**Cracking Statistics Interviews for Data Scientists**

Breaking down the necessary knowledge and question formatting for statistics interviews

In this blog posrt, we are covering another type of interview that an aspiring data scientist has to learn to crack: the statistics interview. A statistics interview is a**technical interview** that evaluates your ability to actually do your job as a data scientist. If you do not ace this interview, you will have a hard time convincing a company that you are a capable data scientist.

The first step to acing any interview is **knowing what to expect**. Statistics is a broad subject, so this post will break down what areas of statistics a data scientist is expected to know for interviews and what types of questions will be asked. We will also discuss tips and general advice throughout.

I also have [a video](https://www.youtube.com/watch?v=65iKT1s9RHw) on this subject if you would prefer to get the information that way!

Let’s start with what areas of statistical knowledge a statistics interview will cover.

**Table of Contents**

* [Areas of Knowledge](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#a48d)  
  - [Probability](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#2f4a)  
  - [Hypothesis Testing](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#fa8e)  
  - [Regression](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#538b)
* [Types of Questions](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#cd58)  
  - [Conceptual Questions](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#da5a)  
  - [Questions Involving Calculations](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#e7c1)  
  - [Coding Questions](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#9d7a)
* [Conclusion](https://towardsdatascience.com/cracking-statistics-interviews-for-data-scientists-711a63bcb375#bafa)

**Areas of Knowledge**

In a statistics interview, you naturally have to demonstrate a lot of technical knowledge, but that does not mean that you need to be prepared to answer any and all questions related to statistics. For data scientists, a statistics interview will typically focus on three areas of knowledge:

* **Probability**
* **Hypothesis Testing**
* **Regression**

**Probability**

For probability questions, you should expect to be asked about probability basics, conditional probability, and probability distribution. **Probability basics** includes expectation, variance, permutation, and combinations, etc.

For **conditional probability**, you will need to know Bayes’ rule, and for **probability** **distribution**, you need to be familiar with some commonly used discrete and continuous distributions, such as binomial, normal, and long-tailed distribution.

This may sound like a lot, but probability is typically the area of knowledge for which you will need the most in-depth knowledge for interviews. There are therefore a lot of things the interviewer could ask you about, and besides, it is always better to over-prepare than to under-prepare.

**Hypothesis Testing**

Moving on to hypothesis testing, for statistics interviews you need to be aware of both the **terminology**, such as power, p-value, and confidence interval, and the different kinds of **testing methods**, parametric tests, such as z-test, t-test, and non-parametric tests, such as chi-squared test, and more.

**Regression**

The final area of knowledge you need for statistics interviews is regression. For regression, you need to be familiar with**linear and multiple regression**. Of the three areas, regression will likely appear the least in interviews, but you still need to make sure you are comfortable with it.

So probability, hypothesis testing, and regression are the areas on which you will be questioned in an interview. Now that we know what subjects the questions will cover, let’s discuss the types of questions you can expect.

**Types of Questions**

There are three different formats that questions take in a statistics interview:

* **Conceptual Questions**
* **Questions Involving Calculations**
* **Coding Questions**

Understanding what types of questions you will be asked is crucial in understanding how to prepare and how to answer. Let’s take a closer look at each type.

**Conceptual Questions**

As the name suggests, conceptual questions are more interested in your ability to**explain concepts** than your ability to do math with those concepts. We can further break this category down into two styles: explaining to a **technical audience** and explaining to a **non-technical audience.**

Before we dive into the different approaches for explaining concepts to a technical versus a non-technical audience, what exactly are conceptual questions? Let’s look at some examples of conceptual questions from each of the areas of knowledge you can expect in a statistics interview.

* Probability: What is the distribution of average time spent per user?
* Hypothesis Testing: Explain p-value and confidence interval to a non-technical audience.
* Regression: What are the assumptions of linear regression?

As you can see from the examples, conceptual questions ask you to define terms and show an understanding of what these things mean in a real-world context. You will likely be asked to explain a term or concept, sometimes for a non-technical audience. Explaining a concept is a bit of a vague request though, so what should you include in your answers to these conceptual questions?

**Explaining to a Technical Audience**

Let’s start by looking at some steps for explaining concepts to a technical audience. It can be easy to dismiss the importance of this as a data scientist. After all, if the audience has a technical background, they should understand your explanation with no problems right?

While a technical audience should have an easier time following the terminology and general idea of your explanations, if your answer is not organized or if it deals with a more obscure concept, it can still be difficult to follow. You still need to take steps to ensure that you can explain the concepts clearly. Here are the steps I recommend:

1. **Start with some context.** When or where is this terminology used?
2. **Define that concept**. Even when explaining to a technical person, you want to keep the definition easy to understand. Try to not sound like a high-level textbook. Your ability to explain things in simple terms shows a higher level of understanding.
3. For concepts that can be represented by numbers, you might want to **explain what changes in value mean**. What does it mean when this concept has a larger or smaller value?
4. This step is optional. You can finish by talking about**how this concept is applied in practice.** Think about questions such as why is this concept widely used or why is it important to data science.

To see these steps in action, I recommend checking out [one of my videos](https://www.youtube.com/watch?v=Allap_hrjyo) on how to explain the top 5 statistical concepts in interviews.

**Explaining to a Non-Technical Audience**

These steps hopefully give you some clear talking points and structure when you are asked to explain or define a concept to a technical audience, but how does a non-technical audience differ? As you saw earlier in our examples, you may be asked to explain a concept in layman terms or to a non-technical audience, this requires you to explain things more intuitively.

Using**examples and analogy**is a great way to explain terminology to a non-technical audience. Try to make connections to things that a layman would be more familiar with to explain what is unfamiliar.

It is also crucial to **avoid using technical terms** when explaining things to a non-technical audience. For example, If you use terms like hypothesis testing, null hypothesis, or alternative hypothesis when explaining the concept of the power of a test, you will only confuse your audience.

With all conceptual questions, the goal should be to **keep your explanations clear and structured.** Remember that even with a technical audience, you want your explanation to be as easy to understand as possible as this shows a deeper understanding of the concepts.

**Questions Involving Calculations**

Understanding the concepts and terminology is wonderful and necessary for a data scientist, but you also have to be able to **do the math that goes along with those concepts**. A statistics interview will include questions that involve actual calculations.

Questions involving calculations could require you to simply know what to use to solve a problem. For example, a conditional probability question might require you to use Bayes’ rule to calculate it, but the question itself may not mention Bayes’ rule. This is evaluating if you can correctly identify what methods and steps you need to solve a particular problem, basically if you know HOW to do the problem.

A question involving calculations could go much further though. It could ask you to write the equation and provide the exact answer. This would be evaluating not only if you know how to do the problem but also if you actually can do the math correctly.

For example, the problem might say “We have a total of 100 coins, which includes 99 fair coins and 1 biased coin that has a probability of getting heads 100%. If you choose a random coin and flip it 10 times and all 10 times are heads, what’s the probability that the coin is the biased coin?” This problem not only requires you to know that you need to use Bayes’ rule, but you also have to find the numerical answer here.

Some more examples of questions that involve calculations would be:

* “What’s the probability of getting two heads among 10 tosses of a fair coin?”
* “Given two groups of users, compare the click-through rates and draw conclusions as to whether the two click-through rates are the same”.
* “Can you lay out the testing steps and draw conclusions?”

The first question is an example of a question dealing with probability distribution, and the second is hypothesis testing. The final example is something you might be asked after getting the results of another question. I recommend [this video](https://www.youtube.com/watch?v=IY7y-t30UJc&list=PLY1Fi4XflWStljP1tzfAfU_Qn0wHzhzYm&index=3) for a better look at hypothesis testing questions.

To summarize, questions involving calculations are thus one way that the interviewer evaluates whether you can deliver on your knowledge. These questions are an opportunity to show that you have not only knowledge but **skills** as well. The final type of question, the coding question, gives you even more opportunity to demonstrate your skills.

**Coding Questions**

If conceptual questions are all about your understanding of concepts then coding questions are all about your **implementation skills**. These questions will require you to not only know about theories but to also implement those theories.

For example, a coding question that deals with probability might ask you to design and run a coin simulation problem. Check out [this video](https://www.youtube.com/watch?v=-SANBbv0-Hw&list=PLY1Fi4XflWStljP1tzfAfU_Qn0wHzhzYm&index=5) for a more in-depth look at this type of problem. Another example would be a hypothesis testing coding question, which will likely have you writing code in R or Python to calculate the results.

These examples and indeed all coding questions, let you demonstrate an ability to get results. Understanding what to do is one thing, but **showing that you can do it**is crucial for acing an interview. Of course, the best way to prepare for coding questions is to**practice coding**. I recommend the 10 days of statistics by HackerRank to get started on some statistics problems.

**Conclusion**

Generally speaking, statistics interviews are a very skill-based type of interview. You need to be prepared to show thorough knowledge and capability in these interviews.

So to recap, there are three areas of statistical knowledge you will need to prepare for in interviews as a data scientist:

* **Probability**
* **Hypothesis Testing**
* **Regression**

You may notice that we did not talk about causal inference, even though it is a domain closely related to statistics. If you would like to learn about causal inference in data science interviews, make sure to subscribe to [this channel](https://www.youtube.com/channel/UCAWsBMQY4KSuOuGODki-l7A) to get updates on future content.

Besides the three areas of knowledge there are also three types of questions you can expect in a statistics interview:

* **Conceptual questions**
* **Questions involving calculations**
* **Coding questions**

All together these three types of questions evaluate both your understanding and skill with statistical problems you would face as a data scientist.

With this outline, you now know what you need to know and prepare to ace statistics interviews as a data scientist. Good luck!