# Summarizing and Visualizing Data Part II

Elie Gurarie

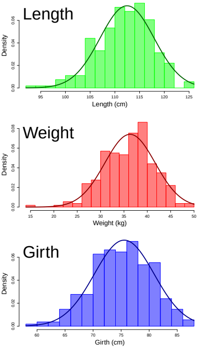
Biol 799 - Lecture 2
January 2, 2017

January 2, 2017

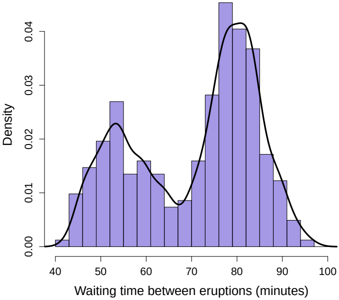
## Concepts

- Types of distributions: **skewed**, **right-skewed**, **left-skewed**, **multi-modal**
- Summary statistics
  - Measures of center: **arithmetic mean**, **median**, **geometric mean**
  - Measures of spread: **variance**, **standard deviation**, **quantiles**,
- Visualizing Data
  - boxplots
  - scatterplots
  - more!

# Distributions: **Unimodal**, **“symmetric”**




# Distributions: **Bimodal**



# Distributions: **Right-skewed**

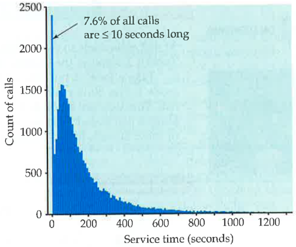
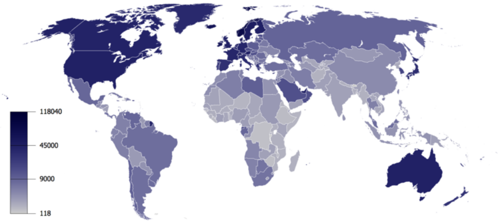


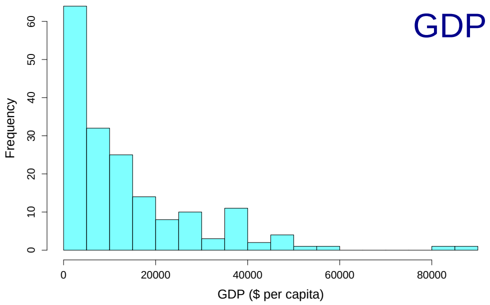

**FIGURE 1.10** The distribution of call lengths for 31,492 calls to a bank's customer service center, for Example 1.15. The data show a surprising number of very short calls. These are mostly due to representatives deliberately hanging up in order to bring down their average call length.

Most values bunched LOW, but a few values very LARGE

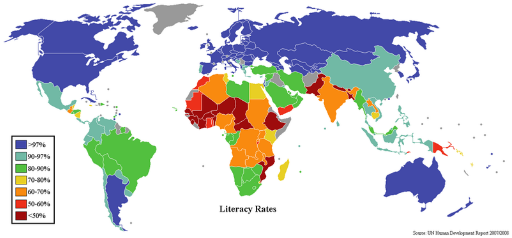
# Distributions: per capita GDP



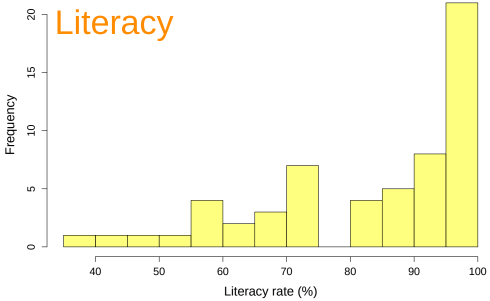
## Distributions: per capita GDP



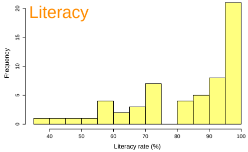

## Consider countries: Literacy

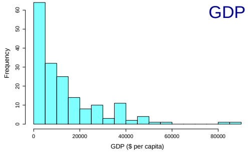


# Consider countries: **Left-skewed**




Most values bunched HIGH, fewer values very LOW.

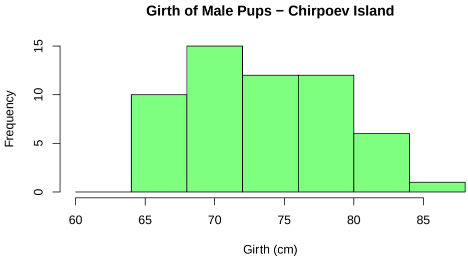



**Skewness**

is a measure of **assymetry** in a distribution.

Question: How do describe the MIDDLE and the SPREAD of a distribution?




$G_i$: 69, 76, 72, 73.5, 72, 75, 67, 77, 71, 74, 79, 77, 67.5, 71, 76, 74, 84, 77.5, 67, 80, 67, 78, 76, 70, 81, 77, 69, 66, 71, 77, 88, 71.5, 67, 76, 70, 78, 80.5, 69.5, 72.5, 79, 73, 74.5, 73, 65.5, 66.5, 72, 80, 82, 83, 71, 71, 70, 82, 78, 64.5, 66

# Definition of **Mean**

Given $\{X_1, X_2, X_3, ...X_n\}$, the mean is defined as:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} \quad (1)$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i \quad (2)$$

The mean is also known as: **arithmetic mean** or, non-technically, **average**.

## Definition of **Mean**

So ...

$\sum G_i = 69 + 76 + 72 + 73.5 + 72 + 75 + 67 + 77 + 71 + 74 + 79 + 77 + 67.5 + 71 + 76 + 74 + 84 + 77.5 + 67 + 80 + 67 + 78 + 76 + 70 + 81 + 77 + 69 + 66 + 71 + 77 + 88 + 71.5 + 67 + 76 + 70 + 78 + 80.5 + 69.5 + 72.5 + 79 + 73 + 74.5 + 73 + 65.5 + 66.5 + 72 + 80 + 82 + 83 + 71 + 71 + 70 + 82 + 78 + 64.5 + 66$

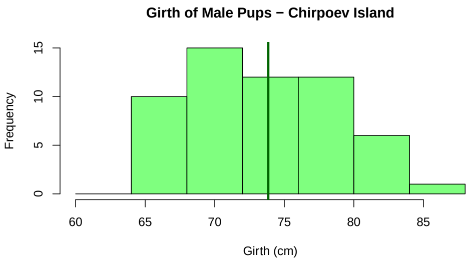
$n = 56$

$\overline{G} = \frac{1}{n} \sum_{i=1}^{n} G_i = \frac{1}{56} \times 4135.5 = 73.85$

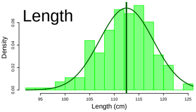
**R code: calculating means**

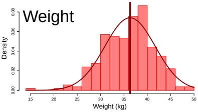
```
> mean(Girth)
[1] 73.84821
> sum(Girth)/length(Girth)
[1] 73.84821
```

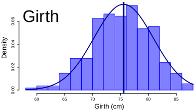
## Means for different kinds of data



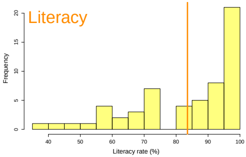
The **mean** is most *mean*ingful for unimodal, symmetric distributions

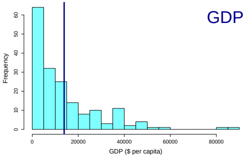


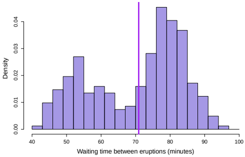




The **mean** is somewhat less *mean*ingful for multimodal or asymmetric distributions



## Another measure of center: **median**

Definition: The **median** is the point of a distribution that splits the data into two equal sized halves.

## Median

First, order the data:

$G_{(i)} =$ 64.5, 65.5, 66, 66, 66.5, 67, 67, 67, 67, 67.5, 69, 69, 69.5, 70, 70, 70, 71, 71, 71, 71, 71, 71.5, 72, 72, 72, 72.5, 73, 73, 73.5, 74, 74, 74.5, 75, 76, 76, 76, 76, 77, 77, 77, 77, 77.5, 78, 78, 78, 79, 79, 80, 80, 80.5, 81, 82, 82, 83, 84, 88

Note the notation:

- $X_{(i)}$ is the minimum value of $X$
- $X_{(n)}$ is the maximum value of $X$.

# Median

Second, find the point that splits the data in half.

- If $n$ is *odd*, you take point: $\widetilde{X} = X_{(n+1)/2}$.
- If $n$ is *even*, you take the mid-point between: $\widetilde{X} = \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right)$.

Example - Pup Girth: $n = 56$, so we take data points 28 and 29, and average.

**Male girth** = 64.5, 65.5, 66, 66, 66.5, 67, 67, 67, 67, 67.5, 69, 69, 69.5, 70, 70, 70, 71, 71, 71, 71, 71, 71.5, 72, 72, 72, 72.5, 73, 73, 73.5, 74, 74, 74.5, 75, 76, 76, 76, 76, 77, 77, 77, 77, 77.5, 78, 78, 78, 79, 79, 80, 80, 80.5, 81, 82, 82, 83, 84, 88
**Male girth** = 64.5, 65.5, 66, 66, 66.5, 67, 67, 67, 67, 67.5, 69, 69, 69.5, 70, 70, 70, 71, 71, 71, 71, 71, 71.5, 72, 72, 72, 72.5, 73, **73**, **73.5**, 74, 74, 74.5, 75, 76, 76, 76, 76, 77, 77, 77, 77, 77.5, 78, 78, 78, 79, 79, 80, 80, 80.5, 81, 82, 82, 83, 84, 88

So $\widetilde{G} = (73 + 73.5)/2 = 73.25$. (Compare to $\overline{G} = 73.85$.)

# Medians

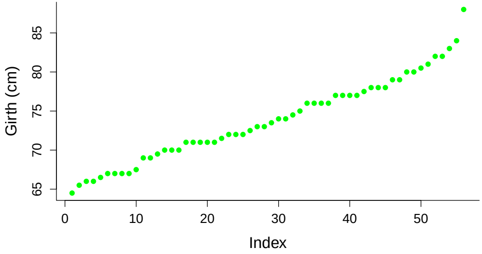
**Male girth** = 64.5, 65.5, 66, 66, 66.5, 67, 67, 67, 67, 67.5, 69, 69, 69.5, 70, 70, 70, 71, 71, 71, 71, 71, 71.5, 72, 72, 72, 72.5, 73, **73**, **73.5**, 74, 74, 74.5, 75, 76, 76, 76, 76, 77, 77, 77, 77, 77.5, 78, 78, 78, 79, 79, 80, 80, 80.5, 81, 82, 82, 83, 84, 88

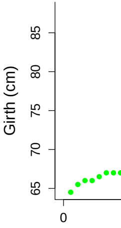
So $\widetilde{G} = (73 + 73.5)/2 = 73.25$.

**R code: calculating median**

```
> median(Girth)
[1] 73.25
> # creating a customized "median function" - is an exercise
```

## Median of Girth




## Geometric mean

Given $\{X_1, X_2, X_3, ...X_n\}$, the *geometric mean* is defined as:

$$\check{X} = (X_1 \cdot X_2 \cdot X_3 \cdot ... \cdot X_n)^{1/n} \quad (3)$$

$$= \left( \prod_{i=1}^{n} X_i \right)^{1/n} \quad (4)$$

A little bit of math:

$$\log(\check{X}) = \log \left( \left( \prod_{i=1}^{n} X_i \right)^{1/n} \right)$$

$$= \frac{1}{n} \log \left( \prod_{i=1}^{n} X_i \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log X_i$$

So ...

$$\check{X} = e^{\overline{\log X}}$$

## Geometric mean

**R code: calculating geometric means**

```
> X <- 1:170
> prod(X)^(1/length(X))
[1] 63.83567
> # But note that things break down at high numbers.
> X <- 1:171
> prod(X)^(1/length(X))
[1] Inf
> # This is where logarithms are most useful!
> exp(mean(log(X)))
[1] 64.20457
```

## Comparison of measures

Example: $X = \{1, 2, 3, 4, 5\}$

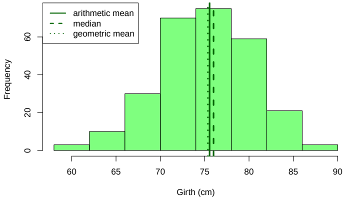
| Measure | Notation | Result |
|---|---|---|
| Arithmetic Mean | $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ | 3 |
| Median | $\widetilde{X} = X_{(n/2)}$ or $X_{(n+1/2)}$ | 3 |
| Geometric Mean | $\breve{X} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$ | 2.67 |

## Comparison of measures

Example: $X = \{1, 10, 100, 1000, 10000\}$

| Measure | Notation | Result |
|---|---|---|
| Arithmetic Mean | $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ | 2222.2 |
| Median | $\widetilde{X} = X_{(n/2)}$ or $X_{(n+1/2)}$ | 100 |
| Geometric Mean | $\check{X} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$ | 100 |

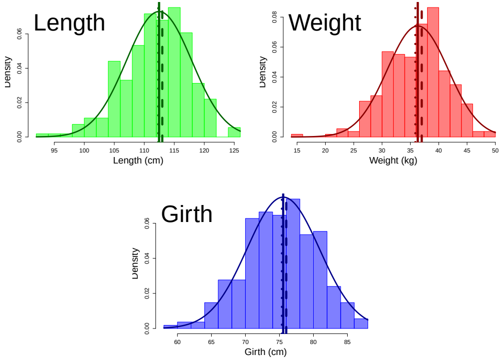
## Compare median and mean



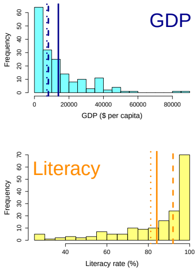

**R code: adding lines to histogram**

```
> GirthMales <- Pups$Girth[Pups$Sex == "M"]
> hist(GirthMales , col=rgb(0,1,0,.5), breaks=20)
> abline(v = median(GirthMales), lwd=3)      # draws a vertical line
> abline(v = mean(GirthMales), lwd=3, lty=2)
> geomean <- function(x) exp(mean(log(x)))
> abline(v = geomean(GirthMales), lwd=3, lty=3)
> legend("topleft", lwd=3, lty=1:3,
         legend=c("arithmetic mean", "median", "geometric mean"))
```

# Median and mean for symmetric distributions




# Median and mean for asymmetric distributions




| Measure | GDP | Literacy |
|---|---|---|
| $\overline{X}$ | \$13,871 | 84.10% |
| $\widetilde{X}$ | \$8,080 | 91.95% |
| $\check{X}$ | \$7,228 | 81.16% |

**R code: Tabulating Measures of Center**

```
> # create a customized function
> getCenters <- function(X)
+ {
+   M1 <- mean(X, na.rm=TRUE)
+   M2 <- median(X, na.rm=TRUE)
+   M3 <- exp(mean(log(X), na.rm=TRUE))
+   # naming the output elements is useful later
+   return(c(mean=M1, median=M2, geo.mean=M3))
+ }
> # make a data frame
> data.frame(GDP = getCenters(GDP),
+              Literacy= getCenters(Literacy))
               GDP Literacy
mean     13871.492 84.09965
median    8080.000 91.95000
geo.mean  7228.001 81.16408
```

### Measure of spread: Variance

Given data $x_1, x_2, x_3 ... x_n$, the **population variance** is

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

and the **sample variance** is

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

We will talk about the difference between these and where the $n-1$ comes from later in the course. For now, we mostly use the "sample variance", because we typically do not assume that we have measured the entire population.

### Measure of spread: Standard Deviation

**Population standard deviation:**

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

**Sample standard deviation:**

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

## Measures of spread

Some girth data: G = 69, 76, 72, 73.5, 72, 75, 67, 77, 71, 74, 79, 77, 67.5, 71, 76 Mean: $\overline{G} = 73.133$.

| | $G$ |
|---|---|
| 1 | 69.00 |
| 2 | 76.00 |
| 3 | 72.00 |
| 4 | 73.50 |
| 5 | 72.00 |
| 6 | 75.00 |
| 7 | 67.00 |
| 8 | 77.00 |
| 9 | 71.00 |
| 10 | 74.00 |
| 11 | 79.00 |
| 12 | 77.00 |
| 13 | 67.50 |
| 14 | 71.00 |
| 15 | 76.00 |

| | $G$ | $G - \overline{G}$ |
|---|---|---|
| 1 | 69.00 | -4.13 |
| 2 | 76.00 | 2.87 |
| 3 | 72.00 | -1.13 |
| 4 | 73.50 | 0.37 |
| 5 | 72.00 | -1.13 |
| 6 | 75.00 | 1.87 |
| 7 | 67.00 | -6.13 |
| 8 | 77.00 | 3.87 |
| 9 | 71.00 | -2.13 |

# Measures of spread

Mean: $\overline{G} = 73.133$.

| | $G$ | $G - \overline{G}$ | $(G - \overline{G})^2$ |
|---|---|---|---|
| 1 | 69.00 | -4.13 | 17.08 |
| 2 | 76.00 | 2.87 | 8.22 |
| 3 | 72.00 | -1.13 | 1.28 |
| 4 | 73.50 | 0.37 | 0.13 |
| 5 | 72.00 | -1.13 | 1.28 |
| 6 | 75.00 | 1.87 | 3.48 |
| 7 | 67.00 | -6.13 | 37.62 |
| 8 | 77.00 | 3.87 | 14.95 |
| 9 | 71.00 | -2.13 | 4.55 |
| 10 | 74.00 | 0.87 | 0.75 |
| 11 | 79.00 | 5.87 | 34.42 |
| 12 | 77.00 | 3.87 | 14.95 |
| 13 | 67.50 | -5.63 | 31.73 |
| 14 | 71.00 | -2.13 | 4.55 |
| 15 | 76.00 | 2.87 | 8.22 |
| $\Sigma$: | 1097 | 0 | 183 |

**Variance**:
$s_G^2 = 183/(15-1) = 13.07 \text{ cm}^2$

**Standard devation**:
$s_G = \sqrt{13.07\text{cm}^2} = 3.617$ cm.


# Measures of spread

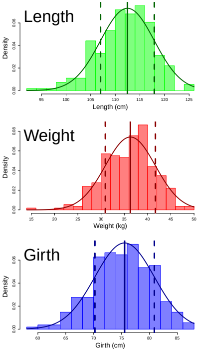
**Variance**:
$s_G^2 = 183/(15-1) = 13.07 \text{ cm}^2$

**Standard devation**:
$s_G = \sqrt{13.07\text{cm}^2} = 3.617$ cm.


**R code: variance and standard deviation**

```
> var(g)
[1] 13.0881
> # by "hand"
> sum((g-mean(g))^2)/(length(g)-1)
[1] 13.0881
> # but not:
> sum((g-mean(g))^2)/(length(g))
[1] 12.21556
> # standard deviation
> sd(g)
[1] 3.617747
> sqrt(var(g))
[1] 3.617747
```

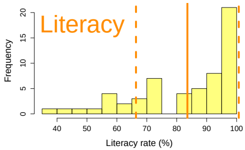
# Median and mean for symmetric distributions

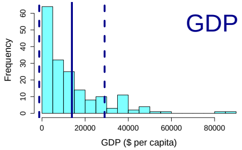



| | $\overline{x}$ | $s$ |
|---|---|---|
| Length | 112.46 | 5.46 |
| Weight | 36.30 | 5.38 |
| Girth | 75.55 | 5.33 |

**Some features of sample standard deviation...**

- They are the most common measurement of spread for a distribution
- They are always positive and have the same units as the measurements (unlike variance)
- Points that are more distant from the mean have a larger contribution to the standard deviation.
- For "normal" distributions, the range: $\overline{x} - 2s_x$ to $\overline{x} + 2s_x$ includes *about* 95% of the observations.




**A Warning**

As with means ($\overline{x}$), standard deviations are most meaningful for symmetric and "normal" distributions.

## Measures of Spread II: Quartiles

- Take ordered observations.
- Separate them into 4 groups of equal size.
- Report: $Q_0$, $Q_{25}$, $Q_{50}$, $Q_{75}$ and $Q_{100}$
  - (split the differences between neighboring observations)

Male girth = 64.5, 65.5, 66, 66, 66.5, 67, 67, 67, 67, 67.5, 69, 69, 69.5, 70, 70, 70, 71, 71, 71, 71, 71, 71.5, 72, 72, 72, 72.5, 73, 73, 73.5, 74, 74, 74.5, 75, 76, 76, 76, 76, 77, 77, 77, 77, 77.5, 78, 78, 78, 79, 79, 80, 80, 80.5, 81, 82, 82, 83, 84, 88

64.5 65.5 66.0 66.0 66.5 67.0 67.0 67.0 67.0 67.5 69.0 69.0 69.5 70.0
70.0 70.0 71.0 71.0 71.0 71.0 71.0 71.5 72.0 72.0 72.0 72.5 73.0 73.0
73.5 74.0 74.0 74.5 75.0 76.0 76.0 76.0 76.0 77.0 77.0 77.0 77.0 77.5
78.0 78.0 78.0 79.0 79.0 80.0 80.0 80.5 81.0 82.0 82.0 83.0 84.0 88.0

64.5 65.5 66.0 66.0 66.5 67.0 67.0 67.0 67.0 67.5 69.0 69.0 69.5 70.0
70.0 70.0 71.0 71.0 71.0 71.0 71.0 71.5 72.0 72.0 72.0 72.5 73.0 73.0
73.5 74.0 74.0 74.5 75.0 76.0 76.0 76.0 76.0 77.0 77.0 77.0 77.0 77.5
78.0 78.0 78.0 79.0 79.0 80.0 80.0 80.5 81.0 82.0 82.0 83.0 84.0 88.0

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 64.5 | 70 | 73.25 | 77.75 | 88 |

# Quartiles

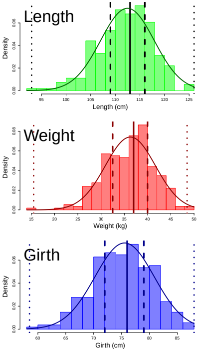
64.5 65.5 66.0 66.0 66.5 67.0 67.0 67.0 67.0 67.5 69.0 69.0 69.5 70.0
70.0 70.0 71.0 71.0 71.0 71.0 71.0 71.5 72.0 72.0 72.0 72.5 73.0 73.0
73.5 74.0 74.0 74.5 75.0 76.0 76.0 76.0 76.0 77.0 77.0 77.0 77.0 77.5
78.0 78.0 78.0 79.0 79.0 80.0 80.0 80.5 81.0 82.0 82.0 83.0 84.0 88.0

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 64.5 | 70 | 73.25 | 77.75 | 88 |

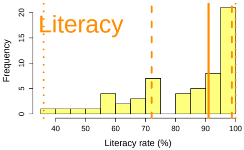
**R code: quantiles**

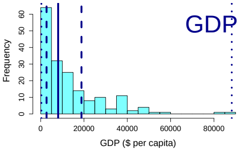
```
>quantile(Girth)
   0%    25%    50%    75%   100%
64.5     70  73.25  77.75     88
```

# Median and mean for symmetric distributions




| | weight | girth | length |
|---|---|---|---|
| 0% | 15.50 | 58.50 | 93.00 |
| 25% | 32.50 | 72.00 | 109.00 |
| 50% | 37.00 | 76.00 | 113.00 |
| 75% | 40.00 | 79.00 | 116.00 |
| 100% | 48.50 | 88.00 | 126.00 |



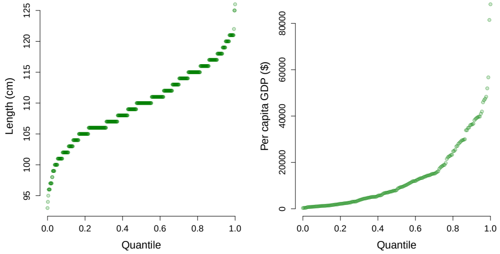




| | Literacy | GDP |
|---|---|---|
| 0% | 36.00 | 329 |
| 25% | 72.00 | 2721 |
| 50% | 91.00 | 8080 |
| 75% | 98.75 | 18841 |
| 100% | 100.00 | 88222 |

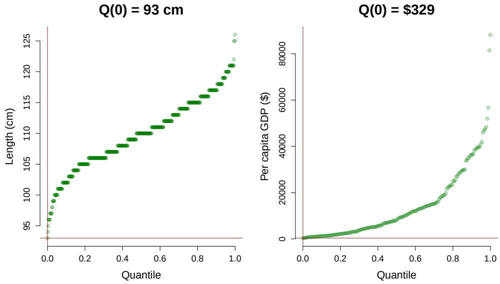
**Quantiles...**

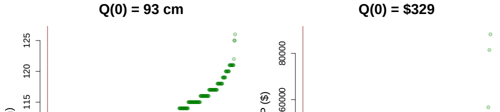
provide a *flexible*, *empirical*, *robust* description of ANY kind of distribution.

# Quantiles vs. Quartiles

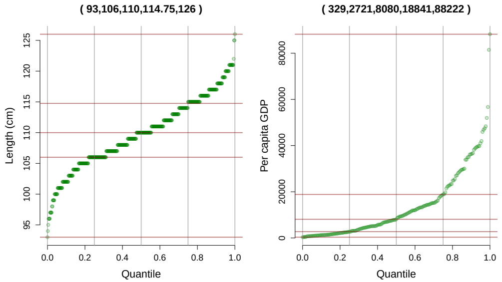



# Quantiles vs. Quartiles



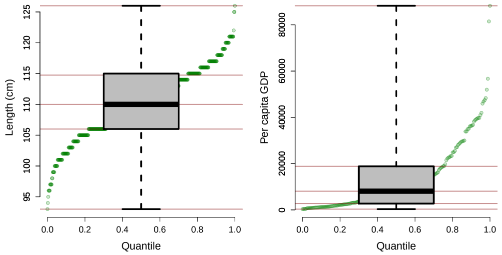



# Quantiles vs. Quartiles




The **quartiles** are the 0th, 25th, 50th, 75th and 100th **quantile**: Q(0), Q(25), Q(50), Q(75), Q(100).
Note that: Q(0) = minimum, Q(50) = median, Q(100) = maximum.

# **Boxplots** (aka **Box-and-whiskers plots**)




The **box** spans Q(25)-Q(75) aka **the inter-quartile range** (IQR).
The **whiskers** span Q(0)-Q(100) - the **range**

## Summary statistics: Review

**Measures of center**

| | |
|---|---|
| mean | `mean(X)` |
| median | `median(X)` |
| geometric mean | `exp(mean(log(X)))` |

**Measures of spread**

| | |
|---|---|
| sample variance | `var(X)` |
| sample standard deviation | `sd(X)` |
| range | `range(X)` |
| | `c(min(X), max(X))` |
| | `quantile(X, c(0,1))` |
| quartiles | `quantile(X)` |
| inter-quartile range | `quantile(X, c(.25, .75))` |

## Getting summary statistics for different groups

**R code: The long way**

```
> mean(Length[Island == "Chirpoev"])
[1] 109.9798
> mean(Length[Island == "Antsiferov"])
[1] 110.57
> mean(Length[Island == "Lovushki"])
[1] 109.38
> mean(Length[Island == "Lovushki" & Sex == "M"])
[1] 112.3
```

But this is very tedious!

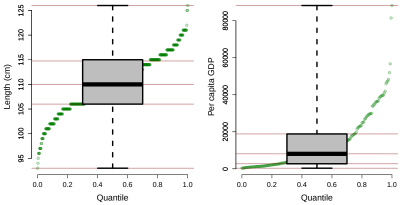
**R code: The slick way**

```
>   tapply(Length, Island, mean)
Antsiferov   Chirpoev   Lovushki    Raykoke   Srednova
  110.5700   109.9798   109.3800   110.7600   108.5152
>   tapply(Length, paste(Island, Sex), sd)
Antsiferov F Antsiferov M   Chirpoev F   Chirpoev M   Lovushki F
    4.837619     5.607867     5.383314     5.574495     4.046717
  Lovushki M    Raykoke F    Raykoke M   Srednova F   Srednova M
    6.071849     4.734624     4.677942     4.454222     5.339355
```

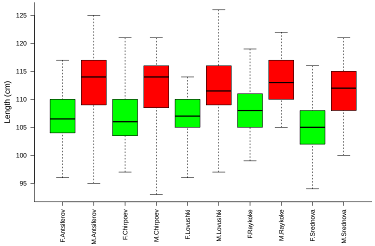
# Visualizing relationships: Boxplots

**Boxplots**

- Allow you to easily see if the distribution is “skewed”.
- Allow you to easily compare distributions of different groups.
- Allow us to visualize relationships between *quantitative* and *categorical* variables

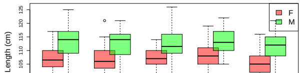



# Boxplot examples

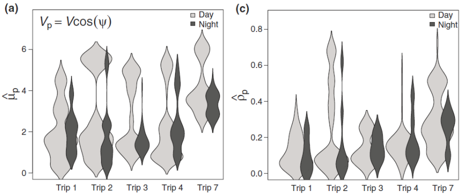



**R code: boxplots**

```
boxplot(Length ~ Sex + Island)
```



# Violin plots




*Ecology Letters*, (2009) **12**: 395–408 doi: 10.1111/j.1461-0248.2009.01293.x

LETTER

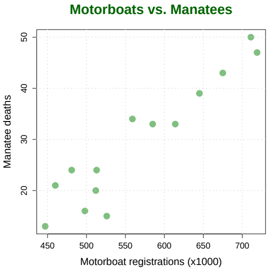
**A novel method for identifying behavioural changes in animal movement data**

Eliezer Gurarie,[1]* Russel D. Andrews[2] and Kristin L. Laidre[3]

# Manatees and motorboats

| Year | Motorboat Registrations (thousands) | Manatee Motorboat Deaths |
|---|---|---|
| 1977 | 447 | 13 |
| 1978 | 460 | 21 |
| 1979 | 481 | 24 |
| 1980 | 498 | 16 |
| 1981 | 513 | 24 |
| 1982 | 512 | 20 |
| 1983 | 526 | 15 |
| 1984 | 559 | 34 |
| 1985 | 585 | 33 |
| 1986 | 614 | 33 |
| 1987 | 645 | 39 |
| 1988 | 675 | 43 |
| 1989 | 711 | 50 |
| 1990 | 719 | 47 |

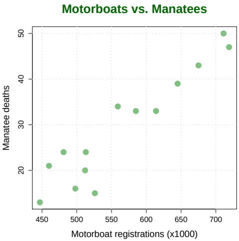
# Manatees and motorboats: Scatterplot




**Note**: Typically, the **explanatory** variable is on the x-axis, the **response** is on the y-axis.

# Manatees and motorboats: Scatterplots

allow us to visually characterize the relationships between *continuous/quantitative* variables.




Identify:

- **direction** (positive/negative),
- **form** (linear/non-linear),
- **strength** (strong/weak)

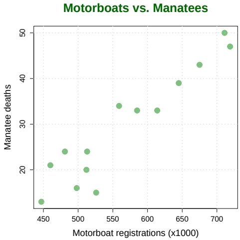
Of a relationship

**R code: scatterplot**

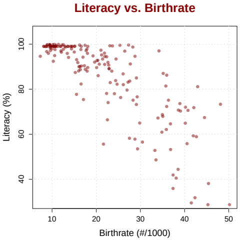
```
plot(Motorboats, Deaths)
```

# **Direction** of relationship

### Positive relationship

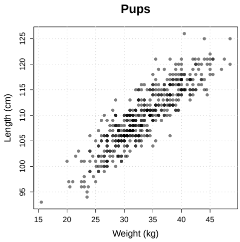



### Negative relationship

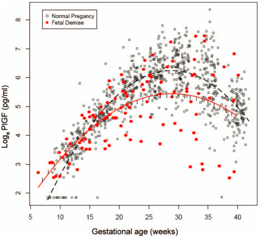



# **Form** of relationship

### Linear relationship



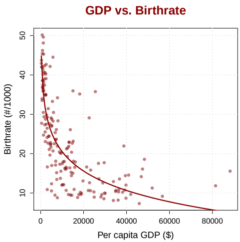

### Non-linear relationship



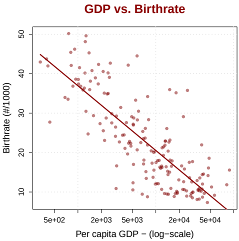

Romero et al. (2010) The Journal of Maternal-Fetal and Neonatal Medicine 23(12): 1384Ű1399

# **Form** of relationship

**Non-linear relationship**




**Non-linear ... linearized!**

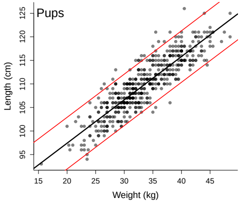



**R code:log transformation**

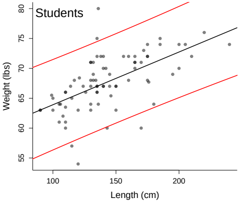
```
plot(GDP, Birthrate, log="x")
```

# **Strength** of relationship

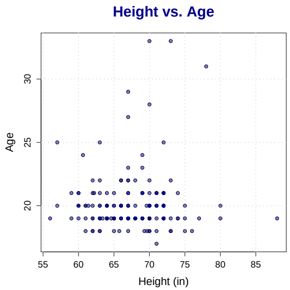
**Stronger relationship**




**Weaker relationship**



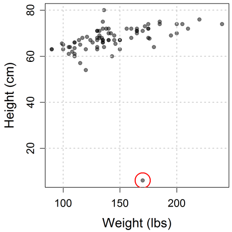
## **No** relationship




Knowing your height tells me basically nothing about your age.

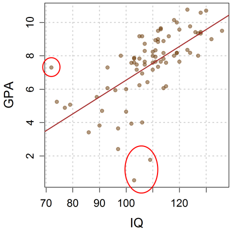
## Scatterplots help identify outliers

**Data-entry error**




Weight: 6 kg?

**Informative outliers**



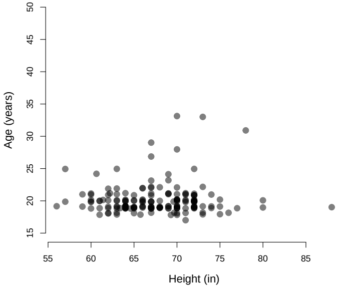

Over- and underachievement

# Scatterplots show 2-dimensions

| Year | Motorboat Registrations (thousands) | Manatee Motorboat Deaths |
|---|---|---|
| 1977 | 447 | 13 |
| 1978 | 460 | 21 |
| 1979 | 481 | 24 |
| 1980 | 498 | 16 |
| 1981 | 513 | 24 |
| 1982 | 512 | 20 |
| 1983 | 526 | 15 |
| 1984 | 559 | 34 |
| 1985 | 585 | 33 |
| 1986 | 614 | 33 |
| 1987 | 645 | 39 |
| 1988 | 675 | 43 |
| 1989 | 711 | 50 |
| 1990 | 719 | 47 |

# Many datasets have MANY dimensions!

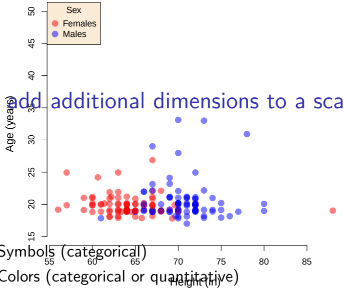
| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Section | Sex | Age | Height (in) | Weight (lbs) | Shoe Size | Right/Left-Handed |
| 2 | AA | Male | 25 | 74 | 205 | 10.5 | R |
| 3 | AA | Female | 19 | 64 | 105 | 7 | R |
| 4 | AA | Male | 19 | | 150 | 13 | R |
| 5 | AA | Male | 20 | 74 | 185 | 12 | R |
| 6 | AA | Female | 18 | 66 | 110 | 8.5 | R |
| 7 | AA | Female | 20 | 63 | 90 | 6.5 | R |
| 8 | AA | Male | 20 | 68 | 132 | 9.5 | R |
| 9 | AA | Female | 21 | 62 | 108 | 6.5 | R |
| 10 | AA | Male | 21 | | | | R |
| 11 | AA | Male | 19 | 70 | 200 | 12 | R |
| 12 | AA | Male | 32 | 71 | 130 | 10.5 | R |
| 13 | AA | Female | 19 | 68 | 117 | 7.5 | R |
| 14 | AA | Male | 21 | 70 | 142 | 10 | R |
| 15 | AA | Male | 19 | 74 | 200 | 11.5 | R |
| 16 | AA | Male | 19 | 68 | 130 | 10 | R |
| 17 | AA | Male | 19 | 68 | 160 | 8.5 | R |
| 18 | AA | Male | 20 | 74 | 175 | 11 | R |
| 19 | AA | Male | 20 | 68 | 134 | 9.5 | L |
| 20 | AA | Male | 28 | 6 | 170 | 10 | R |




**R code**

```
plot(h2, a2, ylim=c(15,50), col=rgb(0,0,0,.5), pch=19)
```

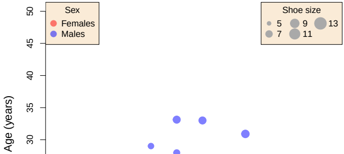
## How do we add additional dimensions to a scatterplot?




- With Symbols (categorical)
- With Colors (categorical or quantitative)
- With Sizes (quantitative)
- Also ... animations (especially for processes in time)

**R code**

```
cols <- c(rgb(1,0,0,.5), rgb(0,0,1,.5))
plot(h2, a2, ylim=c(15,50), col=cols[Sex], pch=19)
```



## Commentary...

In times like this when unemployment rates are up to 13% and income has fallen by 5% and suicide rates are climbing I get so *angry* that the government is wasting money on things like collection of statistics.
*Hans Rosling quoting a radio talkshow guest.*


http://www.gapminder.org/