Statistical Inference

Foundations for Inference

Behnam Bahrak

Spring 2020

Nearly Normal Sampling Distributions

- The sample mean is not the only point estimate for which the sampling distribution is nearly normal:

sample mean	$\bar{x}$
difference between sample means	$\bar{x}_1 - \bar{x}_2$
sample proportion	$\hat{p}$
difference between sample proportions	$\hat{p}_1 - \hat{p}_2$

- Some point estimates follow distributions other than the normal distribution, and some scenarios require statistical techniques that we will cover later.

Unbiased Estimator

- An important assumption about point estimates is that they are unbiased, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimates.
- That is, an unbiased estimate does not naturally over or underestimate the parameter, it provides a "good" estimate.
- The sample mean is an example of an unbiased point estimate, as well as others we just listed.

Confidence intervals for nearly normal point estimates

- A confidence interval based on an unbiased and nearly normal point estimate is

$$point\ estimate \pm z^* SE$$

where z^* is selected to correspond to the confidence level, and SE represents the standard error.

- Remember that the value $z^* SE$ is called the margin of error.

ME: Margin of error / SE: Standard Error

Example

- A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show (an American late-night TV show). The standard error of this estimate is 0.014. Estimate the 95% confidence interval for the proportion of college graduates who watch The Daily Show.

$\hat{p} = 0.33$
$SE = 0.014$

$$\hat{p} \pm z^* SE = 0.33 \pm 1.96 \times 0.014 = 0.33 \pm 0.027$$

$$(0.303, 0.357)$$

Hypothesis testing for nearly normal point estimates

- Like with confidence intervals, we can apply the same framework for hypothesis testing to different estimators, as long as the estimator is unbiased and has a nearly normal sampling distribution.

- So if that's the case, we can use the z-statistic as our test statistic, that we always calculate as a point estimate minus the null value, kind of like the observed minus the mean, divided by some standard error:

$$Z = \frac{point\ estimate\ - null\ value}{SE}$$

Example

- The 3rd NHANES collected body fat percentage (BF%) and gender data from 13,601 subjects ages 20 to 80. The average BF% for the 6,580 men in the sample was 23.9, and this value was 35.0 for the 7,021 women. The standard error for the difference between the average male and female BF%s was 0.114. Do these data provide convincing evidence that men and women have different average BF%s. You may assume that the distribution of the point estimate is nearly normal.

1. Set the hypotheses

$H_0: \mu_{men} = \mu_{women}$ $H_A: \mu_{men} \neq \mu_{women}$

2. Calculate the point estimate

$\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

Example

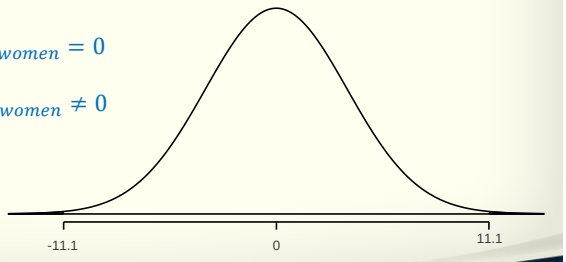
3. Check conditions:

We may assume that the distribution of the point estimate is nearly normal.

4. Draw sampling distribution, shade p-value:

$H_0: \mu_{men} = \mu_{women} \rightarrow \mu_{men} - \mu_{women} = 0$

$H_A: \mu_{men} \neq \mu_{women} \rightarrow \mu_{men} - \mu_{women} \neq 0$



Example

5. Calculate test statistics and p-value, make a decision

$$Z = \frac{-11.1 - 0}{0.114} = -97.36$$

$$p - value \approx 0 \quad \rightarrow \quad \text{Reject } H_0$$

These data provide convincing evidence that the average BF% of men and women are different.

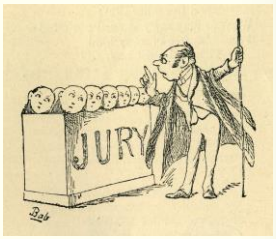
Decision Error

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	Type 1 error
	H_A true	Type 2 error	✓

- Type 1 error is rejecting H_0 when H_0 is true.
- Type 2 error is failing to reject H_0 when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis test as a trial

- If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:
  - H_0 : Defendant is innocent
  - H_A : Defendant is guilty



- Which type of error is being committed in the following circumstances?
  - Declaring the defendant innocent when they are actually guilty: *Type 2 error*
  - Declaring the defendant guilty when they are actually innocent: *Type 1 error*

Which error is the worse error to make?

- Which error is the worst error to make?
  - Type 2 : Declaring the defendant innocent when they are actually guilty
  - Type 1 : Declaring the defendant guilty when they are actually innocent

"better that ten guilty persons escape than that one innocent suffer"

William Blackstone

Type 1 Error Rate

- We reject H_0 when the p-value is less than 0.05 (α = 0.05).
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error}) = P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α: increasing α increases the Type 1 error rate.

Choosing α

- If Type 1 Error is dangerous or especially costly, choose a small significance level (e.g. 0.01)

Goal: we want to be very cautious about rejecting H_0, so we demand very strong evidence favoring H_A before we would do so.

- If Type 2 Error is relatively more dangerous or more costly, choose a higher significance level (e.g. 0.1)

Goal: we want to be cautious about failing to reject H_0 when the null is actually false.

Truth vs. Decision Table

goal: keep α and β low

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	Type 1 error, α
	H_A true	Type 2 error, β	$1-\beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level).
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β.
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$.

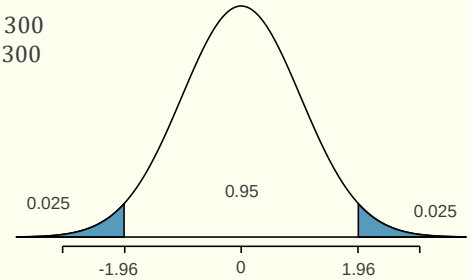
Type 2 Error Rate

- If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?
  - The answer is not obvious.
  - If the true population average is very close to the null value, it will be difficult to detect a difference (and reject H_0).
  - If the true population average is very different from the null value, it will be easier to detect a difference.
  - Clearly, β depends on the effect size (δ), difference between point estimate and null value: $\delta = \bar{x} - \mu_0$

Significance vs. Confidence Level

two sided HT with α = 0.05 ⇔ 95% confidence interval

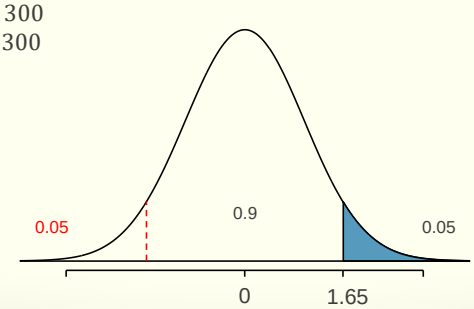
$H_0: \mu = 300$
$H_A: \mu \neq 300$



Significance vs. Confidence Level

one sided HT with α = 0.05 ⇔ 90% confidence interval

$H_0: \mu = 300$
$H_A: \mu > 300$



Agreement of CI and HT

- A two sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1 - \alpha$.
- A one sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1 - (2 \times \alpha)$.
- If H_0 is rejected, a confidence interval that agrees with the result of the hypothesis test should not include the null value.
- If H_0 is failed to be rejected, a confidence interval that agrees with the result of the hypothesis test should include the null value.

Question

- All else held equal, will the p-value be lower if $n = 100$ or $n = 10{,}000$?
  - $n = 10{,}000$

$$\bar{x} = 50$$
$$s = 2$$
$$H_0 : \mu = 49.5$$
$$H_A : \mu > 49.5$$

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25$$

- When we're thinking about practical significance, we focus on the effect size.

Example

- Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

- When n is large, even small deviations from the null (small effect sizes), which may be considered practically insignificant, can yield statistically significant results.

n زیاد بزرگ باعث میشه که اگه به راحتی H_A رو ببینیم

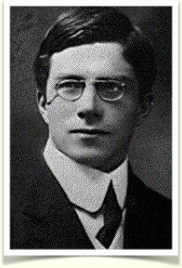
Statistical vs. Practical Significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.
- The sample size is something the researcher has control over, because we can decide how many observations we want to sample.
  - when you see highly statistically significant results make sure that you also inquire whether the effect size is reported and what the sample size is as well.

Statistical vs. Practical Significance

- In order to make sure that our findings don't suffer from the problem of being statistically significant, but not practically significant, we often do a priori analysis before we actually do the data collection to figure out, based on characteristics of the variable you are studying, how many observations to collect.
  - The last thing you want to do is having to find out you either don't have enough or you have too many observations

R.A. Fisher: "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."



Sir Ronald Fisher (1890-1962)

Example

- A municipal employee has to audit the parking tickets issued by city parking officers to determine the number of tickets that found to be improperly issued.
- In past years, the number of improperly issued tickets per officer had a normal distribution with mean $\mu = 380$ and $\sigma = 35.2$.
- Due to a recent change in the city's parking regulations, the employee suspects that the mean number of improperly issued tickets has increased.
- An audit of 50 randomly selected officers showed an average of 390 improper tickets.
- Use the sample data given here and $\alpha = 0.01$ to test the employee's suspicion.

Hypothesis Test

1. **Set the hypotheses**

μ = average number of tickets

$H_0: \mu = 380 \qquad H_A: \mu > 380$

2. **Calculate the point estimate**

$\bar{x} = 390 \ , \ n = 50$

3. **Check conditions**

1. random & 50 < 10% of all parking officers → independence

2. $n > 30$ & sample not skewed → nearly normal sampling distribution

4. **Calculate test statistic**

$\bar{x} \sim N(\mu = 380 \ , SE = \frac{\sigma}{\sqrt{n}} = \frac{35.2}{\sqrt{50}} \approx 4.98)$

test statistic: $Z = \frac{390-380}{4.98} = 2.01$ → p-value $\approx 0.02 > 0.01$ → Fail to reject H_0

نتیجه گیری با داده آزمایش دارد صورت!

Power

- What is the power of the test?
- To compute the power of a test, we need another parameter called actual mean or μ_a:

$$\text{Type 2 error} = \beta = P(\text{Fail to reject } H_0 | \mu = \mu_a)$$

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \ | \mu = \mu_a)$$

- In other words, what is the power of the test for detecting the difference between actual mean (μ_a) and the mean that we are testing (μ_0 or null value)?
- The plot of power versus μ_a is called the **power curve**.

Question

- For parking ticket example, in which case β is smaller?
  1) The actual mean is $\mu_a = 400$
  2) The actual mean is $\mu_a = 387$

- If the null hypothesis is $H_0: \mu = 380$, the probability of incorrectly accepting H_0 will depend on how close the actual mean is to 380.
- If the actual mean number of improperly issued tickets is 400, we would expect β to be much smaller than if the actual mean is 387.

Computing Power

- Let us suppose that the actual mean number of improper tickets is 395 per officer. What is power?

$$\mu_a = 395 \rightarrow \bar{x} \sim N(\mu = 395, SE = \frac{\sigma}{\sqrt{n}} = \frac{35.2}{\sqrt{50}} \approx 4.98)$$

$$\alpha = 0.01, \text{one sided test} \rightarrow z_\alpha = 2.33 \ (P(Z > 2.33) = 0.01)$$

- We reject H_0 if the z-statistics $z = \frac{\bar{x}-\mu_0}{SE}$ is larger than $z_\alpha = 2.33$

$$\text{power} = P\left(\frac{\bar{x} - 380}{4.98} > 2.33 \,\middle|\, \bar{x} \sim N(\mu = 395, SE = 4.98)\right)$$

Computing Power

$$\text{power} = P\left(\frac{\bar{x} - 380}{4.98} > 2.33 \,\middle|\, \bar{x} \sim N(\mu = 395\,, SE = 4.98)\right)$$

$$\text{power} = P(\bar{x} > 380 + 2.33 \times 4.98 = 391.6 \,|\, \bar{x} \sim N(395\,, 4.98))$$

$$= P\left(Z > \frac{391.6 - 395}{4.98}\right) = P(Z > -0.98) = 0.8365$$

Computing Power