PHP 2500 Introduction to Biostatistics

Problem Set One Solutions (Updated)

---

1. (Pagano #7, p30) (a) discrete, (b) continuous, (c) continuous, (d) discrete

   (Pagano #9, 15, p30) self explanatory

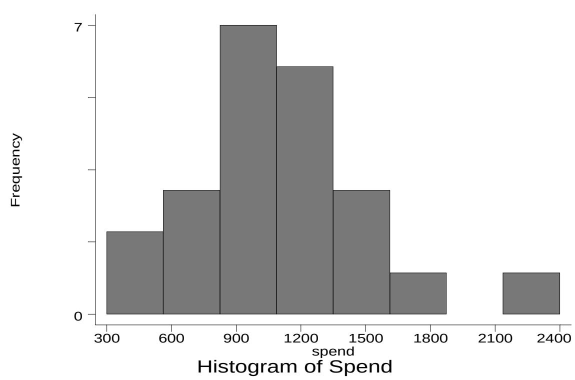
2. (Pagano #8, p30)

   (a) Sort by Per Capita Expenditure:

```
. sort spend                              . gsort - spend
. list                                    . list
           country     spend                         country     spend
  1.        Greece       371               1. United states       2354
  2.      Portugal       464               2.         Canada       1683
  3.         Spain       644               3.    Switzerland       1376
  4.       Ireland       658               4.         Sweden       1361
  5.   New Zealand       820               5.        Iceland       1353
  6.       Britain       836               6.         France       1274
  7.       Denmark       912               7.         Norway       1234
  8.       Belgium       980               8.        Germany       1232
  9.     Australia      1032               9.     Luxembourg       1193
 10.         Japan      1035              10.    Netherlands       1135
 11.         Italy      1050                11.        Austria      1093
 12.       Finland      1067              12.        Finland       1067
 13.       Austria      1093              13.          Italy       1050
 14.   Netherlands      1135              14.          Japan       1035
 15.    Luxembourg      1193              15.      Australia       1032
 16.       Germany      1232              16.        Belgium        980
 17.        Norway      1234              17.        Denmark        912
 18.        France      1274              18.        Britain        836
 19.       Iceland      1353              19.    New Zealand        820
 20.        Sweden      1361              20.        Ireland        658
 21.   Switzerland      1376              21.          Spain        644
 22.        Canada      1683              22.       Portugal        464
 23. United states      2354             23.         Greece       371
```

   (b) Histogram

```
. hist spend, frequency bin(8) xlabel(300(300)2400) title ("Histogram of Spend")
```



2. (c) The shape of the histogram is fairly 'bell-shaped', with maybe an extended tail towards higher per Capita Expenditure. Note that the shape of your histogram will depend on the number of bins.

   (d) Mean = 1093.783, Variance (S.D.) = 170,223.41 (412.58)
   Range = 1983, Median 1067.

3. (Pagano #6, p59)
   (a) Mean = 25.9, Median = 24, Modes(s) = 12 & 24, Range = 95.9,
   IQR(stata)= 39-2.25=36.75, IQR(us)=36-4=32,
   Standard deviation = 27.4

   (b) You can verify that if you subtract the mean, $\bar{x} = 25.9115$, from each of these thirteen numbers, the results will sum to zero. Or you can use algebra:

$$\begin{aligned}\sum_{i=1}^{13}(x_i - \bar{x}) &= \sum_{i=1}^{13} x_i - \sum_{i=1}^{13} \bar{x} \\ &= \sum_{i=1}^{13} x_i - 13\bar{x} \\ &= 13\frac{\sum_{i=1}^{13} x_i}{13} - 13\bar{x} = 13\bar{x} - 13\bar{x} = 0\end{aligned}$$

4. (a) The Mean daily rainfall in 1999 was probably greater in Providence than in Tucson.

   (b) The mode daily rainfall in 1999 for both cities is zero. The most frequent amount of rain is none even though the exact number of days without rain could be different between the two cities.

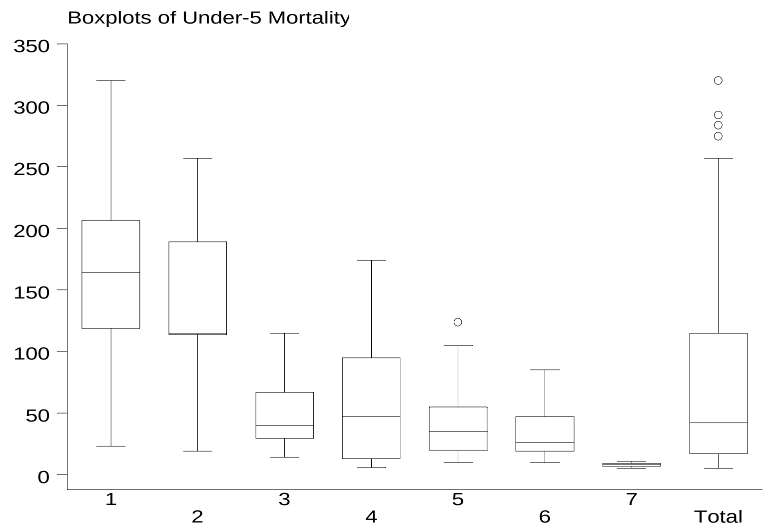
   (c) The median daily rainfall in 1999 for both cities is zero. This is because it rains less than half of the days during the year. Thus, the 50% percentile has to be zero.

5. (a) Smallest mean -- Europe (averaging smaller numbers)
   Largest median -- Africa (histogram shows 50% is greatest)
   Smallest Standard deviation – Europe (smallest spread in data)

   (b) The mean and median are nearly equal when the data are distributed nearly symmetrically, as they are in Africa. They differ when the numbers are skewed, as they are in Asia, because the mean depends heavily on outliers. For Asia, the median will be less than the mean because of the skewness.

6. (a)

```
. graph box mortality, over(region, total label(alternate)) cap(10)
title(Boxplots of Under-5 Mortality) ylab(0(50)350)
```



Key:
1) Sub-Saharan Africa
2) North Africa & Middle East
3) South Asia
4) East Asia and Pacific
5) Latin America and Carribbean
6) Europe, Commonwealth, and Baltic
7) Industrial

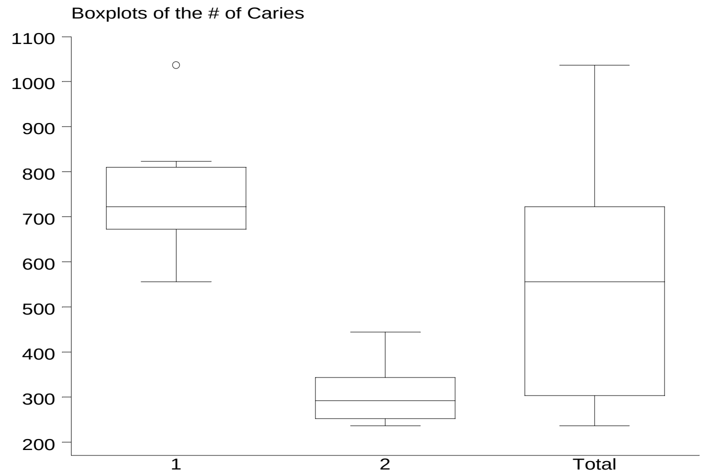
(a) Notice that groups 1 and 2 have similarly high Under5-mortality with a wide spread, while groups 3,4,5, and 6 all tend to be more compact and less variable. Finally, the industrial nations tend to have little variability and a low Under-5 mortality.

(b) (i) Sub-Saharan
(LH =128, M=164, UH =203, IQR=75, LF =15.5, UF =315.5)
Niger is an in the Sub-Saharan.

(ii) Middle East
(LH =34, M=40, UH =63, IQR=29, LF =0, UF =106.5)
Sudan and Yemen are outliers in the Middle east

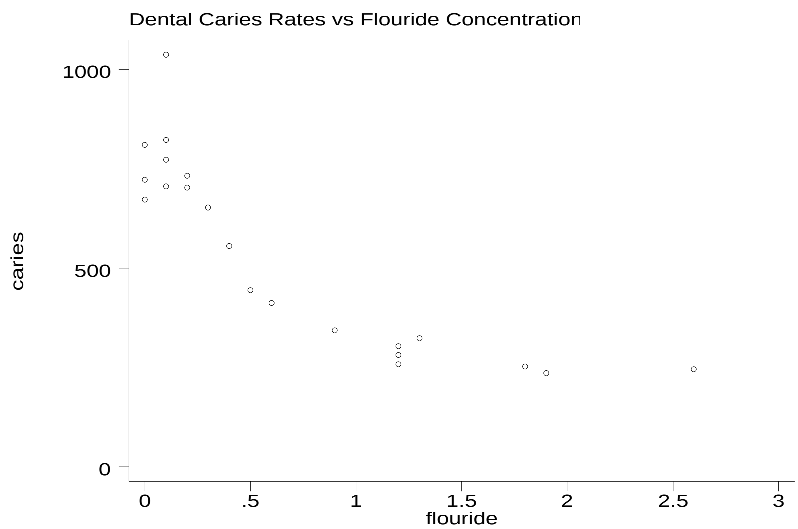
7. (a)

```
. graph box caries, over(group, total) title(Boxplots of the # of
Caries) ylab(200(100)1100)
```



(b)

```
. graph twoway scatter caries  fluoride, title(Dental Caries Rates vs
Flouride Concentration) ylab(0(100)1000) xlab(0(0.5)3) msymbol(oh)
```



7. (b) The box plot shows that the caries rates are consistently higher in the low-fluoride group than the high-fluoride group. The second graph reveals that the caries rates drop smoothly with increasing fluoride concentration. The rate of decrease is greatest when concentration is very low, and it diminishes as concentration rises – by the time the fluoride concentration gets up to 1.5 or 2 ppm the curve appears to be pretty flat.

(c) About 250 carries per 100 children, or 2.5 per child.

(d) It looks like an increase in fluoride concentration from 0.25 to 1.25 will cut the caries rate nearly in half.

(e) An increase in fluoride concentration from 2.0 to 3.0 would appear to have little effect.

8. (a) Crude Birth rate: 87,202/6,060943 = 0.0144 14.39 per 1000
   (b) Crude death rate: 53,804/6,060943 = 0.0089 8.88 per 1000
   (c) Infant Mortality rate: 596/87,202 = 0.0068 6.83 per 1000

9. The death rate from all causes must have been much greater in Guatemala City than in Lima. Specifically, from the facts that were given, we can calculate that the death rate from all causes in Guatemala City was 12.9 per 1000, while in Lima it was only 5.4 per 1,000, so that the Guatemala City rate was more than double the Lima rate (12.9/5.4 = 2.39).

   The calculation goes like this: We're given that for Guatemala City

   % of all deaths due to cancer = $\dfrac{CancerDeaths(GC)}{TotalDeaths(GC)} = 0.135$ (13.5%)

   death rate for cancer = $\dfrac{CancerDeaths(GC)}{Popultaion(GC)} = 0.001736$ (173.6/100,000)

   Dividing the second by the first, we find that:

   All-cause DR = $\dfrac{TotalDeaths(GC)}{Popultaion(GC)} = 0.001736/0.135 = 1285.9/100{,}000$

   or 12.9 per 1000.

   A similar calculation shows that the rate in Lima was 5.4 per 1000.

10. (a) The first explanation that comes to mind is that the quality of care has improved. Later you might wonder if the population of babies has changed, which might lead you to do the analysis in (b).

    (b) The weight-specific mortality rates (per 100 babies, in order of ascending Birth Weight categories) in 1953-57 were

92.1,42.9,12.9, and 6.0, and the corresponding rates in 1968-72 were 91.8, 55.7,13.7, and 4.0.

These changes cannot explain the 40% drop in the crude mortality rate. If we adjust to the 1953-57 weight distribution, we find that the adjusted rate for 1968-72 is [(0.918)(126)+(0.557)(112)+(**0.435**)(241)+(0.040)(601)]/1080 =**0.284**, or **28.4** deaths per 100 babies, slightly <u>greater</u> than the 1953-57 rate (which was 21.4 per 100).

(c) The decline in the crude mortality rate can be explained entirely by the changes in the weight distribution. (There were relatively small changes in the weight-specific weights, and those changes were not all in the same direction.) The mortality rate in the smallest weight group was just as high in 1970 (92 per 1000) as it was in 1955. But such babies made up a much smaller proportion of the population in 1980, and that is what brought the crude rate down.

(d) SMR = (observed deaths)/(Expected deaths) = crude/IA rates. SMR(1953-57)= 0.2139/0.2839=0.75 , SMR(1968-72) = 1

```
. dstdize death number weight_ca, by(year) base(1953)

----------------------------------------------------------
-> year= 1953
                         -----Unadjusted-----  Std.
                                Pop.  Stratum  Pop.
  Stratum       Pop.     Cases  Dist. Rate[s] Dst[P]  s*P
----------------------------------------------------------
        1        126       116  0.117 0.9206  0.117 0.1074
        2        112        48  0.104 0.4286  0.104 0.0444
        3        241        31  0.223 0.1286  0.223 0.0287
        4        601        36  0.556 0.0599  0.556 0.0333
----------------------------------------------------------
Totals:         1080       231    Adjusted Cases:    231.0
                                      Crude Rate:   0.2139
                                   Adjusted Rate:   0.2139
                       95% Conf. Interval: [0.1960, 0.2318]


----------------------------------------------------------
-> year= 1968
                         -----Unadjusted-----  Std.
                                Pop.  Stratum  Pop.
  Stratum       Pop.     Cases  Dist. Rate[s] Dst[P]  s*P
----------------------------------------------------------
        1         49        45  0.040 0.9184  0.117 0.1071
        2        106        59  0.086 0.5566  0.104 0.0577
        3         92        40  0.075 0.4348  0.223 0.0970
        4        985        39  0.800 0.0396  0.556 0.0220
----------------------------------------------------------
Totals:         1232       183    Adjusted Cases:    349.8
```

```
                                         Crude Rate:   0.1485
                                      Adjusted Rate:   0.2839
                             95% Conf. Interval: [0.2568, 0.3110]

Summary of Study Populations:
     year            N      Crude     Adj_Rate      Confidence Interval
 --------------------------------------------------------------------------
     1953          1080   0.213889    0.213889    [  0.195970,    0.231808]
     1968          1232   0.148539    0.283919    [  0.256844,    0.310994]

. di 0.2139/0.2839
.75343431
```

(e) If we adjust to the 1968-72 rates, we find that the adjusted rate for 1953-57 is
[(0.918)(126)+(0.557)(112)+(**0.435**)(241)+(0.040)(601)]/1080
=**0.284**, or **28.4** deaths per 100 babies,
slightly <u>greater</u> than the 1953-57 rate (which was 21.4 per 100). This is the same as (b).

11. (a) 1940: 15820/131670 = 1.201 per 1000
1986: 469330/241097 = 1.947 per 1000
(1.947/1.201 = 1.62 ; 62% greater in 1986)

(b) 1986 population concentrated more in older age groups.

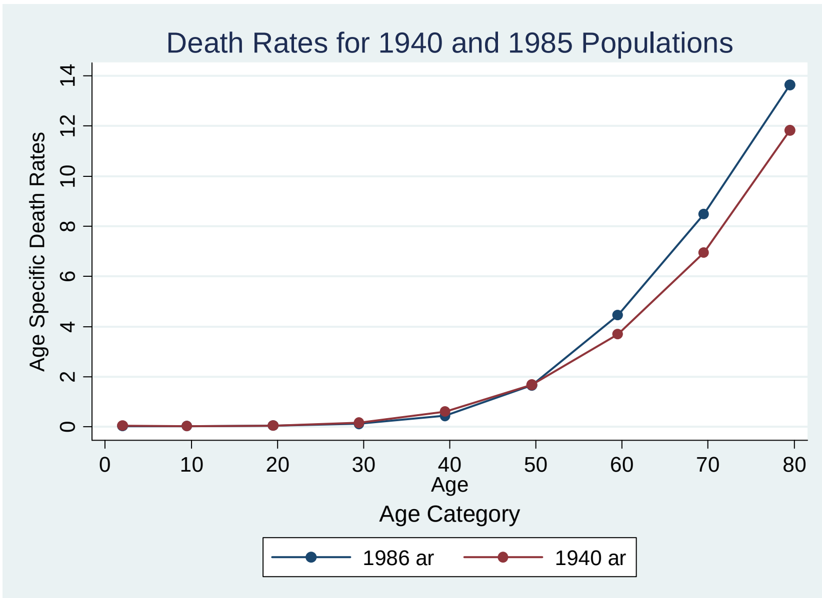
(c) There is a relationship – the death rate rises sharply in the older age groups.

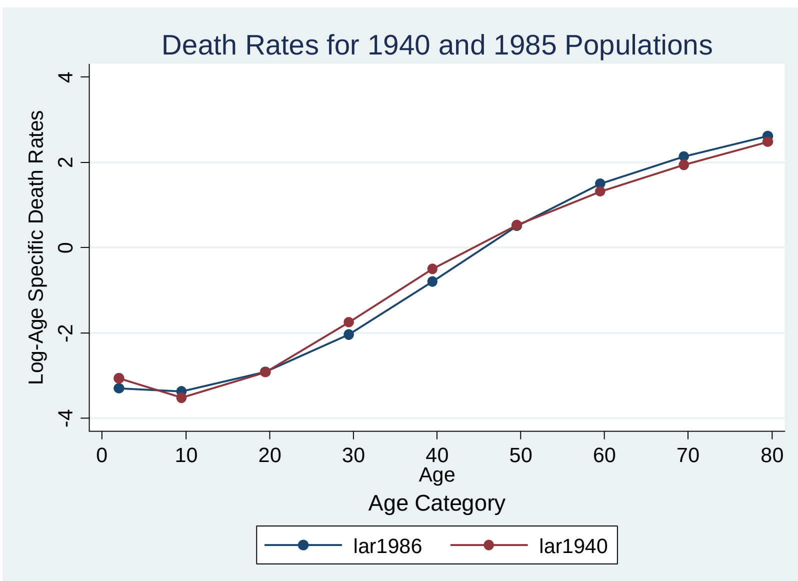
11. (d) Yes, because (i) rate changes with age, and (ii) the two populations have somewhat different age distributions.

(e) Age-adjusted rates: 1940 1.201 per 1000 (= crude rate)
1986 1.339 per 1000
(now only 11% greater than 1940)

(f) 1940 Age-adjusted rate = Crude
(because 1940 population was used as the standard)
1986 Age adjusted rate < Crude
(because rate increase with age, and 1940 population used as standard has “younger” age distribution.)

(g) The attached graphs show that the age-specific rate in the two populations follow the same overall pattern, increasing rapidly with age. For this reason it is appropriate to adjust. This is a situation where all “reasonable” choices of a standard population will give a ratio of adjusted rates (1986/1940) that is greater than 1, but much smaller than the ratio of crude rates. For example, we saw above that with the 1940 population as the standard, the ratio of adjusted rates is 1.11. If instead we use 1986 as the standard, the ratio of adjusted rates is 1.14. Both show 1986 is 10-15% higher, and both are much less than the ratio of crude rates (part (a)), where 1986 was 62% higher than 1940.





Graph one Stata commands:

```
. graph twoway connected ar1986 Age || connected ar1940 Age,
    ylab(0(2)14) xlab(0(10)80) title("Death Rates for 1940 and 1986
    Populations") l2("Age Specific Death Rates") b2("Age Category")
```

Graph two Stata commands:

```
. graph twoway connected lar1986 Age || connected lar1940 Age,
    ylab(-4(2)4) xlab(0(10)80) title("Death Rates for 1940 and 1985
    Populations") l2("Log-Age Specific Death Rates") b2("Age
    Category")
```

11. (h) SMR = (observed deaths)/(Expected deaths) = crude/IA rates.
SMR(1940)= 1 (reference) , SMR(1986) = 1.339 / 1.71=0.7821.

(i) The 1986 death rates increase faster than the 1940 rate in older age groups.

(j) Using 1940 rates as the standard:
Age-adjusted rates: 1940 1.201 per 1000 (= crude rate)
1986 1.71 per 1000
(now only 42% greater than 1940 but less than the original 62%)

Note: attached is my Stata log for this problem.

Stata log for problem 10 ps#1, (Pagano #15, p93)

```
use "E:\classes\BC213\data\pagano15.dta", clear

. * Ok this data I type in by hand
. list

        year       pop      dead      age
  1.    1940     10541       494        2
  2.    1940     22431       667      9.5
  3.    1940     23922      1287     19.5
  4.    1940     21339      3696     29.5
  5.    1940     18333     11198     39.5
  6.    1940     15512     26180     49.5
  7.    1940     10572     39071     59.5
  8.    1940      6377     44328     69.5
  9.    1940      2643     31279     79.5
 10.    1986     18152       666        2
 11.    1986     33860      1165      9.5
 12.    1986     39021      2115     19.5
 13.    1986     42779      5604     29.5
 14.    1986     33070     14991     39.5
 15.    1986     22815     37800     49.5
 16.    1986     22232     98805     59.5
 17.    1986     17332    146803     69.5
 18.    1986     11836    161381     79.5
```

. generate freq=pop/131670

. replace freq=pop/241097 if year==1986

```
(9 real changes made)

. * to get the age specific rates
. generate ar=dead/pop

. list

        year       pop      dead      age       freq        ar
  1.    1940     10541       494        2   .0800562   .0468646
  2.    1940     22431       667      9.5   .1703577   .0297356
  3.    1940     23922      1287     19.5   .1816815   .0537998
  4.    1940     21339      3696     29.5   .1620643    .173204
  5.    1940     18333     11198     39.5   .1392345   .6108111
  6.    1940     15512     26180     49.5   .1178097   1.687726
  7.    1940     10572     39071     59.5   .0802916   3.695706
  8.    1940      6377     44328     69.5   .0484317   6.951231
  9.    1940      2643     31279     79.5   .0200729   11.83466
 10.    1986     18152       666        2   .0752892   .0366902
 11.    1986     33860      1165      9.5   .1404414   .0344064
 12.    1986     39021      2115     19.5   .1618477   .0542016
 13.    1986     42779      5604     29.5   .1774348   .1309988
 14.    1986     33070     14991     39.5   .1371647   .4533111
 15.    1986     22815     37800     49.5     .09463   1.656805
 16.    1986     22232     98805     59.5   .0922118    4.44427
 17.    1986     17332    146803     69.5   .0718881   8.470056
 18.    1986     11836    161381     79.5   .0490923   13.63476
```

Stata log (continued)

```
. * But I cannot calculate adjusted rates easily with data in this form
. * check this out -- I'll reshape the data

. reshape groups year 1940 1986

. reshape vars ar freq dead pop

. reshape cons age

. reshape wide

. * You'll need to do this in Stat to get the full effect.
. list

Observation 1

      pop1986      18152    dead1986         666         age           2
     freq1986   .0752892      ar1986    .0366902     pop1940       10541
     dead1940        494    freq1940    .0800562      ar1940    .0468646


Observation 2

      pop1986      33860    dead1986        1165         age         9.5
     freq1986   .1404414      ar1986    .0344064     pop1940       22431
     dead1940        667    freq1940    .1703577      ar1940    .0297356


Observation 3

      pop1986      39021    dead1986        2115         age        19.5
     freq1986   .1618477      ar1986    .0542016     pop1940       23922
     dead1940       1287    freq1940    .1816815      ar1940    .0537998


Observation 4

      pop1986      42779    dead1986        5604         age        29.5
     freq1986   .1774348      ar1986    .1309988     pop1940       21339
     dead1940       3696    freq1940    .1620643      ar1940     .173204


Observation 5

      pop1986      33070    dead1986       14991         age        39.5
     freq1986   .1371647      ar1986    .4533111     pop1940       18333
     dead1940      11198    freq1940    .1392345      ar1940    .6108111


Observation 6

      pop1986      22815    dead1986       37800         age        49.5
     freq1986     .09463      ar1986    1.656805     pop1940       15512
     dead1940      26180    freq1940    .1178097      ar1940    1.687726


Observation 7

      pop1986      22232    dead1986       98805         age        59.5
     freq1986   .0922118      ar1986     4.44427     pop1940       10572
     dead1940      39071    freq1940    .0802916      ar1940    3.695706
```

```
Observation 8

     pop1986        17332    dead1986       146803         age          69.5
    freq1986     .0718881      ar1986     8.470056     pop1940          6377
    dead1940        44328    freq1940     .0484317      ar1940      6.951231


Observation 9

     pop1986        11836    dead1986       161381         age          79.5
    freq1986     .0490923      ar1986     13.63476     pop1940          2643
    dead1940        31279    freq1940     .0200729      ar1940      11.83466

. ** Neat huh?

. * ok now the data is in the correct form
. * that is, the populations are side by side like in the book
```

. *** for the direct adjustment**

```
. generate DA1986=ar1986*freq1940

. sum DA1986

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
  DA1986 |       9    .1487696   .1635541   .0029373   .4102191

. ** notice Dadj rate=9*0.1487696 = 1.3389
```

. *** similarly for indirect adjustment**

```
. generate IA1986=ar1940*freq1986

. sum IA1986

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
  IA1986 |       9     .190236   .2267338   .0035284   .5809903

. ** notice Iadj rate=9*0.190236 = 1.7121

. ** enjoy !!
```