# Appendix
# Class 6, 18.05
# Jeremy Orloff and Jonathan Bloom

## 1 Introduction

In this appendix we give more formal mathematical material that is not strictly a part of 18.05. This will not be on homework or tests. We give this material to emphasize that in doing mathematics we should be careful to specify our hypotheses completely and give clear deductive arguments to prove our claims. We hope you find it interesting and illuminating.

## 2 With high probability the density histogram resembles the graph of the probability density function:

We stated that one consequence of the law of large numbers is that as the number of samples increases the density histogram of the samples has an increasing probability of matching the graph of the underlying pdf or pmf. This is a good rule of thumb, but it is rather imprecise. It is possible to make more precise statements It will take some care to make a sensible and precise statement, which will not be quite so sweeping.

Suppose we have an experiment that produces data according to the random variable $X$ and suppose we generate $n$ independent samples from $X$. Call them

$$x_1, x_2, \ldots, x_n.$$

By a bin we mean a range of values, i.e. $[x_k, x_{k+1})$. To make a density histogram of the data we divide the range of $X$ into $m$ bins and calculate the fraction of the data in each bin.

Now, let $p_k$ be the probability a random data point is in the $k$th bin. This is this probability for an indicator (Bernoulli) random variable $B_{k,j}$ which is 1 if the $j$th data point is in the bin and and 0 otherwise.

**Statement 1.** Let $\bar{p}_k$ be the fraction of the data in bin $k$. As the number $n$ of data points gets large the probability that $\bar{x}_k$ is close to $p_k$ approaches 1. Said differently, given any small number, call it $a$ the probability $P(|\bar{p}_k - p_k| < a)$ depends on $n$, and as $n$ goes to infinity this probability goes to 1.

**Proof.** Let $\bar{B}_k$ be the average of $B_{k,j}$. Since $E(B_{k,j}) = p_k$, the law of large number says exactly that

$$P(|\bar{B}_k - p_k| < a) \quad \text{approaches 1 as } n \text{ goes to infinity.}$$

But, since the $B_{k,j}$ are indicator variables, their average is exactly $\bar{p}_k$, the fraction of the data in bin $k$. Replacing $\bar{B}_k$ by $\bar{p}_k$ in the above equation gives

$$P(|\bar{p}_k - p_k| < a) \quad \text{approaches 1 as } n \text{ goes to infinity.}$$

This is exactly what statement 1 claimed.

**Statement 2.** The same statement holds for a finite number of bins simultaneously. That is, for bins 1 to $m$ we have

$P(\,(|\bar{B}_1 - p_1| < a),\,(|\bar{B}_2 - p_2| < a),\,\ldots,\,(|\bar{B}_m - p_m| < a)\,)$ approaches 1 as $n$ goes to infinity.

**Proof.** First we note the following probability rule, which is a consequence of the inclusion exclusion principle: If two events $A$ and $B$ have $P(A) = 1 - \alpha_1$ and $P(B) = 1 - \alpha_2$ then $P(A \cap B) \geq 1 - (\alpha_1 + \alpha_2)$.

Now, Statement 1 says that for any $\alpha$ we can find $n$ large enough that $P(|\bar{B}_k - p_k| < a) > 1 - \alpha/m$ for each bin separately. By the probability rule, the probability of the intersection of all these events is at least $1 - \alpha$. Since we can let $\alpha$ be as small as we want by letting $n$ go to infinity, in the limit we get probability 1 as claimed.

**Statement 3.** If $f(x)$ is a continuous probability density with range $[a, b]$ then by taking enough data and having a small enough bin width we can insure that with high probability the density histogram is as close as we want to the graph of $f(x)$.

**Proof.** We will only sketch the argument. Assume the bin around $x$ has width is $\Delta x$. If $\Delta x$ is small enough then the probability a data point is in the bin is approximately $f(x)\Delta x$. Statement 2 guarantees that if $n$ is large enough then with high probability the fraction of data in the bin is also approximately $f(x)\Delta x$. Since this is the area of the bin we see that its height will be approximately $f(x)$. That is, with high probability the height of the histogram over any point $x$ is close to $f(x)$. This is what Statement 3 claimed.

**Note.** If the range is infinite or the density goes to infinity at some point we need to be more careful. There are statements we could make for these cases.

## 3 The Chebyshev inequality

One proof of the LoLN follows from the following key inequality.

The Chebyshev inequality. Suppose $Y$ is a random variable with mean $\mu$ and variance $\sigma^2$. Then for any positive value $a$, we have

$$P(|Y - \mu| \geq a) \leq \frac{\mathrm{Var}(Y)}{a^2}.$$

In words, the Chebyshev inequality says that the probability that $Y$ differs from the mean by more than $a$ is bounded by $\mathrm{Var}(Y)/a^2$. Morally, the smaller the variance of $Y$, the smaller the probability that $Y$ is far from its mean.

**Proof of the LoLN:** Since $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X)/n$, the variance of the average $\bar{X}_n$ goes to zero as $n$ goes to infinity. So the Chebyshev inequality for $Y = \bar{X}_n$ and fixed $a$ implies that as $n$ grows, the probability that $\bar{X}_n$ is farther than $a$ from $\mu$ goes to 0. Hence the probability that $\bar{X}_n$ is within $a$ of $\mu$ goes to 1, which is the LoLN.

**Proof of the Chebyshev inequality:** The proof is essentially the same for discrete and continuous $Y$. We'll assume $Y$ is continuous and also that $\mu = 0$, since replacing $Y$ by

$Y - \mu$ does not change the variance. So

$$\begin{aligned} P(|Y| \geq a) &= \int_{-\infty}^{-a} f(y)\,dy + \int_{a}^{\infty} f(y)\,dy \leq \int_{-\infty}^{-a} \frac{y^2}{a^2} f(y)\,dy + \int_{a}^{\infty} \frac{y^2}{a^2} f(y)\,dy \\ &\leq \int_{-\infty}^{\infty} \frac{y^2}{a^2} f(y)\,dy = \frac{\mathrm{Var}(Y)}{a^2}. \end{aligned}$$

The first inequality uses that $y^2/a^2 \geq 1$ on the intervals of integration. The second inequality follows because including the range $[-a, a]$ only makes the integral larger, since the integrand is positive.

## 4 The need for variance

We didn't lie to you, but we did gloss over one technical fact. Throughout we assumed that the underlying distributions had a variance. For example, the proof of the law of large numbers made use of the variance by way of the Chebyshev inequality. But there are distributions which do not have a variance because the sum or integral for the variance does not converge to a finite number. For such distributions the law of large numbers is still true, but the proof is harder. In 18.05 we won't have to worry about this, but if you go deeper into statistics this may become important.