PLSC 502 – Autumn 2016
Data: Structure and Measurement

September 1, 2016

Rectangular Data

i	X_1	X_2	...	X_K
1	X_{11}	X_{21}	...	X_{K1}
2	X_{12}	X_{22}	...	X_{K2}
3	X_{13}	X_{23}	...	X_{K3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X_{1N}	X_{2N}	...	X_{KN}

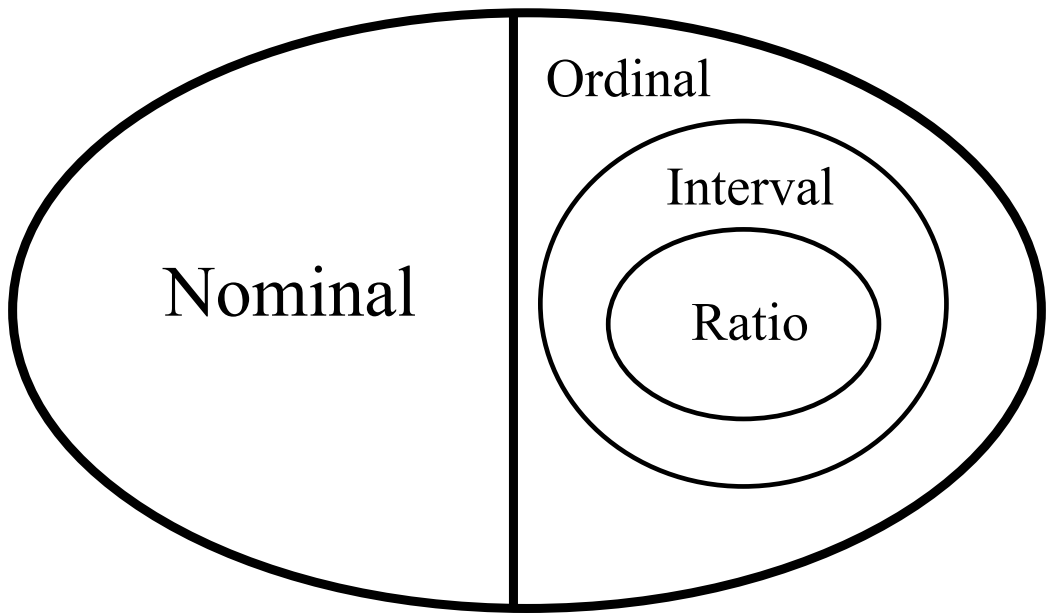
with indices:

$$i \in \{1, 2, 3, ...N\}$$

$$k \in \{1, 2, 3, ...K\}$$

Levels of Measurement

- Nominal
- Ordinal
- Interval
- Ratio



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, degrees F
Ratio	Counts of things	Height, weight, distance, etc.

Cross-Sectional Data: 1997 Baseball Survey

```
> select<-c("respon","age","female","followbaseball","DH_appr")
> head(DH[select],8)

  respon age female followbaseball DH_appr
1      1  65 Female              0      NA
2      2  63   Male              1       1
3      3  56 Female              1      NA
4      4  24 Female              0      NA
5      5  47   Male              0      NA
6      6  81 Female              1      NA
7      7  28   Male              1       1
8      8  76   Male              1       0
```

Time Series Data: SCOTUS Clerks

```
> select<-c("Term","female","white","top5law","lcclerk")
> head(Clerks[select],15)

   Term    female      white   top5law    lcclerk
1  1953 0.0000000 100.000000 44.444447 12.5000000
2  1954 0.0000000 100.000000 64.705887 44.4444470
3  1955 0.0000000 100.000000 76.470589 41.6666640
4  1956 0.0000000 100.000000 55.555557 20.0000000
5  1957 0.0000000 100.000000 58.823532 30.0000020
6  1958 0.0000000 100.000000 57.894737 27.2727280
7  1959 0.0000000 100.000000 61.111111 44.4444470
8  1960 0.0000000 100.000000 66.666672  7.1428576
9  1961 0.0000000 100.000000 55.555557 21.4285720
10 1962 0.0000000 100.000000 71.428574 21.4285720
11 1963 0.0000000 100.000000 78.947372 25.0000000
12 1964 0.0000000 100.000000 62.500000  8.3333340
13 1965 0.0000000 100.000000 70.000000 43.7500000
14 1966 5.8823528 100.000000 52.941177 33.3333360
15 1967 0.0000000  95.238098 66.666672 44.4444470
```

Panel/TSCS Data

$$X_{it} \in X = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1T} \\ X_{21} \\ X_{22} \\ \vdots \\ X_{NT-1} \\ X_{NT} \end{pmatrix}$$

Panel/TSCS Data: Countries, 1946-1999

```
> select<-c("country","ccode","year","gdppc","polity","region","coldwar")
> Panel<-Panel[order(Panel$ccode,Panel$year),] # sort
> Panel[1:200,select]
     country ccode year     gdppc polity region coldwar
9664      US     2 1946        NA     10      1       1
9665      US     2 1947        NA     10      1       1
9666      US     2 1948        NA     10      1       1
9667      US     2 1949        NA     10      1       1
9668      US     2 1950  1915.000     10      1       1
9669      US     2 1951  2196.000     10      1       1
9670      US     2 1952  2300.000     10      1       1
.
.
.
9706      US     2 1988 20848.000     10      1       1
9707      US     2 1989 22192.000     10      1       1
9708      US     2 1990 23218.000     10      1       0
9709      US     2 1991 23639.000     10      1       0
.
.
.
9715      US     2 1997 30468.000     10      1       0
9716      US     2 1998 31776.000     10      1       0
9717      US     2 1999        NA     10      1       0
2676              2 2000        NA     10      1       0
3886  CANADA    20 1946        NA     10      1       1
3887  CANADA    20 1947        NA     10      1       1
3888  CANADA    20 1948        NA     10      1       1
3889  CANADA    20 1949        NA     10      1       1
3890  CANADA    20 1950  1544.000     10      1       1
3891  CANADA    20 1951  1717.000     10      1       1
```

Relational Data

$$X_{ij} \in X = \begin{pmatrix} X_{12} \\ X_{13} \\ \vdots \\ X_{1N} \\ X_{23} \\ X_{24} \\ \vdots \\ X_{2N} \\ X_{34} \\ X_{35} \\ \vdots \\ X_{N-2,N-1} \\ X_{N-2,N} \\ X_{N-1,N} \end{pmatrix}$$

Relational Data: Country "Dyads" (1968)

```
> select<-c("ccode1","ccode2","dyadid","dem1","dem2","allies","distance")
> Dyads[1:300,select]
    ccode1 ccode2 dyadid dem1 dem2 allies distance
1        2     20   2020   10   10      1        0
2        2     40   2040   10   -7      0     1135
3        2     41   2041   10   -9      1     1437
4        2     42   2042   10   -3      1     1477
5        2     51   2051   10   10      0     1446
6        2     52   2052   10    8      1     2176
.
.
.
126      2    840   2840   10    5      1     8570
127      2    850   2850   10   -7      0    10172
128      2    900   2900   10   10      1     9916
129      2    920   2920   10   10      1     8759
130     20     40  20040   10   -7      0     1586
131     20     41  20041   10   -9      0     1869
132     20     42  20042   10   -3      0     1893
133     20     51  20051   10   10      0     1897
134     20     52  20052   10    8      0     2547
135     20     53  20053   10   NA      0     2426
.
.
.
259     20    900  20900   10   10      0    10019
260     20    920  20920   10   10      0     9009
261     40     41  40041   -7   -9      0      722
262     40     42  40042   -7   -3      0      868
263     40     51  40051   -7   10      0      506
.
.
.
```

Missing Data

Why?

- Observation doesn't exist
- Data don't exist for that observation
- Data exist, but are *impossible* to measure
- Data exist, but were not measured

Three types:

- Missing completely at random ("MCAR"),
- Missing at random ("MAR"), and
- Informatively (or "non-ignorably") missing.

Missing Data: What To Do?

- Listwise deletion
- Interpolation / replacement values
- Imputation-based approaches

Data Tips

- **Use descriptive variable names**.
  - Spell it out.
  - Use "directional" names.
- **Be consistent in naming variables**.
- **Label everything**.
- **Never overwrite anything**.
- **Log everything** (or use reproduceable code).

Reproduceable Code

From PLSC502-DayTwo-2016.R:

```
#####################################
# PLSC 502 -- Fall 2016
#
# Day Two materials
#####################################

library(RCurl)

# DH data:

temp<-getURL("https://raw.githubusercontent.com/PrisonRodeo/PLSC502-2016-git/master/Data/DH.csv")
DH<-read.csv(text=temp, header=TRUE)
rm(temp)

select<-c("respon","age","female","followbaseball","DH_appr")
head(DH[select],8)
```