Summary of Week 2

Monday

- Expectation / Variance
  - ↳ allow us to summarize important aspects of distribution
- Binomial distribution
  - ↳ example of iid sum model of election / survey

Wednesday

- Histograms in python
- Properties of Binomial and large N limit

Friday

- LLN, CLT
  - ↳ allows us to easily describe dist of iid sum
- Normal dist and density
  - ↳ density is mathematical idealization of histogram

This week

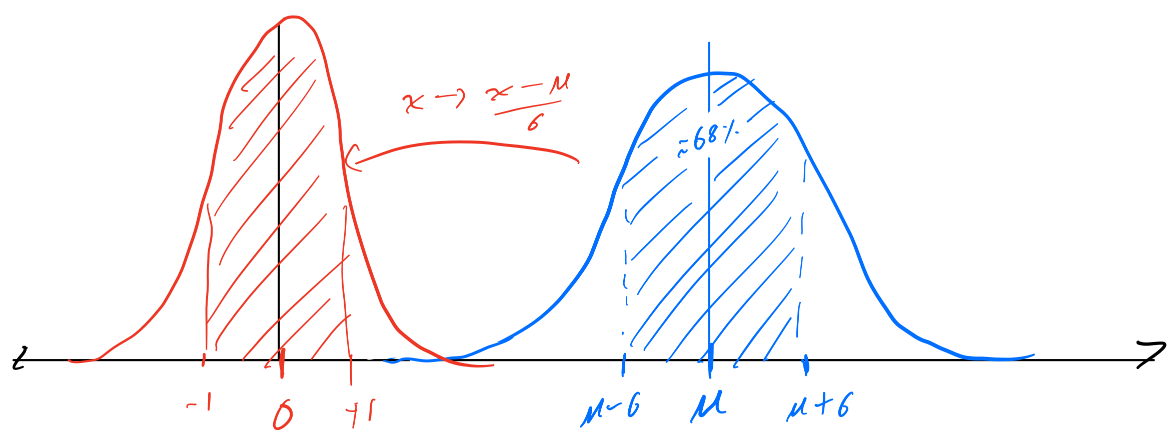
- linear regression Model basics
  - ↳ covariance, least squares
- working w/ tabular data in python

Properties of Normal random variables (ch 4.6)

$\underline{X} \sim \text{Normal}(\mu, \sigma^2)$ or $\underline{X} \sim N(\mu, \sigma^2)$

is a normal r.v. and has density

$$f_{\underline{X}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} = \frac{1}{\sigma}\phi((x-\mu)/\sigma)$$



Combining Normal r.v.s (theorem 4.6.1)

Let $Z_1 \sim N(0,1)$, μ, σ constants. Then let

$$X = \sigma Z_1 + \mu \implies \mathbb{E}[X] = \sigma \mathbb{E}[Z_1] + \mu$$
$$= \mu$$
$$Var(X) = \sigma^2 Var(Z_1) = \sigma^2$$

1) X is normal $\implies X \sim N(\mu, \sigma^2)$

Let $X_i \sim Normal(\mu_i, \sigma_i^2)$ be independent

2) Then $$Y = \sum_{i=1}^{N} X_i \sim N\left(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2\right)$$

- might talk about theorem 4.6.2 later
- will talk about chi-squared in week 4

Note on CLT: CLT says $Z_1 = \frac{S - \mu N}{\sqrt{N}\sigma} \to N(0,1)$

meaning $P(a < Z < b) \to \int_a^b \phi(x)dx$

not technically true

$S = \sqrt{N}\sigma Z_1 + \mu N$, $\quad S \cancel{\to} N(\mu N, \sqrt{N}\sigma)$

still, often think of CLT as saying S is approximately Normal

Linear Regression Model (Ch 10.)

related variables (Def 10.1.1) = <u>not</u> independent

linear regression w/ one predictor (Ex 10.1.1)

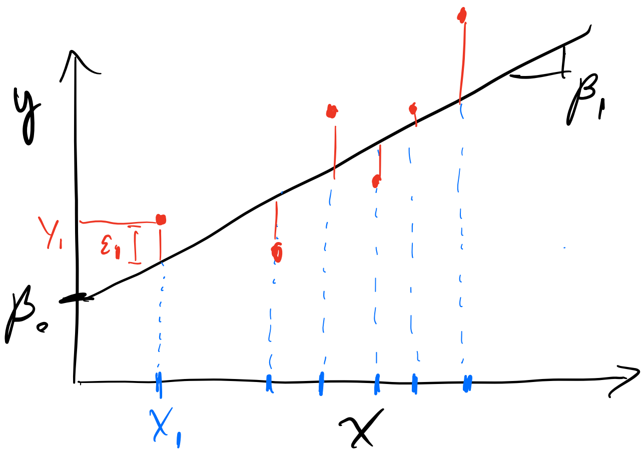
Let X be any r.v. and

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Can also write

"error" or "residual" term

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$



Very important model because simple linear relationship + Normal error $\Rightarrow$ we can easily estimate $\beta_0, \beta_1, \sigma^2$

X = <u>predictor</u> variable (e.g. weight of dog (quantitative) or breed (qualitative))

Y = <u>response</u> variable (e.g. lifespan)

Meaning of parameters

Param	formula	description	units
β_0	$\mathbb{E}[Y \mid X=0]$	avg. of Y when $X=0$	units of Y
β_1	$\mathbb{E}[Y \mid X=x+1] - \mathbb{E}[Y \mid X=x]$ (x can be anything)	avg. change in Y when we change X by unit	$\frac{\text{units of } Y}{\text{units of } Y}$
σ^2	$\operatorname{var}(Y \mid X=x)$ (x can be anything)	var of Y for fixed x	units of Y^2

Example (model of hieght)

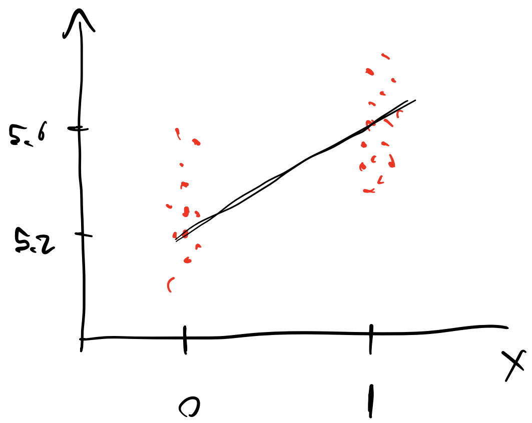
$$X \sim \text{Bernoulli}(1/2) \quad (1 = \text{male}, 0 = \text{female})$$

$$Y \mid X \sim N(5.2 + 0.55X,\ 0.25^2) \quad (\text{hight in inches})$$

Note: $\mathbb{E}[Y \mid X=0] = 5.2 = \beta_0$

$\mathbb{E}[Y \mid X=1] = 5.75 = \beta_1$

$\operatorname{var}(Y \mid X=1) = \operatorname{var}(Y \mid X=0) = 0.25^2$

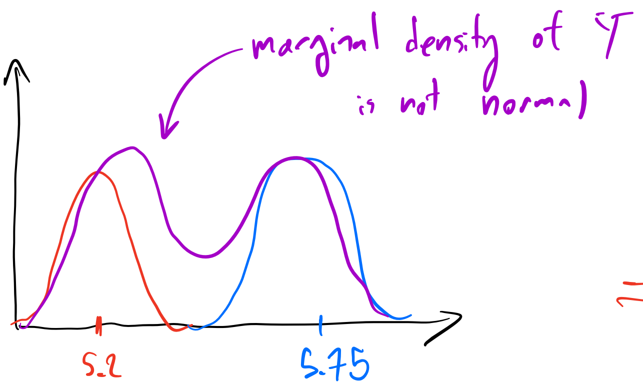


$\Rightarrow$ in order to estimate β_0, β_1, σ we merely need to compute mean and variance within each group

Q: what is prob. male > 6.25 ft
7ft?

$6.25 = 5.75 + 0.25 \times 2$

= mean of male hieght + 2 standard devs



$\Rightarrow P(Y > 6.1 \mid X=1) =$  $\approx 2\%$

Let X have mean and variance μ_X and σ_X

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma)$$

Marginal mean and variance

$$\mathbb{E}[Y] = \mathbb{E}[\beta_0 + X\beta_1 + \varepsilon] = \beta_0 + \beta_1 \mu_X$$

$$\mathrm{Var}(Y) = \underbrace{\beta_1^2 \sigma_X^2}_{} + \sigma^2$$

more spread of x points $\rightarrow$ more y variation

larger slope $\rightarrow$ more y variation

Least Squares

Question: How do we estimate slope?

In principle we could use sample avg.:

$$\beta_1 = \mathbb{E}[Y \mid X = x+1] - \mathbb{E}[Y \mid X = x]$$

$$\approx \overline{Y \mid X = x+1} - \overline{Y \mid X = x} \quad \text{for some } x$$

Won't work so well if we only have one value for each x

Covariance (Def 3.3.3)

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

$$= \mathbb{E}[XY] - \mu_X \mu_Y \qquad (\text{unit of } X \cdot Y)$$

↑ theorem 3.3.3

Lets calculate cov(X,Y) for linear regression model

Need to compute $\mathbb{E}[XY]$. Can either tower property (HW 3B) or

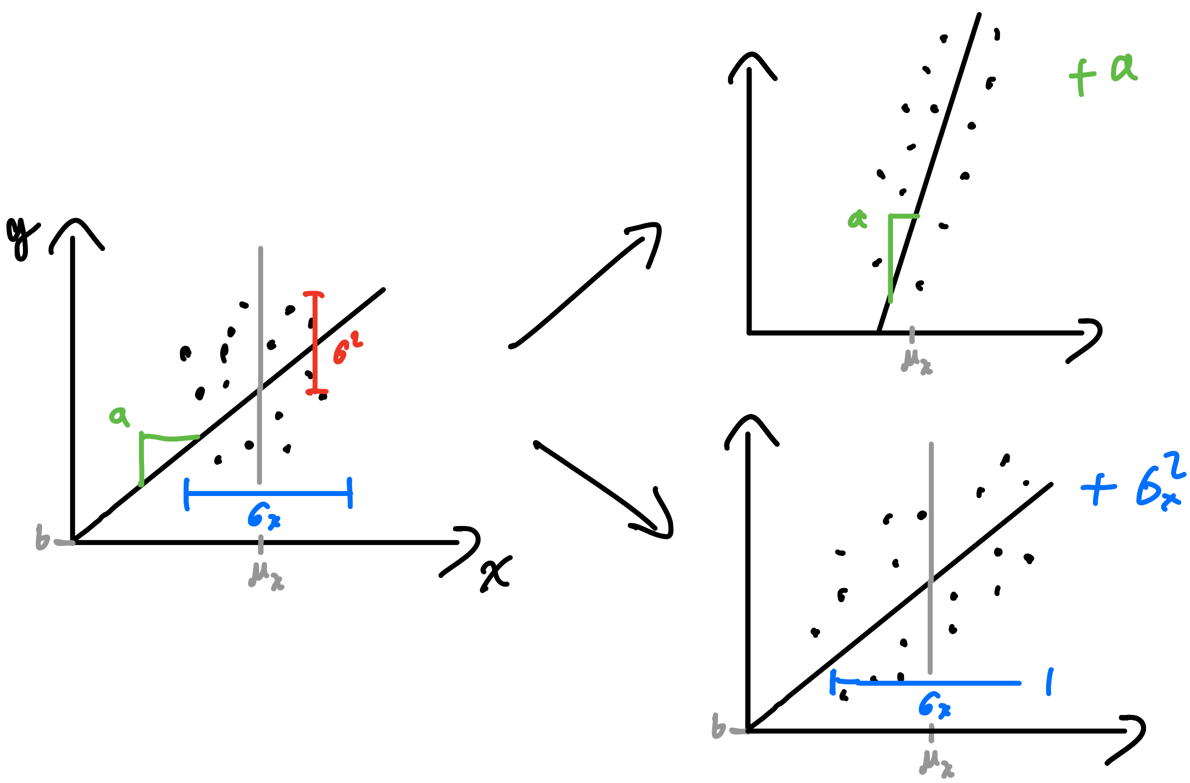
$$\mathbb{E}[XY] = \mathbb{E}[X(\beta_0 + \beta_1 X + \varepsilon)]$$
$$= \mathbb{E}[\beta_0 X] + \beta_1 \mathbb{E}[X^2] + \mathbb{E}[X\varepsilon]$$
$$= \beta_0 \mu_X + \beta_1 \left(Var(X) + \mu_X^2\right) + \mathbb{E}[X]\mathbb{E}[\varepsilon]$$

$$= \beta_0 \mu_X + \beta_1 \sigma_X^2 + \beta_1 \mu_X^2$$

$$\mathbb{E}[X]\mathbb{E}[Y] = \mu_X \mathbb{E}[\beta_0 + \beta_1 X + \varepsilon]$$
$$= \mu_X \beta_0 + \mu_X^2 \beta_1$$

$$\Rightarrow \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \sigma_X^2 \beta_1 \quad \star$$

check units



Formula for covariance gives us a way to express slope from samples:

Samples $(X_1, Y_1), \ldots, (X_N, Y_N)$

Not $\sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])]$

$$\approx \overline{(X - \overline{X})^2} \approx \frac{1}{N} \sum_{i=1}^{N} (X_i - \overline{X})^2$$

technically should by more on th later

$$\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \approx \frac{1}{N} \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})$$

$$\Rightarrow \boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_1 (X_i - \overline{X})^2}}$$

important!

Next, we can estimate β_0:

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mu_X$$

$$\approx \boxed{\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}}$$

These are formula in theorem 10.3.1 derived differently

We can also estimate σ by

$$\hat{\sigma} = \overline{\left(Y - \left(\hat{\beta}_0 + X\hat{\beta}_1\right)\right)^2}$$