

# LECTURE NOTES 6

## 1 Hoeffding's inequality

The main drawback was that Mill's inequality only applies to Gaussian random variables. Another commonly useful exponential concentration inequality applies to bounded random variables. This is called Hoeffding's inequality.

**Hoeffding's inequality:** Suppose that  $X_1, \dots, X_n$  are independent and that,  $a_i \leq X_i \leq b_i$ , and  $\mathbb{E}[X_i] = \mu$ . Then for any  $t > 0$ ,

$$\mathbb{P}(|\bar{X} - \mu| \geq t) \leq 2 \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

where  $\bar{X}_n = n^{-1} \sum_i X_i$ . When  $a \leq X_i \leq b$  this becomes

$$\mathbb{P}(|\bar{X} - \mu| \geq t) \leq 2 \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

We will not prove this but the proof is in the book.

Hoeffding's inequality looks a bit different from the other inequalities we have seen yesterday, but let us rearrange it a bit. Equivalently,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i\right| \geq t \sqrt{\frac{\sum_{i=1}^n (b_i - a_i)^2}{n^2}}\right) \leq 2 \exp(-2t^2).$$

This is more like the earlier inequalities, but notice that we don't really have the standard deviation any more. That said, if  $a_i \leq X_i \leq b_i$  then  $\text{Var}(X_i) \leq (b_i - a_i)^2$ .

**Exercise:** Prove the above fact.

So that:

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \leq \frac{\sum_{i=1}^n (b_i - a_i)^2}{n^2},$$

and this will allow us to interpret Hoeffding's inequality in a more familiar way. Roughly, it says that the probability that the sample average is more than  $t$  standard deviations from its expectation is at most  $\exp(-2t^2)$ .

Let us now use Hoeffding's inequality in our case study example of coin tosses. There each random variable is between  $-1$  and  $1$  so we have that by Hoeffding's inequality:

$$\mathbb{P}(|\bar{X} - \mu| \geq t) \leq 2 \exp(-2nt^2).$$

Observe once again this inequality is similar to Chebyshev's but is much tighter.

## 2 Confidence Intervals

We will cover confidence intervals this in much more detail later on but here is a quick preview. Hoeffding's inequality gives us a simple way to create a confidence interval for a binomial parameter  $p$ . Let  $\hat{p} = n^{-1} \sum_i X_i$  be the fraction of tosses that are heads. Fix  $\alpha > 0$  and let,

$$t = \sqrt{\frac{1}{2n} \log(2/\alpha)}.$$

By Hoeffding's inequality,

$$\mathbb{P}(|\hat{p} - p| \geq t) \leq 2 \exp(-2nt^2) = \alpha.$$

Let  $C = [\hat{p} - t, \hat{p} + t]$ . Then,

$$\mathbb{P}(p \notin C) \leq \alpha$$

where  $C$  is a random set and  $p$  is fixed. Hence, the random interval  $C$  traps the true parameter value  $p$  with probability at least  $1 - \alpha$ ; we call  $C$  a  $1 - \alpha$  confidence interval.

## 3 The Weak Law of Large Numbers

The weak law of large numbers essentially assures us that the average of independent and identically distributed random variables “converges” to the expectation.

We will assume that  $X_1, \dots, X_n$  are i.i.d with  $\text{Var}(X) < \infty$ . Let  $\mu = \mathbb{E}[X_i]$ ,  $\sigma^2 = \text{Var}[X_i]$  and  $\bar{X}_n = n^{-1} \sum_i X_i$ . Then the weak law of large numbers says that for any  $\epsilon > 0$ ,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \rightarrow 0,$$

as  $n \rightarrow \infty$ .

This type of convergence is quite common and has a name: it is called convergence in probability. So the weak LLN tells us that the sample average converges to the population average in probability. This is called the weak law because convergence in probability is often referred to as weak convergence. We will discuss convergence more systematically soon.

**Proof:** The proof is a simple consequence of Chebyshev's inequality. We have that for any positive  $\epsilon$ , via Chebyshev's inequality:

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\text{Var}(X)}{n\epsilon^2} \rightarrow 0,$$

as  $n \rightarrow \infty$ .

## 4 Introduction to Stochastic Convergence

The weak LLN is an example where we tried to reason about the limiting behaviour of a sequence of random variables:

$$Y_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

In statistics we commonly estimate parameters using data, and in this case our estimate forms a sequence of random variables (as we collect more data). We would like to reason about the limiting behaviour of our estimates (do they “converge” to the truth, what is their distribution around the truth and so on). This is what we refer to as **large sample theory**.

Suppose we have a sequence of random variables  $Y_1, \dots, Y_n$ , and another random variable  $Y$ . Let  $F_n$  denote the CDF of  $Y_n$ , and let  $F$  be the CDF of  $Y$ .

The two most basic forms of stochastic convergence are:

1. **Convergence in Probability:** We say the sequence  $Y_n$  converges to  $Y$  if for every  $\epsilon > 0$ ,

$$\mathbb{P}(|Y_n - Y| \geq \epsilon) \rightarrow 0,$$

as  $n \rightarrow \infty$ . An important example is the weak law. An importance special case is when  $X = c$  for some constant  $c$ . we say the sequence  $Y_n$  converges to  $c$  if for every  $\epsilon > 0$ ,

$$\mathbb{P}(|Y_n - c| \geq \epsilon) \rightarrow 0,$$

as  $n \rightarrow \infty$ . We write

$$Y_n \xrightarrow{P} c.$$

2. **Convergence in Distribution:** The sequence  $Y_n$  converges in distribution to  $Y$  if

$$\lim_{n \rightarrow \infty} F_n(t) = F(t),$$

at all points  $t$  where  $F$  is continuous where  $F_n(t) = P(Y_n \leq t)$  and  $F(t) = P(Y \leq t)$ . We write

$$Y_n \rightsquigarrow Y \quad \text{or} \quad Y_n \xrightarrow{d} Y.$$

An important example of this type of convergence is the central limit theorem. We will see this in much more detail but roughly the central limit theorem says that the

average of i.i.d. RVs, rescaled appropriately converges in distribution to a standard normal distribution, i.e. that

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

If you think about it, this is a really stunning result. You do not assume anything about the RVs except their first two moments need to exist, and you conclude that the average approaches a Gaussian distribution.