Welcome and introduction to statistics

(Week 01 lecture notes)

Wei Q. Deng

July 4th 2018

A little about me

- Wei Q. Deng
- Current PhD in Statistics, MSc in Statistical Genetics and BSc in Mathematics and Statistics
- Research interests: principal component analysis, classification and clustering, machining learning heuristics, statistical computing and graphics/data visualization, statistical genetics.

Overview of Today's Lecture

- First half: course syllabus and statistics
  - statistics as a data science
  - population and sample
  - parameters and statistics
- Second half: summarizing data (graphically and through descriptive statistics)
  - descriptive statistics (idea of a centre/location, spread/scale of a distribution)
  - graphical representations (bar, histogram, boxplot, scatterplot, qqplot)
  - information from the graphs (outliers, model assumptions, dependence)
  - summarizing two or more variables simultaneously
  - data transformation

Notes about syllabus

- **Course syllabus** is available on blackboard and `https://weiakanedeng.github.io/STA248H1_2018Summer/`.
- **Class schedule**
  - Monday 18:00-21:00 in **SS 2118** (no lecture on Monday, August 6 (Civic holiday))
  - Wednesday 18:00-21:00 in **SS 2118**.
  - Your TA Tianle Chen will be covering two lectures the week of July 30th.
- **Office Hours**
  - hours: Monday 4-6pm, Wednesday 4-6pm (starts from 2nd week)
  - location: **103A** Steward Building (149 College Street
  - Other times by appointment only.
- **Textbook(s)**
  - No official textbook for the course.
  - Reference (recommended)
    - *Open Intro Statistics*, 3rd edition `https://www.openintro.org/stat/textbook.php`
    - Chapter 2-3 can be used as review for STA247
    - topics in Chapter 1,4,5,7 and some of 8 are covered in this course
    - *Probability and Statistics for Engineering and the Sciences* **9**th Edition by Jay L. Devore

Notes about syllabus

- All course material (syllabus, lecture slides, practice problems and solutions) will be posted on portal and `https://weiakanedeng.github.io/STA248H1_2018Summer/`.
- Portal contains a **Discussion Board** that you can use to communicate with others, but it will not be monitored.
- For all inquiries, please come to office hours or email me at `wei.deng@mail.utoronto.ca`.
- For general annoucements and test information, I will be sending the entire class email through portal. Please make sure that your mail.utoronto.ca account is configured correctly and check it frequently

Course Objectives

- This course covers some theory for the most commonly used statistical methods.
- The main goal is to develop problem solving skills with emphasis on thinking with data, constructing models for data analysis and assessing quality of analyses.
- Learnt to do data analysis using R or Python via DataCamp modules.

Statistics

- most statistical methods: **estimation** and **prediction**
- **theoretical** statisticians focus on the methodology and theory that methods for estimation or prediction have desirable properties (such as unbiasedness and consistency)
- **applied** statisticians focus on the actual execution of these methods on real data

Applied Statistics

The process of scientific investigations can be summarized in four stages [1]:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Topics related to Stage 2 are covered in Survey Sampling and Design of Experiments.

Here (and most of the other applied statistics courses) focus on stages 3-4, the process of learning from data.

[1] OpenIntro page 7

Applied Statistics

The collected data might be from one of the following methods:

- survey
- observation
- experiment
- secondary data

Applied Statistics vs. data science

The data science is more rigourous in its treatment of

- data preparation
- data presentation
- data analysis

while traditional statistics deal with data analysis via statistical modelling alone.

Modern data (such as those collected from social media, netflix, or genomics, image data etc.) can be very heterogenous, complex and even misleading without proper data preparation.

Statistics: to think with data.

- In an ideal world, we have the entire **population** of data.
- In practice, we only have a subset of it, i.e. a **sample** of the data.

An important concept is random sampling, in which samples are taken from population at random.

- This ensures the collection of samples are **representative** of the population, and statistics calculated from samples can enjoy certain properties.
- If we believe the sample collected truly arise from the population, then any information we extract from the sample, should give us some idea of the information we can obtain from the entire population.

Statistics, but what is a statistic?

A **statistic** $f(\mathbf{x}_n)$ is a function of the sample data ($\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$).

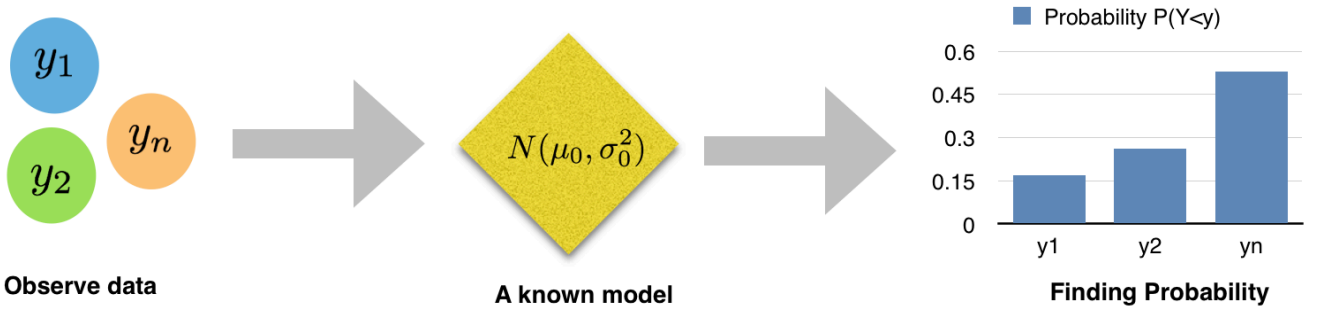
As $n \to \infty$, some statistic could approach the population value.

On the other hand, a statistical **parameter** (broadedly referred to as θ) is constant associated with the population ($\mathbf{x} = (x_1, x_2, \ldots, x_n, \ldots)$).

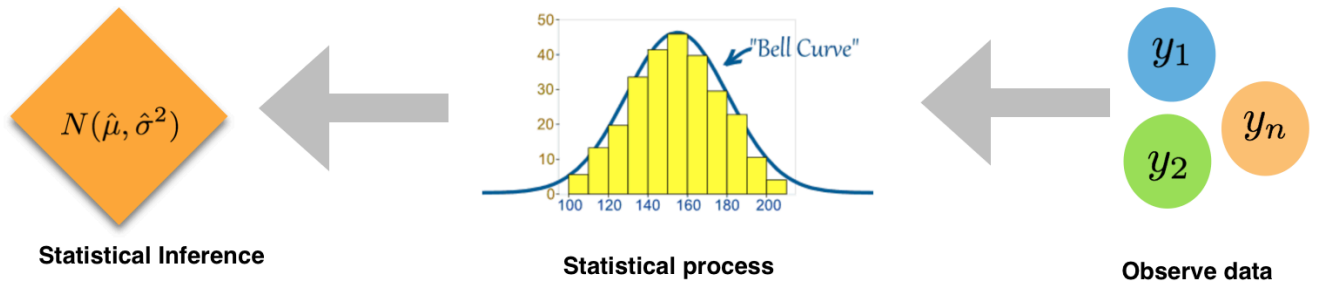
Statisticians often use **statistics** ($f(\mathbf{x}_n)$) to estimate **parameters** (θ), which could also be referred to as **estimators** ($\hat{\theta}$).

Connection to other courses

- Introduction to probability (eg. STA247/257:learn several distributions, know how to find mean, variance, etc.)



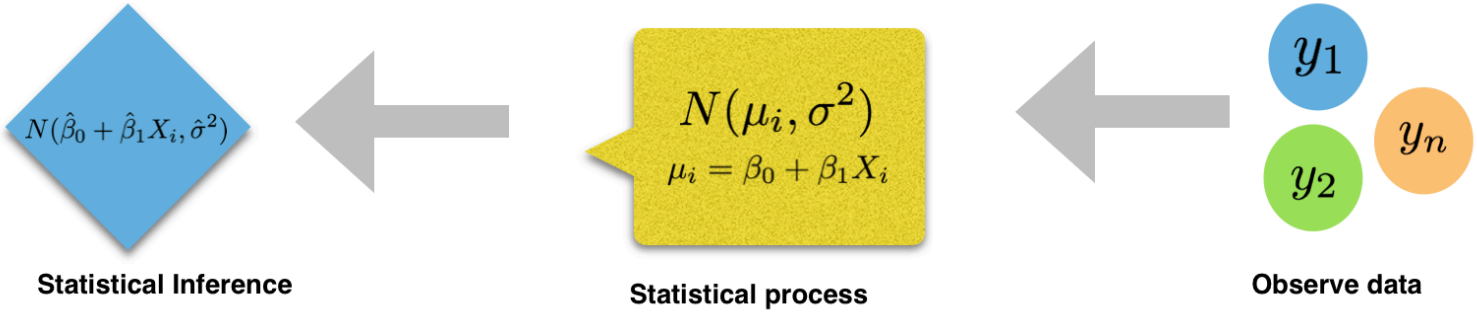
- Introduction to statistical inference (eg, STA248/STA261: know how to estimate model parameter θ, CI,hypothesis testing, etc)



Connection to other course

Our course ends on linear regression, which will be explored in more details in STA302 and more general forms of regression in STA303 (logistic and Poisson).

- STA302: methods of data analysis I (the major topic is on linear regression)



Let's look at some data!

In R:

```
library(MASS)
data(Cars93)
print(Cars93[1:5,1:5])
```

```
##   Manufacturer   Model    Type Min.Price Price
## 1        Acura Integra   Small      12.9  15.9
## 2        Acura  Legend Midsize      29.2  33.9
## 3         Audi      90 Compact      25.9  29.1
## 4         Audi     100 Midsize      30.8  37.7
## 5          BMW    535i Midsize      23.7  30.0
```

```
#head(Cars93)
```

In Python:

```
import pandas
Cars93 = pandas.read_csv("Cars93.csv")
head = Cars93.head()
print(head)
```

```
##   Manufacturer    Model     Type  Min.Price  Price  Max.Price  MPG.city  \
## 0        Acura  Integra    Small       12.9   15.9       18.8        25
## 1        Acura   Legend  Midsize       29.2   33.9       38.7        18
## 2         Audi       90  Compact       25.9   29.1       32.3        20
## 3         Audi      100  Midsize       30.8   37.7       44.6        19
## 4          BMW     535i  Midsize       23.7   30.0       36.2        22
##
##    MPG.highway             AirBags DriveTrain      ...       Passengers  \
## 0           31                None      Front      ...                5
## 1           25  Driver & Passenger      Front      ...                5
## 2           26         Driver only      Front      ...                5
## 3           26  Driver & Passenger      Front      ...                6
## 4           30         Driver only       Rear      ...                4
##
##    Length  Wheelbase  Width  Turn.circle Rear.seat.room  Luggage.room  Weight  \
## 0     177        102     68           37           26.5            11    2705
## 1     195        115     71           38           30.0            15    3560
## 2     180        102     67           37           28.0            14    3375
## 3     193        106     70           37           31.0            17    3405
## 4     186        109     69           39           27.0            13    3640
##
##     Origin           Make
## 0  non-USA  Acura Integra
## 1  non-USA   Acura Legend
## 2  non-USA        Audi 90
## 3  non-USA       Audi 100
## 4  non-USA       BMW 535i
##
## [5 rows x 27 columns]
```

What types of variables are present in this dataset?

Generally, two main types of variables: categorical and continuous.

- Categorical variables
  - Nominal variable (no order), e.g., Manufacturer or Model
  - Ordinal variable (ordered), e.g., Cylinders
- Continuous variables
  - Interval variable (a unit difference has the same meaning), e.g. price
  - Ratio variable (different meanings depending on the value), e.g. temperature in Kelvin.

Why different types of variables are used/needed?

You can also think of these as different levels of resolutions to the measurements.

For example, grades of students in STA130,

which could be collected as a **nominal variable** with two levels:

- middle range grades
- extreme high or low grades

You can also be more specific and define an **ordinal variable** "grade" with three levels:

- excellent
- adequate
- poor

A **continuous** grade could be the actual marks received out of 100, and thus is an interval variable.

Why different types of variables are used/needed? (cont'd)

More on interval and ratio variables:

- for interval variables, the zero is relative, for ratio, there is an absolute zero
- the same information can be represented either as an interval or ratio variable (year of birth or age)
- the distinction is more important in specific applications (e.g. clinical epidemiology)

What types of variables do statistical programs use?

In R:

- factor (categorical with levels)
  - nominal
  - ordinal
- numeric
  - integer (discrete as supposed to continuous)
  - continuous

In python:

- string
- integer (discrete as supposed to continuous)
- float (with digits)

```
class(1)
```

```
## [1] "numeric"
```

```
print(type(3.4))
```

```
## <type 'float'>
```

Back to the data

Categorical variables:

```
names(Cars93)[sapply(Cars93, is.factor)]
```

```
## [1] "Manufacturer"    "Model"           "Type"            "AirBags"
## [5] "DriveTrain"      "Cylinders"       "Man.trans.avail" "Origin"
## [9] "Make"
```

Continuous variables:

```
names(Cars93)[sapply(Cars93, is.numeric)]
```

```
##  [1] "Min.Price"          "Price"              "Max.Price"
##  [4] "MPG.city"           "MPG.highway"        "EngineSize"
##  [7] "Horsepower"         "RPM"                "Rev.per.mile"
## [10] "Fuel.tank.capacity" "Passengers"         "Length"
## [13] "Wheelbase"          "Width"              "Turn.circle"
## [16] "Rear.seat.room"     "Luggage.room"       "Weight"
```

A quick summary

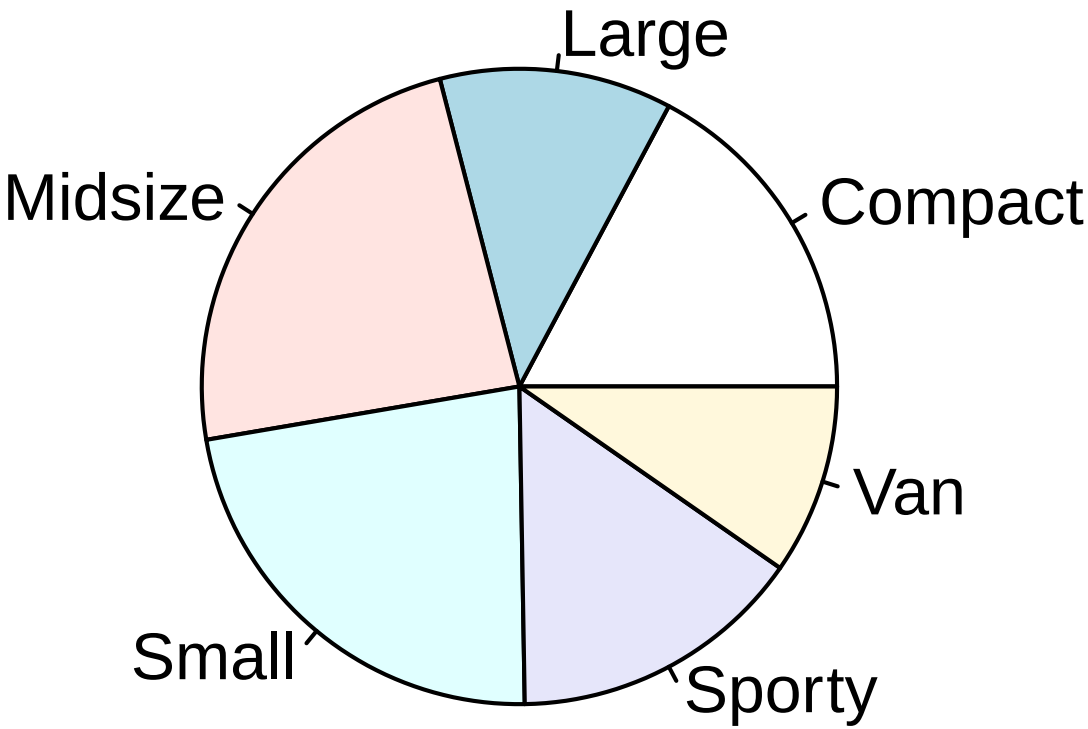
- Different types of variables require different graphical representations
- Different types of variables contain different levels of details
- Different types of variables require different statistical techniques to analyze (next a few lectures; STA302 and STA303)

How to graphically represent each type of variable?

categorical variables

a pie chart

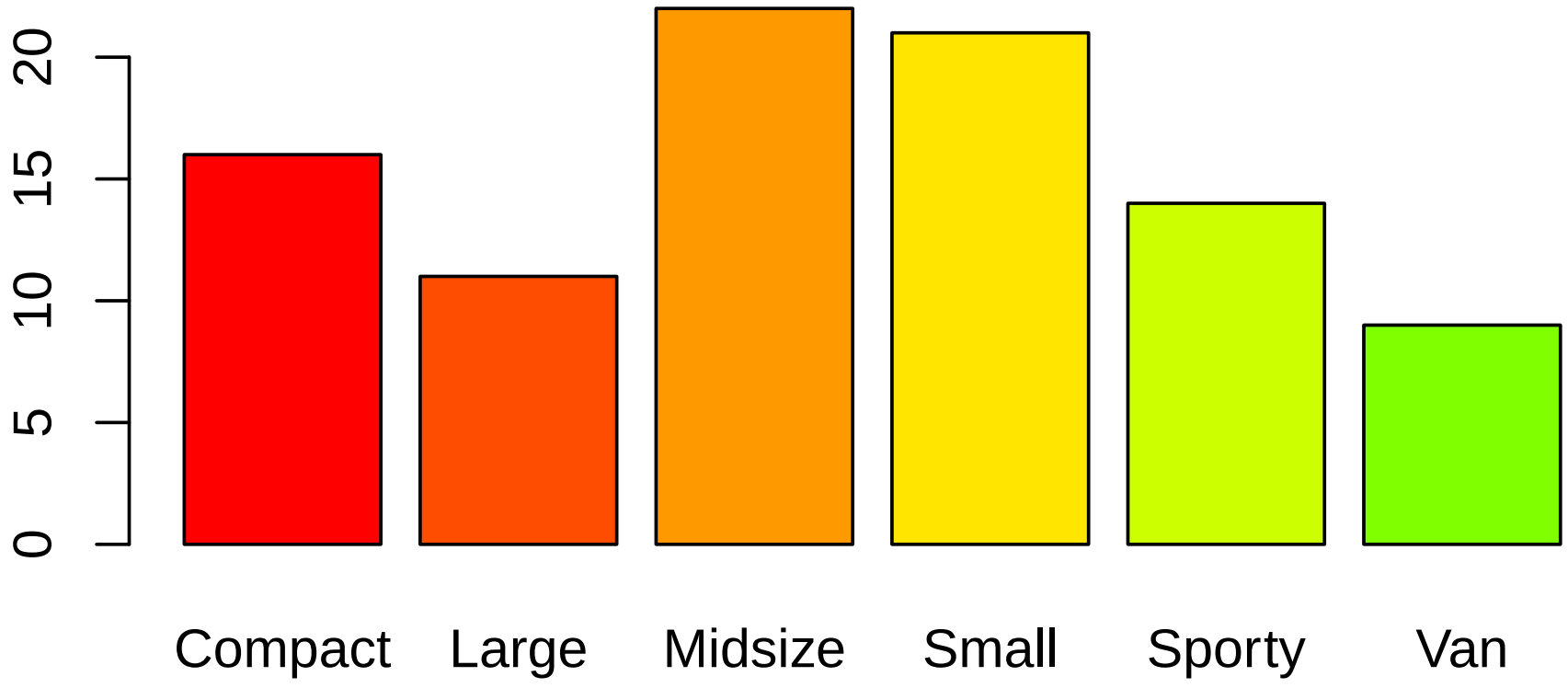
```
pie.price <- table(Cars93$Type)/length(Cars93$Type)
names(pie.price) <- names(table(Cars93$Type))
pie(pie.price)
```



categorical variables (cont'd)

or a bar plot

```
barplot(table(Cars93$Type), col = rainbow(20))
```



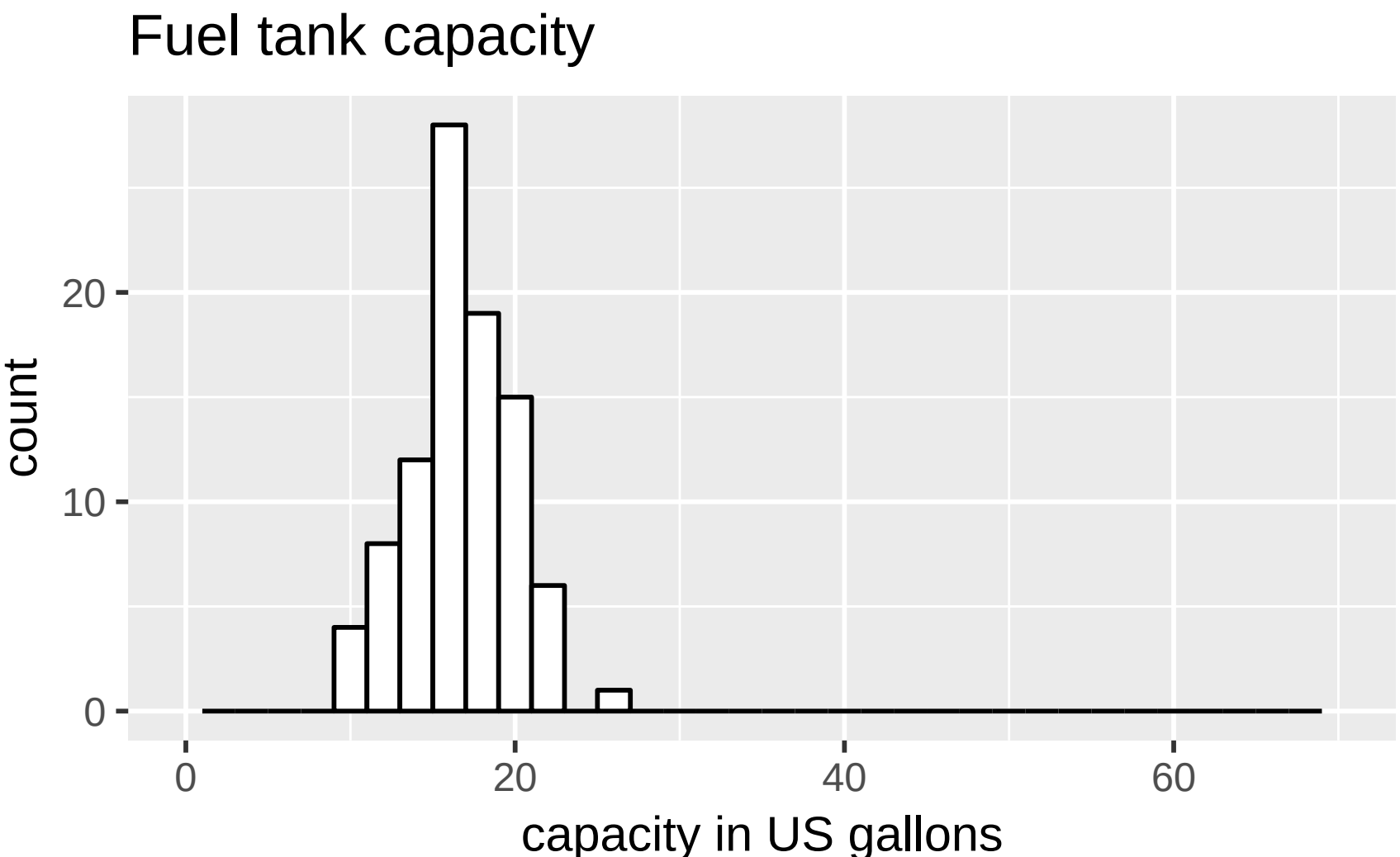
categorical variables (cont'd)

- pie chart: difficult to interpret when many categories are present
  - when a few categories
  - one or two dominate categories
  - should include percentages
- bar plot: a direct comparison and more frequently used in science
  - bar height corresponds to percentages or counts

continuous variables

a histogram

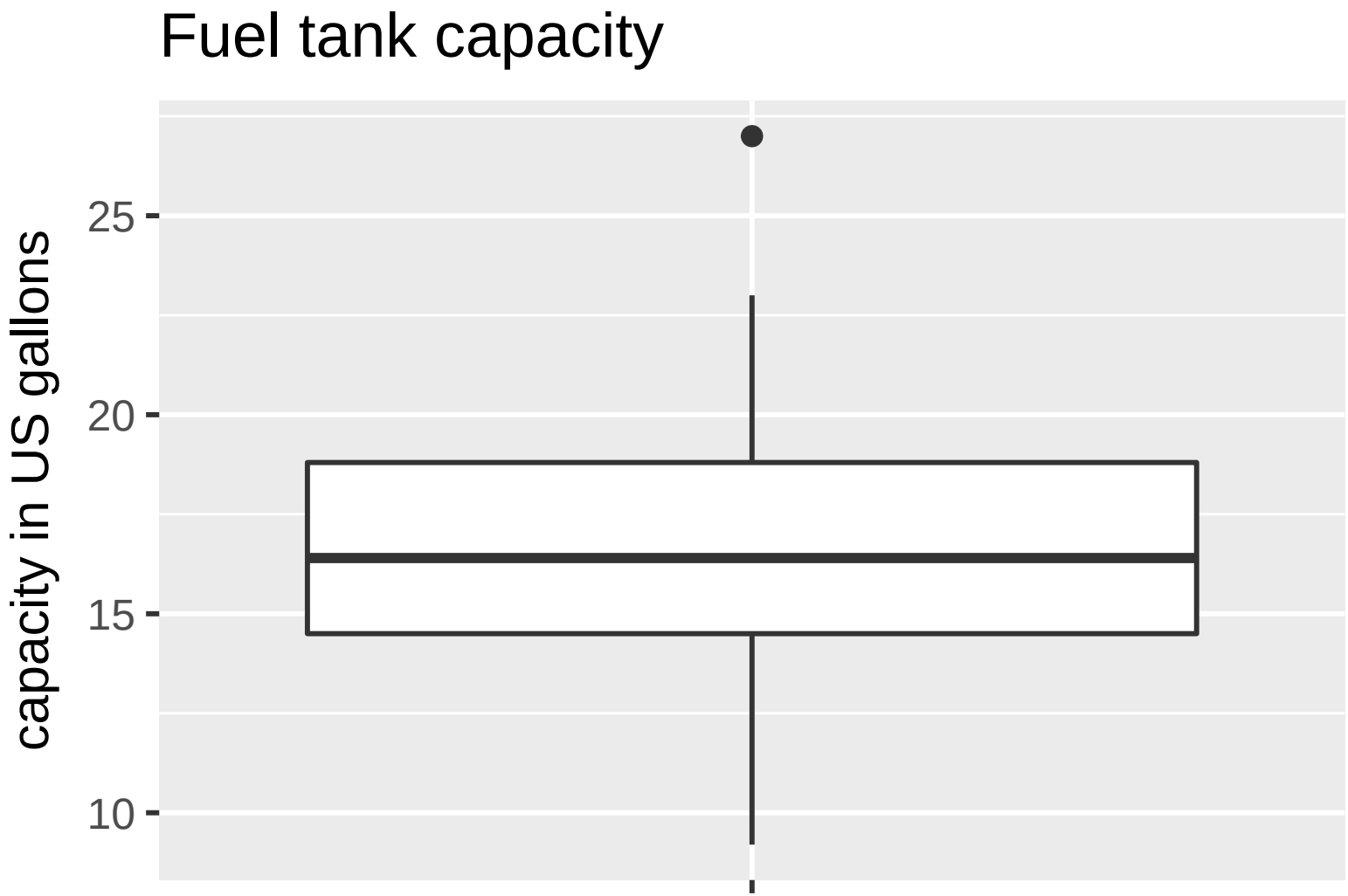
```
library(ggplot2)
ggplot(data=Cars93, aes(x=Fuel.tank.capacity)) +
  geom_histogram(binwidth=2, colour="black", fill="white") +
  xlab("capacity in US gallons")  + ggtitle("Fuel tank capacity") +
  xlim(0,70)
```



continuous variables (cont'd)

or a boxplot:

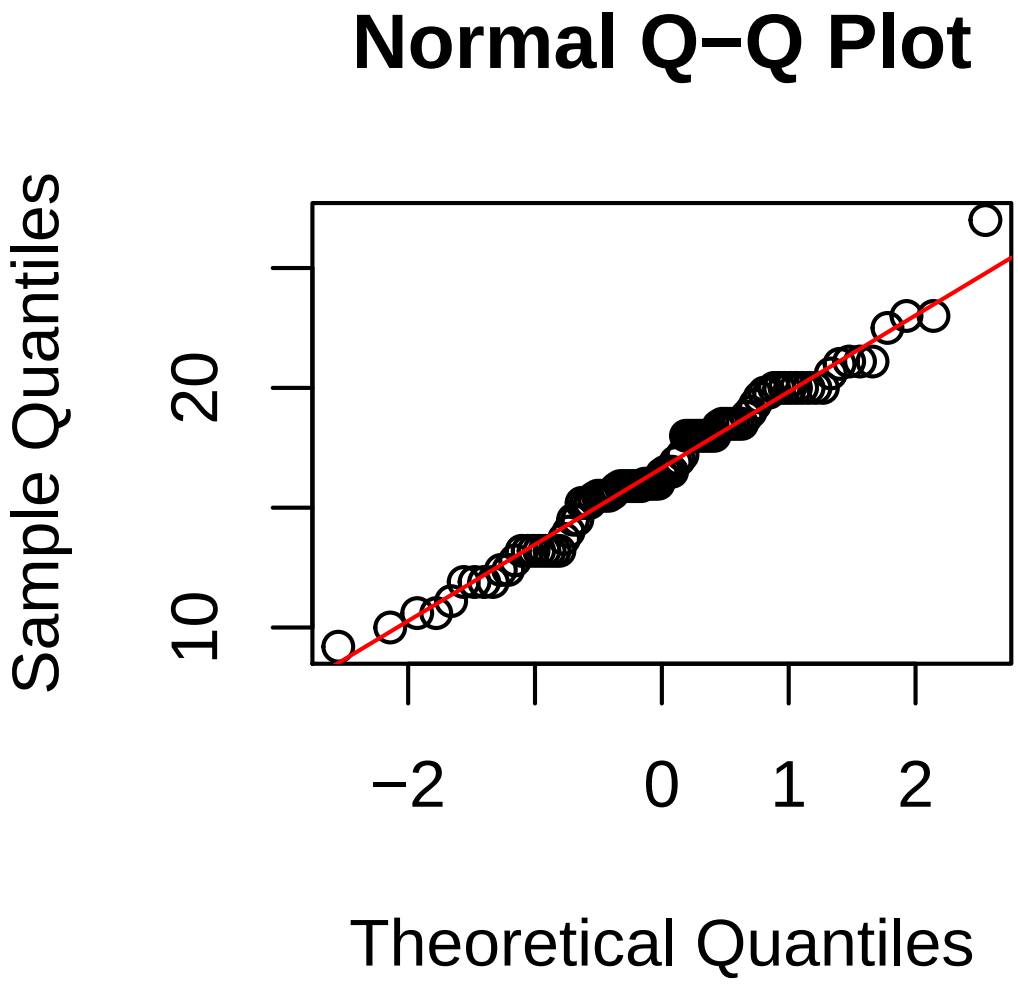
```
ggplot(data = Cars93, aes(x = "", y = Fuel.tank.capacity)) +
  geom_boxplot() + xlab("") + ylab("capacity in US gallons") +
  ggtitle(paste("Fuel tank capacity"))
```



continuous variables (cont'd)

suspecting this might be following a normal distribution

```
qqnorm(Cars93$Fuel.tank.capacity)
qqline(Cars93$Fuel.tank.capacity, col=2)
```



continuous variables (cont'd)

all three types are commonly used:

- histogram: a direct visual on the overall distribution (sample density)
- boxplot: useful for spotting anomalies
- qqplot: useful for a direct comparison with another distribution

How to summarize information in each type of variable numerically?

Think about what kind of information we can/want to extract from a variable.

For example, car type:

```
print(Cars93$Type[1:10])
```

```
##  [1] Small   Midsize Compact Midsize Midsize Midsize Large   Large
##  [9] Midsize Large
## Levels: Compact Large Midsize Small Sporty Van
```

```
print(Cars93$Cylinders[1:10])
```

```
##  [1] 4 6 6 6 4 4 6 6 6 8
## Levels: 3 4 5 6 8 rotary
```

```
print(Cars93$Fuel.tank.capacity[1:10])
```

```
##  [1] 13.2 18.0 16.9 21.1 21.1 16.4 18.0 23.0 18.8 18.0
```

Descriptive statistics

- descriptive statistics are by definition, statistics, calculated from the samples.
- they provide a concise summary on the sample data distribution

Why look at a statistic instead of the entire collection of data?

- to learn data structure by an initial data examination
- to help spot anomaly in data collection/recording (data quality control)
- sometimes the summary statistic is the quantity we are interested in (e.g. sample mean)
- to help make a decision on the next steps (statistical models)

Descriptive statistics for categorical variables

- central tendency
  - mode: category with the highest occurence/frequency
- overall distribution
  - frequencies
  - counts
- anomlies
  - missing value
  - outliers

Back to the data

```
table(Cars93$Cylinders)
```

```
##
##      3      4      5      6      8 rotary
##      3     49      2     31      7      1
```

```
table(Cars93$Cylinders)/sum(table(Cars93$Cylinders))
```

```
##
##          3          4          5          6          8     rotary
## 0.03225806 0.52688172 0.02150538 0.33333333 0.07526882 0.01075269
```

```
table(Cars93$Type)
```

```
##
## Compact   Large Midsize   Small  Sporty     Van
##      16      11      22      21      14       9
```

```
table(Cars93$Type)/sum(table(Cars93$Type))
```

```
##
##    Compact      Large    Midsize      Small     Sporty        Van
## 0.17204301 0.11827957 0.23655914 0.22580645 0.15053763 0.09677419
```

Descriptive statistics for continuous variables

- central tendency
  - mean
  - median
- spread/dispersion
  - standard deviation/variance
  - interquartile range
  - range
  - absolute mean/median deviation
- overall distribution/shape
  - skewness
  - kurtosis
  - modality
- anomlies
  - missing value
  - outliers

measure of central tendency in numerical variables

- mean: $\bar{x}$ the arithmetic mean of all observations
- median: $\tilde{x}$ the most central value [2]

[2] when there is an even number of observations, you take the average of the middle two

computing central tendency measures

In R:

```
print(c(mean(Cars93$Luggage.room), median(Cars93$Luggage.room)))
```

```
## [1] NA NA
```

```
print(c(mean(Cars93$Luggage.room, na.rm=T), median(Cars93$Luggage.room, na.rm=T)))
```

```
## [1] 13.89024 14.00000
```

In Python

```
import pandas
Cars93 = pandas.read_csv("Cars93.csv")
import numpy
print("median", numpy.median(Cars93[["Luggage.room"]]))
```

```
## ('median', 14.0)
```

```
print("mean", numpy.mean(Cars93[["Luggage.room"]]))
```

```
## ('mean', Luggage.room    13.890244
## dtype: float64)
```

measure of symmetry in shape: skewness

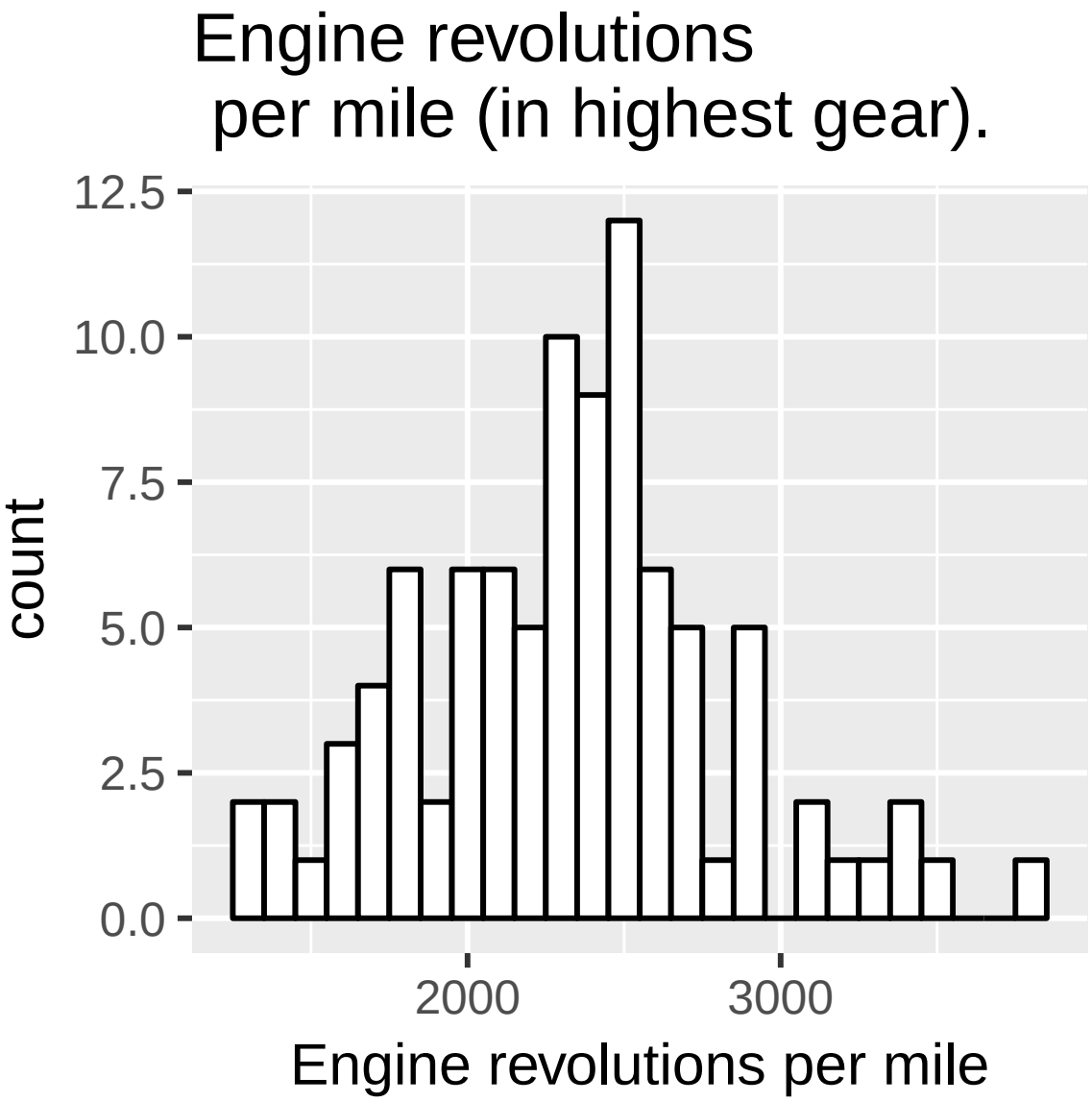
- If the distribution is symmetric, then mean and median tend to be really close.
- On the other hand, if the mean is less or greater than the median, it could indicate a **skewed** distribution, e.g. income of household (mean $<$ median)
- Median and mean both capture the centre, but median is a more **robust** measure as it is not sensitive to **outliers** or extreme values.

A more formal definition of skewness:

- **skew to the left**: median is greater than the mean
- **skew to the right**: median is less than the mean

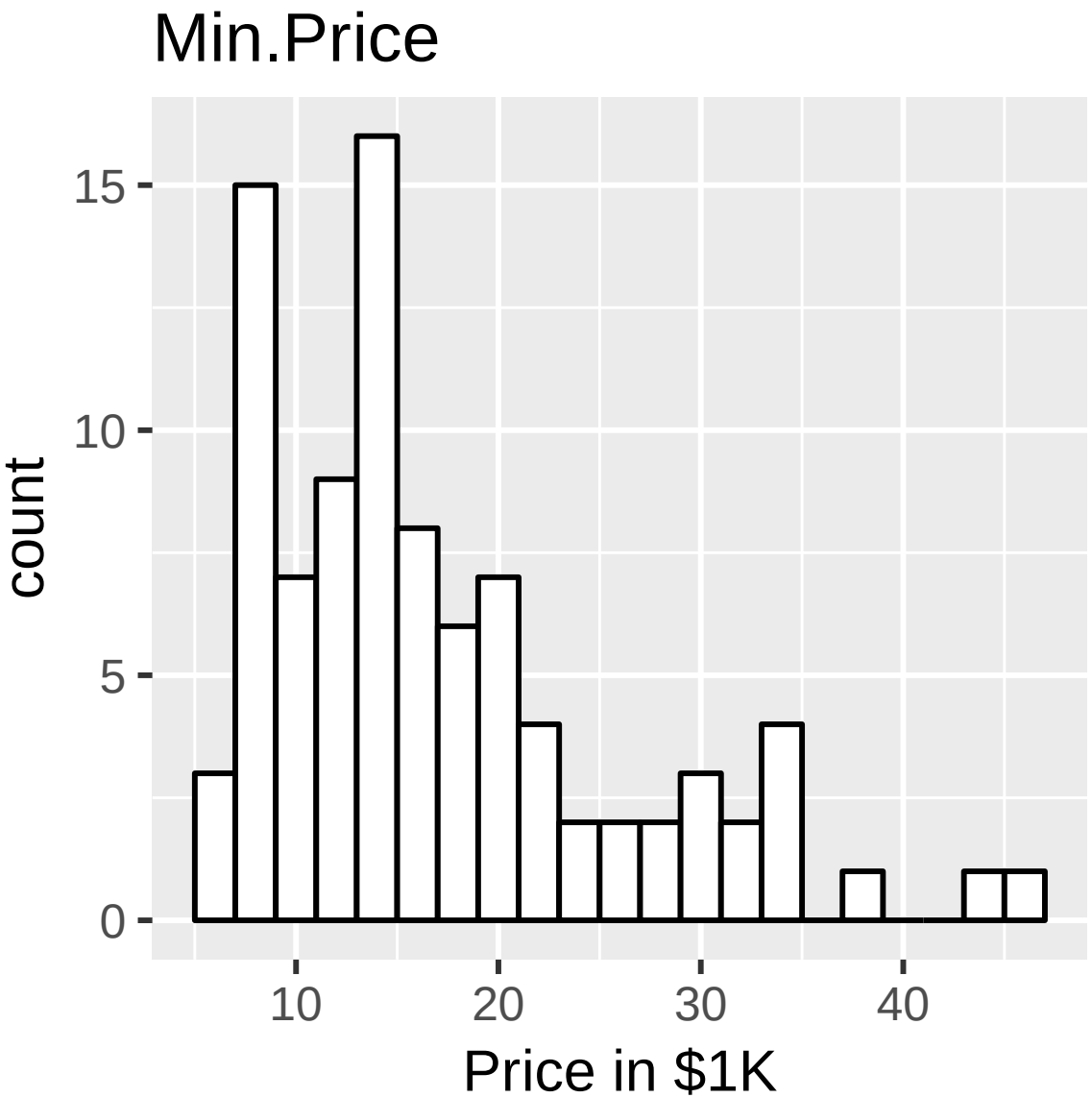
skew to the left

```
library(ggplot2)
ggplot(data=Cars93, aes(x=Rev.per.mile)) +
  geom_histogram(binwidth=100, colour="black", fill="white") +
  xlab("Engine revolutions per mile")  +
  ggtitle("Engine revolutions \n per mile (in highest gear).")
```



skew to the right

```
library(ggplot2)
ggplot(data=Cars93, aes(x=Min.Price)) +
  geom_histogram(binwidth=2, colour="black", fill="white") +
  xlab("Price in $1K")  + ggtitle("Min.Price")
```



measure of dispersion (or spread)

Several measures are typically used:

- Interquartile range (IQR): a measure of variability, based on dividing a data set into quartiles
- Standard deviation/Variance:

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

- Absolute mean/median deviation:

$$\sum_{i=1}^{n}|x_i - \bar{x}| \text{ or } \sum_{i=1}^{n}|x_i - \tilde{x}|$$

- Range: minimum to maximum
- Coefficient of variation:

$$\frac{\sigma}{\mu}$$

Quantiles and other percentiles

- The 25th percentile is known as the first quartile ($Q1$)
- The 10th percentile is known as the first decile ($D1$)

There are *three* quartiles ($Q1, Q2, Q3$) correspond to 25th, 50th, and 75th percentile of the data, dividing the data to **four** equal parts.

Interquartile range (IQR):

$$Q3 - Q1$$

There are *nine* deciles ($D1, \ldots, D9$) correspond to 10th, ..., 90th percentile of the data, dividing the data to **ten** equal parts.

Typically used when we care more about the tails of the data.

Connection between IQR and SD for a standard normal random variable

Let x be a standard normally distributed random variable, i.e. $\mathrm{x} \sim \mathcal{N}(0, 1)$.

We want to find out what the first quartile is by solving the following:

$$Pr(\mathrm{x} < a) = Pr(\frac{\mathrm{x} - \mu}{\sigma} < \frac{a - \mu}{\sigma}) = 0.25$$

alternatively

$$Pr(\mathrm{x} > a) = 1 - 0.25 = 0.75$$

or since we know $a < 0$ because it is to the left of the mean (which is zero), we can also find

$$Pr(\mathrm{x} < -a) = 1 - 0.25 = 0.75$$

You can find a or $-a$ by

- using a standard normal table (know how to do this)
- use R
- use Python

using a standard normal table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

using R or Python

```
qnorm(0.75)
```

```
## [1] 0.6744898
```

```
qnorm(0.25)
```

```
## [1] -0.6744898
```

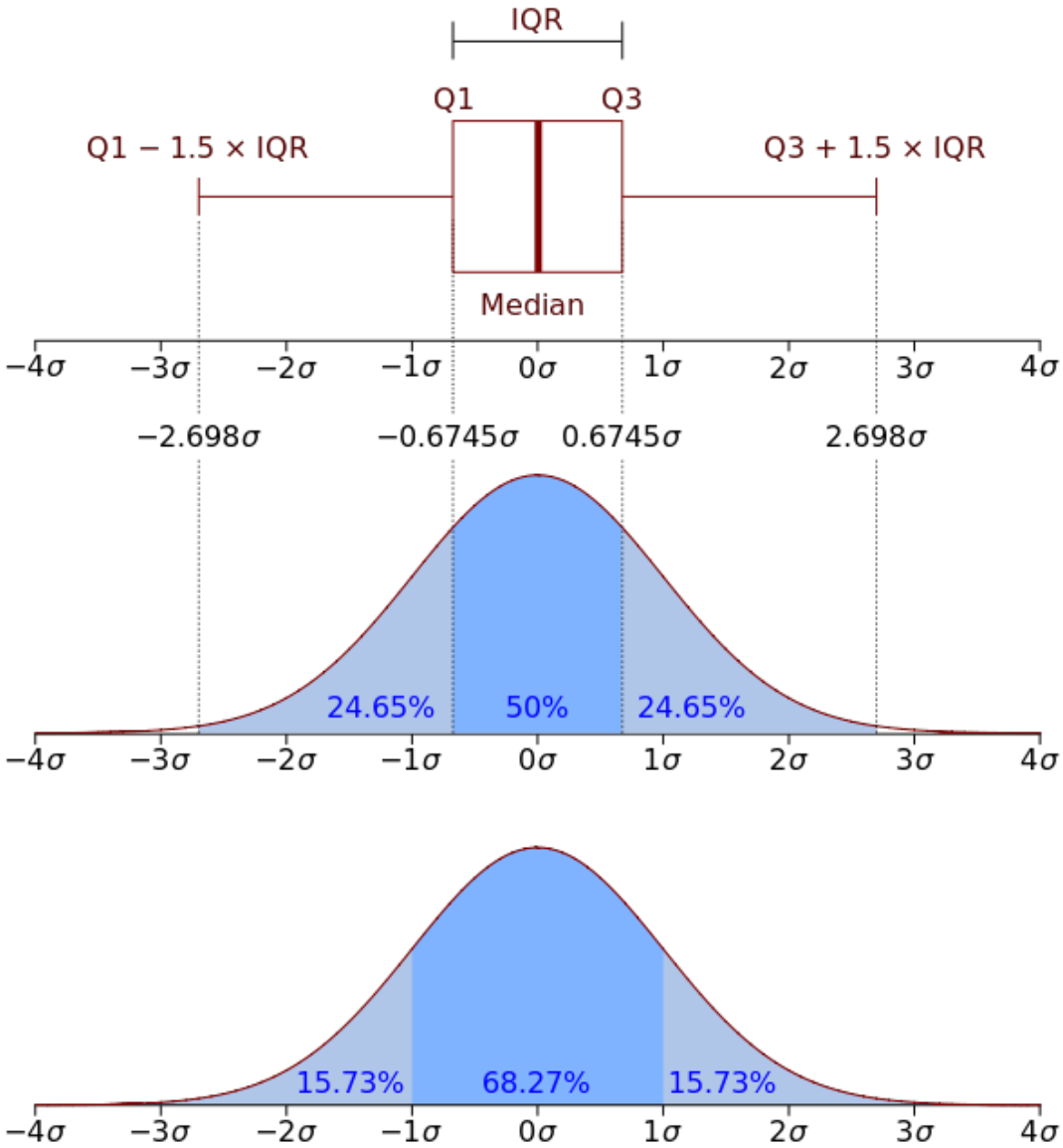
```
from scipy.stats import norm
a = norm.ppf(0.75)
print(a)
```

```
## 0.674489750196
```

```
b = norm.ppf(0.25)
print(b)
```

```
## -0.674489750196
```

Connection between IQR and SD for a standard normal random variable (cont'd)



[3] https://en.wikipedia.org/wiki/Interquartile_range

Connection between IQR and SD for a standard normal random variable (cont'd)

- $a \sim 0.675$, which means that $0.675 \times 2 = 1.35 = 1.35\sigma$ corresponds to the IQR.
- similarly, we see that 4σ roughly corresponds to the inter 95%-quantile range.
- to find the inter q-quantile range that $a\sigma$ corresponds to, you need to find the quantile at $q + z$ and $q - z$, where $q + 2 * z = 1$.

Connection between IQR and SD for a standard normal random variable (cont'd)

you should try to work this out both ways

- given an inter-quantile range q, find the number of multiples of σ

e.g. what is the value of a such that the region $\pm a\sigma$ corresponds to an inter 99%-quantile range?

- given a multiple of $k\sigma$, find the inter-quantile range q.

e.g. what is the inter-quantile range that corresponds to the region of $\pm 3\sigma$?

Coefficient of variation:

$$\frac{\sigma}{\mu}$$

is defined so the measure is dimensionless.

- for ratio scale variable only
- for log-normal distributed data this is roughly constant

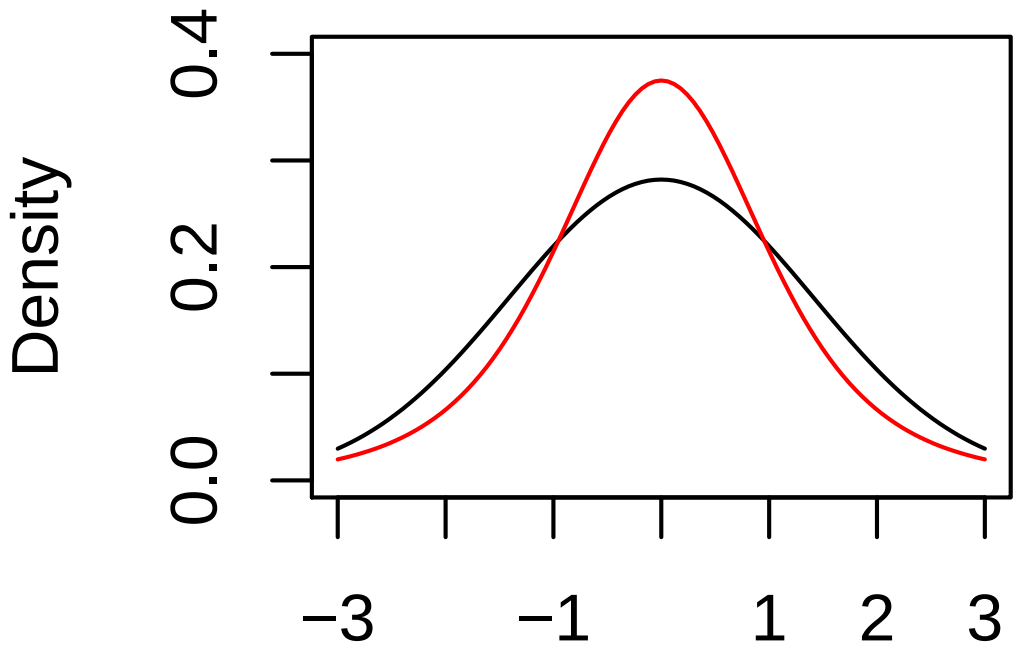
Similar idea for a robust dispersion measure, consider

$$\text{quartile coefficient of dispersion} = \frac{Q3 - Q1}{Q3 + Q1}$$

measure of shape - the tails: kurtosis

For distributions with similar mean, standard deviation and both symmetric, another possible difference in distribution arises in the tail.

```
curve(dnorm(x, sd=sqrt(2)), -3,3, ylim=c(0, 0.4), xlab="", ylab = "Density")
curve(dt(x, df=4), -3,3, col=2, add=T)
```




```
mean(rnorm(1000, sd=sqrt(2))); sd(rnorm(1000, sd=sqrt(2)));
```

```
[1] 0.005282994
```

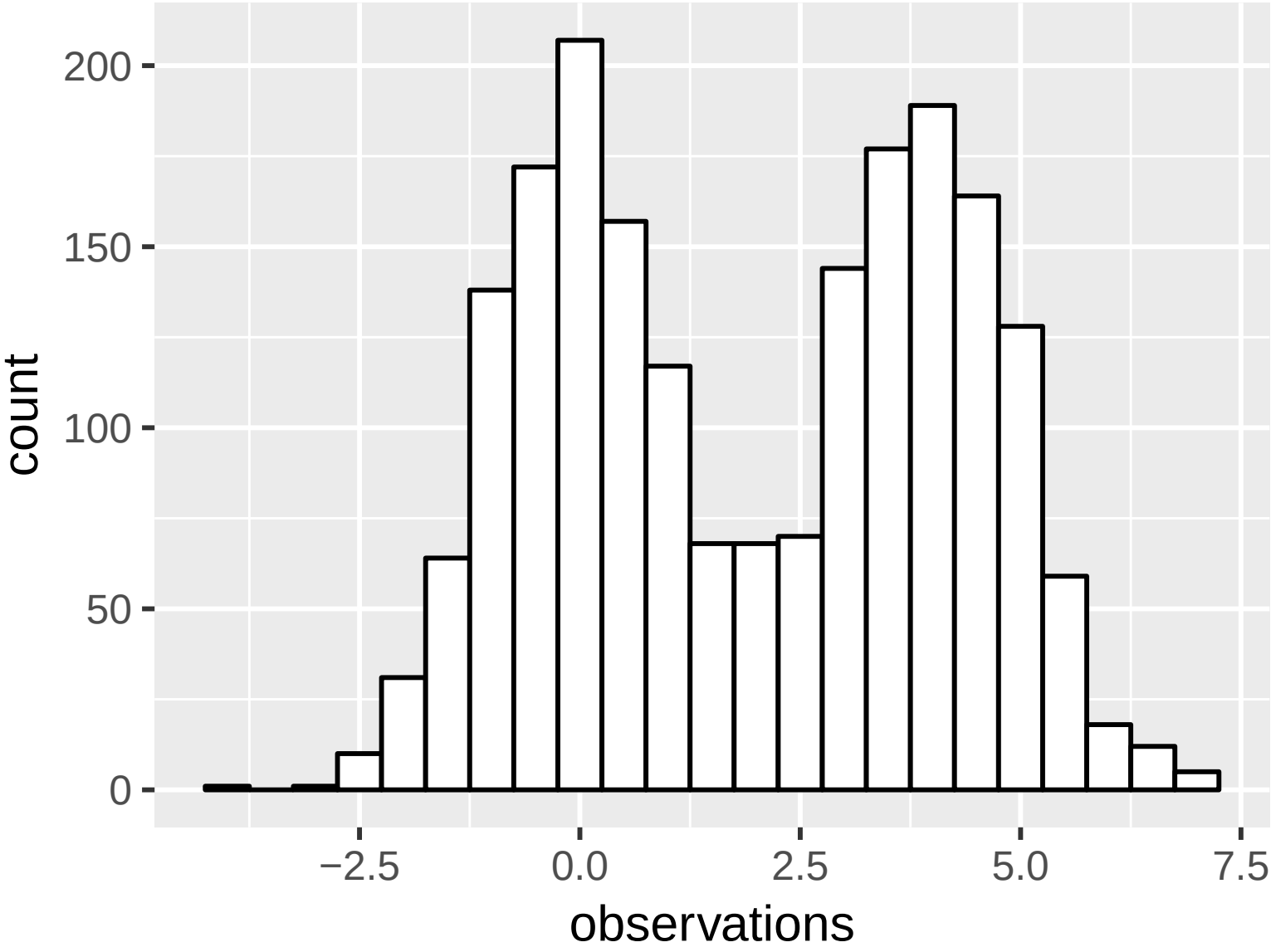
```
[1] 1.441243
```

```
mean(rt(1000, df=4)); sd(rt(1000, df=4));
```

```
[1] 0.00473542
```

measure of shape - the peaks: modality

Sometimes, it is possible to have two or more peaks in the distribution.



connecting the shape to a normal random variable

- A normal random variable has only one peak
- Due to the symmetry, the median and mean are very close
- For a normal random variable, any centralized moments higher than the 2nd (variance), is either 0 (odd moments) or a multiple of the variance (even moments). So it suffices to look at only the first two.
- In other words, a sample measure of centralized skewness or kurtosis that deviates from the expected values (0 or $3\sigma^4$) can indicate non-normality in the data.

what about descriptive statistics for two or more variables?

- Discrete vs. Discrete: Cross-tabulations and contingency tables
- Discrete vs. Continuous: Graphical representation via boxplots
- Continuous vs. Continuous: Graphical representation via scatterplots
- All possible combinations: Quantitative measures of dependence such as a range of correlation coefficients

Discrete vs. discrete

```
table(Cars93$Type, Cars93$Origin)
```

```
##
##              USA non-USA
##    Compact     7       9
##    Large      11       0
##    Midsize    10      12
##    Small       7      14
##    Sporty      8       6
##    Van         5       4
```

Discrete vs. Continuous

```
tapply(Cars93$Fuel.tank.capacity, Cars93$Origin, "median")
```

```
##     USA non-USA
##   16.45   15.90
```

```
tapply(Cars93$Fuel.tank.capacity, Cars93$Origin, "mean")
```

```
##      USA  non-USA
## 17.05208 16.25111
```

```
tapply(Cars93$Fuel.tank.capacity, Cars93$Origin, "sd")
```

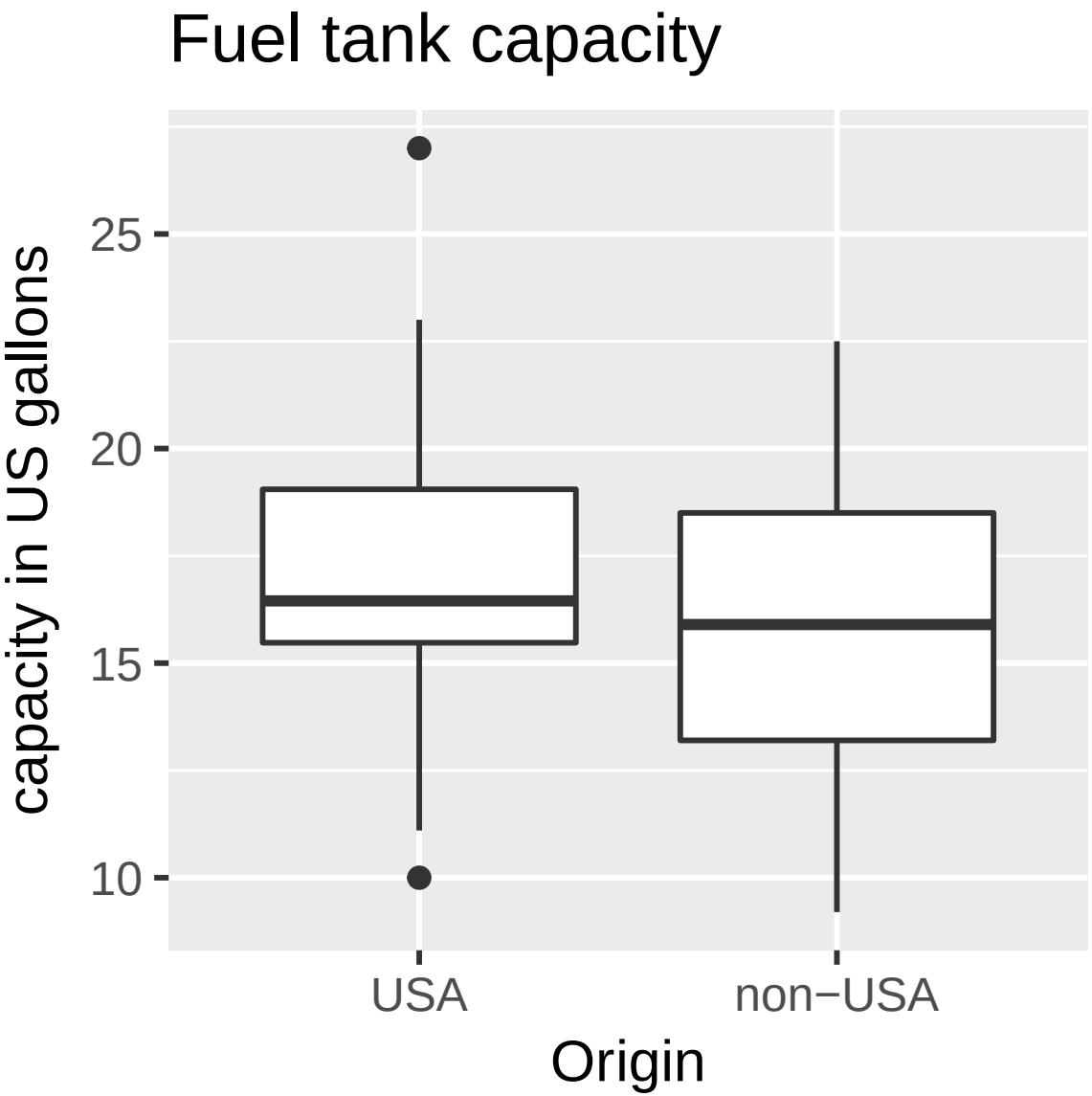
```
##      USA  non-USA
## 3.157698 3.390671
```

```
tapply(Cars93$Fuel.tank.capacity, Cars93$Origin, "IQR")
```

```
##     USA non-USA
##   3.575   5.300
```

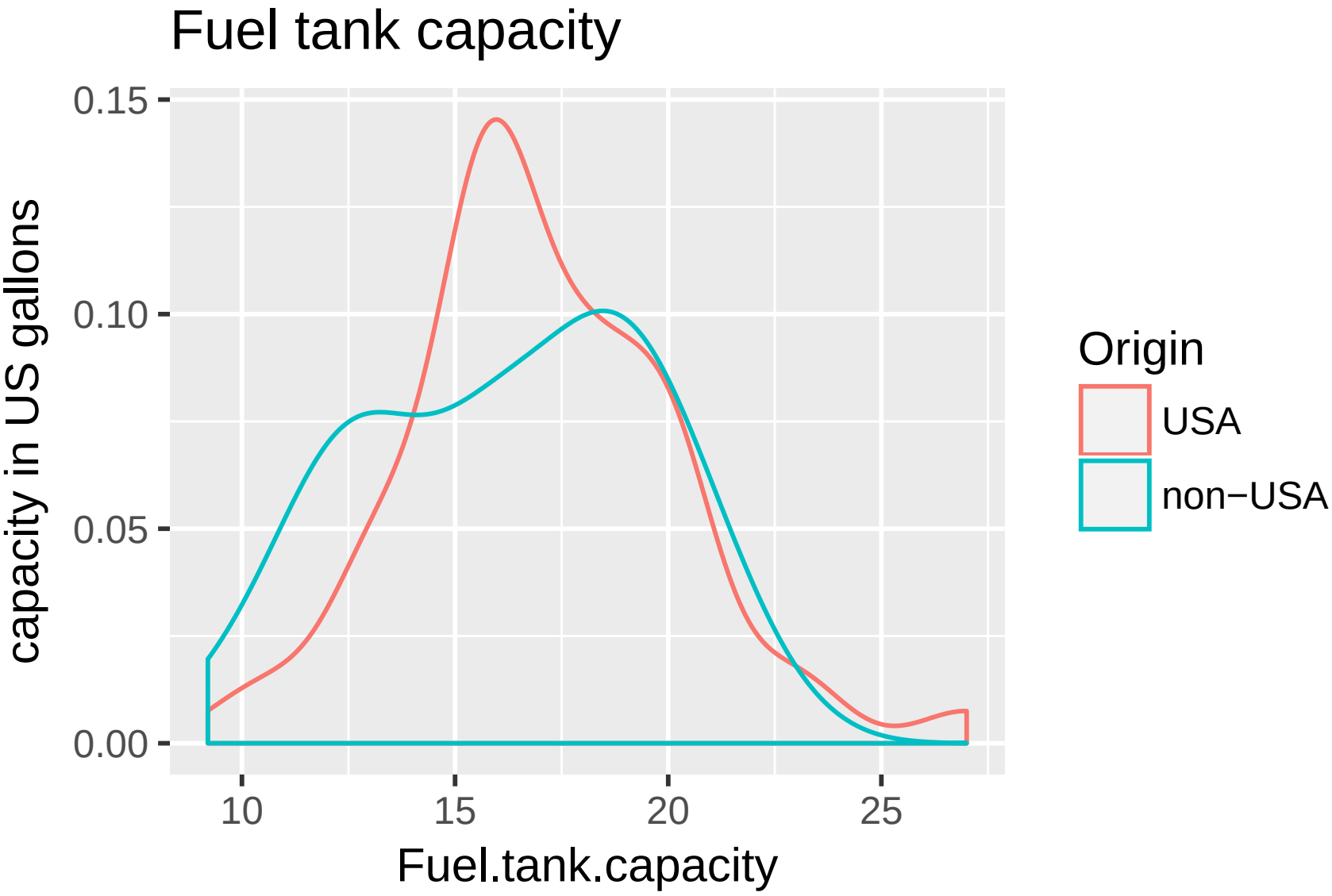
Discrete vs. Continuous

```
ggplot(data = Cars93, aes(x = Origin, y = Fuel.tank.capacity)) +
               geom_boxplot() + xlab("Origin") + ylab("capacity in US gallons") + ggtitle(paste("Fuel tank capacity"))
```



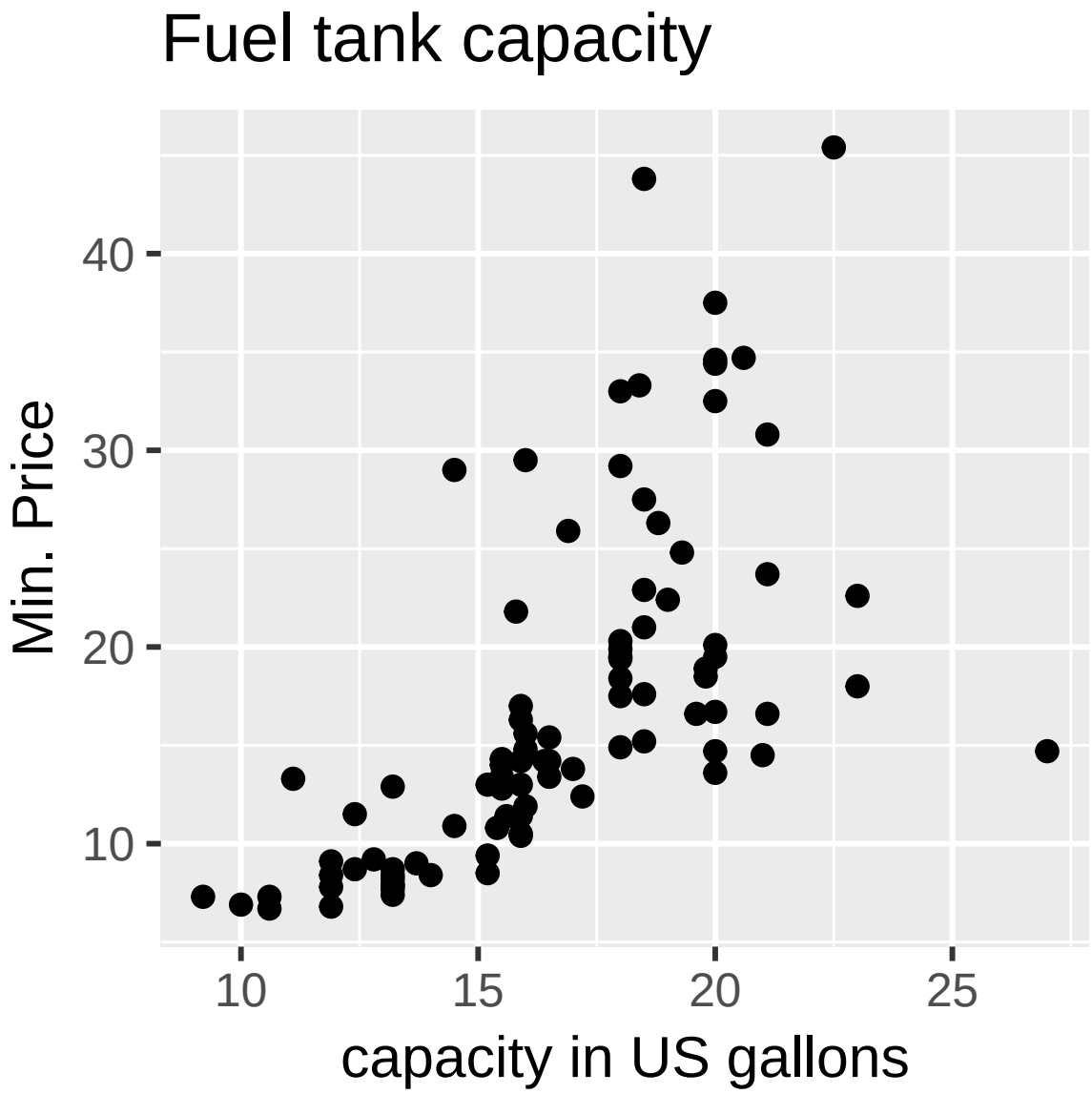
Discrete vs. Continuous

```
ggplot(data=Cars93, aes(x=Fuel.tank.capacity, colour = Origin)) +
               geom_density() + ylab("capacity in US gallons") + ggtitle(paste("Fuel tank capacity"))
```



Continuous vs. Continuous

```
ggplot(data=Cars93, aes(x=Fuel.tank.capacity, y = Min.Price)) +
                geom_point() + xlab("capacity in US gallons") + ylab("Mi
```



recap of this class

- know that statistics is the science of learning from data
- know how to summarize data
  - what are the appropriate descriptive statistics to use
  - what are the appropriate graphical display to summarize information
  - able to detect non-normality in data (outliers, distribution)