Lecture 2

Summarizing the Sample

WARNING: Today's lecture may bore some of you…

It's (sort of) not my fault…I'm required to teach you about what we're going to cover today.

I’ll try to make it as exciting as possible…

But you’re more than welcome to fall asleep if you feel like this stuff is too easy

Lecture Summary

- Once we obtained our sample, we would like to summarize it.
- Depending on the type of the data (numerical or categorical) and the dimension (univariate, paired, etc.), there are different methods of summarizing the data.
  - Numerical data have two subtypes: discrete or continuous
  - Categorical data have two subtypes: nominal or ordinal
- Graphical summaries:
  - **Histograms**: Visual summary of the sample distribution
  - **Quantile-Quantile Plot**: Compare the sample to a known distribution
  - **Scatterplot**: Compare two pairs of points in X/Y axis.

Three Steps to Summarize Data

1. Classify sample into different type

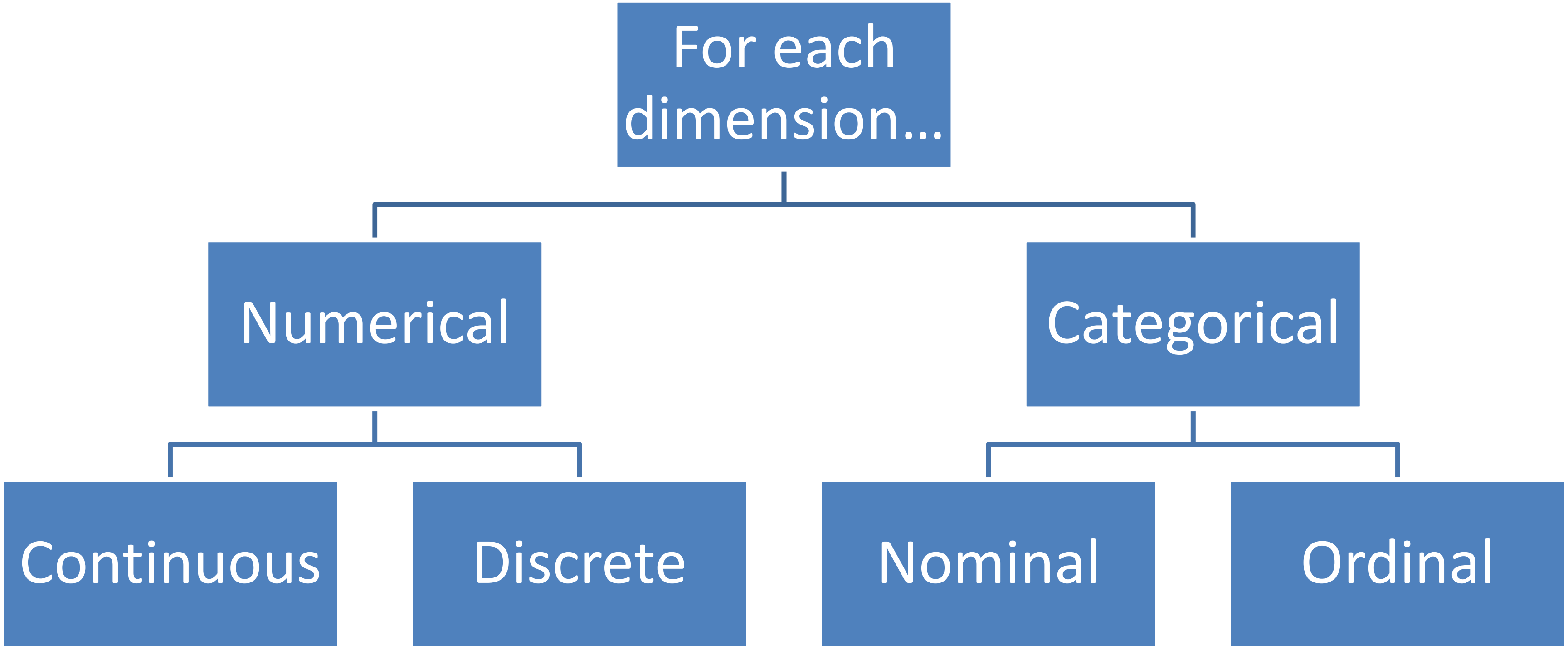
2. Depending on the type, use appropriate numerical summaries

3. Depending on the type, use appropriate visual summaries

Data Classification

- Data/Sample: $(X_1, \ldots, X_n)$
- Dimension of X_i (i.e. the number of measurements per unit i)
  - Univariate: one measurement for unit i (height)
  - Multivariate: multiple measurements for unit i (height, weight, sex)
- For each dimension, X_i can be numerical or categorical
- Numerical variables
  - Discrete: human population, natural numbers, (0,5,10,15,20,25,etc..)
  - Continuous: height, weight
- Categorical variables
  - Nominal: categories have no ordering (sex: male/female)
  - Ordinal: categories are ordered (grade: A/B/C/D/F, rating: high/low)

Data Types


For each dimension…
Numerical
Categorical
Continuous
Discrete
Nominal
Ordinal

Summaries for numerical data

- Center/location: measures the "center" of the data
  - Examples: sample mean and sample median
- Spread/Dispersion: measures the "spread" or "fatness" of the data
  - Examples: sample variance, interquartile range
- Order/Rank: measures the ordering/ranking of the data
  - Examples: order statistics and sample quantiles

Summary	Type of Sample	Formula	Notes
Sample mean, $\hat{\mu}, \bar{X}$	Continuous	$\frac{1}{n}\sum_{i=1}^{n} X_i$	• Summarizes the "center" of the data • Sensitive to outliers
Sample variance, $\widehat{\sigma^2}, S^2$	Continuous	$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$	• Summarizes the "spread" of the data • Outliers may inflate this value
Order statistic, $X_{(i)}$	Continuous	i[th] largest value of the sample	• Summarizes the order/rank of the data
Sample median, $X_{0.5}$	Continuous	If n is even: $\frac{\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right)}{2}$ If n is odd: $X_{\left(\frac{n}{2}+0.5\right)}$	• Summarizes the "center" of the data • Robust to outliers
Sample α quartiles, X_α $0 \leq \alpha \leq 1$	Continuous	If $\alpha = \frac{i}{n+1}$ for $i = 1, \dots, n$: $X_\alpha = X_{(i)}$ Otherwise, do linear interpolation	• Summarizes the order/rank of the data • Robust to outliers
Sample Interquartile Range (Sample IQR)	Continuous	$X_{0.75} - X_{0.25}$	• Summarizes the "spread" of the data • Robust to outliers

Multivariate numerical data

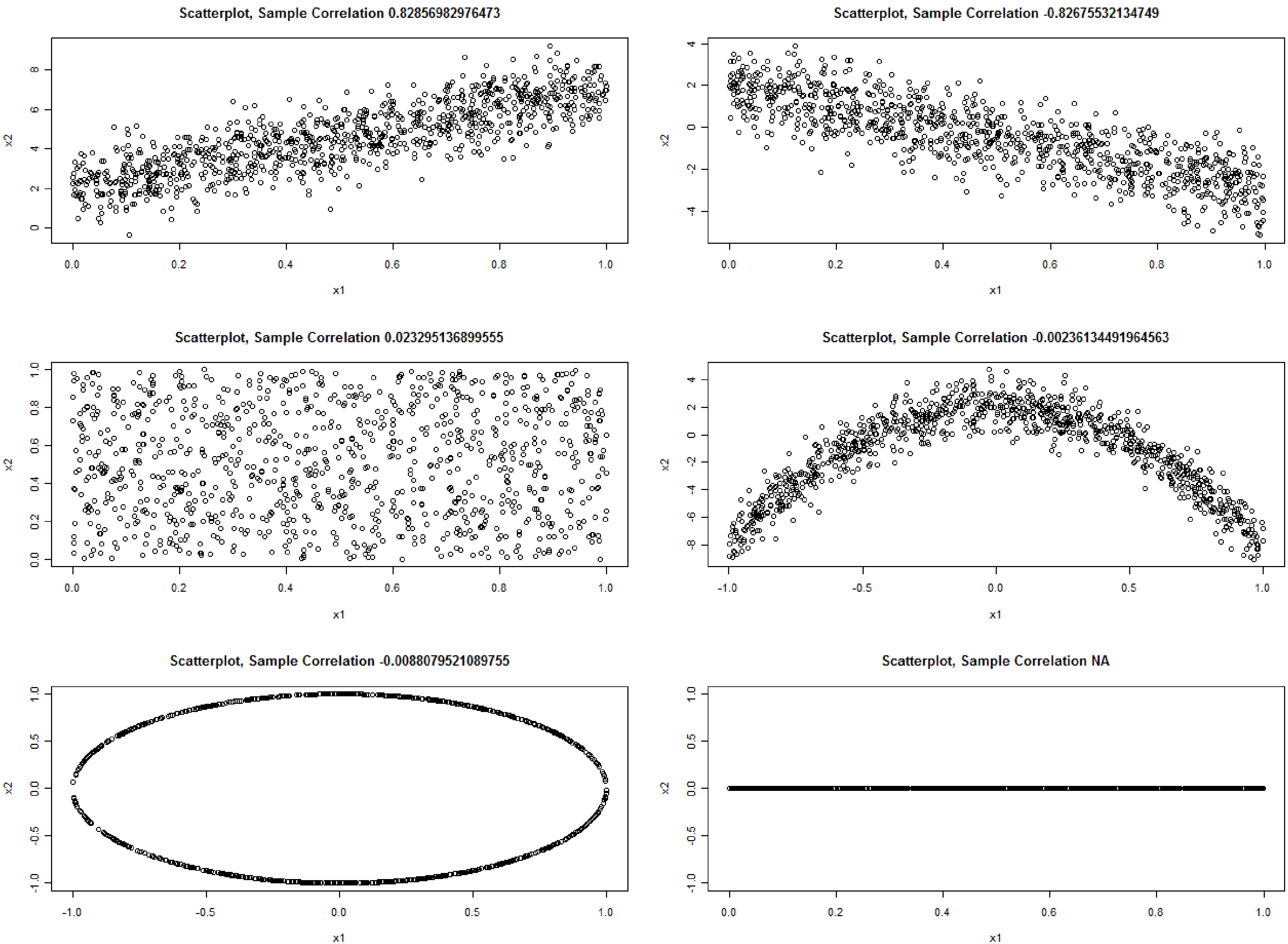
- Each dimension in multivariate data is univariate and hence, we can use the numerical summaries from univariate data (e.g. sample mean, sample variance)
- However, to study two measurements and their relationship, there are numerical summaries to analyze it
- Sample Correlation and Sample Covariance

Sample Correlation and Covariance

- Measures linear relationship between two measurements, X_{i1} and X_{i2}, where $X_i = (X_{i1}, X_{i2})$

- $\hat{\rho} = \frac{\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2)}{(n-1)\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}}$
  - $-1 \leq \hat{\rho} \leq 1$
  - Sign indicates proportional (positive) or inversely proportional (negative) relationship
  - If X_{i1} and X_{i2} have a perfect linear relationship, $\hat{\rho} = 1$ or -1

- Sample covariance
=$\hat{\rho}\hat{\sigma}_{X_1}\hat{\sigma}_{X_2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$


Scatterplot, Sample Correlation 0.82856982976473
Scatterplot, Sample Correlation -0.82675532134749
Scatterplot, Sample Correlation 0.023295136899555
Scatterplot, Sample Correlation -0.00236134491964563
Scatterplot, Sample Correlation -0.0088079521089755
Scatterplot, Sample Correlation NA
x1
x2

How about categorical data?

Summaries for categorical data

- Frequency/Counts: how frequent is one category
- Generally use tables to count the frequency or proportions from the total
- Example: Stat 431 class composition

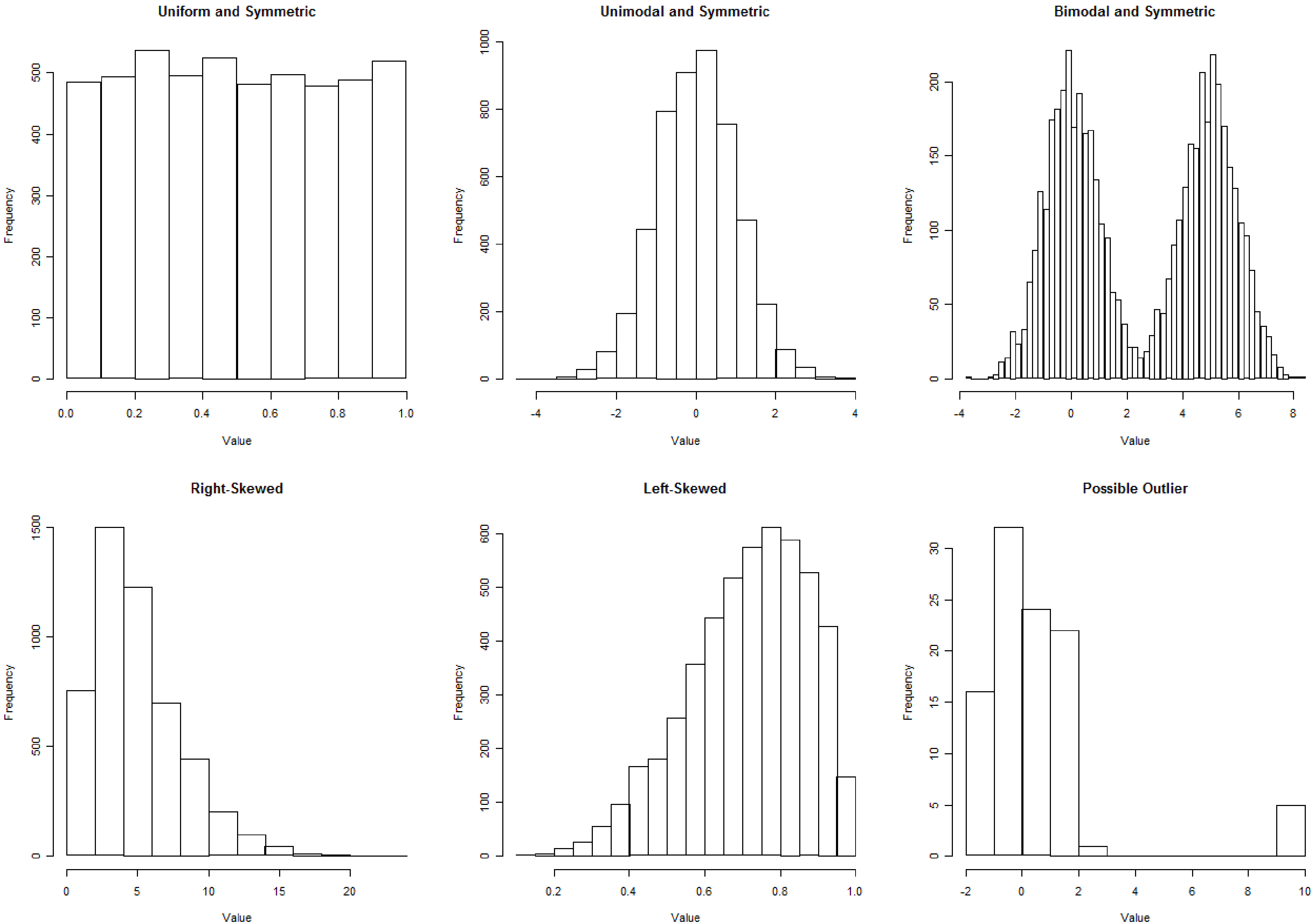
	Undergrad	Graduate	Staff
Counts	17	1	2
Proportions	0.85	0.05	0.1

Are there visual summaries of the data?

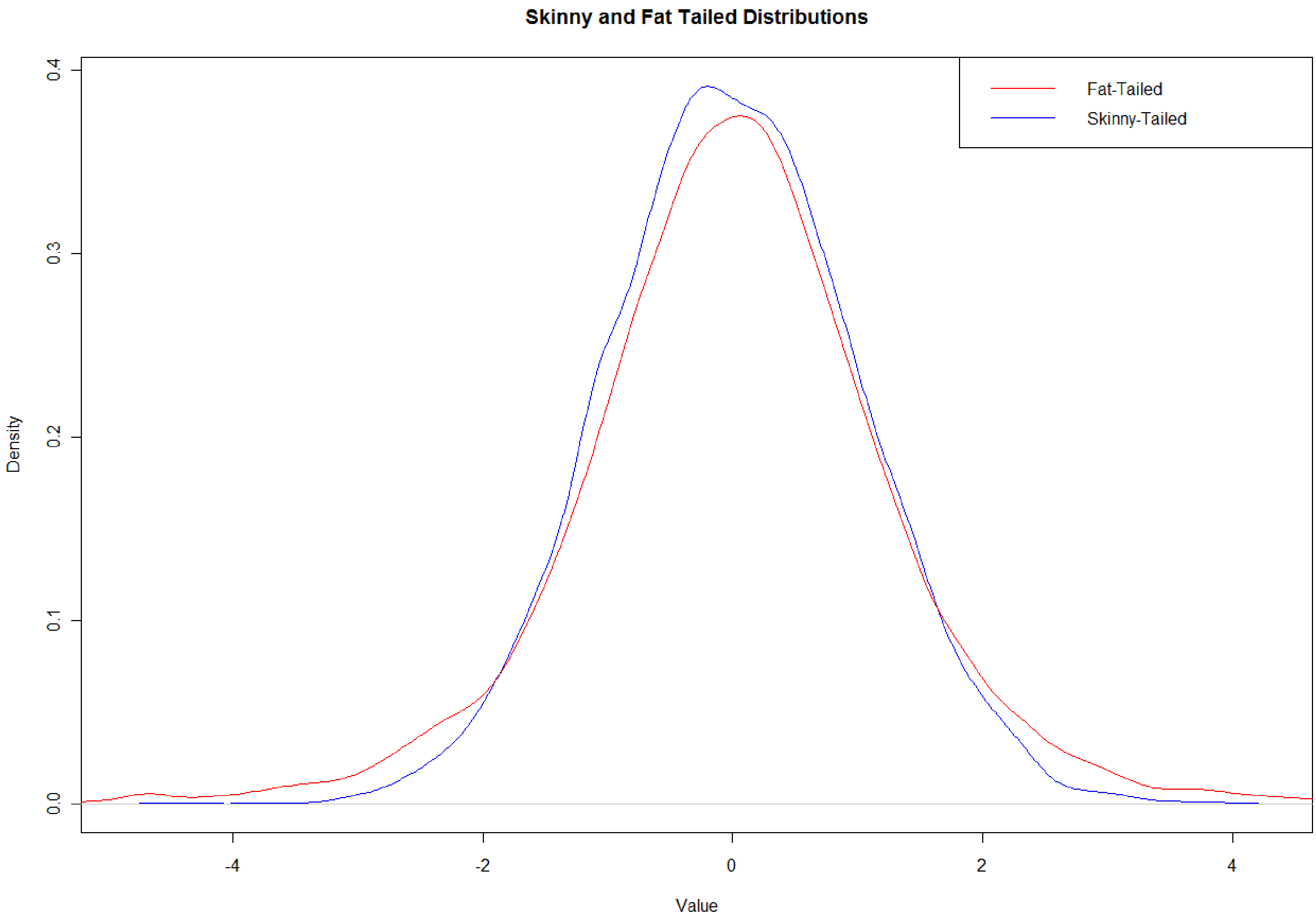
Histograms, boxplots, scatterplots, and QQ plots

Histograms

- For numerical data
- A method to show the "shape" of the data by tallying frequencies of the measurements in the sample
- Characteristics to look for:
  - Modality: Uniform, unimodal, bimodal, etc.
  - Skew: Symmetric (no skew), right/positive-skewed, left/negative-skewed distributions
  - Quantiles: Fat tails/skinny tails
  - Outliers



Uniform and Symmetric
Frequency
0 100 200 300 400 500
0.0 0.2 0.4 0.6 0.8 1.0
Value
Unimodal and Symmetric
Frequency
0 200 400 600 800 1000
-4 -2 0 2 4
Value
Bimodal and Symmetric
Frequency
0 50 100 150 200
-4 -2 0 2 4 6 8
Value
Right-Skewed
Frequency
0 500 1000 1500
0 5 10 15 20
Value
Left-Skewed
Frequency
0 100 200 300 400 500 600
0.2 0.4 0.6 0.8 1.0
Value
Possible Outlier
Frequency
0 5 10 15 20 25 30
-2 0 2 4 6 8 10
Value



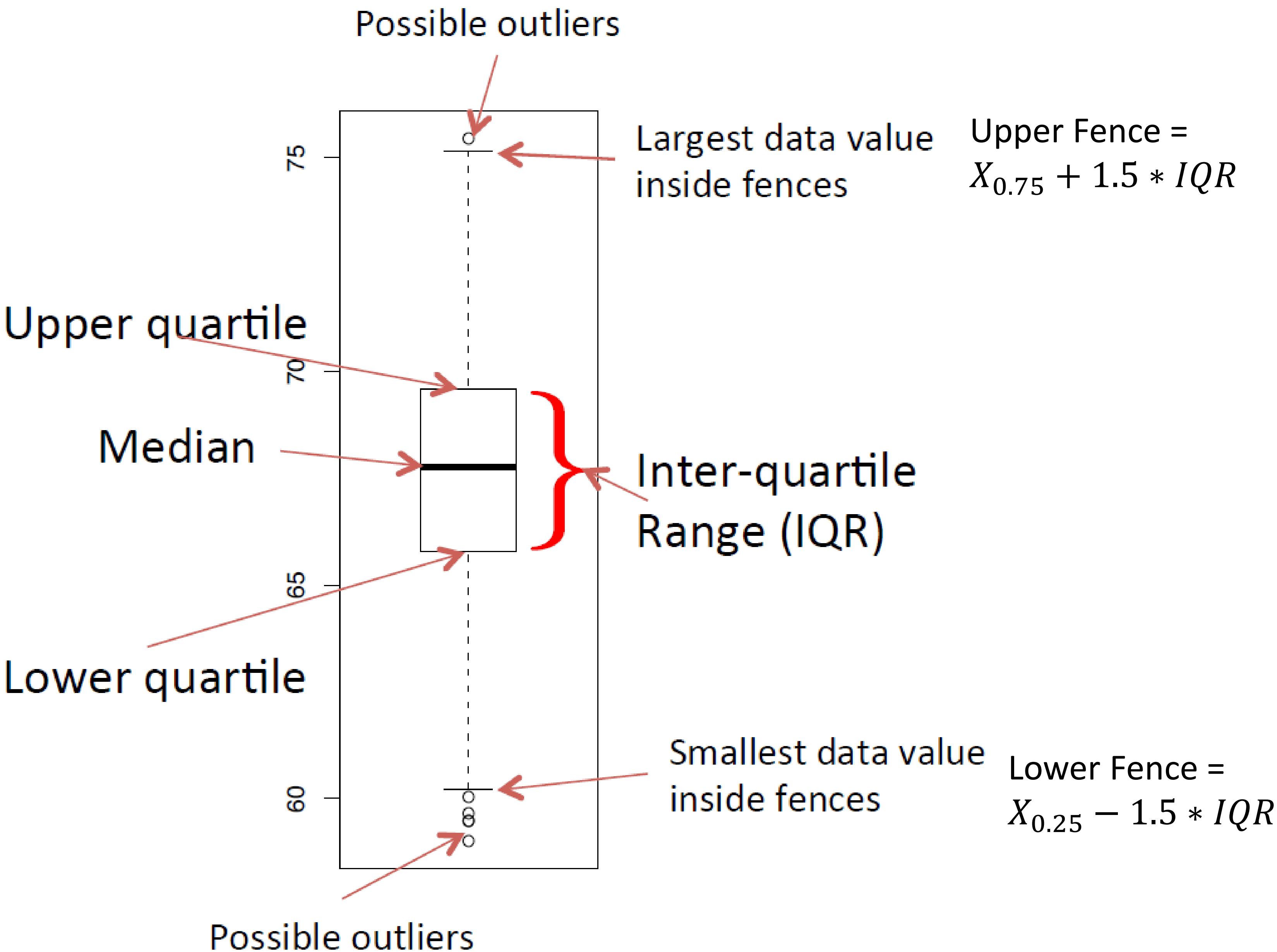
Skinny and Fat Tailed Distributions
Fat-Tailed
Skinny-Tailed
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Value

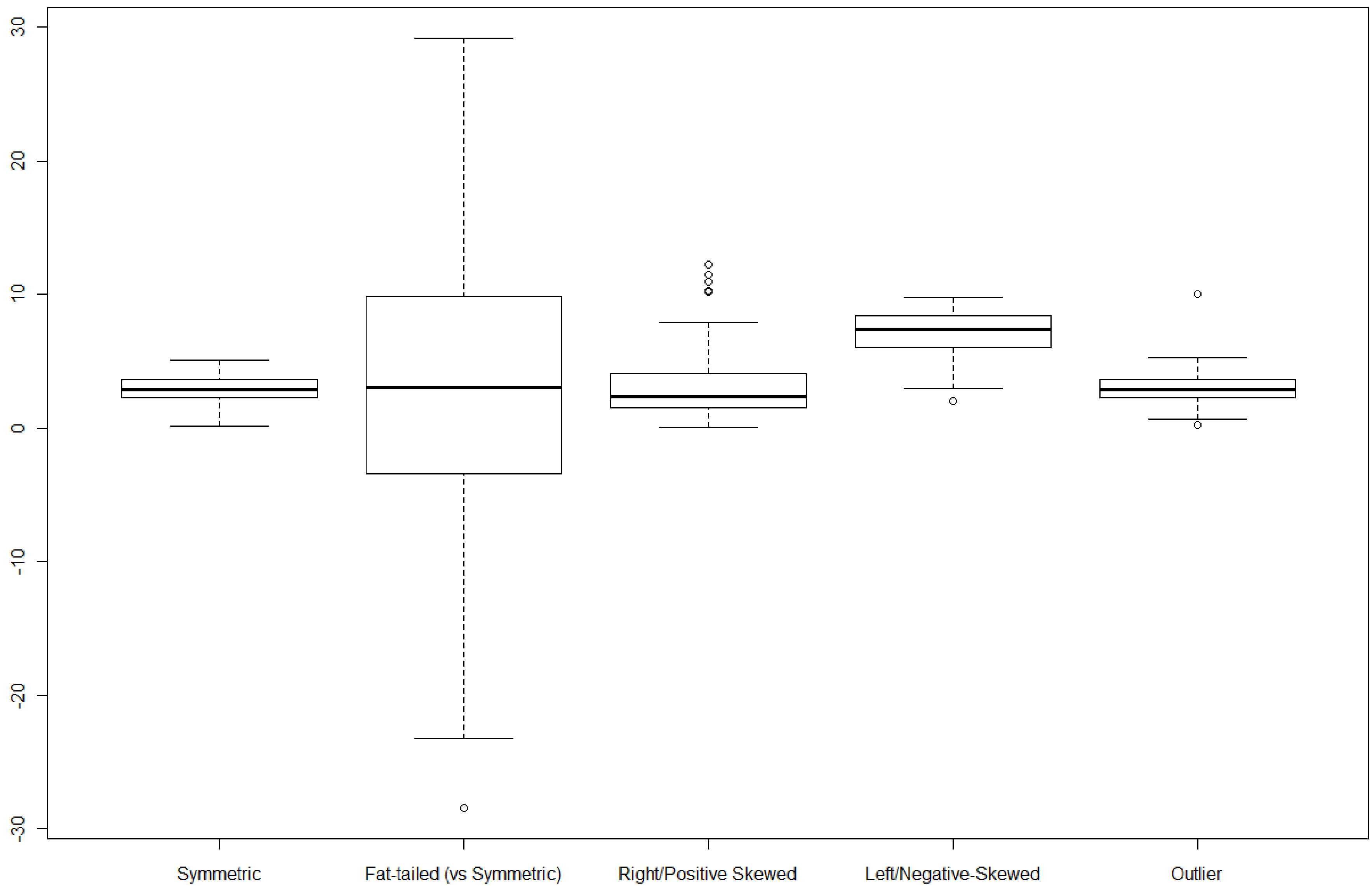
Boxplots

- For numerical data

- Another way to visualize the "shape" of the data. Can identify…
  - Symmetric, right/positive-skewed, and left/negative-skewed distributions
  - Fat tails/skinny tails
  - Outliers

- However, boxplots cannot identify modes (e.g. unimodal, bimodal, etc.)


Possible outliers
Largest data value inside fences
Upper Fence = $X_{0.75} + 1.5 * IQR$
75
70
65
60
Upper quartile
Median
Inter-quartile Range (IQR)
Lower quartile
Smallest data value inside fences
Lower Fence = $X_{0.25} - 1.5 * IQR$
Possible outliers

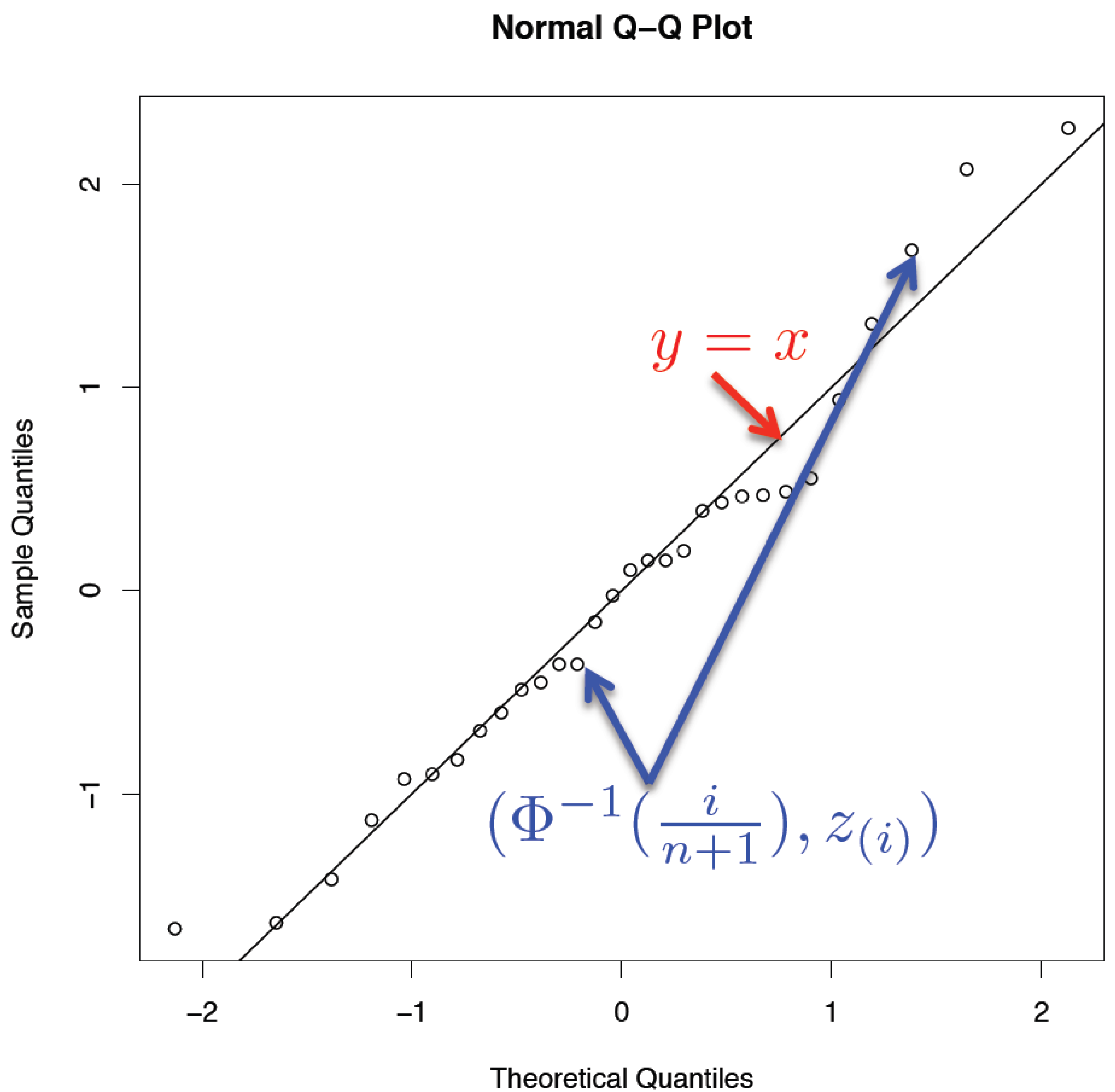

30
20
10
0
-10
-20
-30
Symmetric
Fat-tailed (vs Symmetric)
Right/Positive Skewed
Left/Negative-Skewed
Outlier

Quantile-Quantile Plots (QQ Plots)

- For numerical data: visually compare collected data with a known distribution

- Most common one is the Normal QQ plots
  - We check to see whether the sample follows a normal distribution
  - This is a common assumption in statistical inference that your sample comes from a normal distribution

- Summary: If your scatterplot "hugs" the line, there is good reason to believe that your data follows the said distribution.

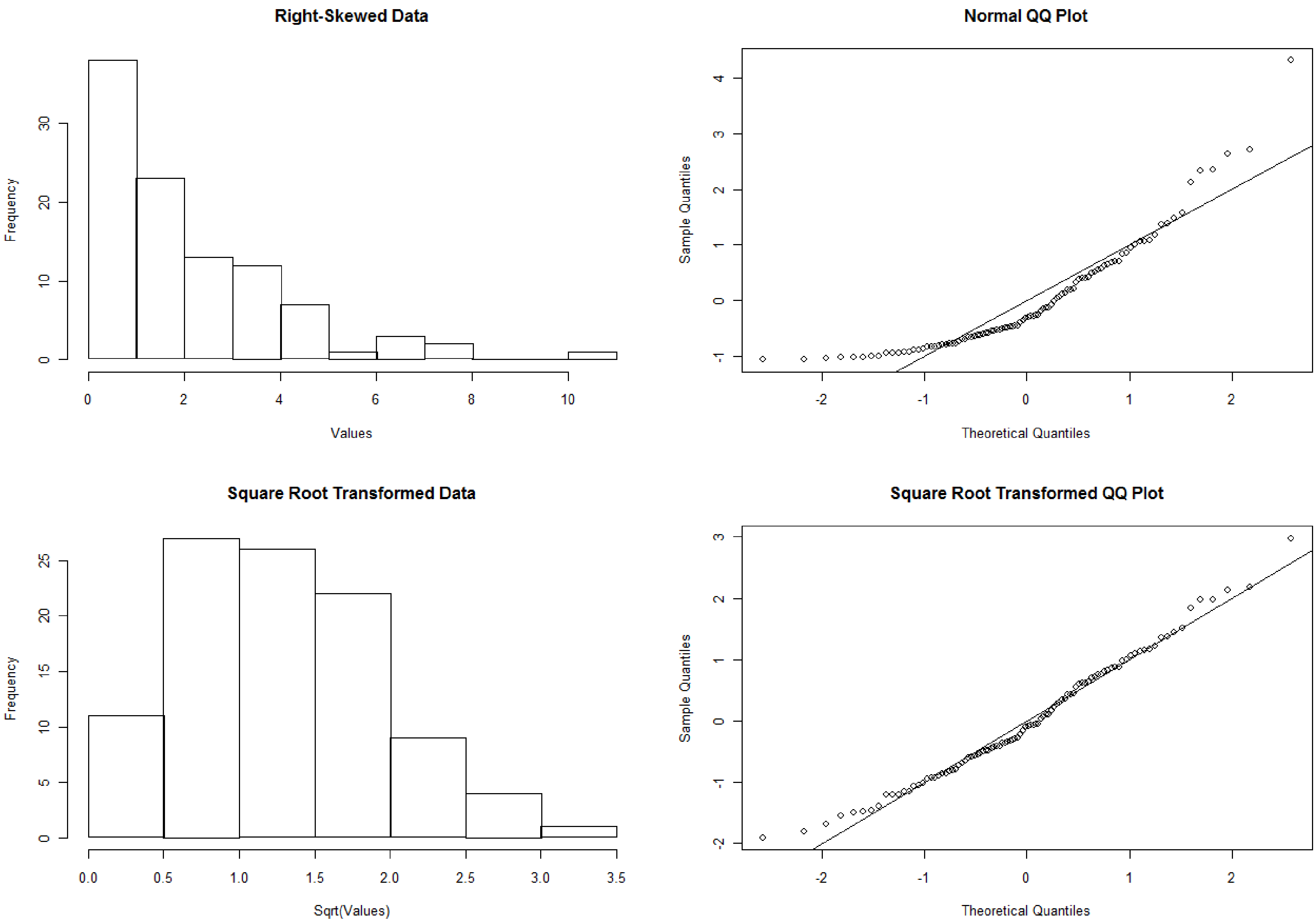

Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
$y = x$
$(\Phi^{-1}(\frac{i}{n+1}), z_{(i)})$

Making a Normal QQ plot

1. Compute z-scores: $\mathrm{Z_i} = \frac{X_i - \bar{\mathrm{X}}}{\hat{\sigma}}$
2. Plot $\frac{i}{n+1}$th theoretical normal quantile against ith ordered z-scores (i.e. $\left(\Phi\left(\frac{i}{n+1}\right)^{-1}, Z_{(i)}\right)$
   - Remember, $Z_{(i)}$ is the $\frac{i}{n+1}$ sample quantile (see numerical summary table)
3. Plot $Y = X$ line to compare the sample to the theoretical normal quantile

If your data is not normal…

- You can perform transformations to make it look normal
- For right/positively-skewed data: Log/square root
- For left/negatively-skewed data: exponential/square


Right-Skewed Data
Frequency
0
10
20
30
0
2
4
6
8
10
Values
Normal QQ Plot
Sample Quantiles
-1
0
1
2
3
4
-2
-1
0
1
2
Theoretical Quantiles
Square Root Transformed Data
Frequency
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Sqrt(Values)
Square Root Transformed QQ Plot
Sample Quantiles
-2
-1
0
1
2
3
-2
-1
0
1
2
Theoretical Quantiles

Comparing the three visual techniques

Histograms

- <u>Advantages</u>:
  - With properly-sized bins, histograms can summarize any shape of the data (modes, skew, quantiles, outliers)
- <u>Disadvantages</u>:
  - Difficult to compare side-by-side (takes up too much space in a plot)
  - Depending on the size of the bins, interpretation may be different

Boxplots

- <u>Advantages</u>:
  - Don’t have to tweak with “graphical” parameters (i.e. bin size in histograms)
  - Summarize skew, quantiles, and outliers
  - Can compare several measurements side-by-side
- <u>Disadvantages</u>:
  - Cannot distinguish modes!

QQ Plots

- <u>Advantages</u>:
  - Can identify whether the data came from a certain distribution
  - Don’t have to tweak with “graphical” parameters (i.e. bin size in histograms)
  - Summarize quantiles
- <u>Disadvantages</u>:
  - Difficult to compare side-by-side
  - Difficult to distinguish skews, modes, and outliers

Scatterplots

- For multidimensional, numerical data:
  $\mathrm{X_i} = (X_{i1}, X_{i2}, \dots, X_{ip})$

- Plot points on a p dimensional axis

- Characteristics to look for:
  - Clusters
  - General patterns

- See previous slide on sample correlation for examples. See R code for cool 3D animation of the scatterplot

Lecture Summary

- Once we obtain a sample, we want to summarize it.
- There are numerical and visual summaries
  - Numerical summaries depend on the data type (numerical or categorical)
  - Graphical summaries discussed here are mostly designed for numerical data
- We can also look at multidimensional data and examine the relationship between two measurement
  - E.g. sample correlation and scatterplots

Extra Slides

Why does the QQ plot work?

- You will prove it in a homework assignment ☺

- Basically, it has to do with the fact that if your sample came from a normal distribution (i.e. $X_i \sim N(\mu, \sigma^2)$), then $Z_i = \frac{X_i - \bar{X}}{\hat{\sigma}} \sim t_{n-1}$ where t_{n-1} is a t-distribution.

- With large samples ($n \geq 30$), $t_{n-1} \approx N(0,1)$. Thus, if your sample is truly normal, then it should follow the theoretical quantiles.

- If this is confusing to you, wait till lecture on sampling distribution

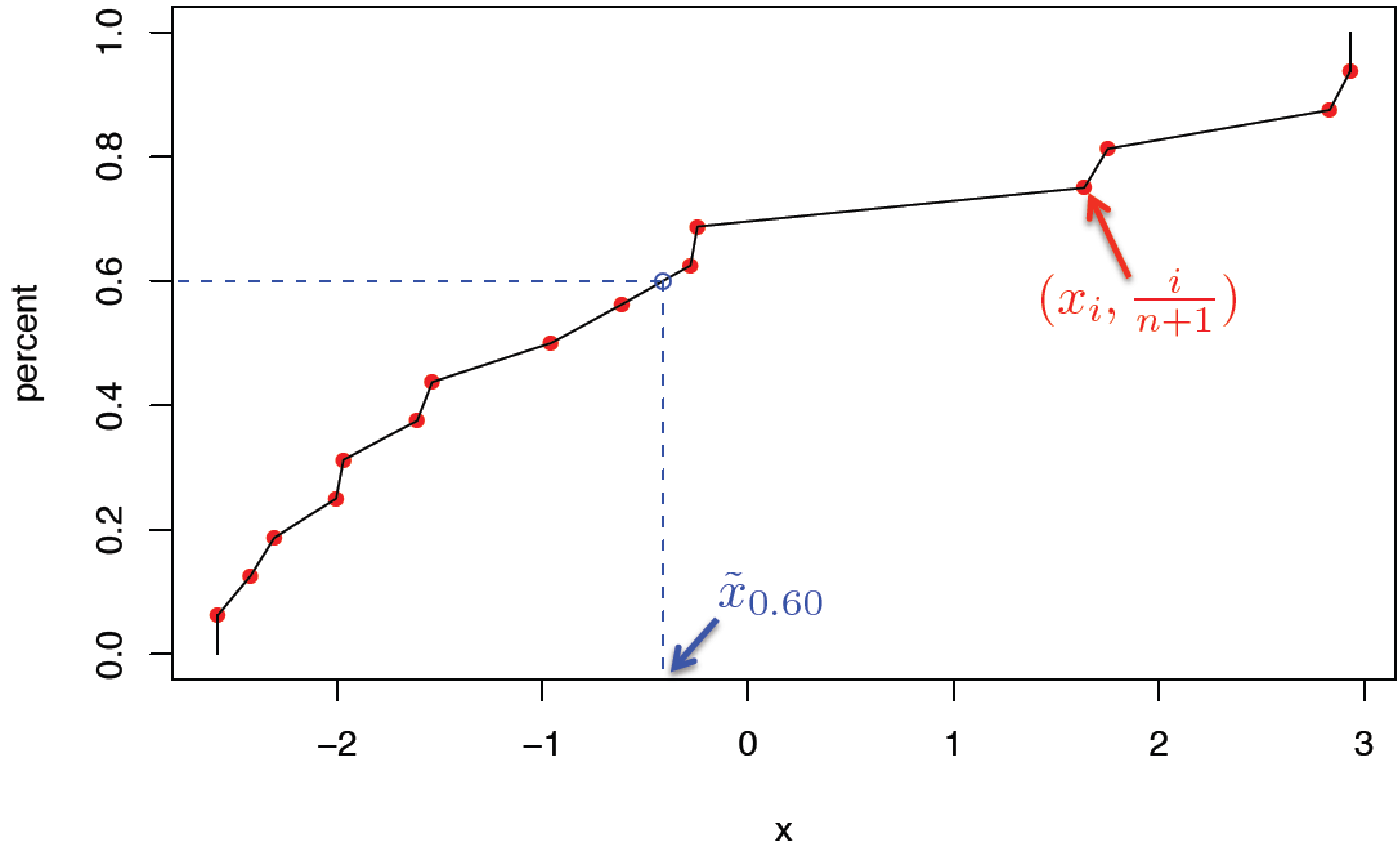
Linear Interpolation in Sample Quantiles

If you want an estimate of the sample quantile that is not $\frac{i}{n+1}$, then you do a linear interpolation

1. For a given α, find $i = 1, \ldots, n$ such that $\frac{i}{n+1} \leq \alpha \leq \frac{i+1}{n+1}$

2. Fit a line, $y = a * x + b$, with two points $\left(X_{(i)}, \frac{i}{n+1}\right)$ and $\left(X_{(i+1)}, \frac{i+1}{n+1}\right)$.

3. Plug in y as your α and solve for x. This x will be your X_α quantile.



Schematic plot of sample quantile definition