Lecture 18 – Standard Deviation and Correlation
Announcements

- **Lab07 – Normal Distribution and Variance of Sample Means (short)**
  - Due Wednesday 11/23

- **Homework 7 - Confidence Intervals, Resampling, the Bootstrap, and the Central Limit Theorem**
  - Due Thursday 11/24
  - Not the shortest

- **Homeworks:**
  - Run all cells before submitting

- **Dropping 2 homeworks and labs**
Confidence Intervals
95% Confidence Interval

- Interval of estimates of a parameter
- Based on random sampling
- 95% is called the confidence level
  - Could be any percent between 0 and 100
  - Higher level means wider intervals

- The confidence is in the process that gives the interval:
  - It generates a “good” interval about 95% of the time
Confidence Intervals & Hypothesis Tests
Null hypothesis: Population average = x
Alternative hypothesis: Population average ≠/ x
Cutoff for P-value: ρ%

Method:
- Construct a (100-ρ)% confidence interval for the population average
- If x is not in the interval, reject the null
- If x is in the interval, can’t reject the null
Data Science in this course

- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)

- Inference
  - Make reliable conclusions about the world
  - Statistics is useful

- Prediction
  - Informed guesses about unseen data
Center & Spread
Questions/Goals

- 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?
Average and the Histogram
The average (mean)

Data: 2, 3, 3, 9

\[
\text{Average} = \frac{2+3+3+9}{4} = 4.25
\]

- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly
Relation to the histogram

- The average depends only on the **proportions** in which the distinct values appears.

- The average is the **center of gravity** of the histogram.

- It is the point on the horizontal axis where the histogram balances.
Average as balance point

- Average is 4.25
Question

- What list produces this histogram?
Question

- What list produces this histogram?

1, 2, 2, 3, 3
3, 4, 4, 5
Question

- What list produces this histogram?
  - 1, 2, 2, 3, 3
  - 3, 4, 4, 5

- Average?
Question

- What list produces this histogram?

1, 2, 2, 3, 3  
3, 4, 4, 5  

- Average?
  - 3
What list produces this histogram?

1, 2, 2, 3, 3
3, 4, 4, 5

Average?
• 3

Median?
Question

- What list produces this histogram?
  - 1, 2, 2, 3, 3
  - 3, 4, 4, 5

- Average?
  - 3

- Median?
  - 3
Question 2

- Are the medians of these two distributions the same or different? Are the means the same or different? If you say “different,” then say which one is bigger.
Answer 2

- **List 1**
  - 1, 2, 2, 3, 3, 3, 4, 4, 5

- **List 2**
  - 1, 2, 2, 3, 3, 3, 4, 4, 10

- **Medians** = 3
- **Mean(List1)** = 3
- **Mean (List 2)** = 3.55556
Comparing Mean and Median

- **Mean**: Balance point of the histogram
- **Median**: Half-way point of data; half the area of histogram is on either side of median
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail.
Question

- Which is bigger, median or mean?
Standard Deviation
Defining Variability

- **Plan A**: “biggest value - smallest value”
  - Doesn’t tell us much about the shape of the distribution

- **Plan B**:
  - Measure variability around the mean
  - Need to figure out a way to quantify this
Standard deviation (SD) measures roughly how far the data are from their average.

- SD = root mean square of deviations from average

Steps: 5 4 3 2 1

- SD has the same units as the data.
Why use Standard Deviation

- 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 plus or minus a few SDs”

- The second reason:
  - Relation with the bellshaped curve
  - Discuss this later in the lecture
How big are most 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 $1 - \frac{1}{z^2}$
the proportion of values in the range “average ± z SDs” is at least $1 - \frac{1}{z^2}$
the proportion of values in the range “average ± z SDs” is at least $1 - \frac{1}{z^2}$

<table>
<thead>
<tr>
<th>Range</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>average ± 2 SDs</td>
<td>at least $1 - \frac{1}{4}$ (75%)</td>
</tr>
</tbody>
</table>
Chebyshev’s Bounds

the proportion of values in the range “average ± z SDs” is at least $1 - 1/z^2$

<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>
</tbody>
</table>
Chebyshev’s Bounds

the proportion of values in the range “average ± z SDs” is at least 1 - 1/z^2

<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>
</tbody>
</table>
Chebyshev’s Bounds

the proportion of values in the range “average ± z SDs” is at least 1 - 1/z^2

<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>

True no matter what the distribution looks like
Understanding HW05 Results

Statistics:
Minimum: 7.5
Maximum: 29.0
Mean: 24.55
Median: 25.0
Standard Deviation: 3.96

- At least 50% of the class had scores between 20.59 and 28.51
- At least 75% of the class had scores between 16.62 and 32.47
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
What whole numbers are closest to

(1) Average age

(2) The SD of ages
(1) Average age is close to 27 (standard unit here is close to 0)

(2) The SD is about 6 years (standard unit at 33 is close to 1. $33 - 27 = 6$)
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
Normal Distribution
Equation for the normal curve

\[
\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty
\]
The Standard Normal Curve
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”
<table>
<thead>
<tr>
<th>Percent in Range</th>
<th>All Distributions</th>
<th>Normal Distributions</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
Central Limit Theorem

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 the sample average) is roughly normal.
We often only have a sample.

We care about sample averages because they estimate population averages.

The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) is connected to random sample averages.

CLT allows us to make inferences based on averages of random samples.
Correlation
To predict the value of a variable:

- Identify (measurable) attributes that are associated with that variable
- Describe the relation between the attributes and the variable you want to predict
- Then, use the relation to predict the value of a variable
Visualizing Two Numerical Variables

- **Trend**
  - Positive association
  - Negative association

- **Pattern**
  - Any discernible “shape” in the scatter
  - Linear
  - Non-linear

Visualize, then quantify
The Correlation Coefficient $r$

- Measures **linear** association
- Based on standard units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*