• Correlation and covariance
• Introduction to the Poisson Distribution
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:
  \[
  \sigma_q^2 = \left( \frac{\partial q}{\partial x} \delta x \right)^2 + \left( \frac{\partial q}{\partial y} \delta y \right)^2 + 2 \frac{\partial q}{\partial x} \frac{\partial q}{\partial y} \sigma_{xy}
  \]
  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 <5% (significant) or <1% (highly significant)
### CORRELATION: A NON-PHYSICS EXAMPLE


<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?

![Graphs showing correlation between life expectancy and various factors: physicians, smoking rate, and GDP per capita.]

Calculate linear correlation coefficient

\[
 r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]

- For life expectancy versus physicians, \( r = -0.40 \)
- For life expectancy versus smoking, \( r = -0.17 \)
- For life expectancy versus GDP per capita, \( r = -0.04 \)
IS IT SIGNIFICANT?

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></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></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></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.
COUNTING EXPERIMENTS: THE POISSON DISTRIBUTION

- Start with a large amount of radioactive material with a very long half-life (compared to the experiment)
- Measure the decays with a detector at a fixed distance
- Measure the number of decays in five 100-second intervals
- Should the five intervals all have the same number of decays?
- Should there be a consistent trend?
COUNTING EXPERIMENTS

- Should there be a consistent trend?
- NO! The long half-life assures that over the time of the experiment, the decay rate isn’t changing significantly.

- Should the five intervals all have the same number of decays?
- NO! The decay is a random process.
A POISSON PROCESS

- Assume a decay rate of 0.01 decays/sec
- On average, will get 1 decay/100 seconds.
  - Sometimes you will get zero
  - Sometimes you will get two
  - Rarely you will get three or more
- What is the probability of getting exactly \( n \) decays in 100 seconds?
- Answer: Poisson distribution. Applies to any discrete process where events occur randomly at a constant rate.
THE POISSON DISTRIBUTION

- Characteristics the process has to have:
  - Asymmetry! Can’t have fewer than zero counts, so a symmetric function like a gaussian can’t describe this process. Also must vanish for n<0.
  - Discreteness: Gaussian describes a variable that can take on a continuum of values. There’s no such thing as 0.48 events, so Poisson must give probabilities (not probability density) of discrete outcomes.
  - Summing: we could view a process with a mean of $2\mu$ as a collection of two processes with a mean of $\mu$ (events from two halves of the same source, for example). So the following sum rule must be true:

$$P_{2\mu}(3) = P_\mu(0)P_\mu(3) + P_\mu(1)P_\mu(2) + P_\mu(2)P_\mu(1) + P_\mu(3)P_\mu(0)$$

$$= 2P_\mu(0)P_\mu(3) + 2P_\mu(1)P_\mu(2)$$

- Notation: If mean is $\mu$, let probability of finding 3 events be $P_\mu(3)$. If mean is $2\mu$, then probability of finding 3 events is $P_{2\mu}(3)$. 
THE POISSON DISTRIBUTION

\[ P_\mu(n) = e^{-\mu} \frac{\mu^n}{n!} \]

- This is obviously asymmetric, doesn’t allow negative events, and is obviously discrete (factorial is only defined for nonnegative integers). Note that the mean \( \mu \) doesn’t have to be an integer.
- The sum rule works too: try it out at home.
- Note that \( \mu^n \) factor keeps probability down for too few events; \( n! \) factor keeps it down for too many events.
# GAUSS VS. POISSON

## GAUSSIAN

\[ P_{\mu,\sigma}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- Symmetric
- Continuous \((x\) is real)
- Mean = \(\mu\); most probable point = \(\mu\)
- Standard deviation = \(\sigma\)
- Distribution describes results of measurements with accuracy \(\sigma\)

## POISSON

\[ P_{\mu}(n) = e^{-\mu} \frac{\mu^n}{n!} \]

- Asymmetric \((P_{\mu}(n) \geq 0)\)
- Discrete \((n\) is integer)
- Mean = \(\mu\); most probable point \(\leq \mu\)
- Standard deviation = \(\sqrt{\mu}\)
- Distribution describes results from “counting experiment”