Standard deviation
normal approximation for data
areas under the normal curve
percentiles

Introduction to probability
conditional probability
multiplication rule

Quiz 3.
Standard Deviation.

- Measure of spread.

- SD of a list of numbers measures how far away they are from the mean.

\[ \begin{align*}
\text{Small SD} & \quad \text{compared to} \quad \text{large SD}
\end{align*} \]

Most observations are within 1 - 2 SD of average.

≈ 68% of data lies within 1 SD of mean.

≈ 95% of data lies within 2 SD.
Computing the SD.

1. RMS of the deviations from the mean.
   \[
   \begin{align*}
   \text{data} &= 20, 10, 15, 15 \\
   \text{mean} &= \frac{20 + 10 + 15 + 15}{4} = \frac{60}{4} = 15
   \end{align*}
   \]

2. Compute the deviations from the mean.
   \[
   \begin{align*}
   \text{deviations} &= 20-15, 10-15, 15-15, 15-15 \\
   &= 5, -5, 0, 0
   \end{align*}
   \]

3. RMS of the deviations
   \[
   \begin{align*}
   \text{RMS} &= \sqrt{\frac{5^2 + (-5)^2 + 0^2 + 0^2}{4}} \\
   &= \sqrt{\frac{25 + 25 + 0 + 0}{4}} = \sqrt{\frac{50}{4}} = \sqrt{12.5} = 3.5
   \end{align*}
   \]

Scientific calculators - caution. Often divide by (\# data - 1) in step 3. We call this SD*:

\[
\text{SD} = \sqrt{\frac{\text{\# data} - 1}{\text{\# data}}} \times \text{SD*}
\]
Normal Approximation for Data.

Very many data sets — the histogram can be approximated by the Normal Curve:

\[ y = \frac{100\%}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \]

Areas under the curve represent \%.

Mean Zero
Symmetric.
Area under the curve between $-1$ and $+1$ is $68\%$. 

percent per standard unit

standard units
The area is 95%. 

Standard units are used. 

Area is negligible.
Areas under the normal curve.

-1 to +1  68%
-2 to +2  95%
-3 to +3  99.7%

Use standard units so that the same normal curve can describe many data sets.

standard units: are # of SD away from the mean.
A group of women had a mean height 63.5 in. S.D. 3 in.

A particular woman was 69 inches tall.

In standard units, 69 in is 5.5 inches above mean.

\[
\text{5.5 in} = \frac{5.5}{3} = 1.8 \text{ S.D. above the mean}
\]

69 in \(z\) is 1.8 in standard units.

To convert to standard units, subtract the mean and divide by S.D.
Area under any histogram is always 100%.

distribution

Area under both curves is 100%.

- Peak value depends on the SD.

in standard units.
Q: What % of women were between 60.5 and 65 inches tall?

\[ \% \text{ in given by this area.} \]

Data Scale

Convert to standard units.

\[
\begin{align*}
60.5 & \rightarrow \frac{60.5 - 63.5}{3} = -1.33, \\
65 & \rightarrow \frac{65 - 63.5}{3} = \frac{2.5}{3} = 0.83.
\end{align*}
\]

Corresponding area on curve in standard units.

\[ 0.83 \]
Areas under the Normal curve in standard units are in the table in back of the book.

- Table gives areas in \( \pm \) symmetric intervals.

\[ z = 1 \text{ from table, area } = 68.27\% \]

\[ z = 0.5 \text{ area } = 38\% \]

We were interested in the interval

\(-1 \text{ to } 0.83\).

Use: **Symmetry.**

\( \text{total area } = 100\% \)

\( \text{areas under symmetric intervals} \) to get areas under arbitrary regions.
Area under the curve from 0 to 1.

Area from 0 to 1 is half of area from -1 to 1 (by symmetry).

Tail area.

What proportion of data is larger than a given value.

Complementary area.

Symmetric area. 86.64% from table.
area of two tails together is
\[ 100 - 86.64 \]

one tail has area
\[ \frac{100 - 86.64}{2} \]

\[ \frac{1}{2} \times 68.27 + \frac{1}{2} \times 59. \]

\[ = 63.5\% . \]
Use symmetry to get any area under the normal curve.
Histogram of Rainfall in Guarico, Venezuela
Percentiles.

When the histogram doesn’t follow the normal curve, describing it by means so is not appropriate.

E.g. rainfall data

Normal approximation says that many days had negative rainfall.

- Use percentiles.

  → Choose a data value.

  - The % of the data up to + including the data value gives the percentile.

  → Choose a percentile

  - Count up the data items in order, from lowest, and stop when you’ve counted the specified % of the data.

  The data value is the one corresponding to that percentile.
What percentile corresponds to the median?

50th percentile

25th percentile = 1st quartile
75th percentile = 3rd quartile

A useful measure of spread is

\[
\text{Inter-quartile range} = \text{(data value corresponding to 75th percentile)} - \text{(data value corresponding to 25th percentile)}
\]

More robust to outliers than the SD
as a measure of spread.
Probability

- quantifying chance.
  - using probabilities
  - mathematics to calculate quantities associated with uncertain events.

- what do we mean by chance?
- how do we interpret probabilities?

Die - When we roll it many times each side should come up as often as any other.
2-spot should appear \( \frac{1}{6} \) of the time.

Definition: the chance of an event is the percentage of the time the event is expected to occur, when the process is repeated over and over, independently, and under the same conditions.

Frequency definition of probability.