FTDS // INFERENTIAL STATISTICS

Hacktiv8 DS Curriculum Team

Phase 0 Learning Materials

Hacktiv8 DS
Curriculum
Team

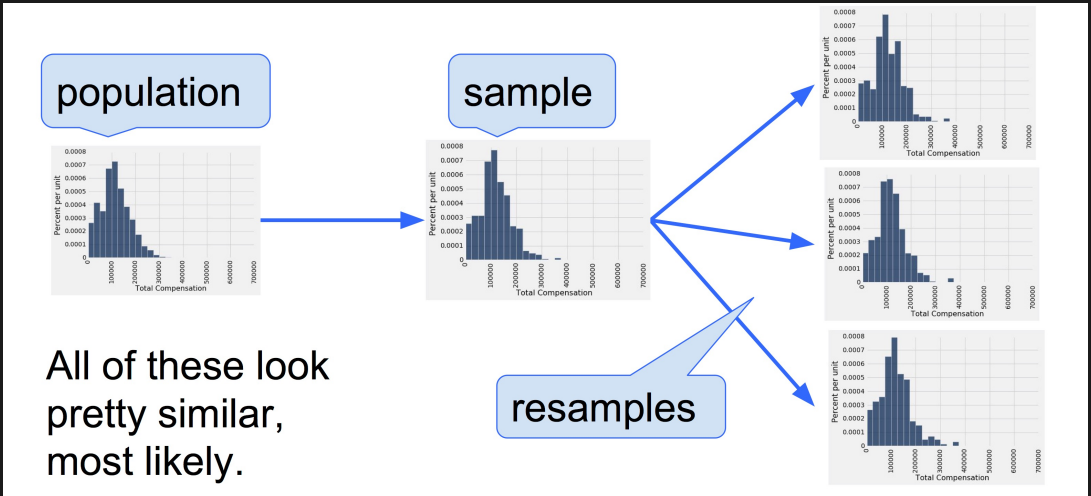
Contents

Brief of Inferential Statistics

“The purpose of statistical inference is to draw conclusions about a population on the basis of data obtained from a sample of that population. Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population”.

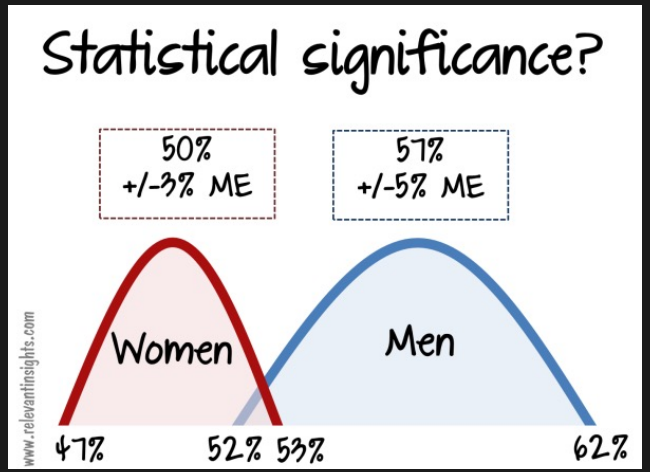
Resampling

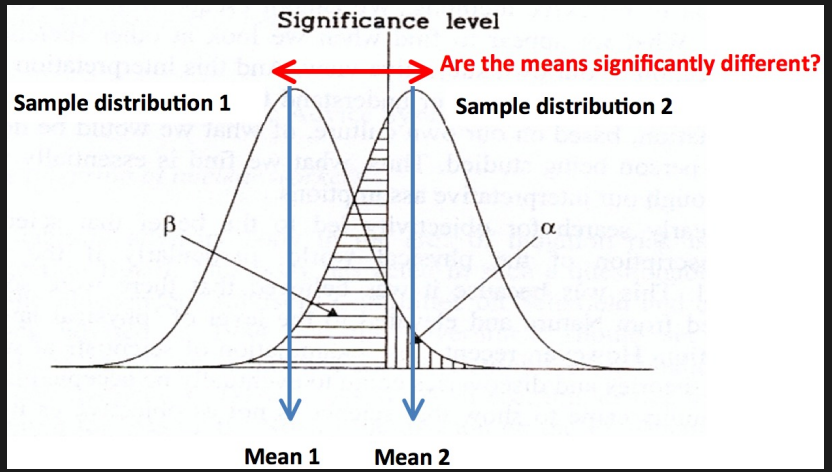
Resampling in statistics means to repeatedly sample values from observed data, with a general goal of assessing random variability in a statistic. **Randomness is probably one of the most powerful phenomena that we have at hand when dealing with Statistics. It may not be so apparent, but under the right scope, it can help us uncover hidden patterns in data.**



Statistical Significance

Statistical significance refers to whether any differences observed between groups being studied are "real" or whether they are simply due to chance.

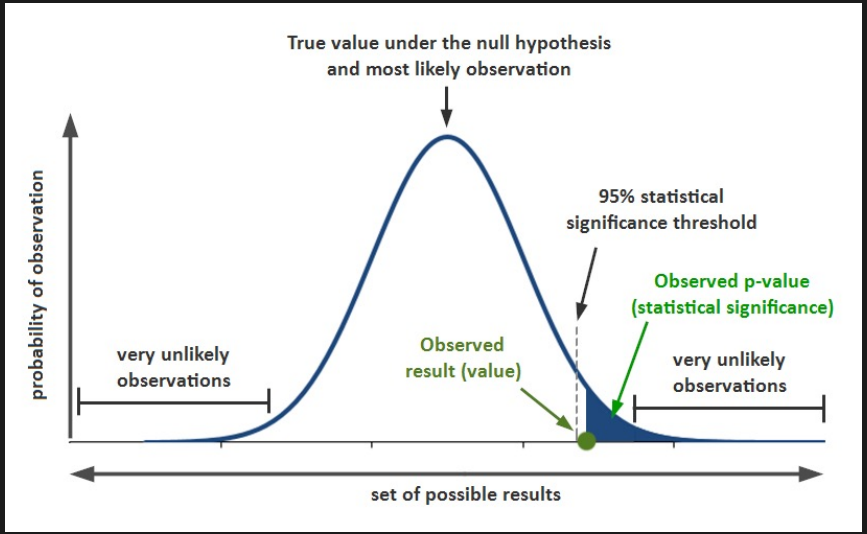




P-Value

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true).

The level of statistical significance is often expressed as a *p*-value between 0 and 1.



Hypothesis Testing

Definition of Terms:

•Null hypothesis (H0): A hypothesis associated with a theory one would like to prove.
•Alternative hypothesis (H1): A hypothesis associated with a contradiction to a theory one would like to prove.
•Critical value: The threshold value of the test statistic for rejecting the null hypothesis.

Important Points:

•Hypotheses are mutually exclusive.
•The p-value is the probability that a given or more significant result, would occur under the null hypothesis.
•We can only reject or not reject H0 in favor of the alternate hypothesis, thus we cannot accept H1 in any condition.

Single-Sample One-Sided

Formula of Hypotheses:

$$H_0: \mu \leq x$$
$$H_1: \mu > x$$

Single-Sample Two-Tailed

Formula of Hypotheses:

$$H_0: \mu = x$$
$$H_1: \mu \neq x$$

Two Sample Independent

We want to hypothesize whether two population mean difference is statistically significant.

Formula of Hypotheses:

Two Tail test

$$H_0: \mu_a = \mu_b$$
$$H_1: \mu_a \neq \mu_b$$

Single tail test

$$H_0: \mu_a \leq \mu_b$$
$$H_1: \mu_a > \mu_b$$

Paired Tests

The event on observed groups are related between each other. Technically, it tests whether the changes tend to be in the positive or negative direction.

T-Test

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.

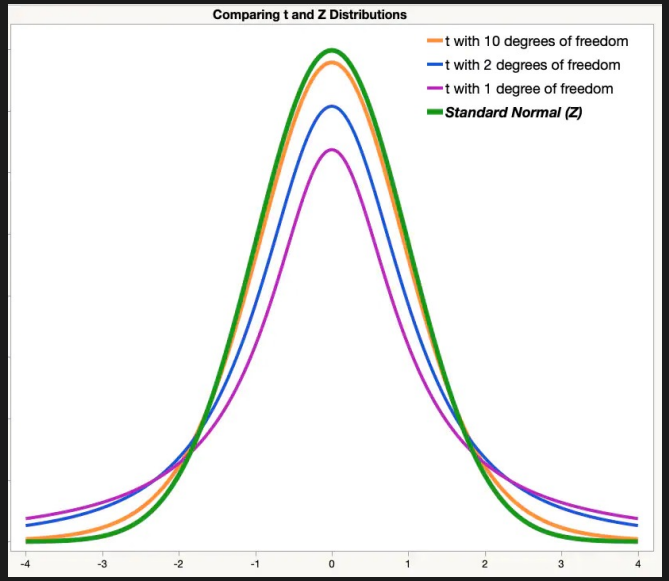
Assumptions:
1. Scale of measurement applied to the data collected follows a continuous or ordinal scale.
2. Of a simple random sample, the data is collected from a representative, randomly selected portion of the total population.
3. The data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. Homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

T-Test and Z-Test

A t-test can be regarded as a Z-test with limited sample size.

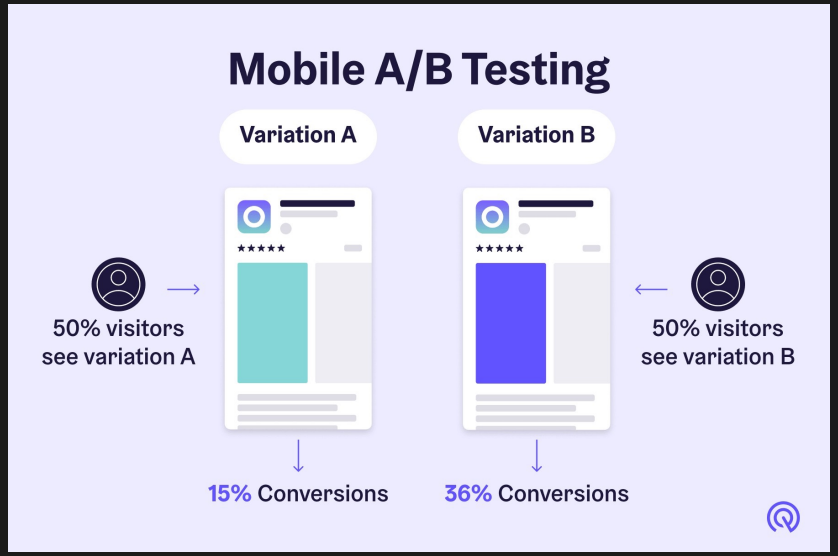
With more samples, the Student-T distribution will be closer to a Normal or Z distribution.

However, if we do not know the variance/standard deviation of the population then we use t-test.



A/B Testing

A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.



"The ABCs of A/B Testing - Pardot". Pardot. Retrieved 2016-02-21.

ANOVA

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. ANOVA can be used for comparison between multiple groups

We can compute an ANOVA table using `statsmodel`.

	df	sum_sq	mean_sq	F	PR(>F)
Page	3.0	831.4	277.133333	2.739825	0.077586
Residual	16.0	1618.4	101.150000	NaN	NaN

F statistic is based on the ratio of the variance across group means (i.e., the treatment effect) to the variance due to residual error. The higher this ratio, the more statistically significant the result.

If the data follows a normal distribution, then statistical theory dictates that the statistic should have a certain distribution. Based on this, it is possible to compute a p-value

Chi Square

The chi-square test, or Fisher's exact test, is used when you want to know whether an effect is for real or might be the product of chance. A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.

H_0: In the population, the two categorical variables are unrelated

H_a: In the population, the two categorical variables are related

Chi Square

Assumptions:

1. The data in the cells should be frequencies, or counts of cases rather than percentages or some other transformation of the data.
2. The levels (or categories) of the variables are mutually exclusive. That is, a particular subject fits into one and only one level of each of the variables.
3. Each subject may contribute data to one and only one cell in the $\chi 2$.
4. The study groups must be independent.
5. There are 2 variables, and both are measured as categories, usually at the nominal level. However, data may be ordinal data. Interval or ratio data that have been collapsed into ordinal categories may also be used.
6. The value of the cell expected should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than This assumption is most likely to be met if the sample size equals at least the number of cells multiplied by 5. Essentially, this assumption specifies the number of cases (sample size) needed to use the $\chi 2$ for any number of cells in that $\chi 2$.

External References

Colab Link — Visit Here