Introductory Statistics Lectures

Measures of Variation

Descriptive Statistics III

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:48:30 2009)

Contents

1 Measures of Variation

1.1 Introduction

Histograms are important tools for image processing. By doing statistics on the individual pixel values we can analyze an image, attempt to recognize features, or just make it "look better".

center overall brightness of the image (mean pixel value)

variation image contrast

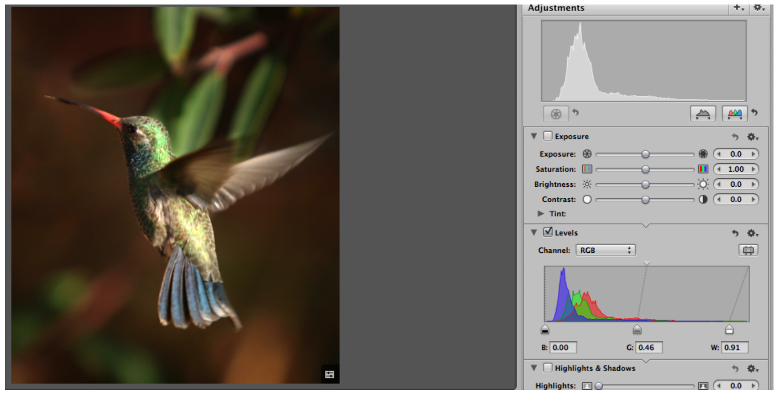
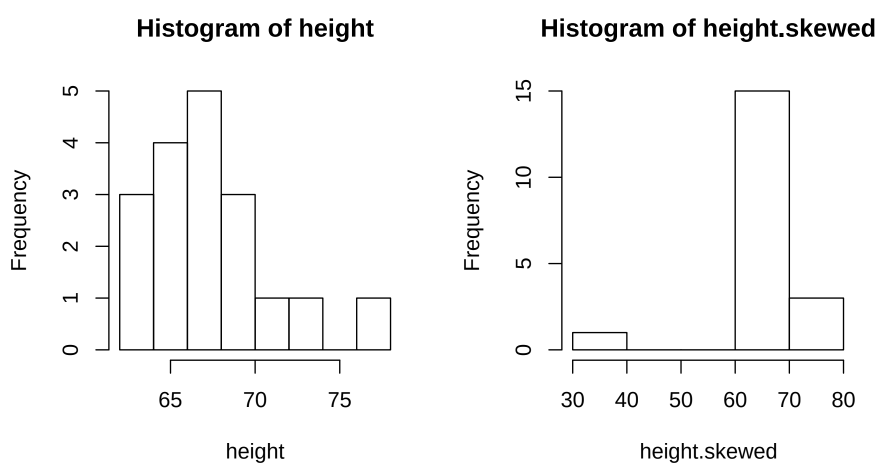


Figure 1: Image and its histogram.

Please welcome our special visitor to the class today:
Add Mini-Me (Verne Troyer) 2’ 8” to the class:

```
R: load("ClassData.RData")
R: height = class.data$height
R: height.skewed = c(height, 32)
R: par(mfrow = c(1, 2))
R: hist(height)
R: hist(height.skewed)
```



1.2 Measures of variation

Variation

Figure 2: Mini-Me. (2' 11").

How can we measure how much the data **varies** from the center? We need quantitatively describe the **dispersion** in the data.

RANGE

RANGE. DEFINITION 1.1

The max - min of a data set.

$$\boxed{\text{range} = \max - \min} \quad (1)$$

Easy to compute but very susceptible to outliers.

RANGE: R COMMAND

`max(x)-min(x)`

Where `x` is a vector. Note that typing `range(x)` in R will return the min and max values.

Need better measure of **average** variation from center.

Example 1. Range and skewed distributions:

```
R: max(height) − min(height)
```

```
[1] 15
R: max(height.skewed) - min(height.skewed)
[1] 45
```

STANDARD DEVIATION

DEFINITION 1.2

STANDARD DEVIATION: σ, s.
average variation from the mean value.

$$\text{standard deviation population: } \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \tag{2}$$

$$\boxed{\text{standard deviation sample: } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}} \tag{3}$$

Has same units as x.
Let $z = \{1, 2, 3\}$.

Question 1. Assume z represents a sample, find the standard deviation of z

Question 2. Assume z represents a population, find the standard deviation of z

R COMMAND

SAMPLE STANDARD DEVIATION:
`sd(x)`
Where `x` is a vector.

Question 3. If x represents a population how would we compute its standard deviation?

Example 2. Given that $x = \{9, 5, 6, 4\}$ find the standard deviation of x. This is easy in R!

```
R: x = c(9, 5, 6, 4)
R: sd(x)
[1] 2.1602
```

Example of standard deviation

Example 3. Effect of skewed distributions on standard deviation

```
R: sd(height)
[1] 3.8370
R: sd(height.skewed)
[1] 8.9805
```

Question 4. Use R to find the **sample** standard deviation of $x = \{1, 3, 14\}$

VARIANCE. DEFINITION 1.3

The square of the standard deviation.

population σ^2

sample s^2

Unbiased measure of sample variation. Does not have same units as x.

SAMPLE VARIANCE: R COMMAND

`var(x)`

Where `x` is a vector.

Example 4. Given that $x = \{9, 5, 6, 4\}$ find the sample variance of x. This is easy in R!

Can either use square of standard deviation, or use var(x).

```
R: x = c(9, 5, 6, 4)
R: s = sd(x)
R: s^2
[1] 4.6667
R: var(x)
[1] 4.6667
```

Quickly estimating s

RANGE RULE OF THUMB. DEFINITION 1.4

Quick rough estimate of standard deviation:

$$\boxed{\text{range rule of thumb: } s \approx \frac{\text{range}}{4}} \tag{4}$$

Question 5. Use the range rule to estimate s for $x = \{9, 5, 6, 4\}$

EMPIRICAL RULE

DEFINITION 1.5

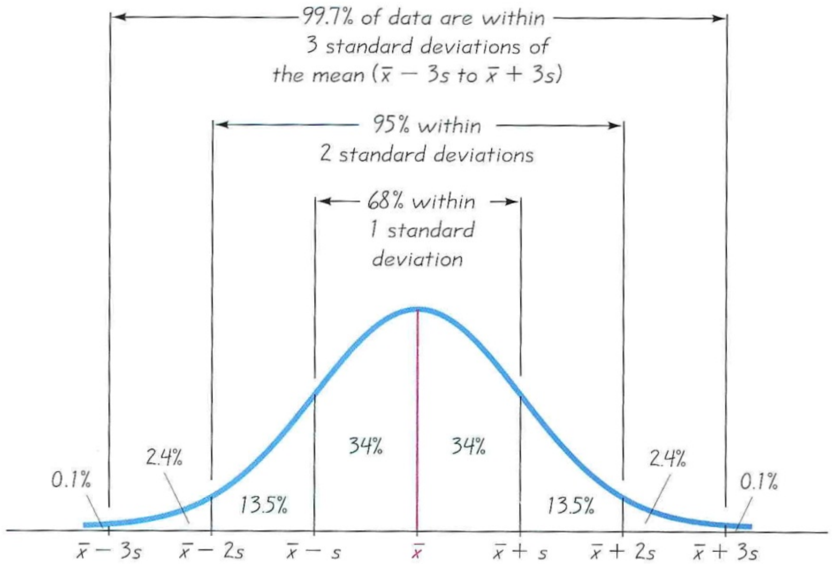
EMPIRICAL RULE.

For data that has an approximately bell shaped distribution:
68% of values fall within $\bar{x} \pm s$
95% of values fall within $\bar{x} \pm 2s$
99.7% of values fall within $\bar{x} \pm 3s$

Know the Empirical Rule!

Question 6. If a population has $\mu = 25$ and $\sigma = 2$, what interval contains approximately 95% of the population's values?

Question 7. If a population has $\mu = 25$ and $\sigma = 2$, what interval contains approximately 99.7% of the population's values?



CHEBYSHEV'S THEOREM. DEFINITION 1.6

The proportion p (fraction) of **any** set of data lying within K standard deviations of the mean is:

$$\text{Chebyshev's Theorem: } p \geq 1 - \frac{1}{K^2} \tag{5}$$

Example 5. Empirical rule states that 0.997 of data lies within 3 standard deviations, Chebyshev's theorem states that for all data sets at least the following proportion is within 3 standard deviations:

```
R: K = 3
R: 1 - 1/K^2
[1] 0.88889
```

Interpreting standard deviation

USUAL VALUES. DEFINITION 1.7

within $\bar{x} \pm 2s$. 95% of values in bell shape dist.

UNUSUAL VALUES. DEFINITION 1.8

outside $\bar{x} \pm 2s$. 5% of values in bell shape dist.

COMPARING VARIATION: COEFFICIENT OF VARIATION

COEFFICIENT OF VARIATION: CV. DEFINITION 1.9

standard deviation normalized by the mean. (unit-less)

$$\text{sample: } CV = \frac{s}{\bar{x}} \tag{6}$$

Useful for comparing variation.[1]

Question 8. Find the coefficient of variation for $x = \{1, 3, 14\}$

Example 6. For our class data, is there more variability in student heights or student work hours?

```
R: cv.height = sd(height)/mean(height)
R: cv.height
[1] 0.05675
R: cv.work = sd(class.data$work_hours)/mean(class.data$work_hours↩
   )
R: cv.work
[1] 0.8084
```

Question 9. Which has more variability, student heights or student work hours?

1.3 Summary

1. Standard deviation: σ, s. Variance is standard deviation squared: s^2.
2. Empirical Rule: $\bar{x} \pm s$: 68%, $\bar{x} \pm 2s$: 95%, $\bar{x} \pm 3s$: 99.7%,
3. Usual values: within $\bar{x} \pm 2s$
4. Coefficient of variation:
$$CV = \frac{s}{\bar{x}}$$

A measure of center is not enough, you need to know the **variation** in the population.

[1] The CV is often expressed as a percentage, however it is not limited to the range of 0-100%. Therefore, it makes more sense to leave it as a coefficient.