Lecture 4 (Ch. 1)

Something completely different, eventhough it may not look different

We have been talking about data, and histograms of data.
A histogram pertains to data.

But There is something else That looks like a hist, but it's NOT:

Distribution

A Huge and Tricky concept

A dist. is a purely mathematical Thing That has nothing to do with data. So, for now, forget data (and hists).

Example: $y \sim f(x) \sim e^{-\frac{1}{2}x^2}$

Technically, This f(x) is not a distribution! See next page.
But it's good enough to make The important point That a dist. is a purely mathematical Thing (ie. a function), not a histogram.

Don't be tempted to Think of a dist as a "fit" to a hist. It's not!

The variety of shapes for dists is similar to That of hists.
They even have the same names (bell-shaped, ---).
This can add to The confusion between Them. Beware!

Big picture:
The bridge we are attempting to make between sample and population, has hists on the sample side and dists on the population side. Said differently, hists are used to represent the sample/data, while dists are used to represent (almost define) the population. Later, we are going to learn how to tell something about the population/distribution from a sample/histogram by comparing them.
But, for now, and most importantly, think of hists and dists as completely unrelated.

Here is the precise definition of a distribution.

Defn: A distribution, $f(x)$, $p(x)$, must satisfy:

1)	$f(x) \geq 0$	$p(x) \geq 0$
2)	$\int_{-\infty}^{\infty} f(x)\,dx = 1$	$\sum_{\text{all } x} p(x) = 1$

eg. x = Computer Brand

$\sum_x p(x) = p(\text{Mac}) + p(\text{Dell}) + p(\text{HP}) + \cdots = 1$

For x = Continuous, $f(x)$ is called The prob. density function (pdf)

For x = Discrete or Categ, $p(x)$ " " prob. mass function (pmf)

→ Generally, $f(x)$ and $p(x)$ are called distributions.

$\boxed{\text{Example}}$: $f(x) = e^{-\frac{1}{2}x^2}$, $-\infty < x < \infty$, is not a dist, because

$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx$ = remind yourself how to do such integrals $= \sqrt{2\pi} \neq 1$

So, $\boxed{f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}$ is a dist. (also $f(x) \geq 0$ ✓)

This $f(x)$ is very famous. It's called The standard Normal pdf

$\boxed{\text{Example}}$: $f(x) = k x^8 (1-x)$, $0 < x < 1$ dist?

$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 k x^8 (1-x)\,dx = k \cdot \frac{1}{90} \neq 1$ unless $k = 90$.

So, $f(x) = 90 x^8 (1-x)$ is a distr. (also $f(x) \geq 0$ ✓)

All of The above examples have been for x = cont.

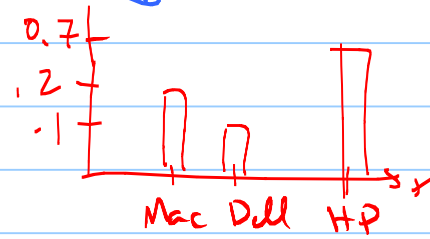
$\boxed{x = \text{categorical}}$ (Harder because one cannot write formulas. Instead, use Tables or charts)

E.g. x = "computer Brand"

x	Mac	Dell	HP
$p(x)$	0.2	0.1	0.7

$p(x) \geq 0$ ✓

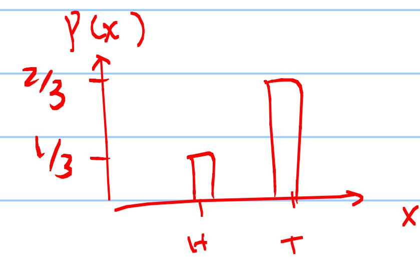
$\sum_x p(x) = 1$ ✓



E.g. x = "state of an unfair coin"

e.g.

x	H	T
$p(x)$	$1/3$	$2/3$



If we encode $\overset{x = \text{Categ}}{\overbrace{x = H, T}}$ as $\overset{x = \text{Discrete}}{\overbrace{x = 1, 0}}$, Then $p(x) = \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{1-x}$

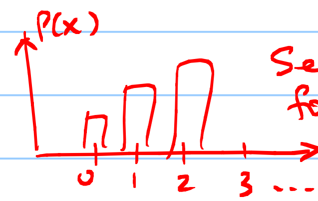
That dist. is a special case of $p(x) = (\pi)^x (1-\pi)^{1-x}, \; x = 0, 1, \; 0 < \pi < 1,$

Called The $\boxed{\text{Bernoulli dist.}}$

$\boxed{x = \text{Discrete}}$ (Easier because we can write formulas)

E.g. x = "number of heads out of n tosses of a fair coin."

$$p(x) = \frac{n!}{x!(n-x)!} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x}, \; x = 0, 1, \ldots, n$$



See lab for shape.

This is a special case of The $\boxed{\text{Binomial dist.}}$ which we will derive later.

$\boxed{\text{IMPORTANT WARNING}}$: The plots above are NOT histograms; They are distributions. Two very different Things.

hist. refers to sample (from data); dist. refers to pop. (from Math).

Recall the connection between $\boxed{\text{hists}}$ and prob (or prop.).

If x = Discrete/Catg. $pr(a \le x \le b) = \sum_x$ height of rel. freq. hist at x

Eg. x = "computer type" $\in$ {Mac, Dell, HP}

$pr(x = \text{Mac or Dell})$ = height at Mac + height at Dell.

If x = cont., $pr(a < x < b)$ = some kind of area under hist.

Recall $pr(x=a) = 0$

Similarly for $\boxed{\text{dists}}$: Just change "hist" to dist, above.

One difference: Because dists are mathematical functions, we can find the areas (probs) with nice sums or integrals:

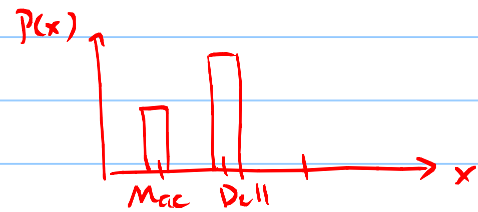
If x = Discrete/Catg.

distr.

$$pr(x \in \{\ldots\}) = \sum_{x \in \{\ldots\}} p(x)$$

prob.

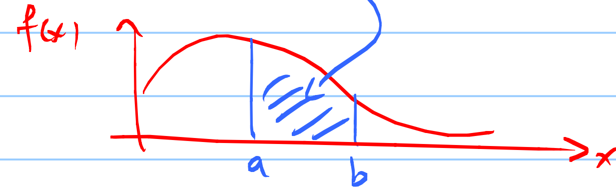
Eg. $= p(\text{Mac}) + p(\text{Dell})$



If x = Cont.

distr.

$$pr(a < x < b) = \int_a^b f(x)\,dx$$



$\boxed{\text{Eg.}}$ for $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

prob

$$pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \text{some number}$$

In all of the above, probability simply refers to the proportions of times that something happens. So prob = prop!

Key points: (Sample vs. pop.) (hist. vs. dist.)

hw-lect4-1:
Consider the Bernoulli dist. with parameter pi:

$$p(x) = \pi^{x}(1-\pi)^{1-x}, \quad x=0,1 \qquad 0<\pi<1$$

a) Show that it's a distribution (prob. mass function).
b) Find the prob that x=1.

hw-lect4-2:
Based on data, we have observed that x is between 0 and 1/2 about 25% of the time. Which of the following is the more reasonable distribution from which our data may have come? Show work (always!)

A) $f(x) = 2x \quad 0 \leq x \leq 1$ B) $f(x) = e^{-x} \quad 0 \leq x < \infty$

hw_lect4_3
What's the prob of getting 1 or 2 boys in a sample of size 10, taken from a population in which the proportion of boys is exactly 50%?