PLSC 476: Empirical Legal Studies

Christopher Zorn

January 26, 2021

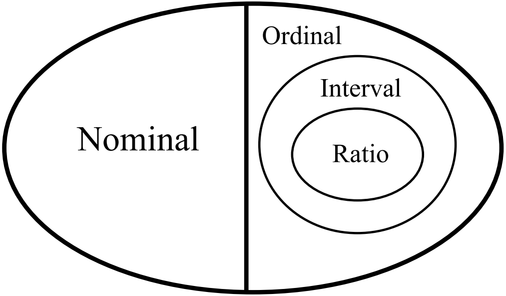
Logistics

Details:

- Syllabus is on the Github repository (https://github.com/PrisonRodeo/PLSC476-SP2021-git)
- Three broad course "themes":
  - Introduction / review software, statistics, etc.
  - Empirical work on courts and judges
  - Empirical analysis of (and in) the practice of law
- Research modules (4 @ 15% each):
  - Module #1 will be "common" (assigned the end of this week)
  - Modules #2-4 will be your choice
  - More details will be posted soon

Levels of Measurement

- Nominal (classification)
- Ordinal (order)
- Interval (equal intervals)
- Ratio ("true zero")



Variables: Discrete vs. Continuous

Examples of Variables, by Type and Level of Measurement

Level of Measurement	Discrete	Continuous
Nominal	$\{Blonde, Brunette, Redhead\}$	n/a
Ordinal	Social Class (Upper, middle, lower)	n/a
Interval	Year	Temperature (in degrees F)
Ratio	Counts of things	Height, weight, distance, etc.

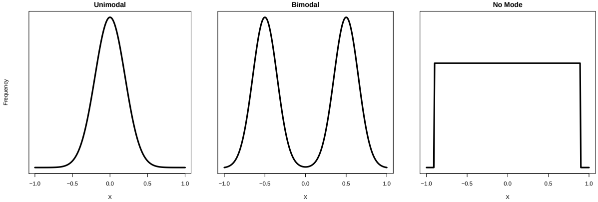
Arithmetic Mean (minimizes squared deviations):

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Median (minimizes absolute deviations):

$$\begin{aligned}\check{X} &= \text{“middle observation” of } X \\ &= \text{50th percentile of } X.\end{aligned}$$

Mode (most frequently-occurring value):



Variation: Range and Percentiles

Range:

$$\text{Range}(X) = \max(X) - \min(X)$$

The k**th percentile** is the value of the variable below which k percent of the observations fall

- 50th percentile $= \check{X}$
- 0th percentile $=$ minimum(X)
- 100th percentile $=$ maximum(X)

Variance and S.D.

Variance:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Standard deviation:

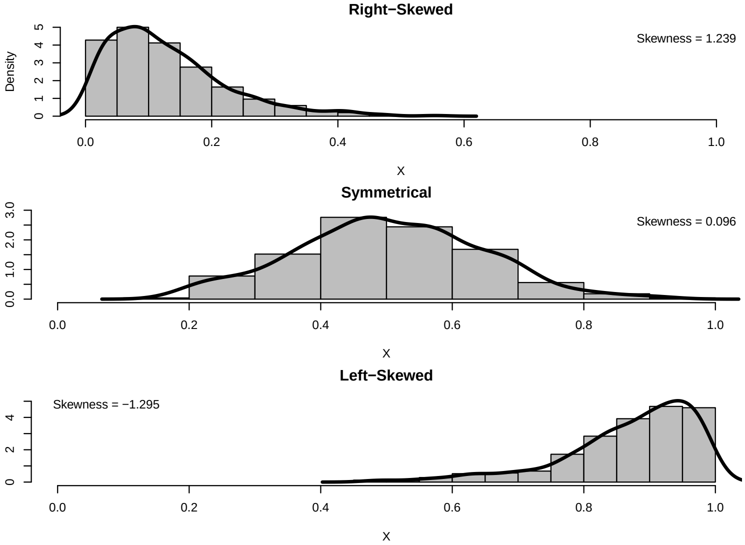
$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Typically:

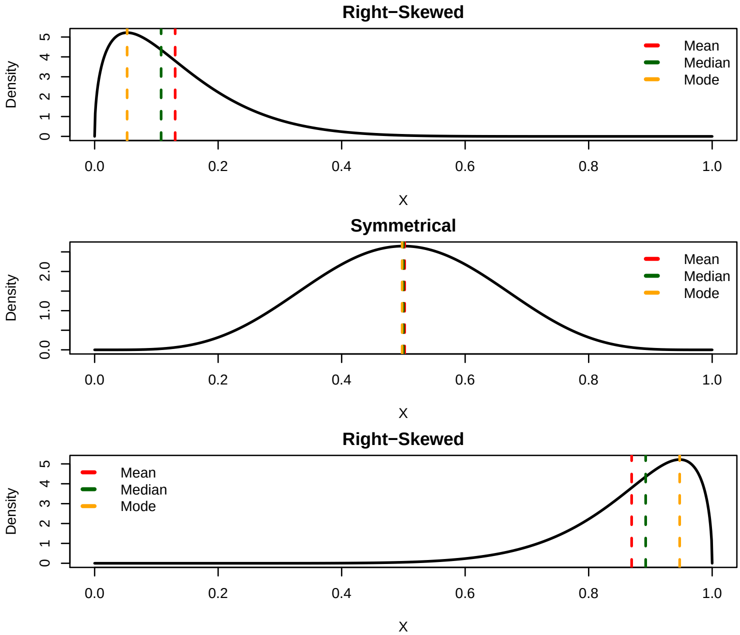
$$\begin{aligned}\mu_3 &= \frac{M_3^2}{\sigma^3} \\ &= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right]^{3/2}}\end{aligned}$$

- Skewness $= 0 \rightarrow$ symmetrical
- Skewness $> 0 \rightarrow$ "positive" (tail to the right)
- Skewness $< 0 \rightarrow$ "negative" (tail to the left)

Skewness Illustrated



Means, Medians, Modes, and Skewness



Dichotomous / “Binary” Variables

Defined as:

$$D \in \{0, 1\}$$

Central Tendency:

$$\begin{aligned} \text{Mean } \bar{D} &= \widehat{\Pr(D = 1)} \\ \text{Median} &= \text{Mode} \end{aligned}$$

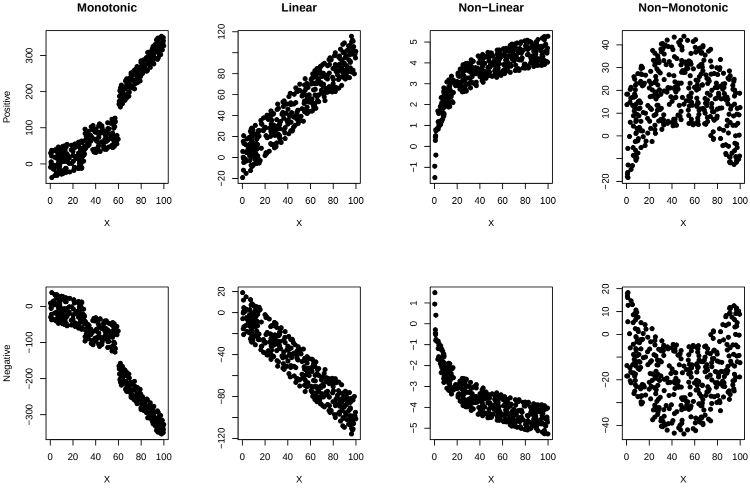
Variance:

$$\sigma_D^2 = \bar{D} \times (1 - \bar{D})$$

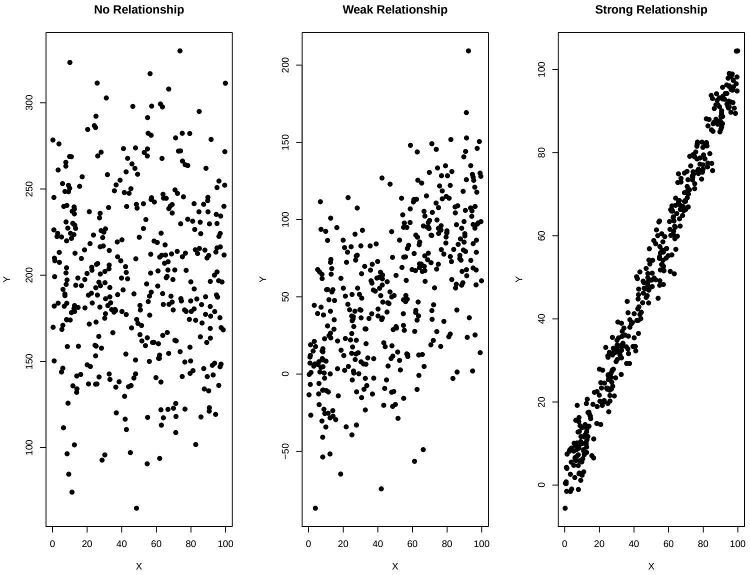
and so SD:

$$\sigma_D = \sqrt{\bar{D} \times (1 - \bar{D})}$$

Types of Relationships



Strength of Relationships



Tabular Methods “Crosstabs”

- Requires *nominal*- or *ordinal*-level data...
- Rows / columns denote categories (or intervals) of Y and X respectively
- Cell entries indicate frequencies of observations that meet both conditions...
- Levels of Measurement:
  - Nominal categories = no indication of “direction”
  - Ordinal categories should appear in order
  - Continuous variables require “binning”...
  - Are related to statistics (e.g., χ^2)

Statistical Measures of Association

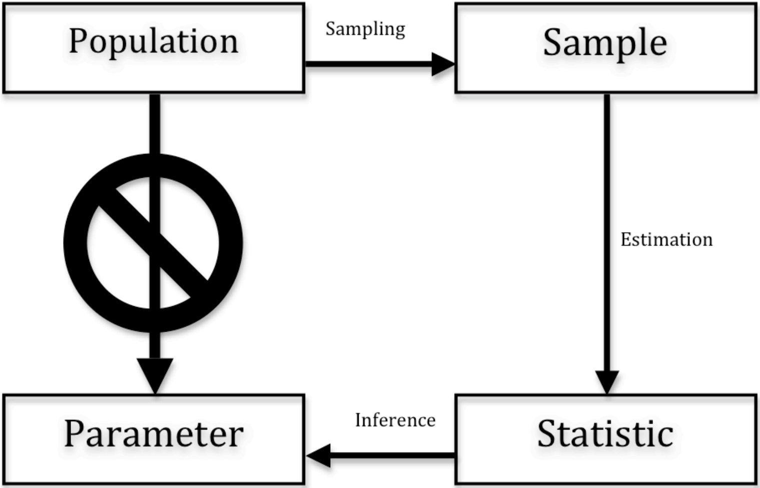
The general idea:

- If two variables X and Y are unrelated, then we should see an "even" distribution of cases on each, irrespective of the values of the other
- If we observe something other than such an "even" distribution, then the variables are not unrelated
- Formally: No association means $f(Y|X) = f(Y)$

Measures of Association, by Levels of Measurement

		X			
		Nominal	Binary	Ordinal	Interval/Ratio
Y	Nominal	χ^2	χ^2	χ^2	t-test (and η)
	Binary	χ^2	ϕ, Q	γ, τ_c	t-test
	Ordinal	χ^2	γ, τ_c	γ, τ_a, τ_b	Spearman's ρ
	Interval / Ratio	t-test (and η)	t-test	Spearman's ρ	r (+ regression)

Statistical Inference



Moving parts:

- A *null hypothesis*, usually denoted H_0
- an *alternative* (or *research*) *hypothesis* H_a or H_1
- a *test statistic* $\theta =$ f(sample data $\mathbf{X}$)
- a *rejection region* for the null in the space of the sample statistic

Type I and Type II Errors:

- **Type I error**: rejecting a *true* null hypothesis (think of this as a "false positive")
- **Type II error**: failing to reject a *false* null hypothesis (think of this as a "false negative")

	Reality / Population	
Test Statistic / Sample	H_a	H_0
H_a	Correct	Type I error
H_0	Type II Error	Correct

Example: English Premier League (EPL) Table

```
> print(EPL)
                  Team Rank Points Matches Win Draw Loss Goals GoalsAgainst GoalDifference
     Manchester United    1     40      19  12    4    3    36           25             11
       Manchester City    2     38      18  11    5    2    31           13             18
        Leicester City    3     38      19  12    2    5    35           21             14
             Liverpool    4     34      19   9    7    3    37           22             15
     Tottenham Hotspur    5     33      18   9    6    3    33           17             16
               Everton    6     32      17  10    2    5    28           21              7
       West Ham United    7     32      19   9    5    5    27           22              5
           Aston Villa    8     29      17   9    2    6    31           18             13
               Chelsea    9     29      19   8    5    6    33           23             10
           Southampton   10     29      18   8    5    5    26           21              5
               Arsenal   11     27      19   8    3    8    23           19              4
          Leeds United   12     23      18   7    2    9    30           34             -4
        Crystal Palace   13     23      19   6    5    8    22           33            -11
         Wolverhampton   14     22      19   6    4    9    21           29             -8
               Burnley   15     19      18   5    4    9    10           22            -12
      Newcastle United   16     19      19   5    4   10    18           32            -14
 Brighton & Hove Albion   17     17      19   3    8    8    22           29             -7
                Fulham   18     12      18   2    6   10    15           27            -12
    West Bromwich Albion   19     11      19   2    5   12    15           43            -28
      Sheffield United   20      5      19   1    2   16    10           32            -22
```

EPL Data Summary

```
> summary(EPL)
     Team                Rank           Points         Matches          Win       
 Length:20          Min.   : 1.00   Min.   : 5.00   Min.   :17.0   Min.   : 1.0  
 Class :character   1st Qu.: 5.75   1st Qu.:19.00   1st Qu.:18.0   1st Qu.: 5.0  
 Mode  :character   Median :10.50   Median :28.00   Median :19.0   Median : 8.0  
                    Mean   :10.50   Mean   :25.60   Mean   :18.5   Mean   : 7.1  
                    3rd Qu.:15.25   3rd Qu.:32.25   3rd Qu.:19.0   3rd Qu.: 9.0  
                    Max.   :20.00   Max.   :40.00   Max.   :19.0   Max.   :12.0  

      Draw            Loss           Goals        GoalsAgainst   GoalDifference 
 Min.   :2.00   Min.   : 2.0   Min.   :10.00   Min.   :13.00   Min.   :-28.00  
 1st Qu.:2.75   1st Qu.: 5.0   1st Qu.:20.25   1st Qu.:21.00   1st Qu.:-11.25  
 Median :4.50   Median : 7.0   Median :26.50   Median :22.50   Median :  4.50  
 Mean   :4.30   Mean   : 7.1   Mean   :25.15   Mean   :25.15   Mean   :  0.00  
 3rd Qu.:5.00   3rd Qu.: 9.0   3rd Qu.:31.50   3rd Qu.:29.75   3rd Qu.: 11.50  
 Max.   :8.00   Max.   :16.0   Max.   :37.00   Max.   :43.00   Max.   : 18.00  
```

Alternative Summary

```
> describe(EPL)
               vars  n  mean    sd median trimmed   mad min  max range  skew kurtosis   se
Team*             1 20   NaN    NA     NA     NaN    NA Inf -Inf  -Inf    NA       NA   NA
Rank              2 20 10.50  5.92   10.5   10.50  7.41   1   20    19  0.00    -1.38 1.32
Points            3 20 25.60  9.65   28.0   26.12  8.90   5   40    35 -0.40    -0.86 2.16
Matches           4 20 18.50  0.69   19.0   18.62  0.00  17   19     2 -0.92    -0.49 0.15
Win               5 20  7.10  3.29    8.0    7.19  2.97   1   12    11 -0.33    -1.05 0.74
Draw              6 20  4.30  1.75    4.5    4.19  1.48   2    8     6  0.18    -0.86 0.39
Loss              7 20  7.10  3.48    7.0    6.81  2.97   2   16    14  0.61    -0.05 0.78
Goals             8 20 25.15  8.40   26.5   25.62  8.90  10   37    27 -0.35    -1.16 1.88
GoalsAgainst      9 20 25.15  7.16   22.5   24.75  6.67  13   43    30  0.60    -0.19 1.60
GoalDifference   10 20  0.00 13.59    4.5    1.00 16.31 -28   18    46 -0.39    -1.14 3.04
```

In the EPL,

- wins are worth three points,
- draws are worth one point, and
- losses are worth zero points.

If (on average) teams are "balanced," then each team can expect to score

$$\frac{\{(0.5 \times 1) + [(0.25 \times 3) + (0.25 \times 0)]\}}{2} = 1.25$$

points per game. Do they?

Hypothesis test for $\overline{PPG} = 1.25$:

```
> EPL$PPG <- EPL$Points / EPL$Matches
> describe(EPL$PPG)
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 20 1.39 0.53   1.47    1.42 0.57 0.26 2.11  1.85 -0.42    -0.94 0.12


> t.test(EPL$PPG, mu=1.25)

One Sample t-test

data:  EPL$PPG
t = 1.1733, df = 19, p-value = 0.2552
alternative hypothesis: true mean is not equal to 1.25
95 percent confidence interval:
 1.141219 1.636318
sample estimates:
mean of x
 1.388768
```

Hypothesis Testing: Differences Of Means

Q: Do London-area teams score more points than those elsewhere?

Hypothesis test for $\overline{PPG}_{\text{London}} = \overline{PPG}_{\text{Non-London}}$:

```
> LACs<-c("Tottenham Hotspur","West Ham United","Chelsea",
          "Crystal Palace","Fulham","Arsenal")
> EPL$London<-ifelse((EPL$Team %in% LACs==TRUE),1,0)
> table(EPL$London)

 0  1
14  6

> t.test(PPG~London,data=EPL)

Welch Two Sample t-test

data:  PPG by London
t = -0.0098105, df = 13.439, p-value = 0.9923
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4984103  0.4938892
sample estimates:
mean in group 0 mean in group 1
       1.388090        1.390351
```

Q: Do teams that score a lot of goals also allow a lot of goals?

Examine the association between `Goals` and `GoalsAgainst`:

```
> with(EPL, cor(Goals,GoalsAgainst))
[1] -0.5218317
```

```
> with(EPL, cor.test(Goals,GoalsAgainst))

Pearson's product-moment correlation

data:  Goals and GoalsAgainst
t = -2.5953, df = 18, p-value = 0.01828
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7834396 -0.1031246
sample estimates:
       cor
-0.5218317
```

Next time: Data Visualization