MATH 5010 –Foundations of Statistical Theory and Probability

❖ Inequalities and Identities

Instructor: He Wang
Department of Mathematics
Northeastern University

Probability Inequalities

- Boole’s inequality
- Bonferroni inequalities

- Markov’s inequality
- Chebyshev’s inequality

- Chernoff bounds
- Hoeffding’s inequality

- Cauchy-Schwarz inequality
- Hölder's Inequality
- Minkowski’s Inequality

- Jensen’s inequality

Introduction

"Probability inequalities" provide **bounds** on the probabilities of certain events, especially those involving deviations of random variables from their expected values.

They allow us to understand the likelihood of events **without** needing exact calculations or knowledge of the underlying probability distribution

These inequalities are widely used in theoretical analysis, limit theorems, and applications such as statistical learning, algorithm analysis, and risk assessment.

Boole's inequality

Suppose$(S, \mathcal{B}, P)$ is a probability space, and $E_1, E_2, \ldots \in F$ are events. Then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i)$$

Boole's inequality says that the probability that **at least** one of the events happens is **no greater** than the **sum** of the probabilities of the events in the collection.

This is from a generalize of the property:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Inclusion-Exclusion formula

Suppose$(S, \mathcal{B}, P)$ is a probability space, and $E_1, E_2, \ldots, E_n \ \in \ F$ are events.

Define: $$S_1 := \sum_{i=1}^{\infty} P(E_i)$$

$$S_2 := \sum_{i<j}^{\infty} P\big(E_i \cap E_j\big)$$

$$S_k := \sum_{i_1<\cdots<i_k}^{\infty} P\big(E_{i_1} \cap \cdots \cap E_{i_k}\big)$$

Then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = S_1 - S_2 + S_3 - \cdots + (-1)^{n-1} S_n$$

Bonferroni inequalities

Suppose$(S, \mathcal{B}, P)$ is a probability space, and $E_1, E_2, \dots, E_n \ \in \ F$ are events.

Boole's inequality can be extended to get lower and upper bounds on the probability of unions of events.

For **odd** $k \in \{1,2, \dots, n\}$

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{j=1}^{k} (-1)^{j-1} S_j$$

For **even** $k \in \{1,2, \dots, n\}$

$$P\left(\bigcup_{i=1}^{n} E_i\right) \geq \sum_{j=1}^{k} (-1)^{j-1} S_j$$

Example:

Place n distinguishable balls into m distinguishable boxes at random ($n > m$).

Let E be the event that a box is empty.

Goal: Find $P(E)$.

The sample space can be described as

$$S = \{w = (w_1, \dots, w_n) \mid 1 \leq w_i \leq m\}$$

So, $P(w) = \frac{1}{m^n}$

Denote $E_l \coloneqq \{w \mid w_i \neq l \text{ for all } i = 1, \dots, n\}$ for $l = 1,2, \dots, m$.

Then $E = E_1 \cup E_2 \cup \cdots \cup E_m$

Question: What is E_m?

We can see that for any m we have

$$\mathbb{P}\left(E_{i_1} \cup \ldots \cup E_{i_k}\right) = \frac{(m-k)^n}{k^n} = \left(1 - \frac{k}{m}\right)^n$$

By inclusion-exclusion property

$$\mathbb{P}(E) = m\left(1 - \frac{1}{m}\right)^n - \binom{m}{2}\left(1 - \frac{2}{m}\right)^n + \ldots + (-1)^{m-2}\binom{m}{m-1}\left(1 - \frac{m-1}{m}\right)^n$$

This expression is quite complicated.

But if we use Bonferroni inequalities we see that

$$m\left(1 - \frac{1}{m}\right)^n - \binom{m}{2}\left(1 - \frac{2}{m}\right)^n \leqslant \mathbb{P}(E) \leqslant m\left(1 - \frac{1}{m}\right)^n$$

This gives a good estimate when n is large compared to m

For example, if $m = 10$ then $10 \cdot (0.9)^n - 45 \cdot (0.8)^n \leqslant \mathbb{P}(E) \leqslant 10 \cdot (0.9)^n$

In particular, for $n = 50$, then $45(0.8)^{50} = 0.00064226146$, which is the difference between the left and right sides of the estimates. This gives a rather good estimate.

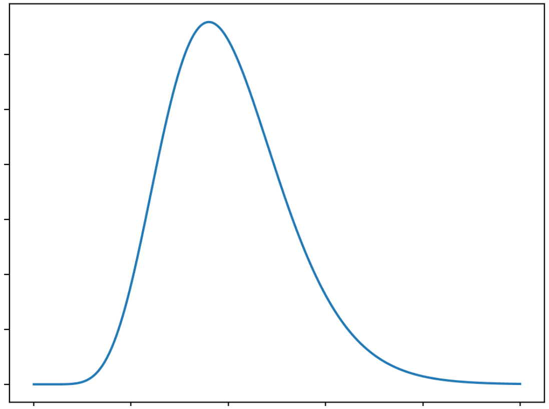
Question: given a random variable Z with expectation $E[Z]$,

How likely is Z to be close to its expectation?

How close is it likely to be?

The next few inequalities give tools for computing bounds of

$P(Z \leq E[Z] - t)$ and $P(Z \geq E[Z] + t)$



- **Markov's inequality**

Theorem (Markov's inequality)

Let X be a non-negative random variable. Then, for any $a > 0$,

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

Proof: (for continuous X with pdf $f(x)$)

$$\begin{aligned}
E[X] &= \int_0^\infty x f(x)\, dx \\
&= \int_0^a x f(x)\, dx + \int_a^\infty x f(x)\, dx \\
&\geq \int_a^\infty x f(x)\, dx \\
&\geq \int_a^\infty a f(x)\, dx \\
&= a \int_a^\infty f(x)\, dx \\
&= a P\{X \geq a\}
\end{aligned}$$

General Markov's (Chebychev's) inequality

More generally, let $g(x)$ be a non-negative function. For any $r > 0$,

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r}$$

Proof is the same as Markov's inequality.

Example,

Suppose $X \sim Binom\ (n, p)$. Then $E(X) = np$. By Markov's inequality

For $p < q < 1$,

$$P(X \geq qn) \leq \frac{E(X)}{qn} = \frac{p}{q}$$

Example: (Universal bound for deviation $|X - \mu|$)

Let X be a random variable with mean μ and variance σ^2.

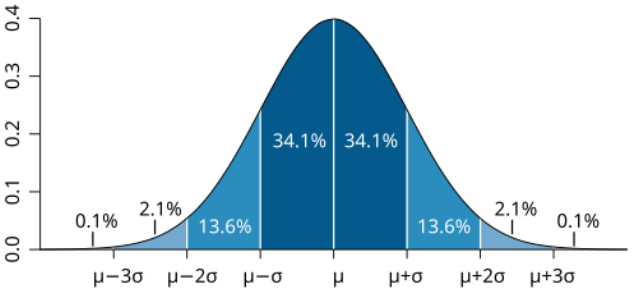
Let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$

$$P\left(\frac{(X-\mu)^2}{\sigma^2} \geq t^2\right) \leq \frac{1}{t^2}\, E\left(\frac{(X-\mu)^2}{\sigma^2}\right) = \frac{1}{t^2}$$

So, $P(|X-\mu| \geq t\sigma) \leq \frac{1}{t^2}$ and $P(|X-\mu| < t\sigma) \geq 1 - \frac{1}{t^2}$

For example, when $t = 2$

$$P(|X-\mu| \geq 2\sigma) \leq 0.25$$



So there is at least a 75% chance that a random variable will be within 2σ of its mean (no matter what the distribution of X).

- **Chebyshev's inequality**

Essentially all other bounds on the probabilities

$$P(Z \leq E[Z] - t) \qquad \text{and} \qquad P(Z \geq E[Z] + t)$$

are variations on Markov's inequality.

The first variation uses second moments-the variance-of a random variable rather than simply its mean, and is known as Chebyshev's inequality.

- **Chebyshev’s inequality**

Theorem (Chebyshev’s inequality)

Let X be a random variable with finite variance. Then for any $k > 0$,

$$P(|X - E[X]| \geq k) \leq \frac{Var(X)}{k^2}$$

This results shows that the difference between a random variable and its expectation is controlled by its variance.

Informally we can say that it shows how far the random variable is from its mean on average.

Proof:

Let $Y = \left(X - E(X)\right)^2$ be an non-negative random variable.

Apply Markov's inequality to Y. For $k \geq 0$,

$$P\left(\left(X - E(X)\right)^2 \geq k^2\right) \leq \frac{E\left[\left(X - E(X)\right)^2\right]}{k^2}$$

So,

$$P(|X - E(X)| \geq k) \leq \frac{E\left[\left(X - E(X)\right)^2\right]}{k^2} = \frac{Var(X)}{k^2}$$

Example,

Suppose $X \sim Binom\,(n,p)$. Then $E(X) = np$. By Markov's inequality

For $p < q < 1$,

$$P(X \geq qn) \leq \frac{E(X)}{qn} = \frac{p}{q}$$

Now, By Chebyshev's inequality with $k = (q-p)n$

$$P(X \geq qn) = P(X - np \geq (q-p)n)$$

$$\leq P(|X - np| \geq (q-p)n)$$

$$\leq \frac{Var(X)}{\big((q-p)n\big)^2} = \frac{p(1-p)n}{\big((q-p)n\big)^2} = \frac{p(1-p)}{(q-p)^2 n}$$

Chernoff Bounds

Recall moment generating function $M_X(t) \coloneqq E[\exp(tX)]$.

Chernoff bounds use of moment generating functions in an essential way to give exponential deviation bounds.

Theorem: Let X be a random variable with moment generating function $M_X(t)$. For any $a \in \mathbb{R}$,

$$P(X \geq a) \leq \min_{t \geq 0} e^{-ta} M_X(t)$$

$$P(X \leq a) \leq \min_{t \geq 0} e^{-ta} M_X(t)$$

Proof: Apply Markov's inequality to e^{tX}.

Example,

Suppose $X \sim Binom\ (n, p)$. Then $E(X) = np$.

$$M_X(t) = \left(pe^t + (1-p)\right)^n$$

The Chernoff bounds give, for $p < q < 1$,

$$P(X \geq qn) \leq \min_{t \geq 0} e^{-tqn}\left(pe^t + (1-p)\right)^n$$

By calculus, the minimum of $g(t) \coloneqq e^{-tqn}\left(pe^t + (1-p)\right)^n$ on $(0, \infty)$ is achieved at t_* such that

$$e^{t_*} = \frac{q(1-p)}{(1-q)p}$$

So, the minimum value is

$$
\begin{aligned}
g\left(t_*\right) &= \left(\frac{q\left(1-p\right)}{\left(1-q\right)p}\right)^{-qn}\left(p\frac{q\left(1-p\right)}{\left(1-q\right)p}+\left(1-p\right)\right)^{n} \\
&= \left(\frac{q\left(1-p\right)}{\left(1-q\right)p}\right)^{-qn}\left(\frac{1-p}{1-q}\right)^{n} = \left(\frac{p}{q}\right)^{qn}\left(\frac{1-p}{1-q}\right)^{-qn}\left(\frac{1-p}{1-q}\right)^{n} \\
&= \left(\frac{p}{q}\right)^{qn}\left(\frac{1-p}{1-q}\right)^{(1-q)n}.
\end{aligned}
$$

Thus the Chernoff bound gives

$$
P(X \geq qn) \leq \left(\frac{p}{q}\right)^{qn}\left(\frac{1-p}{1-q}\right)^{(1-q)n}
$$

Comparison of Markov's, Chebyshev's inequalities and Chernoff bounds

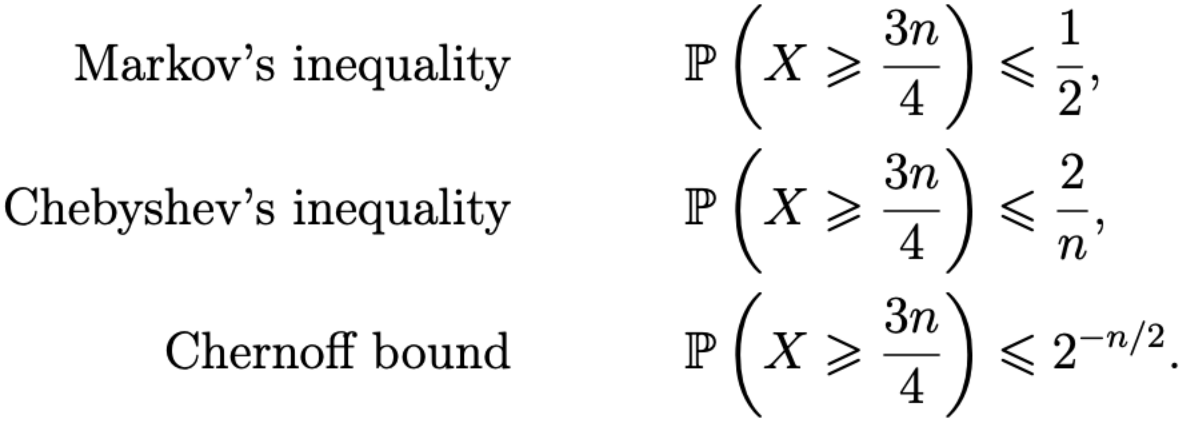
Suppose $X \sim Binom\,(n, p)$.

$$\begin{array}{rl} \text{Markov's inequality} & \mathbb{P}\left(X \geqslant qn\right) \leqslant \dfrac{p}{q}, \\ \text{Chebyshev's inequality} & \mathbb{P}\left(X \geqslant qn\right) \leqslant \dfrac{p\left(1-p\right)}{\left(q-p\right)^2 n}, \\ \text{Chernoff bound} & \mathbb{P}\left(X \geqslant qn\right) \leqslant \left(\dfrac{p}{q}\right)^{qn} \left(\dfrac{1-p}{1-q}\right)^{(1-q)n}. \end{array}$$

Markov's inequality gives a bound independent of n.
Chernoff bound is the strongest with exponential convergence to 0 as $n \to \infty$.
For example, for $p = 1/2$ and $q = 3/4$ we have

$$\begin{array}{rl} \text{Markov's inequality} & \mathbb{P}\left(X \geqslant \dfrac{3n}{4}\right) \leqslant \dfrac{2}{3}, \\ \text{Chebyshev's inequality} & \mathbb{P}\left(X \geqslant \dfrac{3n}{4}\right) \leqslant \dfrac{4}{n}, \\ \text{Chernoff bound} & \mathbb{P}\left(X \geqslant \dfrac{3n}{4}\right) \leqslant \left(\dfrac{16}{27}\right)^{n/4}. \end{array}$$

For example, for $p = 1/3$ and $q = 2/3$ we have



$$\begin{array}{rl} \text{Markov's inequality} & \mathbb{P}\left(X \geqslant \frac{3n}{4}\right) \leqslant \frac{1}{2}, \\ \text{Chebyshev's inequality} & \mathbb{P}\left(X \geqslant \frac{3n}{4}\right) \leqslant \frac{2}{n}, \\ \text{Chernoff bound} & \mathbb{P}\left(X \geqslant \frac{3n}{4}\right) \leqslant 2^{-n/2}. \end{array}$$

Chernoff bound is a sharper bound than the first- or second-moment-based tail bounds such as Markov's inequality or Chebyshev's inequality.

More examples:

1. $Z \sim Normal(0, \sigma^2)$

2. S random sign variable. $P(S) = \begin{cases} 1/2 & X = 1 \\ 1/2 & X = -1 \\ 0 & \end{cases}$

3. $X = S_1 + \cdots + S_n$

Chernoff bounds for summations

The important property is that Chernoff bounds **compatible** with summations, which is a consequence of the moment generating function.

Assume that Z_i are independent. Then we have that

$$M_{Z_1+\cdots+Z_n}(t) = \prod_{i=1}^{n} M_{Z_i}(t)$$

This means that when we calculate a Chernoff bound of a sum of i.i.d. variables, we need only calculate the moment generating function for one of them.

Assume $E(Z_i) = 0$, then

$$P\left(\sum_{i=1}^{n} Z_i \geq a\right) \leq \frac{\prod_{i=1}^{n} E[\exp(tZ_i)]}{e^{at}} = (E[e^{tZ_i}])^2 e^{-at}$$

Hoeffding's inequality -Concentration of mean

Recall that Chebyshev's inequality

$$P(|X - E[X]| \geq \epsilon) \ \leq \frac{Var(X)}{\epsilon^2}$$

Suppese $X_1, \ldots, X_n$ is an IID sequence with $Var(X_n) = \sigma^2$.
Apply Chebyshev's inequality to sample mean $\bar{X}_n$:

$$P(|\bar{X}_n - E[\bar{X}_n]| \geq \epsilon \ ) \leq \frac{\sigma^2}{n\epsilon^2}$$

Theorem (Hoeffding's inequality)

Suppose $0 \leq X_n \leq 1$, then for any $\epsilon > 0$,

$$P(|\bar{X}_n - E[\bar{X}_n]| \geq \epsilon \ ) \leq 2e^{-2n\epsilon^2}$$

Result can be scaled to $a \leq X_n \leq b$

Applications:

Hoeffding’s inequality gives a concentration of the order of exponential (actually it is often called a Gaussian rate) so the convergence rate is much faster than the one given by the Chebyshev’s inequality.

Obtaining such an exponential rate is useful for analyzing the property of an estimator.

Many modern statistical topics, such as high-dimensional problem, nonparametric inference, semi-parametric inference, and empirical risk minimization all rely on a convergence rate of this form

- **Cauchy-Schwarz inequality**

The Cauchy-Schwarz inequality is an application of linear algebra.

Theorem: Suppose $X \ and \ Y$ are two random variables, then

$$E(XY)^2 \leq EX^2 \cdot EY^2$$

and the equality holds if and only if $X = aY$ for some constant a ∈ R

Define **inner product** on *vector space of random variables*:

$$\langle X, Y \rangle \coloneqq E(XY)$$

Cauchy-Schwarz inequality can be easily generalized to random vectors $\vec{X}$ and $\vec{Y}$.

$$\langle \vec{X}, \vec{Y} \rangle \coloneqq E\left(\vec{X}^T \vec{Y}\right) = E(X_1 Y_1 + \cdots + X_n Y_n)$$

Example:

Use Cauchy-Schwartz inequality to prove $Coor(X,Y) \leq 1$. Moreover, equality hold if and only if there are constants $a, b \in \mathbb{R}$ such that $X = a + bY$.

Use normalized random variables

$$U := \frac{X - E(X)}{\sqrt{Var(X)}} \qquad V := \frac{Y - E(Y)}{\sqrt{Var(Y)}}$$

Use Cauchy-Schwartz inequality

$$|E[UV]| \leq \sqrt{E[U]^2 E[V]^2} = 1$$

$$Coor(X,Y) = E[UV]$$

In addition, equality only if $U = kV$, which is

$$X = k\sqrt{Var(X)}\left(\frac{Y - E(Y)}{\sqrt{Var(Y)}}\right) + E(X) = a + bY$$

Hölder's Inequality and Minkowski’s

Theorem: (Hölder's Inequality). Let $X\ and\ Y$ be any two random variables, and let $p\ and\ q$ satisfy $\frac{1}{p}+\frac{1}{q}=1$. Then,

$$|E[XY]| \leq \boldsymbol{E}|\boldsymbol{XY}| \leq (E|X|^{p})^{1/p}(E|Y|^{q})^{1/q}$$

Cauchy-Schwarz inequality is a special case when $p=q=2$

Theorem: (Minkowski's Inequality) Let $X\ and\ Y$ be any two random variables.

For $1 \leq p < \infty$

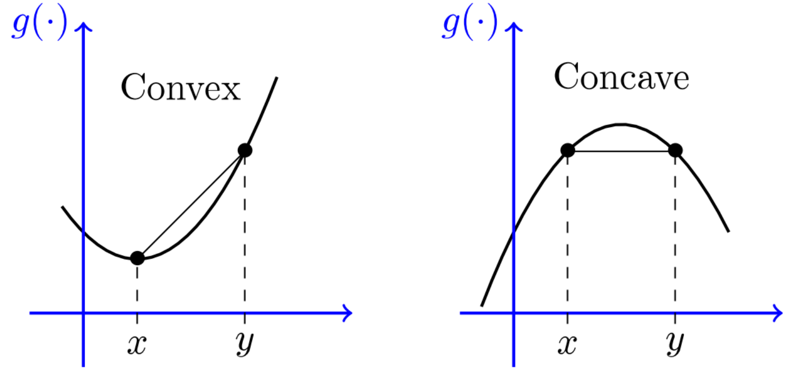
$$E[|X+Y|^{p}]^{1/p} \leq E[|X|^{p}]^{\frac{1}{p}} + E[|X|^{q}]^{1/q}$$

➢ Jensen’s inequality

Jensen’s inequality relates the value of a convex (or concave) function of an expectation to the expectation of that function.

A function $g : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** on $[a, b]$ if for each $x, y \in [a, b]$ and each $\lambda \in [0, 1]$ we have

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



A function g is **concave** if $-g$ is convex.

For a convex function g, the property holds for any *convex linear combination* of points in $[a, b]$, that is

$$g\ (\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq\ \lambda_1 g\ (x_1) + \cdots + \lambda_n g\ (x_n)$$

$$\lambda_1 + \cdots + \lambda_n = 1$$

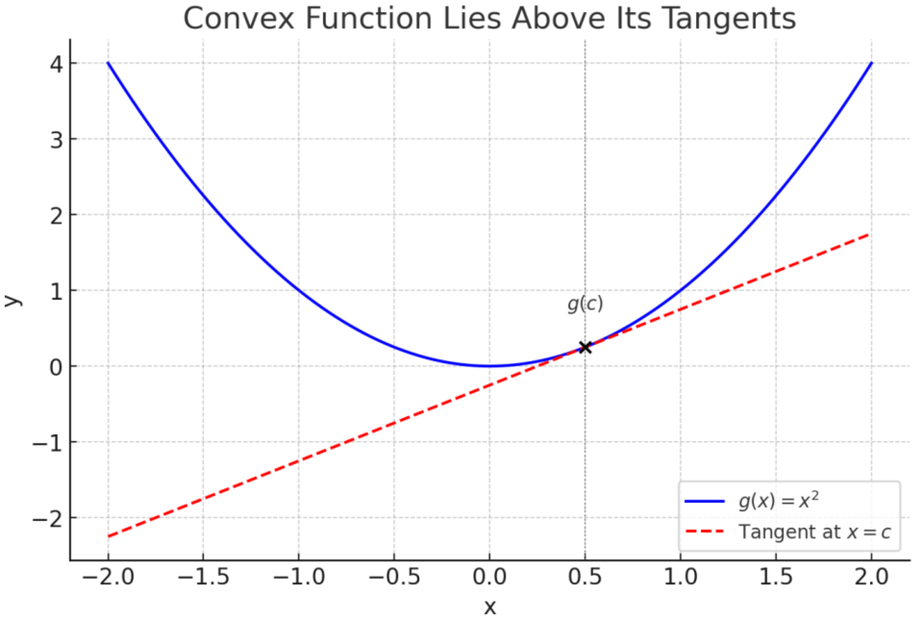
$$0 \leq \lambda_1, \dots, \lambda_n \leq 1$$

$$x_1, \dots, x_n \in [a, b]$$

If g is twice differentiable, then, g is convex if $g''\ (x)\ >\ 0$ for all $x\ \in\ [a, b]$

Theorem(Convex functions lie above tangents)

Suppose $a < c < b$ and $g : [a, b] \to \mathbb{R}$ be convex. Then there exist $A, B \in \mathbb{R}$ such that $g(c) = Ac + B$ and for all $x \in [a, b]$ we have $g(x) > Ax + B$.



Jensen’s inequality

Theorem: Suppose X is a random variable such that $P\ (a \le X \le b)\ =\ 1.$

If $g : \mathbb{R}\ \to \mathbb{R}$ is convex on $[a, b]$, then

$$E[g(X)] \ge g(E[X])$$

If $g : \mathbb{R}\ \to \mathbb{R}$ is concave on $[a, b]$, then

$$E[g(X)] \le g(E[X])$$

Proof: Suppose $a < E[X] < b$, denote $c = E[X]$, then

$$g(X) \ge AX + b \text{ and } g(E[X]) = AE[X] + B \text{ for some } A, B \in \mathbb{R}$$

So,

$$g(E[X])\ =\ AE[X] + B = E\ (AX + B) \le\ Eg\ (X)$$

Example: (Arithmetic-geometric mean inequality)

Suppose $a_1, \ldots, a_n$ are positive numbers, and X is a discrete random variable with the mass density

$$f_X(a_k) = \frac{1}{n} \text{ for } k = 1, \ldots, n$$

Note that the function $g\ (x)\ =\ -\log x$ is a convex function on $(0, \infty)$. Jensen's inequality gives that

$$-\log\left(\frac{1}{n}\sum_{k=1}^{n} a_k\right) = -\log(E[X]) \leq E[-\log X] = -\frac{1}{n}\sum_{k=1}^{n} \log a_k$$

Exponentiating this we get

$$\frac{1}{n}\sum_{k=1}^{n} a_k \geq \left(\prod_{k=1}^{n} a_k\right)^{1/n}$$

Example

Suppose $p > 1$, then the function $g\ (x) = |x|^p$ is convex.

$$E[|X|^p] \geq |E[X]|^p$$

for any random variable X such that $E[X]$ is defined.

For $p = 2$,

$$E[X^2] \geq (E[X])^2$$

So, $E[X^2] - (E[X])^2 \geq 0$

Equality Relations:

For some specific distributions, like Normal, Chi Squared, Poisson, Negative binomial, there exist **equalities relations:** (See [CB] Section 3.6 for more detail.)

1. If $X \sim Poisson(\lambda)$, $\quad P(X = x+1) = \frac{\lambda}{x+1} P(X = x)$,

2. *Let* $X_{\alpha,\beta}$ *denote a* $gamma(\alpha, \beta)$ *random variable, with pdf* $f(x|\alpha, \beta)$, *where* $\alpha > 1$. *Then for any constants a and b,*

$$P(a < X_{\alpha,\beta} < b) = \beta\left(f(a|\alpha,\beta) - f(b|\alpha,\beta)\right) + P(a < X_{\alpha-1,\beta} < b).$$

3. Stein's Lemma: *Let* $X \sim \mathrm{n}(\theta, \sigma^2)$, *and let g be a differentiable function satisfying* $\mathrm{E}\,|g'(X)| < \infty$. *Then*

$$\mathrm{E}\left[g(X)(X-\theta)\right] = \sigma^2 \mathrm{E} g'(X).$$

Example Stein's Lemma makes calculation of higher-order moments quite easy. For example, if $X \sim Normal\ (\theta, \sigma^2)$, then

$$\begin{aligned}
\mathrm{E}X^3 &= \mathrm{E}X^2(X - \theta + \theta) \\
&= \mathrm{E}X^2(X - \theta) + \theta \mathrm{E}X^2 \\
&= 2\sigma^2 \mathrm{E}X + \theta \mathrm{E}X^2 \qquad (g(x) = x^2, g'(x) = 2x) \\
&= 2\sigma^2\theta + \theta(\sigma^2 + \theta^2) \\
&= 3\theta\sigma^2 + \theta^3.
\end{aligned}$$

4. *Let χ_p^2 denote a chi squared random variable with p degrees of freedom.*

If expectations exist, For any function $h(x)$.

$$\mathrm{E}h(\chi_p^2) = p\mathrm{E}\left(\frac{h\left(\chi_{p+2}^2\right)}{\chi_{p+2}^2}\right)$$

5. **(Hwang)** *Let $g(x)$ be a function with $-\infty < \mathrm{E}g(X) < \infty$ and $-\infty < g(-1) < \infty$.*

If $X \sim \text{Poisson}(\lambda)$,

$$\mathrm{E}\left(\lambda g(X)\right) = \mathrm{E}\left(Xg(X-1)\right).$$

If $X \sim$ **negative binomial**(r, p),

$$\mathrm{E}\left((1-p)g(X)\right) = \mathrm{E}\left(\frac{X}{r+X-1}g(X-1)\right).$$

Example: If $X \sim Poisson(\lambda)$, take $g(x) = x^2$, use above result,

$$\begin{aligned}\mathrm{E}(\lambda X^2) &= \mathrm{E}\left(X(X-1)^2\right)\\ &= \mathrm{E}(X^3 - 2X^2 + X).\end{aligned}$$

$$\begin{aligned}\mathrm{E}X^3 &= \lambda \mathrm{E}X^2 + 2\mathrm{E}X^2 - \mathrm{E}X\\ &= \lambda(\lambda+\lambda^2) + 2(\lambda+\lambda^2) - \lambda\\ &= \lambda^3 + 3\lambda^2 + \lambda.\end{aligned}$$

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition (Section 3.6. Section 4.7)
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- **Book 3. Introduction to Probability**. C.M. Grinstead and J.L. Snell. American Mathematical Society, 2012
- **Book 4. Introduction to Probability Models**, S. Ross, 12th edition (published by Academic Press).

Online books:

https://www.probabilitycourse.com/