Toolkit

Lecture 2 – Σ's of iid variables, Gaussians, CLT,

[[A very typical scenario in analysis of algs: You have sthg (alg?) which "succeeds" w.p. p, you run it n times indep., you want to understand the total # of successes]]

Let $X_1, X_2, \ldots, X_n$ be i.i.d. [[indep & ident. distrib.]] r.v.'s with $\Pr[X_i = 1] = p$, $\Pr[X_i = 0] = q = 1-p$. [["Bernoulli"]]

Let $S_n = X_1 + \cdots + X_n$. [[try to understand it...]]

$E[S_n] = \underline{np} = E[X_1] + \cdots + E[X_n]$ [[linearity of expec didn't need X_i's indep]]

$Var[S_n] = \underline{np(1-p)}$

Rec: · $Var[Y] = E[(Y-\mu)^2]$ $\mu = E[Y]$, $= E[Y^2] - E[Y]^2$.
· $Var[Y+Y'] = Var[Y] + Var[Y']$ if indep. [ex]
· $Var[cY] = c^2 Var[Y]$
· $Var[X_i] = p - p^2 = p(1-p)$.

[Rule of life: if you ever have a r.v., consider making it mean 0 & variance 1.]

mean 0: $S_n \overset{E[S_n]\ (\mu)}{\underbrace{- np}}$

Var 1: $Z_n := \dfrac{S_n - np}{\text{stddev}[S_n]\ (\sigma)}$ $\left(\leftarrow \sqrt{Var[S_n - np]} = \sqrt{Var[S_n]}\right)$ $\rightarrow \sqrt{np(1-p)}$

[[lost no info:]] $\Pr[S_n \le u] = \Pr[\sigma Z_n + \mu \le u]$
$= \Pr\left[Z_n \le \dfrac{u-\mu}{\sigma}\right]$

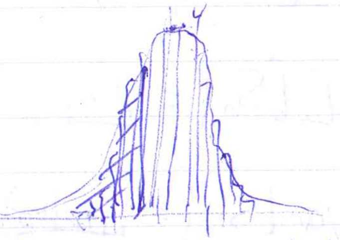
[[So study Z_n]]

eg: $p=\frac{1}{2}$ [[coinflips]]

$$\frac{X_1+\cdots+X_n-\frac{1}{2}n}{\frac{1}{2}\sqrt{n}} = \frac{1}{\sqrt{n}}\Big(\underbrace{2X_1-1+\cdots\cdots+2X_n-1}_{\text{indep }\pm1\text{ rvs.}}\Big)$$

rec: $\Pr[Z_n=0] = \underset{\sqrt{\frac{2}{\pi}}}{\Theta}\left(\frac{1}{\sqrt{n}}\right)$ [[Stirling]]

[[Plot the histogram of Z_n:]]
[[For any p. Even for $\Pr[X_i=1]=p_i$ for diff p_i's.]]



[[Appears to be "converging" to a fixed continuous distrib.]]
[[the famous Bell curve / Gaussian / normal r.v.]]

CLT: For any iid $X_1, X_2, \ldots$ [[not nec 0/1-valued]],

$$Z_n \rightarrow Z \leftarrow \text{"standard Gaussian"}$$

in that $\forall u \in \mathbb{R}$, $\Pr[Z_n \leq u] \xrightarrow{n\to\infty} \Pr[Z \leq u]$.

rem: This thm. useless. [[For practical purposes. No info on speed of cvgc. Also iid is a little lame]].

def: "$Z \sim N(\overset{\text{mean of }Z}{0},\overset{\text{var of }Z}{1})$" $\equiv$ "Z is a std. normal / Gaussian" means Z is a continuous r.v. with pdf

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

[[You're like: ew, gross? No. It's awesome. Learn to love it.]]

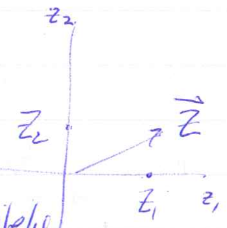
[[basically, want e^{-z^2}, make it a pdf, w/ var 1]] consts

THE FACT [[about Gaussians]] [[from which all else is derived]]:

Let $\vec{Z} = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$, where $Z_1, \ldots, Z_d \sim N(0,1)$ iid.

[[random vector]]

Then $\vec{Z}$'s distrib. is rotationally symmetric.



why? pdf of $\vec{Z}$ @ $\vec{z} = (z_1, \ldots, z_d)$

$= \phi(z_1)\phi(z_2)\cdots\phi(z_d)$ [[∵ indep.]]

$= \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\underbrace{(z_1^2 + \cdots + z_d^2)}_{\|\vec{z}\|^2}/2}$

[[equally likely anywhere on this circle; not equally likely radius, but ...]]

[[PDF of $\vec{Z}$ @ $\vec{z}$ is simply a fcn of length of $\vec{z}$, not angle.]]

Cor 1: $\boxed{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}}$ ⊛ $= 1$: it's really a pdf.

hmwk/omit?

Pf: [[Look at 2 indep gaussians]]. Let $f(z_1, z_2) = \frac{1}{2\pi} e^{-(z_1^2 + z_2^2)/2}$

Area under surf = ⊛^{2} ← [[int by vertical slices]]

so why is it 1?

Int by horz slices: [[seq. of disks]] $\frac{1}{2\pi}\int_0^1 \pi \cdot \underbrace{2\ln\tfrac{1}{h}}_{r^2}\, dh = 1$ [[ex]]

Rem: $\int\phi$ needs a trick, but $\int h(x)\phi(x)$ usually doable by int-by-parts; e.g., if $h(x) = x^k$, $k \in \mathbb{Z}^+$

Cor 2: Sum of indep Gaussians is Gaussian

[[If you believe sthg like the CLT, that sum of indep rv's → something (when standardized), then sums, indeed lin-combs of that sthg must have the same distrib.]]

[[This Cor 2 is why the limit is Gaussian. In turn, Cor 2 is basically b/c of the key fact.]]

def: Let $Z \sim N(0,1)$. Let $\mu, \sigma \in \mathbb{R}$ [[constants]]. Let $Y = \mu + \sigma Z$.

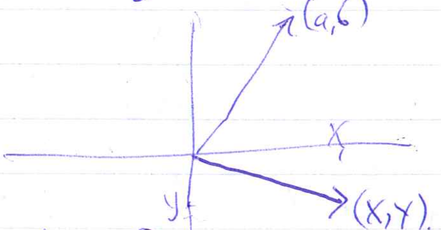
rem: $E[Y] = \underline{\mu}$. $Var[Y] = \underline{\sigma^2}$.

We call Y a Gaussian [[non-standard one]] too, write $\underline{Y \sim N(\mu, \sigma^2)}$.

[[non-important exs]] [[What is Y's pdf? Look up/derive it. Not important; just std'ize]]

cor2: Let $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$. indep. Let $Z = aX + bY$.
Then $Z \sim \underline{N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)}$.

pf: WOLOG (why?) $X, Y \sim N(0,1)$.
Want $\underset{Z=}{aX + bY} \sim N(0, a^2 + b^2)$.



$Z = (a,b) \cdot (X,Y)$ [[kinda like len of (X,Y)'s proj on (a,b); $\text{len}(a,b) \cdot \text{len}(X,Y) \cdot \cos(\text{angle})$.]]

(X,Y) rotationally symm. $\Rightarrow$ WOLOG can rotate (a,b) to x-axis
$$\downarrow$$
$$\underline{(\sqrt{a^2+b^2}, 0)}$$

Then $Z = (\sqrt{a^2+b^2}, 0) \cdot (X,Y) = \sqrt{a^2+b^2}\, X$
$\sim N(0, \sqrt{a^2+b^2}^2)$ by def^n.

□

[[Return now to CLT. As I said, it's kinda useless as stated, mostly b/c no error bounds.]]

Berry–Esseen Thm [[= CLT w/ error bounds]]

Let $X_1, \dots, X_n$ be indep. Assume $E[X_i] = 0$ $\forall i$, $Var[X_i] (= E[X_i^2]) = \sigma_i^2$, $\sum_i \sigma_i^2 = 1$ (WOLOG, why?)

Let $S = X_1 + \dots + X_n$ [[so $E[S] = \underline{0}$, $Var[S] = \underline{1}$.]]

Then $\forall u \in \mathbb{R}$, $\left| \Pr[S \le u] - \Pr[\underset{N(0,1)}{Z} \le u] \right| \le O(1) \cdot \beta$

where $\beta = \sum_{i=1}^{n} E[|X_i|^3]$.

$O(1)$: .56 [Shevtsova '10], .5514 (?B)

[[Is that error small?!]] Yes, in most cases

E.g., to study n coin flips, let

$$X_i = \begin{cases} +1/\sqrt{n} & \text{w.p. } 1/2 \\ -1/\sqrt{n} & \text{w.p. } 1/2. \end{cases}$$

[[mean 0: ✓]] $\sigma_i^2 = E[X_i^2] = \frac{1}{n}$, $\sum \sigma_i^2 = 1$ ✓.

$E[|X_i|^3] = \frac{1}{n^{3/2}}$, so $\beta = \underline{\frac{1}{\sqrt{n}}}$. [[☺]]

$\therefore$ $\forall u$, $\left| \Pr[S \le u] - \Pr[Z \le u] \right| \le \frac{.56}{\sqrt{n}}$

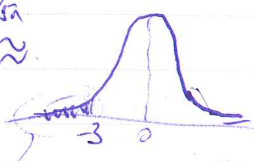
$$\Pr\left[\underbrace{\frac{\#\text{heads} - \overset{n-H}{\#\text{tails}}}{\sqrt{n}} \le u}\right]$$

$$\frac{2H - n}{\sqrt{n}} \le u$$

$$H \le \frac{n}{2} + u \cdot \frac{\sqrt{n}}{2}$$

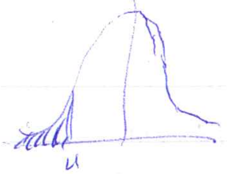
eg. $\Pr[\#H \le \frac{n}{2} - 1.5\sqrt{n}]$

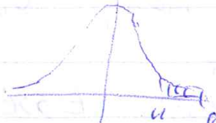
$\overset{.56/\sqrt{n}}{\approx}$



$\approx .001$. [[!]]

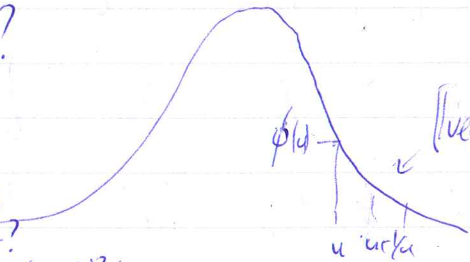
[[How do you get .001? Type into computer: Maple, WA, etc.]]

def: [[cdf of Gaussian]] $\Phi(u) = \int_{-\infty}^{u} \phi \overset{\Pr[Z \le u]}{\overset{\|}{=}}$ 

[[complem. cdf]] $\overline{\Phi}(u) = \int_{u}^{\infty} \phi \underset{\Pr[Z \geq u]}{\underset{\|}{=}}$  $= \Phi(-u)$.

[[less annoying b/c small for large pos u]]

Asymptotics of $\overline{\Phi}(u)$??



[[very rapidly decr.]]

$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$

What δ s.t. $\phi(u+\delta) \approx \frac{\phi(u)}{\text{const}}$?

$\delta = \frac{1}{u}$: $= \frac{1}{\sqrt{2\pi}} e^{-(u+\frac{1}{u})^2/2} \approx \frac{1}{e}\phi(u)$.

$\uparrow$ $-\frac{u^2}{2} - 1 - \frac{1}{2u^2}$

[[so look at rects of base $u, u+\frac{1}{u}, u+\frac{2}{u}+\cdots$]]

~~w~~ areas $\frac{\phi(u)}{u}, \frac{\phi(u)/e}{u}, \frac{\phi(u)/e^2}{u}, \ldots$

sum to $\Theta(\phi(u)/u)$.

prop: $\overline{\Phi}(u) = \Theta(\phi(u)/u)$. In fact, $\sim \frac{\phi(u)}{u}$.

In fact, $(\frac{1}{u} - \frac{1}{u^3})\phi(u) \le \boxed{\overline{\Phi}(u) \le \frac{1}{u}\phi(u)}$, $(u \geq 0)$

pf: ~~$\overline{\Phi}(u) \geq$~~ $\int_{u}^{\infty} \underset{x}{\wedge} \phi(x)\,dx \geq \int_{u}^{\infty} u\,\phi(x)\,dx$

$\|$

$\phi(u)$. b/c $\frac{d}{dx}\phi(x) = -x\phi(x)$

$\therefore -\phi(x)$ antideriv of $x\phi(x)$.

ex/hmwk.