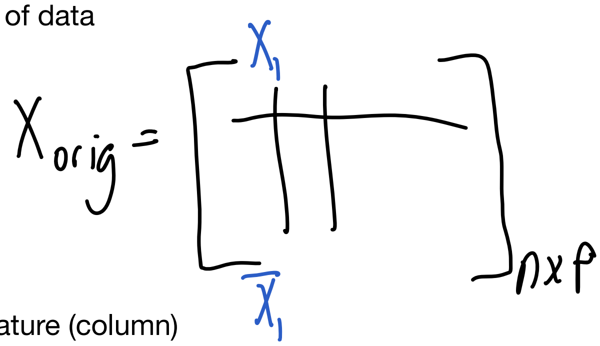
- Recap PCA
  - Transform higher dimensional data to lower dimensional data
  - Linear transformation
  - Applications
    - Used to visualize clusters of data
  - Steps
    - Start with original data X



    - Subtract mean of each feature (column)

$n \times r$ (usually 2)

    - Covariance matrix

$$A = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) \\ cov(x_2, x_1) & cov(x_2, x_2) \end{bmatrix}$$

    - Compute eigenvalues

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

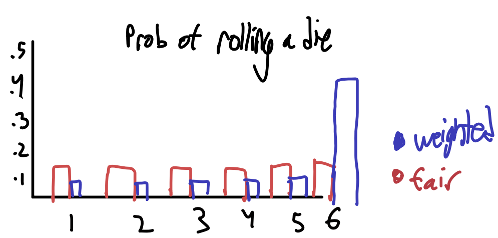
    - Compute transformed data

$$T_{n \times r} = X_{n \times p} W_{p \times r}$$

- Statistics
  - Test to see if a method is *significantly* better than some other method
  - Useful concepts
    - Distributions
      - Probability distributions

Probability mass function (pmf): $p(x)$

$$\sum_{x \in vals(x)} p(x) = 1$$



- Expected value
  - Weighted average

$$E(x) = \sum_{x \in Vals(x)} x * p(x)$$

ex for die $X = \{1, 2, \ldots, 6\}$

$$\sum_{x \in \{1,\ldots,6\}} x * p(x) = 1 * \frac{1}{6} + 2 * \frac{1}{6} + \cdots + 6 * \frac{1}{6}$$

$$= \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6}$$

$$= \frac{21}{6} = 3.5$$

So $E(x) = 3.5$

Weighted die: $p(1, \ldots, 5) = 0.1$

$p(6) = 0.5$

$$E(x) = \sum_{x \in Vals(x)} x * p(x) = 1 * \frac{1}{10} + 2 * \frac{1}{10} + 3 * \frac{1}{10} + 4 * \frac{1}{10} + 5 * \frac{1}{10} + 6 * \frac{5}{10}$$

$$= \frac{1}{10} + \frac{2}{10} + \frac{3}{10} + \frac{4}{10} + \frac{5}{10} + \frac{30}{10}$$

$$= \frac{45}{10} = 4.5$$

- Sample (empirical) mean:
  - If you don't know probabilities for weighted die, you can roll it a bunch of times

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Variance

$$Var(x) = E[\underbrace{(X - \overset{E(X)}{\mu})^2}_{\text{spread}}] = \sum_{x \in Vals(x)} (x-\mu)^2 p(x)$$

Fair die

$$E[X_{fair}] = 3.5$$

$$Var(X_{fair}) = \sum_{x \in Vals(x)} (x-3.5)^2 \frac{1}{6} = \frac{1}{6}\left[(1-3.5)^2 + \dots + (6-3.5)^2\right] \approx 2.92$$

Weighted die

$$p(1,\dots,5) = \frac{1}{10}$$
$$p(6) = \frac{1}{2}$$

$$E[X_w] = 4.5$$

$$Var(X_w) = \sum_{x \in Vals(x)} (x-4.5)^2 p(x) = \frac{1}{10}\left[(1-4.5)^2 + \dots + (5-4.5)^2\right] + \frac{1}{2}(6-4.5)^2 = 3.25$$

- Sample (empirical) variance

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - Trying to estimate population variance based on sample

- Central limit theorem
  - If X1, X2,…, Xn are samples from a population with expected value μ and finite variance σ^2, and $\overline{X}n$ is the sample mean, then:

$$Z = \lim_{n \to \infty} \left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right)$$

mean

variance

is a standard normal distribution N(0,1)

Central Limit theorem: if the population is not normal, when n is large, the sample mean is approximately normal

-Even if the original data isn't normal, the mean of the samples (Xbar) will always converge to a normal distribution as n approaches infinity

If you don't normalize the distribution to Z as N(0,1) then we end up with: $N\left(\bar{X}_n, \frac{\sigma^2}{n}\right)$, with $\frac{\sigma^2}{n}$ as sample variance, $\frac{\sigma}{\sqrt{n}}$ as standard deviation, and $\bar{X}_n$ as mean. Its the same, just dealing with means of samples instead of z values for the data.

- Z score
  - Tells us how many standard deviation the sample mean is away from the expected value
  - A z score of 0 means the sample mean is the same as expected value

- Hypothesis testing
  - H0: null hypothesis (eg die is fair)
  - H1: alternative hypothesis (eg die is weighted towards higher values)
  - Apply central limit theorem:

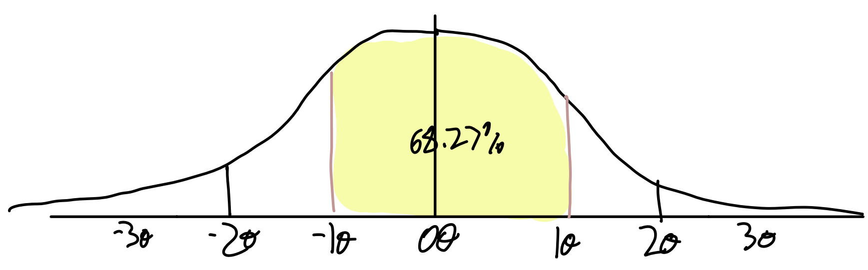
$$z\text{-score} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \qquad \text{sample: } n=20, \ \bar{X}_n = 4.2$$

$$z\text{-score(test statistic)} = \frac{4.2-3.5}{2.92/\sqrt{20}} \approx 1.83$$

- P-value
  - Probability of observing a result as or more extreme than our under the null hypothesis
  - Probability density function

$$pdf = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

area under curve within some range



66.27% fall within 1σ of Mean

• Hypothesis testing

test-statistic = 1.83

$$\text{p-value} = \int_{1.83}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \approx 0.033$$

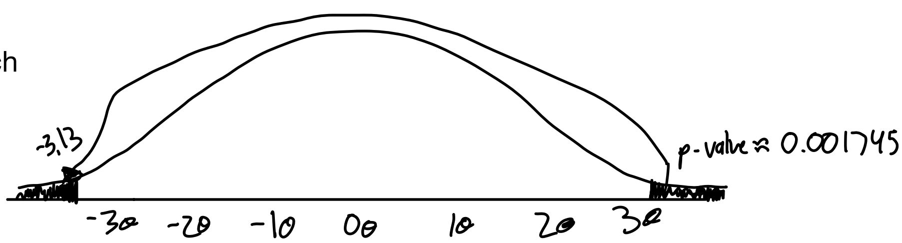
- Usually compare p-value to alpha value of 0.05 (significance level)
- Since p-value = 0.033 <= 0.05, we reject the null hypothesis

Handout 17

1. E[x] = 0.5
2. Var(X) = 1/2[(1-.5)^2+(0-.5)^2] = 0.25
3. Sample mean $\bar{X}_n$ = 54/80 = 0.675

4. Test statistic $= \sqrt{80}\left(\frac{0.675-0.5}{\sqrt{0.25}}\right) = 3.13$

5. Sketch



6. Since p-value = 0.001745 <= 0.05, we reject the null hypothesis