Viva La Correlación!

- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var(X)Var(Y)}}}$$

  - Note: $-1 \leq \rho(X, Y) \leq 1$
  - Correlation measures <u>linearity</u> between X and Y
  - $\rho(X, Y) = 1 \quad \Rightarrow$ Y = aX + b where $a = \sigma_y/\sigma_x$
  - $\rho(X, Y) = -1 \quad \Rightarrow$ Y = aX + b where $a = -\sigma_y/\sigma_x$
  - $\rho(X, Y) = 0 \quad \Rightarrow$ absence of <u>linear</u> relationship
    - But, X and Y can still be related in some other way!
  - If $\rho(X, Y) = 0$, we say X and Y are "uncorrelated"
    - Note: Independence implies uncorrelated, but <u>**not**</u> vice versa!

Fun with Indicator Variables

- Let I_A and I_B be indicators for events A and B

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \qquad I_B = \begin{cases} 1 & \text{if } B \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

  - $E[I_A] = P(A)$, $E[I_B] = P(B)$, $E[I_A I_B] = P(AB)$
  - $\text{Cov}(I_A, I_B) = E[I_A I_B] - E[I_A]\,E[I_B]$

    $= P(AB) - P(A)P(B)$

    $= P(A \mid B)P(B) - P(A)P(B)$

    $= P(B)[P(A \mid B) - P(A)]$
  - $\text{Cov}(I_A, I_B)$ determined by $P(A \mid B) - P(A)$
  - $P(A \mid B) > P(A) \Rightarrow \rho(I_A, I_B) > 0$
  - $P(A \mid B) = P(A) \Rightarrow \rho(I_A, I_B) = 0$ (and $\text{Cov}(I_A, I_B) = 0$)
  - $P(A \mid B) < P(A) \Rightarrow \rho(I_A, I_B) < 0$

Can't Get Enough of that Multinomial

- Multinomial distribution
  - n independent trials of experiment performed
  - Each trials results in one of m outcomes, with respective probabilities: $p_1, p_2, \ldots, p_m$ where $\sum_{i=1}^{m} p_i = 1$
  - X_i = number of trials with outcome i

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \ldots p_m^{c_m}$$

  - E.g., Rolling 6-sided die multiple times and counting how many of each value {1, 2, 3, 4, 5, 6} we get
  - Would expect that X_i are negatively correlated
  - Let's see... when $i \neq j$, what is Cov(X_i, X_j)?

Covariance and the Multinomial

- Computing Cov(X_i, X_j)
  - Indicator $I_i(k)$ = 1 if trial *k* has outcome *i*, 0 otherwise

$$E[I_i(k)] = p_i \qquad X_i = \sum_{k=1}^{n} I_i(k) \qquad X_j = \sum_{k=1}^{n} I_j(k)$$

  - $\mathrm{Cov}(X_i, X_j) = \sum_{a=1}^{n}\sum_{b=1}^{n} \mathrm{Cov}(I_i(b), I_j(a))$
  - When *a* ≠ *b*, trial *a* and *b* independent: $\mathrm{Cov}(I_i(b), I_j(a)) = 0$
  - When *a* = *b*: $\mathrm{Cov}(I_i(b), I_j(a)) = E[I_i(a)I_j(a)] - E[I_i(a)]E[I_j(a)]$
  - Since trial *a* cannot have outcome *i* and *j*: $E[I_i(a)I_j(a)] = 0$

$$\mathrm{Cov}(X_i, X_j) = \sum_{a=b=1}^{n} \mathrm{Cov}(I_i(b), I_j(a)) = \sum_{a=1}^{n} (-E[I_i(a)]E[I_j(a)])$$

$$= \sum_{a=1}^{n} (-p_i p_j) = -np_i p_j \qquad \Rightarrow X_i \text{ and } X_j \text{ negatively correlated}$$

Multinomials All Around

- Multinomial distributions:
  - Count of strings hashed into buckets in hash table
  - Number of server requests across machines in cluster
  - Distribution of words/tokens in an email
  - Etc.
- When *m* (# outcomes) is large, p_i is small
  - For equally likely outcomes: p_i = 1/m

$$\mathrm{Cov}(X_i, X_j) = -np_i p_j = -\frac{n}{m^2}$$

  - Large $m \Rightarrow X_i$ and X_j very mildly negatively correlated
  - Poisson paradigm applicable