Limit Theorems

Recall, that in the very beginning of the course, probability was defined in the frequentist sense as the limiting long-run frequency of the occurrence of an event. We now want to make this idea more precise. To do this, we need to study sequences of random variables $X_1, X_2, \ldots$ and how they behave in the limit. But we first have to clarify what we mean by the "limit" or "convergence" of such a sequence.

Math Definitions: Recall some definitions of limits you have encountered in your previous math (not stats) classes:

(a) Limit of a sequence: We call a the limit of sequence (a_n) and write $\lim_{n\to\infty} a_n = a$ if for every $\epsilon > 0$ there exists N such that for all $n \geq N$ we have $|a_n - a| < \epsilon$.

(b) Limit of a function: We say that a function $f(x)$ has the limit L at the point c ($\lim_{x\to c} f(x) = L$) if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(x) - L| < \epsilon$ for every x with $|x - c| < \delta$.

But what does convergence mean for random variables? For instance, a sequence of random variables could converge to a constant (random variable with variance zero) or they could converge to a random variable with a known distribution (but non-zero variance). There are different kinds of convergence (with different names) that one could meaningfully define for random variables.

- CONVERGENCE IN DISTRIBUTION

  That is the random variables in the sequence $X_1, X_2, \ldots$ become better and better modeled by a specific probability distribution. More precisely, if $F_n(x)$ is the CDF of X_n, then

  $$\lim_{n\to\infty} F_n(x) = F(x), \quad \text{for every } x \in \mathbb{R}$$

  We will see an example of this type of convergence in the Central Limit Theorem.

- CONVERGENCE IN PROBABILITY

  A sequence of random variables $X_1, X_2, \ldots$ is said to converge in probability to a random variable X if for all $\epsilon > 0$

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

  That is, it becomes less and less likely that the random variables in the sequence will differ from the target random variable (on the same sample space). Sometimes, in the probability literature, this type of convergence is called WEAK CONVERGENCE. Convergence in probability implies convergence in distribution. The reverse is only true if X is a constant. We will see an example of convergence in probability in the weak law of large numbers.

- Almost Sure Convergence

  This is the strongest type of convergence for a sequence of random variables. It means for a sequence of random variables $X_1, X_2, \ldots$ and another random variable X that

  $$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

  We will see an example of almost sure convergence in the strong law of large numbers.

Almost sure convergence implies convergence in probability which in turn implies convergence in distribution. Much more detail about the different types of convergence for sequences of random variables will be provided in Math 164. We will begin by proving two results that are not yet limit theorems but they are helpful in proving limit theorems.

Theorem: Markov's inequality

If X is a random variable that takes only nonnegative values, then for any value $a > 0$

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

Proof:

Theorem: Chebyshev's inequality

If X is a random variable with finite mean μ and variance σ^2, then for any value $k > 0$

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

Proof:

Example 91. Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.

(a) What can be said about the probability that this week's production will exceed 75?

(b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

We are now ready to prove our first limit theorem for a sequence of random variables.

Theorem: The weak law of large numbers
Let $X_1, X_2, \ldots$, be a sequence of independent and identically distributed random variables, each having a finite mean $E[X_i] = \mu < \infty$. Then, for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0 \qquad \text{as } n \to \infty$$

Or, in other words, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to the population mean μ in probability as the sample size n increases.

Proof:

Historical Note: The first version of this law was stated in the 16^{th} century by the Italian mathematician Gerolamo Cardano (left). It took Jacob Bernoulli (middle) over 20 years to prove a special form of the law of large numbers for Bernoulli random variables in his famous work "Ars Conjectandi" published in 1713. The more general form stated above was proved by the Russian mathematician Khinchin (right) in 1929.

The Central Limit Theorem

The Central Limit Theorem is the most famous and important result in probability theory. It states that the sum (or average) of many independent and identically distributed random variables is approximately Normal.

Theorem: Central Limit Theorem

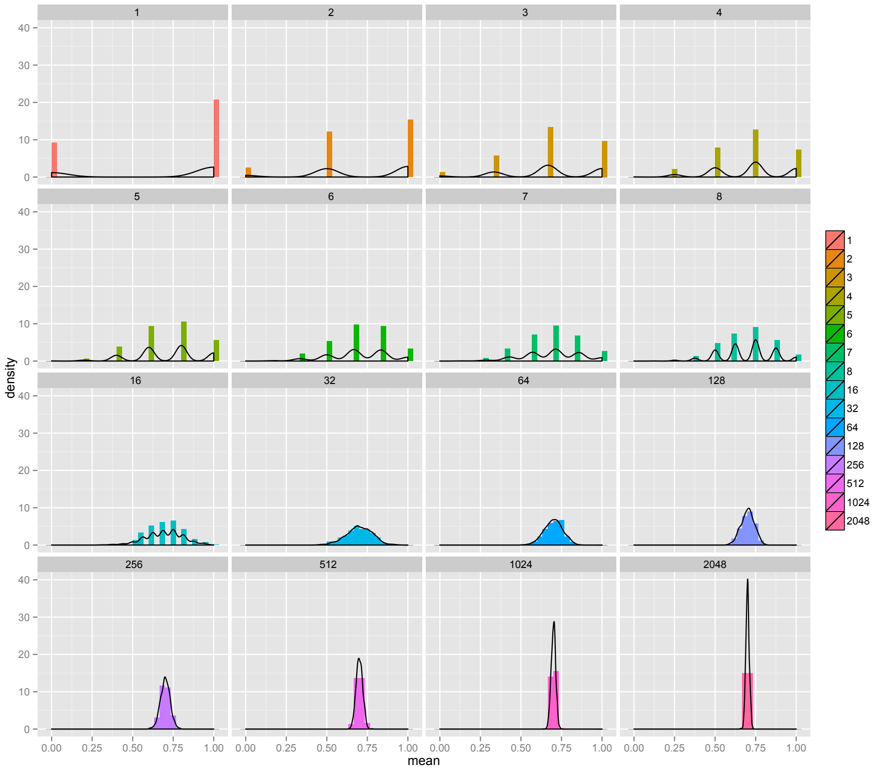
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the random variable

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

converges to a standard Normal random variable in distribution. That is

$$\lim_{n\to\infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) = \Phi(a)$$

Example 92. Below, you can see a plot of the probability mass function of a Bernoulli random variable with $p = 0.6$ (upper left panel). The other panels show sums of n independent Bernoulli($p = 0.6$) random variables. The black curves are the density estimates of the distributions. One can see that the distributions look more and more Normal with increasing n. The density peak becomes centered at 0.6 and the variance decreases.

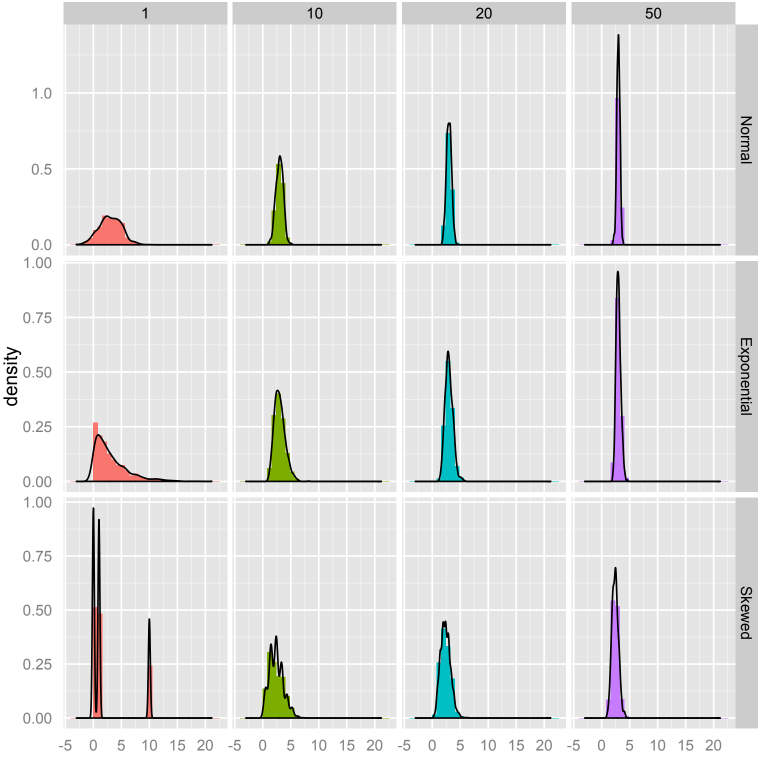


We will first take a closer look at the statement made by the Central Limit Theorem before we turn to the proof. The result is stated for a standardized sum. Alternatively we could make a statement about the distribution of the sum or average of many IID random variables which each have mean μ and finite variance $\sigma^2 < \infty$.

$$\boxed{\begin{array}{lrcl} \text{Sum:} & \sum_{i=1}^{n} X_i & \sim & \text{Normal}(n\mu, n\sigma^2) \\ \text{Average:} & \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i & \sim & \text{Normal}(\mu, \frac{\sigma^2}{n}) \end{array}}$$

Note: The Central Limit Theorem is a limit statement. That is, it tells us about the approximate distribution of the sum or average of n IID random variables if n is large. But how large is large? Would it be ok to use the Central Limit Theorem approximation if $n = 2, 10, 20, 100$? How large n has to be depends on the distribution of the individual X_i's. If this distribution is already normal, then $n = 2$ is enough, since we know that the sum (and average) of independent normal random variables is normal. If the distribution of the X_i is close to Normal (for instance Binomial with p close to 0.5), then a small n ($n \sim 10$) will already yield a pretty good approximation. If the original distribution of the X_i is very non-normal (skewed or bimodal), then n has to be larger (sometimes more than 50) before the approximation becomes good.

Example 93. Shown below are the distributions of the averages of $n = 1, 10, 20, 50$ random variables with normal, exponential, and strongly skewed distributions.



Example 94. * The amounts of automobile losses reported to an insurance company are mutually independent, and each loss is uniformly distributed between 0 and 20,000. The company covers each loss subject to a deductible of 5,000. Calculate the probability that the total payout on 200 reported losses is between 1,000,000 and 1,200,000.

Example 95. I have a 20-sided fair die. Suppose I roll the die 50 times and compute the average outcome. What is the probability that the average falls between 10 and 11? How often should I roll the 20-sided fair die to be at least 90% certain that the average of the outcomes will be between 10 and 11?

To prove the Central Limit Theorem, we will first need a Lemma that relates convergence of moment generating functions to convergence in distribution that is stated here without proof.

Lemma: Let $Z_1, Z_2, \ldots$ be a sequence of random variables having cumulative distribution functions $F_{Z_n}(a)$ and moment generating functions $M_{Z_n}(t)$, $n \geq 1$, and let Z be a random variable having cumulative distribution function $F_Z(a)$ and moment generating function $M_Z(t)$. If

$$\lim_{n\to\infty} M_{Z_n}(t) \to M_Z(t) \quad \text{for all } t$$

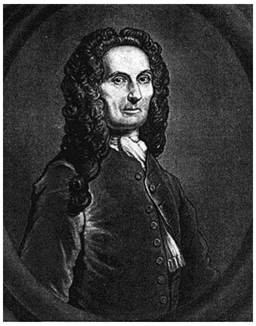
then

$$\lim_{n\to\infty} F_{Z_n}(a) \to F_Z(a) \quad \text{for all } a \text{ at which } F_Z \text{ is continuous}$$

Proof: Central Limit Theorem

Historical Note: The first version of the Central Limit Theorem was stated and proved by Abraham De Moivre (left) in 1733 for Bernoulli random variables with $p = 0.5$. The theorem was generalized to arbitrary p by Pierre-Simon Laplace (middle) in 1812. Many statisticians contributed to make the statement of the CLT more general: among them Cauchy, Bessel, Poisson and Chebyshev. Among Chebyshev's students were Markov and Lyapunov and it was Aleksandr Lyapunov (right) who in 1901 first provided the proof of the Central Limit Theorem in its most general form. In the 1920s statisticians working on refining the proof and providing simpler versions for specific cases were Cramér, Lévy, Lindeberg, von Mises, and Pólya. The version stated and proved here is known as the Lindeberg-Lévy version.

The Strong Law of Large Numbers



In the weak law of large numbers we have seen that the random variable "sample mean" converges to the population mean in probability for a growing sample of IID random variables. The strong law of large numbers makes a stronger statement, namely that this convergence is also true almost surely.

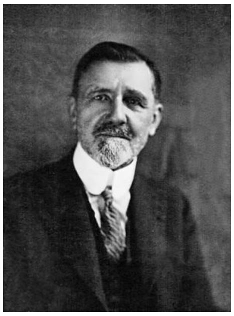
Theorem: Strong Law of Large Numbers

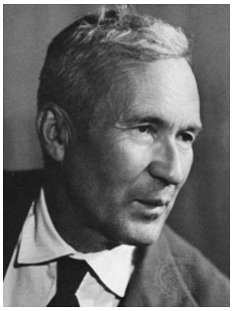
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each with mean $\mu < \infty$. Then,

$$P\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1$$

Proof:

Historical Note: The first person to prove the strong law of large numbers for the special case of Bernoulli random variables was Emile Borel (left) in 1909. Franceso Paolo Cantelli (middle) extended this result to general IID random variables. However, neither Borel nor Cantelli's arguments were flawless. In addition to Borel and Cantelli, Markov, Chebyshev, Khinchin and Kolmogorov (right) contributed to the complete proof. Today, the theorem as stated above is usually attributed to Kolmogorov.





Remark: Recall, that in the very beginning of this course, before the axioms of probability were even introduced, we understood probability (in the frequentist sense) as the long run frequency of an event. The strong law of large numbers gives justification to this idea and allows to make this definition of probability more precise.

Let E be an event (a set of outcomes of an experiment) and use $P(E)$ to denote the probability that E occurs in any particular trial. Define the indicator random variable

$$X_i = \begin{cases} 1 & E \text{ occurs on the } i^{th} \text{ trial} \\ 0 & \text{otherwise} \end{cases}$$

Then, by the strong law of large numbers with probability one

$$\frac{X_1 + \cdots + X_n}{n} \to E[X] = P(E)$$

Here, $X_1 + \cdots + X_n$ represents the number of times that the event E occurs in the first n trials.

Note: This fact, that probability can be thought of as empirical frequency if the number of trials is large, is what is used extensively in computer simulation studies for probabilistic algorithms (Math 167PS, Math 267).