DS 3003: Descriptive Statistics

Spring 2022

Youmi Suk

School of Data Science, University of Virginia

1. 18. 2022

Descriptive Statistics

- **Location**: mean, median, mode, quantiles
- **Dispersion**: standard deviation, variance, range, interquartile range; covariance, correlation

Example Data

```r
library(haven) # or use package foreign
dat <- read_sav("income_exmpl.sav")

head(dat)
```

```
## # A tibble: 6 × 6
##           sex   age          edu          occ  oexp income
##     <dbl+lbl> <dbl>    <dbl+lbl>    <dbl+lbl> <dbl>  <dbl>
## 1 1 [female]    62 1 [low]      1 [low]         35    953
## 2 0 [male]      32 3 [high]     3 [high]         6   1224
## 3 0 [male]      56 2 [medium]   3 [high]        36   1466
## 4 1 [female]    63 2 [medium]   2 [medium]      38   1339
## 5 0 [male]      20 1 [low]      1 [low]          3   1184
## 6 1 [female]    38 2 [medium]   2 [medium]      12   1196
```

Example Data

- We want to summarize the `income` variable.
- Check the `income` variable.

```
length(dat$income)
```

```
## [1] 1922
```

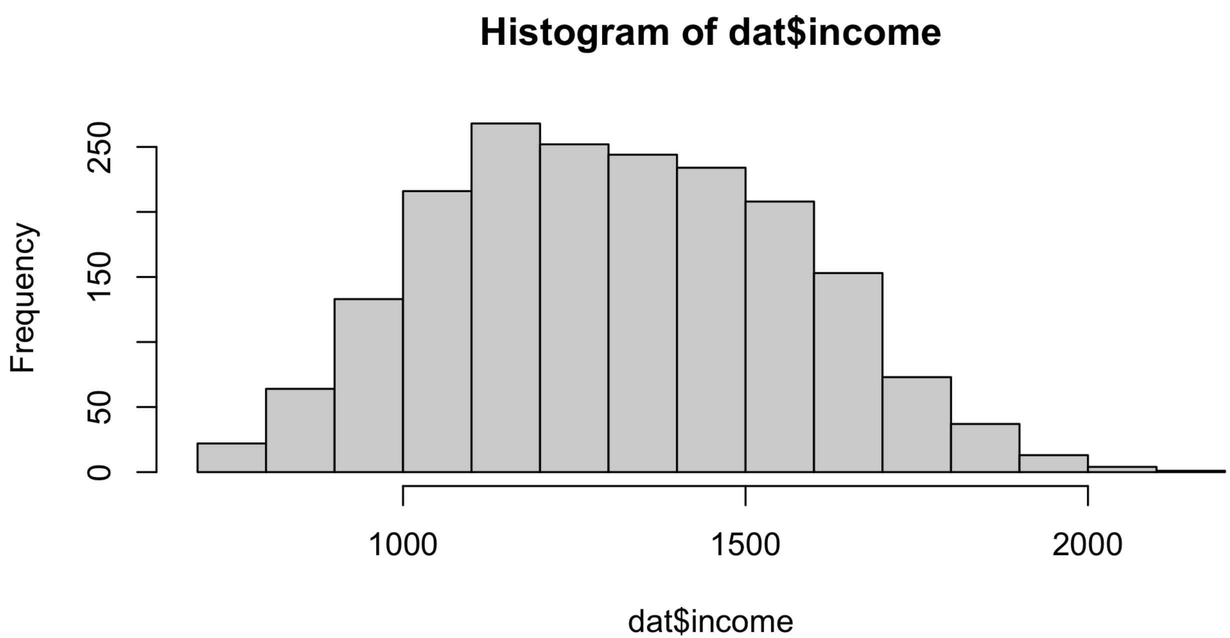
```
head(dat$income, 10)
```

```
##  [1]  953 1224 1466 1339 1184 1196  951 1039 1438 1000
```

Example Data

- We want to summarize the `income` variable.
- Check the `income` variable.

```
hist(dat$income)
```



Mean

- The mean is the average of a data set, i.e.,

$$\bar{X} = \frac{\sum X_i}{N}$$

```
mean(dat$income)
```

```
## [1] 1313.145
```

Median

- The median is the middle of the set of numbers.

```
median(dat$income)
```

```
## [1] 1304
```

Mode

- The mode is the most common number in a data set.
- R does not have a standard in-built function to compute mode. Thus, we can create a user function to compute mode of a data set in R.

```
getmode <- function(x) {
   uniqv <- unique(x)
   uniqv[which.max(tabulate(match(x, uniqv)))]
}

getmode(dat$income)
```

```
## [1] 1235
```

Quantiles

- Ordered observations $Y_{(1)}, Y_{(2)}, \ldots, Y_{(N)}$ are partitioned into q groups of equal size. The (observed) values which separate the q groups are called quantiles.
- **Quartiles**: $q = 4$ equally sized groups consisting of 25% of observations each
  - $Q_1 = Y_{.25}$: The first quartile is the smallest observation for which holds that 25% of all observations are smaller or equal to it.
  - $Q_2 = Y_{.50}$: The second quartile is the smallest observation for which holds that 50% of all observations are smaller or equal to it.
  - $Q_3 = Y_{.75}$: The third quartile is the smallest observation for which holds that 75% of all observations are smaller or equal to it.

Quantiles

The construction principle is similar for other quantiles, e.g.,

- **Quintiles**: $Y_{.2}, Y_{.4}, Y_{.6}$, and $Y_{.8}$ partition all observations into $q = 5$ equally sized groups consisting of 20% of observations each.
- **Deciles**: $Y_{.1}, Y_{.2}, \ldots, Y_{.8}, Y_{.9}$ partition all observations into $q = 10$ equally sized groups consisting of 10% of observations each.
- **Percentiles**: more generally, Y_p is the p percentile; Y_p is the smallest value for which holds that at least p of observations are smaller than or equal to Y_p.

```r
quantile(dat$income, probs=c(0.25, 0.5, 0.75))
```

```
##     25%     50%     75% 
## 1117.00 1304.00 1505.75
```

```r
quantile(dat$income, probs=0.11)
```

```
##    11% 
## 995.31
```

Standard Deviation and Variance

$$Var = \frac{\sum(X_i - \bar{X})^2}{N-1}$$

$$SD = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}}$$

```
var(dat$income)
```

```
## [1] 65312.13
```

```
sd(dat$income)
```

```
## [1] 255.5624
```

Range and Interquartile Range

```
range(dat$income)
```

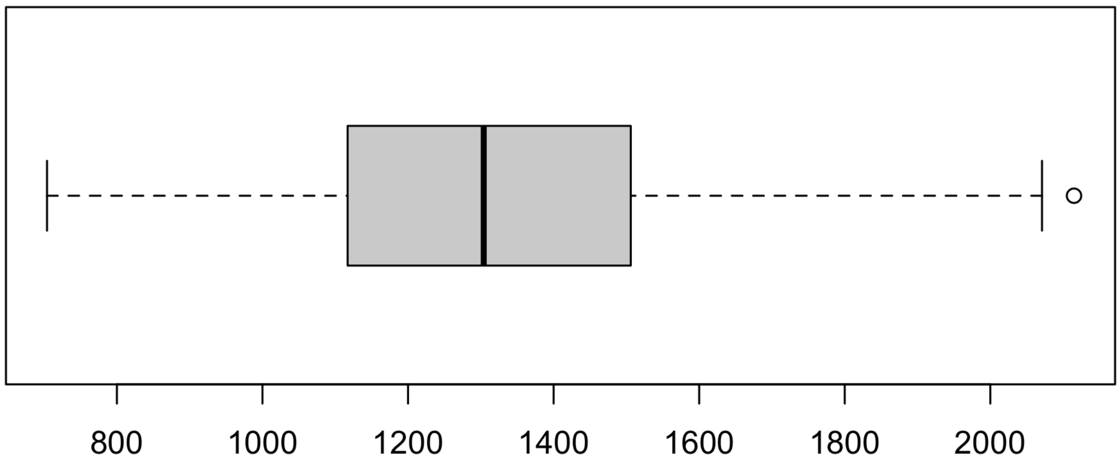
```
## [1]  704 2115
```

```
IQR(dat$income)
```

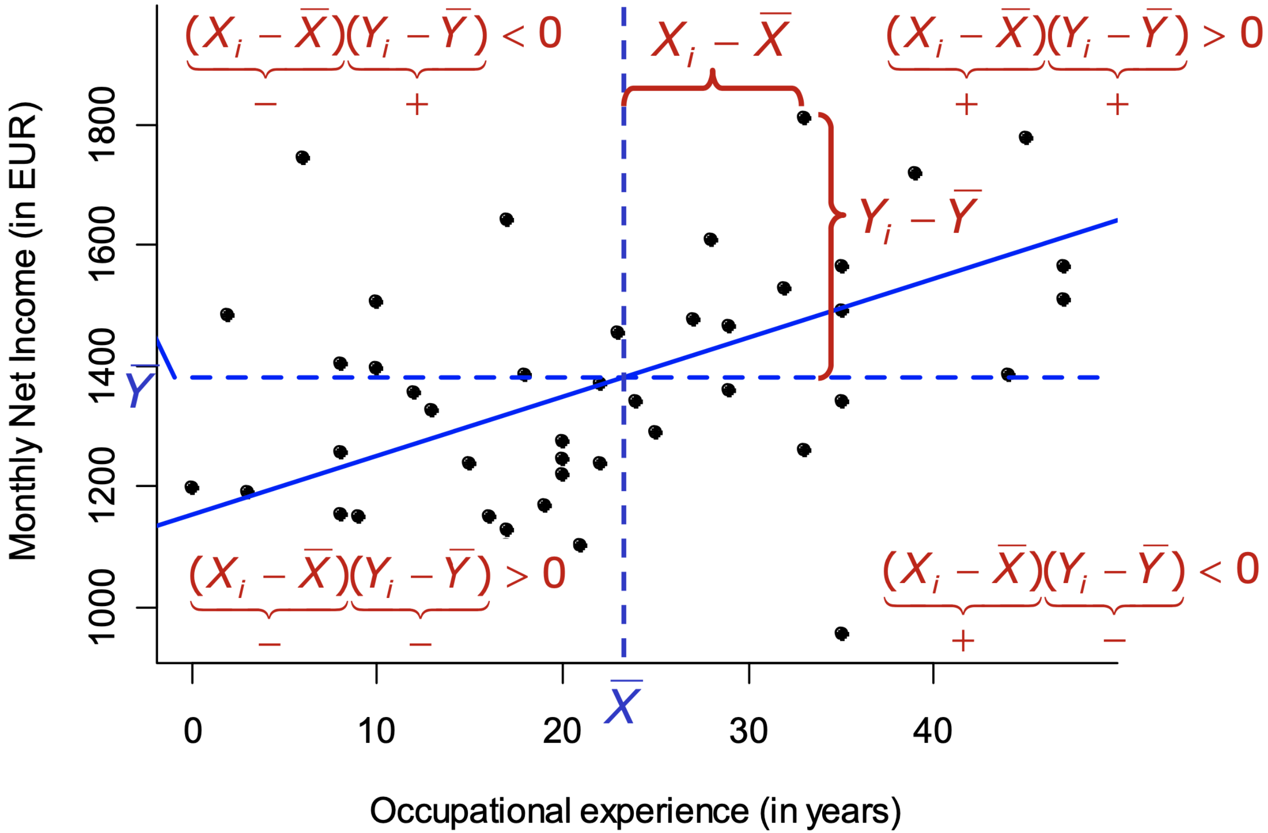
```
## [1] 388.75
```

Range and Interquartile Range: Boxplot

```
boxplot(dat$income, horizontal=TRUE)
```



Covariance



$(X_i - \bar{X})(Y_i - \bar{Y}) < 0$
$-$
$+$
$X_i - \bar{X}$
$(X_i - \bar{X})(Y_i - \bar{Y}) > 0$
$+$
$+$
$Y_i - \bar{Y}$
$\bar{Y}$
$(X_i - \bar{X})(Y_i - \bar{Y}) > 0$
$-$
$-$
$(X_i - \bar{X})(Y_i - \bar{Y}) < 0$
$+$
$-$
Monthly Net Income (in EUR)
1000
1200
1400
1600
1800
$\bar{X}$
0
10
20
30
40
Occupational experience (in years)

Covariance

- Covariance measures the co-variation of X and Y, i.e., to what extent and in which direction does Y co-vary with X?

$$S_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

Covariance

- Again, S_{XY} measures the co-variation of X and Y, i.e., to what extent and in which direction does Y co-vary with X?
    - $S_{XY} > 0$: positive relation (Y increases with increasing X); slope of the regression line is positive.
    - $S_{XY} < 0$: negative relation (Y decreases with increasing X); slope of the regression line is negative.
    - $S_{XY} = 0$: no relation (Y varies independent of X; X is independent of Y); slope of the regression line is zero.

Correlation

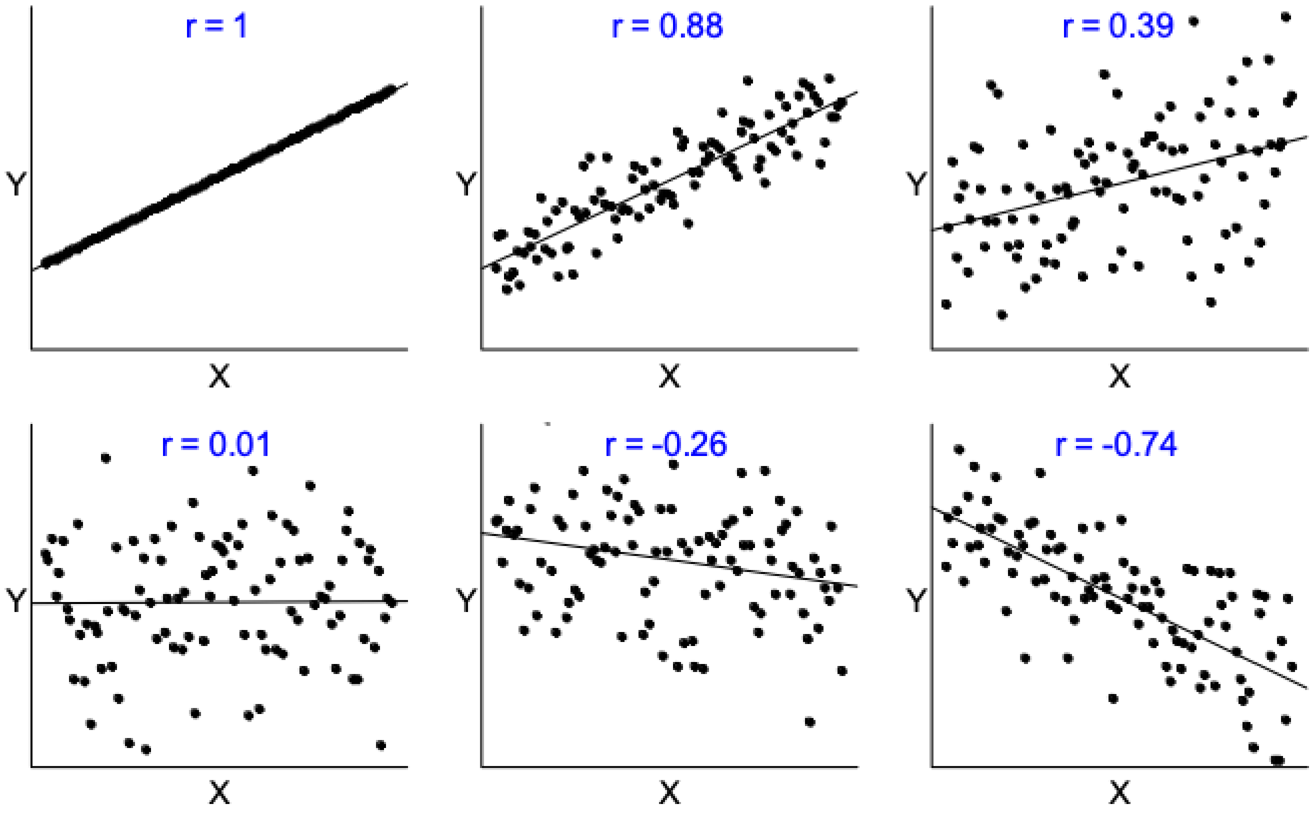
- The covariance depends on the units of measure-ment and is frequently not easy to interpret.
- However, we can use the covariance to construct a standardized measure that indicates the strength of the linear relationship between two continuous variables X and Y: the correlation coefficient r_{XY}.

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}}{\sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}}\sqrt{\frac{\sum(Y_i - \bar{Y})^2}{N-1}}}$$
$$= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

Correlation

- The correlation coefficient represents a standardized covariance and is always between –1 and +1.
- We use the correlation coefficient for assessing the strength of the linear relationship between X and Y:
  - $r_{XY} = 1$ indicates a perfect positive linear relationship.
  - $r_{XY} = -1$ indicates a perfect negative linear relationship.
  - $r_{XY} = 0$ indicates that there is no (linear) relationship.

Correlation



Correlation

- The correlation coefficient is appropriate for (almost) linear relationships. Whenever relations deviate from linearity the linear correlation is misleading.
- Examples where the simple correlation coefficient is not really informative: