Hypothesis Testing with Probability Distributions

Hypothesis testing is a part of statistics in which you use trends in data to determine the validity of a running hypothesis about a set of data by assessing the probability of a measured result under the working hypothesis. If this probability is low enough, namely below some particular “significance level”, the working hypothesis can be scrapped in place of a new hypothesis. There is a lot of notation around this topic, which I will list here:

$\mathrm{X} - \text{A random variable (holds any value, but is unknown)}$
$\mathrm{A, B} - \text{An equation or inequality in the form } \mathrm{X} = \mathrm{x}$
$\mathrm{A'} - \text{The compliment to A, all other possibilities than those in which A is the case}$
$\mathrm{B(n, p)} - \text{A binomial distribution with n trials at p probability per single trial}$
$\mathrm{N}(\mu, \sigma^2) - \text{A normal distribution with mean } \mu \text{ and variance } \sigma^2$
$\mathrm{P}(\lambda) - \text{A Poisson distribution with an average of } \lambda \text{ occurences of an event in a constant interval}$
$\mathrm{X \sim F(args)} - \text{A random variable X distributed over some distribution function F}$
$\mathrm{P(X = x)} - \text{The probability of random variable being equal to a variable x under an independent test}$
$\mathrm{P(A \cup B)} - \text{The probability of A or B, } \mathrm{P(A) + P(B) - P(A \cap B)}$
$\mathrm{P(A \cap B)} - \text{The probability of A and B, } \mathrm{P(A) * P(B|A)} \text{ or } \mathrm{P(B) * P(A|B)}$
$\mathrm{P(A|B)} - \text{The probability of A given that B is known to be the case}$
$\mathrm{H_0} - \text{The null hypothesis, the current working theory under which beliefs lie}$
$\mathrm{H_1} - \text{The alternate hypothesis, a new theory which can take the place of } \mathrm{H_0} \text{ when proved to be the case}$
$\alpha - \text{The significance level, the probability of a set of test outcomes } \mathrm{P(X \leq x \cup X \geq x)} \text{ in which } \mathrm{H_0} \text{ can be rejected}$

Binomial Distribution:

A binomial curve is used in probability for random variables under a set few conditions:

- There is an event occurrence, or trial, which can be triggered and has a random outcome.
- The random outcome has two possible complimentary states, usually A and A’, success and failure.
- The probability of outcome A is fixed and unchanging for each individual event.
- One events outcome does not affect the probability of other occurrences of the event.
- X is a random variable distributed across a binomial curve with parameters n and p.
- X is the number of “successes”, in n trials, at an individual probability of p.

The binomial expansion which forms the probability is as such:

$$(\mathrm{a}+\mathrm{b})^c = \sum_{d=0}^{c}\binom{c}{d} * (a)^d * (b)^{c-d} = \sum_{x=0}^{n}\binom{n}{x} * (p)^x * (1-p)^{n-x} = \big(p+(1-p)\big)^n = 1 = 100\%$$

This formula is written as $\mathrm{X \sim B(n, p)}$. This formula is a summation of the probability of x successes out of n trials, at p probability per trial, for every possible value of x, hence, this formula sums to 1 (100% of outcomes). Due to this summation, the input to the sigma can be taken out, instead using a single value for x. This formula can be broken down in a comprehensive way, with arbitrary variables n = 6, x = 4, p = 3/5:

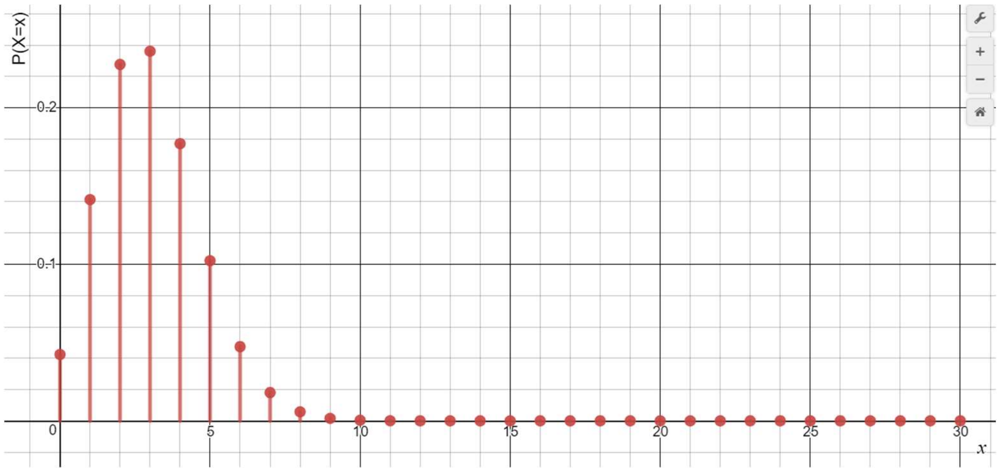
$$\binom{n}{x} * (p)^x * (1-p)^{n-x} = {}_6C_4 * \frac{3^4}{5} * \frac{2^2}{5} = 15 * \left(\frac{3}{5} * \frac{3}{5} * \frac{3}{5} * \frac{3}{5}\right) * \left(\frac{2}{5} * \frac{2}{5}\right) = \frac{4860}{15625} = 31.104\%$$

As multiplication can be visualised as “and”, and the nCr function represents the number of possible combinations of r choices from a set of n objects, this formula is clearly shown as the probability for the 6 choices: a success and then a success and then a success and then a success, and then a failure and then a failure, or any of the other possible orders of this ratio between successes and failures. This helps justify the requirement for the probability of each event to be independent of each other, as the equation for the “and”ing of multiple probabilities is actually $\mathrm{P(A)} * \mathrm{P(B|A)}$, and $\mathrm{P(B)} = \mathrm{P(B|A)}$ only when the two events are independent. This probability, with one integer value for x, can be written as $\mathrm{P(X = x)}$. For a range of values, you can simply sum a range of values for x somewhere within the range $0 \leq \mathrm{x} \leq \mathrm{n}$. Due to the integer nature of the input to the binomial function, finding a range of values for $\mathrm{X \leq x}$ is the same as finding the area under the curve between 0, and the point x, or finding the integral of the distribution between 0 and x. This integral probability is what the significance level is compared to, where $\alpha \leq \mathrm{P(X \leq x \cup X \geq x)}$, the inequality direction being based on the hypothesis $\mathrm{H_1}$ compared to the working theory $\mathrm{H_0}$.

A regular hypothesis test with a binomial distribution can look like this:

> There is a working hypothesis, H_0, that the probability of a manufacturing defect in any given individual product of a particular type is 10% (1/10). Out of a test batch of 30 products, 6 of them turn out to be defected. With a significance level of 5%, can it be shown that the estimated probability of a defect is under the true value?

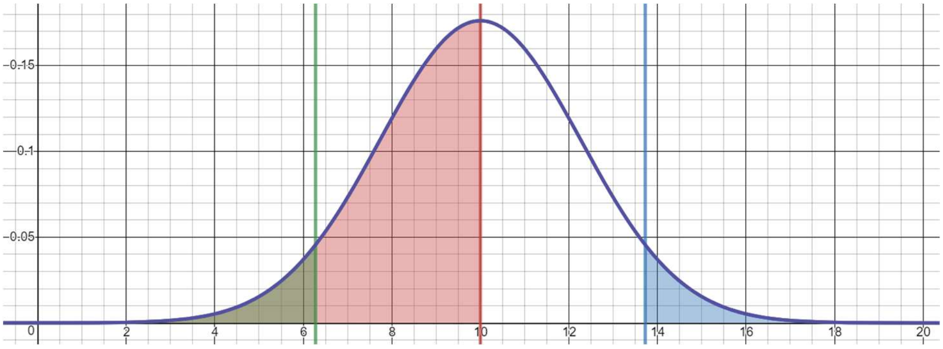
We need a binomial distribution for this, as it is an example of discrete events (each product made), with a fixed probability (10%) of a “success” (defect), and a “failure” (not defect), under a certain number of trials (30). A binomial distribution under these parameters is written as B(30, 0.1), and the random variable X, the number of defects, can be defined by X ~ B(30, 0.1). A graph of the actual value of X against the probability of said value would look like so:



From this graph, it is clear that the most likely outcome is 3 defects, and this is definitely the maximum point, as 10% of 30 is exactly 3, not any decimal close to it. If the mean fell exactly between 2 integer values, either value would be equally likely. Because a binomial distribution’s random variable can only be integer, a cumulative distribution graph is easy to make, by summing all previous integer heights to create a new integer height. Around half of the probability falls in the 3 bars to the left of $x = 3$, the other ~50% falling under the right 27 points. Because $P(X \leq x)$ is the sum of the heights of all points leading up to x, it can be seen as the integral of the distribution, between 0 and x.

We want to find that this value for the probability, 10%, is an understatement. To do this, we need to find that the probability of 6 defects, or any values higher than 6 for X, is under 5%, the significance level. In other words, we need to find whether or not the sum of the probabilities that make up the range of outcomes the measured value is on the bound of, is low enough to say that H_0 is incorrect. To do this, we need to work out the sum of the probability for 6, 7, 8, 9, 10…, 29, and 30 defects. It is much easier to calculate the values for which $X < 6$, then subtract this probability from 1, to get whatever remaining probability there is. Doing this for the data in the question, we get that there is a 92.68% chance of there being less than 6 defects, and thus 7.32% chance of 6 or more defects. This value is greater than 5%, meaning this range of values is too likely under the current hypothesis, meaning it cannot be discarded. To find the number of defects required to disprove H_0, you can keep reducing the lower bound of the range $6 \leq x \leq 30$, until the likelihood of the number of defects falling within that range being that lower bound is less than 5%.

Modelling a curve through this distribution helps show the integral properties of this area. As you can see on the graph, the red area (the integral between 0 and the mean), makes up 50% of the graph, and the green and blue sections show the outer portions of the graph with areas of 0.05 (the 5% significance level). In this case, any values from 0 to 6, or 14 to 20 will allow for H_0 to be scrapped, and the values from 7 to 13 do not. This theory will be used for all 3 types of distributions, but can only have integer values for binomial and Poisson, as both output discrete random variables.


0.15
0.1
0.05
0
2
4
6
8
10
12
14
16
18
20