Lecture 9 (Ch. 2)

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. sample mean).

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

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

Examples of statistics for location are: The x for the i^{th} 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

Examples of statistics for spread are:

- Sample Range pros/cons?
- (Sample) variance $= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (standard deviation (same units as $\bar{x}$)) S_{xx}; $(x_i - \bar{x})$ deviation.

$\sim$ Average of (deviations)^2

$s \sim$ "typical" spread / deviation.

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

Example:

$x = c(1, 3, 8)$ 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

$s = \sqrt{13}$ cm

In summary 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: $s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$ ← Because $\sum_i (x_i - \bar{x}) = 0$

"funny Average" — deviation

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 $S_{xx} \equiv \sum_i^n (x_i - \bar{x})^2$.

Finally, note that all of these measures have the word "sample," reminding you that they pertain to sample/data, not population.

1 for-loop, instead of 2 for-loops.

Useful & fast formula for computing s^2:

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

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

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

Step through this.

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

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

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

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

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

Example

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

$$s^2 = \frac{3}{2}\left[\frac{74}{3} - 16\right] = \frac{3}{2}\,\frac{74-48}{3} = \frac{26}{2} = 13$$

Q1: s^2, as defined by $s^2 = \frac{1}{n-1}\sum_i^n (x_i-\bar{x})^2$, can be written as

A) $\frac{n}{n-1}\overline{(x-\bar{x})^2}$ B) $\frac{n}{n-1}\left(\overline{x-\bar{x}}\right)^2$ C) $\frac{n}{n-1}\left(x-\bar{x}\right)^2$

$$s^2 = \frac{1}{n-1}\,n\,\underbrace{\frac{1}{n}\sum(x_i-\bar{x})^2}_{=\overline{(x-\bar{x})^2}}$$

hw-lect 9-1

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 9-2

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

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

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

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