Descriptive statistics
Measures of central tendency

MODE
MEDIAN
MEAN
Mode

The MODE is the most frequent observation

It is the only meaningful measures for *nominal data/categorical data*
Mode: Example

<table>
<thead>
<tr>
<th>Case</th>
<th>ColorEye</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Blue</td>
</tr>
<tr>
<td>2</td>
<td>Brown</td>
</tr>
<tr>
<td>3</td>
<td>Brown</td>
</tr>
<tr>
<td>4</td>
<td>Green</td>
</tr>
<tr>
<td>5</td>
<td>Black</td>
</tr>
<tr>
<td>6</td>
<td>Blue</td>
</tr>
<tr>
<td>7</td>
<td>Green</td>
</tr>
<tr>
<td>8</td>
<td>Green</td>
</tr>
<tr>
<td>9</td>
<td>Black</td>
</tr>
<tr>
<td>10</td>
<td>Green</td>
</tr>
</tbody>
</table>

What’s the Mode?

NB! A distribution can be bimodal, i.e. with two peaks.
Median: the midpoint of the dataset

Half of the cases are *above* the median, half of the cases are *below* it.

It is suitable for ordinal data

Example 1:

The syntactic abilities of the aphasic subject are:

1 = Severely impaired
2 = Impaired
3 = Lightly impaired
4 = Mostly preserved
5 = Unimpaired
**THE MEDIAN** $M$

The **median** $M$ is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.

2. If the number of observations $n$ is odd, the median $M$ is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the bottom of the list.

3. If the number of observations $n$ is even, the median $M$ is the mean of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.
Mean

The arithmetic average.

It is suitable for numeric variables.
THE MEAN $\bar{x}$

To find the mean $\bar{x}$ of a set of observations, add their values and divide by the number of observations. If the $n$ observations are $x_1$, $x_2$, $\ldots$, $x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$
Descriptive statistics
Measures of variation 1

MINIMUM/MAXIMUM
RANGE
Measure of variation 1

- They are never suitable for nonnumeric variables!
- They are very useful to understand with a first glance if the data distribution is normal.
Measure of variation 1

- **Minimum/Maximum**
  - The lowest and the highest value

- **Range**
  - The difference between the minimum and the maximum value
Descriptive statistics
Measures of variation 2

QUARTILES
INTERQUARTILE RANGE
BOX-N-WHISKERS
X-ILE’S: Quartiles

- Quartiles: C

37 68 78 90
49 71 79 90
54 71 79 90
56 73 83 92
60 75 83 94
64 76 85 95
65 77 87 96
65 77 88 97

- Q1: 1st quartile: divides between 1st & 2nd groups
- Q2: 2nd quartile: divides data between the 2nd & 3rd (=median!)
X-ILE’S: Interquartile range

\[ q_3 - q_1 \]

Is the center where half of the scores lie

\[
\begin{align*}
37 & 68 & 78 & 90 \\
49 & 71 & 79 & 90 \\
54 & 71 & 79 & 90 \\
56 & 73 & 83 & 92 \\
60 & 75 & 83 & 94 \\
64 & 76 & 85 & 95 \\
65 & 77 & 87 & 96 \\
65 & 77 & 88 & 97 \\
\end{align*}
\]
Descriptive

The 5 number summary

1) Minimum
2) Q1
3) Median
4) Q3
5) Maximum

They are summarized by the Box & Whiskers plots (Boxplots)
“Boxes” show $q_3 - q_1$,
The midline is the median.
“Whiskers” show first and last quartiles.
Descriptive statistics
Measures of variation 3

DEVIATION
VARIANCE
STANDARD DEVIATION
DEVIATION & VARIANCE

DEVIATION
Is the difference between the observation and the mean

VARIANCE
Average square of deviation
THE STANDARD DEVIATION \( s \)

The variance \( s^2 \) of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of \( n \) observations \( x_1, x_2, \ldots, x_n \) is

\[
s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}
\]

or, in more compact notation,

\[
s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2
\]

The standard deviation \( s \) is the square root of the variance \( s^2 \):

\[
s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}
\]
Understanding when a distribution is normal

USE DESCRIPTIVE STATISTICS

DENSITY CURVES

LOOK FOR MEAN SKEWNESS, OUTLIERS AND VARIANCE

USE THE 68-95-99.7% RULE
Density curves

- Can be imagined as an histogram with a smooth approximation to the bars of the histograms.
- It is always above the horizontal axis
- Has area=1 underneath it
- Gives an idea if the distribution is normal, or skewed or with outliers
The normal distribution

- The density curve of a normal distribution is
  - Symmetric
  - Unimodal
  - Bell-shaped

The flatness of the curve will depend on the SD
THE 68–95–99.7 RULE

In the normal distribution with mean $\mu$ and standard deviation $\sigma$:

- Approximately 68% of the observations fall within $\sigma$ of the mean $\mu$.
- Approximately 95% of the observations fall within $2\sigma$ of $\mu$.
- Approximately 99.7% of the observations fall within $3\sigma$ of $\mu$. 

*Definition, pg 71
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company*
A z-score shows the distance from the mean in number of SDs.
Z-scores

- Tom got 112 on a language score and Henry got 105:
- Suppose: Mean=108
  SD=10

\[ z_{112} = \frac{112 - 108}{10} = 0.4 \]
\[ z_{105} = \frac{112 - 105}{10} = -0.3 \]
THE STANDARD NORMAL DISTRIBUTION

The **standard normal distribution** is the normal distribution $N(0, 1)$ with mean 0 and standard deviation 1.

If a variable $X$ has any normal distribution $N(\mu, \sigma)$ with mean $\mu$ and standard deviation $\sigma$, then the standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

has the standard normal distribution.
Sampling
A simple random sample of size $n$ consists of $n$ individuals from a population such as every set of $n$ individuals has an equal chance to be the actual sample selected.

In principle there could be many SRS which are a representation of a population.

We take one specific SRS and from that we draw conclusions on the population (again the Hp concerns the population which is studied through our SRS)
MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

Let \( \bar{x} \) be the mean of an SRS of size \( n \) from a population having mean \( \mu \) and standard deviation \( \sigma \). The mean and standard deviation of \( \bar{x} \) are

\[
\begin{align*}
\mu_{\bar{x}} &= \mu \\
\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}}
\end{align*}
\]

Definition, pg 361
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

If a population has the \( N(\mu, \sigma) \) distribution, then the sample mean \( \bar{x} \) of \( n \) independent observations has the \( N(\mu, \sigma / \sqrt{n}) \) distribution.

Definition, pg 362a
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company
Central Limit Theorem relates sample means to likely population mean. To understand it, imagine all the possible samples one might use, and all those sample means—the distribution of the sample means.

CENTRAL LIMIT THEOREM

Draw an SRS of size $n$ from any population with mean $\mu$ and finite standard deviation $\sigma$. When $n$ is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$
The CLT at work. The larger the size of the sample (n), the more normal the distribution.
Z-TEST

Given a RANDOMLY SELECTED SAMPLE, we know

- **Distribution** it is one of a normally distributed population of samples
- **Mean**: the mean of such samples will be the population mean
- **Standard deviation**: the standard deviation of the sample means (the STANDARD ERROR) decreases in proportion of the square root of the sample size. That is why the larger the sample size is the smaller the SE will be.

- These facts allow us to reason about the population.
- The reasoning will always include a **probability** that population has a mean of a given size.
- An essential assumption is that the sample is randomly selected. We can’t correct for biased data—even unintentionally biased.
Example z-test

- \[ z = \frac{x - \mu}{\sigma/\sqrt{n}} \]

You suspect that CALL programs may be effective for young children (since they can be initiated before reading, and look like computer games, need little supervision, ...).

- You have a standard test for English proficiency, where \( \mu = 70; \sigma = 14 \). You apply the same test to 49 randomly chosen schoolchildren who’ve had a CALL program at home for three years. Result: \( x = 74 \)
- We apply the formula → \( z = 2 \)
- From the tables: for \( z = 2 \) the area on the left of \( z \) is 0.98 → there is only 2% probability that the sample mean would be this high by chance.
Statistical confidence & Confidence Intervals
• The goal is to estimate a population parameter.
• To assess some evidence in favor of some claim on a population.
• Given that usually we study a population via a random sample we have to allow some confidence to our claims.
Confidence intervals

- $\sigma_{\text{mean}} = \frac{100}{\sqrt{500}} = 4.5$; remember 68-95-99.7 rule?

$s \text{ population}$

Sample size

Density curve of $\bar{x}$

Probability = 0.95

$\mu - 9, \mu, \mu + 9$

$\pm 2\sigma$

Figure 6.2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company
CONFIDENCE INTERVAL

A level $C$ confidence interval for a parameter is an interval computed from sample data by a method that has probability $C$ of producing an interval containing the true value of the parameter.
For a standard normal curve

Area = $C$

Area = $\frac{1-C}{2}$

Area = $\frac{1-C}{2}$

$-z^*$

0

$z^*$
CONFIDENCE INTERVAL FOR A POPULATION MEAN

Choose an SRS of size $n$ from a population having unknown mean $\mu$ and known standard deviation $\sigma$. The **margin of error** for a level $C$ confidence interval for $\mu$ is

$$m = z^* \frac{\sigma}{\sqrt{n}} \quad \text{SE}$$

Here $z^*$ is the value on the standard normal curve with area $C$ between the critical points $-z^*$ and $z^*$. The level $C$ **confidence interval** for $\mu$ is

$$\bar{x} \pm m$$

This interval is exact when the population distribution is normal and is approximately correct for large $n$ in other cases.
Figure 6-6
*Introduction to the Practice of Statistics, Fifth Edition*
© 2005 W.H. Freeman and Company
Confidence intervals

Ways of looking at confidence intervals

■ You have a sample and you want specific info on μ: between which and which value does it lie (with some level of certainty)?

■ 95 % of the time, averages of repeated samples are between –m and +m (margin of error)

■ 95 % of samples will catch the true μ in the interval –m, +m; 5 % will not! So we accept that we are wrong 5 % of the time, when μ is somewhere else.
Confidence intervals

Improving estimation

- Increasing the interval: more confidence that you have captured true $\mu$ (but margin of error becomes larger)
- Increasing number of observations $\rightarrow$ area around $\mu$ becomes smaller (better estimation)
- Decreasing $\sigma$, for instance by choosing a homogeneous subpopulation, or by improving sensitivity of measurement technique
Figure 6-5
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company