Making decisions using stats

Consider some real-world scenarios:

- Pfizer-BioNtech argues their COVID-19 vaccine is safe and effective for adolescents, based on results from clinical trials. Should the government authorize its use?
- A credit card company wants to determine which of two promotional offers will be more attractive to customers. They conduct a pilot study with a random sample of 5000 customers.
- A company wants to determine where to focus its efforts and resources to increase the diversity of its employees. How should they approach this task?

A common thread in these situations is the need to make a decision in the presence of uncertainty.

Confidence intervals and hypothesis tests provide a statistical framework for making such decisions, within risk tolerances acceptable to the stakeholders.

Exercise

A credit card company is planning a new promotional offer for current cardholders, in which they will give double airline miles for each dollar spent. To avail this offer, a cardholder must go online and register for it. To test the effectiveness of the campaign, the company sent out offers to a random sample of 50,000 cardholders. Of those, 1184 registered.

1. Construct a 90% confidence interval for the true proportion of cardholders who will register for the offer.

2. If the acceptance rate is $\leq 2\%$, the campaign won't be worth the expense. Given your confidence interval, what would you recommend?

3. Carry out a hypothesis test to determine whether the company should move forward

Chapters 17 & 19: Hypothesis tests
[Test of significance]

Objective

(1) To determine whether (& when) a sampled statistic "significantly" differs from some expected value.

(2) To learn concepts and methodology for carrying out hypothesis tests.

Concept briefs:

* Hypothesis = Proposed or estimated value of a statistic or parameter.

* Null hypothesis (H_0) = Accepted, or known, or default value of the parameter.

* Alternative hypothesis (H_A) = Always says H_0 is false.

* 2-tailed vs 1-tailed H_A = 2 types of H_A: either it says true value is different from H_0 (2-tailed), or it also specifies direction - larger/smaller (1-tailed). This is also known as 2-sided vs 1-sided probability.

* Sampling distribution model based on H_0 = H_0 is taken to be true. So it is the basis on which we apply CLT to get sampling dist. model for H_A.

* P-value = The probability that sampling variability accounts for the $\hat{p}$ observed in our sample (if we take the null hypothesis as fact).

* Statistically significant = When P-value of the estimated statistic drops below some threshold, it is considered statistically significant.

* Significance level [α-level] = The threshold value for "statistically significant."

* One proportion z-test = General term for hypothesis tests for proportions.

Illustration:

* Pharmaceutical company wants FDA to approve new headache medication.

* They conduct an experiment with 100 headache patients and find that 74% are cured of their headache within 2 hours of taking the medication.

* They create a 95% confidence interval & claim the true proportion of headaches cured within 2 hours is between 68% and 80%.

* Should the FDA approve their headache medication?

Extend above illustration:

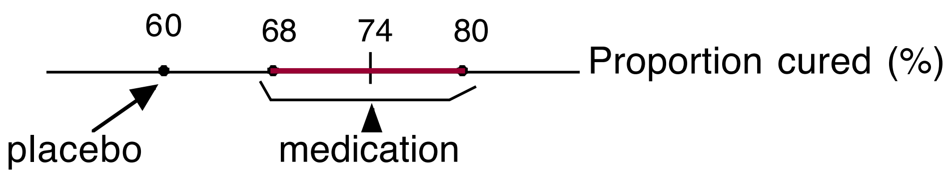
* A placebo group of headache patients was found to have a 60% cure rate within 2 hours.

* How does this affect one's assessment of the headache medication?

Statistical inference: Two strategies

Q: How to infer whether headache medication works better than placebo?

(1) Confidence interval



Inference: Placebo cure rate is outside the confidence interval. Therefore, medication does work (better than the placebo).

(2) Hypothesis test

* Default assumption: Medication does not work. Only reason for different cure rate is sampling variability.

* To prove the contrary: Must show that the probability of seeing a cure rate this far from the placebo in a random sample is very, very small.

Hypothesis Test Recipe

Step1: Write a clear statement of both hypotheses (H_0, H_A). [E.g., H_0: $p=p_0$, H_A: $p \neq p_0$ OR $p > p_0$ OR $p < p_0$]

Step2: Verify CLT assumptions for the sampled statistic ($\hat{p}$).

Step3: Find, sketch and label the sampling distribution model for the statistic.

Step4: Show your statistic ($\hat{p}$) on the sketch and shade in the p-value. Be careful to clarify whether you require 1-sided or 2-sided probability.

Step5: Find the z-score for your statistic.

Step6: Use z-score to find p-value from standard normal table.

Step7: Write a sentence or two as the conclusion from your test. It should tell whether to retain or reject the null hypothesis based on the given (or assumed) significance level.

Simpler Way to View These Steps

Every hypothesis test is expected to contain the following components:

* Hypotheses [step1]
* Model & assumptions [steps 2-3]
* Computations [steps 4-6]
* Interpretation/conclusion [step7]

Here is an exercise whose solution is discussed on the next page.

Exercise 14, pg. 469: The seller of a loaded die claims that it will favor the outcome 6. We don't believe the claim, and roll the die 200 times to test an appropriate hypothesis. Our P-value turns out to be 0.03. Which conclusion is correct, and why:

(a) There is a 3% chance the die is fair.

(b) There is a 97% chance the die is fair.

(c) There is a 3% chance that a loaded die could randomly produce the results we observed, so it is reasonable to conclude the die is fair.

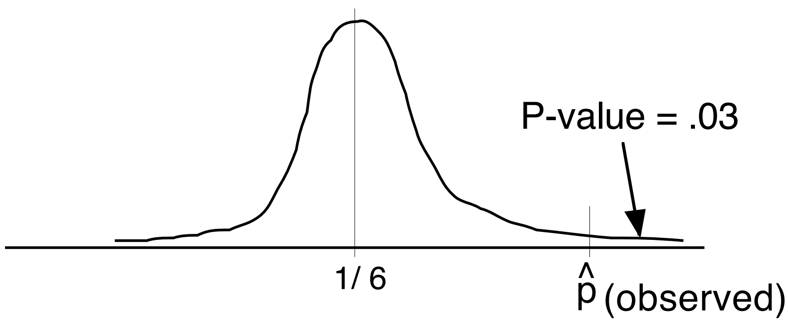
(d) There is a 3% chance that a fair die could randomly produce the results we observed, so it is reasonable to conclude the die is loaded.

Exercise 14, pg. 469

Here is detailed reasoning to help show (d) is the only appropriate conclusion:

* Null hypothesis is H_0: $p = 1/6$ (i.e., probability of 6 in a fair die is 1/6).
* Alt. hypothesis is H_A: $p > 1/6$
* We get some $\hat{p}$ from the experiment of rolling it 200 times.

Sampling distribution of $\hat{p}$



* So, the probability of observing this $\hat{p}$ (or larger) is 0.03, if H_0 is true.
* In other words, there is a 3% chance of this outcome randomly occurring with a fair die.

=

= –

Chapter 18: 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

<u>Also known as: 1-sample t-interval</u>

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.

I recommend review of the following exercises related to t-models:

Pg. 497: Ex. 3-6 and 9-10.

Exercise 24, pg. 499

(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.