Chapter 6[*]

September 2, 2018

In this lecture we introduce a number of approximation results.

1 Probability inequalities

There is an adage in probability that says that behind every limit theorem lies a probability inequality (i.e., a bound on the probability of some undesired event happening).

Since a large part of probability theory is about proving limit theorems, people have developed a bewildering number of inequalities.

1.0.1 Markov's inequality

Theorem 1.1. *For a nonnegative r.v $X(\geq 0)$, and any $u > 0$,*

$$\mathbb{P}(X > u) \leq \frac{\mathbb{E}(X)}{u}.$$

So if $\mathbb{E}(X)$ is small and we know $X \geq 0$, then X must be near zero with high probability. (Note that the inequality is *not* true if X can be negative.)

Proof. The proof is really simple: Suppose that X has a p.d.f f_X. Then,

$$u\,\mathbb{P}(X > u) = u \int_u^\infty f_X(t)dt \leq \int_u^\infty t f_X(t)dt \leq \int_0^\infty t f_X(t)dt = \mathbb{E}(X).$$

□

[*] Notes for Chapter 6 of DeGroot and Schervish adapted from Giovanni Motta's, Bodhisattva Sen's and Martin Lindquists notes for STAT W4109/W4105.

1.0.2 Chebyshev's inequality

Theorem 1.2. *Let X be a r.v such that* $\mathrm{Var}(X) < \infty$*. Then, for any* $u > 0$*,*

$$\mathbb{P}(|X - \mathbb{E}(X)| > u) \leq \frac{\mathrm{Var}(X)}{u^2},$$

i.e.,

$$\mathbb{P}\left(\frac{|X - \mathbb{E}(X)|}{\sigma(X)} > u\right) \leq \frac{1}{u^2},$$

where $\sigma(X) = \sqrt{\mathrm{Var}(X)}$*.*

Proof. Just look at the (nonnegative) r.v $(X - \mathbb{E}(X))^2$, and apply Markov's inequality. □

So if the variance of X is really small, X is close to its mean with high probability.

1.0.3 Cauchy-Schwarz inequality

For two random variables X and Y with a joint distribution,

$$|\mathrm{Cov}(X, Y)| \leq \sigma(X)\sigma(Y),$$

that is, the correlation coefficient is bounded between -1 (X and Y are anti-correlated) and 1 (perfectly positively correlated).

Exercise 1: Complete the proof. (Done is class before!)

Another way to express this inequality is

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

1.0.4 Chernoff's inequality

$$\mathbb{P}(X > u) = \mathbb{P}(e^{sX} > e^{su}) \leq e^{-su}\mathbb{E}(e^{sX}) = e^{-su}M_X(s).$$

So the m.g.f controls the size of the tail of the distribution — yet another surprising application of the m.g.f idea.

The really nice thing about this bound is that it is easy to deal with sums of independent r.v's (recall our discussion above of m.g.f.'s for sums of independent r.v.'s).

Exercise 2: Derive Chernoff's bound for sums of *independent* r.v's (i.e., derive an upper bound for the probability that $\sum_i X_i$ is greater than u, in terms of M_{X_i}).

Note that

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > u\right) = \mathbb{P}\left(e^{s\sum_{i=1}^{n} X_i} > e^{su}\right) \leq e^{-su}\mathbb{E}(e^{s\sum_{i=1}^{n} X_i}) = e^{-su}\prod_{i=1}^{n} M_{X_i}(s).$$

The other nice thing is that the bound is exponentially decreasing in u, which is much stronger than Chebyshev. (On the other hand, since not all r.v's have m.g.f's, Chernoff's bound can be applied less generally than Chebyshev.)

Note that the bound holds for all s simultaneously, so if we need as tight a bound as possible, we can use

$$P(X > u) \leq \inf_s e^{-su} M(s),$$

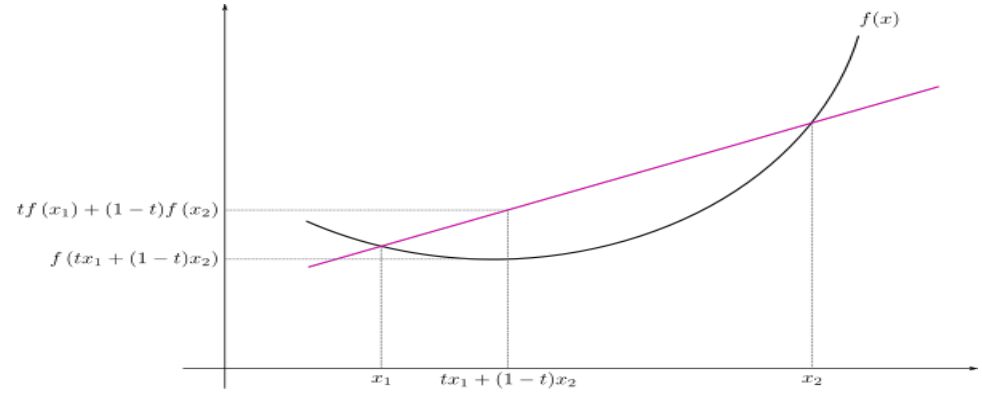
i.e., we can minimize over s.

1.0.5 Jensen's inequality

This inequality is more geometric. Think about a function $g : I \to \mathbb{R}$ which is "curved upward", that is,

$$g''(u) \geq 0,$$

for all u. Such a g is called "convex" (downward-curving functions are called "concave").



Definition: A function g is *convex* if for any x, y and $\alpha \in [0,1]$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y),$$

More generally, a convex function is bounded above by its chords.

It can be shown that if g is convex, then g lies above any line that touches g at some point, called a tangent line.

Theorem 1.3. *Let I be an open interval, and g is a convex function on I. Suppose that X is a r.v such that $\mathbb{P}(X \in I) = 1$, and $\mathbb{E}(X) < \infty$, then*

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X)).$$

That is, the average of $g(X)$ is always greater than or equal to g evaluated at the average of X.

Proof. As g is convex, we know that there exists $c > 0$ such that

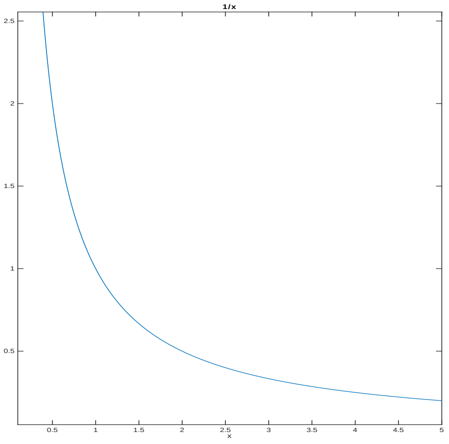
$$g(\mathbb{E}(X)) + c(x - \mathbb{E}(X)) \leq g(x)$$

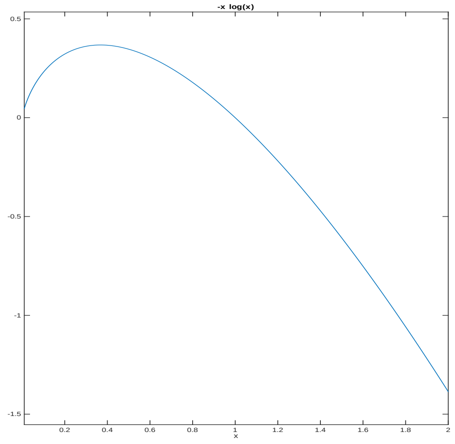
for all $x \in I$, and so,

$$g(\mathbb{E}(X)) + c(X - \mathbb{E}(X)) \leq g(X).$$

The result follows by taking expectations. □

Exercise 3: What does this inequality tell you about the means of $1/X$ and $-X \log X$?





Solution: Note that $1/x$ is convex on $(0, \infty)$. Thus, for any $X \geq 0$,

$$\frac{1}{\mathbb{E}(X)} \leq \mathbb{E}\left(\frac{1}{X}\right).$$

The function $-x \log x$ is concave on $(0, \infty)$. Thus for any nonnegative r.v X,

$$-\mathbb{E}(X) \log(\mathbb{E}(X)) \geq \mathbb{E}(-X \log X).$$

1.1 Limit theorems

1.1.1 Properties of the sample mean

Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d r.v with mean μ and variance $\sigma^2 < \infty$. Let

$$\bar{X}_n := \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

be the sample average (or mean).

Theorem 1.4. $\mathbb{E}(\bar{X}_n) = \mu$ *and* $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.

Proof. Observe that

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} \cdot n\mu = \mu.$$

Also,

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

□

1.1.2 (Weak) law of large numbers

Theorem 1.5. *Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d r.v's with finite mean μ. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\mathbb{E}(X) - \frac{1}{n}\sum_{i=1}^{n} X_i\right| > \epsilon\right) \to 0 \qquad \textit{as } n \to \infty.$$

This says that if we take the sample average of n i.i.d r.v's, the sample average will be close to the true average.

Chebyshev's simple inequality is enough to prove this fundamental result in probability theory.

Proof. We will prove the result under the assumption that $\mathrm{Var}(X) < \infty$.

Fix $\epsilon > 0$. By Chebyshev's inequality

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| > \epsilon) \le \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0, \qquad \text{as } n \to \infty.$$

□

Remember, the LLN does not hold for all r.v's: remember what happened when you took averages of i.i.d Cauchy r.v's?

Exercise 4: What goes wrong in the Cauchy case?

Hoeffding's inequality

Theorem 1.6. *Let $X_1, \ldots, X_n$ be i.i.d observations such that $\mathbb{E}(X_i) = \mu$ and $\mathbb{P}(a \leq X_i \leq b) = 1$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}.$$

1.2 Stochastic convergence concepts

1.2.1 Convergence in probability

In the above, we say that the sample mean $\frac{1}{n}\sum_{i=1}^n X_i$ *converges in probability* to the true mean.

More generally, we say r.v's $\{Z_n\}_{n=1}^\infty$ converge to Z *in probability*, and write

$$Z_n \xrightarrow{\mathbb{P}} Z,$$

if for every $\epsilon > 0$,

$$\mathbb{P}(|Z_n - Z| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

This is equivalent to saying that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Z_n - Z| \leq \epsilon) = 1.$$

The weak LLN is called "weak" because it asserts convergence in probability, which turns out to be a somewhat "weak" sense of stochastic convergence, in the mathematical sense that there are "stronger" forms of convergence — that is, it's possible to find sequences of r.v's which converge in probability but not in these stronger senses.

Exercise: Distribute n balls independently at random into n boxes. Let N_n be the number of empty boxes. Show that N_n/n converges in probability and identify the limit.

Solution: Note that

$$N_n = I_1 + \ldots + I_n,$$

where $I_i = I_{\{i\text{'th box is empty}\}}$. But the I_i's are not independent, so we cannot immediately apply WLLN.

Nevertheless, show that

$$\mathbb{E}(I_i) = \left(1 - \frac{1}{n}\right)^n, \qquad \mathbb{E}(I_i I_j) = \left(1 - \frac{2}{n}\right)^n, \text{ for } i \neq j.$$

This would show that

$$\mathbb{E}(N_n) = n\left(1 - \frac{1}{n}\right)^n, \quad \mathbb{E}(N_n^2) = \mathbb{E}(N_n) + \sum_{i \neq j} \mathbb{E}(I_i I_j) = n\left(1 - \frac{1}{n}\right)^n + n(n-1)\left(1 - \frac{2}{n}\right)^n.$$

Thus, letting $Y_n := N_n/n$, show that

$$\mathrm{Var}(Y_n) \to 0.$$

Therefore, $Y_n \xrightarrow{\mathbb{P}} e^{-1}$.

Theorem 1.7. Continuous functions of random variables. *If $Z_n \xrightarrow{\mathbb{P}} b$ and if g is a function that is continuous at b, then*

$$g(Z_n) \xrightarrow{\mathbb{P}} g(b).$$

1.2.2 Convergence in distribution

We discussed convergence of r.v's above; it's often also useful to think about convergence of distributions.

But the above notion of convergence does not help us approximate probabilities associated with random variables.

Example: Suppose you toss a coin n times and let Y be the number of heads. Suppose $n = 100$. What is $\mathbb{P}(40 \leq Y \leq 60)$? When $n = 1000$, what is $\mathbb{P}(490 \leq Y \leq 510)$?

Note that the inequalities gives us bounds that can be quite loose.

Definition: We say a sequence of r.v's with c.d.f's $F_n(u)$ *converge in distribution* to F if

$$\lim_{n \to \infty} F_n(u) = F(u)$$

for all u such that F is continuous at u (here F is itself a c.d.f).

Exercise 5: Explain why do we need to restrict our attention to continuity points of F. (Hint: think of the following sequence of distributions: $F_n(u) = I(u \geq 1/n)$, where the "indicator" function of a set A is one if $x \in A$ and zero otherwise.)

It's worth emphasizing that convergence in distribution — because it only looks at the c.d.f. — is in fact **weaker** than convergence in probability. For example, if p_X

is symmetric, then the sequence $X, -X, X, -X, \ldots$ trivially converges in distribution to X, but obviously doesn't converge in probability.

Also, if $U \sim \mathrm{Unif}(0,1)$, then the sequence

$$U, 1-U, U, 1-U, \ldots$$

converge in distribution to a uniform distribution. But obviously they do not converge in probability.

Exercise 6: Prove that convergence in probability actually is stronger, that is, implies convergence in distribution.

Central limit theorem

The second fundamental result in probability theory, after the LLN, is the CLT.

Theorem 1.8. *If $X_1, X_2, \ldots$ are i.i.d with mean zero and variance 1, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \overset{\mathcal{D}}{\to} N(0,1),$$

where $N(0,1)$ is the standard normal distribution. More generally, the usual rescalings tell us that, for $X_1, X_2, \ldots$ are i.i.d with mean μ and variance σ^2

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}(X)) \overset{\mathcal{D}}{\to} N(0,1).$$

Thus we know not only that (from the LLN) the distribution of the sample mean approaches the degenerate distribution on $\mathbb{E}(X)$, but moreover (from the CLT) we know exactly what this distribution looks like, asymptotically, if we take out our magnifying glass and zoom in on $\mathbb{E}(X)$, to a scale of $n^{-1/2}$.

In this sense the CLT is a stronger result than the WLLN: it gives more information about what the asymptotic distribution actually looks like.

You can see the applet http://onlinestatbook.com/stat_sim/sampling_dist/index.html

---

One thing worth noting: keep in mind that the CLT really only tells us about the distribution of $\bar{X}_n$ in the local neighborhood

$$(\mathbb{E}(X) - n^{-1/2}c, \mathbb{E}(X) + n^{-1/2}c)$$

— think of this as the mean plus or minus a few standard deviations.

But this does *not* imply that, say, for $\epsilon > 0$ small, (X_i's i.i.d. mean 0 and variance 1)

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq -\epsilon\right) = \mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \leq -\sqrt{n}\epsilon\right) \sim \int_{-\infty}^{-\sqrt{n}\epsilon} \phi(x)dx \quad \textbf{not true};$$

a different asymptotic approximation typically holds for the "large deviations," the tails of the sample mean distribution[1].

1.3 More on stochastic convergence

So, as emphasized above, convergence in distribution can drastically simplify our lives, if we can find a simple approximate (limit) distribution to substitute for our original complicated distribution.

The CLT is the canonical example of this; the Poisson theorem is another. What are some general methods to prove convergence in distribution?

1.3.1 Delta method

The first thing to note is that if Y_n converge in distribution or probability to a constant c, then $g(Y_n) \overset{\mathcal{D}}{\to} g(c)$ for any continuous function $g(\cdot)$.

Exercise 7: Prove this, using the definition of continuity of a function: a function $g(u)$ is continuous at u if for any possible fixed $\epsilon > 0$, there is some (possibly very small) $\delta > 0$ such that $|g(u+v) - g(u)| < \epsilon$, for all v such that $-\delta < v < \delta$. (If you're having trouble, just try proving this for convergence in probability.)

So the LLN for sample means immediately implies an LLN for a bunch of functions of the sample mean, e.g., if X_i are i.i.d with $\mathrm{Var}(X) < \infty$, then

$$\left(\prod_{i=1}^{n} e^{X_i}\right)^{1/n} = e^{\frac{1}{n}\sum_{i=1}^{n} X_i} \overset{\mathbb{P}}{\to} e^{\mathbb{E}(X)}.$$

Of course, $e^{\mathbb{E}(X)}$ should not be confused with $\mathbb{E}(e^X)$; in fact:

Exercise 8: Which is greater, $\mathbb{E}(e^X)$ or $e^{\mathbb{E}(X)}$? Give an example where one of $\mathbb{E}(e^X)$ or $e^{\mathbb{E}(X)}$ is infinite, but the other is finite.

We can also "zoom in" to look at the asymptotic distribution (not just the limit point) of $g(Z)$, whenever g is sufficiently smooth.

[1] See e.g., Large deviations techniques and applications, Dembo and Zeitouni '93, for more information.

Theorem 1.9. *Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of r.v's and let Z be a r.v with a continuous c.d.f F^*. Let $\theta \in \mathbb{R}$, and let $a_1, a_2, \ldots,$ be a sequence such that $a_n \to \infty$. Suppose that*

$$a_n(Z_n - \theta) \xrightarrow{\mathcal{D}} F^*.$$

Let g be a function with a continuous derivative such that $g'(\theta) \neq 0$. Then

$$a_n \frac{g(Z_n) - g(\theta)}{g'(\theta)} \xrightarrow{\mathcal{D}} F^*.$$

Proof. We will only give an outline of the proof (think $a_n = n^{1/2}$, if Z_n as the sample mean). As $a_n \to \infty$, Z_n must get close to θ with high probability as $n \to \infty$.

As $g(\cdot)$ is continuous, $g(Z_n)$ will be close to $g(\theta)$ with high probability.

Let's say $g(\cdot)$ has a Taylor expansion around θ, i.e.,

$$g(Z_n) \approx g(\theta) + g'(\theta)(Z_n - \theta),$$

where we have ignored all terms involving $(Z_n - \theta)^2$ and higher powers.

Then if

$$a_n(Z_n - \theta) \xrightarrow{\mathcal{D}} Z,$$

for some limit distribution F^* and a sequence of constants $a_n \to \infty$, then

$$a_n \frac{g(Z_n) - g(\theta)}{g'(\theta)} \approx a_n(Z_n - \theta) \xrightarrow{\mathcal{D}} F^*.$$

□

In other words, limit distributions are passed through functions in a pretty simple way. This is called the **delta method** (I suppose because of the deltas and epsilons involved in this kind of limiting argument), and we'll be using it a lot.

The main application is when we've already proven a CLT for Z_n,

$$\sqrt{n}\frac{Z_n - \mu}{\sigma} \xrightarrow{\mathcal{D}} N(0,1),$$

in which case

$$\sqrt{n}(g(Z_n) - g(\mu)) \xrightarrow{\mathcal{D}} N(0, \sigma^2(g'(\mu))^2).$$

Exercise 9: Assume $n^{1/2} Z_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. What is the asymptotic distribution of

1. $g(Z_n) = (Z_n - 1)^2$?
2. What about $g(Z_n) = Z_n^2$? Does anything go wrong when applying the delta method in this case? Can you fix this problem?

1.3.2 M.g.f method

What if the r.v we're interested in, Y_n, cannot be written as $g(Z_n)$, i.e., a nice function of an r.v we already know converges?

Are there methods to prove limit theorems directly?

Here we turn to our old friend the m.g.f. It turns out that the following generalization of the m.g.f invertibility theorem we quoted above is true:

Theorem 1.10. *The distribution functions F_n converge to F if:*

- *the corresponding m.g.f.'s $M_n(s)$ and $M(s)$ exist (and are finite) for all $s \in (-z, z)$, for all n, for some positive constant z.*
- *$M_n(s) \to M(s)$ for all $s \in (-z, z)$.*

So, once again, if we have a good handle on the m.g.f's M_n, we can learn a lot about the limit distribution. In fact, this idea provides the simplest way to prove the CLT.

Proof. (of Theorem 1.8) Assume X_i has mean zero and unit variance; the general case follows easily, by the usual re-scalings.

Now let's look at $M_n(s)$, the m.g.f of $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$. If X_i has m.g.f $M_X(\cdot)$, then $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ has m.g.f

$$M_n(s) = \mathbb{E}\left(e^{s\sum_{i=1}^n X_i/\sqrt{n}}\right) = \mathbb{E}\left(\prod_{i=1}^n e^{sX_i/\sqrt{n}}\right) = \left[M_X\left(\frac{s}{\sqrt{n}}\right)\right]^n.$$

Now let's make a Taylor expansion. We know that $M_X(0) = 1, M_X'(0) = 0$, and $M_X''(0) = 1$. (Why?) So we can write

$$M_X(s) = 1 + s^2/2 + o(s^2)$$

for s around 0. Thus,

$$M_n(s) \approx \left\{1 + \frac{s^2}{2n} + o\left(\frac{s^2}{2n}\right)\right\}^n \to e^{s^2/2}.$$

Now recall that $e^{s^2/2}$ is the m.g.f of a standard normal r.v, and then appeal to our general convergence-in-distribution theorem for m.g.f's. □