Hypothesis Testing and Sampling

Collin Nolte

February 8, 2022

Review and Preview

What we did last time:

- Sampling distribution of sample mean

$$\overline{X} \sim N\left(\mu, \sigma^2/n\right)$$

- Intervals like $\overline{x} \pm \hat{\sigma}/\sqrt{n}$ give us interval of values for specified *coverage probability* $(1-\alpha)$
- Sample mean and intervals themselves are results of random process

What we do today:

- Framing hypotheses and interpreting p-values
- Drawing conclusions
- Reconciliation with sampling distribution

Basic Hypothesis Testing

The formal process of scientific investigation

1. Define the *null hypothesis* as a declarative, unambiguous statement
2. Collect observational or experimental data
3. Compare the results to what would have been expected based on the null hypothesis (statistical inference)
4. Either *reject* or *fail to reject* the null hypothesis based on the *strength of the evidence*

Null hypothesis

Typically, these are built around parameters in a distribution, with the "naught" subscript used to identify it. Common ones include:

- Testing a specific parameter, $H_0 : \mu = \mu_0$
- Group comparisons, $H_0 : \mu_A - \mu_b = \mu_0 = 0$
- Odds or relative risk, $H_0 : \theta = \theta_0 = 1$

Often, this null takes the assumption of no effect or no change, i.e., difference between groups is 0, or odds ratio is equal to. Once we have observed our data, we compute a test statistic and compare it to a null model parameter, i.e., we compare $\overline{x}$ against μ_0

Considering the evidence

"Is this difference due to chance, or is the null hypothesis incorrect?"

We know that because of randomness, our observations will never be equal to the null, and we will never know the absolute truth. As a consequence, inference is framed in terms of probabilities.

If our null hypothesis, H_0, is true, what is the probability that we observe the given data? This is what is reported with a *p-value*.

$$p = P(\text{observed data} \mid H_0)$$

p-values

p-values have become a heated topic in statistical inference due to how easily they can be misinterpreted. Here are some key points that we will come back to:

- A p-value *is not* the probability that the null hypothesis is false
- A p-value *is not* the probability of an observation being produced by random chance alone
- A p-value *does not* tell us the magnitude of difference or effect
- A p-value *must* be taken in the context of the study; a p-value of 0.05 is completely arbitrary
- A p-value *is* a probabilistic statement relating observed data to a hypothesis

A possible scenario

Suppose we have some null hypothesis, $H_0 : \mu = \mu_0$ that we wish to test

We start by collecting a random sample of our population of size n and compute the mean and standard deviation, giving us estimates of $\overline{x}$ and $\hat{\sigma}$

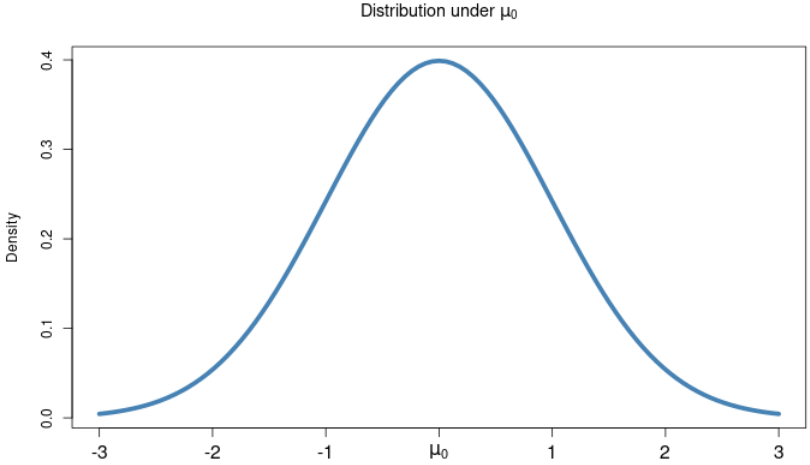
From the CLT, we know that $\overline{x} \sim N\left(\mu, \sigma^2/n\right)$, which we can use to construct intervals of likely values (coverage probability)

Let's keep the measure of variance that we found, $\hat{\sigma}^2$, but instead let's center this distribution at μ_0

We will finish by considering $\overline{x}$ as a draw from this distribution

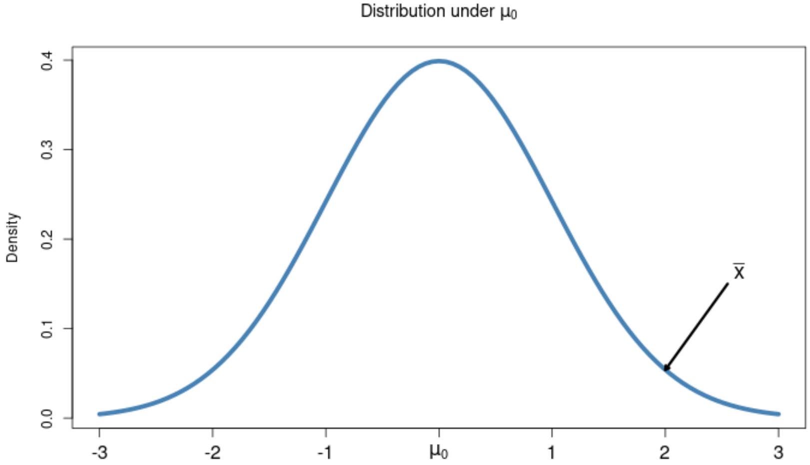
Null Distribution

We begin with a sampling distribution centered at μ_0



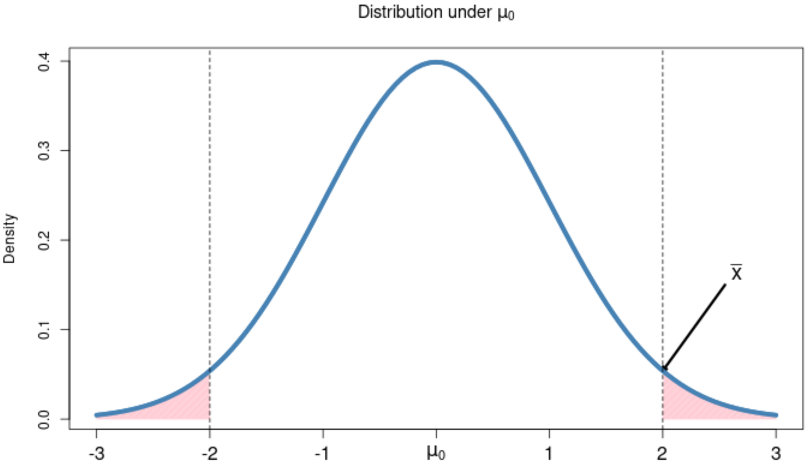
Null Distribution

Then pretend that our observed $\overline{x}$ was drawn from *this* distribution



Null Distribution

Given that $\mu = \mu_0$, what is the probability that we draw the observation $\overline{x}$ or something greater. The area in the pink region is our *p-value*



Null Distribution and p-value

In a way, this is similar to what was done when considering the intervals for our *coverage probability*

Before, we found $\overline{x}$ and $\hat{\sigma}$ and, using these, constructed intervals of varying length and asked, "What is the probability of drawing samples *within* this interval?"

Now, the statement is framed slightly differently. Given our null distribution and an observation of $\overline{x}$, we are now asking, "If my intervals are the length between μ_0 and $\overline{x}$, what is the area *outside* my coverage probability?"

In other words, "Assuming μ_0 is true, what is the probability of observing $\overline{x}$ or something greater?"

$$p = P(\text{observed data} \mid H_0)$$

Null Distribution and *p*-value, cont.

To reiterate a few things:

- The *p*-value is not telling us if our null hypothesis is incorrect or not
- The *p*-value is also not telling us the probability of the data
- The *only* thing a *p*-value is telling us is a probabilistic relationship between our observed sample mean and a hypothetical distribution

Let's now step back and start considering how we might use this information to draw conclusions

Revisting the CLT

We've noted several times that the CLT provides that (approximately)

$$\overline{X} \sim N\left(\mu, \sigma^2/n\right)$$

However, with some arithmetic, we can simplfy this just a bit

$$\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) \sim N(0,1)$$

Substituting $Z = \sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)$ gives us a *standardized* test statistic that is significantly easier to work with

An analytic approach

Let's say now that we want to find an interval for this test statistic, Z, such that, when randomly sampled, our coverage probability is equal to $(1 - \alpha)$. As this is symmetric about zero, we might say we are looking for the *critical value* z_α such that

$$P(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha$$

Being that $Z \sim N(0, 1)$, we know everything there is to know about Z; computing these values is relatively simple to do on a computer for any given value of α.

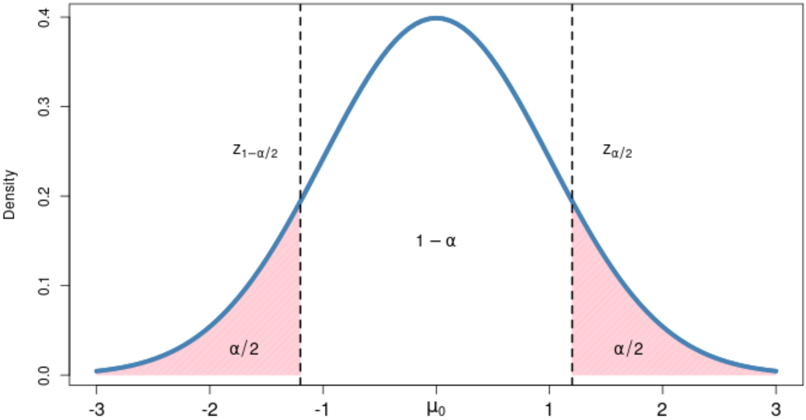
An analytic approach, cont.

With this in place, let's make another substitution

$$
\begin{aligned}
1 - \alpha &= P(-z_\alpha \leq Z \leq z_\alpha) \\
&= P\left(-z_\alpha \leq \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) \leq z_\alpha\right) \\
&= P\left(-z_\alpha \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq z_\alpha \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)
\end{aligned}
$$

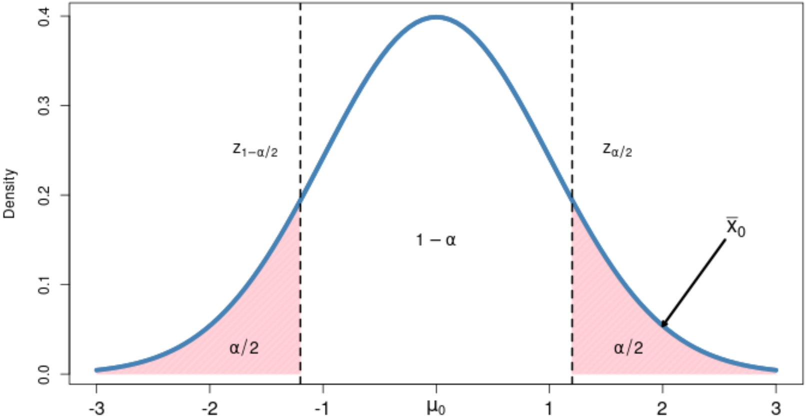
And violá, we have an interval $\overline{X} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$ that will contain μ $(1-\alpha)\%$ of the time

Confidence Intervals



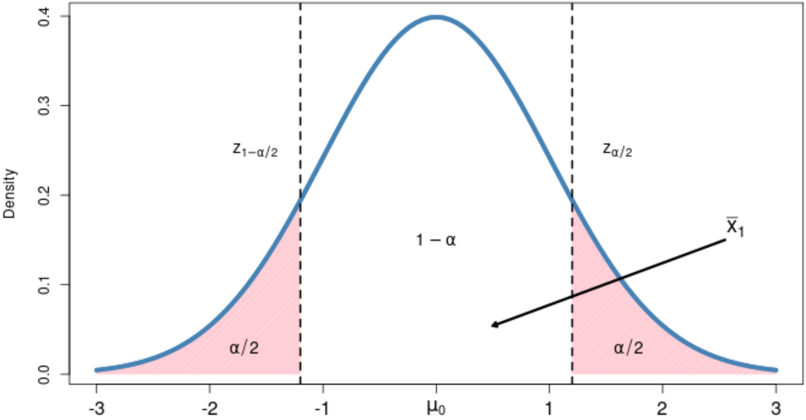
Confidence Intervals

Reject $H_0 : \mu = \mu_0$
p-value $< \alpha$



Confidence Intervals

Fail to reject $H_0 : \mu = \mu_0$
p-value $> \alpha$



Drawing Conclusions

In actuality, a null hypothesis is either true or false, and based on the data, we may reject or fail to reject this null. As a consequence, there are two ways in which we might make a mistake.

Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct $(1 - \alpha)$	Incorrect Type II Error (β)
Reject H_0	Incorrect Type I Error (α)	Correct $(1 - \beta)$

- Type I error $= P(\text{Reject } H_0 | H_0 \text{ true}) =$ false alarm
- Type II error $= P(\text{Fail to reject } H_0 | H_A \text{ true}) =$ missed opportunity

Controlling Errors

While all mistakes aren't great, some are worse than others, and the design of our study can influence which errors are more likely to occur.

The *Type I* error can be controlled by setting the level of significance, α. The smaller the value of α, the more evidence required to reject H_0. In other words, we can require the p-value to be such that $p < \alpha$

The *Type II* error is controlled by of β. The quantity $1 - \beta$ is called the *power* of a study. More powerful studies have lower probabilities of Type II errors

Unfortunately, these values are often in conflict: if we always reject the null, we will never commit a Type II error. Similarly, if we never reject the null, the probability of a Type I error is zero. Obviously, neither is ideal

Review so far

- A distribution, governed by parameters, describes the mechanism by which our data are generated
- Null hypothesis (H_0) given in terms of distribution parameters
- Data is collected and test statistic computed
- p-value generated by comparing test statistic to model parameter, indicating probability of observation assuming the null hypothesis is true, i.e., $p = P(\text{observed data} \mid H_0 \text{ is true})$
- Reject or fail to reject H_0
    - Type I error (α): probability of incorrectly rejecting H_0 when H_0 is true
    - Type II error (β): probability of failing to reject H_0 when H_0 is false