TOPICS

- Fitting a line to data points
- Fitting exponentials and other generic functions
- Covariance and correlation
FITTING A LINE

• Example from photoelectric effect calibration:

Measure voltage on capacitor by discharging through galvanometer converting charge to displacement so \( V = \frac{1}{kC} D \) where \( V \) is calibration voltage, \( D \) is galvanometer deflection, and \( 1/kC \) converts \( D \) to \( V \). Here \( V \) is independent \((x)\) and \( D \) is dependent \((y)\) so use \( D = kC \cdot V \)

<table>
<thead>
<tr>
<th>Voltage (V)</th>
<th>Deflection (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>0.2</td>
<td>1.1</td>
</tr>
<tr>
<td>0.4</td>
<td>2.3</td>
</tr>
<tr>
<td>0.6</td>
<td>3.7</td>
</tr>
<tr>
<td>0.8</td>
<td>4.5</td>
</tr>
<tr>
<td>1.0</td>
<td>6.0</td>
</tr>
<tr>
<td>1.2</td>
<td>7.3</td>
</tr>
<tr>
<td>1.4</td>
<td>8.5</td>
</tr>
<tr>
<td>1.6</td>
<td>9.4</td>
</tr>
<tr>
<td>1.8</td>
<td>10.9</td>
</tr>
<tr>
<td>2.0</td>
<td>12.1</td>
</tr>
</tbody>
</table>
FITTING A LINE

- Use least-squares formula from Taylor Sec. 8.2
- Fit the N measurements to the form \( y = A + Bx \)
- \( x \leftrightarrow V \) and \( y \leftrightarrow D \)
- Assume that errors on \( x \) are negligible compared to errors on \( y \)

\[
A = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\Delta}, \quad B = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{\Delta}, \quad \Delta = N \sum x_i^2 - (\sum x_i)^2
\]

\[
\sum x_i = 0.0 + 0.2 + 0.4 \ldots 2.0 = 11.0 \text{ V}
\]
\[
\sum x_i^2 = 0.0 + 0.2^2 + 0.4^2 \ldots 2.0^2 = 15.4 \text{ V}^2
\]
\[
\sum y_i = 0.0 + 1.1 + 2.3 \ldots 12.1 = 65.8 \text{ cm}
\]
\[
\sum x_i y_i = 0.0 \cdot 0.0 + 0.2 \cdot 1.1 + 0.4 \cdot 2.3 \ldots 2.0 \cdot 12.1 = 92.48 \text{ V} \cdot \text{cm}
\]
FITTING A LINE

\begin{align*}
\Delta &= N \sum x_i^2 - \left( \sum x_i \right)^2 = 11 \times 15.4 - 11.0^2 = 48.4 \text{ V}^2 \\
A &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\Delta} = \frac{15.4 \times 65.8 - 11.0 \times 92.48}{48.4} = -0.08 \text{ cm} \\
B &= \frac{N \sum x_i y_i - \sum x_i \sum y_i}{\Delta} = \frac{11 \times 92.48 - 11.0 \times 65.8}{48.4} = 6.06 \text{ cm/V} = kC
\end{align*}

A is consistent with zero: this is a good check, since in this case the intercept should be zero as D should be directly proportional to V.
FITTING A LINE

From spread of measurements, find uncertainty on $y$, $A$, and $B$:

$$\sigma_y = \sqrt{\frac{\sum(y_i - A - Bx_i)^2}{N - 2}} = 0.14 \text{ cm} = \sigma_D$$

$$\sigma_A = \sigma_y \sqrt{\frac{\sum x_i^2}{\Delta}} = 0.14 \times \sqrt{\frac{15.4}{48.4}} = 0.08 \text{ cm}$$

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}} = 0.14 \times \sqrt{\frac{11}{48.4}} = 0.07 \text{ cm/V}$$

Calibration constant $kC$:

$$kC = (6.06 \pm 0.07) \text{ cm/V}$$
Some functions can be easily “converted” to linear and then fitted to a line.

Exponential decay shows up in many places, especially radioactive decays.

Decay rate drops with time:

$$N(t) = N_0 \exp\left(-\frac{t}{\tau}\right)$$

where \( \tau \) is the mean lifetime; half-life is related by factor of \( \ln(2) \)

However, \( \ln(N) = \ln(N_0) - \frac{t}{\tau} \) which is linear!

So, can do a linear fit to \( \ln(N) \) and extract \( \tau \)
FITTING TO AN EXPONENTIAL

- Start with \( N(t) = N_0 e^{-\lambda t} \)
- We want the decay constant \( \lambda \)
- Take natural logarithm to get \( \ln N = \ln N_0 - \lambda t \)
- Identify \( y \) as \( \ln N \), \( x \) as \( t \), \( A \) as \( \ln N_0 \) and \( B \) as \( \lambda \)
- Use previous formulas to find \( \lambda \)
FITTING TO A MORE GENERAL FUNCTION

- Possible to fit a completely general function $f(x)$ where $f$ is dependent on parameters $A, B, C, \ldots$

- Obtain $N$ measurements $y_i$ for $N$ values of $x$

- Construct the chi-squared

$$
\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2}
$$

- Scan over possible values of the parameters $A, B, C, \ldots$ and find the values that cause $\chi^2$ to be minimized

- These values are the most likely values of the parameters. Uncertainties on the parameters are determined from how far you have to scan away from the best values to cause $\chi^2$ to increase by 1.
How do we know if we got a good fit (i.e., does the function f describe the data adequately)?

Can ask how small the minimum $\chi^2$ value was. Smaller values indicate a better fit (fewer, smaller deviations between the data points and the best fit function).

Actual numerical expectations for depend on the number N of data points and the number of parameters (A, B, C, ...) in the function f. More on this in Chapter 12.
COVARIANCE AND CORRELATION

• **Take a desired measurement** $q(x,y)$

• **Error propagation** says

$$
\delta q = \sqrt{\left(\frac{\partial q}{\partial x} \delta x\right)^2 + \left(\frac{\partial q}{\partial y} \delta y\right)^2}
$$

• Making two assumptions here: errors are gaussian and $x$ and $y$ are uncorrelated.

• If there is a correlation, the error on $q$ can be either larger or smaller than our estimate.

• Sometimes (epidemiology, other complex systems) the correlation itself is something interesting to learn.
AN EXAMPLE OF CORRELATED VARIABLES: THE $K^0$ MASS

- **Incoming** $K^+$ hits stationary neutron, producing proton and $K^0$

- The $K^0$ travels a short distance and decays to $\pi^+\pi^-$

- Need to measure angle $\theta_T$ between pions

- Also need to measure $\theta_{\pm}$, angles between pions and $K^0$

- Measuring $\theta_T$ is easy, but $K^0$ direction can be hard if its path is short
AN EXAMPLE OF CORRELATED VARIABLES: THE $K^0$ MASS

- If direction of the blue line is wrong, then $\theta_+$ and $\theta_-$ will be wrong by equal amounts but in opposite directions.

- $\theta_T = \theta_+ + \theta_-$ will still be OK.

- Thus we can say that our measurements of $\theta_+$ and $\theta_-$ will be correlated (actually anticorrelated).
COVARIANCE AND ERROR PROPAGATION

- Degree of correlation can be determined from **covariance**: $\sigma_{xy}$

- Experimentally: $\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$

- Correct propagation of errors for $q(x, y)$ is now:
  
  \[ \sigma_q^2 = \left( \frac{\partial q}{\partial x} \delta x \right)^2 + \left( \frac{\partial q}{\partial y} \delta y \right)^2 + 2 \frac{\partial q}{\partial x} \frac{\partial q}{\partial y} \sigma_{xy} \]

**Note:** When the variables were independent, this term was 0.
COVARIANCE AND ERROR PROPAGATION

- If the covariance $\sigma_{xy}$ is small compared to $\delta x$ or $\delta y$, then the error reduces as expected to standard addition in quadrature.

- In case of maximum positive correlation (worst case scenario), addition is linear. Thus, $\sigma_q \leq \left| \frac{\partial q}{\partial x} \right| \sigma_x + \left| \frac{\partial q}{\partial y} \right| \sigma_y$

- $\sigma_{xy}$ can be positive or negative (in $K^0$ case would find negative $\sigma_{xy}$).

- If the covariance $\sigma_{xy}$ is negative, then the error on the result is actually smaller than from addition in quadrature!
COVARIANCE VS. CORRELATION

• The covariance \( \sigma_{xy} \) can be normalized to create a correlation coefficient \( r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \)

• \( r \) can vary between \(-1\) and \(1\).
  - \( r = 0 \) indicates that the variables are uncorrelated;
  - \(|r| = 1\) means the variables are completely correlated (i.e. knowing the value of \(x\) completely determines the value of \(y\)).

• **Sign** indicates direction of covariance:
  - **positive** means that large \(x\) indicates \(y\) is likely large;
  - **negative** means that large \(x\) indicates \(y\) is likely small.
CORRELATION AS A MEASUREMENT

• Sometimes the correlation itself is interesting.

• In order to establish the significance of the correlation, need to ask what r is, as well as how many measurements were taken to establish it.

• For given r and N, look in table (Taylor Appendix C) to find out probability of randomly measuring a particular correlation value.

• Random probabilities: is there evidence of linear correlation between the two variables?
  • <5% = significant evidence
  • <1% = highly significant evidence
CORRELATION: A NON-PHYSICS EXAMPLE

From http://hdr.undp.org/statistics/data/ take nations with human development index in top 25 and at least 10 million people

<table>
<thead>
<tr>
<th>Country</th>
<th>Physicians per 100K people</th>
<th>Smoking rate (%)</th>
<th>Life Expectancy (years)</th>
<th>GDP per capita (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>249</td>
<td>19</td>
<td>80.2</td>
<td>26,300</td>
</tr>
<tr>
<td>Canada</td>
<td>209</td>
<td>22</td>
<td>79.9</td>
<td>27,100</td>
</tr>
<tr>
<td>Belgium</td>
<td>418</td>
<td>24</td>
<td>78.8</td>
<td>29,100</td>
</tr>
<tr>
<td>USA</td>
<td>549</td>
<td>23</td>
<td>77.3</td>
<td>37,600</td>
</tr>
<tr>
<td>Japan</td>
<td>201</td>
<td>29</td>
<td>81.9</td>
<td>33,700</td>
</tr>
<tr>
<td>Netherlands</td>
<td>329</td>
<td>28</td>
<td>78.3</td>
<td>31,500</td>
</tr>
<tr>
<td>UK</td>
<td>166</td>
<td>27</td>
<td>78.3</td>
<td>30,300</td>
</tr>
<tr>
<td>France</td>
<td>329</td>
<td>27</td>
<td>79.4</td>
<td>29,400</td>
</tr>
<tr>
<td>Italy</td>
<td>606</td>
<td>26</td>
<td>80.0</td>
<td>25,500</td>
</tr>
<tr>
<td>Germany</td>
<td>362</td>
<td>35</td>
<td>78.7</td>
<td>29,100</td>
</tr>
<tr>
<td>Spain</td>
<td>320</td>
<td>32</td>
<td>79.5</td>
<td>20,400</td>
</tr>
<tr>
<td>Greece</td>
<td>440</td>
<td>38</td>
<td>78.2</td>
<td>15,600</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>330.7</strong></td>
<td><strong>28.4</strong></td>
<td><strong>78.85</strong></td>
<td><strong>24,929</strong></td>
</tr>
</tbody>
</table>
WITH WHAT DOES LIFESPAN CORRELATE? DOCTORS? SMOKING? MONEY?

- Calculate linear correlation coefficient
  \[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \]
- For life expectancy versus physicians, \( r = -0.40 \)
- For life expectancy versus smoking, \( r = -0.17 \)
- For life expectancy versus GDP per capita, \( r = -0.04 \)
The linear correlation coefficients of life expectancy versus physicians, smoking, and GDP per capita are $-0.40$, $-0.17$, and $-0.04$, respectively.

- Life expectancy is most closely correlated with number of physicians (but negatively correlated).
- How significant are the correlations?
Appendix C gives probability to find \( r > r_0 \) for \( N \) measurements of two uncorrelated variables:

<table>
<thead>
<tr>
<th>( N )</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>100</td>
<td>77</td>
<td>56</td>
<td>37</td>
<td>22</td>
<td>12</td>
<td>5.1</td>
<td>1.6</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>100</td>
<td>76</td>
<td>53</td>
<td>34</td>
<td>20</td>
<td>9.8</td>
<td>3.9</td>
<td>1.1</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>100</td>
<td>75</td>
<td>51</td>
<td>32</td>
<td>18</td>
<td>8.2</td>
<td>3.0</td>
<td>0.8</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- 12 measurements are 20% likely to have \(|r| > 0.4\)
- Using linear extrapolation, the probabilities for \( r = -0.17 \) and \( r = -0.04 \) are 61% and 90% so are likely uncorrelated
- General rule: < 5% is significant, < 1% is highly significant

**NO!**