Masters Bridge Program **Summer 2021**

Session 3: Convergence theorems

Instructor: Bryan Liu

We go out in the world and collect data. In a simplistic model of the world, let the observed data-points be denoted $X_1, ..., X_n$, and we view the data as coming from random draws of a distribution F. If each data point is drawn independently from all others, then we say $X_1, ..., X_n$ are drawn *i.i.d.* from F; here, "i.i.d." stands for *independently and identically distributed*. Compactly, we write

$$X_1, ..., X_n \overset{iid}{\sim} F.$$

In most applications, the distribution F is unknown. The job of the statistician is to use the data in order to estimate some properties of the distribution F. For example, we might be interested in the mean of the distribution F, and we could estimate the unknown mean of F with the sample mean $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$.

Why is this a good estimate?

The law of averages. Suppose $X_1, ..., X_n \overset{iid}{\sim} F$. Let $\mu = \mathbb{E}(X_1)$ be the mean of the distribution F. Define the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

Then for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \epsilon) \to 1 \text{ as } n \to \infty. \tag{3.1}$$

Exercise 1. Use Chebychev's inequality (see prevous session) to prove the law of averages.

IN CLASS: demo with simulated coin flips.

3.1 Normal approximations

3.1.1 Approximation to the binomial distribution

Recall that a random variable X is *binomially distributed* with parameters (n, p) if

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{3.2}$$

This random variable models the number of heads if a coin is flipped n times, with each flip having probability p of landing heads.

Below, we plot the distribution of X for $p = 0.2$, $n = 10$, $n = 20$, $n = 50$, and $n = 100$.

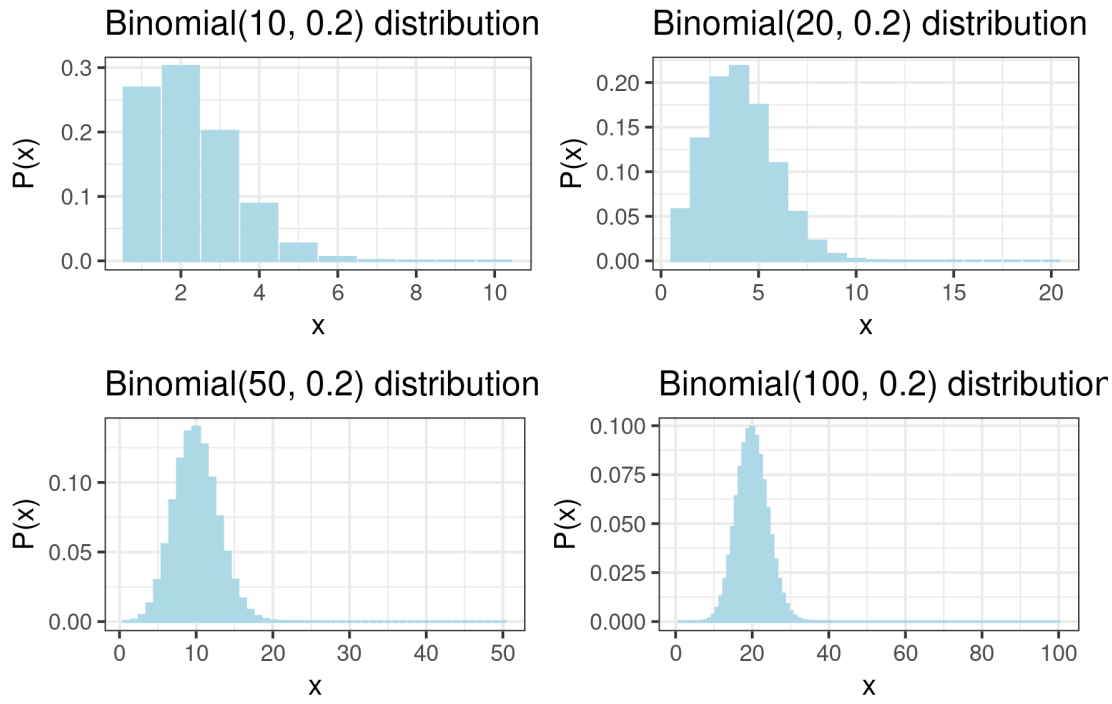

Figure 3.1: Biniomial distributions for p fixed at $p = 0.2$ while n varies.

What do you notice about the Binomial distributions in Figure 3.1?

The normal approximation to the binomial distribution. Let X be binomially distributed with parameters (n, p). If $np(1-p)$ is large, then

$$\mathbb{P}(X \in [a, b]) \approx \int_l^u \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2} \, dx,$$

where

- $\mu = np$
- $\sigma^2 = np(1-p)$
- $l = a - \mu - 0.5\sigma$
- $u = b - \mu + 0.5\sigma$.

In other words, X is approximately follows a normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$.

Note: the extra ± 0.5 is called the continuity correction, and helps make better approximations. It does not make a big difference, unless a and b are close.

Exercise 2 (Airline overbooking)**.** An airline estimates that about 90% of passengers who reserve seats will show up for their flight. On a particular flight with 300 seats, the airline accepts 324 reservations. Assume all passengers show up independently of each other. Use the normal approximation to the binomial distribution to compute the probability that the flight will be overboooked.

3.1.2 Confidence intervals

The normal approximation above allows us to construct confidence intervals. As a concrete example, suppose we want to estimate the proportion of voters in the United States who will vote for the Democratic Party in the upcoming election. Let p be the proportion of registered voters who will vote Democratic.

The true proportion p is an unknown quantity that we would like to estimate. To form an estimate, we randomly subsample $n = 200$ individuals from the set of registered voters and ask about their voting preferences. Suppose that in the sample, 110 out of the 200 individuals responded that they would vote Democratic. Let us estimate p with $\hat{p} = 110/200 = 0.55$. How good is the estimate $\hat{p}$?

We will use the normal approximation to construct a confidence interval using the following steps:

1. We model the number of Democratic respondents as $n\hat{p} \sim$ Binomial(n, p). In our polling example, $n = 200$ and p is unknown.
2. Using the normal approximation, we make the simplification that $n\hat{p} \sim \mathcal{N}(np, np(1-p))$.
3. By the shifting and scaling properties of a Normal, observe that
$$Z = \frac{n\hat{p} - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1).$$
4. For a standard normal variable, note that
$$\mathbb{P}(-1.96 \leq Z \leq 1.96) = 0.95.$$
5. Plug in our definition of Z, and re-arrange:
$$\mathbb{P}\left(-1.96 \leq \frac{n\hat{p} - np}{\sqrt{np(1-p)}} \leq 1.96\right) = 0.95. \iff \tag{3.3}$$
$$\mathbb{P}\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95 \tag{3.4}$$
6. The above display 3.4 says that the interval
$$\left[\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}\ ,\ \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right] \tag{3.5}$$

will cover the true, unknown p with probability 0.95.

However, the formula 3.5 still depends on the unknown p. We plug in $\hat{p}$ for p to construct our final confidence interval,

$$\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Applying this formula to our polling example with $\hat{p} = 0.55$ and $n = 200$, our confidence interval is $[0.51, 0.59]$.

Discussion: why is the Binomial model in step 1 appropriate here? In what scenarios might it be inappropriate?

3.1.3 Central limit theorem

The Normal approximation is not unique to limited to distributions. More generally, we have

The central limit theorem. Suppose $X_1, ..., X_n \overset{iid}{\sim} F$. Let $\mu = \mathbb{E}(X_1)$, be the mean of the distribution F and $\sigma^2 = \text{Var}(X_1)$ be its variance. Assume $\sigma^2 < \infty$. Then the random variable

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

is approximately distributed as $\mathcal{N}(0, 1)$. In other words

$$\mathbb{P}(Z \in [a, b]) \approx \int_a^b \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}x^2}\ dx,$$

for all $a \leq b \in \mathbb{R}$.

Exercise 3 (CLT-based confidence intervals)**.** Suppose I want to know if a certain GRE prep course actually improves GRE scores for participants. I select a sample of 50 students who took this course, and for each student, I compute their difference in scores before and after the course. The sample average of these differences was +5, with a sample standard deviation of 8.

(a) Use the central limit theorem to construct a 95% confidence interval for the change in GRE scores.

(b) Is there statistical evidence that students perform better on GRE at the end of this course?

(c) Can we conclude that this GRE course is helpful for students? Why or why not?

3.2 Poisson approximation

In the section above, we saw that if $np(1-p)$ is large, then a normal distribution can be used to approximate the binomial distribution. However, the normal approximation will not be appropriate

if $1/p$ is on the same order of magnitude as n. See Figure 3.2. Do you see above why the normal approximation is not be appropriate here?

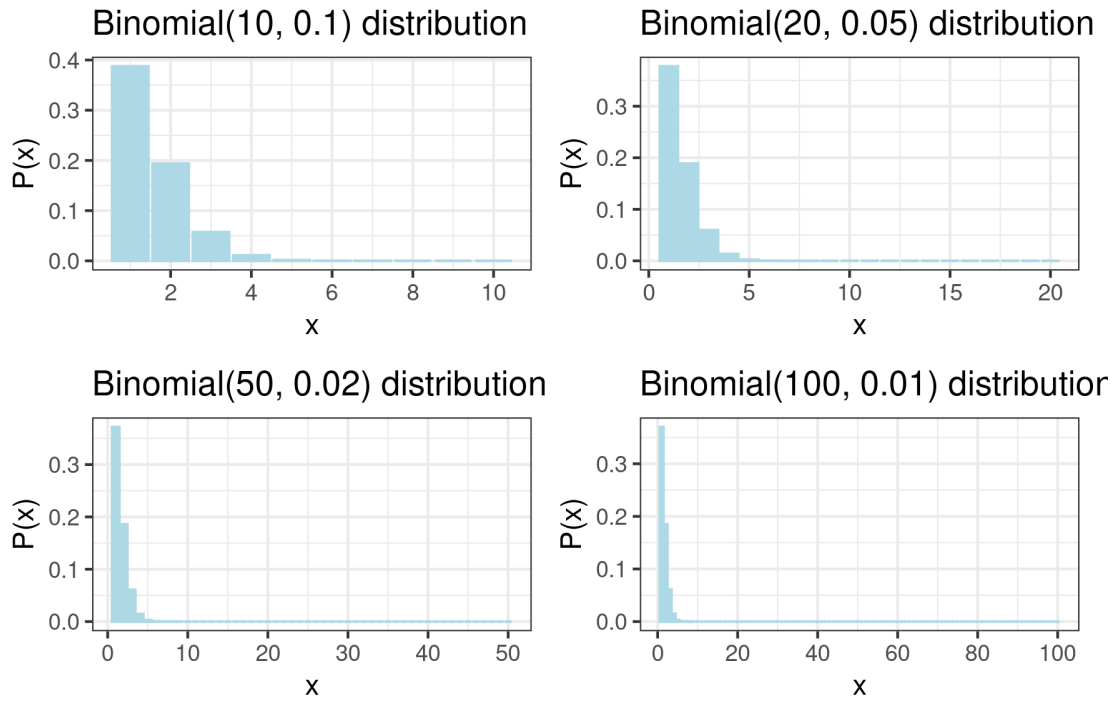

Figure 3.2: Biniomial distributions for np fixed at $np = 1$, while n varies.

A more appropriate approximation in such scenarios is the Poisson distribution.

The Poisson approximation to the binomial distribution. Let X be binomially distributed with parameters (n, p). If n is large and p is small, then X is approximately Poisson distributed with mean $\lambda = np$. That is,

$$\mathbb{P}(X = k) \approx \exp^{-\lambda} \frac{\lambda^k}{k!}$$

In particular, the approximation becomes exact as $n \to \infty$ with $p = \lambda/n$.

This is sometimes called the "law of small numbers" or the "law of rare events," since convergence happens as p gets small with increasing n.

Example 3.1 (Raindrops in a bucket)**.** I place a bucket outside in a rainstorm. The number of raindrops falling in my bucket is well-modeled with a Poisson distribution. The total number of raindrops falling from the sky n is large, but the probability p of any raindrop falling in my bucket is small.

Exercise 4. Suppose that a manufactoring process has a 1% chance of producing a defective item.

Use the Poisson approximation to compute the probability that in a batch of 200 items, there are two or more defective items.

3.2.1 Poisson arrivals

The Poisson distribution is often used to model the number of occurences of some event during an interval of time. Examples include:

- The number of car accidents at an intersection in a given week.
- The number of earthquakes occuring in the Bay area in a given month.
- The number of customers arriving at a coffee shop between noon and 1pm.

In this subsection, we will use the Poisson approximation to the binomial to justify why (or why not) the Poisson distribution is appropriate in these examples.

For concreteness, let us take the last example, the number of arrivals at a coffee shop. Suppose from historical data, we observe that on average, $\lambda = 30$ customers arrive between noon and 1pm.

To see why the Poisson distribution is apt, divide the time between noon and 1pm into $(1/n)$-hour intervals. Assume

- At most one arrival can occur during any $(1/n)$-hour interval.
- Whether or not an arrival occurs during a $(1/n)$-hour interval happens with probability λ/n.
- The occurance of arrivals in each interval is independent of all other intervals.

With the above assumptions, the total number of arrivals in a one-hour period is distributed as Binomial$(n, \lambda/n)$ (do you see why?).

Taking $n \to \infty$, the law of rare events shows that the number of arrivals is Poisson with mean λ.

So when are these assumptions appropriate? The first is mostly non-controversial, since the intervals get smaller and smaller as $n \to \infty$. The second says that the *rate of arrivals are constant* in this one-hour interval. This assumption would not be met if, for example, there is a large undergraduate lecture that ends at 12:30pm, resulting in a surge of customers at 12:35pm. The last assumption says that *arrivals must be independent* – this assumption would not be met if, for example, customers tend to arrive in pairs or groups.

This does not mean that the Poisson distribution is useless as a model for the world. Often, these simple distributions form the building blocks for more complex models. We start with simple models, and iteratively refine them, either by incorporating domain expertise or through data driven algorithms. All model building exercises depend on careful analysis of the assumptions – and these assumptions may be rejected based on either domain knowledge or observed data.

Exercise 5 (Arrival times)**.** Often, we are also interested not only in the number of arrivals, but also when these arrivals occur. Show that if the number of arrivals in t-units of time follows a Poisson distribution with mean λt, then the *time* of the first arrival follows an exponential distribution with rate λ.

3.3 Proofs

Time permitting, prove the Normal and Poisson approximations to the binomial ...