Descriptives

LPO 9951 | Fall 2020

PURPOSE

Describing the data in your sample is one of the most important steps in reporting on your research. A reader that has a clear understanding of the patterns in your data will be able to readily understand your more complex analyses.

The key to descriptive statistics turns out to be the humble conditional mean: the mean of the dependent variable at various levels of the independent variable. Master the conditional mean and how to display it, and everyone will always remember your pa pers and presentations.

```
. capture log close                       // closes any logs, should they be open

. set linesize 90

. log using "descriptives.log", replace    // open new log
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/descriptives.log
  log type:  text
 opened on:  28 Oct 2020, 11:16:50

. clear all                               // clear memory


. graph drop _all
```

HEADER

Incidental to the lesson today, but important to set up correctly is the header. Notice that the plot and table files types are saved in global macros. With a quick switch at the top of the file, you can change the file format of the plots and tables th at Stata saves. Very handy.

```
. global datadir "../data/"

. global plotdir "../plots/"

. global tabsdir "../tables/"

. global gtype png
```

```
. global ttype rtf

. set scheme s1color

. use plans2, clear

. recode bypared  (1=1 "Less than HS") ///
                                  (2=2 "HS") ///
                                  (3/5=3 "Some College or Associate") ///
                                  (6=4 "BA") ///
                                  (6/8=5 "Adv Degree") , gen(bypared2)

(9655 differences between bypared and bypared2)

. la var bypared2 "Parental Education"

. recode f1psepln (1/2 = 1 "No plans") ///                                    (3/4 = 2 "V

(13995 differences between f1psepln and newpln)

. label var newpln "PS Plans"

. svyset psu [pw = f1pnlwt], strat(strat_id) singleunit(scaled)

     pweight: f1pnlwt
         VCE: linearized
  Single unit: scaled
     Strata 1: strat_id
         SU 1: psu
        FPC 1: <zero>
```

Tables

Every manuscript should include a table of descriptive statistics, listing the mean and standard error or standard deviation of every variable to be used in the dataset. In addition, tables should be used to convey crosstabs of two categorical variables . Most of your papers will also eventually include tables for regression results. Tables should be used sparingly for describing data: your best bet is almost always graphics.

For many categorical variables, however, tables may be your only option. In that case you need to think hard about two things:

1. How can I best show patterns in the conditional mean of my dependent variable at different levels of my independent variables?
2. How can I best show relationships among key independent variables?

Principles for displaying data

Tufte (2001) lists the following principles for describing data using graphics. He says they should:

- Show the data
- Induce the viewer to think about the substance rather than about the methodology, graphic design, the technology production, or something else.
- Avoid distorting what the data have to say.
- Present many numbers in a small space.
- Make large datasets coherent.
- Encourage the eye to compare different pieces of data.
- Reveal the data at several levels of detail, from a broad overview to fine structure.
- Serve a reasonably clear purpose: description, exploration, tabulation, or decoration.
- Be closely integrated with the statistical and verbal descriptions of a dataset.

Describing variation and central tendency in continuous variables

Plots

The two key tools for describing variation and central tendency in a continuous variable are the kernel density plot and the histogram. A histogram should be your first choice for most variables: the key decisions will be the number of bins or the frequ ency of the plot. Histograms can also be combined across levels using the onewayplot command.

The basic histogram is shown here.

```
. histogram bynels2m, name(hist_bynels2m, replace)  ///
     xtitle("NELS-1992 Scale-Equated Math Score") ///
     bin(25) /// we can try different bin widths
     percent

(bin=25, start=.1471, width=.025824)


. graph export hist_bynels2m.$gtype, name(hist_bynels2m) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/hist_bynels2m.png written in
```

At the extreme end of the histogram is the "spike" plot, which has a single line for every level of the underlying variable.

```
. spikeplot bynels2m, name(spike_bynels2m,replace) ///
```

```
    xtitle("NELS-1992 Scale-Equated Math Score") ///
        color(blue*.5%25)
```

```
. graph export spike_bynels2m.$gtype, name(spike_bynels2m) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/spike_bynels2m.png written in
```

Kernel density plots are a key tool for describing a continuous variable. The density of the variable can be compared to standard densities for visual comparison, like in the first plot below. Kernel density plots can be particularly illuminating when d isplayed across multiple levels of a categorical variable, as in the second plot below.

```
. kdensity bynels2m, name(kd_bynels2m, replace) ///
    xtitle("NELS Math Scores") ///
    n(100) ///
    bwidth(.025) ///
        color("98 47 117*.7%30") ///
        recast(area) ///
    normal ///
        kernel(gaussian) ///
    normopts(lpattern(dash) color(black))
```

```
. graph export kd_bynels2m.$gtype, name(kd_bynels2m) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/kd_bynels2m.png written in PN

. kdensity bynels2m if bypared == 2, name(kd_bynels2m_cond, replace) ///
    xtitle("NELS Math Scores") ///
    n(100) ///
    bwidth(.025) ///
        recast(area) ///
        color("216 171 76* .5%50") ///
    addplot(kdensity bynels2m if bypared > 5, ///
            n(100) ///
            bwidth(.025) ///
            color("98 47 117*.7%50") ///
                        recast(area)) ///
    legend(label(1 "Parent Ed=HS") label(2 "Parent Ed=BA+")) ///
    note("") ///
    title("")
```

```
. graph export kd_bynels2m_cond.$gtype, name(kd_bynels2m_cond) replace
```

```
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/kd_bynels2m_cond.png written

. histogram bystexp, name(bar_bystexp) ///
    percent ///
    addlabels ///
    xlabel(-1 1 2 3 4 5 6 7, ///
           value ///
           angle(45) ///
           labsize(vsmall) ///
           ) ///
    addlabopts(yvarformat(%4.1f)) ///
    xtitle("")

(bin=41, start=-1, width=.19512195)


. graph bar ,over(bystexp, ///
                              sort(1) ///
                                  descending ///
                                  label(angle(45) labsize(small) ) ///
                                  ) ///
        blabel(bar,format(%4.1f)) ///
        bar(1, color(blue*.5)) ///
        ytitle("")



. graph hbar ,over(bystexp, ///
                              sort(1) ///
                                  descending ///
                                  label(labsize(small) ) ///
                                  ) ///
        blabel(bar,format(%4.1f)) ///
        bar(1, color(blue*.5)) ///
        ytitle("")



. graph export bar_bystexp.$gtype, name(bar_bystexp) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/bar_bystexp.png written in PN
```

Pie Charts

No.

Tables

For tables describing continuous variables, the industry standard is a table of means and standard errors or standard deviations. Below is a table of means and standard errors, nicely formatted.

```
. svy: mean bynels2m bynels2r byses1 byses2 amind asian black hispanic white female
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     361        Number of obs   =     15,512
Number of PSUs   =     751        Population size =  3,210,779
                                  Design df       =        390

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    bynels2m |   .4485513   .0026715       .443299    .4538036
    bynels2r |   .2947283   .0017403      .2913067    .2981498
      byses1 |   .0050858   .0145406     -.0235019    .0336736
      byses2 |   .0049615   .0143462     -.0232441    .0331671
       amind |   .0096097   .0021007      .0054797    .0137397
       asian |   .0386928   .0025431      .0336929    .0436926
       black |   .1373778    .006713      .1241797    .1505759
    hispanic |   .1510179   .0084864      .1343331    .1677028
       white |   .6219287   .0099563      .6023541    .6415033
      female |   .4973999    .005431      .4867221    .5080776
--------------------------------------------------------------

. estimates store my_mean

. esttab my_mean using means_se.$ttype, ///    // . means all in current memory    not

(output written to `"means_se.rtf"')

. tabstat bynels2m bynels2r byses1 byses2, stat(sd) save

   stats |  bynels2m  bynels2r    byses1    byses2
---------+----------------------------------------
      sd |  .1353664  .0939987  .7429628  .7502604
--------------------------------------------------

. mat mysd = r(StatTotal)
```

```
. estadd  matrix mysd: my_mean

. esttab my_mean using means_sd.$ttype, ///
     not ///
     replace ///
     nostar ///
     label ///
     main(b) ///
     aux(mysd) ///                          // NOTE: aux = standard deviations
     nonumbers ///
     nonotes ///
     addnotes("Standard deviations in parentheses")

(output written to `"means_sd.rtf"')

. estpost svy: tabulate bystexp
(running tabulate on estimation sample)

Number of strata   =       361                 Number of obs     =     15,236
Number of PSUs     =       751                 Population size   =  3,210,779
                                               Design df         =        390

----------------------
how far   |
in school |
student   |
thinks    |
will      |
get-compo |
site      | proportion
----------+-----------
 Don't Kn |      .0967
 Less tha |      .0089
       HS |      .0688
     2 yr |      .0657
  4 yr No |      .0369
 Bachelor |       .361
  Masters |      .2002
 Advanced |      .1619
          |
    Total |          1
----------------------
  Key:  proportion  =  cell proportion

saved vectors:
             e(b) =  cell proportions
```

```
              e(se) =  standard errors of cell proportions
              e(lb) =  lower 95% confidence bounds for cell proportions
              e(ub) =  upper 95% confidence bounds for cell proportions
            e(deff) =  deff for variances of cell proportions
            e(deft) =  deft for variances of cell proportions
            e(cell) =  cell proportions
           e(count) =  weighted counts
             e(obs) =  number of observations

. esttab . using proportions.$ttype, ///
    not ///
    replace ///
    nostar ///
    label ///
    main(b) ///
    aux(se) ///
    nonotes ///
    nonumbers ///
    addnotes("Linearized standard errors in parentheses")

(output written to `"proportions.rtf"')

. graph twoway scatter bynels2m bynels2r, name(sc_math_read)


. graph export sc_math_read.$gtype, name(sc_math_read) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/sc_math_read.png written in P

. preserve                                  // preserve data

. sample 10                                 // sample random 10%
(14,544 observations deleted)

. graph twoway scatter bynels2m bynels2r, name(sc_math_read_10) ///
   ytitle("NELS Math Scores") ///
   xtitle("NELS Reading Scores") ///
   msize(tiny)



. graph export sc_math_read_10.$gtype, name(sc_math_read_10) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/sc_math_read_10.png written i

. restore                                   // restore data
```

You can condition on another variable in order to add another level to your

scatter plot. This can be done both with use of `if` statements, as in the first plot below, and `by` statements, as in the second plot.

```
. preserve                                  // preserve data

. sample 25                                 // sample random 25%
(12,120 observations deleted)

. graph twoway (scatter bynels2m byses1 if urm == 0, ///
                    mcolor(blue*.5%25) ///
                    msize(tiny) ///
                                    ) ///
                || scatter bynels2m byses1 if urm == 1, ///
                    mcolor(orange*.5%25) ///
                    msize(tiny) ///
                    ytitle("NELS Math Scores") ///
                    xtitle("SES") ///
                    legend(order(1 "Non-Minority" 2 "Underrep Minority")) ///
                    name(sc_complex, replace)



. graph export sc_complex.$gtype, name(sc_complex) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/sc_complex.png written in PNG

. restore

. graph twoway scatter bynels2m byses1, by(bystexp,note("")) ///
    msize(*.05) ///
        mcolor(dknavy) ///
    ytitle("NELS Math Scores") ///
    xtitle("SES") ///
    note("") ///
    name(sc_cond, replace)



. graph twoway scatter bynels2m byses1, ///
                        subtitle(, ring(0) pos(11) nobexpand fcolor(white%1) lstyle(n
                        by(bystexp, total note("")) ///
    msize(*.05) ///
        mcolor(dknavy) ///
    ytitle("NELS Math Scores") ///
    xtitle("SES") ///
    note("") ///
    name(sc_cond2, replace)
```

```
. graph export sc_cond2.$gtype, name(sc_cond2) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/sc_cond2.png written in PNG f
```

You can also run a scatter plot across levels of a categorical variable if you suspect the underlying relationship may not be the same in each level of the categorical variable. The `matrix` plot helpfully with plot each combination of include d varibles against each other to produe a sort of "small multiples" correlation plot. I mostly don't recommend running a matrix plot.

```
. graph matrix bynels2m bynels2r byses1 byses2, name(matrix_plot) msize(vtiny)
```

```
. graph export matrix_plot.$gtype, name(matrix_plot) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/matrix_plot.png written in PN
```

Scatterplot of proportions

If you have a binary dv, you can do a scatterplot. What you need to do is calculate the proportion of the sample in ranges of another continuous variable. A standard solution here is to use percentiles of another variable. Below I calculate the proporti on of students who plan to go to college for each percentile of math scores, then plot the result. This can be a bit "too neat" so be careful.

```
. xtile pct_math = bynels2m, nq(100)

. egen fouryr_avg= mean(fouryr) , by(pct_math)

. graph twoway scatter fouryr_avg pct_math, ///
                        xtitle("Math Score Pctile") ///
                        ytitle("Pr(Plan to go to 4yr)") ///
                        mcolor(blue*.5%50)
```

Describing relationships between a categorical and a continuous variable

Plots

There are multiple options for plotting the relationship between a categorical and a continuous variable. A particularly useful option is to plot the continuous variable as a series of boxplots, one for each level of the categorical variable.

Boxplots

For boxplots to be effective, they should be sorted by the median of the dependent variable. This contrast is shown in the two figures below.

```
. graph box bynels2m, over(bypared2, ///
                           label(alternate ///
                                 labsize(tiny) ///
                                 ) ///
                           ) ///
                     name(box1, replace)
```

```
. graph export box1.$gtype, name(box1) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/box1.png written in PNG forma
```

```
. graph box bynels2m, over(bypared2, ///
                           label(alternate ///
                                 labsize(tiny) ///
                                 ) ///
                           sort(1) ///
                           ) ///
                     name(box2)
```

```
. graph export box2.$gtype, name(box2) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/box2.png written in PNG forma
```

Bar Plots

Bar plots are always a great option, particularly for policy audiences. Below I go over some options when using bar plots.

```
. graph hbar bynels2m bynels2r [pw=bystuwt], ///
                 over(bystexp, sort(bynels2m) descending) ///
                 ytitle("Test Scores") ///
                 legend(order(1 "Math Scores" 2 "Reading Scores"))  ///
                 blabel(bar,format(%9.2f)) ///
                 bar(1, color(orange*.5)) bar(2, color(blue*.5))
```

```
. graph hbar bynels2m bynels2r [pw=bystuwt], ///
                 over(bystexp, sort(bynels2m) descending) ///
```

```
                    ytitle("Test Scores") ///
                    legend(order(1 "Math Scores" 2 "Reading Scores"))  ///
                    blabel(bar,format(%9.2f)) ///
                    bar(1, color(orange*.5)) bar(2, color(blue*.5)) ///
                    name(barplot1)
```

```
. betterbarci bynels2m bynels2r [pw=bystuwt], ///
                         over(bypared2)
```

```
. statplot bynels2m bynels2r, over(bystexp,sort(1) descending) over(bysex)  name(statp
```

Dot plots

Dot plots can also be useful for plotting the measure of central tendency across groups. In this case, we'll produce two plots, one each for reading and math scores, and then combine them into a single graphic.

```
. graph dot bynels2m, over(bypared2, ///
                           label(alternate ///
                                 labsize(tiny) ///
                                 ) ///
                           ) ///
                      ytick(0(.10).80) ///
                      ylabel(0(.1).8) ///
                      ytitle("Math Scores") ///
                      marker(1, msymbol(O) mcolor(dknavy)) ///
                      name(dot_math,replace)
```

```
. graph save dot_math.gph, replace
(file dot_math.gph saved)

. graph dot bynels2r, over(bypared2, ///
                           label(alternate ///
                                 labsize(tiny) ///
                                 ) ///
                           ) ///
                      ytick(0(.10).80) ///
                      ylabel(0(.1).8) ///
                      ytitle("Reading Scores") ///
                      marker(1, msymbol(O) mcolor(orange*.5)) ///
                      name(dot_read,replace)
```

```
. graph save dot_read.gph, replace
(file dot_read.gph saved)

. graph combine dot_math.gph dot_read.gph, ///
    name(dot_both, replace) ///
        cols(1)



. graph export dot_both.$gtype, name(dot_both) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/dot_both.png written in PNG f
```

Describing relationships between two categorical variables

Plots

The basic tool for comparing two categorical variables is the crosstabulation. In a crosstabulation we take a look at counts of the sample that are identified by their presence in cells created by the two categorical variables. There are several tools f or plotting categorical variables, including tabplots, jittered plots, and heatmaps.

Tabplots

Below are examples of a tabplot, with both two and three dimensions.

```
. tabplot bypared2 newpln, name(tabplot1, replace) ///
    percent(bypared2) ///
    showval ///
    subtitle("")



. graph export tabplot1.$gtype, name(tabplot1) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/tabplot1.png written in PNG f

. tabplot bypared2 newpln, by(bysex) ///
    percent(bypared2) ///
    showval ///
    subtitle("") ///
    name(tabplot2, replace)
```

```
. graph export tabplot2.$gtype, name(tabplot2) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/tabplot2.png written in PNG f

. graph twoway scatter f1psepln bypared, name(jitterplot) ///
    jitter(5) ///
    msize(vtiny) ///
        mcolor(dknavy)



. graph export jitterplot.$gtype, name(jitterplot) replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/jitterplot.png written in PNG

. tddens bypared f1psepln, title("") ///
    xtitle("Parent's Level of Education") ///
    ytitle("PS PLans")


. graph export heatmap.$gtype, replace
(file /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/heatmap.png written in PNG fo
```

When checking crosstabulations, we can produce two-way tables that include survey weights in the command itself.

```
. table byrace2 f1psepln [pw = bystuwt], by(bysex) contents(freq) row

------------------------------------------------------------------------------
sex-composi |
te and      |
RECODE of   |
byrace      |
(student^s  |
race/ethnic |
ity-composi |         f1 post-secondary plans right after high school
te)         | don^t plan to contin  don^t know or planni  vocational, technica
------------+-----------------------------------------------------------------
male        |
    Am.Ind. |             959.9478              2,165.51             1,850.23
   Asian/PI |             955.4428              3,010.54             4,576.58
      Black |             3,280.14              13,926.5             22,419.1
   Hispanic |             2,481.69              23,987.1             25,572.5
Multiracial |             2,885.66              7,204.85             6,076.28
      White |               27,037                74,634               91,041
            |
      Total |             37,599.8               124,929              151,536
```

```
------------+-------------------------------------------------------------
female      |
    Am.Ind. |                              961.4319              1,422.26
   Asian/PI |             182.5247             1,706.65             1,229.69
      Black |             1,682.59             9,165.85               12,928
   Hispanic |             1,098.54             12,487.4             16,983.4
Multiracial |             916.7742             2,516.75             2,778.86
      White |             5,139.84             37,381.5             44,545.3
            |
      Total |             9,020.27             64,219.6             79,887.6
-------------------------------------------------------------------------------


-------------------------------------------------------------------------------
sex-composi |
te and      |
RECODE of   |
byrace      |
(student^s  |
race/ethnic |
ity-composi |         f1 post-secondary plans right after high school
te)         | two-year community c  four-year college or  early hs grad attend
------------+-------------------------------------------------------------
male        |
    Am.Ind. |             1,168.03              8,487.8             509.1148
   Asian/PI |             13,906.7             41,635.3             1,415.14
      Black |             33,505.1              121,834             8,161.93
   Hispanic |             68,736.3             82,651.8                7,870
Multiracial |             12,485.4             31,158.9             1,156.67
      White |              172,672              531,000             18,210.1
            |
      Total |              302,473              816,768               37,323
------------+-------------------------------------------------------------
female      |
    Am.Ind. |             2,538.59              6,913.7             198.1191
   Asian/PI |             9,878.74             47,395.6             1,796.86
      Black |             48,116.9              124,183             5,876.88
   Hispanic |             76,304.8              114,582             7,564.25
Multiracial |             14,070.6             36,070.2             3,290.11
      White |              195,002              606,845             20,989.9
            |
      Total |              345,912              935,989             39,716.1
-------------------------------------------------------------------------------
```

Of course, if we want to use a table in a paper, we should use `esttab`.

```
. estpost svy: tabulate byrace2 newpln, row percent se
(running tabulate on estimation sample)
```

```
Number of strata   =       361                  Number of obs     =     13,055
Number of PSUs     =       750                  Population size   =  2,908,622
                                                Design df         =        389

-------------------------------------------------
RECODE of |
byrace    |
(student^ |
s         |
race/ethn |
icity-com |               PS Plans
posite)   | No plans  VoTech/C      4 yr     Total
----------+--------------------------------------
  Am.Ind. |    16.17     26.93     56.89       100
          |  (3.647)   (4.454)   (5.413)
          |
 Asian/PI |    4.746     23.67     71.59       100
          |  (.8931)   (1.933)   (2.107)
          |
    Black |     7.31      29.9     62.79       100
          |  (.7736)   (1.368)   (1.487)
          |
 Hispanic |    9.732     43.95     46.32       100
          |  (.8412)   (1.494)   (1.699)
          |
 Multirac |    12.27     30.13      57.6       100
          |   (1.72)   (2.436)   (2.591)
          |
    White |    8.147     28.26     63.59       100
          |  (.3851)   (.8152)   (.9036)
          |
    Total |    8.369      30.6     61.04       100
          |  (.3292)   (.6266)   (.7225)
-------------------------------------------------
  Key:  row percentage
        (linearized standard error of row percentage)

  Pearson:
    Uncorrected   chi2(10)        =  253.9129
    Design-based  F(9.11, 3544.12)=   17.6793     P = 0.0000

Note: Variance scaled to handle strata with a single sampling unit.

saved vectors:
             e(b) =  row percentages
```

```
            e(se) =  standard errors of row percentages
            e(lb) =  lower 95% confidence bounds for row percentages
            e(ub) =  upper 95% confidence bounds for row percentages
          e(deff) =  deff for variances of row percentages
          e(deft) =  deft for variances of row percentages
          e(cell) =  cell percentages
           e(row) =  row percentages
           e(col) =  column percentages
         e(count) =  weighted counts
           e(obs) =  number of observations

row labels saved in macro e(labels)

. eststo racetab

. esttab racetab using race_tab.$ttype, ///
     replace ///
     nostar ///
     nostar ///
     unstack ///
     nonotes ///
         nomtitles ///
         nonumbers ///
     varlabels(`e(labels)') ///
     eqlabels(`e(eqlabels)')

(output written to `"race_tab.rtf"')

. log close
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s1-09-descriptives/descriptives.log
  log type:  text
 closed on:  28 Oct 2020, 11:18:21
-------------------------------------------------------------------------------

. exit
```