Lecture 12:

Inequalities and limit theorems

Cauchy-Schwarz

Cauchy-Schwarz

Theorem 10.1.1 (Cauchy-Schwarz) For any r.v.-s X, Y with finite variances:

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

Proof: For any real t, we have:

$$0 \leq E(Y - tX)^2 = E(Y^2) - 2tE(XY) + t^2E(X^2)$$

- this gives infinitely many inequalities (for different t-s). But the best (tightest) one is given at the minimum of the r.h.s. – where the derivative = 0, which is at $t^* = E(XY)/E(X^2)$.

If X, Y are uncorrelated, $E(XY) = E(X)E(Y)$ – depends only on marginal exp-s. In general, calculating $E(XY)$ requires knowledge of the joint distr. CS gives a bound in terms of 2nd moments.

Cauchy-Schwarz

Theorem 10.1.1 (Cauchy-Schwarz) For any r.v.-s X, Y with finite variances:

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

If $E(X) = E(Y) = 0$ (centred r.v.-s) – CS says the correlation is between -1 and 1.

CS can be applied in creative ways: by writing $X = X \cdot 1$, CS tells $|E(X \cdot 1)| \leq \sqrt{E(X^2)\,E(1)}$, which gives $E(X^2) \geq (EX)^2$.

Cauchy-Schwarz

Theorem 10.1.1 (Cauchy-Schwarz) For any r.v.-s X, Y with finite variances:

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

Example 10.1.3 (Second moment method). Let X – non-negative r.v., and we want an upper bound on $P(X = 0)$.

Rewrite $X = X \cdot I(X > 0)$ – with an indicator of X being positive.

Then $E(X) \leq \sqrt{E(X^2)\,E(I(X > 0))}$. By the fundamental bridge, this gives $P(X > 0) \geq \dfrac{(EX)^2}{E(X^2)}$, or equivalently: $P(X = 0) \leq \dfrac{\text{Var}(X)}{E(X^2)}$

Cauchy-Schwarz

Example 10.1.3 (Second moment method). For a non-negative r.v. we have $P(X=0) \le \dfrac{\text{Var}(X)}{E(X^2)}$.

Let $X = I_1 + \ldots + I_n$ – sum of n uncorrelated indicator r.v.-s.

Let $p_j = E(I_j)$. Then:

$$\text{Var}(X) = \sum_{j=1}^{n} \text{Var}(I_j) = \sum_{j=1}^{n} (p_j - p_j^2) = \sum_{j=1}^{n} p_j - \sum_{j=1}^{n} p_j^2 = \mu - c$$

Recall that $E(X^2) = \text{Var}(X) + (EX)^2 = \mu^2 + \mu - c$, and we have:

$P(X=0) \le \dfrac{\text{Var}(X)}{E(X^2)} = \dfrac{\mu - c}{\mu^2 + \mu - c} \le \dfrac{1}{\mu + 1}$ – for such an r.v. we can say that “the larger the mean – the less the chance of $X = 0$”.

Cauchy-Schwarz

Example 10.1.3 (Second moment method). For a non-negative r.v. $X = I_1 + \ldots + I_n$ we have $P(X=0) \leq \dfrac{1}{E(X)+1}$.

Suppose there are 14 people in a room. How likely is it that there are 2 people with the same birthday or birthdays one day apart?

This is much harder than the birthday paradox, but we can use this bound – let X = (# of "near birthday" pairs).

Using indicator r.v.-s, $E(X) = \binom{14}{2}\dfrac{3}{365} \approx 0.748$

So $P(X=0) \leq \frac{1}{E(X)+1} \approx 0.572,$ while the true $P(X=0) \approx 0.46,$

so our bound is consistent

Jensen

Jensen

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

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

Recall that to test convexity/concavity one can take the 2nd derivative:

g – convex	$g''(x) \geq 0$
g – concave	$g''(x) \leq 0$

Jensen

Theorem 10.1.5 (Jensen). Let X – r.v.

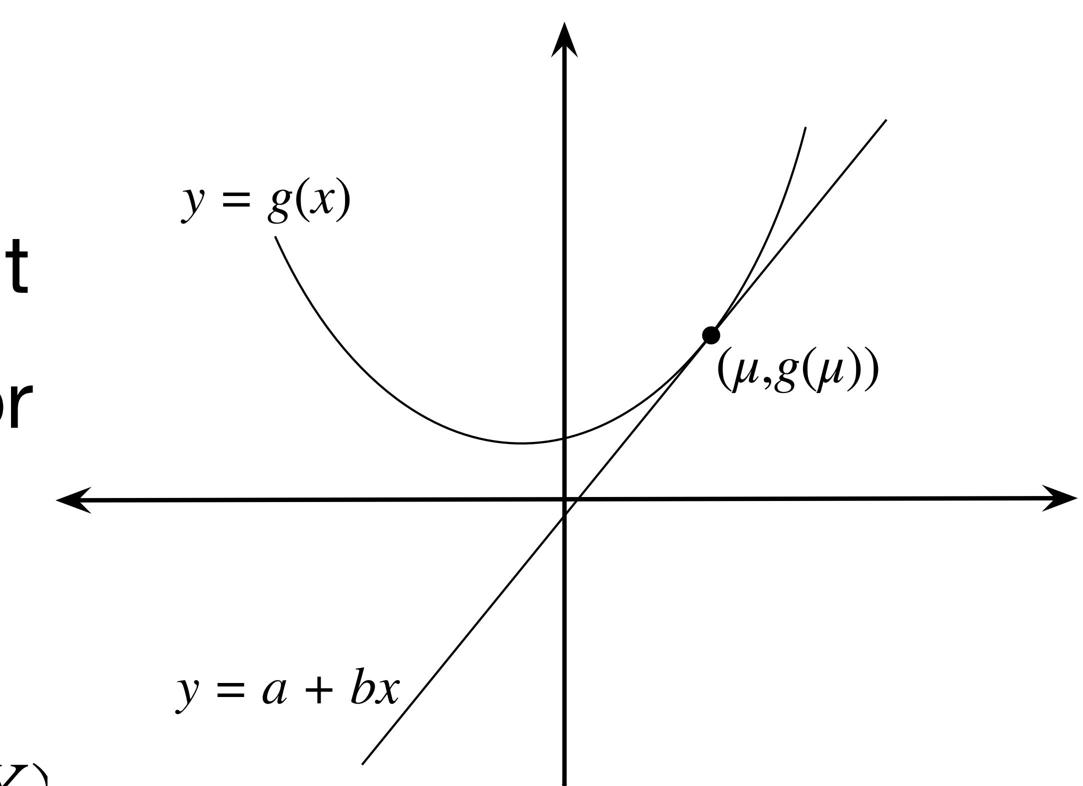
If g – convex function, then $E(g(X)) \geq g(E(X))$.

If g – concave function, then $E(g(X)) \leq g(E(X))$.

The only way the equality can hold is $g(X) = a + bX$ (with proba 1).

Proof: Let g be convex – then all its tangent lines lie below g. Let $\mu = E(X)$ and consider the tangent at $(\mu, g(\mu))$ – it is unique if g is diff., if not – take any. For this tangent $a + bx$, $g(x) \geq a + bx$ for any x. Taking expectation of both sides:

$E(g(X)) \geq E(a + bX) = a + bE(X) = g(E(X))$



Jensen

Here are some cases of Jensen:

- $g(x) = x^2$ is convex, so $E(X^2) \geq (EX)^2$ – recall Cauchy-Schwarz.

- In St.Petersburg paradox we found $E(2^N) > 2^{EN}$ for $N \sim$ FS$(1/2)$. This agrees with Jensen, since $g(x) = 2^x$ is convex – but moreover, it tells that the direction of inequality doesn't depend on the distribution of N.

- $E|X| \geq |EX|$

- $E(1/X) \geq 1/(EX)$ for positive r.v.s X

- $E(\log X) \leq \log(EX)$ for positive r.v.s X

Jensen

Example 10.1.6 (Bias of sample std). Let $X_1, \ldots, X_n$ – be i.i.d. r.v.s with variance σ^2. We've seen that sample variance S_n^2 (with $n-1$ in denominator) is an ***unbiased*** estimator for σ^2 – $E(S_n^2) = \sigma^2$.
But for std, however:

$$E(S_n) = E(\sqrt{S_n^2}) \leq \sqrt{E(S_n^2)} = \sigma$$

– sample std tends to ***underestimate*** the true std!

How biased it is depends on the distribution, there is no universal way to fix this (as with dividing by $n-1$ instead of n in the variance). Fortunately, for large samples this bias is typically small.

Jensen

Example 10.1.7 (Entropy). The **surprise** of learning that an event happened with prob p is defined as $\log_2(1/p)$, measured in **bits** (event of prob $1/2$ has surprise of 1 bit, low proba = high surprise).

Let X be a discrete r.v. taking values $a_1, \ldots, a_n$ with probas $p_1, \ldots, p_n$ (so $p_1 + \ldots + p_n = 1$). The **entropy** of X is the average surprise of learning the value of X:

$$H(X) = \sum_{j=1}^{n} p_j \log_2(1/p_j)$$

- note that it only depends on the probabilities, not the values a_j.

Using Jensen, let's show that $p_j = 1/n \ \forall j$ has maximum entropy!

Jensen

Proof: For $X \sim$ DUnif$(a_1, \ldots, a_n)$, $H(X) = \sum_{j=1}^{n} \frac{1}{n} \log_2 n = \log_2 n$.

Let's make an r.v. Y that takes values $1/p_1, \ldots, 1/p_n$ with probabilities $p_1, \ldots, p_n$.
Then $H(Y) = E(\log_2(Y))$ and, clearly, $E(Y) = n$.
By Jensen,

$$H(Y) = E(\log_2(Y)) \leq \log_2(E(Y)) = \log_2(n) = H(X)$$

and since the entropy of an r.v. depends only on the probabilities p_j, not the specific values the r.v. takes – it is unchanged if we change the support from $1/p_1, \ldots, 1/p_n$ to $a_1, \ldots, a_n$. So X has largest possible entropy of all r.v.s with support on n points!

Jensen

Example 10.1.8 (Kullback-Leibler divergence). Let $\mathbf{p} = (p_1, \ldots, p_n)$ & $\mathbf{q} = (q_1, \ldots, q_n)$ be probability vectors (nonnegative and sum to 1) – the same support.

The KL-divergence between $\mathbf{p}$ and $\mathbf{q}$ is:

$$KL(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{n} p_j \log_2(1/r_j) - \sum_{j=1}^{n} p_j \log_2(1/p_j)$$

– the difference between average surprises when the actual probabilities are $\mathbf{p}$, but we instead are working with $\mathbf{q}$ (i.e., true $\mathbf{p}$ is unknown, and $\mathbf{q}$ is our current guess for it).

Show that KL-divergence is non-negative.

Jensen

Proof: By properties of logs,

$$KL(\mathbf{p}, \mathbf{q}) = -\sum_{j=1}^{n} p_j \log_2\left(\frac{q_j}{p_j}\right)$$

Let Y be a r.v. that takes values q_j/p_j with probabilities p_j, so $KL(\mathbf{p}, \mathbf{q})$ is its negative average surprise: $-E(\log_2(Y))$.

By Jensen,

$$KL(\mathbf{p}, \mathbf{q}) = -E(\log_2(Y)) \geq -\log_2(E(Y)) = -\log_2(1) = 0$$

with equality iff $\mathbf{p} = \mathbf{q}$. So we're more surprised on average when working with wrong probabilities than when working with correct ones.

Markov, Chebyshev, Chernoff

Markov, Chebyshev, Chernoff

Theorem 10.1.10 (Markov). For any r.v. X and constant $a > 0$,

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

Proof: Let $Y = |X|/a$. We need to show $P(Y \geq 1) \leq E(Y)$. Note that $I(Y \geq 1) \leq Y$, since if $I = 0$, this reduces to $0 \leq Y$, and if $I = 1$, this reduces to $1 \leq Y$, which is the argument of the indicator. Taking expectations of both sides, we have Markov's inequality.

Let X – income of a randomly selected individual from a population. If $a = 2E(X)$ – then Markov says $P(X \geq 2E(X)) \leq 1/2$ – its impossible for more than half to make twice the average income. Similarly, $P(X \geq 3E(X)) \leq 1/3$, etc.

Markov, Chebyshev, Chernoff

Theorem 10.1.10 (Chebyshev). Let X have mean μ and variance σ^2. Then for any $a > 0$,

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

Proof: By Markov's inequality,

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

Substituting $c\sigma$ with $c > 0$ for a, Chebyshev takes form:

$$P(|X-\mu| \geq c\sigma) \leq 1/c^2$$

– e.g. there can't be more than 25% chance of being 2std-s or more from the mean.

Markov, Chebyshev, Chernoff

Theorem 10.1.12 (Chernoff). For any r.v. X and constants $a > 0$ and $t > 0$,

$$P(X \geq a) \leq \frac{E(e^{tX})}{e^{ta}}$$

Proof: $g(x) = e^{tx}$ is invertible and strictly increasing, so by Markov's inequality we have

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}}$$

It might be not clear what Chernoff has to offer that Markov couldn't, but actually the r.h.s. can be optimised w.r.t. t to give the tightest upper bound.

Markov, Chebyshev, Chernoff

Example 10.1.13 (Bounds on Normal tail probability). $Z \sim \mathcal{N}(0,1)$. By the 68-95-99% rule, we know that $P(|Z| > 3) \approx 0.003$. Let's compare that to bounds from Markov, Chebyshev and Chernoff.

1) Markov: using $E|Z| = \sqrt{2/\pi}$,

$$P(|Z| > 3) \leq \frac{E|Z|}{3} = \frac{1}{3} \cdot \sqrt{\frac{2}{\pi}} \approx 0.27$$

2) Chebyshev: $P(|Z| > 3) \leq 1/9 \approx 0.11$

3) Chernoff: $P(|Z| > 3) = 2P(Z > 3) \leq 2e^{-3t}E(e^{tZ}) = 2e^{-3t}e^{t^2/2}$ – minimized at $t = 3$, which gives $P(|Z| > 3) \leq 2e^{-9/2} \approx 0.022$.