9 ESTIMATION AND HYPOTHESIS TESTING

In Chapter 7 we introduce the question of how an i.i.d. sample $X_1, X_2, \ldots, X_n$ from an unknown distribution may be used to estimate aspects of that distribution. In Chapter 8 we saw how the sample statistics behave asymptotically. In this chapter we look at some specific examples where various parameters of the distribution such as μ and σ are unknown, and the sample is used to estimate these parameters.

For example, suppose there is a coin which we assume has a probability p of showing heads each time it is flipped. To gather information about p the coin is flipped 100 times. The results of these flips are viewed as i.i.d. random variables $X_1, X_2, \ldots, X_{100}$ with a Bernoulli(p) distribution. Suppose $\sum_{n=1}^{100} X_n = 60$, meaning 60 of the 100 flips showed heads. How might we use this to infer something about the value of p?

The first two topics we will consider are the "method of moments" and the "maximum likelihood estimate". Both of these are direct forms of estimation in the sense that they produce a single-value estimate for p. A benefit of such methods is that they produce a single prediction, but a downside is that the prediction they make is most likely not exactly correct. These methods amount to a statement like "Since 60 of the 100 flips came up heads, we predict that the coin should come up heads 60% of the time in the long run". In some sense the 60% prediction may be the most reasonable one given what was observed in the 100 flips, but it should also be recognised that 0.6 is unlikely to be the true value of p.

Another approach is that of the "confidence interval". Using this method we abandon the hope of realising a specific estimate and instead produce a range of values in which we expect to find the unknown parameter. This yields a statement such as, "With 90% confidence the actual probability the coin will show heads is between 0.52 and 0.68". While this approach does not give a single-valued estimate, it has the benefit that the result is more likely to be true.

Yet another approach is the idea of a "hypothesis test". In this case we make a conjecture about the value of the parameter and make a computation to test the credibility of the conjecture. The result may be a statement such as, "If the coin had a 50% chance of showing heads, then observing 60 heads or more in 100 flips should occur less than 3% of the time. This is a rare enough result, it suggests that the 50% hypothesis of showing heads is inaccurate".

For all of these methods, we will assume that the sample $X_1, X_2, \ldots, X_n$ are i.i.d copies of a random variable X with a probability mass function or probability density function $f(x)$. For brevity, we shall often refer to the distribution X, by which we will mean the distribution of the random variable X. We shall further assume that $f(x)$ depends on one or more unknown parameters $p_1, p_2, \ldots, p_d$ and emphasise this using the notation $f(x \mid p_1, p_2, \ldots, p_d)$. We may abbreviate this $f(x \mid p)$ where p represents the vector of all parameters $(p_1, \ldots, p_d) \in \mathcal{P} \subset \mathbb{R}^d$ for some $d \geq 1$. The set $\mathcal{P}$ may be all of $\mathbb{R}^d$ or some proper subset depending on the nature of the parameters.

9.1 NOTATIONS AND TERMINOLOGY FOR ESTIMATORS

Definition 9.1.1. *Let $X_1, X_2, X_3, \ldots, X_n$ be an i.i.d. sample from the population with distribution X. Let $g : \mathbb{R}^n \to \mathbb{R}$. Then $g(X_1, X_2, \ldots, X_n)$ is defined as a "point estimator" from the sample and the value from a particular realisation is called an "estimate".*

In practice the function g is chosen keeping in mind the parameter of interest. We have seen the following in Chapter 7.

Example 9.1.2. Let $E[X] = \mu$. Let $g : \mathbb{R}^n \to \mathbb{R}$ be given by

$$g(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Then $g(X_1, X_2, \ldots, X_n)$ is the (now familiar) sample mean and it is an estimator for μ. We also saw that $E[g(X_1, X_2, \ldots, X_n)] = \mu$ regardless of the true value of μ and we called such an estimator an unbiased estimator. ■

As noted in Chapter 7, we can view this as estimating the first moment of a distribution by the first moment of the empirical distribution based on a sample. A generalization of this method is known as the method of moments.

9.2 METHOD OF MOMENTS

Let $X_1, X_2, \ldots, X_n$ be a sample with distribution X. Assume that X is either has probability mass function or probability density function $f(x \mid p)$ depending on parameter(s) $p = (p_1, \ldots, p_d)$. For $d \geq 1$. Let $m_k : \mathbb{R}^n \to \mathbb{R}$ be given by

$$m_k(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

Notice that $m_k(X_1, X_2, \ldots, X_n)$ is the k-th moment of the empirical distribution based on the sample $X_1, X_2, \ldots, X_n$, which we will refer to simply as the k-th moment of the sample.

Let $\mu_k = E[X^k]$, the k-th moment of the distribution X. Since the distribution of X depends on $(p_1, p_2, \ldots p_d)$ one can view μ_k as a function of p, which we can make explicit by the notation $\mu_k \equiv \mu_k(p_1, p_2, \ldots p_d)$. The method of moments estimator for $(p_1, p_2, \ldots p_d)$ is obtained by equating the first d moments of the sample to the corresponding moments of the distribution. Specifically, it requires solving the d equations in d unknowns given by

$$\mu_k(p_1, p_2, \ldots p_d) = m_k(X_1, X_2, \ldots, X_n), \quad k = 1, 2, \ldots d.$$

for $p_1, p_2, \ldots p_d$. There is no guarantee in general that these equations have a unique solution or that it can be computed, but in practice it is usually possible to do so. The solution will be denoted by $\hat{p}_1, \hat{p}_2, \ldots \hat{p}_d$ which will be writen in terms of the realised values for $m_k, k = 1, \ldots, d$. We will now explore this method for two examples.

Example 9.2.1. Suppose $X_1, X_2, \ldots, X_{10}$ is an i.i.d. sample with distribution Binomial(N, p) where neither N nor p is known. Suppose the empirical realisation of these variables is 8, 7, 6, 11, 8, 5, 3, 7, 6, 9. One can check that the average of these values is $m_1 = 7$ while the average of their squares is $m_2 = 53.4$. Since $X \sim$ Binomial (N, p) the probability mass function is given by ???. We have previously shown that

$$E[X] = Np \text{ and } E[X^2] = \text{Var}[X] + E[X]^2 = Np(1-p) + N^2p^2.$$

Thus, the method of moments estimator for (N, p) is obtained by solving

$$7 = m_1 = \hat{N}\hat{p} \text{ and } 53.4 = m_2 = \hat{N}\hat{p}(1-\hat{p}) + \hat{N}^2\hat{p}^2.$$

Using elementary algebra we see that

$$\begin{aligned} \hat{N} &= \frac{m_1^2}{m_1 - (m_2 - m_1^2)} \approx 19 \\ \hat{p} &= \frac{m_1 - (m_2 - m_1^2)}{m_1} \approx 0.371. \end{aligned}$$

The method of moments estimates that the distribution from which the sample came is Binomial$(19, 0.371)$. As we noted at the beginning of the chapter, we may wish to restrict the parameters based on the context of the problem. Since the N value is surely some integer, the estimate of $\hat{N}$ was rounded to the nearest meaningful value in this case. ■

EXAMPLE 9.2.2. Suppose our quantity of interest X has a Normal (μ, σ^2) distribution. Therefore our probability density function is given by

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ x \in \mathbb{R}.$$

Let $X_1, X_2, X_3, \ldots, X_n$ be an i.i.d. sample from the population. We have shown that

$$E[X] = \mu \text{ and } E[X^2] = \text{Var}[X] + E[X]^2 = \mu^2 + \sigma^2.$$

The method of moments estimator for μ, σ is found by solving

$$m_1 = \mu \text{ and } m_2 = \mu^2 + \sigma^2.$$

from which

$$\begin{aligned} \hat{\mu} &= m_1 = \overline{X} \text{ and} \\ \hat{\sigma} &= \sqrt{m_2 - m_1^2} = \sqrt{\frac{n-1}{n}}\, S. \end{aligned}$$

■

It happens that the method of moment estimators may not be very reliable. For instance in the first example the estimate for p could be negative, occurring when the sample mean is smaller than the sample variance. Such defects can be somewhat rectified using moment matching and other techniques (see [CasBer90]).

9.3 MAXIMUM LIKELIHOOD ESTIMATE

Let $n \geq 1$, $p = (p_1, p_2, \ldots, p_d) \in \mathbb{R}^d$ and $X_1, X_2, \ldots, X_n$ be a sample from the population described by X. Assume that X either has probability mass function or probability density function denoted by $f(x \mid p)$ depending on parameter(s) $p \in \mathcal{P} \subset \mathbb{R}^d$.

DEFINITION 9.3.1. *The likelihood function for the sample* $(X_1, X_2, \ldots, X_n)$ *is the function* $L : \mathcal{P} \times \mathbb{R}^n \to \mathbb{R}$ *given by*

$$L(p; X_1, \ldots X_n) = \prod_{i=1}^{n} f(X_i \mid p).$$

For a given $(X_1, X_2, \ldots, X_n)$*, suppose* $\hat{p} \equiv \hat{p}(X_1, X_2, \ldots, X_n)$ *is the point at which* $L(p; X_1, \ldots X_n)$ *attains its maximum as a function* p*. Then* $\hat{p}$ *is called the maximum likelihood estimator of* p *(or abbreviated as MLE of* p*) given the sample* $(X_1, X_2, \ldots, X_n)$*.*

One observes readily that the likelihood function is the joint density or joint mass function of $(X_1, X_2, \ldots, X_n)$. The MLE $\hat{p}$ obtained is the most likely value of the parameter p, given that it is the value at which f is maximised for the given realisation of $(X_1, X_2, \ldots, X_n)$.

EXAMPLE 9.3.2. Let $p \in \mathbb{R}$ and $(X_1, X_2, \ldots, X_n)$ be from a population distributed as Normal with mean p and variance 1. Then the likelihood function is given by

$$L(p; X_1, \ldots X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - p)^2}{2}} = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{\sum_{i=1}^{n}(X_i - p)^2}{2}}.$$

To find the MLE, treating the given the realisation $X_1, X_2, \ldots, X_n$ as fixed, one needs to maximise L as a function of p. This is equivalent to finding the minimum of $g : \mathbb{R} \to \mathbb{R}$ given by

$$g(p) = \sum_{i=1}^{n} (X_i - p)^2.$$

Method 1: Since $g(p) = \sum_{i=1}^{n}(X_i - \overline{X})^2 + (\overline{X} - p)^2$ (See Exercise 9.3.1) and first term is always non-negative, the minimum of g will occur when

$$p = \overline{X}.$$

Method 2: The second method is to find the MLE using differential calculus. As g is a quadratic in p, it is differentiable at all p and

$$g'(p) = -2\sum_{i=1}^{n}(X_i - p).$$

As the coefficient of p^2 in g is n (which is positive), and g is quadratic, the minimum will occur in the interior when $g'(p) = 0$. This occurs when p is equal to $\frac{1}{n}\sum_{i=1}^{n} X_i$. So the MLE of p is given by

$$\hat{p} = \overline{X}.$$

■

EXAMPLE 9.3.3. Let $p \in (0,1)$ and $(X_1, X_2, \ldots, X_n)$ be from a populatioon distributed as Bernoulli (p). Now the probability mass function f can be written as

$$f(x \mid p) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} p^x(1-p)^{1-x} & \text{if } x \in \{0,1\} \\ 0 & \text{otherwise.} \end{cases}$$

Then the likelihood function is given by

$$L(p; X_1, \ldots X_n) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

To find the MLE, treating the given the realisation $X_1, X_2, \ldots, X_n$ as fixed, one needs to maximise L as a function of p. One needs to use calculus to find the MLE but differentiating L is cumbersome. So we will look at the logarithm of L (called the log likelihood function).

$$\begin{aligned} T(p; X_1, \ldots X_n) &= \ln L(p; X_1, \ldots X_n) \\ &= \begin{cases} \ln\left(\frac{p}{1-p}\right)a + n\ln(1-p) & \text{if } \sum_{i=1}^{n} X_i = a, 0 < a < n \\ n\ln(1-p) & \text{if } \sum_{i=1}^{n} X_i = 0 \\ n\ln(p) & \text{if } \sum_{i=1}^{n} X_i = n \end{cases} \end{aligned}$$

Therefore, in the first case, differentiating and settting it to zero $p = \frac{a}{n}$. This in fact can be verified to be the global maximum. In the second case T is a decreasing function of p and maximum occurs at $p = 0$. In the final case t is an increasing function of p and maximum occurs at $p = 1$. Therefore we can conclude that

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

■

As a final example, let us revisit Example 9.2.1, where we considered a Binomial distribution with both parameters unknown.

EXAMPLE 9.3.4. Suppose $X_1, X_2, \ldots, X_n$ is an i.i.d. sample with distribution Binomial(N, p) where neither N nor p is known. The likelihood function is given by

$$L(N, p; X_1, \ldots, X_n) = \prod_{i=1}^{k}\binom{N}{x_i}p^{X_i}(1-p)^{N-X_i}$$

...

■

EXERCISES

Ex. 9.3.1. Show that for any real numbers $p, x_1, x_2, \dots, x_n$

$$\sum_{i=1}^{n}(x_i - p)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + (\overline{x} - p)^2.$$

9.4 CONFIDENCE INTERVALS

Let $X_1, X_2, \dots, X_n$ be an i.i.d. sample from a distribution X with unknown mean μ. The sample mean $\overline{X}$ is an unbiased estimator for μ, but the empirical value of the sample mean will likely differ from μ by some unknown amount. Suppose we want to produce an interval, centered around $\overline{X}$ which we can be fairly certain contains the true average μ. This is known as a "confidence interval" and we explore how to produce such a thing in two different settings below.

9.4.1 *Confidence Intervals when the standard deviation σ is known*

Let X have a probability mass function or probability density function $f(x \mid \mu)$ where the distribution X has an unknown expected value μ, but a known standard deviation σ. Let $X_1, X_2, \dots, X_n$ be an i.i.d. sample from the distribution X. Let $\beta \in (0, 1)$ denote a "confidence level". We want to find an interval width a such that

$$P(|\overline{X} - \mu| < a) = \beta.$$

That is, the sample mean $\overline{X}$ will have a probability β of differing from the true mean μ by no more than the quantity a.

Let

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}.$$

Then $E[Z] = 0$ and $Var[Z] = 1$. Observe,

$$\begin{aligned}
\beta = P(|\overline{X} - \mu| < a) = P(|\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}| < \frac{\sqrt{n}a}{\sigma}) \\
= \quad P(|Z| < \frac{\sqrt{n}a}{\sigma}) \\
= \quad P(-\frac{\sqrt{n}a}{\sigma} < Z < \frac{\sqrt{n}a}{\sigma})
\end{aligned}$$

If X has a normal distribution, then $Z \sim Normal(0, 1)$ by Example 6.3.12. If the distribution of X is unknown, but n is large, then by the Central Limit Theorem (Theorem 8.4.1), Z should still be roughly normal, so we can make a valid assumption that $Z \sim Normal(0, 1)$. When n, σ, and β are known, one can use the Normal tables (Table D.2) to find the unknown interval width a. The interval $(\overline{X} - a, \overline{X} + a)$ is then known as "a β confidence interval for μ". The interpretation being that the random sample of size n from the distribution X should produce confidence intervals that include the correct value of μ 95% with probabilty β.

EXAMPLE 9.4.1. Suppose X has a normal distribution with known standard deviation $\sigma = 3.0$ and an unknown mean μ. A sample $X_1, X_2, \dots, X_{16}$ of i.i.d. random variables is taken from distribution X. The sample average of these 16 values comes out to be $\overline{X} = 10.2$. What would be a 95% confidence interval for the actual mean μ?

In this case $\beta = 0.95$ so we must find the value of a for which

$$P(|\overline{X} - \mu| < a) = 0.95.$$

From the computation above, this is equivalent to the equation

$$P(-\frac{4a}{3} < Z < \frac{4a}{3}) = 0.95$$

where $Z \sim Normal(0,1)$. Using the normal table, this is equivalent to $\frac{4a}{3} \approx 1.96$, and so $a \approx 1.47$. In other words, a 95% confidence interval for the actual mean of the distribution X is $(8.73, 11.67)$.

It should be noted that the only random variable in the expression $P(|\overline{X} - \mu| < a) = 0.95$ is the $\overline{X}$ variable. The interpretation is that random samples of size $n = 16$ from the distribution X should produce confidence intervals that include the correct value of μ, 95% of the time. ■

9.4.2 *Confidence Intervals when the standard deviation σ is unknown*

In most realistic situations the standard deviation σ would be unknown and would have to be estimated from the sample standard deviation S. In this case a confidence interval may still be produced, but an approximation via a normal distribution is insufficient.

Let X have normal distribution with density function $f(x \mid \mu, \sigma)$ where μ and σ are unknown. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the distribution X. Let $\overline{X}$ be the sample mean and the sample variance be denoted by S^2. Let $\beta \in (0,1)$ denote a confidence level. As before, we want to find an interval width a such that $P(|\overline{X} - \mu| < a) = \beta$.

Let

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S}.$$

From Corollary 8.1.10, $T \sim t_{n-1}$. In a simliar fashion as the previous example,

$$\begin{aligned}
\beta = P(|\overline{X} - \mu| < a) &= P(|\frac{\sqrt{n}(\overline{X} - \mu)}{S}| < a) \\
&= P(|T| < \frac{\sqrt{n}a}{S}) \\
&= P(-\frac{\sqrt{n}a}{S} < T < \frac{\sqrt{n}a}{S})
\end{aligned}$$

where $T \sim t_{n-1}$. Since n, S, and β are known, this equation can be solved to find the unknown interval width a using the t-distribution.

Example 9.4.2. Suppose X has a normal distribution with unknown mean μ and unknown standard deviation σ. A sample $X_1, X_2, \ldots, X_{16}$ of i.i.d. random variables with distribution X is taken. The sample average of these 16 values comes out to be $\overline{X} = 10.2$ and the sample standard deviation is $S = 3.0$. What would be a 95% confidence interval for the actual mean μ?

In this case $\beta = 0.95$ so we must find the value of a for
$P(|\overline{X} - \mu| < a) = 0.95$. This is equivalent to the equation

$$P(-\frac{4a}{3} < T < \frac{4a}{3}) = 0.95$$

Using the t-distribution, this is equivalent to $\frac{4a}{3} \approx 2.13$, and so $a \approx 1.60$. In other words, a 95% confidence interval for the actual mean of the distribution X is $(8.6, 11.80)$.

It is useful to compare this answer to the result from Example 9.4.1. Note that despite the similarity of the mean and standard deviation, the 95% confidence interval based on the t-distribution is a bit wider than the confidence interval based on the normal distribution. The reason is that in Example 9.4.1 the standard deviation was known exactly, while in this example the standard deviation needed to be estimated from the sample. This introduces an additional source of random error into the problem and thus the confidence interval must be wider to ensure the same likelihood of containing the true value of μ. ■

Example 9.4.3. Consider the example given at the start of the chapter. A coin is flipped 100 times with a result of 60 heads and 40 tails. What would be a 90% confience interval for the actual probability p that the coin shows heads on any given flip?

We represent a flip by $X \sim \text{Bernoulli}(p)$. From the i.i.d. sample $X_1, X_2, \ldots, X_{100}$ we have $\hat{p} = 60/100 = 0.6$. Despite the fact that σ is unknown, it would be inapprorpiate to use a t-distribution for the confidence interval in this case because X is far from a normal distribution. But we may still appeal to the Central Limit Theorem and accept the sample as providing a reasonable estimate for the standard deviation. That is, if $X \sim \text{Bernoulli}(0.6)$, then $\sigma = SD[X] = \sqrt{0.24}$. Using this approximation $\sigma \approx \sqrt{0.24}$ we may proceed as before...... ■

EXERCISES

Ex. 9.4.1. are t intervals always larger than Normal

9.5 HYPOTHESIS TESTING

The idea of hypothesis testing is another approach to comparing an observed quantity from a sample (such as $\overline{X}$) to an expected result based on an assumption about the distribution X. There are many different types of hypothesis tests. What follows is far from an exhausive list, but we explore some particular forms of hypothesis testing built around four familiar random distributions – the z-test, the t-test, the F-test, and the χ^2-test.

For any hypothesis test a "null hypothesis" is a specific conjecture made about the nature of the distribution X. This is compared to an "alternate hypothesis" that specifies a particular manner in which the null may be an inaccurate assumption.

A computation is then performed based on the differences between the sample data and the result which would have been expected if the null hypothesis were true. This computation results in a quantity called a "P-value" which describes the probability that sample would be at least as far from expectation as was actually observed. The nature of this comparison between observation and expectation varies according to the type of test performed and the assumptions of the null hypothesis. A small P-value is an indication that the sample would be highly unusual (casting doubt on the null hypothesis), which a large P-value indicates that the sample is quite consistent with the assumptions of the null.

9.5.1 *The z-test: Test for sample mean when σ is known*

Suppose $X \sim Normal(\mu, \sigma^2)$ where μ is an unknown mean, but σ is a known standard deviation. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the distribution X. Select as a null hypothesis the assumption that $\mu = c$ for some value c less than the observed average $\overline{X}$. Since the sample average $\overline{X}$ is larger than the assumed mean, the assumption $\mu > c$ may be an appropriate alternate hypothesis. If the null is true, how likely is it we would have seen a sample mean as large as the observed value $\overline{X}$?

To answer this question, we assume the empirical values of the sample $X_1, X_2, \ldots, X_n$ are known and let $Y_1, Y_2, \ldots, Y_n$ be an i.i.d. sample from the same distribution X. The Y_j variables effectively mimic the sampling procedure, an idea that will be commonly used through all tests of significance we consider.

We then calculate $P(\overline{Y} \geq \overline{X})$ where $\overline{Y}$ is viewed as a random variable and $\overline{X}$ is taken as the (deterministic) observed sample average. The $\overline{X}$ statistic calculated from the observed data is known as the "test statistic". The probability $P(\overline{Y} \geq \overline{X})$ desribes how likely it is that the test statistic would be at least as far away from μ as what was observed. This probability can be computed precisely because the distribution of $\overline{Y}$ is known exactly. Specifically, if the null hypothesis is true, then

$$Z = \frac{\sqrt{n}(\overline{Y} - c)}{\sigma} \sim Normal(0, 1).$$

So

$$\begin{aligned} P(\overline{Y} \geq \overline{X}) &= P(\frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} \geq \frac{\sqrt{n}(\overline{X} - c)}{\sigma}) \\ &= P(Z \geq \frac{\sqrt{n}(\overline{X} - c)}{\sigma}). \end{aligned}$$

In practice, before a sample is taken, a "signficance level" $\alpha \in (0, 1)$ is typically selected. If $P(\overline{Y} \geq \overline{X}) < \alpha$ then the sample average is so far from the assumed mean c that the assumption $\mu = c$ is judged to be incorrect. This is known as "rejecting the null hypothesis". Alternatively if $P(\overline{Y} \geq \overline{X}) \geq \alpha$ then the sample mean is seen as consistent with the assumption $\mu = c$ and the null hypothesis is not rejected.

Example 9.5.1. Suppose X has a normal distribution with known standard deviation $\sigma = 3.0$. A sample $X_1, X_2, \ldots, X_{16}$ of i.i.d. random variables is taken, each with distribution X. If the observed sample mean is $\overline{X} = 10.2$, what conclusion would a z-test reach if the null hypothesis assumes $\mu = 9.5$ (against an alternate hypothesis $\mu > 9.5$) at a signifiance level of $\alpha = 0.05$? ■

Under the null hypothesis, $Y_1, Y_2, Y_3, \ldots, Y_{16} \sim Normal(9.5, 3.0)$ and independent, so

$$\begin{aligned} P(\overline{Y} \geq \overline{X}) &= P(\overline{Y} \geq 10.2) \\ &= P(\frac{4(\overline{Y} - 9.5)}{3} \geq \frac{4(10.2 - 9.5))}{3} \\ &= P(Z \geq \frac{4(10.2 - 9.5)}{3}) \approx 0.175 \end{aligned}$$

where the final approximation is made using the fact that $Z \sim Normal(0, 1)$. The 0.175 figure is the P-value. Since it is larger than our significance level of $\alpha = 0.05$ we would not reject null hypothesis. Put another way, if the $\mu = 9.5$ assumption is true, the sampling procedure will produce a result at least as large as the sample average $\overline{X} = 10.2$ about 17.5% of the time. This is common enough that we cannot reject the $\mu = 9.5$ assumption.

Example 9.5.2. Make the same assumptions as in Example 9.5.1, but this time test a null hypothesis that $\mu = 8.5$ (with an alternate hypothesis $\mu > 8.5$ and a significance level of $\alpha = 0.05$). Under the null hypothesis $Y_1, Y_2, Y_3, \ldots, Y_{16} \sim Normal(8.5, 3.0)$ and are independent, so

$$\begin{aligned} P(\overline{Y} \geq \overline{X}) &= P(\overline{Y} \geq 10.2) \\ &= P(Z \geq \frac{4(10.2 - 8.5)}{3}) \approx 0.012 \end{aligned}$$

Since 0.012 is less than $\alpha = 0.05$ the null hypothesis would be rejected and the test would reach the conclusion that the true mean μ is some value larger than 8.5. Put another way, if the $\mu = 8.5$ assumption is true, the sampling procedure will produce a result as large as the sample average $\overline{X} = 10.2$ only about 1.2% of the time. This is rare enough that we can reject the hypothesis that $\mu = 8.5$. ■

For a large sample, a z-test is commonly used even without the assumption that X has a normally distribution. This is justified by appealing to the Central Limit Theorem.

Example 9.5.3. Suppose a programmer is writing an app to identify faces based on digital photographs taken from social media. She wants to be sure that the app makes an accurate identification more than 90% of the time in the long run. She takes a random sample of 500 such photos and her app makes the correct identification 462 times - a 92.4% success rate. The programmer is hoping that this is an indication her app has a better than 90% success rate in the long run. However it is also possible the long term success rate is only 90%, but that the app happened to overperform this bar on the 500 photo sample. What does a z-test say about a null hypothesis that the app is only 90% accurate (compared to an alternate hypothesis that the app is more than 90% accurate with a significance level of $\alpha = 0.05$)?

The random variables in question are modeled by a Bernoulli distribution, as the app either reaches a correct conclusion or it does not. Under the null hypothesis $Y_1, Y_2, \ldots, Y_{500} \sim Bernoulli(0.9)$ and are independent. The sample proportion does not precisely have a normal distribution, but the Central Limit Theorem implies that the standardized quantity

$$\frac{Y_1 + \cdots + Y_{500} - 450}{\sqrt{45}}$$

should have approximately a $Normal(0, 1)$ distribution. Therefore

$$\begin{aligned} P(\frac{Y_1 + \cdots + Y_{500}}{500} &\geq \frac{X_1 + \cdots + X_{500}}{500}) \\ &= P(\frac{Y_1 + \cdots + Y_{500} - 450}{\sqrt{45}} \geq \frac{X_1 + \cdots + X_{500} - 450}{\sqrt{45}}) \\ &\approx P(Z \geq \frac{462 - 450}{\sqrt{45}}) \approx 0.03 \end{aligned}$$

Since 0.03 is less than $\alpha = 0.05$ the null hypothesis would be rejected and the test would reach the conclusion that the success rate for the app is greater than 90%. ■

The examples above concern tests on the right hand tail of a normal curve. That is, they test a null hypothesis $\mu = c$ against an alternate hypothesis $\mu > c$. It is also possible to perform a test on the left hand tail (testing a null hypothesis $\mu = c$ against an alternate hypothesis $\mu < c$) and even a two-tailed test (testing a null $\mu = c$ against an alternate $\mu \neq c$), an example of which follows below.

EXAMPLE 9.5.4. Suppose X has a normal distribution random variable with unknown mean and $\sigma = 6$. Suppose $X_1, X_2, \ldots, X_{25}$ is an i.i.d. sample taken with distriubtion X and that $\overline{X} = 6.2$. What conclusion would a z-test reach if the null hypothesis assumes $\mu = 4$ against an alternate hypothesis $\mu \neq 4$ at a signifiance level of $\alpha = 0.05$? Since the alternate hypothesis doesn't specify a particular direction in which the null may be incorrect, the appropriate probability to compute is

$$P(|\overline{Y} - 4| \geq |\overline{X} - 4|),$$

the probability that the absolute distance of a sample from the anticipated mean of 4.0 is larger than what was actually observed.

$$\begin{aligned} P(|\overline{Y} - 4| \geq |\overline{X} - 4|) &= 1 - P(|\overline{Y} - 4| < 2.2) \\ &= 1 - P(-2.2 \leq \overline{Y} - 4 \leq 2.2) \\ &= 1 - P(\frac{5(-2.2)}{6} \leq \frac{5(\overline{Y} - 4)}{6} \leq \frac{5(2.2)}{6}) \\ &= 1 - P(\frac{-11}{6} \leq Z \leq \frac{11}{6}) \approx 0.0668 \end{aligned}$$

since $Z \sim Normal(0, 1)$. As 0.0668 is slightly above the required significance level $\alpha = 0.05$ the test would not reject the null hypothesis. ■

9.5.2 *The t-test: Test for sample mean when σ is unknown*

As in the case of confidence intervals, when σ is unknown and estimated from the sample standard deviation S, an adjustment must be made by using the t-distribution.

Suppose X is known to be normally distributed with $X \sim Normal(\mu, \sigma^2)$ where μ and σ are unkonwn. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the distribution X. Select as a null hypothesis that $\mu = c$ and select $\mu > c$ as an alternate hypothesis. Regard $X_1, \ldots, X_n$ as empirically known and let $Y_1, \ldots, Y_n$ be i.i.d. random variables which mimic the sampling procedure.

Under the null $Y_1, \ldots, Y_n \sim Normal(c, \sigma^2)$ and are independent, from Corollary 8.1.10,

$$\frac{\sqrt{n}(\overline{Y} - c)}{S} \sim t_{n-1}$$

and so

$$\begin{aligned} P(\overline{Y} \geq \overline{X}) &= P(\frac{\sqrt{n}(\overline{Y} - c)}{S} \geq \frac{\sqrt{n}(\overline{X} - c)}{S}) \\ &= P(T \geq \frac{\sqrt{n}(\overline{X} - c)}{S}). \end{aligned}$$

The other aspects of the hypothesis test are the same except that the t-distriubtion must be used to calculate this final probability. As with the z-test, this could be performed as a one-tailed or a two-tailed test depending on the appropriate alternate hypothesis.

EXAMPLE 9.5.5. Suppose X has a normal distribution with unknown standard deviation. A sample $X_1, X_2, \ldots, X_{16}$ of i.i.d. random variables is taken, each with distribution X. The sample standard deviation is $S = 3.0$. What conclusion would a t-test reach if the null hypothesis assumes $\mu = 9.5$ at a signifiance level of $\alpha = 0.05$ (against the alternative hypothesis that $\mu > 9.5$)? ■

Under the null hypothesis, $Y_1, Y_2, \ldots, Y_{16} \sim Normal(9.5, \sigma)$ and independent, so

$$\begin{aligned} P(\overline{Y} \geq \overline{X}) &= P(\overline{Y} \geq 10.2) \\ &= P(\frac{4(\overline{Y} - 9.5)}{3} \geq \frac{4(10.2 - 9.5))}{3}) \\ &= P(T \geq \frac{4(10.2 - 9.5)}{3}) \\ &= P(T \geq \frac{14}{15}) \approx 0.183 \end{aligned}$$

since $T \sim t_{15}$. As $0.183 > \alpha = 0.05$ the null hypothesis would not be rejected.

It is informative to compare this to Example 9.5.1 which was identical except that it was assumed that $\sigma = 3.0$ was known exactly rather than estimated from the sample. Note that the P-value in the case of the t-test (0.183) was slightly larger than in the case of the z-test (0.175). The reason is the use of S, a random variable, in place of σ, a deterministic constant. This adds an additional random factor into the computation and therefore makes larger deviations from the mean somewhat more likely.

9.5.3 *A critical value approach*

An alternate way to view the tests above is to focus on a "critical value". Such a value is the dividing line beyond which the null will be rejected. If a test is being performed with a significance level α, then we can determine ahead of time where this line is and immediately reach a conclusion from the value of the test statistic without calculating a P-value. To demonstrate this we will redo Example 9.5.1 using this approach. In that example, X had a normal distribution with known standard deviation $\sigma = 3.0$ and $Y_1, Y_2, \ldots, Y_{16}$ were i.i.d. with distribution X. The null hypothesis assumed $\mu = 9.5$ while the alternate hypothesis was $\mu > 9.5$. The test had a signifiance level of $\alpha = 0.05$.

To find the critical value, we begin with the same computation as in Example 9.5.1, but keep $\overline{X}$ as a variable.

$$\begin{aligned} P(\overline{Y} \geq \overline{X}) &= P(\frac{4(\overline{Y} - 9.5)}{3} \geq \frac{4(\overline{X} - 9.5)}{3}) \\ &= P(Z \geq \frac{4(\overline{X} - 9.5)}{3}) \end{aligned}$$

Whether or not the null is rejected depends entirely on whether this probability is above or below the significance level $\alpha = 0.05$, so the relevant question is what value of c ensures $P(Z \geq c) = 0.05$. This is something that can be calculated using Normal tables Table D.2 and in fact, $c \approx 1.645$. We solve the equation

$$\frac{4(\overline{X} - 9.5)}{3} = 1.645$$

which yields $\overline{X} \approx 10.73$. The figure 10.73 is the critical value. It is the dividing line we were seeking; if the sample average is above 10.73 the null hypothesis will be rejected while a sample average less than 10.73 will not cause the test to reject the null. For this particular example it was assumed that the observed sample average was $\overline{X} = 10.2$ which is why the null hypothesis was not rejected.

9.5.4 *The χ^2-test : Test for sample variance*

Suppose instead of an inquiry about an average, we are interested in the variability in a population. Suppose $X \sim Normal(\mu, \sigma^2)$ with unknown σ. As with the previous hypothesis tests, we assume the empirical values of the sample $X_1, X_2, \ldots, X_n$ are known, and let $Y_1, Y_2, \ldots, Y_n$ be an i.i.d. sample with distribution X which mimics the sampling procedure. Select as a null hypothesis the assumption that $\sigma = c$ (and take $\sigma > c$ as an alternate hypothesis). How likely is it the sampling procedure would produce a sample standard deviation as large as S_X?

We wish to calculate $P(S_Y \geq S_X)$, the probability a sample would produce a standard deviation at least as large as what was observed. Under the null hypothesis, from Theorem 8.1.9

$$\frac{(n-1)}{c^2}S_Y^2 \sim \chi^2_{n-1}.$$

Therefore,

$$\begin{aligned} P(S_Y \geq S_X) &= P(S_Y^2 \geq S_X^2) \\ &= P(\frac{(n-1)}{c^2}S_Y^2 \geq \frac{(n-1)}{c^2}S_X^2) \\ &= P(W \geq \frac{(n-1)}{c^2}S_X^2) \end{aligned}$$

which may be calculated since $W \sim \chi^2_{n-1}$ and n, c, and S_X are known.

Example 9.5.6. Suppose X is normally distributed with unknown standard deviation σ. Let $X_1, X_2, \ldots, X_{16}$ be an i.i.d. sample with distriubtion X and a sample standard deviation $S_X = 3.5$. What conclusion would a χ^2-test reach if the null hypothesis assumes $\sigma = 3$, with an alternate hypothesis that $\sigma > 3$, and a signifiance level of $\alpha = 0.05$?

The null hypothsis ensures

$$\begin{aligned} P(S_Y \geq S_X) &= P(\frac{15}{\sigma^2}S_Y^2 \geq \frac{15}{\sigma^2}S_X^2) \\ &= P(W \geq \frac{15}{9}(3.5)^2) \approx 0.157 \end{aligned}$$

since $W \sim \chi^2_{15}$. So there is about a 15.7% chance that a sample of size 16 would produce a standard deviation as large as 3.5. As this P-value is larger than α the null hypothesis is not rejected. ■

The shapes of the normal distribution and the t-distribution are symmetric about their means. This implies that when considering an interval centered at the mean, the two tail probabilities are always equal to each other. For example, if $Z \sim Normal(\mu, \sigma^2)$, then $P(Z \geq \mu + c) = P(Z \leq \mu - c)$ regardless of the value of c. In particular, when carrying out a computation for a hypothesis test,

$$\begin{aligned} P(|Z-\mu| \geq c) &= P(Z \geq \mu + c) + P(Z \leq \mu - c) \\ &= 2P(Z \geq \mu + c) \end{aligned}$$

since both tails have the same probability. However, this is not so for the chi-squared distribution. When performing a two-tailed test involving a distribution which is not symmetric, the interval selected is the one which has equal tail probabilities, each of which equal half of the confidence level. Due to this fact it is usually best to use a critical value approach.

Example 9.5.7. Suppose X is normally distributed with unknown standard deviation σ. Let $X_1, X_2, \ldots, X_{16}$ be an i.i.d. sample with distriubtion X and a sample standard deviation $S_X = 3.5$, under the null hypothesis. What conclusion would a χ^2-test reach if the null hypothesis assumes $\sigma = 3$, with an alternate hypothesis that $\sigma \neq 3$, and a signifiance level of $\alpha = 0.05$?

As in the previous example, we let $Y_1, \ldots, Y_{16}$ replicate the sampling procedure and use that $\frac{15}{\sigma^2}S_Y^2$ has a χ^2_{15} distribution. With a significance level of $\alpha = 0.05$ the critical points will be the values of the χ^2_{15} distribution that correspond to tail probabilities of 2.5%. It may be calculated that if $W \sim \chi^2_{15}$, then $P(W \leq 6.26) \approx 0.025$ while $P(W \geq 27.49) \approx 0.025$. As is readily observed 6.26 is the 0.025-th quantile and 27.49 is the 0.975-th quantile. They define the low and high critical values beyond which would be considered among the 5% most unusual occurances for W. The corresponding obsevered value in the sample is $\frac{15}{9}(3.5)^2 \approx 20.42$ does not put it among this unusual 5% mark, so the null hypothesis would not be rejected.

■

9.5.5 *The two-sample z-test: Test to compare sample means*

Hypothesis tests may also be used to compare two samples to each other to see if the populations they were derived from were similar. This is of particular use in many applications. For instance: are the political opinions of one region different from another?; or are test scores at one school better than those at another school? These questions could be approached by taking random samples from each population and comparing them with each other.

Suppose $X_1, X_2, \dots, X_{n_1}$ is an i.i.d. sample from a distribution $X \sim Normal(\mu_1, \sigma_1^2)$ and suppose $Y_1, Y_2, \dots, Y_{n_2}$ is an i.i.d. sample from a distribution $Y \sim Normal(\mu_2, \sigma_2^2)$ independent of the X_j variables. Assume σ_1 and σ_2 are known, but μ_1 and μ_2 are not. How might we test a null hypothesis that $\mu_1 = \mu_2$ against an alternative hypothesis $\mu_1 \neq \mu_2$?

If the null hypothesis were true $\mu_1 - \mu_2 = 0$. We could calculate $\overline{X} - \overline{Y}$ and determine if this difference was close enough to 0 to make the null plausible. As usual, we mimic the sampling procedure, this time with both samples. Let $V_1, \dots, V_{n_1}$ be an i.i.d. sample with distribution X and let $W_1, \dots, W_{n_2}$ be an i.i.d. sample with distribution Y independent of the V_j variables. We would then calculate

$$P(|\overline{V} - \overline{W}| \geq |\overline{X} - \overline{Y}|),$$

the probability that the difference of sample averages would be at least as large as what was observed.

As $\overline{V} \sim Normal(\mu_1, \frac{\sigma_1^2}{n_1})$ and $\overline{W} \sim Normal(\mu_2, \frac{\sigma_2^2}{n_2})$. Under the null hypothesis the mean of $\overline{V} - \overline{W}$ is zero and they are independent with each having normal distribution. By Theorem 6.3.13 $\overline{V} - \overline{W} \sim Normal(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. Therefore,

$$\begin{aligned} P(|\overline{V} - \overline{W}| \geq |\overline{X} - \overline{Y}|) &= P(|\frac{\overline{V} - \overline{W}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}| \geq \frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}) \\ &= P(|\ Z\ | \geq \frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}) \\ &= 2P(Z \leq -\frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}) \end{aligned}$$

were $Z \sim Normal(0, 1)$.

Example 9.5.8. Suppose a biologist wants to know if the average weights of adult members of a species of squirrel in one forest is the same as an identical species in a different location. Historical data suggests that the weights of the species have a standard deviation $\sigma = 10$ grams and the biologist is willing to use this assumption for his computations. Suppose he takes a sample of 30 squirrels from each location is is willing to regard these as independent i.i.d. samples from the respective populations. The first sample average is 122.4 grams and the second sample average is 127.6 grams. What conclusion would a two-sample z-test reach testing a null hypothesis that the population averages are the same against a alternate hypothesis that they are different at a signifiance level of $\alpha = 0.05$?

Nothing in the statement of the problem suggests that an assumption about the normality of the populations, so we will need to appeal to the Central Limit Theorem and be content that this is a decent approximation. Let X and Y represent the distributions of weights of the two populations. Under the null hypothesis these distributions have equal means $\mu = \mu_X = \mu_Y$ and $\sigma = 10$ for both distirubtions. Let $V_1, \dots, V_{30}$ and $W_1, \dots, W_{30}$ be i.i.d samples from populations X and Y respectively. Observe that

$$\overline{V} - \overline{W} = \frac{1}{30} \sum_{i=1}^{30} (V_i - W_i).$$

Now $V_1 - W_1, V_2 - W_2, V_3 - W_3, \dots, V_{30} - W_{30}$ are i.i.d. with zero mean and standard deviation $\sqrt{10^2 + 10^2} = 10\sqrt{2}$. By the Central Limit Theorem, the distribution of

$$\frac{\sqrt{30}(\overline{V}-\overline{W})}{10\sqrt{2}}$$

is approximately standard normal. Therefore,

$$\begin{aligned} P(|\overline{V}-\overline{W}| \geq 5.2) &= P(|\frac{\sqrt{30}(\overline{V}-\overline{W})}{10\sqrt{2}}| \geq \frac{\sqrt{30}(5.2)}{10\sqrt{2}}) \\ &= P(Z \geq 0.52\sqrt{15}) \approx 0.0440 \end{aligned}$$

where $Z \sim Normal(0,1)$.

Since the P-value falls below the significance level, we would reject the null hypothesis and conclude that the populations have different average weights. ■

9.5.6 *The F-test: Test to compare sample variances.*

Let $X_1, X_2, \ldots, X_{n_1}$ be an i.i.d. sample from a distribution $X \sim Normal(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be an i.i.d. sample (independent of the X_j variables) from a distribution $Y \sim Normal(\mu_2, \sigma_2^2)$. Suppose we wish to test the null hypothesis $\sigma_1 = \sigma_2$ against the alternate hypothesis $\sigma_1 \neq \sigma_2$.

Let $V_1, \ldots, V_{n_1}$ replicate the sample from X and $W_1, \ldots, W_{n_2}$ replicate the sample from Y. Let S_V^2 and S_W^2 denote the respective sample variances. Based on the previous examples it may be tempting to perform a test based on the probability that $|S_V^2 - S_W^2|$ is as large as the observed difference $|S_X^2 - S_Y^2|$. However, the random variable $S_V^2 - S_W^2$ does not have a distribution we have already considered. Instead, we will look at the ratio $\frac{S_V^2}{S_W^2}$. If the null hypothesis is true, this value should be close to 1 and the ratio has the benefit of being related to a familiar F-distribution.

The random variables $\frac{(n_1-1)}{\sigma_1^2}S_V^2$ and $\frac{(n_2-1)}{\sigma_2^2}S_W^2$ are independent and by Theorem 8.1.9 have the distributions $\chi^2_{n_1-1}$ and $\chi^2_{n_2-1}$ respectively. Therefore from Example 8.1.7 the ratio $\frac{S_V^2 \cdot \sigma_2^2}{S_W^2 \cdot \sigma_1^2}$ has a $F(n_1 - 1, n_2 - 1)$ distribution, and under the null hypothesis, this ratio simplifies to

$$\frac{S_V^2}{S_W^2} \sim F(n_1-1, n_2-1).$$

Since this is a distribution for which we may compute associated probabilities, we may use it to perform the hypothesis test. As the F distribution is not symmetric, we take a critical values approach.

EXAMPLE 9.5.9. Suppose $X_1, X_2, \ldots, X_{30}$ is an i.i.d. sample from a distribution $X \sim Normal(\mu_1, \sigma_1^2)$ and suppose $Y_1, Y_2, \ldots, Y_{25}$ is an i.i.d. sample from a distribution $Y \sim Normal(\mu_2, \sigma_2^2)$ independent of the X_j variables. If $S_X^2 = 11.4$ and $S_Y^2 = 5.1$, what conclusion would an F-test reach for null hypothesis suggesting $\sigma_1 = \sigma_2$, an alternate hypothesis suggesting $\sigma_1 \neq \sigma_2$, and a signifiance level of $\alpha = 0.05$?

From the computation above, if $R = \frac{S_V^2}{S_W^2}$ is the ratio of the sample variances, then $R \sim F(29, 24)$. From this, it may be calculated that if $P(R \leq 0.464) \approx 0.025$ while $P(R \geq 2.22) \approx 0.025$. Since the observed ratio $\frac{11.4}{5.1} \approx 2.24$ is outside of the interval $(0.464, 2.22)$, the null hypothesis would be rejected. ■

9.5.7 *A χ^2-test for "goodness of fit"*

A more common use of the χ^2 is for something called a "goodness of fit" test. In this case we seek to determine whether the distribution of results in a sample could plausibly have come from a distribution specified by a null hypothesis. The test statistic is calculated by comparing the observed count of data points within specified categories relative to the expected number of results in those categories according to the assumed null.

Specifically, let X is a random variable with finite range $\{c_1, c_2, \ldots, c_k\}$ for which $P(X = c_j) = p_j > 0$ for $1 \leq j \leq k$. Let $X_1, X_2, \ldots, X_n$ be the empirical results of a sample from the distribution X and let

$Y_j = |\{j : X_j = c_j\}|$. That is, Y_j is the number of data points in the sample that resulted in the c_j outcome. Then the statistic $\chi^2 = \sum\limits_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j}$ has approximately the $\chi^2(k-1)$ distribution. Note that the np_j term is the expected number of observations of the outcome c_j, so the numerator of the fractions in the χ^2 computation measures the squared difference between the observed and expected counts.

A formal proof of the statement above requires a deeper understanding of linear algebra than we assume as a prerequiste for this text, but below we demonstrate its truth in the special case where $k = 2$ and we provide the formal technical details in the appendix. The approximation itself relies on the Central Limit Theorem and as with that theorem, larger values of n will tend to lead to a better approximation. Before proceeding to the proof, we present an example will help illustrate the use of the test.

Example 9.5.10. The student body at an undergraduate university is 20% seniors, 24% juniors, 26% sophomores, and 30% freshman. Suppose a researcher takes a sample of 50 such students. Within the sample there are 13 seniors, 16 juniors, 10 sophmores, and 11 freshmen. The researcher claims that his sampling procedure should have produced independent selections from the student body, with each student equally likely to be selected. Is this a plausible claim given the observed results?

If the claim is true (which we take as the null hypothesis), then selecting an individual for the sample should be like the empirical result from a random variable X with a range of { senior, junior, sophomore, freshman } where the probabilities of each outcome are the percentages desribed above. For instance, $p_{senior} = P(X = \text{“}senior\text{”}) = 0.2$. The expected value of results based on the null hypothesis is then 10 seniors, 12 juniors, 13 sophomores, and 15 freshmen. So,

$$\chi^2 = \frac{(13-10)^2}{10} + \frac{(16-12)^2}{12} + \frac{(10-13)^2}{13} + \frac{(11-15)^2}{15} \approx 3.99$$

Notice that the χ^2 statistic can never be less than zero, and that when observed results are close to what was expected, the resulting fraction is small and does not contribute much to the sum. It is only when there is a relatively large discrepency between observation and expectation that χ^2 will have a large value.

Since there were four categories, if the null hypothesis is correct, this statistic resulted from a $\chi^2(3)$ distribution. To see if such a thing is plausible, we let $W \sim \chi^2(3)$ and calculate that $P(W \geq 3.99) \approx 0.2625$. The researcher's claim seems plausible. According to the χ^2-test, samples this far from expectation should be observed around 26% of the time. ■

To give a bit more insight into this test, consider each term individually. Each of the variables $Y_1, Y_2, \ldots, Y_n$ is a binomial random variable. For example Y_1 represents the number of times the outcome c_1 is observed in n trials, when $P(X = c_1) = p_1$, so $Y_1 \sim \text{Binomial}(n, p_1)$. Therefore $E[Y_1] = np_1$ and $SD[Y_1] = \sqrt{np_1(1-p_1)}$. From the Central Limiit Theorem, the normalized quantity $\frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}$ has approximately the Normal$(0,1)$ distribution, and therefore its square $\frac{(Y_1-np_1)^2}{np_1(1-p_1)}$ has approximately a $\chi^2(1)$ distribution. Except for the $(1-p_1)$ term in the denominator, this is the first fraction in the sum of our test statistic. The additional factor in the denominator is connected to the reason the resulting distribution has $n-1$ degrees of freedom instead of n degrees of freedom; the variables $Y_1, Y_2, \ldots, Y_n$ are dependent. Untangling this dependence in the general case is complicated, but if $k = 2$ we can prove a rigorous statement without the use of linear algebra.

Theorem 9.5.11. *Let X be a random variable with finite range $\{c_1, c_2\}$ for which $P(X = c_j) = p_j > 0$ for $j = 1, 2$. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with distribution X and let $Y_j = |\{j : X_j = c_j\}|$. Then $\chi^2 = \sum\limits_{j=1}^{2} \frac{(Y_j - np_j)^2}{np_j}$ has the same distribution as $(\frac{Z - E[Z]}{SD[Z]})^2$ where $Z \sim$ Binomial(n, p_1)*

Proof - To simplify notation, let $p = p_1$ and note that $p_2 = 1 - p$. Also, let $N = Y_1$ and note $Y_2 = n - N$. Then,

$$\begin{aligned}
\chi^2 &= \frac{(N-np)^2}{np} + \frac{((n-N)-n(1-p))^2}{n(1-p)} \\
&= \frac{N^2 - 2npN + n^2p^2}{np(1-p)} \\
&= (\frac{N-np}{\sqrt{np(1-p)}})^2
\end{aligned}$$

Since $N = Y_1 \sim \text{Binomial}(n, p)$, the result follows.

The central limit theorem guarantees $\frac{N-np}{\sqrt{np(1-p)}}$ converges in distribution to a Normal$(0, 1)$ diistribution, the χ^2 quantity will have approximately a $\chi^2(1)$ distribution for large values of n.