Topic 2
Describing Distributions with Numbers
Measuring Spread

Outline

Five Number Summary

The five number summary is

- the minimum
- the first quartile, Q_1.
  - the median of the lower half of the data,
- the median
- the third quartile, Q_3,
  - the median of the upper half of the data
- and the maximum

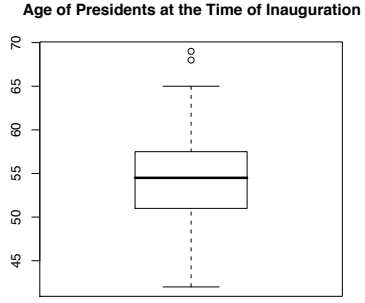
These values, along with the mean, are given in R using `summary(x)`. Returning to the data set on the age of presidents:

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  42.00   51.00   54.50   54.64   57.25   69.00
```

Boxplots

We can display the five number summary using a boxplot

```
> boxplot(age, main = c("Age of Presidents at the Time of Inauguration"))
```



Define the interquartile range $IQR = Q_3 - Q_1$. Then R sets outliers as those observations at lest $\frac{3}{2}IQR$ above Q_3 or below Q_1.

Sample Variance and Standard Deviation

The sample variance averages the square of the differences from the mean

$$\text{var}(x) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{x}(x - \bar{x})^2 n(x).$$

The sample standard deviation, s_x, is the square root of the sample variance. The standard deviation has the same units as the observations.

NB. We shall soon learn the issues involved with division by $n-1$ rather than n.

Sample Variance and Standard Deviation

Example. For the data set on *Bacillus subtilis* data, $\bar{x} = 498/200 = 2.49$.

length x	frequency $n(x)$	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 n(x)$
1.5	18	-0.99	0.9801	17.6418
2.0	71	-0.49	0.2401	17.0471
2.5	48	0.01	0.0001	0.0048
3.0	37	0.51	0.2601	9.6237
3.5	16	1.01	1.0201	16.3216
4.0	6	1.51	2.2801	13.6806
4.5	4	2.01	4.0401	16.1604
sum	200			90.4800

So the sample variance $s_x^2 = 90.48/199 = 0.4547$ and standard deviation $s_x = 0.6743$.

Sample Variance and Standard Deviation

Exercise. For the experiment of flipping a fair coin 16 times, our instinct is that the mean number of heads is 8. Here we compare the two sums of squares

$$SS(\bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad SS(8) = \sum_{i=1}^{n}(x_i - 8)^2$$

on the following 25 outcomes of this experiment:

5 5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9 9 9 10 10 10 11 12 13

x	$n(x)$	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 n(x)$	$x - 8$	$(x - 8)^2$	$(x - 8)^2 n(x)$
5							
6							

Complete the table and determine $SS(\bar{x})$ and $SS(8)$.

Sample Variance and Standard Deviation

Notice that $SS(\bar{x}) < SS(8)$.

- We would like to compute the variation about 8, the value that our intuition tells us is the true mean.
- In many circumstances, we do not have such intuition. Thus, we doing the best we can by computing $\bar{x}$.
- In this case, the variation about the sample mean is smaller than the variation about what may be called a true mean.
- Thus, division of $\sum_{i=1}^{n}(x_i - \bar{x})^2$ by n systematically underestimates the variance.
- We can be compensated for this by dividing by something small than n. Later, we will give a criterion to expelling why the choice is $n - 1$.

Quadratic Identity

For a linear transformation $y_i = a + bx_i$ for the observations x_i we have the quadratic identity for the variance:

$$\text{var}(a + bx) = b^2\text{var}(x)$$

Because it is one of the most frequently used and useful identities in all of statistics, we give a derivation.

$$\begin{aligned}
\text{var}(y) &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}((a + bx_i) - (a + b\bar{x}))^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(b(x_i - \bar{x}))^2 = b^2\text{var}(x).
\end{aligned}$$

Quantiles

A single observation, say 76 on a exam, gives little information about the performance on the exam. One way to include more about this observation would be to give the value of the empirical cumulative distribution function. Thus,

$$F_n(76) = 0.6771$$

tells us that about 68% of the exam scores were at or below 76. This is sometimes reported by saying that 76 is the 0.6771 quantile for the exam scores.

We can determine this value using the R command `quantile`. For the ages of presidents at inauguration, we have that the 72% quantile is 57 year old.

```
> quantile(age,0.72)
72%
 57
```

Standardized Variables

A second way to evaluate a score of 76 is to related it to the mean.

- If $\bar{x} = 71$. Then, we might say that the exam score is 5 points above the mean.
  - If the scores are spread out, then 5 points is just a little above average.
  - If the scores are tightly spread, then 5 points is quite a bit above average.
- Thus, for comparisons, we will sometimes use the standardized version of x_i,

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

- The observations z_i have mean 0 and standard deviation 1.
  - z_i is also called the standard score, z-value, z-score, and the normal score.
- An individual z-score, z_i, gives the number of standard deviations an observation x_i is above (or below) the mean. Thus, the standardized score has no units.

Transformations

Exercise. For a linear transformation $y_i = a + bx_i$ for the observations x_i, fill in the table that give the value of the summary statistic for the transformed data y based on the observations x. Consider what occurs if $b < 0$.

x-statistic	y-statistic
median	
variance	
standard deviation	
first quartile	
third quartile	
interquartile range	