Lecture 8
(Ch. 2)

Dealing with ambiguity (e.g., how wide is a curve)
Dealing with precise syntax (e.g., words and their order matters)
Population (the truth)
Sample (our observed data)
Random Variable
Types of data (well-defined, but ambiguous)
Histogram (its interpretation and uses)
Probability from histogram
Distribution (its interpretation and uses)
Probability from distribution
Named distributions
The random variable "template" associated with each distribution
Standardization (for Normal and other distributions)
Percentile/quantile (for sample and/or dist)
boxplots (uses and interpretation)
Derivation and application of Binomial and Poisson.

Summary of what you have learned, so far.

Time to quantify some of our qualitative ideas.

In prev. chapters we played with histograms of sample/data and distributions of (random) variables (cont. and discrete/Categ.).

Hists and dists are pillars of statistics. Hists describe the data/sample, while dists describe the population. One question we often ask is this: how likely is it that my data/hist came from, say, a normal dist with params $\mu = \dots$, $\sigma = \dots$?

One way to compare histograms with distributions is in terms of their summary measures. For example, we can compare the "location" of the Normal distr. (μ) with the "location" of the histogram.
Or the "width" of the former (σ) with the width of the histogram.

So, we need some measure of location and width for both histograms and distributions (for any dist, not just Normal).

	hist/sample	dist/pop
measure of location	?	?
measure of width	?	?

The first comparison between sample and pop. will happen soon (when we do qq-plots), and then more fully in the 2nd half of 390.

Measures of location for hist/sample:

The x for the ith case

- sample mean : $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

pros/cons

- sample median : $\tilde{x}$ = middle of the ordered data.

data

Measures of spread for hist./sample:

- Sample Range

standard deviation (same units as $\bar{x}$)

pros/cons

Data

- Sample variance $\equiv s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\underbrace{x_i - \bar{x}}_{\text{deviation.}})^{2}$

$s \sim$ "average" (typical) deviation.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - \bar{x}\underbrace{\sum_{i=1}^{n} 1}_{n} = 0.$$

$\boxed{\text{Example}}$: $x = c(1, 3, 8)$ cm

$s = \sqrt{13}$ cm

$\bar{x} = \frac{1}{3}(1+3+8) = 4$ cm

$s^2 = \frac{1}{3-1}\left[(1-4)^2 + (3-4)^2 + (8-4)^2\right] = \frac{1}{2}(9+1+16) = 13$ cm^2

Again: A lot of statistics is about explaining/understanding this variance. Recall, if there were no variance/change, we wouldn't say that we even have any data.

In short, we will use The following summary measures for location and spread of data:

Sample mean : $\bar{x} = \frac{1}{n} \sum_i^n x_i$

sample variance : $s^2 = \underbrace{\frac{1}{n-1} \sum_i^n}_{\text{"funny Average"}} (\underbrace{x_i - \bar{x}}_{\text{deviation}})^2$ ← Because $\sum_i (x_i - \bar{x}) = 0$

Then s will be another measure of spread, and it's even better Than s^2, because s has The same physical dimension as x itself. So, we can write Things like $\bar{x} \pm s$ as a way of summarizing a histogram.

Important: Interpretation of $\bar{x}$ is typical x

" " s_x (or s^2) " typical deviation of x.

In some problems where The $\frac{1}{n-1}$ is not important, one focuses on $S_{xx} = \sum_i^n (x_i - \bar{x})^2$, ie. just The numerator of s^2.

Finally, note That all of These measures have The word "Sample," reminding you That They pertain to sample/data, not pop./distr.

FYI

For The mathematically-inclined: If you Think of x_i as The components of an n-vector, Then after you "center" each component (ie. subtract $\bar{x}$), s is proportional to The magnitude of That vector.

Another way of computing s^2.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\underbrace{\sum_i x_i}_{n\bar{x}} + (\bar{x})^2 \overbrace{\sum_i 1}^{n}\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2\right]$$

$$= \frac{1}{n-1}\left[n\overbrace{\left(\frac{1}{n}\sum x_i^2\right)}^{\overline{x^2}} - n(\bar{x})^2\right] = \frac{n}{n-1}\left[\overline{(x^2)} - (\bar{x})^2\right]$$

In Summary:

$$s^2 = \frac{1}{n-1}\sum_i (x_i-\bar{x})^2 \quad \text{"Defining formula"}$$

$$s^2 = \frac{n}{n-1}\left[\overline{x^2} - (\bar{x})^2\right] \quad \text{"Computational formula"}$$

- sometimes more useful
- always faster (1 vs. 2 loops)
  - Not too important.

Example

$x = c(1, 3, 8) \to x^2 = c(1, 9, 64) \to \overline{x^2} = \frac{74}{3}$

$s^2 = \frac{3}{2}\left[\frac{74}{3} - 16\right] = \frac{3}{2}\,\frac{74-48}{3} = \frac{26}{2} = 13$ (same as above).

Keep the "big picture" in mind: We are looking for

		sample/hist.	pop./distr.
measure of	location	Sample mean $\bar{x} = \frac{1}{n}\sum_i^n x_i$ $\bar{x} \sim$ "typical x"	?
	spread	Sample variance (s^2) Sample std. dev. (s) $s^2 = \frac{1}{n-1} \underbrace{\boxed{\sum_i^n (x_i - \bar{x})^2}}_{S_{xx}}$ $= \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)$ $s \sim$ "typical deviation in x"	?

s, s^2, and S_{xx} are all measures of spread.

Now, we need to come-up with corresponding things in the $\boxed{\text{pop.}}$

So, switch to distributions ($p(x)$, $f(x)$). No Data/Sample!

1) Distribution mean (or Expected Value) $= \overbrace{\mu_x = E[x]}^{\text{different symbols for the same thing.}} = \begin{cases} \sum_x x \; p(x) \\ \int x \; f(x) \, dx \end{cases}$

Motivation: Even though we are now in the realm of math $(p(x), f(x))$ not data, just to motivate the defn of $E[x]$, consider the following "population" $\{3, 2, 2, 1, 3, 2, 3, 1, 2, 2\}$.

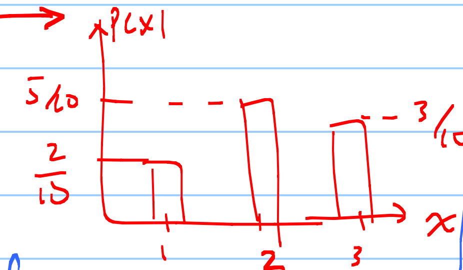
"avg." $= \frac{1}{10}[3 + 2 + 2 + \cdots]$

$= \frac{1}{10}[3(3) + 5(2) + 2(1)]$

$= \frac{3}{10}(3) + \frac{5}{10}(2) + \frac{2}{10}(1) = \sum_{\forall x} p(x) \cdot x$

$\leftarrow$ distinct values of x

where



Compare :

Sample mean : $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\sim$ typical x (in dist. / pop)

distr. mean (Expected Value) : $\mu_x = E[x] = \sum_x x \; p(x), \int_{-\infty}^{\infty} x \, f(x) \, dx$ $\sim$ typical deviation in x.

the book drops the x on μ_x, but then μ can be confused with the parameter of the Normal distr.

As examples, let's find the mean of Normal & poisson, below.

FYI

$E[x]$ does not mean that E is a function of x. In fact, E is a $\sum_x$ or an $\int_{-\infty}^{\infty} dx$, and so it is not a function of x.
$E[x]$ simply means that you need $p(x)$ or $f(x)$ to find it.
See examples, next left; There is no x-dependence in $\mu_x = E[x]$.

The mean of $N(\mu, \sigma)$:

$$\mu_x \text{ (or } E[x]) = \int_{-\infty}^{\infty} x f(x)\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Change of var.

$$\frac{x-\mu}{\sigma} = z$$

$$\frac{dx}{\sigma} = dz$$

$$= \frac{1}{\sqrt{2\pi\cancel{\sigma^2}}} \int_{-\infty}^{\infty} (\mu + \sigma z)\, e^{-\frac{1}{2}z^2}\, \cancel{\sigma}\, dz$$

$$= \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz}_{=1} + \sigma \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z e^{-\frac{1}{2}z^2} dz}_{=0}$$

Either look-up in Table of integrals, or note That z is odd, while The $\int$ goes from $-\infty$ to $+\infty$.

$$\mu_x = \mu$$

Now, you can see why The μ parameter of The normal distr. is called it's mean.

You can also see why The subscript on μ_x is important!

Mean of Poisson(λ):

$$\mu_x = E[x] = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \overset{\text{change of var.}}{=} \cdots = \lambda \underbrace{\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}}_{1 = \sum_x p(x)} = \lambda$$

Now you can see why The λ param. of Poisson is called its mean.

$\boxed{\text{Warning}}$: Don't confuse $\bar{x}$, μ_x

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

distr. mean

→ For $N(\mu, \sigma)$, $\mu_x = \mu$

→ For Poisson(λ), $\mu_x = \lambda$

→ For other distributions, μ_x will be other Things (next!)

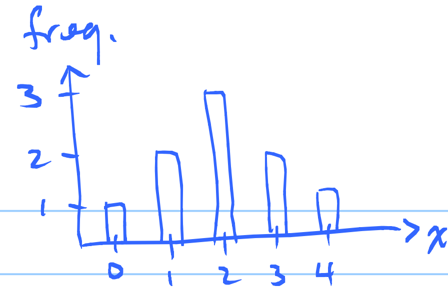
hw_lect8_1

Consider the adjacent histogram. $\bar{x}$

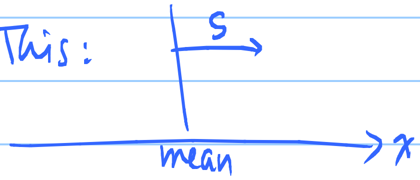
a) Compute/find the sample mean x_bar. Show work.

b) Find The sample std. dev. s, using the defining formula.

c) Draw the mean and the sample std. dev. on the histogram.



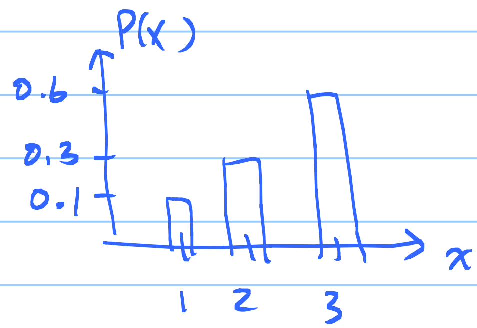
like This:

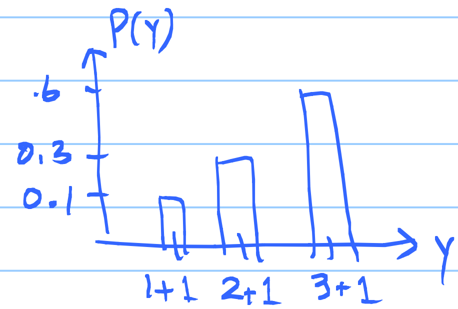


hw_lect8-2

(a) For the bottom/left distribution, find the distr. mean mu_x = E[x].

(b) For the bottom/right distribution, find the distr. mean mu_y = E[y]. Note that y = x + 1.





hw_lect8_3:

To understand our formulas, it is often useful to see what they say if our data are just a bunch of 0's and 1's. So, suppose our data consists of n0 zeros and n1 ones. Note: n = n0 + n1 .

a) Find the sample mean

b) Show that the sample variance is (n0 * n1)/(n(n-1))

hw-lect8-4

For the exponential distribution with parameter lambda, find the mean.

Hint: You may use this integral: $\int_0^\infty y e^{-y}\, dy = 1.$