STAT 111

Recitation 5

Mo Huang

Email: mohuang@wharton.upenn.edu

Office Hours: Wednesdays 3:00 - 4:00 pm, JMHH F96

Slides (adapted from Gemma Moran): `github.com/mohuangx/STAT111-Fall2018`

October 12, 2018

Two-Standard-Deviation Rule

- From the chart:

$$P(Z < -1.96) = 0.025, \quad P(Z > 1.96) = 0.025.$$

- Then:

$$P(-1.96 < Z < 1.96) = 0.95.$$

- Approximate $1.96 \approx 2$ and "unstandardize":

$$P\left(-2 < \frac{X-\mu}{\sigma} < 2\right) = 0.95$$

$\Rightarrow$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95.$$

$$P(\mu - 2.576\sigma < X < \mu + 2.576\sigma) = 0.99.$$

- The probability that a normal random variable is within 2 (2.576) standard deviations of the mean is 95% (99%).

Central Limit Theorem

The Central Limit Theorem:

- Suppose $X_1, X_2, \ldots, X_n$ are *iid* with mean μ and variance σ^2.
- Then, for large n

$$T_n \sim N(n\mu, n\sigma^2) \quad \text{and} \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

no matter the distribution of the individual X_i

- Allows approximation of all distributions using the normal distribution if you know the mean and variance.

Note: if $X_1, \ldots, X_n$ are normally distributed, then this applies for *all* n, not just large n.

Central Limit Theorem: Example

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Binomial(1, \theta)$.

- For each X_i, $Mean(X_i) = \theta$ and $Var(X_i) = \theta(1 - \theta)$.

- The sum is: $T_n = X_1 + X_2 + \cdots + X_n$.

- The proportion is:

$$P = \frac{X_1 + \cdots + X_n}{n}.$$

- For large n,

$$T_n \sim N\left(n\theta, n\theta[1 - \theta]\right)$$

$$P \sim N\left(\theta, \frac{\theta(1 - \theta)}{n}\right)$$

Central Limit Theorem: Problem

- Suppose you are rolling a fair die 1000 times. Calculate the numbers A and B such that the average of the 1000 rolls is between A and B with probability approximately 0.95. You may assume the mean of one roll is 3.5 and the variance is $35/12$.

$$Mean(X_i) = 3.5, \quad Var(X_i) = 35/12$$

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(3.5, \frac{35}{12000}\right) \quad \text{by CLT}$$

$$A = -2\sigma + \mu = -2\sqrt{35/12000} + 3.5 \approx 3.392$$

$$B = 2\sigma + \mu = 2\sqrt{35/12000} + 3.5 \approx 3.608$$

Statistics

- We have finished the first half of the course on probability. Now, we move on to statistics.
- Statistics is used to make inductive statements about some phenomenon (coin-flipping, dice rolling) **after** observing data.
- Three main activities of statistics:
  1. Estimating numerical values of a parameter or parameters.
  2. Assessing accuracy of these estimates.
  3. Testing hypotheses about the numerical values of parameters.
- Example: Suppose flip a coin 1,000 times and observe 700 heads.
  1. How do I estimate the probability θ of achieving a head?
  2. How accurate is my estimate of θ?
  3. Is this a fair coin ($\theta = 0.5$)?

Estimation of a parameter: Binomial parameter θ

- Recall a binomial random variable $X \sim Bin(n, \theta)$. How do we estimate the probability of success θ?
- An intuitive estimator for θ is $p = x/n$, the **observed** proportion of successes.
- Consider the random variable P, the proportion of successes **prior** to performing the experiment.
  - $Mean(P) = \theta$ so p is "shooting at the right target". p is then referred to as an **unbiased** estimate of θ.
- Difference between estimate and estimator:
  - **Estimate:** A function of the observed data used to estimate a given parameter. Ex: p.
  - **Estimator:** The random variable whose realization is the estimate. Ex: P.
- To investigate the precision of an estimate, we need to consider the random variable estimator.

Precision of an estimate: Binomial parameter θ

- Precision of p as an estimate of θ depends on the variance of random variable P.

- By the CLT, two standard deviation rule, and approximations (see pg. 40-41), we get the approximate 95% confidence interval for θ as

$$p \pm 2\sqrt{p(1-p)/n}$$

- We can further approximate the 95% confidence interval with $p(1-p) \leq 1/4$ to get

$$p \pm \sqrt{1/n} \tag{66}$$

- Correspondingly, the 99% confidence interval is

$$p \pm 2.576\sqrt{p(1-p)/n} \approx p \pm 1.288\sqrt{1/n}$$

Example

In the 2017-2018 NBA season, Lebron James shot 531 free throws and made 388. We want to estimate the probability θ that Lebron James makes a free throw.

Q1: What is the estimate for θ?

$$p = x/n = 388/531 = 0.7307$$

Q2: Calculate the 95% confidence interval for θ using the approximate 95% interval formula (66):

$$p \pm \sqrt{1/n} = 0.7307 \pm \sqrt{1/531} = 0.7307 \pm 0.0434$$

Q3: What is the sample size if we want the width of the confidence interval to be 0.02?

$$\begin{aligned} \text{We want } \sqrt{1/n} &= 0.01 = 0.02/2. \\ 1/n &= 0.01^2 \\ n &= 1/0.01^2 = 10000 \end{aligned}$$