Lecture 2:

kiss example:

$n = 124$

$80 : R$

$$\frac{80}{124} = 0.695$$

$X_1, \ldots, X_n$

$$X_i = \begin{cases} 1 & i\text{th couple turn right} \\ 0 & \text{o.w.} \end{cases}$$

$$X_i \overset{iid}{\sim} \text{Ber}(p)$$

$$\frac{X_1 + \ldots + X_n}{n} = \bar{X}_n$$

Let's assume $p = 1/2$.

$$\mathbb{P}(\bar{x}_1 \geq 0.695) = \mathbb{P}(\underbrace{\textstyle\sum X_i}_{\sim \text{Bin}(n,p)} \geq 0.695 \cdot n)$$

Use CLT (Central Limit Thm) → need to know μ & σ.

$$X_i = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{w/ prob } 1-p \end{cases}$$

$$\mathbb{E}[\bar{X}_n] = p \qquad \mathbb{V}[\bar{X}_n] = \frac{\mathbb{V}[X_i]}{n} = \frac{p(1-p)}{n}$$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$$

$\sqrt{\mathbb{V}(x)}$

p score.

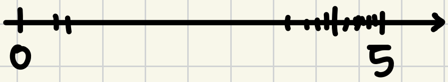
$$\mathbb{P}(\bar{x}_n \geq 0.695) = \mathbb{P}(\bar{x}_n - \mu > 0.695 - \mu)$$

$$= \mathbb{P}\left(\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{0.695 - \mu}{\sigma/\sqrt{n}}\right) = \mathbb{P}\left(Z > \frac{0.695 - \mu}{\sigma/\sqrt{n}}\right)$$

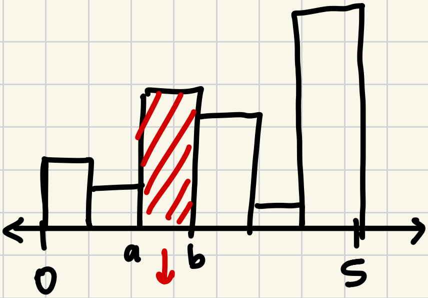
$$= 0.003.$$

where $Z \sim N(0,1)$.

$X_1, \dots, X_{233}$



← dot plot (lots of dots per each decimal (continuous))



← histogram (make intervals & count per interv.)

gpas of 18.650

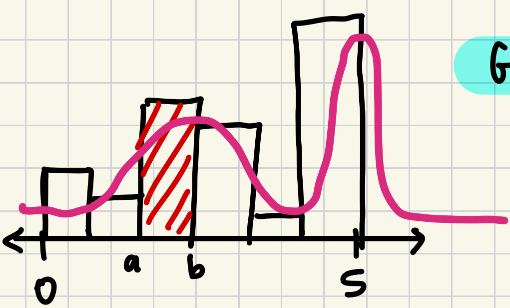
area = proportion of data that is in the bin.

$$\text{area} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(a \le X_i < b) \rightarrow (b-a)\,\text{height} = \text{prop.}$$

↳ add if in range

$$h = \frac{\text{prop.}}{b-a}$$

* histogram readability depends on def. of intervals.

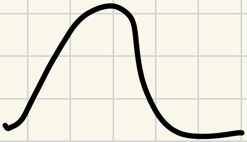


Gaussian = getting curve out of histogram.

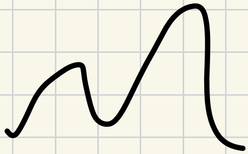
KDE

Analyzing graphs:

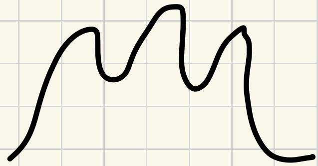
unimodal



bimodal



multi modal



symmetric

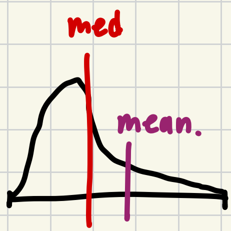
right-skewed

(price of houses, income)

left-skewed

(gpas, hopefully)

* mean (average, μ)
* standard deviation (std, σ)
* median $\rightarrow$ $\frac{1}{n}\sum \mathbb{1}(x_i \leq \text{median}) = \frac{1}{2}$



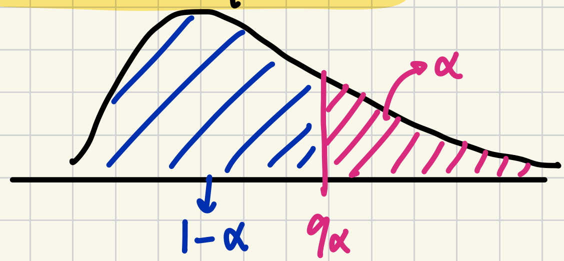
* range
* quantiles : $q_\alpha = (1-\alpha)$ quantile $= 100(1-\alpha)$ percentile

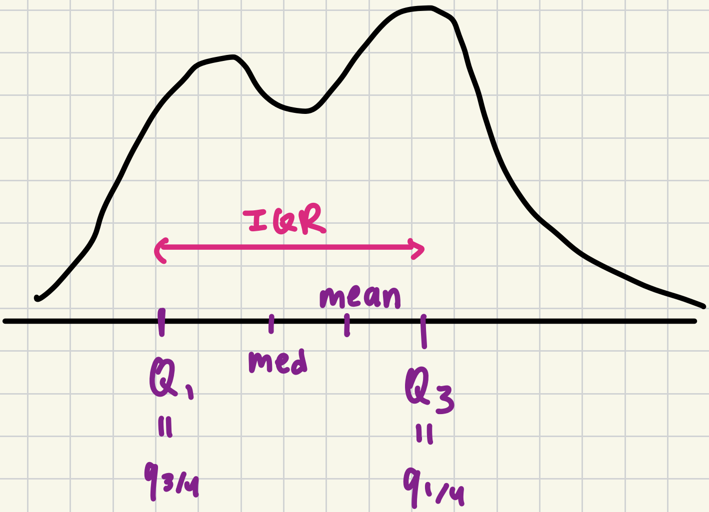
$$\frac{1}{2}\sum \mathbb{1}(x_i \leq q_\alpha) = 1-\alpha$$

1st quartile $= q_{3/4}$

third quartile $= q_{1/4}$

IQR $= Q_3 - Q_1$
(interquartile range)





mean vs med.

- mean is not robust bc 1 bad pt. can ruin whole mean.
- med, Q_1, Q_3, IQR are more robust bc they are not impacted by outliers

x_i is outlier if $x_i > Q_3 + 1.5\,IQR$ or $x_i < Q_1 - 1.5\,IQR$