Lecture 7 (Ch. 1-2)

Last time we derived The Binomial dist.

$$p(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \qquad x = 0, 1, 2, \cdots, n$$

" prob $(X=x)$

" prop. of $X=x$.

short for $\frac{n!}{x!\,(n-x)!}$

" n Choose x "

Look at The clicker qz we skipped to remind you what are all The proportions :

A sample is taken from a population of boys & girls. The binomial mass function provides The proportion of _____ with Certain characteristics.

A) all Samples B) people in The sample C) people in The pop.
Boy or Girl Boy or Girl

D) None of The above

Now, for large n, The n! gets nasty.
But, sometimes π is small. So, consider This limit:

$n \to \infty$, $\pi \to 0$ [ie. rare events] while $n\pi =$ Const. $\equiv \lambda$.

$$\binom{n}{x} \pi^x (1-\pi)^{n-x} \xrightarrow[\text{approx.}]{} \frac{e^{-\lambda} \lambda^x}{x!} = p(x) \text{ of Poisson with param } \lambda.$$

(Table III)

No n! , No π. Just λ (average of someThing, next!)

In The example, since we know n & π, we can compute λ :

$$\lambda = n \cdot \pi = 100(.005) = 0.5$$

Then prop. $(X=0) \simeq \frac{e^{-\lambda}\lambda^0}{0!} = e^{-.5} = \boxed{0.6065}$

exact answer .6058 ← see last lect.

Similarly for prop $(X=1, 2, 3), \ldots \approx$ exact answers from binomial.

Although I derived Poisson as a large-n (small-π) limit of binomial, it turns out that some problems can be solved with Poisson, quite independently of Binomial, when you have λ (average rate) but not n or π.

Examples of data that follow the poisson distr:

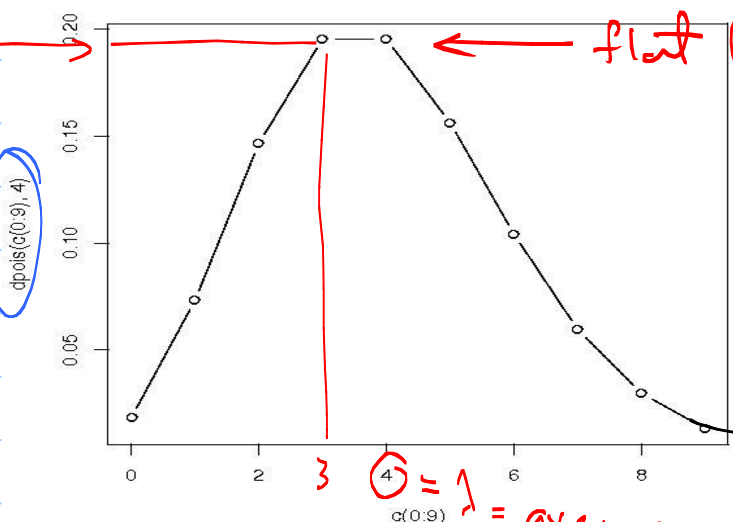
- \# of bombs dropped over London per block.
- \# of potholes per unit length of roads.
- \# of crashes (cars, planes, buildings) per year.
- \# of people arriving at a website per unit time. $= X$

Eg. An avg. of 4 people arrive at a website per hour. What's the prob. that 3 people arrive per hour?

proportion of hours

Assume X = poisson with $\lambda = 4$ people/hr.

$$P(X=3) = e^{-4}\, 4^{3}/3! = 0.19$$



flat (see hw).

Lab

poisson with $\lambda=4$ for $x=0,1,\cdots 9$

3

$4 = \lambda$ = average arrival rate

Ch. 2 (time to quantify!)

In the prev. chs we played with histograms of sample data and distributions of (random) variables (cont. and discrete).

Histograms and distributions are the pillars of statistics.
In statistics, we describe the population in terms of distributions, and then ask: "Based on the histogram of my sample (data), could the sample have come from, say, a Normal distribution with parameters $\mu = 13$, $\sigma = 3$?"
If "No," then we know something about the population.

One way to compare the hist. with the distr. is in terms of their summary measures. For example, we compare the "location" of the distribution (e.g. μ) with the "location" of the histogram (e.g. $\bar{x}$)
Or, the "width" of the distr. (e.g. σ) with "width" of the histogram (e.g. s).

The location/width/... of a distr. are (usually) one of its parameters.
" " " " histogram are called its statistic.

In short, one compares parameters with statistics. Later Ch. 7, ...

Examples of statistics for location are: The x for the ith case

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

&

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

data

Examples of statistics for spread are:

pros/cons?

- Sample Range

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

- (sample) variance $= s^2 = \frac{1}{\underbrace{n-1}_{(c)}} \overbrace{\sum_{i=1}^{n} (\underbrace{x_i - \bar{x}}_{\text{deviation.}})^{2}}^{S_{xx}}$

(a)

$\sim$ Average of (deviations)^2

$s \sim$ "typical" spread / 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

---

only FYI

For the mathematically-inclined: If you think of x_i as the components of an n-vector, then after you "center" each components (ie. subtract $\bar{x}$), s^2 is proportional to the magnitude of that vector.

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 " typical deviation of x.

In some problems where The $\frac{1}{n-1}$ is not important, one focuses on $S_{xx} \equiv \sum_i^n (x_i - \bar{x})^2$, 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 population

---

Q1: Suppose n=2. I.e. There are only 2 observations x_1, x_2. Then s^2 is equal to

A) $\frac{1}{2-1}(x_1^2 + x_2^2)$ B) $\frac{1}{2-1}(x_1 + x_2)^2$ $\boxed{\text{C) } \frac{1}{2}(x_2 - x_1)^2}$ D) none of The above.

$$s^2 = \frac{1}{2-1}\sum_i^2 (x_i - \bar{x})^2 = \left[x_1 - \tfrac{1}{2}(x_1+x_2)\right]^2 + \left[x_2 - \tfrac{1}{2}(x_1+x_2)\right]^2$$

$$= \left[\tfrac{1}{2}(x_1 - x_2)\right]^2 + \left[\tfrac{1}{2}(x_2 - x_1)\right]^2 = \tfrac{1}{2}(x_2 - x_1)^2 \propto (\text{difference})^2$$

Note That s is a generalization of The concept of difference.

hw-lect 7-1

Consider The examples of Poisson in lecture.

a) Find another example (google, books, ...) That qualifies as a Poisson variable. Call it x, and define it clearly.

b) Assume, or even guess, what The value of The λ parameter may be for your example. Remember, λ is The average x. State That value, with The correct units.

c) plot The poisson dist. with That value of λ (by R or by hand)

d) Compute $p(x=0)$, and interpret it.

hw-lect 7-2

The poisson mass function in The website example is "flat" at The top, ie. $p(x)$ has the same value at $x=3$ and $x=4$. Show That, quite generally, The poisson mass function has the same value at $x=\lambda$ (ie. at The average) and at $x=(\lambda-1)$.

Not assigned in summer 17; but do it if you like R.

In R, write code to

a) take a sample of size 100 from a normal distribution with parameters mu = -1, sigma = 2, and compute the sample mean and sample standard deviation for that sample.

b) make the density scale histogram for the data in part a, and overlay the density function itself on the histogram.

Note that sample mean and sample std dev correctly correspond to the location and the width of the histogram and distribution. If you don't see this agreement, take another sample and repeat - it may be that the first sample you took was "weird."

hw-lect 7-3

Consider the sequence of observations $x_1, x_2, ..., x_n$, for which we can easily compute the sample mean, denoted $\bar{x}_n$. The subscript denotes the sample size used for computing the mean. Now, if a new observation is made, say x_{n+1}, we don't have to recompute the new sample mean $\bar{x}_{n+1}$ from **all** of the x_i measurements, because it turns out $\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{x_{n+1}}{n+1}$, which you may have already shown in a different hw. In other words, the new sample mean can be computed from the old sample mean and the new observation, using this formula. ~~Here, prove~~ the analogous formula for sample variance: is

$$s_{n+1}^2 = \frac{n-1}{n}s_n^2 + \frac{(x_{n+1} - \bar{x}_n)^2}{n+1},$$

where s_{n+1}^2, is the sample variance of $(n+1)$ observations, and s_n^2 is the sample variance for the first n observations. Here, starting from the defining formula for var. show

$$s_{n+1}^2 = \frac{1}{n}\sum_i^{n+1}\left[(x_i - \bar{x}_n) + \left(\frac{1}{n+1}\bar{x}_n - \frac{x_{n+1}}{n+1}\right)\right]^2$$

BTW, the expression you are asked to prove here is an intermediate step for proving $s_{n+1}^2 = \frac{n-1}{n}s_n^2 + \cdots$, but I'm not asking you to prove that one.