributions

- Histograms
- Normal Probability Plots (QQ-Plots)
- **Box Plots**

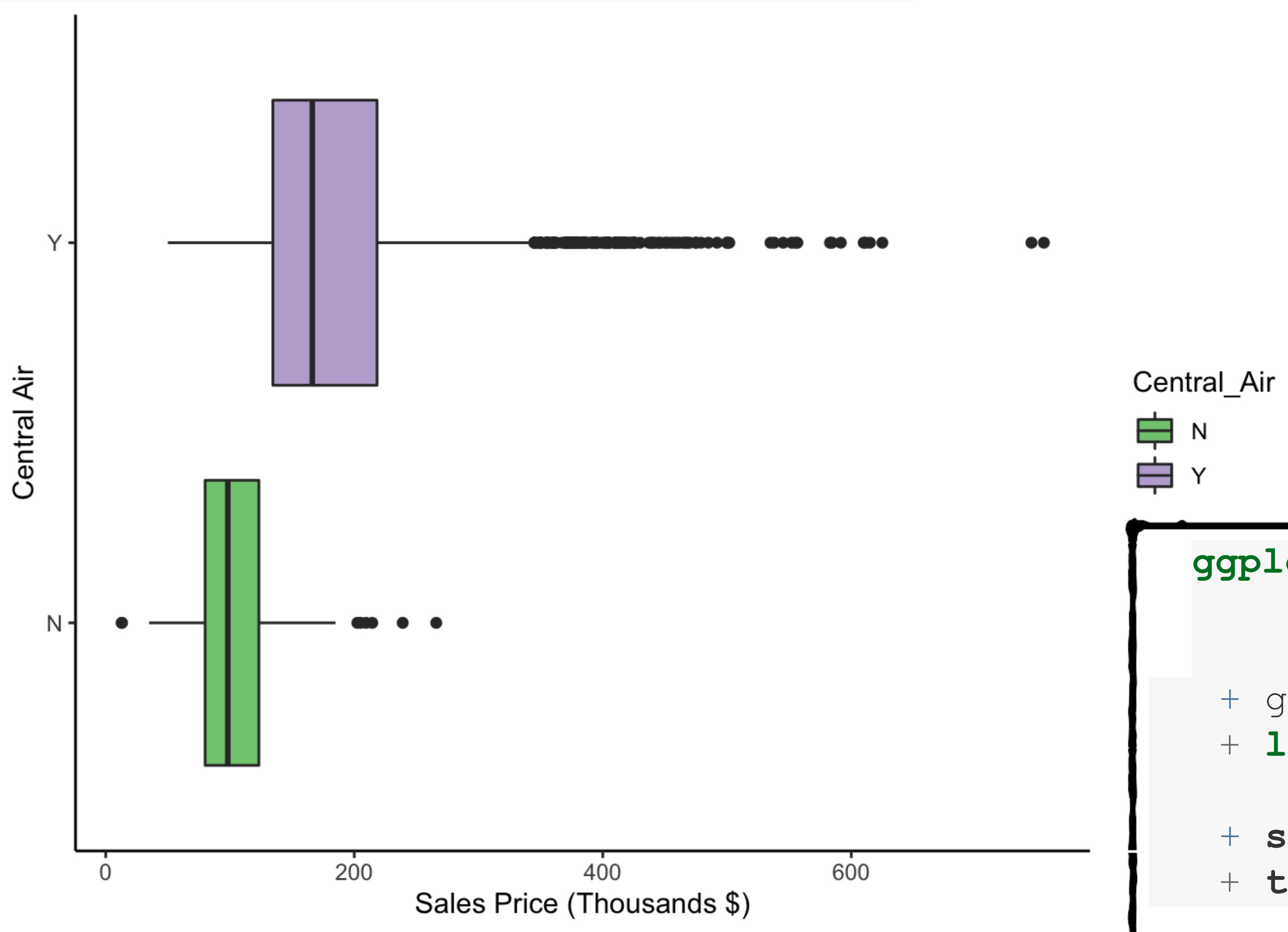


Graphical Displays of Distributions

- Histograms
- Normal Probability Plots (QQ-Plots)
- **Box Plots**



Graphical Displays of Distributions



- Histograms
- Normal Probability Plots (QQ-Plots)
- **Box Plots**

```
ggplot(data = ames, aes(y = Sale_Price/1000,
                        x = Central_Air,
                        fill = Central_Air))
  + geom_boxplot()
  + labs(y = "Sales Price (Thousands $)",
         x = "Central Air")
  + scale_fill_brewer(palette="Accent")
  + theme_classic() + coord_flip()
```

Describing Distributions

- Center/Location
- Spread/Variation
- Shape
- **Anomalous Observations**

Defining Anomalous Observations

Standard Deviations From the mean

For symmetric distributions and particularly for the normal distribution - common to consider observations more than 3 standard deviations from the mean as anomalous

Box-Plot Definition

Box plots define outliers as any points that lie more than 1.5*IQR above the third quartile or less than 1.5*IQR below the first quartile.

More Definitions to Come!

There are many methods to investigate and label anomalous observations in a dataset. Stay tuned for more!

Lab 1

Don’t forget to take the lab check on Moodle!

Introduction to Statistical Inference

MSA 2023

Point Estimates

- We want to learn about an entire population
- We take a representative sample and calculate sample statistics
- Sample statistics will have some error, they are *estimates* of their population parameter counterparts.

estimates the population mean,

estimates the population std. dev.,

Variability among samples

Average Price = $130,011

Average Price = $127,987

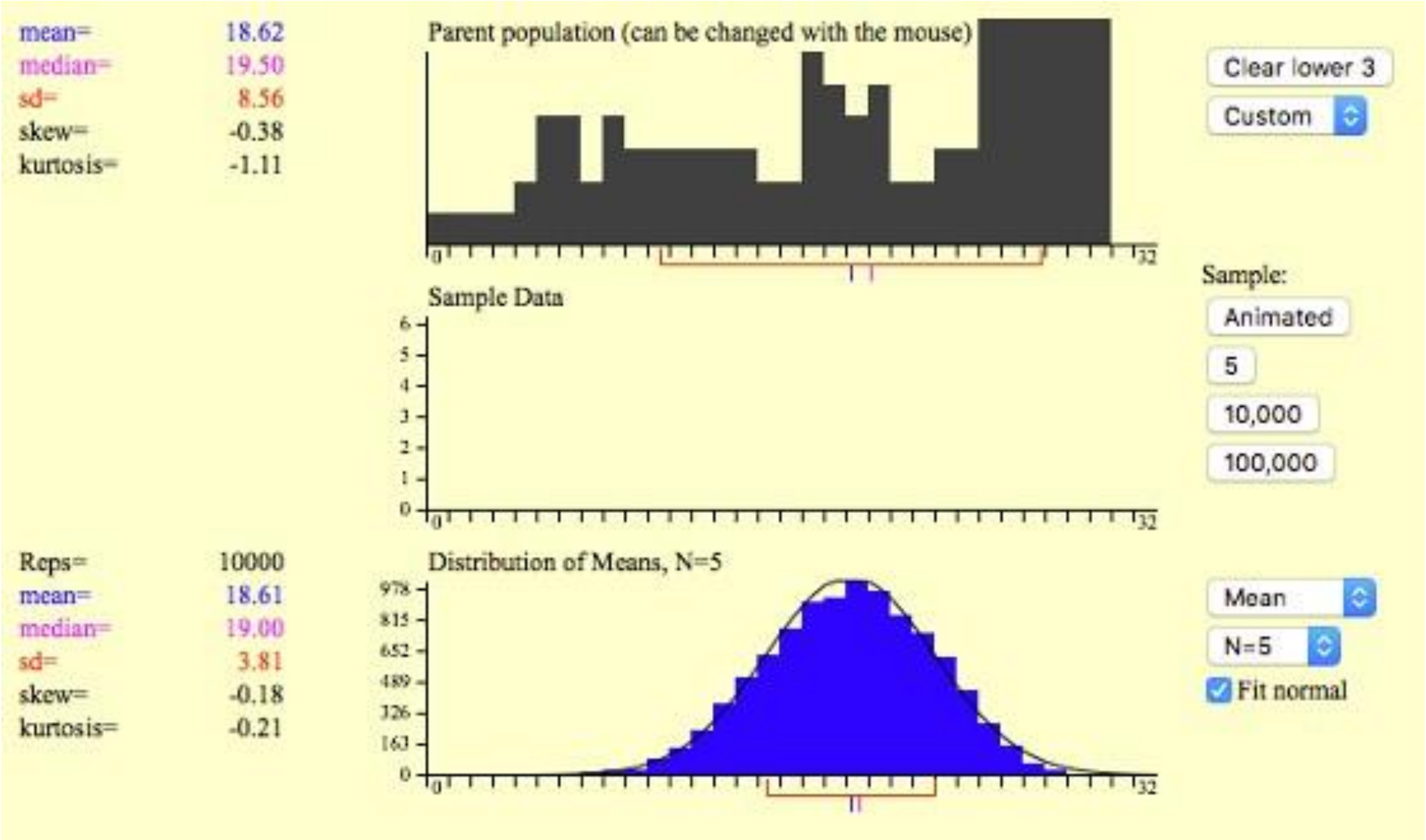
Average Price = $131,125

- What if we take a different sample?
  - We'd have a different sample mean!
- Can we have a margin of error for our estimate?
  - Yes, via Central Limit Theorem.

Central Limit Theorem: The distribution of sample means is approximately normal, regardless of the population distribution's shape, if the sample size is large enough.

Interactive Demo: Central Limit Theorem

http://onlinestatbook.com/stat_sim/sampling_dist/



Average Price = $130,011

Average Price = $127,987

Average Price = $131,125

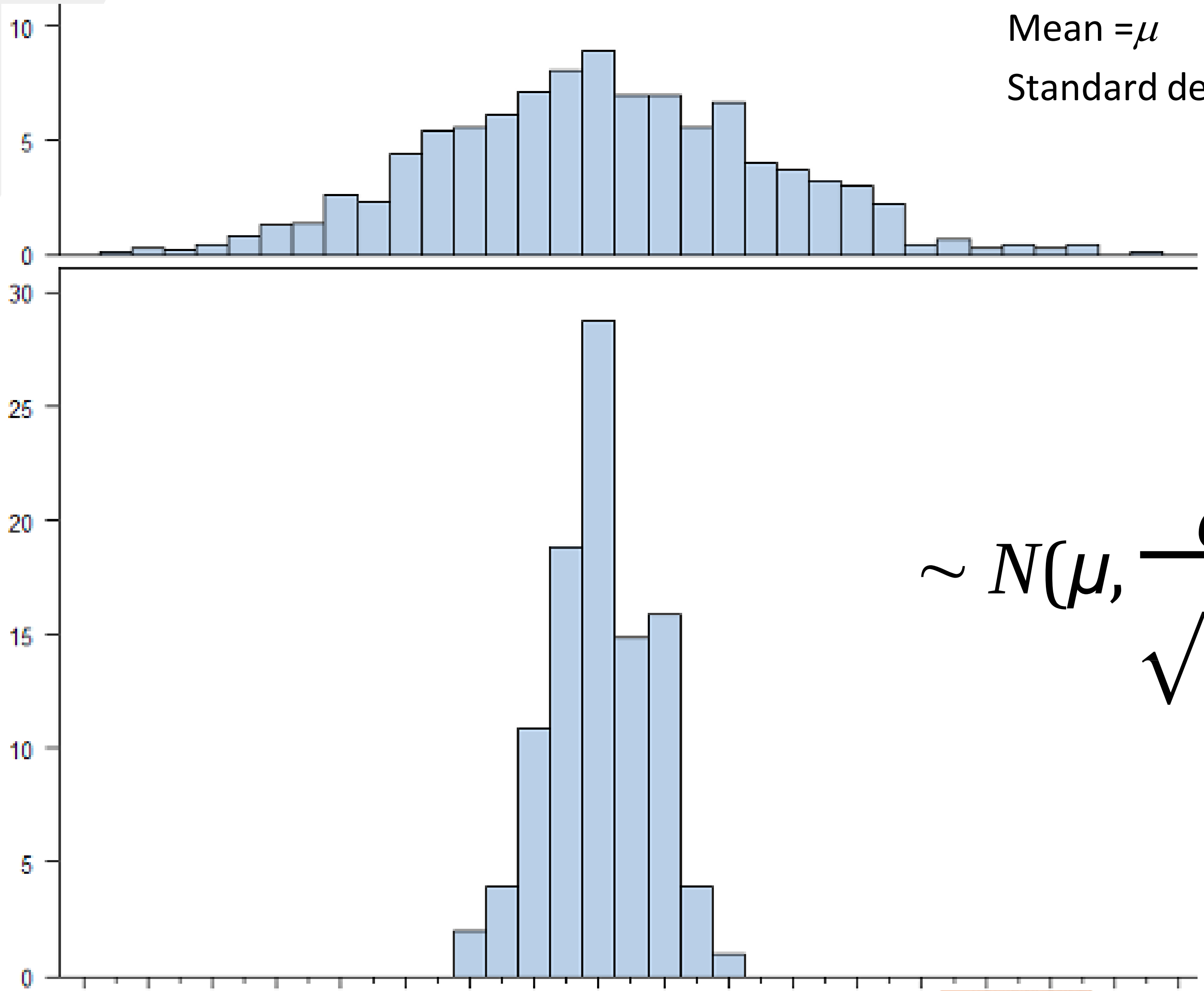
Standard Error of the Mean

- **Standard error** measures the variability of your estimate
  - If you were to re-sample the data and compute the new sample average many times, how much variability might you expect in your results?
- **Different from standard deviation**
  - Sample **standard deviation** ( ) is a measure of the **variability in your data**
  - **Standard error** of the mean ($s_{\bar{x}}$) is a measure of the estimated **variability of the sample means**.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Central Limit Theorem: Example

Distribution of Population



Mean $=\mu$

Standard dev. $=\sigma$

Distribution of sample means (n=10)

$$\sim N(\mu, \frac{\sigma}{\sqrt{10}})$$

Standard Error of the Mean

- Variability about a statistics, for sample mean $\bar{x}$, we will denote this as $s_{\bar{x}}$
- Can be used to construct MOE by going the correct amount of standard error from the estimate
- For example, a 95% confidence interval for the meanis calculated by $\bar{x} - 2s_{\bar{x}}, \bar{x} + 2s_{\bar{x}}$

Average Price = $130,011

Average Price = $127,987

Average Price = $131,125

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Average Price = $130,011

Average Price = $127,987

Average Price = $131,125

Confidence Intervals

- Percent of confidence about true parameter
- A 95% confidence interval represents a range of values within which you are 95% “confident” that the true population mean exists.

$$(\bar{x} - t s_{\bar{x}}, \bar{x} + t s_{\bar{x}})$$

is the *t*-value corresponding to the confidence level, and *n*-1 degrees of freedom, where *n* is the sample size.

t-value varies with sample size and level of confidence

Confidence Intervals

Average Price = $130,011

Average Price = $127,987

Average Price = $131,125

- The t-value indicates the number of standard errors from the mean for MOE
- Is found by looking at the quantile from a t-distribution (in r, this is done by qt(amount in one tail, df, lower.tail=T)

$$(\bar{x} - t s_{\bar{x}}, \bar{x} + t s_{\bar{x}})$$

is the *t*-value corresponding to the confidence level, and *n*-1 degrees of freedom, where *n* is the sample size.

t-value varies with sample size and level of confidence

Hypothesis Test

- **Confidence intervals provided us with information about the variability around the statistic**
- **Hypothesis test is designed to investigate if we can prove that the true population value is significantly different than an assumed value**
- **Two hypotheses: Null hypothesis and Alternative hypothesis**

Judicial Analogy* for Hypothesis Test

In a court of law:

- **Innocence is assumed.**
- **Evidence is collected**
- **If <u>sufficient</u> evidence found (beyond a "reasonable doubt") we reject the assumption of innocence.**
  => Find guilt.
- **If <u>sufficient</u> evidence NOT found, fail to reject assumption of innocence.**

*Assumes a judicial system free from oppression and systemic bias

Hypothesis Test Procedure

- Start with an initial ("null") hypothesis (H_0) about a parameter of interest, assume it to be true.
- Fix an **acceptable significance level** , representing the likelihood that you incorrectly reject his null hypothesis (α).
- The alternative hypothesis (H_a) is the logical opposite
- Collect data, compute statistic of interest
- Determine the probability that you would have observed a statistic as extreme or more extreme as the one you did **if H_0 is true**
  - This is called your p-value.
- If your p-value is $\leq \alpha$, you **reject** null hypothesis
- If your p-value is > α , you **fail to reject** that null hypothesis

Hypothesis Test Example: Coin Flips

- Your friend comes back from vacation, wants to play a new betting game using a coin he got overseas.
- After losing 3 rounds, you hypothesize there is something special about this coin.
- A *null* hypothesis has to be something that we can concretely describe. ("The coin is fair")
- The *alternative* hypothesis is usually what you're trying to demonstrate. ("The coin is unfair")
- You'd be satisfied with a conclusion at 5% significance level

  (i.e. $\alpha = 0.05$ => there is a 5% chance you incorrectly accuse your friend of cheating)

Hypothesis Test Example: Coin Flips

- Start with an initial ("null") hypothesis (H_0) about a parameter of interest, assume it to be true. **H_0: P(Heads) =0.5**
- Fix an **acceptable significance level ,** representing the likelihood that you incorrectly reject his null hypothesis. **= 0.05**
- The alternative hypothesis (H_a) is the logical opposite **H_a: P(Heads) ≠ 0.5**
- Collect data, compute statistic of interest **Flip coin many times, Compute proportion of heads.**
- Determine the probability that you would have observed a statistic as extreme or more extreme as the one you did **if H_0 is true** **How many Heads should we expect for a fair coin??**
  - This is called your p-value.
- If your p-value is $\leq \alpha$, you **reject** null hypothesis
- If your p-value is > α , you **fail to reject** that null hypothesis

A Simulation Study

- We can programmatically simulate the flipping of a fair coin.
- Choose a value of “Heads” or “Tails” randomly with equal likelihood.
  One fair coin flip:

```
sample(c('Heads','Tails'), 1)
```

- Now, flip the coin n times:

```
n <- 100
outcomes <- sample(c('Heads','Tails'), n, replace=T)
```

- Calculate how many ‘Heads’ in those 100 coin tosses:

```
sum(outcomes=="Heads")
```

Hypothesis Testing

- Given what we've observed, what conclusions might we be able to draw about the population at large?
- The foundation of hypothesis testing is **an initial assumption that we try to refute** with evidence.
- It is ***not*** **a proof** in the mathematical sense. **All hypothesis tests have a level of significance (can make a type I or type II error).**

A Simulation Study

- Say we got 58 Heads in this experiment. Will we always get 58 Heads? Of course not. You probably got something different.
- Let's repeat this experiment many times, and see what the *distribution* of the number of heads in 100 coin tosses is expected to look like.
- The vector `number_heads` will record how many Heads were observed from 100 tosses across 10,000 experiments.

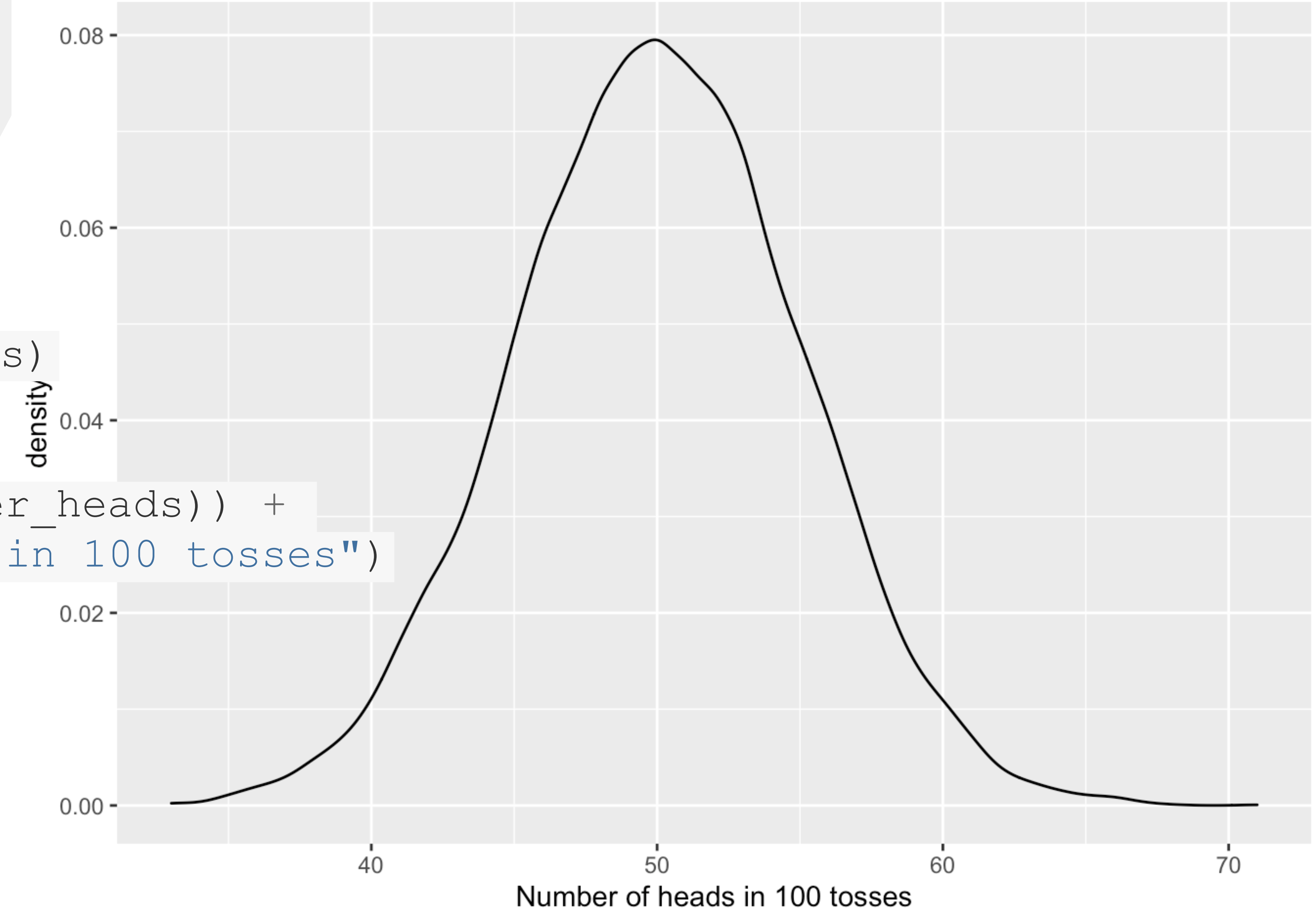
```
T <- 10000
n <- 100
set.seed(11)
number_heads <- vector()
for(i in 1:T){
  outcomes <- sample(c('Heads','Tails'), n, replace=T)
  number_heads[i] <- sum(outcomes=="Heads")
}
```

A Simulation Study

The distribution of the number of heads observed

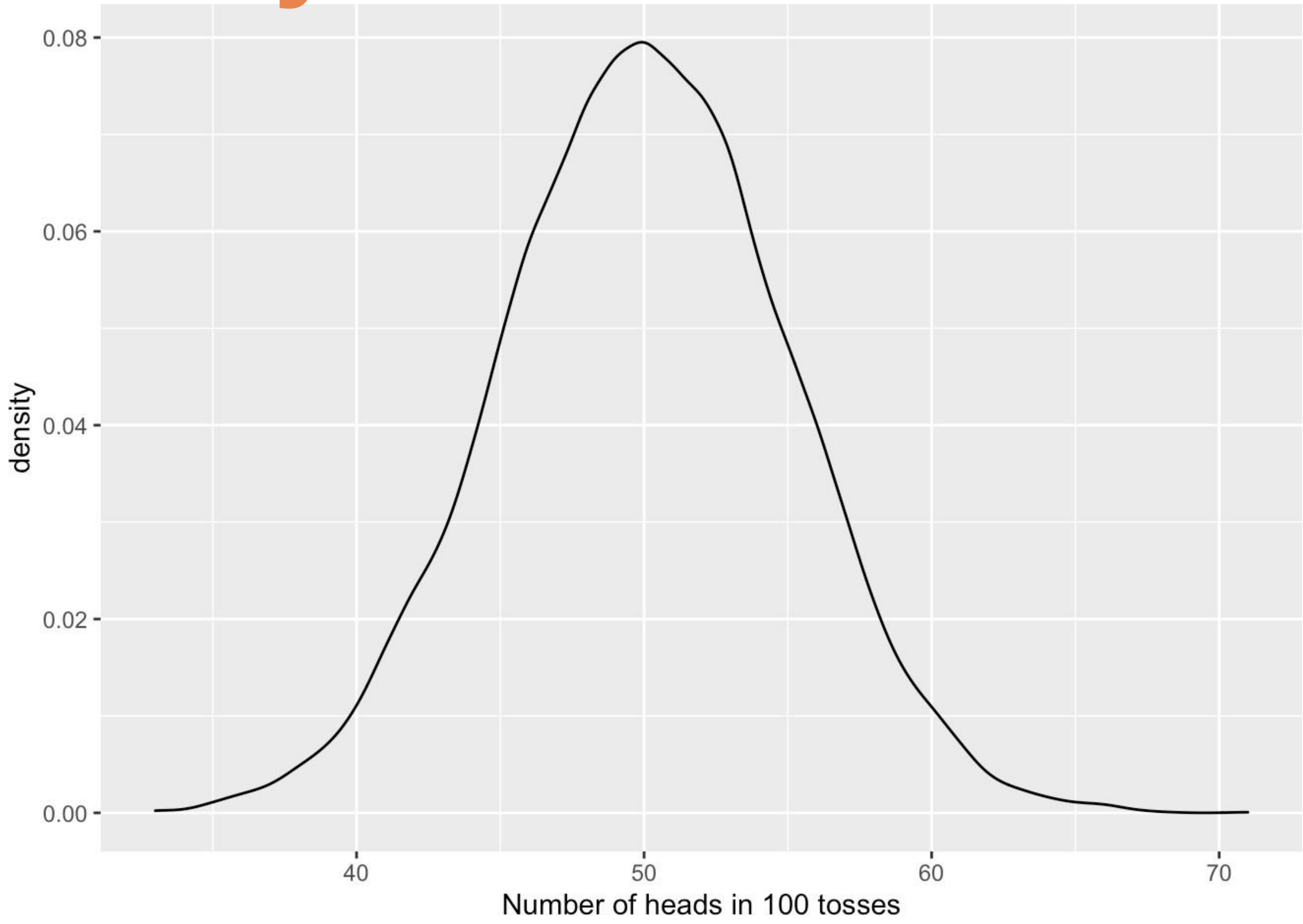
```r
df <- data.frame(number_heads)

ggplot(data = df) +
  geom_density(aes(x = number_heads)) +
  labs(x = "Number of heads in 100 tosses")
```



Simulation Study: Conclusions

- It would be *unusual* to get 71 heads with a fair coin.
- But not impossible! We got that *once* - and only once anything that extreme.
- Simulated p-value for 71 here would be 1/10000 = 0.0001




```
summary(df$number_heads)
```

One-Sample t-tests

Testing a mean against a hypothesized value

t-tests for the mean

- To test the null hypothesis $H_0 : \mu = \mu_0$, we calculate the Student's t statistic value:

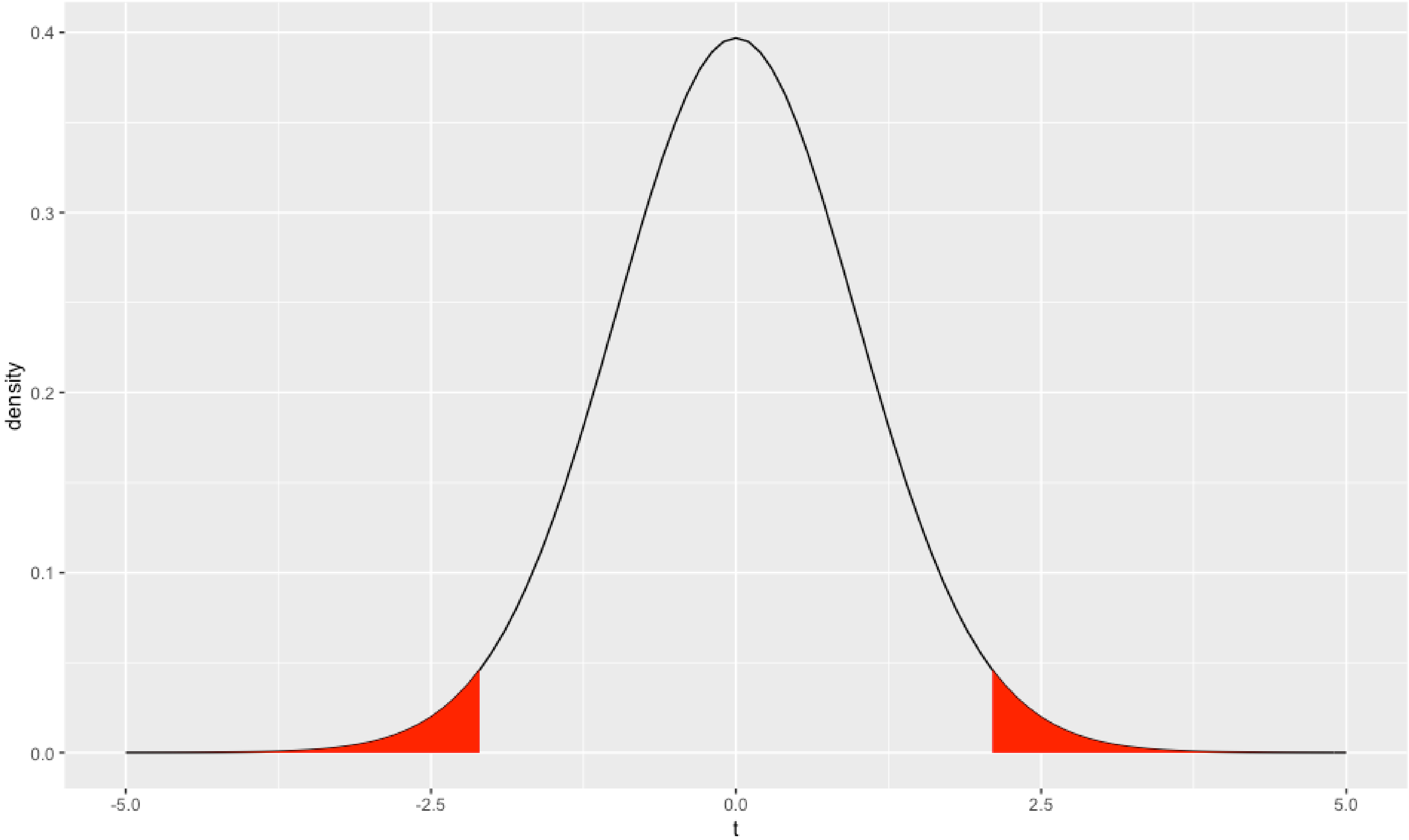
$$t = \frac{(\bar{x} - \mu_0)}{s_{\bar{x}}}$$

"number of standard deviations away from the hypothesized value"

- The null hypothesis is rejected when the t-value is more extreme than one would expect to happen by chance when H_0 is true

Two-sided t-tests for the mean

Rejection region for two-sided hypothesis test: t can be either positive or negative



Two-sided t-tests for the mean in R

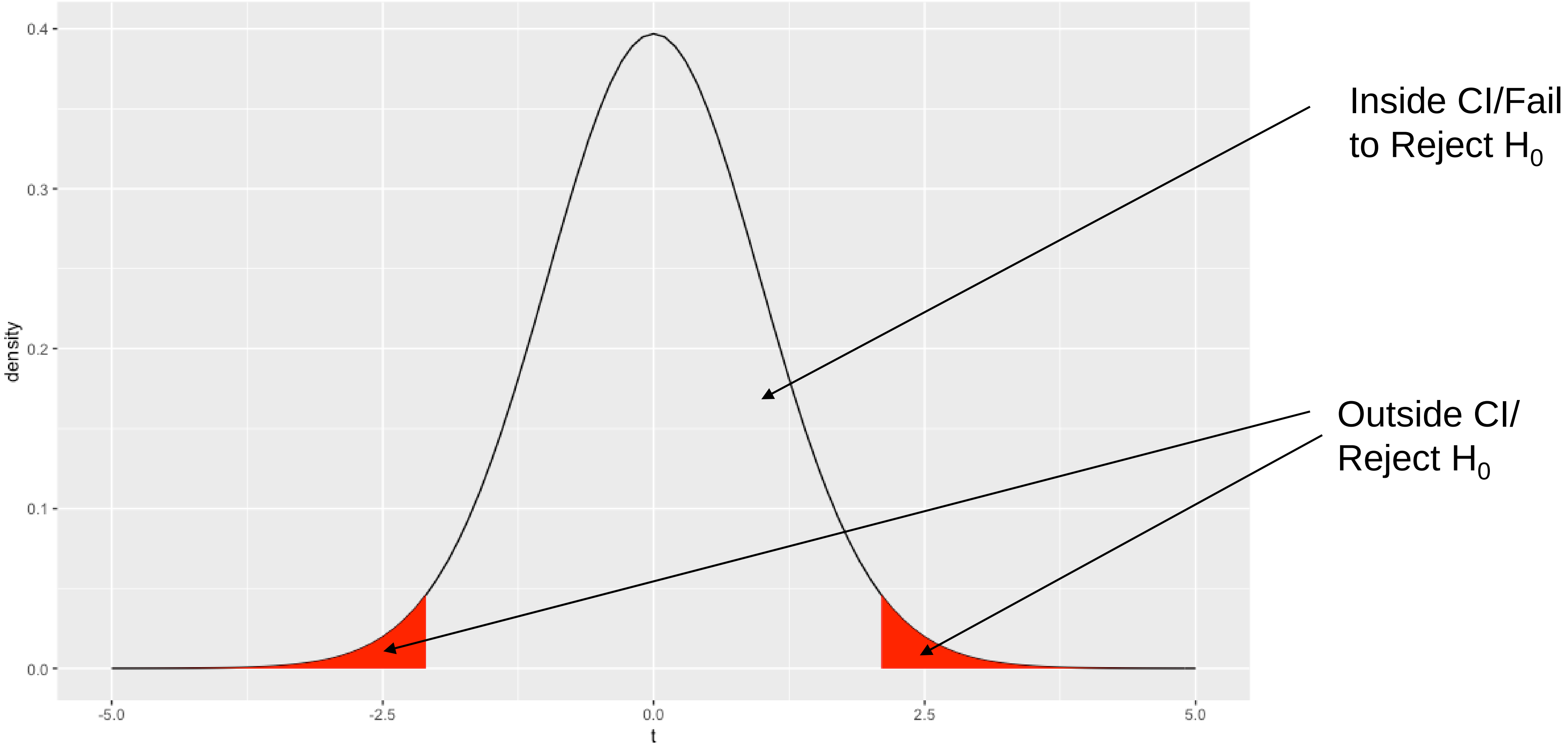
Test the null hypothesis that the mean sale price of homes is $178,000:

```r
t.test(ames$Sale_Price, mu = 178000)
```

```
##
##  One Sample t-test
##
## data:  ames$Sale_Price
## t = 1.8945, df = 2929, p-value = 0.05825
## alternative hypothesis: true mean is not equal to 178000
## 95 percent confidence interval:
##  177902.3 183689.9
## sample estimates:
## mean of x
##  180796.1
```

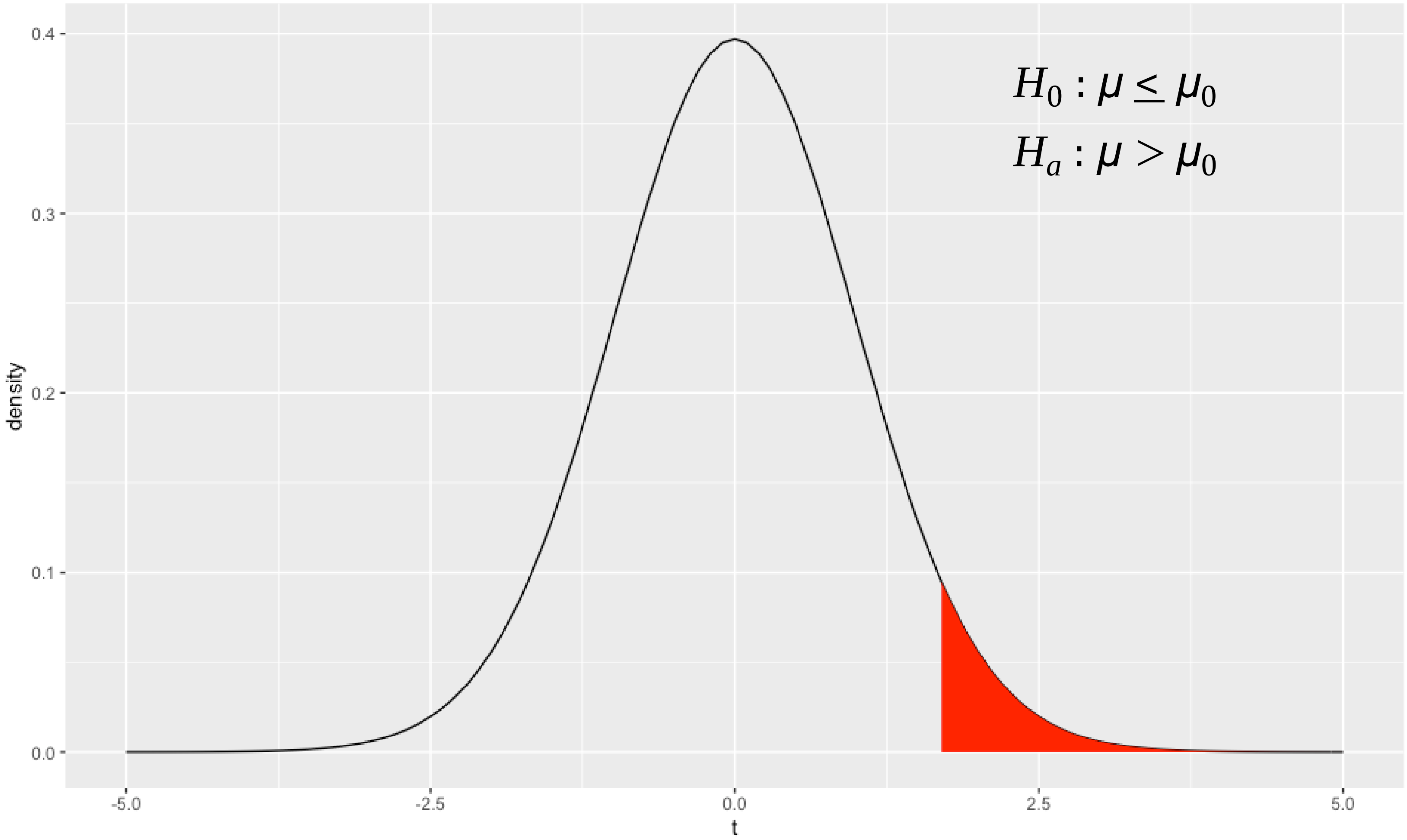
Two-sided t-tests and CI for mean

Relationship between Two-sided hypothesis test and CI



One-sided t-tests for the mean

Rejection region for one-sided hypothesis test



One-sided t-tests for the mean in R

Test the null hypothesis that the mean sale price of homes is < $178,000:

```
t.test(ames$Sale_Price, mu = 178000, alternative = 'greater')
```

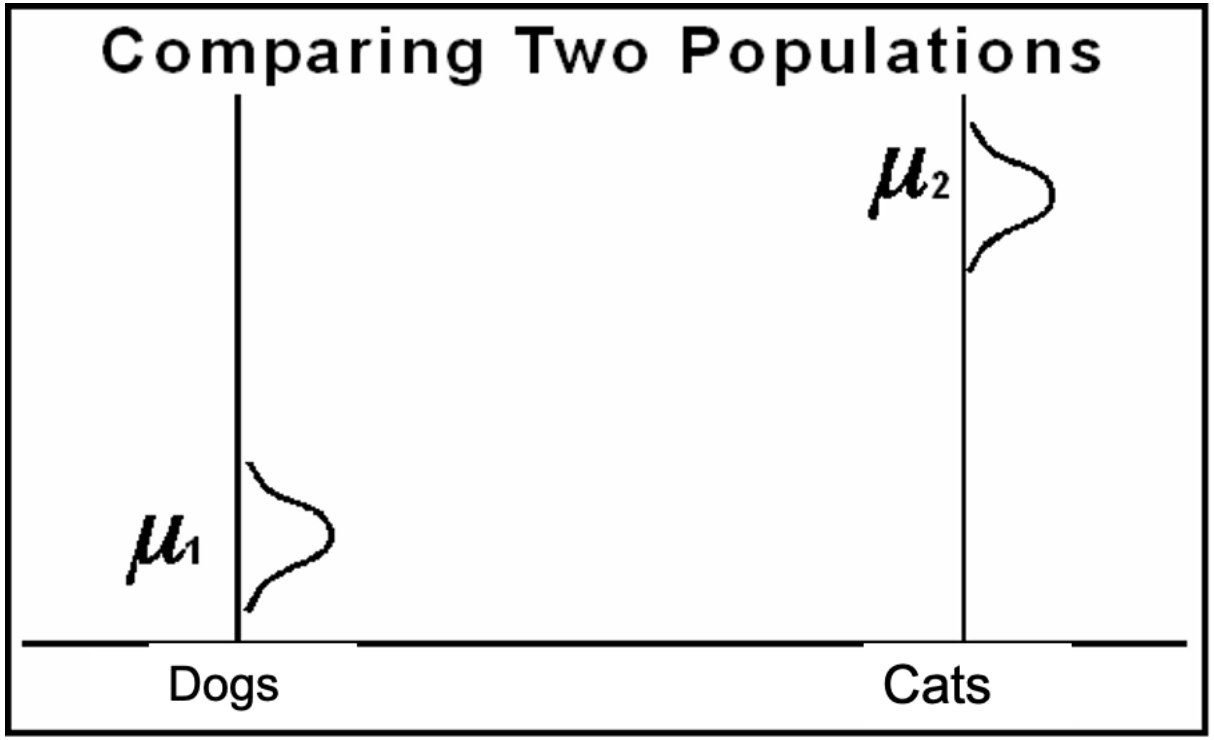
```
##
##  One Sample t-test
##
## data:  ames$Sale_Price
## t = 1.8945, df = 2929, p-value = 0.02913
## alternative hypothesis: true mean is greater than 178000
## 95 percent confidence interval:
##  178367.7      Inf
## sample estimates:
## mean of x
##  180796.1
```

Two-Sample t-tests

Testing the difference between two means

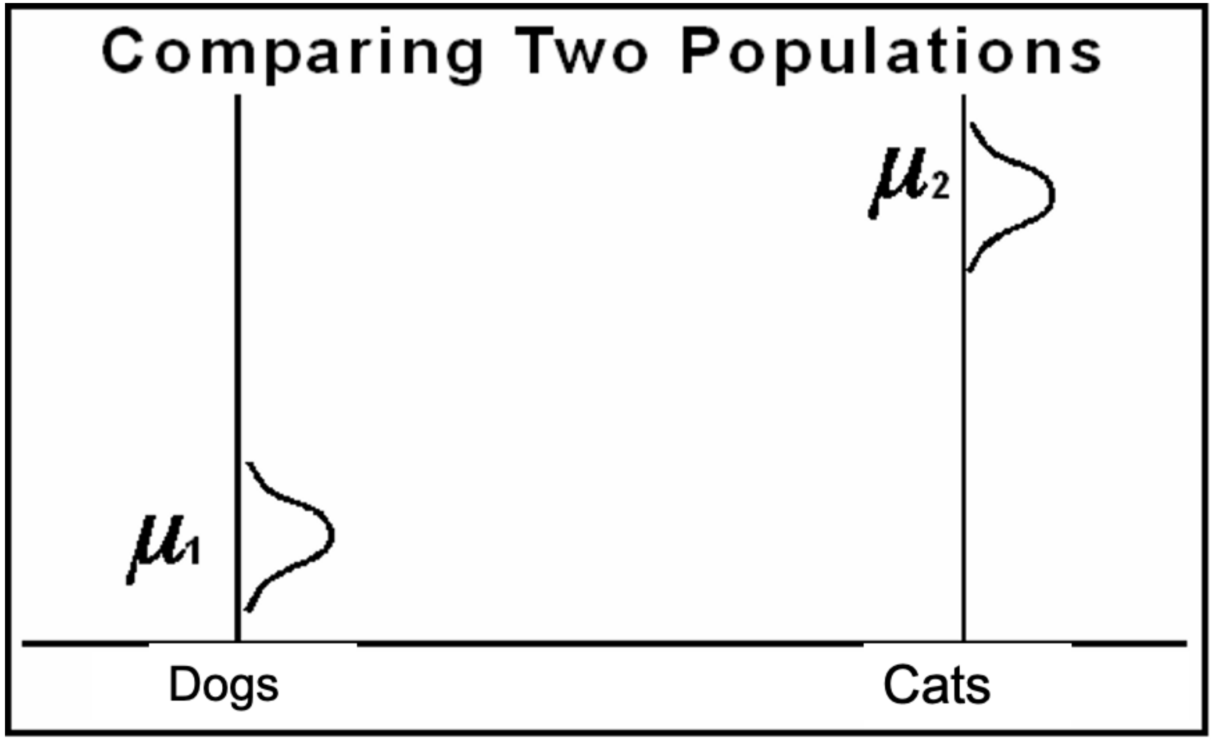
Assumptions

- Independent observations
- Normally distributed data for each group
- Equal variances for each group



Assumptions

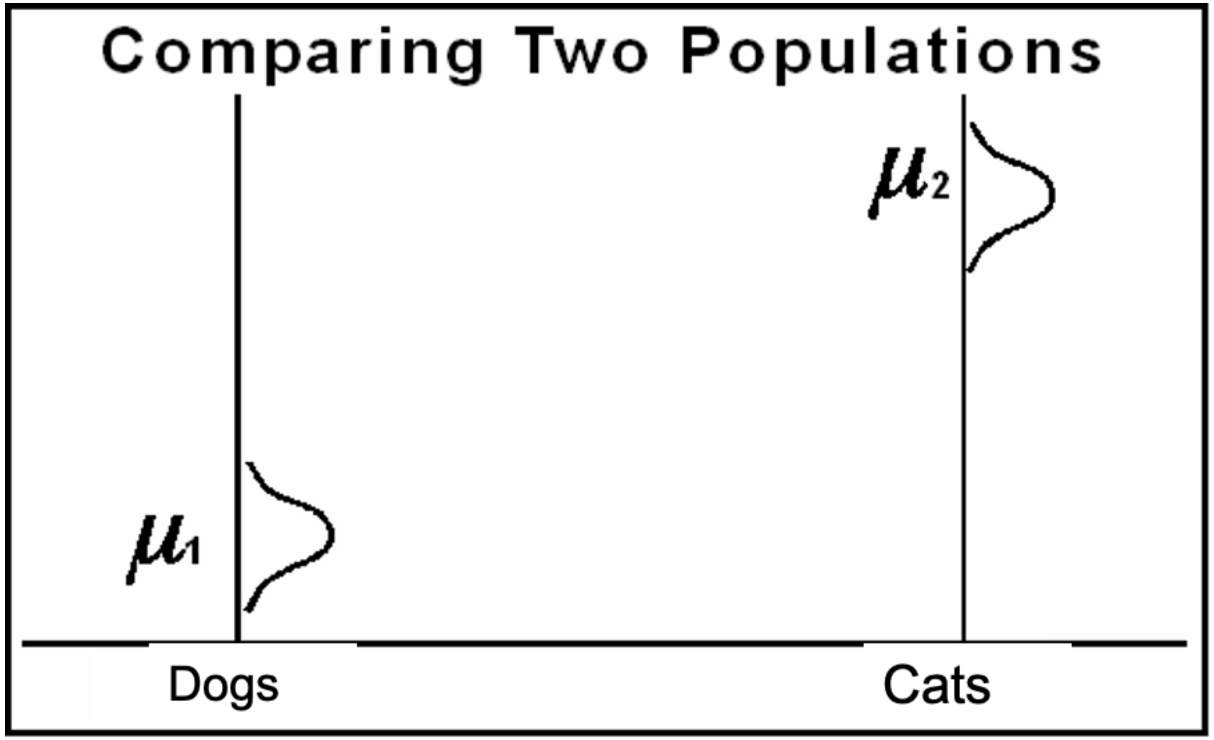
- Independent observations
- **Normally distributed data for each group**
- Equal variances for each group



Assumptions

- Independent observations
- Normally distributed data for each group
- **Equal variances for each group**

Tested formally with F-Test
to determine which t-test to use:
Equal Variance: Pooled Variance t-test
Unequal Variance: Satterthwaite's t-test



F-Test for Equality of Variances

Requires that both populations are normally distributed!

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

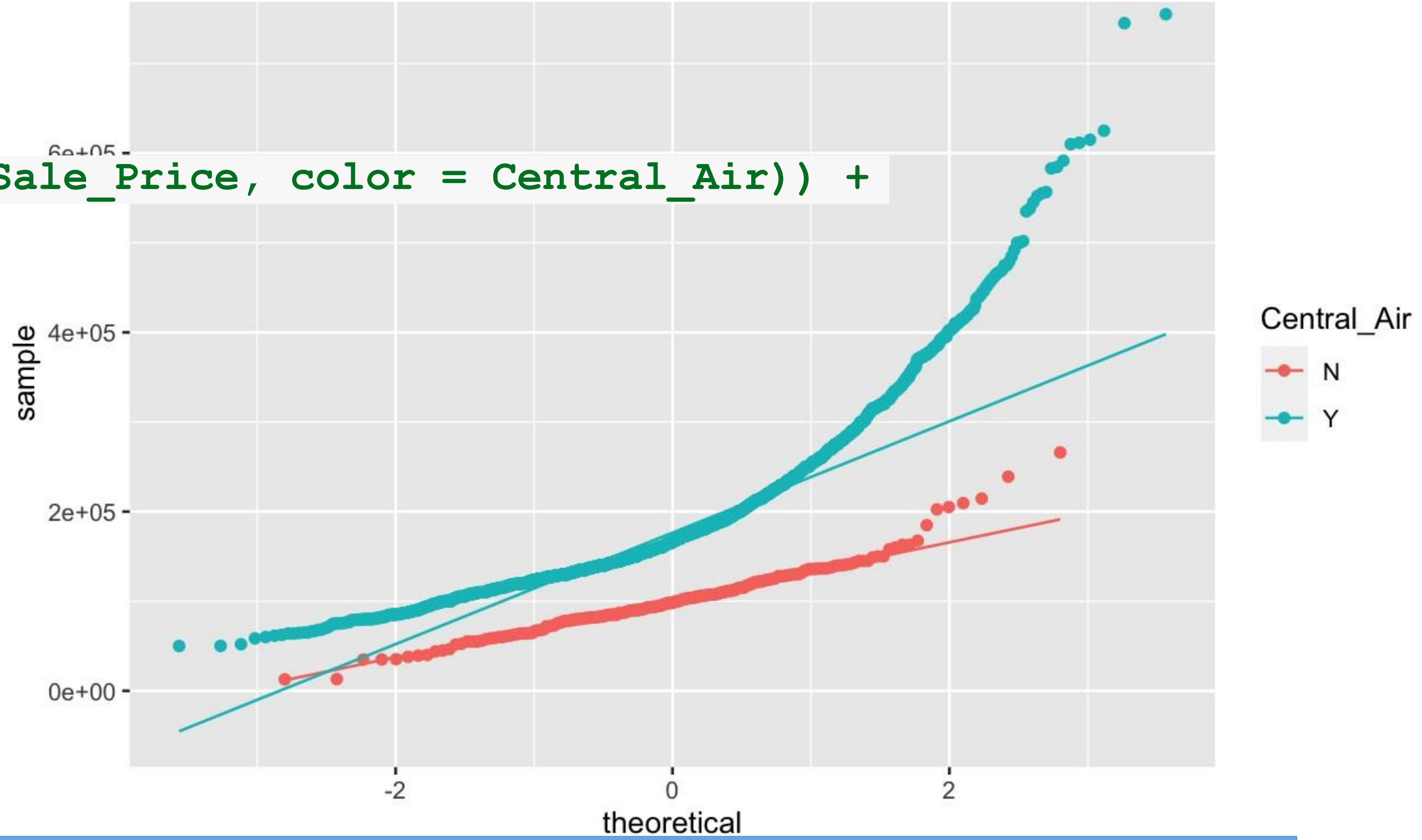
$$F = \frac{\max(s_1^2,\ s_2^2)}{\min(s_1^2,\ s_2^2)}$$

Two-sample t-test in R:

Are the mean sale prices of houses with and without central air the same?

1. Verify normality

```
ggplot(data = ames, aes(sample = Sale_Price, color = Central_Air)) +
      stat_qq() +
      stat_qq_line()
```



This doesn’t look like normality in both groups. When Central_Air=‘No’, it’s closer.

Two-sample t-test in R:

Are the mean sale prices of houses with and without central air the same?

2. Check Equality of Variances

```
var.test(Sale_Price ~ Central_Air, data = ames)
```

Reject null of equal variances and conclude that the variances are NOT equal.

```
##  F test to compare two variances
##
## data:  Sale_Price by Central_Air
## F = 0.2258, num df = 195, denom df = 2733, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1854873 0.2800271
## sample estimates:
## ratio of variances
##          0.2257977
```

Wait a minute. Don’t we need normality for the F-Test?!? Yes!
This situation calls for a nonparametric test. We’ll proceed for sake of illustration

Two-sample t-test in R:

Are the mean sale prices of houses with and without central air the same?

3. Perform two-sample t-test

```
t.test(Sale_Price ~ Central_Air, data = ames, var.equal = FALSE)
```

```
##  Welch Two Sample t-test
##
## data:  Sale_Price by Central_Air
## t = -27.433, df = 336.06, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -90625.69 -78498.92
## sample estimates:
## mean in group N mean in group Y
##        101890.5        186452.8
```

Reject the null hypothesis that the means are equal and conclude that the mean Sale Price of homes with and without Central Air are different.

Nonparametric Test: Wilcoxon Rank

Are the median sale prices of houses with and without central air the same?*

3. Perform Wilcoxon rank test (when normality assumption fails):

```
wilcox.test(Sale_Price ~ Central_Air, data = ames)
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  Sale_Price by Central_Air
## W = 63164, p-value < 2.2e-16
## alternative hypothesis: true location shift
## is not equal to 0
```

Conclude that the median Sale Prices of homes with and without Central Air are different.

* The actual conclusions of this test depend on the shape of the underlying data. See the table in Section 1.5.2 of the text for details.

Lab 2

Don’t forget to take the lab check on Moodle!