13) Law of Large Numbers

13.1 Averages vary less.

Consider a probability experiment and a random variable X being the number on top of the die.

Then consider n independant repetitions and thus n independant random variables, $X_1, X_2, \cdots, X_n$

In our example X_i represents the number that comes up on the i^{th} throw of the die.

Assume there is no change in experimental conditions and therefore all X_i have same distribution. Furthermore outcome of one throw does not affect outcome of another throw and hence all X_i's are all independant.

This situation is so common, there is a name for it

(Defn on next page)

Defn 13.1: Let X be a random variable. A collection $X_1, \cdots, X_n$ of independant random variables that all have the same distribution as X is called i.i.d. sample from the distribution of X of size n. (i.i.d sample stands for independantantly and identically distributed).

The average:

$$\boxed{\hat{X}_n = \frac{(X_1 + X_2 + X_3 + \cdots + X_n)}{n}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is called the sample mean.

Another name for iid sample is random sample.

Consider the experiment as before with n = 30.
In figure 13.1 in notes, the values jump a lot.
The red line shows cumulative averages of the values. These behave much more predictably.

The cumulative averages appear to converge to the expectation: $E[X] = 3.5$

↳ represented by dotted line in fig 13.1

We now prove that the averages converges to expectation of X.

$$E[\hat{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$ by linearity of expectations Thm 10.2

Since all X_i are part of an iid sample, they all have same distribution as X, hence the same expected value as X,

$$E[X_i] = E[X] \qquad \forall\, 1 \le i \le n$$

So

$$E[\hat{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X]$$

$$= \frac{n\,E[X]}{n} = E[X]$$

$$\Rightarrow \boxed{E[\hat{X}_n] = E[X]}$$

■

The variance is

$$\mathrm{Var}(\hat{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Thm 10.5 with covariance of 0 since $X_i \perp\!\!\!\perp X_j$ for $i \neq j$

Since all X_i are part of an iid sample, they all have same distribution as X, hence the same variance value as X,

$$\mathrm{Var}(X_i) = \mathrm{Var}(X) \qquad \forall\ 1 \leq i \leq n$$

$$\mathrm{Var}(\hat{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X)$$

$$= \frac{n\,\mathrm{Var}(X)}{n^2} = \frac{\mathrm{Var}(X)}{n}$$

$$\Rightarrow \boxed{\mathrm{Var}(\hat{X}_n) = \frac{\mathrm{Var}(X)}{n}}$$

So what we basically showed is that:

- As n gets large the sample mean deviates less from the expected value

↳ as a consequence variance gets smaller

13.2 Chebychev's inequality

Previously we showed that variance of sample mean goes down as 1/n.

Given the intuitive understanding of the variance as a measure for likelihood that random variable deviates from its mean

But we need to provide a formal basis for that understanding. This is provided by the chebychev's inequality

→ So we can think of it as the probability of Sample mean going to be far away from true value of the mean is getting smaller, more mathematically

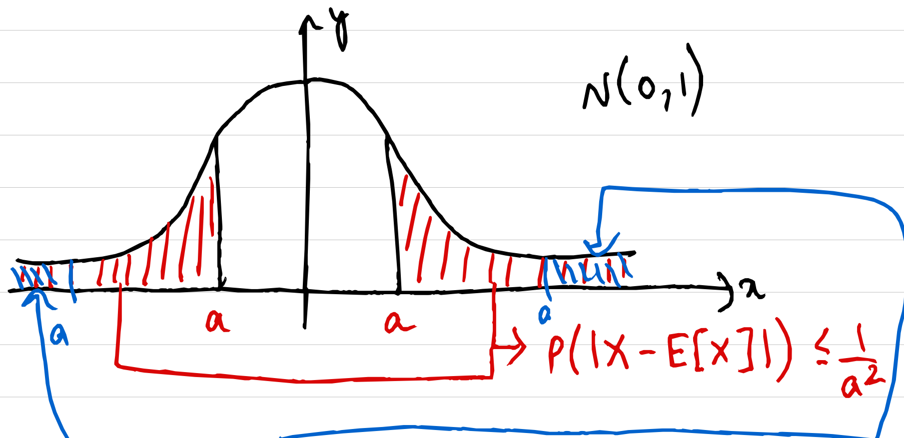
$$P(|X - E[X]|)$$

Theorem 13.2: **(Chebychev's inequality)**

Let X be a random variable, and let $a \in \mathbb{R}$ with $a > 0$. Then

$$\boxed{P(|X - E[X]| \geq a) \leq \frac{1}{a^2} Var(x)}$$

Intuition



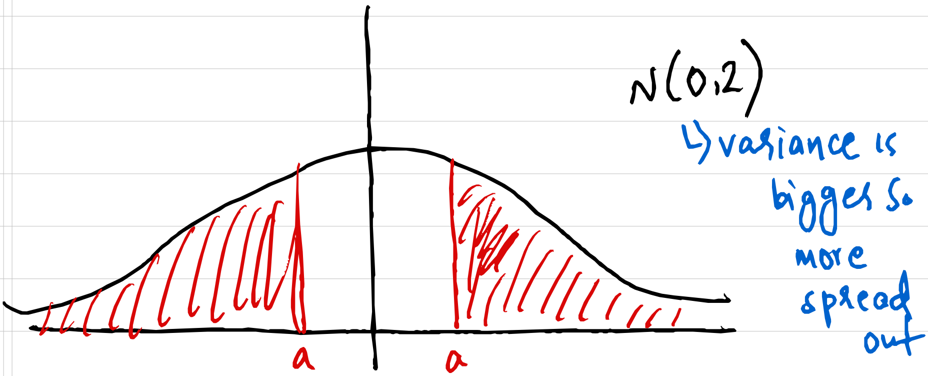
$P(|X - E[x]|)$ is smaller than $P(|X - E[x]|)$

So as a gets bigger, probability gets smaller.

So in $P(|X - E[x]| \geq a) \leq \frac{1}{a^2} Var(x)$

gets smaller — gets bigger

Similarly as $Var(X)$ gets bigger, $P(|X-E[X]| \geq a)$ gets bigger



For same value of a in the case of $N(0,1)$, the area / probability is bigger in $N(0,2)$ than for $N(0,1)$ as <u>variance is bigger</u>

Variance goes up $\Rightarrow$ probability goes up

proof: Where X is continuous random variable.
Let $E[X] = \mu$

Then

$$Var(X) = E[(X-\mu)^2] \qquad \text{by defn of variance}$$

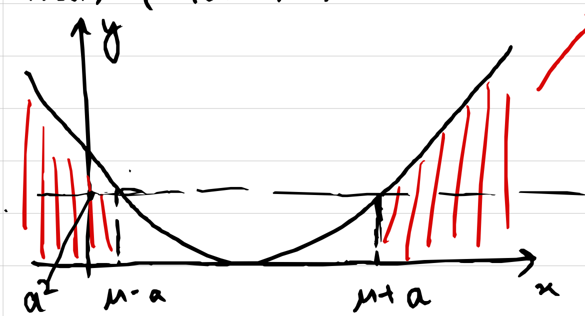
$$= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx \qquad \text{by Thm 7.11}$$

$$\geq \int_{|x-\mu| \geq a} (x-\mu)^2 f_X(x)\, dx$$

integration over whole of $\mathbb{R}$ is bigger than integration/area over a smaller region or subset of $\mathbb{R}$, here all x s.t. $|x-\mu| \geq a$.

reason for $|x-\mu| \geq a$

graph of $(x-\mu)^2$

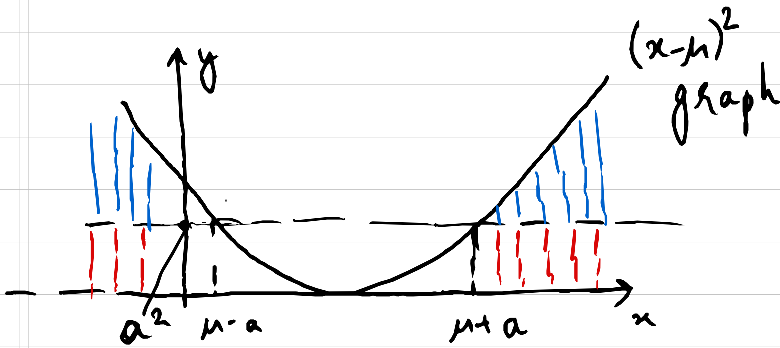


integrating over this region

$$\int_{|x-\mu| \geq a} (x-\mu)^2 f_X(x)$$

Now we have

$$\operatorname{Var}(X) \geq \int_{|x-\mu| \geq a} (x-\mu)^2 f_X(x)\, dx$$



In region $|x-\mu| \geq a$ of graph of $(x-\mu)^2$,

(*1) $$(x-\mu)^2 \geq a^2 \qquad \forall x \in \{x \in \mathbb{R} \mid |x-\mu| \geq a\}$$

$$\Rightarrow \quad f_X(x)(x-\mu)^2 \geq f_X(x)\, a^2$$

$$\Rightarrow \int_{|x-\mu| \geq a} f_X(x)(x-\mu)^2\, dx \geq \int_{|x-\mu| \geq a} f_X(x)\, a^2\, dx$$

by domination property of integrals over the region $\{x \mid |x-\mu| \geq a\}$ where (*1) is valid

So

$$\mathrm{Var}(X) = E[(X-\mu)^2]$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx$$

$$\geq \int_{|x-\mu|\geq a} (x-\mu)^2 f_X(x)\, dx$$

$$\geq \int_{|x-\mu|\geq a} a^2 f_X(x)\, dx = a^2 \int_{|x-\mu|\geq a} f_X(x)\, dx$$

$$\Rightarrow \mathrm{Var}(X) \geq a^2 \int_{|x-\mu|\geq a} f_X(x)$$

$$\Rightarrow \mathrm{Var}(X) \geq a^2 P(|X-\mu| \geq a)$$ by Thm 5.3

(since we are integrating a range of values)

$$\Rightarrow P(|X - E[X]|) \leq \frac{1}{a^2} \mathrm{Var}(X)$$

$\blacksquare$

Corollary 13.3: Let X be a random variable with finite expectation $E[X] = \mu$ and finite variance σ^2. Let $k \in \mathbb{R}$ with $k > 0$. Then

$$\boxed{p(|X - E[X]| \geq k\,sd(X)) \leq \frac{1}{k^2}}$$

and thus

$$\boxed{p(|X - E[X]| < k\,sd(X)) \geq 1 - \frac{1}{k^2}}$$

Example 13.4: Assume that probability for yellow smartie is $p_Y = 1/8$. As in Example 10.7, let Y be number of yellow smarties in a box of n smarties. Let $n = 40$.
You would expect $E[Y] = np_Y = 40/8 = 5$ yellow smarties.

Use chebychev's inequality to get an upperbound on the probability to get 11 or more yellow smarties.

Solution: Because $E[Y] = 5$, we can write the event $\{Y \geq 11\}$ equivalently as

$$\{|Y - E[Y]| \geq 6\}$$

Apply chebychev's inequality using $Var(x) = \frac{35}{8}$

$$P(Y \geq 11) = P(|Y - E[Y]| \geq 6)$$

$$\leq \frac{1}{6^2} Var(Y) = \frac{1}{36} \cdot \frac{35}{8} \approx 0.12$$

$$\Rightarrow P(|Y - E[Y]| \geq 6) \leq 0.12$$

So probability of getting 11 yellow smartie or more is no more than about 12%.

Calculating probability $P(Y \geq 11)$ with $Y \sim Bin(40, 1/8)$

$$P(Y \geq 11) = 1 - F_Y(10)$$

$$\approx 0.008$$

This shows that upperbound from chebychev's inequality is not that good.

13.3 Law of Large Numbers

Theorem: 13.5 For any $n \in \mathbb{N}$, let $X_1, \dots, X_n$ be an iid sample from a distribution with finite expectation, $E[X] = \mu$ and finite variance $Var(X) = \sigma^2$. Then

1) Weak law of large numbers:

$$\boxed{\lim_{n \to \infty} P(|\hat{X}_n - \mu| \geq \varepsilon) = 0 \quad \text{for any } \varepsilon > 0}$$

(convergence of probability)

2) Strong law of large numbers

$$\boxed{P(\lim_{n \to \infty} \hat{X}_n - \mu) = 1}$$

(The $\bar{X}_n$ converges to μ almost surely as $n \to \infty$)

proof: 1) proof of weak law

We have that

$$E[\hat{X}_n] = E[X_i] = \mu \qquad \text{by iid}$$

Also

$$Var(\hat{X}_n) = \frac{Var(X_i)}{n} = \frac{\sigma^2}{n}$$

From Chebychev's inequality (Thm 13.2), we have

$$P(|\hat{X}_n - \mu| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} Var(\bar{X}_n) = \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow P(|\hat{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Taking limit as $n \to \infty$ on both sides

$$\lim_{n\to\infty} P(|\hat{X}_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$$

$$\Rightarrow \lim_{n\to\infty} P(|\hat{X}_n - \mu| \geq \varepsilon) = 0$$

$\blacksquare$

Intuition: The intuition of weak law of large numbers is that if you take a large enough sample, so in the limit $n \to \infty$, then the sample mean is going to be really close to true value, So the probability that the expectation is far away is going to 0, i.e. $P(|\bar{X}_n - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$

weak law of large numbers is a limit of a probability

strong law of large numbers is the probability of a limit.

13.4 Consequence of the law of large numbers

Discuss how we can estimate probability of any event A by performing independent repetitions of probability experiment.

The intuitive idea is that the probability of the event could be approximated by the relative frequency with which event occurs in sample.

To formalise this intuition we are going to use the **indicator random variables for event A.**

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} = \mathbb{1}_A(\omega)$$

To understand utility of indicator random variables, in this context, we calculate its expectation:

$$E[\mathbb{1}_A] = 1 . P(\mathbb{1}_A = 1) + 0 . P(\mathbb{1}_A = 0)$$

$$= 1 . P(\{\mathbb{1}_A = 1\})$$

$$= 1 . P(A) = P(A)$$

$$\Rightarrow \quad E[\mathbb{1}_A] = P(A)$$

We see that probability of any event can be expressed in terms of its indicator random variable.

We already know from law of large numbers how to estimate expectations from an iid sample and so this will allow us to estimate probability of events from an iid sample.

To estimate $P(A)$,

Take iid sample $X_1, X_2, \dots, X_n$ from $X = \mathbb{1}_A$

The sample mean

$$\hat{X}_n = \frac{(X_1 + X_2 + \cdots + X_n)}{n}$$

is equal to first n repetitions of the probability experiment in which A occurs.

Also

$$E[\hat{X}_n] = E[X_i] = P(\mathbb{1}_A = 1) = P(A)$$

From laws of large numbers

$$\lim_{n \to \infty} \hat{X}_n = E[\hat{X}_n] = P(A)$$

almost surely. (by strong law of large numbers)

Given that we can estimate probabilities of any event, we can also estimate probability of the distribution function F_X of X because $F_X(x)$ is just the probability of the event $\{X \le x\}$. Thus $F_X(x)$ will be approximately equal to

$$\boxed{1/n \text{ times the number of } X_i \text{ less than or equal to } x.}$$

We can furthermore estimate the probability density with a histogram.
↳ as seen in R practicals.

We approximate the probability density at a point x using the number of sample values that lie in a small interval $[x-h, x+h]$ around that point, for h small.

$$\boxed{\begin{aligned} f_X(x) &\approx \frac{1}{2h} P(X \in [x-h, x+h]) \\ &\approx \frac{1}{2h} \cdot \frac{1}{n} \cdot \text{numbers of } X_i \text{ that lie in } [x-h, x+h] \end{aligned}}$$

A histogram shows bars for many such small intervals giving an approximation to $f_X(x)$