MSBA 7002

Business Analytics

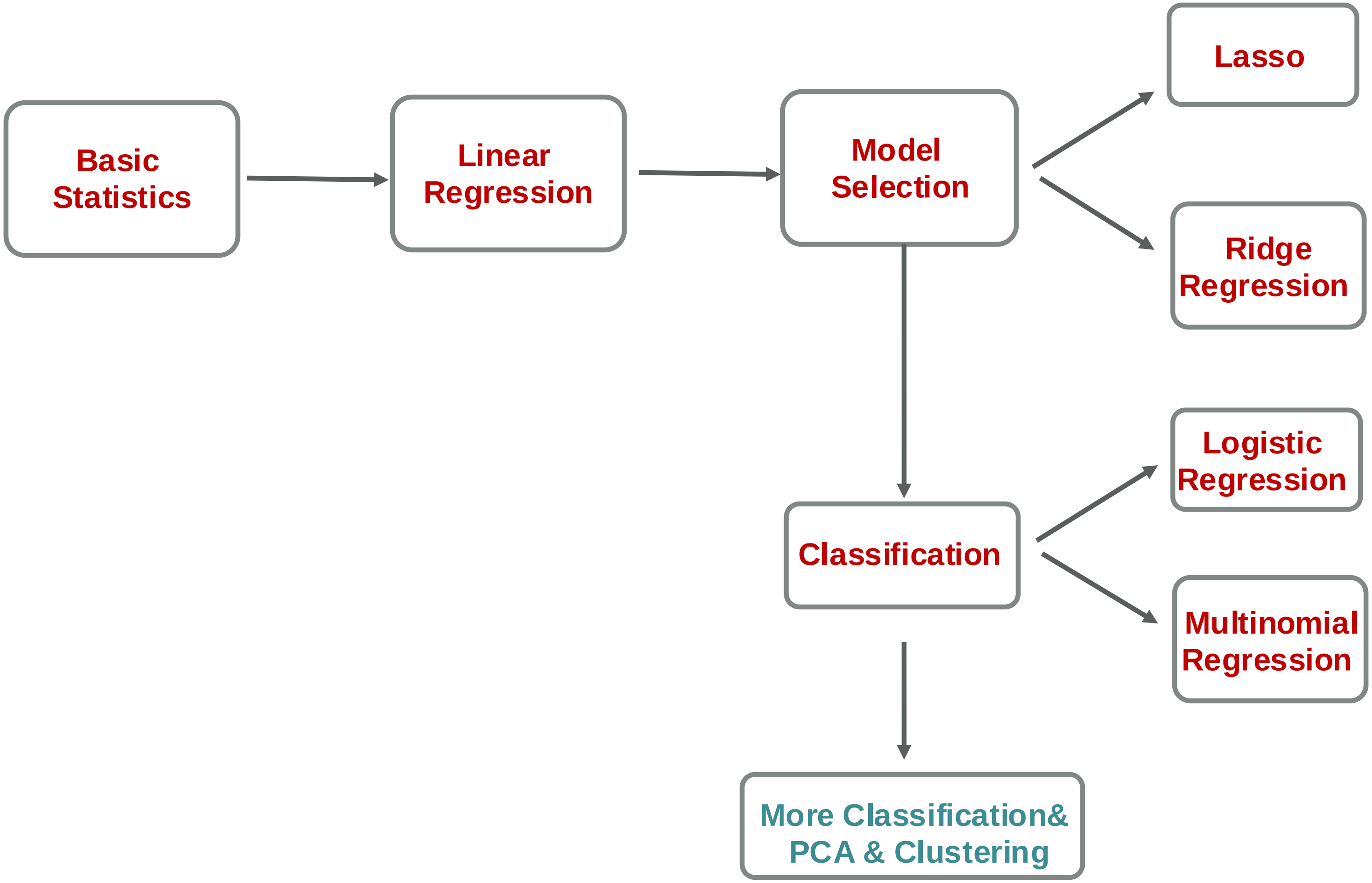
Zhanrui Cai

Assistant Professor in Analytics and Innovation
Faculty of Business and Economics
University of Hong Kong

Business Analytics

- Turn data into information/value.
  - Business managers need to make decisions.
  - They need to make the most informed decisions and generate value.
- Decision-making under uncertainty.
  - Most of the decisions are based on guesses, rather than “facts”.
  - How to make the “best” guess possible as well as how to measure the accuracy of their guesses.

Course structure



Basic Statistics

Linear Regression

Model Selection

Lasso

Ridge Regression

Classification

Logistic Regression

Multinomial Regression

More Classification& PCA & Clustering

Course structure by topics

- Basic Statistics
  - Confidence interval
  - Hypothesis testing
- Supervised Learning
  - Linear regression (with model selection)
  - Logistic regression
  - Multinomial regression
  - Poisson regression
  - Linear discriminant analysis
  - Support vector machine
- Unsupervised Learning
  - Principal component analysis
  - Clustering

Basic Statistics

- Confidence interval
- Hypothesis testing

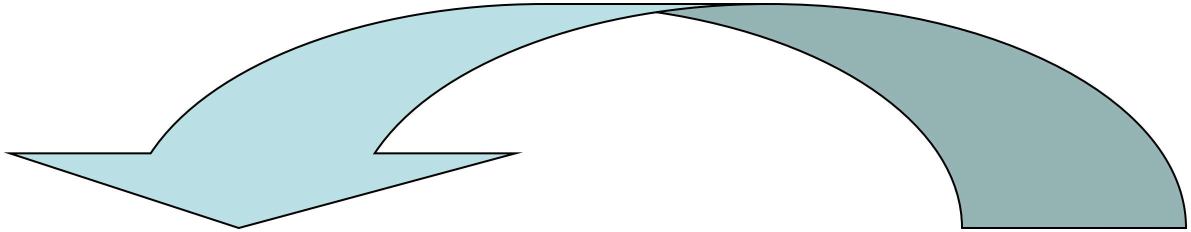
Why do we want to learn statistics?

Why do we want to learn statistics?

- Statistics teaches us how to quantify uncertainty.

The Population-Sample Paradigm

Statistical Inference



Population

Sample

Fixed,
But unknown

- Population summary
  - Pop. Mean (μ)
  - Pop. SD (σ)
  - Pop. Proportion (p)
- Parameters

- Sample summary
  - Sample Mean ($\overline{X}$)
  - Sample SD (s)
  - Sample Proportion ($\hat{p}$)
- Statistics

Random &
uncertain

Sample statistics are guesses for population parameters, but they are unlikely to be exactly correct, so we need to quantify the *error*.

Point Estimation

- A point estimator draws inferences about a population by estimating the unknown parameter using a single value.

- For example, using the sample mean $\bar{X}$ to estimate the population mean μ: height, weight, etc.

- For example, using the sample proportion $\hat{p}$ to estimate the population proportion p: political polls, market popularity, etc.

Confidence Intervals

Why a point estimate is not enough?

- Drawbacks:

1. It is almost certain that the estimate will be wrong based on a single sample

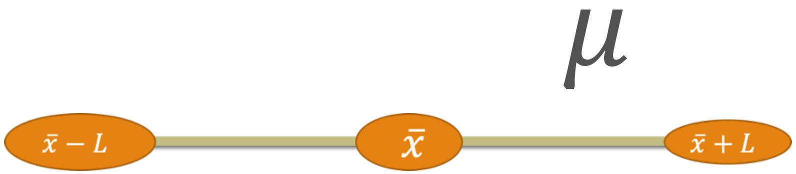
2. What if we want to know how close this estimator is to the true parameter μ (mean) and p (proportion)?

3. Intuitively, larger samples will produce more accurate results, but point estimators alone does not fully reflect the effect of large sample size

Confidence Interval

- Most often, it is very informative to say
  - I don’t know exactly what the mean is, but I am fairly confident that it is between XXX and YYY.
  - For example, I don’t really know what the mean is, but I am almost certain that it is between 814.90 and 815.08.

- This is a *Confidence Interval*
  - Much more informative and realistic than just stating that ``I estimate the mean to be 814.99.”

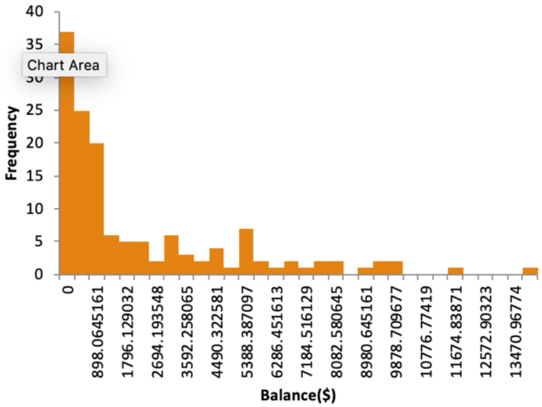


Motivating Example

- To launch an affinity credit card, the contemplated launch process proposes sending pre-approved applications to N= 100,000 alumni of a large university (population)
- Two parameters of the population determine whether the card will be profitable:
  - p, the proportion who will return the application
  - μ, the average monthly balance carried by those who accept the card

- To estimate the parameters, the credit card issuer sent pre-approved application to a sample of 1,000 alumni. Of these, 140 accepted the offer and received a card

Number of offers	1000
Number accepted	140
Proportion who accepted	$\hat{p} = 0.14$
Average balance	$\overline{x} = 1990.5$
SD of balance	$s = 2833.33$



Computing the Interval

- Because $\bar{X}$ follows approximately normal distribution,

$$\mathrm{P}(\mu - 2SE(\bar{X}) \leq \bar{X} \leq \mu + 2SE(\bar{X})) \approx 0.95$$

- Alternatively,

$$\mathrm{P}(\bar{X} - 2SE(\bar{X}) \leq \mu \leq \bar{X} + 2SE(\bar{X})) \approx 0.95$$

- Suppose we take a sample and compute $\bar{X}$ and $SE(\bar{X})$. Then, we can state that μ is somewhere between

$$\bar{X} - 2SE(\bar{X}) \text{ and } \bar{X} + 2SE(\bar{X}),$$

- and be (approximately) 95% sure that we are correct.

General Confidence Interval

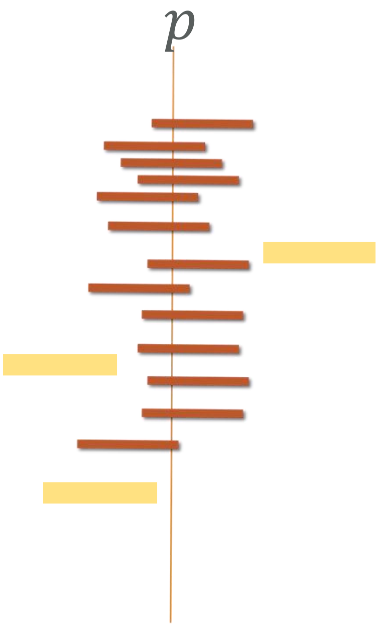
- In general, the confidence interval will be

$$[\bar{X} - z * SE(\bar{X}),\ \bar{X} + z * SE(\bar{X})]$$

- where z determines how sure we are of being correct.
- Some common choices for z:
  - 90% interval: **z=1.645**
  - 95% interval: **z=1.96**
  - 99% interval: **z=2.57**
- Note the trade-off between the size of interval and probability of being correct
  - Margin of error: $z * SE(\bar{X})$, half width of the interval

Statistical Interpretation

- We say “we are 95% confident that…”
- But NEVER say “the probability that the true proportion p is between 12% and 16% is 0.95”
  - For a realized confidence interval, the true mean p is either in there or not
- Statistically, it really means: if you line up the 95% confidence intervals from many, many samples, 95% of these intervals would cover the population parameter p



Back to the Credit Card Example

- The monthly balance for the $n = 140$ customers:
  - $\bar{x} = 1990.5$ and $s = 2833.33$
- Since n>30, we treat the estimated standard deviation as the true population standard deviation.
- The Confidence Interval formula is $\bar{x} \pm z * \frac{s}{\sqrt{n}}$
- The 95% CI for the balance is:

Confidence Interval for the Population Proportion

- A 100(1-*α*)% normal confidence interval for *p* is
- $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- The 95% confidence interval for the proportion of people who accepts the credit card is:

Margin of Error

- $L = z\frac{\sigma}{\sqrt{n}}$,or $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$…,
- Three factors affect the margin of error
  - Confidence level: 90%, 95%, 99%
  - Sample size: n
  - The population variation or the sample variation

Q & As

- For the same sample, the width of a confidence interval will be

A. Narrower for 99% confidence than 95% confidence

B. Wider for a sample size of 100 than for a sample size of 50

C. Narrower for 90% confidence than 95% confidence

Q & As

- With the same confidence level, as standard deviation increases, samples size need to ________ to achieve a specified margin of error

  A. Increase

  B. Decrease

  C. Remains the same

Confidence Interval

- Provides range of values based on observations from one sample
- Stated in terms of confidence level
- Examples: 90% CI for the μ, 95% CI for the p

Hypothesis Testing

Procedures for Statistical Inferences

- Point estimation
  - ``I think the population parameter is XXX.”
- Confidence interval
  - ``I am 95% confident that the population mean is between XXX and YYY.”
- Hypothesis testing
  - ``I think your claim that no more than 10% of the clicks are fraudulent is not valid.”
  - Using statistics to decide which of two possibilities is the truth given imperfect information
  - Hypothesis: a statement/claim/belief about the parameter
    - Two hypotheses: two possibilities (complement of each other)

Realty Agency Expansion

A realty agency manages rental properties, and is considering expanding into the Denver metropolitan area.

To justify the costs of opening a new office, the agency needs rents in the area to be more than $500 per month.

Lower rental generates smaller fees that would make the office unprofitable.

Are rents in Denver high enough to justify the cost of the move?

Realty Agency Expansion

Managerial decision: Expand vs. Don't expand

Population: all rental properties in the area, mean rent μ

- Is $\mu > 500$?

Data collection: 45 houses, sample mean $\bar{X}$ = $647.33

The data suggests that the mean rent μ is above $500.

Hence Expand!

(Wait! Is the evidence strong enough?)

Concepts of Hypothesis Testing

- Two hypotheses:
  - H_a: the alternative hypothesis
    - The statement we hope or suspect is true.

    $H_a : \mu > \$500$ (one-sided alternative)
  - H_0: the null hypothesis
    - The statement of “no effect” or “no difference”.
    - The statement we try to find evidence against.

    $H_0 : \mu = \$500$
- Usually one would decide on H_a first
  - Sometimes, make sense to consider $H_a : \mu \neq \$500$ (two-sided alternative)

Basic Ideas of Hypothesis Testing

To “prove” or “establish” some claim statistically based on the data collected in a sample

- Prove by contradiction

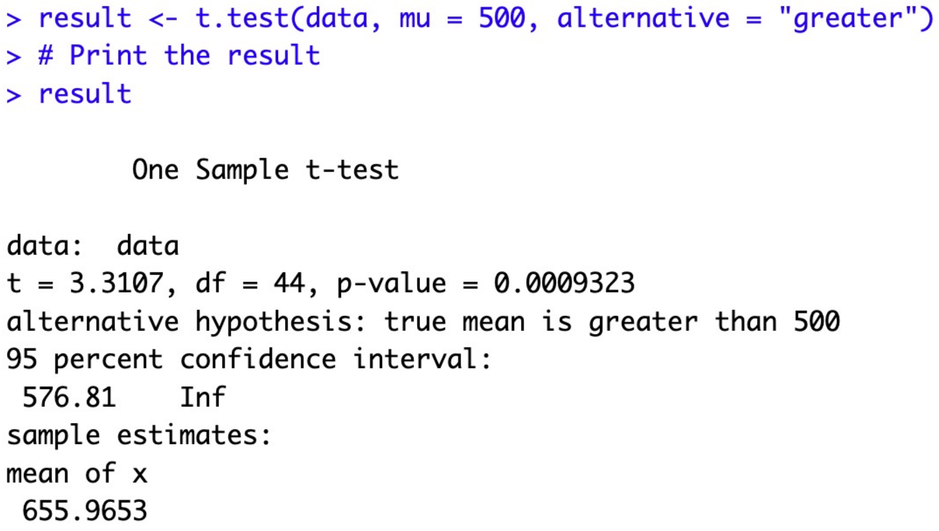
- Assume that the *opposite* is true
  - This “opposite” is your Null Hypothesis H_0

- Given that H_0 is true, calculate the probability for you to see what you ``saw” (i.e. the data)
  - This probability is called the *p-value*, which is the probability of seeing ``data this unusual” due to random chance alone.

- If the p-value is very small, then probably what you assume – H_0 – is incorrect.
  - Then, reject H_0 and conclude that what you claim is true.
  - Otherwise, can not reject H_0 and the data do not support your claim.

Denver Rent: One-sample T Test

- Test H_0: $\mu = 500$ versus H_a: $\mu > 500$



```
> result <- t.test(data, mu = 500, alternative = "greater")
> # Print the result
> result

	One Sample t-test

data:  data
t = 3.3107, df = 44, p-value = 0.0009323
alternative hypothesis: true mean is greater than 500
95 percent confidence interval:
 576.81    Inf
sample estimates:
mean of x
 655.9653
```

- The p-value is 0.009 or (1 out of 111)!
  - In other words, if the mean is not above $500, you would need to collect 111 such samples, before you can see the sample mean this high above $500!
- We are convinced that $\mu > 500$! (Are we 100% correct?)

P-value

- The probability of observing *the data at least this unusual as what we saw*, assuming that H_0 is true.
  - In the direction of the alternative
- The amount of statistical evidence that supports H_0
  - The smaller the *p-value*, the less evidence for H_0, or the more evidence for H_a

❑ Small p-values indicate:

- false H_0 or something rare happened; statistical practice is to believe in the former, hence reject H_0

Rent: P(observing $\bar{x} > \$647$) = .0009

- If μ were \$500, observing $\bar{x} > \$647$ would occur in only 1 out of 1111 samples!
- So we reject H_0. (There is risk of making Type I Error! But Prob<0.0009)

Test Statistic

- A test is based on a statistic, which estimates the parameter that appears in the hypotheses
  - Point estimate

- Values of the estimate far from the parameter value in H_0 give evidence against H_0.
- H_a determines which direction will be counted as far from the parameter value.

- Commonly, the test statistic has the form

  T=(estimate-hypothesized value)/(standard deviation of the estimate)

One-Sample T Test: Test Statistic

- Hypothesized value μ_0 for the mean parameter.
- i.e. H_0: $\mu = \mu_0$.
- Estimate $\bar{X}$ with observed value $\bar{x}$, and estimated standard deviation $s/\sqrt{n}$

- Test statistics

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

One-Sample T Test: p-value

- State null and alternative hypothesis

$$H_0: \mu = \mu_0 \quad vs. \quad H_a: \quad \begin{array}{l} \mu \neq \mu_0 \\ \mu > \mu_0 \\ \mu < \mu_0 \end{array}$$

- p-value equals, assuming H_0 holds

$$2\mathrm{P}(T \geq |t|)$$

$$\mathrm{P}(T \geq t)$$

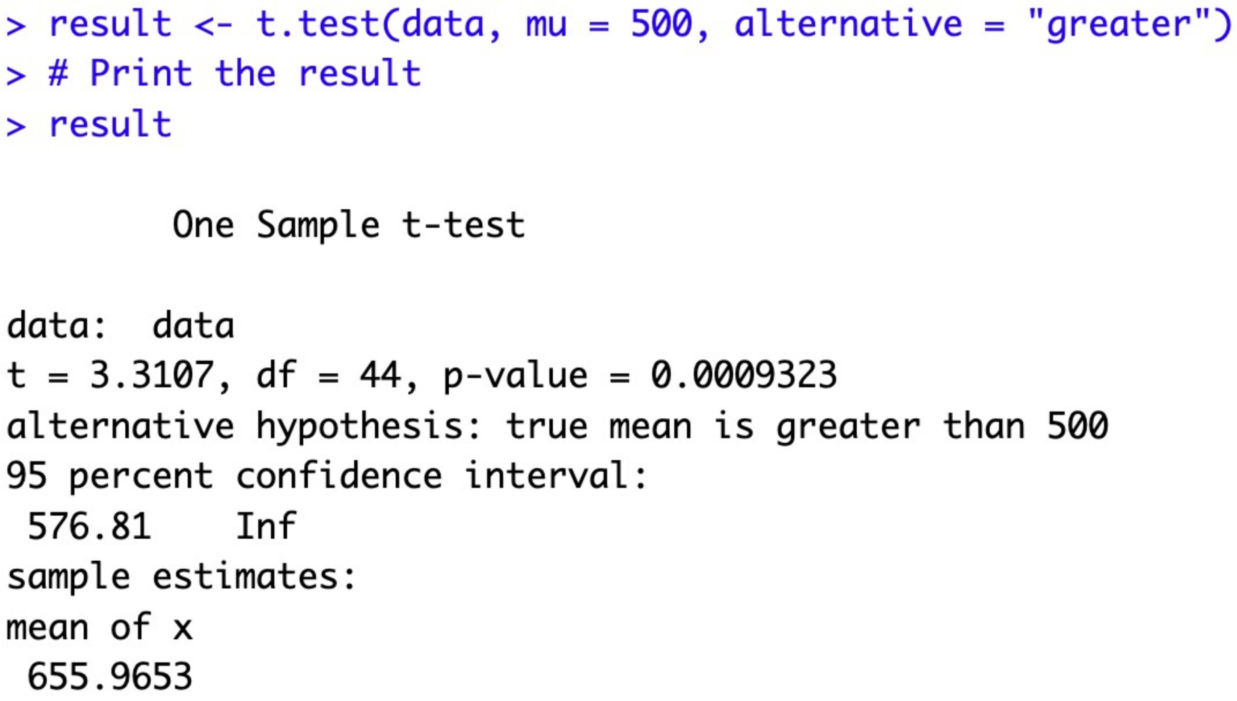
$$\mathrm{P}(T \leq t)$$

How to compute the p-value for a one-sample t test in R?

- Nowadays we can just ask an LLM….

- Let’s give it a try.

How to compute the p-value for a one-sample t test in R?



```
> result <- t.test(data, mu = 500, alternative = "greater")
> # Print the result
> result

	One Sample t-test

data:  data
t = 3.3107, df = 44, p-value = 0.0009323
alternative hypothesis: true mean is greater than 500
95 percent confidence interval:
 576.81    Inf
sample estimates:
mean of x
 655.9653
```

Legal System: Type I and Type II Errors

		Decision	
		Acquit	*Convict*
Truth	*Innocent* H_0	Correct	Type I error
	Guilty H_a	Type II error	Correct

- Given evidence (partial information), the jury is always at risk of making a mistake.

- Type I error (wrongly sentence an innocent person)
  - Safe strategy to remove Type I error is to let everybody go free
- Type II error (wrongly let a guilty person go free)
  - Safe strategy to remove this error is to convict everybody

- Tradeoff between the two errors
  - Unless with perfect information

Hypothesis Testing: Type I and Type II Errors

		Decision	
		H_0 true	H_a true
Truth	H_0	Correct	Type I error
	H_a	Type II error	Correct

- Denver Rent Example:
  - Type I error: Reject H_0 and claim profitable when it’s not
  - Type II error: Fail to reject H_0 and miss opportunity
  - Which error has the higher expected cost?

- Analogy to Legal System
  - A hypothesis testing between being innocent and guilty
  - H_0: innocent, H_a: guilty
  - Type I error: more severe

Common Practice in Hypothesis Testing

- Limit the chance of a Type I Error to a chosen level α
  - referred to as *significance level*
  - upper bound on Type I error
  - commonly set at 5%

- Reject H_0 when the p-value <= α

- If so, we claim that the data support the alternative H_a at level α, or
  - The data are statistically significant at level α

α and P-value

- P-value and significance level α :
  - Reject H_0 if p-value <= α
  - Do not reject H_0 if p-value > α.
  - Denver Rent: p-value=*0.0009*< *0.05*= α; hence reject $H_0: \mu = 500$

- When is the evidence against H_0 stronger?
  - *Large P-value or small P-value?*
  - The smaller the *P-value*, the stronger the evidence against H_0 and in favor of the alternative H_a.

- When is it easier to reject H_0?
  - *Large* α *or small* α *?*
  - We need a lot more evidence to reject H_0 for small α than for large α.

Example: Click Fraud

A retailer pays a hosting site for each click on an ad that brings customers to its website. Recently, however, the retailer suspects that many of these clicks have been generated by automated systems designed to imitate real customers.

The online host has promised that no more than 10% of the clicks are imitations.

To learn more, the retailer hired a service to identify fraudulent clicks.

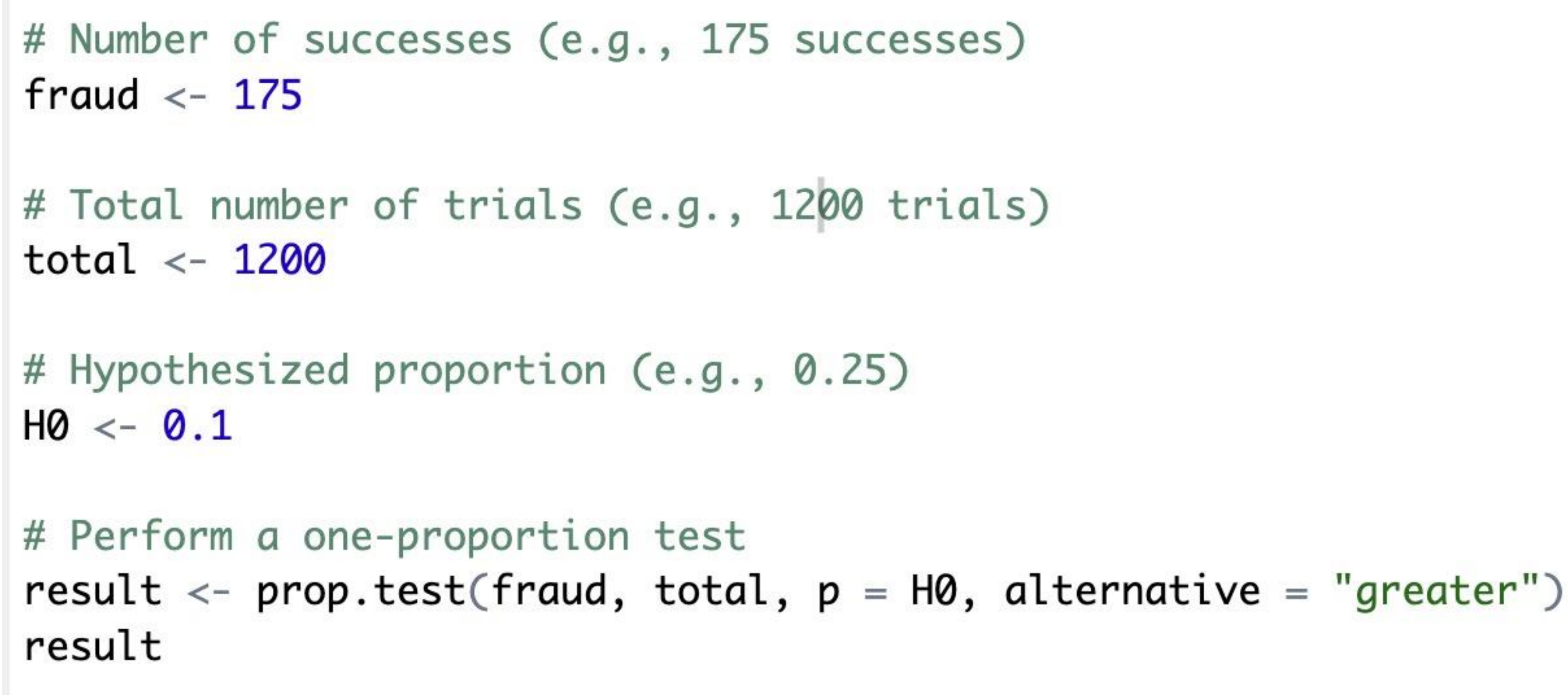
In a sample of 1,200 clicks, the service identified 175 computer-generated fraudulent clicks.

How conduct one-sample proportion test in R?

- Nowadays we can just ask an LLM….
- Let’s give it a try.

Click Fraud: One-Sample Prop Test

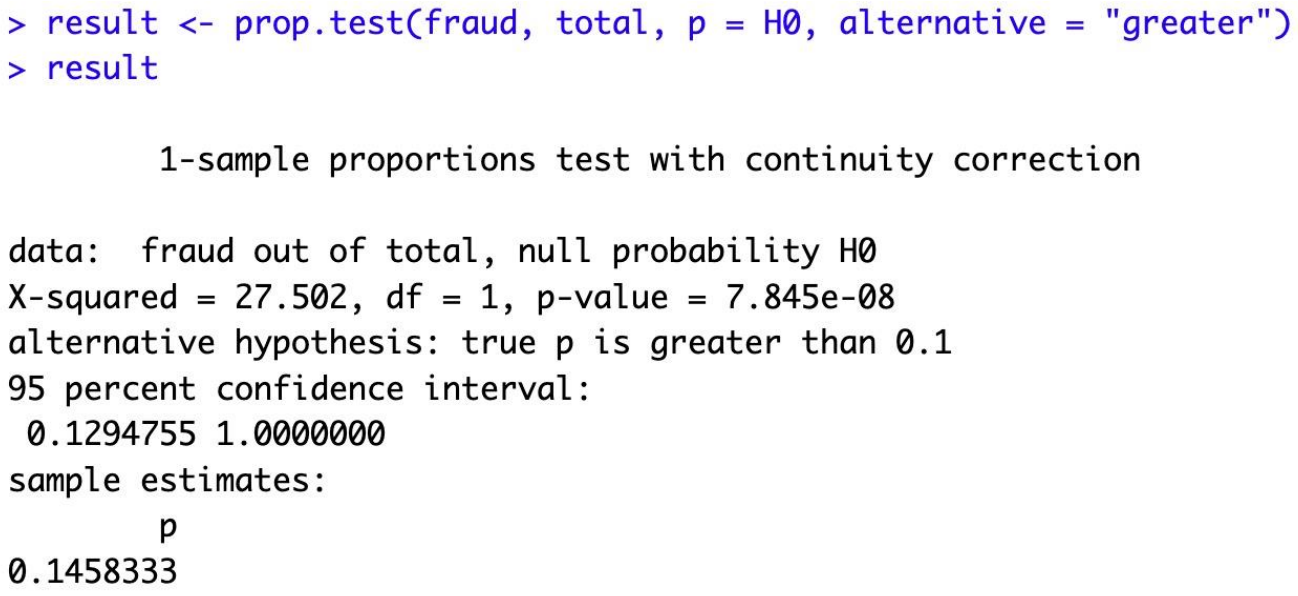
- p : proportion of fraudulent clicks
- $H_0 : p \leq 0.1$ versus $H_a : p > 0.1$



```
# Number of successes (e.g., 175 successes)
fraud <- 175

# Total number of trials (e.g., 1200 trials)
total <- 1200

# Hypothesized proportion (e.g., 0.25)
H0 <- 0.1

# Perform a one-proportion test
result <- prop.test(fraud, total, p = H0, alternative = "greater")
result
```

Click Fraud: One-Sample Prop Test

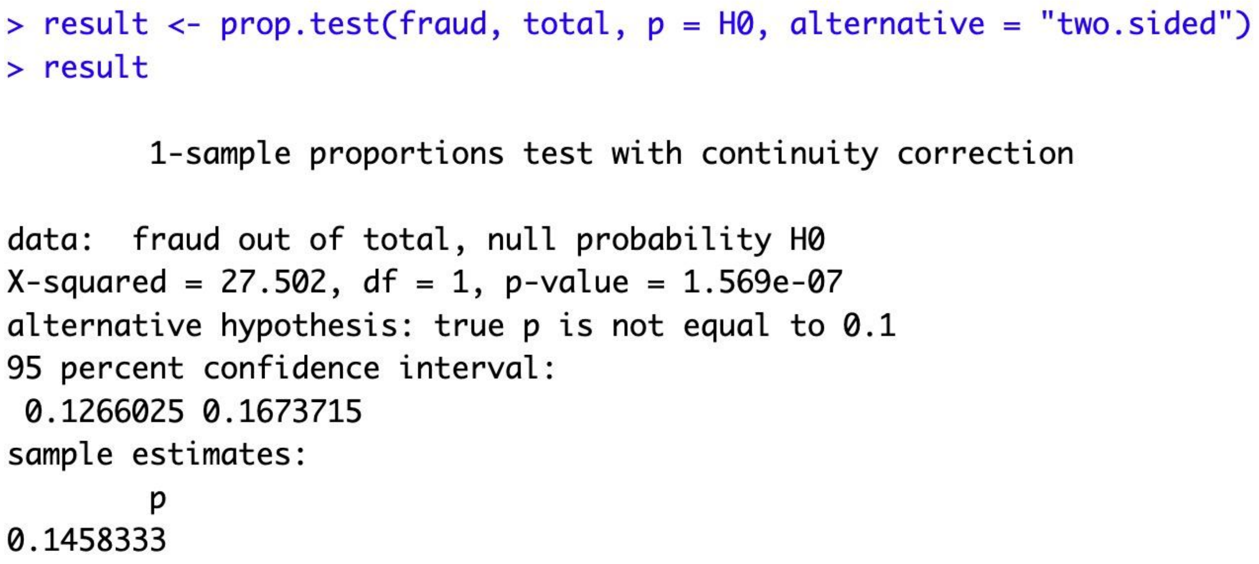
- $H_0: p \leq 0.1$ versus $H_a: p > 0.1$



```
> result <- prop.test(fraud, total, p = H0, alternative = "greater")
> result

	1-sample proportions test with continuity correction

data:  fraud out of total, null probability H0
X-squared = 27.502, df = 1, p-value = 7.845e-08
alternative hypothesis: true p is greater than 0.1
95 percent confidence interval:
 0.1294755 1.0000000
sample estimates:
        p 
0.1458333 
```

- Since the p-value is less than 0.05, we can reject the claim that $p \leq 0.1$ at significance level 0.05.

Click Fraud: confidence interval



```
> result <- prop.test(fraud, total, p = H0, alternative = "two.sided")
> result

        1-sample proportions test with continuity correction

data:  fraud out of total, null probability H0
X-squared = 27.502, df = 1, p-value = 1.569e-07
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
 0.1266025 0.1673715
sample estimates:
        p
0.1458333
```

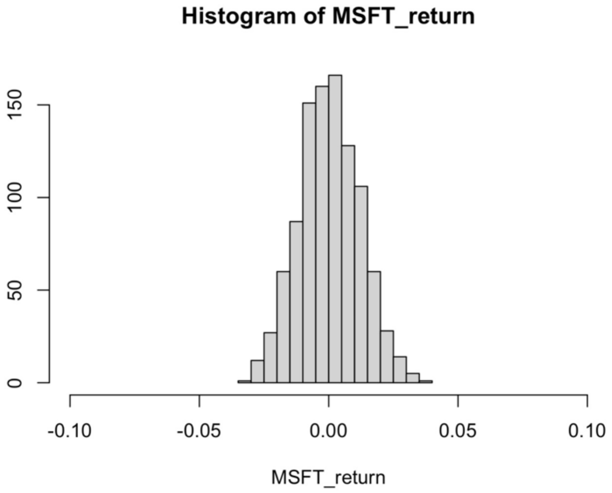
- Another solution: 95% confidence interval for p
  - 95% CI: [0.127, 0.167], which is above 0.10.
  - The host’s claim is not valid, which is consistent with the earlier testing conclusion.

Two-sided Hypothesis Tests

A two-sided hypothesis test detects a deviation in *either* direction from a claimed specific value for the population parameter. Confidence intervals provide an alternate method that can be used to test such hypotheses.

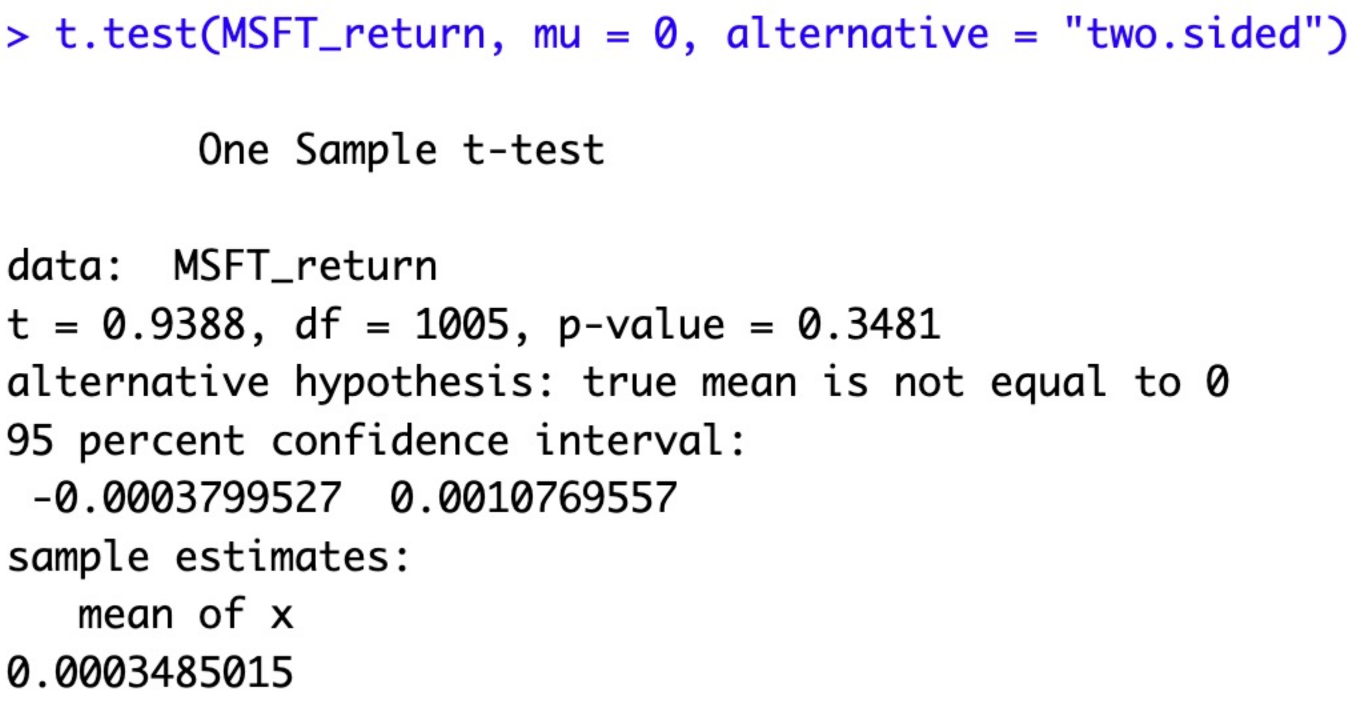
2020 - 2024 Microsoft Returns: mean return μ

H_0: $\mu = 0$ versus H_a: $\mu \neq 0$



Can we reject the null under 5% significance level?

Microsoft Stock Return Example



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

        One Sample t-test

data:  MSFT_return
t = 0.9388, df = 1005, p-value = 0.3481
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.0003799527  0.0010769557
sample estimates:
   mean of x
0.0003485015
```

Confidence Interval (CI) & 2-Sided Test

Two observations from the Stock Returns example:

- The p-value is 0.35. At 5% level, one can not reject H_0: $\mu = 0$. Thus it is possible that $\mu = 0$
- The 95% CI for μ is [-0.00038, 0.00108], which includes 0.
  - So, it is possible that $\mu = 0$

Equivalence between CI and 2-Sided Test!

Equivalence between CI and 2-Sided Tests

$H_0\colon \mu = \mu_0$ versus $H_a\colon \mu \neq \mu_0$

A level α 2-sided test

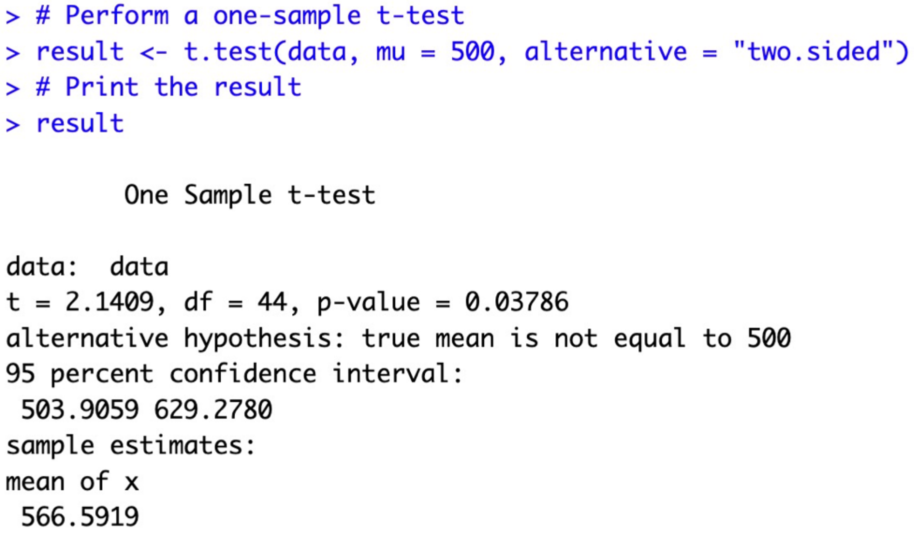
- Rejects H_0 when the value μ_0 falls outside a level 1 - α confidence interval for μ.
- Can't reject H_0 when the value μ_0 falls inside the CI.

CI can be used to test 2-sided hypotheses:

- Calculate the 1 - α level confidence interval
- Then
  - if μ_0 falls outside the interval, reject the null hypothesis;
  - otherwise, can't reject the null hypothesis.

Example: Denver Rent

- Test H_0: $\mu = \$500$ versus H_a: $\mu \neq \$500$



```
> # Perform a one-sample t-test
> result <- t.test(data, mu = 500, alternative = "two.sided")
> # Print the result
> result

	One Sample t-test

data:  data
t = 2.1409, df = 44, p-value = 0.03786
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 503.9059 629.2780
sample estimates:
mean of x
 566.5919
```

- What about H_a: $\mu \neq \$600$?

- What about H_a: $\mu \neq \$750$?

- What is the range of values for μ that H_0 can not be rejected?

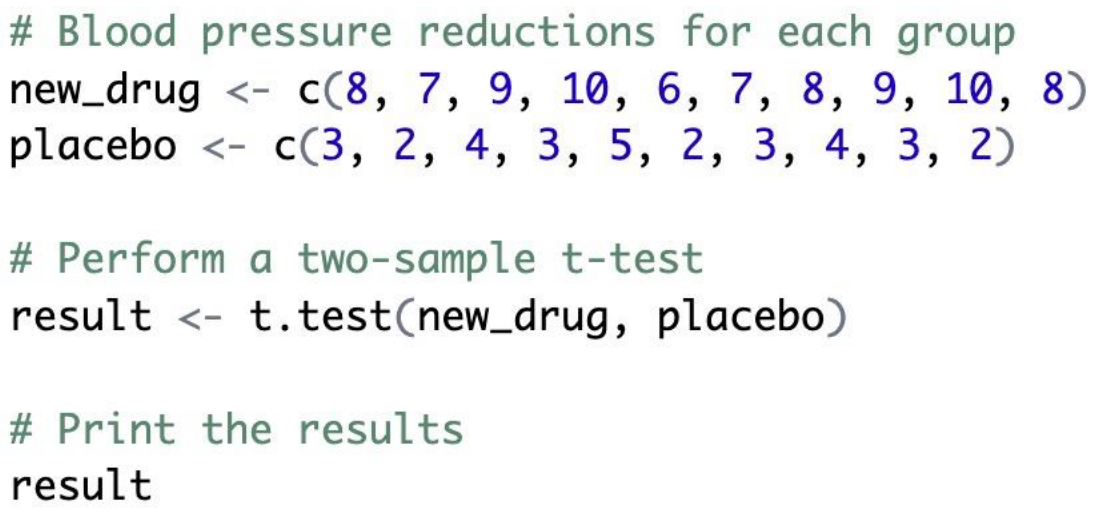
Example: Two sample t test

- A pharmaceutical company is testing the effectiveness of a new drug designed to lower blood pressure. They conduct a clinical trial with two groups of participants: one group receives the new drug, and the other group receives a placebo.

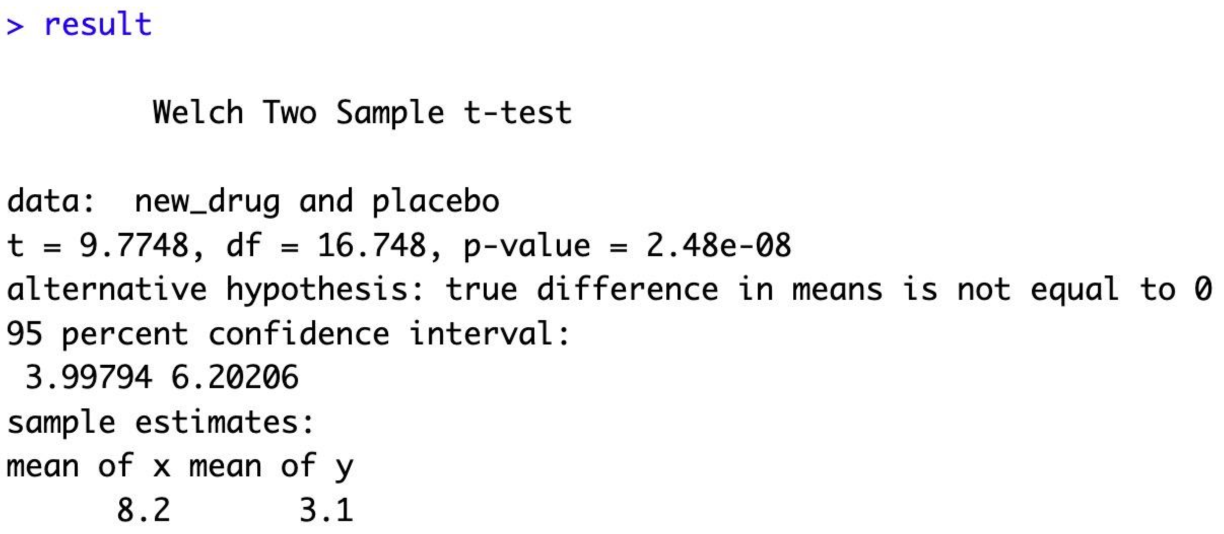
- **Group 1 (New Drug)**: Blood pressure reductions:
  - 8, 7, 9, 10, 6, 7, 8, 9, 10, 8

- **Group 2 (Placebo)**: Blood pressure reductions:
  - 3, 2, 4, 3, 5, 2, 3, 4, 3, 2

Example: Two sample t test



```r
# Blood pressure reductions for each group
new_drug <- c(8, 7, 9, 10, 6, 7, 8, 9, 10, 8)
placebo <- c(3, 2, 4, 3, 5, 2, 3, 4, 3, 2)

# Perform a two-sample t-test
result <- t.test(new_drug, placebo)

# Print the results
result
```



```
> result

	Welch Two Sample t-test

data:  new_drug and placebo
t = 9.7748, df = 16.748, p-value = 2.48e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.99794 6.20206
sample estimates:
mean of x mean of y 
      8.2       3.1 
```

Parts of a Hypothesis Test

- All hypothesis tests work the same way
  - State the null and alternative hypotheses
  - Compute the p-value
  - Interpret the results

- The differences are only in how the hypotheses are stated and the p-value computed.

- The p-value measures how much evidence there is against the null. A small p-value says the data are inconsistent with H_0, so that we should reject it.

- How small p-value needs to be depends on the consequence of making a mistake.