Statistics is the art of summarization. It is an art because, for any given set of data, there are many ways of summarizing that data.

We'll focus on some fairly standard techniques for summarizing sets of data called:

Descriptive statistics

Descriptive statistics are usually a single number, whose purpose is to describe the data;

Examples include:

Range
Number of Observations
Average (mean)
Proportion (fraction, frequency)

As a general rule, there is not a statistical or probability model associated descriptive statistics.

This is good and bad. Why?

Summarization

Review: Types of data

- Categorical – Nominal and Ordinal
- Numerical – Discrete and Continuous

Different types of data require will require different descriptive statistics!

Example:

Proportion of females vs. Average Gender

Average Height vs. Proportion of height

There is no rule for determining which summary measure should be used. Use the context in which the data arose for a guide.

DATA: Male Life Expectancy by Country
(from 1993 Demographic Yearbook)

	Country	Years	**Notation**
1.	Canada	73.02	x_1
2.	Costa Rica	72.89	x_2
3.	Cuba	72.74	x_3
4.	United States	72.00	x_4
5.	Jamaica	71.41	x_5
6.	Bermuda	70.23	x_6
7.	Panama	69.78	x_7
8.	Bahamas	68.32	x_8
9.	Aruba	68.30	x_9
10.	Barbados	67.15	x_{10}
11.	Nicaragua	64.80	x_{11}
12.	Mexico	62.10	x_{12}
13.	Greenland	60.40	x_{13}
14.	Haiti	54.95	x_{14}
15.	El Salvador	50.74	x_{15}

Note: ***Lower case*** letters, such as x_9 or y_{22}, always represent observed data.

Measures of Central Tendency (Location)

1. Mean -- object (average -- verb)

Observations: $x_1, x_2, \ldots, x_n$

The Mean is defined as: $$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example: Mean Male Life Expectancy

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{15} x_i}{15}$$

$$= \frac{(73.02 + 72.89 + \cdots + 50.74)}{15} = 66.59$$

Note: The mean is *very* sensitive to the magnitude of the observations

Measures of Central Tendency (Location)

2. Weighted Average

(Observations: x_1, x_2, …, x_n)

Let $w_1,\ldots,w_n$ each be a number between 0 and 1, such that $w_1+\ldots+w_n=1$.

The weighted average is defined as

$$\bar{x} = \sum_{i=1}^{n} w_i \times x_i = \left(w_1 \times x_1 + \cdots + w_n \times x_n\right)$$

Some familiar weights:

- $w_i = 1/n$ gives the normal average (mean)

For the Male Life Expectancy data:

- $w_1,\ldots,w_{14} = 0$ and $w_{15} = 1$ gives the minimum
- $w_2,\ldots,w_{15} = 0$ and $w_1 = 1$ gives the maximum

Every average is a weighted average. We just leave off the 'weighted' phrase if all the weights are equal.

Measures of Central Tendency (Location)

3. Median

ORDERED observations: $x_1, x_2, \ldots, x_n$

$\Rightarrow$ Ordered implies $x_1 < x_2 < \ldots < x_n$

(Note: our data set is already ordered, but in the opposite direction, i.e. x_1 is the largest and x_n is the smallest)

The median is defined as the 50^{th} percentile of the observations.

'middle most number'
'half the data is below, half above'

n odd: Median is the $((n+1)/2)^{th}$ observation

n even: Median is the *average* of the $(n/2)^{th}$ observation and the $(n/2+1)^{th}$ observation

Example: Median Male Life Expectancy

15 is odd: the Median observation is x_8 = 68.32
{Bahamas b/c (15+1)/2=8}

Now throw out El Salvador (observation x_{15}) so the data set only has 14 observations.

Now the median is

14 is even: The median is $(x_7 + x_8)/2$ = 69.05

x_7 b/c 14/2=7, and

x_8 b/c (14/2+1)=8, and

(69.78 + 68.32)/2 = 69.05

Example: Median Male Life Expectancy (continued)

Mean is now: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{14} x_i}{14} = 67.72$

Unlike the mean, the median is *not* sensitive to the magnitude of the observations!

Regardless of magnitude for El Salvador (x_{15}), the Median Male Life Expectancy remains the same (in both data sets)!

However the Mean Male Life Expectancy depends heavily on the magnitude of the Male Life Expectancy in El Salvador (x_{15})

(Trick) Question:

Which do you prefer, mean or median?

Measures of Central Tendency (Location)

4. Centiles (percentiles)

ORDERED observations: $x_1, x_2, \ldots, x_n$

$\Rightarrow$ Ordered implies $x_1 < x_2 < \ldots < x_n$

(Note: our data set is already ordered, but in the opposite direction, i.e. x_1 is the largest and x_n is the smallest)

The z^{th} centile is the observation that is greater than z% of the data and less than (100-z)% of the data.

Rules for calculating the z^{th} centile can be found on page 44 of Pagano. But nobody does them, so just find a computer and make it do the work.

Stata command for percentiles:

centile *varname*, centile(10,20,33,78,94)

Example: Male Life Expectancy data in Stata

I entered the data by hand into stata

```
. list
          country          life
  1.    El Salvador       50.74
  2.    Haiti             54.95
  3.    Greenland         60.4
  4.    Mexico            62.1
  5.    Nicaragua         64.8
  6.    Barbados          67.15
  7.    Aruba             68.3
  8.    Bahamas           68.32
  9.    Panama            69.78
 10.    Bermuda           70.23
 11.    Jamaica           71.41
 12.    United States     72
 13.    Cuba              72.74
 14.    Costa Rica        72.89
 15.    Canada            73.02
```

```
. summarize life

Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
life     |      15    66.58867    6.81074      50.74      73.02
```

Example: Male Life Expectancy data in Stata

```
. centile life, centile(1,5,10,25,50,75,90,95,99)

                                                  -- Binom. Interp. --
Variable |     Obs  Percentile   Centile        [95% Conf. Interval]
---------+--------------------------------------------------------------
    life |      15           1         50.74       50.74       54.45343*
         |                   5         50.74       50.74       61.01954*
         |                  10        53.266       50.74       64.02599*
         |                  25          62.1       51.47319    68.31608
         |                  50         68.32       62.58105    71.89488
         |                  75            72       68.60591    72.99736
         |                  90         72.942       71.57914    73.02*
         |                  95         73.02        72.47032    73.02*
         |                  99         73.02        72.90533    73.02*

* Lower (upper) confidence limit held at minimum (maximum) of
sample
```

Actually the rules for the 25^{th} and 75^{th} percentile are easy:

25^{th} centile = ((n+1)/4)th observation
round up

75^{th} centile = (3*(n+1)/4)th observation
round down

Measures of Central Tendency (Location)

5. Mode

Observations: $x_1, x_2, \ldots, x_n$

The mode is defined as

the most frequent observation.

Example: Median Male Life Expectancy

As the data stand, each observation is unique,
so there is no mode (or each observation is a mode).

But if we round the Male Life Expectancy data to the nearest integer (see next page),

the mode is 73 ($x_1 = x_2 = x_3 = 73$)

Rounded DATA: Male Life Expectancy by Country (from 1993 Demographic Yearbook)

```
. generate rlife=round(life,1)

. list
```

	country	life	rlife
1.	Canada	73.02	73
2.	Coata Rica	72.89	73
3.	Cuba	72.74	73
4.	United States	72	72
5.	Jamaica	71.41	71
6.	Bermuda	70.23	70
7.	Panama	69.78	70
8.	Bahamas	68.32	68
9.	Aruba	68.3	68
10.	Bardados	67.15	67
11.	Nicaragua	64.8	65
12.	Mexico	62.1	62
13.	Greenland	60.4	60
14.	Haiti	54.95	55
15.	El Salvador	50.74	51

Measures of Dispersion (Variation)

1. Range

The range is defined as the difference between the maximum and minimum observation.

Example: Range of Male Life Expectancy

$$\text{Range} = x_1 - x_{15} = 73.02 - 50.74 = 22.28$$

2. Interquartile Range (IQR)

The Interquartile Range is defined as the difference between the 75^{th} centile observation and 25^{th} centile observation.

$$\text{IQR} = x_4 - x_{12} = 72 - 62.1 = 9.9$$

Measures of Dispersion (Variation)

2. Variance and Standard Deviation

The variance can be thought of as the *average* squared deviation of the observations from the sample mean.

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

However, the variance is in *squared x* units, unlike to the mean $\bar{x}$, which is in regular units of measurement.

Therefore, it is easier to talk about, $s = \sqrt{s^2}$, which is known as the **standard deviation (s)**. Now, units of measurement for s are the same as $\bar{x}$.

Example: Male Life Expectancy

```
. egen xbar=mean(life)
. generate diff=life-xbar
. generate diff2=diff^2
. list
        country        life        xbar        diff        diff2
  1.    Canada         73.02       66.59        6.43        41.36
  2.    Coata Rica     72.89       66.59        6.30        39.701
  3.    Cuba           72.74       66.59        6.15        37.84
  4.    United States  72          66.59        5.41        29.28
  5.    Jamaica        71.41       66.59        4.82        23.25
  6.    Bermuda        70.23       66.59        3.64        13.26
  7.    Panama         69.78       66.59        3.19        10.18
  8.    Bahamas        68.32       66.59        1.73        2.998
  9.    Aruba          68.3        66.59        1.71        2.929
 10.    Bardados       67.15       66.59        0.56        0.315
 11.    Nicaragua      64.8        66.59       -1.79        3.199
 12.    Mexico         62.1        66.59       -4.49        20.15
 13.    Greenland      60.4        66.59       -6.19        38.30
 14.    Haiti          54.95       66.59       -11.64       135.46
 15.    El Salvador    50.74       66.59       -15.85       251.18

. summ xbar diff diff2

Variable |     Obs        Mean    Std. Dev.       Min          Max
---------+--------------------------------------------------------
    xbar |      15    66.58867           0    66.58867     66.58867
    diff |      15   -1.53e-06    6.81074   -15.84867     6.431328
   diff2 |      15    43.29376    66.34464     0.31510     251.1803

. summarize life

Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
    life |      15    66.58867    6.81074        50.74      73.02
```

Notice: $s^2 = (6.811)^2 = 46.39 = 43.29376*(15/14)$

Example: Thinking about variability

Sample 1: 66, 66, 66, 67, 67, 67, 68, 69

$\bar{x} = 67$ s = 1.069

Sample 2: 52, 53, 61, 67, 71, 72, 78, 82

$\bar{x} = 67$ s = 10.98

Sample 3: 43, 44, 50, 54, 67, 90, 91, 97

$\bar{x} = 67$ s = 22.58

- All three samples have the same mean but *different* amounts of variability.

- Often a single summary measure will not do!

Measures of Dispersion (Variation)

3. Coefficient of Variation (cv)

The Coefficient of Variation relates the standard deviation to the mean.

$$cv = \frac{s}{\bar{x}}$$

Sometimes the CV is expressed in percentages.

Measures of Dispersion (Variation) for Grouped Data

Pagano and Gauvreau take some time to discuss how to calculate the 'mean' and 'variance' of grouped data.

I think this is a bad idea. We should not make a habit of treating categorical data as if it was continuous.

Example: Pagano page 51

Cholesterol Level	Midpoint	Number of Men
80-119	99.5	13
120-159	139.5	150
160-199	179.5	442
200-239	219.5	299
240-279	259.5	115
280-319	299.5	34
320-359	339.5	9
360-399	379.5	5
Total	Total	1067

Measures of Dispersion (Variation) for Grouped Data

Pagano averages the midpoint of the categories, pretending that the midpoints were actual observations. He gets a mean of **198.8**.

This is a dangerous practice because it implies that you have more information than you really do.

For example, instead averaging the midpoint of the categories, average the lower boundaries to a mean of **179.3**. If the upper boundary is averaged the mean is **218.3**.

So all we can really say is that the true mean is between **179.3** and **218.3**.

(Grouped variance has similar problems.)

Why not use the table itself as the summary statistic?

Moments

Moments are average deviations about the mean. They characterize or summarize the distribution.

$$m_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)}{n}$$

$$m_2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}$$

$$m_3 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^3}{n}$$

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

etc..

Why are moments useful?

1st moment (m_1)

Automatically sums to zero

$$m_1 = 0 \quad \text{(by definition)}$$

2nd moment (m_2)

- average squared deviation from the mean
- each terms is non-negative
- related to sample variance

$$S^2 = m_2 \times n/(n-1)$$

3rd moment (m_3)

- Average cubed deviation about the mean
- Each term negative or positive
- Does not automatically sum to zero
- Measures *symmetry* about the mean

When $m_3=0$, then the distribution is *symmetric* about the mean

Skewness is measured by the 3rd moment

$$\text{Skewness } (\gamma) = m_3 / (m_2)^{3/2}$$

(standardized so that skewness does not depend on the units of measure)

Variable	mean	median	IQR	SD	Skewness
DBP 24	61.8	61	8	7.6	0.1
FIRI	11.3	9.5	9.8	8.0	1.7
CPR	2.5	1.8	2.1	2.3	2.5