Probability to Statistics: Polling Application

Suppose we want to figure out the preference of a population. For example, we want to know the fraction $p$ of the number of democrats in California. How many people should we ask before we get a reliable answer? Let $X_i$ model the answer of a person, where $X_i = 1$ if the person is a democrat and 0 otherwise. Then $X_i = 1$ with probability $p$.

We want to estimate the parameter $p$ in the model from observing the outcome of the random experiment, which is a standard problem in statistics. An estimator takes the data collected from polling and outputs an estimate $\hat{p}$ of the parameter $p$. The data in this problem is the $n$ responses $X_1, \ldots, X_n$ of the people. We estimate $\hat{p}$ from the $n$ observations, so $\hat{p}$ will be a function of the data. A reasonable estimate here is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

the fraction of the people polled who say they are democrats. How reliable is this estimate? $p$ is not a random variable, but $\hat{p}$ is a random variable because it is a function of the $X_i$s which are random variables (our data). We start by computing the expectation of $\hat{p}$:

$$E[\hat{p}] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{np}{n} = p$$

On the average, our estimator will give $p$, which is a desirable property. When $E[\hat{p}] = p$, we say our estimator is unbiased.

Although $E[\hat{p}] = p$, there is statistical randomness in $\hat{p}$, depending on the answers we got in our polling. Our hope is that the estimate $\hat{p}$ becomes more accurate as the number of people we polled increases. To see if this is true, we compute the variance of $\hat{p}$. The variance of $\hat{p}$ can be written as

$$\text{Var}(\hat{p}) = \text{Var}(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n^2} \text{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2} np(1-p) = \frac{1}{n} p(1-p),$$

where we have used the variance of a binomial distribution which we previously computed ($\sum X_i \sim Bin(n,p)$). The standard deviation is

$$\sqrt{\frac{p(1-p)}{n}}.$$

So indeed, the random fluctuation goes to zero as the number of people we polled grows!

Sum of Independent Random Variables

The above variance calculation implies that as we poll more and more people our answer becomes more reliable: the variance of $\hat{p}$ decreases when $n$ increases. As the amount of data increases, the variance shrinks.
by \( \frac{1}{n} \). The distribution of \( \hat{p} \) is shrinking closer and closer around the mean. Reliability increases because the variance is decreasing.

We seek to understand this point better: Why does the variance decrease?

Suppose we have two random variables \( X \) and \( Y \), and we want to compute \( \mathbb{E}[X + Y] \). By linearity of expectation, we know \( \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \). This says that the relation between the two random variables does not affect the expectation of their sum. Remember the problem of randomly handing out homeworks to students. Suppose \( X \) indicates when the first student gets his homework, and \( Y \) indicates when the second student gets his homework. Then whether or not the first student gets his homework back affects the outcome for the second student. Therefore there is obviously a dependence between \( X \) and \( Y \). Nevertheless it is still true that \( \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \).

On the other hand, the variance of \( X + Y \) is affected by the dependence between the variables. In general, \( \text{Var}(X + Y) \neq \text{Var}(X) + \text{Var}(Y) \).

The polling example was a special case of this result applied to the binomial distribution. In that example we have \( \text{Var}(\sum_i X_i) = np(1-p) \), where \( p(1-p) \) is the variance of each of the \( X_i \). So in the polling example, the variance of the sum is the sum of the variances. If the \( X_i \)s are all strongly positively correlated (for example they survey people that are related to each other), the results are not independent in this case. All the probability will be around \( \sum_i X_i = n \) and \( \sum_i X_i = 0 \). So the variance of \( \sum X_i \) will be of the order of \( n^2 \). We are interested in \( \frac{1}{n^2} \text{Var}(\sum_i X_i) \), which will not go to zero with \( n \).

Application of Chebyshev’s inequality to polling

In the polling example discussed above, we calculated the variance of the estimator \( \hat{p} \) to indirectly justify the fact that the estimator becomes more reliable as the number of samples increases. Here we will define reliability more precisely and use Chebyshev’s inequality to connect the variance of the estimator to reliability. The main question we want to answer is how many samples do we need to get a certain level of reliability.

We first need to define what we mean by reliable estimator. A good estimator \( \hat{p} \) will be close to \( p \) with high probability. For example, we would like an interval of width 0.1 around \( \hat{p} \) to contain \( p \) with 95% probability. More specifically, we want the interval around \( \hat{p} \) to satisfy \( \mathbb{P}(|\hat{p} - p| > 0.1) \leq 0.05 \). If this property holds, then \( [\hat{p} - 0.1, \hat{p} + 0.1] \) is said to be the 95% confidence interval.

We connect this probability and the variance by using Chebyshev’s inequality:

\[
\mathbb{P}(|\hat{p} - p| > 0.1) \leq \frac{\text{Var}(\hat{p})}{0.1^2}.
\]

We would like the right hand side to be \( \leq 0.05 \). The variance of \( \hat{p} \) can be written as

\[
\text{Var}(\hat{p}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} np(1-p) = \frac{1}{n} p(1-p),
\]

where we have used the variance of a binomial distribution which we previously computed (\( \sum_i X_i \sim Bin(n, p) \)).
Therefore, we need
\[ \frac{p(1-p)}{0.1^2} \leq 0.05, \]
or equivalently:
\[ n \geq \frac{p(1-p)}{0.05 \times (0.1)^2}. \] (1)

If we knew \( p \), then (1) would tell us how many people we need to poll. But the problem is that we don’t know \( p \). Observe though that the maximum value of \( p(1-p) \) is 0.25 (achieved at \( p = 0.5 \)). So if we are a bit conservative and choose \( n \) such that
\[ n \geq \frac{0.25}{0.05 \times (0.1)^2} = 500, \]
then the inequality (1) will be satisfied regardless of the true value of \( p \).

Law of Large Numbers

The increasing reliability of the estimator in the polling example is a special case of one of the most important theorems in probability, the law of large numbers:

**Theorem 10.1: [Law of Large Numbers]** Let \( X_1, X_2, \ldots, X_n \) be random variables each having the same distribution with the common expectation \( \mu = \mathbb{E}[X_i] \). Suppose every pair of the random variables is independent. Define \( A_n = \frac{1}{n} \sum_{i=1}^{n} X_i \). Then for any \( a > 0 \), we have
\[ \mathbb{P}[|A_n - \mu| \geq a] \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty. \]

**Proof:** Let \( \text{Var}(X_i) = \sigma^2 \) be the common variance of the r.v.’s; we assume that \( \sigma^2 \) is finite\(^1\). With this (relatively mild) assumption, the law of large numbers is an immediate consequence of Chebyshev’s Inequality. For, as we have seen above, \( \mathbb{E}[A_n] = \mu \) and \( \text{Var}(A_n) = \frac{\sigma^2}{n} \), so by Chebyshev we have
\[ \mathbb{P}[|A_n - \mu| \geq a] \leq \frac{\text{Var}(A_n)}{a^2} = \frac{\sigma^2}{na^2} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty. \]

This completes the proof. \( \Box \)

In English, the law of large numbers says that the sample average \( A_n \) of \( n \) random variances converges to the mean of the random variables as \( n \) grows to infinity. It is the basis of statistics and simulations, among other things. For example, in your labs, doing many multiple trials of a simulation and then averaging the results of the multiple trials to calculate the fraction of successful trials is a good estimate of the probability of success of a trial, because of the law of large numbers.

The Poisson Distribution

The Poisson distribution is a very widely accepted model for so-called “rare events”, such as misconnected phone calls, radioactive emissions, crossovers in chromosomes, etc. Here, we motivate this distribution

\(^1\)If \( \sigma^2 \) is not finite, the LLN still holds but the proof is much trickier.
with the following example. Suppose Verizon wants to deploy a wireless network in a city. At any time, the number of people making calls is random and cannot be predicted ahead of time. There are many paying customers in the network, and all of them can potentially make a call during the same period of time. However, only a very small fraction of them actually will. Verizon needs to size the capacity of its network such that it can support a large number of users at any given point. Suppose an average call lasts for one minute. If we want to know how many people are using the same network at the same time, we need the number of people that initiate a call within the same minute. Focusing on a window of one minute, we are interested in the random variable which counts the number of calls initiated within that minute. What is a reasonable distribution to model this random variable?

Let $X$ be the random variable that counts the number of calls initiated within a minute. To model the random variable $X$ we need to make some reasonable assumptions. First, we need to collect some statistics. A useful quantity to measure is the rate at which calls are coming in. We count the number of calls generated in $i$ minutes, and divide it by the number of minutes $i$ to get a rate of call arrivals per minute. Let $\lambda$ be the rate of call arrivals (with unit the number of calls per minute). Then $\lambda = E[X]$, the expected number of arrivals in a minute. We would like to figure out the rest of the distribution.

Let us focus on that one minute interval and divide it into $n$ very small subintervals. The first assumption we make on these subintervals is that the chance that we have more than one initiation within a very small subinterval is negligible. People are making the calls separately, and the chance of having a collision is small. More specifically, the first assumption we make is:

1- The probability of having more than one arrival in a very small interval is negligible. Two events can happen in a small interval: either no initiations are made, or exactly one.

Let $X_i$ be the number of call initiations in the $i$th interval. Then $X = X_1 + \cdots + X_n$, where

$$X_i = \begin{cases} 0 & \text{if no calls are initiated in the } i\text{th sub-interval}, \\ 1 & \text{otherwise.} \end{cases}$$

What should the probability of $X_i = 1$ be? We need to express it in terms of $\lambda$ and $n$. This probability must be $\frac{\lambda}{n}$ because $E[X] = \lambda = E[X_1] + \cdots + E[X_n]$. (Here we are assuming that each of the $X_i$’s have the same distribution.) We are still missing the understanding of the relationship between the $X_i$ variables. A reasonable model is that the event that a call is initiated in interval $i$ is independent from the event that a call is initiated in interval $j$. Therefore, we make the following second assumption:

2- The events of call initiation in different sub-intervals are mutually independent.

The random variable $X$ counts the number of call initiations in the 1-minute interval. Under the two assumptions, $X$ becomes a binomial random variable: $X \sim Bin(n, \frac{\lambda}{n})$. The parameters of this binomial are related together by $n$. We now have that

$$P(X = i) = \binom{n}{i} \left( \frac{\lambda}{n} \right)^i \left( 1 - \frac{\lambda}{n} \right)^{n-i}, \quad i = 0, 1, \ldots, n.$$

To obtain the Poisson distribution, we need to take the limit of the above expression as $n$ goes to infinity. To do that, we write,

$$P(X = i) = \frac{n(n-1) \cdots (n-i+1)}{i!n^i} \lambda^i \left( 1 - \frac{\lambda}{n} \right)^{n-i}$$
For fixed $i$, as $n$ tends to infinity,
\[
\frac{n(n-1)\ldots(n-i+1)}{n^i} \to 1
\]
\[
(1 - \frac{\lambda}{n})^{n-i} \to e^{-\lambda}
\]
and hence,
\[
P(X = i) \to \frac{\lambda^i e^{-\lambda}}{i!}
\]
giving the Poisson distribution of the random variable $X$. By choosing a large $n$, it is thus reasonable to model the number of call initiations during a one-minute period to be Poisson with parameter $\lambda$.

**Definition 10.1 (Poisson distribution):** A random variable $X$ for which
\[
P(X = i) = \frac{\lambda^i e^{-\lambda}}{i!} \quad \text{for } i = 0, 1, 2, \ldots
\]
is said to have the Poisson distribution with parameter $\lambda$. This is abbreviated as $X \sim \text{Poiss}(\lambda)$.

Let’s verify that this distribution sums to 1.
\[
\sum_{i=0}^{\infty} P(X = i) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}
\]
But, using the Taylor series expansion of $e^\lambda$, we know that $e^\lambda = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$. Thus,
\[
\sum_{i=0}^{\infty} P(X = i) = e^{-\lambda} e^\lambda = 1
\]
Let’s compute the expectation of the Poisson distribution.
\[
E[X] = \sum_{i=0}^{\infty} i \times P(X = i) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!}
\]
The term with $i = 0$ can be ignored as it is 0.
\[
E[X] = \sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!}
\]
\[
E[X] = \sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-1)!}
\]
Changing index of summation to $j = i - 1$, we get,
\[
E[X] = \lambda \sum_{j=0}^{\infty} \frac{\lambda^j e^{-\lambda}}{j!}
\]
But now the sum is 1 as seen above when confirming that the distribution sums to 1. Thus,
\[
E[X] = \lambda
\]
Thus, the calculations for the expectation match the way we motivated the Poisson distribution in the first place. By similar calculations, it can be shown that $\text{Var}(X) = \lambda$. 

EE 178, Spring 2018, Lecture 10