STAT 512: Statistical Inference **Autumn 2020**

Lecture 3: Expectation and basic asymptotic theories

Instructor: Yen-Chi Chen

Reference: Casella and Berger Chapter 2 and 3.

3.1 Expectation

For a function $g(x)$, the expectation of $g(X)$ is

$$\mathbb{E}(g(X)) = \int g(x)dF(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx, & \text{if } X \text{ is continuous} \\ \sum_x g(x)p(x), & \text{if } X \text{ is discrete} \end{cases}.$$

In the simplest case $g(x) = x$,

$$\mathbb{E}(X) = \int xdF(x) = \begin{cases} \int_{-\infty}^{\infty} xp(x)dx, & \text{if } X \text{ is continuous} \\ \sum_x xp(x), & \text{if } X \text{ is discrete} \end{cases}.$$

is known as the the mean (expectation) of a R.V. X. Let $\mu = \mathbb{E}(X)$, the variance of X is $\mathsf{Var}(X) = \mathbb{E}((X-\mu)^2)$. The mean is a common measure of the center of a distribution and the variance is a common measure of the spread of a distribution.

The m-th moment of a random variable X is

$$\mathbb{E}(X^m).$$

Let $\mu = \mathbb{E}(X)$ be the mean/first moment of X, the m-th *centered* moment of X is

$$\mathbb{E}((X-\mu)^m).$$

Thus, the variance is the second centered moment.

Example.

- $X \sim \mathsf{Binomial}(n,p)$. Then $\mathbb{E}(X) = np$ and $\mathsf{Var}(X) = np(1-p)$.
- $X \sim \mathsf{Geometric}(p)$. Then $\mathbb{E}(X) = 1/p$ and $\mathsf{Var}(X) = (1-p)^2/p$.
- $X \sim \mathsf{Poisson}(\lambda)$. Then $\mathbb{E}(X) = \lambda$ and $\mathsf{Var}(X) = \lambda$.
- $X \sim \mathsf{Normal}(\mu,\sigma^2)$. Then $\mathbb{E}(X) = \mu$ and $\mathsf{Var}(X) = \sigma^2$.
- $X \sim \mathsf{Exponential}(\lambda)$. Then $\mathbb{E}(X) = 1/\lambda$ and $\mathsf{Var}(X) = 1/\lambda^2$.
- $X \sim \mathsf{Gamma}(\alpha,\lambda)$. Then $\mathbb{E}(X) = \alpha/\lambda$ and $\mathsf{Var}(X) = \alpha/\lambda^2$.
- $X \sim \mathsf{Beta}(\alpha,\beta)$. Then $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$ and $\mathsf{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- $X \sim \mathsf{Uniform}(a,b)$. Then $\mathbb{E}(X) = (a+b)/2$ and $\mathsf{Var}(X) = (b-a)^2/12$.

Expectation are decomposable:

$$\mathbb{E}\left(\sum_{j=1}^{k} c_j g_j(X)\right) = \sum_{j=1}^{k} c_j \cdot \mathbb{E}(g_j(X_i)).$$

Note that the above equality holds even if X_i's are dependent.

When a set of random variables $X_1, \cdots, X_n$ are independent, then

$$\mathbb{E}\left(X_1 \cdot X_2 \cdots X_n\right) = \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) \cdots \mathbb{E}(X_n).$$

In fact, you can also prove that

$$\mathbb{E}\left(g_1(X_1) \cdot g_2(X_2) \cdots g_n(X_n)\right) = \mathbb{E}(g_1(X_1)) \cdot \mathbb{E}(g_2(X_2)) \cdots \mathbb{E}(g_3(X_n)).$$

For two random variables X and Y with their mean being μ_X and μ_Y and variance being σ_X^2 and σ_Y^2. The *covariance*

$$\mathsf{Cov}(X,Y) = \mathbb{E}((X-\mu_x)(Y-\mu_y)) = \mathbb{E}(XY) - \mu_x\mu_y$$

and the (Pearson's) correlation

$$\rho(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sigma_x\sigma_y}.$$

When two R.V. are not independent, we have

$$\mathsf{Var}(X \pm Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) \pm 2\mathsf{Cov}(X,Y).$$

The independence implies the covariance (and correlation) is 0, i.e.,

$$X \perp Y \Rightarrow \mathsf{Cov}(X,Y) = 0.$$

As a result, if $X \perp Y$,

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y).$$

A more general result is that for independent random variables $X_1, \cdots, X_n$, we have

$$\mathsf{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \cdot \mathsf{Var}(X_i).$$

Example (Binomial). Here we illustrate how the above properties can be useful in computing the variance of some distributions. Consider $X \sim \mathsf{Binomial}(n,p)$. By the definition of a Binomial distribution, we can rewrite $X = Y_1 + Y_2 + \cdots + Y_n$, where each Y_i is an independent Bernoulli random variable with parameter p. Thus,

$$\mathsf{Var}(X) = \mathsf{Var}(Y_1 + Y_2 + \cdots + Y_n) = \sum_{i=1}^{n} \mathsf{Var}(Y_i) = np(1-p).$$

3.2 Moment generating function (MGF)

Moment generating function (MGF) is a powerful function that describes the underlying features of a random variable. The MGF of a RV X is

$$M_X(t) = \mathbb{E}(e^{tX}).$$

Note that M_X may not exist. When M_X exists in a neighborhood of 0, using the fact that

$$e^{tX} = 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots,$$

we have

$$M_X(t) = 1 + t\mu_1 + \frac{t^2\mu_2}{2!} + \frac{t^3\mu_3}{3!} + \cdots,$$

where $\mu_j = \mathbb{E}(X^j)$ is the j-th moment of X. Therefore,

$$\mathbb{E}(X^j) = M^{(j)}(0) = \left.\frac{d^j M_X(t)}{dt^j}\right|_{t=0}$$

Here you see how the moments of X is generated by the function M_X.

For two random variables X, Y, *if their MGFs are the same, then the two random variables have the same CDF.* Thus, MGFs can be used as a tool to determine if two random variables have the identical CDF. Note that the MGF is related to the Laplace transform (actually, they are the same) and this may give you more intuition why it is so powerful.

The MGF has some interesting properties:

- **Location-scale.** $M_{aX+b}(t) = \mathbb{E}(e^{(aX+b)t}) = e^{bt}\mathbb{E}(e^{atX}) = e^{bt}M_X(at)$.
- **Multiplicity.** $M_{X+Y}(t) = \mathbb{E}(e^{(X+Y)t}) = \mathbb{E}(e^{Xt}e^{Yt})$. Thus,

$$X \perp Y \Rightarrow M_{X+Y}(t) = \mathbb{E}(e^{Xt}e^{Yt}) = \mathbb{E}(e^{Xt})\mathbb{E}(e^{Yt}) = M_X(t)M_Y(t).$$

Example (Bernoulli and Binomial). Let $X \sim \mathsf{Ber}(p)$. Its MGF is $M_X(t) = \mathbb{E}(e^{tX}) = pe^t + (1-p)$. Let $Y \sim \mathsf{Bin}(n,p)$. Using the fact that we can express it as $Y = X_1 + \cdots + X_n$, where each X_i is independent Bernoulli R.V. with parameter p. Its MGF is

$$M_Y(t) = \prod_{i=1}^{n} M_{Z_i}(t) = (pe^t + (1-p))^n.$$

Example (Poisson). Let $X \sim \mathsf{Poisson}(\lambda)$. Then its MGF is

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x=0} e^{tx}\frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda}\underbrace{\sum_{x}\frac{[\lambda e^t]^x}{x!}}_{=e^{\lambda e^t}} = e^{\lambda(e^t-1)}.$$

Example (Exponential). Let $X \sim \mathsf{Exp}(\lambda)$. Then its MGF is

$$M_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda - t}$$

for $t < \lambda$.

Example (Normal). Let $X \sim N(\mu, \sigma^2)$. Then you can show that (exercise)

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

You can use the fact that the MGF uniquely determines a distribution to show that any addition of normals is still normal.

Remark (characteristic function). A more general function than MGF is the characteristic function. Let i be the imagination number. The characteristic function of a RV X is

$$\phi_X(t) = \mathbb{E}(e^{itX}).$$

When X is absolutely continuous, the characteristic function is the Fourier transform of the PDF. The characteristic function always exists and when two RVs have the same characteristic function, the two RVs have identical distribution.

3.2.1 Multivariate MGF

The MGF can be defined for a random vector. Consider $X = (X_1, \cdots, X_d) \in \mathbb{R}^d$ be a random vector. Then its MGF will be a function of d augments

$$M_X(t) = \mathbb{E}(e^{t^T X}),$$

where $t = (t_1, \cdots, t_d) \in \mathbb{R}^d$.

Example. Let X be a multivariate normal $MVN(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance matrix. Namely, each component $X_i \sim N(\mu_i, \Sigma_{ii})$ and the covariance $\mathsf{Cov}(X_i, X_j) = \Sigma_{ij}$. Then its MGF will be

$$M_X(t) = e^{t^T\mu + \frac{1}{2}t^T\Sigma t}.$$

Using this, you can show that a linear tranformation

$$Z = b + AX \sim MVN(b + A\mu, A\Sigma A^T).$$

Example (Normal plus Normal). Here we show that the MGF provide a simple way to see that the addition of two normal random variable still leads to normal random variable. Let X, Y be two normal random variable such that their joint distribution is MVN with mean (μ_1, μ_2) and covariance matrix Σ. Consider $Z = X + Y$. To see why Z is still normal, consider its MGF:

$$M_Z(t) = \mathbb{E}(e^{tZ}) = \mathbb{E}(e^{tX+tY}) = M_{X,Y}(t, t),$$

which is the MGF of the normal vector (X, Y) with the augment (t, t). Thus,

$$M_Z(t) = M_{X,Y}(t, t) = e^{t(\mu_1+\mu_2)+\frac{1}{2}t^2(\Sigma_{11}+\Sigma_{22}+2\Sigma_{12})},$$

which is the MGF of a normal random variable with mean $\mu_1 + \mu_2$ and variance $\Sigma_{11} + \Sigma_{22} + 2\Sigma_{12} = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X, Y)$.

3.3 Convergence Theory

Let $F_1, \cdots, F_n, \cdots$ be the corresponding CDFs of $Z_1, \cdots, Z_n, \cdots$. For a random variable Z with CDF F, we say that Z_n **converges in distribution** (a.k.a. converge weakly or converge in law) to Z if for every x,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

In this case, we write

$$Z_n \overset{D}{\to} Z, \quad \text{or } Z_n \overset{d}{\to} Z.$$

Namely, the CDF's of the sequence of random variables converge to a the CDF of a fixed random variable.

For a sequence of random variables $Z_1, \cdots, Z_n, \cdots$, we say Z_n **converges in probability** to another random variable Z if for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = 0$$

and we will write

$$Z_n \overset{P}{\to} Z$$

Remark (convergence almost surely). For a sequence of random variables $Z_1, \cdots, Z_n, \cdots$, we say Z_n **converges almost surely** to a random variable Z if

$$P(\lim_{n\to\infty} Z_n = Z) = 1$$

or equivalently,

$$P(\{\omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}) = 1.$$

We use the notation

$$Z_n \overset{a.s.}{\to} Z$$

to denote convergence almost surely. Note that almost surely convergence implies convergence in probability. Convergence in probability implies convergence in distribution.

Examples.

- Let $\{X_1, X_2, \cdots,\}$ be a sequence of random variables such that $X_n \sim N\left(0, 1 + \frac{1}{n}\right)$. Then X_n converges in distribution to $N(0,1)$.
- Let $\{X_1, X_2, \cdots\}$ be a sequence of random variables such that $X_i \sim N(0, 1/n)$. Then $X_n \overset{P}{\to} 0$, i.e., it converges in probability to 0. Also, the random variable $\sqrt{n}X_n \overset{D}{\to} N(0,1)$.
- Let $\{X_1, X_2, \cdots\}$ be a sequence of random variables such that

  $$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = 1) = \frac{1}{n}.$$

  Then $X_n \overset{P}{\to} 0$.

Sometimes, one may be thinking that the convergence in probability/distribution may imply convergence in *expectation*. But this is not true! Here is an example that it converges in probability to 0 but its expectation diverges.

Example (diverging expectation but convergence in probability). Consider a sequence of RVs $X_1, X_2, \cdots$, such that

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = n^2) = \frac{1}{n}.$$

Then you can easily verify that $X_n \overset{P}{\to} 0$. However, if you compute the expectation,

$$\mathbb{E}(X_n) = n \to \infty.$$

So the expectation is in fact diverging. Later we will see that converge in expectation does imply convergence in probability (Markov's inequality).

3.3.1 Weak Law of Large Numbers

We write $X_1, \cdots, X_n \sim F$ when $X_1, \cdots, X_n$ are IID (independently, identically distributed) from a CDF F. In this case, $X_1, \cdots, X_n$ is called a *random sample*.

Theorem 3.1 (Markov's inequality) *Let X be a non-negative RV. Then for any $\epsilon > 0$,*

$$P(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}.$$

A feature of the Markov inequality is that it implies that *converges in expectation ⇒ convergence in probability.* Also, the Markov's inequality implies the following useful result, known as the Chebyshev's inequality.

Theorem 3.2 (Chebyshev's inequality) *Let X be a RV with finite variance. Then for any $\epsilon > 0$,*

$$P(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathsf{Var}(X)}{\epsilon^2}.$$

The proof of the Chebyshev's inequality is a direct application of the Markov's inequality. The Chebyshev's inequality shows that for a sequence of random variables with equal mean but a vanishing variance, this sequence converges in probability to the mean. When applying to the sample mean, it becomes the famous (weak) law of large numbers.

Theorem 3.3 (Weak Law of Large Numbers) *Let $X_1, \cdots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$. If $\mathbb{E}|X_1| < \infty$ and $\mathsf{Var}(X_1) = \sigma^2 < \infty$, the sample average*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

converges in probability to μ. i.e.,

$$\bar{X}_n \xrightarrow{P} \mu.$$

Proof: Using the property of sample mean, one can easily show that

$$\mathsf{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Thus, by the Chebyshev's inequality

$$P(|\bar{X}_n - \mu| > t) \leq \frac{\sigma^2}{nt^2} \to 0,$$

which completes the proof.

■

The above theorem is also known as Weak Law of Large Numbers. In fact, we do not need to assume the existence of variance–this condition can be relaxed (but the proof will become much more complicated). Note that there is something called the strong law of large number, which states the convergence in terms of 'almost surely convergence'.

3.3.2 Central Limit Theorem

Theorem 3.4 (Central Limit Theorem) *Let $X_1, \cdots, X_n$ be IID random variables with $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathsf{Var}(X_1) < \infty$. Let $\bar{X}_n$ be the sample average. Then*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Note that $N(0,1)$ is also called standard normal random variable*.*

Proof: Let $Z = \sqrt{n}(\bar{X}_n - \mu)$. Proving the problem is equivalent to showing that $Z \to N(0, \sigma^2)$.

Note that we can rewrite Z as

$$Z = \sqrt{n}(\bar{X}_n - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i,$$

where each Y_i has mean 0 and variance σ^2 and are IID. Thus, the MGF of Z is

$$M_Z(t) = \mathbb{E}(e^{tZ}) = \mathbb{E}\left(e^{\frac{t}{\sqrt{n}}\sum_{i=1}^{n} Y_i}\right) = \mathbb{E}\left(e^{\frac{t}{\sqrt{n}}Y_1}\right)^n = M_{Y_1}(t/\sqrt{n})^n. \tag{3.1}$$

Note that we use the fact that $Y_1, \cdots, Y_n$ are IID in the last equality.

Now we analyze $M_{Y_1}(t/\sqrt{n})$:

$$M_{Y_1}(t/\sqrt{n}) = \mathbb{E}(e^{tY/\sqrt{n}}) = 1 + \frac{t}{\sqrt{n}}\underbrace{\mathbb{E}(Y)}_{=0} + \frac{t^2}{2n}\underbrace{\mathbb{E}(Y^2)}_{=\sigma^2} + \text{smaller order term}.$$

The remaining terms are smaller order because we are under $n \to \infty$. Denoting the smaller order term as $o(1)$, using the above expansion, we can see that

$$M_{Y_1}(t/\sqrt{n})^n = \left(1 + \frac{t^2\sigma^2}{2n + o(1)}\right)^n \to e^{\frac{1}{2}t^2\sigma^2}.$$

Thus, Equation (3.1) will be approaching

$$M_Z(t) = M_{Y_1}(t/\sqrt{n})^n = \left(1 + \frac{t^2\sigma^2}{2n + o(1)}\right)^n \to e^{\frac{1}{2}t^2\sigma^2},$$

which is the MGF of a normal random variable with mean 0 and variance σ^2. So we have proved the desired result.

■

Note that there are other versions of central limit theorem that allows dependent RVs or infinite variance using the idea of 'triangular array' (also known as the Lindeberg-Feller Theorem). However, the details are beyond the scope of this course so we will not pursue it here.

3.3.3 Other useful theorems

Continuous mapping theorem: Let g be a continuous function.

- If a sequence of random variables $X_n \overset{D}{\to} X$, then $g(X_n) \overset{D}{\to} g(X)$.
- If a sequence of random variables $X_n \overset{p}{\to} X$, then $g(X_n) \overset{p}{\to} g(X)$.

Slutsky's theorem: Let $\{X_n : n = 1, 2, \cdots\}$ and $\{Y_n : n = 1, 2, \cdots\}$ be two sequences of RVs such that $X_n \overset{D}{\to} X$ and $Y_n \overset{p}{\to} c$, where X is a RV c is a constant. Then

$$\begin{aligned}
X_n + Y_n &\overset{D}{\to} X + c \\
X_n Y_n &\overset{D}{\to} cX \\
X_n / Y_n &\overset{D}{\to} X/c \quad (\text{if } c \neq 0).
\end{aligned}$$

We will these two theorems very frequently when we are talking about the maximum likelihood estimator.

Why do we need these notions of convergences? The convergence in probability is related to the concept of statistical consistency. An estimator is statistically consistent if it converges in probability toward its target population quantity. The convergence in distribution is often used to construct a confidence interval or perform a hypothesis test.

3.4 Concentration inequality

In addition to the above two theorems, we often use the concentration inequality to obtain convergence in probability. Let $\{X_n : n = 1, 2, \cdots\}$ be a sequence of RVs. For a given $\epsilon > 0$, the concentration inequality aims at finding the function $\phi_n(\epsilon)$ such that

$$P(|X_n - \mathbb{E}(X_n)| > \epsilon) \leq \phi_n(\epsilon)$$

and $\phi_n(\epsilon) \to 0$. This automatically gives us convergence in probability. Moreover, the *convergence rate* of $\phi_n(\epsilon)$ with respect to n is a central quantity that describes how fast X_n converges toward its mean.

Example: concentration of a Gaussian mean. The Markov's inequality implies a useful bound on describing how fast the sample mean of a Gaussian converges to the population mean. For simplicity, we consider a sequence of mean 0 Gaussians: $X_1, \cdots, X_n \sim N(0, \sigma^2)$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. It is known that $\bar{X}_n \sim N(0, \sigma^2/n)$. Then

$$\begin{aligned}
P(\bar{X}_n > \epsilon) &= P(e^{\bar{X}_n} > e^{\epsilon}) \\
&= P(e^{s\bar{X}_n} > e^{s\epsilon}) \\
&\leq \frac{\mathbb{E}(e^{s\bar{X}_n})}{e^{s\epsilon}} \qquad \text{by Markov's inequality} \\
&\leq e^{\frac{1}{2n}\sigma^2 s^2 - s\epsilon} \qquad \text{by the MGF of Gaussian}
\end{aligned}$$

for any positive number s. In the exponent, it is a quadratic function of s and the maximal occurs at $s = \frac{n\epsilon}{\sigma^2}$, leading to

$$P(\bar{X}_n > \epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}.$$

The same bound holds for the other direction $P(\bar{X}_n < -\epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}$. So we conclude

$$P(|\bar{X}_n| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}$$

or more generally,

$$P(|\bar{X}_n - \mathbb{E}(X_1)| > \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}.$$

A bound like the above is often referred to as a *concentration inequality*.

Example (concentration of a maximum). Let $X_1, \cdots, X_n$ be IID standard normal random variables $N(0, \sigma^2)$. Define $Z_n = \max\{|X_1|, \cdots, |X_n|\}$ be the maximal number among them. Intuitively, we know that when $n \to \infty$, Z_n should be diverging since we are taking the maximum of more and more values. But it is possible to find an increasing sequence $\gamma_n \to \infty$ such that Z_n/γ_n will not diverge (in probability). How do we find such γ_n? A simple approach is based on the concentration inequality. Using the result from previous example, we know that for a single random variable X_i (replace the sample mean by the mean of a single RV), we have

$$P(|X_i| > \epsilon) \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}.$$

With this, we can bound

$$\begin{aligned} P(Z_n > \epsilon) &= P(\max\{|X_1|, \cdots, |X_n|\} > \epsilon) \\ &\leq \sum_{i=1}^n P(|X_i| > \epsilon) \qquad \text{(maximum is over } \epsilon \Rightarrow \text{one of them must hold)} \\ &\leq 2ne^{-\frac{\epsilon^2}{2\sigma^2}}. \end{aligned}$$

Thus, as long as we can choose a sequence $\epsilon = \epsilon_n$ such that

$$2ne^{-\frac{\epsilon_n^2}{2\sigma^2}} \to \delta$$

for some constant $0 < \delta < 1$, we can bound how fast Z_n diverge. Solving this gives us a single rule $\epsilon_n = \sigma\sqrt{2\log(2n) - 2\log(\delta)}$. This leads to the choice of $\gamma_n = \sigma\sqrt{2\log n}$, which gives a characterization on how fast Z_n diverges.

3.4.1 Concentration of mean

Let $X_1, \cdots, X_n \sim F$ be a random sample such that $\sigma^2 = \mathsf{Var}(X_1)$. Using the Chebyshev's inequality, we know that the sample average $\bar{X}_n$ has a concentration inequality:

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

However, when the RVs are bounded, there is a stronger notion of convergence, as described in the following theorem.

Theorem 3.5 (Hoeffding's inequality) *Let $X_1, \cdots, X_n$ be IID RVs such that $0 \leq X_1 \leq 1$ and let $\bar{X}_n$ be the sample average. Then for any $\epsilon > 0$,*

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

Before proving the Hoeffding's inequality, we first introduce the following lemma:

Lemma 3.6 *Let X be a random variable with $\mathbb{E}(X) = 0$ and $a \leq X \leq b$. Then*

$$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8}$$

for any positive number t.

Proof: We will use the fact that $x \mapsto e^{tx}$ is a convex function for all positive t. Recall that a function $g(x)$ is a convex function if for any two point $a < b$ and $\alpha \in [0,1]$,

$$g(\alpha a + (1-\alpha)b) \leq \alpha g(a) + (1-\alpha)g(b).$$

Because $X \in [a,b]$, we define α_X to

$$X = \alpha_X b + (1-\alpha_X)a.$$

This implies

$$\alpha_X = \frac{X-a}{b-a}$$

Using the fact that $x \mapsto e^{tx}$ is convex,

$$e^{tX} \leq \alpha_X e^{tb} + (1-\alpha_X)e^{ta} = \frac{X-a}{b-a}e^{tb} + \frac{b-X}{b-a}e^{ta}.$$

Now taking the expectation in both sides,

$$\mathbb{E}(e^{tX}) \leq \frac{\mathbb{E}(X)-a}{b-a}e^{tb} + \frac{b-\mathbb{E}(X)}{b-a}e^{ta} = \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = e^{g(s)}, \tag{3.2}$$

where $s = t(b-a)$ and $g(s) = -\gamma s + \log(1-\gamma+\gamma e^s)$ and $\gamma = -a/(b-a)$. Note that $g(0) = g'(0) = 0$ and $g''(s) \leq 1/4$ for all positive s. Using Taylor's theorem,

$$g(s) = g(0) + sg'(0) + \frac{1}{2}s^2 g''(s^*)$$

for some $s^* \in [0,s]$. Thus, we conclude $g(s) \leq \frac{1}{2} \times s^2 \times \frac{1}{4} = \frac{1}{8}s^2$.

Then equation (3.2) implies

$$\mathbb{E}(e^{tX}) \leq e^{g(s)} \leq e^{\frac{s^2}{8}} = e^{\frac{t^2(b-a)^2}{8}}.$$

■

Now we formally prove the Hoeffding's inequality.

Proof:

We first prove that $P\left(\bar{X}_n - \mu \geq \epsilon\right) \leq e^{-2n\epsilon^2/(b-a)^2}$.

Let $Y_i = X_i - \mu$. Because the exponential function is monotonic, for any positive r,

$$\begin{aligned}
P\left(\bar{X}_n - \mu \geq \epsilon\right) &= P\left(\bar{Y}_n \geq \epsilon\right) \\
&= P\left(\sum_{i=1}^n Y_i \geq n\epsilon\right) \\
&= P\left(e^{\sum_{i=1}^n Y_i} \geq e^{n\epsilon}\right) \\
&= P\left(e^{t\sum_{i=1}^n Y_i} \geq e^{tn\epsilon}\right) \\
&\leq \frac{\mathbb{E}(e^{t\sum_{i=1}^n Y_i})}{e^{tn\epsilon}} \quad \text{by Markov's inequality} \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1} \cdot e^{tY_2} \cdots e^{tY_n}) \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1}) \cdot \mathbb{E}(e^{tY_2}) \cdots \mathbb{E}(e^{tY_n}) \\
&= e^{-tn\epsilon}\mathbb{E}(e^{tY_1})^n \\
&\leq e^{-tn\epsilon}e^{nt^2(b-a)^2/8} \text{ by Lemma 3.6.}
\end{aligned}$$

Because the above inequality holds for all positive t, we can choose t to optimize the bound. To get the bound as sharp as possible, we would like to make it as small as possible. Thus, we need to find t such that

$$-tn\epsilon + nt^2(b-a)^2/8$$

is minimized. Taking derivatives with respect to t and set it to be 0, we obtain

$$t_* = \frac{4\epsilon}{(b-a)^2}$$

and

$$-t_*n\epsilon + nt_*^2(b-a)^2/8 = -2n\epsilon^2/(b-a)^2.$$

Thus, the inequality becomes

$$P\left(\bar{X}_n - \mu \geq \epsilon\right) \leq e^{-t_*n\epsilon}e^{nt_*^2(b-a)^2/8} = e^{-2n\epsilon^2/(b-a)^2}.$$

The same proof also applies to the case $P\left(\bar{X}_n - \mu \leq \epsilon\right)$ and we will obtain the same bound. Therefore, we conclude that

$$P\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}.$$

■

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.

Note that the exponential rate may also be used to obtain an almost sure convergence via the Borel-Cantelli Lemma.

Example: consistency of estimating a high-dimensional proportion. To see how the Hoeffding's inequality is useful, we consider the problem of estimating the proportion of several binary variables. Suppose that we observe IID observations

$$X_1, \cdots, X_n \in \{0,1\}^d.$$

$X_{ij} = 1$ can be interpreted as the i-th individual response 'Yes' in j-th question. We are interested in estimating the proportion vector $\pi \in [0,1]^d$ such that $\pi_j = P(X_{ij} = 1)$ is the proportion of 'Yes' response in j-th question in the population. A simple estimator is the sample proportion $\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_d)^T$ such that

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}.$$

When d is much smaller than n, it is easy to see that this is a good estimator. However, if $d = d_n \to \infty$ with $n \to \infty$, will $\hat{\pi}$ still be a good estimator of π? To define a good estimator, we mean that *every proportion* can be estimated accurately. A simple way to quantify this is the vector max norm:

$$\|\hat{\pi} - \pi\|_{\max} = \max_{j=1,\cdots,d} |\hat{\pi}_j - \pi_j|.$$

We consider the problem of estimating π_j first. It is easy to see that by the Hoeffding's inequality,

$$P(|\hat{\pi}_j - \pi_j| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Thus,

$$\begin{aligned} P(\|\hat{\pi}-\pi\|_{\max} > \epsilon) &= P\left(\max_{j=1,\cdots,d} |\hat{\pi}_j - \pi_j| > \epsilon\right) \\ &\leq \sum_{j=1}^{d} P(|\hat{\pi}_j - \pi_j| > \epsilon) \\ &\leq 2de^{-2n\epsilon^2}. \end{aligned} \tag{3.3}$$

Thus, as long as $2de^{-2n\epsilon^2} \to 0$ for any fixed ϵ, we have the statistical consistency. This implies that we need

$$\frac{\log d}{n} \to 0,$$

which allows the number of questions/variables to increase a lot faster than the sample size n!