STATISTICS INTERVIEW CHEAT SHEET

EMMA ding

EMMA ding

emmading.com

info@datainterviewpro.com

Statistics Interview Cheat Sheet

This summary walks you through some of the **most common statistics interview problems** and gives you steps for addressing them. Examples in this summary will help you understand what to expect in a statistics interview and give you some practice addressing each question type.

There are three main question types in a statistics interview:

(1) Conceptual questions

(2) Questions involving calculations

(3) Implementation questions

> 🎉 By cracking questions in this cheat sheet, you are able to solve over **40%** of statistics interview questions!

Part 1: Conceptual Questions

Conceptual questions are more about your ability to explain concepts clearly than your ability to do math calculations. We can further break this category down into two response styles: explaining to a **technical** audience and explaining to a **non-technical** audience.

📌 Steps to explain to a technical audience (e.g., a data scientist and an engineer):

1. Start with some **context**. For example, when or where is this terminology used?
2. Provide a **definition** of the concept. Even when explaining a concept to a technical person, you want to keep the definition easy to understand. Try NOT to sound like Wikipedia or an advanced textbook. Your ability to explain things in simple terms actually shows a higher level of understanding.
3. Next, for concepts that can be represented by numbers, you might want to explain what **changes** in a particular value mean. For example, what does a higher p-value mean?
4. The final step is optional. You can finish your response by talking about how this concept is applied in practice. Think about questions such as "Why is this concept widely used?" or "Why is this concept important to data science?"

📍 How to explain to a non-technical audience:

1. Use **examples** and **analogies**, which are 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.
2. 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.
3. As with all conceptual questions, the goal should be to keep your explanations clear and structured.

💡 Below is a list of conceptual questions ranked by frequency. We'll provide sample answers to explain those concepts in a technical way, followed by an explanation to a non-technical audience.

▼ P-value

It's commonly used in hypothesis testing to connect the dots between observation and conclusion. It is a conditional probability that measures the probability of obtaining results at least as extreme as those observed in the given sample, given that the null hypothesis is true. When we say "at least as extreme," we mean "containing at least as much evidence in favor of the alternative hypothesis."

A low p-value indicates less support for the null hypothesis. In practice, we often choose 0.05 as the cutoff value. P-value < 0.05 denotes fairly strong evidence against the null hypothesis which means the null hypothesis can be rejected. P-value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected. P-value equals 0.05 is the marginal value indicating it is possible to go either way, but this value is rarely encountered in practice.

The p-value is commonly used in A/B testing when we have a treatment and a control group and we want to test whether a metric is different between those groups. Now suppose we run the experiment and observe different metric values in each group. The smaller the p-value is, the more we are convinced that there is a difference between the two.

▼ **Explain the p-value to a non-technical audience**

Your friend claims that the average height of adults in your town is 175 cm and you decide to gather some data to see if he's right. You randomly select 30 adults from the town, record their height, and average the measurements. The average probably won't be exactly 175 cm, and if it's at all close, your friend will probably claim that the difference is just due to random chance.

The p-value allows you to quantify how likely your friend's counter-argument is. Could the difference really be due to chance, or is his explanation unlikely based on the size of the difference and the number of individuals you analyzed?

Let's imagine that the average height in your sample is 172 cm. Then the p-value has the following interpretation: given the average height in the population really is 175cm, the p-value is the probability of sampling 30 individuals with an average height that differs from 175 cm by 3 cm or more. A very small p-value means that your data is very unlikely if the average height in the population is 175 cm; therefore, it's more plausible that the average height in the population is NOT, in fact, 175 cm.

▼ Type I Error and Type II Error

Type I error refers to the situation in which we conclude there is a difference when, in fact, there isn't. If we use A/B testing as an example, a type I error occurs when we conclude—based on experimental data—that the two groups differ when, in reality, they don't.

Holding all else constant, we would prefer a test with a lower type I error rate because that would mean we make fewer mistakes when there are no underlying differences.

A type II error occurs when we mistakenly accept a false null hypothesis; i.e., we conclude that the observed differences are not statistically meaningful when, in fact, there is a real systematic difference between groups.

Holding all else constant, we would prefer a test with a lower type II error rate because that would mean we are more likely to find real differences.

Tip to remember these two concepts: As explained in this Cross Validated post - since "false" and "negative" have similar meanings, a type II error is a "false negative" or "false false," because it contains two falses. By comparison, a Type I error is a "false positive" that has only one "false" in it.

▼ Power

Statistical power is used in a binary hypothesis test. It is the probability that a test correctly rejects the null hypothesis when the alternative hypothesis is true. To put it in another way, statistical power is the likelihood that a test will detect an effect when the effect is present.

The higher the statistical power, the better the test is. It is commonly used in experimental design, to calculate the minimum sample size required so that one can reasonably detect an effect.

✅ We've just explained Type I Error, Type II Error, and power in a technical way; let's now describe them to a **non-technical audience**. The key is to use an intuitive example to explain them. Below is an example; feel free to come up with your own!

A person wants to test if he is infected by COVID, so we can break the problem into two scenarios.

In the first scenario, he really does have COVID. In this case, the power is the probability that his test comes back positive. In contrast, the type II error rate is the probability his test comes back negative. A type II error is problematic because he should quarantine and seek out medical treatment, but he doesn't know that he needs to.

In the second scenario, he does not have COVID. In this scenario, the type I error rate is the probability that his test comes back positive even though he doesn't have COVID. A type I error is problematic because he needlessly quarantines and seeks out medical treatment when he doesn't need to.

▼ What are the assumptions of linear regression?

There are 4 assumptions of linear regression, the first assumption of linear regression is about the relationship between x and y, the independent and dependent variables, and the remaining three assumptions are about the errors or residuals. You can remember them with the acronym **LINE** (technically all of the below statements must be true conditional on x):

- **L: The mean of y is linear in x (potentially including transformations of x, such as x2)**
- **I:** The residuals are statistically **independent**
- **N:** The residuals are **normally** distributed; this one is less important in large samples
- **E:** The residuals have **equal variance** (homoscedasticity)

▼ What are the differences between t-tests vs z-tests? How to choose which test to use?

Both t-tests and z-tests can be used to test whether a population mean is equal to a particular value or whether the means from two different populations are equal.

With a z-test, your test statistic follows a normal distribution under the null hypothesis. With a t-test, by contrast, the test statistic follows the student t-distribution because it involves estimating the population variance.

Considering the population mean, we can use either the z-test or the t-test only if the sample mean is normally distributed, which is possible in two cases: the initial population is normally distributed, or the sample size is large enough (n > 30) that we can apply the Central Limit Theorem.

If the condition above is satisfied, then we need to decide which type of test is more appropriate to use. In general, we use Z-tests if the population variance is known and t-tests when it is unknown. In practice, we usually don't know the population variance, so t-tests are more common.

Additionally, if the sample size is very large (n > 200), we can use the Z-test even if we don't know the population variance. This approach is reasonable because with a large sample size, the degrees of freedom for the t-distribution will be so large that the distribution will almost perfectly resemble a normal distribution, so the z-test (using the sample variance in place of the population variance) will be almost exactly equivalent to the t-test.

Considering the population proportion, we can use a Z-test (but not a t-test) when each of the numbers of successes and the number of failures is at least 10.

▼ Central Limit Theorem (CLT)

The CLT states that the sample means of a sufficiently large number of iterates of independent random variables will be approximately normally distributed regardless of the underlying distribution. There are two important assumptions of the CLT: one is that the random variables are independent and identically distributed (i.i.d.). And the second is that the variables have finite variance.

$$\text{sample mean} \sim N(\mu, \frac{\sigma^2}{n})$$

So basically, under some independence and variance assumptions, the sampling distribution of the means follows a normal distribution no matter what the underlying distribution of the population is.

▼ Confidence Interval

A confidence Interval is used when we want to quantify how confident we are about a given estimate. The confidence interval is for the true value but we never know what the true value is. That's why we gather data: to create a reasonable estimate of the true—but unknown—value.

The confidence interval is a range of numbers with an accompanying confidence level. The confidence level is the probability that a confidence interval generated from a new (but identically distributed) data set will contain the true value. Most data scientists like to use a confidence level of 95%, but others values are common too.

Higher confidence levels require wider confidence intervals. Gathering more data will typically make your confidence intervals narrower because the additional samples give you more information about the true value.

▼ Explain confidence intervals to a non-technical audience

Confidence intervals measure the degree of uncertainty in an estimate. For example, suppose we want to estimate the average height of all men in the U.S. We can estimate this value by randomly selecting a sample of, say, 30 men and recording their heights. Then we can use the average of their 30 heights as an estimate of the average height of all men in the U.S.

Of course, our estimate won't be perfect, but that's where confidence intervals come in. Let's say our estimate is 178 cm. Then we might generate a 95% confidence interval of 170 to 186 cm. The number 95% means that if we repeated the procedure many times, then 95% of the intervals would contain the true average height across all men in the U.S.

▼ Covariance vs. Correlation Coefficient

Both of them measure the linear relationship between two variables. The unit of covariance is obtained by multiplying the units of the two variables, but the correlation coefficient has no units; you would get the same correlation coefficient if you used different units for the original variables (as long as the new units are linearly related to the old units). The absolute value of the covariance is no greater than the product of the individual standard deviations. The correlation coefficient, on the other hand, is always between -1 and 1.

	Covariance	Correlation Coefficient
Measure	The **direction** of the linear relationship between two variables.	The **strength** of the linear relationship between two variables.
Equation	$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$	$r_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
Unit	Product of the units of the two variables	None
Range	$[-SD(X)SD(Y), SD(X)SD(Y)]$	[-1, 1]

Part 2: Calculation Questions

🌈 Over 80% of calculation questions involve combinatorics and probability. The equations below are helpful to review before interviews because they can be used to answer most calculation questions.

Combinatorics

▼ **Rule of Sum**

- Rule for counting things using addition
- If there are *m* ways to arrange *a*, *n* ways to arrange *b*, and *a* and *b cannot* happen simultaneously, then the number of ways to arrange *a* or *b* is *m* + *n*:

$$\text{Ways to do a or b} = m + n$$

▼ **Rule of Product**

- Rule for counting arrangements using multiplication
- Given *a* and *b*, which can happen simultaneously, and n ways to do *a* and *m* ways to do *b,* then *n* x *m* ways exist to do both:

$$\text{Ways to do a and b} = n * m$$

Probability

▼ **Bayes' Theorem (Bayes' rule)**

Bayes' rule is one of the most important rules in probability. It deals with conditional probabilities and provides a rule for changing existing beliefs based on new data.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

- A, B are events
- $P(A|B)$: the probability of A occurring given that B is true
- $P(B|A)$: the probability of B occurring given that A is true
- $P(A)$, $P(B)$: independent probabilities of A and B

▼ **Expectation**

The expected value of a random variable with a finite number of outcomes is a weighted average of all possible outcomes.

$$E[X] = \sum_{n=1}^{\infty} x_n * P(X = x_n)$$

▼ **Binomial Distribution**

Number of successes among n independent and identically distributed Bernoulli trials. Each trial has the same probability of getting a success p. It is commonly used to model the number of successes in a sample of size n drawn with replacement from a population of size n.

$$P(X = k) = \binom{N}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$E[X] = np, SD(X) = \sqrt{np(1-p)}$$

Part 3: Implementation Questions

Most implementation questions are about t-tests, so it's essential to review one-sample and two-sample t-tests as well as Welch's t-test.

▼ One-Sample T-test

Let $X_1, ..., X_n$ be a small random sample $(n \leq 30)$ sample from a normal population with mean μ.

If the population of differences is approximately normal then we can construct our t-statistic. Under H_0

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Where $\bar{X}$ is the sample mean, μ_0 is a constant, and s is the sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$.

A $(1-\alpha) * 100\%$ confidence interval for μ is:

$$\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

If the sample size is large, we could instead use: $\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

▼ Two-Sample T-test with Equal Variances

When the two groups variances are believed to be equal, we use a t-test based on a pooled variance estimate.

▼ t-statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2,\alpha/2}$$

Where $\bar{X}_i$ is the sample mean of sample i, for $i \in \{1, 2\}$, and the pooled sample variance s_p^2 is defined by $s_p^2 = \frac{SS_1+SS_2}{n_1+n_2-2}$ with $SS_i = \sum_{j=1}^{n_i}\left(X_j - \bar{X}_i\right)^2$.

▼ **Confidence interval**

A level $(1-\alpha)*100\%$ confidence interval for the difference between two means:

$$\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2,\alpha/2} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

▼ Welch's T-test with Unequal Variances

If the two standard deviations are not similar (one is more than twice of the other) but the other assumptions for the 2-sample t-test hold, then we can use Welch's t-test, which employs an unpooled standard error.

▼ **t-statistic**

Unpooled standard error: $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{df,\alpha/2}$$

where $s_1 = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{1i}-\bar{X_1}\right)^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{2i}-\bar{X_2}\right)^2}{n_2-1}}$

and $df = \frac{(s_1^2/n_1+s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1)+(s_2^2/n_2)^2/(n_2-1)}$.

▼ **Confidence interval**

A level $(1-\alpha)*100\%$ confidence interval for the difference between two means:

$$\bar{X}_1 - \bar{X}_2 \pm t_{df,\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

where $s_1 = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{1i}-\bar{X_1}\right)^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{2i}-\bar{X_2}\right)^2}{n_2-1}}$

and $df = \frac{(s_1^2/n_1+s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1)+(s_2^2/n_2)^2/(n_2-1)}$.