Chapter 1

Probability Theory

1.1 Binomial Distribution

$$(x_1 + x_2)^n = \sum_{k=0}^{n} \binom{n}{k} x_1^k x_2^{n-k}$$

1.2 Basic Probability

A fiar 6-sided die is rolled 5 times. What is the probability of exactly two 3's?

1.2.1 Outcomes

Divide favorable outcomes by possible outcomes.

$$= \frac{\binom{5}{2} \cdot 5^3}{6^5}$$

1.2.2 Raw Probability

Find the probability of getting a favorable outcome.

$$\binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$$

1.3 Basics of Expected Value

$$\text{expected value} = \sum_{\text{results}} (\text{value})(\text{probability})$$

For a die,

$$\langle k \rangle = \frac{1}{6}(1+2+3+4+5+6) = \frac{6 \cdot 7}{6 \cdot 2} = 3.5$$

$$\langle k^2 \rangle = \frac{1}{6}(1^2 + 2^2 + \cdots + 6^2) = \frac{6 \cdot 7 \cdot 13}{6} frac16 = \frac{91}{6} \neq \langle k \rangle^2$$

1.3.1 Expectation of Square vs Square of the Expectation

Consider,

$$\langle (k - \langle k \rangle)^2 \rangle = \langle k^2 - 2k\langle k \rangle + \langle k \rangle^2 \rangle = \langle k^2 \rangle - \langle 2k\langle k \rangle \rangle + \langle \langle k \rangle^2 \rangle$$

$$\langle k^2 \rangle - \langle k \rangle^2$$

Since the LHS is ≥ 0, this value is > 0 so $\langle k^2 \rangle > \langle k \rangle^2$. The variance is equal to this LHS value: $\sigma_k^2 = \langle (k - \langle k \rangle)^2 \rangle$. So, for the die, $\sigma_k = \sqrt{\frac{91}{6} - \frac{49}{4}}$. The probability of being in a standard deviation of a expected value is $P(2 \leq k \leq 5) = \frac{2}{3}$. If this distribution was normal, this value would be $\approx 68.2\%$.

1.3.2 Independence and Products

Given that k_1 and k_2 are two independent measurements, determine $k_1 + k_2$ and $\sigma^2_{k_1+k_2}$.

$$k_1 + k_2 = k_1 + k_2$$

$$\sigma^2_{k_1+k_2} = (k_1 + k_2)^2 - k_1 + {k_2}^2$$

$$= k_1^2 + 2k_1k_2 + k_2^2 - ({k_1}^2 + 2k_1k_2 + {k_2}^2)$$

$$= k_1^2 - {k_1}^2 + 2k_1k_2 - 2k_1k_2 + k_2^2 - {k_2}^2$$

Two outcomes are independent if and only if $k_1k_2 = k_1k_2$ always. So, given independence, we can simplify

$$\sigma^2_{k_1+k_2} = \sigma^2_{k_1} + \sigma^2_{k_2}$$

The value $k_1k_2 - k_1k_2$ measures the correlation between k_1 and k_2.

1.4 Multinomial Distribution

This can be used to model distributions with more than 2 objects. Consider n objects being placed in m boxes. The number of ways to place r_1 in box 1, r_2 in box 2, $\cdots$, and r_m in box m is

$$\binom{n}{r_1 r_2 \cdots r_m} = \frac{n!}{r_1! r_2! \cdots r_m!}; \sum_{i=1}^{m} r_i = n$$

Representing the distribution,

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{r_1 + r_2 + \cdots r_m = n} \binom{n}{r_1 r_2 \cdots r_m} x_1^{r_1} x_2^{r_2} \cdots x_m^{r_m}$$

1.4.1 Application

A fair 6-sided die is rolled four times. k_1 is the number of 3's and k_2 is the number of 5's.

$$k_1 = \sum_{k=0}^{4} \binom{4}{k} k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{4-k}$$

Taking a derivative of the binomial expansion and multiplying by x_1,

$$x_1 \frac{\partial}{\partial x_1}(x_1 + x_2)^n = n x_1 (x_1 + x_2)^{n-1} = \sum_{k=0}^{n} \binom{n}{k} k x_1^k x_2^{n-k}$$

Applying this,

$$k_1 = 4 \cdot \frac{1}{6} = \frac{2}{3}$$

1.5 Experimentation

A certain quantity is measured n times with the results $k_1, k_2, \cdots k_n$. Asume the expected value of k is $\bar{k}$ (unknown) and its standard deviation in σ_k (unknown).

$$k_{\text{mean}} = \frac{1}{n} \sum_{i=1}^{n} k_i$$

Note that $k_{\text{mean}} \neq \bar{k}$. However,

$$k_{\text{mean}} = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{1}{n} \sum_{i=1}^{n} \bar{k} = \bar{k}$$

Note that the expected value of both k and k_{mean} is $\bar{k}$. So, let's analyze the standard deviation,

$$\sigma^2_{k_1+k_2+\cdots+k_n} = \sum_{i=1}^{n} \sigma_i^2 = n\sigma_k^2 \Rightarrow \sigma_{\sum} = \sqrt{n}\sigma k \Rightarrow \sigma_{\text{mean}} = \sqrt{n}\frac{\sigma_k}{n} = \frac{\sigma_k}{\sqrt{n}}$$

Thus, taking the mean keeps the same expected value but divides the std. dev. by $\sqrt{n}$. Note that we don't know the values of $\bar{k}$ and σ_k. So, let's calculate $\sigma_{k_{\text{mean}}}$.

$$\begin{aligned}
\sum_{i=1}^{n}(k_i - k_{\text{mean}})^2 = \sum_{i=1}^{n} (k_i - k_{\text{mean}})^2 &= n(k_1 - k_{\text{mean}})^2 \\
= n(k_1 - \bar{k})^2 - 2(k_1 - \bar{k})(k_{\text{mean}} - \bar{k}) &+ (k_{\text{mean}} - \bar{k})^2 \\
= n\left[\sigma_k^2 + \frac{\sigma_k^2}{n} - 2(k_1 - \bar{k})(k_{\text{mean}-\bar{k}})\right]
\end{aligned}$$

Since k_1 and k_{mean} are dependent, let's look at k_2.

$$(k_1 - \bar{k})(k_{\text{mean}} - \bar{k}) = \frac{\sigma_k^2}{n}$$

Plugging this in,

$$\sum_{i=1}^{n}(k_i - k_{\text{mean}})^2 = n\left[\sigma_k^2 + \frac{\sigma_k^2}{n} - 2\frac{\sigma_k^2}{n}\right] = (n-1)\sigma_k^2$$

So,

$$\frac{1}{n-1}\sum_{i=1}^{n}(k_i - k_{\text{mean}})^2$$

1.6 Large n

Suppose we roll a fair six-sided die 6000 times. What is the probability a 2 comes up between 990 and 1050 times?

$$P = \sum_{k=990}^{1050}\binom{6000}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{6000-k}$$

This is computationally intensive, so we can approximate this instead with an integral. Generalizing, say there are n rolls and a p probability. Using Sterling's approximation, $\ln n! \approx n\ln n - n + \frac{1}{2}\ln(2\pi n)$,

$$\ln\binom{n}{k}p^k(1-p)^{n-k} = \ln n! - \ln k! - \ln(n-k)! + k\ln p + (n-k)\ln(1-p)$$

$$\approx n\ln n - n + \frac{1}{2}\ln(2\pi n) - k\ln k + k - \frac{1}{2}\ln(2\pi k) - (n-k)\ln(n-k)$$

$$+n - k - \frac{1}{2}\ln(2\pi(n-k)) + k\ln p + (n-k)\ln(1-p)$$

We want to look at this for large n. Let $k = xn$.

$$\ln P_k \approx n\ln n - xn\ln(xn) - n(1-x)\ln(n-xn) + \frac{1}{2}\ln\frac{n}{2\pi xn^2(1-x)}$$

$$+xn\ln p + n(1-x)\ln(1-p)$$

$$= n\ln n - xn\ln n - xn\ln x - n(1-x)\ln n - n(1-x)\ln(1-x)$$

$$+\frac{1}{2}\ln\frac{1}{2\pi x(1-x)} - \frac{1}{2}\ln n + xn\ln p + n(1-x)\ln(1-p)$$

Cancelling and rearranging,

$$= n\left[x\ln p - x\ln x + (1-x)\ln(1-p) - (1-x)\ln(1-x)\right] + \frac{1}{2}\ln\frac{1}{2\pi nx(1-x)}$$

$$= n\left[x\ln\frac{p}{x} + (1-x)\ln\frac{1-p}{1-x}\right] + \frac{1}{2}\ln\frac{1}{2\pi nx(1-x)}$$

Similar to asymptotic expansions, let's look at the maximum. For large n, the last term is negligible. Note that when $x = p$, the derivative and this expression vanish.

$$\frac{\partial^2}{\partial x^2} \Rightarrow -\frac{1}{x} - \frac{1}{1-x} = \frac{-1}{x(1-x)}$$

So,

$$\ln P_k \approx -\frac{n}{2}\frac{(x-p)^2}{p(1-p)} + \frac{1}{2}\ln\frac{1}{2\pi np(1-p)}$$

Substituting back to k,

$$= -\frac{1}{2n}\frac{(k-pn)^2}{p(1-p)} + \frac{1}{2}\ln\frac{1}{2\pi n(p)(1-p)}$$

Remember that $\sigma_k = np(1-p)$ from the binomial distribution:

$$P_k \approx \frac{\exp\left[-\frac{(k-np)^2}{2\sigma_k^2}\right]}{\sqrt{2\pi\sigma_k^2}}$$

This is the bell curve and is valid for large n. We also note that $np = k$ Rewriting this, we get

$$\approx \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\left(\frac{k-k}{\sigma_k}\right)^2}$$

So, back to our example,

$$P = \sum_{k=990}^{1050} \binom{6000}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{6000-k} \approx \int_{989.5}^{1050.5} \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\left(\frac{k-1000}{\sigma_k}\right)^2} dk$$

Since numerical integrations don't work that well with large values, we can substitute to rescale with $u = \frac{k-k}{\sigma_k}; du = \frac{dk}{\sigma_k}$. Recall that this u is the z^* score from statistics.

$$= \int_{z_{min}}^{z_{max}} \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} du$$

1.7 Birthday Problem

There are n problem in a room with random birthdays (none born on Feb. 29). How large must n be in order that the probability that at least two share the same birthday exceeds $\frac{1}{2}$.

1.7.1 Solution

Suppose we choose some fixed birthdays and then assign them:

$$P = 1 - \binom{365}{n} n! \left(\frac{1}{365}\right)^n$$

From a multinomial perspective, the probability of all different days is

$$\binom{365}{n, 365-n}\binom{n}{1, 1, \cdots, 1}\left(\frac{1}{365}\right)^n$$

This generalizes well. Consider the case for one pair,

$$\binom{365}{1, n-2, 366-n}\binom{n}{2, 1, 1\cdots 1}\left(\frac{1}{365}\right)^n$$

For two pairs,

$$\binom{365}{2, n-4, 367-n}\binom{n}{2, 2, 1, 1\cdots 1}\left(\frac{1}{365}\right)^n$$