Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2018

Probability & Statistics:
Intro, summary statistics, probability

Statistics is the science of learning from experience, especially experience that arrives a little bit at a time. The earliest information science was statistics, originating in about 1650. This century has seen statistical techniques become the analytic methods of choice in biomedical science, psychology, education, economics, communications theory, sociology, genetic studies, epidemiology, and other areas. Recently, traditional sciences like geology, physics, and astronomy have begun to make increasing use of statistical methods as they focus on areas that demand informational efficiency, such as the study of rare and exotic particles or extremely distant galaxies.

Most people are not natural-born statisticians. Left to our own devices we are not very good at picking out patterns from a sea of noisy data. To put it another way, we are not too good at picking out non-existent patterns that happen to suit our purposes. Statistical theory attack the problem from both ends. It provides optimal methods for finding a real signal in a noisy background, and also provides strict checks against the overinterpretation of random patterns.

- Efron & Tibshirani, Introduction to the Bootstrap, 1998
**Some history…**

- 1600’s: Early notions of data summary/averaging
- 1700’s: Bayesian prob/statistics (Bayes, Laplace)
- 1920’s: Frequentist statistics for science (e.g., Fisher)
- 1940’s: Statistical signal analysis and communication, estimation/decision theory (e.g., Shannon, Wiener, etc)
- 1950’s: Return of Bayesian statistics (e.g., Jeffreys, Wald, Savage, Jaynes…)
- 1970’s: Computation, optimization, simulation (e.g., Tukey)
- 1990’s: Machine learning (large-scale computing + statistical inference + lots of data)
- Since 1950’s!: statistical neural/cognitive models

---

**Scientific process**

- Observe / measure data
- Generate predictions, design experiment
- Summarize/fit model(s), compare with predictions
- Create/modify hypothesis/model
Descriptive statistics: Central tendency

- We often summarize data with the average. Why?
- Average minimizes the squared error (as in regression!):
  \[ \mu(x) = \arg \min_c \frac{1}{N} \sum_{n=1}^{N} (x_n - c)^2 = \frac{1}{N} \sum_{n=1}^{N} x_n \]
- Generalize: minimize \( L_p \) norm: \[ \arg \min_c \left[ \frac{1}{N} \sum_{n=1}^{N} |x_n - c|^p \right]^{1/p} \]
  - minimize \( L_2 \) norm: median, \( m(x) \)
  - minimize \( L_0 \) norm: mode
  - minimize \( L_\infty \) norm: midpoint of range
- Issues: outliers, asymmetry, bimodality
- How do we choose?
Descriptive statistics: Dispersion

- Sample standard deviation
  \[
  \sigma(\bar{x}) = \min_c \left[ \frac{1}{N} \sum_{n=1}^{N} (x_n - c)^2 \right]^{1/2}
  = \left[ \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu(\bar{x}))^2 \right]^{1/2}
  \]

- Mean absolute deviation (MAD) about the median
  \[
  d(\bar{x}) = \frac{1}{N} \sum_{n=1}^{N} |x_n - m(\bar{x})|
  \]

- Quantiles
Descriptive statistics: Dispersion

Summary statistics (e.g., sample mean/var) can be interpreted as *estimates of model parameters*. To formalize this, we need tools from *probability*…

\[
\text{data} \rightarrow \text{histogram} \rightarrow \text{probability distribution}
\]

\[
\{x_n\} \quad \{c_k, h_k\} \quad p(x)
\]
You pick a family at random and discover that one of the children is a girl. What are the chances that the other child is a girl?
Statistical Middleville

In Middleville, every family has two children, brought by the stork.

The stork delivers boys and girls randomly, with family probability \((BB, BG, GB, GG) = (0.2, 0.3, 0.2, 0.3)\).

In a survey of 100 of the Middleville families, 32 have two girls, 23 have two boys, and the remainder one of each.

You pick a family at random and discover that one of the children is a girl.

What are the chances that the other child is a girl?

Probability basics (outline)

- distributions: discrete and continuous
- expected value, moments
- cumulative distributions. Quantiles, Q-Q plots, drawing samples.
- transformations: affine, monotonic nonlinear
let $X$, $Y$, $Z$ be random variables
they can take on values (like ‘heads’ or ‘tails’; or integers 1-6; or real-valued numbers)

let $x$, $y$, $z$ stand generically for values they can take, and denote events such as $X = x$

write the probability that $X$ takes on value $x$ as $P(X = x)$, or $P_X(x)$, or sometimes just $P(x)$

$P(x)$ is a function over values $x$, which we call the probability “distribution” function (pdf) (for continuous variables, “density”)

---

**Probability distributions**

Discrete random variable

$0 < P(x_i) < 1, \ \forall i$

$\sum_i P(x_i) = 1$

Continuous random variable

$0 < p(x)$

$\int_{-\infty}^{\infty} p(x)\,dx = 1$
Example distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Probability Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>A not-quite-fair coin</td>
<td></td>
</tr>
<tr>
<td>Roll of a fair die</td>
<td></td>
</tr>
<tr>
<td>Sum of two rolled fair dice</td>
<td></td>
</tr>
<tr>
<td>Clicks of a Geiger counter, in a fixed time interval</td>
<td></td>
</tr>
<tr>
<td>Horizontal velocity of gas molecules exiting a fan</td>
<td></td>
</tr>
</tbody>
</table>

Expected value - discrete

\[ E(X) = \sum_{i=1}^{N} x_i p(x_i) \]  
[the mean, \( \mu \)]

More generally:

\[ E(f(X)) = \sum_{i=1}^{N} f(x_i) p(x_i) \]
Expected value - continuous

\[ E(x) = \int x \, p(x) \, dx \]  \hspace{1cm} \text{[mean, } \mu \text{]}

\[ E(x^2) = \int x^2 \, p(x) \, dx \]  \hspace{1cm} \text{["second moment", } m_2 \text{]}

\[ E((x-\mu)^2) = \int (x-\mu)^2 \, p(x) \, dx \]  \hspace{1cm} \text{[variance, } \sigma^2 \text{]}

\[ = \int x^2 \, p(x) \, dx - \mu^2 \]  \hspace{1cm} \text{[equal to } m_2 \text{ minus } \mu^2 \text{]}

\[ E(f(x)) = \int f(x) \, p(x) \, dx \]  \hspace{1cm} \text{["expected value of } f \text{"]}

Note: this is an inner product, and thus linear:

\[ E(af(x) + bg(x)) = aE(f(x)) + bE(g(x)) \]
Multi-variate probability

- joint distributions
- marginals (integrating)
- conditionals (slicing)
- Bayes’ rule (inverse probability)
- statistical independence (separability)
- linear transformations
Joint and conditional probability - discrete

- $P(\text{Ace})$
- $P(\text{Heart})$
- $P(\text{Ace} \& \text{Heart})$
- $P(\text{Ace} | \text{Heart})$
- $P(\text{not Jack of Diamonds})$
- $P(\text{Ace} | \text{not Jack of Diamonds})$

"Independence"
Joint distribution (continuous)

\[ p(x, y) \]

Marginal distribution

\[ p(x) = \int p(x, y) \, dy \]
Conditional probability

\[ p(A \mid B) = \frac{p(A \& B)}{p(B)} \]

Neither A nor B

Conditional distribution

\[ p(x, y) \quad p(x \mid y = 68) \]
Conditional distribution

\[ p(x|y = 68) = \frac{p(x, y = 68)}{\int p(x, y = 68) \, dx} \]

More generally:
\[ p(x|y) = \frac{p(x, y)}{p(y)} \]

Bayes’ Rule

\[ p(A \, | \, B) = \frac{\text{probability of } A \text{ given that } B \text{ is asserted to be true}}{\frac{p(A \, & \, B)}{p(B)}} \]
\[ p(A \, & \, B) = p(B) \cdot p(A \, | \, B) \]
\[ = p(A) \cdot p(B \, | \, A) \]
\[ \Rightarrow p(A \, | \, B) = \frac{\frac{p(B \, | \, A) \cdot p(A)}{p(B)}}{p(B)} \]

\[ p(A \, | \, B) = \frac{p(B \, | \, A) \cdot p(A)}{p(B)} \]
Bayes’ Rule

\[ p(x|y) = \frac{p(y|x)p(x)}{p(y)} \]

(a direct consequence of the definition of conditional probability)

In general, the marginals for different Y values differ. When are they the same? In particular, when are all conditionals equal to the marginal?
Statistical independence

Random variables $X$ and $Y$ are statistically independent if (and only if):

$$p(x, y) = p(x)p(y) \quad \forall \ x, y$$

[Note: for discrete distributions, this is an outer product!]

Independence implies that all conditionals are equal to the corresponding marginal:

$$p(x \mid y) = p(x, y) / p(y) = p(x) \quad \forall \ x, y$$

Sums of RVs

Let $Z = X + Y$. Since expectation is linear:

$$E(X + Y) = E(X) + E(Y)$$

In addition, if $X$ and $Y$ are independent, then

$$E(XY) = E(X)E(Y)$$

$$\sigma_z^2 = E \left( (X + Y) - \left( \mu_x + \mu_y \right) \right)^2 = \sigma_x^2 + \sigma_y^2$$

and $p_z(z)$ is a convolution of $p_x(x)$ and $p_y(y)$
Mean and variance

- Mean and variance summarize the centroid/width
- Translation and rescaling of random variables
- Mean/variance of weighted sum of random variables
- The sample average
  - ... converges to true mean (except for bizarre distributions)
  - ... with variance $\sigma^2/N$
  - ... most common choice for an estimate ...

Central limit for a uniform distribution...

10k samples, uniform density (sigma=1)

\[
\frac{u_1 + u_2 + u_3 + u_4}{\sqrt{4}}
\]

\[
\frac{1}{\sqrt{10}} \sum_{i=1}^{10} u_i
\]
Central limit for a binary distribution...