Statistics

CHAPTER 6

Statistics is a core component of any data scientist's toolkit. Since many commercial layers of a data science pipeline are built from statistical foundations (for example, A/B testing), knowing foundational topics of statistics is essential.

Interviewers love to test a candidate's knowledge about the basics of statistics, starting with topics like the Central Limit Theorem and the Law of Large Numbers, and then progressing on to the concepts underlying hypothesis-testing, particularly p-values and confidence intervals, as well as Type 1 and Type II errors and their interpretations. All of those topics play an important role in the statistical underpinning of A/B testing. Additionally, derivations and manipulations involving random variables of various probability distributions are also common, particularly in finance interviews. Lastly, a common topic in more technical interviews will involve utilizing MLE and/or MAP.

Topics to Review Before Your Interview

Properties of Random Variables

For any given random variable X, the following properties hold true (below we assume X is continuous, but it also holds true for discrete random variables).

The expectation (average value, or mean) of a random variable is given by the integral of the value of X with its probability density function (PDF) fx (x):

$$\mu = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

and the variance is given by:

$$Var(X) = E\left[\left(X - E[X]\right)^2\right] = E\left[X^2\right] - \left(E[X]\right)^2$$

The variance is always non-negative, and its square root is called the standard deviation, which is heavily used in statistics.

$$\sigma = \sqrt{Var(X)} = \sqrt{E\left[\left(X - E[X]\right)^2\right]} = \sqrt{E\left[X^2\right] - \left(E[X]\right)^2}$$

The conditional values of both the expectation and variance are as follows. For example, consider the case for the conditional expectation of X, given that $Y = y$:

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx$$

For any given random variables X and Y, the covariance, a linear measure of relationship between the two variables, is defined by the following:

$$Cov(X,Y) = E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right] = E[XY] - E[X]E[Y]$$

and the normalization of covariance, represented by the Greek letter ρ, is the correlation between X and Y:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

All of these properties are commonly tested in interviews, so it helps to be able to understand the mathematical details behind each and walk through an example for each.

For example, if we assume X follows a Uniform distribution on the interval $[a, b]$, then we have the following:

$$f_X(x) = \frac{1}{b-a}$$

Therefore the expectation of X is:

$$E[X] = \int_a^b x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}$$

Although it is not necessary to memorize the derivations for all the different probability distributions, you should be comfortable deriving them as needed, as it is a common request in more technical interviews. To this end, you should make sure to understand the formulas given above and be able to apply them to some of the common probability distributions like the exponential or uniform distribution.

Law of Large Numbers

The Law of Large Numbers (LLN) states that if you sample a random variable independently a large number of times, the measured average value should converge to the random variable's true expectation. Stated more formally,

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \rightarrow \mu, \text{ as } n \rightarrow \infty$$

This is important in studying the longer-term behavior of random variables over time. As an example, a coin might land on heads 5 times in a row, but over a much larger n we would expect the proportion

of heads to be approximately half of the total flips. Similarly, a casino might experience a loss on any individual game, but over the long run should see a predictable profit over time.

Central Limit Theorem

The Central Limit Theorem (CLT) states that if you repeatedly sample a random variable a large number of times, the distribution of the sample mean will approach a normal distribution regardless of the initial distribution of the random variable.

Recall from the probability chapter that the normal distribution takes on the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

with the mean and standard deviation given by μ and σ respectively.

The CLT states that: $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \rightarrow\sim N\left(\mu, \frac{\sigma^2}{n}\right)$; hence $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

The CLT provides the basis for much of hypothesis testing, which is discussed shortly. At a very basic level, you can consider the implications of this theorem on coin flipping: the probability of getting some number of heads flipped over a large n should be approximately that of a normal distribution. Whenever you're asked to reason about any particular distribution over a large sample size, you should remember to think of the CLT, regardless of whether it is Binomial, Poisson, or any other distribution.

Hypothesis Testing

General Setup

The process of testing whether or not a sample of data supports a particular hypothesis is called hypothesis testing. Generally, hypotheses concern particular properties of interest for a given population, such as its parameters, like μ (for example, the mean conversion rate among a set of users).

The steps in testing a hypothesis are as follows:

1. State a null hypothesis and an alternative hypothesis. Either the null hypothesis will be rejected (in favor of the alternative hypothesis), or it will fail to be rejected (although failing to reject the null hypothesis does not necessarily mean it is true, but rather that there is not sufficient evidence to reject it).
2. Use a particular test statistic of the null hypothesis to calculate the corresponding p-value.
3. Compare the p-value to a certain significance level α.

Since the null hypothesis typically represents a baseline (e.g., the marketing campaign did not increase conversion rates, etc.), the goal is to reject the null hypothesis with statistical significance and hope that there is a significant outcome.

Hypothesis tests are either one- or two-tailed tests. A one-tailed test has the following types of null and alternative hypotheses:

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu < \mu_0 \text{ or } H_1 : \mu > \mu_0$$

whereas a two-tailed test has these types: $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis, and μ is the parameter of interest.

Understanding hypothesis testing is the basis of A/B testing, a topic commonly covered in tech companies' interviews. In A/B testing, various versions of a feature are shown to a sample of different users, and each variant is tested to determine if there was an uplift in the core engagement metrics.

Say, for example, that you are working for Uber Eats, which wants to determine whether email campaigns will increase its product's conversion rates. To conduct an appropriate hypothesis test, you would need two roughly equal groups (equal with respect to dimensions like age, gender, location, etc.). One group would receive the email campaigns and the other group would not be exposed. The null hypothesis in this case would be that the two groups exhibit equal conversion rates, and the hope is that the null hypothesis would be rejected.

Test Statistics

A test statistic is a numerical summary designed for the purpose of determining whether the null hypothesis or the alternative hypothesis should be accepted as correct. More specifically, it assumes that the parameter of interest follows a particular sampling distribution under the null hypothesis.

For example, the number of heads in a series of coin flips may be distributed as a binomial distribution, but with a large enough sample size, the sampling distribution should be approximately normally distributed. Hence, the sampling distribution for the total number of heads in a large series of coin flips would be considered normally distributed.

Several variations in test statistics and their distributions include:

1. *Z*-test: assumes the test statistic follows a normal distribution under the null hypothesis
2. *t*-test: uses a student's t-distribution rather than a normal distribution
3. Chi-squared: used to assess goodness of fit, and to check whether two categorical variables are independent

Z-Test

Generally the *Z*-test is used when the sample size is large (to invoke the CLT) or when the population variance is known, and a *t*-test is used when the sample size is small and when the population variance is unknown. The *Z*-test for a population mean is formulated as:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

in the case where the population variance σ^2 is known.

t-Test

The *t*-test is structured similarly to the *Z*-test, but uses the sample variance s^2 in place of population variance. The *t*-test is parametrized by the degrees of freedom, which refers to the number of independent observations in a dataset, denoted below by $n - 1$:

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

where $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

As stated earlier, the t-distribution is similar to the normal distribution in appearance but has larger tails (i.e., extreme events happen with greater frequency than the modeled distribution would predict), a common phenomenon, particularly in economics and Earth sciences.

Chi-Squared Test

The Chi-squared test statistic is used to assess goodness of fit, and is calculated as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed value of interest and E_i is its expected value. A Chi-squared test statistic takes on a particular number of degrees of freedom, which is based on the number of categories in the distribution.

To use the squared test to check whether two categorical variables are independent, create a table of counts (called a contingency table), with the values of one variable forming the rows of the table and the values of the other variable forming its columns, and check for intersections. It uses the same style of Chi-squared test statistic as given above.

Hypothesis Testing for Population Proportions

Note that, due to the CLT, the *Z*-test can be applied to random variables of any distribution. For example, when estimating the sample proportion of a population having a characteristic of interest, we can view the members of the population as Bernoulli random variables, with those having the characteristic represented by "1s" and those lacking it represented by "0s". Viewing the sample proportion of interest as the sum of these Bernoulli random variables divided by the total population size, we can then compute the sample mean and variance of the overall proportion, about which we can form the following set of hypotheses:

$$H_0 : \hat{p} = p_0 \text{ versus } H_1 : \hat{p} \neq p_0$$

and the corresponding test statistic to conduct a *Z*-test would be: $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

In practice, these test statistics form the core of A/B testing. For instance, consider the previously discussed case, in which we seek to measure conversion rates within groups A and B, where A is the control group and B has the special treatment (in this case, a marketing campaign). Adopting the same null hypothesis as before, we can proceed to use a *Z*-test to assess the difference in empirical population means (in this case, conversion rates) and test its statistical significance at a predetermined level.

When asked about A/B testing or related topics, you should always cite the relevant test statistic and the cause of its validity (usually the CLT).

p-values and Confidence Intervals

Both p-values and confidence intervals are commonly covered topics during interviews. Put simply, a p-value is the probability of observing the value of the calculated test statistic under the null hypothesis assumptions. Usually, the p-value is assessed relative to some predetermined level of significance (0.05 is often chosen).

In conducting a hypothesis test, an α, or measure of the acceptable probability of rejecting a true null hypothesis, is typically chosen prior to conducting the test. Then, a confidence interval can also be calculated to assess the test statistic. This is a range of values that, if a large sample were taken, would contain the parameter value of interest $(1-\alpha)\%$ of the time. For instance, a 95% confidence interval would contain the true value 95% of the time. If 0 is included in the confidence intervals, then we cannot reject the null hypothesis (and vice versa).

The general form for a confidence interval around the population mean looks like the following, where the term is the critical value (for the standard normal distribution):

$$\mu \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

In the prior example with the A/B testing on conversion rates, we see that the confidence interval for a population proportion would be

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

since our estimate of the true proportion will have the following parameters when estimated as approximately Gaussian:

$$\mu = \frac{np}{n} = p, \sigma^2 = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

As long as the sampling distribution of a random variable is known, the appropriate p-values and confidence intervals can be assessed.

Knowing how to explain p-values and confidence intervals, in technical and nontechnical terms, is very useful during interviews, so be sure to practice these. If asked about the technical details, always remember to make sure you correctly identify the mean and variance at hand.

Type I and II Errors

There are two errors that are frequently assessed: type I error, which is also known as a "false positive," and type II error, which is also known as a "false negative." Specifically, a type I error is when one rejects the null hypothesis when it is correct, and a type II error is when the null hypothesis is not rejected when it is incorrect.

Usually $1-\alpha$ is referred to as the confidence level, whereas $1-\beta$ is referred to as the power. If you plot sample size versus power, generally you should see a larger sample size corresponding to a larger power. It can be useful to look at power in order to gauge the sample size needed for detecting a significant effect. Generally, tests are set up in such a way as to have both $1-\alpha$ and $1-\beta$ relatively high (say at 0.95 and 0.8, respectively).

In testing multiple hypotheses, it is possible that if you ran many experiments — even if a particular outcome for one experiment is very unlikely — you would see a statistically significant outcome at least once. So, for example, if you set $\alpha = 0.05$ and run 100 hypothesis tests, then by pure chance you would expect 5 of the tests to be statistically significant. However, a more desirable outcome is to have the overall α of the 100 tests be 0.05, and this can be done by setting the new α to α/n, where n is the number of hypothesis tests (in this case, $\alpha/n = 0.05/100 = 0.0005$). This is known as Bonferroni correction, and using it helps make sure that the overall rate of false positives is controlled within a multiple testing framework.

Generally, most interview questions concerning Type I and II errors are qualitative in nature — for instance, requesting explanations of terms or of how you would go about assessing errors/power in an experimental setup.

MLE and MAP

Any probability distribution has parameters, so fitting parameters is an extremely crucial part of data analysis. There are two general methods for doing so. In maximum likelihood estimation (MLE), the goal is to estimate the most likely parameters given a likelihood function: $\theta_{MLE} = \arg\max L(\theta)$, where $L(\theta) = f_n(x_1 \ldots x_n|\theta)$.

Since the values of X are assumed to be i.i.d., then the likelihood function becomes the following:

$$L(\theta) = \prod_{i-1}^{n} f(x_i \mid \theta)$$

The natural log of $L(\theta)$ is then taken prior to calculating the maximum; since log is a monotonically increasing function, maximizing the log-likelihood log $L(\theta)$ is equivalent to maximizing the likelihood:

$$\log L(\theta) = \sum_{i-1}^{n} \log f(x_i \mid \theta)$$

Another way of fitting parameters is through maximum a posteriori estimation (MAP), which assumes a "prior distribution:"

$$\theta_{MAP} = \arg\max g(\theta) f(x_1 \ldots x_n|\theta)$$

where the similar log-likelihood is again employed, and $g(\theta)$ is a density function of θ.

Both MLE and MAP are especially relevant in statistics and machine learning, and knowing these is recommended, especially for more technical interviews. For instance, a common question in such interviews is to derive the MLE for a particular probability distribution. Thus, understanding the above steps, along with the details of the relevant probability distributions, is crucial.

40 Real Statistics Interview Questions

Easy

6.1. Uber: Explain the Central Limit Theorem. Why it is useful?

6.2. Facebook: How would you explain a confidence interval to a non-technical audience?

6.3. Twitter: What are some common pitfalls encountered in A/B testing?

6.4. Lyft: Explain both covariance and correlation formulaically, and compare and contrast them.

6.5. Facebook: Say you flip a coin 10 times and observe only one heads. What would be your null hypothesis and p-value for testing whether the coin is fair or not?

6.6. Uber: Describe hypothesis testing and p-values in layman's terms?

6.7. Groupon: Describe what Type I and Type II errors are, and the trade-offs between them.

6.8. Microsoft: Explain the statistical background behind power.

6.9. Facebook: What is a Z-test and when would you use it versus a t-test?

6.10. Amazon: Say you are testing hundreds of hypotheses, each with t-test. What considerations would you take into account when doing this?

Medium

6.11. Google: How would you derive a confidence interval for the probability of flipping heads from a series of coin tosses?

6.12. Two Sigma: What is the expected number of coin flips needed to get two consecutive heads?

6.13. Citadel: What is the expected number of rolls needed to see all six sides of a fair die?

6.14. Akuna Capital: Say you're rolling a fair six-sided die. What is the expected number of rolls until you roll two consecutive 5s?

6.15. D.E. Shaw: A coin was flipped 1,000 times, and 550 times it showed heads. Do you think the coin is biased? Why or why not?

6.16. Quora: You are drawing from a normally distributed random variable $X \sim N(0, 1)$ once a day. What is the approximate expected number of days until you get a value greater than 2?

6.17. Akuna Capital: Say you have two random variables X and Y, each with a standard deviation. What is the variance of $aX + bY$ for constants a and b?

6.18. Google: Say we have $X \sim$ Uniform(0, 1) and $Y \sim$ Uniform(0, 1) and the two are independent. What is the expected value of the minimum of X and Y?

6.19. Morgan Stanley: Say you have an unfair coin which lands on heads 60% of the time. How many coin flips are needed to detect that the coin is unfair?

6.20. Uber: Say you have n numbers 1...n, and you uniformly sample from this distribution with replacement n times. What is the expected number of distinct values you would draw?

6.21. Goldman Sachs: There are 100 noodles in a bowl. At each step, you randomly select two noodle ends from the bowl and tie them together. What is the expectation on the number of loops formed?

6.22. Morgan Stanley: What is the expected value of the max of two dice rolls?

6.23. Lyft: Derive the mean and variance of the uniform distribution U(a, b).

6.24. Citadel: How many cards would you expect to draw from a standard deck before seeing the first ace?

6.25. Spotify: Say you draw n samples from a uniform distribution U(a, b). What are the MLE estimates of a and b?

Hard

6.26. Google: Assume you are drawing from an infinite set of i.i.d random variables that are uniformly distributed from (0, 1). You keep drawing as long as the sequence you are getting is monotonically increasing. What is the expected length of the sequence you draw?

6.27 Facebook: There are two games involving dice that you can play. In the first game, you roll two dice at once and receive a dollar amount equivalent to the product of the rolls. In the second game, you roll one die and get the dollar amount equivalent to the square of that value. Which has the higher expected value and why?

6.28. Google: What does it mean for an estimator to be unbiased? What about consistent? Give examples of an unbiased but not consistent estimator, and a biased but consistent estimator.

6.29. Netflix: What are MLE and MAP? What is the difference between the two?

6.30. Uber: Say you are given a random Bernoulli trial generator. How would you generate values from a standard normal distribution?

6.31. Facebook: Derive the expectation for a geometric random variable.

6.32. Goldman Sachs: Say we have a random variable $X \sim D$, where D is an arbitrary distribution. What is the distribution F(X) where F is the CDF of X?

6.33. Morgan Stanley: Describe what a moment generating function (MGF) is. Derive the MGF for a normally distributed random variable X.

6.34. Tesla: Say you have N independent and identically distributed draws of an exponential random variable. What is the best estimator for the parameter λ?

6.35. Citadel: Assume that $\log X \sim N(0, 1)$. What is the expectation of X?

6.36. Google: Say you have two distinct subsets of a dataset for which you know their means and standard deviations. How do you calculate the blended mean and standard deviation of the total dataset? Can you extend it to K subsets?

6.37. Two Sigma: Say we have two random variables X and Y. What does it mean for X and Y to be independent? What about uncorrelated? Give an example where X and Y are uncorrelated but not independent.

6.38. Citadel: Say we have $X \sim$ Uniform(−1, 1) and $Y = X\text{^}2$. What is the covariance of X and Y?

6.39. Lyft: How do you uniformly sample points at random from a circle with radius R?

6.40. Two Sigma: Say you continually sample from some i.i.d. uniformly distributed (0, 1) random variables until the sum of the variables exceeds 1. How many samples do you expect to make?

40 Real Statistics Interview Solutions

Solution #6.1

The Central Limit Theorem (CLT) states that if any random variable, regardless of distribution, is sampled a large enough number of times, the sample mean will be approximately normally distributed. This allows for studying of the properties for any statistical distribution as long as there is a large enough sample size.

The mathematical definition of the CLT is as follows: for any given random variable X, as n approaches infinity,

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \rightarrow\sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

At any company with a lot of data, like Uber, this concept is core to the various experimentation platforms used in the product. For a real-world example, consider testing whether adding a new feature increases rides booked in the Uber platform, where each X is an individual ride and is a Bernoulli random variable (i.e., the rider books or does not book a ride). Then, if the sample size is sufficiently large, we can assess the statistical properties of the total number of bookings, as well as the booking rate (rides booked / rides opened on app). These statistical properties play a key role in hypothesis testing, allowing companies like Uber to decide whether or not to add new features in a data-driven manner.