

EPI 809

(Biostatistics II)

Introduction:

⇒ Syllabus

⇒ GitHub (Scripts and other materials)

⇒ D2L (INCLASS, HW, Grades)

Types of Random Variables

- ⇒ **Quantitative (usually continuous)**. Can take on an (conceptually) infinite number of values, and there is a notion of order and distance between values (e.g., body height)
- ⇒ **Categorical**; take on a finite number of values (e.g., presence/absence of disease).
- ⇒ Categorical RVs can be divided into various subtypes (binary, ordinal, multinomial)

Example

CPS5 data set (Goldberger 1998, adapted from Berndt 1991)

<http://www.hup.harvard.edu/features/golint/CPS5.txt>.

⇒ This data comprise information from 528 people surveyed in 1985.

⇒ The variables included in the data set are:

- years of education (integer, but can be analyzed as continuous)
- years of experience in the labor market (integer, but can be analyzed as continuous)
- wage USD/hour (continuous)
- Sex (Male/Female, binary)
- Region (South/non-South, multinomial)
- Marital status (married/ not married)

⇒ We illustrate regression analysis to quantify effects of education on wages, after accounting for differences due to sex, and region.

The first rows of the CPS5 data set...

| education | south | ehtnicGroup | female | married | experience | unionized | hourlyWage |
|-----------|-------|-------------|--------|---------|------------|-----------|------------|
| 10 | 0 | White | 0 | 1 | 27 | 0 | 9 |
| 12 | 0 | White | 0 | 1 | 20 | 0 | 5.5 |
| 12 | 0 | White | 1 | 0 | 4 | 0 | 3.8 |
| 12 | 0 | White | 1 | 1 | 29 | 0 | 10.5 |
| 12 | 0 | White | 0 | 1 | 40 | 1 | 15 |
| 16 | 0 | White | 1 | 1 | 27 | 0 | 9 |
| 12 | 0 | White | 1 | 1 | 5 | 1 | 9.57 |
| 14 | 0 | White | 0 | 0 | 22 | 0 | 15 |
| 8 | 0 | White | 0 | 1 | 42 | 0 | 11 |

Quantitative

- Education
- Experience and
- Hourly-wage

Multinomial:

- Ethnicity

Binary:

- Region (south=1)
- Sex (female=1)
- Married (yes=1)
- Unionizes (yes=1)

I-Loading the data in R from an text-file

```
DATA=read.table(file='~/Dropbox/STATCOMP/2018/wage.txt',header=T)
```

II-Checking and inspecting the data set

```
dim(DATA)  # returns number of rows and columns  
str(DATA)  # describe the data type and structure  
head(DATA) # shows the first rows  
tail(DATA) # same for last rows  
fix(DATA)  # shows data in an spreadsheet format  
summary(DATA) # summaries by variable
```

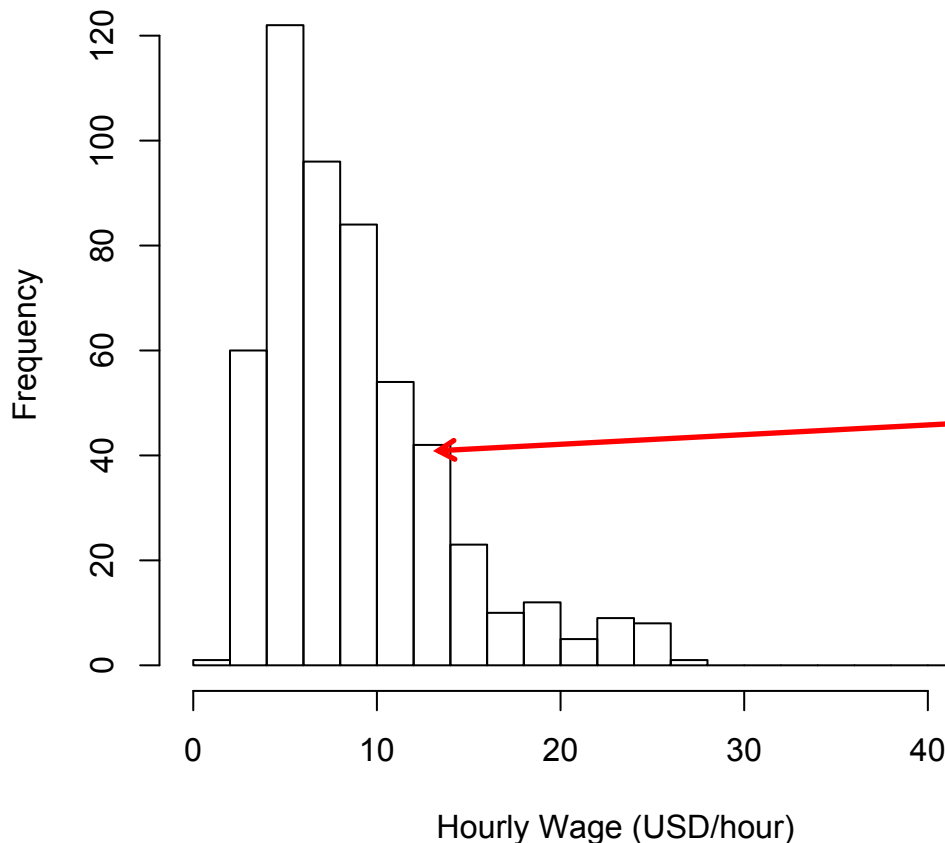
III-Univariate & Bi-variate descriptive analysis

Hourly Wages

```
hist(DATA$Wage, 30)  
summary(DATA$Wage)
```

```
> summary(DATA$Wage)  
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
  1.750   5.250   7.790   9.048  11.250  44.500
```

Histogram



The variable on the x-axis is divided into bins of equal width.

Here, each bin has \$2 increments.

The height of each bar reflects the number of subjects falling in each bin.

Here, 40 people make \$12-14 USD/hr

Outlier?

The distribution appears to be skewed

Univariate Descriptive Statistics & Graphs: Continuous RVs

```
table(DATA$Education)
```

```
> table(Education)
```

```
Education
```

| | | | | | | | | | | | | |
|---|---|----|----|----|----|-----|----|----|----|----|----|----|
| 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 3 | 5 | 15 | 12 | 17 | 27 | 218 | 37 | 56 | 13 | 70 | 24 | 31 |

```
>
```

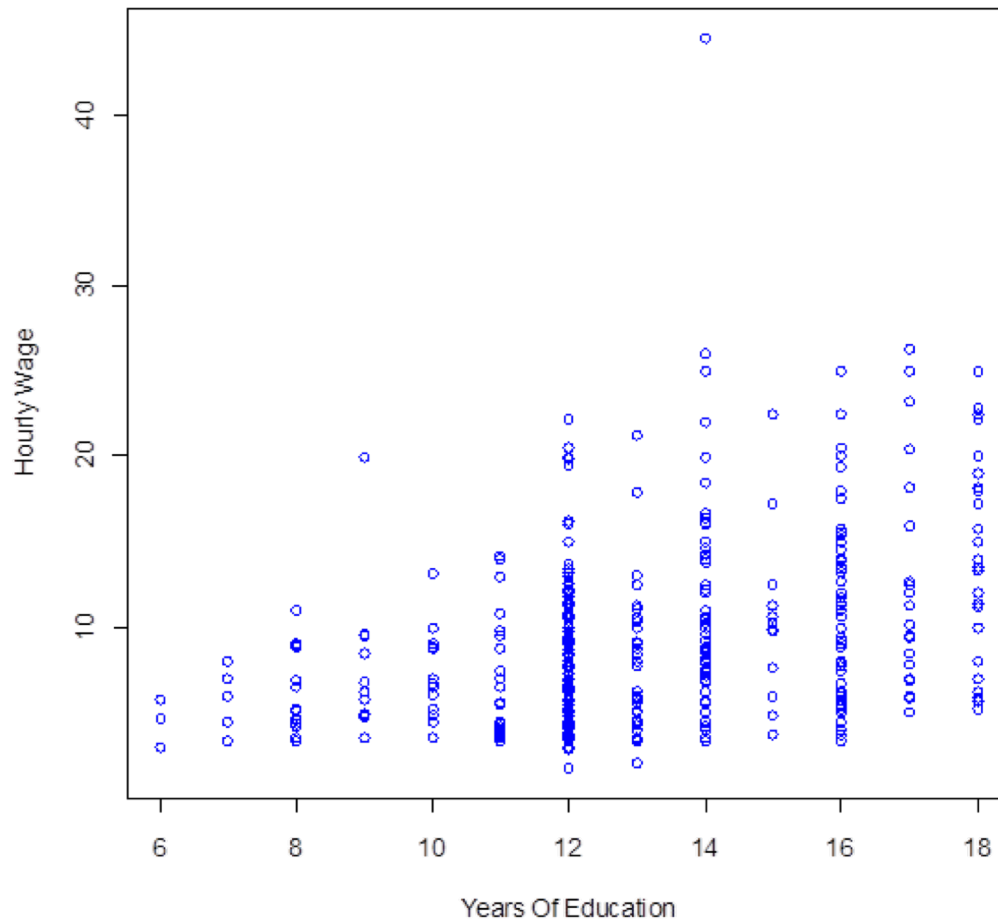
```
> summary(DATA$Education)
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|-------|---------|-------|
| 6.00 | 12.00 | 12.00 | 13.09 | 15.00 | 18.00 |

Bi-variate analysis

Bi-variate Descriptive Statistics (Quantitative variables)

```
> plot(Wage~Education, col='blue',data=DATA)
```

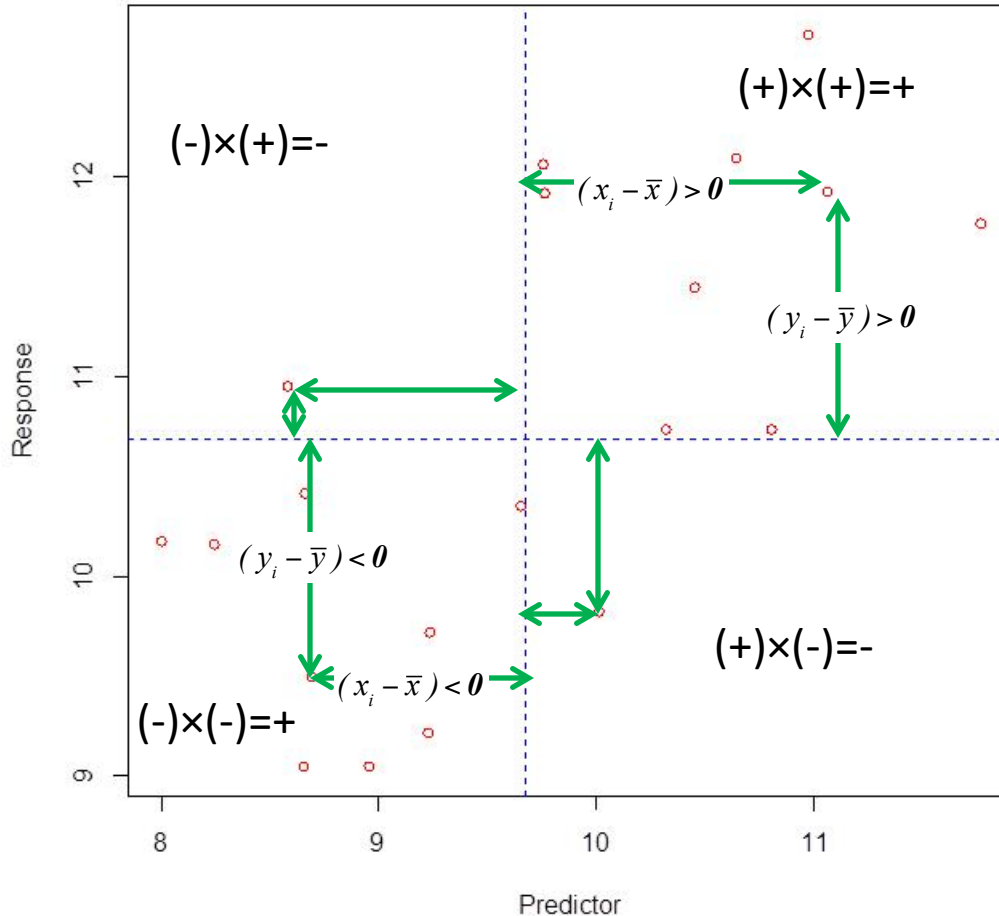


Key Concepts we will discuss

- Variance & Co-variance
- Correlation

Quantifying Linear Dependence Between Two Continuous Variables

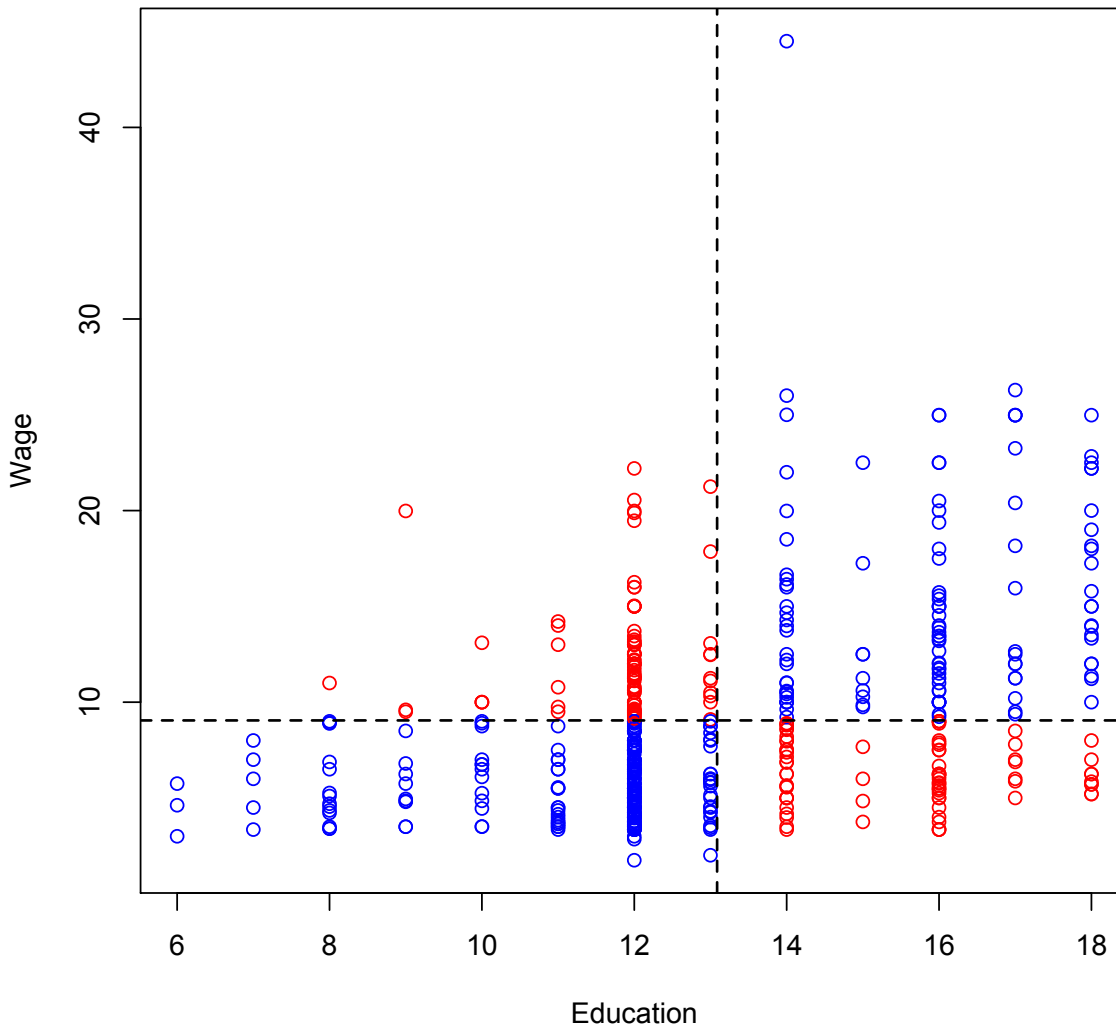
Quantifying Co-variability (co-variance)



| Observation | x_i | y_i | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|-------------------|--------|--------|-------------------|-------------------|----------------------------------|
| 1 | 10.65 | 12.09 | 0.97 | 1.40 | 1.37 |
| 2 | 11.76 | 11.77 | 2.09 | 1.08 | 2.26 |
| 3 | 10.80 | 10.73 | 1.13 | 0.04 | 0.05 |
| 4 | 9.23 | 9.22 | -0.44 | -1.47 | 0.65 |
| 5 | 10.46 | 11.45 | 0.78 | 0.76 | 0.59 |
| 6 | 9.23 | 9.72 | -0.44 | -0.97 | 0.43 |
| 7 | 8.66 | 10.42 | -1.01 | -0.27 | 0.28 |
| 8 | 10.97 | 12.72 | 1.30 | 2.02 | 2.63 |
| 9 | 9.76 | 11.92 | 0.09 | 1.23 | 0.11 |
| 10 | 8.66 | 9.05 | -1.02 | -1.64 | 1.67 |
| 11 | 11.06 | 11.93 | 1.39 | 1.24 | 1.72 |
| 12 | 8.25 | 10.16 | -1.43 | -0.53 | 0.76 |
| 13 | 8.69 | 9.50 | -0.98 | -1.19 | 1.18 |
| 14 | 10.01 | 9.82 | 0.34 | -0.87 | -0.29 |
| 15 | 10.32 | 10.74 | 0.65 | 0.05 | 0.03 |
| 16 | 8.58 | 10.95 | -1.09 | 0.26 | -0.28 |
| 17 | 9.65 | 10.35 | -0.02 | -0.34 | 0.01 |
| 18 | 8.95 | 9.05 | -0.72 | -1.64 | 1.18 |
| 19 | 8.00 | 10.18 | -1.67 | -0.52 | 0.86 |
| 20 | 9.76 | 12.06 | 0.08 | 1.37 | 0.11 |
| Sum | 193.45 | 213.82 | 0.00 | 0.00 | 15.31 |
| Average | 9.67 | 10.69 | 0.00 | 0.00 | 0.765 |
| Sample Covariance | | | | | 0.806 |

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Quantifying Co-variability (co-variance)



```
x=DATA$Education
```

```
y=DATA$Wage
```

```
meanX=mean(x)
```

```
meanY=mean(y)
```

```
dx=(x-meanX)
```

```
dy=(y-meanY)
```

```
signColor=ifelse(sign(dx)*sign(dy)==1,  
                 'blue','red')
```

```
plot(Wage~Education,data=DATA,  
     col=signColor);
```

```
abline(h=meanY,lwd=1.5,lty=2)
```

```
abline(v=meanX,lwd=1.5,lty=2)
```

Variance & Covariance

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Sample co-variance: Measures Association between Y and X.

$$\begin{aligned} S_{XX} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

Sample variance: Measures Variability.

Sample variance is simply the sample covariance of a variable with itself.

Covariance is location invariant but scale dependent

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Location Invariant: adding a constant to any of the variables does not change the sample co-variance (why? Because S_{xy} is based on deviations from the sample mean)

Scale dependent: multiplying any of the variables by a constant will alter the S_{xy} .

Proof: let $Z = a + b \times Y$

$$\begin{aligned} S_{XZ} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \left(a + by_i - \frac{1}{n} \sum_{j=1}^n a + by_j \right) \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \left(a + by_i - \frac{n}{n} a - \frac{1}{n} \sum_{j=1}^n by_j \right) \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) (by_i - b\bar{y}) \\ &= b \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y}) \\ &= b S_{XY} \end{aligned}$$

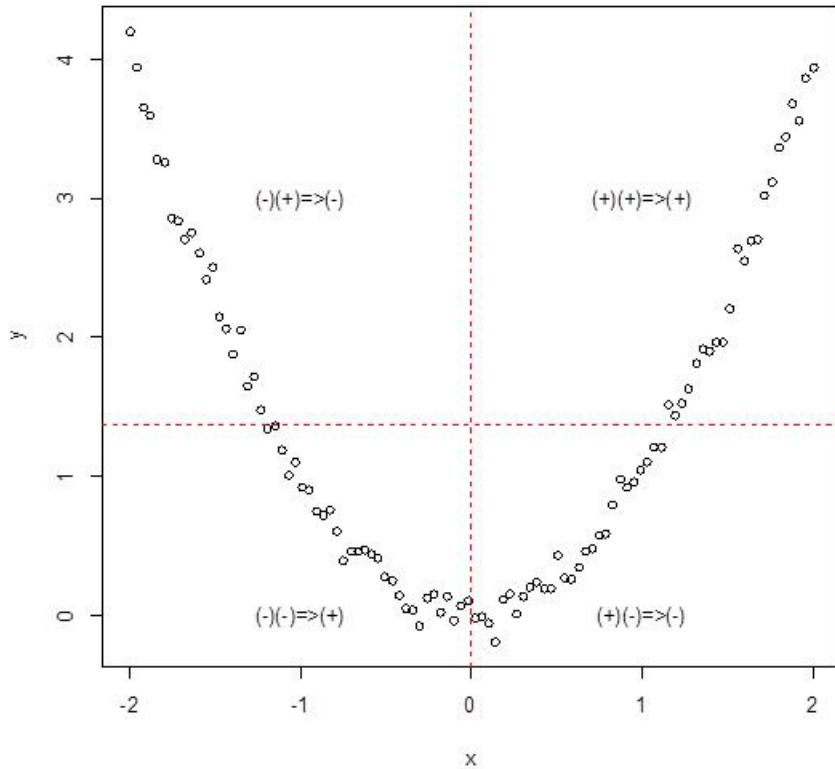
Co-variance is location invariant but scaled-dependent.

Summary: Sample Covariance

- ⇒ Measures patterns of (linear) association between two variables
- ⇒ S_{xy} is computed using cross-products of deviations from the sample mean.
- ⇒ Therefore, it is location invariant (adding a constant to either x or y does not change covariance).
- ⇒ However, it is scale dependent.
- ⇒ Next we will discuss a standardized version of the co-variance that is scale invariant: Pearson's Product Moment Correlation Coefficient.

Covariance Versus Independence

Covariance Vs Independence



- \Rightarrow Absence of covariance does not imply independence.
- \Rightarrow Two random variables are independent if the joint distribution $p(X,Y)$ is equal to the product of the marginal distributions, that is $p(X,Y)=p(X)p(Y)$.
- \Rightarrow Independence implies that the means, variances (and all the moments) of one of the RV does not depend on the other.
- \Rightarrow Absence of covariance implies absence of dependence of the mean $E[Y|X]$ only in a linear sense.

Pearson's Correlation Coefficient

Correlation: a scale-invariant measure of linear association.

$$R_{yx} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

$$-1 \leq R_{yx} \leq 1$$

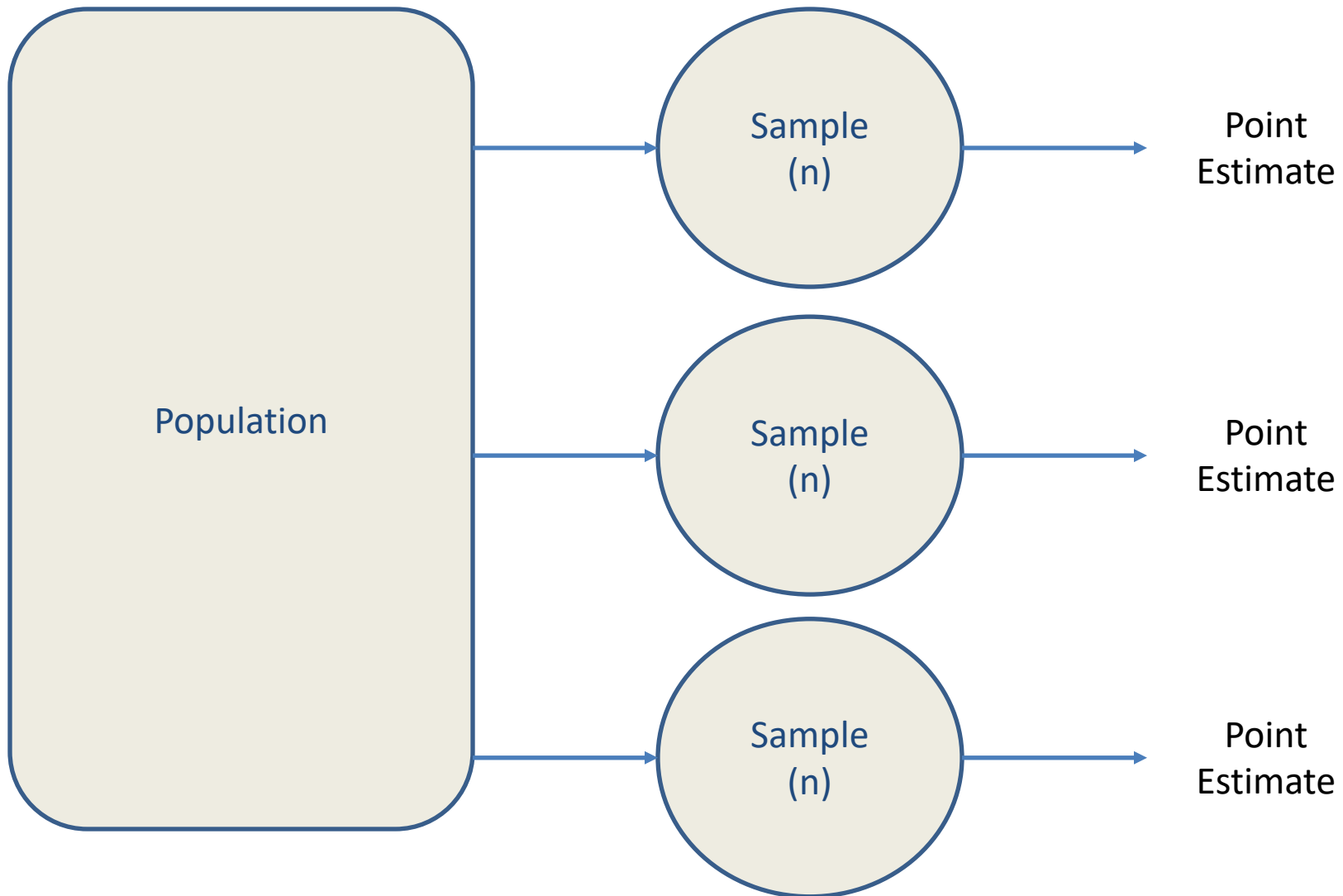
Pearson's product-moment correlation:

- Measures linear association
- Absence of correlation does not mean independence.
- It is scale and location invariant
- COR=0 indicates absence of linear association,
- Positive correlation means that as X increases so does Y,
- Negative correlation means that as X increases Y decreases
- COR = +/- 1 implies perfect linear dependence between X and Y.

Inference on the correlation coefficient

Statistical Inference

(& Conceptual Repeated Sampling from the Population)



Statistical Inference

(& Conceptual Repeated Sampling from the Population)

- ⇒ θ , a population parameter of unknown value (e.g., correlation coefficient)
- ⇒ $\hat{\theta}$, a point-estimate derived from a finite sample (e.g., the correlation of two variables in a sample)
- ⇒ $\hat{\theta}$ informs us about θ , but we have uncertainty about θ because $\hat{\theta}$ was derived from a finite sample.
- ⇒ How do we quantify such uncertainty?
- ⇒ We often use Standard Errors (SE), which represent the standard deviation of the estimator over conceptual repeated sampling.
- ⇒ We can also build confidence intervals, intervals that over conceptual repeated sampling will contain the true population parameter x% of the time (x=confidence)
- ⇒ How do we estimate SEs and CIs?

Inference on the correlation coefficient

Standard Error:

$$SE(r) = \sqrt{\frac{1 - r^2}{n - 3}}$$

t-stat:

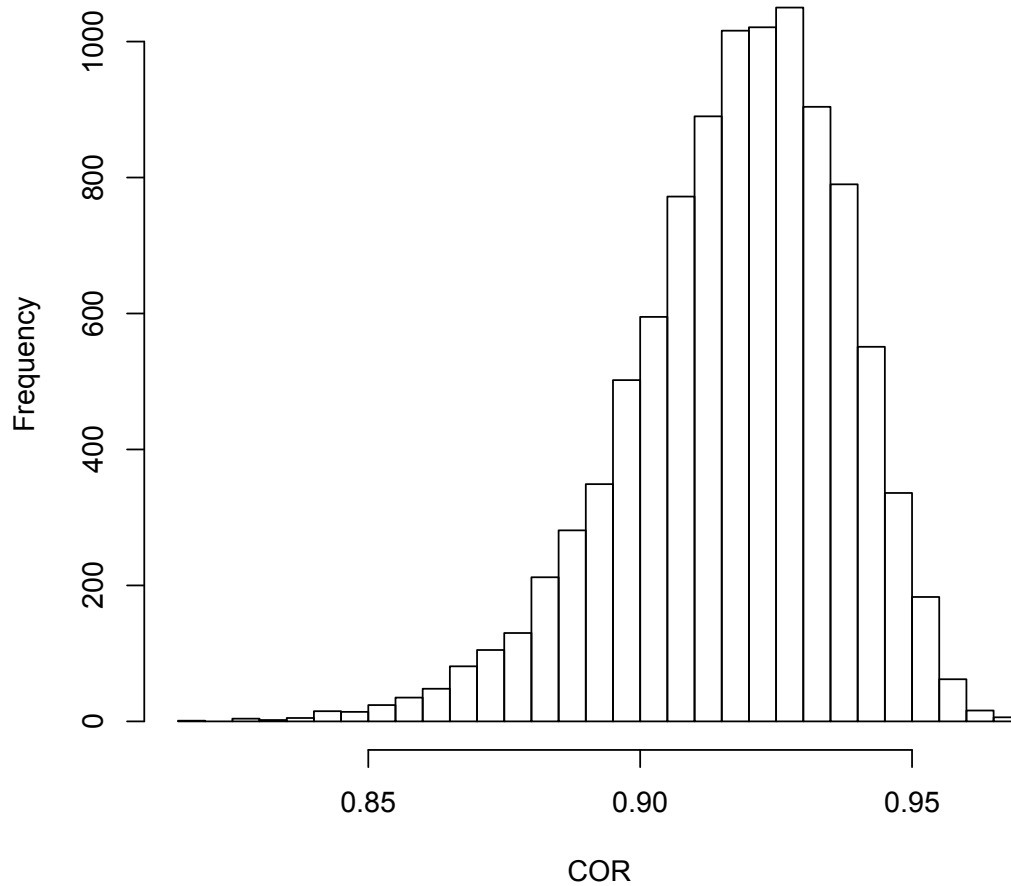
$$tstat = \frac{r}{\sqrt{\frac{1 - r^2}{n - 3}}}$$

For n reasonably large (e.g., >50) and intermediate coefficient, we can safely assume that r follows a normal distribution.

However, when n is small and/or the correlation coefficient is close to either 1 or -1, the sampling distribution is not symmetric.

Distribution of the correlation coefficient

Histogram of COR



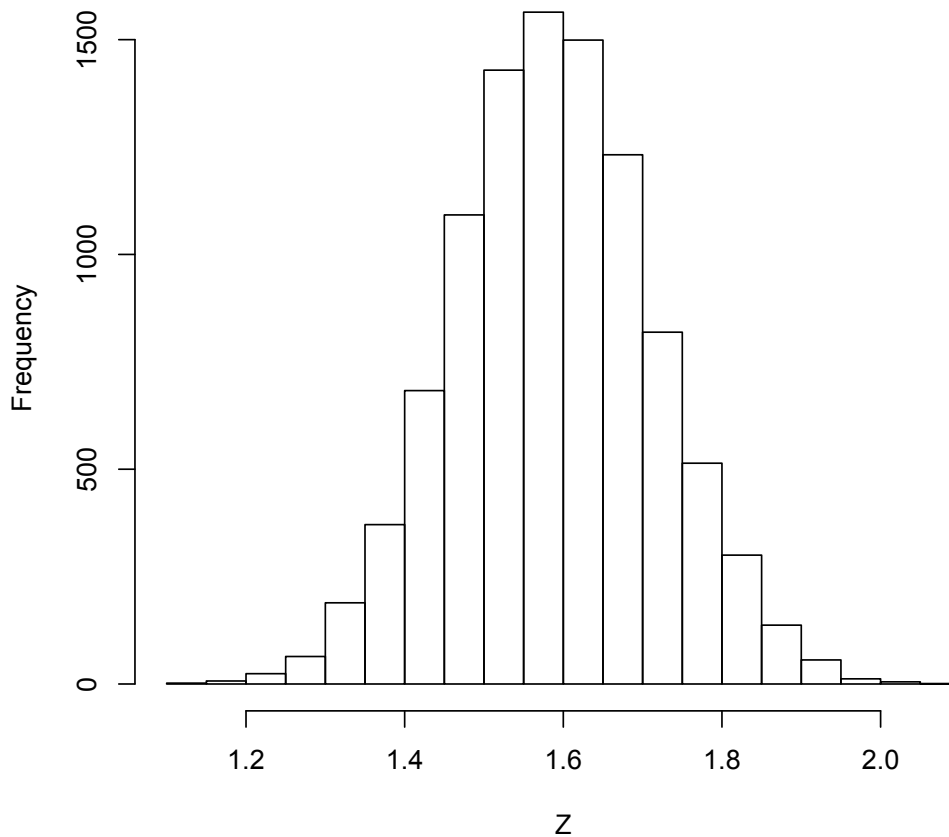
```
# Monte Carlo
nRep=10000
COR=rep(NA,nRep)
for(i in 1:10000){
  x=runif(n)
  y=x+rnorm(n)/8
  COR[i]=cor(x,y)
}

hist(COR,30)
```

Fisher's Z-transform

$$Z = 0.5 \times \log\left(\frac{1+r}{1-r}\right)$$

Histogram of Z



```
# Monte Carlo
nRep=10000
COR=rep(NA,nRep)
for(i in 1:10000){
  x=runif(n)
  y=x+rnorm(n)/8
  COR[i]=cor(x,y)
}
Z=0.5*log((1+COR)/(1-COR))
hist(Z,30)
```

Fisher's Z-transform

$$Z = 0.5 \times \log \left(\frac{1+r}{1-r} \right) \sim N \left(0, \frac{1}{n-3} \right)$$

We can compute a CI for Z, then map-back the interval into r .

$$Z = 0.5 \times \log \left(\frac{1+r}{1-r} \right)$$

$$r = \frac{e^{2Z} - 1}{1 + e^{2Z}}$$

Github:

<https://github.com/gdlc/EPI809/blob/master/CORREATION.md>