Statistics

What is a p-value?
When is it used?

Statistics

What is a hypothesis test?
Give an example and provide necessary steps to arrive at an answer.

The p-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.

The p-value is widely used in statistical hypothesis testing, specifically in null hypothesis significance testing. In this method, as part of experimental design, before performing the experiment, one first chooses a model (the null hypothesis) and a threshold value for p, called the significance level of the test, traditionally 5% or 1% [6] and denoted as α. If the p-value is less than or equal to the chosen significance level (α), the test suggests that the observed data is inconsistent with the null hypothesis, so the null hypothesis must be rejected. However, that does not prove that the tested hypothesis is true. When the p-value is calculated correctly, this test guarantees that the Type I error rate is at most α. For typical analysis, using the standard α = 0.05 cutoff, the null hypothesis is rejected when p <= .05 and not rejected when p > .05.

A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population. A hypothesis test examines two opposing hypotheses about a population: the null hypothesis and the alternative hypothesis.

For example, let's say we'd like to test the assumption that men are taller than and women at the Galvanize campuses.
Steps:
1) State the null and alternate hypotheses
The null hypothesis (typically described as H0) is that the average heights of the men and women are not significantly different, while the alternate hypotheses (H1 or Ha) is that men are taller than women.
2) Calculate the t (or z) statistic
The t-value quantifies how big the difference in the means is compared tvariation in the data. It's basically the difference in the means quantified in multiples of the standard error.
3) Using the t value and the degrees of freedom, find the corresponding probability p.
Based on the alternate hypothesis, a decision about using a single tailed probability (as in this case for greather than) or a double-tailed probability (means not equal) is desired to get the right value of p.
4) Compare p to the desired Type I error rate, alpha.
Alpha, the significance level, is typically 0.025 to 0.05 for a 95% confidence level for in double and single sided tests, respectively. If p <= alpha, the null hypothesis can be rejected, otherwise it can't.
5) State the conclusion.

Statistics

What is statistical power?
What affects statistical power?

Statistics

Give examples where the median is a better measure of centrality than the mean.

Statistical power is the likelihood that a study will detect an effect when an effect exists (or, that a false null hypothesis will be rejected). Statistical power is inversely related to the probability of making a Type II error (β). Power = 1 – β.

As statistical power increases the probability of making a Type II error (a false negative) decreases. Though there is no formal standard for power, (1-β) = 0.8 is often used.

Generally statistical power depends on the desired significance criterion, the magnitude of the effect of interest in the population (effect size), and the sample size used to detect the effect. Increasing the sample size is often the easiest way to increase the power.

Many formulae exist for determining the correct sample size for a desired statistical power, depending on the application.

The median is a better measure than the mean when data are skewed.
The median is a better measure in cases where outliers shift the centrality of the mean.

Statistics

What is A/B testing?

Give a couple of examples where it might be used.

Statistics

What type of distribution models the period of time between events occurring at an average rate (assuming each event occurs independently of the last event)?

Give a couple real life examples that would be modeled by the above distribution.

A/B testing is:
* Two-sample hypothesis testing
* Randomized exposure of test subjects to two variants: A and B
* A is control, B is the variation

Examples:
* In website design, A/B testing will determine if changes to the website changed (hopefully increased) the click through rate.
* In the case of marketing materials, where a letter to a prospective customer will end with either “Sale ends March 8, use discount code M8 online to claim” or “Sale ends soon, use code SL to claim”

The exponential distribution.

Examples include:
The probability that you’ll get a phone call in the next hour given you get a call on average once every 4 hours.
The probability someone will pass you at a street corner in the next five minutes given that 10 people walk by that street corner every hour.

Statistics

What are common methods for dealing with missing data?
What are the advantages and disadvantages of each method?

Statistics

What is the Central Limit Theorem?
Where is it often employed in data science?

Common imputation methods of dealing with unknown or missing values include:

* Removing entire observations containing one or more unknown values
  Advantage: easy; Disadvantage: decreasing power and losing data
* Filling in unknown values with the average of the existing values (mean imputation)
  Advantage: easy; Disadvantage: diminishes utility of correlations that use the variable that's imputed
* k-nearest-neighbors: Use the values of clustered neighbors to to fill in missing data points
  Advantage: more representative, Disadvantage: more computationally expensive

The CLT states that the arithmetic mean of a sufficiently large number of iterates of independent random variables will be approximately normally distributed regardless of the underlying distribution. i.e: the sampling distribution of the sample mean is normally distributed.

In statistics notation speak:
$X \sim \_(\mu, \sigma^2)$ where _ is any distrubution, and by the CLT:

$\bar{X} \sim N(\mu, \sigma^2/n)$ where N is a normal distribution and n is the number of samples drawn.

The CLT is used in hypothesis testing and confidence intervals.

The process of bootstrapping is like the process of sampling for the CLT but can be used to generate statistics other than the mean.

Statistics

Contrast Frequentist and Bayesian statistics.
Explain how determining the average height of adult women in the US would be approached by these two paradigms.

Statistics

Explain Bayes rule in words and then write out its formula.

In a Frequentist framework, there are true fixed parameters that describe a population. So in the case of average height of women in the US, there is one true answer to this (e.g 5' 7"). Now when this population is sampled, that's where probability and confidence intervals enter the picture.

In a Bayesian framework, distribuions are associated with parameters that descibe the population. A Bayesian would start with a "prior" distribution that describes his/her present state of knowledge (say that the height of women follows a normal distribution centered on 5' 6" with a standard deviation of 6"). Then the Bayesian would collect data. This data would be used to update the prior distribution to get a new distribution – the posterior distribution. Statements about probability and confidence intervals reference this posterior distribution.

In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on conditions that might be related to the event.

$P(A|B) = P(B|A) * P(A) / P(B)$
where:
P(A) and P(B) are the probabilities of events A and B without regard to each other.
P(B|A) is the probability of event B occuring given that event A occurred,
P(A|B) is the probability of event A occurring given that event B occurred. It's what we are trying to find.

these distributions are often referred to as:
P(B|A) is the "likelihood."
P(A) is the "prior."
P(B) is the "normalizing constant", sometimes referred to as the "evidence."
P(A|B) is the "posterior."