Descriptive Statistics

Michael Luu

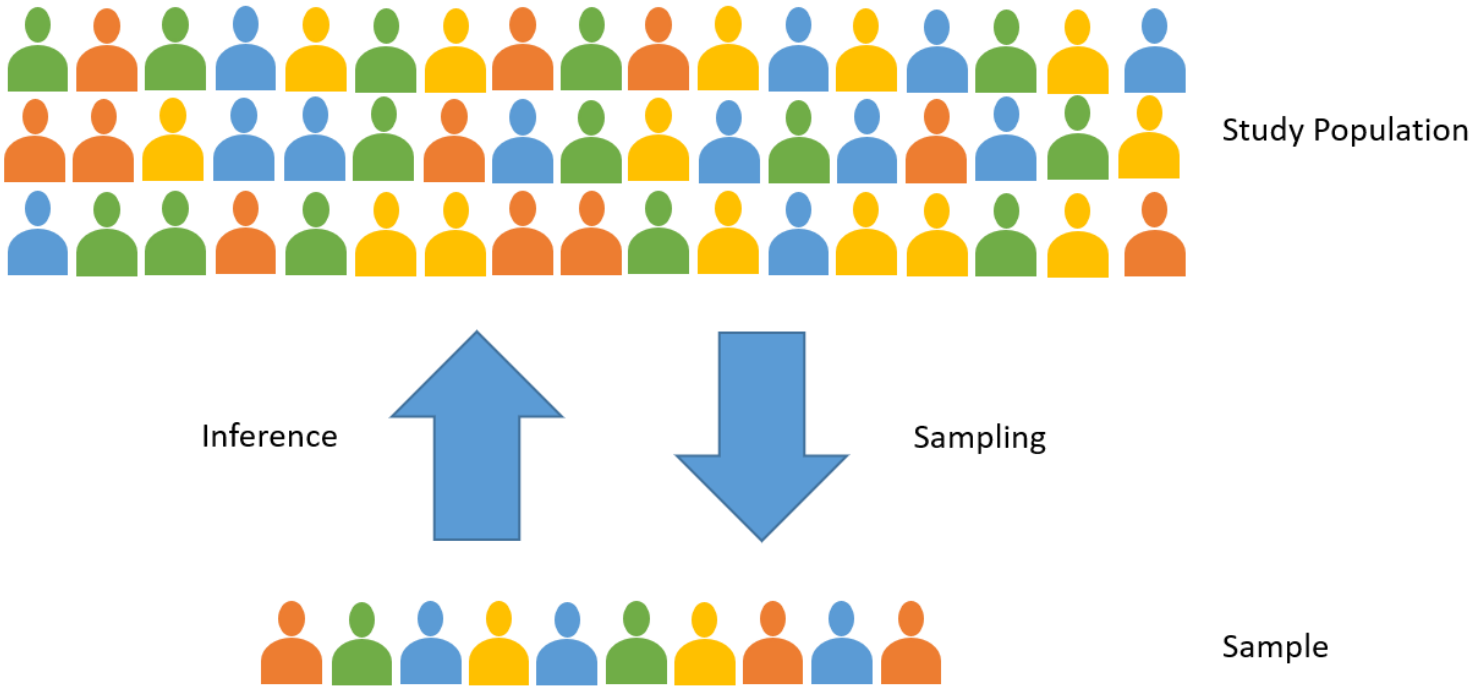
Biostatistics and Bioinformatics Research Center

Samuel Oschin Comprehensive Cancer Institute

Cedars Sinai Medical Center

10/03/2018

Introduction



Introduction

What is descriptive statistics ?

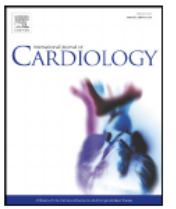
- **Descriptive statistics** is a collection of statistical measures and tools used to give us a better sense of the data we have in front of us (Sample)
- Not to be confused with **inferential statistics** where we are trying to reach conclusions that extend beyond the immediate data we have available (Population).

Introduction

Contents lists available at ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard



Inverse association of MRI-derived native myocardial T1 and perfusion reserve index in women with evidence of ischemia and no obstructive CAD: A pilot study

Jaime L. Shaw [a,b], Michael D. Nelson [a,c], Janet Wei [a,c], Manish Motwani [d], Sofy Landes [c], Puja K. Mehta [c], Louise E.J. Thomson [c,d], Daniel S. Berman [a,d,e], Debiao Li [a,b,e], C. Noel Bairey Merz [c,e], Behzad Sharif [a,b,e,*]

[a] Biomedical Imaging Research Institute, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, United States
[b] Department of Bioengineering, University of California Los Angeles, CA, United States
[c] Women's Heart Center, Cedars-Sinai Heart Institute, Los Angeles, CA, United States
[d] Department of Imaging, Cedars-Sinai Medical Center, Los Angeles, CA, United States
[e] David Geffen School of Medicine, University of California Los Angeles, CA, United States

Introduction

Background

Magnetic resonance "native T1" mapping has been shown to be capable of characterizing abnormal micro circulation in patients with coronary artery disease (CAD). However the potential role of native T1 as an imaging marker and its association with indices of diastolic function or vasodilator-induced myocardial ischemia have not been explored.

Sample Population

- Twenty-two female patients with INOCA and twelve female reference controls with matching age and body-mass index were studied.

Introduction

id	female	inoca	age	bmi	bsa	hypertension	type_ii_diabetes	ever_smoker	dyslipidemia	family_hist_cad	beta_blocker	calc_chan_blocker	nitrates	aspirin	ace_inhib	statin	chest_pain	dyspnea
1	1	0	50.25464	24.30644	1.845874	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	76.62715	22.81795	1.780491	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	58.67225	21.50502	1.581906	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	41.57575	21.90134	1.450523	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	37.76118	19.94607	1.610652	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	47.15073	30.62259	2.123460	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	47.04402	29.34906	2.002971	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	45.81539	25.19129	1.755049	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	53.07733	27.12801	1.703724	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	0	40.55920	24.60504	1.705182	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	0	55.08457	35.66520	2.059184	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	0	36.77779	22.96200	1.740983	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	1	45.32944	27.94519	2.029789	1	0	0	1	1	1	0	0	1	0	1	1	1
14	1	1	55.47515	30.03496	1.905659	0	0	0	0	1	0	0	0	1	0	1	0	1
15	1	1	51.35964	32.55694	2.038608	1	0	0	0	1	0	0	0	1	0	1	1	1
16	1	1	62.67208	27.13848	1.596034	1	0	1	1	1	1	0	1	1	0	1	1	1
17	1	1	48.14824	13.08458	1.775631	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	57.60344	19.66378	1.594483	1	1	1	1	1	1	1	1	1	1	1	1	1
19	1	1	67.42029	30.77893	1.895389	0	0	0	0	1	0	0	0	1	0	0	0	1
20	1	1	37.45983	28.24550	1.798313	1	1	1	1	1	1	1	1	1	1	1	1	1
21	1	1	27.44932	21.87651	1.716444	1	0	0	1	1	1	0	0	1	0	1	1	1
22	1	1	49.89915	27.52762	2.070524	1	0	0	1	1	1	0	0	1	0	1	1	1
23	1	1	45.13656	22.73758	1.820579	0	0	0	0	0	0	0	0	0	0	0	0	0
24	1	1	39.95662	29.98099	1.976931	1	1	1	1	1	1	1	1	1	1	1	1	1
25	1	1	41.80441	18.55740	1.933198	0	0	0	0	0	0	0	0	0	0	0	0	0
26	1	1	38.15933	30.82834	1.615689	0	0	0	0	1	0	0	0	0	0	0	0	0
27	1	1	61.36625	30.02312	1.648442	0	0	0	0	1	0	0	0	0	0	0	0	1
28	1	1	68.53563	28.76848	1.783414	1	0	1	1	1	1	0	1	1	0	1	1	1
29	1	1	65.84178	28.45242	1.543655	1	0	1	1	1	1	0	0	1	0	1	1	1
30	1	1	41.86844	30.08427	1.891409	1	0	1	1	1	1	1	1	1	0	1	1	1
31	1	1	69.99245	23.87461	1.513025	0	0	0	0	1	0	0	0	1	0	0	0	1
32	1	1	44.40067	27.50471	2.105045	0	0	0	0	0	0	0	0	0	0	0	0	0
33	1	1	78.75432	32.87136	1.671929	0	0	0	0	0	0	0	0	0	0	0	0	0
34	1	1	58.56695	35.86424	2.115811	0	0	0	0	0	0	0	0	0	0	0	0	0

Introduction

Why Descriptive Statistics ?

- Provides an understanding of the underlying sample population
- Simplifies large amounts of data to a simpler summary
- Identifies potential measurement errors or mistakes

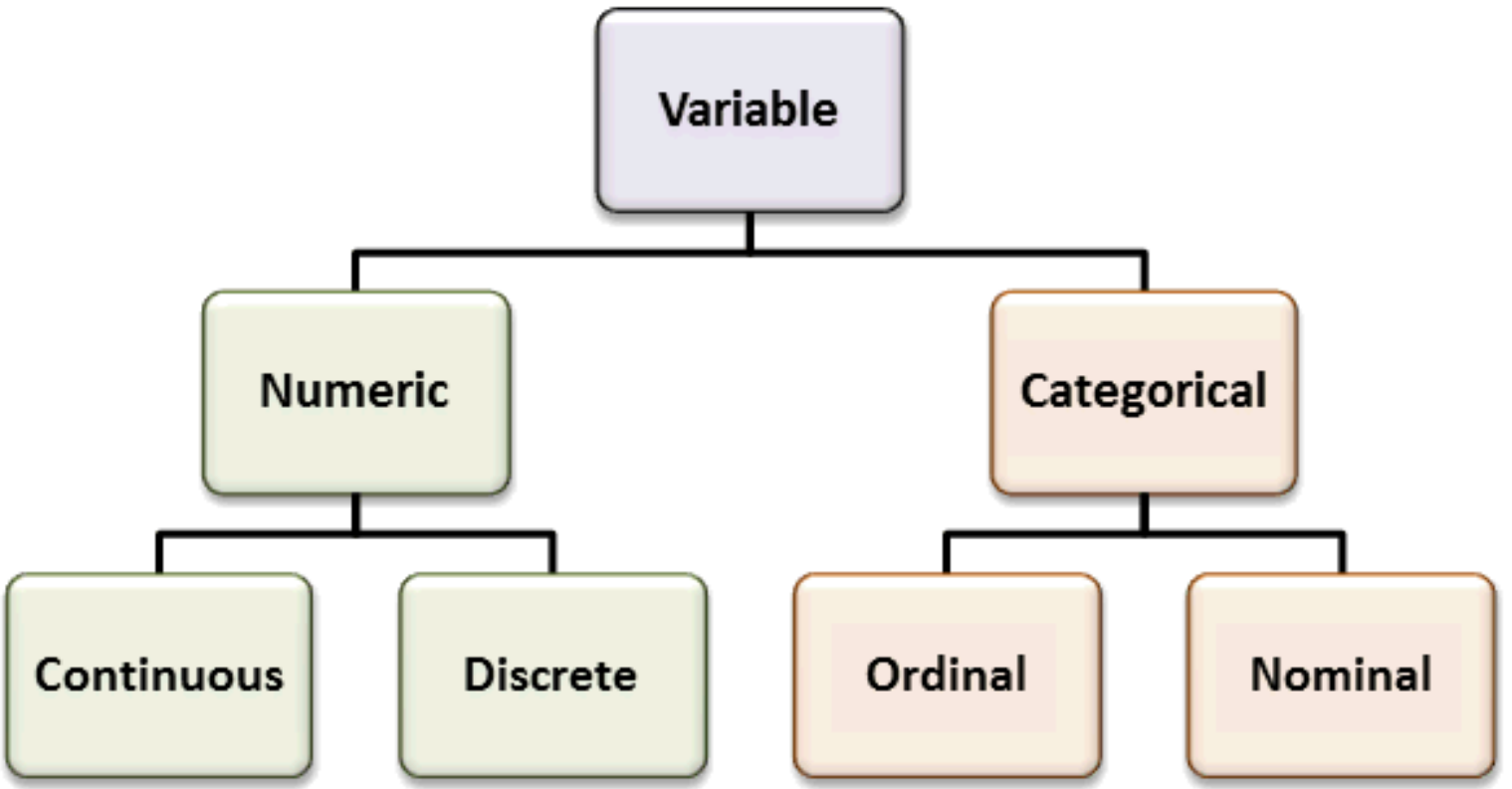
Introduction

Table 1
Baseline clinical and demographic characteristics of the study population.

Variable	Reference controls	INOCA subjects	*p* Value
Number	12	22	
Females	12 (100%)	22 (100%)	
Age (years)	49.2 ± 11	52.6 ± 13	0.436
BMI (kg/m^2)	25.5 ± 4.5	27.2 ± 5.3	0.344
BSA (m^2)	1.78 ± 0.20	1.82 ± 0.19	0.506
Risk Factors:			
Hypertension	0	11 (50%)	
Type II diabetes	0	3 (14%)	
Ever smoker	0	7 (32%)	
Dyslipidemia	0	10 (45%)	
Family history of CAD	0	16 (73%)	
Medications:			
Beta blockers	0	10 (45%)	
Calcium channel blockers	0	4 (18%)	
Nitrates	0	6 (27%)	
Aspirin	0	14 (64%)	
ACE inhibitors	0	3 (14%)	
Statins	0	12 (55%)	
Symptoms:			
Chest pain	0	11 (50%)	
Dyspnea	0	15 (68%)	

Values are mean ± standard deviation, or number (percentage).
BMI = body mass index; BSA = body surface area; CAD = coronary artery disease; INOCA = ischemia and no obstructive CAD; ACE = Angiotensin-converting enzyme.

Types of Data



Types of Data

Quantitative (Numeric)

Variable that has been measured on a numeric or quantitative scale

Continuous

- Can theoretically take on an infinite number of values - accuracy is limited only by the measuring instrument
  - e.g. age, BMI, BSA, height, weight, etc..

Discrete

- Numerical variables that are measured and can only be whole numbers
  - e.g. age, heart rate, number of medication taken, number of relapses, etc..

Types of Data

Qualitative (Categorical)

Variables that are typically not directly measured by an instrument, and are based on observations

Ordinal

- Variables that have an inherent hierarchical order to the relationship among the different categories
  - e.g. pain scores, stage of cancer, education level, etc..

Nominal

- Variables that are "named" or classified into one or more qualitative groups
- Do not have a sense of ordering between the different categories
  - e.g. risk factors, types of medications consumed, types of symptoms experienced, surgical outcomes, blood type, gender, etc..

Types of Data

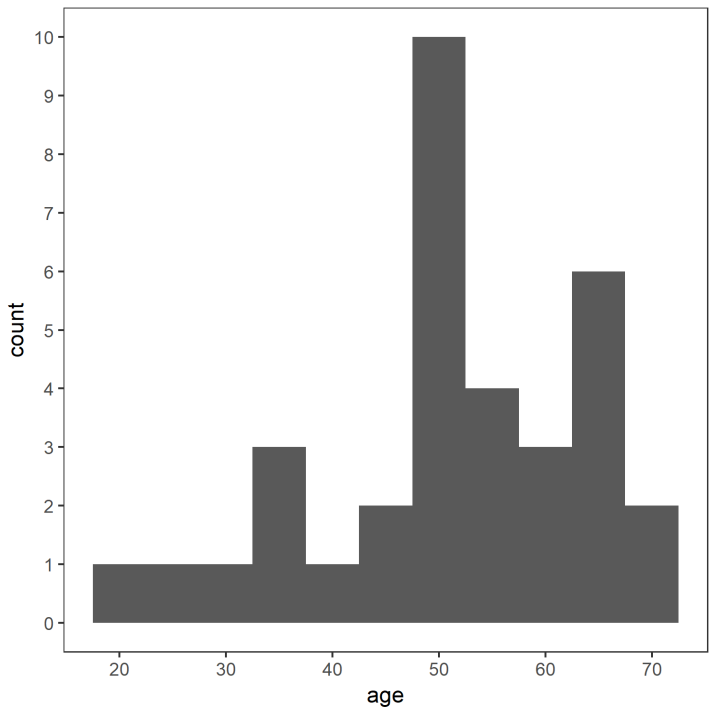
Why do we need to identify the types of data ?

In statistics we have specialized tools or measures to handle different type of data

Quantitative

Graphical Summarization

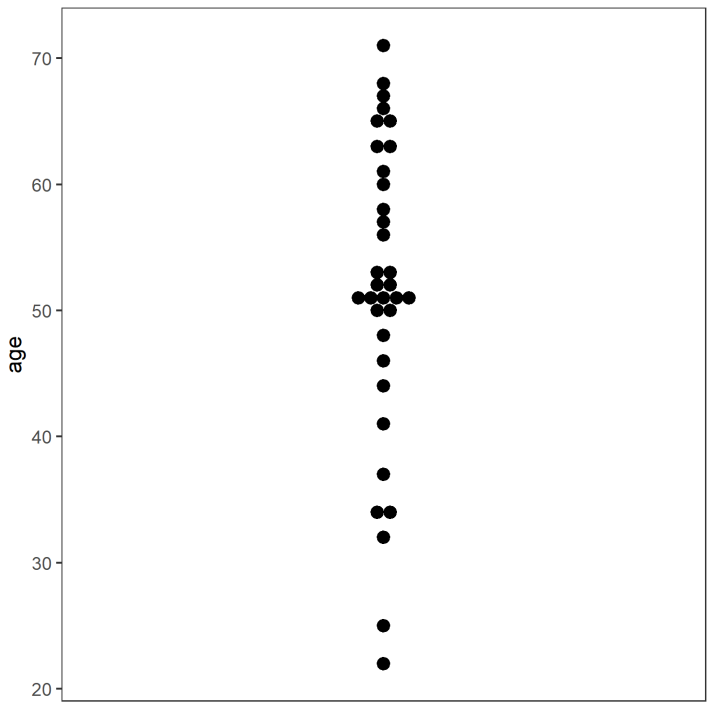
Histograms



- Histograms allows us to group similar values into 'bins' with specific widths
- Provides intuitive sense of both central tendency and dispersion
- The y axis describes the frequency (counts) of the occurrences of values in each bin

Graphical Summarization

Dot plots



- Dot plots allows us to have a closer look at each observation as represented by individual dots
- Also provides an intuitive sense of both central tendency and dispersion
- Only useful for relatively smaller sets of data

Numerical Summarization

Measures of Location

- Mean
- Median
- Mode

Measures of Variability or Dispersion

- Minimum and Maximum
- Percentiles / Interquartile Range (IQR)
- Standard Deviation

Measures of Location

Mean

- The sample mean is the most commonly used and readily understood measure of central tendency.
- The sample mean can be defined as:

$$\bar{x} = \frac{\sum x_i}{n}$$

Mean

Example

Reference

```
##  [1] 34 34 37 41 48 50 51 52 53 56 66 67
```

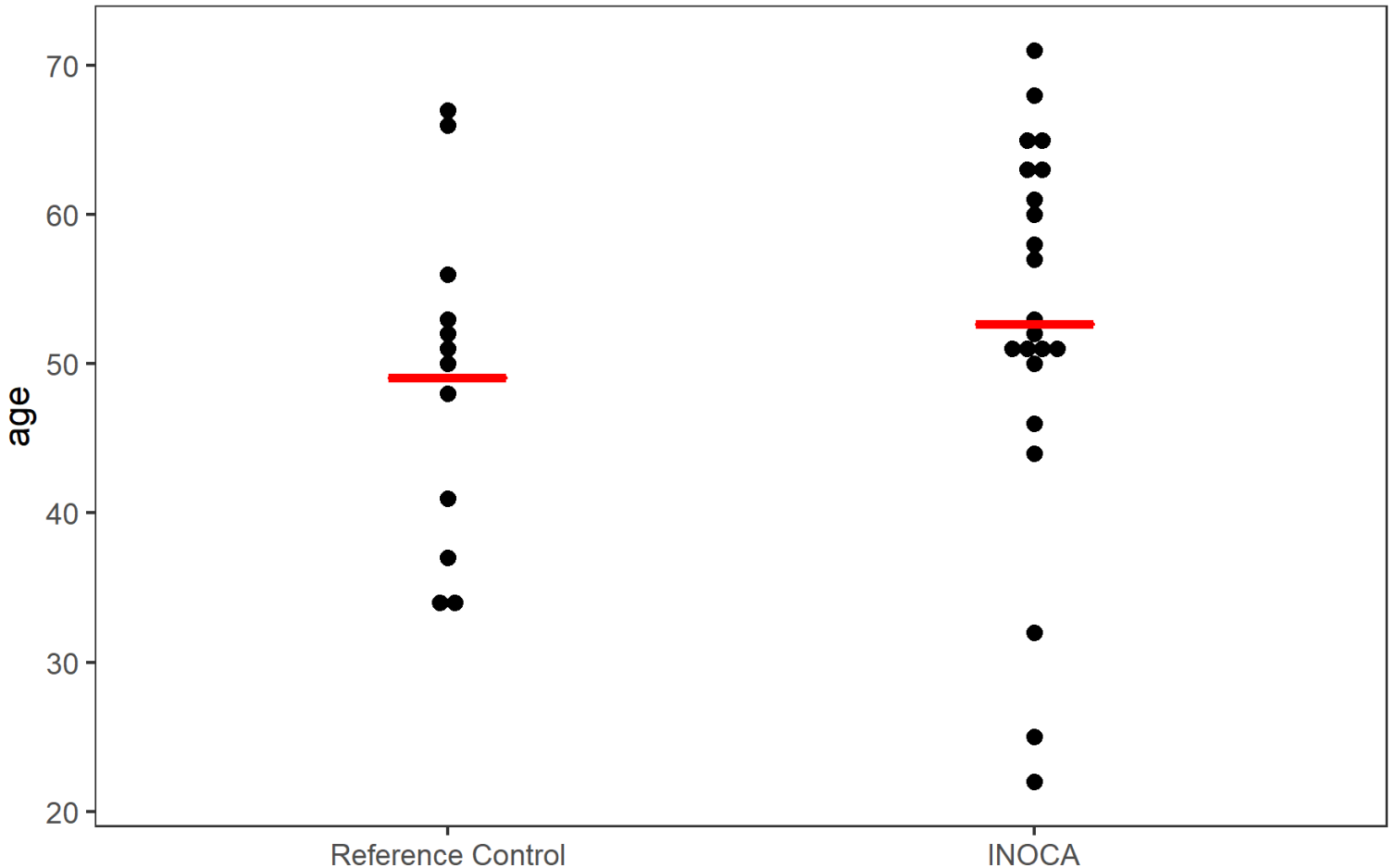
- The total sum of the reference age is 589
- The total number of measurements is 12
- The mean is 49.08

INOCA

```
##  [1] 22 25 32 44 46 50 51 51 51 51 52 53 57 58 60 61 63 63 65 65 68 71
```

- The total sum of the reference age is 1159
- The total number of measurements is 22
- The mean is 52.68

Mean



Measures of Location

Median

- The median is the midpoint of the values
  - We begin by ranking the data from smallest to largest
  - The midpoint value is the point at which half the observations are above the value and half the observations are below the value *(50 percentile*).
  - If there are two 'middle' values then the median is the average of the two mid values

Median

Example

Reference

```
##  [1] 34 34 37 41 48 50 51 52 53 56 66 67
```

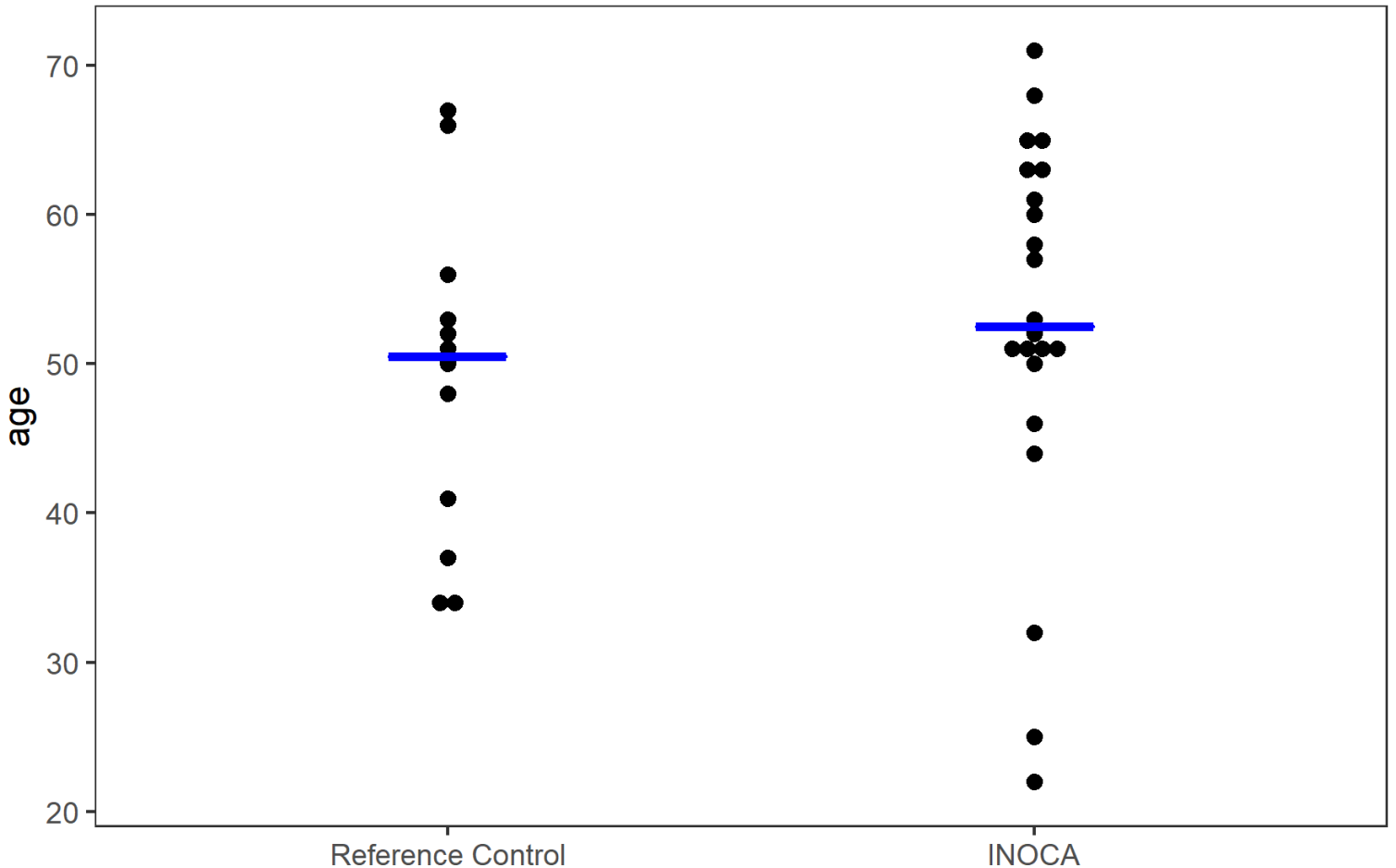
- The median is 50.5

INOCA

```
##  [1] 22 25 32 44 46 50 51 51 51 51 52 53 57 58 60 61 63 63 65 65 68 71
```

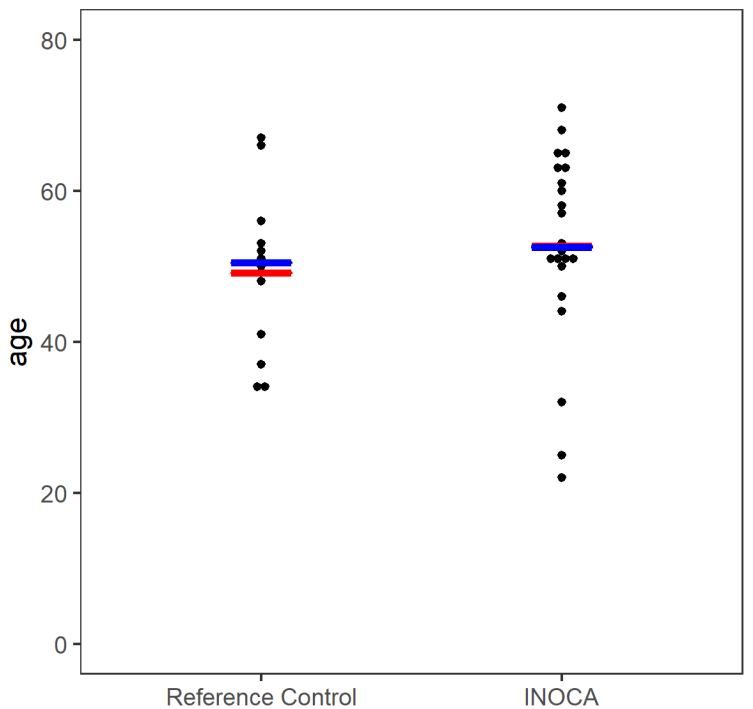
- The median is 52.5

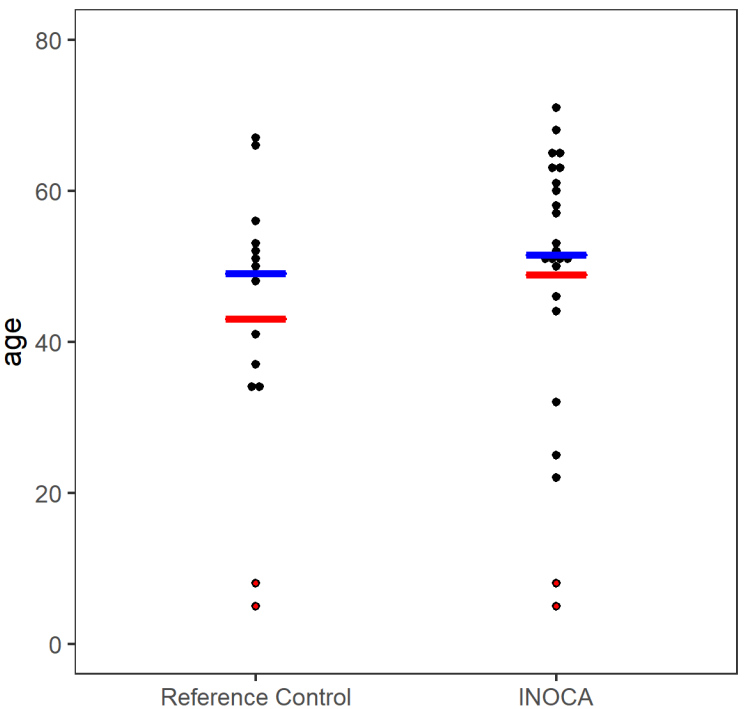
Median



Mean vs Median

- The mean and median both measure central tendency.
- However ...
  - The mean is more susceptible to outliers in the data. e.g. *unusual values*
  - The median is more robust to outliers





Measures of Location

Mode

- The mode is the value that appears most often in a set of values.
- Not always a measure of central tendency
- The mode is only useful for discrete values or continuous values with limited digits of accuracy

Mode

Age vs BMI

Age (Discrete)

```
##  [1] 22 25 32 34 34 37 41 44 46 48 50 50 51 51 51 51 51 52 52 53 53 56 57
## [24] 58 60 61 63 63 65 65 66 67 68 71
```

BMI (Continuous)

```
##  [1] 16.51343 17.45843 20.02502 20.46802 21.60509 21.72114 22.03183
##  [8] 22.71402 23.41656 23.54479 23.98646 24.41757 24.96775 25.19321
## [15] 25.31125 25.61932 25.69124 26.04075 26.87650 27.66826 27.79021
## [22] 27.96446 28.85647 29.43411 29.48031 29.88986 30.30300 30.47500
## [29] 30.71306 32.79919 33.34373 33.51282 36.18887 38.37826
```

Mode

Example

Reference

```
##  [1] 34 34 37 41 48 50 51 52 53 56 66 67
```

```
##
## 34 37 41 48 50 51 52 53 56 66 67
##  2  1  1  1  1  1  1  1  1  1  1
```

- The mode is 34

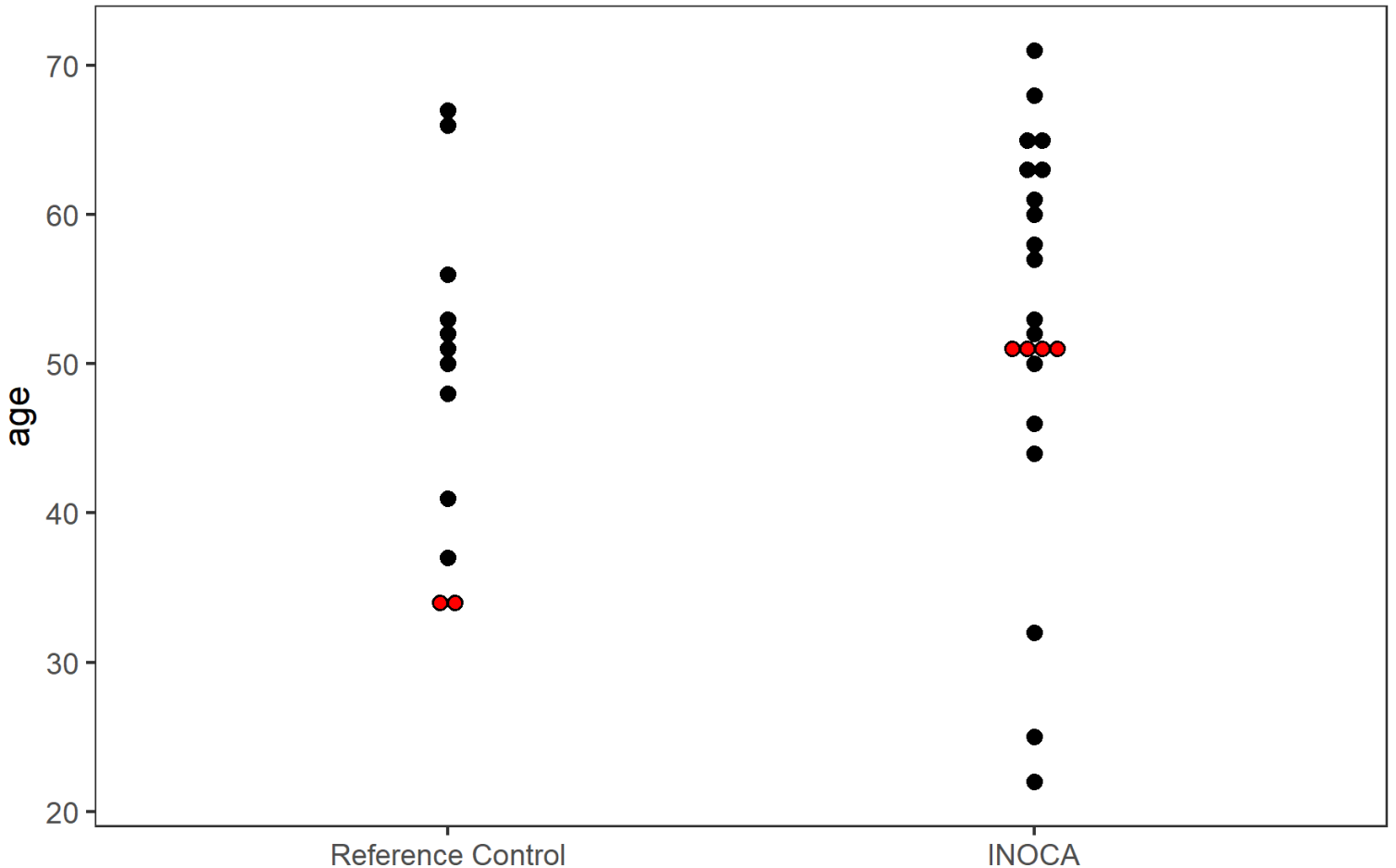
INOCA

```
##  [1] 22 25 32 44 46 50 51 51 51 51 52 53 57 58 60 61 63 63 65 65 68 71
```

```
##
## 22 25 32 44 46 50 51 52 53 57 58 60 61 63 65 68 71
##  1  1  1  1  1  1  4  1  1  1  1  1  1  2  2  1  1
```

- The mode is 51

Mode



Measures of Variability or Dispersion

Minimum and Maximum

A measure of dispersion and is defined as the smallest and largest value

Example

Reference

```
##  [1] 34 34 37 41 48 50 51 52 53 56 66 67
```

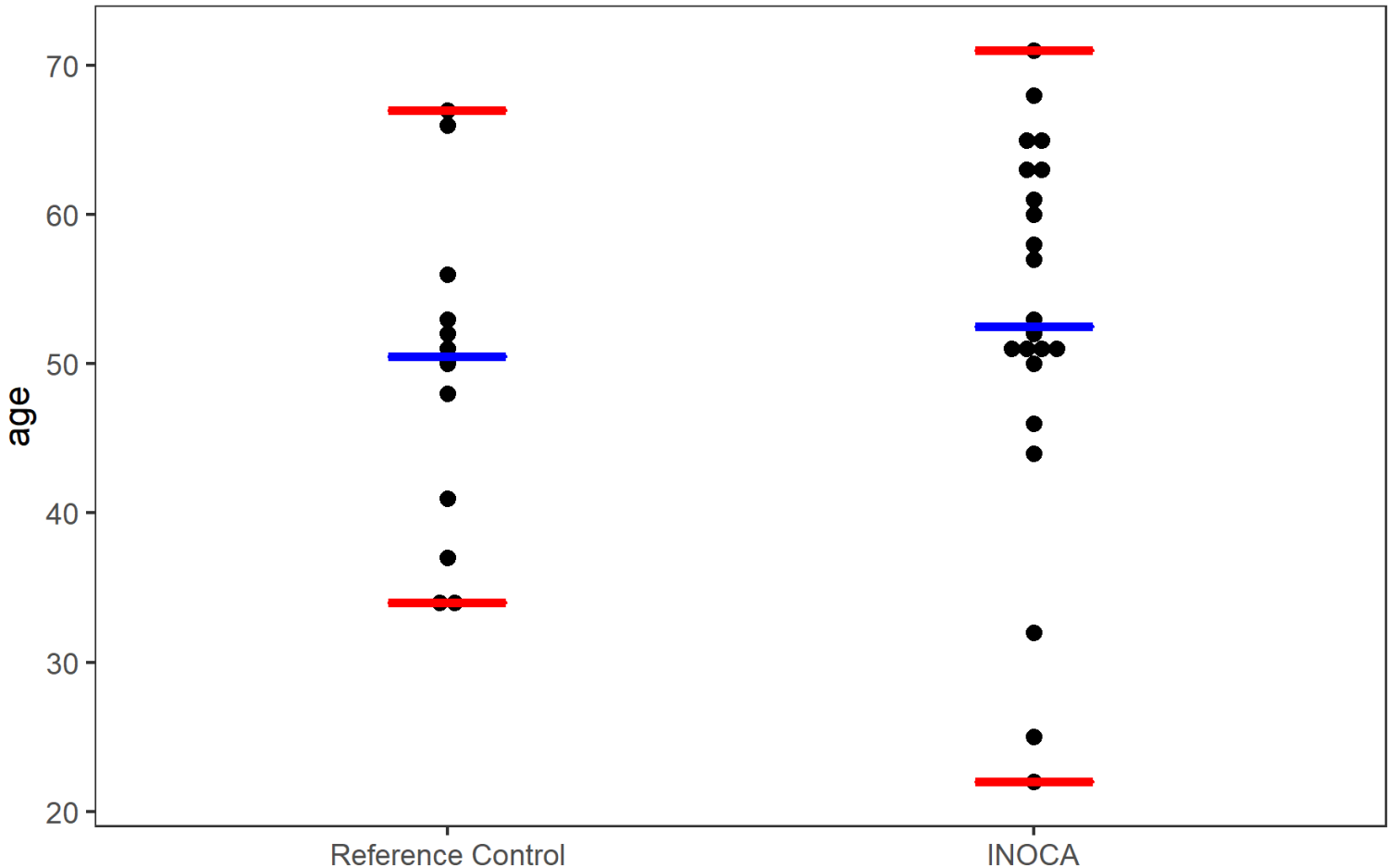
- The min is 34 and max is 67

INOCA

```
##  [1] 22 25 32 44 46 50 51 51 51 51 52 53 57 58 60 61 63 63 65 65 68 71
```

- The min is 22 and max is 71

Minimum and Maximum



Measures of Variability or Dispersion

Percentiles / Interquartile Range (IQR)

- The interquartile range is defined as the range between the 25th and 75th percentiles

$$IQR = Q_3 - Q_1$$

- It is commonly denoted after presenting the median
- The interquartile range can be used to describe the spread of the data. As the spread of the data increases, the IQR becomes larger.
- It is also used to build box plots.

Percentiles / Interquartile Range (IQR)

Example

Reference

```
##  [1] 34 34 37 41 48 50 51 52 53 56 66 67
```

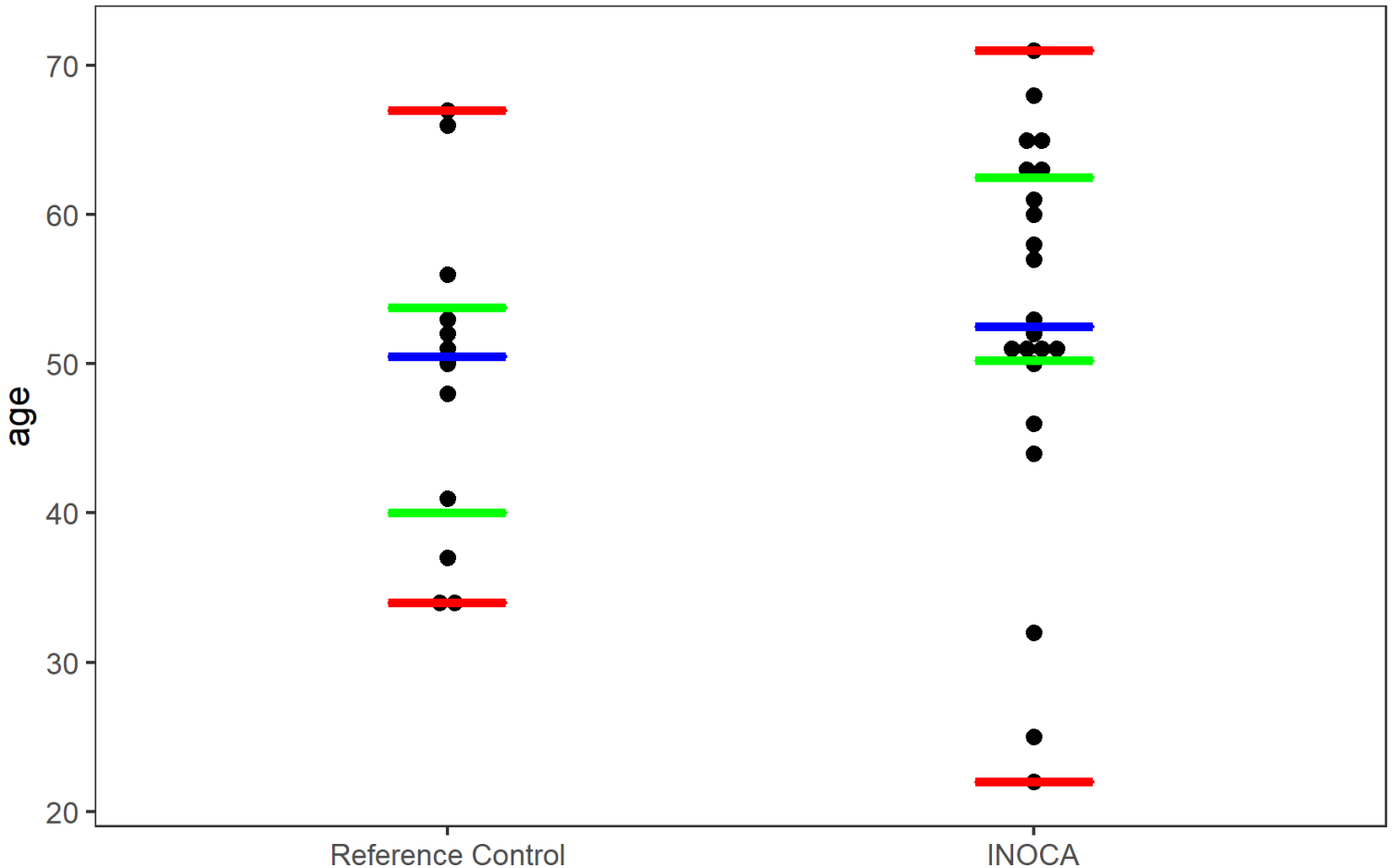
- The 25th percentile is 40 and 75th percentile is 53.75
- The IQR is 13.75

INOCA

```
##  [1] 22 25 32 44 46 50 51 51 51 51 52 53 57 58 60 61 63 63 65 65 68 71
```

- The 25th percentile is 50.25 and 75th percentile is 62.5
- The IQR is 12.25

Percentiles / Interquartile Range (IQR)



Measures of Variability or Dispersion

Standard Deviation

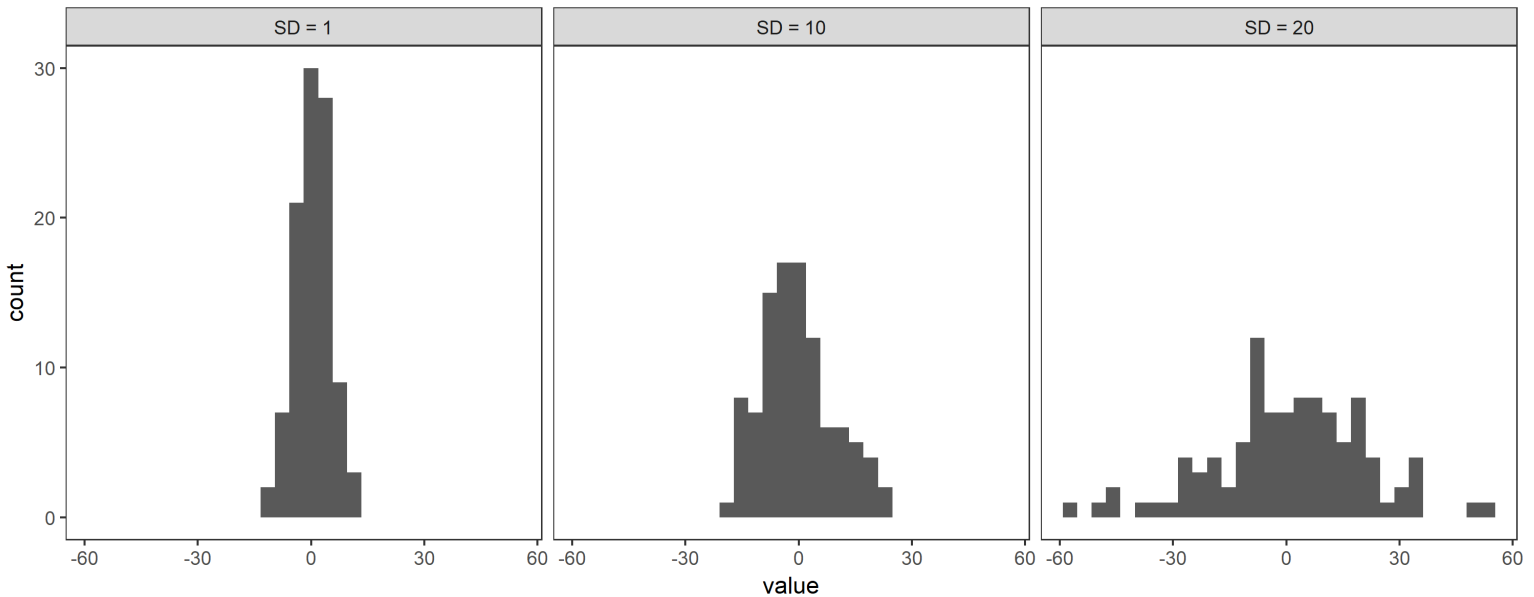
The standard deviation is a measure of how spread out the data are about the mean.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Measures of Variability or Dispersion

Standard Deviation

All of the figures below have the same exact mean (mean = 0) with varying standard deviations



Presenting / Describing your data

You should always include **BOTH** a *measure of central tendency* and *measure of dispersion* when presenting your data

The choice is dependent on the distribution of your data

Symmetric Distribution

- mean ± SD
- median (25%, 75% quantiles)
- median (IQR)
- median (min - max)

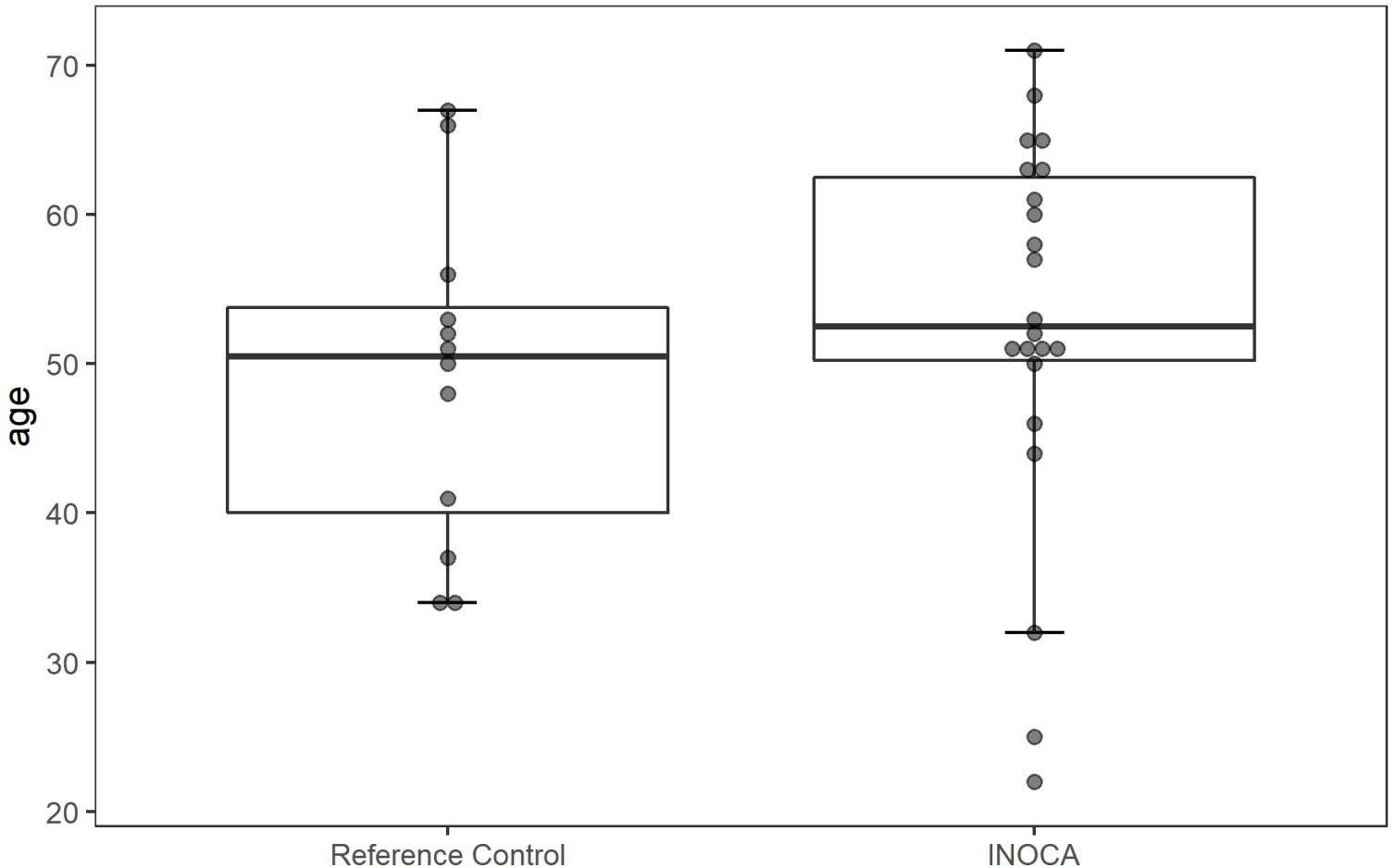
Asymmetric Distribution

- median (25%, 75% quantiles)
- median (IQR)
- median (min - max)

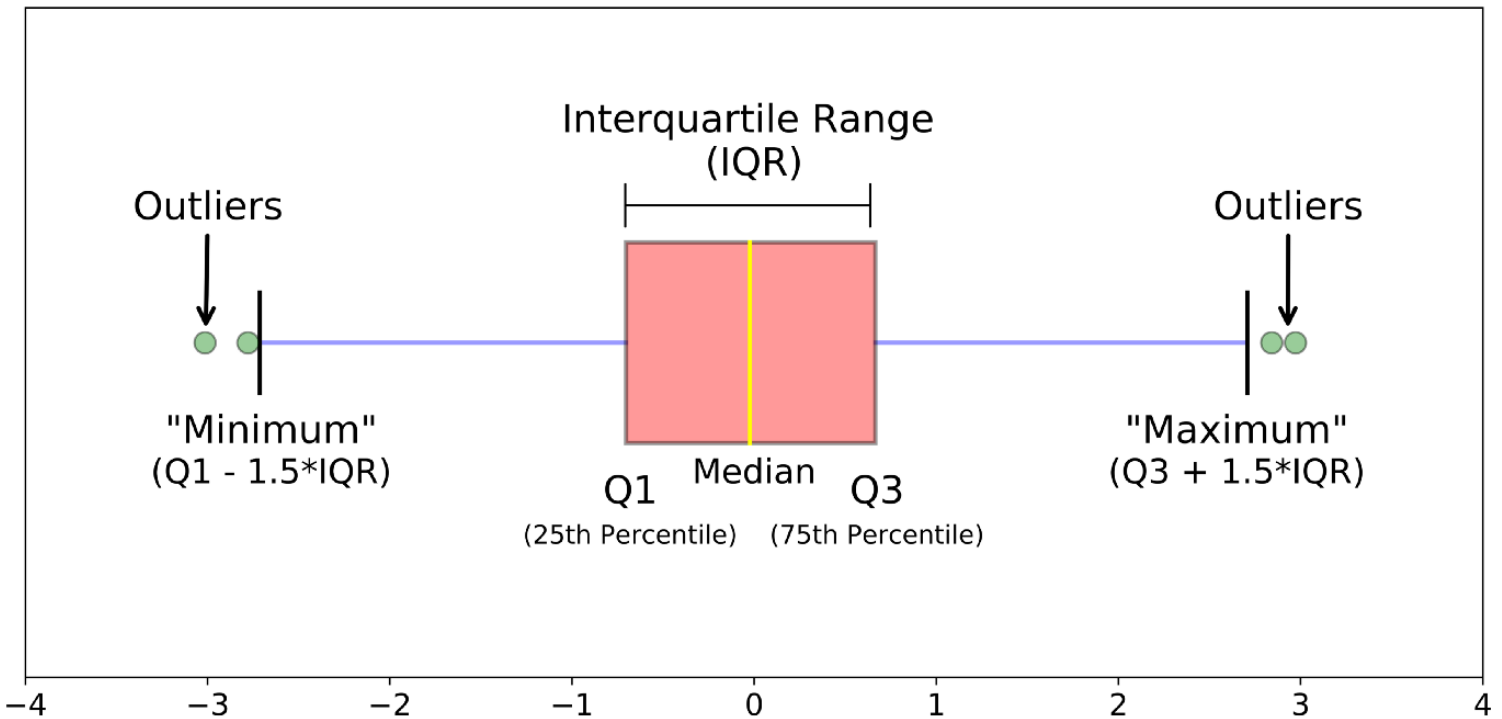
Graphical Summarization

- Histograms
- Dot plots
- Box plots

Box plots



Box plots


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Qualitative

Types of Data

Qualitative (Categorical)

Variables that are typically not directly measured by an instrument, and are based on observations

Ordinal

- Variables that have an inherent hierarchical order to the relationship among the different categories
  - e.g. pain scores, stage of cancer, education level, etc..

Nominal

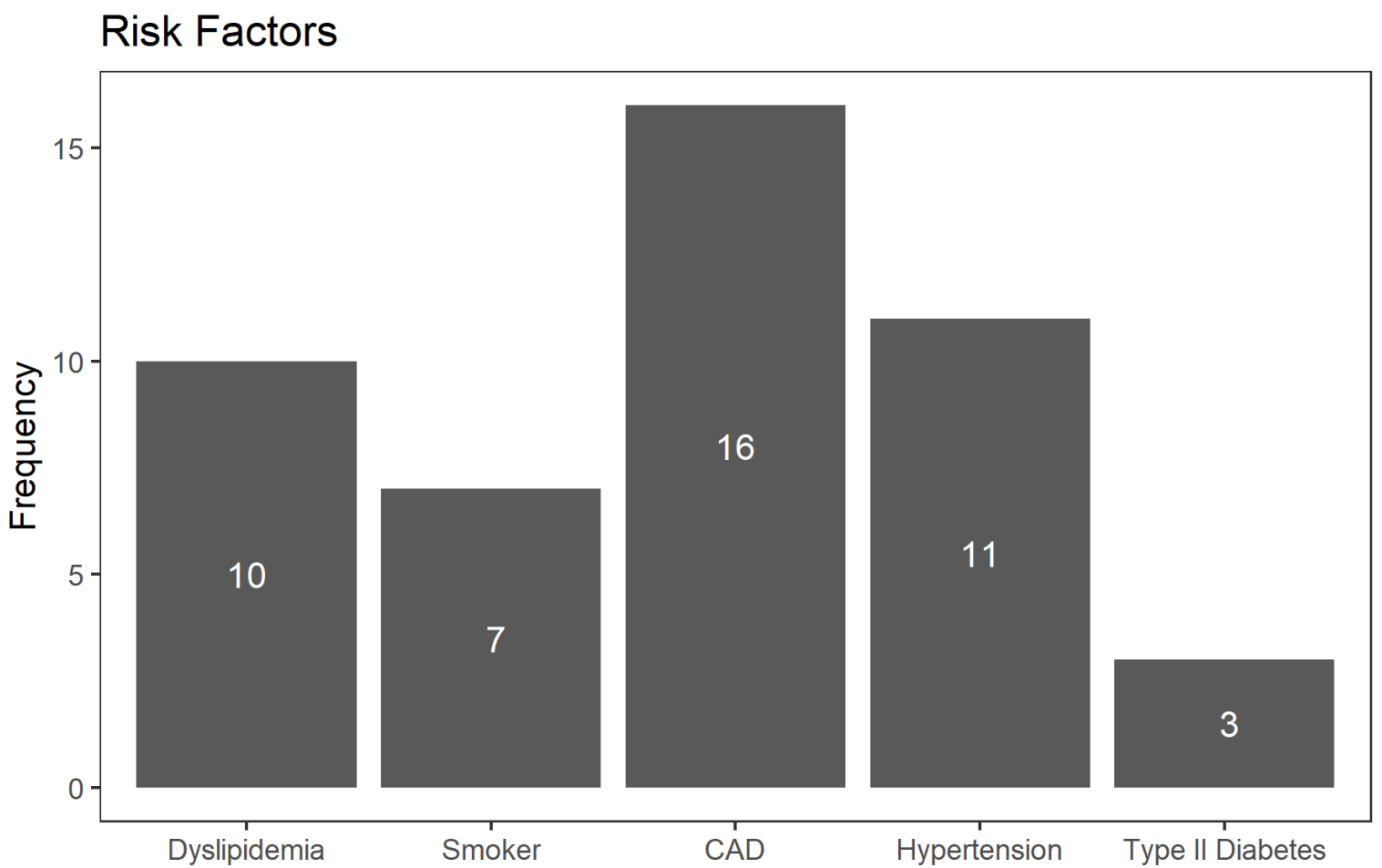
- Variables that are "named" or classified into one or more qualitative groups
- Do not have a sense of ordering between the different categories
  - e.g. risk factors, types of medications consumed, types of symptoms experienced, surgical outcomes, blood type, gender, etc..

Summarizing Qualitative Data

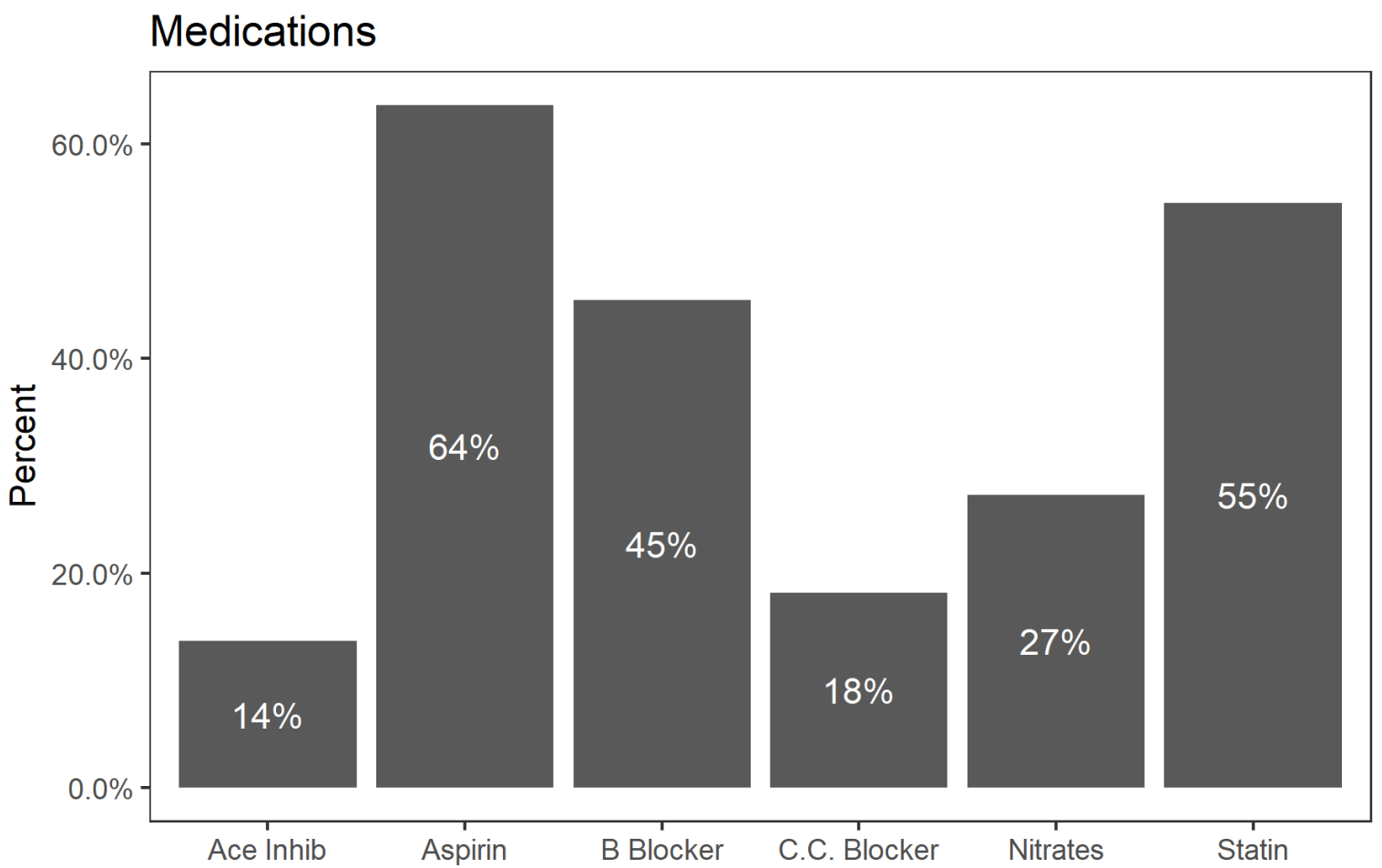
The primary method of summarizing qualitative data is frequency **counts** and **percentages**

- Graphical Methods
  - Bar Plots

Bar plots

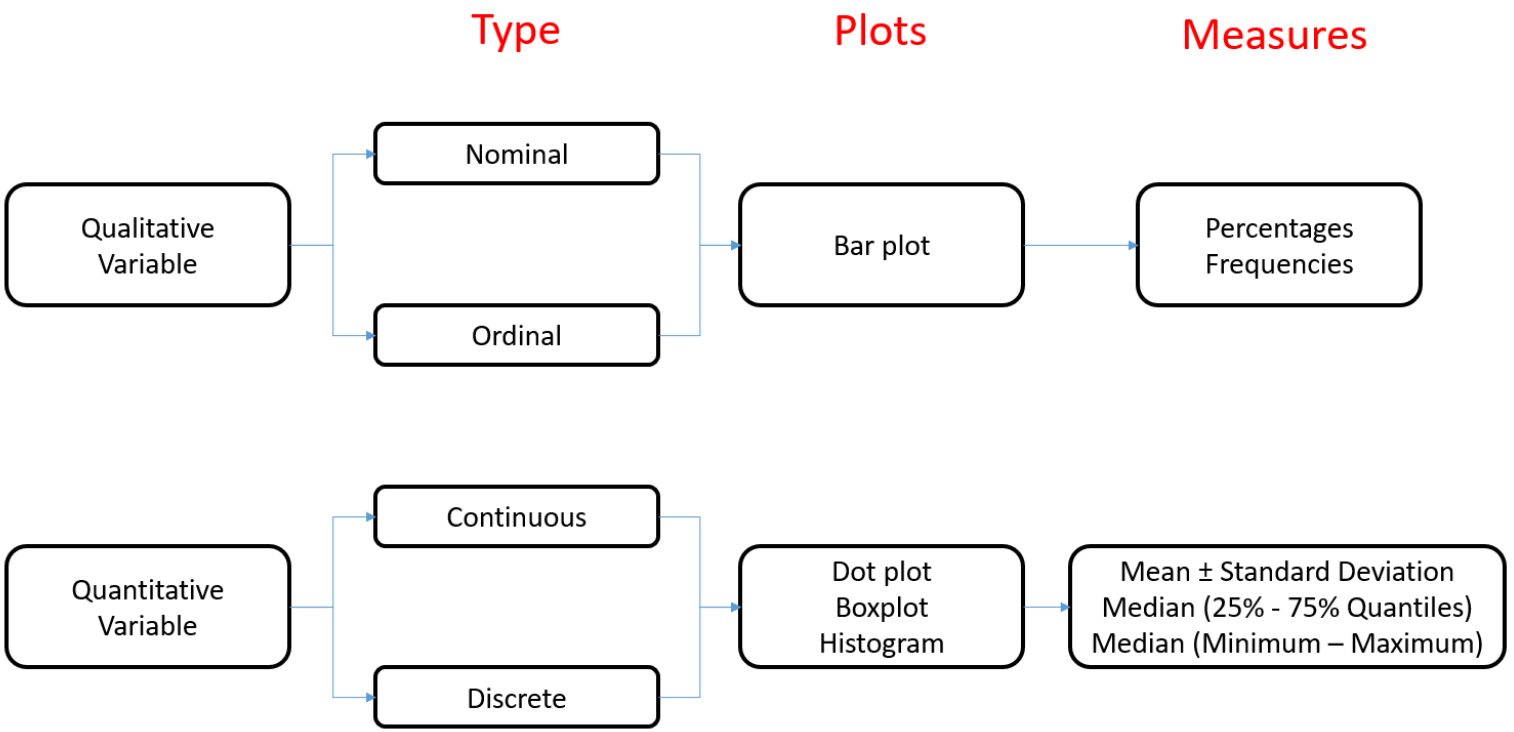

Risk Factors
Frequency
15
10
5
0
10
7
16
11
3
Dyslipidemia
Smoker
CAD
Hypertension
Type II Diabetes

Barplots


Medications
Percent
60.0%
40.0%
20.0%
0.0%
14%
64%
45%
18%
27%
55%
Ace Inhib
Aspirin
B Blocker
C.C. Blocker
Nitrates
Statin

Summary

Summary


Type
Plots
Measures
Qualitative Variable
Nominal
Ordinal
Bar plot
Percentages
Frequencies
Quantitative Variable
Continuous
Discrete
Dot plot
Boxplot
Histogram
Mean ± Standard Deviation
Median (25% - 75% Quantiles)
Median (Minimum – Maximum)