Lecture 19

Correlations

Multinomial correlations

Sharkey's casino

Spurious correlations

Variance of a sum

Square root law

Covariance and correlation

- $Cov(X,Y) \;=\; E\{(X-\mu_X)\,(Y-\mu_Y)\}$
- $Cov(X,Y) \;=\; E\{XY\} - \mu_X\mu_Y$
- $Var(X+Y) \;=\; Var(X) \;+\; Var(Y) \;+\; 2Cov(X,Y)$
- X, Y independent $\Rightarrow\; Cov(X,Y) = 0$
- $Corr(X,Y) \;=\; Cov(X,Y)/\sigma_X\sigma_Y$
- $-1 \leq Corr(X,Y) \leq +1$

Variance of a sum:

$$Var\left(\sum_k X_k\right) = \sum_k Var(X_k) + 2\sum_{j<k} Cov(X_j, X_k)$$

$$\left[\sum_k X_k - E\left\{\sum_k X_k\right\}\right]^2 = \left[\sum_k X_k - \sum_k \mu_k\right]^2$$

$$= \left[\sum_k (X_k - \mu_k)\right]^2$$

$$= \sum_k (X_k - \mu_k)^2 + 2\sum_{j<k} (X_j - \mu_j)(X_k - \mu_k)$$

- Addition rule for variances: Independence OR lack of correlation implies $Var\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n Var(X_k)$.

Bilinearity of Covariance

$$Cov\left(\sum_j a_j X_j,\ \sum_k b_k Y_k\right) = \sum_j \sum_k a_j b_k\ Cov(X_j, Y_k)$$

- i.e. linear combinations of uncorrelated variables are uncorrelated.

PROOF: As usual, by algebra.

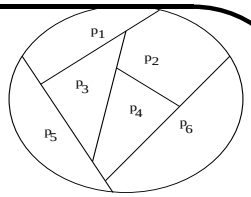
Sampling without replacement

- Sample, without replacement, from a $\{x_1, x_2, \ldots, x_N\}$.
  $E\{X_1\} = \sum_{k=1}^{N} x_k/N \equiv \mu$,
  $Var(X_1) = \sum_k (x_k - \mu)^2/N \equiv \sigma^2$.
- Set $S_n = X_1 + \ldots + X_n = n\bar{X}_n$.

$$
\begin{aligned}
Var(S_n) &= \sum_k Var(X_k) + \sum_{j \neq k} Cov(X_j, X_k) \\
&= n\sigma^2 + n(n-1)Cov(X_1, X_2) \\
\text{But,} \quad Var(S_N) &= N\sigma^2 + N(N-1)Cov(X_1, X_2) = 0\ ! \\
\Rightarrow \quad Cov(X_1, X_2) &= -\sigma^2/(N-1) \\
\Rightarrow \quad Var(S_n) &= n\sigma^2 \left[1 - \frac{n-1}{N-1}\right], \quad Var(\bar{X}_n) = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]
\end{aligned}
$$

NOTE, $Corr(X_1, X_2) = -1/(N-1)$, <u>independent</u> of the x_k!

Multinomial correlations



- $m \geq 2$ categories, $n \geq 1$ items chosen at random, with replacement, $p_k = P\{\text{Item of type } k \text{ chosen}\}$, $N_k =$ Number of type k chosen.

$$P\{N_1 = n_1, \ldots, N_m = n_m\} \;=\; \frac{n!}{n_1! n_2! \cdots n_m!}\, p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

- $Corr(N_i, N_j)$ should be negative, and close to -1 if m is small.
- <u>Hard</u> to compute from joint probabilities.
- Recall: $N_i \sim \text{Bin}(n, p_i)$, $N_i + N_j \sim \text{Bin}(n, p_i + p_j)$

$$\begin{aligned} Var(N_i + N_j) &= Var(N_i) + Var(N_j) + 2Cov(N_i, N_j) \\ \Rightarrow \quad Cov(N_i, N_j) &= \left(Var(N_i + N_j) - Var(N_i) - Var(N_j)\right)/2 \end{aligned}$$

$$\boxed{Cov(N_i, N_j) \;=\; [Var(N_i + N_j) - Var(N_i) - Var(N_j)]\,/2}$$

$$\begin{aligned} Cov(N_i, N_j) &= \Big[n(p_i + p_j)(1 - p_i - p_j) \\ &\qquad - \; np_i(1 - p_i) \; - \; np_j(1 - p_j)\Big]/2 \\ &= -np_ip_j \end{aligned}$$

Note

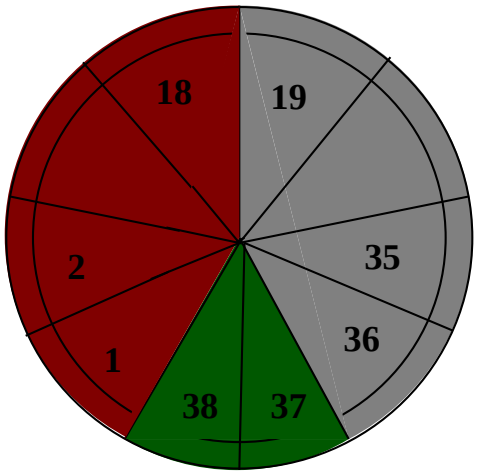
$$Var(N_i) \;=\; np_i(1 - p_i), \qquad Var(N_j) \;=\; np_j(1 - p_j)$$

to get

$$\begin{aligned} Corr(N_i, N_j) &= \frac{-np_ip_j}{\sqrt{np_i(1 - p_i)\; np_j(1 - p_j)}} \\ &= -\sqrt{\frac{p_ip_j}{(1 - p_i)(1 - p_j)}} \end{aligned}$$

Sharkey's casino

- N_R reds, N_B blacks, in n spins
- r = proportion of red slots
  b = proportion of black slots
  $1 - r - b$ = proportion of green slots
- In Las Vegas, $r = b = 18/38$.



Find $Corr(N_R, N_B)$

- Note: This should be negative for $r \approx b \approx 0.5$.
  e.g. $r = b = 0.5 \Rightarrow N_R = n - N_B \Rightarrow Corr(N_R, N_B) = -1$
- This multinomial, with $p_1 = r,\ p_2 = b,\ p_3 = 1 - r - b$
- From multinomial result

$$Corr(N_R, N_B) = \frac{-rb}{\sqrt{r(1-r)}\,\sqrt{b(1-b)}} \quad (= -0.9 \text{ if } r = b = 18/38)$$

What happens if n is random?

- Replace n by a <u>random variable</u> N.
- Since we expect $N_R \approx rN$, and $N_B \approx bN$,
  if N varies enough this should lead to positive correlation
  between N_R and N_B
- In fact – See lecture 31

$$Corr(N_R, N_B)$$
$$= \frac{rb[Var(N) - E(N)]}{\sqrt{r^2 Var(N) + r(1-r)E(N)]\ [b^2 Var(N) + b(1-b)E(N)]}}$$

- If $Var(N) = 0$ get same answer as before.
- e.g. If $E(N) = 1,000,\ SD(N) = 100,\ r = b = 18/38$, then
  $Corr(N_R, N_B) = +0.81$

Building correlations

- Take X, Y, Z independent.
- Assume $\mu_X = \mu_Y = \mu_Z = 0$, and general variances $\sigma_X^2, \sigma_Y^2, \sigma_Z^2$.
- Set $U = X + Z$, $V = Y + Z$.
- $\mu_U = \mu_V = 0$, $\sigma_U^2 = \sigma_X^2 + \sigma_Z^2$, $\sigma_V^2 = \sigma_Y^2 + \sigma_Z^2$.
- $Cov(U, V) = E\{UV\} = E\{(X + Z)(Y + Z)\} = \sigma_Z^2$.

$$\begin{aligned} Corr(U, V) &= \frac{Cov(U, V)}{\sigma_U \sigma_V} \\ &= \frac{\sigma_Z^2}{\sqrt{(\sigma_X^2 + \sigma_Z^2)(\sigma_Y^2 + \sigma_Z^2)}} \end{aligned}$$

Sample averages

- $(X_1, \ldots, X_n)$ is a sample of independent observations of a random variable X (e.g. Sampling with replacement)
- $E\{X\} = \mu$, $Var(X) = \sigma^2$.
- Define the <u>sample mean</u> $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$

THEN

- $E\{\bar{X}_n\} = E\{\sum_{k=1}^{n} X_k\}/n = [\sum_{k=1}^{n} E\{X_k\}]/n = \mu$
- $Var(\bar{X}_n) = Var(\sum_{k=1}^{n} X_k/n)$
  . $= n^{-2}\sum_{k=1}^{n} Var(X_k) = \sigma^2/n$

$$E\{\bar{X}_n\} = \mu \qquad Var(\bar{X}_n) = \sigma^2/n$$