Basic probability primer

Jimmy Jin
UNC-Chapel Hill

Last updated: 12/12/13

Contents

1 Inequalities

1.1 Jensen's inequality

This unassuming little inequality is arguably the most important inequality in all of statistics. Indeed, as you read on further you will see that the proofs of many other inequalities can be elegantly phrased in terms of Jensen's inequality (e.g. Young's inequality in the section on Holder's inequality).

The key to Jensen's inequality is convexity–relating the expected value of a convex function to the convex function of the expected value. Therefore we first review some important properties of convex functions:

Definition. (Convex function)

Let I be an open subset of $\mathbb{R}$. Let $g : I \to \mathbb{R}$. We say g is **convex** if, $\forall x, y \in I$ and $\forall \alpha \in [0, 1]$ we have

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

Concave functions are defined analogously.

Note that convexity can be phrased in terms of some other familiar concepts.

Lemma 1.1. (Sufficient conditions for convexity)

Let I be an open subset of $\mathbb{R}$. Let $g : I \to \mathbb{R}$. If either

1. g' is nondecreasing and continuous on I, or
2. $g'' \geq 0$ on I.

then g is convex.

Here is the key property of convex functions which will give us Jensen's inequality in one effortless stroke:

Lemma 1.2. If ψ is a convex function on an open subset $I \subset \mathbb{R}$, then for any $x_0 \in I$ there exists a support line l_0 such that

$$l_0(x) \leq \psi(x) \quad \forall x \in I \quad \text{and} \quad l_0(x_0) = \psi(x_0)$$

Proof. Convexity gives us two facts:

1. For any $h > 0$, applying the definition of convexity with $\alpha = 1/2$ gives:
$$\frac{\psi(x) - \psi(x - h)}{h} \leq \frac{\psi(x + h) - \psi(x)}{h}$$
2. For any $h_1 > h_2$, applying the definition of convexity with $\alpha = h_2/h_1$ to points x and $x - h_1$ gives:
$$\frac{\psi(x) - \psi(x - h_1)}{h_1} \leq \frac{\psi(x) - \psi(x - h_2)}{h_2}$$

$$\frac{\psi(x+h_1)-\psi(x)}{h_1} \geq \frac{\psi(x+h_2)-\psi(x)}{h_2}$$

By (2) the sequences are monotone so their limits as $h \to 0^+$ exist, so define:

$$\psi'_-(x) = \lim_{h\to 0^+} \frac{\psi(x)-\psi(x-h)}{h} \quad \text{and} \quad \psi'_+(x) = \lim_{h\to 0^+} \frac{\psi(x+h)-\psi(x)}{h}$$

By (1), $\psi'_-(x) \leq \psi'_+(x)$ for any x. Then, for fixed z, we can choose some $a \in [\psi'_-(z), \psi'_+(z)]$ and define the line ℓ by:

$$l_z(x) = \psi(z) + a(x-z)$$

Clearly $\ell(z) = \psi(z)$ and by monotonicity of limits, it is easy to see that

$$l_z(x) \leq \psi(x) \quad \forall x \in I$$

□

Theorem 1.3. (Jensen's inequality)

Let g be convex on I, an open subset of $\mathbb{R}$. Let $x \in I$ with probability 1 and assume $X, g(X)$ have finite expectation. Then

$$g(EX) \leq E(g(X))$$

Proof. First note that $EX \in I$. So let $\ell(x) = ax + b$ be the support line for $g(\cdot)$ at EX. Then by the definition of support line we have

1. $\ell(EX) = g(EX)$
2. $\ell(x) \leq g(x) \quad \forall x \in I$

Taking expectations in (2) above, we obtain:

$$Eg(x) \;\geq\; E\ell(x) \;=\; E(ax+b) \;=\; aEX + b \;=\; \ell(EX)$$

Then noting (1) above, we are done.

□

Jensen's inequality has many wide-ranging uses. One unexpected use is the proof of the fact that the geometric mean of a positive series is less than or equal to the arithmetic mean of the same series:

Theorem 1.4. Let $a_1, \ldots, a_n > 0$. Define the arithmetic mean and geometric mean by

$$\text{AM} = \frac{1}{n}\sum_{i=1}^{n} a_i, \quad \text{GM} = \left(\prod_{i=1}^{n} a_i\right)^{1/n}$$

Then GM $\leq$ AM.

Proof. Let X be a discrete random variable taking values in $\{a_i, \ldots, a_n\}$, with probabilities $\mathbb{P}(X = a_i) = 1/n$.

Clearly, $\mathbb{E}X = AM$. We show that $\log \text{GM} \leq \log \text{AM}$. By Jensen,

$$\mathbb{E}(\log X) \leq \log \mathbb{E}X = \log \text{AM}$$

So we only need to show that $\mathbb{E}(\log X) = \log \text{GM}$. Note:

$$\begin{aligned} \log \text{GM} = \log \left(\prod a_i \right)^{1/n} &= \frac{1}{n} \log(\prod a_i) \\ &= \frac{1}{n} \sum \log a_i \\ &= E \log X \end{aligned}$$

□

Another use of Jensen's inequality involves a result concerning the expected value of the log likelihood ratio of a discrete random variable:

Theorem 1.5. Let $X \in S \subset \mathbb{R}$ be a discrete random variable. Let p be the pmf of X and q be any other pmf defined on S.

We consider the likelihood functions $p(X)$ and $q(X)$. Noting that these are random variables, we then have that

$$E\left[\log \frac{p(X)}{q(X)}\right] \geq 0$$

Proof. First note that

$$\log \frac{p(x)}{q(x)} \geq 0 \iff p(x) \geq q(x)$$

Assume WLOG that $p(x) > 0 \; \forall x \in S$. Then

$$E\left[\log \frac{p(X)}{q(X)}\right] = \sum_{x \in S} p(x) \log \frac{p(x)}{q(x)}$$

Note that this sum $= +\infty$ if $q(x) = 0$ for some $x \in S$. So assume that $q(x) > 0 \; \forall x \in S$. Then we have

$$\begin{aligned} \sum_{x \in S} p(x) \log \frac{p(x)}{q(x)} &= -\sum_{x \in S} p(x) \log \frac{q(x)}{p(x)} \\ &= -E\left[\log \frac{q(x)}{p(x)}\right] \\ &\geq -\log E\left[\frac{q(x)}{p(x)}\right] \quad \text{(by Jensen)} \\ &= -\log \left[\sum_{x \in S} q(x)\right] \\ &= -\log(1) = 0 \end{aligned}$$

□

1.2 Holder's inequality

Holder's inequality is an important tool which relates the expected value of a product to the product of expected values. The original statement of the inequality is more general and phrased in terms of integrals and the L_p-norm of a function. Here we present a more specific form in terms of expected value.

We shall see that Holder's inequality follows immediately from Young's inequality:

Lemma 1.6. (Young's inequality)

Let $a, b \geq 0$ and $1 < p, q < \infty$ be such that $1/p + 1/q = 1$. Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality iff $a^p = b^q$.

Proof. First note that if either or both of a or b are zero, then the result is trivial. So assume $a, b > 0$.

$$\begin{aligned} ab &= \exp(\log a + \log b) \\ &= \exp\left(\frac{1}{p}\log a^p + \frac{1}{q}\log b^q\right) \\ &\leq \frac{1}{p}\exp(\log a^p) + \frac{1}{q}\exp(\log b^q) \quad \text{(by convexity)} \\ &= \frac{1}{p}a^p + \frac{1}{q}b^q \end{aligned}$$

and since e^x is strictly convex, then equality holds only when $\log a^p = \log b^q$, i.e. when $a^p = b^q$.

□

To simplify the presentation of the statement and proof of the inequality, we first define a special case of the L_p-norm of a function:

Definition. (L_p-norm of a random variable)

For $p \geq 1$, the **L_p-norm of a random variable** X is

$$\|X\|_p = [E|X|^p]^{1/p}$$

Theorem 1.7. (Holder's inequality)

Let $1 < p, q < \infty$ satisfy the cosntraint $1/p + 1/q = 1$, and let X, Y be random variables such that $E|X|^p, E|Y|^q < \infty$. Then

$$|E(XY)| \leq E|XY| \leq \|X\|_p \|Y\|_q$$

Proof. Define

$$a = \frac{|X|}{\|X\|_p}, \quad b = \frac{|Y|}{\|Y\|_q}$$

Note that a, b are both non-negative for any $\omega \in \Omega$. Applying our lemma, we obtain

$$\frac{1}{p}\frac{|X|^p}{E|X|^p} + \frac{1}{q}\frac{|Y|^q}{E|Y|^q} \geq \frac{|XY|}{\|X\|_p \|Y\|_q}$$

Taking expectations of both sides, we obtain

$$1 = \frac{1}{p} + \frac{1}{q} \geq \frac{E|XY|}{\|X\|_p \|Y\|_q}$$

□

The well-known Cauchy-Schwarz inequality for random variables is a special case of Holder's inequality when $p = q = 2$:

Corollary 1.8. (Cauchy-Schwarz inequality)

Let X, Y be r.v.'s such that $EX^2, EY^2 < \infty$. Then

$$|E(XY)| \leq E|XY| \leq [EX^2]^{\frac{1}{2}}[EY^2]^{\frac{1}{2}}$$

1.3 Independent copies and association inequalities

The association inequalities are used in the context of two different functions of the same random variable. The inequalities relate the expected value of the product of those functions to the product of the expected value of those functions.

First we review some basic definitions and properties behind the concept of independent copies, which will be used to prove the inequalities.

Definition. (Equal in distribution)

Let X, Y be random variables. X and Y are **equal in distribution** (written $X \stackrel{d}{=} Y$) if X and Y have the same distribution.

More specifically, let $X : \Omega \to \mathbb{R}$ with probability measure P and $Y : \Omega \to \mathbb{R}$ with probability measure P'. Then $X \stackrel{d}{=} Y$ iff

1. $F_X(u) = F_y(u) \quad \forall u \in \mathbb{R}$
2. $P(X \in B) = P'(Y \in B) \quad \forall B \in \mathbb{R}$

The key point is that random variables can be equal in distribution without being exactly equal, suggesting the concept of independent copies of a random variable. To illustrate this, consider the following example:

Example. Let X and X' be independent $N(0, 1)$ random variables. Consider the new random variable Y defined by:

1. $Y = X$. Then $Y \stackrel{d}{=} X$ and the variables are **not** independent.
2. $Y = X'$. Then $Y \stackrel{d}{=} X$ but the variables **are** independent.

Furthermore, we have an important result relating the variance and covariance of random variables to the expected value of a function of those random variables and their independent copies.

Theorem 1.9. Let X, Y be random variables with $EX^2, EY^2 < \infty$. Furthermore let X' be a copy of X which is independent of both X and Y. Then

1. $\text{Var}(X) = \frac{1}{2}E(X - X')^2$
2. $\text{Cov}(X) = E(XY - X'Y)$

Proof. Not given.

□

Now the main result:

Theorem 1.10. (Association inequalities)

Let X be a random variable and let $f, g : \mathbb{R} \to \mathbb{R}$ such that $E|f(X)| < \infty$, $E|g(X)| < \infty$, and $E|f(X)g(X)| < \infty$. Then

1. If f, g are both nondecreasing (or nonincreasing), then

$$E[f(X)g(X)] \geq Ef(X)Eg(X)$$

2. If f is nondecreasing and g is nonincreasing (or v.v.), then

$$E[f(X)g(X)] \leq Ef(X)Eg(X)$$

Proof. We prove the case where f and g are both nondecreasing or both nonincreasing.

Let Y be an independent copy of X. Note that

$$0 \leq [f(x) - f(y)] \cdot [g(x) - g(y)]$$

since f and g are both nondecreasing (nonincreasing). Expanding the RHS and taking expectations, we have

$$0 \leq E[f(X)g(X)] - E[f(X)g(Y)] - E[f(Y)g(X)] + E[f(Y)g(Y)]$$

Note that since X is independent of Y and $X \stackrel{d}{=} Y$, then

1. $E[f(Y)g(Y)] = E[f(X)g(X)]$
2. $E[f(X)g(Y)] = E[f(Y)g(X)] = Ef(X)Eg(Y) = Ef(X)Eg(X)$

And the expression above reduces to

$$0 \leq 2\,E[f(X)g(X)] - 2\,Ef(X)Eg(X)$$

□

1.4 Markov and Chebyshev's inequalities

Markov's inequality, which is the basis for almost every other inequality we learn in this course, has a proof which is surprisingly simple. We shall use indicator functions to give a two-line proof of Markov's inequality.

Remark. Let $\mathbb{1}_A(X), \mathbb{1}_B(X)$ be indicator functions of a random variable X where A, B are measurable sets. Then

1. $\mathbb{1}_A(X) \cdot \mathbb{1}_B(X) = \mathbb{1}_{A \cap B}(X)$
2. $E[\mathbb{1}_A(X)] = P(X \in A)$

Theorem 1.11. (Expected value for non-negative r.v.'s)

Let $X \geq 0$ be a random variable with density function f. Then

$$EX = \int_0^\infty P(X > t)\,\mathrm{d}t$$

Proof.

$$\begin{aligned}
\int_0^\infty P(X > t)\,\mathrm{d}t &= \int_0^\infty \left[\int_t^\infty f(x)\,\mathrm{d}x\right]\,\mathrm{d}t \\
&= \int_0^\infty \left[\int_0^\infty \mathbb{1}_{[t,\infty)}(x) f(x)\,\mathrm{d}x\right]\,\mathrm{d}t \\
&= \int_0^\infty f(x)\left[\int_0^\infty \mathbb{1}_{[t,\infty)}(x)\,\mathrm{d}t\right]\,\mathrm{d}x \quad \textit{(by Fubini)} \\
&= \int_0^\infty x f(x)\,\mathrm{d}x
\end{aligned}$$

□

We now arrive at the main result. Note that this inequality assumes nothing about the random variable except for that it is non-negative.

Theorem 1.12. (Markov's inequality)

Let $X \geq 0$ be a random variable and let $t > 0$. Then

$$P(X \geq t) \leq \frac{EX}{t}$$

Proof. Note:

$$x \geq x \cdot \mathbb{1}_{(t,\infty)}(x) \geq t \cdot \mathbb{1}_{(t,\infty)}(x)$$

Taking expectations, we obtain:

$$\begin{aligned}
EX &\geq E[t \cdot \mathbb{1}_{(t,\infty)}(X)] \\
&= t \cdot P(X > t)
\end{aligned}$$

□

Markov's inequality is sometimes referred to as Chebyshev's inequality, but in this course we present Chebyshev's inequality as a separate bound for random variables **with finite variance**.

Corollary 1.13. (Chebyshev's inequality)

Let X be a random variable such that $EX^2 < \infty$ and let $t > 0$. Then

$$P(|X - EX| > t) \leq \frac{\text{Var}(X)}{t^2}$$

Proof. Apply Markov's inequality to, $P(|X - EX|^2 > t^2)$.

□

1.5 Markov's inequality variants

In this section we introduce a few inequalities that are "tweaks" of Markov's inequality, and introduce a new inequality that is very powerful in machine learning.

The first inequality is simply Markov's inequality applied to the variable $|X|^s$, which gives a bound on a r.v. in terms of its moments. This bound is valid for any variable (not necessarily non-negative) whose s^{th} moment exists:

Theorem 1.14. (Extended Markov)

Let X be any r.v. with $E|X|^s < \infty$. Then:

$$P(|X| > t) = P(|X|^s > t^s) \leq \frac{E|X|^s}{t^s}$$

Proof. Trivial.

□

This next inequality is similar to Chebyshev's inequality in that it involves the variance, but has the added advantage of never exceeding 1 (unlike Chebyshev's inequality).

Theorem 1.15. (Chebyshev-Cantelli)

For any r.v. X with $E[X^2] < \infty$ and $\forall t > 0$,

$$P(X - EX > t) \leq \frac{\text{Var}(X)}{\text{Var}(X) + t^2}$$

Proof. Assume WLOG $\mathbb{E}X = 0$. Note $t = \mathbb{E}(t - X) \leq \mathbb{E}[(t - X) \cdot \mathbb{1}_{X \leq t}]$. Applying the Cauchy-Schwarz inequality to the rightmost quantity, we get:

$$t \leq \left[E(t - X)^2 \cdot E(\mathbb{1}_{X \leq t})^2\right]^{\frac{1}{2}} = \left[(t^2 + EX^2) \cdot P(X \leq t)\right]^{\frac{1}{2}}$$

Then square both sides and rearrange.

□

Finally this last Markov variant demonstrates a key technique–that of introducing an "extra" parameter (in this case, s), obtaining an inequality for each s, and then selecting the one which gives us the tightest bound.

Theorem 1.16. (Chernoff MGF bound)

For any r.v. X and $s > 0$,

$$P(X > t) \leq \inf_{s>0} \left[e^{-st} E(e^{sx}) \right]$$

Proof. Apply Markov's to $P(e^{sx} > e^{st})$ and then take infimum over s. □

1.6 Moment generating functions

Some basic properties of moment generating functions which the reader should already be familiar with:

Lemma 1.17. Let X be a random variable. If $E[e^{s_0|x|}] < \infty$ for some $s_0 > 0$, then $E[e^{sx}] < \infty$ for all $-s_0 \leq s \leq s_0$.

Proof. Properties of the integral.

Definition. (Moment generating function)

Let X be a random variable such that $E[e^{s_0|x|}] < \infty$ for some $s_0 > 0$. Then the **moment generating function of X** is

$$\phi_X(s) = E[e^{sx}] \quad |s| \leq s_0$$

Example. (Some common MGFs)

1. $X \sim \text{Poisson}(\lambda)$: $\phi_X(s) = e^{\lambda(e^s - 1)} \quad s \in \mathbb{R}$
2. $X \sim \text{Exponential}(\lambda)$: $\phi_X(s) = \lambda/\lambda - s \quad s < \lambda$
3. $X \sim \text{Normal}(0, 1)$: $\phi_X(s) = e^{s^2/2} \quad s \in \mathbb{R}$

The moment generating function of a random variable has many nice properties and can be used in quite sophisticated ways in proofs. However, we won't go into those in depth here. We finish by presenting a few simple properties.

Remark. (Taylor expansion of MGF)

If $\phi_X(s)$ exists, then $E|X|^k < \infty \; \forall k \geq 1$ and

$$\phi_X(s) = \sum_{k \geq 0} \frac{s^k E X^k}{k!}$$

Remark. (Differentiation of the MGF)

$$\phi'_X(s) = \frac{d}{ds} E[e^{sX}] = E\left[\frac{d}{ds} e^{sX} \right] = E\left[X e^{sX} \right]$$

and, more generally,

$$\phi_X^{(k)}(s) = E\left[X^k e^{sX}\right]$$

Remark. (Properties of the MGF)

1. $\phi_X^{(k)}(0) = E[X^k]$
2. $\phi_X''(s) \geq 0 \Rightarrow \phi_X(s)$ is convex

We close this section with a brief example concerning sums of independent random variables:

Example. (Sums of independent random variables)

Let $X_1, \ldots, X_n$ be independent where each X_i has an MGF. Define $S_n = X_1 + \ldots + X_n$. Then $\phi_{S_n}(s)$ exists:

$$\begin{aligned}
\phi_{S_n}(s) = E[e^{s \cdot S_n}] &= E\left[s(X_1 + \ldots + X_n)\right] \\
&= E\left[\prod_{i=1}^{n} e^{sX_i}\right] \\
&= \prod_{i=1}^{n} E[e^{sX_i}] \\
&= \prod_{i=1}^{n} \phi_{X_i}(s)
\end{aligned}$$

1.7 Hoeffding's inequality

Hoeffding's inequality is one of the most important inequalities in the machine learning literature. It gives a **exponential** bound for sums of random variables where the increments are bounded. You should ask yourself: why is it good that the bound is exponential?

Theorem 1.18. (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be independent r.v.'s with $a_i \leq X_i \leq b_i$ a.s. Then $\forall\, t > 0$,

$$\begin{aligned}
P(S_n - ES_n \geq t) &\leq \exp\left\{\frac{-2t^2}{\sum (b_i - a_i)^2}\right\} \\
P(S_n - ES_n \leq -t) &\leq \exp\left\{\frac{-2t^2}{\sum (b_i - a_i)^2}\right\}
\end{aligned}$$

And the two-sided bound:

$$P(\left|S_n - ES_n\right| \geq t) \leq 2\, \exp\left\{\frac{-2t^2}{\sum (b_i - a_i)^2}\right\}$$

Not only is the inequality of great use, but its proof also draws on many of the concepts of the previous sections. The main idea of the proof is actually quite

simple. First, we bound the MGFs of the increments using Taylor expansion. Then we plug these bounds into a Chernoff bound for the overall sum.

Lemma 1.19. (MGF bound for Hoeffding)

Let X be an r.v. with $EX = 0$ and $a \leq X \leq b$. Then $\forall s > 0$,

$$E(e^{sX}) \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}$$

Proof. Fix some $s > 0$ and consider the function $f(x) = e^{sx}$. Then for every $x \in [a, b]$, Jensen's inequality gives us:

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

Taking expectations (note $EX = 0$) and defining $p = -a/(b-a)$, we have:

$$Ee^{sX} \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} = \left[1 - p + pe^{s(b-a)}\right]e^{-ps(b-a)} = e^{\phi(u)}$$

Where $u = s(b-a)$ and $\phi(u) = -pu + \log(1-p+pe^u)$. Therefore it is sufficient to show that $\phi(u) \leq s^2(b-a)^2/8$.

Since ϕ is sufficiently smooth, the 1^{st} order Taylor expansion about $u = 0$ with Lagrange remainder is:

$$\phi(u) = \phi(0) + \phi'(0) \cdot u + \frac{u^2}{2}\phi''(c), \quad \text{some } c \in [0, u]$$

Now some calculate show that

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}} \quad \Rightarrow \quad \phi'(0) = 0$$

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \quad = \quad \frac{\alpha\beta}{(\alpha+\beta)^2} \leq \frac{1}{4}$$

Plugging this into the Taylor expansion above shows that

$$\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$$

□

Now the proof of the main result:

Proof. Applying Markov's to $P(S_n - ES_n \geq t) = P(e^{s(S_n - ES_n)}) \geq e^{st})$, we have:

$$\begin{aligned} P(S_n - ES_n \geq t) &\leq e^{-st}E\left\{\exp\left(s\sum_{i=1}^{n}(X_i - EX_i)\right)\right\} \\ &= e^{-st}E\left\{\prod_{i=1}^{n}e^{s(X_i - EX_i)}\right\} \\ &= e^{-st}\prod_{i=1}^{n}E\left\{e^{s(X_i - EX_i)}\right\} \quad \text{(by independence)} \end{aligned}$$

Now applying our lemma to the RHS, we have

$$P(S_n - ES_n \geq t) \leq e^{-st} \cdot \prod_{i=1}^{n} \exp\left\{\frac{s^2(b_i - a_i)^2}{8}\right\}$$
$$= e^{-st} \cdot \exp\left\{\frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right\}$$

Choosing $s = 4t/\sum(b_i - a_i)^2$ completes the proof.

□

1.8 Bennett's and Bernstein's inequalities

Like Hoeffding's inequality, both Bennett's and Bernstein's inequality give exponential bounds for a sum of random variables whose increments are bounded. However, the key difference is that Bennett's and Bernstein's inequality take into account the **variances** of the increments.

Theorem 1.20. (Bennett's inequality) Let $X_1, \ldots, X_n$ be independent with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma_i^2$, and $|X_i| \leq c$ for all i. Then for $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left\{\frac{-n\sigma^2}{c^2} \cdot h\left(\frac{ct}{n\sigma^2}\right)\right\}$$

where $h(u) = (1+u)\log(1+u) - u$, for $u \geq 0$.

The proof of this inequality is actually quite similar to the proof of Hoeffding's inequality. Again we first bound the MGFs of the increments using Taylor expansion, and plug those into a Chernoff-type bound for the overall sum.

Lemma 1.21. (MGF bound for Bennett)

Let X be an r.v. with $\mathbb{E}X = 0, \mathbb{E}X^2 = \sigma^2$, and $|X| \leq c$. Then:

$$\mathbb{E}(e^{sX}) \leq \exp\left\{\frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)\right\}, \quad \text{for all } s > 0$$

Proof. By Taylor series for e^x we have for any $s > 0$,

$$\mathbb{E}(e^{sX}) = \mathbb{E}\left\{1 + sX + \sum_{r=2}^{\infty}\frac{s^r X^r}{r!}\right\} = 1 + \sum_{r=2}^{\infty}\frac{s^r \mathbb{E}(X^r)}{r!}$$

Now by Holder's inequality,

$$\mathbb{E}X^r \;\leq\; \mathbb{E}|X|^r \;\leq\; \mathbb{E}|X|^2|X|^{r-2} \;\leq\; \sigma^2 \cdot c^{r-2}$$

Therefore plugging this into our Taylor expansion and summing, we obtain:

$$\begin{aligned}\mathbb{E}(e^{sX}) &\leq 1 + \sum_{r=2}^{\infty} \frac{s^r \sigma^2 c^{r-2}}{r!} \\ &= 1 + \frac{\sigma^2}{c^2} \sum_{r=2}^{\infty} \frac{(sc)^r}{r!} \\ &= 1 + \frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)\end{aligned}$$

Then apply the identity $1 + x \leq e^x$.

□

Now we plug into a basic Chernoff-type bound to prove the main result:

Proof. (of Bennett's inequality)

Recall that we have $X_1, \ldots, X_n$ be independent with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma_i^2$, and $|X_i| \leq c$ for all i.

Define $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. Then by the basic Chernoff bound for $s > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leq e^{-st} \prod_{i=1}^n \mathbb{E}e^{sX_i} \leq \exp\left\{\frac{n\sigma^2(e^{sc} - 1 - sc)}{c^2} - st\right\}$$

Now this bound holds for all $s > 0$, so we optimize. To ease notation define the function:

$$f(s) = \frac{n\sigma^2(e^{sc} - 1 - sc)}{c^2} - st$$

Some high school calculus shows this to be minimized at:

$$s_0 = c^{-1} \log(1 + tc/n\sigma^2)$$

and substituting into the above expression gives the result.

□

2 Conditional expectation and other stuff

2.1 Intuition behind conditional expectation

Technical note: in what follows, we assume that all expectations exist.

IF X is a random variable then the ordinary expected value $\mathbb{E}X$ represents our best guess of the value of X if we have no prior information. But now suppose that Y is a random variable related to X somehow and that we know the value of Y. Then one might expect that we can use that extra information to our advantage in guessing where X will end up.

This is the idea behind the expectation of X **conditional on** Y, written $\mathbb{E}(X \mid Y)$. The first thing to notice is that the conditional expectation $\mathbb{E}(X \mid Y)$ is actually a random variable (that depends on Y):

> **Example.** (Throwing darts)
>
> A game of darts is being played in another room and we want to guess the landing spot of a dart X. If we have no prior information, then our best guess for the landing spot of the dart might be, say, the the dead center of the board (call it C):
>
> $$\mathbb{E}X = C$$
>
> However, suppose now that every time a dart is thrown, someone watching in the other room tells us whether the dart landed in the upper or lower half of the board. Call this random variable Y, taking values in the set $\{\text{upper}, \text{lower}\}$. If we **condition** on Y, then we can improve our guess for the landing spot:
>
> - If $Y = \text{upper}$, then our best guess for X might be the centerpoint of the upper half of the board: $\mathbb{E}(X \mid Y = \text{upper}) = C_{\text{upper}}$
> - If $Y = \text{lower}$ then our best guess for X might be the centerpoint of the lower half of the board: $\mathbb{E}(X \mid Y = \text{lower}) = C_{\text{lower}}$
>
> Thus $\mathbb{E}(X \mid Y)$, our conditional best guess for the landing spot of the dart given Y, depends on whether $Y = \text{upper}$ or $Y = \text{lower}$. But since Y is a random variable, then $\mathbb{E}(X \mid Y)$ is also a random variable.

The discussion can be summarized as:

> **Fact.** Let X, Y, Z be random variables. Let $Y = a$ be a specific realization of Y. Then:
>
> - $\mathbb{E}(X \mid Y)$ is a random variable (and a function of Y)
> - $\mathbb{E}(X \mid Y, Z)$ is a random variable (and a function of both Y and Z)
> - $\mathbb{E}(X \mid Y = a)$ is **not** a random variable

How do we calculate conditional expectations? Recall that if X is an r.v. then:

1. If X is continuous with pdf $f_X(x)$, then $\mathbb{E}X = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x$

2. If X is discrete with pmf $p_X(x)$, then $\mathbb{E}X = \sum x \cdot p(x)$

The calculations have a natural extension for the case of conditional expectation. Let X and Y be r.v.'s and for now assume that we can calculate the *conditional pdf* or *conditional pmf* of $X \,|\, Y = y$. Then:

1. If $X \,|\, Y = y$ is continuous with conditional pdf $f_X(x \,|\, Y = y)$, then

$$\mathbb{E}(X \,|\, Y = y) = \int_{-\infty}^{\infty} x \cdot f_X(x \,|\, Y = y)\, \mathrm{d}x$$

2. If X is discrete with conditional pmf $p_X(x \,|\, Y = y)$, then

$$\mathbb{E}(X \,|\, Y = y) = \sum x \cdot p(x \,|\, Y = y)$$

Now the natural question is: how can we calculate the conditional pdf or conditional pmf of $X \,|\, Y = y$? With the same notation as above, recall the following (**sometimes** called Bayes' rule):

1. If $f_{X,Y}(x, y))$ is the joint pdf of X and Y, then

$$f_X(x \,|\, Y = y) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

2. If $p_{X,Y}(x, y)$ is the joint pmf of X and Y, then

$$p_X(x \,|\, Y = y) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

2.2 Basic properties of conditional expectation

Unless otherwise specified, X, Y, and Z are random variables and $g(\cdot)$ is a known integrable function of a random variable.

As mentioned in the previous section, conditional expectation differs from ordinary expectation because the random variable we are conditioning on may give us extra information about the random variable we are interested in. Here are two extreme cases:

Fact. (No information)

1. Let Y be **independent** of X. Then $\mathbb{E}(X \,|\, Y) = \mathbb{E}X$
2. Let c be a constant. Then $\mathbb{E}(X \,|\, c) = \mathbb{E}X$

Intuitive explanation: if Y is independent of X then its outcomes give us no information about the outcomes of X. Similarly, a constant has no variation and so cannot reveal any information to us about X. Therefore conditioning on Y or c should not change our best guess of X (relative to having no information).

Fact. (Perfect information)

1. $\mathbb{E}(g(X) \mid X) = g(X)$
2. $\mathbb{E}(g(X) \mid X = x) = g(x)$

Intuitive explanation: conditioning on X means that we already know the value that X has taken, so therefore our (trivial) best "guess" of the value of $g(X)$ is just $g(X)$.

Fact. (Expectation of conditional expectation)

$$\mathbb{E}\Big[\mathbb{E}(X \mid Y)\Big] = \mathbb{E}X$$

Intuitive explanation: Recall that $\mathbb{E}(X \mid Y)$ is a random function of Y. That is, for each (random) realization of Y, we have a best guess of the value of X. This is $\mathbb{E}(X \mid Y)$.

Now if we take expectation over Y of our best guess of X given Y, then we would expect all the different best guesses to average out to the overall best guess given no information.

Fact. (Substitution rule)

Let $g(X, Y)$ be some function of random variables X and Y. Then:

$$\mathbb{E}(g(X, Y) \mid Y = y) = \mathbb{E}(g(X, y) \mid Y = y)$$

Intuitive explanation: If we are trying to guess the outcome of a function of both X and Y but we know that Y will take value $Y = y$, then our best guess of the function value is obtained by fixing $Y = y$ and then guessing over the random values of X.

Fact. ("Taking out what is known")

$$\mathbb{E}(X \cdot g(Y) \mid Y) = g(Y) \cdot \mathbb{E}(X \mid Y)$$

Intuitive explanation: conditioning on Y tells us the value of Y and therefore also $g(Y)$, so we can pull it out of the expectation.

2.3 Probability integral transform

From undergraduate probability we know that a CDF is a non-decreasing function that is right-continuous. These conditions are not strong enough to guarantee that the function will always be invertible. However, we can define a psuedo-inverse function that will always exist for any given CDF:

Definition. (Quantile function/inverse CDF)

Let X be a random variable with CDF F. The **inverse CDF** of X is

$$F^{-1}(y) = \inf\{x : F(x) \geq y\} \quad \text{for } 0 < y < 1$$

Remark. F is non-decreasing, so the set $\{x : F(x) \geq y\}$ is of the form $[F^{-1}(y), \infty)$ or $(F^{-1}(y), \infty)$. But since F is right-continuous, then

$$\{x : F(x) \geq y\} = [F^{-1}(y), \infty)$$

and thus $\forall y \in (0,1)$ and $x \in \mathbb{R}$, $F^{-1}(y) \leq x \iff F(x) \geq y$.

The inverse CDF gives the least value of x for which the probability of X being below x is at least y. One useful application of this is a method that generates values distributed according to any CDF only using values distributed Uniform$(0,1)$.

Theorem 2.1. (Probability integral transform)

Let F be a **continuous** CDF. Then $Y \sim \text{Uniform}(0,1) \Rightarrow F^{-1}(Y) \sim F$.

Proof. Let X be the random variable defined by $F^{-1}(Y)$. Then

$$\begin{aligned} F_X(k) &= P(F^{-1}(Y) \leq k) \\ &= P(Y \leq F(k)) \\ &= F(k) \quad (Y \sim \textit{Uniform}(0,1)) \end{aligned}$$

□

The proof in the opposite direction is has two cases:

Theorem 2.2. (Continuous case)

If X has continuous CDF F, then $\mathbb{P}(F(X) \leq y) = y$ (i.e. $F(X)$ is distributed Uniform$(0,1)$).

Proof. Consider y in the range of $F(X)$. Since F is increasing, F^{-1} is also increasing. Thus:

$$\begin{aligned} \mathbb{P}(F(X) \leq y) &= \mathbb{P}(F^{-1}(F(X)) \leq F^{-1}(y)) \quad (F^{-1} \text{ is increasing}) \\ &= \mathbb{P}(X \leq F^{-1}(y)) \quad (F \text{ is increasing}) \\ &= F(F^{-1}(y)) \\ &= y \quad (y \text{ is in the range of } F) \end{aligned}$$

□

Remark. We elaborate on the equality used above:

$$\mathbb{P}(F^{-1}(F(X)) \leq F^{-1}(y)) = \mathbb{P}(X \leq F^{-1}(y))$$

If F is strictly increasing, then it is clear that $F^{-1}(F(x)) = x$. If it is not, then it might be true that $F^{-1}(F(x)) \neq x$, but then we are in a set of measure zero so the result remains.

Theorem 2.3. (Discrete (general) case)

If X has discrete or continuous CDF F, then $\mathbb{P}(F(X) \leq y) \leq y$

Proof. Let y be in the range of $F(X)$. We show that there may exist a point such that $\mathbb{P}(F(X) \leq y) < y$.

Let y be a value in the range of F that is "jumped over" by F. Then since CDF's are right-continuous, there is a point x_0 such that $F(x_0 - \epsilon) < y < F(x_0)$ for any $\epsilon > 0$. In other words:

$$\{x : F(x) \leq y\} = \{x : F(x) < y\}$$

Follow the steps of the above proof with the strict inequality to arrive at the result.

□

2.4 Moment generating functions

Some basic properties of moment generating functions which the reader should already be familiar with:

Lemma 2.4. Let X be a random variable. If $E[e^{s_0|x|}] < \infty$ for some $s_0 > 0$, then $E[e^{sx}] < \infty$ for all $-s_0 \leq s \leq s_0$.

Proof. Properties of the integral.

Definition. (Moment generating function)

Let X be a random variable such that $E[e^{s_0|x|}] < \infty$ for some $s_0 > 0$. Then the **moment generating function of X** is

$$\phi_X(s) = E[e^{sx}] \quad |s| \leq s_0$$

Example. (Some common MGFs)

1. $X \sim$ Poisson(λ): $\phi_X(s) = e^{\lambda(e^s - 1)} \quad s \in \mathbb{R}$
2. $X \sim$ Exponential(λ): $\phi_X(s) = \lambda/\lambda - s \quad s < \lambda$
3. $X \sim$ Normal$(0, 1)$: $\phi_X(s) = e^{s^2/2} \quad s \in \mathbb{R}$

The moment generating function of a random variable has many nice properties and can be used in quite sophisticated ways in proofs. However, we won't go into those in depth here. We finish by presenting a few simple properties.

Remark. (Taylor expansion of MGF)

If $\phi_X(s)$ exists, then $E|X|^k < \infty \; \forall k \geq 1$ and

$$\phi_X(s) = \sum_{k \geq 0} \frac{s^k E X^k}{k!}$$

Remark. (Differentiation of the MGF)

$$\phi'_X(s) = \frac{d}{ds} E[e^{sX}] = E\left[\frac{d}{ds} e^{sX}\right] = E\left[X e^{sX}\right]$$

and, more generally,

$$\phi_X^{(k)}(s) = E\left[X^k e^{sX}\right]$$

Remark. (Properties of the MGF)

1. $\phi_X^{(k)}(0) = E[X^k]$
2. $\phi_X''(s) \geq 0 \Rightarrow \phi_X(s)$ is convex

We close this section with a brief example concerning sums of independent random variables:

Example. (Sums of independent random variables)

Let $X_1, \ldots, X_n$ be independent where each X_i has an MGF. Define $S_n = X_1 + \ldots + X_n$. Then $\phi_{S_n}(s)$ exists:

$$\begin{aligned}
\phi_{S_n}(s) = E[e^{s \cdot S_n}] &= E\left[s(X_1 + \ldots + X_n)\right] \\
&= E\left[\prod_{i=1}^{n} e^{sX_i}\right] \\
&= \prod_{i=1}^{n} E[e^{sX_i}] \\
&= \prod_{i=1}^{n} \phi_{X_i}(s)
\end{aligned}$$

3 Supplemental exercises

1. If X is a Poisson(λ) random variable, show that for $i < \lambda$

$$\mathbb{P}(X \leq i) \leq \frac{e^{-\lambda}(e\lambda)^i}{i^i}$$

2. If $\mathbb{E}X < 0$ and $\theta \neq 0$ is such that $\mathbb{E}e^{\theta X} = 1$, show that $\theta > 0$.
3. Let X be a random variable such that $\mathbb{P}(|X| > t) \leq ae^{-bt^2}$ for $a \geq 1$ and $b > 0$. Show that:

$$\mathbb{E}|X| \leq \sqrt{\frac{1 + \log a}{b}}$$

Hint: use $\mathbb{E}X^2 = \int_0^\infty \mathbb{P}(X^2 > t)\,\mathrm{d}t$*, and split the integral into two parts.*

4. For random variables X and Z, show that

$$\mathbb{E}[(X - Y)^2] = \mathbb{E}(X^2) - \mathbb{E}(Y^2)$$

where $Y = \mathbb{E}(X \mid Z)$.

5. Show that $\mathrm{Cov}(X, \mathbb{E}(Y \mid X)) = \mathrm{Cov}(X, Y)$.

6. Suppose that conditional on $Y = y$, the random variables X_1 and X_2 are independent with mean y. Show that

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Var}(Y)$$