Stat 400 : Discussion section BD3 and BD4 Handout 7

Subhadeep Paul

April 2, 2013

1 Homework hints and Quiz review

1.1 Revision: Sum and difference of independent normal random variables

1.1.1 Linear combinations

- If $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim N(\mu_2, \sigma_2^2)$ are two *independent* R.Vs then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, and $X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$
- If $Z = aX + bY$, where X and Y are as defined above, i.e Z is linear combination of X and Y , then $Z \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$
- The above property extends to n independent normal random variables, i.e if $X_1, X_2, ...X_n$ are i.i.d normal random variables then $S = X_1 + X_2 + ... + X_n$ is also Normal random variable with $\mu = E(X) = \sum \mu_i$ and $\sigma^2 = Var(X) = \sum \sigma_i^2$

1.1.2 Difference between multiple copies of same object and a random sample

- ***The Butcher's shop problem in homework:*** In that problem the 2 packets of ground beef are not identical copies. They are two randomly chosen objects from a certain population with different weights. So we will denote their weights as X_1 and X_2 and both them will follow the same distribution independently (i.i.d). Another example would be to **pick two students randomly** from the class to measure heights. Their heights may follow same distribution but their observed values will be different. So you can't write the quantity of ground beef in the 2 packets as $2X$. Hence the variance will be $2\sigma^2$ NOT $4\sigma^2$
- ***Price of Coffee in Dunkin Donuts:*** Let's assume price X of a medium coffee in DD varies daily following a normal distribution with certain mean and variance. But on a certain day the price is same for ***all*** medium coffee mugs. So if you order 2 medium coffee, you will treat the price as $2X$. This will follow normal distribution with mean 2μ and variance $4\sigma^2$.

1.2 Home work hints

- *PVC problem*: Total$=X_1 + X_2 - X_3$, where X_i follow normal distribution. X_3 is the length of the overlap region

- *Who does the dishes*: P(Dick rolls 1 in k th attempt)~ geometric. P(Jane rolls 1 in m th attempt)~ geometric. Joint PMF is product of the two probabilities. Then the 3 parts correspond to 3 different relationship between k and m, $k = m$, $k = 2m$ and $m > k$ respectively. The last one means $m = k + 1, k + 2, ...\infty$

- Smoker/Non smoker problem : 3 categories, denote by x, y and z with $z = 7 - x - y$. Joint probability of (x, y, z) follows multinomial distribution.

2 From last work sheet : Basics of Statistics

2.1 Population and Sample

- Statistics is the science of infering from data. What method I need to use depends on my data and the questions I am trying to answer. e.g suppose, my question is what is the mean of heights of students in this class. So the class (of 50 students) is my population and the height is my random variable which follows a certain distribution (say normal). I am interested in finding mean $\mu$of this random variable. Depening upon my time and resources I can take two approaches. 1) Record every one's height and take the average 2) select 10 students randomly, record their heights and take the average.

2.2 "Statistic" and "estimator"

- A random sample is always i.i.d (independently and identically distributed). We compute different functions of the data/sample, which are called "statistic" to "estimate" population characteristics. So Statistic is any function of the data/sample. Examples : the entire data/sample x, the sum of observations $\sum x_i$, sum of squares/absolute values of the observations $\sum x_i^2$, $\sum |x_i|$, sample mean $\bar{x}$, Sample median, mode, quantiles, percentiles, standard deviation, minimum and maximum of the sample, range of sample etc.

- Question. A population parameter, say mean μ for a population that follows normal distribution $N(\mu, 1)$. Answer: all we have is the data. so it must be a function of the data, a "statistic". Now you can use any statistic to estimate μ. The corresponding statistic will be called "estimator" of the parameter μ. You may even use a constant function, i.e say I don't care about the data and my estimator is 6, (I just like that number !) no matter what the data says.

- So there can be thousands of estimator for a parameter. Some estimators are always good, some are occassionally good, some are never good.

2.3 Two results on sample mean $\bar{x}$

- $E(\bar{X}) = \mu$, $Var(\bar{X}) = var(\frac{1}{n}\sum x_i) = \frac{n\sigma^2}{n^2} = \sigma^2/n$. (these is true irrespective of the distribution of X).

- $\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}$ follows a standard normal distrbution under two conditions : (a) n is large X has any distibution, (b) n is any value, X has normal distribution.