More Programming and Statistics Boot Camps?

Prob Set 2: due Fri night 24 Feb

Midterm date closes in 5 day(s)

A total of 36 vote(s) in 33 hours

29 (81% of users) [green bar] Thu, 23 Mar 2017
7 (19% of users) [green bar] Tue, 21 Mar 2017
Brief Summary

- **Expectation value:** \( E[X] = \sum_{s \in S} p(s)X(s) \)

- **Variance:** \( V[X] = \sum_{s \in S} p(s)(X(s) - E[X])^2 \)
  \[ = E[X^2] - (E[X])^2 \]

- **Standard deviation:** \( \sigma[X] = \sqrt{V[X]} \)

- For \( X \) a sum of random variable \( X = \sum_i X_i \), the expectation always satisfies:
  \[ E[X] = \sum_i E[X_i] \]

- If (and only if) the variables \( X \) and \( Y \) are *independent*, then
  \[ E[XY] = E[X]E[Y] \]

- If (and only if) all the variables \( X_i \) are *independent*, then
  \[ V[X] = \sum_i V[X_i] \]
Bernoulli Trial

A Bernoulli trial is a trial with two possible outcomes:

- “success” with probability \( p \),
- and “failure” with probability \( 1 - p \).

Probability \((r\ \text{successes in } N \\text{ trials}) = \binom{N}{r} p^r (1 - p)^{N-r}\)
To analyze a succession of $N$ Bernoulli trials,
let random variable $X_i = 0$ or 1 represent the result of the $i^{\text{th}}$ trial,
where $X_i = 1$ if the trial is a “success” (with probability $p$). Then

$$E[X_i] = (1 - p) \cdot 0 + p \cdot 1 = p$$

Now let $X = X_1 + X_2 + \ldots + X_N$ count the total number of successes.

Using linearity of expected value, the mean satisfies

$$E[X] = \sum_{i=1}^{N} E[X_i] = Np$$
Now recall that in general $V[X_i] = E[X_i^2] - E[X_i]^2$, and here

$$E[X_i^2] = (1 - p) \cdot 0^2 + p \cdot 1^2 = p$$

so using as well $E[X_i]^2 = p^2$ we find

$$V[X_i] = E[X_i^2] - E[X_i]^2 = p - p^2 = p(1 - p)$$

Finally, since the $X_i$ are independent random variables, it follows that the variance satisfies

$$V[X] = \sum_{i=1}^{N} V[X_i] = Np(1 - p)$$

and hence the standard deviation is $\sigma = \sqrt{V[X]} = \sqrt{Np(1 - p)}$. 
For $p = 1/2$ and $N = 3$, these give

$$E[X] = Np = 3/2 \quad \text{and} \quad V[X] = Np(1 - p) = 3/4$$

reproducing the result of the coin flip example from earlier:

**Example:** again flip a coin 3 times, and let $X$ be the number of tails.

$$E[X^2] = \frac{1}{8} 0^2 + \frac{3}{8} 1^2 + \frac{3}{8} 2^2 + \frac{1}{8} 3^2 = 3$$


Here $X = X_1 + X_2 + X_3$, where $X_i$ is the number of tails (0 or 1) for the $i^{th}$ roll, then the $X_i$ are independent variables with $E[X_i] = 1/2$

and $E[X]$ is additive: $E[X] = \sum_i E[X_i] = 3 \cdot \frac{1}{2}$
Again the results

\[ E[X] = Np \quad \text{and} \quad \sigma[X] = \sqrt{V[X]} = \sqrt{Np(1-p)} \]

explain the observation that the probability gets more sharply peaked as the number of trials increases, since the width of the distribution (\( \sigma \)) divided by the average \( E[X] \) behaves as

\[
\frac{\sigma[X]}{E[X]} = \frac{\sqrt{Np(1-p)}}{Np} = \frac{\sqrt{N}}{N} \sqrt{\frac{1-p}{p}} \sim \frac{1}{\sqrt{N}}
\]

a decreasing function of \( N \) (for any fixed \( p \)).
for i,N in enumerate((1,2,4,10,40,80,160,320)):
    plt.subplot(3,3,i+1)
    plt.title('Probability of r sixes in $N$ trials'.format(N))
    plt.xlabel('Number of sixes')
    plt.ylabel('Probability')
    plt.bar(range(N+1), [bern_prob(N,m) for m in range(N+1)])

plt.figure(figsize=(18,12))

plt.show()
By the “central limit theorem” (won’t be proven here), many such distributions always tend for sufficiently large number of trials to a universal form

\[ P(x) \approx \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \]

known as the “gaussian” or “normal” distribution.

Here \( x = r \) = number of successes and mean \( \mu = E[X] \).

[This is properly normalized, with \( \int_{-\infty}^{\infty} dx \, P(x) = 1 \), and also has

\[ \int_{-\infty}^{\infty} dx \, x P(x) = \mu \quad \int_{-\infty}^{\infty} dx \, x^2 P(x) = \sigma^2 + \mu^2 \]

so the above distribution has mean \( \mu \) and variance \( \sigma^2 \).]

Setting \( \mu = Np \) and \( \sigma = \sqrt{Np(1-p)} \) for \( p = 1/6 \) thus gives a good approximation to the distribution of successful rolls of 6 for large number of trials in the example above.
Central limit theorem:

Consider a set of $N$ independent random variables $X_i$, with arbitrary probability distributions $p_i(x)$, with means $\mu_i$ and variances $\sigma_i^2$,

Consider the sum $X = \sum_{i=1}^{N} X_i$, then for sufficiently large $N$, the probability distribution of $X$ will tend to a gaussian (normal) distribution with mean:

$$\mu = \sum_{i=1}^{N} \mu_i$$

and variance:

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2$$