Hypothesis Testing and Sampling

February 09, 2021

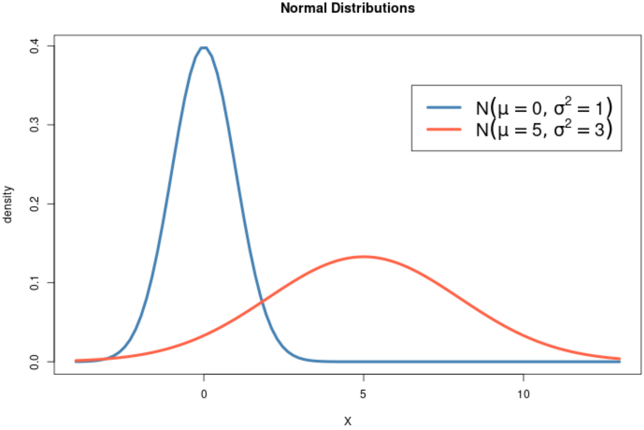
Distributions

A distribution is a function that takes an "event" as an input and returns a probability

- Convenient to think of a physical process, i.e., a "data generating mechanism"
- Governed by set of distribution parameters
- Continuous or discrete
- When we say X follows a distribution, we mean that our observations were generated by a mechanism following a particular form
- $X \sim N(\mu, \sigma^2)$, $X \sim bin(n, p)$

Normal Distribution

$$X \sim N(\mu, \sigma^2) \quad \iff \quad f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



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:

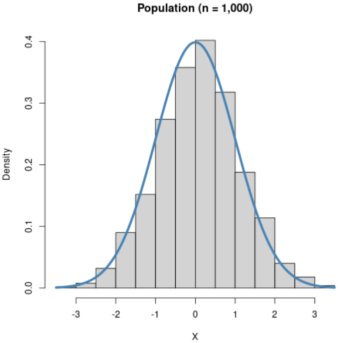
- Testing a specific parameter, $H_0 : \mu = \mu_0$
- Group comparisons, $H_0 : \mu_A - \mu_b = \mu_0 = 0$
- Odds or relative risk, $H_0 : \theta = \theta_0 = 1$

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

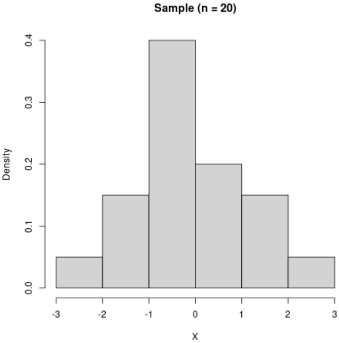
Distribution → Population → Sample

A sample that is randomly drawn from a population will follow the same distribution, though will never be an exact fit.

Here, $\mu = 0$, and $\overline{x} = 0.016$



Population (n = 1,000)



Sample (n = 20)

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

Reject or Fail to Reject

A critical point and a common error must be pointed out: statistical inference is *never used to accept a null hypothesis*. Consider the following:

1. If the power goes out, we will not have an internet connection
2. We do not have an internet connection
3. We conclude that the power must be out

This is the consequence of a logical fallacy assuming biconditional equivalence; here, there are a multitude of other reasons why we may not have internet.

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.

	True State of Nature	
Test Result	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(Reject $H_0 | H_0$ true) $=$ false alarm
- Type II error $=$ P(Fail to reject $H_0 | H_A$ 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 $p < \alpha$

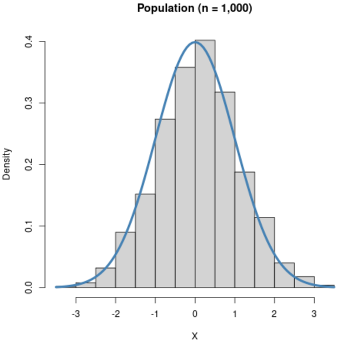
The *Type II* error is controlled by of β. The quantity $1 - \beta$ is called the *power* of a study. More powerful studies have lower probabilities of Type II errors

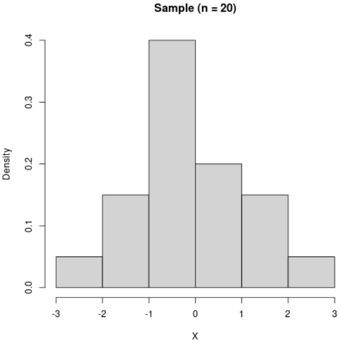
Unfortunately, these values are often in conflict: if we always reject the null, we will never commit a Type II error. Similarly, if we never reject the null, the probability of a Type I error is zero. Obviously, neither is ideal

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

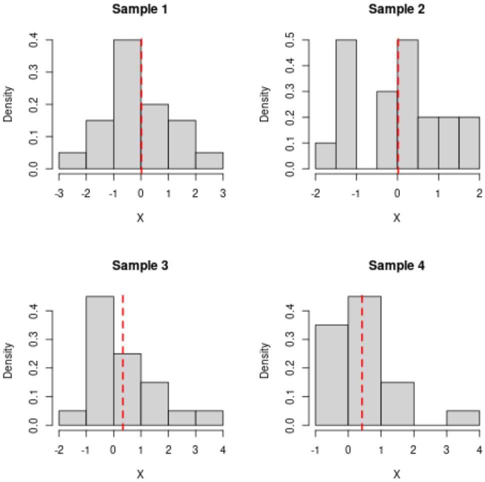
Distribution → Population → Sample





More Samples ($n = 20$)

Each random sample will have a different sample mean, $\overline{x}$



Sample Size and Sampling Distribution

Suppose we have a population of size $N = 1,000$ with $X \sim N(\mu, \sigma^2)$, and we have a sample of size n.

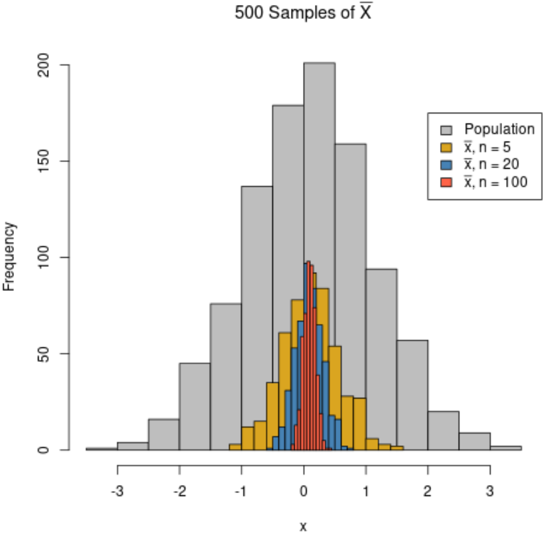
- If $n = 1,000$, i.e., I sample every person in my population, what would we expect the value of $\overline{X}$ to be? How much variability will there be between samples?
- What if $n = 1$, i.e., if $\overline{X} = X_i$?
- $n = 20$?

In particular, we want to ask ourselves two questions:

1. What is the expected value of $\overline{X}$? In other words, if I randomly sample $\overline{X}$ a large number of times, what would I expect the average to be?
2. How much variability will we see if we randomly sample $\overline{X}$? How does this relate to sample size?

Sampling Distribution

Just as our population follows a distribution, so does our test statistic



Central Limit Theorem

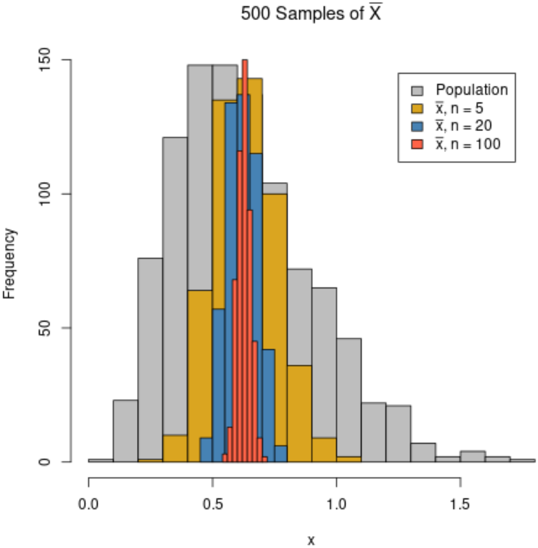
This is perhaps the most single most amazing thing in the field of statistics.

If a collection of random samples $X_1, X_2, \ldots, X_n$ are drawn from a population that has mean μ and finite variance σ^2, then

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

In other words, for large enough n, we have $\overline{X} \sim N(\mu, \sigma^2/n)$. Even more amazingly, this continues to be true for a large number of test statistics AND it does not require that the original distribution be normal

CLT with Gamma distribution


500 Samples of $\overline{X}$
Frequency
x
Population
$\overline{x}$, n = 5
$\overline{x}$, n = 20
$\overline{x}$, n = 100