Stat 202-2015XD-W4-Thursday

Review

The mean and standard deviation of a discrete random variable are calculated from their probability table

Values of X	x_1	x_2	$\cdots$	x_K
Probability	p_1	p_2	$\cdots$	p_K

μ_X = mean $= \sum x_i p_i = x_1 p_1 + \cdots + x_K p_K$

σ_X^2 = Variance $= \sum (x_i - \mu_X)^2 p_i$

$= (x_1 - \mu_X)^2 p_1 + \cdots + (x_K - \mu_X)^2 p_K$

σ_X = std dev $= \sqrt{\sigma_X^2}$

Formulas

All:

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

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

Independent:

$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$

$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$ ← not a typo

All

ρ correlation

$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho\sigma_X\sigma_Y$

$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$

Linear transformation (all)

$\mu_{a+bx} = a + b\mu_x$

$\sigma^2_{a+bx} = b^2\sigma^2_x$

Mistake in wednesday's notes

Values of x	0	1	2	3
Probability	1/8	3/8	3/8	1/8

$\sigma^2_x = 3/4$ not $9/16$

$\sigma_x = \sqrt{3/4}$

$\mu_x = 0 \cdot 1/8 + 1 \cdot 3/8 + 2 \cdot 3/8 + 3 \cdot 1/8 = 3/2$

$\sigma^2_x = (0 - 3/2)^2 \cdot 1/8 + (1 - 3/2)^2 \cdot 3/8$

$+ (2 - 3/2)^2 \cdot 3/8 \cdot (3 - 3/2)^2 \cdot 1/8$

$= 3/4$

$\sigma^2_x = \sqrt{3/4}$

Why are these formulas useful?

Well in Chapter 5 (and else) we are going to need the mean and variance of a Binomial Random Variable

remember this was like tossing an unfair coin many times where the probability of "success" (heads) was p and probability of "failure" (tails) was 1-p.

(Remember Binomial calculator from exam.)

The Binomial RV is the count of # of successes.

Well I could just give you the formulas for mean and variance

But they would be very bizzare

It wouldn't make sense why they were true

With the formulas shown today I can actually explain why these formulas are true.

They will be similarly helpful in other situations.

Lets derive the mean and variance of the random variable "number of heads in three tosses of a coin" that we have seen earlier

We have already derived them

$\mu_X = 3/2 \quad \sigma_X^2 = 3/4$ But it was hard and our method will get harder with large values of n

Now we will use a method that works easily for large n,

Consider that X is the sum of three independent RVs each one the number of heads in the toss of one coin,

$$X = Y_1 + Y_2 + Y_3$$

Values of Y_i	0	1
probability	1/2	1/2

Mean $\mu_{Y_i} = 0 \cdot 1/2 + 1 \cdot 1/2 = 1/2$

$$\sigma_{Y_i}^2 = (0-1/2)^2 \cdot 1/2 + (1-1/2)^2 \cdot 1/2$$
$$= \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{4}$$

Pg 5

Now we apply our rules

$$\mu_X = \mu_{Y_1} + \mu_{Y_2} + \mu_{Y_3} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}$$

check ✓

Because Y_i independent

$$\sigma_X^2 = \sigma_{Y_1}^2 + \sigma_{Y_2}^2 + \sigma_{Y_3}^2 = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

check ✓

Now in general

Values of Y_i	0	1
Probability	$1-p$	p

$$\mu_{Y_i} = 0 \cdot (1-p) + 1 \cdot p = p$$

$$\sigma_{Y_i}^2 = (0-p)^2(1-p) + (1-p)^2 \cdot p$$

$$= p^2(1-p) + (1-p)^2 \cdot p$$

$$= p(1-p)[p + 1 - p]$$

$$= p(1-p)$$

$$\mu_X = \mu_{Y_1} + \cdots + \mu_{Y_n} = np$$

Because Independent

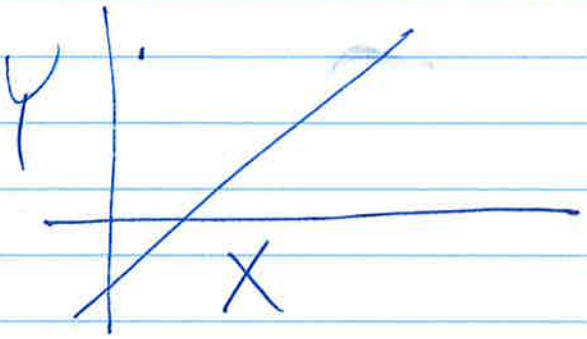
$$\sigma_X^2 = \sigma_{Y_1}^2 + \cdots + \sigma_{Y_n}^2 = np(1-p)$$

$$\sigma_X = \sqrt{np(1-p)}$$

Notes about Correlation between x and y

If $\rho = 1$ x and y are related linearly with positive slope

$$y = mx + b \qquad m > 0$$



Data look like this

$r = 1$

$3y = 4 + 5x$

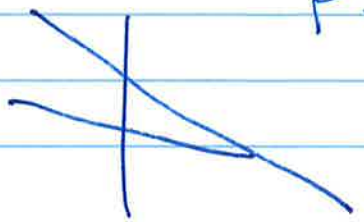
$x - y = 2$

] both have correlation $\rho = 1$

Likewise $y = mx + b \qquad m < 0$

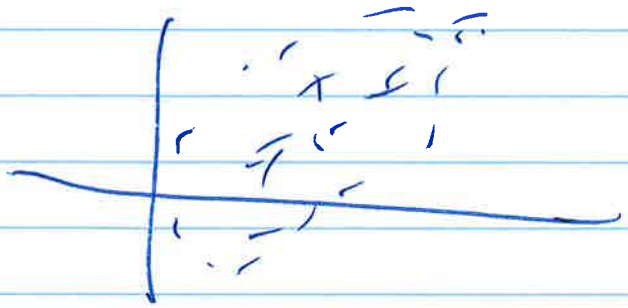
$\rho = -1$

eg. $y + x = 3$



If $0 < |\rho| < 1$ there is scatter between x and y

When values are simulated get a scatterplot with



$r = \rho$ approximately

If $\rho = 0$ there is no linear relationship between x and y

If x and y are independent

$\rho = 0$

If $\rho = 0$ x and y could be but are not necessarily independent.

Independent mean knowing x tells you nothing about y and vice-versa

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 <u>random variable</u>

The probability distribution of such a statistic is its <u>sampling distribution</u>

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 — Q-Q Plot looks more like a line (unless they both a line)

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

The sample mean is approximately the mean of the underlying distribution

But every time 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

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 are two names for the same thing.

Thus $\mu_{\bar{x}} = \frac{1}{n}(\mu_{x_1} + \cdots + \mu_{x_n})$

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

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

$= \mu$

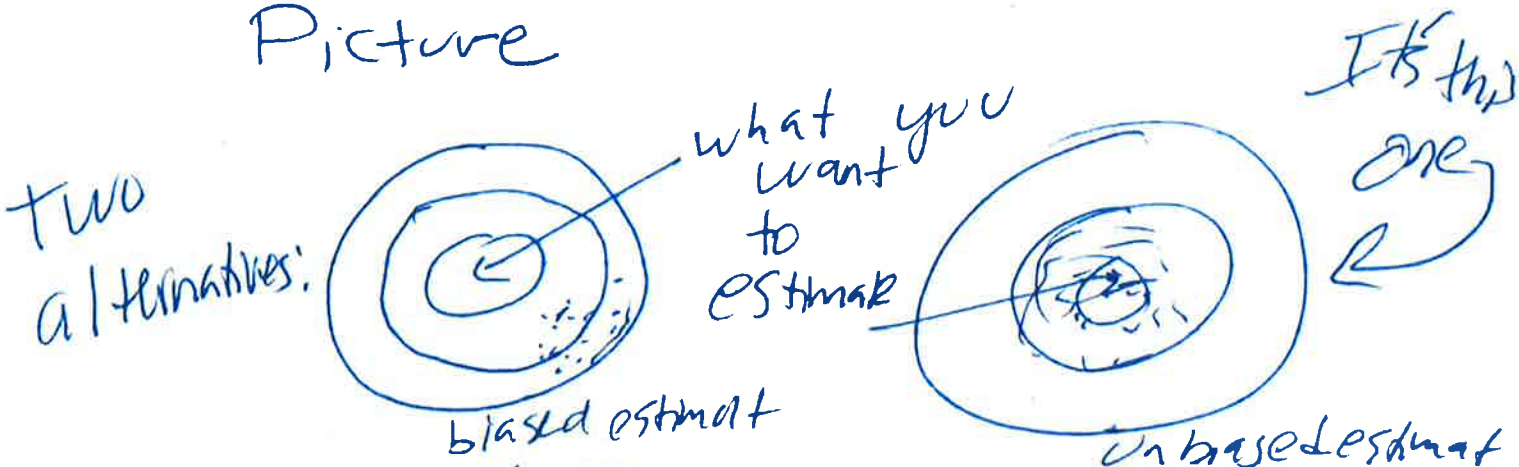
The mean of the sampling distribution $\mu_{\bar{x}}$ is the same as the mean of the population.

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



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 bull's 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

$$\sigma_{\bar{x}}^2 = \left(\frac{1}{n}\right)^2 \left(\sigma_{x_1}^2 + \sigma_{x_2}^2 + \ldots + \sigma_{x_n}^2\right)$$

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

σ^2: 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

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

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

$\uparrow$

$\sqrt{n} > 1$

of the sampling distribution

So the standard deviation $\sigma_{\bar{x}}$ is smaller 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 ~~contradiction~~ law of large numbers)

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

$\bar{X}$ is approximately normal

(proportion ~~number~~ of heads in n tosses