Econometrics 1 *Applied Econometrics with R*

Lecture 6: Review of Statistics

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼1726

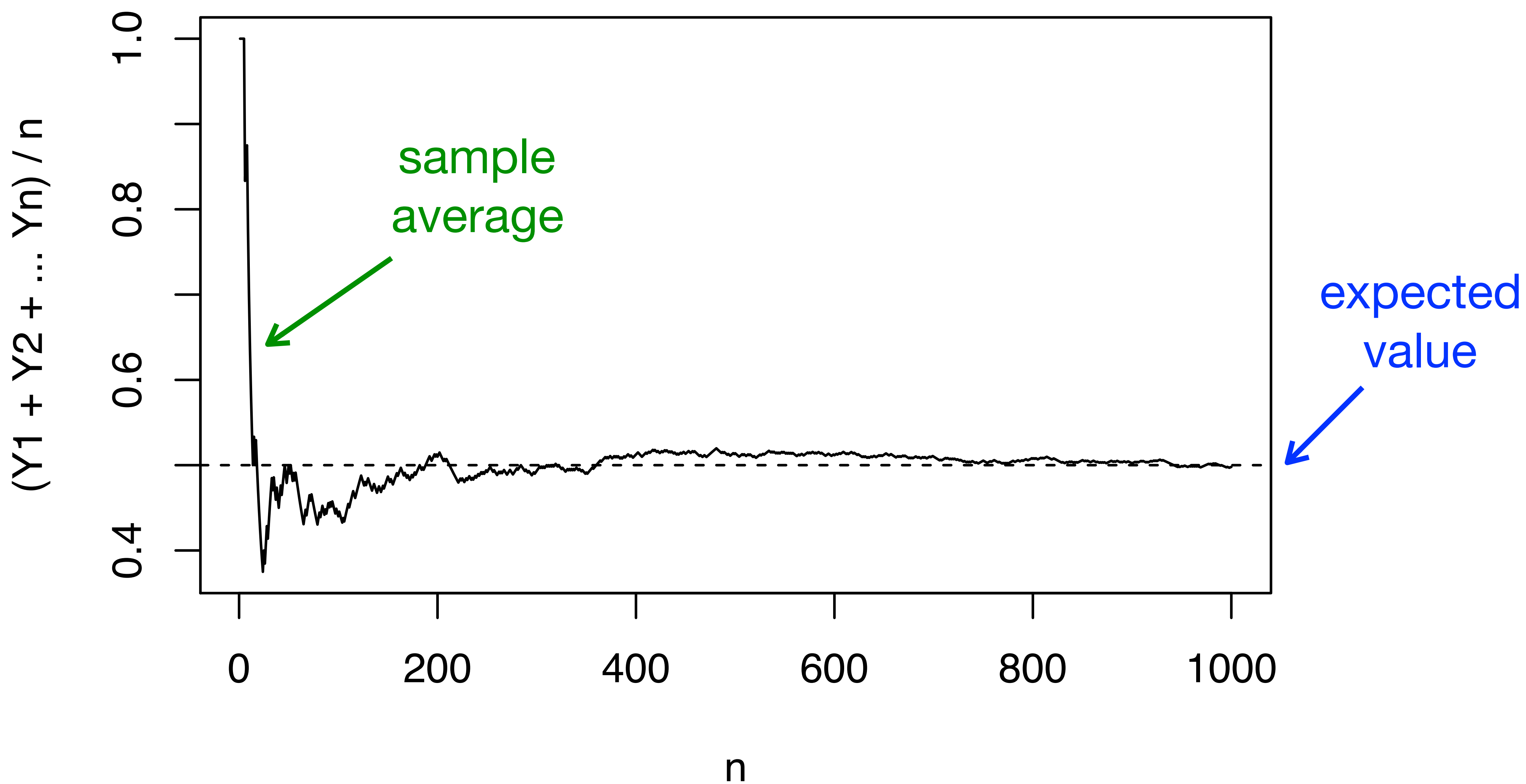
E-mail: huangjp@szu.edu.cn

Tel: (0755) 2695 0548

Office hour: Mon./Tue. 13:00-14:00

Demonstrating the LLN

- The sample mean of n Bernoulli random variables (flipping a fair coin n times).



Basics of Statistics

What does statistics do?

- A typical statistical question:

  What is the mean of the distribution of earnings of recent college graduates?

- It is too expensive, sometimes even impossible, to do an exhaustive survey to know the answer of such questions.

- Though we cannot know the exact answer, we can use statistical methods to reach tentative conclusions — to draw statistical inferences — about characteristics of the full population based on simple *random sampling* data.

What does statistics do?

- Statistics (statistical tools) helps us answer questions about unknown characteristics of distributions in *population* of interest.

- Three most used statistical methods:

  Estimation
  Hypothesis testing
  Confidence intervals

Estimation

Estimation: a "best guess" numerical value

- Suppose you want to know the mean value of Y (that is, μ_Y) in a population.

- You have a sample of n i.i.d. observations Y_1, …, Y_n .

- A natural way of estimate the value of μ_Y is to calculate the sample mean $\overline{Y}$, which is an *estimator* of μ_Y .

- There are many other possible estimators, for example, the first observation Y_1.

Estimator and estimates

- An **estimator** is a *function* of a sample of data to be drawn randomly from a population. Thus it is a random variable.

- An **estimate** is the numerical *value* of the estimator when it is actually computed using data from a specific sample. Thus it is a nonrandom number.

Desirable characteristics of an estimator $\widehat{\mu}_Y$

- *Unbiasedness* — The bias of $\widehat{\mu}_Y$ is $\mathrm{E}(\widehat{\mu}_Y) - \mu_Y$. We say $\widehat{\mu}_Y$ is an unbiased estimator of μ_Y if $\mathrm{E}(\widehat{\mu}_Y) = \mu_Y$.

- *Consistency* — $\widehat{\mu}_Y$ is a consistent estimator of μ_Y if

$$\widehat{\mu}_Y \xrightarrow{p} \mu_Y$$

  as the sample size increases.

- *Efficiency* — Let $\tilde{\mu}_Y$ be another estimator of μ_Y. Suppose both $\tilde{\mu}_Y$ and $\widehat{\mu}_Y$ are unbiased. $\widehat{\mu}_Y$ is more efficient than $\tilde{\mu}_Y$ if $\mathrm{var}(\widehat{\mu}_Y) < \mathrm{var}(\tilde{\mu}_Y)$.

Properties of $\overline{Y}$

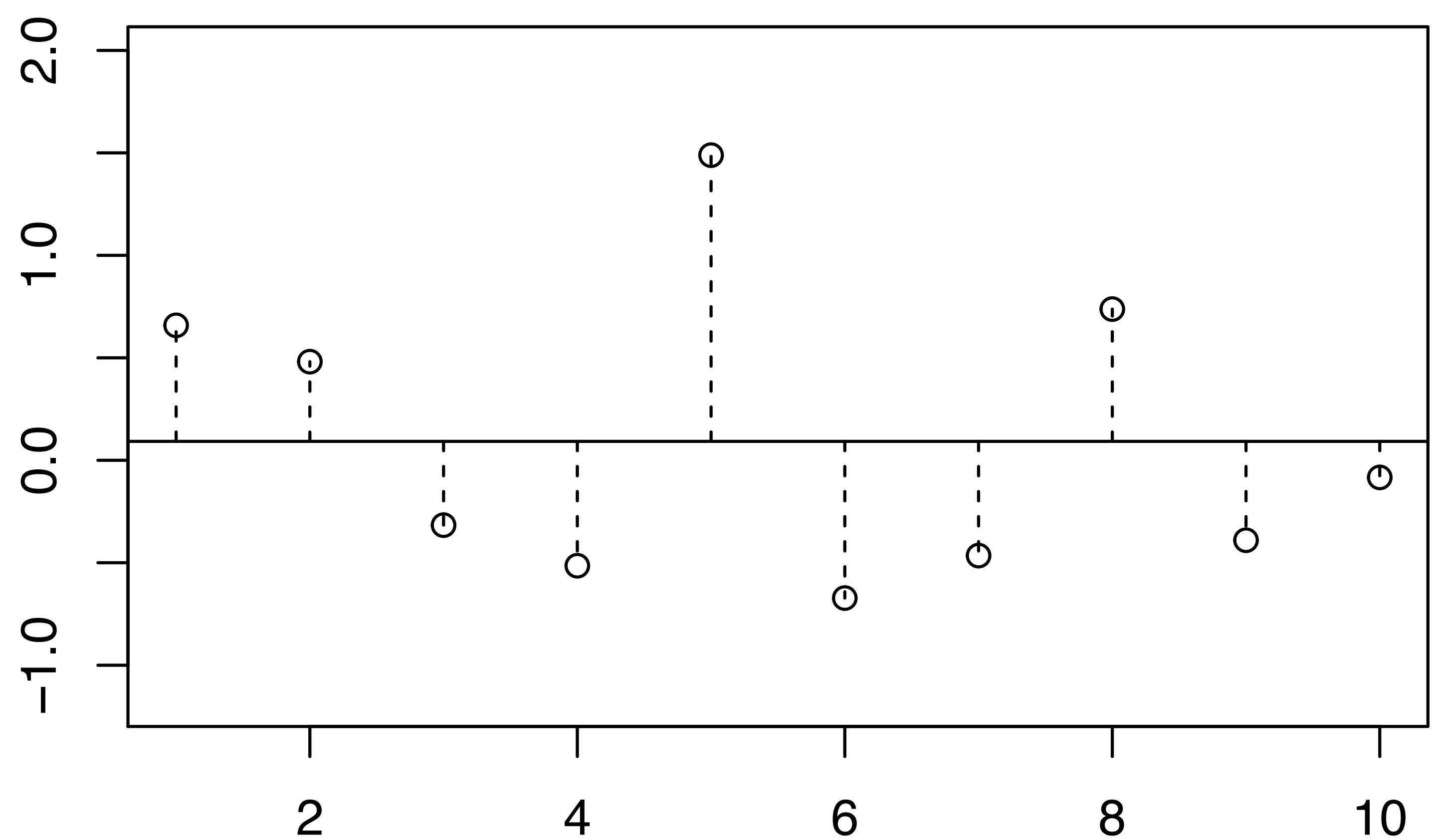
- $\overline{Y}$ is unbiased.
- $\overline{Y}$ is consistent (the law of large numbers).
- $\overline{Y}$ is the **B**est **L**inear **U**nbiased **E**stimator (**BLUE**) of μ_Y. That is, it is the most efficient (best) estimator among all estimators that are unbiased and are linear functions of Y_1, …, Y_n .
- $\overline{Y}$ is the least squares estimator of μ_Y .

Least squares estimator

- The estimator m that minimizes function

$$\sum_{i=1}^{n}(Y_i - m)^2$$

is called the least squares estimator.



Practice

- Write a program to find the least squares estimator of the mean of a standard normal distributed population using 100 random samples, and compare it with $\overline{Y}$.

  Hint:
  1. Generate 100 standard normal random samples.
  2. Make a guess of an interval that the population mean may drop in.
  3. Construct a grid on that interval.
  4. Each value of the grid can be seen as a candidate of the least squares estimator.

```r
y <- rnorm(100)
ybar <- mean(y)

mstep <- 0.0001
mgrid <- seq(-5, 5, mstep)
        # candidates of LSE
msls <- rep(0, length(mgrid))
        # sum of linear squares

for (i in 1:length(mgrid)) {
  msls[i] <- sum((y - mgrid[i])^2)
}

mindex <- which.min(msls)
        # the index of the minimum in "msls"
lse <- mgrid[mindex]
        # the least squares estimator
```

Hypothesis Testing

Hypothesis

- A hypothesis here is a question about the characteristics of the population that can be answered by “yes” or “no”.

- **Null hypothesis** — the specific hypothesis to be tested.

  H_0: Is the population mean equal to 0?

- **Alternative hypothesis** — the hypothesis which is true if the null hypothesis is not.

  H_1: Is the population mean not equal to 0?

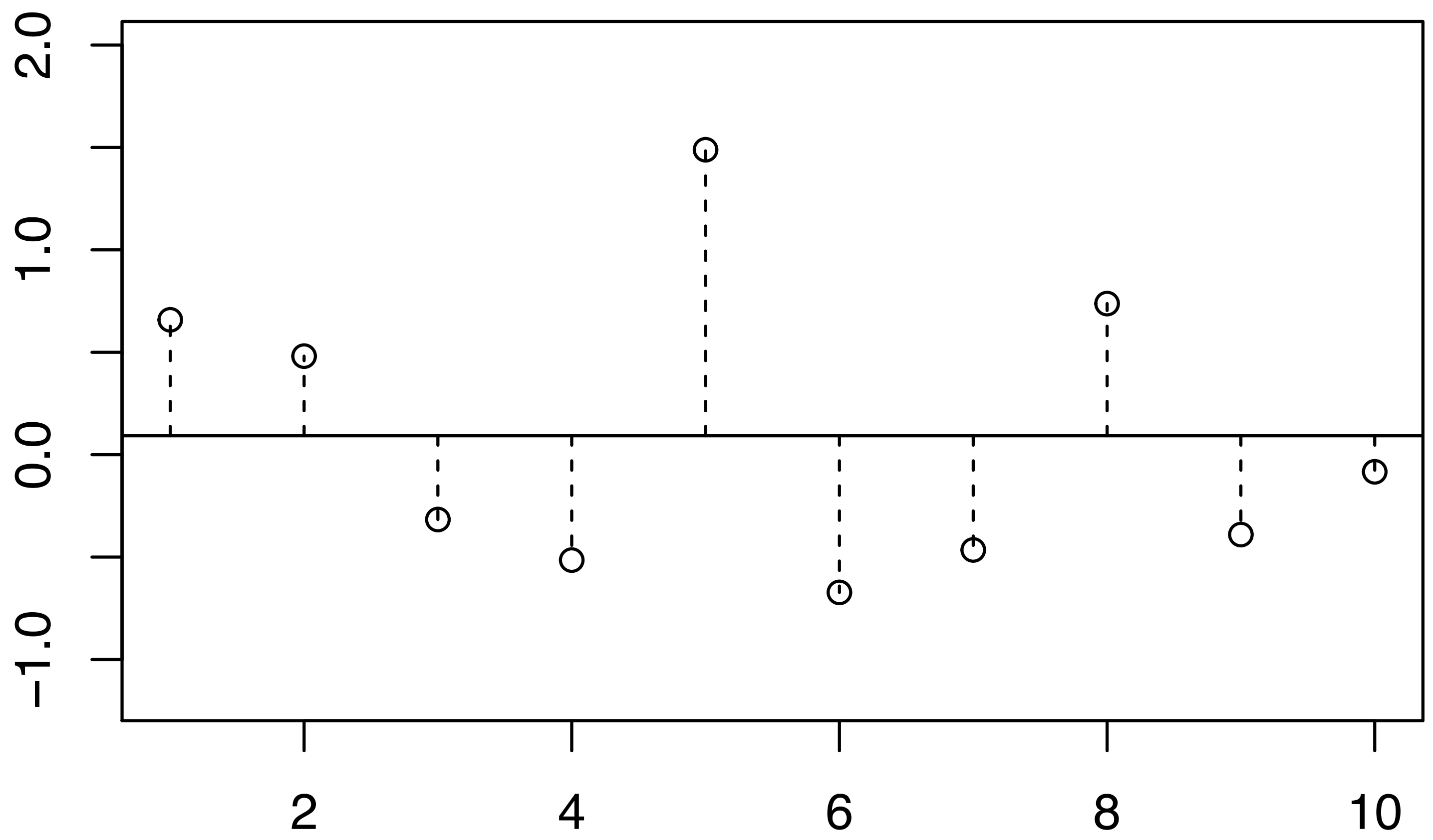
Hypothesis testing

- To either *reject* or *not reject* the null hypothesis based on a random sample of data.
- The null hypothesis is *rejected* if the alternative hypothesis is likely to be true.
- If the alternative hypothesis is not true, it means that the null hypothesis is failed to be rejected by the current sample.

→ Either the null hypothesis is true,
or the evidence is not strong enough to reject it.

Example

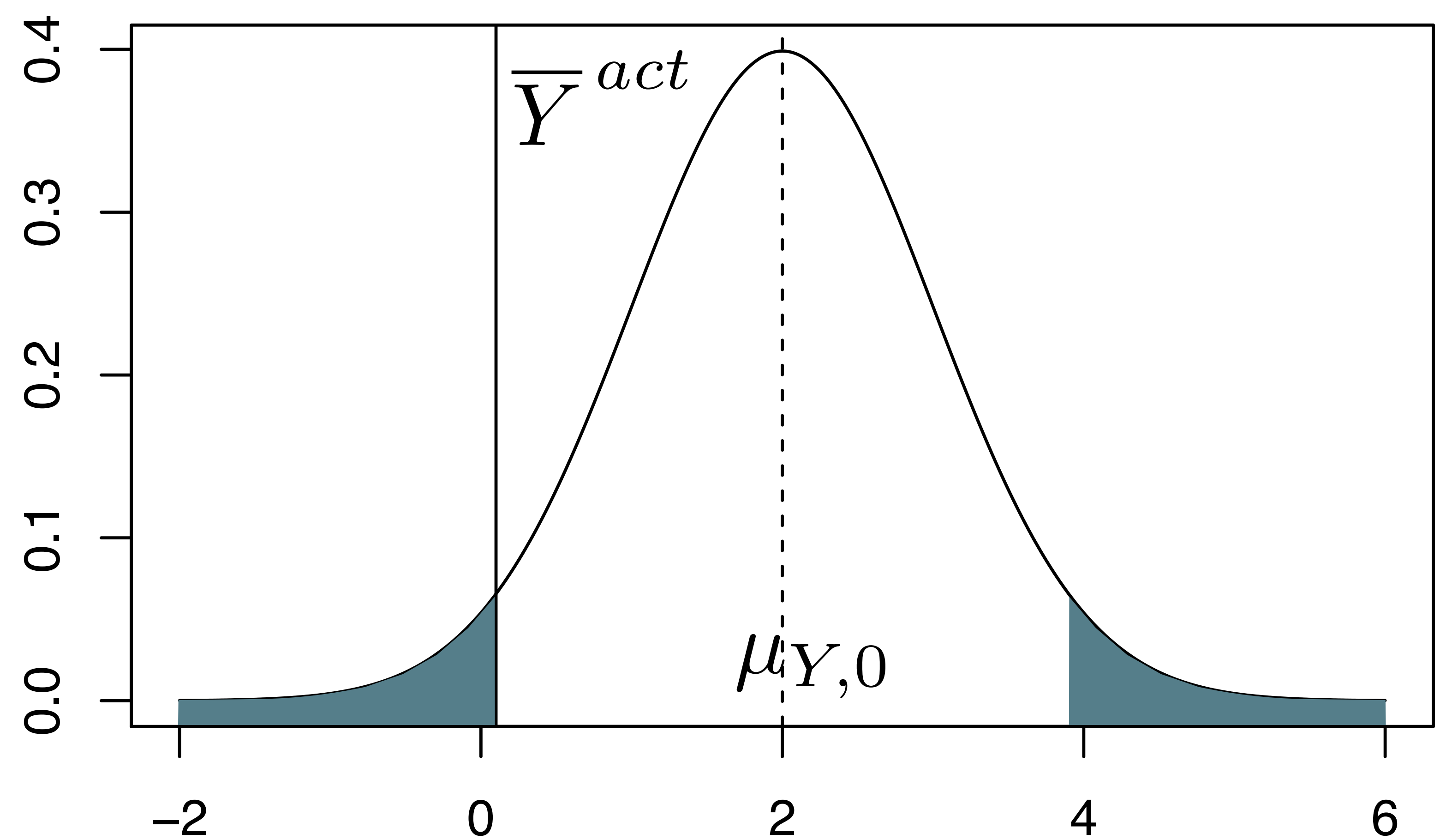
- H_0: $\mathrm{E}(Y) = 2$
- H_1: $\mathrm{E}(Y) \neq 2$



The p-value

- The p-value, also called the *significance probability*, is the probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming the null hypothesis is correct.

$$H_0 : \mathrm{E}(Y) = \mu_{Y,0}$$



The *p*-value

- If the *p*-value is extremely small, it means that our observation is very unlikely to be obtained from the population, assuming the null hypothesis is correct. Thus, the null hypothesis is very likely to be wrong.

- Let $\overline{Y}^{act}$ denote the actual computed sample average of the data set at hand. $H_0 : \mathrm{E}(Y) = \mu_{Y,0}$

$$p\text{-value} = \Pr_{H_0}\left[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\right]$$

- How to evaluate the *p*-value depends on whether the population variance is known.

Calculation the *p*-value when σ_Y is known

- Under the central limit theorem and the null hypothesis,

$$\overline{Y} \sim N(\mu_{Y,0}\,,\, \sigma_Y^2/n)$$

- So the standardized r.v. $\dfrac{\overline{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}$ has a standard normal distribution.

$$p\text{-value} = \Pr_{H_0}\left[\left|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right]$$

$$= 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right)$$

When σ_Y is unknown

- If σ_Y is unknown, we need to estimate it before calculation the *p*-value.

- Sample variance — an estimator of the population variance.

- Sample standard deviation — an estimator of the population standard deviation.

- Standard error of the sample mean — an estimator of of the standard deviation of the sampling distribution of the sample mean.

- The sample variance s_Y^2

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- The sample standard deviation is s_Y

- The standard error of $\overline{Y}$

$$SE(\overline{Y}) = s_Y/\sqrt{n}$$

Calculation the *p*-value when σ_Y is unknown

- We can use s_Y instead of σ_Y

$$p\text{-value} = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right) = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}\right|\right)$$

- The *t*-statistic

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})}$$

Large-sample distribution of the *t*-statistic

- By the central limit theorem, when the sample size n is large, the distribution of t is well approximated by the standard normal distribution $N(0,1)$.

- The *p*-value can be rewritten in terms of the *t*-statistic actually computed:

$$t^{act} = \frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}$$

$$p\text{-value} = 2\Phi\left(-\left|t^{act}\right|\right)$$

General procedure of hypothesis testing for the population mean

1. Specify a null hypothesis and an alternative hypothesis

$$H_0 : \mathrm{E}(Y) = \mu_{Y,0} \qquad H_1 : \mathrm{E}(Y) \neq \mu_{Y,0}$$

2. Calculate the *t*-statistic and the *p*-value

3. Choose a level of significance — e.g. 5%

4. If the *p*-value < 0.05, the null hypothesis is rejected.

   The corresponding critical value to 0.05 is 1.96 under the standard normal distribution, we can also say the null hypothesis is rejected if $|t^{act}| > 1.96$.

Practice

- The R command for this kind of hypothesis testing is `t.test(…)`

- Read the help of `t.test` and learn how to use it.

- Generate a sample of 100 observations that follows a chi-squared distribution with 3 degrees of freedom.

- Test the hypotheses

$$H_0 : \mathrm{E}(Y) = \mu_{Y,0} \qquad H_1 : \mathrm{E}(Y) \neq \mu_{Y,0}$$

where $\mu_{Y,0} = 1.5$. Find the estimate, the *t*-statistic, and the *p*-value. What is your conclusion?

```
> y <- rchisq(100, 3)
> mu0 <- 1.5
> ty <- t.test(y, mu = mu0)
        # with null hypothesis "mu = 0"

> ty

   One Sample t-test

data:  y
t = 4.5992, df = 99, p-value = 1.256e-05
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 1.975775 2.697797
sample estimates:
mean of x
 2.336786
```

The `t.test` command use an exact Student t distribution instead of normal approximation. We address this point later.

Confidence Intervals

Confidence intervals

- What can we say about the population mean without specifying a hypothesis?

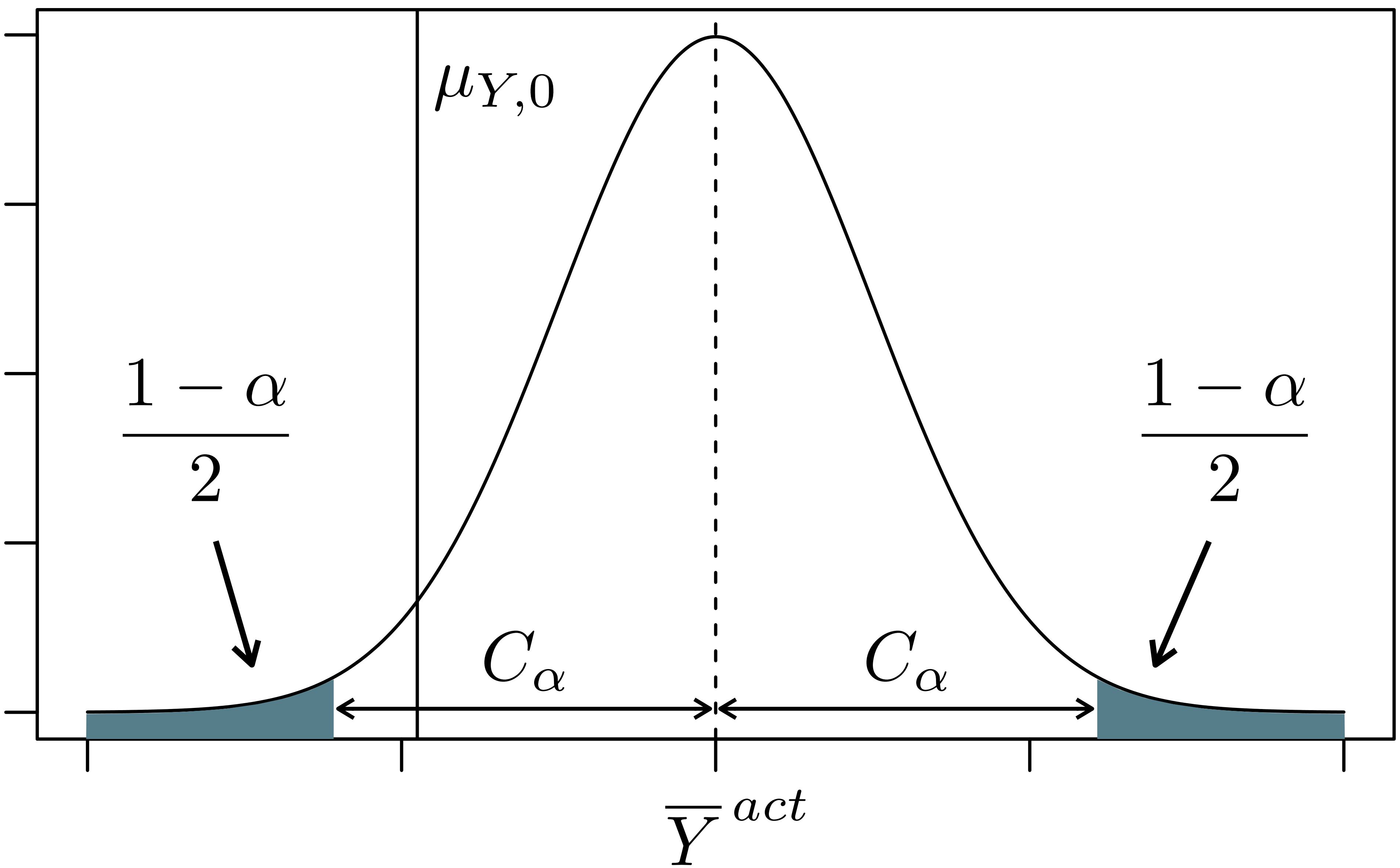
- It is possible to *use data from a random sample* to construct a set of values that contains the true population mean μ_Y with a certain pre-specified probability. This set is called a *confidence set*, and the pre-specified probability is called the *confidence level*.

- The confidence set for μ_Y is an interval, thus it is also called a *confidence interval*.

Confidence interval v.s. *p*-value

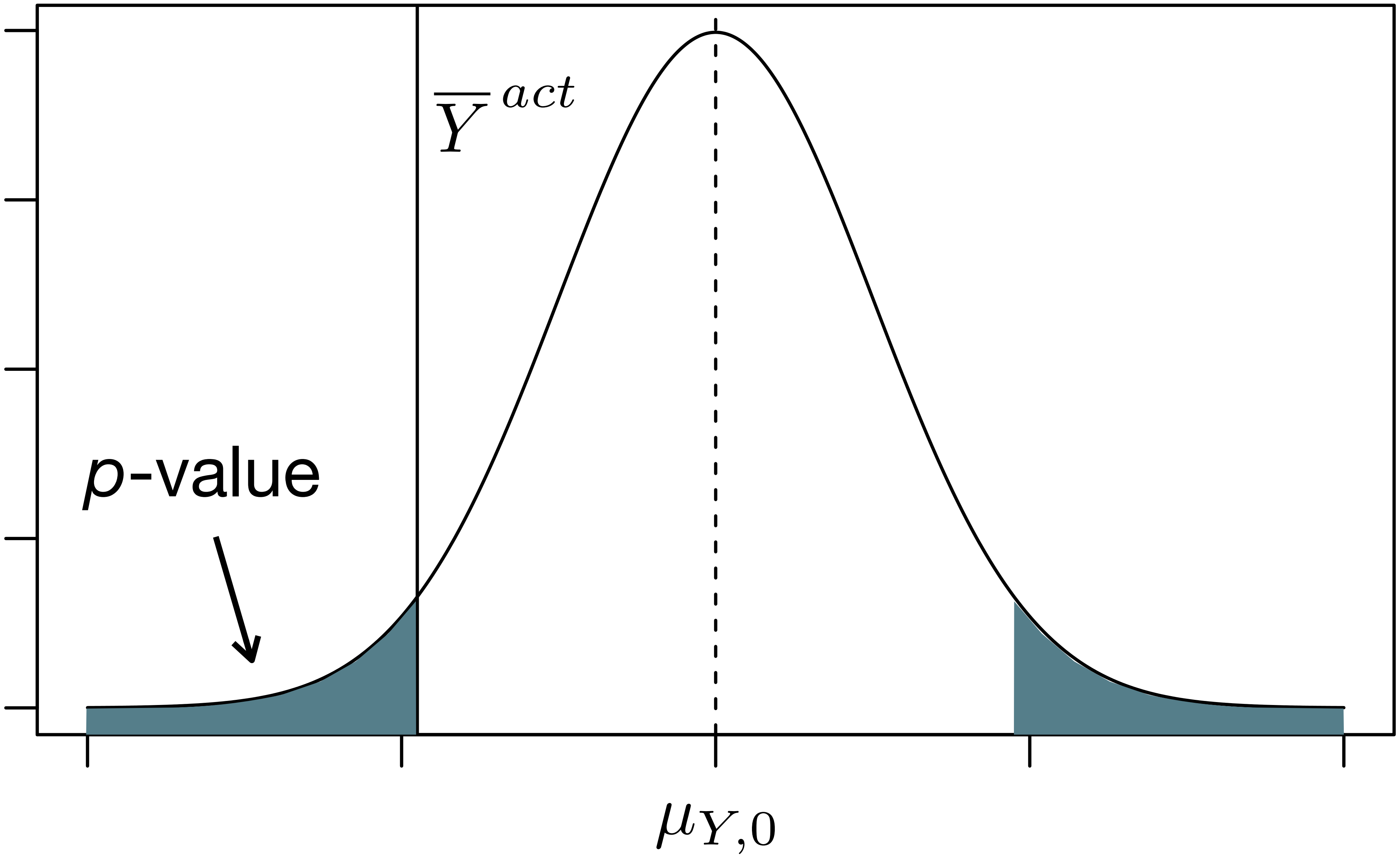
α confidence interval of μ_Y is

$$[\overline{Y} - C_\alpha \times SE(\overline{Y}), \overline{Y} + C_\alpha \times SE(\overline{Y})]$$



Confidence interval v.s. *p*-value

$$p\text{-value} = \Pr_{H_0}\left[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{\,act} - \mu_{Y,0}| \right]$$



Confidence intervals when sample size is large

- α confidence interval of μ_Y

$$[\overline{Y} - C_\alpha \times SE(\overline{Y}), \overline{Y} + C_\alpha \times SE(\overline{Y})]$$

- If the sample size is large, by the central limit theorem, C_α is calculated from the standard normal distribution.

$$C_\alpha = \Phi^{-1}(\alpha + \tfrac{1-\alpha}{2})$$

Confidence intervals when sample size is large

- The 95% confidence interval of μ_Y

$$[\overline{Y} - 1.96SE(\overline{Y}), \overline{Y} + 1.96SE(\overline{Y})]$$

- The 90% confidence interval of μ_Y

$$[\overline{Y} - 1.64SE(\overline{Y}), \overline{Y} + 1.64SE(\overline{Y})]$$

- The 99% confidence interval of μ_Y

$$[\overline{Y} - 2.58SE(\overline{Y}), \overline{Y} + 2.58SE(\overline{Y})]$$

The *p*-value and Confidence Interval when Sample is Small*

The *t*-statistic when sample size is small

- The *t*-statistic

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})}$$

  is approximated by normal distribution using the central limit theorem for large samples.

- If the sample size is small, say $n < 30$, the use of central limit theorem may leads to a poor approximation, and the exact distribution of the *t*-statistic depends on the population distribution of Y, which can be complicated.

The *t*-statistic when sample size is small

- If the population distribution is known to be normal, the *t*-statistic has an exact Student *t* distribution with n - 1 degrees of freedom.

- For small samples with normal population distribution assumed, the hypothesis testing (*p*-value) and confidence intervals (C_α) should be evaluated with the Student *t* distribution.

Practice

- Generate a sample of 25 observations that follows a normal distribution with mean being 1.2 and variance being 4.

- Calculate the 80% confidence interval of the population mean from the generated sample using the exact formula for small samples (*t* distribution).

- Repeat the above quest but use the large sample approximation (normal distribution), compare the result with the previous.

```r
y <- rnorm(25, 1.2, 2)
estimate <- mean(y)
se <- sd(y) / sqrt(length(y))

# using exact t distribution
df_t <- length(y) - 1
confint_t <- c(estimate - qt(0.9, df_t) * se,
                 estimate + qt(0.9, df_t) * se)

# using normal approximation
confint_n <- c(estimate - qnorm(0.9) * se,
                 estimate + qnorm(0.9) * se)
```

References

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.