These lectures

• Lecture 1 : basic descriptive statistics

• Lecture 2 : searching for correlations

• Lecture 3 : hypothesis testing and model-fitting

• Lecture 4 : Bayesian inference
Lecture 2: searching for correlations

- Correlation coefficient and its error
- How to quantify the significance of a correlation
- Bootstrap error estimates
- Non-parametric correlation tests
- Common pitfalls when searching for correlations
- Comparing two distributions
What is a correlation?

- Two variables are **correlated** if they share a statistical dependence / relationship.

- For example, measurements of temperature at noon and 1 pm every day are correlated, because they both lie consistently above the mean daily temperature.

- Correlations between variables are important because they indicate some **underlying physical relationship** between those variables.
What is a correlation?

Uncorrelated variables!
What is a correlation?

Trick for generating correlated variables:

\[ x, z \text{ unit gaussian variables} \]

\[ y = x.r + z \left(1 - r^2\right)^{1/2} \]

\((x, y)\) have corr. coefficient \(r\)
What is a correlation?

- Example in astronomy: black-hole / bulge relation

![Graph showing correlation between black hole mass and bulge velocity dispersion.](image)
Correlation coefficient

- Describes the strength of the correlation between \((x, y)\)

- Means: \((\mu_x, \mu_y)\)

- Standard deviations: \((\sigma_x, \sigma_y)\)

- Definition of correlation coefficient:

\[
\rho = \frac{\langle (x - \mu_x)(y - \mu_y) \rangle}{\sigma_x \sigma_y} = \frac{\langle xy \rangle - \mu_x \mu_y}{\sigma_x \sigma_y}
\]

\[
\langle xy \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy P(x, y) \, dx \, dy
\]
Correlation coefficient

\[ \rho = \frac{\langle (x - \mu_x)(y - \mu_y) \rangle}{\sigma_x \sigma_y} = \frac{\langle xy \rangle - \mu_x \mu_y}{\sigma_x \sigma_y} \]

- **No correlation** \([P(x,y)\) separable into \(f(x) g(y)\)] :
  \[ \langle xy \rangle = \langle x \rangle \langle y \rangle = \mu_x \mu_y \quad \rho = 0 \]

- **Complete correlation** :
  \[ y = C' x \quad \rho = +1 \]

- **Complete anti-correlation** :
  \[ y = - C' x \quad \rho = -1 \]

- **Possible range is** \(-1 \leq \rho \leq +1\)
The dark energy puzzle

Lies, damn lies and statistics

Why was this poor statistics?  Correlation is not the same as causation.  Other dietary or lifestyle habits could be a third variable!  [ ... more examples later ... ]
We can estimate the Pearson product-moment correlation coefficient as:

\[ r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}} \]

\[ r = \frac{\sum_{i=1}^{N} x_i y_i - N \bar{x} \bar{y}}{(N - 1) \sqrt{\text{Var}(x) \text{Var}(y)}} \]

c.f. \[ \rho = \frac{\langle xy \rangle - \mu_x \mu_y}{\sigma_x \sigma_y} \]

The possible range of values is

\[ -1 \leq r \leq +1 \]
Estimating the correlation coefficient

- If the correlation is statistically significant:
  - $0 < |r| < 0.3$ is a “weak correlation”
  - $0.3 < |r| < 0.7$ is a “moderate correlation”
  - $0.7 < |r| < 1.0$ is a “strong correlation”
Estimating the correlation coefficient

- **Assumption**: \((x,y)\) are drawn from a bivariate Gaussian distribution about an underlying linear relation:

\[
P(x, y) = \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} \right] \right\}
\]

\[
2\pi\sigma_x\sigma_y \sqrt{1 - \rho^2}
\]

- If this model is true, then the uncertainty in the measured value of \(r\) is

\[
\sigma(r) = \sqrt{\frac{1 - r^2}{N - 2}}
\]
Estimating the correlation coefficient

Example 1: who discovered the distance-redshift relation, Lemaître (1927) or Hubble (1929)?

arXiv:1106.3928 etc.

“A Hubble Eclipse: Lemaître and Censorship”

David L. Block, School of Computational & Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa.

Abstract. One of the greatest discoveries of modern times is that of the expanding Universe, almost invariably attributed to Hubble (1929). What is not widely known is that the original treatise by Lemaître (1927) contained a rich fusion of both theory and of observation. The French paper was meticulously censored when printed in English — all discussions of radial velocities and distances (and the very first empirical determination of “H”) were omitted. Fascinating insights are gleaned from a letter recently found in the Lemaître archives. An appeal is made for a Lemaître Telescope, to honour the discoverer of the expanding universe.
• Example 1: who discovered the distance-redshift relation, Lemaitre (1927) or Hubble (1929)?
Part (a) : For each dataset, find the Pearson product-moment correlation coefficient and its error

- Lemaitre : $r = 0.38 \pm 0.15$
- Hubble : $r = 0.79 \pm 0.13$
Is a correlation significant?

• What is the probability of obtaining the measured value of \( r \) if the true correlation is zero? (also depends on \( N \))

• In order to determine whether the correlation is significant, calculate

\[
t = r \sqrt{\frac{N - 2}{1 - r^2}}
\]

• This obeys the Student’s t probability distribution with number of degrees of freedom \( \nu = N - 2 \)

• Consult tables (2-tailed test) with these two numbers
Is a correlation significant?

Student's t distribution

\[ t = r \sqrt{\frac{N - 2}{1 - r^2}} \]

- Distribution of t when \( \rho = 0 \)
- Number of data points is crucial!
Is a correlation significant?

- Standard tables list the critical values that \( t \) must exceed, as a function of \( \nu \), for the hypothesis that the two variables are unrelated to be rejected at a particular level of statistical significance (e.g. 95%, 99%)

"Two-tailed test": correlation or anti-correlation
Is a correlation significant?

- Part (b): Determine the statistical significance of the correlation

- Lemaitre: $t=2.60$, $\nu=40$, $\text{prob}=1.3\times10^{-2}$ (2.5 sigma)

- Hubble: $t=6.03$, $\nu=22$, $\text{prob}=4.5\times10^{-6}$ (4.6 sigma)
• The regression line is the linear fit that minimizes the sum of the squares of the y-residuals

• With intercept \([y = a + b x]\) :

\[
b = \frac{\sum_{i=1}^{N} x_i y_i - N \overline{x} \overline{y}}{(N - 1) \sigma_x^2} = r \frac{\sigma_y}{\sigma_x}
\]

\[a = \mu_y - b \mu_x\]

• Without intercept \([y = b x]\) :

\[
b = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}
\]
• Part (c) : Determine linear least-squares regression lines of the form $V = HD$ and $v = H'D + C$

• Lemaitre : $H=414.9$, $H'=221.7$, $C=316.5$
  Hubble : $H=423.9$, $H'=453.9$, $C=-40.4$
• Aside: Hubble and Lemaitre both found values of $H_0 \sim 420$ km/s/Mpc with independent techniques! How could they both be wrong? [Example of statistical bias]

• Today would probably indicate confirmation bias, but Hubble didn’t even cite Lemaitre’s result!

• Lemaitre: assumed galaxy apparent magnitude was standard candle - scuppered by Malmquist bias

• Hubble: used “brightest stars” as standard candles, but could not distinguish brightest star from HII region (systematic error bias due to aperture effect)
Bootstrap errors and probabilities

- **Bootstrap statistics** allow us to determine parameter errors and probability distributions using just the data.

- If we have $N$ data points, repeatedly draw at random samples of $N$ points (with replacement).

- Re-compute the parameter of interest for each bootstrap sample.

- The **distribution** of the re-computed parameters estimates the uncertainty in the measurement from the original sample.
Bootstrap errors and probabilities

• Applied to our example: create 1000 bootstrap samples and measure the correlation coefficients.

![Histogram showing correlation coefficients with 68% confidence intervals.](image)
Bootstrap errors and probabilities

- Applied to our example: create 1000 bootstrap samples and do a linear regression fit.
Non-parametric correlation coefficient

- If we do not want to assume \((x,y)\) are drawn from a bivariate Gaussian we can use a non-parametric correlation test.

- Let \((X_i, Y_i)\) be the rank of \((x_i, y_i)\) in the overall order such that \(1 \leq (X_i, Y_i) \leq N\).

- Find Spearman rank correlation coefficient

\[
r_s = 1 - 6 \frac{\sum_{i=1}^{N} (X_i - Y_i)^2}{N^3 - N}
\]

- Compare to standard tables with \(\nu = N - 2\).
Non-parametric correlation coefficient

- Standard tables list the critical values that $r_s$ must exceed, as a function of $n_u$, for the hypothesis that the two variables are unrelated to be rejected at a particular level of statistical significance (e.g. 95%, 99%)

<table>
<thead>
<tr>
<th>df</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>.900</td>
<td>1.000</td>
<td>1.000</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>.829</td>
<td>.886</td>
<td>.943</td>
<td>1.000</td>
</tr>
<tr>
<td>7</td>
<td>.714</td>
<td>.786</td>
<td>.893</td>
<td>.929</td>
</tr>
<tr>
<td>8</td>
<td>.643</td>
<td>.738</td>
<td>.833</td>
<td>.881</td>
</tr>
<tr>
<td>9</td>
<td>.600</td>
<td>.683</td>
<td>.783</td>
<td>.833</td>
</tr>
<tr>
<td>10</td>
<td>.564</td>
<td>.643</td>
<td>.746</td>
<td>.794</td>
</tr>
<tr>
<td>12</td>
<td>.506</td>
<td>.591</td>
<td>.712</td>
<td>.777</td>
</tr>
<tr>
<td>14</td>
<td>.456</td>
<td>.544</td>
<td>.645</td>
<td>.715</td>
</tr>
<tr>
<td>16</td>
<td>.425</td>
<td>.506</td>
<td>.601</td>
<td>.665</td>
</tr>
<tr>
<td>18</td>
<td>.399</td>
<td>.475</td>
<td>.564</td>
<td>.625</td>
</tr>
<tr>
<td>20</td>
<td>.377</td>
<td>.450</td>
<td>.534</td>
<td>.591</td>
</tr>
<tr>
<td>22</td>
<td>.359</td>
<td>.428</td>
<td>.508</td>
<td>.562</td>
</tr>
<tr>
<td>24</td>
<td>.343</td>
<td>.409</td>
<td>.485</td>
<td>.537</td>
</tr>
<tr>
<td>26</td>
<td>.329</td>
<td>.392</td>
<td>.465</td>
<td>.515</td>
</tr>
<tr>
<td>28</td>
<td>.317</td>
<td>.377</td>
<td>.448</td>
<td>.496</td>
</tr>
<tr>
<td>30</td>
<td>.306</td>
<td>.364</td>
<td>.432</td>
<td>.478</td>
</tr>
</tbody>
</table>
Part (e): Determine the Spearman rank cross-correlation coefficient and its statistical significance

- Lemaitre: $r_s = 0.42$, $\nu = 40$, $prob = 6.0e-3$ (2.8 sigma)
- Hubble: $r_s = 0.80$, $\nu = 22$, $prob = 3.4e-6$ (4.7 sigma)

[Results very similar to Pearson product-moment correlation coefficient, but fewer assumptions]
Issues with correlations

- **Selection effects** leading to spurious correlations, for example **Malmquist bias**
Issues with correlations

- Is the correlation driven by a small number of outliers, so is not robust?

Mean, variance, correlation coefficient and regression line are all identical

“Rule of thumb”
Issues with correlations

- Correlation does not necessarily imply causation
- Sleeping in your shoes causes you to wake up with a headache! [third variable - you were probably having a few drinks the night before...]
- Ice cream sales cause drowning! [third variable - hotter weather means more people are at the beach...]
- Having grey hair causes cancer! [third variable - age...]
- Obesity causes global warming! [third variable - everyone is getting richer...]
Lies, damn lies and statistics

“‘We don’t accept the idea that there are harmful agents in tobacco’”
[Phillip Morris, 1964]

Why was this poor statistics? Cigarette companies were attempting to invoke a “third variable”. But correlation does sometimes imply causation, if it can be demonstrated by independent lines of evidence.
Comparing two distributions

- Are two samples drawn from the same distribution?
- Example 2: samples of flux densities measured at random positions [A] and galaxy positions [B]
Comparing two distributions

- Part (a): Are their means consistent?

- Calculate $t$ statistic and no. of degrees of freedom:

$$t = \frac{|\mu_x - \mu_y|}{\sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}}$$

$$\nu = \frac{\left(\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}\right)^2}{\frac{\sigma_x^4}{N_x(N_x-1)} + \frac{\sigma_y^4}{N_y(N_y-1)}}$$

- Compare to Student's $t$ distribution

- $t = 0.31$, $\nu = 661.5$, prob of consistency = 0.76

- Small print: assumes $(x,y)$ are normally-distributed populations
Part (b): Are the full distributions consistent?

The Kolmogorov-Smirnov test considers the maximum value of the absolute difference between the cumulative probability distributions.

Max diff = $D = 0.067$

$$\nu = \frac{N_x N_y}{N_x + N_y} = 165.6$$

$$Q = \left( \sqrt{\nu} + 0.12 + \frac{0.11}{\sqrt{\nu}} \right) D = 0.876$$

Prob(Q) = 0.427