3. Describing Univariate Distributions (II)

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

Last time

Describing univariate distributions: categorical and quantitative variables

- Frequency and relative frequency distributions
- Bar graphs (categorical) and histograms (interval/ratio)
- Stem-and-leaf plots, pie graphs
- Describing the shape (symmetry, skewness) of a distribution
- Measures of central tendency: mean, median, mode
- Properties of summation operator

This lecture

Describing univariate distributions, cont.

- Measures of variability ("dispersion" or "spread")
- Measures of skewness
- Quantiles, z-scores
- Box plots, interquartile range

Data transformations

- Linear and nonlinear transformations
- Effects on descriptive statistics

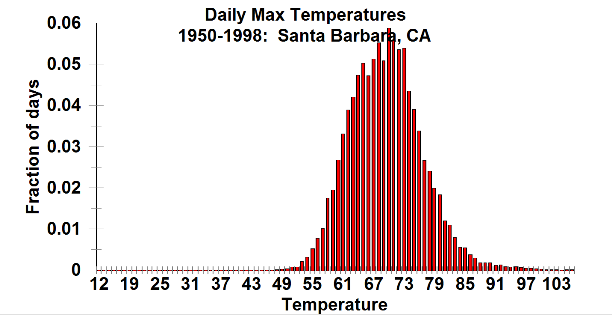
Measures of variability

Measures of **variability** are intended to capture how "spread out" the data points of the distribution are.

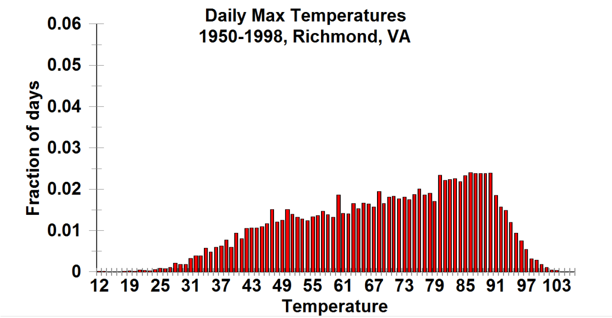
Two distributions can have the same measure of central tendency but very different dispersion. Example: consider mean daily high temperatures for two cities over 49 years:

- Richmond, VA: 69.0
- Santa Barbara: 69.5

Measures of variability



Measures of variability



Measures of variability

Common measures of variability, dispersion, or “spread” include:

- Range
- Variance
- Standard deviation
- Interquartile range (IQR)

Range

The **range** is the difference between the largest and smallest observed values in a distribution

```
. sum achmat08, detail

                 math achievement in eighth grade
-------------------------------------------------------------
      Percentiles      Smallest
 1%        38.55          36.61
 5%        41.89          37.14
10%       44.185           37.2       Obs                 500
25%        49.42          37.24       Sum of Wgt.         500

50%        56.18                      Mean           56.59102
                        Largest       Std. Dev.      9.339608
75%        63.74           77.2
90%       68.935           77.2       Variance       87.22827
95%        73.33           77.2       Skewness       .1133238
99%         77.2           77.2       Kurtosis       2.242742
```

For `achmat08`: 77.2 - 36.61 = 40.59

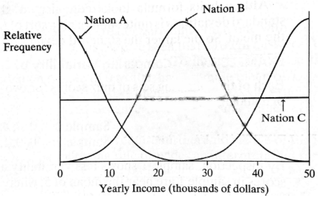
Range

The range is clearly sensitive to extreme values, since it is defined by the largest and smallest values.

```
. sum expinc, detail

                   expected income at age 30
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%        20000              0
10%        25000              0       Obs                 459
25%        30000              0       Sum of Wgt.         459

50%        40000                      Mean           51574.73
                        Largest       Std. Dev.      58265.76
75%        55000         250000
90%        80000         250000       Variance       3.39e+09
95%       100000         500000       Skewness       10.89864
99%       250000        1000000       Kurtosis       162.8027
```

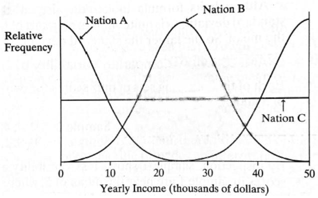

Range

Another disadvantage of the range is that it ignores all of the data beyond the maximum and minimum. In the following figure, the three income distributions have the same range (and mean):



Range

- Nation B: incomes tend to be close to the mean (low variability)
- Nation A: incomes tend to be far from the mean (high variability)



Variance

A more meaningful measure of variability would reflect distance from *every* data point to the mean. Consider the "average deviation from the mean":

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

Variance

Table 2.3: Deviations above and below the mean

	W_i	$(W_i - \bar{W})$	Deviation Totals
Tampa Bay	62	10	
Boston	62	10	
Oakland	61	9	
Baltimore	58	6	
Detroit	58	6	
Texas	56	4	
Cleveland	55	3	
NY Yankees	54	2	50
Kansas City	50	-2	
Seattle	49	-3	
LA Angels	48	-4	
Toronto	47	-5	
Minnesota	45	-7	
Chi White Sox	40	-12	
Houston	35	-17	-50

Variance

- Problem: the sum of all deviations from the mean *will always be zero*—the negative deviations from the mean will cancel out the positive deviations.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum x_i - \frac{1}{n}n\bar{x} = \bar{x} - \bar{x} = 0$$

- Recall characterization of the mean as a "center of gravity," where deviations above the mean are exactly balanced by deviations below the mean.

Variance

An alternative: the "average squared deviation from the mean," or variance (s^2):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- the numerator is the **sum of squares**
- the denominator is $n-1$ rather than 1 (explained later)

Variance

Table 2.4: Variance and standard deviation of A.L. wins

	W_i	$(W_i - \bar{W})$	$(W_i - \bar{W})^2$
Tampa Bay	62	10	100
Boston	62	10	100
Baltimore	58	6	36
NY Yankees	54	2	4
Toronto	47	-5	25
Detroit	58	6	36
Cleveland	55	3	9
Kansas City	50	-2	4
Minnesota	45	-7	49
Chi White Sox	40	-12	144
Oakland	61	9	81
Texas	56	4	16
Seattle	49	-3	9
LA Angels	48	-4	16
Houston	35	-17	289
SUM	-	0	918
Variance (s^2)	-	-	(918/14)=65.57
Std. Dev (s)	-	-	$\sqrt{65.57} = 8.1$

Note: $\bar{W} = 52$

Variance

- A variance of zero means there is no variation in x
- It is hard to interpret the variance in isolation, but it could be used to compare two distributions of a similar measure
- The magnitude of s^2 depends on the units of measurement. E.g.: variance of income in *cents* will be higher than the variance of income in *dollars*
- s^2 is hard to interpret (it is the average "squared deviation from the mean")

Standard deviation

The **standard deviation** (s) restores the variance measure to units of the original variable:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

- If it helps, think of this as the "average" or "typical" deviation of x from its mean (though this is <u>not</u> strictly correct).
- s^2 and s are always positive, unless there is no variance in x (in which case they are both zero)

Standard deviation

Table 2.4: Variance and standard deviation of A.L. wins

	W_i	$(W_i - \bar{W})$	$(W_i - \bar{W})^2$
Tampa Bay	62	10	100
Boston	62	10	100
Baltimore	58	6	36
NY Yankees	54	2	4
Toronto	47	-5	25
Detroit	58	6	36
Cleveland	55	3	9
Kansas City	50	-2	4
Minnesota	45	-7	49
Chi White Sox	40	-12	144
Oakland	61	9	81
Texas	56	4	16
Seattle	49	-3	9
LA Angels	48	-4	16
Houston	35	-17	289
SUM	-	0	918
Variance (s^2)	-	-	(918/14)=65.57
Std. Dev (s)	-	-	$\sqrt{65.57} = 8.1$

Note: $\bar{W} = 52$

Standard deviation

The variance and standard deviation are obtained using `summarize` with the `detail` option:

```
. sum achmat08, detail

                  math achievement in eighth grade
-------------------------------------------------------------
      Percentiles      Smallest
 1%        38.55          36.61
 5%        41.89          37.14
10%       44.185           37.2       Obs                 500
25%        49.42          37.24       Sum of Wgt.         500

50%        56.18                      Mean           56.59102
                        Largest       Std. Dev.      9.339608
75%        63.74           77.2
90%       68.935           77.2       Variance       87.22827
95%        73.33           77.2       Skewness       .1133238
99%         77.2           77.2       Kurtosis       2.242742
```

Can use `summarize` alone for standard deviation. Also, `tabstat` with `stat(sd)`

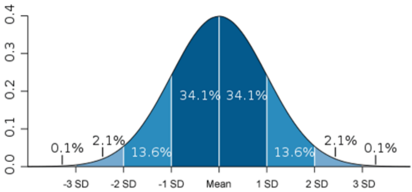
Standard deviation

```
. sum expinc, detail

                    expected income at age 30
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%        20000              0
10%        25000              0       Obs                 459
25%        30000              0       Sum of Wgt.         459

50%        40000                      Mean           51574.73
                        Largest       Std. Dev.      58265.76
75%        55000         250000
90%        80000         250000       Variance       3.39e+09
95%       100000         500000       Skewness       10.89864
99%       250000        1000000       Kurtosis       162.8027
```


Standard deviation

- The variance and standard deviation are especially sensitive to outliers/extreme values, since the variance formula *squares* deviations from the mean
- With normal (bell-shaped) distributions, s provides a good rule of thumb (the "Empirical Rule"):
  - About 68% of observations lie within 1 s.d. of the mean
  - About 95% of observations lie within 2 s.d. of the mean
  - Nearly all (99%) lie within 3 s.d. of the mean
  - Example: IQ mean of 100, standard deviation of 15

Standard deviation



Coefficient of variation

The **coefficient of variation** (CV) expresses the standard deviation as a percentage of the mean:

$$CV = \frac{s}{\bar{x}} * 100$$

- Must have $\bar{x} \neq 0$. If $\bar{x} < 0$, can use the absolute value of $\bar{x}$.
- The CV adjusts for the scale of units used, allowing for more appropriate comparisons. For example, the CV is often used to measure inequality in school spending across states or districts. Comparing s across these groups would be misleading if, say, the *level* of spending was higher in some districts/states than others.

Standard deviation

- Choice of dispersion measure should be made carefully—be aware of extreme values / outliers that may influence or distort the picture
  - The IQR is an alternative (shown later)

- One possibility when outliers exist: calculate statistics with outliers *dropped* (but be forthcoming about this—e.g. discuss both sets of results—and understand why outliers exist)

Skewness statistic

The **skewness statistic** is a summary of the positive or negative skewness of a distribution:

$$G = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^3}{s^3}$$

- Notice this is based on *cubed* deviations from the mean (can be positive or negative).
- Positive for positively skewed distributions and negative for negatively skewed distributions (zero for symmetric)

Skewness statistic

The **standard error of the skewness** is:

$$\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

- The concept of a standard error is introduced later
- For now: if the skewness statistic is more than 2x the standard error of the skewness, we can say the distribution is "significantly skewed"
- Can calculate yourself in Stata using `display`, or install the user-written `summskew` command (see Github for ado file).

Skewness statistic

The skewness statistic is reported using `summarize` with `detail`:

```
. sum achmat08, detail

                 math achievement in eighth grade
-------------------------------------------------------------
      Percentiles      Smallest
 1%        38.55          36.61
 5%        41.89          37.14
10%       44.185           37.2       Obs                 500
25%        49.42          37.24       Sum of Wgt.         500

50%        56.18                      Mean           56.59102
                        Largest       Std. Dev.      9.339608
75%        63.74           77.2
90%       68.935           77.2       Variance       87.22827
95%        73.33           77.2       Skewness       .1133238
99%         77.2           77.2       Kurtosis       2.242742
```

- *achmat08* has a slight positive skew
- Std err of skewness not reported here by Stata, but `display sqrt((6*500*499)/(498*501*503))` yields 0.109. The skewness statistic is *not* more than twice the std err of the skewness

Quantiles

Quantiles are cutpoints that divide a distribution into continuous intervals of equal size. Common quantiles include:

- **Quartiles**: 4 equally sized groups
- **Deciles**: 10 equally sized groups
- **Terciles**: 3 equally sized groups
- **Percentiles**: 100 equally sized groups

If there are n observations in a dataset, *quartiles* will include $n/4$ observations in each group. Note the term "quantile" is often used to refer to both the cutpoint and the groups themselves.

Quantiles

In real datasets, there may not be the right number of data points to divide into equally sized quantiles. For example, if $n = 10$, you can't divide the 10 observations into 4 equally sized groups. So what rule do you use?

If dividing the data into q groups, and you want to find the kth quantile ($k \leq q$), find the observation ranked $n * k/q$ when ranking the data points from smallest to largest.

- If this is an *fractional* value, take the next-largest ranked observation.
- If this is an *integer* value, take the midpoint of that ranked observation and the next-largest one.

Quantiles

Example: suppose $n = 74$

- 5th percentile is the 74*(5/100) = 3.7th obs (round to 4)
- 10th percentile is the 74*(10/100) = 7.4th obs (round to 8)
- 25th percentile is the 74*(25/100) = 18.5th obs (round to 19)
- Median is the 74*(50/100) = 37th obs (take avg of 37th and 38th)

This operationalizes the pth percentile as the smallest value that is greater than p% of the observations.

Percentiles

Can often determine percentiles from cumulative frequency distributions:

```
. tabulate famsize
```

family size	Freq.	Percent	Cum.
2	9	1.80	1.80
3	52	10.40	12.20
4	199	39.80	52.00
5	142	28.40	80.40
6	55	11.00	91.40
7	21	4.20	95.60
8	9	1.80	97.40
9	13	2.60	100.00
Total	500	100.00	

Percentiles

For quantitative variables, can use `summarize` with `detail` option in Stata to get common percentiles:

```
. sum achrdg08, detail

              reading achievement in eighth grade
-------------------------------------------------------------
      Percentiles      Smallest
 1%      37.525          35.74
 5%      40.705          35.82
10%       43.71          37.17       Obs                 500
25%      49.975          37.29       Sum of Wgt.         500

50%      56.445                      Mean           56.04906
                        Largest      Std. Dev.      8.829726
75%      63.195          70.55
90%       69.15          70.55       Variance       77.96406
95%       70.55          70.55       Skewness      -.1485071
99%       70.55          70.55       Kurtosis       2.191284
```

Alternative 1: `tabstat` with `stat(p25 p50 etc)`.

Alternative 2: create a variable that contains a percentile of *varname* using `egen`. E.g., `egen` *`varnamep10`*`=pctile(`*`varname`*`), p(10)`

Percentiles

A third alternative uses `centile` *`varname`*`, centile(`*`p`*`)`.

Note, however, that `centile` uses a linear interpolation rather than the "next highest rank" method. This will often provide a slightly different answer.

- `centile` will calculate the rank $P = (n+1)p/100$ where p is the desired percentile.
- If P is fractional, `centile` will interpolate between the two observations ranked $int(P)$ and $int(P)+1$.
- If P is an integer, it will use that ranked observation for the pth percentile.

Percentiles

The **five number summary** is a report of the minimum, Q1, median, Q3, and maximum. In Stata, use `summarize, detail` or `tabstat`:

```
. tabstat achrdg08 achmat08 achsci08 achsls08, stat(min p25 p50 p75 max)

   stats |  achrdg08  achmat08  achsci08  achsls08
---------+----------------------------------------
     min |     35.74     36.61     34.94      29.2
     p25 |    49.975     49.42     48.69     49.39
     p50 |    56.445     56.18      55.4     54.76
     p75 |    63.195     63.74     62.19      61.2
     max |     70.55      77.2     80.01      76.7
--------------------------------------------------
```


Percentiles

The pth percentile is sometimes defined as the value that is greater than or equal to p% of the observations (vs. "greater than").

This can make a difference in some datasets. In the limit—with continuous distributions—there will be no difference in these definitions.

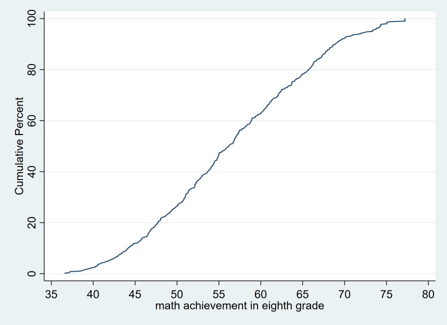
Empirical CDF ("ogive")

- A graph showing percentiles on the vertical axis and values of x on the horizontal axis is an **ogive**, or **empirical CDF** (cumulative distribution function)

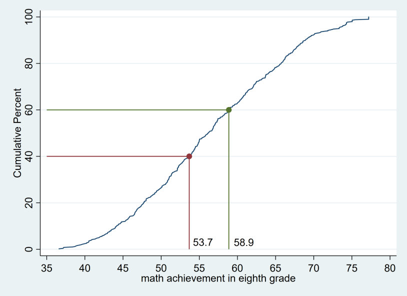
- This graph in Stata requires the extra step of creating a new variable with the cumulative relative frequency, using `cumul`:
  - `cumul achmat08, gen(cumachmat08)`
  - `replace cumachmat08=cumachmat08*100`
  - `sort cumachmat08`
  - `twoway line cumachmat08 achmat08, ytitle("Cumulative Percent")`

- Alternative: `ssc install distplot`

Empirical CDF ("ogive")



Empirical CDF ("ogive")



- The 40th percentile is approximately 53.7
- The 60th percentile is approximately 58.9

Z-scores

Percentiles are a measure of *position* in a distribution. Another is the **z-score** (or standardized score). The z-score for a particular value of x (x_i) is defined as:

$$z_i = \frac{x_i - \bar{x}}{s}$$

(Details later in this lecture).

Interquartile range

- The **interquartile range** is the difference between the upper and lower quartiles (75th and 25th percentiles)—a measure of dispersion that is robust to extreme values/outliers
- Pull quartiles from `summarize, detail` or `tabstat`, or use the `centile` command.
- You can also use the `iqr` stat option in `tabstat`

Interquartile range

```
. tabstat achrdg08 achmat08 achsci08 achsls08, stat(min p25 p50 p75 max)

   stats |  achrdg08   achmat08   achsci08   achsls08
---------+----------------------------------------
     min |     35.74      36.61      34.94       29.2
     p25 |    49.975      49.42      48.69      49.39
     p50 |    56.445      56.18       55.4      54.76
     p75 |    63.195      63.74      62.19       61.2
     max |     70.55       77.2      80.01       76.7
--------------------------------------------------
```

IQR = 61.2 – 49.39 = 11.81

IQR = 63.195 – 49.975 = 13.22

Interquartile range

Compare the IQR ($25,000) to s ($58,265)

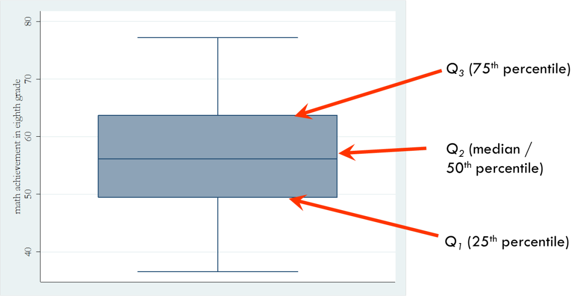
```
. sum expinc30, detail

                   expected income at age 30
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%        20000              0
10%        25000              0       Obs                 459
25%        30000              0       Sum of Wgt.         459

50%        40000                      Mean           51574.73
                        Largest       Std. Dev.      58265.76
75%        55000         250000
90%        80000         250000       Variance       3.39e+09
95%       100000         500000       Skewness       10.89864
99%       250000        1000000       Kurtosis       162.8027
```


Box plots

A **boxplot** (or box-and-whiskers plot) shows features of a variable's distribution: median, Q3, Q1, and extreme values (tails)

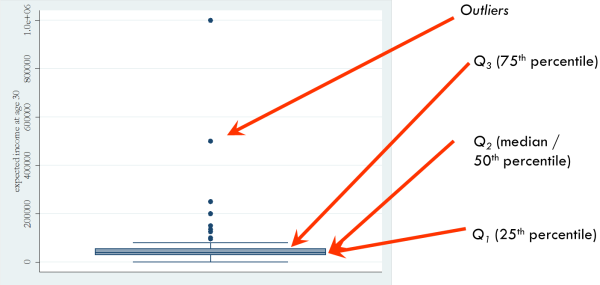
- The tails are "whiskers" that extend to the maximum and minimum in the data *if* there are no outliers (defined here as 1.5 IQR lengths above Q3 or below Q1.
- If there are outliers, the whiskers extend to the observation closest to (but not beyond) these thresholds (the highest and lowest **adjacent value**). Outliers are shown as points or asterisks (*).
- In Stata: `graph box` *`varname`*

Box plots



- Median=56.18, Q1=49.42, Q3=63.74, IQR=14.32
- Outliers would be above $(63.74 + (1.5 * 14.32)) = 85.22$ or below $(49.42 - (1.5 * 14.32)) = 27.94$

Box plots



- Outliers would be above $(55,000 + (1.5 * 25,000)) = \$92,500$ or below $(30,000 - (1.5 * 25,000)) = -\$7,500$
- Stata can suppress the outliers with the `nooutsides` option

Transforming variables

Transforming a variable simply means expressing it another way. Useful in many circumstances:

- Changing units of measure (e.g. feet to inches)
- Putting an observed value into the context of its distribution (e.g., "is 160 a high LSAT score?")
- Reduce the skewness of a variable—to work with a more symmetric variable
- Compare distributions on different scales
- Combining multiple variables into one index

Transforming variables

The z-score was one example of a transformation: from the units of the original scale to "standard deviation units"

- Example: Boston Red Sox have 62 wins, where the mean in the A.L. is 52. Their z-score is $(62 - 52)/8.1 = 1.235$. Their number of wins is $z = 1.235$ standard deviations above the mean.

Transforming variables

- Transformations can be linear or nonlinear; monotonic or non-monotonic.
- Monotonic transformations preserve the order of the data points, while non-monotonic transformations do not. We will only consider the former.

Transforming variables

A **linear transformation** involves only addition, subtraction, multiplication, or division, applied to the original variable x_0:

$$x_1 = a + (b * x_0)$$

- a can be positive (addition) or negative (subtraction)
- $b \neq 0$, but can be negative and $|b|$ can be < 1 (i.e., division)

Transforming variables

In Stata, `generate` (or `gen`) is the all-purpose command for creating new variables. Example: expected income, in thousands of dollars

```
. gen expinc30b = (expinc30 / 1000)
(41 missing values generated)

. sum expinc*

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    expinc30 |       459    51574.73    58265.76          0    1000000
   expinc30b |       459    51.57473    58.26576          0       1000
```

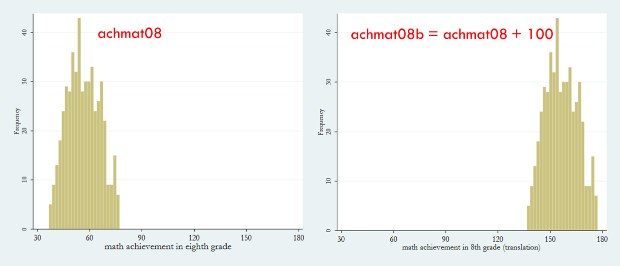

Translation

The simplest linear transformation is a **translation** where $b = 1$ and a is positive or negative:

$$x_1 = a + x_0$$

Shifts the distribution to the right or left. E.g. $achmat08r = achmat08 + 100$

Translation

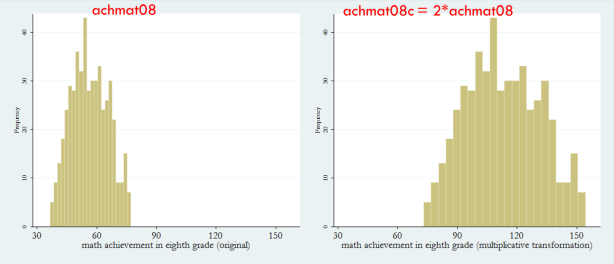


Multiplicative transformation

- The translation did not affect the *shape* or *spread* (variability) of the distribution, but did affect its position, shifting it to the left or right.
- Now consider a **multiplicative** transformation, where $a = 0$ but $b \neq 0$ and $b \neq 1$:

$$x_1 = \left(b * x_0\right)$$

Multiplicative transformation

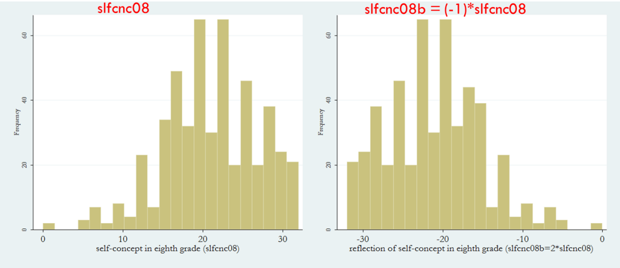


Multiplicative transformation

- The multiplicative transformation has affected *both* the location of the distribution *and* its spread.
  - The distribution is "stretched" whenever $|b| > 1$
  - The distribution is "compressed" whenever $|b| < 1$
- While the spread is affected, *relative* distance between points is preserved. *Shape* (symmetry, skewness) is preserved unless $b < 0$, in which case the distribution is flipped to its mirror image. A **reflection** multiplies the original variable by -1.

Multiplicative transformation

Reflection:



Transforming variables

How will a linear transformation of a variable affect its:

- Mean?
- Variance?
- Standard deviation?
- Range?
- IQR?
- Skewness?

Transforming variables

How a linear transformation affects the mean ($\bar{x}_{new}$ vs. $\bar{x}_{old}$):

$$\bar{x}_{old} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x}_{new} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i)$$

Transforming variables

How a linear transformation affects the mean ($\bar{x}_{new}$ vs. $\bar{x}_{old}$):

$$\bar{x}_{new} = \frac{1}{n}\sum_{i=1}^{n} a + \frac{1}{n}\sum_{i=1}^{n} bx_i$$

$$\bar{x}_{new} = a + b\bar{x}_{old}$$

Transforming variables

In other words, the mean of the new variable ($\bar{x}_{new}$) is equal to the old mean ($\bar{x}_{old}$) multiplied by b, and increased (or decreased) by a. Also true for median, mode, and percentiles.

Transforming variables

Example: use a linear transformation to express Fahrenheit temperature in Celsius.

- The conversion from F to C is: $x_C = -17.78 + 0.556 * x_F$ (a linear transformation)

- The mean daily high in Santa Barbara (in Fahrenheit) is $\bar{x}_F = 69.5$

- So then: $\bar{x}_C = -17.78 + 0.556 * 69.5 = 20.86$

Transforming variables

How a linear transformation affects the variance (s^2_{new} vs. s^2_{old}):

$$s^2_{old} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s^2_{new} = \frac{1}{n-1}\sum_{i=1}^{n}(a + bx_i - (a + b\bar{x}_{old}))^2$$

$$s^2_{new} = \frac{1}{n-1}\sum_{i=1}^{n}(bx_i - b\bar{x}_{old})^2$$

Transforming variables

How a linear transformation affects the variance (s^2_{new} vs. s^2_{old}):

$$s^2_{new} = \frac{1}{n-1}\sum_{i=1}^{n}b^2(x_i - \bar{x}_{old})^2$$

$$s^2_{new} = b^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_{old})^2$$

$$s^2_{new} = b^2 s^2_{old}$$

Transforming variables

- In other words, the variance of the transformed variable (s^2_{new}) is equal to the variance of the original variable (s^2_{old}), multiplied by b^2.
- The standard deviation of the transformed variable is:
$$s_{new} = \sqrt{s^2_{new}} = \sqrt{b^2 s^2_{old}} = |b| s_{old}$$
- Notice a does not affect the variance or standard deviation.
- The range and IQR of the transformed variable are equal to the range and IQR of the original variable, multiplied by $|b|$

Transforming variables

The skewness statistic is unaffected by a linear transformation *unless* b is negative. Then the skewness of the transformed variable is $(-1) * G_{old}$

Standardized variables

Standardized variables are an example of a linear transformation

In general a standardized variable applies a linear transformation to a variable to give it a predefined mean (c) and standard deviation (v)

	Variable	Mean
Original	x	$\bar{x}$
Transformed	$x - \bar{x} + c$	c

	Variable	Standard deviation
Original	x	s
Transformed	$\frac{v}{s}x$	v

Standardized variables

Examples of standardized variables:

- SAT sections: mean 500, sd 100
- GRE sections: 130-170, mean 150, sd 8
- IQ score: mean 100, sd 15
- MCAT section: formerly mean 8, sd 2; new overall mean 500
- z-score: mean 0, sd 1

Z-score transformation

It is easy to go from standard score to original scale score, given the mean and sd:

$$z = \frac{x - \bar{x}}{s}$$

$$x = \bar{x} + sz$$

Example:

- *achtmat08*: mean 56.59 and sd 9.34
- $z = 0.89$ would be $x = 56.59 + 9.34(0.89) = 64.9$

Z-score transformation

There are several ways to obtain z-scores in Stata. Manually, can obtain mean and sd, and then create the z-score using `gen`

```
. sum expinc30

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    expinc30 |        459    51574.73    58265.76          0    1000000

. sum achmat08

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    achmat08 |        500    56.59102    9.339608      36.61       77.2

. gen zachmat08 = (achmat08 - 56.59102) / 9.339608

. sum zachmat08

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   zachmat08 |        500    1.04e-08           1  -2.139385   2.206621
```

Can also use `egen` with `std` function

Z-score transformation

Alternatively, `sum` saves its results in "`r()`" macros:

```
. sum achmat08

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    achmat08 |        500    56.59102    9.339608      36.61       77.2

. return list

scalars:
                  r(N) =  500
              r(sum_w) =  500
               r(mean) =  56.59102010345459
                r(Var) =  87.22826930507507
                 r(sd) =  9.339607556266756
                r(min) =  36.61000061035156
                r(max) =  77.19999694824219
                r(sum) =  28295.5100517273

. gen zachmat08b = (achmat08 - r(mean)) / r(sd)

. sum zachmat08b

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  zachmat08b |        500   -9.77e-10           1  -2.139385   2.206621
```


Nonlinear transformations

- A *nonlinear* transformation involves some mathematical operation other than addition, subtraction, multiplication, and division
- There are lots of examples, but common ones include logarithmic (log), square root, inverse hyperbolic sine transformations

Nonlinear transformations

- Nonlinear transformations alter the shape (symmetry, skewness) of a distribution. In fact, this is often one of the purposes of a nonlinear transformation: to reduce the skewness in a distribution

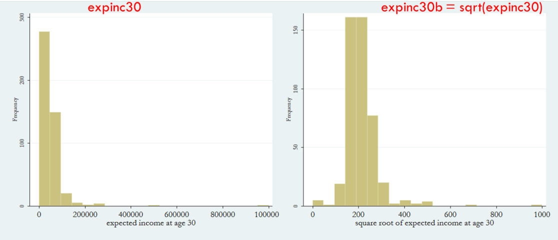
- Many statistical procedures assume the variable has a normal or approximately normal distribution. In practice this assumption is violated. Transformations can help.

Square root transformation

- Consider the square root transformation of expected income at age 30: `gen expinc30b=sqrt(expinc30)`

- Skewness statistics: $G_{old} = 10.898$, $G_{new} = 3.7$

Square root transformation



Square root transformation

- The square root transformation reduces the magnitude of large values of the original variable by more than it does small values. Examples:
- Original value of 4 becomes 2 (half the size)
- Original value of 100 becomes 10 (1/10 the size)

Square root transformation

- The square root is not defined for negative numbers. If the original variable has negative values, one can first apply a translation ($+a$) that ensures the minimum value is above zero (preferably equal to 1)
- Values between 0 and 1 become larger when applying the square root, unlike values above 1
- Log transformations have a similar effect on the distribution of a variable as the square root

Log transformation

- Recall that logarithms are exponents and depend on the base
- e.g. $\log_2 8 = 3$
- In this example, 2 is the base. $\log_2 8$ asks: "to what power does one take 2 to get 8?"
- The base can be any positive number other than 1

Log transformation

x value	Base 2	"log points"
2	$\log_2 2$	1
4	$\log_2 4$	2
8	$\log_2 8$	3
16	$\log_2 16$	4
32	$\log_2 32$	5
64	$\log_2 64$	6

x value	Base 10	"log points"
10	$\log_{10} 10$	1
100	$\log_{10} 100$	2
1000	$\log_{10} 1000$	3
10000	$\log_{10} 10000$	4

Log transformation

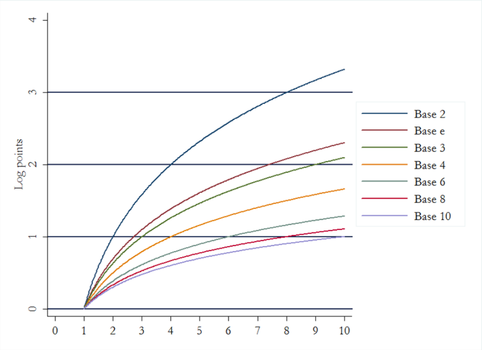
- On the $\log_2$ scale: each point is associated with 2 to that power
  - Values of 2, 4, 8, 16, ... → 1, 2, 3, 4
  - Larger values reduced more than smaller ones
- On the $\log_{10}$ scale: each point is associated with 10 to that power
  - Values of 10, 100, 1000, ... → 1, 2, 3
  - Transformation has a larger effect

Log transformation

- Notice it takes larger and larger increases in x to go up by one unit (one "log point"). The square root has a similar property.

- The "natural logarithm" is a commonly used log function, where the base is 2.718 (e). The natural log is denoted $\ln(x)$.

Log transformation

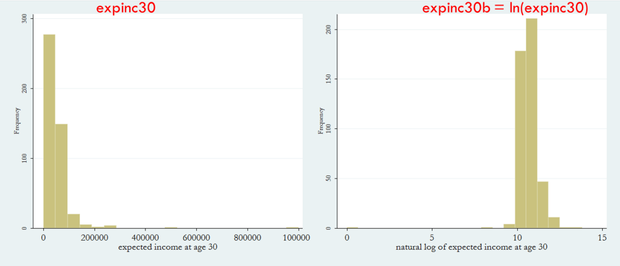


Log transformation

- Note the logarithm is only valid for $x \geq 1$. (The square root was only valid for $x \geq 0$)

- If x doesn't meet this condition, you can do a simple linear transformation first (e.g. add a constant a to all values such that the minimum $x \geq 1$.

Log transformation



Skewness statistics: $G_{old} = 10.898$, $G_{new} = -6.76$

Other types of transformations

Other types of transformations:

- "collapsing" or re-coding variables—creating a categorical variable from a continuous variable with many possible values
- ranking
- combining variables—e.g. a composite score (math + reading + social studies + science) or an index of multiple outcomes (Kling, Liebman, & Katz, 2007)

Note: be aware of how missing values affect your calculation!

Recoding into ordinal groups

Example: *unitmath*, units of math classes taken, ranges from 1-6 but has fractional values. Can set up groups/intervals

```
. table unitmath
```

units in mathematics (naep)	Freq.
1	1
1.5	2
1.75	1
2	36
2.38	1
2.5	10
2.75	1
2.81	1
2.97	1
3	121
3.5	29
3.6	1
3.63	1
3.96	1
4	229
4.08	1
4.29	1
4.5	22
4.99	2
5	31
5.5	2
5.99	1
6	4

Recoding into ordinal groups

```
. recode unitmath (1/1.99 = 1) (2/2.99 = 2) (3/3.99 = 3) (4/4.99 =
4) (5/100 = 5), gen(unitmathc)
(82 differences between unitmath and unitmathc)

. table unitmathc

----------------------
RECODE of |
unitmath  |
(units in |
mathemati |
cs        |
(naep))   |      Freq.
----------+-----------
        1 |          4
        2 |         50
        3 |        153
        4 |        255
        5 |         38
----------------------
```


Lecture 4

- Basic rules of probability; probability as relative frequency
- Conditional probability and Bayes' Rule
- Probability distributions: discrete and continuous
- Binomial distribution
- Normal distribution