Probability Theory

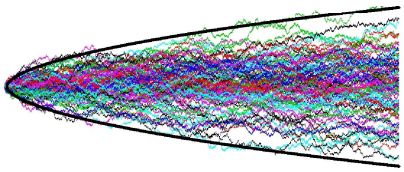
Instructor: Jodin Morey moreyj@lemoyne.edu

Previous Lecture

- Change of Vars in 1D

10 - Inequalities and Limit Theorems

What should we do if we can't calculate a **probability** or **expectation** exactly? Simulate it, bound it, approximate it.



- **Simulate it** using random computer models. (Monte Carlo. No provable guarantees)

- **Bound it** using inequalities. (§10.1. Provable guarantees! The desired quantity is in a certain range)

- **Approximate it** using limit theorems. (§10.2/§10.3, provides probabilities the desired quantity is in any particular range)

§10.1 - Inequalities

Recall: if X and Y are uncorrelated, then $E(XY) = E(X)E(Y)$.
But in general, calculating $E(XY)$ (like we do w/covariance) requires knowledge of the *joint* distr of X and Y.

If we don't know it, the Cauchy-Schwarz inequality lets us *bound* $E(XY)$ in terms of the marginal second moments $E(X^2)$ and $E(Y^2)$.

Thm (Cauchy-Schwarz): For any X and Y w/finite variances, $|E(XY)| \le \sqrt{E(X^2)E(Y^2)}$.

Recall from calculus: $\vec{x} \cdot \vec{y} = |\vec{x}||\vec{y}|\cos\theta$. Or equivalently: $|\vec{x} \cdot \vec{y}| = \sqrt{\vec{x} \cdot \vec{x}}\sqrt{\vec{y} \cdot \vec{y}}\cos\theta \le \sqrt{(\vec{x} \cdot \vec{x})(\vec{y} \cdot \vec{y})}$ (for $0 \le \theta \le 90$)
(dot product is playing the role of expectation in this version of Cauchy-Schwarz)

Proof. For any t, we have: $0 \le E((Y - tX)^2) = E(Y^2) - 2tE(XY) + t^2E(X^2)$.

Where did t come from? The idea is to introduce t to get a continuous function.

This allows us to take a derivative w/respect to t to find extremas of the expression.
This will allow us to find the tightest bound possible for $E(XY)$.

Differentiating the RHS with respect to t and setting it equal to 0, we find an extremum (a minimum, since the second derivative $2E(X^2)$ is positive) when $t = \frac{E(XY)}{E(X^2)}$, resulting in the tightest bound.

Substituting in this value of t, we have: $0 \le E(Y^2) - 2\left(\frac{E(XY)}{E(X^2)}\right)E(XY) + \left(\frac{E(XY)}{E(X^2)}\right)^2 E(X^2)$

$$0 \le E(X^2)E(Y^2) - E(XY)^2 \qquad \text{(simplifying)}$$

$$E(XY)^2 \le E(X^2)E(Y^2) \qquad \text{(rearranging)}$$

$$|E(XY)| \le \sqrt{E(X^2)E(Y^2)}, \qquad \text{(square rooting)}$$

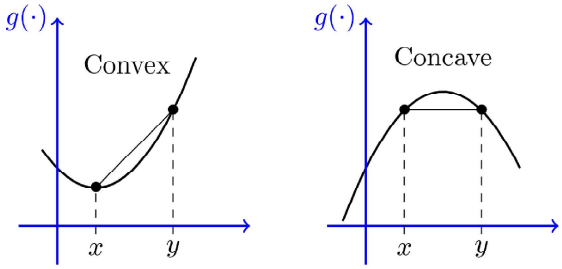
we have the Cauchy-Schwarz inequality. ■

Jensen: an Inequality for Curvature

For nonlinear functions g, $E(g(X))$ may be very different from $g(E(X))$.

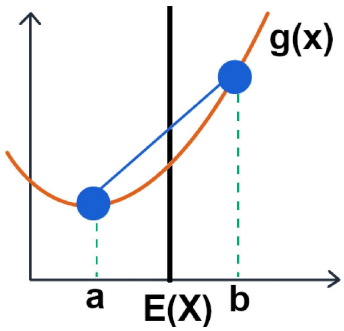
If g is either a convex or a concave function, Jensen's inequality tells us exactly which of $E(g(X))$ and $g(E(X))$ is greater.

Often we can take the second derivative to test for convexity/concavity.

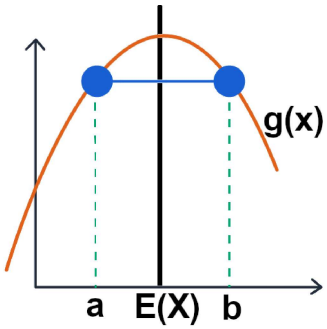


Thm (Jensen's Ineq): Given X, if g is convex, then $E(g(X)) \ge g(E(X))$. If g is a concave, then $E(g(X)) \le g(E(X))$.
The only way equality can hold is if there are constants a,b such that $g(X) = a + bX$ w/prob 1.

Ex: Let X be discrete taking on values a,b with prob $\frac{1}{2}$ each. Let g be convex or concave. Then:

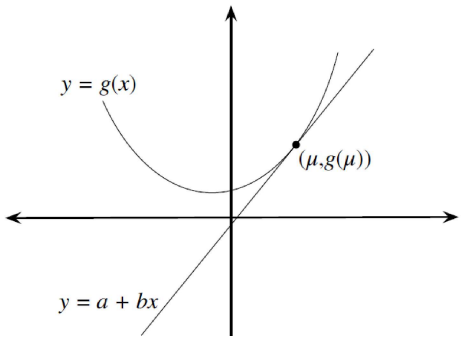


g is convex, so avg of $g(a),g(b)$ is greater than $g(E(X))$.



g is concave, so avg of $g(a),g(b)$ is less than $g(E(X))$.

Proof of thm: If g is convex, then all lines that are tangent to g lie below g (see figure).
In particular, let $\mu := E(X)$, and consider the tangent line at the point $(\mu, g(\mu))$.



Denoting this tangent line by $a + bx$, we have $g(x) \geq a + bx$ for all x by convexity, so $g(X) \geq a + bX$.

Taking the expectation of both sides, $E(g(X)) \geq E(a + bX) = a + bE(X) = a + b\mu = g(\mu) = g(E(X))$, as desired.

If g is concave, then $h = -g$ is convex, so we can apply what we just proved to h to see that the inequality for g is reversed from the convex case.

Lastly, assume that equality holds in the convex case. Let $Y = g(X) - a - bX$.

Then Y is nonnegative w/$E(Y) = 0$, so $P(Y = 0) = 1$ (even a tiny nonzero chance of $Y > 0$ occurring would make $E(Y) > 0$).

So equality holds if and only if $P(g(X) = a + bX) = 1$.

For the concave case, we can use the same argument with $Y = a + bX - g(X)$. ■

Ex (**Jensen's Inequality**): Let $g(x) = x^2$. What does Jensen say about $g(E(X))$, $E(g(X))$?

Solution: Since g is convex (its second derivative is 2), Jensen's inequality says $E(X^2) \geq (E(X))^2$.

Note, we can verify this since we already know variances are nonnegative: $E(X^2) - (E(X))^2 \geq 0$.

A few examples:

- $E|X| \geq |E(X)|$,
- $E(\frac{1}{X}) \geq \frac{1}{E(X)}$, for positive X,
- $E(\ln X) \leq \ln(E(X))$, for positive X.

Ex (**Which is Larger**?): Given positive X and Y. Which is larger?

a. $E(X^3)$ or $(E(X))^3$

Since $g(x) = x^3$ is convex when $x > 0$, then $E(X^3) \leq (E(X))^3$.

b. $-e^{E(X)}$ or $E(-e^x)$

Since $g(x) = -e^x$ is concave, then $E(-e^x) \leq -e^{E(X)}$.

Markov, Chebyshev, Cherno: bounds on tail probabilities

The following inequalities provide bounds on the prob of a rv taking on an "extreme" value in the right or left tail of a distr.

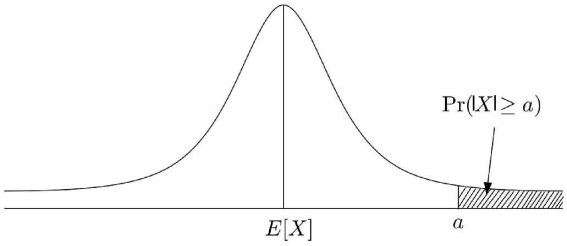


Figure : Markov's Inequality bounds the probability of the shaded region.

Limit prob of insurance payout X

Thm (**Markov's Ineq**). For any X and constant $a > 0$, $P(|X| \geq a) \leq \frac{E|X|}{a}$.

Proof. Let $Y = \frac{|X|}{a}$. Then Markov can be rewritten: $P(\frac{|X|}{a} \geq 1) \leq \frac{E|X|}{a} = \frac{1}{a}E|X| = E\left(\frac{|X|}{a}\right) = E(Y)$.

So, we need to show: $P(Y \geq 1) \leq E(Y)$.

Note: $I(Y \geq 1) \leq Y$. Why?

First, notice $0 \leq \frac{|X|}{a} = Y$. Now there are two options: $I(Y \geq 1) = 0$ or $I(Y \geq 1) = 1$.

If $I(Y \geq 1) = 0$ then since Y is nonnegative, we have $I(Y \geq 1) \leq Y$.

Next, if $I(Y \geq 1) = 1$ then $I(Y \geq 1) = 1 \leq Y$ (because the indicator says so).

Taking the expectation of both sides, we have $P(Y \geq 1) \leq E(Y)$. ■

For an intuitive interpretation, let X be the income of a randomly selected individual from a population.

Taking $a = 2E(X)$, Markov's inequality says: $P(X \geq 2E(X)) \leq \frac{1}{2}$. (no need for $|\cdot|$ since income is positive)

That is: it's impossible for more than half the population to make at least twice the average income.

If we put some conditions on X, we can get an even better estimate.

Thm (**Chebyshev**). Let X have mean μ and variance σ^2. Then for any $a > 0$, $P(|X - \mu| \leq a) \leq \frac{\sigma^2}{a^2}$.

Proof. By Markov's inequality, $P(|X - \mu| \geq a) = P((X - \mu)^2 \geq a^2) \leq \frac{E(X-\mu)^2}{a^2} = \frac{\sigma^2}{a^2}$. ■

Thm (Chernoff). For any X and constants $a > 0$ and $t > 0$, $P(X \geq a) \leq \frac{E(e^{tX})}{e^{ta}}$.

Proof. Note the transformation $g(x) = e^{tx}$ is invertible and strictly increasing.

So by Markov's inequality, we have $P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}}$. ■

Chernoff has two very nice features:

- RHS can be optimized over t to give the tightest upper bound.
- If MGF of X exists, then the numerator in the bound is the MGF, and some of the useful properties of MGFs can come into play.

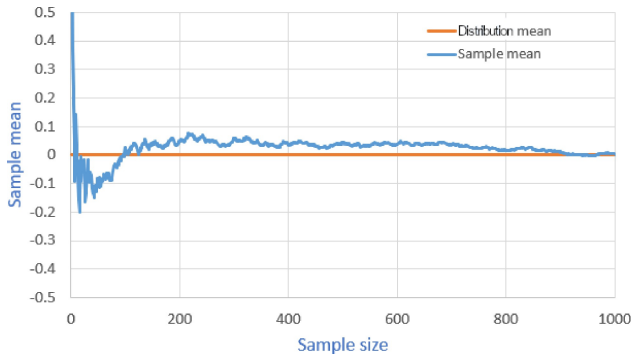
☣ A bound is not an approximation! $P(X \geq a) \leq \frac{E(e^{tX})}{e^{ta}}$, but $P(X \geq a)$ might be a *lot* less than $\frac{E(e^{tX})}{e^{ta}}$.

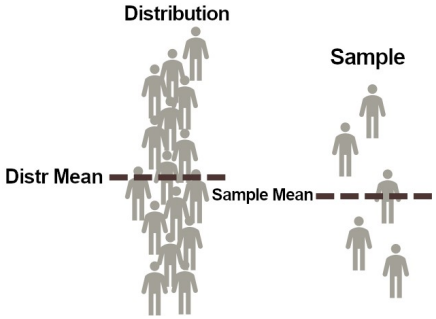
Harvard Video: youtube.com/watch?v=UtXK_EQ3Pow&list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo&index=28

---

§10.2 - Law of Large Numbers (LLN)

If you are sampling from some distr, and you want to know how many samples you should take before you can trust your sample mean to accurately represent the distr's mean, we will need some new tools.





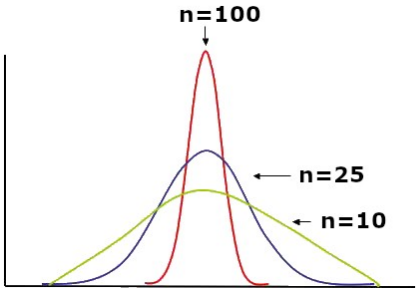
For iid $X_1, X_2, \ldots, X_n$ w/mean μ and finite variance σ^2, let $\overline{X}_n = \frac{X_1+X_2+\ldots+X_n}{n}$ be the **sample mean** of X_1 thru X_n.

To describe the behavior of $\overline{X}_n$ as the sample size n grows, we introduce: the Law of Large Numbers (LLN) and the Central Limit Thm (CLT).

Observe the expected value of $\overline{X}_n$ is μ:

$$E(\overline{X}_n) = \tfrac{1}{n}E(X_1 + X_2 + \ldots + X_n) = \tfrac{1}{n}(E(X_1) + \ldots + E(X_n)) = \mu.$$

And $Var(\overline{X}_n) = \frac{1}{n^2}(Var(X_1) + \ldots + Var(X_n)) = \frac{\sigma^2}{n}$.



Var decreases as n increases

An individual sample mean $\overline{X}_n$ will usually not equal μ.
However, we can ensure a sample's mean will be close to μ by increasing sample size.

To understand the following theorem, imagine taking an infinite number of samples X_i from a larger population (individuals from the world), and writing down their height $\hat{s}_1 := (X_1, X_2, \dots) = (72, 80, \dots)$.

Now imagine doing it again.
This 2nd time, due to randomness, you will obviously not choose the same people in the same order:
$\hat{s}_2 = (X_1, X_2, \dots) = (83, 56, \dots)$.

Now imagine all the different infinite vectors you can create this way. This is our sample space $S = \{\hat{s}_1, \hat{s}_2, \dots\}$ for rv $\overline{X}_n(\hat{s})$.

Notice you could theoretically randomly get $\hat{s} = (X_1, X_2, \dots) = (65, 65, 65, 65, 65, \dots)$, but the prob is zero (because the vectors are infinite).

Thm (**Strong Law of Large Numbers**, **SLLN**). The sample mean $\overline{X}_n$ converges to the true mean μ pointwise, w/prob 1.
Recalling that rvs are functions from the sample space S to $\mathbb{R}$. This form of convergence says that $\overline{X}_n(\hat{s}) \to \mu$ for each infinite vector $\hat{s} \in S$, except the convergence is allowed to fail on some set $B_0 \subseteq S$ of exceptions, as long as $P(B_0) = 0$.
In short, $P(\overline{X}_n \to \mu) = 1$.

The confusing part of this thm is the bit about pointwise convergence.

Ex (**Pointwise Convergence**): Imagine flipping a coin an infinite number of times.  . . .

Now think about all the different ways this could turn out. Calling heads 1, and tails 0.

The sample space could include vectors like $\hat{s}_1 = \{01110001101001101110011011110\dots\}$ and $\hat{s}_2 = \{10111000110100100111000001\dots\}$, and infinitely many more.

When you flip your coin an infinite # of times, you end up generating one of these $\hat{s}$ from the sample space.

SLLN says that with prob 1, $\overline{X}_n(\hat{s}) \to \frac{1}{2}$.

You might say to yourself, but what if I get the following: $\hat{s}_0 = \{0000000000000\dots\}$ (all tails), and therefore $\overline{X}_n(\hat{s}) \to 0$.

SLLN admits to this possibility, but tells us that $\hat{s}_0 \in B_0$, and therefore the probability of this occurring is zero.

Thm (**Weak Law of Large Numbers**, **WLLN**): For all $\varepsilon > 0$, $P(|\overline{X}_n - \mu| > \varepsilon) \to 0$ as $n \to \infty$.
(This form of convergence is called convergence in probability.)

Proof. Fix $\varepsilon > 0$. Then, by Chebyshev's inequality, $P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$.

As $n \to \infty$, the RHS goes to 0, and thus the LHS must as well. ■

In simulations, statistics, and science, every time we use the average after repeating some process, or after sampling from a larger population, and we use this average to approximate the true average, we are implicitly using the LLN.

Harvard Video: youtube.com/watch?v=OprNqnHsVIA&list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo&index=29

What did we learn?

- Cauchy-Schwarz Inequality, $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$.
- Jensen's Inequality: convex $E(g(X)) \geq g(E(X))$; concave $E(g(X)) \geq g(E(X))$.
- Bounds on tail probabilities.
  - Markov: $P(|X| \geq a) \leq \frac{E|X|}{a}$.
  - Chebyshev: $P(|X - \mu| \leq a) \leq \frac{\sigma^2}{a^2}$.
  - Cherno: $P(X \geq a) \leq \frac{E(e^{tX})}{e^{ta}}$.
- Strong Law of Large #s (SLLN): $\overline{X}_n \to \mu$ pointwise, with prob 1.
- Weak Law of Large #s (WLLN): For all $\varepsilon > 0$, $P(|\overline{X}_n - \mu| > \varepsilon) \to 0$ as $n \to \infty$.