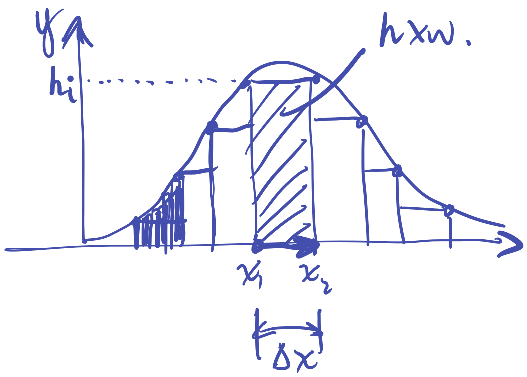
10/11/20

XDAS 2020 Class Notes – Normal Distribution

Reimann Sums ⇒ approx. area under curve.

⇒ For some function of x, $y = f(x)$



$w = \Delta x = x_2 - x_1$

$h = y = f(x)$

$x_i \in \{x_1 \ldots x_n\}$

- Divide x into bins, $\Delta x = x_2 - x_1$ n = total # bins.
- Pick some x_i b/w x_1 & x_2: $x_1 \le x_i \le x_2$ e.g. $x_i = x_1$
- Height × width $h_i \cdot \Delta x$ for each x_i approx. of area under curve for $[x_i + \Delta x]$ where h_i is height at x_i

Now, total area under curve is approximated by:

$$F(x) \approx \sum_{i=1}^{n} \underbrace{f(x_i)}_{\text{height}} \underbrace{\Delta x}_{\text{width}}$$

⇒ sum up areas of boxes $f(x_i) \cdot \Delta x = h_i \cdot w$

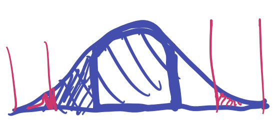
⊛ The integral is the limit area as Δx get smaller & smaller:

$$F(x) \equiv \lim_{\Delta x \to 0} \sum f(x) \Delta w = \int_{-\infty}^{\infty} f(x)\, dx$$ $\boxed{\text{Indefinite integral}}$

⇒ **Definite Integral**

- The area under curve spanning interval $[a, b]$ from $x = a$ to $x = b$:   (exclusive   ] inclusive

$$\int_a^b f(x)\,dx = F(x)\Big|_a^b = F(b) - F(a) = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx$$



PDF & CDF : $f(x)$ & $F(x)$

- For probability distributions, total area is by definition $= 1$.

Discrete: $\sum_{\forall x} f(x) = 1$

- $f(x)$ = height at x
- $x \in X = \{x_1, \ldots x_n\}$

Continuous: $\int_{-\infty}^{\infty} f(x)\,dx = 1$

- $f(x) = 0 \;\forall$ continuous variables

- Total probability that $X \leq x$ is (lower-tail)

Discrete $\Rightarrow F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$

Continuous $\Rightarrow F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$ (t is a variable for integration)

Discrete distributions

- Total prob that $a \leq X \leq b$

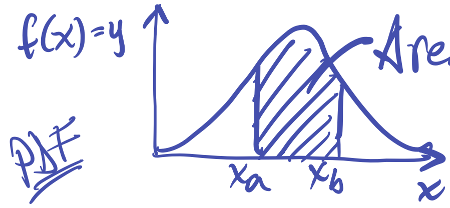
$$Pr(x_a < X \leq x_b) = \sum_{x_i = x_a + 1}^{x_b}$$

- Note that x_i ranges from x_{a+1} to x_b & $x_b > x_a$
- To get interval w/ a inclusive use $F(x_b) - F(x_{a-1}) = P(x_a \leq X \leq x_b)$

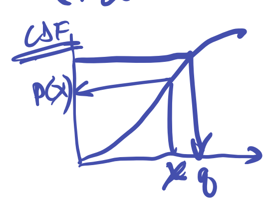
Continuous Distributions

- Total prob $x_a < X \leq x_b$

$$Pr(a < x \leq b) = \int_a^b f(x)\,dx = F\Big|_{-\infty}^{b} - F\Big|_{-\infty}^{a} = F(x_b) - F(x_a)$$



Area $= P(x_a \leq X \leq x_b)$



$p = Pr[X \leq x]$

$q = x$ such that $Pr[X \leq x] = p$

The Normal Distribution

$$f(x) = \underbrace{\frac{1}{\sigma\sqrt{2\pi}}}_{③}\, \underbrace{e^{\overbrace{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}^{②}}}_{①} \quad ?$$

⇒ What controls the shape & location & size of resulting curve?

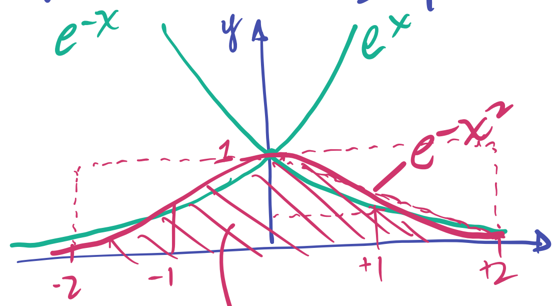
① e^{-x^2} ⇒ bell shape.

$e^0 = 1$

$e^{-0^2} = 1$

$e^{-1^2} = 1/e = 1/2.73..$

$e^{-(-1)^2} = 1/e = 1/2.73$



$\sqrt{\pi} = 1.77$. Area

② Exponent ⇒ width & offset : $\left[-\frac{1}{2}\boxed{\left(\frac{x-\mu}{\sigma}\right)^2}\right]$

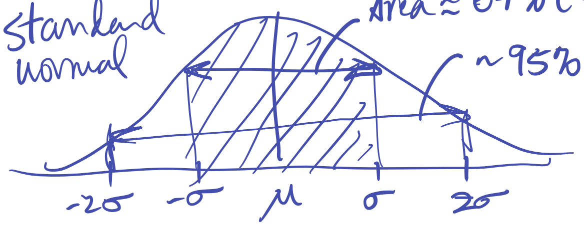
$$Z = \frac{x-\mu}{\sigma} \Rightarrow x = \underbrace{\sigma Z}_{\text{spread}} + \underbrace{\mu}_{\text{location}}$$

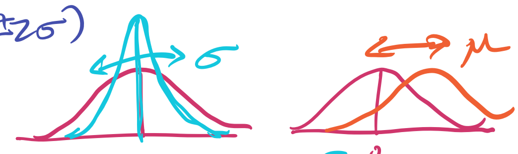
$\mu = 0$
$\sigma = 1$
Total area = 1

standard normal

Area ≈ 67% (2/3) · ±1σ

~95% (±2σ)





CLT & $SE_{\bar{\mu}} \Rightarrow \frac{\sigma}{\sqrt{n}}$

③ Coefficient ⇒ scaling factor $\left[\frac{1}{\sigma\sqrt{2\pi}}\right]$

⊕ Area under curve = 1 ⇒ Total Probability!

(Derives from 2D solution using polar coord's)

Normal approximation of binomial

$$X \sim BIN(n,p) \xrightarrow[p \simeq 0.5]{n \text{ large}} N(\mu, \sigma) \quad \begin{cases} \mu = np \\ \sigma = np(1-p) = npq \end{cases}$$

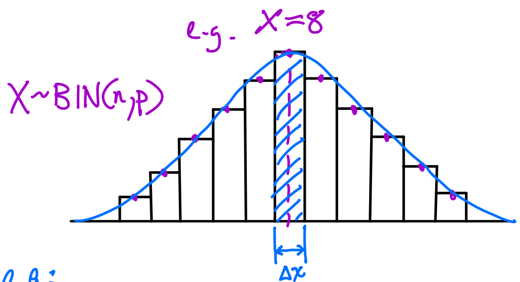
(not too large or too small)

$(q = 1-p)$

$\Rightarrow np \geq 5$ & $qp \geq 5$

Continuity correction for normal approximation of a binomial

- To approximate a discrete distribution with a continuous one, we need to compute the area under a curve across a range of values (since the area under a single point is 0):



Binomial distribution for $\begin{cases} n=16 \\ p=0.5 \end{cases}$

$\mu = Exp(X) = np = 8$

$\sigma = np(1-p) = 16 \cdot 0.5^2 = 4 \Rightarrow s.d. = 2$

e.g.:

In order to compute $Pr[X=8]$ using the normal approx, we need to get the area under a normal curve across an interval that spans $X=8$, and is adjacent to $X=7$ & $X=9$. The width of $\Delta X = 1$, so we need to get $Pr[8-0.5 < X < 8+0.5] = Pr[7.5 < X < 8.5]$.

More generally, we need to follow these rules:

	Binomial (discrete)	Normal (continuous)	
	$P(X=x)$	$P(x-0.5 < X < x+0.5)$	
lower tail	$P(X \leq x)$	$P(X < x+0.5)$	x inclusive
	$P(X < x)$	$P(X < x-0.5)$	x exclusive
upper tail	$P(X \geq x)$	$P(X > x-0.5)$	x inclusive
	$P(X > x)$	$P(X > x+0.5)$	x exclusive