Lecture 3 (Ch. 1)

Here is one more use of a histogram that should interest all of you. Just concentrate on the hist: you will learn about the rest, later.

All of this suggests that attending 390 lectures is associated with higher test grades. This is from only 4 quarters, but the same pattern exists for every quarter!
Now: A huge, but tricky, change in topic!

Distributions.

A histogram pertains to data.

But there is something else (called distribution) that looks like a histogram, but is not. In fact, distributions have nothing to do with data. So, for now, forget about data.

Later, we will use distributions to "describe" the population.

Example: \( y = f(x) \sim e^{-x^2} \)

For continuous \( x \), \( f(x) = \text{density function} \)

For discrete \( x \), \( p(x) = \text{mass function} \)

To be more precise,

**Definition:** A density function \( f(x) \), or mass function \( p(x) \) must satisfy

1) \( f(x) \geq 0 \) \( \iff \) \( p(x) \geq 0 \)

2) \( \int_{-\infty}^{\infty} f(x) \, dx = 1 \) \( \iff \) \( \sum_{x} p(x) = 1 \) (i.e. \( \sum_{i=1}^{n} p_i = 1 \))

E.g. \( p(\text{apple}) + p(\text{orange}) + \cdots = 1 \)
\( x = \text{fruit} + \text{type} \).
Like density histograms, density functions and mass functions can be used for computing (mathematical) proportions.

E.g. \( f(x) \uparrow \)

\[
\int_a^b f(x) \, dx \quad \frac{\int_a^b f(x) \, dx}{\int_{-\infty}^{\infty} f(x) \, dx} = \text{area} = \text{area}
\]

The proportion of \( x \) values between \( a, b \quad (a < b) \)?

\[
\frac{\int_a^b f(x) \, dx}{\int_{-\infty}^{\infty} f(x) \, dx}
\]

E.g. \( f(x) = k \times^8 \left(1-x^2\right) \), \( 0 < x < 1 \). \( \text{prop} \left(a < x < b\right) = \) ?

\( f(x) = \text{density if} \quad \int_{-\infty}^{\infty} f(x) \, dx = 1 \), \( f(x) > 0 \)

\[
\int_{-\infty}^{\infty} f(x) \, dx = k \int_0^1 x^8 \left(1-x^2\right) \, dx = k \cdot \frac{1}{10} \quad \Rightarrow \quad k = 90
\]

Then \( \text{prop. between } a \text{ and } b = \int_a^b 90 \times^8 \left(1-x^2\right) \, dx \)

Important: histograms and dists are completely unrelated (for now).

Later, we will use distributions to "describe" the population, and histograms to describe samples, and ask if a sample could have come from some population.
Examples of dists (mass functions and density functions):

\[ f(x) \quad \text{cont.} \]

\[ f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} & -\infty < x < \infty \\ f \geq 0 & \end{cases} \]

Called "Standard normal distr".

Examples (\(x = \text{categorical}\))

\(x = 3\) food items.

<table>
<thead>
<tr>
<th>(x)</th>
<th>nuts</th>
<th>cheese</th>
<th>wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p(x))</td>
<td>0.2</td>
<td>0.1</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\[ \sum p(x) = 1 \]

Note: There is no data anywhere here. These are not histograms or formula.

\(x\) = "State of a fair coin" \(p(x)\)

\[ p(x) = \begin{cases} 0.6 & x = T \\ 0.4 & x = H \end{cases} \]

\((\text{Bernoulli distr.})\)

\(x\) = "Number of heads out of \(n\) tosses of a fair coin."

Then,

\[ p(x) = \frac{n!}{x!(n-x)!} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x} \quad (\text{Binomial distr.}) \]

We will derive this \(p(x)\), later.

This \(p(x)\) can be used to describe the population of \(x\) later, we will replace the \(1/2\) with other values.
we even talk about mean (and median, ...) of a variable, \( x \), whose distr is \( f(x) \), but even those have nothing to do (yet) with mean (and median...) of data.

**Mean of \( x \) (or of \( f(x) \)):**

\[
\text{mean of } x \ (\text{or of } f(x)) = \int_{-\infty}^{\infty} x f(x) \, dx
\]

**Median of \( x \) (---):**

\[
\text{median of } x \ (---) = \int_{-\infty}^{\text{median}} f(x) \, dx = \frac{1}{2} = \int_{\text{median}}^{\infty} f(x) \, dx
\]

**Mode ---:**

\[
\frac{df(x)}{dx} \bigg|_{\text{mode}} = 0
\]

Again, the computation of these quantities requires \( f(x) \), or \( px \), i.e., the density/mass functions.

The corresponding quantities for data are computed differently, but are called by the same names, a poor but common practice.

Once again: histograms \( \leftrightarrow \) sample/data

\[ \text{distributions} \leftrightarrow \text{population} \]