Lecture 3 (Ch. 1)

We are talking about hists and dists

counts of data — a mathematical function

For now, They are completeley different Things. In stats, hists are used to describe data/sample, and dists are used to describe The population.

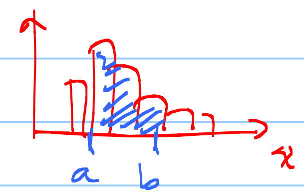
Now, we need to talk about a few other features of hists, because They allow us to relate hists to dists.

Two variations on histograms are

Relative freq histograms: Rel. freq. = freq. / total sample size.

Density scale histograms: Rel. freq. / bin size.

Density histograms have a nice property:



Area = proportion of something = probability of something.

E.g. area for $a<x<b$ = prob. of $a<x<b$.

$\underbrace{\hspace{6em}}_{\text{Important!}}$

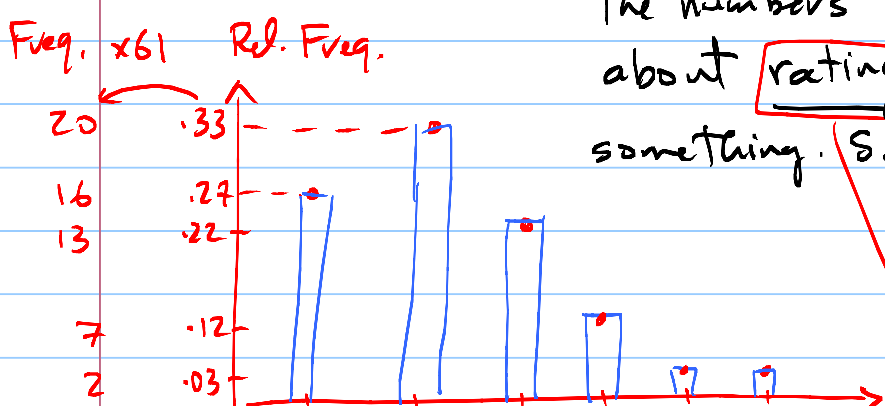
Caren Marzban Other SP16

Form G: Lecture -- Assignments "61" surveyed "124" enrolled

Question	Excellent	Very Good	Good	Fair	Poor	Very Poor	Median
The course as a whole:	27%	33%	22%	12%	3%	3%	3.80
Textbook overall:	33%	30%	27%	10%	0%	0%	3.94
Instructor overall:	50%	28%	10%	7%	2%	3%	4.50
Instructor's contribution:	42%	27%	15%	8%	3%	3%	4.22
Instuctor's interest:	53%	26%	7%	5%	2%	7%	4.56
Amount learned:	39%	27%	20%	8%	3%	2%	4.09
Relevance and usefulness of homework:	37%	17%	27%	12%	3%	3%	3.75

For median calculation: 5 = Excellent 4 = Very Good 3 = Good 2 = Fair 1 = Poor 0 = Very Poor

The numbers on each row say something about rating, i.e. how the students rated something. So, The random variable is rating.



Rating

Ex VG G F P VP

5 4 3 2 1 0

quantitative (discrete)

qualitative (categorical)

median ?#!

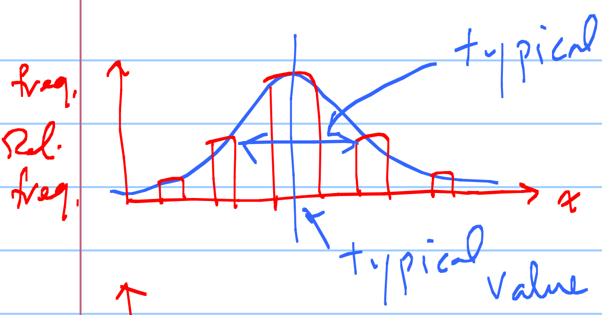
median = 3.80

So, median (and mean) make sense only for quantitative data.

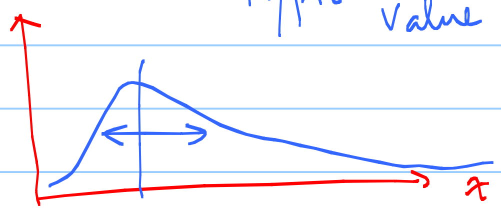
Interpert: Before { center ~ 3 or 4; spread ~ 1, 1.5; shape ~ not symmetric } Now

$\boxed{\text{area}}$ = .27 + .33 = 60% of students say Excellent or Very Good etc.

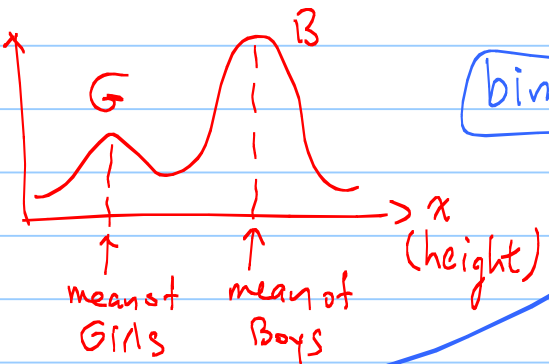
These are some other shapes that you may come across
(These shapes arise in both hists and dists)



bell-shaped



skewed (to the right)

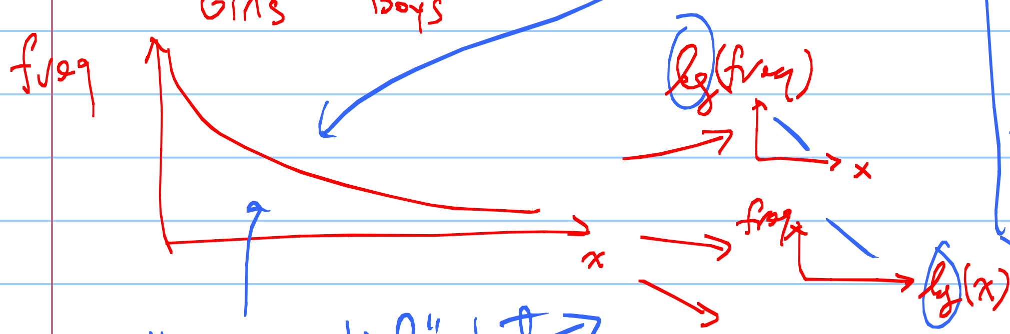


bimodal

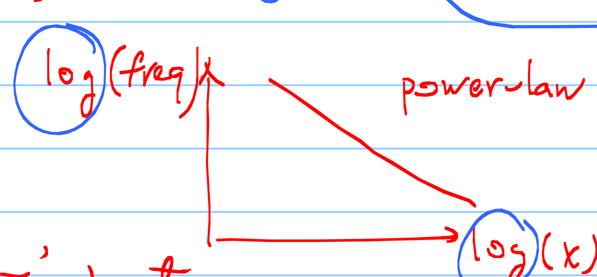
NOT some variable as a function of time!

NOT some variable (e.g. demand) as a function of some other variable (e.g. supply).

A hist is a plot of freq. (or rel. freq, ...) of different values of ONE variable.



"exponential" but ⇒ - looking



power-law

Eg.

x = magnitude of earthquakes

= population of cities, on the planet

= length of words, in a book

= casualties of wars, for different wars

$\log(\text{freq}) = m \log(x) + b$

$= \log x^m + \log b' = \log(b' x^m)$

$\therefore \text{freq} \sim x^m$

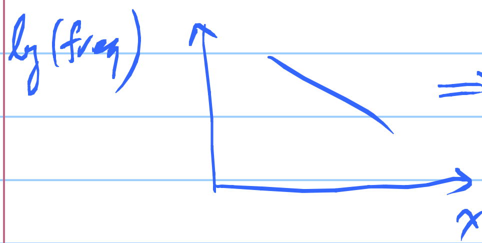
leave out all constants

Q1: Suppose you are looking at an exponential-looking hist.,and then note that if you plot th log of the frequencies vs. x, then the histogram looks linear with a negative slope. what is the relationship between freq. and x?

a) $freq \sim x$ $\quad$ b) $\boxed{freq \sim e^{-x}}$ $\quad$ c) $e^{-freq} \sim x$ $\quad$ d) $e^{-freq} \sim e^{-x}$

When poll is active, respond at **PollEv.com/marzban**

Text **MARZBAN** to **22333** once to join



$$\Rightarrow \quad \log(freq) = -mx + b$$

$$freq = e^{-mx+b} = e^{-mx} \cdot e^{b}$$

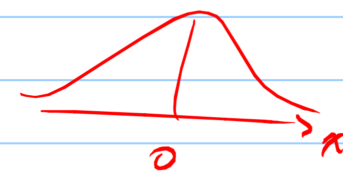
$$freq \sim e^{-x}$$

So, this is how you can identify a <u>real</u> exponential hist by plotting the log of the freqs and seeing straight line.

Now, back to $\boxed{\text{dists:}}$

Last time I gave The example $y = f(x) \sim e^{-\frac{1}{2}x^2}$

But, technically That's not quite a dist.



Here is The precise definition:

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 continuous x, $f(x)$ is called a density function
For discrete x, $p(x)$ is called a mass function
} generally called distribution

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

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

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

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

observed or sample

Recall that areas under hists ~ proportion of times ... ~ prob. of ...

based on data i.e. for sample

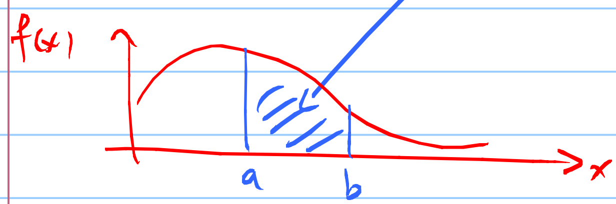
Theoretical or population

Similarly, areas under dists ~ proportion of times ... ~ prob. of ...

based on distribution i.e. for population

proportion/prob. of $a < x < b$

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

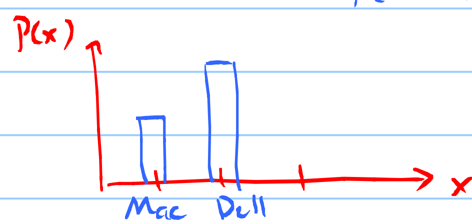


prop./prob. of $x = \{\ldots\}$

E.g. Mac, Dell, ...

$$\sum_{x=\{\}} p(x)$$

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



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

$$\underbrace{\text{prop}}_{\text{or prob.}}(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx = \text{whatever}$$

$\boxed{\text{E.g.}}$ for $f(x) = 90\,x^8(1-x)$

$$\underbrace{\text{prop}}_{\text{or prob.}}(a < x < b) = \int_a^b 90x^8(1-x)\,dx = \text{whatever}$$

hw-lect3-1

For each of the following shapes, come-up with atleast 1 example of a quantity x (a random variable) whose histogram you expect to be approximately

a) Bell-shaped (symmetric)
b) skewed (one way or the other)
c) exponential - looking
d) Bimodal

Describe the quantity clearly, and explain in words why you expect the particular shape. If you have data to support your expectation, then go ahead and show the histogram. (For this problem, x may be continuous or discrete.)

hw-lect3-2: In the above lecture note, there exists atleast one random variable that when considered as quantitative, has an exponential-looking histogram. Identify one of them, and plot its rel. freq. hist. (By hand).

hw-lect3-3

Consider the density function $f(x) = \begin{cases} 0 & \text{else} \\ a(-x^3 + x^2 + x + 2), & 0 < x < 2 \end{cases}$

a) First, determine a to make sure $f(x)$ is a density function.

b) Compute the prob. that x will be between 0 and 1.

c) Use R to plot $f(x)$. Include code and figure.

This problem is basically an exercise in calculus.