Advanced Course in Marketing 07
Statistical inference and tests

Takumi Tagashira

Email: takumi.tagashira@r.hit-u.ac.jp

Aim of this session

- It is easy to run statistical tests in R.

- Understanding the meaning of the statistical principles and interpreting results correctly is more valuable than remembering analytical commands.
- This session briefly covers introductory statistics.
- If you haven't studied introductory statistics, I highly recommend studying individually or taking relevant courses.
- See distributed slides for more detailed and precise explanation.

Reference

- OpenIntro Statistics (Introductory)
  - https://www.openintro.org/book/os/
- DeGroot and Schevish “Probability and Statistics”, Pearson.

Agenda

- Introductory statistics (Skip)
  - Random variable, probability distribution, expectation and variance
  - Statistical inference
  - Statistical test (population mean)
- Examples of statistical tests
  - Mean comparison test
  - Analysis of Variance (ANOVA)

Probability

- Probability：
  - Proportion of times an outcome would occur, and it takes values between 0 and 1 (inclusively).

Random variable and distribution

- Random variable:
  - Variable defined on a sample space (a collection or a set of possible outcomes)

  Each possible value (outcome) corresponds to a certain probability.
- Suppose we throw a fair die and let x denote the value of the dice roll.
- x can be considered as a variable that can take the integer from 1 to 6.
  - In this condition, the sample space refers to six sample points corresponding to the dice roll. Each sample point has its probability.

x	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

- A rule that assigns probabilities to corresponded outcomes ($P(x)$ as a function of x) is called probability distribution or probability distribution function.

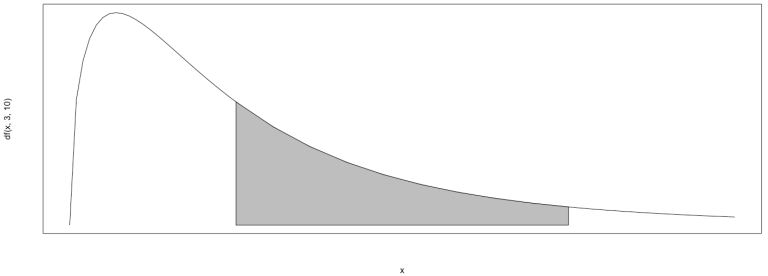
Discrete and continuous random variable

- Discrete random variable:
  - Random variables that only take values in a discretized manner.
  - E.g. dice roll
- Continuous random variable:
  - Random variables that take any values continuously in a particular range.
- We can identify a particular probability for a corresponded value for discrete random variables.
  - E.g. The probability of rolling a 1 on a 6-sided die.
- However, for continuous variables, the probability of taking a particular value is zero.
  - Since there are an infinite number of possible values.
  - If the probability is assigned to each possible value, the sum of the probabilities would be infinite.

Continuous variable and p.d.f.

- The probability is assigned to the interval of possible values in continuous random variables.
- The possible values of continuous correspond to probability density.
  - i.e. The relationship between variable x and the probability density is defined as the probability density function (p.d.f) $f(x)$.
- Suppose that a continuous random variable x has a PDF $f(x)$.
- The probability of x taking an interval [a, b] is as follows.

$$P(a \leq x \leq b) = \int_a^b f(x)\, d$$

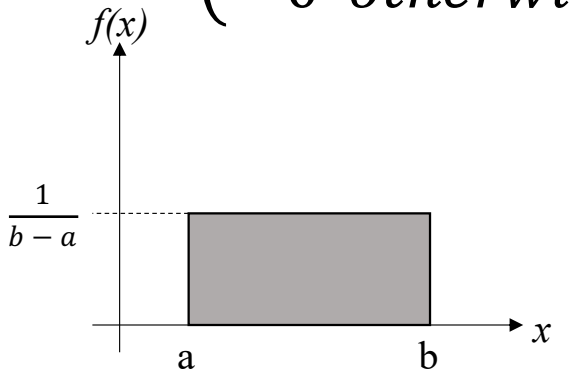


*We can use this relationship to identify a and b from pre-determined $f(x)$.

Uniform distribution

- Suppose x is a random variable that follows uniform distribution between [a, b].
- The p.d.f. of the uniform distribution can be shown as follows:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 \ \ otherwise \end{cases}$$



Example: Uniform distribution

- A uniform distribution with the interval between [-1, 3] is as follows:

$$f(x) = \begin{cases} \frac{1}{4} & -1 \leq x \leq 3 \\ 0 & otherwise \end{cases}$$

- The probability that x takes the value between [0,2] can be defined as:

$$P(0 \leq x \leq 2) = \int_0^2 \frac{1}{4}\, dx = \left[\frac{x}{4}\right]_0^2 = \frac{1}{2} - 0 = \frac{1}{2}$$

Expectation

- Expectation can be considered as theoretical mean value that involves probability arguments.
- Expectation of x $(E(x))$ with probability distribution $P(x)$ can be defined as:
- $E(x) = \sum_x xP(x) = \mu$.
  - What is the expectation of the dice roll?
- Expectation of x $(E(x))$ with PDF $f(x)$ can be defined as:
- $E(x) = \int_{-\infty}^{\infty} xf(x)dx = \mu$.

Example: Uniform distribution

- Consider a uniform distribution with interval [a, b]:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 \;\; otherwise \end{cases}$$

- The expectation of x can be calculated as follows:

$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{a}^{b} x \cdot \frac{1}{b-a}\, dx = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)}$$

$$= \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}$$

Characteristics of expectation and variance

- Expectation:
- Let a denote a constant and $g(x)$ and $h(x)$ are functions of x.
  - $E(a) = a$
  - $E[ag(x)] = aE[g(x)]$
  - $E[g(x) + h(x)] = E[g(x)] + E[h(x)]$
- Thus, variance σ^2 can be defined as follows.
- $\sigma^2 = E[\{x - E(x)\}^2] = E[(x - \mu)^2] = E[x^2] - E[x]^2$
  - i.e. Difference between the expectation of squared value of x and squared value of expectation.

Statistical inference

Population and sample

- When we are interested in a particular group for a study, we usually collect information from a part of the group rather than from the entire group.
  - Whole group：Population
  - A part：Sample
- Statistical analyses infer the characteristics of population from samples based on the principles of probability.
  1. Modelling the population based on probability distribution.
  2. Samples are viewed as random variables that follows the probability distribution.

Parameter

- Parameters refer to values that represent population characteristics
  - E.g. Population mean (μ), variance (σ^2) etc.
- Researchers are interested in parameters.
  - However, parameters are usually unknown and unobservable.
- Thus, we rely on sample information to infer the population characteristics.
  - i.e. statistical inference
- Random sampling procedures are required for statistical inference fundamentally.
  - Random sample: Samples that are extracted in situations where each observation has a **uniformly equal probability** of being extracted.
  - Random samples are identically, independently distributed (iid).

Inference and statistics

- An approach that tries to capture population through the sample information is called **statistical inference**.
- **Statistic** refers to the calculable value (formula) with sample observations.
- An **estimator** is a statistic for statistical inference and the actual calculated result is called an estimated value.

Estimation types

1. Point estimation
   - Using sample data to calculate a ‘single value’ that is to guess or predict an unknown parameter.
   - E.g., Sample mean is a point estimator of population mean (μ).
2. Interval estimation
   - Using sample data to calculate an ‘interval’ that include the unknown parameter with a particular probability.
   - Capturing an interval considering statistical error.
   - E.g. confidence interval of population mean

Point estimate

- Examples of parameters and the corresponding estimators.
  - Population mean (μ): Sample mean ($\bar{X}$)
  - Population variance (σ^2): (Unbiased) sample variance (S^2)
  - Population standard deviation (σ): (Unbiased) sample s.d. (S)

- Does a calculated sample mean represent the true parameter?

Point estimator and error

- We sometimes misunderstand that the calculated estimated value is the sole true value.
- However, estimators are composed of sampled information.
  - There is a gap (error) between estimated value and the unknown parameter.
- E.g. Suppose we collected 100 random samples of Hitotsubashi university students and the average monthly income across the sample was 0.
  - Do you think the result equals to the parameter?
- Estimators that are defined based on random variables are also random variables.

Desirability of estimators

- The reliability of estimators are judged based on several characteristics of estimators:
- An estimator $\hat{\theta}$ is desirable to take a closer value with unknown parameter θ.
- Desirable characteristics:
  - Unbiasedness: $E(\hat{\theta}) = \theta$
  - Consistency: $\hat{\theta}$ gets closer to θ when the sample size is large
  - Efficiency: Variance of $\hat{\theta}$ (smaller, better)
- We will skip the details.

(Sup) Central Limit Theorem

- As the sample size n increases, the distribution of $\bar{X}$ can be approximated by normal distribution $N(\mu, \sigma^2/n)$.
  - Skewness of the original distribution of $\bar{X}$ influences the required sample size for approximation.
- We often apply this theorem to a standardized statistic.

Application of CLT

- Consider the following standardized random variable Z

- $Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$

- Distribution function of Z converges to the distribution function of standard normal distribution $N(0,1)$ when the sample size is sufficiently large.
  - Distribution function of $N(0,1)$ is often written as $\Phi(z)$.
- This helps statistical estimations and tests.

Brief summary

- We infer parameters through estimators.
- However, estimators that are defined based on random variables are random variables.
- Sample mean has desirable features:
  - Unbiasedness: $E(\bar{X}) = \mu$
  - Law of large number: Sample mean converges to population mean with large sample size

Try sampling and estimation

- Let’s see the following features:
  - Sample mean and error
  - Effect of sample size

- Let’s role a die!
- What is the expectation of dice roll?

Dice roll in R

```
install.packages("knitr")

die <- 1:6
d <- sample(die,size=1,replace = TRUE)
d
```

- What was the sample mean?

Dice roll in R II

- Repeat the the trial of rolling a dice ten times, three times.
- You probably obtain different values.
  - You may observe the same values or 3.5, but it was a coincidence.
- Estimators that are defined by random variables are also random variables.

```
set.seed(352)
d1 <- sample(die,size=10,replace = TRUE)
d2 <- sample(die,size=10,replace = TRUE)
d3 <- sample(die,size=10,replace = TRUE)
d_mean <- matrix(c(mean(d1),mean(d2),mean(d3)),nrow = 1)
colnames(d_mean) <- c("d1 mean", "d2 mean", "d3 mean")
knitr::kable(d_mean, caption = ”Comparison of sample means", align = "ccc")
```

Sample mean and sample size

- Law of large number:
  - As the sample size is sufficiently large, the probability that sample mean $\bar{X}$ equals population mean μ converges to one.

```
set.seed(541)
d10 <- sample(die,size=10,replace = TRUE)
d100 <- sample(die,size=100,replace = TRUE)
d1000 <- sample(die,size=1000,replace = TRUE)
d_lln <- matrix(c(mean(d10),mean(d100),mean(d1000)),nrow = 1)
colnames(d_lln) <- c("Mean (10times)", "Mean (100times)", "Mean (1000times)")
knitr::kable(d_lln, caption = "Comparison of sample means II", align = "ccc")
```

Interval estimation

Interval estimation

- Although estimators satisfy unbiasedness, point estimates involve a gap (error) between unknown parameters.
- We want to identify the reliability of the estimated values (e.g., sample mean) concerning the parameter.
- We apply interval estimation to calculate an interval that includes the unknown focal parameter (e.g., the population means) with a certain probability (i.e., confidence level).
- Confidence interval with zz%:
  - Identifying an interval [xx, yy] that satisfies “the interval from xx to yy includes the parameter with the confidence level zz%.”

Confidence interval example

- Consider that you are working at a light bulb manufacturer.
  - Assume that mean length of product lifetime is 1700 (hours).
- A new product model was developed but the mean length of product lifetime is yet unknown.
- The product lifetimes for newer and older models follow normal distribution and the standard deviation is $\sigma = 180$ (hour).

Confidence interval example II

- We randomly picked up 16 samples from new product and observed the length of the lifetime.

  1873 1685 2275 1760 1769 2176 1748 1760
  1994 1473 1715 1771 1784 1684 2038 1850

  - This data can be interpreted as the observations of $n = 16$ random samples $(X_1, \ldots, X_{16})$from normal population $N(\mu, 180^2)$
  - Sample mean: $\bar{X} = 1835$
  - Unbiased sample standard deviation: $s = 200$
- Is the sample mean (1835) sufficiently different from 1700?

Confidence interval in R

- When we conduct t-test (more on the later) , conf.int can be calculated.
- Conf.int: Confidence interval
- 95% interval with the light bulb example is as follows (see slides for details).
- What does confidence interval mean?

```
bulb <- c(1873, 1685, 2275, 1760, 1769, 2176, 1748, 1760, 1994, 1473, 1715,
1771, 1784, 1684, 2038, 1850)

bulb_ci <- t.test(bulb)
bulb_ci$conf.int

attr(,"conf.level")
```

Interpretation of the results

- The results show that the [1728.235, 1941.140] is the 95% confidence interval.
- It means that this interval contains the unknown parameter with the probability of 0.95.
- Thus, the duration of new product lifetime seems longer than the previous product type (1700).

Confidence interval interpretation

- An interpretation of “the probability that the focal parameter is in this interval is 95%” is **inappropriate**.
- It is important to understand that the two ends of the interval, calculated based on the sample mean (random variable), are also random variables.
- Confidence level represents the probability that the computed interval includes the true parameter.
  - Confidence interval captures the **procedure’s** reliability that collects samples and computes the interval.

Sup.) Confidence intervals

- Intuition of (95%) confidence interval
  - Consider a case where the same procedure of taking a sample from a population and constructing a 95% interval of its mean is repeated 100 times. 95% confidence interval implies that the calculated interval will include population mean 95 times out of 100 times.
- e.g. 95% confidence interval (CI) = [5.00, 10.00] means…
  - Suppose we conduct the same procedure (research) 100 times, the calculated confidence interval includes the parameter 95 times.
  - The current result might be wrong (with the probability of 5%).
- Confidence intervals are informative to interpret the strength of the effects or differences.
  - Especially in business field.

Statistical tests

Aims of statistical test

- Statistical tests aim to judge whether the proposed statements (hypotheses) are acceptable.
- Key terms:
  - Hypothesis
  - Test statistic
  - Errors in statistical decisions
  - Significance level

Statistical test

- Steps of statistical test

1. Developing (Null and Alternative) hypotheses.
2. Choosing a statistic (e.g. t or z value) to test the hypothesis.
   - i.e. test statistic
3. Finding critical values based on a specific significance level α.
   - Determining the threshold of the acceptance region (i.e. critical values) and rejection region of the test.
4. Assuming the null hypothesis is correct and checking whether the calculated statistic is in acceptance or rejection region.

Hypotheses in statistical analysis

- Hypotheses that corresponds research questions
  - i.e. (Theoretical or operational) Hypotheses
  - e.g. “Males have higher buying intention than females”

  →Comparing the mean values of males and females.
- Essential hypotheses for statistical tests
- Null hypothesis (H_0):
  - Hypothesis capturing the characteristic of parameter.
  - This hypothesis is the foundation of statistical test (Rejection or Acceptance).
  - e.g. Male’s buying intention equals to female’s.
- Alternative hypothesis:
  - This hypothesis will be supported when H_0 is rejected.
  - e.g. Male’s buying intention does not equal to female’s.

Hypotheses in statistical analysis

- Hypotheses that corresponds research questions
  - i.e. (Theoretical or operational) Hypotheses
  - e.g. “Males have higher buying intention than females”

  →Comparing the mean values of males and females.
- Essential hypotheses for statistical tests
- Null hypothesis (H_0):
  - Hypothesis capturing the characteristic of parameter
  - This hypothesis is the foundation of statistical test (Rejection or Acceptance).
  - e.g. $\mu_m = \mu_f$
- Alternative hypothesis:
  - This hypothesis will be supported when H_0 is rejected.
  - e.g. $\mu_m \neq \mu_f$.

Hypotheses

- The followings are critical points of hypotheses in the marketing research context.

1. The null hypothesis should work as the foundation of the test.
2. Ensure what statement (alternative hypothesis) will be supported once the null hypothesis is rejected.
3. Coherence between null, alternative, and operational hypotheses.

- i.e., Researchers need to employ a research/analytical method that verifies the operational hypotheses based on null/alternative hypotheses.

Population mean test

- We will use the example of the product lifetime example again.

Light bulb again

- Recall the light bulb manufacturer example.
- A new product was developed but the mean length of product lifetime is yet unknown.
  - We want statistically test whether the new product lifetime has been changed from the old version.

- Let μ denote the mean length of lifetime of new light bulb and we can develop following two hypotheses.
  - H0: The new product lifetime is not different from previous one.
  - H1: The new product lifetime is different from previous one.

Population mean test

- Recall the light bulb manufacturer example.
- A new product was developed but the mean length of product lifetime is yet unknown.
  - We want statistically test whether the new product lifetime has been changed from the old version.

- Let μ denote the mean length of lifetime of new light bulb and we can develop following two hypotheses.
  - H0: $\mu = 1700$
  - H1: $\mu \neq 1700$

Test of population mean

- We randomly picked up 16 samples from new product and observed the length of the lifetime.

1873 1685 2275 1760 1769 2176 1748 1760
1994 1473 1715 1771 1784 1684 2038 1850

- This data can be interpreted as the observations of $n = 16$ random samples $(X_1, \ldots, X_{16})$from normal population $N(\mu, 180^2)$
  - Sample mean: $\bar{X} = 1835$
  - Unbiased sample standard deviation: $s = 200$
- Is the sample mean (1835) sufficiently different from 1700?
- If so, we will reject H_0. We will accept H_0, otherwise.

Let R analyze it.

```
##t-test (e.g., mean = 1700)
bulb <- c(1873, 1685, 2275, 1760, 1769, 2176, 1748, 1760, 1994, 1473, 1715,
1771, 1784, 1684, 2038, 1850)

t.test(bulb, alternative = "two.sided", mu = 1700)
```

Interpretation of the results

```
> t.test(bulb, alternative = "two.sided", mu = 1700)

        One Sample t-test

data:  bulb
t = 2.6968, df = 15, p-value = 0.01657
alternative hypothesis: true mean is not equal to 1700
95 percent confidence interval:
 1728.235 1941.140
sample estimates:
mean of x
 1834.688
```

t-value

P-value is less than 10% significance level.
H0 (μ=1700) was rejected.

Confidence interval

Practical (intuitive) interpretation

- Very intuitive interpretation approach:

1. Define a test approach that satisfies the coherence between alternative hypothesis and research (operational) hypothesis.
2. Check the estimated results and p-value.
3. We can reject the null hypothesis when p-value is small (e.g., lower than 0.05 or 0.01).
4. We can conclude that the alternative (operational) hypothesis is supported.

What is p-value?

- When we conduct statistical tests in R (or other statistical software), we obtain "**p-value**".
  - P-value and significance level are slightly different.
- Type 1 error is a key concept to understand the meaning of p-value.

Decision errors

- There are two types of errors in statistical tests.
  - Type 1 and type 2 error

	H_0 is True	H_0 is false
Rejecting H_0	Type 1 error	✓
Accepting H_0	✓	Type 2 error

Decision errors and significance level

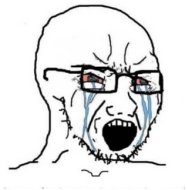
- Type 1 error
  - Rejecting the null hypothesis when H_0 is true.
  - e.g. Considering a medicine is effective while it is not actually.
- Type 2 error
  - Accepting the null hypothesis when the alternative is true.
  - e.g. Considering a medicine is ineffective while it is.

- Significance level (α)
  - The probability that type 1 error occurs.

Intuition of statistical test

- Considering the risk of type 1 error vs. type 2 error,
- We put more focuses on Type 1 error.
- Significance level α represents the probability that type 1 error occurs.

- Statistical tests are interpreted as a process based on the proof of contradiction with using observed data with admitting the probability of type 1 error (α).



Then, what is p-value?

- p-value is the significance level (i.e., probability that type 1 error occurs) when we employ the calculated value of statistics as the critical value.
- More precisely, the relationship can be explained based on the rejection region and the significance level.
  - E.g. There are cases where the null hypothesis can be rejected with the 5% level, but it cannot be with the 1% level.
  - Based on the computed test statistic value, we can find the limit of the significance level, where the null hypothesis cannot be rejected if the significance level is reduced anymore. This limit is called the p-value.

Interpretation of the results

- When we reject the null hypothesis, we can propose some deterministic conclusion supporting the alternative hypothesis.
  - Sample calculation of the test statistic shows that we can conclude the null hypothesis condition rarely occurs while admitting the error probability.
- On the other hand;
- When we cannot reject the null hypothesis, we cannot say that the null hypothesis condition is proven.
- The conclusion should be, “We cannot say the parameter is not μ_0.”
  - Consider the probability of type 2 error.
  - The significance level focuses on the probability of type 1 error, and the test procedure doesn’t focus on the probability that we accept the null hypothesis when the alternative is true.
- Let β denote the probability of type 2 error occurs and $1-\beta$ is called the test power.
  - Test power：The probability of rejecting H0 when H1 is true.

Test of population mean

- Intuition of test statistic of population mean is as follows

- Test statistic= $\frac{Point\ estimate - Null\ value}{Std.Dev.(or\ error)}$

- When we know the population variance, we can assume that test statistic follows standard normal distribution (Z).
- When we don't know the population variance, we can use a estimator of standard deviation (standard error) and assumes t-distribution with (n-1) degree of freedom.

Summary of mean test

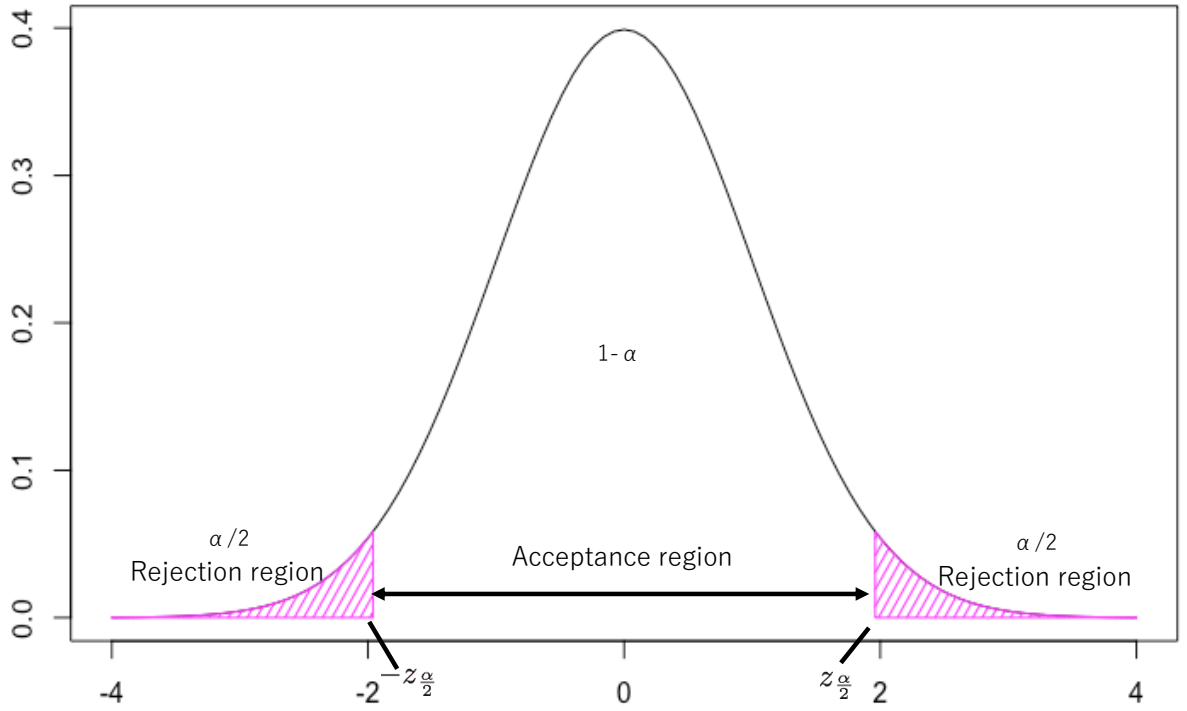
- Let $X_1,\ldots,X_n$ and μ_0 denote n random samples from normal population $N(\mu,\sigma^2)$ and the value of parameter assuming the null hypothesis, respectively.

$$H_0 : \ \mu = \mu_0, \quad H_1 : \ \mu \neq \mu_0$$

- The test statistic: $Z = \frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}}$
- When $|Z|$ is larger than a specific value c (i.e. sufficiently different from 0), we will reject the null hypothesis H0.
  - Where c is called critical value and determined by statistical calculation.
- $|Z| > c \rightarrow$ reject H_0
- $|Z| \leq c \rightarrow$ accept H_0
- When population variance (σ^2) is unknown we assume t-distribution for tests (i.e., t-test)

Two-tailed test

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$



Exercise (assignment)

- Conduct statistical tests and calculate confidence interval of “monetary” in the “idpos_customer.csv” dataset.
- Execute analysis and answer the following questions regarding “monetary” through quiz tab in manaba.
  1. Test whether the mean of monetary is 15,000 (yen) the unknown variance, and report sample mean, t-value, p-value, and your conclusion.
  2. Calculate a 95% confidence interval of monetary with t-distribution and report the results and the interpretation.
- Deadline: 10 October
  - By 23:59, 9 October