Terence Shin, MSc, MBA

Summary

This article provides 40 statistics interview problems and answers for data

40 Statistics Interview Problems and Answers for Data Scientists

A resource to brush up your statistics knowledge for your interview!

Photo from admiralmarkets.com

> ***Be sure to*** subscribe here ***or to my*** exclusive newsletter ***to never miss another article on data science guides, tricks and tips, life lessons, and more!***

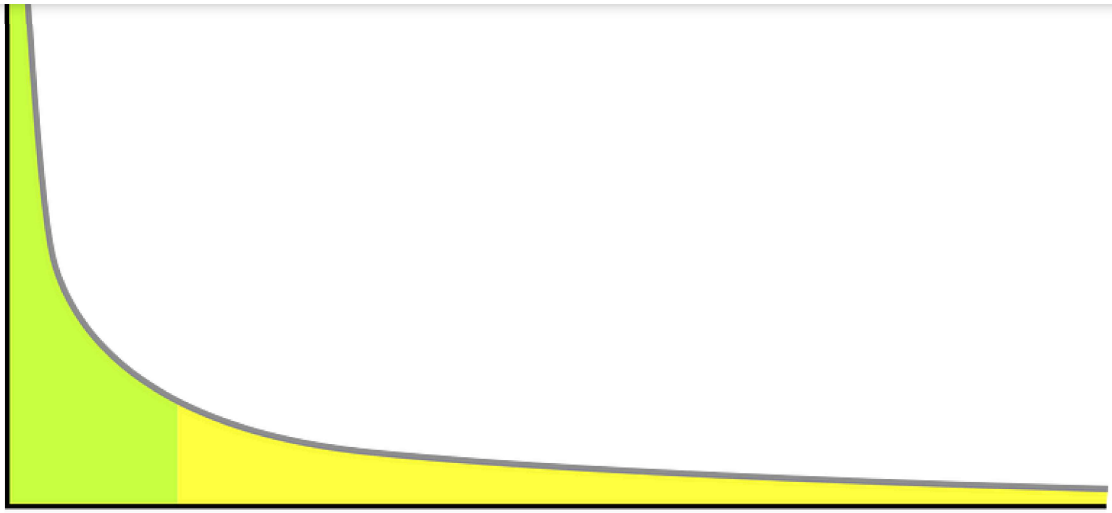
Given the popularity of my articles, Google's Data Science Interview Brain Teasers, Amazon's Data Scientist Interview Practice Problems, Microsoft Data Science Interview Questions and Answers, and 5 Common SQL Interview Problems for Data Scientists, I collected a number of statistics data science interview questions on the web and answered them to the best of my ability.

fundamental statistics knowledge is essential to be successful. So, I crawled the web and found **forty** statistics interview questions for data scientists that I will be answering. Here we go!

1. How do you assess the statistical significance of an insight?

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

2. Explain what a long-tailed distribution is and provide three examples of relevant phenomena that have long tails. Why are they important in classification and regression problems?

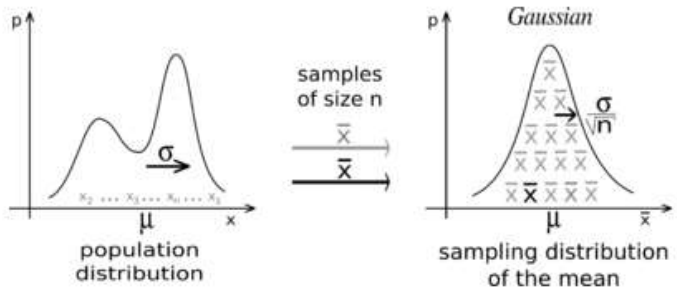


Example of a long tail distribution

A **long-tailed distribution** is a type of heavy-tailed distribution that has a tail (or tails) that drop off gradually and asymptotically.

3 practical examples include the power law, the Pareto principle (more commonly known as the 80–20 rule), and product sales (i.e. best selling products vs others).

It's important to be mindful of long-tailed distributions in classification and regression problems because the least frequently occurring values make up the majority of the population. This can ultimately change the way that you deal with outliers, and it also conflicts with some machine learning techniques with the assumption that the data is normally distributed.



Central Limit Theorem explained visually

From Wikipedia

Statistics How To provides the best definition of CLT, which is:

> *"The central limit theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size gets larger no matter what the shape of the population distribution." [1]*

The central limit theorem is important because it is used in hypothesis testing and also to calculate confidence intervals.

4. What is the statistical power?

'Statistical power' refers to the power of a binary hypothesis, which is the probability that the test rejects the null hypothesis given that the alternative

$$Power = P(reject\ Null\ |alternative\ is\ true)$$

5. Explain selection bias (with regard to a dataset, not variable selection). Why is it important? How can data management procedures such as missing data handling make it worse?

Selection bias is the phenomenon of selecting individuals, groups or data for analysis in such a way that proper randomization is not achieved, ultimately resulting in a sample that is not representative of the population.

Understanding and identifying selection bias is important because it can significantly skew results and provide false insights about a particular population group.

Types of selection bias include:

- **sampling bias**: a biased sample caused by non-random sampling
- **time interval**: selecting a specific time frame that supports the desired conclusion. e.g. conducting a sales analysis near Christmas.
- **exposure**: includes clinical susceptibility bias, protopathic bias, indication bias. *Read more* here.

- **attrition**: attrition bias is similar to survivorship bias, where only those that 'survived' a long process are included in an analysis, or failure bias, where those that 'failed' are only included
- **observer selection**: related to the Anthropic principle, which is a philosophical consideration that any data we collect about the universe is filtered by the fact that, in order for it to be observable, it must be compatible with the conscious and sapient life that observes it. [3]

Handling missing data can make selection bias worse because different methods impact the data in different ways. For example, if you replace null values with the mean of the data, you adding bias in the sense that you're assuming that the data is not as spread out as it might actually be.

> ***Be sure to*** subscribe here ***or to my*** exclusive newsletter ***to never miss another article on data science guides, tricks and tips, life lessons, and more!***

6. Provide a simple example of how an experimental design can help answer a question about behavior. How does experimental data contrast with observational data?

Observational data comes from observational studies which are when you observe certain variables and try to determine if there is any correlation.

control certain variables and hold them constant to determine if there is any causality.

An example of experimental design is the following: split a group up into two. The control group lives their lives normally. The test group is told to drink a glass of wine every night for 30 days. Then research can be conducted to see how wine affects sleep.

7. Is mean imputation of missing data acceptable practice? Why or why not?

Mean imputation is the practice of replacing null values in a data set with the mean of the data.

Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.

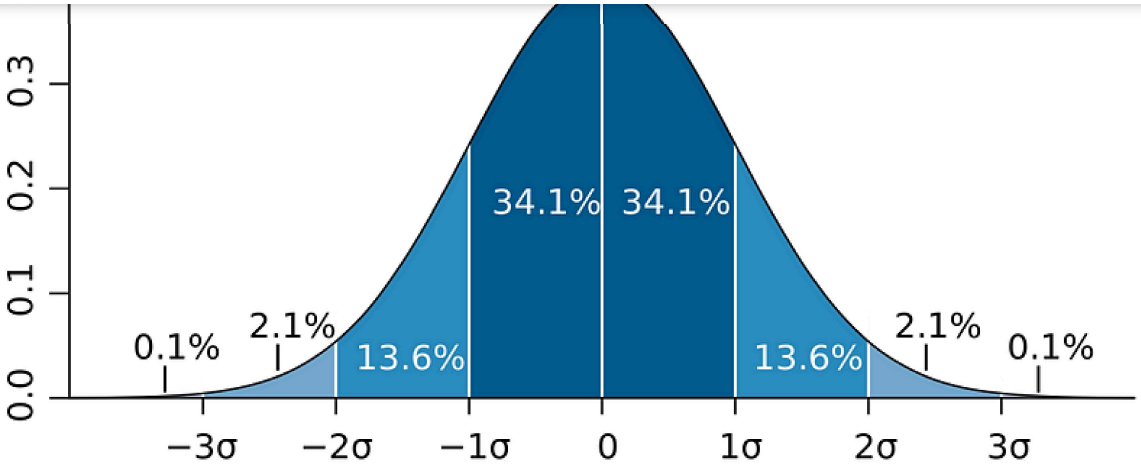
Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

dataset. Also, explain what an inlier is and how you might screen for them and what would you do if you found them in your dataset.

An **outlier** is a data point that differs significantly from other observations.

Depending on the cause of the outlier, they can be bad from a machine learning perspective because they can worsen the accuracy of a model. If the outlier is caused by a measurement error, it's important to remove them from the dataset. There are a couple of ways to identify outliers:

Z-score/standard deviations: if we know that 99.7% of data in a data set lie within three standard deviations, then we can calculate the size of one standard deviation, multiply it by 3, and identify the data points that are outside of this range. Likewise, we can calculate the z-score of a given point, and if it's equal to +/- 3, then it's an outlier. Note: that there are a few contingencies that need to be considered when using this method; the data must be normally distributed, this is not applicable for small data sets, and the presence of too many outliers can throw off z-score.



Interquartile Range (IQR): IQR, the concept used to build boxplots, can also be used to identify outliers. The IQR is equal to the difference between the 3rd quartile and the 1st quartile. You can then identify if a point is an outlier if it is less than Q1–1.5*IRQ or greater than Q3 + 1.5*IQR. This comes to approximately 2.698 standard deviations.

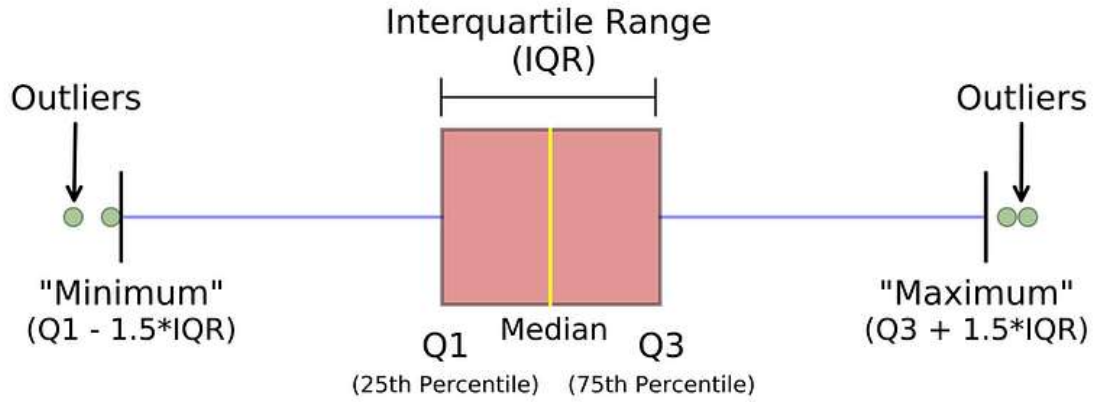


Photo from Michael Galarnyk

Other methods include DBScan clustering, Isolation Forests, and Robust Random Cut Forests.

An **inlier** is a data observation that lies within the rest of the dataset and is unusual or an error. Since it lies in the dataset, it is typically harder to identify than an outlier and requires external data to identify them. Should you identify any inliers, you can simply remove them from the dataset to address them.

9. How do you handle missing data? What imputation techniques do you recommend?

There are several ways to handle missing data:

- Delete rows with missing data
- Mean/Median/Mode imputation
- Assigning a unique value
- Predicting the missing values
- Using an algorithm which supports missing values, like random forests

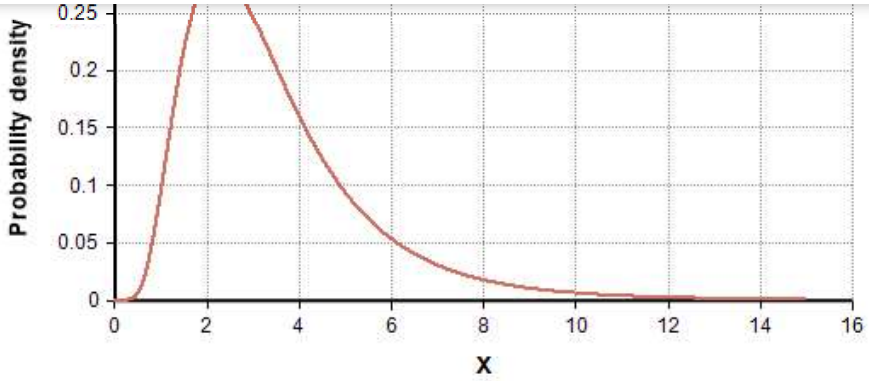
The best method is to delete rows with missing data as it ensures that no bias or variance is added or removed, and ultimately results in a robust and

start with and the percentage of missing values is low?

10. You have data on the duration of calls to a call center. Generate a plan for how you would code and analyze these data. Explain a plausible scenario for what the distribution of these durations might look like. How could you test, even graphically, whether your expectations are borne out?

First I would conduct EDA — Exploratory Data Analysis to clean, explore, and understand my data. *See my article on EDA* here. As part of my EDA, I could compose a histogram of the duration of calls to see the underlying distribution.

My guess is that the duration of calls would follow a lognormal distribution (see below). The reason that I believe it's positively skewed is because the lower end is limited to 0 since a call can't be negative seconds. However, on the upper end, it's likely for there to be a small proportion of calls that are extremely long relatively.



Lognormal Distribution Example

You could use a QQ plot to confirm whether the duration of calls follows a lognormal distribution or not. *See* here *to learn more about QQ plots*.

11. Explain likely differences between administrative datasets and datasets gathered from experimental studies. What are likely problems encountered with administrative data? How do experimental methods help alleviate these problems? What problem do they bring?

Administrative datasets are typically datasets used by governments or other organizations for non-statistical reasons.

Administrative datasets are usually larger and more cost-efficient than experimental studies. They are also regularly updated assuming that the

the data that one may want and may not be in the desired format either. It is also prone to quality issues and missing entries.

12. You are compiling a report for user content uploaded every month and notice a spike in uploads in October. In particular, a spike in picture uploads. What might you think is the cause of this, and how would you test it?

There are a number of potential reasons for a spike in photo uploads:

- A new feature may have been implemented in October which involves uploading photos and gained a lot of traction by users. For example, a feature that gives the ability to create photo albums.
- Similarly, it's possible that the process of uploading photos before was not intuitive and was improved in the month of October.
- There may have been a viral social media movement that involved uploading photos that lasted for all of October. Eg. Movember but something more scalable.
- It's possible that the spike is due to people posting pictures of themselves in costumes for Halloween.

The method of testing depends on the cause of the spike, but you would conduct hypothesis testing to determine if the inferred cause is the actual cause.

friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle?

You can tell that this question is related to Bayesian theory because of the last statement which essentially follows the structure, "What is the probability A is true **given** B is true?" Therefore we need to know the probability of it raining in London on a given day. Let's assume it's 25%.

P(A) = probability of it raining = 25% P(B) = probability of all 3 friends say that it's raining P(A|B) probability that it's raining given they're telling that it is raining P(B|A) probability that all 3 friends say that it's raining given it's raining = $(2/3)^3 = 8/27$

Step 1: Solve for P(B) P(A|B) = P(B|A) * P(A) / P(B), can be rewritten as P(B) = P(B|A) * P(A) + P(B|not A) * P(not A) P(B) = $(2/3)^3$ * 0.25 + $(1/3)^3$ * 0.75 = 0.25*8/27 + 0.75*1/27

Step 2: Solve for P(A|B) P(A|B) = 0.25 * (8/27) / (0.25*8/27 + 0.75*1/27) P(A|B) = 8 / (8 + 3) = 8/11

Therefore, if all three friends say that it's raining, then there's an 8/11 chance that it's actually raining.

14. There's one box — has 12 black and 12 red cards, 2nd box has 24 black and 24 red; if you want to draw 2 cards at random from one of the 2 boxes, which box has the higher probability of getting the same color? Can you tell intuitively why the 2nd box has a higher probability

The box with 24 red cards and 24 black cards has a higher probability of getting two cards of the same color. Let's walk through each step.

Let's say the first card you draw from each deck is a red Ace.

This means that in the deck with 12 reds and 12 blacks, there's now 11 reds and 12 blacks. Therefore your odds of drawing another red are equal to 11/(11+12) or 11/23.

In the deck with 24 reds and 24 blacks, there would then be 23 reds and 24 blacks. Therefore your odds of drawing another red are equal to 23/(23+24) or 23/47.

Since 23/47 > 11/23, the second deck with more cards has a higher probability of getting the same two cards.

15. What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?

against a random choice targeting model, in other words, it tells you how much better your model is at predicting things than if you had no model.

KPI: stands for Key Performance Indicator, which is a measurable metric used to determine how well a company is achieving its business objectives. Eg. error rate.

Robustness: generally robustness refers to a system's ability to handle variability and remain effective.

Model fitting: refers to how well a model fits a set of observations.

Design of experiments: also known as DOE, it is the design of any task that aims to describe and explain the variation of information under conditions that are hypothesized to reflect the variable. [4] In essence, an experiment aims to predict an outcome based on a change in one or more inputs (independent variables).

80/20 rule: also known as the Pareto principle; states that 80% of the effects come from 20% of the causes. Eg. 80% of sales come from 20% of customers.

16. Define quality assurance, six sigma.

desired level of quality by minimizing mistakes and defects.

Six sigma: a specific type of quality assurance methodology composed of a set of techniques and tools for process improvement. A six sigma process is one in which 99.99966% of all outcomes are free of defects.

17. Give examples of data that does not have a Gaussian distribution, nor log-normal.

- Any type of categorical data won't have a gaussian distribution or lognormal distribution.
- Exponential distributions — eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

18. What is root cause analysis? How to identify a cause vs. a correlation? Give examples

Root cause analysis: a method of problem-solving used for identifying the root cause(s) of a problem [5]

Correlation measures the relationship between two variables, range from -1 to 1. **Causation** is when a first event appears to have caused a second event. Causation essentially looks at direct relationships while correlation can look at both direct and indirect relationships.

one causes another. Instead, it's because both occur more when it's warmer outside.

You can test for causation using hypothesis testing or A/B testing.

19. Give an example where the median is a better measure than the mean

When there are a number of outliers that positively or negatively skew the data.

20. Given two fair dices, what is the probability of getting scores that sum to 4? to 8?

There are 4 combinations of rolling a 4 (1+3, 3+1, 2+2): P(rolling a 4) = 3/36 = 1/12

There are combinations of rolling an 8 (2+6, 6+2, 3+5, 5+3, 4+4): P(rolling an 8) = 5/36

21. What is the Law of Large Numbers?

The Law of Large Numbers is a theory that states that as the number of trials increases, the average of the result will become closer to the expected value.

22. How do you calculate the needed sample size?

$$ME = t * \frac{S}{\sqrt{n}} \quad -or- \quad ME = z * \frac{\sigma}{\sqrt{n}}$$

Formula for margin of error

You can use the margin of error (ME) formula to determine the desired sample size.

- t/z = t/z score used to calculate the confidence interval
- ME = the desired margin of error
- S = sample standard deviation

23. When you sample, what bias are you inflicting?

Potential biases include the following:

- **Sampling bias:** a biased sample caused by non-random sampling
- **Under coverage bias:** sampling too few observations
- **Survivorship bias:** error of overlooking observations that did not make it past a form of selection process.

There are many things that you can do to control and minimize bias. Two common things include **randomization**, where participants are assigned by chance, and **random sampling**, sampling in which each member has an equal probability of being chosen.

25. What are confounding variables?

A confounding variable, or a confounder, is a variable that influences both the dependent variable and the independent variable, causing a spurious association, a mathematical relationship in which two or more variables are associated but not causally related.

26. What is A/B testing?

A/B testing is a form of hypothesis testing and two-sample hypothesis testing to compare two versions, the control and variant, of a single variable. It is commonly used to improve and optimize user experience and marketing.

27. Infection rates at a hospital above a 1 infection per 100 person-days at risk are considered high. A hospital had 10 infections over the last 1787 person-days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard.

given timeframe, this is a Poisson distribution question.

$$P(k \text{ events in interval}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The probability of observing k events in an interval

Null (H0): 1 infection per person-days Alternative (H1): >1 infection per person-days

k (actual) = 10 infections lambda (theoretical) = (1/100)*1787 p = 0.032372 or 3.2372% *calculated using .poisson() in excel or ppois in R*

Since p-value < alpha (assuming 5% level of significance), we reject the null and conclude that the hospital is below the standard.

28. You roll a biased coin (p(head)=0.8) five times. What's the probability of getting three or more heads?

Use the General Binomial Probability formula to answer this question:

$$P(k \text{ out of } n) = \frac{n!}{k!\,(n-k)!} * p^k (1-p)^{(n-k)}$$

General Binomial Probability Formula

P(3 or more heads) = P(3 heads) + P(4 heads) + P(5 heads) = **0.94 or 94%**

29. A random variable X is normal with mean 1020 and a standard deviation 50. Calculate P(X>1200)

Using Excel… p =1-norm.dist(1200, 1020, 50, true) **p= 0.000159**

30. Consider the number of people that show up at a bus station is Poisson with mean 2.5/h. What is the probability that at most three people show up in a four hour period?

x = 3 mean = 2.5*4 = 10

using Excel…

p = poisson.dist(3,10,true) **p = 0.010336**

31. An HIV test has a sensitivity of 99.7% and a specificity of 98.5%. A subject from a population of prevalence 0.1% receives a positive test result. What is the precision of the test (i.e the probability he is HIV positive)?

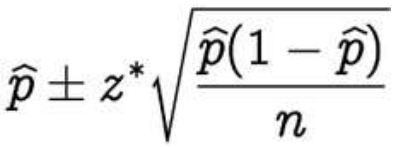
$$PV+ = \frac{\text{Prevalence} \times \text{Sensitivity}}{(\text{Prevalence} \times \text{Sensitivity}) + \{(1 - \text{Prevalence}) \times (1 - \text{Specificity})\}}$$

Precision = Positive Predictive Value = PV PV = (0.001*0.997)/[(0.001*0.997)+((1−0.001)*(1−0.985))] PV = 0.0624 or 6.24%

See more about this equation here.

32. You are running for office and your pollster polled hundred people. Sixty of them claimed they will vote for you. Can you relax?

- Assume that there's only you and one other opponent.
- Also, assume that we want a 95% confidence interval. This gives us a z-score of 1.96.



Confidence interval formula

p-hat = 60/100 = 0.6 z* = 1.96 n = 100 This gives us a confidence interval of [50.4,69.6]. Therefore, given a confidence interval of 95%, if you are okay with the worst scenario of tying then you can relax. Otherwise, you cannot relax until you got 61 out of 100 to claim yes.

- Since this is a Poisson distribution question, mean = lambda = variance, which also means that standard deviation = square root of the mean
- a 95% confidence interval implies a z score of 1.96
- one standard deviation = sqrt(115) = 10.724

Therefore the confidence interval = 115+/- 21.45 = [93.55, 136.45]. Since 99 is within this confidence interval, we can assume that this change is not very noteworthy.

35. Consider influenza epidemics for two-parent heterosexual families. Suppose that the probability is 17% that at least one of the parents has contracted the disease. The probability that the father has contracted influenza is 12% while the probability that both the mother and father have contracted the disease is 6%. What is the probability that the mother has contracted influenza?

Using the General Addition Rule in probability: P(mother or father) = P(mother) + P(father) — P(mother and father) P(mother) = P(mother or father) + P(mother and father) — P(father) P(mother) = 0.17 + 0.06–0.12 P(mother) = 0.11

36. Suppose that diastolic blood pressures (DBPs) for men aged 35–44 are normally distributed with a mean of 80 (mm

70?

Since 70 is one standard deviation below the mean, take the area of the Gaussian distribution to the left of one standard deviation.

= 2.3 + 13.6 = 15.9%

37. In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?

$$\overline{X} \pm t\frac{s}{\sqrt{n}}$$

Confidence interval for sample

Given a confidence level of 95% and degrees of freedom equal to 8, the t-score = 2.306

Confidence interval = 1100 +/- 2.306*(30/3) Confidence interval = [1076.94, 1123.06]

pounds. What would the standard deviation of the difference in weight have to be for the upper endpoint of the 95% T confidence interval to touch 0?

Upper bound = mean + t-score*(standard deviation/sqrt(sample size)) 0 = -2 + 2.306*(s/3) 2 = 2.306 * s / 3 s = 2.601903 Therefore the standard deviation would have to be at least approximately 2.60 for the upper bound of the 95% T confidence interval to touch 0.

39. In a study of emergency room waiting times, investigators consider a new and the standard triage systems. To test the systems, administrators selected 20 nights and randomly assigned the new triage system to be used on 10 nights and the standard system on the remaining 10 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 3 hours with a variance of 0.60 while the average MWT for the old system was 5 hours with a variance of 0.68. Consider the 95% confidence interval estimate for the differences of the mean MWT associated with the new system. Assume a constant variance. What is the interval? Subtract in this order (New System — Old System).

See here for full tutorial on finding the Confidence Interval for Two Independent Samples.

Use the t-table with degrees of freedom = n_1+n_2-2

Confidence Interval = mean +/- t-score * standard error *(see above)*

mean = new mean — old mean = 3–5 = -2

t-score = 2.101 given df=18 (20–2) and confidence interval of 95%

$$SE(\bar{x}_1-\bar{x}_2)=S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$S_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

standard error = sqrt((0.^{62}*9+0.^{682}*9)/(10+10–2)) * sqrt(1/10+1/10)
standard error = 0.352

confidence interval = [-2.75, -1.25]

40. To further test the hospital triage system, administrators selected 200 nights and randomly assigned a new triage system to be used on 100 nights and a standard system on the remaining 100 nights. They calculated the nightly median

hours while the average MWT for the old system was 6 hours with a standard deviation of 2 hours. Consider the hypothesis of a decrease in the mean MWT associated with the new treatment. What does the 95% independent group confidence interval with unequal variances suggest vis a vis this hypothesis? (Because there's so many observations per group, just use the Z quantile instead of the T.)

Assuming we subtract in this order (New System — Old System):

$$(\bar{x}_1 - \bar{x}_2) \pm z\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Use Z table for standard normal distribution

confidence interval formula for two independent samples

mean = new mean — old mean = 4–6 = -2

z-score = 1.96 confidence interval of 95%

$$S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

st. error = sqrt((0.52*99+22*99)/(100+100−2)) * sqrt(1/100+1/100)
standard error = 0.205061 lower bound = -2−1.96*0.205061 = -2.40192
upper bound = -2+1.96*0.205061 = -1.59808

confidence interval = [-2.40192, -1.59808]

References

[1] Central Limit Theorem, Definition and Examples in Easy Steps, *Statistics How To*

[2] Power, Statistics, *Wikipedia*

[3] Anthropic principle, *Wikipedia*

[4] Design of experiments, *Wikipedia*

[5] Root cause analysis, *Wikipedia*

> ***If you enjoyed this be sure to*** subscribe here ***or to my*** exclusive newsletter ***to never miss another article on data science guides, tricks and tips, life lessons, and more!***

More Relevant Articles

Amazon's Data Scientist Interview Practice Problems

A walkthrough of some of Amazon's interview questions!

towardsdatascience.com

Microsoft Data Science Interview Questions and Answers!

A walkthrough of some data science questions from a Microsoft Interview

towardsdatascience.com

More Microsoft Data Science Interview Questions and Answers

Another walkthrough of some data science questions from a Microsoft Interview

towardsdatascience.com

As a part of Google's Data Science Interview, they like to ask questions that they call "Problem-Solving" questions...

towardsdatascience.com

5 Common SQL Interview Problems for Data Scientists

Helping you develop your SQL skills to ace any interview

towardsdatascience.com

Data Science | Statistics | Interview | Work | Education

Pinterest ML Internship Summer 2025

Looking for a tech internship for the Summer 2025. I share my recent Interview experience, Questions, Solutions and tips.

6 min read

John Vastola

10 Must-Know Machine Learning Algorithms for Data Scientists

Machine learning is the science of getting computers to act without being explicitly programmed.”—Andrew Ng

7 min read

Youssef Hosni

13 SQL Statements for 90% of Your Data Science Tasks

Structured Query Language (SQL) is a programming language designed for managing and manipulating relational databases. It is widely used by…

15 min read

M

Maxim Gusarov

Do I need to tune logistic regression hyperparameters?

Aren’t we over-committed to optimizing the data science work we do? We are often trying to find the best combination of x, y, z variables…

Zach Quinn

3 Data Science Projects That Got Me 12 Interviews. And 1 That Got Me in Trouble.

3 work samples that got my foot in the door, and 1 that almost got me tossed out.

7 min read

Rina Mondal

Data Cleaning- Exploratory Data Analysis

Data cleaning, also known as data cleansing or data preprocessing, is a crucial step in the data preparation process that involves...

3 min read

Diogo Santos

Why Correlation Isn't Enough: Unlocking Causality with Randomized Controlled Trials

How Randomized Controlled Trials Transform Data into Actionable Insights and Drive Meaningful Change

4 min read

Noor Fatima

ensure that the assumptions of various...

3 min read


Nadeem

Understanding Data Science and Machine Learning Error Metrics

Error metrics are the cornerstone of evaluating machine learning models. They provide insights into a model's performance, weaknesses, and...

5 min read


Vikash Singh

Top Interview Questions and Answers on Bagging Algorithms Every Data Scientist Should Know

If you're preparing for a data science interview, understanding ensemble methods is a must, and Bagging (Bootstrap Aggregating) is one of...

6 min read


Abdur Rahman

Python is No More The King of Data Science

5 Reasons Why Python is Losing Its Crown

7 min read