1 Various Inequalities

Inequalities are useful for bounding quantities that might otherwise be hard to compute. They will also be used in the theory of convergence. Our first inequality is Markov’s inequality.

**Theorem 1 (Markov’s inequality)** Let $X$ be a non-negative random variable and suppose that $\mathbb{E}(X)$ exists. For any $t > 0$,

$$P(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$  

**Proof.** Since $X > 0$,

$$\mathbb{E}(X) = \int_{0}^{\infty} xf(x)dx = \int_{0}^{t} xf(x)dx + \int_{t}^{\infty} xf(x)dx \
\geq \int_{t}^{\infty} xf(x)dx \geq t \int_{t}^{\infty} f(x)dx = tP(X > t).$$

The second part follows by setting $t = k\sigma$. ■

**Theorem 2 (Chebyshev’s inequality)** Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{Var}(X)$. Then,

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad P(|Z| \geq k) \leq \frac{1}{k^2}$$

where $Z = (X - \mu)/\sigma$. In particular, $P(|Z| > 2) \leq 1/4$ and $P(|Z| > 3) \leq 1/9$.

**Proof.** We use Markov’s inequality to conclude that

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

The second part follows by setting $t = k\sigma$. ■

**Theorem 3 (Cauchy-Schwartz inequality)** If $X$ and $Y$ have finite variances then

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}. \quad (1)$$

Recall that a function $g$ is **convex** if for each $x, y$ and each $\alpha \in [0, 1]$,

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

If $g$ is twice differentiable and $g''(x) \geq 0$ for all $x$, then $g$ is convex. It can be shown that if $g$ is convex, then $g$ lies above any line that touches $g$ at some point, called a tangent line. A function $g$ is **concave** if $-g$ is convex. Examples of convex functions are $g(x) = x^2$ and $g(x) = e^x$. Examples of concave functions are $g(x) = -x^2$ and $g(x) = \log x$.  

\mbox{}
Theorem 4 (Jensen’s inequality) If \(g\) is convex, then
\[
\mathbb{E}g(X) \geq g(\mathbb{E}X). \tag{2}
\]
If \(g\) is concave, then
\[
\mathbb{E}g(X) \leq g(\mathbb{E}X). \tag{3}
\]

Proof. Let \(L(x) = ax + bx\) be a line, tangent to \(g(x)\) at the point \(\mathbb{E}(X)\). Since \(g\) is convex, it lies above the line \(L(x)\). So,
\[
\mathbb{E}g(X) \geq \mathbb{E}L(X) = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = L(\mathbb{E}(X)) = g(\mathbb{E}X). \quad \blacksquare
\]

Example: From Jensen’s inequality we see that \(\mathbb{E}(X^2) \geq (\mathbb{E}X)^2\).

Example: KL distance. Define the Kullback-Leibler distance between two densities \(f\) and \(g\) by
\[
D(f, g) = \int f(x) \log \left( \frac{f(x)}{g(x)} \right) dx.
\]
Note that \(D(f, f) = 0\). We will use Jensen to show that \(D(f, g) \geq 0\). Let \(X \sim f\). Then
\[
-D(f, g) = \mathbb{E} \log \left( \frac{g(X)}{f(X)} \right) \leq \log \mathbb{E} \left( \frac{g(X)}{f(X)} \right) = \log \int f(x) \frac{g(x)}{f(x)} dx = \log \int g(x) dx = \log(1) = 0.
\]
So, \(-D(f, g) \leq 0\) and hence \(D(f, g) \geq 0\).

2 Limit Theorems

Some convergence properties are preserved under transformations.

Theorem 5 Let \(X_n, X, Y_n, Y\) be random variables.
(a) If \(X_n \xrightarrow{P} X\) and \(Y_n \xrightarrow{P} Y\), then \(X_n + Y_n \xrightarrow{P} X + Y\).
(b) If \(X_n \xrightarrow{qm} X\) and \(Y_n \xrightarrow{qm} Y\), then \(X_n + Y_n \xrightarrow{qm} X + Y\).
(c) If \(X_n \xrightarrow{P} X\) and \(Y_n \xrightarrow{P} Y\), then \(X_n Y_n \xrightarrow{P} XY\).
Theorem 6 Let $X_n, X, Y_n, Y$ be random variables.

(d) If $X_n \Rightarrow X$ and $Y_n \Rightarrow c$, then $X_n + Y_n \Rightarrow X + c$.
(e) If $X_n \Rightarrow X$ and $Y_n \Rightarrow c$, then $X_n Y_n \Rightarrow cX$.

Theorem 7 Let $X_n, X, Y_n, Y$ be random variables. Let $g$ be a continuous function.

(f) If $X_n \overset{P}{\rightarrow} X$, then $g(X_n) \overset{P}{\rightarrow} g(X)$.
(g) If $X_n \Rightarrow X$, then $g(X_n) \Rightarrow g(X)$.

- Parts (d) and (e) are known as Slutzky’s theorem
- Parts (f) and (g) are known as The Continuous Mapping Theorem.
- It is worth noting that $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$ does not in general imply that $X_n + Y_n \Rightarrow X + Y$.

3 The Law of Large Numbers

The LLN says that the mean of a large sample is close to the mean of the distribution. For example, the proportion of heads of a large number of tosses of a fair coin is expected to be close to 1/2. We now make this more precise.

Let $X_1, X_2, \ldots$ be an iid sample, let $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \text{Var}(X_1)$. Recall that the sample mean is defined as $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ and that $\mathbb{E}(\bar{X}_n) = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$.

Theorem 8

The Weak Law of Large Numbers (WLLN).

If $X_1, \ldots, X_n$ are iid, then $\bar{X}_n \overset{P}{\rightarrow} \mu$. Thus, $\bar{X}_n - \mu = o_P(1)$.

Interpretation of the WLLN: The distribution of $\bar{X}_n$ becomes more concentrated around $\mu$ as $n$ gets large.

Proof. Assume that $\sigma < \infty$. This is not necessary but it simplifies the proof. Using Chebyshev’s inequality,

$$
P \left( |\bar{X}_n - \mu| > \epsilon \right) \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}
$$

which tends to 0 as $n \to \infty$. ■

Theorem 9 The Strong Law of Large Numbers. Let $X_1, X_2, \ldots$ be an iid sample. Let $\mu = \mathbb{E}(X_1)$ and $\mathbb{E}(|X_1|) < \infty$. We have

$$
\bar{X}_n \overset{\text{as}}{\rightarrow} \mu.
$$
We proved in class a version of the SLLN with independent random variables each with bounded fourth moment using the Borel-Cantelli Lemma.

**Theorem 10** Let $X_1, X_2, \ldots$ be a sequence of independent random variables, each having the same finite mean $m$, each having

$$E(X_i - m)^4 \leq \alpha < \infty.$$  

Then

$$P\left\{\lim_{n \to \infty} \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = m\right\} = 1.$$  

**Lemma 11** Let $Z_1, \ldots, Z_n$ be random variables. Suppose that for each $\epsilon > 0$, we have

$$\sum_{n=1}^{\infty} P\{|Z_n - Z| \geq \epsilon\} < \infty.$$  

Then $\{Z_n\}$ converges almost surely to $Z$.

**The Central Limit Theorem.**

The law of large numbers says that the distribution of $\overline{X}_n$ piles up near $\mu$. This isn’t enough to help us approximate probability statements about $\overline{X}_n$. For this we need the central limit theorem.

Suppose that $X_1, \ldots, X_n$ are IID with mean $\mu$ and variance $\sigma^2$. The central limit theorem (CLT) says that $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ has a distribution which is approximately Normal with mean $\mu$ and variance $\sigma^2/n$. This is remarkable since nothing is assumed about the distribution of $X_i$, except the existence of the mean and variance. For instance, the CLT applies even if $X$ is a coin toss. Although a Bernoulli distribution is far from normal, the mean of a sequence of Bernoulli experiments is normally distributed (for a large number of tosses).

**Theorem 12 (The Central Limit Theorem (CLT))** Let $X_1, \ldots, X_n$ be IID with mean $\mu$ and variance $\sigma^2$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{\text{Var}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} Z$$  

where $Z \sim N(0, 1)$. In other words,

$$\lim_{n \to \infty} P(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$
Interpretation: Probability statements about $\overline{X}_n$ can be approximated using a Normal distribution. It’s the probability statements that we are approximating, not the random variable itself.

In addition to $Z_n \Rightarrow N(0, 1)$, there are several forms of notation to denote the fact that the distribution of $Z_n$ is converging to a Normal. They all mean the same thing. Here they are:

\[
\begin{align*}
Z_n & \approx N(0, 1) \\
\overline{X}_n & \approx N\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{X}_n - \mu & \approx N\left(0, \frac{\sigma^2}{n}\right) \\
\sqrt{n}(\overline{X}_n - \mu) & \approx N\left(0, \sigma^2\right) \\
\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} & \approx N(0, 1).
\end{align*}
\]

Recall that if $X$ is a random variable, its moment generating function (MGF) is $\psi_X(t) = Ee^{tX}$. Assume in what follows that the MGF is finite in a neighborhood around $t = 0$.

**Lemma 13** Let $Z_1, Z_2, \ldots$ be a sequence of random variables. Let $\psi_n$ be the MGF of $Z_n$. Let $Z$ be another random variable and denote its MGF by $\psi$. If $\psi_n(t) \to \psi(t)$ for all $t$ in some open interval around 0, then $Z_n \Rightarrow Z$.

Recall

- $n$’th moment: $E(X^n)$
- Central moments: $E((X - \mu)^n)$.

**Proof of the central limit theorem.** Let

\[ Y_i = (X_i - \mu)/\sigma. \]

Then,

\[ Z_n = n^{-1/2} \sum_i Y_i. \]

Let $\psi(t)$ be the MGF of $Y_1$. The MGF of $\sum_i Y_i$ is $(\psi(t))^n$ and MGF of $Z_n$ is $[\psi(t/\sqrt{n})]^n \equiv \xi_n(t)$. Now

\[
\begin{align*}
\psi'(0) & = E(Y_1) = 0 \\
\psi''(0) & = E(Y_1^2) = \text{Var}(Y_1) = 1.
\end{align*}
\]
So,
\[
\psi(t) = \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi'''(0) + \cdots
\]
\[
= 1 + 0 + \frac{t^2}{2} + \frac{t^3}{3!}\psi''(0) + \cdots
\]
\[
= 1 + \frac{t^2}{2} + \frac{t^3}{3!}\psi''(0) + \cdots
\]

Now,
\[
\xi_n(t) = \left[ \psi \left( \frac{t}{\sqrt{n}} \right) \right]^n
\]
\[
= \left[ 1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}\psi''(0) + \cdots \right]^n
\]
\[
= \left[ 1 + \frac{\frac{t^2}{2}}{\frac{1}{2}} + \frac{\frac{t^3}{3!n^{1/2}}}{n}\psi''(0) + \cdots \right]^n
\]
\[
\rightarrow e^{t^2/2}
\]

which is the mgf of a $N(0,1)$. The result follows from the previous Theorem. In the last step we used the fact that if $a_n \rightarrow a$ then
\[
\left(1 + \frac{a_n}{n}\right)^n \rightarrow e^a. \quad \blacksquare
\]

The central limit theorem tells us that
\[
Z_n = \sqrt{n}(X_n - \mu)/\sigma \text{ is approximately } N(0,1).
\]
However, we rarely know $\sigma$. Later, we will see that we can estimate $\sigma^2$ from $X_1, \ldots, X_n$ by
\[
S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.
\]

This raises the following question: if we replace $\sigma$ with $S_n$, is the central limit theorem still true? The answer is yes.

**Theorem 14** Assume the same conditions as the CLT. Then,
\[
T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim N(0,1).
\]

**Proof.** We have that
\[
T_n = Z_n W_n
\]
where
\[
Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \quad \text{and} \quad W_n = \frac{\sigma}{S_n}.
\]
Now \( Z_n \sim N(0, 1) \) and \( W_n \xrightarrow{P} 1 \). The result follows from Slutzky’s theorem.

A detailed derivation on \( W_n \xrightarrow{P} 1 \) is given on board. ■

There is also a multivariate version of the central limit theorem.

**Theorem 15 (Multivariate central limit theorem)** Let \( X_1, \ldots, X_n \) be IID random vectors where

\[
X_i = \begin{pmatrix}
X_{1i} \\
X_{2i} \\
\vdots \\
X_{ki}
\end{pmatrix}
\]

with mean

\[
\mu = \begin{pmatrix}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_k
\end{pmatrix} = \begin{pmatrix}
\mathbb{E}(X_{1i}) \\
\mathbb{E}(X_{2i}) \\
\vdots \\
\mathbb{E}(X_{ki})
\end{pmatrix}
\]

and variance matrix \( \Sigma \). Let

\[
\bar{X} = \begin{pmatrix}
\bar{X}_1 \\
\bar{X}_2 \\
\vdots \\
\bar{X}_k
\end{pmatrix}
\]

where \( \bar{X}_j = n^{-1} \sum_{i=1}^n X_{ji} \). Then,

\[
\sqrt{n}(\bar{X} - \mu) \sim N(0, \Sigma).
\]

**Exercises:** Show that \( \mathbb{E}(\bar{X} - \mu)(\bar{X} - \mu)^T = \Sigma \).

### 4 The Delta Method

If \( Y_n \) has a limiting Normal distribution then the delta method allows us to find the limiting distribution of \( g(Y_n) \) where \( g \) is any smooth function.

**Theorem 16 (The Delta Method)** (CB Theorem 5.5.24) Suppose that

\[
\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \sim N(0, 1)
\]

and that \( g \) is a differentiable function such that \( g'(\mu) \neq 0 \). Then

\[
\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \sim N(0, 1).
\]

In other words,

\[
Y_n \approx N\left( \mu, \frac{\sigma^2}{n} \right)
\]

implies that \( g(Y_n) \approx N\left( g(\mu), \frac{(g'(\mu))^2\sigma^2}{n} \right) \).
There is also a multivariate version of the delta method.

**Theorem 17 (The Multivariate Delta Method)** Suppose that \( Y_n = (Y_{n1}, \ldots, Y_{nk}) \) is a sequence of random vectors such that

\[
\sqrt{n}(Y_n - \mu) \rightsquigarrow N(0, \Sigma).
\]

Let \( g : \mathbb{R}^k \rightarrow \mathbb{R} \) and let

\[
\nabla g(y) = \begin{bmatrix}
\frac{\partial g}{\partial y_1} \\
\vdots \\
\frac{\partial g}{\partial y_k}
\end{bmatrix}.
\]

Let \( \nabla \mu \) denote \( \nabla g(y) \) evaluated at \( y = \mu \) and assume that the elements of \( \nabla \mu \) are nonzero. Then

\[
\sqrt{n}(g(Y_n) - g(\mu)) \rightsquigarrow N(0, \nabla_\mu^T \Sigma \nabla \mu).
\]

**Example 18** Let

\[
\begin{pmatrix}
X_{11} \\
X_{21}
\end{pmatrix}, \quad \begin{pmatrix}
X_{12} \\
X_{22}
\end{pmatrix}, \ldots, \quad \begin{pmatrix}
X_{1n} \\
X_{2n}
\end{pmatrix}
\]

be iid random vectors with mean \( \mu = (\mu_1, \mu_2)^T \) and variance \( \Sigma \). Let

\[
\bar{X}_1 = \frac{1}{n} \sum_{i=1}^{n} X_{1i}, \quad \bar{X}_2 = \frac{1}{n} \sum_{i=1}^{n} X_{2i}
\]

and define \( Y_n = \bar{X}_1 \bar{X}_2 \). Thus, \( Y_n = g(\bar{X}_1, \bar{X}_2) \) where \( g(s_1, s_2) = s_1 s_2 \). By the central limit theorem,

\[
\sqrt{n} \left( \frac{\bar{X}_1 - \mu_1}{\bar{X}_2 - \mu_2} \right) \rightsquigarrow N(0, \Sigma).
\]

Now

\[\nabla g(s) = \begin{bmatrix}
\frac{\partial g}{\partial s_1} \\
\frac{\partial g}{\partial s_2}
\end{bmatrix} = \begin{bmatrix}
s_2 \\
s_1
\end{bmatrix}\]

and so

\[
\nabla_\mu^T \Sigma \nabla \mu = (\mu_2 \mu_1) \begin{pmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{22}
\end{pmatrix} \begin{pmatrix}
\mu_2 \\
\mu_1
\end{pmatrix} = \mu_2^2 \sigma_{11} + 2 \mu_1 \mu_2 \sigma_{12} + \mu_1^2 \sigma_{22}.
\]

Therefore,

\[
\sqrt{n}(\bar{X}_1 \bar{X}_2 - \mu_1 \mu_2) \rightsquigarrow N \left( 0, \mu_2^2 \sigma_{11} + 2 \mu_1 \mu_2 \sigma_{12} + \mu_1^2 \sigma_{22} \right). \]
5 Monotone convergence theorem

A sequence of random variables \( \{X_n\} \nearrow X \) if \( X_1 \leq X_2 \leq \ldots \), and also \( \lim_{n \to \infty} X_n(\omega) = X(\omega) \) for each \( \omega \in \Omega \). That is, the sequence \( \{X_n\} \) converges monotonically to \( X \).

**Theorem 19** *(The monotone convergence theorem.)* Suppose \( X_1, X_2, \ldots \), are random variables with \( \mathbb{E}(X_1) > -\infty \), and \( \{X_n\} \nearrow X \). Then \( X \) is a random variable, and \( \lim_{n \to \infty} \mathbb{E}(X_n) = \mathbb{E}(X) \).

**Remark.** Since expected values are unchanged if we modify the random variable values on sets of probability 0, we still have \( \lim_{n \to \infty} \mathbb{E}(X_n) = \mathbb{E}(X) \) provided \( \{X_n\} \nearrow X \) almost surely, i.e., on a subset of \( \Omega \) having probability 1. Recall if \( \{X_n\} \) is a sequence of random variables, then \( \lim_{n \to \infty} X_n \) exists a.s. means that there exists an event \( N \in \mathcal{B} \), such that \( P(N) = 0 \) and if \( \omega \in N^c \) then

\[
\lim_{n \to \infty} X_n
\]

exists. It also means that for a.a. \( \omega \),

\[
\limsup_{n \to \infty} X_n(\omega) = \liminf_{n \to \infty} X_n(\omega).
\]

We will write \( \lim_{n \to \infty} X_n = X \) a.s. or \( X_n \xrightarrow{a.s.} X \), or \( X_n \to X \) a.s..

**Lemma 20** *If \( X \) is a non-negative random variable, then*

\[
\sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \mathbb{E}[X],
\]

*where \( [X] \) is the greatest integer not exceeding \( X \). In particular, if \( X \) is non-negative-integer valued, then \( \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \mathbb{E}(X) \).*