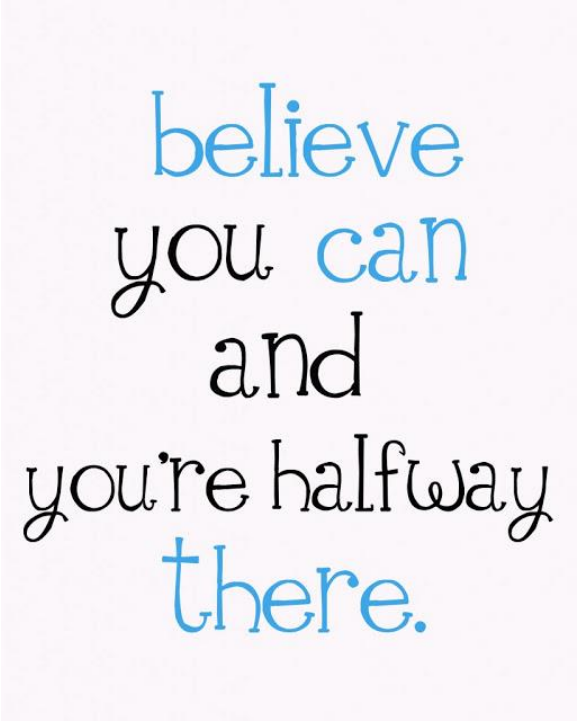
Midterm Review

What to do:

- Read lectures 1-4, chapters 1-10 from the textbook
- Do the assigned exercises from the textbook
- Go over the quiz questions
- Use sample tests to practice
- Use extra TAs' office hours



Topics to review:

Who
What
Why

- **Types of variables**

  - Categorical
    - Bar char
    - Pie chart
    - Two-way table
    - Joint distribution
    - Marginal distribution
    - Conditional distribution
    - Independence

  - Quantitative
    - Dot plot
    - Stemplot
    - Histogram
    - Distribution
      - Shape: symmetric, skewed, unimodal, etc
      - Center: mean, median, mode
      - Spread: range, standard deviation, percentiles, IQR
    - Five-number summary
    - Boxplots
    - Outliers: $1.5 \times IQR$ rule

trimmed mean

Review how linear transformations affect measures of center and spread

Example: Here are the projected numbers (in thousands) of earned degrees in the U.S. in the 2010-2011 academic year, classified by level and by the sex of the degree recipient.

	Bachelor's	Master's	Professional	Doctorate	Total
Female	933	402	51	26	1412
Male	661	260	44	26	991
Total	1594	662	95	52	2403

(a) What proportion of degree recipients are women?

$$\frac{1412}{2403} = 0.5876$$

$$\approx 59\%$$

	Bachelor's	Master's	Professional	Doctorate	Total
Female	933	402	51	26	1412
Male	661	260	44	26	991
Total	1594	662	95	52	2403

(b) What proportion of those who received a professional degree are women?

$$\frac{51}{95} = 0.5368 \approx 54\%$$

(c) Are the events "choose a woman" and "choose a professional degree recipient" independent?

No

$$59\% \neq 54\%$$

$n = 13$

61 $(Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR)$

$(71.5,\ 115.5)$

Example: For given stemplot find a five-number summary, range, standard deviation. Are there any outliers?

6	1
7	
8	7 7 9
9	1 1 2 2 3 8
10	0 1 1

Min Q_1 M Q_3 Max

61 88 92 99 101

(89) (98)

range $= 101 - 61 = 40$

lecture 2 → st. dev. $\approx \frac{\text{range}}{4} = \frac{40}{4} = 10$

$IQR = Q_3 - Q_1 = 99 - 88 = 11$

$Q_1 - 1.5\,IQR = 88 - 1.5 \cdot 11 = 71.5 > 61$

$Q_3 + 1.5\,IQR = 99 + 1.5 \cdot 11 = 115.5 > 101$

61 is the only outlier

- **Density Curve**
  - Normal distribution
    - Shape: symmetric, unimodal, bell-shaped
    - Parameters: mean μ and standard deviation σ
    - 68-95-99.7 Rule
    - Z-scores
    - N(0,1)
    - Z-Table
    - Inverse Normal calculations
    - Normal quantile plots

Example: An English placement examination is given to 900 incoming students. The distribution of examination scores is approximately normal with a mean of 82 and a standard deviation of 5. How many students had test scores between 75 and 85?

$X = \text{scores} \sim N(82, 5)$

$P(75 \le X \le 85) = P(X \le 85) - P(X \le 75)$

$z\text{-score for } 85 = \frac{85-82}{5} = 0.6$

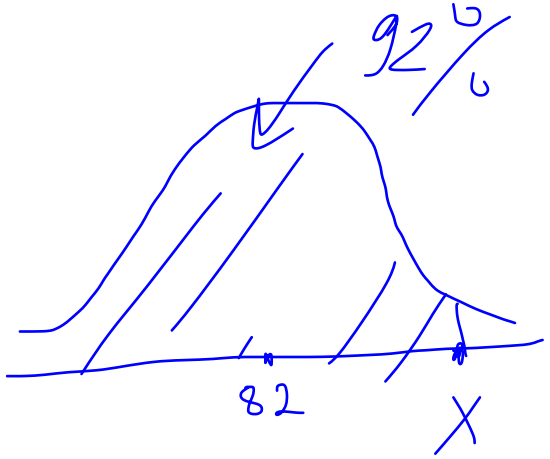
$z\text{-score for } 75 = \frac{75-82}{5} = -1.4$

$= P(Z \le 0.6) - P(Z \le -1.4)$

$= 0.7257 - 0.0808 = 0.6449 \approx 65\%$

$\#\text{ of students} = 900 \cdot 0.6449 \approx 580$

What score is equal to P_{92}, or the 92nd percentile?



$X \sim N(82, 5)$

$P(Z \leq ?) = 0.92$

$Z = 1.405$

$Z = 1.405$

$X = \mu + \sigma Z = 82 + 5 \cdot Z$

$= 82 + 5 \cdot 1.405 \approx 89$

Example: What value is closest to the interquartile range for the standard normal distribution?

(A) 0

(B) 0.5

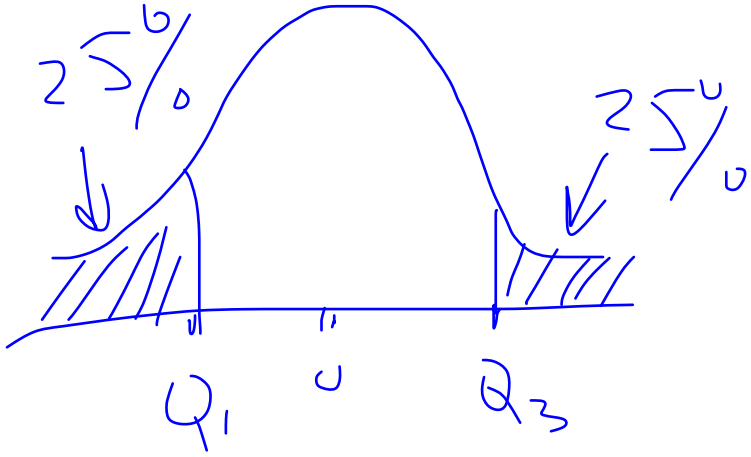
(C) 1.3

(D) 3.0

(E) 2.3

$Z \sim N(0,1)$

$IQR = Q_3 - Q_1$



$Q_1 = -Q_3$

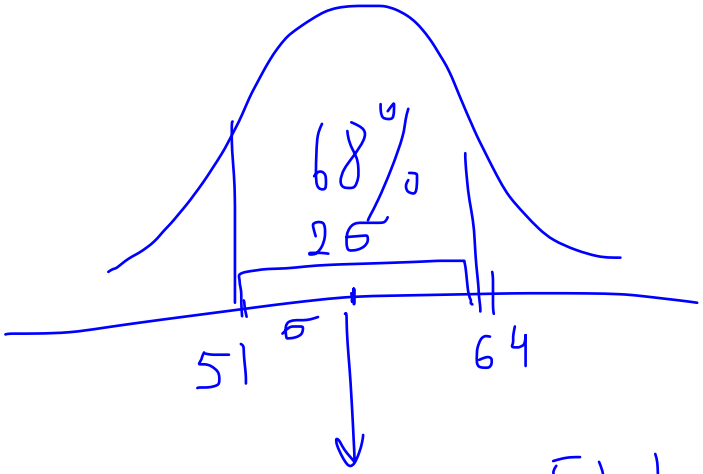
$Q_1 = -0.675$

$Q_3 = 0.675$

$IQR = 0.675 - (-0.675)$

$= 1.35$

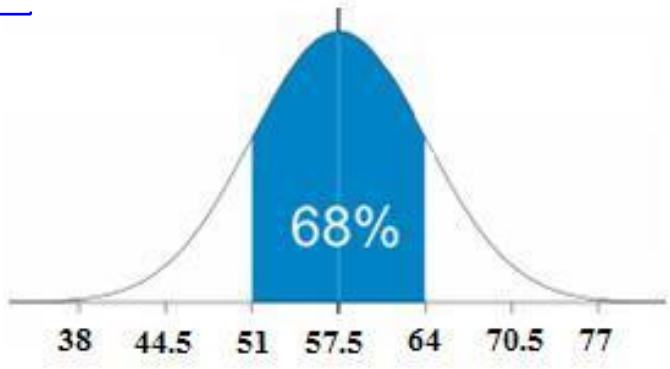
Example: 68% of the marks in a test are between 51 and 64. Assuming these data are normally distributed, what are the mean and standard deviation?



$$\text{mean} = \mu = \frac{51+64}{2} = 57.5$$

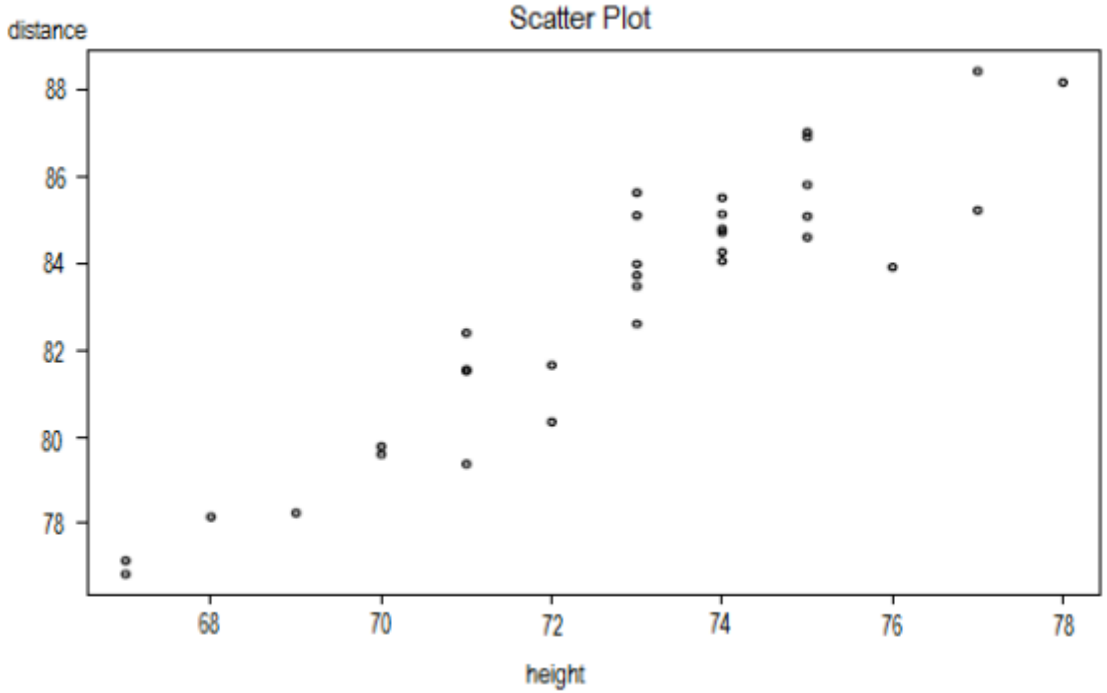
$$2\sigma = 64 - 51 = 13$$

$$\sigma = 13/2 = 6.5$$



- **Linear regression**
  - Scatterplot: overall pattern (form, direction, and strength of the relationship), outliers, clusters
  - Association
    - Variables: explanatory or response?
    - Positively or negatively associated?
  - Correlation
  - Regression line: $y = b_0 + b_1 x$. How to interpret slope b_1, intercept b_0?
  - How to use regression line for prediction?
  - Extrapolation
  - Coefficient of determination
  - Residuals/residual plots
  - Ourliers vs influential observations
  - Transformations

Example: A long jump competition took place recently at a local high school. The coach is interested in performing as well as possible next time, so he is looking at the relationship between height and distance jumped (both measured in inches). He uses height to predict distance. The data are shown:



Which statement is correct about the scatter plot?

(A) There is negative linear relationship between the two variables.
(B) There is no apparent relationship between height and distance.
(C) There is positive linear relationship between the two variables.
(D) As height increases distance decreases.
(E) There is a negative quadratic relationship between height and distance.

Which of the following is true?

(A) Both height and distance are explanatory variables.
(B) Height is the explanatory variable.
(C) Height is the response variable.
(D) Distance is the explanatory variable.

b_0 b_1

The least square solution is: intercept = 6.4285; and slope = 1.0534
Interpret the slope.

distance $= 6.4285 + 1.0534 \cdot$ height

(A) As distance increases by one inch, height increases by 1.0534 inches.
(B) As height increases by one inch, distance increases by 6.4284 inches.
(C) As distance increases by 6.4285 inches, height increases by 1.0534 inches.
(D) As distance increases by one inch, height increases by 6.4285 inches.
(E) As height increases by one inch, distance increases by 1.0534 inches.

For the pair of data (75, 85), what is the residual?

(A) - 0.4335
(B) 160.4335
(C) 0.4335
(D) 0
(E) - 398.1909

$\hat{y} = 6.4285 + 1.0534 \cdot 75$

$= 85.4335$

residual $= y - \hat{y}$

$= 85 - 85.4335$

$= -0.4335$

Example: In a simple linear regression problem, the least squares line is given by $y = 2.15 - 1.75x$, and the coefficient of determination is 0.81. What is the correlation?

$r^2 = 0.81$

$r = \sqrt{0.81}$

-0.9

(A) 0.81

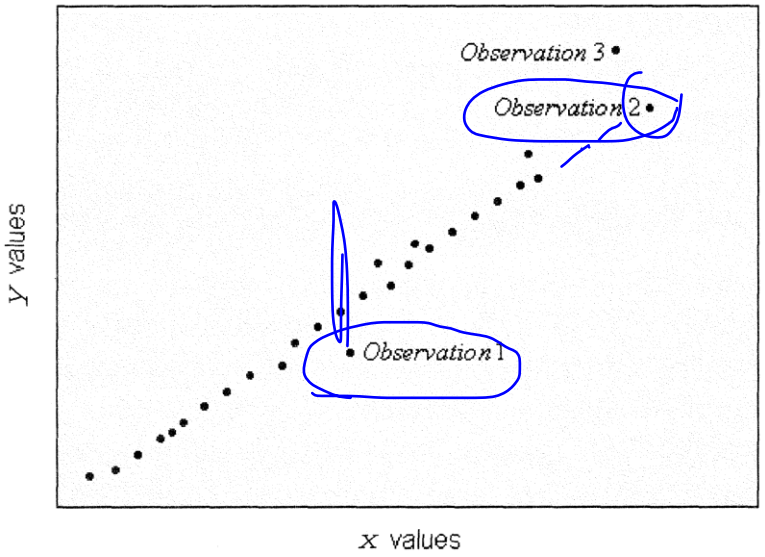
(B) -0.81

(C) 0.9

(D) -0.9

(E) Impossible to determine

Example: Given the graph,



which of the statements are true?

I. Observations 1 and 2 are not influential

II. Observations 2 and 3 have high leverage

III. Observations 2 and 3 have large residuals

IV. Only observation 3 is an outlier

(A) I only

(B) I and II

(C) I, II, and III

(D) IV only

(E) all are correct