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,

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

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: $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \to\sim N\left(\mu, \frac{\sigma^2}{n}\right)$; 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 $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

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-α)% 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-α is referred to as the confidence level, whereas 1-β 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-α and 1-β 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:

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

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:

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

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 $Y = X$^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.

Solution #6.2

Suppose we want to estimate some parameters of a population. For example, we might want to estimate the average height of males in the U.S. Given some data from a sample, we can compute a sample mean for what we think the value is, as well as a range of values around that mean. Following the previous example, we could obtain the heights of 1,000 random males in the U.S. and compute the average height, or the sample mean. This sample mean is a type of point estimate and, while useful, will vary from sample to sample. Thus, we can't tell anything about the variation in the data around this estimate, which is why we need a range of values through a confidence interval.

Confidence intervals are a range of values with a lower and an upper bound such that if you were to sample the parameter of interest a large number of times, the 95% confidence interval would contain the true value of this parameter 95% of the time. We can construct a confidence interval using the sample standard deviation and sample mean. The level of confidence is determined by a margin of error that is set beforehand. The narrower the confidence interval, the more precise the estimate, since there is less uncertainty associated with the point estimate of the mean.

Solution #6.3

A/B testing has many possible pitfalls that depend on the particular experiment and setup employed. One common drawback is that groups may not be balanced, possibly resulting in highly skewed results. Note that balance is needed for all dimensions of the groups — like user demographics or device used — because, otherwise, the potentially statistically significant results from the test may simply be due to specific factors that were not controlled for. Two types of errors are frequently assessed: Type I error, which is also known as a "false positive," and Type II error, also known as a "false negative." Specifically, Type I error is rejecting a null hypothesis when that hypothesis is correct, whereas Type II error is failing to reject a null hypothesis when its alternative hypothesis is correct.

Another common pitfall is not running an experiment for long enough. Generally speaking, experiments are run with a particular power threshold and significance threshold; however, they often do not stop immediately upon detecting an effect. For an extreme example, assume you're at either Uber or Lyft and running a test for two days, when the metric of interest (e.g., rides booked) is subject to weekly seasonality.

Lastly, dealing with multiple tests is important because there may be interactions between results of tests you are running and so attributing results may be difficult. In addition, as the number of variations you run increases, so does the sample size needed. In practice, while it seems technically feasible to test 1,000 variations of a button when optimizing for click-through rate, variations in tests are usually based on some intuitive hypothesis concerning core behavior.

Solution #6.4

For any given random variables X and Y, the covariance, a linear measure of relationship, is defined by the following: $Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

Specifically, covariance indicates the direction of the linear relationship between X and Y and can take on any potential value from negative infinity to infinity. The units of covariance are based on the units of X and Y, which may differ.

The correlation between X and Y is the normalized version of covariance that takes into account the variances of X and Y:

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

Since correlation results from scaling covariance, it is dimensionless (unlike covariance) and is always between -1 and 1 (also unlike covariance).

Solution #6.5

The null hypothesis is that the coin is fair, and the alternative hypothesis is that the coin is biased:

$$H_0 : p_0 = 0.5, H_1 : p_1 \neq 0.5$$

Note that, since the sample size here is 10, you cannot apply the Central Limit Theorem (and so you cannot approximate a binomial using a normal distribution).

The p-value here is the probability of observing the results obtained given that the null hypothesis is true, i.e., under the assumption that the coin is fair. In total for 10 flips of a coin, there are 2^10 = 1024 possible outcomes, and in only 10 of them are there 9 tails and one heads. Hence, the exact probability of the given result is the p-value, which is $\frac{10}{1024} = 0.0098$. Therefore, with a significance level set, for example, at 0.05, we can reject the null hypothesis.

Solution #6.6

The process of testing whether data supports particular hypotheses is called hypothesis testing and involves measuring parameters of a population's probability distribution. This process typically employs at least two groups — one a control that receives no treatment, and the other group(s), which do receive the treatment(s) of interest. Examples could be the height of two groups of people, the conversion rates for particular user flows in a product, etc. Testing also involves two hypotheses — the null hypothesis, which assumes no significant difference between the groups, and the alternative hypothesis, which assumes a significant difference in the measured parameter(s) as a consequence of the treatment.

A p-value is the probability of observing the given test results under the null hypothesis assumptions. The lower this probability, the higher the chance that the null hypothesis should be rejected. If the p-value is lower than the predetermined significance level α, generally set at 0.05, then it indicates that the null hypothesis should be rejected in favor of the alternative hypothesis. Otherwise, the null hypothesis cannot be rejected, and it cannot be concluded that the treatment has any significant effect.

Solution #6.7

Both errors are relevant in the context of hypothesis testing. Type I error is when one rejects the null hypothesis when it is correct, and is known as a false positive. Type II error is when the null hypothesis is not rejected when the alternative hypothesis is correct; this is known as a false negative. In layman's terms, a type I error is when we detect a difference, when in reality there is no significant difference in an experiment. Similarly, a type II error occurs when we fail to detect a difference, when in reality there is a significant difference in an experiment.

Type I error is given by the level of significance α, whereas the type II error is given by β. Usually, $1-\alpha$ is referred to as the confidence level, whereas $1-\beta$ is referred to as the statistical power of the test being conducted. Note that, in any well-conducted statistical procedure, we want to have both α and β be small. However, based on the definition of the two, it is impossible to make both errors small

simultaneously: the larger α is, the smaller β is. Based on the experiment and the relative importance of false positives and false negatives, a data scientist must decide what thresholds to adopt for any given experiment. Note that experiments are set up so as to have both 1-α and 1-β relatively high (say at .95, and .8, respectively).

Solution #6.8

Power is the probability of rejecting the null hypothesis when, in fact, it is false. It is also the probability of avoiding a Type II error. A Type II error occurs when the null hypothesis is not rejected when the alternative hypothesis is correct. This is important because we want to detect significant effects during experiments. That is, the higher the statistical power of the test, the higher the probability of detecting a genuine effect (i.e, accepting the alternative hypothesis and rejecting the null hypothesis). A minimum sample size can be calculated for any given level of power — for example, say a power level of 0.8.

An analysis of the statistical power of a test is usually performed with respect to the test's level of significance (α) and effect size (ie., the magnitude of the results).

Solution #6.9

In a Z-test, your test statistic follows a normal distribution under the null hypothesis. Alternatively, in a t-test, you employ a student's t-distribution rather than a normal distribution as your sampling distribution.

Considering the population mean, we can use either Z-test or t-test only if the mean is normally distributed, which is possible in two cases: the initial population is normally distributed, or the sample size is large enough ($n \geq 30$) 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 variation is *known*, and vice versa: we use *t*-test if the population variation is *unknown*.

Additionally, if the sample size is very large ($n > 200$), we can use the Z-test in any case, since for such large degrees of freedom, *t*-distribution coincides with *z*-distribution up to thousands.

Considering the population proportion, we can use a Z-test (but not *t*-test) where $np_0 \geq 10$ and $n(1-p_0) \geq 10$, i.e., when each of the number of successes and the number of failures is at least 10.

Solution #6.10

The primary consideration is that, as the number of tests increases, the chance that a stand-alone p-value for any of the t-tests is statistically significant becomes very high due to chance alone. As an example, with 100 tests performed and a significance threshold of $\alpha = 0.05$, you would expect five of the experiments to be statistically significant due only to chance. That is, you have a very high probability of observing at least one significant outcome. Therefore, the chance of incorrectly rejecting a null hypothesis (i.e., committing Type I error) increases.

To correct for this effect, we can use a method called the Bonferroni correction, wherein we set the significance threshold to α/m, where m is the number of tests being performed. In the above scenario with 100 tests, we can set the significance threshold to instead be $0.05/100 = 0.0005$. While this correction helps to protect from Type I error, it is still prone to Type II error (i.e., failing to reject the null hypothesis when it should be rejected). In general, the Bonferroni correction is mostly useful when there is a smaller number of multiple comparisons of which a few are significant. If the number of tests becomes sufficiently high such that many tests yield statistically significant results, the number of Type II errors may also increase significantly.

Solution #6.11

The confidence interval (CI) for a population proportion is an interval that includes a true population proportion with a certain degree of confidence $1 - \alpha$.

For the case of flipping heads from a series of coin tosses, the proportion follows the binomial distribution. If the series size is large enough (each of the number of successes and the number of failures is at least 10), we can utilize the Central Limit Theorem and use the normal approximation for the binomial distribution as follows:

$$N\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

where $\hat{p}$ is the proportion of heads tossed in series, and n is the series size. The CI is centered at the series proportion, and plus or minus a margin of error:

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

where $z_{\alpha/2}$ is the appropriate value from the standard normal distribution for the desired confidence level.

For example, for the most commonly used level of confidence, 95%, $z_{\alpha/2} = 1.96$.

Solution #6.12

Let X be the number of coin flips needed to obtain two consecutive heads. We then want to solve for $E[X]$. Let H denote a flip that results in heads, and T denote a flip that results in tails. Note that $E[X]$ can be written in terms of $E[X|H]$ and $E[X|T]$, i.e., the expected number of flips needed, conditioned on a flip being either heads or tails, respectively.

Conditioning on the first flip, we have: $E[X] = \frac{1}{2}(1 + E[X|H]) + \frac{1}{2}(1 + E[X|T])$

Note that $E[X|T] = E[X]$ since if a tail is flipped, we need to start over in getting two heads in a row.

To solve for $E[X|H]$, we can condition it further on the next outcome: either heads (HH) or tails (HT).

Therefore, we have: $E[X|H] = \frac{1}{2}(1 + E[X|HH]) + \frac{1}{2}(1 + E[X|HT])$

Note that if the result is HH, then $E[X|HH] = 0$, since the outcome has been achieved. If a tail was flipped, then $E[X|HT] = E[X]$, and we need to start over in attempting to get two heads in a row. Thus:

$$E[X|H] = \frac{1}{2}(1+0) + \frac{1}{2}(1 + E[X]) = 1 + \frac{1}{2}E[X]$$

Plugging this into the original equation yields:

$$E[X] = \frac{1}{2}\left(1 + 1 + \frac{1}{2}E[X]\right) + \frac{1}{2}(1 + E[X])$$

and after solving we get: $E[X] = 6$. Therefore, we would expect 6 flips.

Solution #6.13

Let k denote the number of distinct sides seen from rolls. The first roll will always result in a new side being seen. If you have seen k sides, where $k < 6$, then the probability of rolling an unseen value will be $(6 - k)/6$, since there are $6 - k$ values you have not seen, and 6 possible outcomes of each roll.

Note that each roll is independent of previous rolls. Therefore, for the second roll ($k = 1$), the time until a side not seen appears has a geometric distribution with $p = 5/6$, since there are five of the six sides left to be seen. Likewise, after two sides ($k = 2$), the time taken is a geometric distribution, with $p = 4/6$. This continues until all sides have been seen.

Recall that the mean for a geometric distribution is given by $1/p$, and let X be the number of rolls needed to show all six sides. Then, we have the following:

$$E[X] = 1 + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + \frac{6}{1} = 6\sum_{p=1}^{6}\frac{1}{p} = 14.7 \text{ rolls}$$

Solution #6.14

Similar in methodology to question 13, let X be the number of rolls until two consecutive fives. Let Y denote the event that a five was just rolled.

Conditioning on Y, we know that either we just rolled a five, so we only have one more five to roll, or we rolled some other number and now need to start over after having rolled once:

$$E[X] = \frac{1}{6}(1 + E[X \mid Y]) + \frac{5}{6}(1 + E[X])$$

Note that we have the following: $E[X \mid Y] = \frac{1}{6}(1) + \frac{5}{6}(1 + E[X])$

Plugging the results in yields an expected value of 42 rolls: $E[X] = 42$

Solution #6.15

Because the sample size of flips is large (1,000), we can apply the Central Limit Theorem. Since each individual flip is a Bernoulli random variable, we can assume that p is the probability of getting heads. We want to test whether p is .5 (i.e., whether it is a fair coin or not). The Central Limit Theorem allows us to approximate the total number of heads seen as being normally distributed.

More specifically, the number of heads seen out of n total rolls follows a binomial distribution since it a sum of Bernoulli random variables. If the coin is not biased ($p = .5$), then the expected number of heads is as follows: $\mu = np = 1000 * 0.5 = 500$, and the variance of the number of heads is given by:

$$\sigma^2 = np(1-p) = 1000 * 0.5 * 0.5 = 250, \sigma = \sqrt{250} \approx 16$$

Since this mean and standard deviation specify the normal distribution, we can calculate the corresponding z-score for 550 heads as follows:

$$z = \frac{550 - 500}{16} = 3.16$$

This means that, if the coin were fair, the event of seeing 550 heads should occur with a < 0.1% chance under normality assumptions. Therefore, the coin is likely biased.

Solution #6.16

Since X is normally distributed, we can employ the cumulative distribution function (CDF) of the normal distribution: $\Phi(2) = P(X \le 2) = P(X \le \mu + 2\sigma) = 0.9772$

Therefore, $P(X > 2) = 1 - 0.977 = 0.023$ for any given day. Since each day's draws are independent, the expected time until drawing an $X > 2$ follows a geometric distribution, with $p = 0.023$. Letting T be a random variable denoting the number of days, we have the following:

$$E[T] = \frac{1}{p} = \frac{1}{.02272} \approx 44 \text{ days}$$

Solution #6.17

Let the variances for X and Y be denoted by $Var(X)$ and $Var(Y)$.

Then, recalling that the variance of a sum of variables is expressed as follows:

$$Var(X + Y) = Var(X) = Var(Y) + 2Cov(X, Y)$$

and that a constant coefficient of a random variable is assessed as follows: $Var(aX) = a^2Var(X)$

We have $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$, which would provide the bounds on the designated variance; the range will depend on the covariance between X and Y.

Solution #6.18

Let $Z = \min(X, Y)$. Then we know the following: $P(Z \le z) = P(\min(X, Y) \le z) = 1 - P(X > z, Y > z)$

For a uniform distribution, the following is true for a value of z between 0 and 1:

$$P(X > z) = 1 - z \text{ and } P(Y > z) = 1 - z$$

Since X and Y are i.i.d., this yields: $P(Z \le z) = 1 - P(X > z, Y > z) = 1 - (1 - z)^2$

Now we have the cumulative distribution function for z. We can get the probability density function by taking the derivative of the CDF to obtain the following: $fz(z) = 2(1 - z)$. Then, solving for the expected value by taking the integral yields the following:

$$E[Z] = \int_0^1 zfz(z)dz = 2\int_0^1 z(1-z)dz = 2\left(\frac{1}{2} - \frac{1}{3}\right) = \frac{1}{3}$$

Therefore, the expected value for the minimum of X and Y is 1/3.

Solution #6.19

Say we flip the unfair coin n times. Each flip is a Bernoulli trial with a success probability of p:

$$x_1, x_2, \ldots x_n, x_i \sim \text{Ber}(p)$$

We can construct a confidence interval for p as follows, using the Central Limit Theorem. First, we decide on our level of confidence. If we select a 95% confidence level, the necessary z-score is $z = 1.96$. We then construct a 95% confidence interval for p. If it does not include 0.5 as its lower bound, then we can reject the null hypothesis that the coin is fair.

Since the trials are i.i.d., we can compute the sample mean for p from a large number of trials:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We know the following properties hold: $E[\hat{p}] = \frac{np}{n} = p$ and $Var = (\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

Therefore, our 95% confidence interval is given by the following: $p \pm z\sqrt{\frac{p(1-p)}{n}}$

Since the true $p = 0.6$, plugging that in and setting the lower bound of the interval equal to 0.5 yields:

$$0.6 - 1.96\sqrt{\frac{0.6(1-0.6)}{n}} = 0.5$$

Solving for n yields 93 flips.

Solution #6.20

Let the following be an indicator random variable: $X_i = 1$ if i is drawn in n turns

We would then want to find the following: $\sum_{i=1}^{n} E[X_i]$

We know that $p(X_i = 1) = 1 - p(X_i = 0)$, so the probability of a number not being drawn (where each draw is independent) is the following:

$$p(X_i = 0) = \left(\frac{n-1}{n}\right)^n$$

Therefore, we have: $p(X_i = 1) = 1 - \left(\frac{n-1}{n}\right)^n$ and by linearity of expectation, we then have:

$$\sum_{i=1}^{n} E[x_i] = nE[X_i] = n\left(1 - \left(\frac{n-1}{n}\right)^n\right)$$

Solution #6.21

Say that we have n noodles. At any given step, we will have one of two outcomes: (1) we pick two ends from the same noodle (which makes a loop), or (2) we pick two ends from different noodles. Let X_n denote a random variable representing the number of loops with n noodles remaining.

The probability of case (1) happening is: $\frac{n}{\binom{2n}{2}} = \frac{1}{2n-1}$

where the denominator represents the number of ends we can choose from the noodles, and the numerator represents the number of cases where we choose the same noodle.

Therefore, the probability of case (2) happening is: $1 - \frac{1}{2n-1} = \frac{2n-2}{2n-1}$

Then, taking case (1) and (2), we have the following recursive formulation for the expectation of the number of loops formed:

$$E[X_n] = \frac{1}{2n-1} + \frac{2n-2}{2n-1} E[X_{n-1}]$$

Plugging in $E[X_1] = 1$ and calculating the first few terms, we can notice the following pattern, for which we can plug in $n = 100$ to obtain the answer:

$$E[X_{100}] = 1 + \frac{1}{3} + \ldots + \frac{1}{2(100)-1} \approx 3.3$$

Solution #6.22

Since we only have two dice, let the maximum value between the two be m. Let

$$X_1, X_2, Y = \max(X_1, X_2)$$

denote the first roll, second roll, and the max of the two. Then we want to find the following:

$$E[Y] = \sum_{i=1}^{6} i * P(Y = 1)$$

We can condition $Y = m$ on three cases: (1) die one is the max roll; (2) die two is the max roll; or (3) they are both the same.

For cases (1) and (2) we have: $P(X_1 = i, X_2 < i) = P(X_2 = i, X_1 < i) = \frac{1}{6} * \frac{i-1}{6}$

"For case (3), where both dice are the maximum:"

$$P(X_1 = X_2 = i) = \frac{1}{6} * \frac{1}{6}$$

Putting everything together yields the following: $E[Y] = \sum_{i=1}^{6} i * \left(\frac{1}{6} * \frac{i-1}{6} * 2 + \frac{1}{6} * \frac{1}{6}\right) = \frac{161}{36}$

A simpler way to visualize this is to use a contingency table, such as the one below:

1	2	3	4	5	6
(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Then the expectation is simply given by:

$$E[Y] = 1 \times \frac{1}{6} + 2 \times \frac{3}{6} + 3 \times \frac{5}{6} + 4 \times \frac{7}{6} + 5 \times \frac{9}{6} + 6 \times \frac{11}{6} = \frac{161}{36} \approx 4.5$$

Solution #6.23

For $X \sim U(a, b)$, we have the following: $f_X(x) = \frac{1}{b-a}$

Therefore, we can calculate the mean as:

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

Similarly, the variance can be as expressed as follows: $Var(X) = E[X^2] - E[X]^2$

Giving us:

$$E[X^2] = \int_a^b x^2 f_X(x)\,dx = \int_a^b \frac{x^2}{b-a}\,dx = \frac{x^3}{3(a-b)}\Big|_a^b = \frac{a^2+ab+b^2}{3}$$

Therefore: $Var(X) = \frac{a^2+ab^2+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$

Solution #6.24

Although one can enumerate all the probabilities, this can get a bit messy from an algebraic standpoint, so obtaining the following intuitive answer is more preferable. Imagine we have aces A1, A2, A3, A4. We can then draw a line in between them to represent an arbitrary number (including 0) of cards between each ace, with a line before the first ace and after the last.

$$|A1|A2|A3|A4|$$

There are 52 – 4 = 48 non-ace cards in a deck. Each of these cards is equally likely to be in any of the five lines. Therefore, there should be 48/5 = 9.6 cards drawn prior to the first ace being drawn. Hence, the expected number of cards drawn until the first ace is seen is 9.6 + 1 = 10.6 cards — we can't forget to add 1, because we need to include drawing the ace card itself.

Solution #6.25

Note that for a uniform distribution, the probability density is $\frac{1}{b-a}$ for any value on the interval [a, b]. The likelihood function is therefore as follows:

$$f(x_1 \ldots, x_n \mid a, b) = \left(\frac{1}{b-a}\right)^n$$

To obtain the MLE, we maximize this likelihood function, which is clearly maximized if $\hat{b}$ is the largest of the samples and a is the smallest of the samples. Therefore, we have the following:

$$\hat{a} = \min(x_1 \ldots, x_n), = \hat{b} \max(x_1 \ldots, x_n)$$

Solution #6.26

Assume that we have an indicator random variable: $X_i = 1$ if the sequence is increasing up to ith element, and otherwise $X_i = 0$.

Then, we calculate the expectation: $E[X_1 + X_2 + \ldots]$. Consider some arbitrary i. In order to draw up to element i, the entire sequence up to i must be monotonically increasing, which means that the following is true: $X_1 < X_2 < \ldots < X_i$. Given that there are n possible sequences of the elements, there is a

$$\frac{1}{i!}$$

chance of X_i being 1. Since each X is i.i.d., we then have: $E[X_1 + X_2 + \ldots] = 1 + \frac{1}{2!} + \ldots = e - 1$

Solution #6.27

One method of solving this problem is the brute force method, which consists of computing the expected values by listing all of the outcomes and associated probabilities and payoffs. However, there exists an easier way of solving the problem.

Assume that the outcome of the roll of a die is given by a random variable X (meaning that it takes on the values 1...6 with equal probability). Then, the question is equivalent to asking, "What is $E[X] \cdot E[X] = E[X]^2$ (i.e., the expected value of the product of two separate rolls), versus $E[X^2]$ (the expected value of the square of a single roll)?"

Recall that the variance of a given random variable X is as follows:

$$Var(X) = E[(X - E[X])^2] = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - E[X]^2$$

Typically, this variance term is exactly the difference between the two sets of die rolls — the two "games" — (the payoff of the second game minus the payoff of the first game). Since the left-hand side is positive, as expected for the value of a squared number, then the right-hand side is also positive. Therefore, it must be the case that the second game has a higher expected value than the first.

Solution #6.28

In both cases, we are dealing with an estimator of the true parameter value. An estimator is unbiased if the expectation of the estimator is the true underlying parameter value. An estimator is consistent if, as the sample size increases, the estimator's sampling distribution converges towards the true parameter value.

Consider the following random variable X, which is normally distributed, and n i.i.d. samples used to calculate a sample mean:

$$X \sim N(\mu, \sigma^2) \text{ and } \bar{x} = \frac{x_1 + x_2 \ldots + x_n}{n}$$

The first sample is an example of an unbiased but not consistent estimator. It is unbiased since $E[x_1] = \mu$. However, it is not consistent since, as the sample size increases, the sampling distribution of the first sample does not become more concentrated with respect to the true mean.

An example of a biased but consistent estimator is the sample variance: $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

It can be shown that $E\left[S_n^2\right] = \frac{n-1}{n}\sigma^2$

The formal proof of the above is called Bessel's correction, but there is an intuitive way to grasp the presence of the term preceding the variance. If we uniformly sample two numbers randomly from the series of numbers 1 to n, we have an $n/n^2 = 1/n$ chance that the two equal the same number, meaning the sampled squared difference of the numbers will be zero. The sample variance will therefore slightly underestimate the true variance. However, this bias goes to 0 as n approaches infinity, since the term in front of the variance, (n-1/n), approaches 1. Therefore, the estimator is consistent.

Solution #6.29

MLE stands for maximum likelihood estimation, and MAP for maximum a posteriori. Both are ways of estimating variables in a probability distribution by producing a single estimate of that variable.

Assume that we have a likelihood function $P(X|\theta)$. Given n i.i.d. samples, the MLE is as follows:

$$MLE(\theta) = \max_{\theta} P(X \mid \theta) = \max_{\theta} \prod_{i}^{n} P(x_i \mid \theta)$$

Since the product of multiple numbers all valued between 0 and 1 might be very small, maximizing the log function of the product above is more convenient. This is an equivalent problem, since the log function is monotonically increasing. Since the log of a product is equivalent to the sum of logs, the MLE becomes the following:

$$MLE_{\log}(\theta) = \max_{\theta} \sum_{i=1}^{n} \log P(x_i \mid \theta)$$

Relying on Bayes rule, MAP uses the posterior $P(\theta|X)$ being proportional to the likelihood multiplied by a prior $P(\theta)$, i.e., $P(X|\theta)P(\theta)$. The MAP for θ is thus the following:

$$MAP(\theta) = \max_{\theta} P(X \mid \theta) = \max_{\theta} \prod_{i}^{n} P(x_i \mid \theta) P(\theta)$$

Employing the same math as used in calculating the MLE, the MAP becomes:

$$MAP_{\log}(\theta) = \max_{\theta} \sum_{i=1}^{n} \log P(x_i \mid \theta) + \log P(\theta)$$

Therefore, the only difference between the MLE and MAP is the inclusion of the prior in MAP; otherwise, the two are identical. Moreover, MLE can be seen as a special case of the MAP with a uniform prior.

Solution #6.30

Assume we have n Bernoulli trials, each with a p probability of success. Altogether, they form a binomial distribution: $x_1, x_2, \ldots, x_n$, $X \sim B(n, p)$ where $x_i = 1$ means success and $x_i = 0$ means failure. Assuming i.i.d. trials, we can compute the sample proportion for $\hat{p}$ as follows:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We know that if n is large enough, then the binomial distribution approximates the following normal distribution:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

where n must be $np \geq 10$, $n(1-p) \geq 10$

Therefore, the value $\hat{p}$ can be used as simulation for a normal distribution. The sample size n must only be large enough to satisfy the conditions above (at least $n = 20$ for $p = .5$), but it is recommended to use a significantly larger n to get the better normal approximation.

Finally, to simulate the standard normal distribution, we normalize $\hat{p}$: $\hat{p}_0 = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$

At this point, we can derive the final formula for our normal random generator: $\bar{x} = \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i - p}{\sqrt{\frac{p(1-p)}{n}}}$

The previous expression can be simplified to the following: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i - np}{\sqrt{np(1-p)}}$

where x_1, ..., x_n is the Bernoulli series we get from the given random generator, with probability of success p.

Solution #6.31

We are seeking the expected value of geometric random variable X as follows: $E[X] = \sum_{k=1}^{\infty} k f_X(k)$

The expression above contains a summation instead of an integral since k is a discrete rather than continuous random variable, and we know the probability mass function of the geometric probability distribution is given by the following: $f_X(k) = (1-p)^{k-1} p$

Therefore, we obtain the expected value of X as follows: $E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1} p$

Since p is constant with respect to k, we separate out p as follows: $E[X] = p\sum_{k=1}^{\infty} k(1-p)^{k-1}$

Note that the term inside the summation is really the following:

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = \sum_{k=1}^{\infty} k(1-p)^{k-1} + \sum_{k=2}^{\infty} k(1-p)^{k-1} + \ldots$$

This simplifies to the following:

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = \left(\frac{1}{p}\right) + (1-p)/p + (1-p)^2/p + \ldots = \frac{1}{p}\left(1 + (1-p) + (1-p)^2 + \ldots\right)$$

$$= \frac{1}{p} \cdot \frac{1}{1-(1-p)} = \frac{1}{p^2}$$

Plugging this back into the equation for the expected value of X yields the following:

$$E[X] = p * \frac{1}{p^2} = \frac{1}{p}$$

Solution #6.32

We can define a new variable $Y = F(X)$, and, hence, we want to find the CDF of y (where y is between 0 and 1 by definition of a CDF): $F_Y(y) = P(Y \leq y)$

Substituting in for Y yields the following: $F_Y(y) = P(F(X) \leq y)$

Applying the inverse CDF on both sides yields the following:

$$F_Y(y) = P(F^{-1}(F(X)) \leq F^{-1}(y)) = P(X \leq F^{-1}(y))$$

Note that the last expression is simply the CDF for: $P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y$

Therefore, we have: $F_Y(y) = y$

Since y falls between 0 and 1, Y's distribution is simply a uniform one from 0 to 1, i.e., $U(0, 1)$.

Solution #6.33

A moment generating function is the following function for a given random variable:

$$M_X(s) = E[e^{sX}]$$

If X is continuous (as in the case of normal distributions), then the function becomes the following:

$$M_X(s) = \int_{-\infty}^{\infty} e^{sX} f_X(x)\,dx$$

Hence, the moment generating function is a function for a given value of s. It is useful for calculating moments, since taking derivatives of the moment generating function and evaluating at $s = 0$ yields the desired moment.

For a normal distribution, recall that: $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

First, taking the special case of the standard normal random variable, we have the following:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Plugging this into the above MGF yields: $M_X(s) = \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2 + sx} dx$

Completing the square yields: $M_X(s) = e^{\frac{s^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2 + sx - \frac{s^2}{2}} dx = e^{\frac{s^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-s)^2}{2}} dx = e^{\frac{s^2}{2}}$

Note that the last step uses the fact that the expression within the integral is a PDF for a normally distributed random variable with mean s and variance 1, and hence the integral evaluates to 1.

To solve for a general random variable, you can plug in $X = \sigma Y + \mu$, where Y is standard normal variable, to yield: $M_Y(s) = e^{s\mu}$, $M_X(s\sigma) = e^{(s^2\sigma^2/2) + s\mu}$

Solution #6.34

Denote the n i.i.d. draws as: x_1, x_2, ..., x_n where, for any individual draw, we have the pdf: $f_X(x_i) = \lambda e^{-\lambda x_i}$

Therefore the likelihood of the data is given by the following:

$$L(\lambda; x_1 \ldots x_n) = \prod_{i=1}^{n} f_X(x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

Taking the log of the equation above to obtain the log-likelihood results in the following:

$$\log L(\lambda; x_1 \ldots x_n) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

Taking the derivative with respect to λ and setting the results to 0 yields: $\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$

Therefore, the best estimate of λ is given by: $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$

Solution #6.35

Define $Y = \log X$. We then want to solve for: $E[e^Y] = E[X]$

Recall that a moment generating function has the following form: $M_Y(s) = E[e^{sY}]$

Therefore, we want the moment generating function for $Y \sim N(0, 1)$, which was derived in problem 33 and has the form: $M_Y(s) = e^{s^2/2}$

Therefore, evaluating at $s = 1$ (since we want the mean) gives: $M_Y(1) = e^{1/2}$ which is the desired answer.

Solution #6.36

Say that the two have two distinct group sizes: n_1 = size of group 1, and n_2 = size of group 2.

Given the means of two groups, μ_1 and μ_2, the blended mean can be found simply by taking a weighted average:

$$\bar{\mu} = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}$$

We know that the blended standard deviation for the total data set has the form:

$$\bar{s} = \sqrt{\frac{\sum_{i=1}^{n_1+n_2} (z_i - \bar{\mu})^2}{n_1 + n_2}}$$

where z_i is the union of the points from both groups.

However, since we are not given the initial data points from the two groups, we have to rearrange this formula by using instead the given variations of these groups, s_1^2 and s_2^2, as follows:

$$\bar{s} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2}{n_1 + n_2}}$$

Applying the Bessel correction, the blended standard deviation for the two groups is as follows:

$$\bar{s} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + n_1(\bar{\mu}_1 - \bar{\mu})^2 + n_2(\bar{\mu}_2 - \bar{\mu})^2}{n_1 + n_2 - 1}}$$

To extend the definition above to subsets, the mean is as follows: $\bar{\mu} = \frac{\sum_{i=1}^{K} n_i \mu_i}{\sum_{i=1}^{K} n_i}$

And the standard deviation is: $\bar{s}_K = \sqrt{\frac{\sum_{i=1}^{K} (n_i - 1)s_i^2 + n_i(\bar{\mu}_i - \bar{\mu})^2}{\sum_{i=1}^{K} n_i - 1}}$

where n_i are the sizes of initial groups, μ_i and s_i are their respective means and standard deviations.

Solution #6.37

Independence is defined as follows: $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x, y. Equivalently, we can use the following definitions: $P(X = x \mid Y = y) = P(X = x)$, $P(Y = y \mid X = x) = P(Y = y)$

When two random variables X and Y are uncorrelated, their covariance, which is calculated as follows, is 0: $Cov(X, Y) = E[XY] - E[X]E[Y]$

For an example of uncorrelated but not independent variables, let X take on values –1, 0, or 1 with equal probability, and let $Y = 1$ if $X = 0$ and $Y = 0$ otherwise. Then we can verify that X and Y are uncorrelated:

$$E(XY) = \frac{1}{3}(-1)(0) + \frac{1}{3}(0)(1) + \frac{1}{3}(1)(0) = 0$$

And $E[X] = 0$, so the covariance between the two random variables is zero. However, it is clear that the two are not independent, since we defined Y in such a way that it obviously depends on X.

$$P(Y = y \mid X = x) \neq P(Y = y)$$

For example, $P(Y = 1 \mid X = 0) = 1$

Solution #6.38

By definition of the covariance, we have: $Cov(X, Y) = Cov(X, X^2) = E[(X - E[X])(X^2 - E[X^2])]$

Expanding terms of the equation above yields: $Cov(X, Y) = E[(X^3 - XE[X^2] - X^2E[X] + E[X]E[X^2])]$

Using linearity of expectation, we obtain: $Cov(X, Y) = E[X^3] - E[X]E[X^2] - E[X^2]E[X] + E[X]E[X^2]$

Since the second and last terms cancel one another, we end up with the following:

$$Cov(X, Y) = E[X^3] - E[X^2]E[X]$$

Here, we conclude that $E[X] = 0$ (based on the definition of X) and that $E[X^3] = 0$ by evaluating the probability density function of X as follows:

$$f_X(x) = \frac{1}{b - a} = \frac{1}{1 - (-1)} = \frac{1}{2}$$

Since we are evaluating X from –1 to 1, we then have: $E[X^3] = \int_{-1}^{1} x^3 f(x)\,dx = \int_{-1}^{1} \frac{1}{2} x^3 dx = 0$

Thus, the covariance between X and Y is 0.

Solution #6.39

This can be proved using the inverse-transform method, whereby we sample from a uniform distribution and then simulate the points on the circle employing the inverse cumulative distribution functions (i.e., inverse CDFs).

We can define a random point within the circle using a given radius value and an angle (and obtain the corresponding x, y values from polar coordinates). To sample a random radius, consider the following. If we sample points from a radius r, we know that there are $2\pi r$ points to consider (i.e., the circumference of the circle). Likewise, if we sample a radius $2r$, there are $4\pi r$ points to consider. Therefore, we have the following probability density function given by the following:

$$f_R(r) = \frac{2r}{R^2}$$

This follows from the CDF, which is given by the ratio of the areas of the two circles: $F_R(r) = \frac{r^2}{R^2}$

Therefore, for the inverse sampling, we want the following: $y = \frac{r^2}{R^2}$

This simplifies to the following: $\sqrt{R^2 y} = r$

Therefore, we can sample $Y \sim U(0, 1)$ and the corresponding radius will be the following:

$$r = R\sqrt{y}$$

For the corresponding angles, we can sample theta uniformly from the range 0 to 2π: $\theta \in [0, 2\pi]$ and then set the following: $x = r\cos(\theta)$, $y = r\sin(\theta)$

Solution #6.40

Let us define: N_t = smallest n such that: $\sum_{i=1}^{n} U_i > t$ for any value t between 0 and 1. Then we want to find: $m(t) = E[N_t]$

Consider the first draw. Assuming that result is some value x, we then have two cases as follows. The first is that $x > t$, in which $N_t = 1$

The second is that $x < t$, necessitating that we sample again, yielding: $N_t = 1 + N_{t-x}$

Putting these two together, we have: $m(t) = 1 + \int_0^1 m(t-x)\,dx$

Employing the following change of variables: $u = t - x$, $du = -dx$

We then substitute and simplify to obtain: $m(t) = 1 + \int_0^t m(u)\,du$

Differentiating both sides, we then obtain: $m'(t) = m(t)$

Since $m(0) = 1$, we then have: $m(t) = e^t$

Since we actually need to find $m(N_{t=1})$, we can plug in $t = 1$ into the equation, which yields the desired result $m(1) = e$.