

Variability

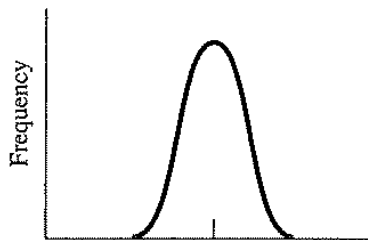
MSc Module 6: Introduction to Quantitative Research Methods
Kenneth Benoit

February 10, 2010

The concept of variability

- ▶ Central tendency only characterizes the *central point* of a distribution
- ▶ Beyond that, however, any distribution also has a certain *spread* or dispersion of values that exists independently from its central tendency
- ▶ This concept is referred to as the variability or *variance* of a distribution
- ▶ Illustration (from LFF) is between the distributions of daily temperatures between the cities of Honolulu and Pheonix. Both have mean temperatures of 75°F , but Pheonix's temperature is *much more variable*

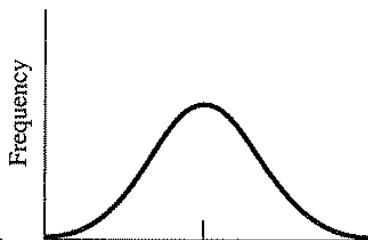
Variability: temperature example



$$\bar{X} = 75^{\circ}$$

$$R = 24^{\circ}$$

Honolulu



$$\bar{X} = 75^{\circ}$$

$$R = 65^{\circ}$$

Phoenix

where R refers to the range of temperatures

Range

- ▶ Range is very simple: it refers the difference between the highest and lowest values in a distribution
- ▶ Formula: $\max(X) - \min(X)$
- ▶ Advantage: *very simple* to compute and to understand
- ▶ Disadvantage: totally dependent on two (extreme) values in your data sample
- ▶ Note: There is a difference between the *empirical range* from a sample, and the *theoretical range* from a variable.
 - ▶ Example: Grades (as a percentage for instance)
 - ▶ Example: Votes (where maximum is defined by the electorate)

Percentiles

- ▶ Percentiles are closely related to the notion of the range, but they assume the data are (and can be) *sorted*
- ▶ Assume a sorted variable X for which $N = 100$
- ▶ The p th percentile is then simply the p th value
- ▶ For N not a factor of 100, interpolation must be computed. No “standard” method for this but frequently:

$$p\text{th percentile} = \frac{p}{100}(N + 1) + 1$$

Percentiles

- ▶ Some percentiles have special designations
 - 1st quartile is the 25th percentile
 - 3rd quartile is the 75th percentile
 - median is the 50th percentile or 2nd quartile
 - deciles describe every 10th percentile, e.g. 90th %ile = 9th decile
- ▶ Applies to any ordinal data, not just interval data, but most commonly used for interval data

Percentiles in R

```
> ## percentiles illustrated
> x <- (round(runif(100)*100)) # create 100 random values 0-100
> xsorted <- sort(x)         # sort the values
> xsorted[c(1,25,50,75,100)] # min; 25th, 50th, 75th %tile; max
[1] 0 28 48 68 100
> summary(xsorted)          # ask R for same information
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.0   28.0   48.0   48.4   68.5   100.0
> summary(x)                # sorting makes no difference
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.0   28.0   48.0   48.4   68.5   100.0
> quantile(x,0.25)         # 25th percentile using quantile()
25%
28
> quantile(x,0.95)        # 75th percentile
95%
88.1
> quantile(x,0.50)        # 50th percentile (median)
50%
48
> median(x)               # median (50th percentile)
[1] 48
> # compute position of 25th %tile using formula
> (p <- (25/100)*(length(xsorted)+1))
[1] 25.25
> x[floor(p)]             # the position before p
[1] 64
> x[ceiling(p)]          # the position after p
[1] 34
```

Relationship of variance to deviations

- ▶ Recall from last week that means can be expressed as *deviations*: $X_i - \bar{X}$
- ▶ One way to assess deviation would be to sum all of the deviations. . .
- ▶ . . . but then the sum would have to be zero — how to get around this?
 - ▶ one way would be to take the *absolute value* of the deviations and sum them: $|X_i - \bar{X}|$
 - ▶ another way would be to *square* the deviations and sum them: $(X_i - \bar{X})^2$

Relationship of variance to deviations

- ▶ We need to somehow rescale this to something comparable across different distributions — so we can look at the size of the *average* (squared) deviation by dividing the sum by N
- ▶ We would then have measured the mean of the (squared) deviations from the mean

Relationship of variance to deviations

- ▶ This defines the **variance** or s^2 :

$$s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

- ▶ Example:

TABLE 4.1 *Squaring Deviations* ($\bar{X} = 5$)

X	$X - \bar{X}$	$(X - \bar{X})^2$
9	+4	16
8	+3	9
6	+1	1
4	-1	1
2	-3	9
1	-4	16
	<hr/>	<hr/>
	0	$\sum (X - \bar{X})^2 = 52$

- ▶ Example in R:

```
> x <- c(9,8,6,4,2,1)
> xdev <- x - mean(x)
> sum(xdev^2)/length(x)
[1] 8.666667
```

From variance to *standard* deviations

- ▶ Problem: Since this was the average sum of *squared* deviations, we are not really sure what metric the answer is in
- ▶ For instance what does 8.67 mean?
- ▶ To return to our original metric, we can take the *square root* of the variance, which yields the **standard deviation** or **s**:

$$s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}}$$

- ▶ From the previous example, $s = \sqrt{8.67} = 2.94$
> (s <- sqrt(s2))
[1] 2.94392

Computational formulas for s^2 and s

- ▶ The deviation method is fine if we have R, but there is an easier method that does not rely on first calculating the deviations

- ▶ For variance:

$$s^2 = \frac{\sum_{i=1}^N X_i^2}{N} - \bar{X}^2$$

- ▶ For the standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^N X_i^2}{N} - \bar{X}^2}$$

- ▶ Example in R (using previous example):

```
> sqrt(sum(x^2)/length(x) - mean(x)^2)
[1] 2.94392
```

The notion of degrees of freedom

- ▶ Suppose we have a sample X of 5 numbers where $\bar{X} = 4$

--	--	--	--	--

- ▶ We know that the sum must be 20, since $\bar{X} = \sum_i X_i / N$
- ▶ Now let's try to assign numbers to each box. The first number could be any number at all. Suppose we choose 2:

2				
---	--	--	--	--

- ▶ Now choose a second number, say 7:

2	7			
---	---	--	--	--

- ▶ Now choose 4 and 0 for third and fourth numbers:

2	7	4	0	
---	---	---	---	--

- ▶ **Question:** What must the last value be?

Degrees of freedom (continued)

- ▶ Once the first four values have been chosen, the last value can only be one number: since the sum must be 10, then

$$X_5 = \sum_{i=1}^N X_i - \sum_{i=1}^{N-1} X_i$$

- ▶ In this case: $20 - (2 + 7 + 4 + 0) = 7$

2	7	4	0	7
---	---	---	---	---

- ▶ With N numbers, we had only $N - 1$ numbers that were free to vary
- ▶ We call this concept **degrees of freedom**, calculated as $N - 1$ for a sample mean
- ▶ **More formally:** Degrees of freedom is the sample size N minus the number of parameters p , or **$df = (n - p)$**

The “sample” variance

- ▶ In practice (in finite samples) we generally compute the variance by dividing not by N , but by $(N - 1)$: the degrees of freedom
- ▶ Very important general formula is then:

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

- ▶ When divided by degrees of freedom instead of N , the quantity is known as the *sample variance*
- ▶ The sample variance is generally used instead of the population variance formula (given previously) because the population variance formula is biased in finite samples

The “sample” variance and standard deviation

- ▶ Typically this is denoted s^2 , although Levin and Fox use s^2 for the population variance
- ▶ Most statistical packages provide s^2 by default, for instance R's `var()` command
- ▶ Sample variance formula is:

$$s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1}$$

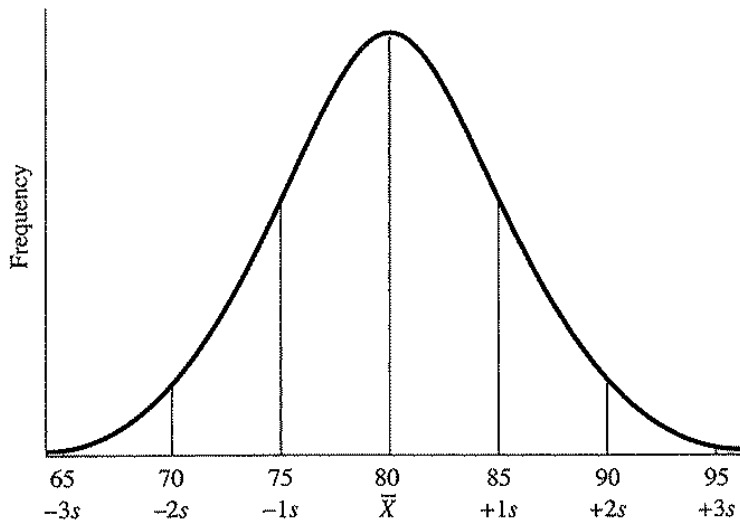
- ▶ As before, $\sigma = \sqrt{\sigma^2}$

```
> var(x)
[1] 10.4
> sqrt(var(x))
[1] 3.224903
> sd(x)
[1] 3.224903
```

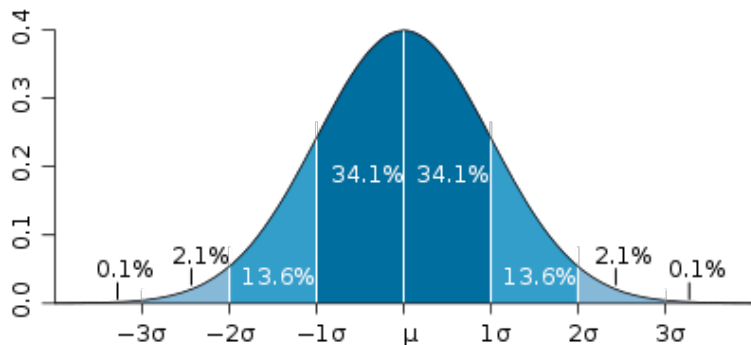
Interpreting standard deviations

- ▶ Standard deviations have the advantage that they are in units of the variable whose distribution we are interested in (unlike the variance)
- ▶ So if our quantity is votes, for instance, we can interpret the SD in terms of votes
- ▶ But the notion of *standardizing* goes even farther: it can be used to compare different distributions because it is standard
- ▶ Let's say we have a distribution X for which $\bar{X} = 80$ and $s = 5$
- ▶ Then 85 is exactly $1s$ above the mean, or a distance of $+1s$
- ▶ Likewise, the value 75 would be $-1s$ below the mean
- ▶ We could continue this reasoning by marking off subsequent deviations in increments of $s = 5$, both positive and negative

Interpreting SDs: example



Interpreting SDs: example



- ▶ Middle two standard deviations represents approximately two-thirds (68.1%) of the area under the normal curve
- ▶ Middle four standard deviations represent approximately 95% (95.4%) of the area under the curve
- ▶ Middle six, or $\pm 3s \approx 99\%$

Standard error of the mean

- ▶ Variance is typically used to *test hypotheses* by making it possible to compare numbers from different distributions
- ▶ This is the cornerstone of statistical analysis: being able to make inferences about a statistic (a quantity computed from sample data) relative to some assumed parameter value
- ▶ This measure is proportional to s^2 , and inversely proportional to N

$$\text{unreliability} \propto \frac{s^2}{N}$$
$$SE_{\bar{x}} = \sqrt{\frac{s^2}{N}}$$

- ▶ This quantity $SE_{\bar{x}}$, is known as the **standard error of the mean**

Interpreting box plots: Beware LFF Figure 4.4

Males: 5 2 7 9 3 4 3 1 3 8
Females: 3 5 7 4 5 6 7 6 5 4

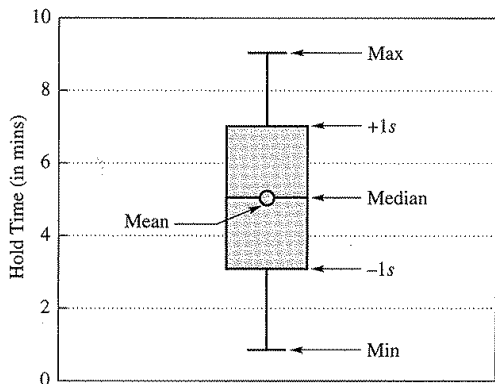


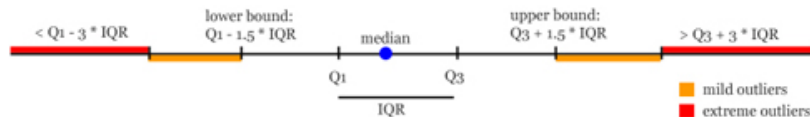
FIGURE 4.4 Box plot of hold time distribution

Normally this is NOT how you would interpret the elements of a box plot

Interpreting box plots

- ▶ at least 25% of all values are below the lower quartile Q1.
- ▶ at least 50% of all values are below (or above) the median.
- ▶ at least 25% of all values are above the upper quartile Q3.
- ▶ The box contains 50% of the data ($Q3 (75\%) - Q1(25\%) = 50\%$).
- ▶ You can read from the size of the box, the distance of the whiskers the distribution of the values.
- ▶ Between the median and the quartiles are 25% of the data ($75\% - 50\% = 25\%$ and $50\% - 25\% = 25\%$), i.e. the position of the median inside the box indicates whether there are more values towards the upper or lower quartile.
- ▶ Not to mention the outliers, which are those values, that are far away from most of the other values.

Interpreting box plots continued



The whiskers mark those values which are minimum and maximum unless these values exceed $1.5 * IQR$. The IQR is the inter quartile range: the distance between Q_1 and Q_3 . If there are observations which are outside $1.5 * IQR$ or even $3 * IQR$ then they are considered as mild and extreme outliers, respectively.