Convergence Concepts

Limit Theory

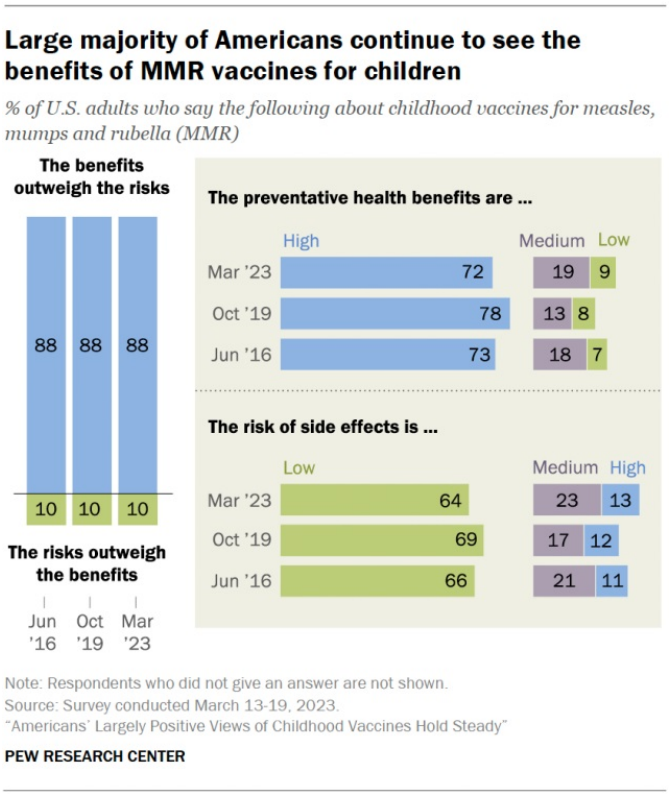
Motivation

By limit theory we mean "what is the behavior our our random variable as some quantity grows?" Usually, we concern ourselves with the sample size (n) being the quantity that grows. We'll look at two major ideas:

-
-

Motivating Example

A Pew Research Center survey of 10,701 U.S. adults was conducted in March 2023. The survey asked participants questions related to their thoughts on vaccination. One question centered around the perceived efficacy of the MMR vaccine.



The Center survey finds 88% of Americans say the benefits of childhood vaccines for measles, mumps and rubella (MMR) outweigh the risks, compared with just 10% who say the risks outweigh the benefits.

The sample proportion of 0.88 is an **estimate of the population proportion**. That is, the actual proportion of U.S. adults that believe the benefits outweigh the risks (call the true proportion p).

Of course this is a single number estimate that would change if we sampled again. We can report the standard deviation of this sample proportion, called a standard error, to give us an idea of the variability in the estimate.

Assuming independence between study participants, we can find an estimated standard error for this sample proportion using techniques learned earlier:

$$\widehat{SE(\hat{p})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx \sqrt{\frac{0.88 * 0.12}{10701}} = 0.0031$$

Two big questions arise:

- First, can we provide a range of values we are 'confident' the true proportion falls in?

- Second, does our estimator $\hat{p}$ get closer to the true value p for larger sample sizes?

These two questions can often be answered by looking at the *limiting* behavior (here as the sample size grows) of the estimator $\hat{p}$.

$\xrightarrow{d}$ Idea

To answer the first question: "Can we provide a range of values for the parameter?", let's consider determining the *sampling distribution* through simulation. A distribution just describes the pattern in which we observe our variable. If we can simulate observing the variable, we can create many *realizations* of $\hat{p}$ to understand the *sampling distribution*.

To do this we need to make some assumptions. Namely:

- We have n *iid* (independent and identically distributed) trials
- A value for the true p

Let's use the app below to consider the sampling distribution when p is 0.9 and n is 10.

Instructions:

- "n: sample size" Slider: Move the slider to the right to increase the sample size
- "p: true value in population" Slider: Move the slider to the right to increase the true proportion
- "Generate a sample proportion": Click this button to add a single randomly generated sample proportion to the plot
- "Generate 100 sample proportions": Click this button to add 100 randomly generated sample proportions to the plot
- "Add +/- 2 Standard Error and Overlay Smoothed Density" Check this bot to add bars corresponding to two standard errors and also add a smoothed density overlayed

Put your thoughts from using the app here!

As long as the distribution is roughly normal, we can see that 0.95 of the distribution falls within two standard errors of p. For 95% of the $\hat{p}$ values we observe, adding and subtracting two standard errors would capture the true p. This means we could use something like

$$\hat{p} \pm 2 * \widehat{SE(\hat{p})}$$

as an interval to *capture* the true p. (Indeed this is the usual basic interval for a proportion!)

$\xrightarrow{p}$ Idea

To answer the second question: "does our estimator $\hat{p}$ get closer to the true value p for larger sample sizes?", we could consider generating sample proportions for ever increasing values of the sample size and seeing how they behave. Using the app below, we can generate many sample proportions for varying n, subtract off the true value of p, and see how that difference changes on the plot.

- Start with one sample proportion at each n and see the behavior of $\hat{p} - p$.
- Now increase the number of sample proportions generated for each n. What aspect of this relationship does this help us understand?
- Increase the sample size you are considering. What happens to the observed difference as n grows?

Instructions:

- "Maximum Sample Size": Enter a number to increase/decrease the largest sample size to consider
- "# samples at each n" Slider: Move the slider to select the number of sample proportions to generate at each given sample size

- "p: true value in population": Move this slider to select the true proportion from the population
- "Create/Update graph": Click this button to create the initial graph or update the graph based off of new selections of the above values

Put your thoughts from using the app here!

We can see that $\hat{p} - p$ seems to get closer to 0. This indicates that $\hat{p}$ is in some sense *converging* to the true value of p!

Now that we some basic intuition, let's formalize what we are talking about.

Definitions

By *limit*, *large-sample*, or *asymptotic* theory we mean we want to understand the behavior of some quantity, usually a *statistic*, as something changes, usually the sample size n. For instance, we will investigate the behavior of the sample mean, $\bar{Y}$, as the sample size grows. We'll look at questions like:

- When the distribution of a statistic (called a **sampling distribution**) is difficult to derive *exactly*, is there a good **approximating** distribution that can be used to get **approximate** probability statements about $\bar{Y}$?
- What value does $\bar{Y}$ *get close to* or *converge to* as the sample size grows?

Answers to these questions will allows us to do inference (confidence intervals and hypothesis tests) and understand the quality of our estimator.

Common Assumptions & Definitions

We often make some assumptions about how we observe our random variables in order to investigate these types of questions. For simplicity, we often assume we have a **random sample**.

Random Sample $Y_1, ..., Y_n$ are a random sample (RS) of size n if the random variables are independent and identically distributed (iid).

We'll often say 'assume we have a random sample' from some distribution or that 'our random variables are iid' from some distribution. These are equivalent ways of stating this assumption.

For the proportion example mentioned previously, we might formally state our assumption as follows:

- Define $X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ individual says the benefits of childhood vaccines for MMR outweigh the risks} \\ 0 & \text{if not} \end{cases}$
- Then $X_i \overset{iid}{\sim} Ber(p)$ where p represents the true proportion of people in the U.S. that believe the benefits outweight the risks
- The random variable $Y = \sum_{i=1}^{n} X_i \sim Bin(n, p)$
- We then often try to use Y or $\hat{p} = Y/n$ to make inference on p.
- Y and $\hat{p}$ are referred to as **statistics**. Note: $\hat{p}$ is also $\bar{X}$!

Statistic A function of $Y_1, Y_2, ..., Y_n$ from a random sample that does not involve any unknown parameters is called a statistic.

Commonly studied statistics:

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- $Y_{(n)}$ = the maximum value from the sample

Quantities that aren't statistics:

- $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$ (since μ is unknown - if we assume μ is known (like when we do in a hypothesis test) then this is a statistic)
- $\frac{(n-1)S^2}{\sigma^2}$ (since σ^2 is unknown)

One type of convergence we'll look at is focused on the pattern in which these statistics are observed, that is, the **sampling distribution** of the statistics.

Recall that a distribution is just the pattern and frequency with which we observe a random variable. With a statistic, we give this distribution the special name of **sampling distribution**. This is because we can think of how that distribution is formed by considering repeated samples from the population, each sample producing the statistic of interest.

Sampling Distribution The distribution of a statistic is called a sampling distribution.

We see that the sampling distribution of $\hat{p}$ looks like a bell curve for some combinations of n and p. If we fix a p and increase n, we will start to see a bell shape for large enough n! Later we'll see that a good **large-sample** distribution for $\hat{p}$ is the Normal distribution with mean p and variance $p(1-p)/n$.

We can see that there may be a distinction between the *actual* distribution, which is a discrete distribution for $\hat{p}$, and an approximating distribution, the Normal distribution for $\hat{p}$. We call these by different names.

Exact Distribution The (sampling) distribution of a quantity that is valid for any sample size (or, occasionally, values of the parameters of the population distribution).

Large-Sample or Approximate Distribution A (sampling) distribution that is reasonable to use for a quantity for a *large* sample size (or occasionally other parameter values).

We use the notation

$$Statistic \overset{\bullet}{\sim} f$$

to denote a large-sample approximating distribution.

In the sample proportion example, we would write

$$\hat{p} \overset{\bullet}{\sim} N(p, p(1-p)/n)$$

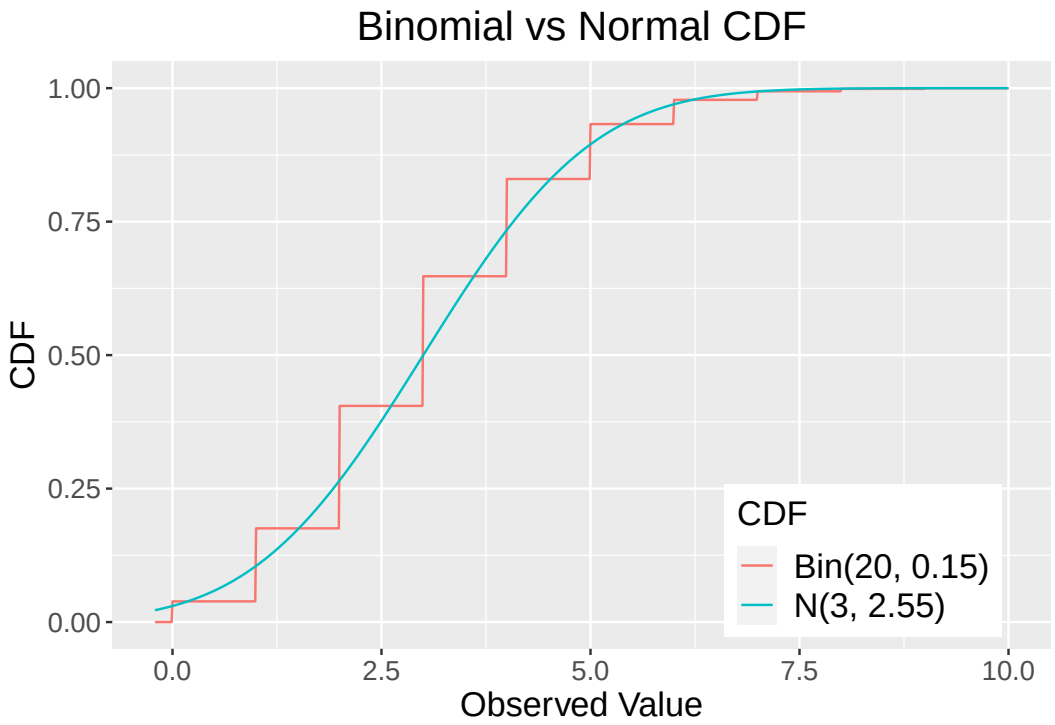
Convergence in Distribution

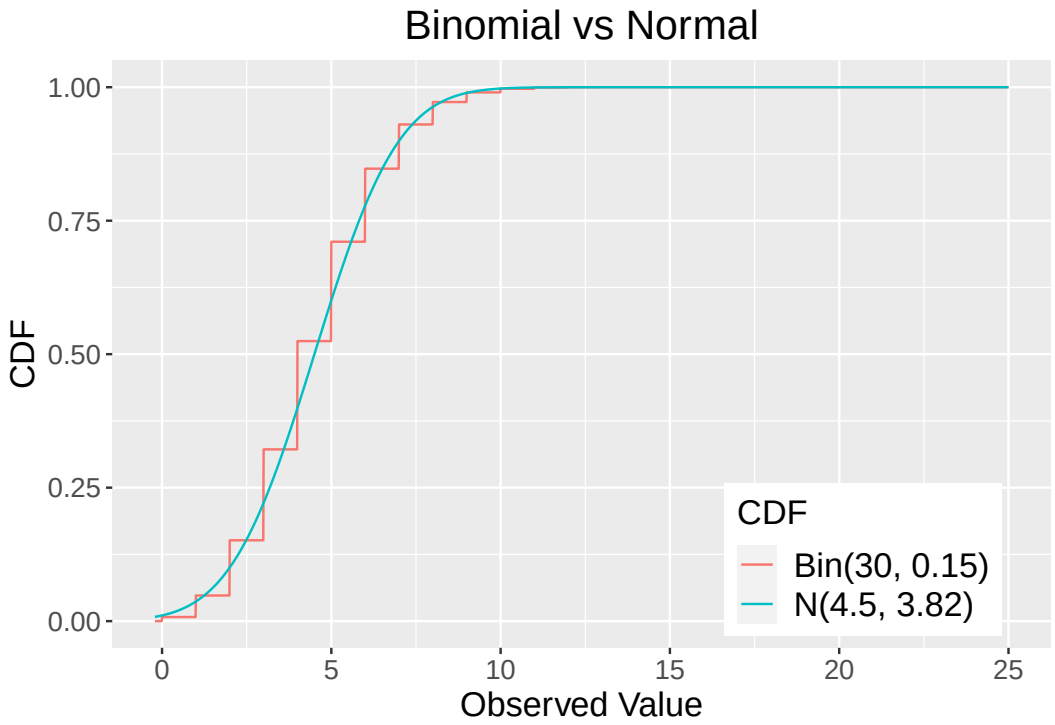
While convergence in distribution can be visually inspected with a histogram, to formally define **convergence in distribution** we use the cumulative distribution function or CDF.

Recall the Cumulative Distribution Function (or CDF) of a random variable Y is defined as

$$F_Y(y) = P(Y \leq y)$$

For our binomial example, we can compare the CDF of binomial random variables to Normal random variables to see that the Binomial is 'converging' to the normal distribution in a sense!





Convergence in Distribution Consider a sequence of random variables $Y_1, ..., Y_n, ...$ with corresponding CDFs $F_{Y_1}(y), ..., F_{Y_n}(y), ...$ Then Y_n converges in distribution to the random variable Y (with CDF $F_Y(y)$) if

$$\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$$

or equivalently

$$\lim_{n\to\infty} |F_{Y_n}(y) - F_Y(y)| = 0$$

(at all points y where $F_Y(y)$ is continuous). We denote this as

$$Y_n \xrightarrow{d} Y$$

The subscript n notation here may be confusing. This is just to show the RV on the left is dependent on the sample size in some way. For our example with a Binomial/sample proportion, the distribution clearly depends on n. We could write the following to be explicit:

$$Y_n \sim Bin(n, p) \text{ and } \hat{p}_n = Y_n/n$$

We'll prove (via the CLT) that the standardized version of these statistics converge to a standard Normal distribution! For example,

$$Z_n = \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} Z \sim N(0,1)$$

Alternatively, for practical purposes we'll equivalently talk about 'large-sample' distributions using the $\overset{\bullet}{\sim}$ notation:

$$\hat{p}_n \overset{\bullet}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

It is sometimes easier to work with MGFs rather than CDFs. In that case, we can use the following result:

Convergence of MGFs Consider a sequence of random variables $Y_1, ..., Y_n, ...$ with corresponding MGFs $m_{Y_1}(t), ..., m_{Y_n}(t), ...$ Then Y_n converges in distribution to the random variable Y (with MGF $m_Y(t)$) if

$$\lim_{n\to\infty} m_{Y_n}(t) = m_Y(t)$$

Proving $\xrightarrow{d}$ using CDFs

Example: Suppose that $Y_i \overset{iid}{\sim} U(0,1)$. That is,

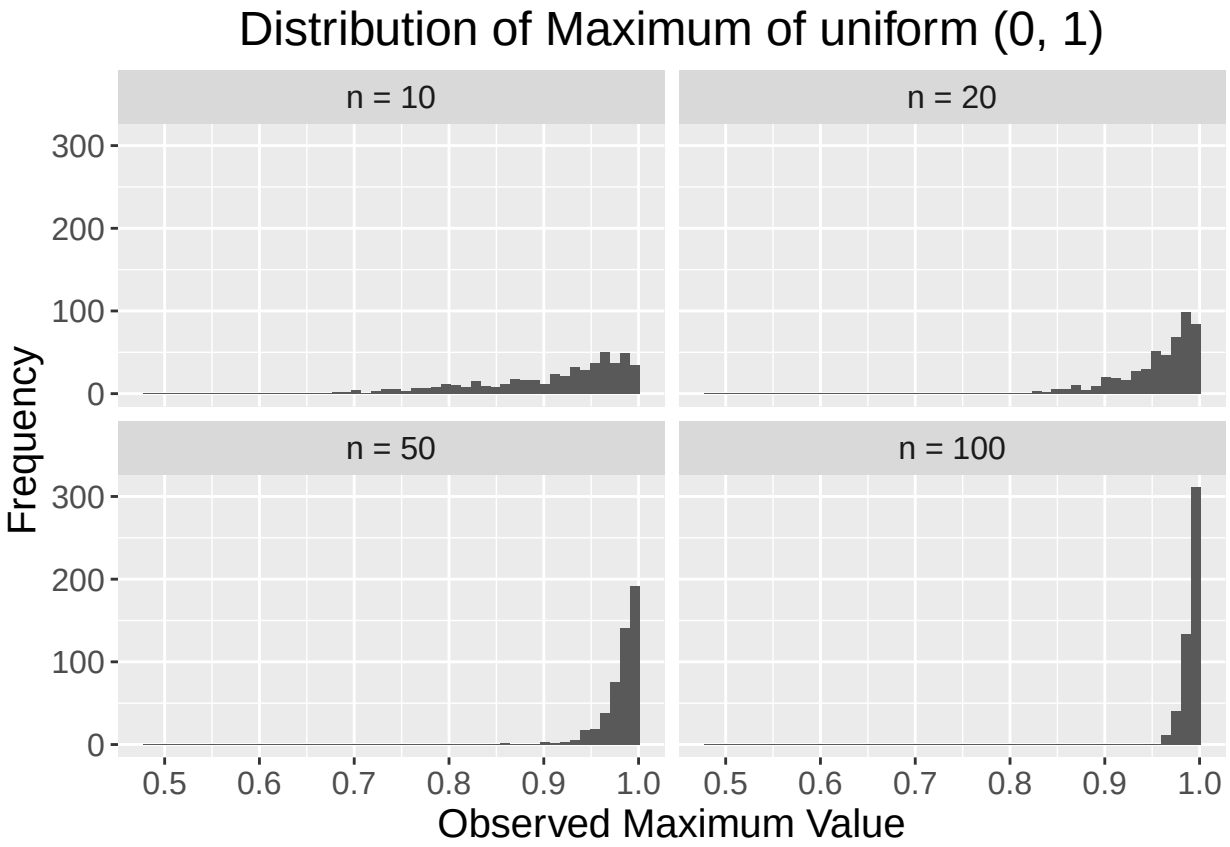
$$f_Y(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & otherwise \end{cases}$$

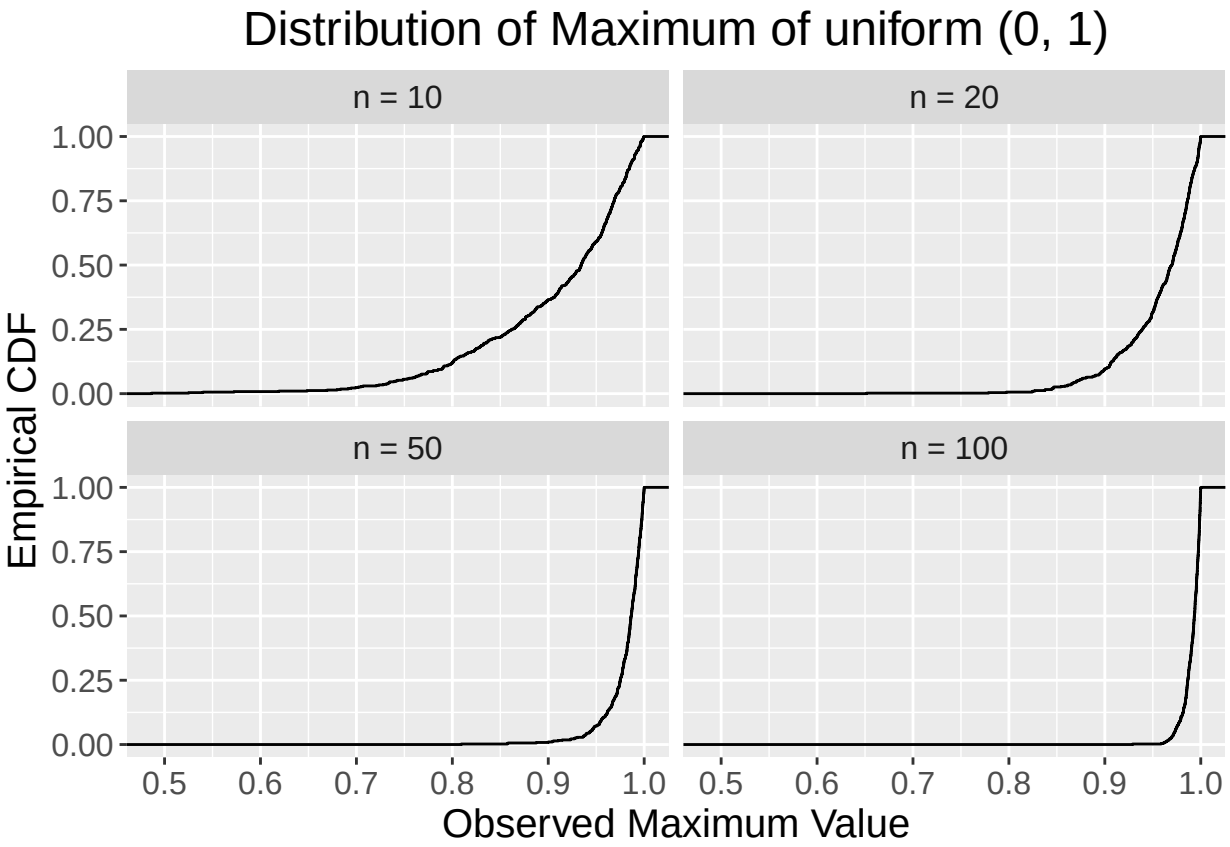
and

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ y & 0 \le y < 1 \\ 1 & y \ge 1 \end{cases}$$

What does the maximum from the sample converge to in distribution as n grows?

Let's generate many samples, find the maximum for each sample, and look at the empirical distribution via a histogram and CDF.


Distribution of Maximum of uniform (0, 1)
n = 10
n = 20
n = 50
n = 100
Frequency
300
200
100
0
0.5
0.6
0.7
0.8
0.9
1.0
Observed Maximum Value



It appears that the distribution converges to a random variable that always takes on 1. We'd say there is a **point mass** at 1. If W is a random variable that always takes on the constant c then

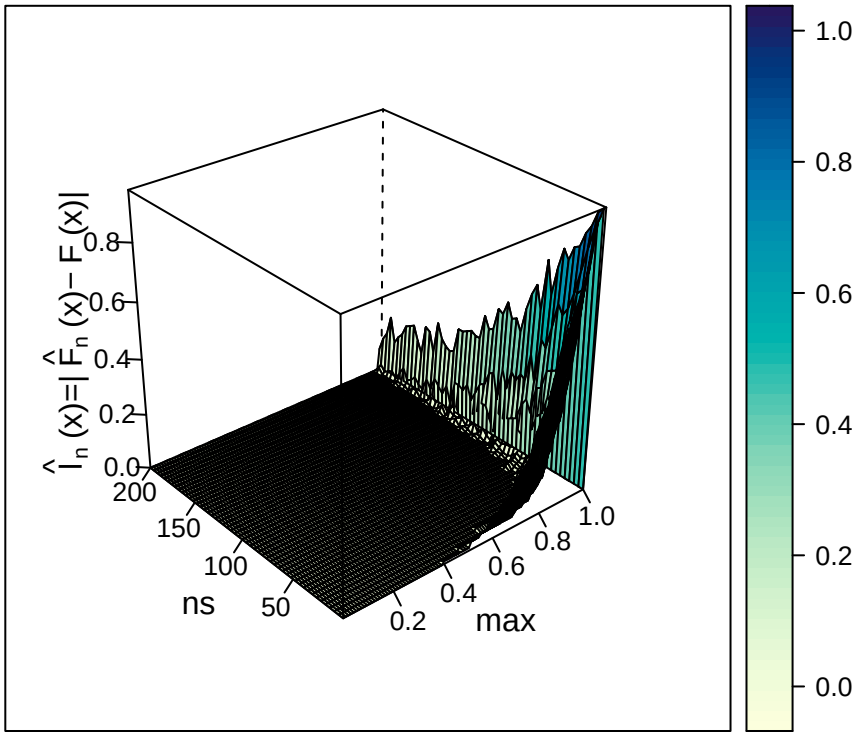
$$f_W(w) = \begin{cases} 1 & w = c \\ 0 & otherwise \end{cases}$$

$$F_W(w) = \begin{cases} 0 & w < c \\ 1 & w \geq c \end{cases}$$

We also said we could look at

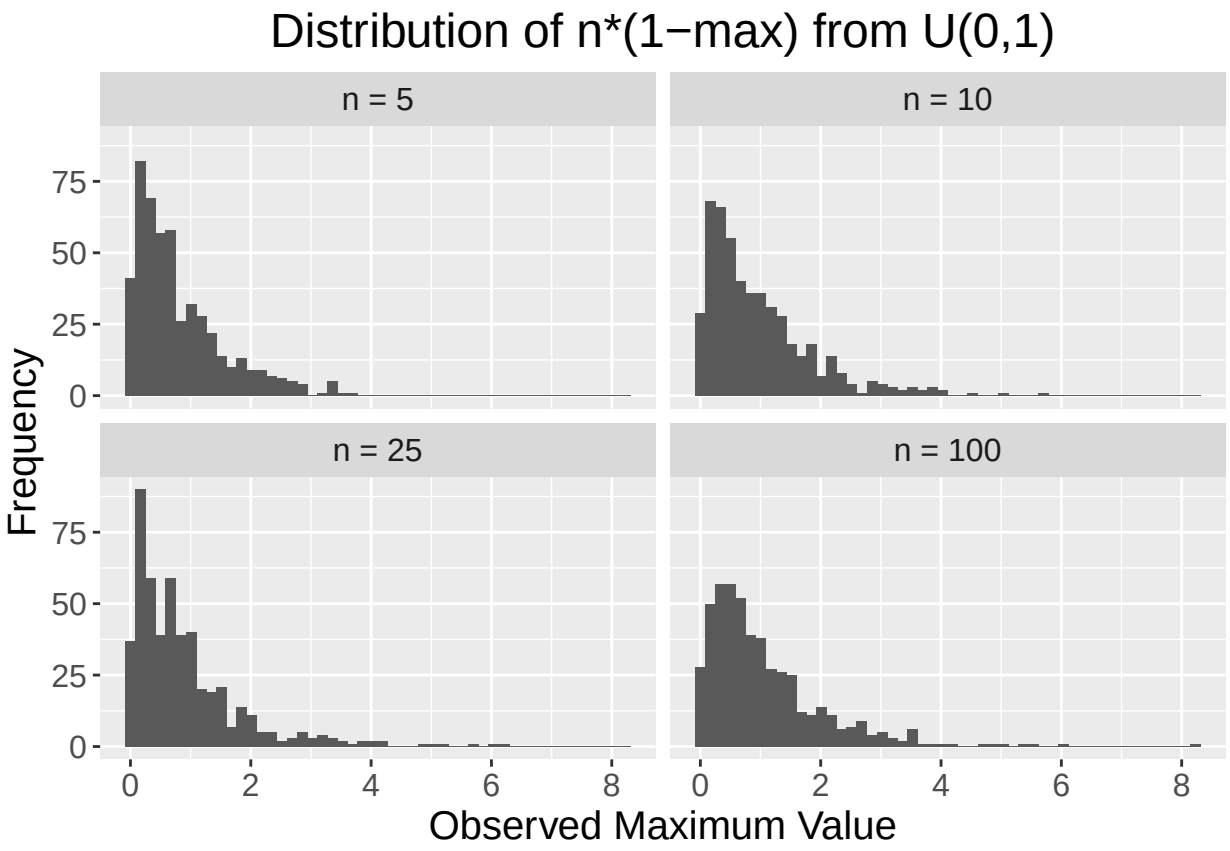
$$\lim_{n \to \infty} |F_{Y_n}(y) - F_Y(y)| = 0$$

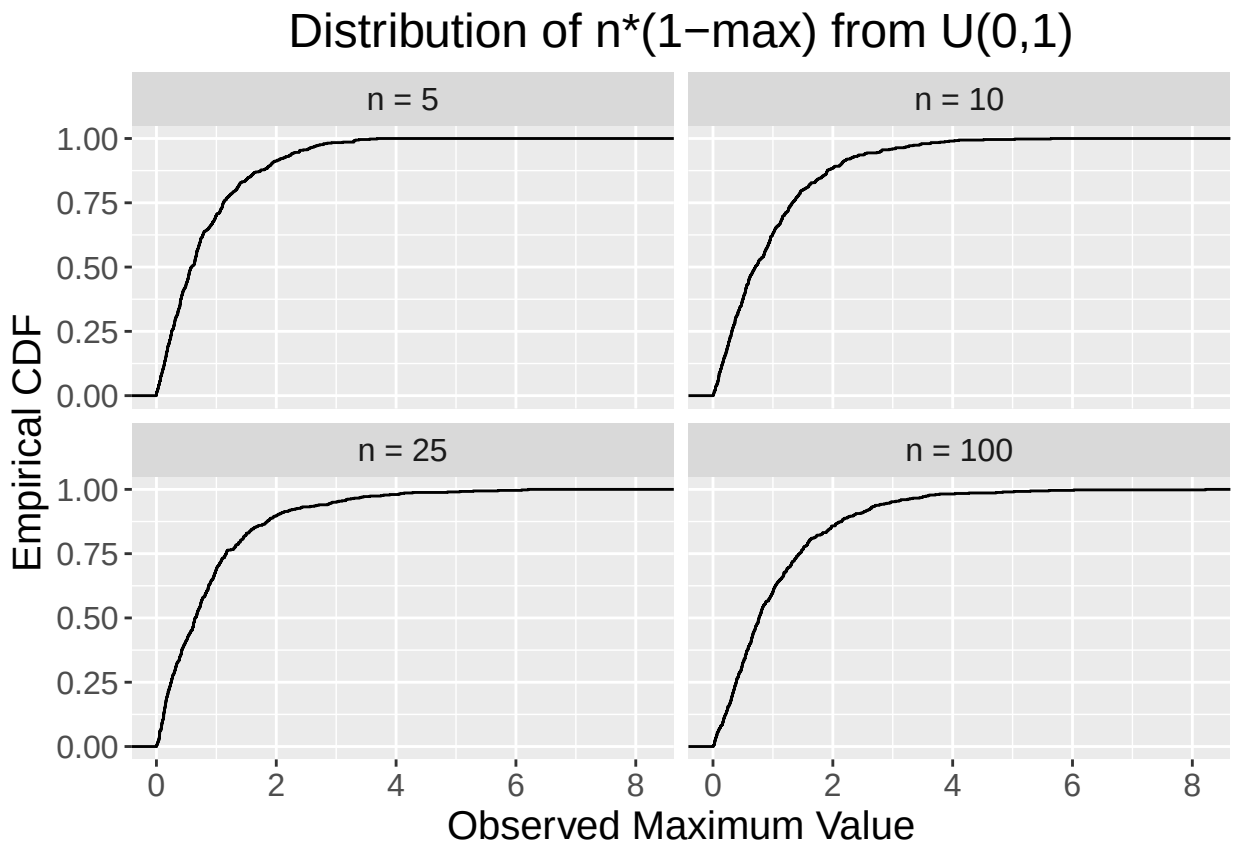
for convergence in distribution. This difference in CDFs can be plotted in three dimensions.



Let's formally prove that the maximum of n iid $U(0,1)$ RVs converges to a RV with a point mass at 1.

Example: Consider again a random sample of $U(0,1)$ RVs. What does $W = n(1 - Y_{(n)})$ converge in distribution to as n grows? Can we describe a rule of thumb for when the approximating distribution is reasonable?

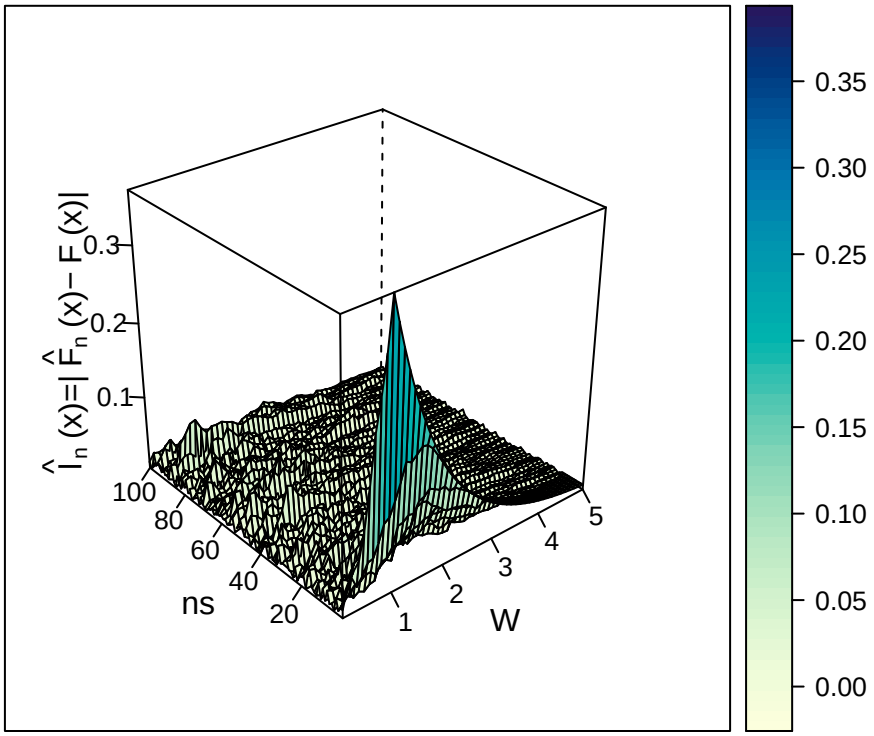

Distribution of n*(1−max) from U(0,1)
n = 5
n = 10
n = 25
n = 100
Frequency
0
25
50
75
0
2
4
6
8
Observed Maximum Value



This one doesn't appear to be converging to a point mass. Let's use the limit of the CDF to determine what this random variable converges to.

Let's compare the distribution of $W = n(1 - Y_{(n)})$ with $X \sim exp(1)$ via a plot of

$$|F_{Y_n}(y) - F_Y(y)|$$



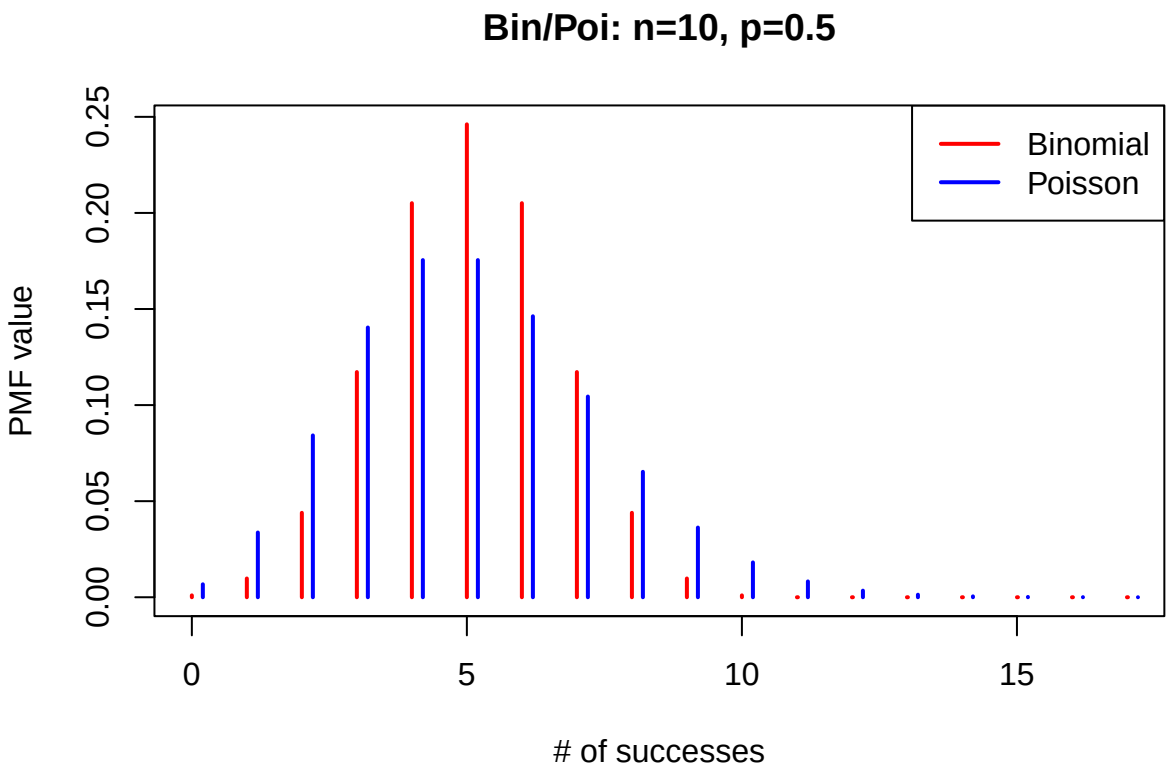
Proving $\xrightarrow{d}$ using MGFs

Example: Suppose $Y \sim Bin(n, p)$ where $np \to \lambda$ as n grows. Show $Y \xrightarrow{d} Poi(\lambda)$.

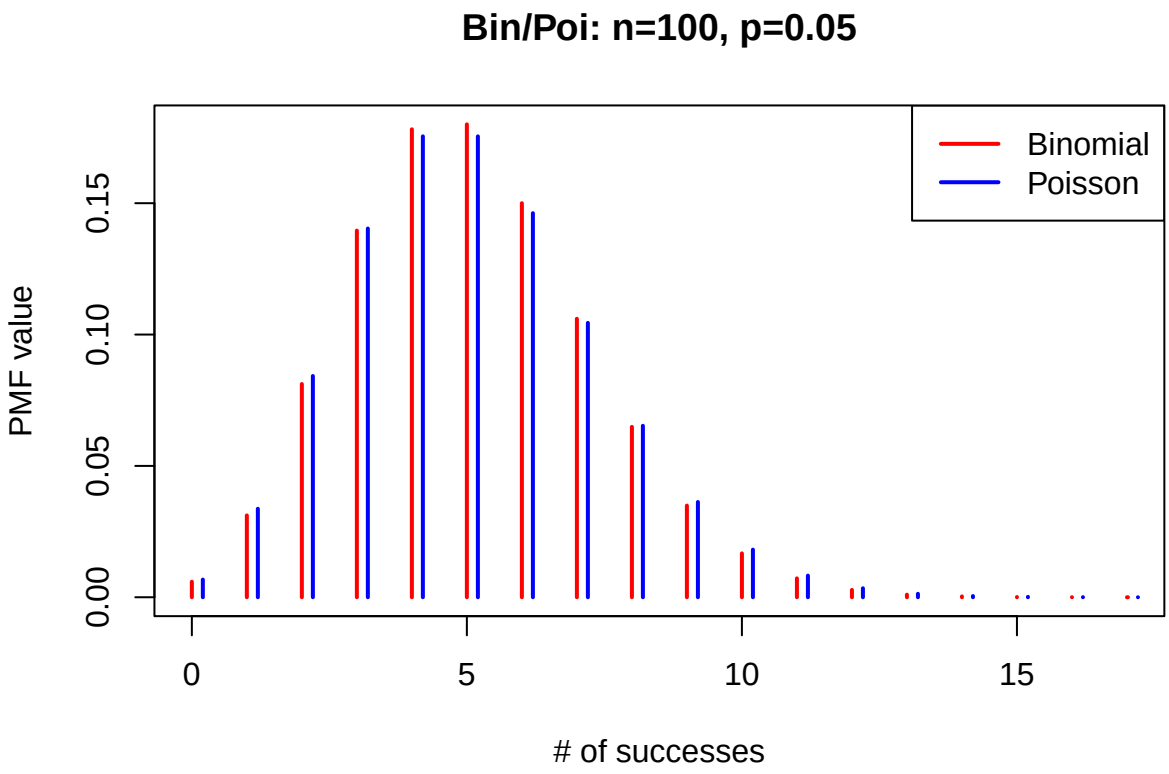
First, let's compare plots to see that the relationship seems to hold. We'll create three different binomial and poisson plots with the same ratio for n and p.

Consider how well do the PMFs match up for the following situations:

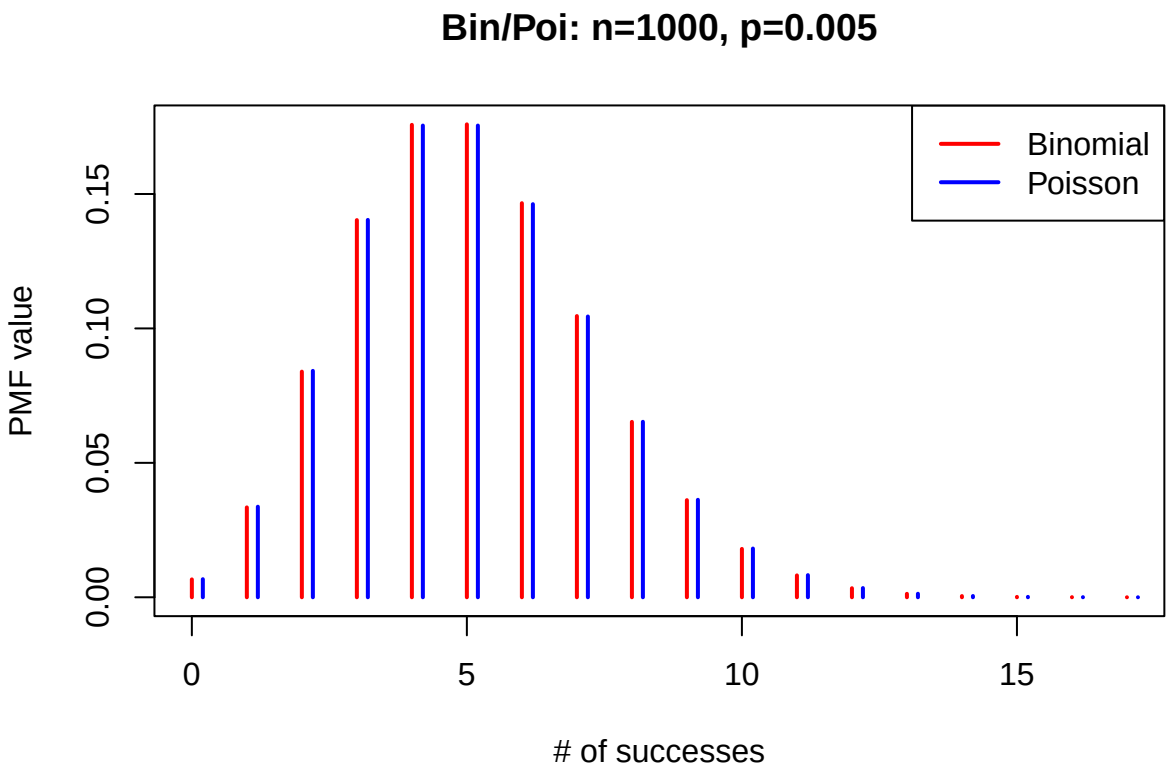
- $n = 10, p = 0.5 \to n * p = 5$



- $n = 100, p = 0.05 \rightarrow n * p = 5$



- $n = 1000, p = 0.005 \rightarrow n * p = 5$



Example: After we learn about the central limit theorem (CLT), we'll see a (relatively) easy way to prove that a $Y \sim Gamma(n, \lambda)$, properly standardized, converges to a standard normal distribution. That is,

$$W = \frac{Y - n/\lambda}{\sqrt{n}/\lambda} \xrightarrow{d} Z \sim N(0,1)$$

We can prove it using MGFs directly. Let's look through the proof below:

Goal: Start with the MGF of the standardized random variable and try to show it converges to a standard normal MGF ($e^{t^2/2}$) as $n \to \infty$.

$$\begin{aligned} m_Z(t) &= E\left(e^{t\left(\frac{Y-n/\lambda}{\sqrt{n}/\lambda}\right)}\right) \\ &= E\left(e^{\frac{t\lambda}{\sqrt{n}}Y}\right) e^{-t\sqrt{n}} \\ &= \left(\frac{1}{1-\frac{(\lambda t)/\sqrt{n}}{\lambda}}\right)^n e^{-t\sqrt{n}} \\ &= \left(\frac{e^{-t/\sqrt{n}}}{1-t/\sqrt{n}}\right)^n \end{aligned}$$

As we want the limit of this quantity as n goes to infinity, consider that this involves n in the term and is raised to the n. We saw the result:

$$\lim_{n\to\infty}(1 + a_n/n)^n = e^a$$

where $\lim_{n\to\infty} a_n = a$. A rewrite can allow us to use this!

$$= \left(1 + \frac{n\left(\frac{e^{-t/\sqrt{n}}}{1-t/\sqrt{n}} - 1\right)}{n}\right)^n$$

Now we can just consider what happens to the numerator of the second term as n grows. That is, we just need to consider

$$\lim_{n\to\infty} n\left(\frac{e^{-t/\sqrt{n}}}{1-t/\sqrt{n}} - 1\right)$$

Using a common denominator and then applying a Taylor series expansion of the e term about 0,

$$e^{-t/\sqrt{n}} = 1 - t/\sqrt{n} + t^2/(2n) - t^3/(3!n^{3/2}) + ...,$$

we can rewrite this as

$$\begin{aligned}
&= \lim_{n\to\infty} n\left(\frac{t^2/(2n) - t^3/(3!n^{3/2}) + ...}{1-t/\sqrt{n}}\right) \\
&= \lim_{n\to\infty} \left(\frac{t^2/2 - t^3/(3!n^{1/2}) + ...}{1-t/\sqrt{n}}\right) \\
&= t^2/2
\end{aligned}$$

Thus, our MGF converges $e^{t^2/2}$. This is the MGF of a standard normal random variable! Therefore,

$$W = \frac{Y - n/\lambda}{\sqrt{n/\lambda^2}} \xrightarrow{d} Z \sim N(0,1)$$

Central Limit Theorem

One of the most important theorems in statistics is the Central Limit Theorem (CLT). The CLT gives us a general result about the large-sample behavior of a sample mean.

Central Limit Theorem (CLT) Suppose that $Y_i \overset{iid}{\sim} f_Y$ where $E(Y) = \mu$ and $Var(Y) = \sigma^2 < \infty$. Define $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $Z \sim N(0,1)$. Then the standardized sample mean converges in distribution to a standard normal random variable.

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$$

Practically, we can say that a good approximating distribution or large-sample distribution for $\bar{Y}$ is

$$\bar{Y} \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$$

CLT Applied to a Sample Proportion **Example:** A common application of the CLT is to the sample proportion from a Binomial experiment.

For example, if $X_i \overset{iid}{\sim} Bin(1, p)$ with mean $E(Y) = p$ and variance $Var(Y) = p(1-p)$.

Define $Y = \sum_{i=1}^{n} X_i$. We know $Y \sim Bin(n, p)$. The sample proportion is then

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

By the CLT, we can say that the sampling distribution of $\hat{p}$ can be well approximated by a Normal distribution with mean $E(X_i) = p$ and variance $Var(X_i)/n = p(1-p)/n$.

We might state this as either

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} Z \sim N(0, 1)$$

or

$$\hat{p} \overset{\bullet}{\sim} N(p, p(1-p)/n)$$

You've likely seen the Normal approximation to the Binomial before. You may even know a rule of thumb for using it. The app below simulates sample proportions from a given binomial distribution. Use the app below to

- Explore the relationship between $\hat{p}$ and the corresponding Normal distribution (the larger the M value the more precisely the graph mimics the exact distribution of $\hat{p}$)
- Either verify the rule of thumb you know or try and come up with a rule of thumb for when the approximation is reasonable

Put your thoughts from using the app here!

CLT Applied to a Sum Recall the result:

- If $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$

That is, a Normal random variable multiplied by a constant is still Normally distributed, just with a different mean and variance. - This tells us that if the CLT is applicable to the sample average, then we can also apply it to the corresponding summation as well!

Under the same assumptions as the CLT, since $n\bar{Y} = \sum_{i=1}^{n} Y_i$ we have the following result:

$$\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sqrt{n}\sigma} \overset{\bullet}{\sim} N(0, 1)$$

or

$$\sum_{i=1}^{n} Y_i \overset{\bullet}{\sim} N(n\mu, n\sigma^2)$$

Example: Based on this result, what is a good large-sample distribution for $Y \sim Bin(n, p)$?

CLT Applied to a Gamma Example: Earlier we proved that $Y \sim Gamma(n, \lambda)$, properly standardized, converges to a standard normal distribution. That is,

$$W = \frac{Y - n/\lambda}{\sqrt{n}/\lambda} \xrightarrow{d} Z \sim N(0,1)$$

Rather than use MGFs, we can apply the CLT in a clever way!

First, note that for $Y \sim gamma(n, \lambda)$ we can think of Y as $Y = X_1 + X_2 + ... + X_n$, where $X_i \overset{iid}{\sim} gamma(1, \lambda)$. Here we know that $E(X_i) = 1/\lambda$ and $Var(X_i) = 1/\lambda^2$. By the CLT applied to a sum we know

$$\frac{\sum_{i=1}^{n} X_i - n(1/\lambda)}{\sqrt{n}(1/\lambda)} = \frac{Y - n/\lambda}{\sqrt{n/\lambda^2}} \overset{\bullet}{\sim} Z$$

where $Z \sim N(0,1)$. That means we can approximate a gamma random variable by

$$Y \overset{\bullet}{\sim} N(n/\lambda, n/\lambda^2)$$

Alternatively, we could apply the CLT to an average instead of a sum. Instead use $X_i \overset{iid}{\sim} gamma(1, \lambda/n)$ then Y can be thought of as an average of these X's

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \sim gamma(n, \lambda)$$

We know that $E(X_i) = n/\lambda$ and $Var(X_i) = n^2/lambda^2$. Since we are now viewing this as an average of iid random variables we can apply the CLT.

$$\frac{\bar{X} - n\lambda}{\sqrt{(n^2/\lambda^2)/n}} = \frac{\bar{X} - n\lambda}{\sqrt{n/\lambda^2}} \xrightarrow{d} Z$$

where $Z \sim N(0,1)$.

Convergence Exploration

- Suppose that $Y \sim Gamma(n, \lambda)$. Or, assume that $Y = \sum_{i=1}^{n} X_i$ where $X_i \overset{iid}{\sim} Gamma(1, \lambda)$.

  $$f_Y(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}$$

  with mean $E(Y) = \alpha/\lambda$ and variance $Var(Y) = \alpha/\lambda^2$.
- We again showed we can approximate a gamma by a Normal distribution.

- Can you develop a rule of thumb around α (n here) and λ for when a Normal distribution may be a reasonable approximation? Remember we look for the following in each graph:
  - Histogram: when does it become a symmetric bell-shape?
  - CDF comparison: When do the CDFs essentially overlap?

Put your thoughts from using the app here!

CLT Importance

Practically, why is the CLT so important?

- The CLT gives us a distribution we can use to find probabilities when we deal with most sample sums and sample means.
- Knowing a large-sample distribution allows us to find (approximate) probabilities when exact probabilities may be too difficult to find.
- This means we can do approximate inference in many cases!

Example:

- Suppose we know σ and we want inference for μ.
- If we have a random sample $Y_1, ..., Y_n$, we know $\bar{Y} \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$ (μ only unknown)
- We can make an approximate claim about μ via a confidence interval derived from an argument similar to that below:

$$\begin{aligned} P(-1.96 < Z < 1.96) &= 0.95 \\ \Leftrightarrow P\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right) &= 0.95 \\ \Leftrightarrow P\left(\bar{Y} - 1.96\sigma/\sqrt{n} < \mu < \bar{Y} + 1.96\sigma/\sqrt{n}\right) &= 0.95 \end{aligned}$$

- That is, there is a 95% probability the RVs $\bar{Y} - 1.96\sigma/\sqrt{n}$ and $\bar{Y} + 1.96\sigma/\sqrt{n}$ capture μ!
- In practice we observe a value for $\bar{Y}$ as $\bar{y}$. We then lose the ability to talk about probability but instead asy we are 95% confident the observed interval contains μ.
- Note: No assumption about Y's distribution made other than finite variance!

Convergence in Probability

Definition

We saw that the estimator of p, $\hat{p}$, from the Binomial example seemed to be observed closer and closer to p for larger sample sizes. Additionally, we saw a good large-sample distribution for $\hat{p}$ is

$$\hat{p} \overset{\bullet}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Does this large-sample distribution support the 'convergence' of $\hat{p}$ to p idea?

More formally, we're going to take on the idea of **convergence in probability to a constant**. First, let's define convergence in probability generally.

Convergence in Probability A sequence of RVs $Y_1, ..., Y_n, ...$ converges in probability to a RV Y if for every $\epsilon > 0$

$$\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0 \iff \lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$$

This is denoted as

$$Y_n \xrightarrow{p} Y$$

We'll mostly care about convergence in probability to a constant, call it c. We can see the definition in this case can be simplied to the following:

$$\lim_{n\to\infty} P(|Y_n - c| < \epsilon) = \lim_{n\to\infty} P(-\epsilon < Y_n - c < \epsilon) = \lim_{n\to\infty} P(c - \epsilon < Y_n < c + \epsilon) = 1$$

$Y_n \xrightarrow{p} c$ if the *probability* we observe Y_n close to c goes to 1 in the limit.

Example - We can visualize this idea.

Assume that $Y_i \overset{iid}{\sim} N(0,1)$. Let's investigate the behavior of

$$X = \frac{1}{n^2}\sum_{i=1}^{n} Y_i$$

To put this in the context of the definition, let's refer to X explicitly as a function of n:

$$X_n = \frac{1}{n^2}\sum_{i=1}^{n} Y_i$$

We want to understand the behavior of X_n as n grows. We'll see that $X_n \xrightarrow{p} 0$, which implies that for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} P(-\epsilon < X_n < \epsilon) = 0$$

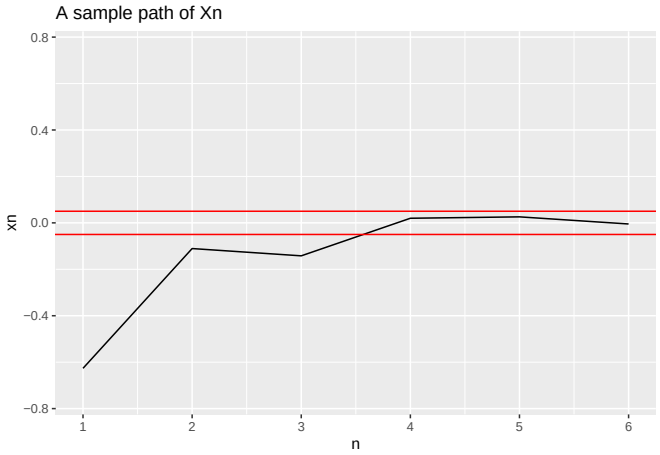
To visualize this, we can consider **sample paths** of X_n. That is, we can look at a particular sequence of y_i's that will generate a sequence of x and see how the values change.

Consider the following 6 values randomly sampled from a $N(0,1)$ and the corresponding sequence of x_n values.

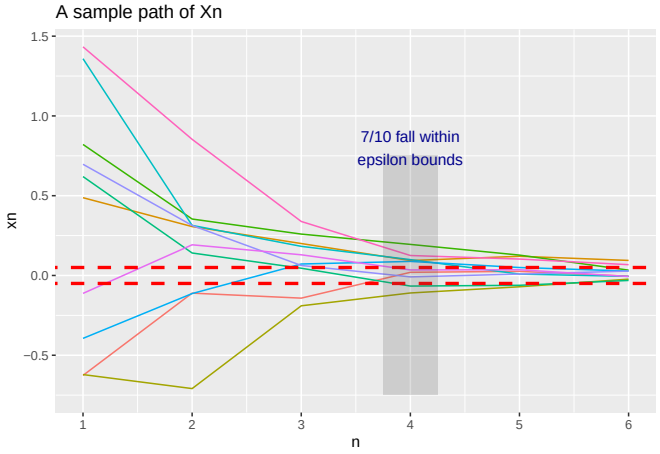
y sequence	x sequence
y_1 = -0.6264538	x_1 = -0.6264538/1^2 = -0.6264538
y_2 = 0.1836433	x_2 = (-0.6264538+0.1836433)/2^2 = -0.1107026
y_3 = -0.8356286	x_3 = (-0.6264538+0.1836433+-0.8356286)/3^2 = -0.1420488
y_4 = 1.5952808	x_4 = (-0.6264538+...+1.5952808)/4^2 = 0.0198026
y_5 = 0.3295078	x_5 = (-0.6264538+...+0.3295078)/5^2 = 0.025854
y_6 = -0.8204684	x_6 = (-0.6264538+...+-0.8204684)/6^2 = -0.0048366

If we consider multiple sample paths, then convergence in probability to 0 of this sequence implies that the proportion of sample paths outside of $\pm\epsilon$ should go to zero.

Let's plot our sample path with an $\epsilon = 0.05$:



Now let's add 9 more sample paths:



What we hope to see is that the proportion of lines falling outside of the ϵ bars goes to 0!

Example - Suppose we have a random sample from a Normal distribution with mean 10 and standard deviation 1. What do you think $W = (\bar{Y})^2$ converges to in probability? Take an educated guess and use the app below to explore!

- Select the value c that you guess W converges to in probability.
- Choose a sample size to go up to (start smaller and then get larger once you have a good idea).
- Select an ϵ range.
- Look for the proportion of lines (50 sample paths are generated) falling outside of the ϵ bars to go to 0!

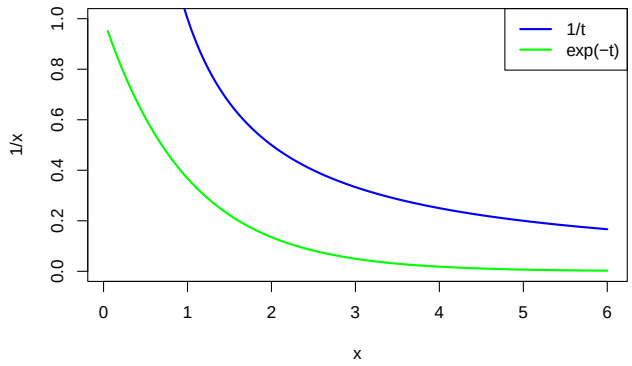
Inequalities

To prove convergence in probability, we'll sometimes rely on some very famous inequalities. These will help us to show the probability goes to 0 or 1.

Markov's Inequality If X is a nonnegative RV (support has no negative values) for which $E(X)$ exists, then for $t > 0$

$$P(X \geq t) \leq \frac{E(X)}{t}$$

Example: If $X \sim exp(1)$ then $P(X \geq t) = e^{-t}$ and $E(X)/t = 1/t$.



Chebychev's Inequality Let X be a RV with mean $= \mu$ and variance $= \sigma^2$, then for $t > 0$

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Example: If $t = \sigma k$ for $k > 0$, we can apply Chebychev's to get

$$P\left(|X - \mu| \geq k\sigma\right) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

For $k = 2$ we have $P(|X - \mu| \geq 2\sigma) \leq 1/4$.

Practically, what can we take home from this?

- At least 75% of a RVs distribution lies within 2 standard deviations of the mean (if these moments exist)
- Regardless of distribution! (if moments exist)
- If $X \sim N(\mu, \sigma^2)$ we know $P(|X - \mu| \geq 2\sigma) \approx 0.05$. The bound isn't always very tight!

WLLN

One of the most important results regarding convergence in probability is called the Law of Large Numbers (LLN).

(Weak) Law of Large Numbers (WLLN) Suppose $Y_i \overset{iid}{\sim} f$ where the mean and variance of Y_i exist. Then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} E(Y) = \mu$$

- Big picture goal is to estimate parameters such as μ
- If we get a RS we know that $\bar{Y}$ will be a 'close' to μ for 'large' samples
- Applies to the **average of any independent random variables with the same finite mean**

Note that the variance assumption is actually not needed but will help us facilitate an easy proof. Let's use our inequalities to prove this result!

Example - Let $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$. What does $\bar{Y}$ converge to? What does $\frac{1}{n}\sum_{i=1}^{n} Y_i^2$ converge to?

Continuity Theorems

The WLLN is also quite useful when combined with the continuity theorem.

Continuity Theorem If $Y_1, Y_2, Y_3, ...$ converge to Y and $g()$ is a continuous function then $g(Y_1), g(Y_2), g(Y_3)...$ converge to $g(Y)$.

Example (exploration example proved) - Suppose we have a random sample from a Normal distribution with mean 10 and standard deviation 1. Consider $W = (\bar{Y})^2$. What does this converge to in probability?

Note: The continuity theorem also works for convergence in distribution!

Example - Suppose that $Y_i \overset{iid}{\sim} gamma(\alpha, \lambda)$. We have that

$$\frac{\bar{Y} - \alpha/\lambda}{\frac{\sqrt{\alpha}}{\lambda\sqrt{n}}} \xrightarrow{d} Z$$

where $Z \sim N(0,1)$. By the continuity theorem we have that

$$\left(\frac{\bar{Y} - \alpha/\lambda}{\frac{\sqrt{\alpha}}{\lambda\sqrt{n}}}\right)^2 \xrightarrow{d} Z^2$$

and recall that a standard Normal squared is distributed as a χ^2_1 or a $gamma(1/2, 1/2)$.

Other Standard Limit Results Work Too!\ Most of the common limit theorem ideas from calculus follow here as well (θ and λ are constants below):

$$\text{If } Y \xrightarrow{p} \theta, X \xrightarrow{p} \lambda \text{ then } Y \pm X \xrightarrow{p} \theta \pm \lambda$$

Example - Consider the 'biased' version of the sample variance, $S_n^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$. Let's show $S_n^2 \xrightarrow{p} \sigma^2$

$\xrightarrow{d}$ & $\xrightarrow{p}$ Relationship

Convergence in probability implies convergence in distribution. However, the converse is not true generally (**convergence in distribution does not imply convergence in probability**).

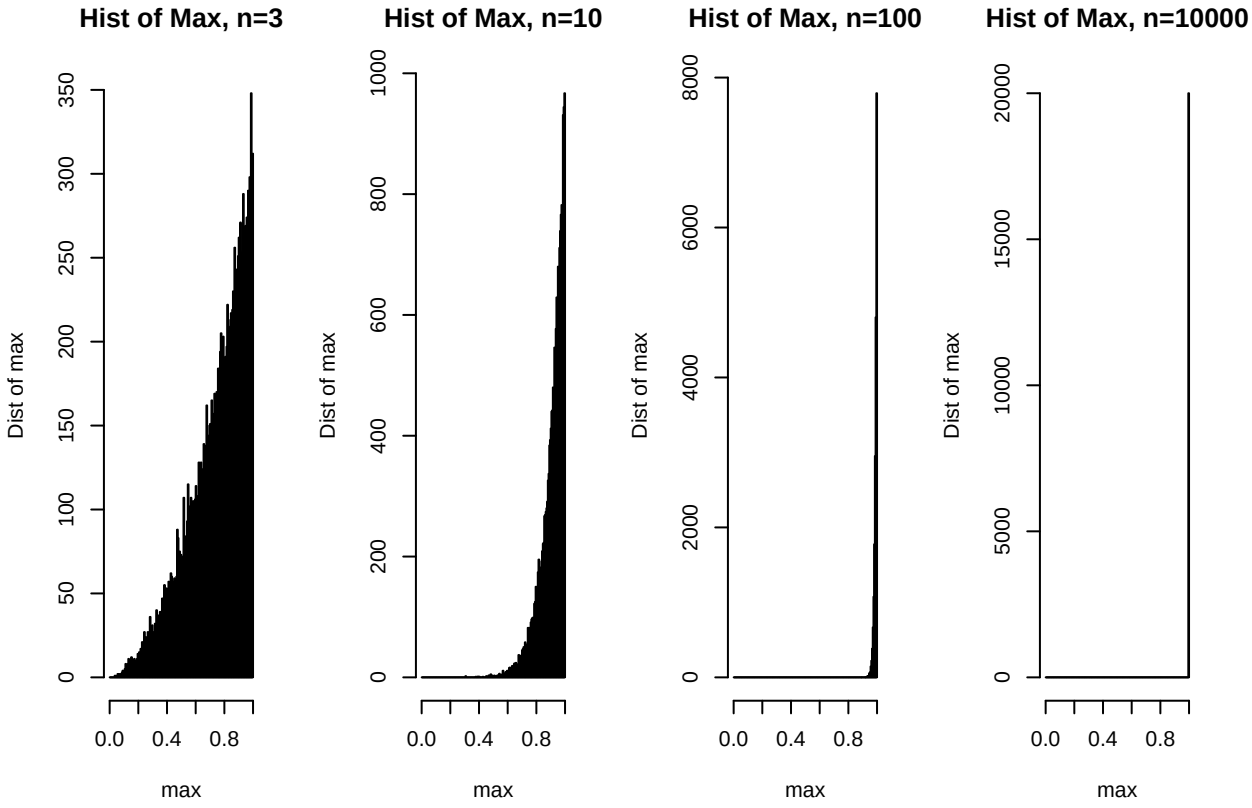
Example - Suppose $X \sim Beta(2,2)$ then $1-X$ is also distributed as Beta(2,2) (recall the symmetry of the Beta distribution with equal α and β).

Define a sequence of RVs to be $X_n = X$ for all n. Then $X_n \xrightarrow{d} 1 - X \sim Beta(2,2)$.

Now consider convergence in probability, does $X_n \xrightarrow{p} 1 - X$?

Convergence in distribution to a constant - If $Y_n \xrightarrow{d} c$ then $Y_n \xrightarrow{p} c$.

Why does it makes sense that convergence in distribution to a constant implies convergence in probability to that constant? Consider our example where we look at the maximum from a random sample of $U(0,1)$ RVs. Below are plots of the distribution of the sample max for varying n values.



Another really useful theorem relating convergence results is called Slutsky's Theorem.

Slutsky's Theorem If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$, then

- $X_n Y_n \xrightarrow{d} aX$
- $X_n + Y_n \xrightarrow{d} X + a$

Slutsky's theorem is extremely useful for creating hypothesis tests and confidence intervals! Recall the example we talked about when discussing the importance of the CLT:

Example:

- Suppose we know σ and we want inference for μ.
- If we have a random sample $Y_1, ..., Y_n$, we know $\bar{Y} \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$ (μ only unknown)
- We can make an approximate claim about μ via a confidence interval derived from an argument similar to that below:

$$\begin{aligned} P(-1.96 < Z < 1.96) &= 0.95 \\ \Leftrightarrow P\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right) &= 0.95 \\ \Leftrightarrow P\left(\bar{Y} - 1.96\sigma/\sqrt{n} < \mu < \bar{Y} + 1.96\sigma/\sqrt{n}\right) &= 0.95 \end{aligned}$$

- That is, there is a 95% probability the RVs $\bar{Y} - 1.96\sigma/\sqrt{n}$ and $\bar{Y} + 1.96\sigma/\sqrt{n}$ capture μ!

Of course, σ won't be known. Slutsky's theorem allows us to substitute a 'consistent' estimator of σ (i.e. an estimator of σ^2 that converges in probability to σ) and obtain a similar result!

Delta Method

A common place where we'd use the CLT, LLN, and Slutsky's theorem together is when looking at **Delta Method Normality**.

Large Sample Normality and the Delta Method Let $Y_1, Y_2, ...$ be a sequence of RVs such that

$$\sqrt{n}(Y_n - \theta_0) \xrightarrow{d} N(0, \sigma^2) \qquad or \qquad Y_n \overset{\bullet}{\sim} N(\theta_0, \sigma^2/n)$$

For a function g and value θ_0 where $g^{'}(\theta_0)$ exists and is not 0 we have

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \xrightarrow{d} N(0, (g^{'}(\theta_0))^2\sigma^2) \qquad or \qquad g(Y_n) \overset{\bullet}{\sim} N(g(\theta_0), (g^{'}(\theta_0))^2\sigma^2/n)$$

Example - Suppose $Y_i \overset{iid}{\sim} gamma(\alpha, \lambda)$. Goal: make inference on $\frac{1}{\mu}$. Provide an approximate distribution for $1/\bar{Y}$ an **estimator** of $1/\mu$.

Example - Let $Y_i \overset{iid}{\sim} Ber(p)$ then $\bar{Y} \overset{\bullet}{\sim} N(p, \frac{p(1-p)}{n})$. Goal: make inference for $\frac{p}{1-p}$ using $\frac{\bar{Y}}{1-\bar{Y}}$.

Example - Suppose $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$ where $E(Y_i) = \mu \neq 0$. Goal: make inference on $\frac{1}{\mu}$. Provide an approximate distribution for $1/\bar{Y}$ an **estimator** of $1/\mu$.

Recap

We have two big ideas:

- convergence in distribution
- convergence in probability

There are two big theorems:

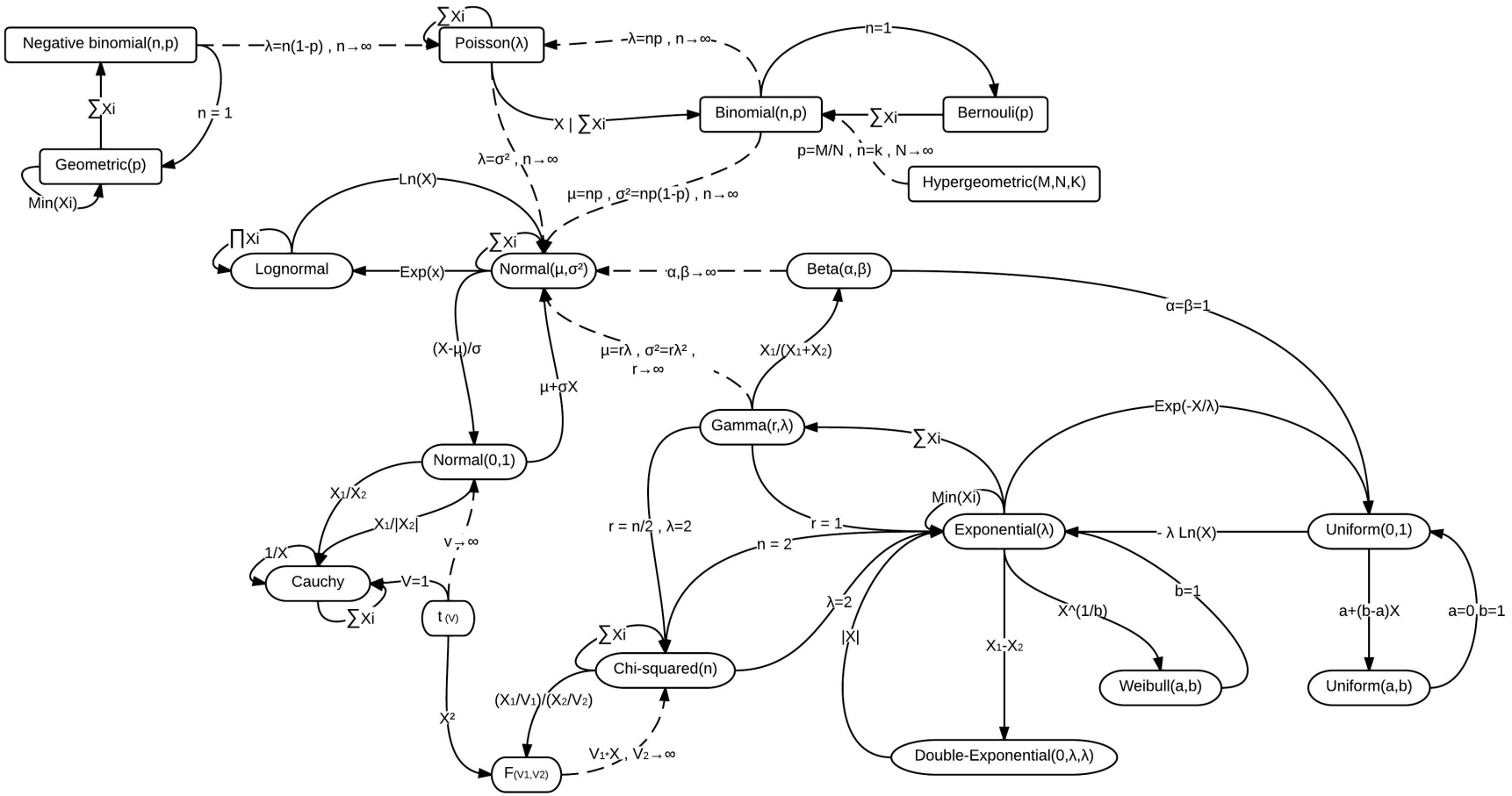
- CLT
- WLLN

Strategies for proving convergence in distribution:

- CLT
- Delta Method Normality
- CDF convergence
- MGF convergence
- Convergence in probability implies convergence in distribution
- Continuity theorem applied to some result

Strategies for proving convergence in probability:

- LLN
- Continuity theorem
- Convergence in distribution to a constant implies convergence in probability
- Resort to the definition of convergence in probability and directly find the probability or use inequalities (Markov's or Chebychev's)


Negative binomial(n,p)
λ=n(1-p) , n→∞
Poisson(λ)
∑Xi
λ=np , n→∞
n=1
∑Xi
n = 1
Geometric(p)
Min(Xi)
X | ∑Xi
Binomial(n,p)
∑Xi
Bernouli(p)
λ=σ² , n→∞
p=M/N , n=k , N→∞
Hypergeometric(M,N,K)
μ=np , σ²=np(1-p) , n→∞
Ln(X)
∏Xi
∑Xi
Lognormal
Exp(x)
Normal(μ,σ²)
α,β→∞
Beta(α,β)
α=β=1
(X-μ)/σ
μ+σX
μ=rλ , σ²=rλ² , r→∞
X1/(X1+X2)
Gamma(r,λ)
∑Xi
Exp(-X/λ)
Normal(0,1)
X1/X2
X1/|X2|
1/X
Cauchy
V=1
∑Xi
v→∞
t(V)
r = n/2 , λ=2
r = 1
n = 2
Min(Xi)
Exponential(λ)
- λ Ln(X)
Uniform(0,1)
λ=2
|X|
X1-X2
X^(1/b)
b=1
a+(b-a)X
a=0,b=1
∑Xi
Chi-squared(n)
Weibull(a,b)
Uniform(a,b)
(X1/V1)/(X2/V2)
X²
V1*X , V2→∞
F(V1,V2)
Double-Exponential(0,λ,λ)