

# Common Distributions and Central Limit Theorem

**Problem:** Describe the generating mechanism and derive the pmf for a random variable that is binomially distributed.

**Problem:** Describe the generating mechanism and derive the pmf for a random variable that is geometrically distributed. [Hint: Use the LLN to reason out the expectation of this distribution.]

**Question:** What are the main features of the Cauchy distribution?

**Problem:** Given two independent normally distributed RV  $X \sim N(\mu_x, \sigma_x)$  and  $Y \sim N(\mu_y, \sigma_y)$ , what can be said about  $Z = X + Y$ ?

**Question:** What is the Central Limit Theorem?

**Problem:** You are playing a spinner game which pays 0, 2 or 9 with  $P(0) = 0.5$ ,  $P(2) = 0.3$ ,  $P(9) = 0.2$  - what is the approximate probability of having \$181 after 64 rounds?

**Problem:** Derive the normal approximation to the binomial distribution and to the estimator for the binomial parameter  $p$ ?

## Contents

### 1 Binomial Distribution

We are performing  $n$  **independent** trials where each trial is a success with probability  $p$ . If the random variable  $X$  is the number of successes in this set of  $n$  trials, then  $X \sim B(n, p)$ .

*Proof.* Since each trial is independent, by the multiplication rule,

$$P(\text{a specific sequence of } k \text{ successes and } n-k \text{ failures}) = p^k(1-p)^{n-k}. \quad (1)$$

Since all specific such sequences are disjoint, the total probability for the event that is the union of them all,  $P(X=k)$ , is the sum of the individual probabilities. The number of such sequences is the number of ways you can pick a committee of  $k$  locations in the sequence to be successes.

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (2)$$

□

## 2 Geometric Distribution

Imagine repeated independent trials with success probability  $p$ . You repeat until the first success, and stop. Let  $T$  denote the total number of trials you perform. Then we say that  $T$  has a geometric distribution with success probability  $p$  i.e.  $T \sim \text{Geom}(p)$ .

$$T = \sum_{i=1}^k I_i, \text{ where all } I_i = 0 \text{ except the last one.} \quad (3)$$

$$P\{T = k\} = P\{I_1 = 0 \cap I_2 = 0 \cap \dots \cap I_k = 1\} = P\{I_1 = 0\} \dots P\{I_k = 1\} \quad (4)$$

$$P\{T = k\} = (1 - p)^{k-1} p \quad (5)$$

If we were to take many many repeated samples of  $T$  and string them all together we would have a very long string of failures punctuated by successes - let us say the combined string is of length  $n$ . Since all the trials are independent and we know the probability of success we can say that the number of successes in our very long string is  $K \sim np$  - but this is also identically the number of individual substrings since each success demarcates the end of a substring. We also know the length of the string will be approximately the number of substrings multiplied by their average length (the expectation),  $n \sim \bar{T}K \implies \bar{T} = 1/p$ .

## 3 Cauchy Distribution

The cauchy distribution is pathological because it has no mean. The average of a sample of  $n$  Cauchy random variables has a Cauchy distribution the average of  $n$  observations has the same distribution as a single observation!

$$f(x; x_0, \gamma) = \frac{1}{\pi \gamma \left[ 1 + \left( \frac{x - x_0}{\gamma} \right)^2 \right]} = \frac{1}{\pi \gamma} \left[ \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \right], \quad (6)$$

where  $x_0$  is the location parameter, specifying the location of the peak of the distribution, and  $2\gamma$  is the scale parameter which specifies the full-width at half-maximum.

## 4 Normal Distribution

A normally distributed RV,  $X \sim \mathcal{N}(\mu, \sigma^2)$ , with mean  $\mu$  and variance  $\sigma^2$  has the following density

$$f(x | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (7)$$

The standard normal distribution has  $\mu = 0$  and  $\sigma = 1$  and in general

$$X \sim N(\mu, \sigma^2) \text{ if } \frac{X - \mu}{\sigma} \sim N(0, 1) \quad (8)$$

A handy rule to remember is that in a normal distribution 65% of the probability mass lies within  $1\sigma$  of the mean, while 95% of the probability mass lies within  $2\sigma$ . A QQ plot is a quick way to check for the normality of data.

### 4.1 Problem: Sum of Normal RVs

We have two independent normally distributed RV  $X \sim N(\mu_x, \sigma_x)$  and  $Y \sim N(\mu_y, \sigma_y)$ . To find the pmf for  $Z = X + Y$  we need to add all the ways that  $Z = z$ ,

$$P(Z = z) = \sum_x P(X = x, Y = z - x) = f_Z(z) = \sum_x f_x(x) f_y(z - x) \quad (9)$$

First of all, using linearity of expectation and the independence of  $X, Y$  (which gives the sum rule for variance) we can say

$$E[Z] = \mu_x + \mu_y \quad (10)$$

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = \sigma_x^2 + \sigma_y^2 \quad (11)$$

Finally, substituting in the analytical forms for  $f_x$  and  $f_y$  we could do some nasty algebra to show that **the sum of two normal variables is itself normally distributed**.

### 4.2 Central Limit Theorem

First recall some thing we already know regarding sample means, and the Law of Large Numbers.

We know just from algebraic properties that the distribution of the  $n$ -sized sample mean (which is thought of as the average of  $n$  iid RVs) has  $E[\bar{X}] = \mu$  and  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ , where  $\mu$  and  $\sigma^2$  are those of the generating distribution of the RV. The LLN tells us that as  $n$  gets large the distribution of the sample mean has an increasing amount of its probability mass right around  $\mu$  (converges in probability to  $\mu$ ). The CLT is going to describe the distributional form of the fluctuations around  $\mu$  during this convergence.

**Central Limit Theorem.** Let  $\{X_1, \dots, X_n\}$  be a sequence of iid RVs drawn from a distribution with mean  $\mu$  and finite variance  $\sigma^2$ . Define  $S_n := \frac{X_1 + \dots + X_n}{n}$ . Then  $S_n$  converges in distribution to  $N(\mu, \sigma^2/n)$ . To avoid talking about converging to a moving target (since  $n$  is changing) we should properly state it as  $\frac{S_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0, 1)$ .

This says that the distribution of the sample mean behaves increasingly like a normal distribution with rapidly diminishing variance. The CLT can instead be framed for the sum of a sequence of RVs rather than their average, so we say for large  $n$  we have  $X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$ . Also note there is another form of this law in which the RV's don't need to be identically distributed or independent, but it comes with more conditions on the strength of the dependence and on how comparable the variances must be.

### 4.3 Problem: Spinner Game After Many Plays

The number of dollars we have after  $n$  plays is the sum of the individual outcomes of each play:  $D_n = \sum_i^n X_i$ . We know the result of each play is iid, and we can calculate for any  $i$  that  $E[X_i] = 2.4$  and  $\text{Var}(X_i) = 11.64$ . So from this we know  $E[D_{64}] = (64)(2.4) = 153$  and  $\text{Var}(D_{64}) = (64)(11.64) = 744.96$ . By the CLT we expect the distribution of  $D_{64}$  to be approximately normal, and the standard deviation is  $\sqrt{744} \sim 27$ . That means that the result \$181 lies just at one standard deviation above  $\mu$  and therefore covers about 84% of the probability mass.

### 4.4 Binomial Approximation to Normal

If we have  $X \sim \text{Binom}(n, p)$  then  $X = \sum_i^n I_i$  where the  $I_i$  are iid indicator variables taking value 1 with probability  $p$ . We know  $E[I] = p$  and  $\text{Var}(I) = p(1-p)$ . Since  $X$  is just a sum of iid RVs, if  $n$  is large we can apply the CLT to say that  $X \sim N(np, np(1-p))$ . If we define a random variable  $\hat{p} = \frac{X}{n}$  then what we see is  $\hat{p} \sim N(p, \frac{p(1-p)}{n})$ . So if given a single value of  $X$  as from a poll or an experiment, we can use  $\hat{p}$  as an estimate of the true success probability  $p$  and the larger  $n$  is the more assured we are that our calculated  $\hat{p}$  is close to the true  $p$ .

It's important to note that the standard deviation has some dependence on  $p$  - as  $p$  approaches 0.5 from either side the size of the standard deviation for fixed  $n$  increases. The more each trial looks like a coin flip the larger your sample size needs to be to achieve a desired standard deviation of  $\hat{p}$ .