Standard Measures of Dispersion

Range

Variance

Standard deviation

Coefficient of variation

Interquartile Range

Graphical Display of Data

The Normal Curve

Normal curve areas

Departures from normality

Graphic methods of evaluation

Example: Larch Population

Pop-A & Pop-B Tree Height

Mean = 50, spread is 40-60 or 20-80

Standard Measures of Dispersion

Range

Range = \( Y_{\text{max}} - Y_{\text{min}} \)

Using Larch Tree Example:

Range_{\text{pop-A}} = 60 - 40 = 20

Range_{\text{pop-B}} = 80 - 20 = 60

Note: VERY sensitive to outliers!
Let's evaluate these data in R. We will first harvest some of the data manipulation power of R to 'create' pop-A & pop-B. There are 3 unique values in A (30,40,50): Create sub-groups a,b,c that contain the observation then concatenate all into one vector A:

```r
> a<-rep(40,c(5))
> a
[1] 40 40 40 40 40
> b<-rep(50,c(20))
> b
[1] 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50
[16] 50 50 50 50 50
> c<-rep(60,c(5))
> c
[1] 60 60 60 60 60
> A<-c(a,b,c)
> A
[1] 40 40 40 40 40 50 50 50 50 50 50 50 50 50
[16] 50 50 50 50 50 50 50 50 50 50 60 60 60 60 60
```

Likewise for B there are 7 groups (N = 54):

```r
> q<-rep(20,c(2))
> r<-rep(30,c(5))
> s<-rep(40,c(10))
> t<-rep(50,c(20))
> u<-rep(60,c(10))
> v<-rep(70,c(5))
> w<-rep(80,c(2))
> B<-c(q,r,s,t,u,v,w)
> B
[1] 20 20 30 30 30 30 30 30 30 30 30 30 30 30 30
[10] 40 40 40 40 40 40 40 40 40 50 50 50 50 50 50
[19] 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50
[28] 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50
[37] 50 60 60 60 60 60 60 60 60 60 60 60 60 60 60
[46] 60 60 70 70 70 70 70 70 70 80 80
```

Now, let's make a pair of histograms to display the data from pop-A and pop-B. For easy comparison, we want them above-below each other on the same X-scale. So we do the following:

1. create a 2 x 1 graph space
2. create each hist & define break points

```r
> par(mfrow=c(2,1))
> hist(A,breaks=c(10,20,30,40,50,60,70,80),
xlim=c(10,80),col="red")
> hist(B,breaks=c(10,20,30,40,50,60,70,80),
xlim=c(10,80),col="yellow")
> par(mfrow=c(1,1))
```
Histogram of A

Histogram of B

Now that we have populations A and B constructed, we can use the base functions built into R to assess the different measures of spread: min, max, and range.

> mean(A)
[1] 50
> min(A)
[1] 40
> max(A)
[1] 60
> range(A)
[1] 40 60
> mean(B)
[1] 50
> min(B)
[1] 20
> max(B)
[1] 80
> range(B)
[1] 20 80

Standard Measures of Dispersion

Variance

\[ S^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1} \]

Takes in to account all values, not just largest and smallest.

Computationally more intensive…
Important Concept

Sum of the squared deviations is known as:
**SUM-OF-SQUARES (or SS)**

$SS / N$ is known as:
**MEAN-SQUARES (or MS)**

* Hold this thought for 5 weeks, this will become a central notion of ANOVA

Standard Measures of Dispersion

**Variance**

Simplified "machine formula":

$$S^2 = \frac{\sum Y_i^2 - (\sum Y_i)^2}{n - 1}$$

Example: $S^2_{pop-A} = 34.48$, $S^2_{pop-B} = 181.13$

**Standard Deviation**

$$S = \sqrt{S^2}$$

$S$ equals the square root of the variance.

Example: $S_{pop-A} = 5.87$, $S_{pop-B} = 13.46$
Standard Measures of Dispersion

Coefficient of Variation

"It is well established that distributions with larger means have greater standard deviations."

Q: How can we then compare the dispersion of one group with the dispersion of another?
A: Compare relative amounts of variation (0-100%).

\[ CV = \frac{S}{\bar{X}} \times 100 \]

Example: CV \_pop-A = 11.74, CV \_pop-B = 26.91

Let's return to the tree snake example from the previous lecture (Ex. 3.1, p.60) and now provide a fuller univariate description of the data:

```r
> Hertz<-c(0.9,1.4,1.2,1.2, + 1.3,2.0,1.4,1.6)
> Hertz
[1] 0.9 1.4 1.2 1.2 1.3 2.0 1.4 1.6
> min(Hertz)
[1] 0.9
> max(Hertz)
[1] 2
> range(Hertz)
[1] 0.9 2.0
> var(Hertz)
[1] 0.106
> sd(Hertz)
[1] 0.3240370
> sd(Hertz)/mean(Hertz)*100 #CV
[1] 23.56633
```
Interquartile Range

You will recall that the second most used descriptor of central tendency was the **median**.

The median is defined as the middle observation of the sample data (i.e., 50th percentile).

The **interquartile range** (IQR) is a widely used measure of spread around the median. The IQR is the middle 50% of the data (i.e., 75th-25th percentile aka quantile). It is the difference between the first and third quartiles of the data. The IQR is most commonly displayed as a **box-plot**.

```r
> quantile(Hertz,.25)
  25%  1.2
> quantile(Hertz,.50)
  50%  1.35
> median(Hertz)
 [1] 1.35
> quantile(Hertz,.75)
  75%  1.45
> IQR(Hertz)
 [1] 0.25
> boxplot(Hertz)
```

**Box-and-whisker Plots**

- **Terminology** (Tukey 1980):
  - **Step** = 1.5 times IQR
  - **Outer Fence** = 2 steps outside hinge
  - **Inner Fence** = 1 step outside hinges
  - **IQR** = difference btw hinges
  - **Adjacent** = area btw IQR & inner fence
  - **Outside** = area btw inner & outer fence
  - **Far out** = area beyond outer fence
  - **Mild Outlier** = variate in outside area (•)
  - **Severe Outlier** = variate far out (♦)
Quick information is available more directly in R using one of two functions: `fivenum` or `summary`.

Fivenum produces the 5 numbers used in many EDA procedures (Q0, Q25, Q50, Q75, Q100).

Summary produces the same information and includes the mean.

```r
> fivenum(Hertz)
[1] 0.90 1.20 1.35 1.50 2.00
> summary(Hertz)
Min. 1st Qu. Median    Mean 3rd Qu.    Max.
0.900   1.200   1.350   1.375   1.450   2.000
```

Returning to our Larch tree example:

```r
> summary(A)
Min. 1st Qu. Median    Mean 3rd Qu.    Max.
40      50      50      50      50      60
> summary(B)
Min. 1st Qu. Median    Mean 3rd Qu.    Max.
20      40      50      50      60      80
> boxplot(A,B, horizontal=TRUE)
```
OUTLINE

Standard Measures of Dispersion
  Range
  Variance
  Standard deviation
  Coefficient of variation
  Interquartile Range

Graphical Display of Data
  The Normal Curve
    Normal curve areas
    Departures from normality
    Graphic methods of evaluation

Displaying Data
  - General Practices -

Classic references on topic:


Graphics Terminology and Graph Construction

Source: Cleveland 1984
Rules for Quality Graph Construction

1. Use 4 scale lines to encapsulate data (not 2)
2. Data space should NEVER intersect scale space
3. Label all axes and include units
4. Keep all data labels outside scale space
5. Keep scale line tick marks to outside
6. Do not overdo tick marks
7. Use a reference line to mark data separation

Rules for Quality Graph Construction

8. Captions need to be comprehensive & informative
   (Figures must “stand alone”)
9. Always use error bars when appropriate
   (Specify what error bars are used)
10. Be careful with aspect ratio (1:1 often best)
11. Minimize scale and maximize data space
    (do not insist on a 0,0 origin)
12. Use log scale to emphasize exponential change

Example Graphs
(Whitlock & Schluter, Chpt. 2)

A quick review of some of the major types of graphs for displaying your data:

Bar graph (simple or grouped)
Histogram*
Box-plot* (box-and-whisker plot)
Cumulative Frequency Distribution (CDF)
Mosaic plot
Scatter plot
Line plot
Map plots
Pie charts
Figure 2.3.1: Bar graph showing the 10 leading causes of death of American adolescents aged 15-19 in 1999. Total number of cases is 51,778. The frequencies are taken from Table 2.3.1.

Figure 2.3.2: Grouped bar graph for reproductive effort and avian males in great tits. The data are from Table 2.3.3, where n = 50 birds.

Figure 2.2.1: The cumulative frequency distribution of the abundance of bird species at Vögelnplatz Nature Reserve (all seasons). Horizontal dotted lines indicate the cumulative frequencies 0.05, 0.50, and 0.95. Vertical dashed line shows the corresponding abundance of bird species abundance 100 and 200. The data n = 40 bird species obtained from Table 2.2.1.
Figure 2.5.2: Mosaic plot for reproductive effort and axiakalulara in great tits. Yellow indicates birds with malaria, orange indicates birds free of malaria. The data are from Table 2.3.1, where n = 65 birds.

Figure 2.6.1: Scatter plot showing the relationship between the ornamentation of male guppies and the average attractiveness of their same-sex number of families n = 36.

Figure 2.6.2: Line graph of firm fur returns to the Hudson's Bay Company from 1752 to 1815.
The graphical method chosen is largely driven by the type of data you have. Summary:

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Graphical method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical data</td>
<td>Bar graph</td>
</tr>
<tr>
<td>Numerical data</td>
<td>Histogram</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Types of data</th>
<th>Graphical method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two categorical variables</td>
<td>Grouped bar graph, mosaic chart</td>
</tr>
<tr>
<td>One numerical variable and</td>
<td>Cumulative frequency distribution</td>
</tr>
<tr>
<td>one categorical variable</td>
<td>plot</td>
</tr>
<tr>
<td>Two numerical variables</td>
<td>Scatter plot, line plot</td>
</tr>
</tbody>
</table>
OUTLINE

Standard Measures of Dispersion
  Range
  Variance
  Standard deviation
  Coefficient of variation
  Interquartile Range

Graphical Display of Data

The Normal Curve
  Normal curve areas
  Departures from normality
  Graphic methods of evaluation

The “Normal” Curve

Basic Properties

• Unimodal
• Symmetrical around mean
• Asymptotic to axis (± ∞)
• Bell-shaped
• Area under the curve = 1
• Inflection points at μ ± σ (~ 2/3 of area)
• 99% of area defined by ± 3 σ

See Chapter 10, W&S text.
The Normal Distribution

Normal Probability Density Function:

\[ Y_i = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \]

\( Y_i \) = height of ordinate or density  
\( \mu \) = mean of distribution  
\( \sigma \) = SD of distribution

Thus, there are an infinite number of such NPDF’s.
Standard Normal Curve

Mathematical properties make this one of the most significant advances in all of statistics!

Deviation from the mean is measured in:
Standard Deviates
Expressing distance from the mean in units of $\sigma$:
Standard Normal Deviates

Standard Normal Deviates

$$Z = \frac{\bar{Y} - \mu}{\sigma}$$

Because of this relationship, if the mean and variance of a population is known, one can calculate a probability associated with an observation $Y$.

> rnorm(10)
[1]  0.4262908  1.6869712 -1.3739154 -1.3569666
[5]  0.2197391  0.1639324 -1.9726047  0.2423546
[9] -0.3706428  0.6827887

> rnorm(10,5,1)
[1] 5.291215 5.543051 3.764690 5.037652 3.858554

> x<-rnorm(100)
> hist(x,freq=FALSE, col='blue')
> curve(dnorm(x), add=TRUE)
Using SND to Determine Probabilities

- Example -

A biologist needs to build a trap to catch rabbits. From years of morphometric analysis, it is well established that cottontail rabbits are known to have:

- mean shoulder width = 3.80 in
- variance around mean = 0.36 in

Q: If the trap door is made to be 5.00 in wide, what percentage of rabbits will be able to make it through the door?

Example (continued):

First determine the standard normal deviate (SND).

\[ Z = \frac{X - \mu}{\sigma} = \frac{5.0 - 3.8}{0.6} = 2.00 \]

Go to Stat Table B (p. 672) and find the area under the curve at \( Z = 2.0 \) (defined as that point and to the right).

The area is 0.0228, thus the area to the left (rabbits) is:

\[ 1 - 0.02275 = 0.97725 (97.725\%) \]

This problem is a snap in R:

\[ > 1 - \text{pnorm}(5.00, \text{mean}=3.80, \text{sd}=0.6) \]
\[ [1] \ 0.02275013 \]
Departures from Normality

A true NDF is largely a theoretical construct.
Extremely large samples may approach a true NDF.
Practically, most statistical samples are not normal.

There are two primary measures of "shape-departure":
- Skewness is asymmetry about the abscissa.
- Kurtosis is vertical shape deflection.

Skewness

Skewness measures the amount of "tailage" or asymmetry.

A long tail to the left is **Negative Skew**.
A long tail to the right is **Positive Skew**.
Kurtosis

Measures the proportion of the distribution in the middle and tails relative to the shoulder.

A "flat-topped" distribution is **Platykurtic**.

A "pointy" distribution is **Leptokurtic**.

Moment Statistics

Both skewness and kurtosis can be quantified using central moment statistics (same as in physics).

The general form of a central moment is the average of the deviations of all items from the mean, each raised to the power of R:

\[
\left( \frac{1}{n} \right) \sum (Y - \bar{Y})^r
\]

Moment Statistics

The **first** central moment is zero by definition (\( \mu \)).

The **second** central moment is the variance (\( \sigma^2 \)).

The **third** central moment is the skewness (\( g_1 \)).

The **fourth** central moment is the kurtosis (\( g_2 \)).
**Moment Statistics**

- **Skewness** -

\[ g_1 = \frac{k_3}{\sqrt{\left(\overline{x}^2\right)^3}} \quad \text{where,} \]
\[ k_3 = \frac{n \sum X_i^3 - 3 \sum X_i \sum X_i^2 + 2 \left(\sum X_i^3\right)}{(n-1)(n-2)} \]

For a true N distribution, \( g_1 \) should be zero.

Negative values of \( g_1 \) are attained for left-skewed distributions.

Positive values of \( g_1 \) are attained for right-skewed distributions.

\(|g_1| > 1.0\) is problematic.

---

- **Kurtosis** -

\[ g_2 = \frac{k_4}{\sqrt{\left(\overline{x}^2\right)^2}} \quad \text{where,} \]
\[ k_4 = \frac{\left[n^4 \sum X_i^4 + 6n^3 \sum X_i^2 + n^2 \left(\sum X_i^2\right)^2 + 3n \sum X_i^2 \sum X_i^2 + n \left(\sum X_i^2\right)^2\right]}{n(n-1)(n-2)(n-3)} \]

For a true N distribution, \( g_2 \) should be zero.

Negative values of \( g_2 \) are attained for platykurtic distributions.

Positive values of \( g_2 \) are attained for leptokurtic distributions.

\(|g_2| > 1.0\) is problematic.

---

**Beta Measures of Symmetry & Kurtosis**

Some authors (e.g., D'Agostino et al. 1990) speak of a population parameter called \( \beta_1 \) as a measure of symmetry (where \( |\beta_1| = 0 \) indicates symmetry).

A parameter designated \( \beta_2 \) is often used as a measure of kurtosis (with \( |\beta_2| = 3 \) indicating non-kurtosis).
Beta Measures of Symmetry & Kurtosis

\[ \sqrt{b_1} = \frac{(N-2)g_1}{\sqrt{N(N-1)}} \]

and

\[ b_2 = \frac{(N-2)(N-3)g_2}{(N+1)(N-1)} + \frac{3(N-1)}{N+1} \]

Moments in R

Moment statistics are not directly available in the BASE package of R. Instead we have to add on a PACKAGE. This is a new and important feature that you need to learn in R...let's go do it.

INSTALL (from the CRAN) and LOAD (note that this is a 2-step process!) the package MOMENTS.

> moment(Hertz, order=1) # mean
[1] 1.375
> moment(Hertz, order=2) # SS
[1] 1.9825
> moment(Hertz, order=3) # aka K3
[1] 2.99575
> moment(Hertz, order=4) # aka K4
[1] 4.737025
> skewness(Hertz) # aka g1 (Ho=0)
[1] 0.6160647
> kurtosis(Hertz) # aka g2 (Ho=3)
[1] 3.079689
Because parametric statistical tests make distributional assumptions, we will discuss later the details of how to test whether your data are normal or not. After all, how non-normal does something have to be to be significantly non-normal?

This concludes the UNIVARIATE description of sample data (Chapters: 1, 2, 3, 10).

From here we will turn to estimation and then hypothesis testing.