Lecture 26

The Normal Curve

Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)
Announcements
Questions for This Week

● How can we quantify natural concepts like “center” and “variability”?

● Why do many of the empirical distributions that we generate come out bell shaped?

● How is sample size related to the accuracy of an estimate?
Standard Deviation (Review)
How Far from the Average?

- Standard deviation (SD) measures roughly how far the data are from their average.

- SD = root mean square of deviations from average:
  5  4  3  2  1

- SD has the same units as the data.
Why Use the SD?

There are two main reasons.

● **The first reason:**
No matter what the shape of the distribution, the bulk of the data are in the range “average ± a few SDs”

● **The second reason:**
Coming up later in this lecture ...
How Big are Most of the Values?

No matter what the shape of the distribution, the bulk of the data are in the range “average ± a few SDs”

Chebyshev’s Inequality
No matter what the shape of the distribution, the proportion of values in the range “average ± z SDs” is at least $1 - 1/z^2$
### Chebyshev’s Bounds

<table>
<thead>
<tr>
<th>Range</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>average ± 2 SDs</td>
<td>at least 1 - 1/4 (75%)</td>
</tr>
<tr>
<td>average ± 3 SDs</td>
<td>at least 1 - 1/9 (88.888…%)</td>
</tr>
<tr>
<td>average ± 4 SDs</td>
<td>at least 1 - 1/16 (93.75%)</td>
</tr>
<tr>
<td>average ± 5 SDs</td>
<td>at least 1 - 1/25 (96%)</td>
</tr>
</tbody>
</table>

**No matter what the distribution looks like**

(Demo)
Standard Units
Standard Units

- How many SDs above average?
- \( z = \frac{\text{value} - \text{average}}{\text{SD}} \)
  - Negative \( z \): value below average
  - Positive \( z \): value above average
  - \( z = 0 \): value equal to average
- When values are in standard units: average = 0, SD = 1
- Chebyshev: At least 96% of the values of \( z \) are between -5 and 5
  
  (Demo)
Discussion Question

Find whole numbers that are close to:

(a) the average age

(b) the SD of the ages

(Demo)
The SD and the Histogram

- Usually, it's not easy to estimate the SD by looking at a histogram.
- But if the histogram has a bell shape, then you can.
The SD and Bell-Shaped Curves

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

(Demo)
The Normal Distribution
The Standard Normal Curve

A beautiful formula that we won’t use at all:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$
Bell Curve

The Standard Normal Curve
Normal Proportions
How Big are Most of the Values?

No matter what the shape of the distribution, the bulk of the data are in the range “average ± a few SDs”

If a histogram is bell-shaped, then

- Almost all of the data are in the range “average ± 3 SDs”
## Bounds and Normal Approximations

<table>
<thead>
<tr>
<th>Percent in Range</th>
<th>All Distributions</th>
<th>Normal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>average ± 1 SD</td>
<td>at least 0%</td>
<td>about 68%</td>
</tr>
<tr>
<td>average ± 2 SDs</td>
<td>at least 75%</td>
<td>about 95%</td>
</tr>
<tr>
<td>average ± 3 SDs</td>
<td>at least 88.888...%</td>
<td>about 99.73%</td>
</tr>
</tbody>
</table>
A “Central” Area

Average ± 2SDs: 95% of the area
Central Limit Theorem
Second Reason for Using the SD

If the sample is
- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample sum
(or of the sample average) is roughly normal

(Demo)