Eco311: Review of Probability and Statistics——Math Refresher or Appendix B and C of Wooldridge's textbook

(Jing Li, Miami University)

What is statistics

1. Statistics is about using sample to understand population
2. We are interested in population, but we cannot get data for everybody. For instance, the government wants to know what's the proportion of people who have got coronavirus. This is a difficult task because it is impossible to give every person a medical test
3. However, it is usually much easier to get data for a sample, which by definition is a subset of population. For instance, we may give every student in this eco311 class a covid test
4. Then statistics tells us how to use the result from a sample to make statistical inference about population. Suppose 1 out of 20 students is tested positive, then we can say about 5 percent of population have the virus
5. We emphasize the word "about" because we may get different result using a different sample (e.g., students in eco317 class can be another sample). One key issue of statistics is accounting for sampling variability

Random Sample

1. Statistics works well when the sample is representative of population. The best sample is random sample or iid sample, which is obtained by randomly selecting members from population. Random sample ensures that everyone has equal chance to be selected.
2. Obviously the sample consisting of eco311 students is not a random sample—for one thing, young college students cannot represent old people who are more vulnerable to virus
3. We can redefine population as all students at MU. Then the eco311 sample should work better
4. If population is still all Americans, we may use random number such as SSN to select the sample, and give those selected people medical tests
5. Lesson: define your population properly, and do not oversell the results from a specific sample

Distribution, Mean and Variance

1. Using statistics jargon, we assume population follow a random distribution. For instance we can use Bernoulli distribution to describe whether a person has virus ($y = 1$) or has no virus ($y = 0$)
2. A random distribution can be characterized by moments
   (a) the first moment is expected value, also called population mean or expectation

$$\mu \equiv E(y) = \texttt{weighted average of possible values} = \sum_j y_j P(y = y_j) \quad (1)$$

   where the weight is probability. The mean value measures the center of distribution

   (b) the second moment is variance

$$\sigma^2 \equiv var(y) = E(y - \mu)^2 = \sum_j (y_j - \mu)^2 P(y = y_j) \quad (2)$$

   Variance measures dispersion of the distribution

Skewness, Kurtosis, Percentile, and Median

1. The third moment is skewness. A distribution has long right (left) tail if skewness is greater (less) than zero. A distribution is symmetric if skewness is zero
2. The fourth moment is kurtosis. A distribution has a tail fatter (thinner) than normal distribution if kurtosis is greater (less) than 3. The height of tail measures the probability of value on the tail
3. The kth percentile is the value below which k percent of the observations may be found
4. Median is 50th percentile

Properties of Expectation

Let c be a constant, and x and y be two random variables. We have following three properties of expectation

$$E(c) = c \tag{3}$$

Proof: $E(c) = cP(y = c) = c$ since $P(y = c) = 1$

$$E(cy) = cE(y) \tag{4}$$

Proof: $E(cy) = \sum_j cy_j P(y = y_j) = c\sum_j y_j P(y = y_j) = cE(y)$

$$E(x+y) = E(x) + E(y) \tag{5}$$

Properties of Variance

$$var(c) = 0 \tag{6}$$

Proof: $var(c) = E(c-\mu)^2 = E(c-c)^2 = 0$

$$var(cy) = c^2 var(y) \tag{7}$$

Proof: $var(cy) = E(cy - c\mu)^2 = c^2 E(y-\mu)^2 = c^2 var(y)$

Covariance and Properties

Covariance measures the linear association between x and y, and is defined as

$$\sigma_{x,y} \equiv cov(x,y) = E[(x-\mu_x)(y-\mu_y)] \tag{8}$$

There are three properties of covariance

$$cov(c,y) = 0 \tag{9}$$

$$cov(y,y) = var(y) \tag{10}$$

$$var(x+y) = var(x) + var(y) + 2cov(x,y) \tag{11}$$

The last property looks similar to this

$$(a+b)^2 = a^2 + b^2 + 2ab$$

Correlation

A bounded measurement of association is correlation (coefficient)

$$\rho \equiv corr(x,y) = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} \tag{12}$$

We can prove that

$$-1 \leq \rho \leq 1 \tag{13}$$

Two variables are perfectly positively (negatively) correlated if ρ equals one (negative one)

Exercise: let k be a constant. Prove that

$$corr(kx,y) = corr(x,y)$$

iid Sample

A sample $(y_1, y_2, \ldots, y_n)$ is iid (independently and identically distributed) sample, which is also random sample, if all three conditions below are satisfied

$$E(y_i) = \mu, \quad var(y_i) = \sigma^2, \quad cov(y_i, y_j) = 0, \ (\forall i, j) \tag{14}$$

Recall that we obtain iid sample by randomly drawing members from the population. IID sample is the best sample

Statistical Inference I: Estimation

1. The population mean is unknown since data for population are unavailable, The most common problem of estimation is to find $\mu =?$
2. In statistics class we use sample mean as the estimate for population mean

$$\bar{y} \equiv \frac{\sum_i^n y_i}{n} \quad (\texttt{sample mean}) \tag{15}$$

   Note that sample mean is a random variable because it varies across samples
3. If we use iid sample, then the average of sample mean is population mean, i.e., sample mean from iid sample is an unbiased estimator

$$E(\bar{y}) = \frac{\sum_i^n E(y_i)}{n} = \frac{\sum_i^n \mu}{n} = \frac{n\mu}{n} = \mu \tag{16}$$

4. The dispersion of sample mean is measured by its variance

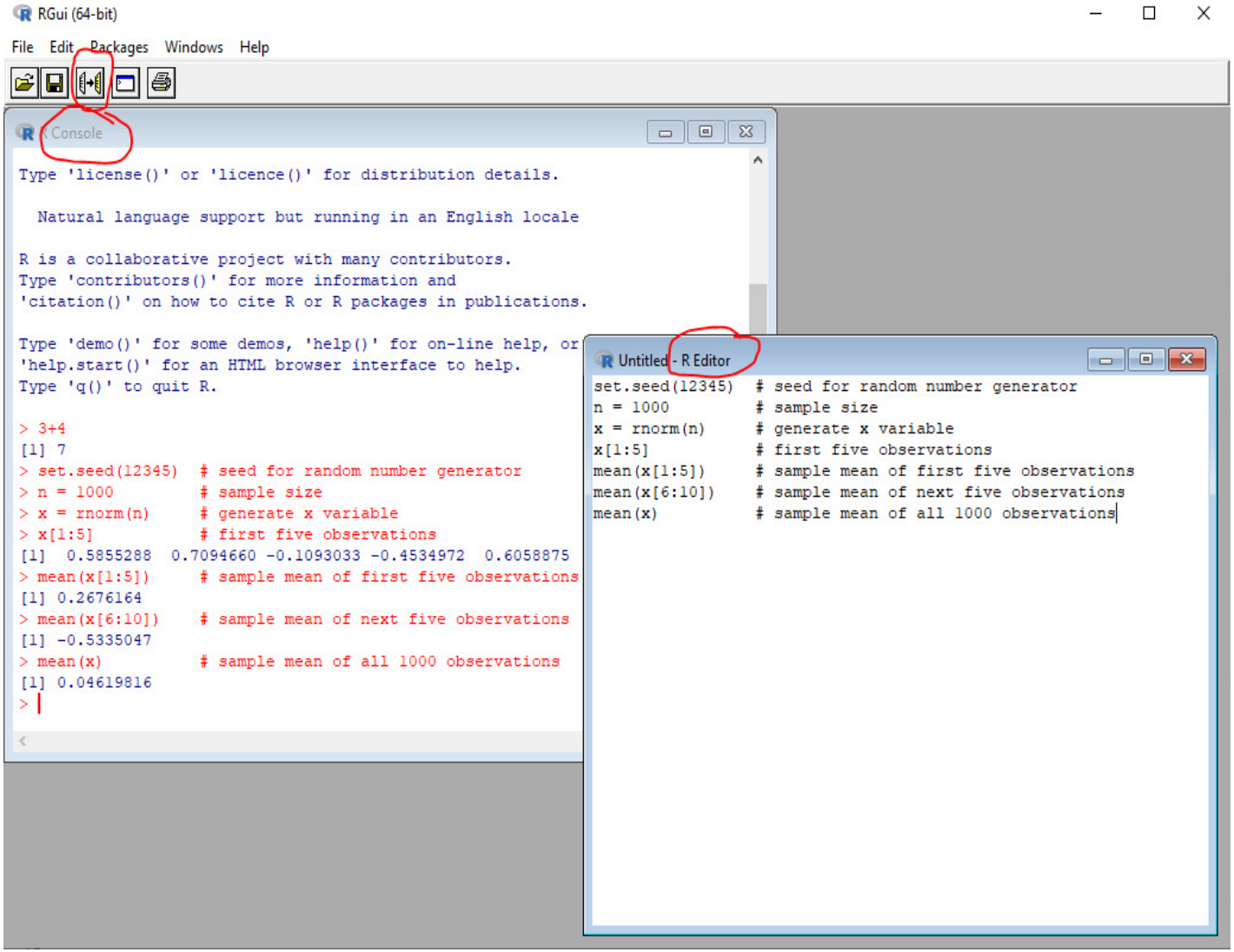
$$var(\bar{y}) = \frac{var(\sum_i^n y_i)}{n^2} = \frac{\sum_i^n \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{17}$$

5. σ is standard deviation; $\frac{\sigma}{\sqrt{n}}$ is standard error

Law of Large Number

1. Notice that as $n \to \infty$, $var(\bar{y})$ approaches 0, see (17). That implies when the sample size increases, sample mean based on iid sample converges to the population mean (a constant). This result is called law of large number
2. So we prefer large sample over small sample, simply because large sample contains more information

R


RGui (64-bit)
File Edit Packages Windows Help
R Console
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> 3+4
[1] 7
> set.seed(12345) # seed for random number generator
> n = 1000 # sample size
> x = rnorm(n) # generate x variable
> x[1:5] # first five observations
[1] 0.5855288 0.7094660 -0.1093033 -0.4534972 0.6058875
> mean(x[1:5]) # sample mean of first five observations
[1] 0.2676164
> mean(x[6:10]) # sample mean of next five observations
[1] -0.5335047
> mean(x) # sample mean of all 1000 observations
[1] 0.04619816
>
Untitled - R Editor
set.seed(12345) # seed for random number generator
n = 1000 # sample size
x = rnorm(n) # generate x variable
x[1:5] # first five observations
mean(x[1:5]) # sample mean of first five observations
mean(x[6:10]) # sample mean of next five observations
mean(x) # sample mean of all 1000 observations

Introduction to R

1. Download R from https://www.r-project.org/
2. Codes are in Editor Window
3. Results are in Console window
4. You can also type command in console window and execute it

R Example: Estimation

```
> set.seed(12345)  # seed for random number generator
> n = 1000         # sample size
> x = rnorm(n)     # generate x variable
> x[1:5]           # first five observations
[1]  0.5855288  0.7094660 -0.1093033 -0.4534972  0.6058875
> mean(x[1:5])     # sample mean of first five observations
[1] 0.2676164
> mean(x[6:10])    # sample mean of next five observations
[1] -0.5335047
> mean(x)          # sample mean of all 1000 observations
[1] 0.04619816
```

Remarks

1. We use simulated data
2. We generate 1000 observations ($n = 1000$) of random values that follow a standard normal distribution. Because data are generated by ourselves, we know the population mean as $\mu = 0$
3. Our first sample consists of the first five observations. The first sample mean is $\bar{x} = 0.268$
4. It is ok that $\bar{x} \neq \mu$, because we are using a sample
5. The second sample mean is $\bar{x} = -0.534$
6. Both sample means are bad estimates as they differ substantially from the true value $\mu = 0$. This finding is not unexpected since we use small samples with only five observations.

Remarks

1. We get a much better sample mean $\bar{x} = 0.05$ using all 1000 observations. This finding is consistent with law of large number
2. Lesson: sample mean is a random variable—sample mean varies from one sample to another. Bad (good) sample can produce bad (good) estimate; in general we prefer using large samples.

Normal Distribution

1. Due to central limit theorem, normal distribution plays a key role in statistics
2. A general normal random variable can be expressed as

$$y \sim N(\mu, \sigma^2) \tag{18}$$

   where μ is the population mean, and σ^2 is the population variance
3. After standardizing y (computing its z-score), we get a standard normal random variable with mean of zero and variance of 1

$$z \equiv \frac{y-\mu}{\sigma} \sim N(0,1) \tag{19}$$

4. Exercise: prove $E(z) = 0, var(z) = 1$
5. In short, z socre measures how many standard deviations a value is from the mean

Normal Distribution

1. Table G.1 of the textbook reports $P(z < \texttt{some value})$. For instance

$$P(z < 1.96) = 0.975 \tag{20}$$

$$P(-1.96 < z < 1.96) = 0.95 \tag{21}$$

   Equation (21) defines the 95 percent confidence interval for a standard normal variable

2. R function **pnorm** reports the probability $P(z < c)$ for given c; function **qnorm** reports c so that $P(z < c)$ equals a given probability. These two functions can substitute Table G.1 of the textbook

3. Exercise: using **qnorm** to find the 90 percent confidence interval for a standard normal variable

R Example

```
> sd(x)
[1] 0.9987476
> #install.packages("moments")
> library("moments")
> skewness(x)
[1] -0.005948778
> kurtosis(x)
[1] 2.975567
> median(x)
[1] 0.04621674
> quantile(x,0.5)
       50%
0.04621674
```

Remarks

1. Using all 1000 observations, the sample skewness -0.006 is close to 0, and sample kurtosis 2.976 is close to 3
2. Because normal distribution is symmetric, the median and mean are close
3. Exercise: Do you think the distribution for household income is symmetric? Which statistic is more relevant—average household income or median household income?
4. Exercise: Find and interpret the 90 percentile of the 1000 observations

Confidence Intervals

1. The 95 percent confidence interval for a general normal variable is

$$P(\mu - 1.96\sigma < y < \mu + 1.96\sigma) = 0.95 \tag{22}$$

Proof: let's subtract population mean and divide by standard deviation for each term in the inequality

$$P(\mu - 1.96\sigma < y < \mu + 1.96\sigma) =$$

$$P\left(\frac{\mu - 1.96\sigma - \mu}{\sigma} < \frac{y - \mu}{\sigma} < \frac{\mu + 1.96\sigma - \mu}{\sigma}\right) = P(-1.96 < z < 1.96) = 0.95 \tag{23}$$

2. Confidence intervals is important because it attaches kind of certainty to randomness. In other words, confidence intervals make a random variable become partially predictable.

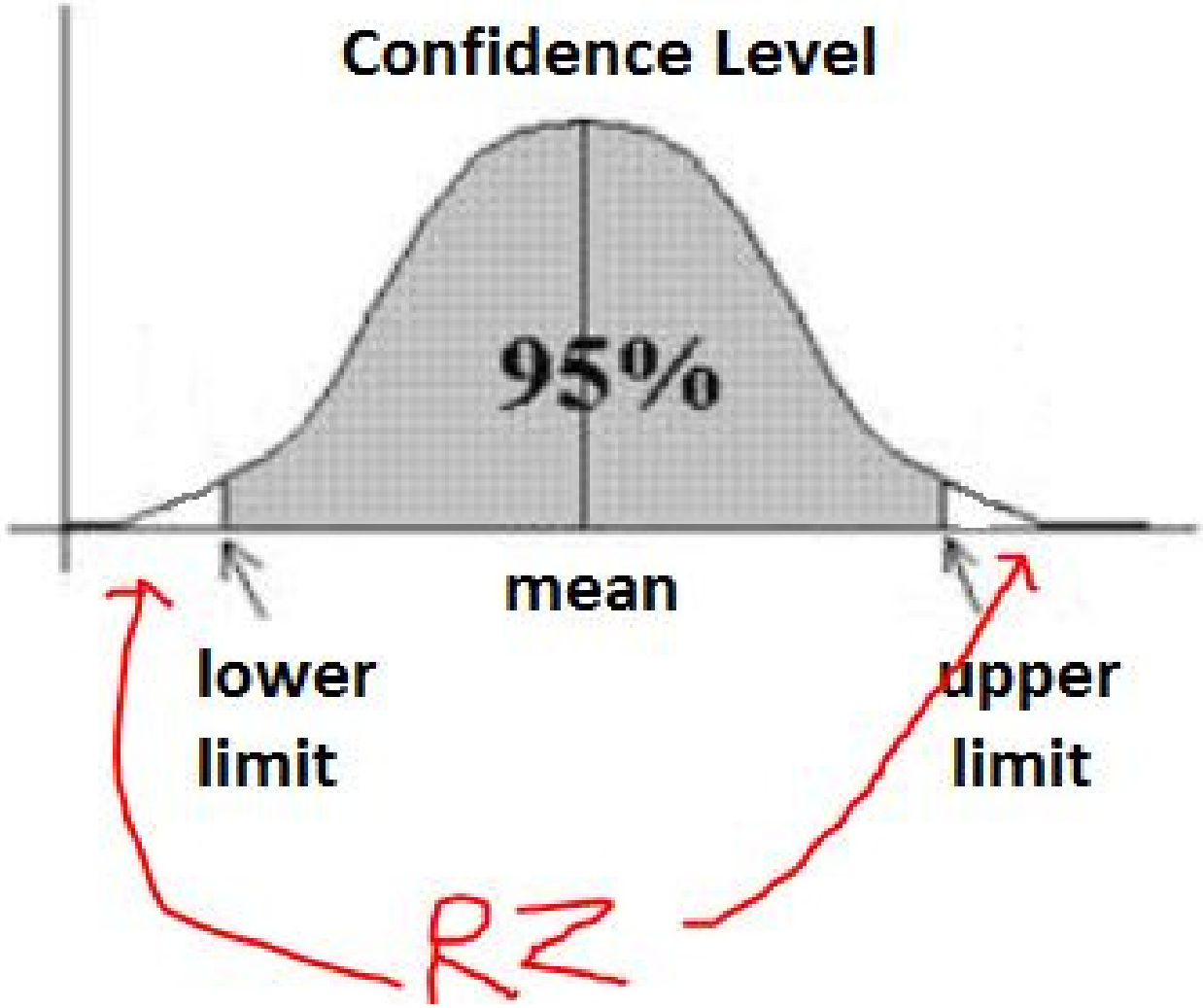
3. We are almost sure (with 95 percent probability) that a general normal random variable takes a value between

$$(\mu - 1.96\sigma,\ \mu + 1.96\sigma) \qquad (\texttt{95 confidence interval}) \tag{24}$$

Rejection Zone

1. The probability that a normal random variable taking a value outside the 95 confidence intervals is only 0.05
2. That is a small probability. So it is unlikely for a normal variable to take those extreme values lying on the two tails
3. In short, tails are "unlikely zone", and hypothesis testing is based on this idea—a hypothesis will be rejected if we end up in the tail part or rejection zone

Confidence Intervals and Rejection Zone (RZ)



Central Limit Theorem (CLT)

1. CLT states that as sample size rises, the sample mean of an iid sample (from any distribution) converges to a normal random variable

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad (as\ n \to \infty) \tag{25}$$

2. The amazing part is, convergence of $\bar{y}$ to normal distribution occurs even though y does not follow normal distribution

3. Do not read too much into CLT: y does not converge to normal distribution. It is $\bar{y}$ that converges to normal distribution when sample rises

4. We get a standard normal distribution after standardizing $\bar{y}$:

$$\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad (as\ n \to \infty) \tag{26}$$

One-Sample T Test

1. The standardized sample mean is nothing but t statistic (t value, t ratio)
2. Under the null hypothesis

$$H_0 : \mu = c \tag{27}$$

t statistic follows a standard normal distribution in large sample ($n > 120$)

$$t - statistic \equiv \frac{\bar{y} - c}{\sigma/\sqrt{n}} = \frac{\bar{y} - c}{se} \sim N(0,1), \quad (when\ n > 120) \tag{28}$$

3. We reject H_0 if the t-statistic is in the tail (rejection zone) and there are three approaches to tell whether that is the case
   (a) (critical value approach): we reject H_0 when $|t| > 1.96$
   (b) (p value approach): we reject H_0 when $2 * P(z > |t|) < 0.05$
   (c) (confidence intervals approach): we reject H_0 when c is outside CI

   The three approaches lead to the same conclusion

R Example 1

```
> t.test(x, mu = 0, alternative = "two.sided")

        One Sample t-test

data:  x
t = 1.4627, df = 999, p-value = 0.1439
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01577881  0.10817513
sample estimates:
 mean of x
0.04619816
```

Remarks

1. We use R function **t.test** to test the null hypothesis $H_0 : \mu = 0$
2. The standard deviation is $\sigma = 0.999$
3. The standard error is $\sigma/\sqrt{n} = 0.999/\sqrt{1000} = 0.03159$
4. The t-statistic is $\frac{\bar{y}-c}{se} = \frac{0.04619816-0}{0.03159} = 1.4627$. We do <u>not</u> reject $H_0 : \mu = 0$ because $|1.4627| < 1.96$ (i.e., t value is not in tail or rejection zone)
5. The p-value is $2P(z > 1.4627) = 2P(z < -1.4627) = 2pnorm(-1.4627) = 0.1439$, greater than 0.05. So we do not reject H_0
6. We do not reject H_0 also because 0 is inside the CI (-0.01577881, 0.10817513)

R Example 2

```
> t.test(x, mu = 1, alternative = "two.sided")

        One Sample t-test

data:  x
t = -30.2, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 -0.01577881  0.10817513
sample estimates:
 mean of x
0.04619816
```

Remarks

1. We can reject $H_0 : \mu = 1$ since t statistic equals -30.2, greater than 1.96 in absolute value. Now the p value is less than 0.05, and confidence intervals do not include hypothesized value 1.
2. Intuitively, we reject $H_0 : \mu = 1$ because the sample mean 0.0462 is far from the hypothesized value 1; we cannot reject $H_0 : \mu = 0$ because the sample mean 0.0462 is close to the hypothesized value 0.
3. We use standard error as yardstick to measure the distance between sample mean and hypothesized value. H_0 is rejected if the gap between sample mean and hypothesized value exceeds $1.96se$ in absolute value