28: Probability Bounds

Lisa Yan and Jerry Cain

November 16, 2020

Quick slide reference

Computing probabilities involving X

If we know the full parameterized distribution:	$X \sim \text{Poi}(5)$	(we can compute any sort of probability on X)
If we don't have the distribution, but we have an i.i.d. sample:	$(3, 4, 1, 6, 0, 2, 3)$	(we can use bootstrapping to compute probabilities on X)
If we know the model but not the parameter + we have an i.i.d. sample:	$(3, 4, 1, 6, 0, 2, 3)$ $X \sim \text{Poi}(\lambda)$	(we can estimate X's parameters and then use the estimated θ to compute probabilities)

Today: Even if we only have a statistic of the sample (e.g., $E[X]$ or $\text{Var}(X)$), we can still <u>bound</u> probabilities of X

Probabilities: What the CLT tells us

$E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$, where X_i i.i.d.

As $n \to \infty$, $X = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$

Central Limit Theorem

Suppose we could observe $E[X]$ and $\text{Var}(X)$.

If we knew that X was a sum of many i.i.d. X_i's:

- By the CLT, X is Normal (for large n)
- Therefore we can compute any probability involving X!

Markov's and Chebyshev's Inequalities

Aside: Inequalities of random variables

Let X and Y be jointly distributed. Suppose that

$$X \leq Y \quad \Longleftrightarrow \quad x \leq y$$

for all possible $X = x, Y = y$
(i.e., with nonzero joint PDF or PMF)

<u>Property</u>

If $X \leq Y$, then $E[X] \leq E[Y]$.

<u>Proof</u>

1.	$Y - X \geq 0$	(for all possible $X = x, Y = y$)
2.	$E[Y - X] \geq 0$	(Expectation)
3.	$E[Y] - E[X] \geq 0$	(Linearity of Expectation)
4.	$E[X] \leq E[Y]$	(rearrange)

Markov's Inequality

Let X be a non-negative random variable ($X \geq 0$). Then

$$P(X \geq a) \leq \frac{E[X]}{a} \quad \text{for all } a > 0$$

Interpret The probability that X is greater than a is bounded by its mean (and a).

Proof

1. Define $I = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}$
2. $I \leq \frac{X}{a}$ (since I is 1 whenever $X \geq a$)
3. $E[I] = P(X \geq a)$ (I is Bernoulli)
4. $E[I] \leq E[X/a] = \frac{E[X]}{a}$ (If $X \leq Y$ then $E[X] \leq E[Y]$)

Chebyshev's Inequality

Let X be a random variable where $E[X] = \mu$, $\text{Var}(X) = \sigma^2$.

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \quad \text{for all } k > 0$$

<u>Interpret</u> The probability that X is further than k from its mean is bounded by its variance (and k).

<u>Proof</u>

1. $(X - \mu)^2 \geq 0$ (i.e, $(X - \mu)^2$ is a non-negative RV)
2. $P\left((X - \mu)^2 \geq k^2\right) \leq \frac{E[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$ (Markov's Inequality with $a = k^2$)
3. $(X - \mu)^2 \geq k^2 \Leftrightarrow |X - \mu| \geq k$ (def. absolute value)
4. $P(|X - \mu| \geq k) \leq \sigma^2/k^2$ (re-define event expressed in 2.)

Bounding happiness

- Suppose you read aggregate survey results of Bhutanese happiness points (h.p.).
- You learn that the average happiness is 86.7 h.p. and variance is 405.62 $(\text{h.p.})^2$.

Let X = the happiness of a Bhutanese person.

1. $P(X \geq 100)$
2. $P(|X - 86.7| \geq 25)$

Bounding happiness

- Suppose you read aggregate survey results of Bhutanese happiness points (h.p.).
- You learn that the average happiness is 86.7 h.p. and variance is 405.62 (h.p.)^2.

Let X = the happiness of a Bhutanese person.

1. $P(X \geq 100) \leq \frac{86.7}{100} = 0.867$

Markov bound:

≤ 86.7%...

In reality (suppose you research more):

≤ 30.1%...

...of Bhutan has ≥100 h.p.

2. $P(|X - 86.7| \geq 25) \leq \frac{405.62}{625} \approx 0.6490$

Chebyshev bound:

≤ 64.90%...

In reality (suppose you research more):

≤ 20.61%...

...of Bhutan has ≥ 111.7 or ≤ 61.7 h.p.

Both inequalities can give very loose bounds,
but they make <u>no assumptions</u> at all about form or distribution of X!

Andrey Andreyevich Markov and Pafnuty Chebyshev

Andrey Andreyevich Markov (1856–1922) was a Russian mathematician.

Andrei Markov, Russian-Canadian pro ice hockey player

Things named after him:

Markov's Inequality, Markov Chains, Hidden Markov Models, Markov Decision Processes, Markov Blanket…

- Markov Chain is the basis for Google's PageRank algorithm
- Also good for reinforcement learning (e.g., robots traveling worlds, simple games)

Pafnuty Lvovich Chebyshev (1821–1894) was also a Russian mathematician.

Vint Cerf, one of "the fathers of the Internet"

- Chebyshev's Inequality is named after him (but actually formulated by colleague Irénée-Jules Bienaymé)
- He was Markov's doctoral advisor (and sometimes credited with first deriving Markov's inequality)
- There is a crater on the moon named in his honor

Jensen's Inequality

Jensen's inequality

Jensen's inequality:

If $g(x)$ is a convex function, then $E[g(X)] \geq g(E[X])$.

Johan Ludwig William
Valdemar Jensen
Danish mathematician
(1859–1925)

Dr. Eggman
from *Sonic the Hedgehog*?

Jensen's inequality

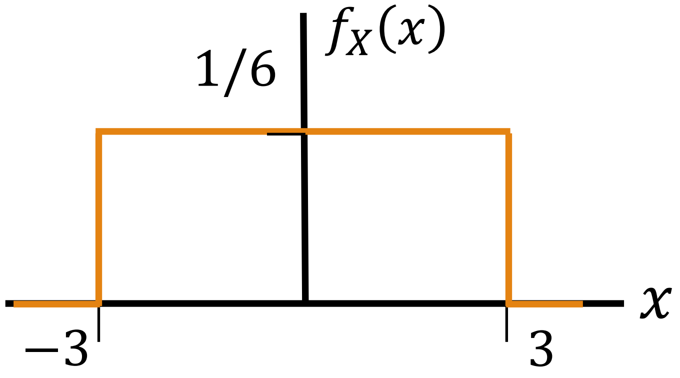
Jensen's inequality:

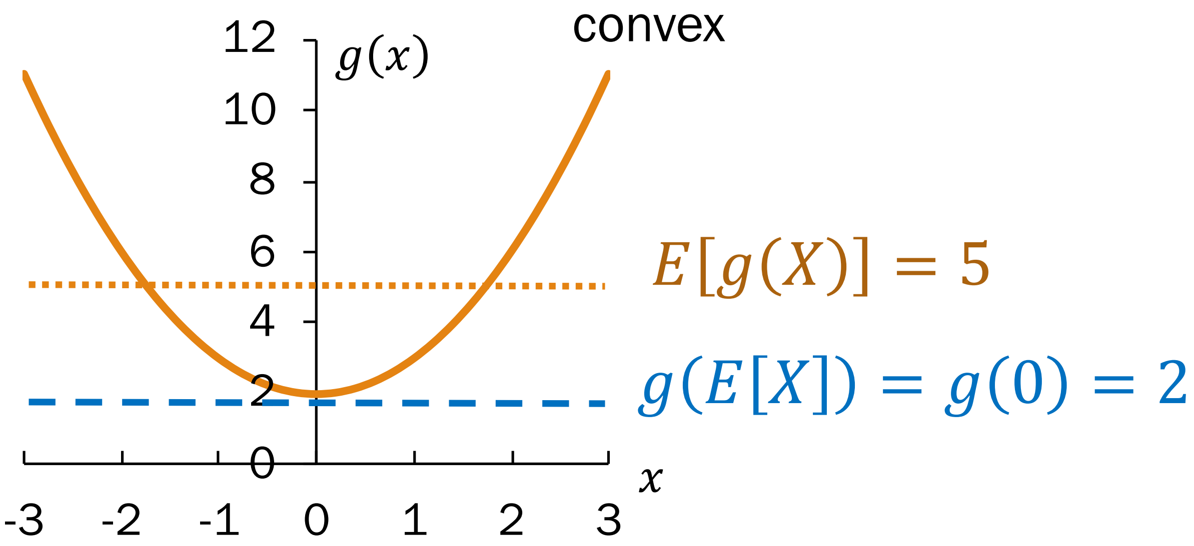
If $g(x)$ is a convex function , then $E[g(X)] \geq g(E[X])$.

def convex function $g(x)$: if $g''(x) \geq 0$ for all x. (Convex = "bowl")

def concave function $g(x)$: if $-g(x)$ is convex.

Let $X \sim \text{Uni}(-3, 3)$.
Define $g(X) = X^2 + 2$.



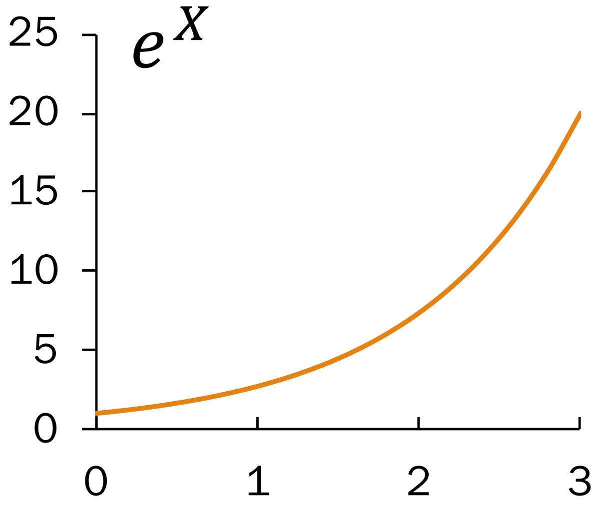


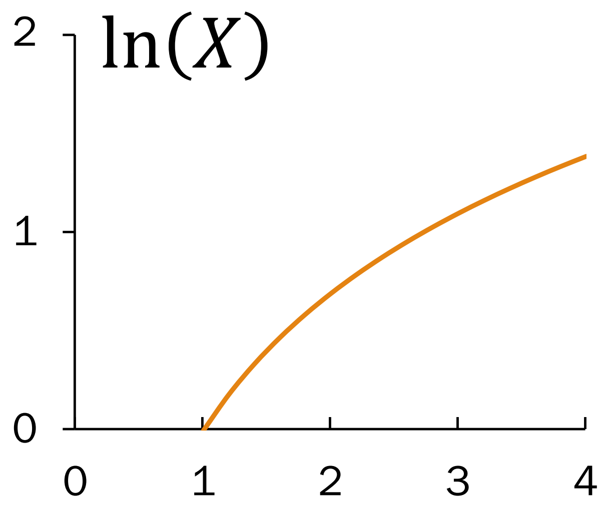
Jensen's quick check

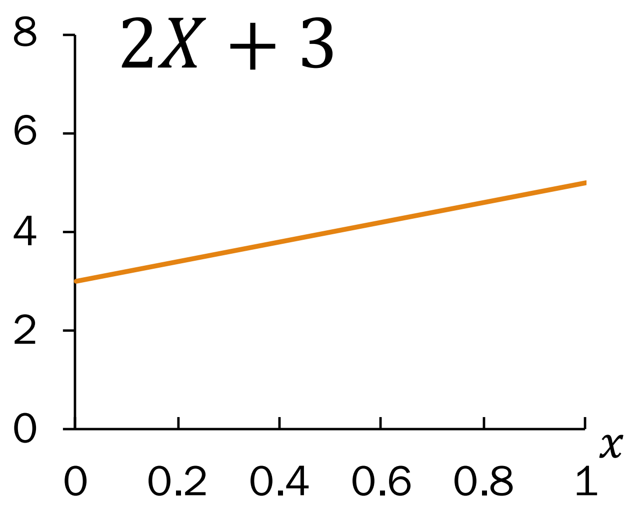
$g(x)$ is convex, $\forall x : g''(x) \geq 0$ $\Rightarrow$ $E[g(X)] \geq g(E[X])$

Let $X \sim$Uniform for the domain of each below graph.

Compare $E[g(X)]$ and $g(E[X])$: (>, <, =)

1.


2.


3.


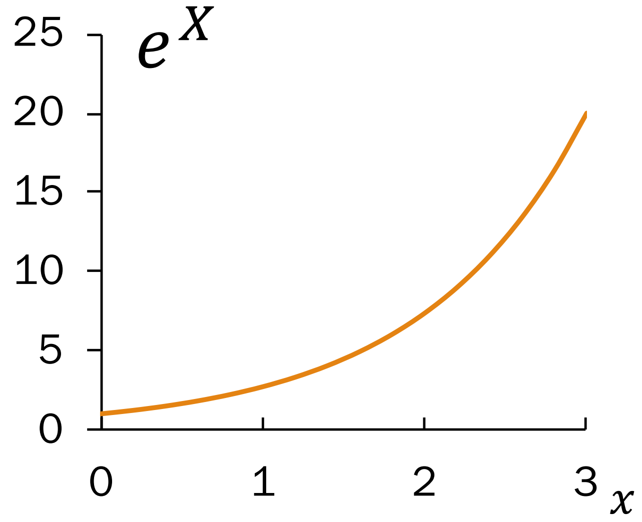
Jensen's quick check

$g(x)$ is convex, $\forall x: \; g''(x) \geq 0$ ⇒ $E[g(X)] \geq g(E[X])$

Let X~Uniform for the domain of each below graph.

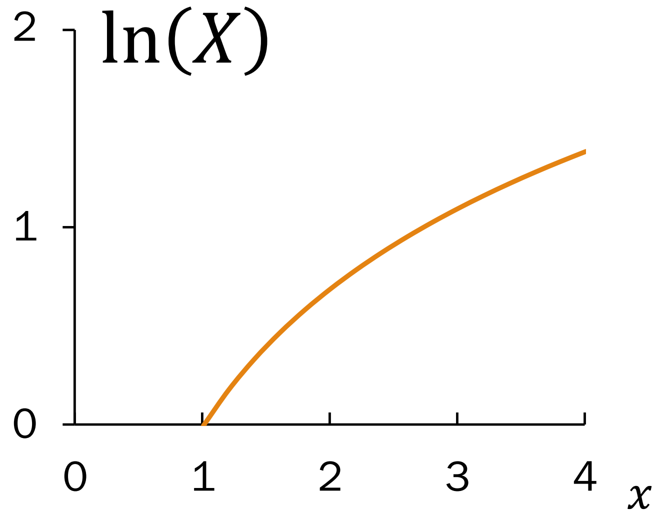
Compare $E[g(X)]$ and $g(E[X])$:   (>, <, =)

1.



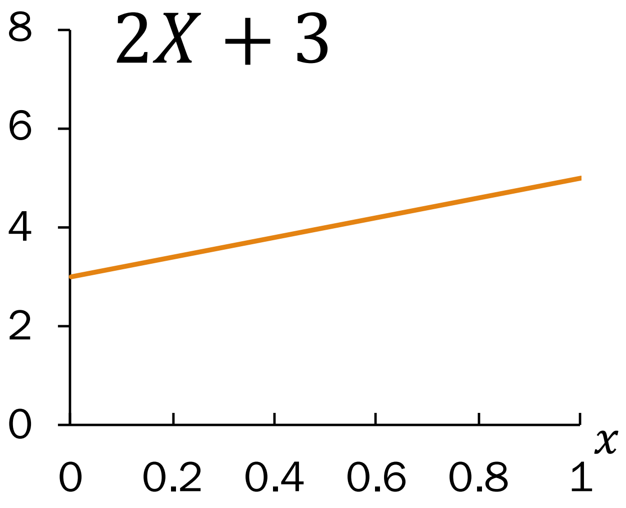
$E[e^X] > e^{E[X]}$

2.



$E[\ln(X)] < \ln(E[X])$

3.



$E[2X + 3] = 2E[X] + 3$

g is both concave and convex only if it is linear.
$E[g(X)] = g(E[X])$ only if $g(x)$ is a linear function.

Why Jensen's is useful

$g(x)$ is convex, $\forall x: \ g''(x) \geq 0$ $\Rightarrow$ $E[g(X)] \geq g(E[X])$

$$SE = \sqrt{\frac{S^2}{n}}$$

Is Standard Error an unbiased estimator?

$$E[S^2] = \sigma^2$$

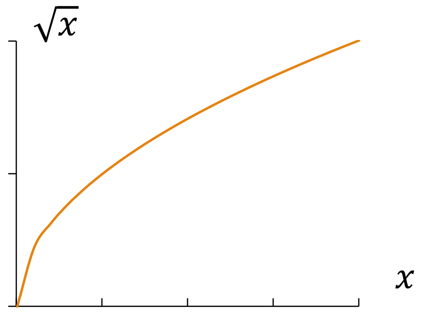
S^2 is an unbiased estimate of σ^2

$$E[S^2/n] = \sigma^2/n$$

Linearity of expectation

$$E\left[\sqrt{S^2/n}\right] < \sqrt{\sigma^2/n}$$

No; on average, it underestimates.



Square root is concave

Jensen's Inequality also used in:

- CS229, EM algorithm: How do we iteratively find the the maximum likelihood or MAP estimates without performing gradient ascent?
- CS228, KL divergence

Laws of Large Numbers

In the limit: What the CLT tells us

Review

$E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$, where X_i i.i.d.

As $n \to \infty$, $\quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Central Limit Theorem

As $n \to \infty$,

- The sample mean $\bar{X}$ **on average** is the population mean μ.
- Often $\bar{X}$ will not be exactly μ; it has a standard deviation of $\sigma/\sqrt{n}$ from μ.

Can we write a probabilistic claim on how close $\bar{X}$ is to μ?
(yes, with the **laws of large numbers**!)

Weak Law of Large Numbers

$E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$, where X_i i.i.d.

$$\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \varepsilon) \longrightarrow 0 \quad \text{for any } \varepsilon \geq 0$$

<u>Interpret</u> As our sample size grows to infinity, it is extremely unlikely that $\bar{X}$ deviates by $\geq \varepsilon$ from the population mean μ.

<u>Proof</u>

1. $P(|\bar{X} - E[\bar{X}]| \geq \varepsilon) \leq \dfrac{\text{Var}(\bar{X})}{\varepsilon^2}$ (Chebyshev's Inequality)
2. $P(|\bar{X} - \mu| \geq \varepsilon) \leq \dfrac{\sigma^2}{n\varepsilon^2}$ (Sum of i.i.d. RVs: $\text{Var}(\bar{X}) = \sigma^2/n$)
3. $0 \leq P(|\bar{X} - \mu| \geq \varepsilon) \leq \dfrac{\sigma^2}{n\varepsilon^2}$ (Probability is a number b/t 0 and 1)
4. $\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \varepsilon) = 0$ ($\lim_{n\to\infty} \sigma^2/(n\varepsilon^2) = 0$)

Strong Law of Large Numbers

$E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$, where X_i i.i.d.

$$P\left(\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu\right) = 1$$

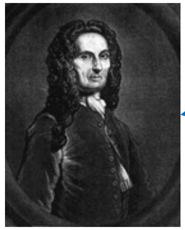
<u>Interpret</u> As our sample size grows to infinity, $\bar{X}$ will approach the population mean μ **with probability 1**.

- "with probability 1": All outcomes that aren't in this event have probability 0. Read more: https://en.wikipedia.org/wiki/Convergence_of_random_variables#Almost_sure_convergence
- Strong Law $\Longrightarrow$ Weak Law, but not vice versa
- Also implies that for any $\varepsilon > 0$, there are a *finite number* of values of n such that Weak Law condition $|\bar{X} - \mu| \geq \varepsilon$ holds

History of LLN and CLT

Central Limit Theorem

1700

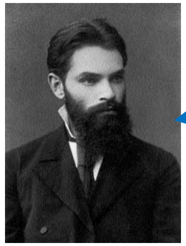


1733: CLT for $X \sim \text{Ber}(1/2)$
Abraham de Moivre

1800

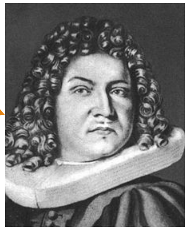
1823: CLT for $\text{Bin}(n, p)$
Pierre-Simon Laplace

1900



1901: Proof of general CLT
Alexandr Lyapunov

Law of Large Numbers



1713: Weak LLN described by Jacob Bernoulli

1835: Poisson calls it “La Loi des Grands Nombres”
(French for “Law of Large Numbers”)

1909: Émile Borel develops Strong LLN for Bernoulli

1928: Andrei Nikolaevich Kolmogorov proves general Strong LLN

Takeaways of LLN

1. **Frequentist definition** of probability

For event E, $P(E) = \lim_{n\to\infty} \frac{n(E)}{n}$

- Define X_i as 1 if E occurs on i-th trial (0 otherwise). $\mu = E[X_i] = P(E)$
- By definition, $X_1 + \cdots + X_n = n(E)$ (# of times E observed), and $\bar{X} = n(E)/n$ (fraction of times E observed)
- By SLLN, $P\left(\lim_{n\to\infty}(\bar{X}) = \mu\right) = 1 \implies P\left(\lim_{n\to\infty}\left(\frac{n(E)}{n}\right) = P(E)\right) = 1$

2. Common misconception (**The Gambler's Fallacy**)

I'm due for a win...!

- LLN only guarantees expectation μ at infinity
- Consider being due for a heads after repeated coin flips

https://en.wikipedia.org/wiki/Gambler%27s_fallacy