STATISTICAL CONCEPTS FOR CLINICAL RESEARCH

Vishwanath Iyer
Ashwini Mathur
Statistical Concepts for Clinical Research

Vishwanath Iyer
Ashwini Mathur

Novartis Healthcare Private Limited
DISCLAIMER

- All opinions expressed are authors and do not reflect the opinions of Novartis
- Material freely available from internet has been used and acknowledged
ACKNOWLEDGEMENTS

- Vinay Mahajan
- Varsha Mahajan
- IDDCR
- Johns Hopkins OpenCourseware
- Andrew Vickers for Acupuncture Data
- All authors of material available freely on the internet
Descriptive Statistics
Background

- **Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial**
  - J Vickers, Rebecca W Rees, Catherine E Zollman, Rob McCarney, Claire Smith, Nadia Ellis, Peter Fisher, Robbert Van Haselen

  - **Objective** To determine the effects of a policy of “use acupuncture” on headache, health status, days off sick, and use of resources in patients with chronic headache compared with a policy of “avoid acupuncture.”
  - **Design** Randomised, controlled trial.
  - **Setting** General practices in England and Wales.
  - **Participants** 401 patients with chronic headache, predominantly migraine.
  - **Interventions** Patients were randomly allocated to receive up to 12 acupuncture treatments over three months or to a control intervention offering usual care.
  - **Main outcome measures** Headache score, SF-36 health status, and use of medication were assessed at baseline, three, and 12 months. Use of resources was assessed every three months.
## Background

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Variable description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Patient ID code</td>
<td>NA</td>
</tr>
<tr>
<td>age</td>
<td>Age</td>
<td>Yrs</td>
</tr>
<tr>
<td>sex</td>
<td>Sex</td>
<td>0 Male, 1 Female</td>
</tr>
<tr>
<td>migraine</td>
<td>Migraine</td>
<td>0 No, 1 Yes</td>
</tr>
<tr>
<td>chronicity</td>
<td>Chronicity (yrs suffering from chronic headache)</td>
<td>Yrs</td>
</tr>
<tr>
<td>group</td>
<td>Treatment Group</td>
<td>0 is control, 1 is acupuncture</td>
</tr>
<tr>
<td>pk1</td>
<td>severity score pack1 (baseline)</td>
<td>NA - Lower is better</td>
</tr>
<tr>
<td>pk5</td>
<td>severity score pack5 (one year followup)</td>
<td>NA - lower is better</td>
</tr>
<tr>
<td>gen1</td>
<td>Pack 1 (baseline) SF36 general health</td>
<td>NA - higher is better</td>
</tr>
<tr>
<td>gen5</td>
<td>Pack 3 SF36 general health</td>
<td>NA - higher is better</td>
</tr>
<tr>
<td>withdrawal_reason</td>
<td>withdrawal reason</td>
<td>NA</td>
</tr>
<tr>
<td>painmedspk1</td>
<td>MQS pack 1</td>
<td>Units of medicine used</td>
</tr>
<tr>
<td>painmedspk5</td>
<td>MQS pack 5</td>
<td>Units of medicine used</td>
</tr>
<tr>
<td>pchange</td>
<td>percent change from baseline in severity score</td>
<td>NA</td>
</tr>
<tr>
<td>response</td>
<td>1 if &gt;35% improvement from baseline</td>
<td>NA</td>
</tr>
</tbody>
</table>
How do you analyze this data?

- What is the effect of acupuncture on headaches?
  - As a preliminary step summarize data at baseline
  - How do you summarize baseline data?
    - What kind of variables are there?
    - What kind of summary statistics and graphics could be used for summarizing this data?
  - How do you do you statistical inference on this data?
    - Sampling variability
    - Confidence Intervals
    - Introduction to Hypothesis Testing
### Types of Variables: Overview

#### Categorical
- **Binary**
  - 2 categories +
  - Dead/alive
  - Treatment/placebo
  - Disease/no disease
  - Exposed/Unexposed

- **Nominal**
  - More categories +
  - Blood type (O, A, B, AB)
  - Marital status
  - Occupation

- **Ordinal**
  - Order matters +
  - Staging in cancer as I, II, III or IV
  - Birth order—1st, 2nd, 3rd, etc.
  - Letter grades (A, B, C, D, F)
  - Ratings on a scale from 1-5
  - Age in categories (10-20, 20-30, etc.)

#### Quantitative
- **Discrete**
  - Counts
  - Time
  - Age
  - Height

- **Continuous**
  - Blood counts
  - Speed of a car
  - Income
  - Time to event

---

Indian Association for Statistics in Clinical Trials
Summary Measures: Overview

- **Center and Location**
  - Mean
  - Median
  - Mode
  - Weighted Mean

- **Other Measures of Location**
  - Percentiles
  - Quartiles

- **Variation**
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
Measures of Central Tendency

Center and Location

Mean
\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]
\[ \mu = \frac{\sum_{i=1}^{N} x_i}{N} \]

Median

Mode

Weighted Mean
\[ \overline{X}_W = \frac{\sum w_i x_i}{\sum w_i} \]
\[ \mu_W = \frac{\sum w_i x_i}{\sum w_i} \]
Mode

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode or several modes

Mode = 9

No Mode
Shape of a Distribution

- Describes how data is distributed
- Symmetric or skewed

**Left-Skewed**
- Mean < Median < Mode
- (Longer tail extends to left)

**Symmetric**
- Mean = Median = Mode

**Right-Skewed**
- Mode < Median < Mean
- (Longer tail extends to right)
Quartiles

- Split Ordered Data into 4 Quarters

- Position of $i$th Quartile

\[
(Q_i) = \frac{i(n + 1)}{4}
\]
Quartiles (continued)

Data in Ordered Array: 11 12 13 16 16 17 17 18 21

Position of $Q_1 = \frac{1(9+1)}{4} = 2.5$

$Q_1 = \frac{(12+13)}{2} = 12.5$

$Q_2 = \text{Median} = 16$

$Q_3 = 17.5$
Box-and-Whisker Plot

Graphical Display of Data Using 5-Number Summary:

\[ X_{\text{smallest}} \quad Q_1 \quad \text{Median} \quad Q_3 \quad X_{\text{largest}} \]
Distribution Shape and Box-and-Whisker Plot

Left-Skewed

Symmetric

Right-Skewed
Measures of Variation

Variation

Variance
- Population Variance
- Sample Variance

Standard Deviation
- Population Standard Deviation
- Sample Standard Deviation

Range

Interquartile Range
Range

- Measure of variation

- Difference between the largest and the smallest observations:

\[
\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}
\]

- Ignores the way in which data are distributed

\[
\text{Range} = 12 - 7 = 5
\]

\[
\text{Range} = 12 - 7 = 5
\]
Interquartile Range

- Interquartile range = 3rd quartile – 1st quartile

Data in Ordered Array: 11 12 13 16 16 17 17 18 21

Interquartile Range = $Q_3 - Q_1 = 17.5 - 12.5 = 5$

Example:

25% 25% 25% 25%

12 30 45 57 70

Interquartile range = $57 - 30 = 27$
Variance

- Important measure of variation
- Shows variation about the mean

Sample variance:

\[
S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}
\]

Population variance:

\[
\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}
\]
Standard Deviation

- Most important measure of variation
- Shows variation about the mean

Sample standard deviation:

\[ S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \]

Population standard deviation:

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \]
Comparing Standard Deviations

Data A

Mean = 15.5
\[ s = 3.338 \]

Data B

Mean = 15.5
\[ s = 0.9258 \]

Data C

Mean = 15.5
\[ s = 4.57 \]
Acupuncture ....

<table>
<thead>
<tr>
<th>Id</th>
<th>Discrete, Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Continuous</td>
</tr>
<tr>
<td>sex</td>
<td>Discrete, Binary</td>
</tr>
<tr>
<td>migraine</td>
<td>Discrete, Binary</td>
</tr>
<tr>
<td>chronicity</td>
<td>Continuous</td>
</tr>
<tr>
<td>group</td>
<td>Discrete, Binary</td>
</tr>
<tr>
<td>pk1</td>
<td>Continuous</td>
</tr>
<tr>
<td>pk5</td>
<td>Continuous</td>
</tr>
<tr>
<td>gen1</td>
<td>Continuous</td>
</tr>
<tr>
<td>gen5</td>
<td>Continuous</td>
</tr>
<tr>
<td>withdrawal_reason</td>
<td>Discrete, Nominal</td>
</tr>
<tr>
<td>painmedspk1</td>
<td>Continuous</td>
</tr>
<tr>
<td>painmedspk5</td>
<td>Continuous</td>
</tr>
<tr>
<td>Pchange</td>
<td>Continuous</td>
</tr>
<tr>
<td>response</td>
<td>Discrete, Binary</td>
</tr>
</tbody>
</table>
### Acupuncture....

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>pk1</th>
<th>gen1</th>
<th>painmedapk1</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td></td>
<td>45.54 Mean</td>
<td>21.46 Mean</td>
<td>26.51 Mean</td>
</tr>
<tr>
<td>chronicity</td>
<td></td>
<td>48.00 Median</td>
<td>20.30 Median</td>
<td>21.00 Median</td>
</tr>
<tr>
<td>Mode</td>
<td></td>
<td>49.00 Mode</td>
<td>10.00 Mode</td>
<td>9.25 Mode</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td>11.05 Standard Deviation</td>
<td>13.77 Standard Deviation</td>
<td>16.16 Standard Deviation</td>
</tr>
<tr>
<td>Sample Variance</td>
<td></td>
<td>122.11 Sample Variance</td>
<td>189.51 Sample Variance</td>
<td>261.25 Sample Variance</td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td>47.00 Range</td>
<td>53.00 Range</td>
<td>88.00 Range</td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
<td>18.00 Minimum</td>
<td>1.00 Minimum</td>
<td>6.75 Minimum</td>
</tr>
<tr>
<td>Maximum</td>
<td></td>
<td>65.00 Maximum</td>
<td>54.00 Maximum</td>
<td>94.75 Maximum</td>
</tr>
<tr>
<td>Count</td>
<td></td>
<td>401.00 Count</td>
<td>401.00 Count</td>
<td>401.00 Count</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>pk5</th>
<th>gen5</th>
<th>painmedapk5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td>19.08 Mean</td>
<td>62.21 Mean</td>
<td>34.25</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>15.00 Median</td>
<td>65.00 Median</td>
<td>18</td>
</tr>
<tr>
<td>Mode</td>
<td></td>
<td>9.50 Mode</td>
<td>70.00 Mode</td>
<td>0</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td>15.61 Standard Deviation</td>
<td>22.65 Standard Deviation</td>
<td>49.44</td>
</tr>
<tr>
<td>Sample Variance</td>
<td></td>
<td>243.72 Sample Variance</td>
<td>512.95 Sample Variance</td>
<td>2444</td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td>87.25 Range</td>
<td>100.00 Range</td>
<td>365</td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
<td>87.25 Minimum</td>
<td>0.00 Minimum</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td></td>
<td>87.25 Maximum</td>
<td>100.00 Maximum</td>
<td>365</td>
</tr>
<tr>
<td>Count</td>
<td></td>
<td>301.00 Count</td>
<td>297.00 Count</td>
<td>301.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>pchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td>0.21</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>0.28</td>
</tr>
<tr>
<td>Mode</td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td>0.52</td>
</tr>
<tr>
<td>Sample Variance</td>
<td></td>
<td>0.27</td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td>4.14</td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
<td>-3.14</td>
</tr>
<tr>
<td>Maximum</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Count</td>
<td></td>
<td>301.00</td>
</tr>
</tbody>
</table>
Acupuncture ....

<table>
<thead>
<tr>
<th>Count</th>
<th>Column Labels</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
<td>Grand Total</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>31</td>
<td>165</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>Acupuncture</td>
<td>33</td>
<td>172</td>
<td>205</td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td>64</td>
<td>337</td>
<td>401</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count</th>
<th>Column Labels</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-Migraine</td>
<td>Migraine</td>
<td>Grand Total</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>13</td>
<td>183</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>Acupuncture</td>
<td>11</td>
<td>194</td>
<td>205</td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td>24</td>
<td>377</td>
<td>401</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count</th>
<th>Column Labels</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;35% Change</td>
<td>&gt; 35% Change</td>
<td>(blank)</td>
<td>Grand Total</td>
</tr>
<tr>
<td>Control</td>
<td>95</td>
<td>45</td>
<td>56</td>
<td>196</td>
</tr>
<tr>
<td>Acupuncture</td>
<td>74</td>
<td>87</td>
<td>44</td>
<td>205</td>
</tr>
<tr>
<td>Grand Total</td>
<td>169</td>
<td>132</td>
<td>100</td>
<td>401</td>
</tr>
</tbody>
</table>
Acupuncture ....

<table>
<thead>
<tr>
<th>age</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>45.54</td>
</tr>
<tr>
<td>Median</td>
<td>48.00</td>
</tr>
<tr>
<td>Mode</td>
<td>49.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>11.05</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>122.11</td>
</tr>
<tr>
<td>Range</td>
<td>47.00</td>
</tr>
<tr>
<td>Minimum</td>
<td>18.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>65.00</td>
</tr>
<tr>
<td>Count</td>
<td>401.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pchange</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.21</td>
</tr>
<tr>
<td>Median</td>
<td>0.28</td>
</tr>
<tr>
<td>Mode</td>
<td>0.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.52</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>0.27</td>
</tr>
<tr>
<td>Range</td>
<td>4.14</td>
</tr>
<tr>
<td>Minimum</td>
<td>-3.14</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.00</td>
</tr>
<tr>
<td>Count</td>
<td>301.00</td>
</tr>
</tbody>
</table>
Sampling Variability and Confidence Intervals
Distributions ...

Density Function – Describes the probability function of a random variable

X Axis – Values that the random variable can take
Y Axis – Probability that the random variable takes that particular value
Distributions ....

<table>
<thead>
<tr>
<th>Z</th>
<th>Within Z SDs of the mean</th>
<th>More than Z SDs above the mean</th>
<th>More than Z SDs above or below the mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>68.27%</td>
<td>15.87%</td>
<td>31.73%</td>
</tr>
<tr>
<td>2.0</td>
<td>95.45%</td>
<td>2.28%</td>
<td>4.55%</td>
</tr>
<tr>
<td>2.5</td>
<td>98.76%</td>
<td>0.62%</td>
<td>1.24%</td>
</tr>
<tr>
<td>3.0</td>
<td>99.73%</td>
<td>0.13%</td>
<td>0.27%</td>
</tr>
</tbody>
</table>
Sampling Variability ....

- Population versus Sample

- The population of interest could be . . .
  - All women between ages 30–40
  - All patients with a particular disease

- The sample is a small number of individuals from the population
  - The sample is a subset of the population
Sampling Variability ....

- *Sample mean* ($\bar{X}$) versus *population mean* ($\mu$)
  - For example, mean blood pressures
  - We know the sample mean $\bar{X}$
  - We don’t know the population mean $\mu$, but we would like to

- **Key Question:**
  - How close is the sample mean to the population mean ?
Sampling Variability ....

- A parameter is a number that describes the population
  - A parameter is a fixed number, but in practice we do not know its value—Example: Population mean

- A statistic is a number that describes a sample of data
  - A statistic can be calculated
  - We often use a statistic to estimate an unknown parameter—Example: Sample mean

- How accurate is the statistic for estimating the parameter?
Sampling Variability ....

- **Errors from biased sampling**
  - The study systematically favors certain outcomes
  - Voluntary response
  - Non-response
  - Convenience sampling

- **Solution: Random sampling**

- **Errors from (random) sampling**
  - Caused by chance occurrence
  - Get a “bad” sample because of bad luck (by “bad” we mean not representative)
  - Can be controlled by taking a larger sample

- **Using mathematical statistics, we can figure out how much potential error there is from random sampling (standard error)**
Sampling Variability ....

- **Example:** Blood pressure study of population of women age 30–40
  - Volunteer - Non-random; selection bias
  - Family members - Non-random; not independent
  - Telephone survey; random-digit dial - Random or non-random sample? Non-response?
  - Clinic population, 100 consecutive patients - Random or non-random sample? (Convenience samples are sometimes assumed to be random)

- **When a selection procedure is biased, taking a larger sample does not help**

- **Non-respondents can be very different from respondents**
Sampling Variability ....

- When a sample is randomly selected from a population, it is called a random sample
  - In a simple random sample, each individual in the population has an equal chance of being chosen for the sample

- Random sampling helps control systematic bias
  - But even with random sampling, there is still sampling variability or error

- If we repeatedly choose samples from the same population, a statistic will take different values in different samples

- If you repeat the study and the statistic does not change much (you get the same answer each time), then it is fairly reliable (not a lot of variability)
Sampling Variability ....

- **Example: Hospital Length of Stay**

  The distribution of the length of stay information for the population of patients discharged from a major teaching hospital in a one year period is a heavily right skewed distribution.
  
  - Mean, 5.0 days, SD 6.9 days; Median, 3 days; Range 1 to 173 days

- Suppose I have a random sample of 10 patients discharged from this hospital and we wish to use the sample information to estimate average length of stay at the hospital.
  
  - The sample mean is 5.7 days? How “good” an estimate is this of the population mean?
Sampling Variability ...

- Suppose we take another random sample of 10 patients... and the sample mean length of stay for this sample is 3.9 days.
- We do this a third time, and get a sample mean of 4.6 days.
- Suppose we did this 200 times.
- If we want to get a handle on the behavior of my sample mean estimate from sample to sample is to plot a histogram of 200 sample mean values.
Sampling Variability ....

- Mathematical statisticians have figured out how to predict what the sampling distribution will look like without actually repeating the study numerous times and having to choose a sample each time.
  - The sampling distribution of a sample statistics will look “normally” distributed – Central Limit Theorem.
  - This happens for sample means, sample proportions, sample mean differences and differences in sample proportions.

- Statisticians have derived formulas to calculate the standard deviation of the sampling distribution: It’s called the standard error of the statistic.
  - If a sampling distribution has a lot of variability (that is, a big standard error), then if you took another sample, it’s likely you would get a very different result.
Sampling Variability ....

- About 95% of the time, the sample mean (or proportion) will be within two standard errors of the population mean (or proportion) - This tells us how “close” the sample statistic should be to the population parameter

- Standard errors (SE) measure the precision of your sample statistic
  - A small SE means it is more precise

- Mathematical statisticians have come up with formulas for the standard error; there are different formulas for:
  - Standard error of the mean (SEM)
  - Standard error of a proportion
Sampling Variability ....

- Standard deviation measures the variability in the population

- Standard error measures the precision of a statistic—such as the sample mean or proportion—as an estimate of the population mean or population proportion

- The standard error of the mean (SEM) is a measure of the precision of the sample mean

  - The smaller SEM is, the more precise \( \bar{x} \) is
  - SEM depends on \( n \) and \( s \)
  - SEM gets smaller if
    - \( s \) gets smaller
    - \( n \) gets bigger

\[
SEM = \frac{s}{\sqrt{n}}
\]
Sampling Variability ...

- How close to the population mean ($\mu$) is the sample mean ($X$)? The standard error of the sample mean tells us!

- If we can calculate the sample mean and estimate its standard error, can that help us make a statement about the population mean?

- The central limit theorem tells us that the sampling distribution is approximately normal given enough data. Additionally, the theorem tell us this sampling distribution should be centered about the true value of the population mean $\mu$. 
Sampling Variability ..... 

- **Sampling distribution** is the distribution of all possible values of $\bar{x}$ from samples of same size, $n$

- The “reverse” is also true—95% of the time $\mu$ will fall within two standard errors of a given $\bar{x}$

- 95% of possible values for $\bar{x}$ will fall within approximately two standard errors of $\mu$

- 95% of the time, the population mean will lie within about two standard errors of the sample mean
  - $\bar{x} \pm 2 \text{SEM}$

- Why is this true?
  - Because of the central limit theorem

---

Indian Association for Statistics in Clinical Trials

Statistics Workshop
Acupuncture ....

<table>
<thead>
<tr>
<th>% Change</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.210389209</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.029828897</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.517512026</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>0.267818697</td>
</tr>
<tr>
<td>Count</td>
<td>301</td>
</tr>
<tr>
<td>Confidence Level(95.0%)</td>
<td>0.058700375</td>
</tr>
<tr>
<td>Lower CI</td>
<td>0.151688834</td>
</tr>
<tr>
<td>Upper CI</td>
<td>0.269089584</td>
</tr>
</tbody>
</table>
Hypothesis Testing
Formulate the question

New question arises

*Trial in breast cancer*

Is Drug A more effective than Drug B?

Clear formulation of the question

- not open question
- precision of primary end-point i.e. what parameter to look at, when, how?
- what exactly do we want to test?

→ Formulation of 2 assumptions: $H_0$ and $H_1$

After 4 months of treatment, considering response rate

$H_0$: Response rate with Drug A $\leq$ Response rate with Drug B

$H_1$: Response rate with Drug A $>$ Response rate with Drug B
The two hypotheses $H_0$ and $H_1$

- $H_0$ and $H_1$ are called hypotheses and do not play the same role.

$H_0$ is called "null hypothesis"

* is the hypothesis we would like to reject

Response rate after 4 months of treatment with Drug A
\[ \leq \text{Response rate after 4 months of treatment with Drug B} \]

$H_1$ is called "alternative hypothesis"

* is the hypothesis we believe in

Response rate after 4 months of treatment with Drug A
\[ > \text{Response rate after 4 months of treatment with Drug B} \]
A decision helper → the statistical test

Objective:
Decide whether $H_0$ can be rejected based on a sample taken from a bigger population.

Hypothesis test = a decision rule

Statistical test
measures the difference between the data and what is expected if the null hypothesis is true

Data

Decision
Accept $H_0$
or
Reject $H_0$
Please order the following steps.

A/ Specify the question of interest

B/ Collect data

C/ Reject or accept $H_0$

D/ Build up the appropriate hypothesis

E/ Perform statistical test
Quiz (Solution)

A/ Specify the question of interest
D/ Build up the appropriate hypothesis
B/ Collect data
E/ Perform statistical test
C/ Reject or accept $H_0$
Quiz

What is a statistical test?

A/ a procedure that helps to make a decision

B/ the description of the null hypothesis

C/ a number that should be below 0.05
Quiz (solution)

What is a statistical test?

A/ a procedure that helps to make a decision

B/ the description of the null hypothesis

C/ a number that should be below 0.05
Possible Errors

- Decision based on sample not on the whole population

Possibility of error between the decision made and reality
In this example, let us suppose we are able to treat all breast cancer patients over the world with Drug A.

-> This would be *ideal world*, but is *never possible*.

This would enable us to calculate the real response rate for Drug A.

Supposing now we were able to treat all breast cancer patients over the world not with Drug A but with Drug B.

**Conclusion:**

Drug A is more effective than Drug B.
When doing a trial, following observations are made:

Drug A treated

Drug A 5 responses out of 9

Drug B treated

Drug B 5 responses out of 9

→ no difference between both!
### Possible Errors (cont.)

- \( H_0: \) RR Drug A \( \leq \) RR Drug B
- \( H_1: \) RR Drug A \( > \) RR Drug B

**Considering response rate after 4 months...**

<table>
<thead>
<tr>
<th>Reality</th>
<th>Decision based on sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never known</td>
<td></td>
</tr>
<tr>
<td>( H_0 ) true</td>
<td>Do not reject ( H_0 )</td>
</tr>
<tr>
<td>( H_1 ) true</td>
<td>Reject ( H_0 )</td>
</tr>
</tbody>
</table>

- \( \alpha \) = Confidence level
- \( 1 - \alpha \) = Power
- \( \beta \) = Type II Error
- \( 1 - \beta \) = Correct
Possible Errors (cont.)

Type I error risk $\alpha$ (false positive)
= *claiming $H_0$ can be rejected when $H_0$ is true in reality*
  - for example, risk to approve a non-effective drug
  - want to keep that risk small, usually 0.05 or 0.01

Type II error risk $\beta$ (false negative)
= *claiming $H_0$ can not be rejected when $H_0$ is not true in reality*
  - for example, risk to *stop development of an active drug*
  - linked to investment
  - usually 0.1 to 0.2
Possible Errors (cont.)

When $H_0$ is true

When $H_1$ is true

Decision: not reject $H_0$

Decision: reject $H_0$

Threshold value/“theoretical” value

Treatment effect

Observed value

Probability
Possible Errors (cont.)

Type I error risk $\alpha = \text{claiming } H_0 \text{ can be rejected when } H_0 \text{ is true in reality}$

Type II error risk $\beta = \text{claiming } H_0 \text{ can not be rejected when } H_0 \text{ is not true}$
Possible Errors (cont.)

What can be done?

- Type I error is controlled and fixed by the design

- Type II error and sample size influence each other
Quiz

You are an FDA reviewer, which risk would you prefer to control?

- approve a non-effective drug

- reject an active drug
Quiz (solution)

You are an FDA reviewer, which risk would you prefer to control?

- approve a non-effective drug
  
  -> claming $H_0$ must be rejected when $H_0$ is true in reality

- reject an active drug
  
  -> claming $H_0$ is true when $H_0$ is not true in reality
What about the P-value?

**Clinical Study report**

<table>
<thead>
<tr>
<th>Efficacy endpoint</th>
<th>Drug A n=154</th>
<th>Drug B N=170</th>
<th>P value</th>
<th>Chi-square test statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical response (palpation)</td>
<td>85 (55%)</td>
<td>61 (36%)</td>
<td>&lt;0.001</td>
<td>12.17</td>
</tr>
</tbody>
</table>

What does that mean?

As we have just seen:

1/ what are the corresponding hypotheses?

2/ what can I conclude?
What about the P-value? (cont.)

Statistical test = * Question-specific

* mechanism or value for decision-making

* Comparison of a calculated value based on sample data to a theoretical value obtained when $H_0$ is true

P-value = probability, when $H_0$ is true in reality, to get a value of a Test statistic at least as extreme as the one observed

- Advantage: not question-specific,
- always same interpretation
What about the P-value? (cont.)

P-value = * probability of getting a value of the test statistic as extreme or more extreme as the observed one by chance only when $H_0$ is true

* quantifies how far my value is from the null hypothesis
What about the P-value? (cont.)

What does that mean?

1/ what are the corresponding hypotheses?

After 4 months of treatment, considering response rate

\( H_0: \) Response rate with Drug A \( \leq \) Response rate with Drug B
\( H_1: \) Response rate with Drug A \( > \) Response rate with Drug B

2/ what can I conclude?

\( P\text{-value}=\text{Probability, when } H_0 \text{ is true in reality, to get a value of a test statistic at least as extreme as the one observed.} \)

\( \rightarrow \) probability small \( \rightarrow \) highly unlikely for \( H_0 \) to be true \( \rightarrow \) Decision=reject \( H_0 \)
What about the P-value? (cont.)

$P$-value = Probability, when $H_0$ is true in reality, to get a value of a test statistic at least as extreme as the one observed.
What about the P-value? (cont.)

If the observed value is 2.1

Threshold/Theoretical value = 3.84

P-value

$P$-value=Probability, when $H_0$ is true in reality, to get a value of a test statistic at least as extreme as the one observed.
What about the P-value? (cont.)

P-value = Probability, when $H_0$ is true in reality, to get a value of a test statistic at least as extreme as the one observed.
What about the P-value? (cont.)

Test statistic when H0 is true

If observed test statistic = 4.1
P-value=0.04

-> difference but less extreme than the previous one

Threshold/ Theoretical value = 3.84

P-value=quantifies how far my value is from the null hypothesis

Probability

Value of the statistical test
Statistical significance and clinical relevance

Less than 3% difference between both groups,
->Statistically significant does not necessarily imply clinically relevant
Avoid testing for sake of testing!
T-test
The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland. He published the test in Biometrika in 1908, but was forced to use a pen name by his employer, who regarded the fact that they were using statistics as a trade secret.

"Small samples are slippery customers whose word is not to be taken as gospel" (Moroney).
t-test
Assumptions

The t-test ANOVA have three assumptions:
- independence assumption (the elements of one sample are not related to those of the other sample),

- normality assumption (samples are randomly drawn from the normally distributed populations with unknown population means; otherwise the means are no longer best measures of central tendency, thus test will not be valid), and

- equal variance assumption (the population variances of the two groups are equal)
t-test

tests: whether the means of two groups are statistically different from each other

The first thing to notice about the three situations is that the difference between the means is the same in all three

1<sup>st</sup> case: moderate variability
2<sup>nd</sup> case: high variability case
3<sup>rd</sup> case: low variability

Clearly, we would conclude that the two groups appear most different or distinct in the bottom or low-variability case. Why?

Because there is relatively little overlap between the two bell-shaped curves. In the high variability case, the group difference appears least striking because the two bell-shaped distributions overlap so much.

To find the differences between scores for two groups:

judge the difference between their means relative to the spread or variability of their scores.
Signal = the difference between the means is the signal that, in this case, we think our program or treatment introduced into the data; 

Noise = a measure of variability that may make it harder to see the group difference

Signal = t-value  ➔ 2 groups are different
Noise

One sample t-test

Two sample t-test
- Equal sample sizes and equal variance
- Unequal sample sizes and equal variance
- Unequal sample sizes and unequal variance

Paired t-test
Testing Statistical Hypothesis
Independent Sample t-Test
1. Determine the Appropriate Test

- If comparing a sample to a population, use one sample tests.

- If comparing two samples in order to draw inferences about group differences in the population use two sample t-test.

  - Here the test statistic is based on a theoretical sampling distribution known as “sampling distribution of the difference between two means”.

\[
M_{\text{diff}} = \bar{X}_1 - \bar{X}_2
\]

  - The standard deviation of such a sampling distribution is referred to as the standard error of the difference.
1. Determine the Appropriate Test

- If the two groups are independent of each other uses independent group t-test.

- If the two groups are not independent of each other use dependent group t-test also known as paired t-test.
3. Determine Whether to Use a One or Two Tailed Test

- If testing for equality of means then two tailed test

- If testing whether one mean greater/smaller than the other then one tailed test
4. Calculating Test Statistics

- For the independent groups t-test the formula is:

\[ t_c = \frac{\bar{X}_1 - \bar{X}_2}{S_{x_1-x_2}} \]

- The numerator is the difference in means between the two samples, and the denominator is the estimated standard error of the difference.
6. Compare the Computed Test Statistic Against a Tabled Value

If \(|t_c| > |t_\alpha|\)         Reject \(H_0\)

If \(p\) value < \(\alpha\)      Reject \(H_0\)
Example of Independent Groups t-tests

- Suppose that we plan to conduct a study to alleviate the distress of preschool children who are about to undergo the finger-stick procedure for a hematocrit (Hct) determination.

- Note: Hct = % of volume of a blood sample occupied by cells.
Twenty subjects will be used to examine the effectiveness of the special treatment.

10 subjects randomly assigned to treatment group.

10 assigned to a control group that receives no special preparation.
### 4. Calculating Test Statistics

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>105</td>
</tr>
<tr>
<td>2</td>
<td>86</td>
<td>95</td>
</tr>
<tr>
<td>3</td>
<td>112</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>115</td>
<td>110</td>
</tr>
<tr>
<td>6</td>
<td>83</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>90</td>
<td>115</td>
</tr>
<tr>
<td>8</td>
<td>94</td>
<td>93</td>
</tr>
<tr>
<td>9</td>
<td>85</td>
<td>107</td>
</tr>
<tr>
<td>10</td>
<td>105</td>
<td>120</td>
</tr>
</tbody>
</table>
6. Compare the Computed Test Statistic Against a Tabled Value

\[ t_c = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{95.5 - 105.0}{5.41} = -1.85 \]

8. \( df = n_1 + n_2 - 2 = 18 \)

9. \( t_{0.05} = 2.10 \)

10. Statistical Decision: Fail to Reject \( H_0 \)
Testing Statistical Hypothesis for Dependent Samples
Example: Two Interventions in Same Patients

- Suppose that we wanted to compare direct and indirect methods of blood pressure measurement in a sample of trauma patients. Blood pressure values (mm Hg) are obtained from 10 patients via both methods:

  - $X_1 = $ Direct method: radial arterial catheter
  - $X_2 = $ Indirect method: the bell component of the stethoscope
4. Calculating Test Statistics

Average of differences

\[ tc = \frac{\bar{D}}{S_D} \]

Standard Deviation of differences

\[ S_D = \frac{SD}{\sqrt{n}} \]

Standard Error of differences

Sample size

Indian Association for Statistics in Clinical Trials
4. Calculating Test Statistics

<table>
<thead>
<tr>
<th>Patients</th>
<th>X1</th>
<th>X2</th>
<th>D</th>
<th>D^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>130</td>
<td>128</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>102</td>
<td>100</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>154</td>
<td>155</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>113</td>
<td>110</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>139</td>
<td>140</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>125</td>
<td>120</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>156</td>
<td>155</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>108</td>
<td>105</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>161</td>
<td>160</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>105</td>
<td>107</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>129.3</td>
<td>128</td>
<td>13</td>
</tr>
</tbody>
</table>
4. Calculating Test Statistics

- Calculate t-statistic from average of differences and standard error of differences

\[ t_c = \frac{\bar{D}}{S_D} \quad S_D = \frac{SD}{\sqrt{n}} = 0.68 \quad t_c = \frac{1.3}{0.68} = 1.90 \]
6. Compare the Computed Test Statistic Against a Tabled Value

\[ \alpha = 0.05 \]

\[ Df = n-1 = 9 \]

\[ t_c = \frac{1.3}{0.68} = 1.90 \]

\[ t_\alpha(df = 9) = 2.26 \quad \text{Two tailed} \]

\[ t_\alpha(df = 9) = 1.83 \quad \text{One tailed} \]

Reject \( H_0 \) if \( t_c \) is greater than \( t_\alpha \).
Excel Example
Section Slide

ANOVA
Purpose of ANOVA

- One-way ANOVA compares three or more groups defined by a single factor.
  - For example, you might compare control, with drug treatment with drug treatment plus antagonist. Or might compare control with five different treatments.

- Some experiments involve more than one factor. These data need to be analyzed by two-way ANOVA or Factorial ANOVA.
  - For example, you might compare the effects of three different drugs administered at two times. There are two factors in that experiment: Drug treatment and time.
Why not do repeated t-tests?

- Rather than using one-way ANOVA, you might be tempted to use a series of t tests, comparing two groups each time. Don’t do it.

- Repeated t-test increase the chances of type I error or multiple comparison problem

- If you are making comparison between 5 groups, you will need 10 comparison of means

- When the null hypothesis is true the probability that at least 1 of the 10 observed significance levels is less than 0.05 is about 0.29
What Does ANOVA Do?

ANOVA involves the partitioning of variance of the dependent variable into different components:
- A. Between Group Variability
- B. Within Group Variability

More Specifically, The Analysis of Variance is a method for partitioning the Total Sum of Squares into two Additive and independent parts.
Definition of Total Sum of Squares or Variance

\[
\sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ij} - X_{..})^2
\]

<table>
<thead>
<tr>
<th>Case</th>
<th>Group 1</th>
<th>Group 2</th>
<th>...</th>
<th>Group p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X_{11}</td>
<td>X_{21}</td>
<td>...</td>
<td>X_{p1}</td>
</tr>
<tr>
<td>2</td>
<td>X_{12}</td>
<td>X_{22}</td>
<td>...</td>
<td>X_{p2}</td>
</tr>
<tr>
<td>3</td>
<td>X_{13}</td>
<td>X_{23}</td>
<td>...</td>
<td>X_{p3}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>X_{1n}</td>
<td>X_{2n}</td>
<td>...</td>
<td>X_{pn}</td>
</tr>
</tbody>
</table>

Summed across all n times p observations.

Grand average
Definition of Between Sum of Squares

\[ n \sum_{j=1}^{p} (\bar{X}_{ij} - \bar{X}_{..})^2 \]

Sum of squared differences of group means from the grand mean is SS_B
Definition of Within Sum of Squares

\[ \sum_{j=1}^{n} \sum_{i=1}^{n} (X_{ij} - \bar{X}_{.j})^2 \]

<table>
<thead>
<tr>
<th>Case</th>
<th>Group 1</th>
<th>Group 2</th>
<th>...</th>
<th>Group p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X_{11}</td>
<td>X_{21}</td>
<td>...</td>
<td>X_{p1}</td>
</tr>
<tr>
<td>2</td>
<td>X_{12}</td>
<td>X_{22}</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X_{13}</td>
<td>X_{23}</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>X_{1n}</td>
<td>X_{2n}</td>
<td>...</td>
<td>X_{pn}</td>
</tr>
</tbody>
</table>

Sum of squared difference of observations from group means
Partitioning of Variance into Different Components

\[
\sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ij} - \bar{X}.)^2 = n \sum_{j=1}^{p} (\bar{X}.j - \bar{X}.)^2 + \sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ij} - \bar{X}.j)^2
\]

- Total sum of squares
- Between groups sum of squares
- Within groups sum of squares
Test Statistic in ANOVA

- $F = \frac{\text{Between group variability}}{\text{Within group variability}}$

- The source of **Within** group variability is the individual differences.

- The source of **Between** group variability is effect of independent or grouping variables.

- Within group variability is sampling error across the cases

- Between group variability is effect of independent groups or variables
4. Calculating Test Statistics

- \( F = \frac{SS_b / df_B}{SS_w / df_w} \)
6. Compare the Computed Test Statistic Against a Tabled Value

\[ \alpha = .05 \]

If \( F_c > F_\alpha \)  Reject \( H_0 \)

If \( F_c > F_\alpha \)  Can not Reject \( H_0 \)
Example

- Suppose we had patients with myocardial infarction in the following groups:
  - Group 1: A music therapy group
  - Group 2: A relaxation therapy group
  - Group 3: A control group

- 15 patients are randomly assigned to the 3 groups and then their stress levels are measured to determine if the interventions were effective in minimizing stress.
Example

- **Dependent Variable**
  - The stress scores. The ranges are from zero (no stress) to 10 (extreme stress)

- **Independent Variable or Factor**
  - Treatment Conditions (3 levels)
<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>
# Sum of Squares for Each Group

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

\[ SS_1 = 20 \quad SS_2 = 10 \quad SS_3 = 16 \]

\[ n_1=5 \quad n_2=5 \quad n_3 = 5 \]
SS Within

\[
\begin{align*}
SS_1 &= \sum (X_{1j} - \bar{X}_1)^2 \\
&= (0 - 3)^2 + (6 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (3 - 3)^2 = 20 \\
SS_2 &= \sum (X_{2j} - \bar{X}_2)^2 \\
&= (1 - 2)^2 + (4 - 2)^2 + (3 - 2)^2 + (2 - 2)^2 + (0 - 2)^2 = 10 \\
SS_3 &= \sum (X_{3j} - \bar{X}_3)^2 \\
&= (5 - 7)^2 + (6 - 7)^2 + (10 - 7)^2 + (8 - 7)^2 + (6 - 7)^2 = 16 \\
SS_{Within} &= 20 + 10 + 16 = 46
\end{align*}
\]
SS Between

\[
SS_{\text{Between}} = 5(3 - 4)^2 + 5(2 - 4)^2 + 5(7 - 4)^2 = 70
\]

- Number of cases
- Grand average
- Group 1 average
- Group 2 average
- Group 3 average
Sum of Squares Total

$$SST_{\text{Total}} = (0 - 4)^2 + (6 - 4)^2 + (2 - 4)^2 + (4 - 4)^2 + (3 - 4)^2 + (1 - 4)^2 + (4 - 4)^2 + (3 - 4)^2 + (2 - 4)^2 + (0 - 4)^2 + (5 - 4)^2 + (6 - 4)^2 + (10 - 4)^2 + (8 - 4)^2 + (6 - 4)^2 = 116$$
Components of Variance

$$SS_{Total} = SS_{Between} + SS_{Within}$$

$$\sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ij} - \bar{X}_{.})^2 = n \sum_{j=1}^{p} (\bar{X}_{.j} - \bar{X}_{.})^2 + \sum_{j=1}^{p} \sum_{i=1}^{n} (X_{ij} - \bar{X}_{.j})^2$$

$$116 = 70 + 46$$
Degrees of Freedom

\[ \text{Df}_{\text{between}} = 3 - 1 \]

\[ \text{Df}_{\text{within}} = 15 - 3 \]

\[ \text{df}_B = k - 1 \]

\[ \text{df}_w = N - k \]
Test Statistic

\[ \text{MS}_{\text{Between}} = \frac{70}{2} = 35 \]

\[ \text{MS}_{\text{Within}} = \frac{46}{12} = 3.83 \]

\[ F_c = \frac{\text{MS}_{\text{Between}}}{\text{MS}_{\text{Within}}} \]

\[ F_c = \frac{35}{3.83} = 9.13 \]

\[ F_{\alpha} = 3.88 \]
Conclusions

\[ F_c = 9.13 \text{ } > \text{ } F_\alpha = 3.88 \]

\[ F_c > F_\alpha \text{ Therefore Reject } H_0 \]
## One-way ANOVA Summary

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>$F_c$</th>
<th>$F_\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>70</td>
<td>2</td>
<td>35</td>
<td>9.13</td>
<td>3.88</td>
</tr>
<tr>
<td>Within</td>
<td>46</td>
<td>12</td>
<td>3.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>116</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Excel Example
Correlation and Regression

- Correlation measures the strength/predictability of a relationship between two variables.
  - Unit-less
  - Ranges between –1 and 1
  - The closer to –1, the stronger the negative linear relationship
  - The closer to 1, the stronger the positive linear relationship
  - The closer to 0, the weaker any positive linear relationship
  - The two variables are treated as equals $r_{xy} = r_{yx}$

---


Statistics Workshop
Correlation ....

Linear relationships

Curvilinear relationships

Strong relationships

Weak relationships


Statistics Workshop
Correlation ...

| Normally distributed | Not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant |

The distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.81

One outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.
Correlation ...

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.13</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

Mean | 9 | 7.5 | 9 | 7.5 | 9 | 7.5 | 9 | 7.5 |
Cor  | 0.816 | 0.816 | 0.816 | 0.816 |
Example 1  Sleeping with one's shoes on is strongly correlated with waking up with a headache. Therefore, sleeping with one's shoes on causes headache. This commits the correlation-implies-causation fallacy, as it prematurely concludes that sleeping with one's shoes on causes headache. A more plausible explanation is that both are caused by a third factor, in this case alcohol intoxication, which thereby gives rise to a correlation.

Example 2  Young children who sleep with the light on are much more likely to develop myopia in later life. Scientific example from a study at the University of Pennsylvania Medical Center. Published in the May 13, 1999 issue of Nature. However, a later study at The Ohio State University did not find that infants sleeping with the light on caused the development of myopia. It did find a strong link between parental myopia and the development of child myopia, also noting that myopic parents were more likely to leave a light on in their children's bedroom. In this case, the cause of both conditions is parental myopia.

Example 3  As ice cream sales increase, the rate of drowning deaths increases sharply. So, ice cream causes drowning. This example fails to recognize the importance of time in relationship to ice cream sales. Ice cream is sold during the summer months at a much greater rate, and it is during the summer months that people are more likely to engage in activities involving water, such as swimming. The increased drowning deaths are simply caused by more exposure to water based activities, not ice cream.

Coincidence: (1) Since the 1950s, both the atmospheric CO2 level & crime levels have increased sharply. Hence, atmospheric CO2 causes crime.

(2) With a decrease in the number of pirates, there has been an increase in global warming over the same period. Therefore, global warming is caused by a lack of pirates.
Correlation does not imply causation
Regression: Purpose of Linear Relationship

- The primary purpose of Regression analysis is the development of an equation that can be used for predicting values on some *Dependent Variable, Y*, given *Independent Variables, X*, for all members of a population.

- One of the most important functions of science is the description of natural phenomenon in terms of ‘functional relationships’ between variables. E.g. If one is given a temperature value in the °C Scale (say X), then the corresponding value in the Fahrenheit Scale (say Y), can be calculated by the formula:
  - Y = 32 + 1.8 X
  - If the °C temperature = 10, the Fahrenheit temperature =:
    - Y = 32 + 1.8 (10) = 32 + 18 = 50
  - Similarly, if the °C = 20, the Fahrenheit temperature must be:
    - Y = 32 + 1.8 (20) = 32 + 36 = 68
  - We can plot this relationship on the usual rectangular system of coordinates.
What is “Linear”?  

- Remember this:
  - \( Y = a + bX \) ?
    - \( X \) = Independent variable
    - \( Y \) = Dependent variable
    - \( b \) = Slope (angle and direction)
    - \( a \) = Intercept (point at which line intersects Y axis)

E.g. \( Y = 2X \), A slope of 2 means that every 1-unit change in \( X \) yields a 2-unit change in \( Y \).
Regression and Prediction

- As a university admissions officer, what GPA would you predict for a student who earns a score of 650 on SAT-V?

- If the relationship between X and Y is not perfect, you should attach error to your prediction.

- Correlation and Regression

- Determining the Line of Best Fit or Regression Line using Least Squares Criterion.
Selection of Regression Line

- Residual or error of prediction = \( (Y - Y') \)
  - Positive or negative

- Regression line, \( Y' = a + bX \), is chosen so that the sum of the squared prediction error for all cases, \( \sum(Y - Y')^2 \), is as small as possible.
Which line has the best “fit” to the data?
Estimating the Coefficients…

- In much the same way we base estimates of $\mu$ on $\bar{x}$, we estimate $\beta_0$ on $b_0$ and $\beta_1$ on $b_1$, the y-intercept and slope (respectively) of the least squares or regression line given by:

$$\hat{y} = b_0 + b_1x$$

- Application of the least squares method and it produces a straight line that *minimizes* the sum of the squared differences between the points and the line.
Least Squares Line...

These differences are called **residuals**.

This line minimizes the sum of the squared differences between the points and the line...

This line minimizes the sum of the squared differences between the points and the line...

\[ y = 0.934 + 2.114x \]
Least Squares Line...

- The coefficients $b_1$ and $b_0$ for the least squares line...

$$\hat{y} = b_0 + b_1 x$$

- ...are calculated as:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$
Calculation of Regression Line (SAT, GPA)

<table>
<thead>
<tr>
<th>Y (GPA)</th>
<th>X (SAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.60</td>
<td>400.00</td>
</tr>
<tr>
<td>2.00</td>
<td>350.00</td>
</tr>
<tr>
<td>2.20</td>
<td>500.00</td>
</tr>
<tr>
<td>2.80</td>
<td>400.00</td>
</tr>
<tr>
<td>2.80</td>
<td>450.00</td>
</tr>
<tr>
<td>2.60</td>
<td>550.00</td>
</tr>
<tr>
<td>3.20</td>
<td>550.00</td>
</tr>
<tr>
<td>2.00</td>
<td>600.00</td>
</tr>
<tr>
<td>2.40</td>
<td>650.00</td>
</tr>
<tr>
<td>3.40</td>
<td>650.00</td>
</tr>
<tr>
<td>2.80</td>
<td>700.00</td>
</tr>
<tr>
<td>3.00</td>
<td>750.00</td>
</tr>
<tr>
<td>30.80</td>
<td>6550.00</td>
</tr>
</tbody>
</table>
Plot of Data (SAT, GPA)

\[ y = 0.0021x + 1.43 \]
Plot of Data (SAT, GPA)

Regression line shows predicted values. Difference between predicted & observed is the residual.

Slope shows change in Y associated to change in one unit of X.

y' = 0.0021x + 1.43

Intercept
Sources of Variation

- The sum of Squares of the Dependent Variable is partitioned into two components:
  - One due to Regression (Explained)
  - One due to Residual (Unexplained)
Partitioning of Sum of Squares

\[
\begin{align*}
SS \text{ Total} & = SS \text{ Regression} + SS \text{ Residual} \\
\Sigma (Y - \bar{Y})^2 & = \Sigma (Y' - \bar{Y})^2 + \Sigma (Y - Y')^2 \\
\Sigma (Y - \bar{Y})^2 & = r^2 \Sigma (Y - \bar{Y})^2 + (1 - r^2) \Sigma (Y - \bar{Y})^2 \\
3.20 & = .25(3.20) + (1-.25)(3.20) \\
3.20 & = .80 + 2.40
\end{align*}
\]
## Testing Statistical Significance of Variance Explained

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>SS</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.80</td>
<td>1</td>
</tr>
<tr>
<td>Residual</td>
<td>2.40</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>3.20</td>
<td>11</td>
</tr>
</tbody>
</table>
Testing Statistical Significance of Variance Explained

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.80</td>
<td>1</td>
<td>0.80</td>
</tr>
<tr>
<td>Residual</td>
<td>2.40</td>
<td>10</td>
<td>0.24</td>
</tr>
<tr>
<td>Total</td>
<td>3.20</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>
## Testing Statistical Significance

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.80</td>
<td>1</td>
<td>0.80</td>
<td>3.33</td>
<td>0.1</td>
</tr>
<tr>
<td>Residual</td>
<td>2.40</td>
<td>10</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3.20</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### A. Testing the proportion of variance due to regression

- $H_0: \text{regression is not significant}$
- $H_a: \text{regression is significant}$
B. Testing the Regression Coefficient

\[ H_0 : \beta = 0 \quad \text{Since the } p > \alpha \text{ Fail to reject } H_0 \]

\[ H_a : \beta \neq 0 \]
Calculation of Predicted Values and Residuals

\[
Y' = 1.42 + 0.0021X
\]

<table>
<thead>
<tr>
<th>Y (GPA)</th>
<th>X (SAT)</th>
<th>Y’</th>
<th>Y-Y’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.60</td>
<td>400.00</td>
<td>2.26</td>
<td>-0.66</td>
</tr>
<tr>
<td>2.00</td>
<td>350.00</td>
<td>2.16</td>
<td>-0.16</td>
</tr>
<tr>
<td>2.20</td>
<td>500.00</td>
<td>2.47</td>
<td>-0.27</td>
</tr>
<tr>
<td>2.80</td>
<td>400.00</td>
<td>2.26</td>
<td>0.54</td>
</tr>
<tr>
<td>2.80</td>
<td>450.00</td>
<td>2.37</td>
<td>0.43</td>
</tr>
<tr>
<td>2.60</td>
<td>550.00</td>
<td>2.58</td>
<td>0.02</td>
</tr>
<tr>
<td>3.20</td>
<td>550.00</td>
<td>2.58</td>
<td>0.62</td>
</tr>
<tr>
<td>2.00</td>
<td>600.00</td>
<td>2.68</td>
<td>-0.68</td>
</tr>
<tr>
<td>2.40</td>
<td>650.00</td>
<td>2.78</td>
<td>-0.38</td>
</tr>
<tr>
<td>3.40</td>
<td>650.00</td>
<td>2.78</td>
<td>0.62</td>
</tr>
<tr>
<td>2.80</td>
<td>700.00</td>
<td>2.89</td>
<td>-0.09</td>
</tr>
<tr>
<td>3.00</td>
<td>750.00</td>
<td>2.99</td>
<td>0.01</td>
</tr>
<tr>
<td>Sum</td>
<td>30.80</td>
<td>6550.00</td>
<td>30.80</td>
</tr>
<tr>
<td>Average</td>
<td>2.57</td>
<td>545.83</td>
<td>2.57</td>
</tr>
</tbody>
</table>
Standard Error of Estimate of Y Regressed on X

\[ Se = \sqrt{\frac{\sum(Y_i - Y')^2}{n-2}} = .49 \]
Interpretation of Standard Error of Estimate

- The average amount of error in predicting GPA scores is 0.49.
- The smaller the standard error of estimate, the more accurate the predictions are likely to be.
Assumptions

- X and Y are normally distributed
Assumptions

- X and Y are normally distributed
- The relationship between X and Y is linear and not curved
Assumptions

- X and Y are normally distributed
- The relationship between X and Y is linear and not curved
- The variation of Y at particular values of X is not proportional to X
Assumptions

1. X and Y are normally distributed

2. The relationship between X and Y is linear and not curved

3. The variation of Y at particular values of X is not proportional to X

4. There is negligible error in measurement of X
Residual Analysis: check assumptions

- The residual for observation $i$, $e_i$, is the difference between its observed and predicted value.

- Check the assumptions of regression by examining the residuals:
  - Examine for linearity assumption.
  - Examine for constant variance for all levels of $X$ (homoscedasticity).
  - Evaluate normal distribution assumption.
  - Evaluate independence assumption.

- Graphical Analysis of Residuals:
  - Can plot residuals vs. $X$.

Skedasis means "Dispersion" => related to variance.  
Homo means "same" 
Hetero means "different" 

A random variable is homoscedastic, if the random variables have finite (constant) variances. 

A sequence of random variables is heteroscedastic, or heteroskedastic, if the random variables have different variances.
Regression Diagnostics…

- There are three conditions that are required in order to perform a regression analysis. These are:
  - The error variable must be normally distributed,
  - The error variable must have a constant variance, &
  - The errors must be independent of each other.

- How can we diagnose violations of these conditions?

  ➞ Residual Analysis, that is, examine the differences between the actual data points and those predicted by the linear equation…
Residual Analysis…

- Recall the deviations between the actual data points and the regression line were called *residuals*. Excel calculates residuals as part of its regression analysis:

- We can use these residuals to determine whether the error variable is nonnormal, whether the error variance is constant, and whether the errors are independent…
Nonnormality…

- We can take the residuals and put them into a histogram to visually check for normality…

…we’re looking for a bell shaped histogram with the mean close to zero. ✓
Heteroscedasticity...

- When the requirement of a constant variance is violated, we have a condition of *heteroscedasticity*.

We can diagnose heteroscedasticity by plotting the residual against the predicted $\hat{y}$. 
If the variance of the error variable ($\sigma^2_e$) is not constant, then we have “heteroscedasticity”. Here’s the plot of the residual against the predicted value of $y$:

There doesn’t appear to be a change in the spread of the plotted points, therefore no heteroscedasticity.
Residual Analysis for Linearity

- **Not Linear**
- **Linear**

Residual Analysis for Homoscedasticity

- Non-constant variance
- Constant variance

Residual Analysis for Independence

- Not Independent
- Independent

Procedure for Regression Diagnostics…

1. Develop a model that has a theoretical basis.
2. Gather data for the two variables in the model.
3. Draw the scatter diagram to determine whether a linear model appears to be appropriate. Identify possible outliers.
4. Determine the regression equation.
5. Calculate the residuals and check the required conditions.
6. Assess the model’s fit.
7. **If the model fits the data**, use the regression equation to predict a particular value of the dependent variable and/or estimate its mean.
The Use of Simple Regression

- Answering Research Questions and Testing Hypothesis
- Making Prediction about Some Outcome or Dependent Variable
- Assessing an Instrument Reliability
- Assessing an Instrument Validity
Acupuncture .....
Proportions
Introduction

- **Categorical Data**: Frequency Counts of observations occurring in response categories

- **Contingency Table**: When you look at the relationship between just 2 categorical variables and cells contain frequencies of outcomes
  - Name given by Karl Pearson in 1904
  - Also called cross-classification tables
  - If one variable has X levels, and the other Y, then have an I x J table

- **2x2 table**: When each of the categorical variables has only 2 levels
Notation

- Let the 2 variables be denoted by $X$ and $Y$
- Let $\pi_{ij}$ be the joint distribution of $X$ and $Y$
- The marginal distributions then are given by $\pi_i+$ and $\pi_+j$
- Assume $Y$ to be response variable, and $X$ to be explanatory variable
- Note: If $X$ is fixed rather than random, then joint distribution concept is no longer meaningful
- Think about distribution of $Y$ for a particular level of $X$
- That is, think in terms of $\pi_{j|i}$
### Notation (continued)

<table>
<thead>
<tr>
<th>X (i)</th>
<th>( \pi_{11} )</th>
<th>( \pi_{12} )</th>
<th>( \pi_{1+} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( (\pi_{1</td>
<td>1}) )</td>
<td>( (\pi_{2</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( \pi_{21} )</td>
<td>( \pi_{22} )</td>
<td>( \pi_{2+} )</td>
</tr>
<tr>
<td></td>
<td>( (\pi_{1</td>
<td>2}) )</td>
<td>( (\pi_{2</td>
</tr>
<tr>
<td>Total</td>
<td>( \pi_{+1} )</td>
<td>( \pi_{+2} )</td>
<td>1.0</td>
</tr>
</tbody>
</table>

where \( \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \)
So, the point of all this is …..

- That we try to see if X and Y are independent

- Number of ways to evaluate independence

- Note, that independence implies \( \pi_{ij} = \pi_{i+} * \pi_{+j} \)

- Which means under independence, one would expect \( \pi_{j|i} = \pi_{+j} \)

- Before we get into testing for independence, lets take a look at estimating some of the following measures:
  
  - Difference in proportions
  - Relative Risk
  - Odds Ratio
Note that ..... 

- All estimates of proportions are based on the cell frequencies

- For e.g., if the cell frequencies are denoted by \{n_{ij}\}, then an estimate of \(\pi_{ij} = p_{ij} = \frac{n_{ij}}{n}\), where \(n\) is the grand total

- Consequently, \(\pi_{j|i} = \frac{n_{ij}}{n_{i+}}\)
Difference in Proportions

- We can compare the 2 rows by looking at the difference in proportions $\pi_{1|1} - \pi_{1|2}$

- Note that since $\pi_{2|1} = 1 - \pi_{1|1}$, above is the same as $-(\pi_{2|1} - \pi_{2|2})$

- Under independence, one would expect this difference to be 0

- So, one could construct a large sample confidence interval for this difference assuming asymptotic normality, and see if the CI contains 0

- Note that one could compare columns rather than rows in a similar manner
Relative Risk

- Sometimes, the difference between proportions itself is not as interesting

- For e.g., the difference between 0.1 and 0.01 maybe very different than the difference between 0.4 and 0.49

- In this case, one might want to look at the relative risk (RR)

- For response 1, the RR is $\pi_{1|1} / \pi_{1|2}$

- Under independence, one expects the RR to be 1

- Note, the RR for response 2 is generally different from 1
Odds Ratio

- Problem with RR is that for a 2x2 table, one gets 2 measures
- Hence a new measure: the Odds Ratio (OR)
- Within row 1, the odds that response is in column 1 rather than column 2 is $\pi_{11} / \pi_{21}$ or $\pi_{11} / \pi_{12}$
- Similarly, within row 2, its $\pi_{12} / \pi_{22}$ or $\pi_{21} / \pi_{22}$
- OR is the ratio of the odds in each row $= \pi_{11} \pi_{22} / \pi_{12} \pi_{21}$
- Under independence, one expects the OR to be 1
- As before, one could construct an asymptotic CI
Odds Ratio versus Relative Risk

- OR can equal any non negative number, as can the RR

- When OR > 1 => subjects in row 1 are more likely to make the first response than are subjects in row 2

- For instance, OR = 4 => the odds of first response are 4 times higher in row 1 than in row 2

- Note, this does not mean $\pi_{1|1} = 4 \times \pi_{1|2}$

- $\text{OR} = \text{RR} \times \left( \frac{1 - \pi_{1|2}}{1 - \pi_{1|1}} \right)$
Example: Cross-classification of Oral Contraceptive Use and Myocardial Infarction

<table>
<thead>
<tr>
<th>OC</th>
<th>MI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>23</td>
<td>34</td>
</tr>
<tr>
<td>No</td>
<td>35</td>
<td>132</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>166</td>
</tr>
</tbody>
</table>

- $\pi_{1|1} = 0.1/0.25 = 0.4$
- $\pi_{2|1} = 0.15/0.25 = 0.6$
- $\pi_{1|2} = 0.21$ & $\pi_{2|2} = 0.79$
- Difference in prop. = 0.19
- RR for MI yes = 0.4/0.21 = 1.9
- => OC users twice as likely to get MI than non-OC users

<table>
<thead>
<tr>
<th>OC</th>
<th>MI</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>No</td>
<td>0.16</td>
<td>0.59</td>
</tr>
<tr>
<td>Total</td>
<td>0.26</td>
<td>0.74</td>
</tr>
</tbody>
</table>

- OR = 0.1*0.59/0.16*0.15 = 2.5
- => the odds of MI are 2.5 times higher among OC users
Other Tests for Independence

- The earlier tests based on the 3 measures relied on the asymptotic distribution of the measures.
- There are other methods to test for independence using statistics based on the cell frequencies themselves.
- To do this, we assume different sampling distributions for the cell frequencies.
- Note, the observed frequencies in the cells are denoted by $n_{ij}$ and the expected frequencies are denoted by $m_{ij}$. 
Poisson Distribution

- Consider \( N = IJ \) cells, and \( n_i \) frequencies in each of these cells
- Since \( n_i \) has to be positive integers, one natural choice for the sampling distribution would be Poisson with expected frequencies \( m_i \)
- Is useful for counts of events that occur randomly over time or space
- However, in most experiments, one has a fixed total \( n \) with the constraint that each of the \( n_i \) has to be less than \( n \)
- Which leads us to the next sampling distribution
Multinomial Distribution

- Conditional on $n$ and with the added constraint $\Sigma n_i = n$, one gets the multinomial as the sampling distribution with expected frequencies $m_i = n^*\pi_i$

- For a 2x2 table, the expected frequencies $m_{ij}$ are given by

  $$n^* \pi_{i+} \pi_{+j}$$

- Look at the statistic $\Sigma (n_{ij} - m_{ij})^2 / m_{ij} \sim \chi^2$ with 1 df where, the $m_{ij}$ are estimated from the data

- This is the $\chi^2$ test for independence – first proposed by Karl Pearson in 1900, though he got the df wrong!

- This was later corrected by Fisher in 1922
Hypergeometric Distribution

- If one conditioned on the row and column totals, as well as the grand total => only 1 cell in a 2x2 table that’s random
- This random cell count follows a hypergeometric distribution
- Fisher’s famous tea drinker example

<table>
<thead>
<tr>
<th></th>
<th>Guess First</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Milk</td>
<td></td>
<td></td>
<td>Milk</td>
</tr>
<tr>
<td>Milk</td>
<td>3</td>
<td>1</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Tea</td>
<td>1</td>
<td>3</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>4</td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>
Things to watch out for….

- Not all 2x2 tables are created equal ….

- Consider the following table on performance of the President

<table>
<thead>
<tr>
<th>First Survey</th>
<th>Second Survey</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>App</td>
<td></td>
</tr>
<tr>
<td>App</td>
<td>794</td>
<td>150</td>
</tr>
<tr>
<td>Disapp</td>
<td>86</td>
<td>570</td>
</tr>
<tr>
<td>Total</td>
<td>880</td>
<td>720</td>
</tr>
</tbody>
</table>

- Note that the respondents are the same in the 2 surveys

- Sort of like matched pairs

- Need other tests like McNemar’s test to analyze
Things to watch out for…. (continued)

- When you have more than 2 dimensional tables

- Let's say you are in a shop to buy hats, and see hats on 2 tables

<table>
<thead>
<tr>
<th>Color</th>
<th>Table 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fit</td>
<td>Not Fit</td>
</tr>
<tr>
<td>Black</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Gray</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Color</th>
<th>Table 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fit</td>
<td>Not Fit</td>
</tr>
<tr>
<td>Black</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Gray</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>26</td>
</tr>
</tbody>
</table>

- The OR for both tables is the same = 1.59!

- One could conclude that in general, black hats fit better.
Things to watch out for… (continued)

- Now, say on the next day, the hats from 2 tables are pooled

<table>
<thead>
<tr>
<th>Color</th>
<th>Pooled Table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fit</td>
</tr>
<tr>
<td>Black</td>
<td>12</td>
</tr>
<tr>
<td>Gray</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
</tr>
</tbody>
</table>

- The OR now is 0.44!

- Called “Simpson’s Paradox”

- Need to adjust for confounders such as tables

- Use tests like the *Cochran-Mantel-Haenszel*
In Conclusion

- Brief overview of Contingency Tables analysis, restricted to 2x2s
- Most of the methods can be extended quite easily to \( R \times C \) tables
- However, extensions to tables that are more than 2-dimensional requires caution
- Lots more that exists than has been covered here, like maximum likelihood estimates, logit, probit and loglinear models and logistic regression
Logistic Regression
Logistic Regression

- Linear regression: two quantitative variables

- Often we have a binary variable, such as response in the acupuncture study
  - LINEAR OR

This can arise from a nonlinear relationship between the and a predictor variable
A given change in $x$ leads to a smaller change in $y$ when $y$ is closer to 0 or to 1.
Logistic Regression

- So the probability that \( y = 1 \) plotted against \( x \), can be a curve and not a straight line.

- The outcome now is continuous, but is still bounded within 0 and 1.

- Could we use a model like this => \( P(Y) = a + bx \).
Logistic Regression

- **Probability and Odds**
  
  Probability = # of successes / total # of attempts
  
  Risk Ratio = Probability in ‘Treatment group’/Probability in ‘Control Group’
  
  Odds = # of successes / # of failures
  
  \[
  \text{Odds} = \frac{P(\text{success})}{1 - P(\text{success})} = \frac{P(\text{success})}{P(\text{failure})}
  \]
  
  Odds Ratio = Odds in ‘Treatment group’/Odds in ‘Control Group’

  \[
  \text{Odds Ratio} = \frac{\text{Odds in 'Treatment group'}}{\text{Odds in 'Control Group'}}
  \]

<table>
<thead>
<tr>
<th></th>
<th>Failure</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>d</td>
<td>c</td>
</tr>
<tr>
<td>Treatment</td>
<td>b</td>
<td>a</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{P(Success/Treatment)} &= \frac{a}{a+b} \\
\text{P(Success/Control)} &= \frac{c}{c+d} \\
\text{Risk Ratio} &= \frac{a/(a+b)}{c/(c+d)} = \frac{a/c}{(c+d)/(a+b)} \\
\text{Odds(Success/Treatment)} &= \frac{a}{b} \\
\text{Odds (Success/Failure)} &= \frac{c}{d} \\
\text{Odds Ratio} &= \frac{a/b}{c/d} = \frac{a*d}{c*d}
\end{align*}
\]
Acupuncture

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>&lt;35% Drop</th>
<th>&gt;35% Drop</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>95</td>
<td>45</td>
<td>140</td>
</tr>
<tr>
<td>Acupuncture</td>
<td>74</td>
<td>87</td>
<td>161</td>
</tr>
<tr>
<td>Grand Total</td>
<td>169</td>
<td>132</td>
<td>301</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>&lt;35% Drop</th>
<th>&gt;35% Drop</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>68%</td>
<td>32%</td>
<td>100%</td>
</tr>
<tr>
<td>Acupuncture</td>
<td>46%</td>
<td>54%</td>
<td>100%</td>
</tr>
<tr>
<td>Grand Total</td>
<td>56%</td>
<td>44%</td>
<td>100%</td>
</tr>
</tbody>
</table>

|                           | Probability of > 35% drop/Acupuncture | 0.54 |
|                           | Probability of > 35% drop/Control     | 0.32 |
| Risk Ratio                |                                         | 1.68 |
| Odds of > 35% drop/Acupuncture |                                         | 1.18 |
| Odds of > 35% drop/Control |                                         | 0.47 |
| Odds Ratio                |                                         | 2.48 |
Logistic Regression

- Odds are convenient because they lie between 0 and infinity. Taking the natural log allows this to vary between $-\infty$ and $+\infty$:
  - $\text{Odds}(Y=1) = a + bx$
  - $\ln[\text{Odds} (Y=1)] = a + bx$

Suppose $X = 1$ is Treatment and $X = 0$ is Control

- $\ln [\text{Odds} (Y=1)/X=1] = a + b$
- $\ln[\text{Odds} (Y=1)/X=0] = a$

- $\ln [\text{Odds} (Y=1)/X=1] - \ln[\text{Odds} (Y=1)/X=0] = b$
- $\ln [\text{Odds} (Y=1)/X=1/\text{Odds} (Y=1)/X=0] = b$
- Odds Ratio = $e^b$

- If $b = 0$ then Odds Ratio = 1 $\iff$ Implies $X$ not related to $Y$
### Acupuncture ...

Call: glm(formula = response ~ group, family = binomial(link = logit), data = Splus, na.action = na.exclude, control = list(epsilon = 0.0001, maxit = 50, trace = F))

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.7470267</td>
<td>0.1799260</td>
<td>-4.151856</td>
</tr>
<tr>
<td>group</td>
<td>0.9088697</td>
<td>0.2395421</td>
<td>3.794197</td>
</tr>
</tbody>
</table>

\[
\exp(\text{fit1$coeff})
\]

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.4737731</td>
</tr>
<tr>
<td>group</td>
<td>2.481516</td>
</tr>
</tbody>
</table>

> anova(fit1,test="Chisq")

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Terms added sequentially (first to last)</th>
<th></th>
<th></th>
<th></th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Df Deviance Resid. Df Resid. Dev</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NULL</td>
<td>300</td>
<td>412.7149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>group</td>
<td>1</td>
<td>14.7487</td>
<td>299</td>
<td>397.9662</td>
</tr>
</tbody>
</table>

Indian Association for Statistics in Clinical Trials
Survival Analysis
Survival Analysis

- Hypothesis: For how many years did patients who had a heart transplant live?

- What is the outcome variable in the above example?
  - outcome variable is time until an event occurs.

- What is the test statistic used here?
Survival Analysis (contd.)

- Statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs.

**Time:**

Start follow-up → Event

Survival time

**Event:** Death, Disease, Relapse, Recovery, Failure

- Examples
  - Mortality studies (death is event).
  - Remission of symptoms (therapy trials).
  - Disease surveillance (when did it occur).
Why censoring occurs?

- No event - before end of the study
- Lost to follow up
- Withdraws
Survival Analysis (contd.)
Survival Function

Let $T = \text{Time of Death (Disease)}$

Survival Function $S(t) = \Pr(\text{alive at time } t) = \Pr(T > t)$

In simple term, this can also be defined as

$$S(t) = \frac{\text{No. of patients surviving longer than } t}{\text{Total number of patients}}$$
Hazard Function

\[ h(t) = \frac{\text{number of patients dying per unit time in the interval } (t, t + \Delta t)}{\text{number of patients surviving at time } t} \]

Hazard function is still not a probability but rate!

Hazard function (estimated) = 4.67 / day
Kaplan Meier Estimate of Survival function

- Calculates the survival function for each individual.
- Non-parametric method and easy to calculate.
- The survival probability can be calculated in the following way:
  - \( p_1 \) = probability of surviving for at least 1 day after transplant.
  - \( p_2 \) = conditional probability of surviving the second day after having survived the first day.
  - so on…
- The cumulative probability of survival time \( t \) is
  \[
  S(t) = p_1 \times p_2 \times \ldots \times p_{t-1} \times p_t
  \]
The Kaplan Meier Estimate of Survival function is used to estimate the probability of survival at a given time point in clinical trials. It is particularly useful for censored data, where some patients may not have reached the event of interest by the end of the study.

### Example: 10 Tumor patients (remission time)

<table>
<thead>
<tr>
<th>t</th>
<th>d_i</th>
<th>n_i</th>
<th>(n_i-d_i)/n_i</th>
<th>S(t) = ∏ [(n_i-d_i)/n_i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
<td>9/10</td>
<td>9/10</td>
</tr>
<tr>
<td>4+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5.7+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6.5</td>
<td>2</td>
<td>7</td>
<td>5/7</td>
<td>(9/10)*(5/7)</td>
</tr>
<tr>
<td>8.4+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>4</td>
<td>3/4</td>
<td>(9/10)<em>(5/7)</em>(3/4)</td>
</tr>
<tr>
<td>10+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2</td>
<td>1/2</td>
<td>(9/10)<em>(5/7)</em>(3/4)*(1/2)</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

d_i = number of events; 
n_i = number at risk

The Kaplan Meier estimate is calculated by multiplying the survival probabilities at each event time. The product is taken over all event times, including censored times.
Median survival time

Kaplan-Meier estimate of median survival time
Survival Analysis for two groups

- Group 1 Treatment
- Group 2 Placebo

Median = 8
Median = 23
Log-rank Test

- Large sample Chi-square test
- Overall comparison of Kaplan-Meier curves
- Observed versus expected cell counts over categories of outcomes
- The Log-rank statistics is given as

\[
\text{Log rank} = \frac{\sum (d_j - e_j)^2}{\sum v(d_j - e_j)} \approx x_1^2 \text{df}
\]

\(d_i\) = no. of deaths
\(e_i\) = expected no. of deaths
Other methods for comparing survival curves

- **Breslow Statistic**
  - Gives greater weight to early observations. It is less sensitive than the Log-Rank test to late events when few subjects remain in the study.

- **Tarone-Ware Statistic**
  - Provides a compromise between the Log-rank and Breslow statistics, with an intermediate weighting scheme. This test maintains power across a wider range of alternatives than do the other two tests.

- **Log-Rank Statistic**
  - Emphasizes failures in the tail of the survival curve, where the number at risk decreases over time, yet equal weight is given to each failure time.
Cox Proportional Hazard model

- To estimate the hazard for an individual
- \( \lambda(t, x) = \lambda_0(t) \exp(\beta x) \)
- \( \ln [\lambda(t, x)] = \ln [\lambda_0(t)] + \beta x \)
Cox Proportional Hazard model (contd.)

- Hazard of a new treatment = \( h(t) \)
- Hazard of a standard treatment = \( h_0(t) \)
- Hazard ratio = \( \frac{h(t)}{h_0(t)} = \exp (\beta_1 X_1) \)
  
- \( h(t) = h_0(t) \exp (\beta_1 X_1) \)

\( X_1 \) → is a time independent covariate
Proportional Hazards Assumption

- The Hazard ratio is constant over time.
- Hazards cross $\Rightarrow$ PROPORTIONAL HAZARDS not met
- Hazards don't cross $\not\Rightarrow$ PROPORTIONAL HAZARDS met
Stratified Cox Model

- Is a modification of Cox proportional hazards model that allows for control by “STRATIFICATION” of the predictor that does not satisfy the proportional hazards assumption.

- The stratified model differ in their baseline hazard functions for each stratum $h_{00}(t)$ and $h_{01}(t)$. For example if the stratum variable is sex then the hazard functions for males and females differ.
Section Slide

Design of Experiments
Things to consider when designing experiments

- Context of the setting
- Objective of the experiment
- Hypotheses to be tested
  - How many
  - What kinds of hypotheses
  - Endpoints needed for the hypotheses to be tested
- Number of times you want to look at the data
  - Decision making at interim time points?
Sample Size
Things required for computing sample size

- Effect size
- Standard deviation
- Errors rate control
- Excel Example
Thank You
Primary care

Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial

Andrew J Vickers, Rebecca W Rees, Catherine E Zollman, Rob McCarney, Claire Smith, Nadia Ellis, Peter Fisher, Robbert Van Haselen

Abstract

Objectives To determine the effects of a policy of “use acupuncture” on headache, health status, days off sick, and use of resources in patients with chronic headache compared with a policy of “avoid acupuncture.”

Design Randomised, controlled trial.

Setting General practices in England and Wales.

Participants 401 patients with chronic headache, predominantly migraine.

Interventions Patients were randomly allocated to receive up to 12 acupuncture treatments over three months or to a control intervention offering usual care.

Main outcome measures Headache score, SF-36 health status, and use of medication were assessed at baseline, three, and 12 months. Use of resources was assessed every three months.

Results Headache score at 12 months, the primary end point, was lower in the acupuncture group (16.2, SD 13.7, n = 161, 34% reduction from baseline) than in controls (22.3, SD 17.0, n = 140, 16% reduction from baseline). The adjusted difference between means is 4.6 (95% confidence interval 2.2 to 7.0; P = 0.0002). This result is robust to sensitivity analysis incorporating imputation for missing data. Patients in the acupuncture group experienced the equivalent of 22 fewer days of headache per year (8 to 38). SF-36 data favoured acupuncture, although differences reached significance only for physical role functioning, energy, and change in health.

Conclusions Acupuncture leads to persisting, clinically relevant benefits for primary care patients with chronic headache, particularly migraine. Expansion of NHS acupuncture services should be considered.

Introduction

Migraine and tension-type headache give rise to notable health, economic, and social costs. Despite the undoubted benefits of medication, many patients continue to experience distress and social disruption. This leads patients to try, and health professionals to recommend, non-pharmacological approaches to headache care. One of the most popular approaches seems to be acupuncture. Each week 10% of general practitioners in England refer patients to acupuncture or practise it themselves, and chronic headache is one of the most commonly treated conditions.

A recent Cochrane review of 26 randomised trials of acupuncture for headache concluded that, although existing evidence supports the value of acupuncture, the quality and amount of evidence are not fully convincing. The review identifies an urgent need for well planned, large scale studies to assess the effectiveness and cost effectiveness of acupuncture under “real” conditions. In 1998 the NHS National Coordinating Centre for Health Technology Assessment commissioned us to conduct such a trial (trial number ISRCTN96537534). Our aim was to estimate the effects of acupuncture in practice; we established an acupuncture service in primary care; we then sought to determine the effects of a policy of “use acupuncture” on headache, health status, days off sick, and use of resources in patients with chronic headache compared with a policy of “avoid acupuncture.” This reflects two real decisions: that made by general practitioners when managing the care of headache patients and that made by NHS entities when commissioning health services.

Methods

The protocol and recruitment methods have been published previously. The study included 12 separate sites consisting of a single acupuncture practice and two to five local general practices. Study sites were located in Merseyside, London and surrounding counties, Wales, and the north and south west of England.

Accrual of patients

Practices searched their databases to identify potential participants. General practitioners then sent letters to suitable patients, providing information about the trial. A researcher at the study centre conducted recruitment interviews, eligibility screening, and baseline assessment by telephone. Patients’ conditions were diagnosed as migraine or tension-type headache, following criteria of the International Headache Society (IHS). Patients aged 18-65 and who reported an average of at least two headaches per month were eligible. Patients were excluded for any of the following: onset of headache disorder less than one year before or at age 50 or older; pregnancy; malignancy; cluster headache (IHS code 3); suspicion that headache disorder had specific aetiology (IHS code 5-11); cranial neuralgias (IHS code 12); and acupuncture treatment in the previous 12 months. Eligible patients completed a baseline headache diary for four weeks. Patients who provided written informed consent, had a mean
Primary care

weekly baseline headache score of 8.75 or more, and completed at least 75% of the baseline diary were randomised to a policy of “use acupuncture” or “avoid acupuncture.” Given a power of 90% and an α of 5%, we estimated that we would require 288 evaluable patients to detect a reduction in headache score of 35% in the acupuncture group, compared with 20% in controls. We assumed a dropout rate of about 25% and planned to randomise 400 patients.

Randomisation
We used randomised minimisation (“biased coin”) to allocate patients. The minimised variables were age, sex, diagnosis (migraine or tension-type), headache score at baseline, number of years of headache disorder (chronicity), and number of patients already allocated to each group, averaged separately by site. We used a secure, password-protected database to implement randomisation, which was thus fully concealed.

Treatment
Patients randomised to acupuncture received, in addition to standard care from general practitioners, up to 12 treatments over three months from an advanced member of the Acupuncture Association of Chartered Physiotherapists. All acupuncturists in the study had completed a minimum of 250 hours of postgraduate training in acupuncture, which included the theory and practice of traditional Chinese medicine; they had practised acupuncture for a median of 12 years and treated a median of 22 patients per week. The acupuncture point prescriptions used were individualised to each patient and were at the discretion of the acupuncturist. Patients randomised to “avoid acupuncture” received usual care from their general practitioner but were not referred to acupuncture.

Outcome assessment
Patients completed a daily diary of headache and medication use for four weeks at baseline and then three months and one year after randomisation. Severity of headache was recorded four times a day on a six-point Likert scale (0-5). The SF-36 health status questionnaire was completed at baseline, three months, and one year. Every three months after randomisation, patients completed additional questionnaires that monitored use of headache treatments and days sick from work or other usual activity. While the study was under way we added an additional end point: we conducted an additional telephone interview at baseline, three months, and one year. The dropout rate was close to that expected and approximately balanced between groups. Patients who dropped out were similar to completers in terms of sex, diagnosis, and chronicity, but they were slightly younger (43 v. 46 years, P = 0.01) and had higher headache score at baseline (29.3 v. 25.6, P = 0.04). Table 1 shows baseline characteristics by group for the 301 patients who completed the trial: the groups are highly comparable. Thirty-one of the patients who withdrew provided three month data, and an additional 45 provided a global assessment. Only 6% of patients (12 in each group) provided no data for headache after randomisation.

Table 2 shows results for medical outcomes for patients completing 12 month follow up. In the primary analysis mean headache scores were significantly lower in the acupuncture group. Scores fell by 34% in the acupuncture group compared with 16% in controls (P = 0.0002). This result was highly robust to sensitivity analysis for missing data (smallest difference between groups 3.85, P = 0.002; see appendix on bmj.com). When we used the prespecified cut-off point of 35% as a clinically significant reduction in headache scores, 22% more acupuncture patients improved than controls, equivalent to a number needed to treat of 4.6 (95% confidence interval 9.3 to 3.0). The difference in days with headache of 1.8 days per four weeks is equivalent to 22 fewer days of headache per year (8 to 38). The effects of acupuncture seemed to be long lasting; although few patients continued to receive acupuncture after the initial three month treatment period (25, 10, and 6 patients received treatment after 3, 6, and 9 months, respectively), headache scores were lower at 12 months than at the follow up after treatment. Medication scores at follow up were lower in the acupuncture group, although differences between groups did not reach significance for all end points. In an unplanned analysis we summed and scaled all medication taken by patients after randomisation and compared groups with adjustment for base-
line scores. Use of medication use fell by 23% in controls but by 37% in the acupuncture group (adjusted difference between groups 15%; 95% confidence interval 3%, 27%; P = 0.01). SF-36 data generally favoured acupuncture (table 3), although differences reached significance only for physical role functioning, energy, and change in health.

Table 2  Headache and medication outcomes. Higher scores indicate greater severity of headache and increased use of medication. Differences between groups are calculated by analysis of covariance. Values are means (SD) unless otherwise indicated

<table>
<thead>
<tr>
<th>End point</th>
<th>Baseline</th>
<th>After treatment (at three months after randomisation)</th>
<th>At 12 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acupuncture</td>
<td>Controls</td>
<td>Acupuncture</td>
</tr>
<tr>
<td>Weekly headache score</td>
<td>24.6 (14.1)</td>
<td>26.7 (16.8)</td>
<td>24.6 (14.1)</td>
</tr>
<tr>
<td>Days of headache in 28 days</td>
<td>15.6 (6.6)</td>
<td>16.2 (8.7)</td>
<td>12.1 (7.2)</td>
</tr>
<tr>
<td>Clinically relevant improvement in score*</td>
<td>65 (41%)</td>
<td>37 (27%)</td>
<td>36 (23%)</td>
</tr>
<tr>
<td>Clinically relevant improvement in frequency†</td>
<td>36 (23%)</td>
<td>17 (13%)</td>
<td>11.0 (13.6)</td>
</tr>
<tr>
<td>Scaled prophylactic medication (weekly)</td>
<td>16.0 (17.6)</td>
<td>13.3 (22.2)</td>
<td>7.9 (17.8)</td>
</tr>
<tr>
<td>Use of any prophylactic medication in 28 days</td>
<td>40 (25%)</td>
<td>45 (32%)</td>
<td>34 (21%)</td>
</tr>
</tbody>
</table>

*As defined in study protocol: 35% or greater improvement in headache score from baseline.
†International Headache Society definition: 50% or greater reduction in days with headache.14
‡Adjusted difference: positive favours acupuncture.
Primary care

Table 3 Health status as scored on the SF-36: values are means (SD)

<table>
<thead>
<tr>
<th>End point</th>
<th>Baseline</th>
<th>After treatment (three months after randomisation)</th>
<th>At 12 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acupuncture</td>
<td>Controls</td>
<td>Difference*</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>n=167: 51.9</td>
<td>n=139: 55.3</td>
<td>3.4 (2.7)</td>
</tr>
<tr>
<td>Role functioning physical</td>
<td>n=167: 64.0</td>
<td>n=139: 59.4</td>
<td>4.6 (3.6)</td>
</tr>
<tr>
<td>Role functioning emotional</td>
<td>n=167: 73.2</td>
<td>n=140: 69.6</td>
<td>3.6 (3.6)</td>
</tr>
<tr>
<td>Energy or fatigue</td>
<td>n=167: 47.9</td>
<td>n=140: 52.2</td>
<td>4.3 (2.0)</td>
</tr>
<tr>
<td>Social functioning</td>
<td>n=167: 71.0</td>
<td>n=140: 73.6</td>
<td>2.6 (2.4)</td>
</tr>
<tr>
<td>Pain</td>
<td>n=167: 59.8</td>
<td>n=140: 66.0</td>
<td>6.2 (3.5)</td>
</tr>
<tr>
<td>General health</td>
<td>n=167: 69.2</td>
<td>n=140: 64.0</td>
<td>5.2 (2.1)</td>
</tr>
<tr>
<td>Health change</td>
<td>n=167: 59.5</td>
<td>n=140: 53.4</td>
<td>6.1 (5.4)</td>
</tr>
</tbody>
</table>

Higher scores indicate better quality of life. Differences between groups are calculated by analysis of covariance.

*Adjusted difference: positive favours acupuncture.

Table 4 Use of resources. Values are means (SD)

<table>
<thead>
<tr>
<th>Resource</th>
<th>Acupuncture</th>
<th>Controls</th>
<th>Difference between groups*</th>
<th>95% CI</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of visits to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General practitioner</td>
<td>1.7 (2.5)</td>
<td>2.3 (3.6)</td>
<td>0.77 (0.8)</td>
<td>0.56 to 1.06</td>
<td>0.10</td>
</tr>
<tr>
<td>Specialist</td>
<td>0.22 (0.9)</td>
<td>0.14 (0.6)</td>
<td>1.13 (1.6)</td>
<td>0.34 to 3.73</td>
<td>0.8</td>
</tr>
<tr>
<td>Complementary therapist</td>
<td>2.0 (7.1)</td>
<td>2.0 (6.6)</td>
<td>0.56 (0.8)</td>
<td>0.18 to 1.72</td>
<td>0.3</td>
</tr>
<tr>
<td>No of days off sick</td>
<td>12.6 (18.9)</td>
<td>18.9 (25.2)</td>
<td>6.3 (16.2)</td>
<td>0.84 (10.8)</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Visits to acupuncturists and physiotherapists are excluded.

*Adjusted difference between groups. Results are expressed as an incident rate ratio—the proportion of events in the acupuncture group compared with controls. Values less than one indicate fewer events in the acupuncture group. For example, the value of 0.77 for visits to general practitioners means that acupuncture patients made 23% fewer visits.

We conducted interaction analyses to determine whether patients responded best to acupuncture. Although improvements in mean headache score over control were much larger for migraine patients (4.9; 95% confidence interval 2.4 to 7.5, n = 294) than for patients who did not meet the criteria for migraine (1.1; 95% confidence interval –2.4 to 4.5, n = 17), the small numbers of patients with tension-type headache preclude us from excluding an effect of acupuncture in this population. The interaction term for baseline score and group was positive and significant (P = 0.004), indicating larger effects of treatment on patients with more severe symptoms, even after controlling for regression to the mean. Predicted improvements in headache score for each quartile of baseline score in acupuncture patients are 22%, 29%, 35%, and 58%; figure 2 shows comparable data for days with headache. Neither age nor chronicity nor sex influenced the results of acupuncture treatment.

Table 4 shows data on use of resources. Patients in the acupuncture group made fewer visits to general practitioners and complementary practitioners than those not receiving acupuncture and took fewer days off sick. Confirming the excellent safety profile of acupuncture, the only adverse event reported was five cases of headache after treatment in four subjects.

Discussion

Main findings

Acupuncture in addition to standard care results in persisting, clinically relevant benefits for primary care patients with chronic headache, particularly migraine, compared with controls. We also found improvements in quality of life, decreases in use of medication and visits to general practitioners, and reductions in days off sick. Methodological strengths of our study include a large sample size, concealed randomisation, and careful follow up. We have maximised the practical value of the trial by comparing the effects of clinically relevant alternatives on a diverse group of patients recruited directly from primary care.

Limitations

Control patients did not receive a sham acupuncture intervention. One hypothesis might be that the effects seen in the acupuncture group resulted not from the physiological action of needle insertion but from the “placebo effect.” Such an argument is not relevant to an assessment of the clinical effectiveness of

Figure 2 Frequency of headache at baseline and after treatment. Red dots are actual values for patients in the acupuncture group; blue squares are for controls. The straight line represents no change; observations above the line improved. The curved lines are regression lines (upper red line for acupuncture, lower blue line for controls) that can be used as predictions. Some outliers have been removed for clarity.
acupuncture because in everyday practice, patients benefit from placebo effects. None the less, good evidence from randomised trials shows that acupuncture is superior to placebo in the treatment of migraine. Furthermore, this study was modelled on Vincent’s earlier double blind, placebo controlled trial in migraine, which makes direct comparison possible. If placebo explained the activity of acupuncture we would expect patients in our control group, who received no treatment, to experience smaller improvements than Vincent’s placebo treated controls, leading to a larger difference between groups. However, improvements in our controls (7.1% from a baseline headache score of 26.7) were similar to those in Vincent’s trial (10.5% from 27.2) and differences between groups are non-significantly smaller in the current trial (4.1 v 8.1). This implies that our findings perhaps cannot be explained purely in terms of the placebo effect. That said, we are unable to rule out such an explanation given our lack of placebo control.

Patients in the trial were not blinded and may therefore have given biased assessments of their headache scores. Measures to minimise bias included minimum contact between trial participants and the study team, extended periods of anonymised diary completion and coaching patients about bias.

The difference between groups is far larger (odds ratio for response 2.5) than empirical estimates of bias from failure to blind (odds ratio 1.2). The similarity of our results to those of the prior blinded study provides further evidence that bias does not completely explain the apparent effects of acupuncture.

Patients recorded all treatments for headache during the course of the study. Use of medication and other therapies (such as chiropractic) was lower in patients assigned to acupuncture, indicating that the superior results in this group were not due to confounding by off-study interventions.

Comparison with other studies

A strength of the current trial is that its results are congruent with much of the prior literature on acupuncture for headache. Effects found in this study that have been previously reported include: differences between acupuncture and control for migraine which increased between follow up after treatment and one year; uncovening effects for tension-type headache; improvements in severity as well as frequency of headaches and increased benefit in patients with more severe headaches.

Conclusion

A policy of using a local acupuncture service in addition to standard care results in persisting, clinically relevant benefits for primary care patients with chronic headache, particularly migraine. Expansion of NHS acupuncture services for headache should be considered.

The views are those of the authors and not that of the NHS. We thank the following for their contributions: Claire Allen was consumer representative; Tim Lancaster provided advice on recruitment methods; Kate Hardy was the study nurse. Acupuncture was provided by Kyriakos Antonakos, Ann Beavis, Reg D’Souza, Joan Davies, Nadia Ellis (who is a coauthor of this paper), Sara Jeravanje, Maureen Lovesey, Bets Mitchell, Alison Nesbitt, Steve Reece, Stephanie Ross, and Hetty Salmon-Roosen.

Contributors: AJV conceived, designed and analysed the study and is its guarantor; RWR, CEZ, CMS, and NE contributed to the original design with particular contributions to outcome assessment (RWR, CMS); patients and treatment (CEZ); acupuncture treatment (NE). RM contributed to design of resource outcome assessment; RM, RVh and PF contributed to development of data collection methods for sensitivity analysis.

Funding: The trial (ISRCTN05357354) was funded by NHS R&I and National Coordinating Centre for Health Technology Assessment (NCHTA) grant: 96/40/15.

Competing interests: NE provides acupuncture as part of her private physiotherapy practice.

Ethical approval: South West Multicentre Research Ethics Committee and appropriate local ethics committees.


(Accepted 21 January 2004)
doi 10.1136/bmj.38029.421863.EB

Integrative Medicine Service, Biostatistics Service, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, NY, NY 10021
Andrew J Vickers assistant attending research methodologist
Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), Social Science Research Unit, Institute of Education, London WC1H 0NS
Rebecca W Rees research officer

Montpelier Health Centre, Bristol BS6 5PT
Catherine E Zollman general practitioner
Department of Psychological Medicine, Imperial College London, London W2 1PD
Rob McCartney research officer
Academic Rheumatology, Westom Education Centre, King's College, London SE5 9RJ
Claire M Smith senior trials coordinator
Department of Health and Social Sciences, Coventry University, Coventry CV1 5FB
Nadia Ellis lecturer
Royal London Homeopathic Hospital, London W11 5PB
Peter Fisher director of research
Robbert Van Haselen deputy director of research
Correspondence to: A J Vickers vickersa@mskcc.org
PGDCD

One program - Multi job opportunities

International Drug Discovery & Clinical Research

IDDCR with its "Integrated Platform" provides Practical training & Resources for global biotech and pharmaceutical companies like ICON, Novartis, Novo Nordisk, Kendle, Quintiles, etc.,

The Post Graduate Diploma in Clinical Development (PGDCD) is a 1 Year program which has been designed as per the requirements of the global Bio-Pharma companies for the various designations in Clinical Research & Development.

- Clinical Data manager
- Clinical SAS Programmer
- Biostatistician
- Data Validation Asst.
- Medical Coder
- Drug safety Officer

This is an exclusive in depth practical program for talented individuals aspiring for a career in Bio-Pharmaceutical industry. All selected candidates will go through the rigorous practical training followed by a Job Placement.

Selection after stringent screening includes written test and personal interview; Limited Seats

Program Highlights

- Only program to offer Hands-on Practical Training in India
- Syllabus designed to fulfill requirements in clinical development Jobs.
- Program led by Subject matter experts with global experience
- Program recognized by Industry
- Includes SAS Institute Inc., (USA) Program
- Adherence and Access to International Associations (IASC, SCDM, ACRP)
- Bank Loan Facility

IDDCR IS THE ONLY COMPANY WHICH HAS A PROVEN PLACEMENT TRACK RECORD OF 96%,

Eligibility

Graduate / Post Graduate with min 60% aggregate, and excellent communication & Inter-Personal skills.

International Drug Discovery & Clinical Research,

# 8-2-682, 5th Floor, Landmark Building, Beside Ohri’s Road No: 12,
Banjara Hills, Hyderabad 500 034, Andhra Pradesh, India.

Tel: +91 40 4005 8899, Mobile: +91 80 0880 0781 Website: http://www.iddcr.com/