PLSC 476: Empirical Legal Studies

Christopher Zorn

September 2, 2025

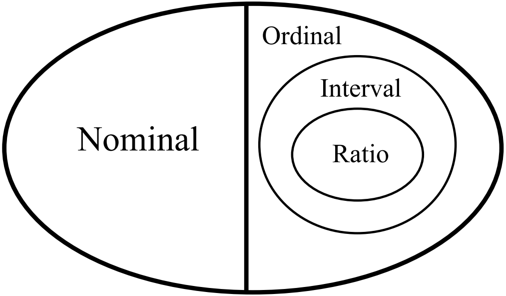
Logistics

Details:

- Syllabus is on the Github repository (`https://github.com/PrisonRodeo/PLSC476-FA2025-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.

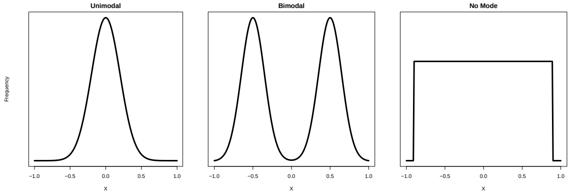
Arithmetic Mean (minimizes squared deviations):

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

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:

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

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:

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

Standard deviation:

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

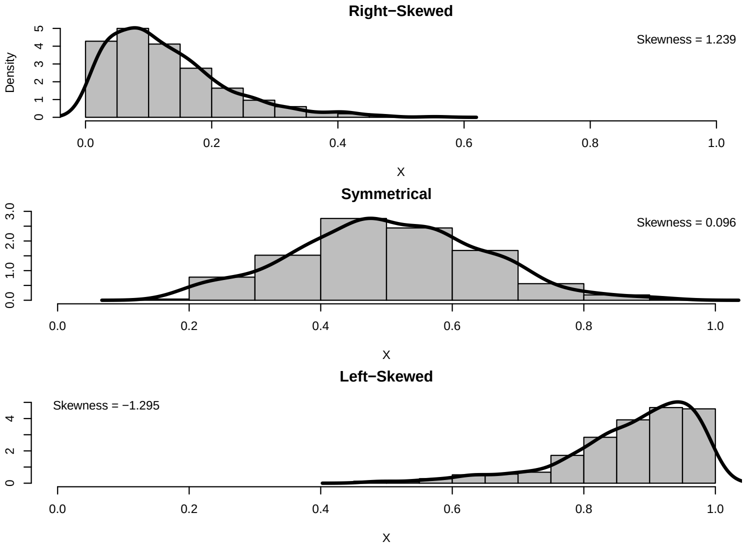
Skewness

Typically:

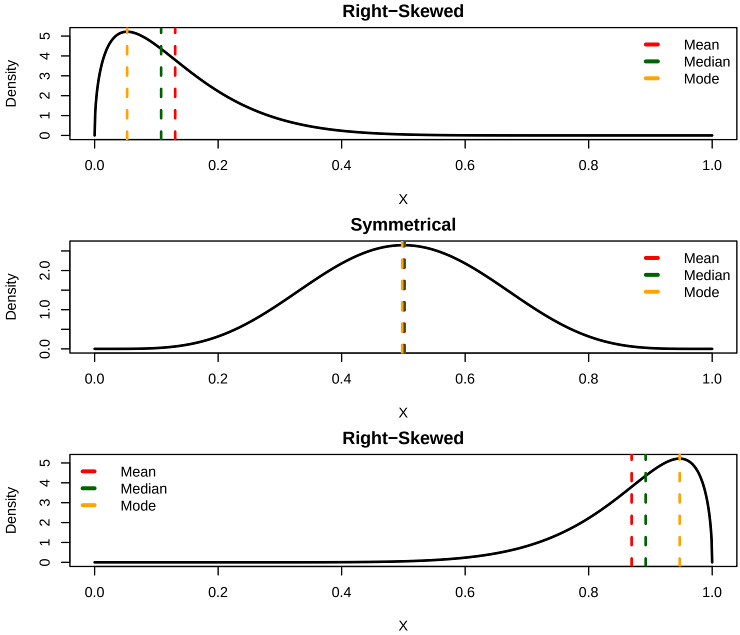
$$\mu_3 = \frac{M_3^2}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right]^{3/2}}$$

- 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:

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

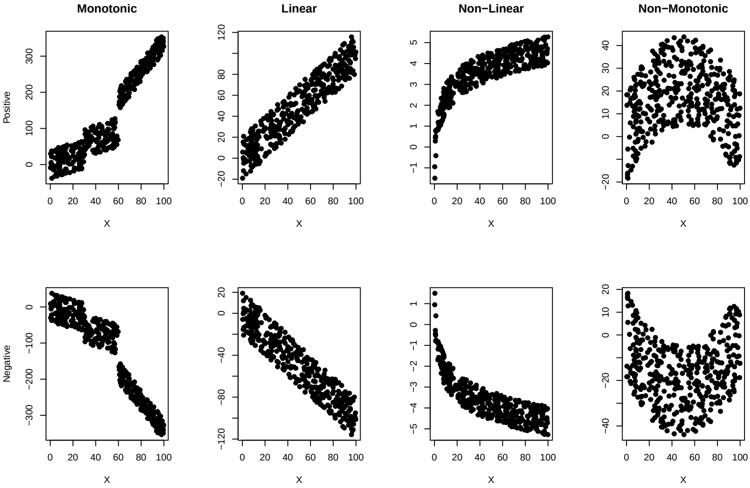
Variance:

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

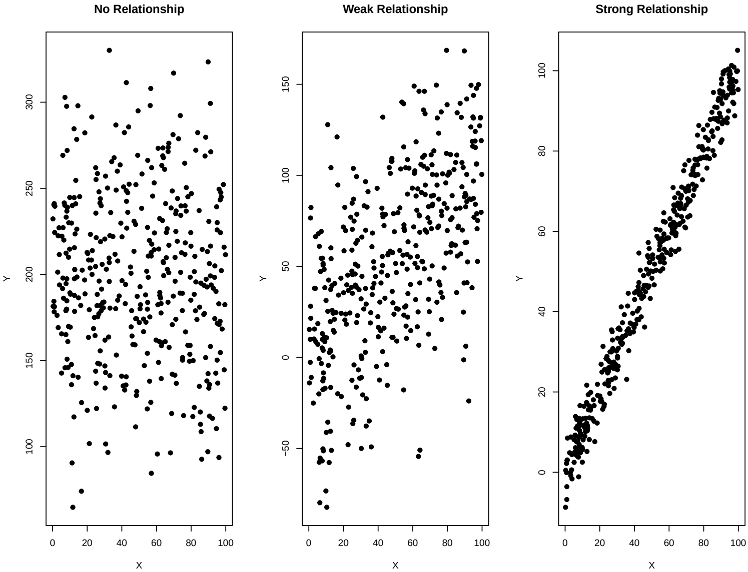
and so SD:

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

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

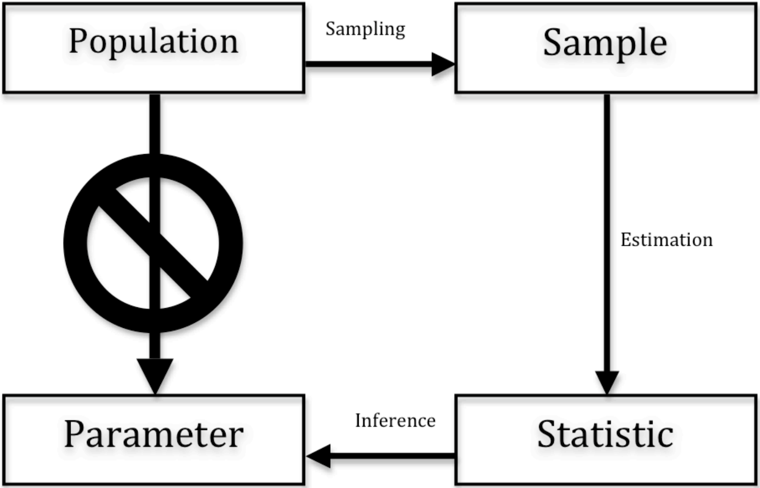
The general idea:

- 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* $\theta = \mathrm{f}(\text{sample data } \mathbf{X})$
- 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: 2024-25 Final English Premier League (EPL)

```
> print(EPL)

   Rank                     Team Matches Win Draw Loss Goals GoalsAgainst GoalDifference Points
1     1                Liverpool      38  25    9    4    86           41             45     84
2     2                  Arsenal      38  20   14    4    69           34             35     74
3     3          Manchester City      38  21    8    9    72           44             28     71
4     4                  Chelsea      38  20    9    9    64           43             21     69
5     5         Newcastle United      38  20    6   12    68           47             21     66
6     6              Aston Villa      38  19    9   10    58           51              7     66
7     7        Nottingham Forest      38  19    8   11    58           46             12     65
8     8 Brighton and Hove Albion      38  16   13    9    66           59              7     61
9     9          AFC Bournemouth      38  15   11   12    58           46             12     56
10   10                Brentford      38  16    8   14    66           57              9     56
11   11                   Fulham      38  15    9   14    54           54              0     54
12   12           Crystal Palace      38  13   14   11    51           51              0     53
13   13                  Everton      38  11   15   12    42           44             -2     48
14   14          West Ham United      38  11   10   17    46           62            -16     43
15   15        Manchester United      38  11    9   18    44           54            -10     42
16   16  Wolverhampton Wanderers      38  12    6   20    54           69            -15     42
17   17        Tottenham Hotspur      38  11    5   22    64           65             -1     38
18   18           Leicester City      38   6    7   25    33           80            -47     25
19   19             Ipswich Town      38   4   10   24    36           82            -46     22
20   20              Southampton      38   2    6   30    26           86            -60     12
```

EPL Data Summary

```
> summary(EPL)

      Rank            Team               Matches       Win             Draw      
 Min.   : 1.00   Length:20          Min.   :38   Min.   : 2.0   Min.   : 5.00  
 1st Qu.: 5.75   Class :character   1st Qu.:38   1st Qu.:11.0   1st Qu.: 7.75  
 Median :10.50   Mode  :character   Median :38   Median :15.0   Median : 9.00  
 Mean   :10.50                      Mean   :38   Mean   :14.3   Mean   : 9.30  
 3rd Qu.:15.25                      3rd Qu.:38   3rd Qu.:19.2   3rd Qu.:10.25  
 Max.   :20.00                      Max.   :38   Max.   :25.0   Max.   :15.00  
      Loss            Goals        GoalsAgainst  GoalDifference      Points    
 Min.   : 4.00   Min.   :26.0   Min.   :34.0   Min.   :-60.0   Min.   :12.0  
 1st Qu.: 9.75   1st Qu.:45.5   1st Qu.:45.5   1st Qu.:-11.2   1st Qu.:42.0  
 Median :12.00   Median :58.0   Median :52.5   Median :  3.5   Median :55.0  
 Mean   :14.35   Mean   :55.8   Mean   :55.8   Mean   :  0.0   Mean   :52.4  
 3rd Qu.:18.50   3rd Qu.:66.0   3rd Qu.:62.8   3rd Qu.: 14.2   3rd Qu.:66.0  
 Max.   :30.00   Max.   :86.0   Max.   :86.0   Max.   : 45.0   Max.   :84.0  
```

Alternative Summary

```
> describe(EPL)

               vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
Rank              1 20 10.50  5.92   10.5   10.50  7.41   1  20    19  0.00    -1.38 1.32
Team*             2 20 10.50  5.92   10.5   10.50  7.41   1  20    19  0.00    -1.38 1.32
Matches           3 20 38.00  0.00   38.0   38.00  0.00  38  38     0   NaN      NaN 0.00
Win               4 20 14.35  6.00   15.0   14.69  5.93   2  25    23 -0.34    -0.73 1.34
Draw              5 20  9.30  2.87    9.0    9.12  2.22   5  15    10  0.52    -0.81 0.64
Loss              6 20 14.35  6.96   12.0   14.00  4.45   4  30    26  0.56    -0.61 1.56
Goals             7 20 55.75 14.71   58.0   56.12 13.34  26  86    60 -0.18    -0.59 3.29
GoalsAgainst      8 20 55.75 14.42   52.5   54.50 12.60  34  86    52  0.70    -0.62 3.22
GoalDifference    9 20  0.00 27.04    3.5    1.69 22.98 -60  45   105 -0.62    -0.28 6.05
Points           10 20 52.35 18.58   55.0   53.44 18.53  12  84    72 -0.46    -0.63 4.15
```

Hypothesis Testing: One Variable

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 Testing: One Variable

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 1.38 0.49   1.45    1.41 0.49 0.32 2.21  1.89 -0.46    -0.63 0.11


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

 One Sample t-test

data:  EPL$PPG
t = 1.2, df = 19, p-value = 0.3
alternative hypothesis: true mean is not equal to 1.25
95 percent confidence interval:
 1.149 1.606
sample estimates:
mean of x
    1.378
```

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 = -0.51, df = 14, p-value = 0.6
alternative hypothesis: true difference in means between group 0
  and group 1 is not equal to 0
95 percent confidence interval:
 -0.5556  0.3438
sample estimates:
mean in group 0 mean in group 1
          1.346           1.452
```

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] -0.7236


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

 Pearson's product-moment correlation

data:  Goals and GoalsAgainst
t = -4.4, df = 18, p-value = 0.0003
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8833 -0.4135
sample estimates:
    cor
-0.7236
```

Next time: Data Visualization