Stat 202 - 2015S - W9 - Wednesday (Pg1)

Review

Probability distribution
described by table (discrete RVs)
or density curve continous

Mean and variances of random variable determined by prob. distribution
(discrete) sum involving elements of table
(continuous) integral involving density curve

Discrete

$\mu_X = \sum_i x_i p_i$

$\sigma_X^2 = \sum_i (x_i - \mu_X)^2 p_i$

Can be easier to calculate means and variances of RV (eg Binomial RV) by expressing it as a sum of simpler RVs and applying following rules

$\mu_{X+Y} = \mu_X + \mu_Y \qquad \mu_{X-Y} = \mu_X - \mu_Y$

$\mu_{a+bX} = a + b\mu_X \qquad \cancel{\mu_{b_1X_1 + b_2X_2} = b_1\mu_{X_1} + b_2\mu_{X_2}}$

~~Answers~~

$\sigma^2_{a+bx} = b^2\sigma^2_x$ (all RVs)

$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y$ (independent

$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y$ (independent)

$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y + 2\rho\sigma_x\sigma_y$ (all RVs)

$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y - 2\rho\sigma_x\sigma_y$ (all RVs)

ρ was the correlation between X any analogous to correlation between variables in data, New Sampling Distributions

Statistics such as means and proportions summarize data

- mean height of 30 people drawn from a population

- the proportion of 30 people drawn from a population who prefer red over blue.

A statistic from a random sample or randomized experiment is a random variable

The probability distribution of such a statistic is its sampling distribution

The population distribution of a variable is the distribution of its values over all members of the population. Also the prob distrib of the variable when we choose one individual at random from the population

§5.1 The sampling distribution of the sample mean

Random phenomenon: Draw a sample of size n from a population

Random Variable: mean of n observations (e.g. heights) from sample, called Sample mean

Facts about sample means

1. Sample means are less variable than individual observations

2. Sample means are more normal than individual observations

In other words the sampling distribution of a sample mean has less spread than the population distribution and is closer to Normal — QQ Plot looks more like a line (unless their both a line)

The mean and standard deviation of $\bar{X}$

The sample mean is approximately the mean of the underlying distribution

But everytime you draw a sample you get a different value (maybe only slightly different for $\bar{X}$)

In other words, there is spread in the sampling distribution of $\bar{X}$

$\bar{X}$ is an estimate of the population mean μ.

"estimate" is a technical term.

What is the mean of $\bar{X}$ and what is the variance / standard deviation?

Let X_i be the observation of X from the i^{th} element of sample

~~Example~~ ~~Random~~ Then $\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$

From Rules $\mu_{\bar{X}} = \frac{1}{n}(\mu_{X_1} + \mu_{X_2} + \ldots + \mu_{X_n})$

But μ_{X_i} (mean of one observation from one individual in sample)

Must be the same as ~~...~~ μ mean of population — they ~~...~~ two names for the same thing.

Thus $\mu_{\bar{X}} = \frac{1}{n}(\mu_{X_1} + \cdots + \mu_{X_n})$

$= \frac{1}{n}(\mu + \mu + \mu + \cdots + \mu)$

$= \frac{1}{n} \cdot n\mu$

$= \mu$

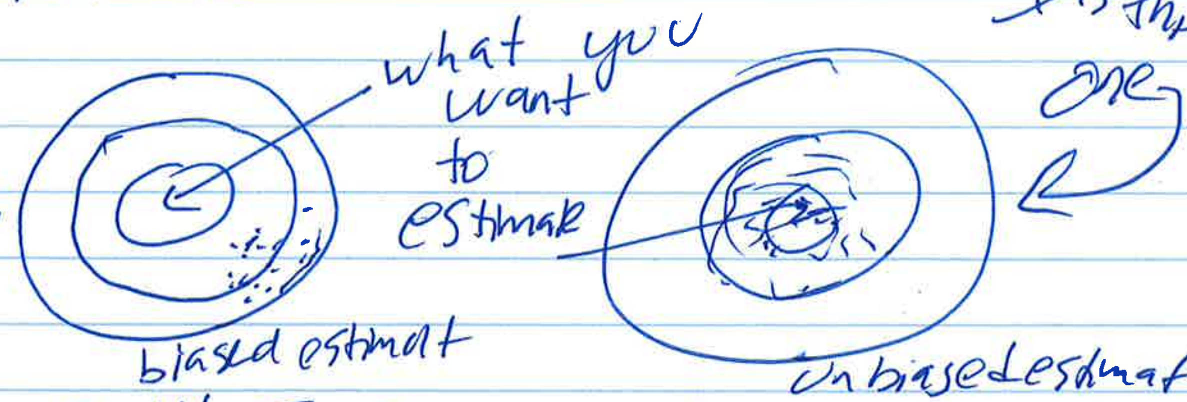
The mean of the sampling distribution $\mu_{\bar{X}}$ is the same as the mean of the population.

Because $\mu_{\bar{X}} = \mu$ and $\bar{X}$ estimates μ

We say $\bar{X}$ is an <u>unbiased estimate</u> of μ (the mean of the estimate happens to be what you are trying to estimate)

Picture

Two alternatives:



Bull's Eye is what you want to estimate. Cloud of points are different sample means each calculated from n observations (independent) of X.

When you say an estimate is unbiased (that's good) but it doesn't tell you how close to the bull's eye the points lie — it tells you that the center (mean) of the sampling distribution is spot on but it doesn't tell you about the spread in distribution — whether all darts fell close to bulls eye or whether they were all over the board,

The variance tells you this, Because the observations are independent we can use the rules for variances

$$\sigma_{\bar{X}}^2 = \left(\frac{1}{n}\right)^2 \left(\sigma_{X_1}^2 + \sigma_{X_2}^2 + \dots + \sigma_{X_n}^2\right)$$

$$= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \dots + \sigma^2\right)$$

population variance

$$= \frac{\sigma^2}{n}$$

Standard deviation $\frac{\sigma}{\sqrt{n}}$

In summary

Let $\bar{X}$ be the mean of an SRS of size n from a population having mean μ and standard deviation σ.

The mean and standard deviation of $\bar{X}$ are

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$\sqrt{n} > 1$ ~~[illegible]~~

So the standard deviation of the sampling distribution $\sigma_{\bar{X}}$ is <u>smaller</u> than the standard deviation of the population

Remember we said sample means are less variable than individual observation

This shows why

If the population distribution is normal $N(\mu, \sigma)$ then the sampling distribution of $\bar{X}$ is also normal $N(\mu, \sigma/\sqrt{n})$

Even if the population distribution is not normal, the sampling distribution of $\bar{X}$ is approximately normal $N(\mu, \sigma/\sqrt{n})$ for large n.

(technically this is only true is σ is finite. I'll bet you didn't know σ could be infinite, The book mentions this.)

This is called The Central Limit Theorem. It is one of the two biggest results in statistics along with the ~~central~~ law of large numbers)

Note that even if RV X is discrete (1 if heads, 0 if tails)

$\bar{X}$ is approximately normal

↳ ~~mean~~ ~~number~~ proportion of heads in n tosses