: Inferences about means

Objective

(1) Learn how/why Standard Error (SE) model for means is less accurate.

(2) Learn how to use student t-distributions.

(3) Learn how to create confidence intervals, and carry out hypothesis tests.

Concept briefs:

* Sampling dist. model for means: Comes from Central Limit Theorem, but this requires knowledge of true SD of the population (often not known).
* Standard error model: Use SD of the sample as approximation to true SD.
* Student t-distribution: New model that replaces normal distribution when using standard error to approximate true SD.
* Degrees of freedom (df): The student t-distribution is not a single model, unlike the normal distribution. Each 'df' corresponds to a different distribution in the student t family. Notation: t_{df} .
* New sampling dist. model for means: When the conditions are met, sample means (standardized) follow the t_{n-1} model with $SE = s/\sqrt{n}$.

  Here: n=sample size, s=sample SD. To standardize we calculate t-scores analogous to z-scores: $t = (\bar{y} - \mu)/SE$ [μ=true population mean].
* Conditions: Sample is random, independent, and approximately normal (i.e., nearly symmetric, unimodal).
* Inference methods: Follow all previous strategies for confidence intervals and hypothesis tests, except use the new sampling dist. model above.

Recap: Sampled data from surveys / experiments

Categorical variable

↓

Summary stats
Counts in each category, OR
Proportions in each category

↓

Simplest case
Only 2 categories ⇒
1 proportion

↓

Use Central Limit Theorem for proportions ⇒
Sampling distribution only depends on 1 thing = true proportion.

Quantitative variable

↓

Summary stats
Shape, center, spread
E.g., mean/SD, median/IQR

↓

Simplest case
Mean & SD

↓

Use Central Limit Theorem for means ⇒
Sampling distribution depends on 2 things = true mean & true SD.

Problem: How to get SE for sample means?

* Standard Error (SE) is an approximation to the Standard Deviation (SD)

* SE idea: If true values not known, replace with sampled values.

* Sampling distribution of means follows: N (True_mean, True_SD / $\sqrt{n}$)

* So we can use: SE = (SD_of_sample) / $\sqrt{n}$.

* This turns out to not be a good approximation.

* To compensate, we must replace normal model with a new model:

t-model, or t-distribution, or student-t model

Confidence interval: Summary of key ideas

Also known as: 1-sample t-interval

Follow the same strategy that we use for proportions:

* Sampling distribution model (check conditions)
* Computations
* Interpretation/conclusion

A key point to keep in mind is that C.I. is totally based on the sample, and is about finding the error margin in the sampled statistic.

Model & conditions

Sampling distribution follows the student t-distribution $T_{n-1}(\mu, SE)$.

Here μ=unknown true mean. $SE=\frac{s}{\sqrt{n}}$ [n=sample size, s=sample SD]

Conditions: (1) Sample is independent, and (2) approximately normal.

Computations

Find the standard error of student t-model

$$SE=\frac{s}{\sqrt{n}}$$ [n=sample size, s=sample SD]

Confidence interval for mean is

$$\bar{y} \pm t^*_{n-1} \times SE$$

sample mean

critical t-score from t-model with df=n-1
(depends on confidence level chosen)

Interpretation/conclusion

We are "C%" confident that the true mean of the population is contained within our C.I. From this it is possible to make inferences and/or compare with other populations.

Hypothesis testing: Summary of key ideas

Also known as: 1-sample t-test

Follow the same strategy that we use for proportions:

* Hypotheses
* Model & conditions
* Computations (with sketch showing model, type of "tail" & P-value)
* Interpretation/conclusion

A key point to keep in mind is that a hypothesis test is totally built upon the null hypothesis. It is the centerpiece of the model & all computations. The sampled statistic only enters the picture at the end, to calculate the P-value.

Hypotheses

Null & alt. hypotheses would have the form

H_0: $\mu = \mu_0$, H_A: $\mu \neq \mu_0$ OR $\mu > \mu_0$ OR $\mu < \mu_0$

Here μ_0 is a specific numerical value that is hypothesized.

Model & conditions

Sampling distribution follows the student t-distribution $T_{n-1}(\mu, SE)$, as before.

Conditions: (1) Sample is independent, and (2) approximately normal.

Computations

* Find standard error of student t: $SE = \frac{s}{\sqrt{n}}$ [n=sample size, s=sample SD]

* Find t-score of the observed sample mean: $t_{n-1} = \frac{\bar{y} - \mu_0}{SE}$

* Find P-value by looking up t-table with the right df (sample size – 1).

Interpretation/conclusion

If P-value is below significance level, there is statistically significant evidence to reject the null - describe what you can infer from this. Otherwise, retain the null & describe what that means.

Student t-model summary

* Looks kind of like a normal model - unimodal, symmetric, bell-shaped.

* However, there are some key differences:

  - It depends upon the sample size (n). Thus, must use a different model for different sample sizes. This is called "degrees of freedom."

  - For large values of "df," it is very close to normal model.

  - For small "df," it differs more significantly from normal: has "fatter" tail, and looks more spread out. Thus, critical t^* values tend to be somewhat larger than corresponding z* values, especially at small "df." [e.g., For 95% confidence z*=1.96, but t_5^*=2.57, t_{10}^*=2.23, t_{50}^*=2.01]

  - Use a t-model with df=n-1 for sample of size n. Use the nearest lower df value if you can't find the specific df you want in the table.

Exercise 24 (See last page for copy of Q)

(a) Check conditions: (1) Random sample?

The 44 weekdays in the sample are consecutive - not randomly selected. We will assume they are representative of all weekdays. 44 is certainly < 10% of all weekdays.

(2) Nearly normal sample data?

No specific info. given. Assume yes. Also, 44 is large enough of a sample that it is okay to proceed even if data is not close to normal.

(b) Steps to create 90% confidence interval:

Confidence interval requires: $\bar{y} \pm t^*_{n-1} \times SE$

Data given in sample: mean, $\bar{y}$=126 dollars

sample std. deviation, s=15 dollars

sample size, n= 44.

* Compute std. error: SE = $\frac{s}{\sqrt{n}} = \frac{15}{\sqrt{44}} = 2.2613$ dollars.

* Get confidence level, df, and t^* value.

  Here we need 90% conf. level. df =n-1 = 43.

  Lookup t-table and get $t^*_{43} = 1.68$

* Create confidence interval: $\bar{y} \pm t^*_{n-1} \times SE$

  126 ± 1.68 x 2.2613 = 122.2 to 129.8 dollars

(c) We can say with 90% confidence that the true mean daily income of the parking garage is between $122.2 and $129.8.

(d) 90% confidence means that 90% of all random (or representative) samples of size 44 will contain the true mean daily income of the parking garage within their confidence interval.

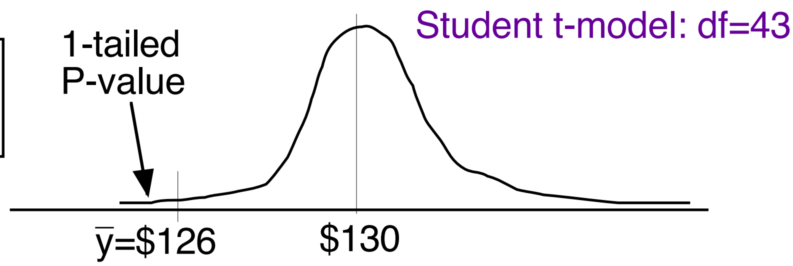
(e) The consultant's claim is not supported by my confidence interval, since a mean of $130 exceeds the upper end of my interval.

Just for fun, here is a hypothesis test to check the same thing

Hypothesis test:

* Null hypothesis - accept consultant's claim H_0: $\mu = \$130$.
* Alternative hypothesis H_A: $\mu < \$130$ [Note that H_A is 1-tailed]
* Assumptions already checked above.
* Sampling distribution model:

$$\boxed{SE = \frac{15}{\sqrt{44}} = \$2.2613}$$



* Observed mean is $\bar{y}$=$126
* Compute t-score: $t = \frac{\bar{y} - \mu}{SE} = \frac{126 - 130}{2.2613} = -1.77$
* Find P-value for $t_{43} < -1.77$ This turns out to be less than 0.05.
* Interpret: There is less than 5% chance that the observed mean could have occurred due to sampling variability if the null hypothesis is true.

  Thus, the true mean daily income of the parking garage is less than $130.

24. Parking Hoping to lure more shoppers downtown, a city builds a new public parking garage in the central business district. The city plans to pay for the structure through parking fees. During a two-month period (44 weekdays), daily fees collected averaged \$126, with a standard deviation of \$15.

a) What assumptions must you make in order to use these statistics for inference?

b) Write a 90% confidence interval for the mean daily income this parking garage will generate.

c) Interpret this confidence interval in context.

d) Explain what "90% confidence" means in this context.

e) The consultant who advised the city on this project predicted that parking revenues would average \$130 per day. Based on your confidence interval, do you think the consultant was correct? Why?