Chapter 9

Sample Statistics

Suppose we have a random variable X, and we collect n independent samples $X_1, X_2, \ldots, X_n$ of this random variable. The probability model is that the samples are random variables $X_1, X_2, \ldots, X_n$ are mutually independent and identically distributed with the same distribution as X. As we discussed in Chapter **??**, the sample mean $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is an approximation to the expected value $\mathbb{E}[X]$ that converges with probability 1 to the true expected value $\mathbb{E}[X]$, by the Strong Law of Large Numbers.

For any finite n, the sample mean M_n is a random variable. This random variable is the sum of n independent random variables, so describing statistical properties such as its PDF if X were a continuous random variable would require computing n-fold convolutions of the PDF $f_X(x)$.

Nevertheless, we know

$$\mathbb{E}\left[M_n - \mathbb{E}[X]\right] = \mathbb{E}[\frac{1}{n}\sum_{k=1}^{n} X_k] - \mathbb{E}[X] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[X_k] - \mathbb{E}[X] = 0$$

from the property that all the X_k are identically distributed. Under the assumption that the random variable X has finite variance σ^2, we can also compute

$$\mathsf{Var}[M_n] = \frac{1}{n^2}\sum_{k=1}^{n} \mathsf{Var}[X_k] = \frac{\sigma^2}{n},$$

because the random variables X_k are independent.

The Central Limit Theorem states that a scaled version of $M_n - \mathbb{E}[X]$ has a CDF that converges to that of a standard Gaussian random variable with mean 0 and variance 1. Specifically, we define $Z_n = \sqrt{n}\frac{M_n - \mathbb{E}[X]}{\sigma}$ as the scaled random variable. Then,

$$\lim_{n\to\infty} \mathbb{P}[\{Z_n \leq x\}] = \Phi(x), \qquad \text{for all } x \in \Re.$$

In this chapter, we are concerned with finite collections of independent samples of a random variable, using these samples to estimate properties of the random variable X. Unlike the limit results of **??**, we want to estimate the accuracy we can obtain from a fixed finite number of samples n. We consider problems in both estimation and detection. For instance, we want to estimate the average height of women in the Boston are by measuring the height of 100 women, uniformly selected from Boston's population. How accurate will our estimate be? As another instance, consider conducting a trial for a new vaccine trial with a test group of 100 subjects and a control group of another 100 subjects. Do the results indicate that the vaccine makes a significant difference, and what confidence do we have in that conclusion?

9.1 Estimation of Mean and Variance

If we don't know the true mean, but can collect independent samples of X, the sample mean M_n is often a reasonable estimator for the true mean $\mathbb{E}[X]$. The sample mean is computed by generating n independent, identically distributed $X_k, k = 1, \ldots, n$, each of which is identically distributed as X. In this case,

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

has mean $\mathbb{E}[X]$, and M_n converges to $\mathbb{E}[X]$ by the Strong Law of Large Numbers. If X has finite variance σ^2, M_n is the sum of independent, identically distributed random variables, and thus has variance $\frac{\sigma^2}{n}$.

Suppose that we would like to estimate the variance σ^2 of X. Assuming the variance is finite, it is obtained as

$$\mathsf{Var}[X] = \sigma^2 = \mathbb{E}[\Big(X - \mathbb{E}[X]\Big)^2].$$

Given knowledge of $\mathbb{E}[X]$, and samples $X_1, \ldots, X_n$, an estimate of the variance can be obtained as

$$\widehat{V}_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mathbb{E}[X])^2.$$

Since the X_k are independent and identically distributed as X, we have

$$\mathbb{E}[\widehat{V}_n] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[(X_k - \mathbb{E}[X])^2] = \frac{1}{n}\sum_{k=1}^{n}\mathsf{Var}[X] = \mathsf{Var}[X].$$

By the Strong Law of Large Numbers, we know $\lim_{n\to\infty}\widehat{V}_n = \mathsf{Var}[X]$ with probability 1.

What if we did not know the mean $\mathbb{E}[X]$, but had only the sample values $X_k, k = 1, \ldots, n$ to estimate the variance? In this case, we may consider estimating the variance by using the sample mean M_n. That is,

$$\overline{V}_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - M_n)^2.$$

This can be simplified as

$$\begin{aligned}\overline{V}_n &= \frac{1}{n}\sum_{k=1}^{n}(X_k^2 - 2X_kM_n + M_n^2)\\ &= \frac{1}{n}\sum_{k=1}^{n}X_k^2 - 2\big(\frac{1}{n}\sum_{k=1}^{n}X_k\big)M_n + M_n^2\\ &= \frac{1}{n}\sum_{k=1}^{n}X_k^2 - 2M_n^2 + M_n^2 = \frac{1}{n}\sum_{k=1}^{n}X_k^2 - M_n^2\end{aligned}$$

Using the previous equation, we can compute the expected value of this estimate as

$$\mathbb{E}[\overline{V}_n] = \mathbb{E}[X^2] - \mathbb{E}[M_n^2] = \mathbb{E}[X]^2 + \mathsf{Var}[X] - \mathbb{E}[X]^2 - \frac{\mathsf{Var}[X]}{n} = \frac{n-1}{n}\mathsf{Var}[X].$$

This shows that the estimate $\overline{V}_n$ is a biased estimate of $\mathsf{Var}[X]$, and underestimates it by a small amount. To compensate for this, one can use the unbiased estimate:

$$V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2.$$

This sample variance is an unbiased estimate of the true variance of X based on the samples $X_1, \ldots, X_n$. Most computer packages compute the sample variance as V_n.

What about an estimate of the standard deviation? While we can generate different estimates for the standard deviation directly, the common definition the sample standard deviation is

$$\widehat{\sigma}_n = \sqrt{V_n}.$$

This guarantees the consistent interpretation that the sample standard deviation is the square root of the sample variance,

9.2 Confidence Intervals for Sample Means

In the press, we read reports that quote statistics such as $57\% \pm 3\%$ of responders prefer brand A to brand B, with confidence interval 95%. How were such numbers calculated? We discuss this in this section.

Assume we have a random variable X, and we collect n independent samples X_k of X. Assume X has finite mean μ and variance σ^2. The sample mean of X, given n samples X_i is $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, which is a random variable. From the previous analysis, we know

$$\mathbb{E}[M_n] = \mathbb{E}[X] = \mu; \; \mathsf{Var}[M_n] = \frac{\mathsf{Var}[X]}{n} = \frac{\sigma^2}{n}.$$

M_n is an estimate of $\mathbb{E}[X]$. Given a small constant α, we want to find an interval $[A, B]$ such that

$$\mathbb{P}[\{A \leq \mathbb{E}[X] \leq B\}|M_n] = 1 - \alpha.$$

If we find such numbers, $B - A$ is called the confidence interval and $1 - \alpha$ the confidence.

Often, we select the interval to be centered about the sample mean M_n, in order to determine how close M_n is to the true mean $\mathbb{E}[X]$. Specifically, consider the event $\{|M_n - \mu| < \epsilon\}$ for some $\epsilon > 0$. Given the statistical properties of M_n, we can compute $\mathbb{P}[\{|M_n - \mu| < \epsilon\}] = q$. We say that the true mean is in the interval $[M_n - \epsilon, M_n + \epsilon]$ with confidence q.

We can use several of the limit theorems from Chapter 8 to estimate these confidence intervals. The variance of M_n is $\frac{\sigma^2}{n}$, which is small for large values numbers of samples n. If we know σ^2, the Chebyshev inequality yields

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Thus, $\mathbb{P}[\{|M_n - \mu| < \epsilon\}] \geq 1 - \frac{\sigma^2}{n\epsilon^2}$, yielding an estimate of the confidence level $q = 1 - \frac{\sigma^2}{n\epsilon^2}$ for fixed values of n, ϵ.

If the random variables X_k are bounded with values in $[a, b]$, we can use Hoeffding's inequality to get an improved confidence level:

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Thus, the true mean is in the interval $[M_n - \epsilon, M_n + \epsilon]$ with confidence level $q = 1 - 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$.

For large n, we can approximate this probability using the Central Limit Theorem (CLT). The CLT states that the random variable $Z = \frac{\sqrt{n}(M_n - \mu)}{\sigma}$ has the distribution of a standard Gaussian random variable, so $Z \sim \mathcal{N}(0, 1)$. Then,

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] = \mathbb{P}[\{|Z| \geq \frac{\sqrt{n}\epsilon}{\sigma}\} = 2(1 - \Phi(\frac{\epsilon\sqrt{n}}{\sigma})) = 2Q(\frac{\epsilon\sqrt{n}}{\sigma}).$$

Thus, the true mean is in the interval $[M_n - \epsilon, M_n + \epsilon]$ with confidence level $q = 1 - 2Q(\frac{\epsilon\sqrt{n}}{\sigma})$.

If we know ϵ and n, we can estimate the confidence level that $|M_n - \mu| < \epsilon$ using the above results. What if we knew ϵ and the desired confidence level $1 - \alpha$, and wanted to know how large n had to be to get that confidence level for that interval?

To answer this, if we are using the CLT, we determine a value T such that $Q(T) = \alpha/2$ using the standard Gaussian CDF table in Appendix C. Then,

$$\mathbb{P}[\{|\frac{\sqrt{n}(M_n - \mu)}{\sigma_X} \leq T\}] \approx 1 - \alpha,$$

or equivalently,

$$\mathbb{P}[|M_n - \mu| \leq \frac{T\sigma_X}{\sqrt{n}}] \approx 1 - \alpha.$$

This translates to the following statement: with confidence level $1-\alpha$, the true expected value μ lies in the interval $[M_n - \frac{T\sigma_X}{\sqrt{n}}, M_n + \frac{T\sigma}{\sqrt{n}}]$. The length of the confidence interval is $2\frac{T\sigma}{\sqrt{n}}$. To determine the number of samples n required for an interval of length 2ϵ, we solve

$$\frac{T\sigma_X}{\sqrt{n}} = \epsilon \Longrightarrow n = (\frac{T^2\sigma^2}{\epsilon^2}).$$

Similarly, we can determine how large n needs to be using other bounds such as the Chebyshev bound or Hoeffding's Inequality. We illustrate this process in the examples below.

Example 9.1
Suppose X is a Bernoulli(0.25) random variable. Assume we collect 100 independent, identically distributed samples of X, denoted as $X_k, k = 1, \ldots, 100$. Define $M_{100} = \frac{1}{100}\sum_{k=1}^{100} X_i$. Estimate the probability $\mathbb{P}[\{|M_{100} - 0.25| > 0.01\}]$.

The variance of a Bernoulli(p) random variable is $p(1-p)$. Thus, the variance of X is $\frac{3}{16}$, and the standard deviation is $\frac{\sqrt{3}}{4}$. Using the Chebyshev Inequality, we obtain

$$\mathbb{P}[\{|M_{100} - 0.25| > 0.01\}] \leq \frac{\frac{3}{16}}{100 \cdot 0.01^2} \leq \frac{3}{0.16} = 18.75$$

which is a useless estimate, as we know probabilities are less than 1. This means we don't have enough samples to estimate the mean of X accurately.

Since Bernoulli random variables take values in $[0,1]$, Hoeffding's inequality yields

$$\mathbb{P}[\{|M_{100} - \mu| \geq 0.01\}] \leq 2e^{-\frac{200(0.01)^2}{1^2}} = 2e^{-0.02}$$

which is also a number greater than one, so it is not a useful bound.

What about the estimate from the Central Limit Theorem? In this case, M_n is approximated by a Gaussian with mean 0.25 and variance $\frac{3}{16\cdot100} = \frac{3}{1600}$. The transformation $Z = \frac{M_n - 0.25}{\sqrt{\frac{3}{1600}}} = \frac{40(M_n - 0.25)}{\sqrt{3}}$ makes Z a standard Gaussian random variable. The event $\{|M_n - 0.25| > 0.01\}$ is equivalent to the event $\{|Z| > \frac{0.4}{\sqrt{3}}\}$. Thus, we can estimate the desired probability as

$$\mathbb{P}[\{|M_n - 0.25| > 0.01\}] \approx \mathbb{P}[\{|Z| > \frac{0.4}{\sqrt{3}}\}] = 2Q(\frac{0.4}{\sqrt{3}}) \approx 0.8174.$$

Example 9.2
Continuing the example 9.1, we would like to estimate the required number of samples n so that the sample mean $M_n \in [\mu - 0.01, \mu + 0.01]$ with confidence 0.95.

Using the Chebyshev inequality, we want

$$\mathbb{P}[\{|M_n - 0.25| > 0.01\}] \leq \frac{\frac{3}{16}}{n \cdot 0.01^2} \leq 0.05$$

Combining the last two equations, we get $n \geq \frac{\frac{3}{16}}{5\cdot(0.01)^3} = \frac{300{,}000}{8} = 37,500$. It is clear why 100 samples were inadequate in the previous example.

Using Hoeffding's inequality yields

$$\mathbb{P}[\{|M_n - \mu| \geq 0.01\}] \leq 2e^{-\frac{2n(0.01)^2}{1^2}} = 2e^{-\frac{2n}{10000}}.$$

Let n be such that

$$2e^{-\frac{2n}{10000}} = 0.05 \iff -\frac{2n}{10000} = \ln(0.025) \iff n = 5000\ln(40) \approx 18,444.$$

Using the Central Limit Theorem, we get the following estimate:

$$\mathbb{P}[\{|M_n - 0.25| > 0.01\}] \approx 2Q(\frac{0.01}{\sqrt{\frac{3}{16n}}}) \leq 0.05.$$

We use the standard Gaussian table in Appendix C to find the value of z^* such that $Q(z) = 0.025$, or equivalently, $\Phi(z) = 0.975$. Looking at the table, we find $z^* = 1.96$. Hence, as long as $\frac{0.01}{\sqrt{\frac{3}{16n}}} > z^*$, we have $Q(\frac{0.01}{\sqrt{\frac{3}{16n}}}) \leq 0.025$. Simplifying the above inequality, we get

$$\frac{0.01\sqrt{16n}}{\sqrt{3}} > 1.96 \Longrightarrow n > (1.96\sqrt{3})^2 \cdot (25)^2 \approx 7203.$$

We see that the estimate obtained from the Central Limit Theorem can give us the required confidence for $M_n \in [\mu - 0.01, \mu + 0.01]$ with a smaller number of samples than the estimate from the Chebyshev Inequality or Hoeffding's inequality.

Example 9.3
Let's ask a different question related to example 9.1: Given that you have collected 1000 samples $X_k, k = 1, \ldots, 1000$, what is the 95% confidence interval around μ for the estimate M_{1000}?

Using the Chebyshev Inequality, we have

$$\mathbb{P}[\{|M_{1000} - 0.25| > \epsilon\}] \leq \frac{\frac{3}{16}}{1000 \cdot \epsilon^2} \leq 0.05$$

This implies

$$\epsilon^2 \geq \frac{6}{1600} \Longrightarrow \epsilon \geq \frac{\sqrt{6}}{40} \approx 0.3873.$$

Using the CLT, we get $Q(\frac{\epsilon}{\sqrt{\frac{3}{16000}}}) = 0.025$, which means that

$$\frac{\epsilon}{\sqrt{\frac{3}{16000}}} = 1.96 \Longrightarrow \epsilon = 1.96\sqrt{3/16000} \approx 0.0268.$$

Using Hoeffding's inequality, we get $e^{-\frac{2000(\epsilon)^2}{1^2}} = 0.025$ which means that $\epsilon^2 = \frac{\ln(40)}{2000} \Longrightarrow \epsilon = 0.0429$.

Example 9.4
We are taking measurements of an unknown distance d, and the measurements are noisy. Hence, we assume that a measurement $X = d + W$, where W is a zero-mean random variable with variance σ^2. Hence, $\mathbb{E}[X] = d$, $\mathsf{Var}[X] = \sigma^2$. We can repeat this measurement n times, resulting in n independent, identically distributed measurements $X_k, k = 1, \ldots, n$. We will estimate d as the sample mean of these measurements, as

$$\hat{d} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Suppose we want to determine how many measurements are needed to obtain 99% confidence interval that the error $|\hat{d} - d| \leq 0.1$? Assuming n is large, so that we use the Central Limit Theorem approximation, so that the random variable $\frac{\sqrt{n}(\hat{d}-d)}{\sigma}$ is approximated by a standard Gaussian random variable with mean 0, variance 1. Using the standard Gaussian table in Appendix C, we determine a value z^* such that $Q(z) = 0.005$, or equivalently, $\Phi(z) = 0.995$. Looking at the table, we find $z^* = 2.575$. This implies that $\mathbb{P}[\{\frac{\sqrt{n}(\hat{d}-d)}{\sigma}| \leq 2.575\}] = 0.99$, or equivalently, $\mathbb{P}[\{|\hat{d} - d| \leq 2.575\frac{\sigma}{\sqrt{n}}\}] = 0.99$.

We want to find n so that the 99% confidence interval is $|\hat{d} - d| \leq 0.1$. Hence, we must select n such that $2.575\frac{\sigma}{\sqrt{n}} \leq 0.1$. This requires $n \geq (25.75)^2\sigma^2$. For $\sigma = 1$, this is approximately 663 samples.

Example 9.5
Suppose we measure the response time X of a service system, and are interested in estimating the mean response time. The 10 measurements we collect are listed in the observation vector $\underline{Y}$ below:

$$\underline{Y} = \begin{bmatrix}41.6 & 41.48 & 42.34 & 41.95 & 41.86 & 42.18 & 41.72 & 42.26 & 41.81 & 42.04\end{bmatrix}^T$$

The sample mean is $M_{10} = 41.924$, which is an approximation of $\mathbb{E}[X]$. Suppose we know $\sigma_X = 0.1$. We want to find a 95% confidence interval for $\mathbb{E}[X]$.

Going to our table for $Q(\cdot)$, we try to find a value T such that $Q(T) = 0.025$. We find that $T \approx 1.96$. Then,

$$\mathbb{P}\left[\{|\mathbb{E}[X] - M_{10}| \leq \frac{1.96(0.1)}{\sqrt{10}} \approx 0.062\right] \approx 0.95.$$

Thus, we say that $\mathbb{E}[M] \in [41.862, 41.986]$ with confidence 95%.

In the previous examples, we assume that we know the variance of X, denoted by σ^2. In many practical situations, we don't know the variance, but have only the observed sample values. We have two approaches for this: one is to use an upper bound on the standard deviation, computed from the properties of the random variable in question. For instance, the variance of a Bernoulli(p) random variable is $p(1-p)$. For any value of p, this number is less than or equal to 0.25, so that the standard deviation is bounded above by 0.5. We can use similar approaches with other types of random variables, provided we have bounds on their parameters.

If the random variable X is bounded with range $R_X \subset [a, b]$, we can use Hoeffding's inequality, which does not require knowledge of the variance, but instead uses knowledge of the bounds on the range of X. Alternatively, we can bound the variance by $\frac{(b-a)^2}{4}$, the largest variance any random variable can have with range $R_X \subset [a, b]$.

A second approach is to use the sample standard deviation as a substitute for the true standard deviation. We illustrate this with examples below.

Example 9.6
We are interested in estimating the probability p that people like bananas. We want a confidence interval of length 0.06 around our estimate, with confidence level 95.5%, corresponding to $T = 2$ standard deviations. How many people do we need to poll, assuming that the opinions of people are independent?

Note that the answer any one person gives is a Bernoulli random variable, which is 1 if they like bananas, and 0 if they don't. We don't allow "I don't know" responses...Thus, if we knew p, the variance in the random variable X corresponding to a response would be $p(1-p)$, which is a number less than 0.25. Let's use this as a bound for the true variance which we don't know. Let the response of person k be X_k, and let $Z_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Then,

$$\mathbb{P}\left[\{|Z_n - p| \leq \frac{2\sqrt{0.25}}{\sqrt{n}}\}\right] \geq 0.955.$$

To get the confidence interval we want, we must have $\frac{2\sqrt{0.25}}{\sqrt{n}} = \frac{1}{\sqrt{n}} = \frac{0.06}{2} = 0.03$. Hence, $\sqrt{n} \approx \frac{100}{3}$, so $n \approx 1112$ persons need to be interviewed. We could reduce this number somewhat by estimating the variance of the specific responses adaptively. By using a bound, we get a conservative number to interview that does not depend on the actual responses.

Note that another bound on the variance is used in Hoeffding's inequality: when the range of X is bounded by $[a, b]$, the variance is bounded by $(b-a)^2/4$.

Example 9.7
Let's return to Example 9.5. Assume we did not know the variance of X. Let M_n denote the mean response time given n observed data. We can estimate the variance using the estimator $V_n = \frac{1}{n-1}\sum_{k=1}^{n}(Y_k - M_n)^2$.

For the data provided in Example 9.5, with $n = 10$ samples, the variance estimate is $V_{10} = 0.081$. Taking the square root yields a sample standard deviation of 0.284.

Now, with only 10 measurements, the 95% confidence interval would be

$$\mathbb{P}\left[\{|\mathbb{E}[X] - M_{10}| \leq \frac{1.96(0.284)}{\sqrt{10}} \approx 0.175\}\right] \approx 0.95.$$

Thus, our confidence interval increases almost by a factor of 3: $\mathbb{E}[X]$ is in the interval [41.75, 42.1] with confidence 95%.

Example 9.8

Here is an example we use in many engineering applications. You are trying to estimate the reliability of a system by using a simulation program that introduces the various random effects that can cause system failures. Note that, in each simulation, the system either fails or not, and hence the outcome of each simulation is a Bernoulli random variable X_k, where $X_k = 1$ indicates success. The reliability we are trying to estimate is $\mathbb{E}[X]$. If we conduct 100 simulations, and the estimated reliability $\hat{p} = M_{100} = 0.95$, and the sample variance V_n is 0.05, what can we say regarding the confidence interval and the level of confidence for this estimate?

Let's look for the 0.955 confidence level interval, corresponding to a threshold of two standard deviations. With $V_n = 0.05$, the sample standard deviation is 0.223. With 100 simulations, the length of the confidence interval is

$$\mathbb{P}[\{|M_{100} - \mathbb{E}[X]| \leq \frac{2 \cdot (0.223)}{\sqrt{100}}\}] \geq 0.955.$$

Thus, our true reliability $\mathbb{E}[X] \in [0.905, 0.995]$ with confidence 95.5%. Note that this is an estimate, because the sample variance V_n was random, and not a bound on the true variance.

What if we increased the number of simulations to 2500? Then our confidence interval tightens significantly, so $\mathbb{E}[X] \in [0.936, 0.964]$ with confidence 95.5%. The important relationship is that the width of the confidence interval is inversely proportional to the square root of the number of simulations.

We conclude this section by referencing some examples illustrating how confidence intervals are used. In 2008, a Gallup survey (https://news.gallup.com/poll/105850/ownership-may-good-wellbeing.aspx) was conducted to determine whether TV ownership was good for wellbeing. People questioned were asked to rate their life on a scale of 0 to 10. Specifically, they were asked: "Please imagine a ladder with 11 steps, numbered zero to 10, where the top represents the best possible life for you, and the bottom represents the worst possible life for you, which step comes closer to the way you feel about your life?" The responses were sorted into those that came from households with TVs, and those that came from households without TVs.

Note that the answers are integers from 0 to 10. Just like we did for Bernoulli replies, we can bound the variance of the responses by the variance of a discrete uniform distribution on $\{0, 1, \ldots, 10\}$, which is 10. Hence, the standard deviation is $\sqrt{10}$. For a population of 810, a 95% confidence level would result in a confidence interval of ± 0.24.

Typical outcomes of this poll are statements such as: "For the European data, one can say with 95% confidence that the true population for wellbeing among those without TVs is between 4.88 and 5.26." This estimate resulted from a sample of 810 persons that did not have TVs in their home. Note that this confidence interval (± 0.19) is narrower than the worst-case interval above, indicating that the Gallup survey used a standard deviation estimate based on the responses that was smaller than the worst-case estimate. Similarly, another statement in the poll was "For those with TVs the 95% confidence interval for well-being is much narrower – between 5.78 and 5.82 – because of the larger sample size." In this case, the poll included 40,267 households with TVs in their home. An increase in the number of samples by a factor of nearly 50 reduced the confidence interval by a factor of close to 7. The ratio is not exactly $\sqrt{n}$ because the estimate of the standard deviation also changed.

Given that 2020 is the year of the U.S. Census, one should note that the U.S. Census Bureau routinely uses confidence levels of 90% in their surveys, which is about 1.645 standard deviations. One survey of the number of people in poverty in 1995 stated a confidence level of 90% for the statistics: "The number of people in poverty in the United States is 35,534,124 to 37,315,094." That means if the Census Bureau repeated the survey using the same techniques, 90 percent of the time the results would fall between 35,534,124 and 37,315,094 people in poverty. The stated figure (35,534,124 to 37,315,094) is the confidence interval. Now you know a little more as to how to interpret such statistical statements that appear in our news reports.

9.3 Sampling Gaussian Random Variables

In the previous sections, we did not assume that the variable X that had n independent, identically distributed samples was Gaussian. For large n, we were able to use properties like the Central Limit Theorem

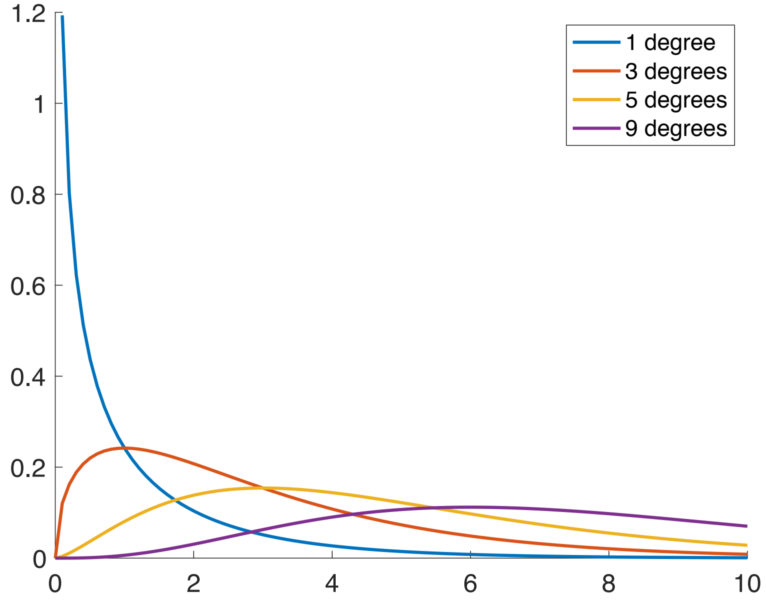

Figure 9.1: PDF of *chi*-squared random variables with different degrees of freedom.

to approximate the distribution of the sample mean as Gaussian, and get confidence intervals for estimates of the sample mean. However, we could not do the same for estimates of the sample variance, or for small number of samples n.

When $X_k, k = 1, \ldots, n$ are Gaussian with mean μ and variance σ^2, the sample mean M_n is Gaussian, and we can use Gaussian properties to get confidence intervals for small values of n. We have

$$\mathbb{P}[\{|M_n - \mu| \geq \epsilon\}] = 2(1 - \Phi(\frac{\epsilon\sqrt{n}}{\sigma})) = 2Q(\frac{\epsilon\sqrt{n}}{\sigma}).$$

What about the sample variance? The estimate of the sample variance is $V_n = \frac{1}{n-1}\sum_{k=1}^n (X_k - M_n)^2$. This random variable is now the sum of squares of random variables. We introduce two new classes of continuous random variables which will be used to analyze properties of the random variance.

Definition 9.1
Let $X_1, \ldots, X_n$ be independent, standard Gaussian random variables with mean 0, variance 1. Define the random variable $Y = X_1^2 + \ldots + X_n^2$. Then, Y is said to be a chi-squared random variable with n degrees of freedom. We write this as $Y \sim \chi^2(n)$.

Figure 9.1 shows the probability density function for Student's t random variables with different degrees of freedom.

We can derive the following properties for $Y \sim \chi^2(n)$:

- $\mathbb{E}[Y] = \sum_{k=1}^n \mathbb{E}[X_k^2] = n$.
- $\mathbb{E}[Y^2] = \sum_{j=1}^n \sum_{k=1}^n \mathbb{E}[X_j^2 X_k^2]$. We can compute each term in the sum as

$$\mathbb{E}[X_j^2 X_k^2] = \begin{cases} \mathbb{E}[X_j^2]\mathbb{E}[X_k]^2 = 1 & j \neq k \\ \mathbb{E}[X_k^4] = 3 & j = k. \end{cases}$$

Thus, $\mathbb{E}[Y^2] = 3n + n^2 - n = 2n + n^2$.

- $\mathsf{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = 2n$.
- Let $U \sim \chi^2(n)$ and $V \sim \chi^2(m)$ be independent random variables. Then, $Y = U + V \sim \chi^2(m+n)$.

Like standard Gaussians, the CDF of chi-squared random variables is tabulated and used to compute probabilities of intervals. An important property of chi-squared random variables is to analyze the statistics of estimates of the sample variance, when the underlying random variables X_k are Gaussian.

Let $X_1, \ldots, X_n$ be independent, identically distributed Gaussian random variables with $X_k \sim \mathcal{N}(\mu, \sigma^2)$. The sample mean and variance are:

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k; \quad V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2.$$

Then, we will show that the random variable $Y = \frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - M_n)^2$ is a $\chi^2(n-1)$ random variable. Moreover, Y and M_n are independent random variables. Note that Y is proportional to the sample variance, as $Y = \frac{n-1}{\sigma^2}V_n$.

Let's first show that Y and M_n are independent random variables. Write $\sigma^2 Y$ as

$$\sigma^2 Y = \sum_{k=1}^{n}(X_k - M_n)^2 = (X_1 - M_n)^2 + \sum_{k=2}^{n}(X_k - M_n)^2 = \Big(\sum_{k=2}^{n}(X_k - M_n)\Big)^2 + \sum_{k=2}^{n}(X_k - M_n)^2$$

where the last equality follows because $\sum_{k=1}^{n}(X_k - M_n) = 0$. We know X_k are i.i.d. and Gaussian. Let's define a linear variable transformation as follows: $W_1 = M_n$; $W_2 = X_2 - M_n$; $W_3 = X_3 - M_n$; $\cdots W_n = X_n - M_n$. This is a linear transformation, so the variables W_k are Gaussian, and zero-mean. Furthermore, the inverse of the transformation is

$$X_2 = W_2 - W_1;\ X_3 = W_3 + W_1; \cdots X_n = W_n + W_1; X_1 = W_1 - W_2 - \cdots - W_n.$$

As a matrix, we write this as

$$\underline{X} = \mathbf{A}\underline{W} = \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \underline{W}.$$

Note that $\det[\mathbf{A}] = n$. Since the X_k are independent, we have

$$f_{\underline{X}}(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\sum_{k=1}^{n}\frac{(x_k-\mu)^2}{2\sigma^2}}.$$

Using the linear transformation, the joint PDF of $\underline{W}$ is Gaussian, and given by

$$f_{\underline{W}}(w_1, \ldots, w_n) = \frac{n}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{(w_1 - \sum_{k=2}^{n} w_k - \mu)^2}{2\sigma^2}} e^{-\sum_{k=2}^{n}\frac{(w_k w_1 - \mu)^2}{2\sigma^2}}.$$

Let's expand and regroup the quadratic in the exponent, as

$$\begin{aligned}(w_1 - \sum_{k=2}^{n} w_k - \mu)^2 + \sum_{k=2}^{n}(w_k - w_1 - \mu)^2 &= w_1^2 - 2w_1\sum_{k=2}^{n}(w_k - \mu) + \Big(\sum_{k=2}^{n} w_k - \mu\Big)^2 + \\ &\quad \sum_{k=2}^{n}(w_k - \mu)^2 + 2w_1\sum_{k=2}^{n}(w_k - \mu) + \sum_{k=2}^{n} w_1^2 \\ &= nw_1^2 + \sum_{k=2}^{n}(w_k - \mu)^2 + \Big(\sum_{k=2}^{n} w_k - \mu\Big)^2\end{aligned}$$

Hence,

$$f_{\underline{W}}(w_1, \ldots, w_n) = \frac{n}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{nw_1^2}{2\sigma^2}} e^{frac \sum_{k=2}^{n}(w_k-\mu)^2 + \left(\sum_{k=2}^{n} w_k - \mu \right)^2 2\sigma^2},$$

which shows that W_1 is independent of $W_2, W_3, \ldots, W_n$.

Observe that $M_n = W_1$, and $Y = \frac{1}{\sigma^2}\left(\sum_{k=2}^{n} W_k^2 + \left(\sum_{k=2}^{n} W_k \right)^2 \right)$. Hence M_n and Y are independent.

To show that Y is a *chi*-squared random variable with $n-1$ degrees of freedom, note the following:

$$U = \sum_{k=1}^{n} \frac{(X_k - \mu)^2}{\sigma^2}$$

is a *chi*-squared random variable with n degrees of freedom. Then,

$$\begin{aligned} U = \sum_{k=1}^{n} \frac{(X_k - M_n + M_n - \mu)^2}{\sigma^2} &= \sum_{k=1}^{n} \frac{(X_k - M_n)^2}{\sigma^2} + 2\sum_{k=1}^{n} \frac{(X_k - M_n)(M_n - \mu)}{\sigma^2} + \sum_{k=1}^{n} \frac{(M_n - \mu)^2}{\sigma^2} \\ &= Y + 2(M_n - \mu)\sum_{k=1}^{n} \frac{(X_k - M_n)}{\sigma^2} + n\frac{(M_n - \mu)^2}{\sigma^2} \\ &= Y + n\frac{(M_n - \mu)^2}{\sigma^2} \end{aligned}$$

where the middle term vanishes because M_n is the sample mean of the X_k. The last term is the square of a standard Gaussian random variable also, because $\mathbb{E}[M_n] = \mu$, $\mathsf{Var}[M_n] = \frac{\sigma^2}{n}$. So, we have $V = Y + Z$, where $V \sim \chi^2(n), Z \sim \chi^2(1)$ and Z is independent of Y. This means that Y is a *chi*-squared random variable with $n-1$ degrees of freedom.

Another standard distribution that is used in statistics is the Student's t-distribution. The CDF of this distribution is also tabulated. Let Z be a standard Gaussian random variable, and let Y be a *chi*-squared distributed random variable with n degrees of freedom, that is independent of Z. Then, the random variable

$$W = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

has a Student's t-distribution with n degrees of freedom, abbreviated as $W \sim T(n)$. Figure 9.2 shows the PDF of a Student's t-distribution with different degrees of freedom, as well as a standard Gaussian PDF. The plots illustrate that the Student's t-distribution approaches a standard Gaussian PDF as the number of degrees of freedom increases.

The following properties of $W \sim T(n)$ are stated without proof:

- For $n > 1$, $\mathbb{E}[W] = 0$. For $n = 1$, $\mathbb{E}[W]$ is undefined.
- For $n > 2$, $\mathsf{Var}[W] = \frac{n}{n-2}$. For $n = 1, 2$, $\mathsf{Var}[W]$ is undefined (infinite).
- For large n, the density of W approaches $\mathcal{N}(0, 1)$.
- The PDF of W is an even function, symmetric about 0.

Why are Students' t-distributions important? Given $X_1, X_2, \ldots, X_n$ i.i.d. Gaussian random variables with mean μ and variance σ^2, and let M_n, V_n denote the sample mean and variance of these variables. Let $\widehat{\sigma} = \sqrt{V_n}$ denote the sample standard deviation. Then,

$$W = \frac{\sqrt{n}(M_n - \mu)}{\widehat{\sigma}} = \frac{\sqrt{n}(M_n - \mu)}{\sqrt{V_n}}$$

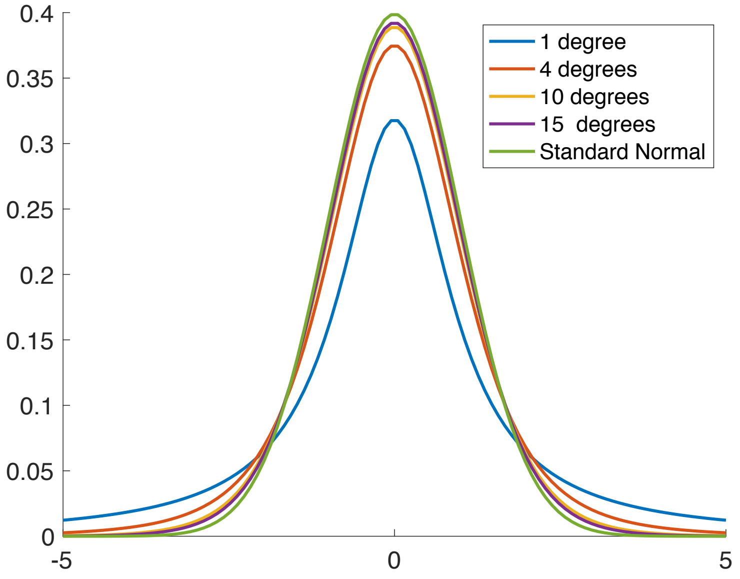

Figure 9.2: PDF of *chi*-squared random variables with different degrees of freedom.

has a Student's t-distribution with $n-1$ degrees of freedom ($W \sim T(n-1)$.)

To see this, note the following:

$$W = \frac{\sqrt{n}(M_n - \mu)}{\sqrt{V_n}} = \frac{\sqrt{n}(M_n - \mu)}{\sigma} \frac{\sigma}{\sqrt{V_n}}.$$

The variable $Z = \frac{\sqrt{n}(M_n - \mu)}{\sigma}$ is a standard Gaussian random variable. The variable $\frac{V_n}{\sigma^2}$ can be written as $\frac{V_n}{\sigma^2} = \frac{1}{n-1}Y$, where Y is a *chi*-squared random variable with $n-1$ degrees of freedom. Hence, the ratio is a Student's t-distribution with $n-1$ degrees of freedom.

We can use this to compute confidence intervals for samples of Gaussian random variables without specifying either the mean or variance of the distribution, as shown in the example below.

Example 9.9
Let's return to the problem of example 9.5, with the additional assumption that response time X of a service system is Gaussian with unknown mean μ and variance σ^2. We collect 10 independent measurements of X, listed in the observation vector $\underline{Y}$ below:

$$\underline{Y} = \begin{bmatrix}41.6 & 41.48 & 42.34 & 41.95 & 41.86 & 42.18 & 41.72 & 42.26 & 41.81 & 42.04\end{bmatrix}^T$$

The sample mean is $M_{10} = 41.924$, which is an approximation of $\mathbb{E}[X]$. The sample variance is 0.0807, and the sample standard deviation $\widehat{\sigma}$ is 0.284.

We want to find a 95% confidence interval for $\mathbb{E}[X]$. We have 10 samples, so $\frac{\sqrt{10}(M_{10}-\mu)}{\sqrt{V_{10}}}\ T(9)$. We use Microsoft Excel or MATLAB to find the value for which the CDF of a $T(9)$ random variable has value 0.975, which is approximately 2.262.

Then,

$$\mathbb{P}\left[\{|\frac{\sqrt{10}(M_{10}-\mu)}{\sqrt{V_{10}}}| \leq 2.262\}\right] = \mathbb{P}\left[\{|M_{10} - \mu| \leq \frac{2.262 \cdot 0.284}{\sqrt{10}} \approx 0.236\}\right] = 0.95.$$

Thus, we say that $\mathbb{E}[M] \in [41.698, 42.160]$ with confidence 95%. The increase in the width of the confidence interval, when compared with the estimate of 9.7, is due to the uncertainty in the estimate of the standard deviation.

We can also get confidence intervals on the sample variance of a normal distribution. The sample variance, based on the sample random variables $X_1, \ldots, X_n$, is

$$V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - M_n)^2,$$

where M_n is the sample mean. We know that the random variable $Y = \frac{(n-1)V_n}{\sigma^2}$ has a *chi*-squared distribution with $n-1$ degrees of freedom. To find a $1-\alpha$ confidence interval, we look at the CDF of $Y \sim \chi(n-1)$ to determine thresholds $t_1, t_2 \geq 0$ such that

$$\mathbb{P}[\{Y \leq t_1\}] = \alpha/2]; \quad \mathbb{P}[\{Y \leq t_2\}] = 1-\alpha/2.$$

This guarantees that

$$\mathbb{P}[\{t_1 \leq Y \leq t_2\}] = \mathbb{P}[\{t_1 \leq \frac{(n-1)\widehat{\sigma^2}}{\sigma^2} \leq t_2\}] = 1-\alpha.$$

We can take inverses to obtain

$$\mathbb{P}[\{\frac{1}{t_1} \geq \frac{\sigma^2}{(n-1)\widehat{\sigma^2}} \geq \frac{1}{t_2}\}] = \mathbb{P}[\{\frac{(n-1)\widehat{\sigma^2}}{t_1} \geq \sigma^2 \geq \frac{(n-1)\widehat{\sigma^2}}{t_2}\}].$$

This gives a $(1-\alpha)$ confidence interval for the true variance σ^2 as $[\frac{(n-1)\widehat{\sigma^2}}{t_2}, \frac{(n-1)\widehat{\sigma^2}}{t_1}]$.

Example 9.10

For the problem of example 9.5, $n = 10$ and the sample variance is 0.0807. To obtain a 95% confidence interval, we compute the thresholds t_1, t_2 for $\alpha = 0.05$ using Microsoft Excel or MATLAB, and obtain $t_1 = 2.700, t_2 = 19.023$. This yields a 95% confidence interval that the true variance $\sigma^2 \in [0.038, 0.269]$. Our sample variance is in this interval, but the interval is large, as n is small.

9.4 Significance Testing based on Sample Statistics

In significance testing, we are interested in determining whether a set of observations show effects that differ significantly from those expected from a nominal model. The nominal model is our null hypothesis H_0, which describes the nominal probability distribution function of the observations. For simplicity, assume Y to be a continuous random variable described by a probability density function $f_{Y|H_0}(y)$. We observe a sample of that random variable, and we are interested in determining whether the sample of the random variable is consistent with the assumed distribution $f_{Y|H_0}(y)$. In contrast to binary hypothesis testing, there is no alternative hypothesis H_1 with a similar probability model for Y. Instead, the alternative is that H_0 is not the correct hypothesis. The question answered by significance testing is whether the observed value of Y is consistent with the hypothesis H_0, or whether the value is inconsistent, so that the hypothesis that Y was generated according to H_0 should be rejected.

The types of error that one makes in significance testing are denoted as Type I and Type II errors. A Type I error occurs when we reject the null hypothesis, declaring that the observed value of Y is inconsistent with the null hypothesis, even though the data was generated according to H_0. This error is a false positive, or a false alarm, using our nomenclature from Chapter **??**. A Type II error occurs when we declare that the observed value is consistent with the null hypothesis, even though it was not generated by a density corresponding to the null hypothesis. This type of error is a false negative, or a missed detection.

To design a test of significance for the null hypothesis, we start with a value of α, called the *level of significance.* We want to design a test such that the probability of false alarm is less than or equal to α. To do this, we select a set $R_0 \subset \Re$ of values such that $\mathbb{P}[\{Y \in R_0|H_0\}] = \alpha$. The significance test declares the value is inconsistent and rejects H_0 if $Y \in R_0$, and fails to reject H_0 if the observed value $Y \notin R_0$.

There are many ways of selecting the set R_0 that satisfy $\mathbb{P}[\{Y \in R_0|H_0\}] = \alpha$. The two most common ways are one-sided tests and two-sided tests. For a typical **one-sided test**, let $F_{Y|H_0}(y)$ denote the cumulative distribution function of the observation Y conditioned on the null hypotheses. We select a value t_α such that $F_{Y|H_0}(t_\alpha) = 1 - \alpha$, and select $R_0 = \{y > t_\alpha\}$. One-sided tests are appropriate for evaluating when the observed value of Y is too large to be consistent with the null hypothesis. In this case, α is the probability of a Type I error, namely, a false alarm. Although these tests appear to focus only on rejection of H_0 for values that are too high, we can extend this to values that are too low by considering the observation $-Y$ instead of Y.

A **two-sided test** is designed to test whether the observed random variable is either too high or too low to be consistent with the null hypothesis. Define t_l, t_h as follows:

$$F_{Y|H_0}(t_l) = \frac{\alpha}{2}; \qquad F_{Y|H_0}(t_h) = 1 - \frac{\alpha}{2}.$$

Define the reject set as $R_0 = \{y : y < t_l \text{ or } y > t_h\}$. Then, $\mathbb{P}[\{Y \in R_0\}|H_0] = \alpha$.

Given an observation $Y = y_0$, the p**-value** of y_0 is defined as the probability, under the null hypothesis, that you will observe a value as extreme or more extreme that y_0. For a one-sided test, the p-value is defined as $1 - F_{Y|H_0}(y_0)$. For a two-sided test, the definition is more nuanced, and depends on the specific nature of the CDF $F_{Y|H_0}(y_0)$; one definition is $2\min\left(F_{Y|H_0}(y_0), 1 - F_{Y|H_0}(y_0)\right)$. If the p-value is smaller than α, the null hypothesis is rejected. This is a different way of implementing the hypothesis test that does not require computing the inverse of the CDF $F_{Y|H_0}(y)$ to obtain a threshold.

Example 9.11

Our probability model for how late the Green Line is in arriving at its scheduled stop on St. Mary's street is that Y, the delay time in minutes, is an exponential random variable with rate parameter $\lambda = 0.5$, so that the expected delay time is 2 minutes. This is our null hypothesis. We are going to measure the observed delay time Y, and we want to design a significance test for hypothesis H_0 at a confidence level of $1 - \alpha = 0.95$, looking for evidence that the null hypothesis is inconsistent with the observed data if the measured delay time is too large.

The appropriate test is a one-sided test, as we are looking for delays that are too large to be consistent with the null hypothesis. Using the properties of exponential random variables, the probability distribution function of Y is

$$F_{Y|H_0}(y) = \begin{cases} 0 & y \le 0 \\ 1 - e^{-0.5y} & y > 0. \end{cases}$$

We want to define the reject set $R_0 = \{y > t_{0.05}\}$ for some threshold value $t_{0.05}$ that gives a confidence level of 0.95. Hence, we want $F_{Y|H_0}(t_{0.05}) = 1 - \alpha = 0.95$. Thus,

$$e^{-0.5t_{0.05}} = 0.05 \Rightarrow t_{0.05} = 5.9915.$$

Hence, our test of significance is $Y > 5.9915$, defining the region of measurements for which the null hypothesis is rejected.

For any measured value $Y = y$, its p-value is computed as $1 - F_{Y|H_0}(y) = e^{-0.5y}$. If the p-value of the measurement is less than the desired significance level $\alpha = 0.05$, the null hypothesis is rejected.

As the above example illustrates, the key to designing a test of significance is to identify the conditional probability distribution of the test statistic Y under the null hypothesis H_0. Using this conditional PDF $F_{Y|H_0}(y)$, we can compute thresholds for the appropriate significance level, and determine the p-values of measured test values $Y = y$.

9.4.1 The One Sample Z-Test

Consider the null hypothesis that the random variable X is a Gaussian random variable with known mean μ and variance σ^2. As an observation, we collect n independent observations of X, and want to accept or reject the hypothesis that the measurements were generated according to the null hypothesis. The one-sample Z test consists of determining whether the batch of n measurements is consistent with null hypothesis.

To make the decision, we use the sample mean of the observations as the statistic Y for the test. Thus, we test the hypothesis that the sample mean of the n observations $X_1, \ldots, X_n$ is consistent with the assumption that the observations were generated according to the null hypothesis.

This type of hypothesis test is known as a one-sample Z-test. Under the null hypothesis H_0, the sample mean $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is a Gaussian random variable with mean μ and variance $\frac{\sigma^2}{n}$. We want to design a test with a level of significance α that the sample mean is different from μ. The appropriate test is a two-sided test, as the sample mean can be either too large or too small.

The random variable $Z = \frac{\sqrt{n}(M_n - \mu)}{\sigma}$, referred to as the Z-statistic, is known to be a standard Gaussian random variable with mean 0 and variance 1. Given the value $M_n = \hat{\mu}_n$, the resulting Z-statistic is $z = \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma}$. The hypothesis test can be expressed in terms of the Z-statistic, as we want to find a threshold $T_{\alpha/2}$ so that $\mathbb{P}[\{|Z| > T_{\alpha/2}\}] = \alpha$, which is the same problem as finding a $1-\alpha$ confidence interval for the estimate M_n. The threshold is computed the same way: we find the value $T_{\alpha/2}$ so that $\Phi(-T_{\alpha/2}) = Q(T_{\alpha/2}) = \alpha/2$, or equivalently $\Phi(T_{\alpha/2}) = 1 - \alpha/2$. For instance, if $\alpha = 0.05$, then $T_{\alpha/2} = 1.96$. Then, if $|z| > T_{\alpha/2}$, the observations do not support the null hypothesis at a level of significance α.

An equivalent way of implementing a Z-test is to compute the p-value of the sample mean M_n, or equivalently, the Z- statistic. The p-value of a measurement is the probability of getting a measurement value that is more extreme than the current measurement. With a two-sided test and a Gaussian null hypothesis, the p-value of $Z = z$ is $\Phi(-z) + (1 - \Phi(z)) = 2\Phi(-|z|)$. If the p-value is less than the level of significance α, then the evidence indicates that the null hypothesis can be rejected at that level of significance. The advantage of this approach is that we don't have to compute the inverse of the standard Gaussian CDF Φ to compute a threshold.

Example 9.12

Assume that a probabilistic model for the weight of a randomly selected male person in the US is a Gaussian random variable measured in pounds, with mean 195, and standard deviation 30. We believe that Canadians have the same weight distribution, so we designed an experiment to weigh 100 randomly selected Canadian males, and compute their average weight, denoted as W_{ave}. Design a statistical test with significance level 0.01 to determine whether the measured W_{ave} supports the null hypothesis that the weight of Canadian males has the same probability model as the weight of US males.

The measured random variable is W_{ave}, which is the average of 100 independent samples of Canadian male weights. To answer the question, we need to compute the probability distribution of W_{ave} under the null hypothesis, given that

$$W_{ave} = \frac{1}{100}(W_1 + W_2 + \ldots + W_{100}).$$

Under the null hypothesis, the W_i's are independent Gaussian random variables, with mean 195 and standard deviation 30.

The Z statistic for this problem is $Z = \frac{10(W_{ave}-195}{30}$. We want to define a two-sided test to accept or reject the null hypothesis with significance level 0.01, we are looking for a threshold $T_{0.005}$ such that, if $|Z| > T_{0.005}$, we will reject the hypothesis with significance level 0.01.

Thus, we need to select $T_{0.005}$ such that $Q(T_0.005) = 1 - \Phi(T_{0.005}) = 0.005$; this implies $T_0.005 = 2.576$.

We reject the null hypothesis with significance level 0.01 whenever $|Z| > 2.576$, or equivalently the average weight difference $|W_{ave} - 195| > 7.728$ pounds. Note the effect of selecting a sample size of 100 persons had in reducing the standard deviation of the test statistic W_{ave}. If we had weighed 9,000 Canadian males, the threshold would be much smaller, as the standard deviation of the sample mean would be 1, and now a smaller difference in average weight would be significant.

For this problem, we can compute the p-value of a measured $W_{ave} = W$, by computing $\mathbb{P}[\{|W_{ave} - 195| > |195 - W|\}|H_0]$ as the probability that the null hypothesis would yield a measurement more extreme than W. This yields a p-value for W of $2Q(\frac{|W-195|}{3})$.

Example 9.13

The lifetime of a certain cell type has been determined to be distributed according to a Gaussian distribution with mean 1570 hours and a standard deviation of 120 hours. You perform an experiment and measure the lifetime of 100 cells, and compute a sample mean lifetime of 1600 hours. Is the sample mean you measure significantly different from the population mean at a significance level of 0.05?

The Z statistic is $z = \frac{\sqrt{100}(1600-1570)}{120} = 2.5$. The p-value of z can be computed from Appendix C as $2\Phi(-2.5) = 0.0124$. Since the p-value is less than the significance level, we can reject the null hypothesis that the experiment lifetimes were sampled from a $\mathcal{N}(1570, 14400)$ distribution.

When the underlying null hypothesis is not a normal distribution, we can still use Z-tests provided that the number of samples n is sufficiently large (e.g. greater than 30). This is because the Z statistic will have an approximately Gaussian distribution, according to the Central Limit Theorem in Chapter **??**.

9.4.2 The One Sample T-Test

In the One Sample Z-Test, the null hypothesis assumed that both the mean and the standard deviation were known. In many applications, these parameters are rarely known. We discuss a different test, where we know the mean but not the standard deviation of the null hypothesis.

As in the Z-Test, we collect n observations of a random variable X, which is assumed under the null hypothesis H_0 to be Gaussian, with known mean μ, but with unknown variance σ^2. We would like to test the hypothesis that the sample mean $M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ is consistent with the null hypothesis at a level of significance α.

Note that we don't have a well-specified PDF for the sample mean. We know that $\mathbb{E}[M_n|H_0] = \mu$, and $f_{M_n|H_0}(x)$ is Gaussian, but we don't know its variance. Let's compute the sample variance V_n, and the sample standard deviation $\widehat{\sigma} = \sqrt{V_n}$ as described earlier. Then, transform M_n to a new random variable known as the T-statistic, as

$$T = \frac{\sqrt{n}(M_n - \mu)}{\widehat{\sigma}}.$$

If the null hypothesis is true, T is distributed according to a Student's t-distribution with $n-1$ degrees of freedom, as shown in Section 9.3. Thus, we know $f_{T|H_0}(t)$, and can perform a test of the null hypothesis with level of significance α.

The Student's t-distribution PDF is symmetric about 0. We use a two-sided test, so we compute threshold $t_{\alpha/2}$ so that $F_{T|H_0}(-t_{\alpha_2}) = \alpha/2$. Then, our decision rule is: if $|T| > t_{\alpha_2}$, we reject hypothesis H_0 at a level of significance α. Otherwise, we don't reject hypothesis H_0.

Equivalently, we compute the p-value of the computed T-statistic $T = t$, as $p = 2 * F(-|t|)$. If $p < \alpha$, we can reject hypothesis H_0 at a level of significance α.

Example 9.14
Consider the problem of example 9.13, except that we don't know the true standard deviation σ^2 of the lifetime of the cells. You perform an experiment and measure the lifetime of 100 cells, and compute a sample mean lifetime of 1600 hours and a sample standard deviation of 120 hours. Is the sample mean you measure significantly different from the population mean at a significance level of 0.05?

Compute the T-statistic:

$$t = \frac{\sqrt{n}(M_n - 1570)}{\widehat{\sigma}} = \frac{10(1600-1570)}{120} = 2.5.$$

The distribution of the T statistic is a Student's t-distribution with 99 degrees of freedom. Looking up the p-value for 2.5 in either MATLAB or Microsoft Excel, it is $2 \cdot 0.00703 = 0.01406$, which is less than 0.05, so the results support rejecting hypothesis H_0 with a level of significance 0.05.

Similarly, the threshold $t_{0.025}$ is 1.984. Since 2.5 is greater than that threshold, the results support rejecting hypothesis H_0.

Suppose we approximated the T- statistic distribution by a standard Gaussian distribution. What would be the corresponding threshold $t_{0.025}$? We have computed this to be 1.96. We see that the threshold using the correct distribution is slightly larger.

9.4.3 Two Samples T- and Z-tests

In one sample tests, we want to evaluate the null hypothesis that a collection of observations is consistent with a prior probability model. In two sample tests, we are interested in evaluating the null hypothesis that two sets of observations are consistent with a common probability model. We begin with the two-sample Z-tests.

Assume we have two Gaussian random variables X, Y, where $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Assume we collect a set of n_1 independent samples $X_1, \ldots, X_{n_1}$ of X, and n_2 independent samples $Y_1, \ldots, Y_{n_2}$ of Y. We want to test the null hypothesis that $\mu_1 = \mu_2$ with a level of significance α.

The sample mean of the first set, $M_{n_1}^{(1)} = \frac{1}{n_1}\sum_{k=1}^{n_1} X_k$, is a Gaussian random variable with mean μ_1 and variance $\frac{\sigma_1^2}{n_1}$. Similarly, the sample mean of the second set, $M_{n_2}^{(2)} = \frac{1}{n_2}\sum_{k=1}^{n_2} Y_k$ is a Gaussian random variable with mean μ_2 and variance $\frac{\sigma_2^2}{n_2}$. Random variables $M_{n_1}^{(1)}, M_{n_2}^{(2)}$ are independent.

Under hypothesis H_0, the difference $M_{n_1}^{(1)} - M_{n_2}^{(2)}$ is a Gaussian random variable with mean 0 and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, as it is the difference of two independent Gaussian random variables. We define the Z-statistic as

$$Z = \frac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Under H_0, Z is a Gaussian random variable with mean 0, variance 1. To evaluate H_0 with a level of significance α, we perform the same test as before: Compute the test statistic $Z = z$ based on the data. Then, compute threshold $t_{\alpha/2}$ such that $\Phi(-t_{\alpha/2}) = \alpha/2$, and determine whether $|z| > t_{\alpha/2}$. If it is, reject the null hypothesis H_0 with level of significance α. Equivalent, compute the p-value $p = 2\Phi(-|t|)$ and reject the null hypothesis with significance level α if $p < \alpha$.

Note that we don't need to know the values of $\mu_1 = \mu_2$ to conduct this Z-test. However, we do need to know the standard deviations of the two sets σ_1 and σ_2.

What if the variances σ_1^2, σ_2^2 were not known? We can use a simple generalization of the one-sample T-test when the unknown variances are assumed to be the same. We know $\frac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$ is a standard Gaussian random variable. We also know that $(n_1 - 1)\frac{V_{n_1}^{(1)}}{\sigma^2} + (n_2 - 1)\frac{V_{n_2}^{(2)}}{\sigma^2}$ is a *chi*-squared random variable with $n_1 + n_2 - 2$ degrees of freedom. Then, the T-statistic can be defined as

$$T = \frac{M_{n_1}^{(1)} - M_{n_2}^{(2)}}{\widehat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\widehat{\sigma} = \sqrt{\frac{(n_1 - 1)V_{n_1}^{(1)} + (n_2 - 1)V_{n_2}^{(2)}}{n_1 + n_2 - 2}}$$

is the pooled variance.

The T- statistic has a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom, and can now be used to accept or reject the null hypothesis with a desired level of significance.

When the variances are unequal and unknown, one can derive a more complex test with approximate numbers of degrees of freedom, known as Welch's t-test. This results in T-statistics that have fractional degrees of freedom. The details can be found in statistics books or in Wikipedia.

Example 9.15

To investigate the effect of a new hay fever drug on driving skills, a researcher studies 24 individuals with hay fever: 12 who have been taking the drug and 12 who have not. All participants then entered a simulator and were given a driving test which assigned a score to each driver as summarized in the table below:

Control	23	15	16	25	20	17	18	14	12	19	21	22
Drug	16	21	16	11	24	21	18	15	19	22	13	24

We want to test the null hypothesis that the drug has no adverse effects in decreasing the average score of the drivers with a lever of significance 0.05. We compute the sample mean and variance for the two groups as $M^{(1)} = 18.5, M^{(2)} = 18.33$, $V^{(1)} = 15.18, V^{(2)} = 17.88$. We assume the variances are the same, since the sampled variances are similar, and compute the pooled variance as $\widehat{\sigma}^2 = 16.53$. Given the mean values, the resulting pooled variance, and the number of samples n_1, n_2, the value of the T-statistic is 0.1004. The one-sided p-value of this T-statistic with 22 degrees of freedom is 0.46, which is much higher than the desired level of significance of 0.05. Thus, we fail to reject the null hypothesis and are 95% confident that any difference between the two groups is due to chance variations.

The two-sample T-tests and Z-tests depend on the assumption that the distribution of the underlying random variables from which the samples are generated is Gaussian. When that assumption is violated, we can still apply the T-tests and Z-tests as appropriate when the number of samples in each group n_1, n_2 are sufficiently large (greater than 30) so that the Central Limit Theorem allows us to use Gaussian distribution approximations for the sample means $M_{n_1}^{(1)}, M_{n_2}^{(2)}$.