Section 4:
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 all too good at picking out non-existent patterns that happen to suit our purposes. Statistical theory attacks 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, 1998]

Historical context

- 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)
- Also, since 1950’s: statistical neural/cognitive models!
Statistics as summary

0.1, 4.5, -2.3, 0.8, -1.1, 3.2, ...

“The purpose of statistics is to replace a quantity of data by relatively few quantities which shall contain as much as possible, ideally the whole, of the relevant information contained in the original data”

- R.A. Fisher, 1934

Descriptive statistics

Data

“Dispersion”

“Central tendency”

Descriptive statistics: Central tendency

- We often summarize data with averages. Why?
- Average minimizes the squared error (as in regression!):

\[
\mu_x = \arg \min_c \frac{1}{N} \sum_{n=1}^{N} (x_n - c)^2 = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{1}{N} \mathbf{x}^T \mathbf{x}
\]

- Generalize: minimize \( L_p \) norm: 
  \[
  \arg \min_c \left[ \frac{1}{N} \sum_{n=1}^{N} |x_n - c|^p \right]^{1/p}
  \]
  - \( p = 1 \) : median, \( m_x \)
  - \( p \to 0 \) : mode (location of maximum)
  - \( p \to \infty \) : midpoint of range
- Issues: outliers, asymmetry, bimodality
Descriptive statistics: Dispersion

- Sample variance (squared standard deviation):
  \[ \sigma^2_x = \min_c \frac{1}{N} \sum_{n=1}^{N} (x_n - c)^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_x)^2 \]
  \[ = \frac{1}{N} \sum_{n=1}^{N} x_n^2 - \mu_x^2 = \frac{1}{N} ||\vec{x}||^2 - \mu_x^2 \]

- Mean absolute deviation (MAD) about the median:
  \[ d_x = \frac{1}{N} \sum_{n=1}^{N} |x_n - m_x| \]

- Quantiles (eg: “90% of data lie in range [1.5 8.2]”)

Descriptive statistics: Multi-D

Data points: \( \{\vec{d}_n\} \quad n \in [1 \ldots N] \)

Sample mean:
\[ \vec{\mu}_d = \frac{1}{N} \sum_{n=1}^{N} \vec{d}_n = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \]

Sample covariance:
\[ C_d = \frac{1}{N} \sum_{n=1}^{N} (\vec{d}_n - \vec{\mu}_d)(\vec{d}_n - \vec{\mu}_d)^T = \frac{1}{N} \sum_{n=1}^{N} \vec{d}_n\vec{d}_n^T - \vec{\mu}_d\vec{\mu}_d^T \]
\[ = \frac{1}{N} \begin{bmatrix} ||\vec{x}||^2 & \vec{x}^T \vec{y} \\ \vec{y}^T \vec{x} & ||\vec{y}||^2 \end{bmatrix} - \begin{bmatrix} \mu_x^2 & \mu_x \mu_y \\ \mu_y \mu_x & \mu_y^2 \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \]

Affine transformations

If \( \vec{b}_n = M \left( \vec{d}_n - \vec{a} \right) \)

then \( \vec{b}_b = M \left( \vec{\mu}_d - \vec{a} \right) \)
\[ C_b = MC_dM^T \]

Special case: “center” and “normalize” the data:

\[ \vec{a} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \quad M = \begin{bmatrix} \frac{1}{\sigma_x} & 0 \\ 0 & \frac{1}{\sigma_y} \end{bmatrix} \]

then \( \vec{b}_b = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad C_b = \begin{bmatrix} \frac{\sigma_{xy}}{\sigma_x \sigma_y} & \frac{\sigma_x}{\sigma_x \sigma_y} \\ \frac{\sigma_x}{\sigma_x \sigma_y} & \frac{\sigma_y}{\sigma_x \sigma_y} \end{bmatrix} \quad \text{[on board]} \)

(Pearson correlation coefficient)
Correlation $r$ captures dependency

... but not slope!

Regression (revisited)

$\hat{y} = \beta \hat{x} + \hat{e}$

Optimal regression line slope:

$$\beta = \frac{\hat{x}^T \hat{y}}{\hat{x}^T \hat{x}} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Error variance:

$$\sigma_e^2 = \sigma_y^2 - 2\beta \sigma_{xy} + \beta^2 \sigma_x^2$$

Expressed as a proportion of $\sigma_y^2$:

$$\frac{\sigma_e^2}{\sigma_y^2} = 1 - \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = 1 - r^2$$

"r-squared" (proportion of variance explained)
**Probability**: an abstract mathematical framework for describing random quantities, or stochastic models of the world

**Statistics**: use of probability to summarize, analyze, interpret data. **Fundamental to all experimental science.**

---

**Probabilistic Middleville**

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

The stork delivers boys and girls randomly, with family probabilities \{\text{BB,BG,GB,GG}\} = \{0.2,0.3,0.2,0.3\}

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.

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?

Univariate Probability (outline)

- distributions: discrete and continuous
- expected value, moments
- transformations: affine, monotonic nonlinear
- cumulative distributions. Quantiles, drawing samples

Frequentist view of probability: limit of infinite data

\[
\begin{align*}
\text{data} \quad & \rightarrow \quad \text{histogram} \quad & \rightarrow \quad \text{probability distribution} \\
\{x_n\} \quad & \rightarrow \quad \{c_k, h_k\} \\
p(x) \quad &
\end{align*}
\]
Probability distributions

Discrete random variable

- $0 \leq P(x_i) \leq 1, \ \forall i$
- $\sum_i P(x_i) = 1$

Continuous random variable

- $0 \leq p(x)\ dx = 1$

Example distributions

- A not-quite-fair coin (Bernoulli)
- Roll of a fair die (uniform)
- Sum of rolls of two fair dice
- Clicks of a Geiger counter, in a fixed time interval (Poisson)
- ...and, time between clicks (exponential)
- Horizontal velocity of gas molecules exiting a fan (Gaussian)

Example distributions [Figure: Sean Owen, Cloudera Engineering]
Frequentist view of probability: limit of infinite data

\[ \mu = \frac{1}{N} \sum_n x_n \]
\[ \mu \approx \frac{1}{K} \sum_k c_k h_k = \vec{c}^T \vec{h} \]
\[ \mu = \int x \, p(x) \, dx \]

Data → Histogram → Probability distribution

\{x_n\} \quad \{c_k, h_k\} \quad p(x)

Expected value (discrete)

\[ E(X) = \sum_{n=1}^{N} x_n \, p(x_n) \]

(the mean, \( \mu \))

More generally: \( E(f(X)) = \sum_{x} f(x_n) \, p(x_n) \)

Expected value (continuous)

\[ E(x) = \int x \, p(x) \, dx \quad \text{["mean", } \mu \text{]} \]
\[ E(x^2) = \int x^2 \, p(x) \, dx \quad \text{["second moment", } m_2 \text{]} \]
\[ E((x-\mu)^2) = \int (x-\mu)^2 \, p(x) \, dx \quad \text{["variance", } \sigma^2 \text{]} \]
\[ = \int x^2 \, p(x) \, dx - \mu^2 \quad \text{[} m_2 \text{ minus } \mu^2 \text{]} \]
\[ E(f(x)) = \int f(x) \, p(x) \, dx \quad \text{["expected value of } f \text{"\]}

Note: expectation is an inner product, and thus linear, so:
\[ E(af(x) + bg(x)) = aE(f(x)) + bE(g(x)) \]
Transformations of random variables

\[ Y = aX + b \]  

“affine” (linear plus constant)

Analogous to sample mean/covariance:

\[ \mu_Y = E(Y) = aE(X) + b = a\mu_X + b \]

\[ \sigma_Y^2 = E\left((Y - \mu_Y)^2\right) = E\left((aX - a\mu_X)^2\right) = a^2\sigma_X^2 \]

Full distribution: \[ p_Y(y) = \frac{1}{a} p_X\left(\frac{y - b}{a}\right) \]

\[ Y = g(X) \]  

“monotonic” (derivative > 0)

\[ p_Y(y) = \frac{p_X\left(g^{-1}(y)\right)}{g'(g^{-1}(y))} \]

Cumulative distributions

\[ c(x) = \int_{-\infty}^{x} p(z) \, dz \]
Confidence intervals

PDF

CDF

mean

P(mean | data, variance)

P(mean < x | data, variance)

90%

10%

mean

Drawing samples - discrete

Drawing samples - continuous

3) Result is uniformly distributed!

[on board]

2) Transform using the cumulative distribution function

1) Sample from \( p(x) \)

[on board]
Drawing samples - continuous

1) Draw uniform sample
2) Transform using the inverse cumulative distribution function
3) This gives a sample from \( p(x) \)!

Multi-variate probability (outline)

- Joint distributions
- Marginals (integrating)
- Conditionals (slicing)
- Bayes’ rule (inverse probability)
- Statistical independence (separability)
- Mean/Covariance
- Linear transformations

Joint and conditional probability - discrete
Conditional probability

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

Neither A nor B

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

\[ p(x, y) \]

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

Conditional distribution

\[ p(x, y) \]

\[ p(x|y = 90) \]

More generally:

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

\[ p(x|y) = p(x, y) / p(y) \]

slice joint distribution

normalize (by marginal)
Bayes’ Rule

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

(a direct consequence of the definition of conditional probability)

### Conditional vs. marginal

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:

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

Special case: Sum of two RVs

Let $Z = X + Y$. From rules for affine transforms:

$$\mu_Z = E(Z) = E(X) + E(Y)$$

$$\sigma_Z^2 = \sigma_X^2 + 2\sigma_{XY} + \sigma_Y^2$$

Special case: if $X$ and $Y$ are independent, then:

$$E(XY) = E(X)E(Y) \quad \text{and thus} \quad \sigma_{XY} = 0$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

$p_Z(z)$ is the convolution of $p_X(x)$ and $p_Y(y)$

Gaussian (a.k.a. “Normal”) densities

Parameterized by mean and stdev:

$$p(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Multi-dimensional generalization:

$$p(\bar{x}) = \frac{1}{\sqrt{(2\pi)^N|C|}} e^{-\frac{(\bar{x} - \bar{\mu})^T C^{-1} (\bar{x} - \bar{\mu})}{2}}$$

mean: $[0.2, 0.8]$

cov: $[1.0, -0.3; -0.3, 0.4]$
Gaussian properties

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

\[ p(\bar{x}) = \frac{1}{\sqrt{(2\pi)^N|C|}} e^{-\frac{(\bar{x}-\mu)^T C^{-1}(\bar{x}-\mu)}{2}} \]

- joint density of indep Gaussian RVs is elliptical \[ \text{easy} \]
- conditionals of a Gaussian are Gaussian \[ \text{easy} \]
- marginals of a Gaussian are Gaussian \[ \text{easy} \]
- product of two Gaussian dists is Gaussian \[ \text{easy} \]
- sum of Gaussian RVs is Gaussian \[ \text{moderate} \]
- the most random (max entropy) density of given variance \[ \text{moderate} \]
- central limit theorem: sum of many RVs is Gaussian \[ \text{hard} \]

Let \( P = C^{-1} \) (the “precision” matrix)

\[
\begin{align*}
p(x_1|x_2 = a) & \propto e^{-\frac{1}{2}[P_{11}(x_1-\mu_1)^2 + 2P_{12}(x_1-\mu_1)(a-\mu_2) + \ldots]} \\
& = e^{-\frac{1}{2}[P_{11}x_1^2 + 2P_{12}(a-\mu_2) + P_{11}(x_1-\mu_1)(a-\mu_2) + \ldots]} \\
& = e^{-\frac{1}{2}(x_1-\mu_1 + \frac{P_{12}}{P_{11}}(a-\mu_2))P_{11}(x_1-\mu_1 + \frac{P_{12}}{P_{11}}(a-\mu_2)) + \ldots}
\end{align*}
\]

Gaussian, with:

\[
\begin{align*}
\mu &= \mu_1 - \frac{P_{12}}{P_{11}}(a-\mu_2) \\
\sigma^2 &= \frac{1}{P_{11}}
\end{align*}
\]

Conditional:

Marginal:

\[ p(x_1) = \int p(\bar{x}) \, d\bar{x}_2 \quad \text{[on board]} \]

Gaussian, with:

\[
\begin{align*}
\mu &= \mu_1 \\
\sigma^2 &= C_{11}
\end{align*}
\]

Generalized marginals of a Gaussian

\[ \tilde{x} \sim N(\tilde{\mu}_z, C_z) \]

\[ z = \tilde{u}^T \tilde{x} \]

\[ p(z) \text{ is Gaussian, with:} \]

\[
\begin{align*}
\mu_z &= \tilde{u}^T \tilde{\mu}_z \\
\sigma^2_z &= \tilde{u}^T C_z \tilde{u}
\end{align*}
\]
Correlation and regression

Correlation implies dependency

… but not slope

… and its absence does not imply independence!

Correlation between variables does not uniquely indicate the shape of their joint distribution

Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are clearly different, and visually distinct.
More extreme examples!

Null Hypothesis: Distribution of normalized dot product of pairs of Gaussian vectors in N dimensions:

\[ (1 - d^2)^{\frac{N-3}{2}} \]

Lack of correlation is favored in N>3 dimensions.

Distribution of angles of pairs of Gaussian vectors:

\[ \sin(\theta)^{(N-2)} \]
Nevertheless, one can find correlation if one looks for it!

Covariation/correlation does not imply causation

- Correlation does not provide a direction for causality. For that, you need additional (temporal) information.
- More generally, correlations are often a result of hidden (unmeasured, uncontrolled) variables…

Example: conditional independence:
\[ p(A,B \mid H) = p(A \mid H) \cdot p(B \mid H) \]

[On board: in Gaussian case, connections are explicit in the precision matrix]

Another example: Simpson’s paradox
Milton Friedman’s Thermostat

O = outside temperature (assumed cold)
I = inside temperature (ideally, constant)
E = energy used for heating

Statistical observations:
- O and I uncorrelated
- I and E uncorrelated
- O and E anti-correlated

Some nonsensical conclusions:
- O and E have no effect on I, so shut off heater to save money!
- I is irrelevant, and can be ignored. Increases in E cause decreases in O.

Statistical summary cannot replace scientific reasoning/experiments!

Summary: Correlation misinterpretations

- Correlation does not imply data lie near a line/plane/hyperplane (subspace), with simple noise perturbations
- Correlation implies dependency, but lack of correlation does not imply independence
- Correlation does not imply causation (temporally, or by direct influence/connection)
- Correlation is a descriptive statistic, and does not eliminate the need for scientific reasoning/experiment!