# MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

# Review

- Last time, finished Module2 Probability Review.
- Today we start Module 3 Confidence Intervals.

We first discuss the sampling distribution of the Mean (Chapter 8). The theoretical results introduced in this chapter is used to derive the confidence interval formulas in the next chapter.

# Chapter8 Sampling Distribution of the Mean

There are three theoretical results that later will be used to derive confidence interval formulas.

- (1) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ ~ N(0,1) whose probability is found through TableA.3 or use pnorm() in R.
- (2) $\frac{(n-1)s^2}{\sigma^2}$ ~ $\chi^2_{n-1}$ whose probability is found through TableA.8 or use pchisq() in R.
- (3) $\frac{\overline{X}-\mu}{s/\sqrt{n}}$ ~ $t_{n-1}$ whose probability is found through TableA.4 or use pt() in R.

# Chapter8 Sampling Distribution of the Mean

Let $X_1, \ldots, X_n$ i.i.d ~ N($\mu$,$\sigma^2$), i.e., a random sample.

*Parameter* $\mu$ is the population mean. To *estimate* it, we use the sample mean $\overline{X}$

- Question: how good is the estimate $\overline{X}$?

  In fact $\overline{X}$ ~ N($\mu$,$\frac{\sigma^2}{n}$), which we will derive later

- Theoretical result (1) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ ~ N(0,1) whose probability is found through TableA.3 or use pnorm() in R.

# Chapter8 Sampling Distribution of the Mean

$X_1, \ldots, X_n$ i.i.d ~ N($\mu$,$\sigma^2$)

*Parameter* $\sigma^2$ is the population variance. It is *estimated* by the sample variance

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

- Theoretical result (2) $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ whose probability is found through TableA.8 or use pchisq() in R.

# Chapter8 Sampling Distribution of the Mean

Recall $\dfrac{Z_0}{\sqrt{(Z_1^2+\cdots+Z_k^2)/k}} = \dfrac{N(0,1)}{\sqrt{\chi_k^2/k}} \sim t_{df=k}$

So long as we have the independence between numerator and denominator, we get

- Theoretical result (3)

$$\frac{N(0,1)}{\sqrt{\chi_{n-1}^2/(n-1)}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}} = \frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$ whose probability is found through TableA.4 or use pt() in R.

# Chapter8 Sampling Distribution of the Mean

Example: An adult human's height is normally distributed with mean 63 inches and standard derivation 7 inches.

(a) What proportion of adults have heights between 56 and 70 inches?

(b) We randomly choose five adults, record their heights as $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ , and take an average $\overline{X} = \frac{X_1+X_2+X_3+X_4+X_5}{5}$.

What is the mean and standard derivation of $\overline{X}$?

What is the probability of $\overline{X}$ falls between 56 and 70 inches?

# Height Example Solutions

X= height of a randomly chosen person ~ N(μ=63,$\sigma^2$= $7^2$)

(a) We learned this calculation in last module

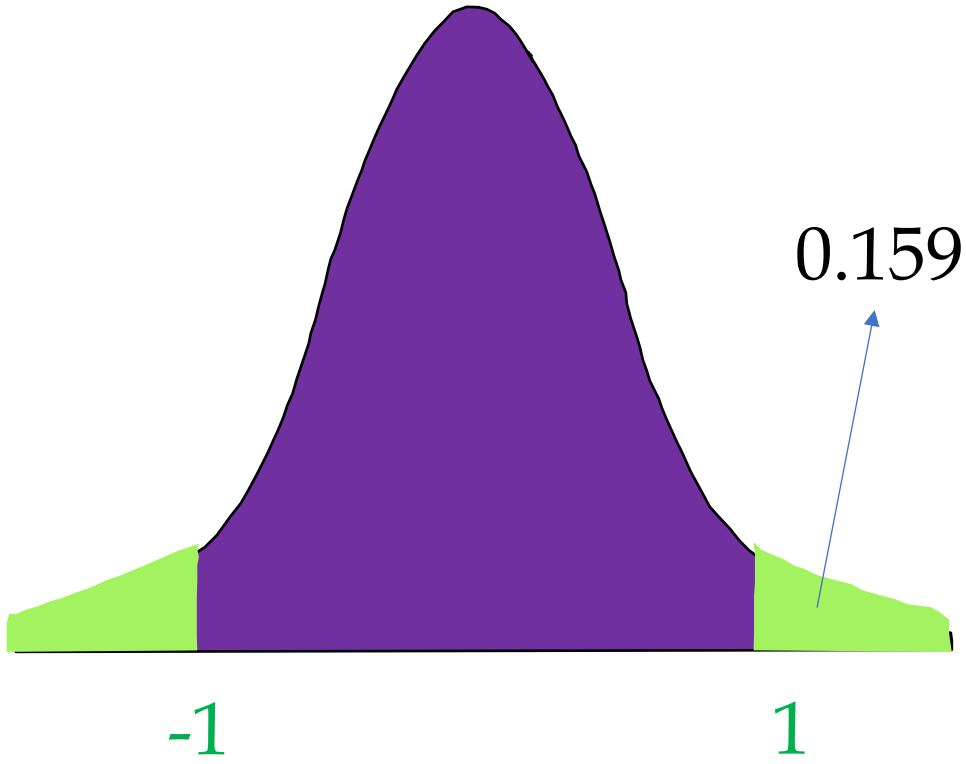
$Z=\frac{X-63}{7}$ ~ N(0,1), use Table A.3

$P(56 \le X \le 70) = P(\frac{56-63}{7} \le Z \le \frac{70-63}{7})$

$= P(-1 \le Z \le 1)$

$= 1 - 2\,P(Z>1)$

$= 1-2(0.159) = 0.682$



# Height Example Solutions

(b) Use properties of Mean and Variance in last module

$$\mathrm{E}(\overline{X}) = E\left(\frac{X_1+X_2+X_3+X_4+X_5}{5}\right) = \frac{1}{5}[E(X_1)+\cdots+E(X_5)] = \frac{1}{5}[63+\cdots+63] = 63$$

$$\mathrm{Var}(\overline{X}) = Var\left(\frac{X_1+X_2+X_3+X_4+X_5}{5}\right) = \left(\frac{1}{5}\right)^2 Var(X_1+\cdots+X_5)$$

$$= \frac{1}{25}[Var(X_1)+\cdots+Var(X_5)] \qquad \text{(Due to i.i.d.)}$$

$$= \frac{1}{25}[49+\cdots+49] = \frac{(5)(49)}{25} = \frac{49}{5}, \qquad \text{So } \mathrm{sd}(\overline{X}) = \sqrt{\frac{49}{5}} = \frac{7}{\sqrt{5}} = 3.13.$$

$$\overline{X} \sim \mathrm{N}(\mu=63, \sigma^2=3.13^2)$$

Then rest of the probability calculation is similar to that in part (a).

# Height Example Solutions

(b) (cont'd) $\overline{X} \sim N(\mu=63, \sigma^2=3.13^2)$

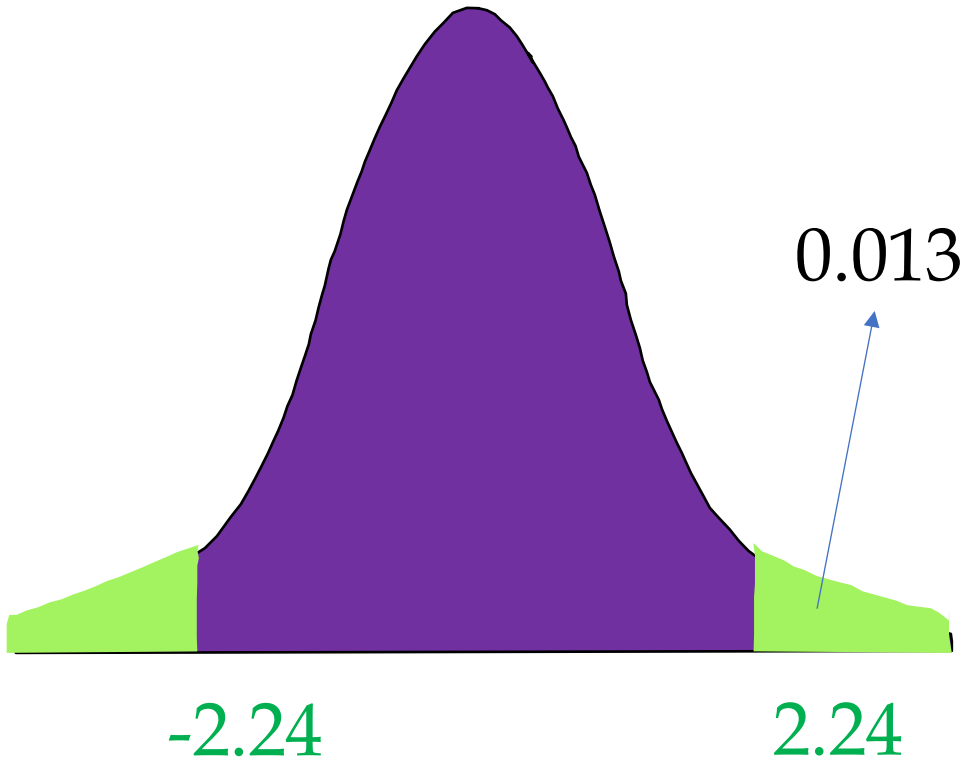
$Z=\frac{\overline{X}-63}{3.13} \sim N(0,1)$, use Table A.3

$P(56 \leq \overline{X} \leq 70) = P(\frac{56-63}{3.13} \leq Z \leq \frac{70-63}{3.13})$

$= P(-2.236 \leq Z \leq 2.236)$

$= 1-2(0.013) = 0.974$

*Notice that this is much higher than* $P(56 \leq X \leq 70) = 0.682$ *in part (a)*



# Theoretical result (1) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

The above example shows how to derive the Theoretical result (1) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, which shows how good the sample mean $\overline{X}$ is as an estimator $\hat{\mu}$ for the true parameter $\mu$.

Let $X_1, \ldots, X_n$ i.i.d with mean $\mu$, and variance $\sigma^2$.

Use properties of Mean and Variance in last module

$$\mathrm{E}(\overline{X}) = E\left(\frac{X_1+\cdots+X_n}{n}\right) = \frac{1}{n}[E(X_1)+\cdots+E(X_n)] = \frac{1}{n}[\mu+\cdots+\mu] = \mu$$

$$\mathrm{Var}(\overline{X}) = Var\left(\frac{X_1+\cdots+X_n}{n}\right) = \left(\frac{1}{n}\right)^2 Var(X_1+\cdots+X_n)$$

$$= \frac{1}{n^2}[Var(X_1)+\cdots+Var(X_n)] \qquad \text{(Due to i.i.d.)}$$

$$= \frac{1}{n^2}[\sigma^2+\cdots+\sigma^2] = \frac{1}{n}\sigma^2, \qquad \text{So } \mathrm{sd}(\overline{X}) = \sqrt{\frac{1}{n}\sigma^2} = \frac{\sigma}{\sqrt{n}}.$$

# Theoretical result (1) $\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$)

$X_1, \ldots, X_n$ i.i.d with mean μ, and variance $\sigma^2$. Estimator $\hat{\mu} = \overline{X}$

$E(\overline{X}) = \mu$. Hence $\hat{\mu} = \overline{X}$ is unbiased for estimating the parameter μ.

$sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$. So the standard error of an estimator $\hat{\mu}$ is $sd(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$.

Hence as $n \to \infty$, $sd(\hat{\mu}) \to 0$. This and the unbiasedness implies that

$$P(|\hat{\mu} - \mu| > \varepsilon) \to 0 \text{ for any } \varepsilon > 0.$$

That is, $\hat{\mu} = \overline{X}$ is a consistent estimator for μ.

Notice that the unbiasedness and consistency derivation above both does not need the assumption that $X_i$s are normally distributed.

If we further assume that $X_i$s are normally distributed, use the property for linear combination of normal R.V.s, then $\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$)

# Theoretical result (1) $\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$)

$X_1$, …, $X_n$ i.i.d with mean μ, and variance $\sigma^2$.

If we further assume that $X_i$s are normally distributed ~ N(μ,$\sigma^2$), then

$\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$) exactly.

What if we do not make this normal assumption?

The Central Limit Theorem (CLT) says that,

$\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$) approximately for large n.

# Illustration of Central Limit Theorem

$X_1, \ldots, X_n$ i.i.d with mean $\mu$, and variance $\sigma^2$.

CLT: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ approximately for large n.

How big the n should be? (Rule of thumb: n≥30)

With homework 3, there is a R code illustrating CLT.

The idea: generate many (B) data sets each of size n.

For each data set, calculate the sample mean $\overline{X}$ .

Use such B sample means to see the distribution of $\overline{X}$ .

# Illustration of Central Limit Theorem

```
### A simulated example illustrating CLT
## Generate n=100 random variables from a Binomial(3, 0.2) distribution
## Have B=400 observations of each variable (we want B data sets each of size n).
n <- 100
B <- 400
rand.data <- matrix(rbinom(B*n, size=3, prob=0.2), nrow=B)
## Create summary variables;
# v1--the first variable, mean2 -- average of the first two variables
# mean10 -- average of the first 10 variables, similarly mean30, etc.
v1 <- rand.data[,1] #The first column (variable)
#Average of the first two columns(variables),
# apply function 'mean' on margin 1 (row-wise function application)
mean2 <- apply(rand.data[,1:2], FUN=mean, MARGIN=1)
#Average of the first ten columns(variables)
mean10<- apply(rand.data[,1:10], FUN=mean, MARGIN=1)
#Average of the first 30, 50, 100 columns (variables)
mean30 <- apply(rand.data[,1:30], FUN=mean, MARGIN=1)
mean50 <- apply(rand.data[,1:50], FUN=mean, MARGIN=1)
mean100 <- apply(rand.data[,1:100], FUN=mean, MARGIN=1)

##Plot the histograms of the summary variables above, overlay with
#   the density plot of the normal distribution in the CLT.
mu    <- 3*0.2                #Binomial mean
sigma <- sqrt(3*0.2*(1-0.2))  #Binomial standard deviation
## Arrange 6 plots in one page 2 by 3
par(mfrow=c(2,3))
# Histogram of v1,
#   then add the normal density curve in red color (col=2)
hist(v1, freq = FALSE, main="histogram when n=1")
curve(dnorm(x, mean=mu, sd=sigma),col=2,add=T)
# Histograms of mean2, mean10, ....
hist(mean2, freq = FALSE, main="histogram when n=2")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(2)),col=2,add=T)
hist(mean10, freq = FALSE, main="histogram when n=10")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(10)),col=2,add=T)
hist(mean30, freq = FALSE, main="histogram when n=30")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(30)),col=2,add=T)
hist(mean50, freq = FALSE, main="histogram when n=50")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(50)),col=2,add=T)
hist(mean100, freq = FALSE, main="histogram when n=100")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(100)),col=2,add=T)
```

# Illustration of Central Limit Theorem

In the example, we generate $V_1, \ldots, V_n$ i.i.d. Binomial(3,0.2).

Get B=400 copies of $V_1, \ldots, V_{100}$ from rbinom(B*n, size=3, prob=0.2)

And arrange them into a 400 by 100 matrix matrix(…, nrow=B)

Then calculate the averages $V_1$, $\frac{V_1+V_2}{2}$, $\frac{V_1+\cdots+V_{10}}{10}$, …, $\frac{V_1+\cdots+V_{100}}{100}$

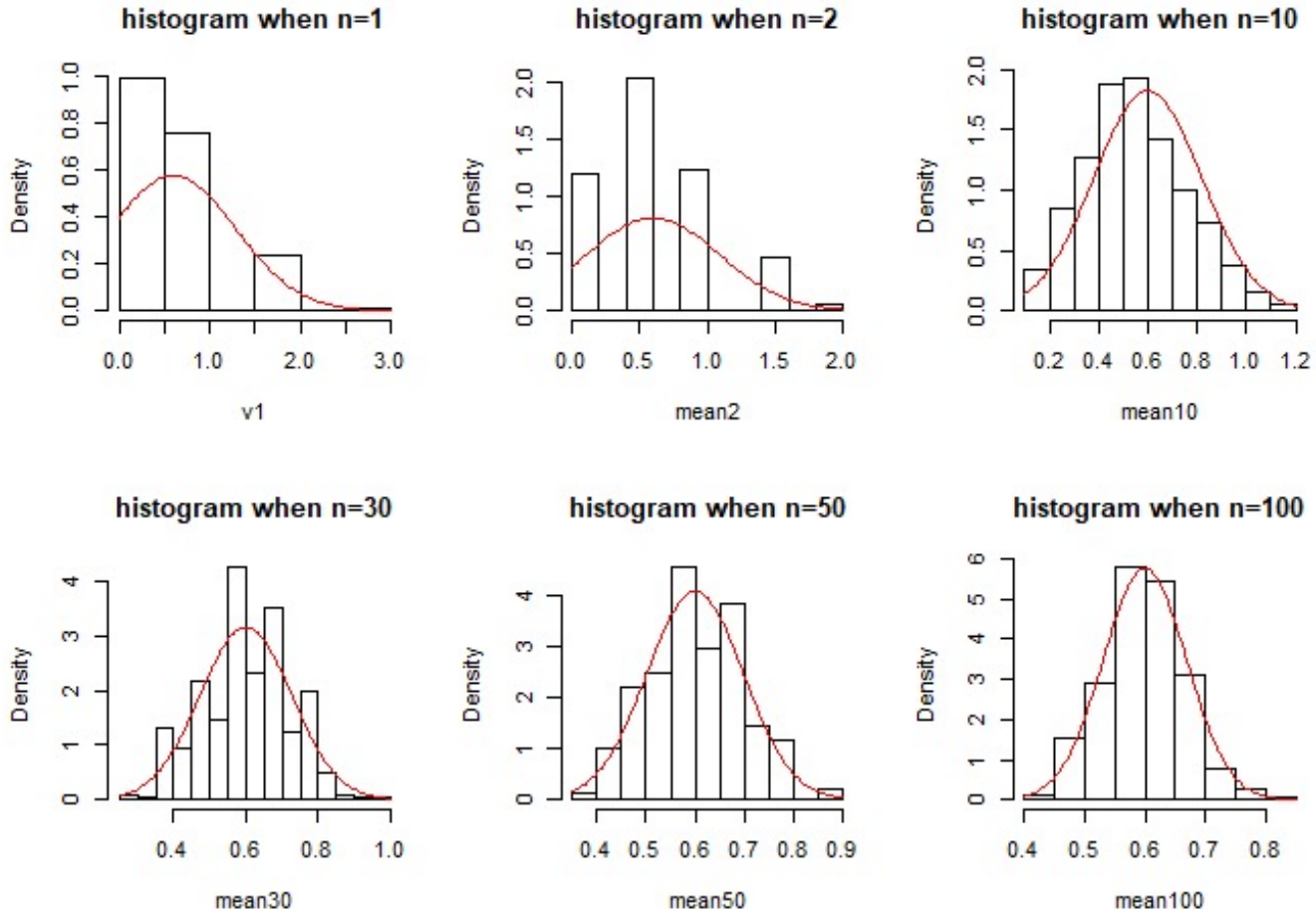
apply(rand.data[,1:?], FUN=mean, MARGIN=1)

This gives a vector of length 400 for values of $\frac{V_1+\cdots+V_?}{?}$

We then plot the histogram of these 400 simulated $\overline{X}$ values, to observe the distribution of $\overline{X}$.

# Illustration of Central Limit Theorem

The distribution of $\overline{X}$ becomes closer to normal as n increase. Skewed at beginning Symmetric bell-curve start around n=30.



# R commands in CLT Illustration example

hist(v1, freq = FALSE, main="histogram when n=1")

freq = FALSE so y-axis displays proportion instead of frequency

main= gives the title of the plot

curve(dnorm(x, mean=mu, sd=sigma),col=2,add=T) this adds a curve of normal pdf to the histogram for easier visual inspection.

col=2 color= 2nd type (red)

add=T the curve is added to plot (overlay), not starting a new plot

dnorm(x, mean=mu, sd=sigma) calculate the normal pdf (density). The mean/variance were matched earlier for best normal approximation

```
mu    <- 3*0.2                    #Binomial mean
sigma <- sqrt(3*0.2*(1-0.2))  #Binomial standard deviation
```

# Illustration of Central Limit Theorem

How big a n is needed for the distribution of $\overline{X}$ be close enough to normal depends on the distribution of the original variable.

One homework problem is to do the above simulation for a few other distributions.

There is also an R applet doing similar illustration.

https://adamding.shinyapps.io/CLTadamding/

# Monte Carlo Simulation

The above example is a Monte Carlo Simulation:

- MC main idea: generate $X_1, \ldots, X_B$ from distribution D.

As $B \to \infty$, the sample distribution $D_B \to D$:

The sample CDF $P_B(X \le x) = \frac{\#(X_i \le x)}{B}$,

Since by probability definition $P(X \le x) = \lim_{B \to \infty} \frac{\#(X_i \le x)}{B}$,

the sample CDF $P_B(X \le x) \to$ CDF $P(X \le x)$ when $B \to \infty$.

- For any population quantity S, its sample version $S_B$ converges to S when $B \to \infty$.

# Monte Carlo Simulation

- MC main idea: generate $X_1, \ldots, X_B$ from distribution D.

Use B as large as your computing power allows, then the sample quantity $S_B(X_1, \ldots, X_B)$ gives a good approximation to the population quantity S of interest.

- When do we use MC simulation?

For simple distribution, we can get analytic explicit answer for the population quantity S. Thus no need.

But for complex relationship, when analytic answers are hard to derive, MC can provide an easy alternative which only requires sufficient computing power.

# Monte Carlo Simulation

- Example: Say X~$D_1$, Y~$D_2$, Z~$D_3$, and let W=$X^2 + \sin(Y/Z)$.

What is E(W) and Var(W)? Analytic solution can be hard to get

- Alternatively, use MC simulation:

generate ($X_1$, $Y_1$, $Z_1$), …, ($X_B$, $Y_B$, $Z_B$).

calculate $W_i = X_i^2 + \sin(Y_i/Z_i)$ for $W_1$, …, $W_B$.

estimate E(W) by $\overline{W} = \frac{1}{n}\sum_{i=1}^{B} W_i$ , estimate Var(W) by $s_W^2 = \frac{1}{n-1}\sum_{i=1}^{B}(W_i - \overline{W})^2$

- Any other quantities about W such as density, CDF, skewedness and kurtosis etc. are all estimated similarly
- In the CLT example, W= $\overline{X}$. We generated B=400 copies such W, and used its histogram to estimate the density of $\overline{X}$.

# Use Theoretical result (1) $\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$)

Example (exercise 8.5.2): For USA adults, the albumin levels in cerebrospinal fluid is normally distributed with mean μ=29.5 mg/100ml and standard derivation σ=9.25mg/100ml. What proportion of the means of samples of size 20 lie above 35 mg/100ml?

Solution: $\overline{X} = \frac{X_1+\cdots+X_{20}}{20}$~ N(μ,$\frac{\sigma^2}{20}$) = N(mean=29.5,var= $\frac{9.25^2}{20} = 4.278$).

P($\overline{X}$ ≥35) = P(Z ≥ $\frac{35-29.5}{\sqrt{4.278}}$) = P(Z ≥ 2.66) =0.004 (from TableA.3)

Question: If you measure 20 persons' albumin levels and get an average of 37mg/100ml. *What should you conclude?*

# Use Theoretical result (1) $\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$) for Confidence Intervals (chapter 9)

We want an interval estimate $(\overline{X} - L, \overline{X} + L)$ to give an indication of uncertainty about point estimator $\hat{\mu} = \overline{X}$ .

Confidence Interval (CI): an interval capturing the true parameter with a certain change in the long term.

Since $\overline{X}$ ~ N(μ,$\frac{\sigma^2}{n}$) , $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ ~ N(0,1). Hence from Table A.3, we get

P(-1.96≤ $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ ≤1.96)=0.95 $\Leftrightarrow$ P(-1.96 $\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le$1.96 $\frac{\sigma}{\sqrt{n}}$)=0.95

$\Leftrightarrow$ P(-1.96 $\frac{\sigma}{\sqrt{n}} \le \mu - \overline{X} \le$1.96 $\frac{\sigma}{\sqrt{n}}$)=0.95

$\Leftrightarrow$ P($\overline{X}$ -1.96 $\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}$ +1.96 $\frac{\sigma}{\sqrt{n}}$)=0.95

So a 95% CI for $\mu$, when $\frac{\sigma}{\sqrt{n}}$ is known, is ($\overline{X}$ -1.96 $\frac{\sigma}{\sqrt{n}}$ , $\overline{X}$ +1.96 $\frac{\sigma}{\sqrt{n}}$).

# Use Theoretical result (1) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ for Confidence Intervals (chapter 9)

- So a 95% CI for $\mu$, when $\frac{\sigma}{\sqrt{n}}$ is known, is

$$(\overline{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \frac{\sigma}{\sqrt{n}}).$$

- Meaning: if we collect many data sets and use the above formula, about 95% of times the interval captures the true parameter $\mu$.

(Notice that for different data sets, $\overline{X}$ is different, thus the CIs are moving around. $\mu$ is fixed.)

# Use Theoretical result (1) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ for Confidence Intervals (chapter 9)

More generally for the (1-α) CI,

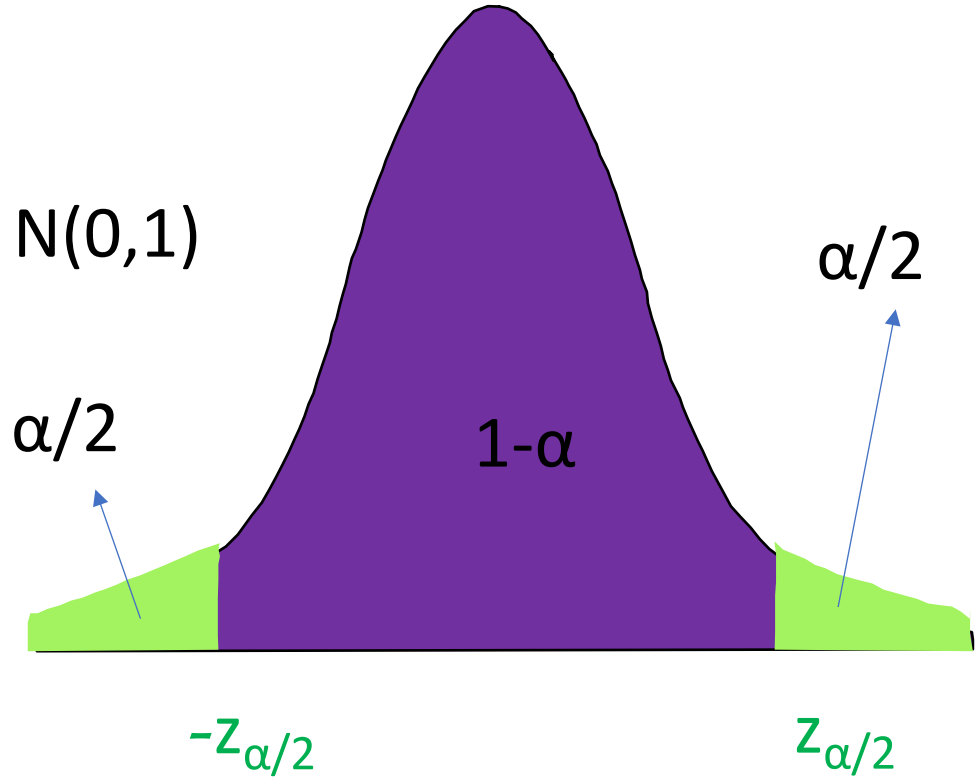
let $z_\alpha$ denote the upper α percentile of N(0,1)

Hence $P(Z \geq z_{\alpha/2}) = \alpha/2$,

$$P(-z_{\alpha/2} \leq Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1-\alpha$$

$$\Leftrightarrow P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu - \overline{X} \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$\Leftrightarrow P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$




So a (1-α) CI for $\mu$, when $\frac{\sigma}{\sqrt{n}}$ is known, is $(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.

In the previous example, $\alpha=0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$ for the 95% CI

# Confidence Intervals (chapter 9)

Example: A random sample of 24 male runners are taken. The sample mean is $\overline{X}$ =60kg. Suppose that the standard deviation of the male runners population is σ=5kg.

Then a 95% CI for the population mean $\mu$ is

$$(60\text{-}1.96\frac{5}{\sqrt{24}}, 60\text{+}1.96\frac{5}{\sqrt{24}}) = (58,62)$$

Then a 90% CI for the population mean $\mu$ is

$$(60\text{-}1.64\frac{5}{\sqrt{24}}, 60\text{+}1.64\frac{5}{\sqrt{24}}) = (58.33, 61.67)$$

The 90% CI is shorter than the 95% CI. Does this make sense?

# Summary

- We discussed about three theoretical results. Derived the first one. Checked the conditions when it holds (either using normal assumption, or using CLT when sample size n is large.)
- Introduced the concept of Monte Carlo simulation. Used MC simulation to illustrate CLT.
- Used the theoretical result to derive confidence interval formula.
- Next time, we will cover the confidence intervals (Chapter 9) in more detail.